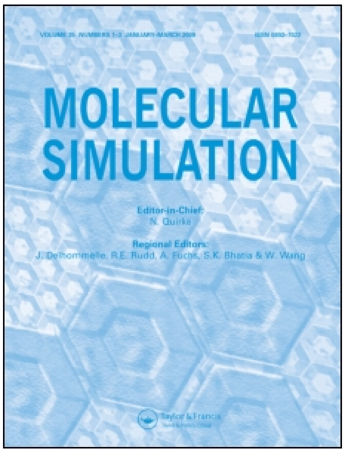


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Pharmacophore mapping, molecular docking and QSAR studies of structurally diverse compounds as CYP2B6 inhibitors

Partha Pratim Roy[a]; Kunal Roy[a]
[a] Department of Pharmaceutical Technology, Division of Medicinal and Pharmaceutical Chemistry, Drug Theoretics and Cheminformatics Laboratory, Jadavpur University, Kolkata, India

Online publication date: 15 October 2010

To cite this Article Roy, Partha Pratim and Roy, Kunal(2010) 'Pharmacophore mapping, molecular docking and QSAR studies of structurally diverse compounds as CYP2B6 inhibitors', Molecular Simulation, 36: 11, 887 — 905

To link to this Article: DOI: 10.1080/08927022.2010.492834

URL: http://dx.doi.org/10.1080/08927022.2010.492834

Molecular Simulation
Vol. 36, No. 11, September 2010, 887–905

Taylor & Francis
Taylor & Francis Group

Pharmacophore mapping, molecular docking and QSAR studies of structurally diverse compounds as CYP2B6 inhibitors

Partha Pratim Roy and Kunal Roy*

Department of Pharmaceutical Technology, Division of Medicinal and Pharmaceutical Chemistry, Drug Theoretics and Cheminformatics Laboratory, Jadavpur University, Kolkata 700 032, India

(*Received 2 January 2010; final version received 28 April 2010*)

Pharmacophore mapping, molecular docking and quantitative structure–activity relationship (QSAR) studies were carried out for a structurally diverse set of 48 compounds as CYP2B6 inhibitors. The generated best pharmacophore hypotheses from the three methods of conformer generation (FAST, BEST and conformer algorithm based on energy screening and recursive buildup) indicate the importance of two features, namely, hydrogen bond acceptor [electron-rich centre] and ring aromaticity. The distance between the two centres of the important features for ideal inhibitors varied from 5.82 to 6.03 Å. The chemometric tools used for the QSAR analysis were genetic function approximation (GFA) and genetic partial least squares. The developed QSAR models indicate the importance of an electron-rich centre, size of molecule, impact of branching and ring system and distribution of charges in the molecular surface. The docking study confirms the importance of an electron-rich centre for binding with the iron atom of the cytochrome enzyme. A GFA model with spline option was found to be the best model based on internal validation as well as the $r^2_{\mathrm{m(overall)}}$ criterion ($Q^2 = 0.772$, $r^2_{\mathrm{m(overall)}} = 0.774$). According to the external prediction statistics ($R^2_{\mathrm{pred}} = 0.876$), another GFA-derived model with spline option outperforms the remaining models.

Keywords: QSAR; pharmacophore; cytochrome 2B6; GFA; G/PLS

1. Introduction

Inhibition of drug metabolising enzymes is one of the major concerns in both clinical practice and drug development. Cytochrome P450s (CYPs) are recognised as the predominant phase I enzymes and probably the most important catalysts among all drug-metabolising enzymes. Human CYP genes are arranged into 18 families and 42 subfamilies, consisting of 59 active genes [1–3]. Human CYP 2B6 is recognised to be of great importance in drug metabolism with notable inter-individual variations in expression and activity [4,5]. Initial studies reported that 2B6 levels were only 0.2% of the total P450 content in human liver microsomes [6,7]. However, other laboratories reported a greater frequency of detection and a higher percentage of 2B6 (2–10%) relative to total P450 content using improved immuno-quantitation techniques [8–12]. Recent studies indicate expression of CYP2B6 in human brain with higher level of expression being observed for smokers, and alcoholics [13].

CYP2B6 accounts for the metabolism of wide structurally diverse chemicals [14] and is particularly susceptible to both induction and inhibition [15–17]. CYP2B6 can metabolise approximately 8% of clinically used drugs ($n > 60$), including cyclophosphamide, ifosfamide, tamoxifen, ketamine, artemisinin, nevirapine, efavirenz, bupropion, sibutramine and propofol [18–22]. CYP2B6 is one of the CYP enzymes which bioactivates several procarcinogens and toxicants [22]. It was reported that substrates of CYP2B6 also have affinity towards other CYP isoforms such as CYP3A4, CYP2C9 and CYP2C19 [23]. The significance of CYP2B6 in drug metabolism or detoxification has been firmly established through a series of studies within the past decade because of high substrate specificity and cross regulation with other enzymes and hepatic transporters [23]. It was also reported in the literature that CYP2B is capable of metabolising 25–30% of known clinical drug substrates for CYP3A4 [14,24,25]. CYP2B6 is expressed in hepatic and extra hepatic tissues and it has recently been suggested as a prognostic factor for prostate cancer which may therefore be clinically relevant [26].

The crucial characteristics of CYP2B6 substrates for necessary interactions with the enzyme are the presence of hydrophobic (HYD) features and hydrogen bond acceptor (HBA) groups [27]. Additional quantitative structure–activity relationship (QSAR) techniques have been applied to the substrates [28,29]. The application of QSAR techniques to CYP2B6 inhibitors has been limited [30]. The applied comparative molecular field analysis study to CYP2B6 inhibitors [30] indicates optimal steric, HYD and HBA features for ideal CYP2B6 inhibitors. Although CYP inhibitors have clinical significance in metabolism-mediated

*Corresponding author. Email: kunalroy_in@yahoo.com

ISSN 0892-7022 print/ISSN 1029-0435 online
© 2010 Taylor & Francis
DOI: 10.1080/08927022.2010.492834
http://www.informaworld.com

drug–drug interactions in general, identification of selective and potent chemical inhibitors is extremely important for delineating the specific roles of particular CYPs in metabolism–detoxification of therapeutics and environmental toxicants. Till date no pharmacophoric models of CYP2B6 inhibitors has been reported. Identification of important molecular features using ligand-based approach is required for the design of CYP2B6 inhibitors. In this work, we have performed pharmacophore and 3D QSAR studies to extract the pharmacophoric features as well as other related properties of interest for ideal CYP2B6 inhibitors. The crystal structure of human CYP450 2B6 genetic variant in complex with the inhibitor 4-(4-chlorophenyl)imidazole has been recently published [31]. We have also attempted molecular docking study for the CYP2B6 inhibitors [30] to explore interactions between the enzyme and the ligands at the molecular level.

2. Method and materials

2.1 The data-set

The CYP2B6 inhibitory activity (Tables 1 and 2) of 48 diverse compounds (Figure 1) reported in the literature [30] has been used as the model data-set for the present study. The biological activity data were represented as IC_{50} (μM) [30]. Here, the activity range of the compounds was of about 5 log units which allowed us to generate meaningful activity-based pharmacophore models. For estimation (prediction) of activity from pharmacophores, the activity values were classified as follows: (i) compounds with IC_{50} (μM) <1 are highly active (represented as $+++$), (ii) compounds with $10 < IC_{50}$ (μM) ≤ 1 are moderately active (represented as $++$), (iii) compounds with $500 < IC_{50}$ (μM) ≤ 10 are marginally active (represented as $+$) and compounds with IC_{50} (μM) $\geqq 500$ are inactive (represented as $-$). The basis of the classification was to find lead compounds and to further modify the lead compounds to active ones. Here, we have developed 3D pharmacophore models for CYP2B6 inhibitors [30] using the Discovery Studio 2.1 software [32]. The activity data (IC_{50} (μM)) were also converted to logarithmic scale [pIC_{50}(M)] and then used for subsequent QSAR analyses as the response variable.

2.2 Development of 3D pharmacophore model

3D pharmacophore model is a ligand-based approach that provides a unique tool for drug design [33]. A 3D pharmacophore is a collection of chemical features in space that are required for a desired biological activity. These may include hydrophobic (HYD) groups, charged/ionisable groups, hydrogen bond donors/acceptors and other features properly assembled in 3D space to reflect structural requirements. An interesting application of pharmacophore-based approaches is that the experimentally determined activity of a set of compounds can be used to drive the generation of pharmacophores, which, once validated, can be used to quantitatively predict the activity of new compounds. Therefore, this approach constitutes a powerful and fast tool to estimate the biological activity of new potential ligands in 3D databases of compounds [34–37].

2.2.1 Training set selection

The selection of training set serves as a crucial aspect in the process of pharmacophore hypothesis generation. On the basis of assumption that the most active compounds share all or most of the required features for binding with the active site, only the active molecules were included in the training set. As the inactive compounds may experience steric hindrance and other disfavoured interactions, these compounds were avoided during pharmacophore generation. The whole data-set was divided in a training set of 22 compounds (approximately 45%) and a test set of 26 compounds. Table 1 shows the compounds selected as the members of the test set.

2.2.2 Diverse conformation generation

Before starting the pharmacophore generation process, conformational analysis of the molecules was performed using the poling algorithm [38]. The poling algorithm eliminates much of the redundancy in conformation generation and improves the coverage of conformational space. The number of conformers generated for each compound was limited to a maximum of 255 with an energy range of 20 kcal/mol. In the present work, conformers were generated using BEST, FAST and CAESAR methods of conformer generation. All the three conformation generation algorithms use poling. The BEST method provides a complete coverage of conformational space by optimising the conformations in both torsional and cartesian space, whereas the FAST generation searches conformations only in the torsion space and takes less time. CAESAR is a new conformation generation method for very fast conformation search [32].

In case of FAST conformation generation method, algorithms were selected depending upon the size of the molecules. When the molecule was too large (number of rotatable bonds more than 30), only one conformation was generated for each possible combination of stereocentres. A quasi-exhaustive systematic search was used to generate conformations for small molecules. The conformational space was composed of discretised rotations about bonds. The conformational space was systematically searched and conformations that had excessive van der Waals clashes were removed. For medium size molecules, a search method with pooling was used. The molecule was split into pieces, a systematic search was performed on

Table 1. Observed and estimated CYP2B6 inhibitory activity of training and test set molecules.

					Activity class[a]			
		IC_{50} (Estimate) (μM)[b]				Estimate		
Sl. No.	IC_{50} (Obs.) (μM) [30]	FAST	BEST	CAESAR	Observed	FAST	BEST	CAESAR
Training set								
1	0.3	0.489	0.429	0.851	+++	+++	+++	+++
3	0.4	0.488	0.429	0.851	+++	+++	+++	+++
4	1.5	2.55	5.223	0.886	++	++	++	+++
5	4.4	47.863	44.668	50.119	++	+	+	+
6	5	3.893	2.543	2.379	++	++	++	++
8	6.8	47.863	44.668	50.119	++	+	+	+
11	22	47.863	44.669	50.119	+	+	+	+
12	35	47.863	44.669	50.119	+	+	+	+
14	37	14.42	10.964	6.498	+	+	+	++
15	39	47.863	44.668	50.119	+	+	+	+
16	45	47.863	44.668	50.119	+	+	+	+
17	50	19.221	37.082	22.27	+	+	+	+
18	67	16.243	24.43	16.702	+	+	+	+
20	96	47.863	44.668	50.119	+	+	+	+
23	150	47.864	44.668	50.119	+	+	+	+
24	310	47.863	44.668	50.119	+	+	+	+
28	420	47.863	44.668	50.119	+	+	+	+
31	480	47.863	44.668	50.119	+	+	+	+
42	3.89	47.863	44.668	50.119	++	+	+	+
43	10	47.863	44.668	50.122	+	+	+	+
45	28.18	47.863	44.668	50.119	+	+	+	+
46	37.15	47.863	44.668	50.119	+	+	+	+
Test set								
2	0.4	0.488	0.42	0.848	+++	+++	+++	+++
7	5.2	2.244	2.5	0.89	++	++	++	++
9	9.1	17.769	35.215	26.24	++	+	+	+
10	13	14.861	31.204	25.802	+	+	+	+
13	37	16.1	36.965	21.759	+	+	+	+
19	95	14.624	15.059	4.205	+	+	+	+
21	140	1256.97	#	#	+	−	−	−
22	150	17.736	35.021	27.236	+	+	+	+
25	350	7.32	13.387	11.223	+	++	+	+
26	370	19.074	38.417	21.654	+	+	+	+
27	390	17.754	35.121	27.33	+	+	+	+
29	420	1263.63	#	#	+	−	−	−
30	440	19.106	38.465	21.577	+	+	+	+
32	630	2.872	2.894	3.329	−	++	++	++
33	670	726.935	#	#	−	−	−	−
34	820	#	#	#	−	−	−	−
35	1100	0.9	0.855	1.179	−	+++	+++	+++
36	1100	7.543	14.016	11.729	−	++	+	+
37	1600	7.701	14.437	12.144	−	++	+	+
38	2800	#	#	#	−	−	−	−
39	4700	#	#	#	−	−	−	−
40	6400	#	#	#	−	−	−	−
41	28000	#	#	#	−	−	−	−
44	19.95	16.096	36.691	31.652	+	+	+	+
47	109.65	19.263	38.828	26.506	+	+	+	+
48	1412.54	15.687	35.73	31.854	−	+	+	+

Note: # = compounds not mapped.[a] Class threshold of $IC_{50} < 500$ μM has been used for active compounds. Compounds have been classified as highly active (IC_{50} (μM) < 1) compounds (represented as +++), moderately active ($10 < IC_{50}$ (μM) ≤ 1) compounds (represented as ++), marginally active ($500 < IC_{50}$ (μM) ≤ 10) compounds (represented as +), inactive (IC_{50} (μM) $\geqq 500$) compounds (represented as −). [b] Calculated on the basis of best hypothesis (hypothesis **3** in each case) for FAST, BEST and CAESAR methods of conformation search.

Table 2. Observed and QSAR model derived CYP2B6 inhibitory activity.

Sl	Obs[a] pIC_{50}(M)	Cal[b]	Cal[c]	Cal[d]
Training set				
1	6.523	6.097	6.201	5.183
2	6.398	5.864	6.128	5.115
3	6.398	6.420	6.185	5.882
5	5.357	5.228	4.839	5.220
6	5.301	5.144	4.839	5.194
7	5.284	5.937	6.101	5.244
8	5.167	5.502	4.839	5.189
9	5.041	3.947	4.126	4.034
10	4.886	4.803	4.839	5.211
11	4.658	4.463	4.449	4.954
13	4.432	3.600	3.637	3.867
15	4.409	4.944	4.839	5.211
16	4.347	4.148	4.181	3.950
17	4.301	3.515	3.803	3.998
18	4.174	3.947	4.169	4.025
19	4.022	5.273	4.839	5.223
20	4.018	3.783	4.386	3.674
21	3.854	3.470	3.780	3.989
22	3.824	3.510	4.181	4.044
24	3.509	3.363	3.311	3.484
25	3.456	3.174	3.738	3.466
27	3.409	4.016	3.228	3.576
30	3.357	3.383	3.408	4.057
31	3.319	3.545	3.780	3.965
32	3.201	3.792	3.904	3.710
33	3.174	3.201	2.871	2.789
34	3.086	3.586	3.369	2.838
35	2.959	3.810	4.205	4.034
36	2.959	3.252	3.347	3.001
37	2.796	3.392	2.840	2.835
38	2.620	2.814	2.813	2.670
39	2.328	2.363	2.782	2.121
41	1.553	2.364	1.460	2.110
43	5.000	4.391	4.839	5.620
44	4.700	3.896	3.650	4.037
47	4.550	4.432	4.437	4.849
Test set				
4	5.824	5.996	6.160	5.925
12	4.456	4.499	4.449	4.744
14	4.432	4.374	4.613	4.601
23	3.824	4.070	4.100	4.227
26	3.432	3.145	2.811	3.224
28	3.377	3.762	3.352	3.760
29	3.377	3.469	3.780	3.948
40	2.194	2.373	2.384	2.136
42	5.410	5.187	4.839	4.944
45	4.550	4.447	4.839	5.668
46	4.430	3.779	3.715	3.364
48	2.850	3.872	3.398	3.354

[a] Observed CYP2B6 inhibitory activities (ref. [30]). [b] Calculated from Equation (6). [c] Calculated from Equation (7). [d] Calculated from Equation (8).

each piece and the pieces were randomly reconnected. Each new conformer was quickly optimised in the torsion space with pooling penalty as to maintain conformational diversity [32].

In general, BEST conformational search method provides the best quality conformation generation with improved conformational coverage relative to the FAST method. It performs more rigorous energy minimisation in both torsional and Cartesian space and uses pooling. The routinely used steps in the BEST method are (i) Conjugate–gradient minimisation in torsion space, (ii) Conjugate–gradient minimisation in Cartesian space and (iii) Quasi-Newton minimisation in Cartesian space. Both (FAST, BEST) use a version of the CHARMm force field for energy calculations and a poling mechanism for forcing the search into unexplored regions of conformer space [32,38,39].

The new algorithm, termed CAESAR, is another method for conformational search. This approach was combined with consideration of local rotational symmetry so that conformer duplicates due to topological symmetry in the systematic search could be efficiently eliminated. In the CAESAR algorithm, the geometry optimisation was replaced by energy pruning on fine torsion grids. The algorithm involves several steps: the first step is recursive partitioning of a molecule tree into the smallest units. Ring structures (such a cyclohexane) or rigid structures (such as $—CH_2—$ and $> C{=}C <$) are considered as the smallest units. The next step is conformation generation from tree node conformation initialisation. The conformations of whole molecule are built up recursively from the smallest fragment. At each level of recursive assembling, the local rotational symmetry check and energy pruning are performed for elimination of high-energy conformations at the earliest possible stages [32,40].

2.2.3 Generation of 3D pharmacophore

For the present work, pharmacophore models were developed using the *HypoGen* module implemented in Discovery Studio 2.1 [32] with the conformers generated for the molecules in the training set ($n = 22$). Predictive pharmacophores were generated in three phases, viz. a constructive, a subtractive and an optimisation phase. In the constructive phase, pharmacophores were generated that were common among the active molecules of the training set. *HypoGen* identified all allowable pharmacophores consisting of up to five features among the two most active compounds and investigated the remaining active compounds in the list. The subtractive phase dealt with the pharmacophores that were created in the constructive phase and the program removed pharmacophores from the data structure which were not likely to be useful. Finally, the optimisation was done using the well-known simulated annealing algorithm. The algorithm applies small perturbations to the pharmacophores created in the constructive and subtractive phases in an attempt to improve the score. All improvements and some detrimental steps based on a probability function

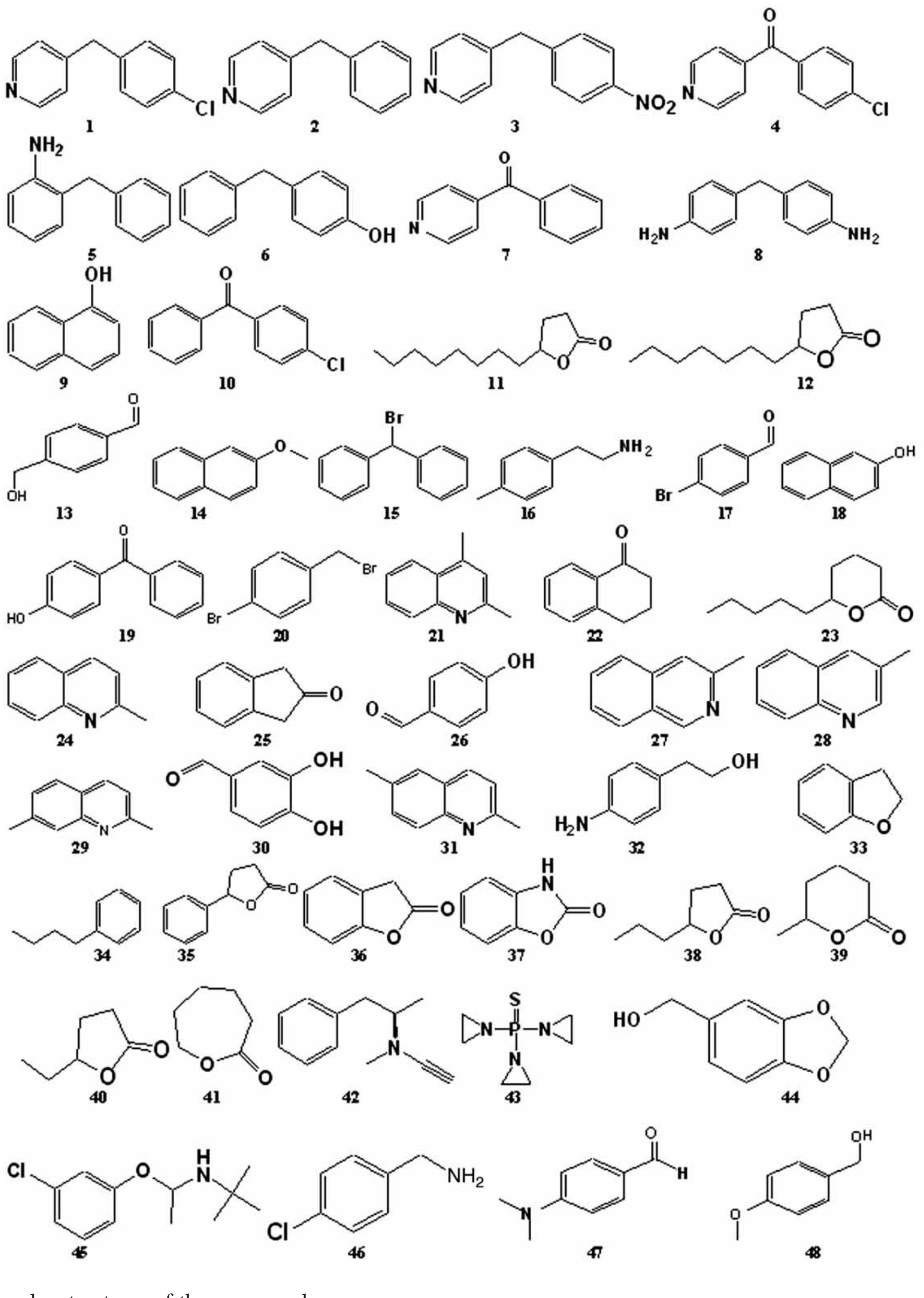


Figure 1. Molecular structures of the compounds.

are accepted and finally the highest scoring pharmacophores are exported.

Ten hypotheses were generated for each of the three sets of conformers (BEST, FAST, CAESAR) used. *HypoGen* allows a maximum of five features in pharmacophore generation. After elimination of the features that do not map the training set molecules, the following two features were selected for subsequent pharmacophore generation: hydrogen bond acceptor (HBA) and 'ring aromatic' (RA). The hypotheses generated were analysed in terms of their correlation coefficients and the cost function values. The *HypoGen* module performs a fixed cost calculation which represents the simple model that fits all the data and a null cost calculation that assumes that there is no relationship in the data-set and that the experimental activities are normally distributed about their average value. A small range of the total hypotheses cost obtained for each of the hypotheses indicates homogeneity of the corresponding hypothesis and that the training set selected for the purpose of pharmacophore generation is adequate. Again, values of total cost close to those of fixed cost are indicative of the fact that the hypotheses generated are statistically robust.

2.2.4 Pharmacophore model validation

Validation of a quantitative model was performed in order to determine whether the developed model was able to identify active structures and forecast their activity precisely. Validation of the obtained pharmacophore models was done using two procedures, viz. Fischer's validation as available in the *HypoGen* module and external validation using the test set prediction method.

2.2.5 Fischer's validation

The statistical significance of the structure–activity correlation is estimated using the Fischer's randomisation test [32]. This is done by scrambling the activity data of the training set molecules and assigning them new values followed by generation of pharmacophore hypothesis using the same features and parameters as those used to develop the original pharmacophore hypothesis. The number of spreadsheets obtained using the randomisation test depends on what level of statistical significance one wants to achieve. At 90% confidence level, nine spreadsheets are generated. The original hypothesis is considered to be generated by mere chance if the randomised data-set results in the generation of a pharmacophore with better correlation than the original one.

2.2.6 Prediction with test set molecules

The purpose of the pharmacophore hypothesis generation is not only to predict the activity of the training set compounds, but also to predict the activities of external molecules. With the objective to verify whether the pharmacophore was able to predict the activity of test set molecules in agreement with the experimentally determined value, the activities of the test set molecules were estimated using the developed pharmacophore models. The conformers generated for the test set molecules ($n = 26$) using FAST, BEST and CAESAR methods were selected and mapped using the corresponding pharmacophore models developed with the training set compounds.

For the performance evaluation of the pharmacophore models, several statistical tests such as recall (or sensitivity), specificity, accuracy, precision and F-measure were used [41,42]. Recall (or sensitivity) and specificity are able to identify the discrimination ability of the pharmacophore model, and accuracy presents the ratio of the correctly discriminated classes. F-measure is a function of recall and precision which indicate the accuracy of real and estimated class, respectively. The calculations of the above parameters according to Fawcett [42] are as follows:

$$\text{Recall} = \frac{\text{TA}}{\text{TA} + \text{FN}}, \tag{1}$$

$$\text{Precision} = \frac{\text{TN}}{\text{FA} + \text{TN}}, \tag{2}$$

$$\text{Specificity} = \frac{\text{TA}}{\text{TA} + \text{FA}}, \tag{3}$$

$$\text{Accuracy} = \frac{\text{TA} + \text{TN}}{\text{TA} + \text{FA} + \text{FN} + \text{TN}}, \tag{4}$$

$$F - \text{measure} = \frac{2(\text{Recall})(\text{Precision})}{\text{Recall} + \text{Precision}}, \tag{5}$$

where TA indicates the number of active compounds correctly classified as active, FA indicates the number of active compounds wrongly classified as active, FN indicates the number of non-active compounds wrongly classified as active, and TN indicates the number of non-active compounds correctly classified as non-active.

2.3 Descriptors for QSAR studies

The QSAR analyses were performed using spatial (radius of gyration, *Jurs* descriptors, area, PMI-mag, density, Vm), thermodynamic (AlogP, AlogP98, molar refractivity (*Molref*)) and structural (MW, hydrogen bond donor, HBA, chiral centres, number of rotatable bonds) and topological descriptors including E-state index descriptors. For the calculation of 3D descriptors, multiple conformations of each molecule were generated using the optimal search as a conformational search method. Each conformer was subjected to an energy minimisation procedure using smart minimiser under open force field to generate the lowest energy conformation for each structure. The charges were calculated according to the Gasteiger method. All the descriptors were calculated using descriptor + module of the Cerius2 version 4.10 software running on a Silicon Graphics workstation [43]. Definitions of all descriptors can be found at the Cerius2 tutorial available at the Accelrys website (http://www.accelrys.com).

2.4 QSAR model development

It was our priority to construct QSAR models which were statistically robust both internally and externally. The main target of any QSAR modelling is that the developed model should be robust enough to be capable of making accurate and reliable predictions of biological activities of new compounds. So, QSAR models which are developed from a training set should be validated using new chemical entities for checking the predictive capacity of the developed models. That is why the original data-set is divided into training and test sets for QSAR model development and validation, respectively. In our study, the whole data-set (n=48) was divided into training ($n = 36$) and test ($n = 12$) sets by k-means clustering techniques based on the standardised topological, structural and thermodynamic variables [44]. This approach (clustering) ensures that the similarity principle can be employed for the activity prediction of the test set [45]. QSAR models were developed using the training set compounds (optimised by Q^2), and then the developed models were validated (externally) using the test set compounds. For the development of the QSAR models, the statistical techniques used were Genetic function approximation (GFA) and Genetic partial least squares (G/PLS).

GFA technique [46,47] was used to generate a population of equations rather than one single equation for correlation between biological activity and physicochemical properties. GFA involves the combination of multivariate adaptive regression splines algorithm with genetic algorithm to evolve population of equations that best fit the training set data. It provides an error measure, called the lack of fit (LOF) score that automatically penalises models with too many features. It also inspires the use of splines as a powerful tool for non-linear modelling. A distinctive feature of GFA is that it produces a population of models (e.g. 100), instead of generating a single model, as do most other statistical methods. The range of variations in this population gives added information on the quality of fit and importance of the descriptors.

The G/PLS algorithm [48,49] was used as an alternative to a GFA calculation. G/PLS is derived from two QSAR calculation methods: GFA and PLS. The G/PLS algorithm uses GFA to select appropriate basis functions to be used in a model of the data and PLS regression as the fitting technique to weigh the basis functions' relative contributions in the final model. Application of G/PLS thus allows the construction of larger QSAR equations while still avoiding overfitting and eliminating most variables.

2.4.1 *Statistical qualities and model validation*

The statistical qualities of the QSAR equations were judged by the parameters such as squared correlation coefficient (R^2) and variance ratio (F) at specified degrees of freedom (df) [50]. For G/PLS equations, least-squares error was taken as an objective function to select an equation, whereas LOF was noted for the GFA-derived equations. The generated QSAR equations were validated by leave-one-out (LOO) cross-validation R^2 (Q^2) and predicted residual sum of squares [51–53] and then were used for the prediction of enzyme-inhibition activity values of the test set compounds. The prediction qualities of the models were judged by statistical parameters such as predictive R^2 (R^2_{pred}), squared correlation coefficient between observed and predicted values of the test set compounds with (r^2) and without (r_0^2) intercept. It was previously shown that use of R^2_{pred} and r^2 might not be sufficient to indicate the external validation characteristics [54]. Thus, an additional parameter $r^2_{\text{m(test)}}$ [defined as $r^2 * (1 - \sqrt{r^2 - r_0^2})$], which penalises a model for large differences between observed and predicted values of the test set compounds, was also calculated. Two other variants [55] of r^2_{m} parameter, $r^2_{\text{m(LOO)}}$ and $r^2_{\text{m(overall)}}$, were also calculated. The parameter $r^2_{\text{m(overall)}}$ is based on prediction of both training (LOO prediction) and test set compounds. It was previously shown [55,56] that $r^2_{\text{m(LOO)}}$ and $r^2_{\text{m(test)}}$ penalise a model more strictly than Q^2 and R^2_{pred}, respectively. As an additional tool for validation, randomisation test was applied to the model development process and the developed models. This method is of two types: process randomisation and model randomisation. In case of process randomisation, the values of the dependent variable are randomly scrambled and variable selection is done freshly from the whole descriptor matrix. In case of model randomisation, the Y column entries are scrambled and new QSAR models are developed using the same set of variables as present in the unrandomised model. The process randomisation test was performed at 90% confidence level for the process, and the developed models were subjected to randomisation test at 99% confidence level. The parameter R_{p}^2 ($R_{\text{p}}^2 = R^2 * \sqrt{R^2 - R_{\text{r}}^2}$) ($R_{\text{r}}^2$ being squared mean correlation coefficient of random models) was also calculated [57] to check whether the models thus developed are not obtained by chance.

2.5 *Molecular docking studies*

The crystal structure of human CYP450 2B6 genetic variant in complex with the inhibitor 4-(4-chlorophenyl) imidazole has been recently published [31]. We have collected the crystal structure from the RCSB protein data bank (http://www.pdb.org) and conducted a docking study for the compounds considered in the present paper using the LigandFit tool available in Discovery Studio 2.1 [32]. Initially, there was a pretreatment process for both the ligands and the enzyme (CYP2B6). For ligand preparation, all the duplicate structures were removed and the options for ionisation change, tautomer generation, isomer generation, Lipinski filter and 3D generator were set true. For enzyme preparation, the whole enzyme was selected and hydrogen atoms were added to it. The pH of the protein was set in the range of 6.5–8.5. Then, we have defined the CYP2B6 enzyme as the total receptor and the active site was selected based on the ligand-binding domain of bound ligand 4-(4-chlorophenyl)imidazole. Then, the pre-existing ligand [4-(4-chlorophenyl)imidazole] was removed and a freshly prepared ligand (compound from the data-set in Table 1) prepared by us was placed. Then, from the receptor–ligand interaction section, LigandFit was chosen. We have used the preprocessed receptor and ligand as inputs. Dreiding was selected as the energy grid. The conformational search of the ligand poses was performed by Monte Carlo trial method. Torsional step size for polar hydrogen was set at 10. The docking was performed with consideration of electrostatic energy. Maximum internal energy was set at 10,000 cal. Pose saving and interaction filters were set as default. Fifty poses were docked for each compound. During the procedure of docking, no attempt was made to minimise the ligand–enzyme complex (rigid docking). After completion of docking, the docked enzyme (protein–ligand complex) was analysed to investigate the type of interactions. Ten docking poses saved for each compound were ranked according to their dock score function. The pose (conformation) having the highest dock score was selected and was analysed to investigate the type of interactions.

3. Results and discussion

3.1 Pharmacophore generation

3.1.1 Pharmacophore development with conformers generated using the FAST method of conformer search

A set of nine pharmacophore hypotheses was generated using the conformers generated from each of the FAST, BEST and CAESAR methods using the 22 training set compounds. The results of the best one hypothesis from each of the FAST, BEST and CAESAR methods together with the pharmacophore features, cost functions and correlation values are listed in Table 3. In the FAST method, the total hypothesis cost, expressed in bits, of the nine best hypotheses varied from 96.192–113.657 (17.47 bits). A 17.47-bit cost range gained over the models, suggesting the existence of a moderate signal generated by the training set [32]. Configuration cost (also known as entropy cost) parameter describes the complexity of the hypothesis space to explore. The value of the configuration cost (6.57 bit) is constant among all the hypotheses. Any value higher than 17 indicates that the correlation from any generated pharmacophore is most likely due to chance and in that case some attention should be given to training set molecules. The difference between null hypothesis and the fixed cost and that between fixed cost and total cost of the best hypothesis (hypothesis **3**) were 27.10 and 16.20 bits, respectively. The predominant features in the generated nine pharmacophore hypotheses were HBA, RA and to a minor extent HYD (for hypothesis **7** only). The first four hypotheses (**1, 2, 3 and 4**) contain two features (RA, HBA). For hypotheses **5, 6, 8** and **9,** the features were RA and HBA, whereas for hypothesis **7**, the features were HBA, HYD and RA. These observations are in agreement with the pharmacophore results for CYP2B6 substrates involving HYD and HBA features [30,58]. The generated best hypothesis **1** (based on correlation) showed moderate value (0.778) of correlation coefficient because of a high root mean square deviation (RMSD) value (1.146) which indicates difference between estimated and measured activities for some of the training set molecules. The reason for lower correlation may be due to lack of common features within the training set molecules due to high structural diversity. Only compounds such as **1**, **3**, **4**, **6**, **14**, **17**, **18** and **23** share both the two features for the generated pharmacophore. For example, compounds such as **11**, **12** and **43** share only HBA features and a large group of compounds such as **15**, **16**, **8**, **5**, **20**, **24**, **28**, **31**, **42**, **45** and **46** share only the RA features in the generated hypothesis. But emphasis was given to test the models for test set prediction. On the basis of the external predictive power of the pharmacophore, hypothesis **3** was selected as the best one among the others. Among the nine pharmacophore hypotheses, all hypotheses except hypothesis **3** were unable to map the inactive molecules in the test set. Hypothesis **3** was capable of mapping 21 molecules out of 26 molecules in the test set. Therefore, we have chosen hypothesis **3** as the best one. The accuracy of hypothesis **3** was 100% when applied to the training set molecules (Table 1). The external statistical measures like recall, precision, specificity, accuracy and *F*-measures for hypothesis **3** were 88.2, 68.75, 75.0, 78.8 and 77.30%, respectively and are listed in Table 4. It was observed that two feature pharmacophore hypotheses gave better external prediction as the test set molecules were small and could not get the appropriate distance for too many features. Hypothesis **3** possesses two pharmacophore features: HBA and RA. The distance between the two centres of the features (HBA, RA) was 5.820 Å (Figure 2). From Figure 2 (most active compound **1**), it is evident that pyridinic nitrogen atom acts as HBA (electron-rich centre)

Table 3. Results of the best pharmacophore hypotheses generated using conformers developed from the FAST, BEST and CAESAR methods of conformer search.

Method of conformer generation	Hypothesis no.	Total cost	Error cost	rms	Fixed cost	Null cost	Features	Correlation (*R*)	Random *R* (±SE)
FAST	**3**	**97.898**	**90.195**	**1.213**	81.694	109.673	**HBA,RA**	**0.738**	0.398 (±0.051)
BEST	**3**	**95.765**	**90.237**	**1.215**	79.517	109.673	**HBA,RA**	**0.738**	0.331 (±0.056)
CAESAR	**3**	**102.765**	**91.82**	**1.272**	84.719	109.673	**HBA,RA**	**0.710**	0.366 (±0.031)

Table 4. Statistical measures for evaluation of the pharmacophore models based on their application on test set compounds.

Method of conformer generation	Recall	Precision	Specificity	Accuracy	F-measure
FAST	0.882	0.688	0.750	0.788	77.30
BEST	0.882	0.688	0.750	0.788	77.30
CAESAR	0.882	0.688	0.750	0.788	77.30

and the other phenyl ring linked with methylene bridge serves as RA, and this is similar for compounds **3** and **4**. For compounds **6** and **18,** the oxygen atom of the —OH group acts as HBA (electron-rich centre), phenyl and naphthalene moieties act as RA features showing significant inhibitory activity. The oxygen of the methoxy group of compound **14** and terminal keto group of compound **17** act as HBA and showed good inhibitory activity. The developed pharmacophore model was subjected to Fischer's randomisation test at 90% confidence level. The experimental activities of the compounds in the training set were permuted nine times and spreadsheets were obtained with the randomised activity data. The average value of the randomised correlation coefficient with standard error for hypothesis **3** was found to be 0.398 (± 0.051) which is much lower than the model correlation coefficient (0.738), indicating that the model was not by chance. Though we got HBA as an important feature in the pharmacophore study, it appears that basically an electron-rich centre capable of formation of co-ordinate bond with the iron of cytochrome is important (see the results of the docking study).

3.1.2 Pharmacophore development with conformers generated using the BEST method of conformer search

Similar to the earlier discussion, nine pharmacophore hypotheses were generated using the training set molecules for the BEST method. The result of the best hypothesis (hypothesis **3**) is listed in Table 3. In the BEST method, the total hypothesis cost, expressed in bits, of the nine best hypotheses varied from 93.8662–115.321 (21.45 bits), suggesting existence of a moderate signal generated by the training set. The configuration cost was 4.392 bits indicating the appropriate selection of training set selection. The difference between null hypothesis and the fixed cost and that between the fixed cost and total cost of the best hypothesis (hypothesis **3**) were 30.16 and 16.25 bits, respectively. The predominant features in the generated hypotheses were RA and HBA supporting previous observation with CYP2B6 substrates involving HYD and HBA features [28,57]. On the basis of the external prediction, hypothesis **3** was selected as the best hypothesis. The generated best hypothesis **1** (based on correlation) showed moderate value (0.773) of correlation coefficient because of a high RMSD value (1.142). Hypothesis **3**

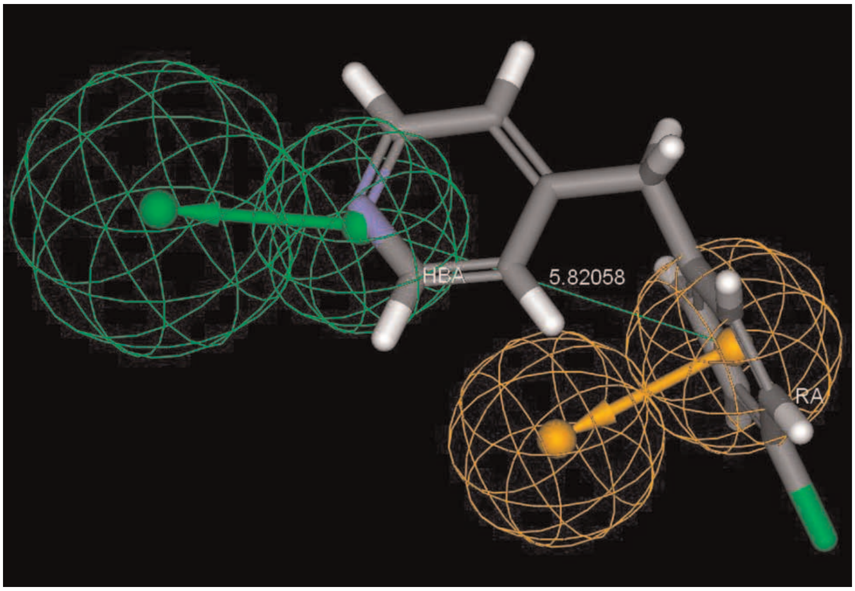


Figure 2. Most active compound (Compound **1**) mapped into the pharmacophore obtained from hypothesis **3** using the training set conformers developed from the FAST method of conformer generation. Shown are RA sphere (orange) feature with vector in the direction of possible pi–pi interaction and HBA (green) feature with vector in the direction of putative hydrogen bond (colour online).

was found to be the best model based on the mapping of test set molecules. The accuracy of hypothesis **3** was 100% when applied to the training set molecules (Table 1). Hypothesis **3** was capable of mapping 18 molecules out of 26 molecules in the test set. The external evaluation parameters for hypothesis **3** are listed in Table 4, indicating validity of the generated hypothesis. The distance between the two centres of the features (HBA, RA) was 6.023 Å (Figure 3). Other observations were similar to the FAST method. Fischer's randomisation test was done at 90% confidence level and the randomised correlation coefficient value with standard error 0.331(± 0.056) was much lower to model correlation coefficient of 0.738 of hypothesis **3** (Table 3).

3.1.3 Pharmacophore development with conformers generated using the CAESAR method of conformer search

By applying CAESAR method, nine pharmacophore hypotheses were generated. The total hypothesis cost, expressed in bits, of the nine best hypotheses varied from 99.4926–106.36 and such a small range (covering only 7 bits) was suggestive of the fact that the generated hypothesis was homogeneous. The configuration cost was 9.594 bits indicating appropriate selection of training set molecules. The difference between null hypothesis and the fixed cost, and that between the fixed cost and total cost of the best hypothesis (hypothesis **3**) were 24.95 and 18.05 bits, respectively. Hypothesis **3** was capable of mapping 18 molecules out of 26 molecules in the test set and was selected as the best hypothesis. The external evaluation parameters for hypothesis **3** are listed in Table 4. Accuracy prediction of training set prediction was 100% for hypothesis **3** (Table 1). The generated best hypothesis **1** (based on correlation) showed moderate value (0.765) of correlation coefficient because of a high RMSD value (1.159). Other results were similar to those of the earlier discussion. The distance between the two centres of the features (HBA, RA) was 6.028 Å (Figure 4). The randomised correlation coefficient value was much lower than the model correlation coefficient value as listed in Table 3.

3.1.4 Overview

The suggested pharmacophores, in all three methods of conformer generation, contain RA and HBA features. These are only hypotheses and may fail in some cases. Compounds **33** and **34** could not be mapped because these do not contain the required pharmacophoric features (according to the suggested hypotheses). However, compounds **35–37**, though having lower experimental values, could be mapped as these contain the required pharmacophoric features. Compound **38** contains one HBA and one HYD feature whereas compounds **39–41** contain HYD features; but none of them contain RA features and hence these could not be mapped.

3.2 QSAR analyses

Membership of compounds in different clusters generated using *k*-means clustering is shown in Table 5. The test set size was set to approximately 25% to the total data-set size [59] and the test set members are shown in Table 2. The following two equations (Equations (6) and (7)) were among the best ones obtained from the GFA (5000 iterations). Both linear and linear spline terms were used for the development of the models. The difference between R^2 and Q^2 values is not very high (less than 0.3) [60] for

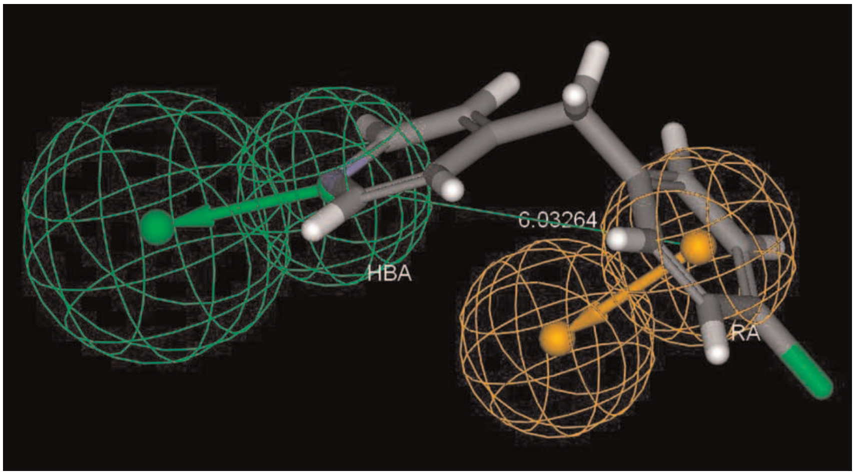


Figure 3. Most active compound (Compound **1**) mapped into the pharmacophore obtained from hypothesis **3** using the training set conformers developed from the BEST method of conformer generation. Shown are RA sphere (orange) feature with vector in the direction of possible pi–pi interaction and HBA (green) feature with vector in the direction of putative hydrogen bond (colour online).

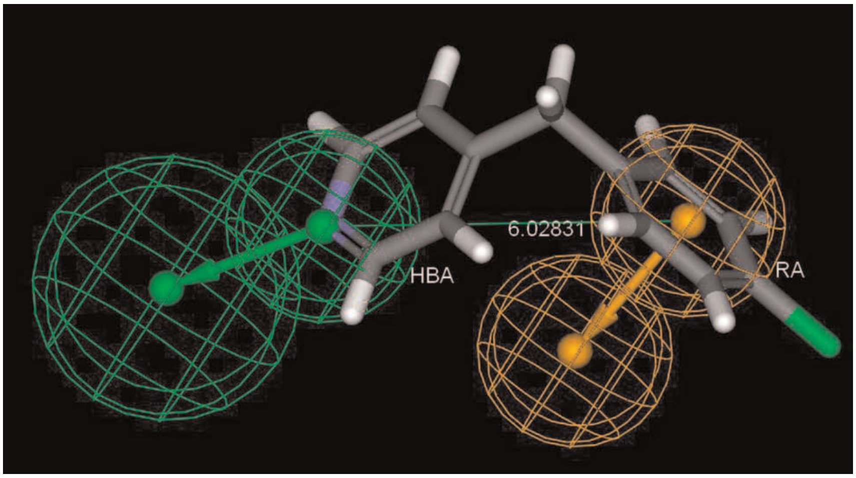


Figure 4. Most active compound (Compound **1**) mapped into the pharmacophore obtained from hypothesis **3** using the training set conformers developed from the CAESAR method of conformer generation. Shown are RA sphere (orange) feature with vector in the direction of possible pi–pi interaction and HBA (green) feature with vector in the direction of putative hydrogen bond (colour online).

all the developed models.

$$\begin{aligned}\mathrm{pIC}_{50} = {} & 5.502(\pm 0.231) - 0.643(\pm 0.086) \\ & < 8.407 -{}^{0}\chi^{\nu} > -3.994(\pm 1.126) \\ & < S\_aaN - 3.980 > -66.190(\pm 16.500) \\ & < 0.043 - Jurs\_\mathrm{FPSA\_3} \\ & > +0.268(\pm 0.082) S\_aaN \end{aligned} \qquad (6)$$

$$n_{\mathrm{Training}} = 36,\ \ \mathrm{LOF} = 0.568,\ \ R^2 = 0.790,$$
$$R_{\mathrm{a}}^2 = 0.762,\ \ F = 29.08(\mathrm{df}\ 4, 31),\ \ Q^2 = 0.725,$$
$$r^2_{\mathrm{m(LOO)}} = 0.723, n_{\mathrm{Test}} = 12,\ \ R^2_{\mathrm{pred}} = 0.843,$$
$$r^2_{\mathrm{m(test)}} = 0.676,\ \ r^2_{\mathrm{m(overall)}} = 0.754.$$

The relative importance of the descriptors according to their standardised regression coefficients is in the following order: $< 8.407 -{}^{0}\chi^{\nu} >>< S\_aaN - 3.980 > >< 0.043 - Jurs\_\mathrm{FPSA\_3} >> S\_aaN$.

The standard errors of regression coefficients are given within parentheses. Equation (6) could explain 76.2% of the variance (adjusted coefficient of variation) while it could predict 72.5% of the variance (LOO predicted variance). When the equation was used to predict the CYP2B6 inhibition potency of the test set compounds, the predicted R^2 (R^2_{pred}) value was found to be 0.843. The r^2_{m} values for the test, training and overall sets were found to be 0.676, 723 and 0.754, respectively.

The zero-order valance-modified connectivity index (${}^{0}\chi^{\nu}$) indicates the number of sub-graphs of zero order that is, therefore, equal to the number of skeletal atoms or vertices. This indicates the size of the molecular skeleton. The term $< 8.407 - {}^{0}\chi^{\nu} >$ indicates that for optimal inhibitory activity of the compounds, the value of ${}^{0}\chi^{\nu}$ should be greater than 8.407. Compounds such as **1**, **3** and **8** showed good inhibitory activity compared to compounds (such as compounds **30**, **37**, **39** and **41**) with lower value of the parameter.

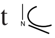
The negative regression coefficient of the term $< S\_aaN - 3.980 >$ indicates that the value of the E-state index of the fragment (*S_aaN*) should be less than or equal to 3.980 for ideal inhibition. The linear term in the equation for the E-state index of the fragment  i.e. *S_aaN* shows positive contribution towards inhibitory activity. Thus this indicates that high value of *S_aaN* parameter facilitates inhibitory activity. Compounds having non-zero value of *S_aaN* are **1–3**, **7**, **21**, **24**, **27** and **31**. From the linear and spline term of *S_aaN*, it was evident that the value of *S_aaN* should be high but less

Table 5. *k*-Means clustering of compounds using standardised descriptors.

Cluster no.	No. of compounds in different clusters	Compounds (Sl nos.) in different clusters																	
1	8	**1**	**2**	**5**	**6**	**7**	**12**	**15**	**42**										
2	15	**13**	**17**	**20**	**25**	**26**	**30**	**33**	**36**	**37**	**38**	**39**	**40**	**41**	**46**	**48**			
3	7	**3**	**4**	**8**	**10**	**11**	**19**	**45**											
4	18	**9**	**14**	**16**	**18**	**21**	**22**	**23**	**24**	**27**	**28**	**29**	**31**	**32**	**34**	**35**	**43**	**44**	**47**

than 3.980. Among the compounds listed above, compounds such as **1**, **3** and **7** show significant inhibitory activity satisfying the above mentioned range of *S_aaN*. The rest of the compounds (such as **21**, **24**, **27** and **31**) showed poor activity due to lower value of the term ${}^{0}\chi^{\nu}$. The results are also supported by the generated pharmacophore hypothesis. In the generated hypothesis, the fragment present in the pyridine nucleus serves as HBA feature (electron-rich centre) in all the three methods of conformer generation for the compounds with higher activity (such as compounds **1**, **3** and **4**).

The next term in Equation (6) is $<0.043 -$ *Jurs*_FPSA_3 $>$. The negative regression coefficient of the term indicates that the value of fractional charged partial surface areas (*Jurs*_FPSA_3) should be more than 0.043 for the required inhibitory activity. *Jurs*_FPSA_3 is derived from the following equation: $\mathrm{FPSA\_3} = \frac{\mathrm{PPSA_3}}{\mathrm{SASA}}$, where PPSA$_3$ (atomic charge weighted positive surface area) is the sum of products of atomic solvent accessible surface areas (SASA) and partial charges q_a^+ over all positively charged atoms ($\mathrm{PPSA_3} = \sum_{a+} q_a^+ . \mathrm{SA}_a^+$). Compounds such as **2**, **5**, **6**, **8** and **9** with *Jurs*_FPSA_3 value greater than 0.043 showed good inhibitory activity. On the other hand **21, 31** and **38–40** with low values of *Jurs*_FPSA_3 showed poor inhibitory activity. Compounds such as **30**, **32** and **37** with high value of *Jurs*_FPSA_3 showed poor inhibitory activity because of the absence of fragment and lower value of ${}^{0}\chi^{\nu}$.

$$\begin{aligned}\mathrm{pIC_{50}} = {} & 3.829(\pm 0.267) - 10.774(\pm 1.772) \\ & < 3.980 - S\_aaN > +10.432(\pm 1.669) \\ & < 4.207 - S\_aaN > -0.093(\pm 0.016) \\ & < 51.395 - Molref > -8.046(\pm 2.827) \\ & < 0.225 -{}^{3}\,\chi_c^{\nu} > \end{aligned} \tag{7}$$

$$n_{\mathrm{Training}} = 36,\ \ \mathrm{LOF} = 0.533,\ \ R^2 = 0.818,$$
$$R_a^2 = 0.795,\ \ F = 34.85(\mathrm{df}\,4, 31),\ \ Q^2 = 0.772,$$
$$r^2_{\mathrm{m(LOO)}} = 0.750, n_{\mathrm{Test}} = 12,\ \ R^2_{\mathrm{pred}} = 0.832,$$
$$r^2_{\mathrm{m(test)}} = 0.749,\ \ r^2_{\mathrm{m(overall)}} = 0.774.$$

Equation (7) was found to be statistically significant with explained variance of 79.5% and LOO predicted variance of 77.2%. When Equation (7) is applied to the test set compounds, the R^2_{pred} value was found to be 0.832. Statistical significance of the model is also indicated by the r^2_{m} parameters listed in Table 6. The relative order of importance of the descriptors: $< 3.980 - S\_aaN >> < 4.207 - S\_aaN >> < 51.395 - Molref >> < 0.225 -{}^{3}\,\chi_c^{\nu} >$.

Two spline terms of the E-state index of the fragment *S_aaN* appearing in Equation (7) were $<3.980 - S\_aaN>$ and $<4.207 - S\_aaN>$. The positive and negative coefficients of $<4.207 - S\_aaN>$ and $<3.980 - S\_aaN>$ terms, respectively, indicate that the values of the *S_aaN* should be between 3.980 and 4.207 for significant inhibitory activity. Compounds such as **1** and **2** having *S_aaN* within this range showed significant inhibitory activity (unlike compound **27**). Considering a further extension of the lower limit of 3.980 for *S_aaN* to less than 3.980 as observed in Equation (6), we got compounds such as **3** and **7** with significant inhibitory activity.

The *Molref* index of a substituent is a combined measure of its size and polarisability. The negative coefficient of $<51.395 - Molref>$ indicates that molar refractivity is conducive for the inhibitory activity when the value is more than 51.395. For optimum activity, the values *Molref* should be less than 51.395 (such as in compounds **1–3**, **5–8**, **10** and **43**). Compounds with lower values of *Molref* such as **33**, **36–39** and **41** showed poor inhibitory activity.

The third-order molecular connectivity index of valance-modified cluster connectivity index (${}^{3}\chi_c^{\nu}$) indicates the impact of branching. The negative coefficient of the term $<0.225 - {}^{3}\chi_c^{\nu}>$ suggests that the value of ${}^{3}\chi_c^{\nu}$ should be more than 0.225 for optimum inhibitory activity (for example compounds **1–3**, **5**, **6**, **8** and **43**). Compounds such as **33**, **34**, **37**, **38** and **41** with low values of ${}^{3}\chi_c^{\nu}$ showed poor inhibitory activity. But compounds (such as **25**, **30**, **35** and **39**) with values of ${}^{3}\chi_c^{\nu}$ greater than 0.225 showed poor inhibitory activity due to lack of aromatic nitrogen and lower *Molref* values.

Equation (8) is one of the best ones obtained from the G/PLS (1000 crossovers, scaled variables, and other default settings respectively). Both linear and linear spline terms were used for development of the models.

Table 6. Comparison of statistical qualities of different QSAR models[a].

Type of statistical analysis	Equation no.	R^2	Q^2	R^2_{pred}	$r^2_{\mathrm{m(test)}}$	$r^2_{\mathrm{m(LOO)}}$	$r^2_{\mathrm{m(overall)}}$
GFA spline	6	0.790	0.725	**0.843**	0.676	0.723	0.754
	7	0.818	**0.772**	0.832	0.749	0.750	**0.774**
G/PLS spline	8	0.751	0.648	0.696	0.654	0.449	0.487

[a] The best values of Q^2, R$^2_{\mathrm{pred}}$ and r$^2_{\mathrm{m(overall)}}$ are shown in bold.

$$\begin{aligned}\mathrm{pIC}_{50} = {} & 5.407 - 0.125 < 54.476 - \mathit{Molref} > \\ & - 0.149 < 23.993 - \mathit{Jurs}\_\mathrm{DPSA}\_3 > \\ & + 0.711 < \mathit{Hbondacceptor} - 2 > \\ & + 0.621 \mathit{S\_dsCH} - 0.087 {}^3\kappa_\mathrm{a} \\ & n_{\mathrm{Training}} = 36,\ \mathrm{LSE} = 0.323,\ R^2 = 0.751, \\ & R^2_\mathrm{a} = 0.728,\ F = 32.26(\mathrm{df}\ 3, 32),\ Q^2 = 0.648, \\ & r^2_{\mathrm{m(LOO)}} = 0.449,\ n_{\mathrm{Test}} = 12,\ R^2_{\mathrm{pred}} = 0.696, \\ & r^2_{\mathrm{m(test)}} = 0.654,\ r^2_{\mathrm{m(overall)}} = 0.487.\end{aligned} \quad (8)$$

The statistical quality of Equation (8) is listed in Table 6. According to the standardised values of the regression coefficients, the relative importance of the variables is in the following order: $<54.476 - \mathit{Molref}> \; > \; < 23.993 - \mathit{Jurs}\_\mathrm{DPSA}\_3 > \; > \; < \mathit{Hbondacceptor} - 2 > \; > \mathit{S\_dsCH} > {}^3\kappa_\mathrm{a}$.

The negative coefficient of the term $<54.476 - \mathit{Molref}>$ indicates that the value *Molref* should be greater than 54.475. The result obtained for Equation (8) is similar to that for Equation (7).

The difference in atomic charge weighted surface area (*Jurs*_DPSA_3) which is the atomic charge weighted positive solvent accessible surface area minus the atomic charge weighted negative solvent accessible surface area is expressed as

$$\mathrm{DPSA}_3 = \mathrm{PPSA}_3 - \mathrm{PNSA}_3,$$

where atomic charge weighted negative surface area (PNSA_3) is the sum of products of atomic SASA and partial charges q^-_a over all negatively charged atoms $\mathrm{PNSA}_3 = \sum_{\mathrm{a}-} q^-_\mathrm{a} \cdot \mathrm{SA}^-_\mathrm{a}$. The negative coefficient of the term $<23.993 - \mathit{Jurs}\_\mathrm{DPSA}\_3>$ indicates that *Jurs*_DPSA_3 has a positive impact when it is more than 23.993. Compounds such as **1**–**3** and **5**–**9** showed significant inhibitory activity and the values of *Jurs*_DPSA_3 for the above compounds are more than 23.993. Similarly, compounds such as **20**, **21**, **24**, **27**, **31**, **33** and **34** showed poor inhibitory activity due to lower values of *Jurs*_DPSA_3. But compounds such as **32, 35**–**39** and **41** with higher value of *Jurs*_DPSA_3 showed poor inhibitory activity due to lower *Molref* value.

The term $<\mathit{Hbondacceptor} - 2>$ with a negative regression coefficient indicates that the number of HBA groups should be 2 or less than 2 for optimum inhibitory activity. However, considering Equations (6) and (7) and also the results of the docking study (*vide infra*), the *Hbondacceptor* group actually indicates contribution of an electron-rich centre such as a nitrogen atom capable of formation of co-ordinate bond with the iron of cytochrome. Compounds such as **1**, **7**, **10** and **11** with two and compounds such as **2** and **6** with one HBA group(s) [i.e. electron-rich centre(s)] showed significant inhibitory activity. Compounds such as **35**–**39** and **41** with two HBA groups [i.e. electron-rich centres] showed poor inhibitory activity due lower *Molref* value.

The E-state index of the fragment $=\mathrm{CH}-$ (*S_dsCH*) has a positive contribution towards the inhibitory activity. Compounds such as **13**, **17**, **30** and **47** have non-zero values for the $=\mathrm{CH}-$ fragment and they showed moderate inhibitory activity.

The third order alpha-modified shape index (${}^3\kappa_\mathrm{a}$) indicates the shape of molecules considering the size differences among the hetero atoms in different valance states and it has detrimental contribution towards the inhibition potency as evidenced by the negative regression coefficient.

The results of process and model randomisation tests are shown in Table 7, which shows that the G/PLS-derived model [Equation (8)] does not fulfil the required criterion of R^2_p (the value being less than 0.5). On the basis of the results on randomisation tests, the GFA-derived Equation (7) is found to be more reliable than the other reported equations.

3.3 Molecular docking study

For validation of the docking study, we removed the preexisting co-crystallised ligand and docked freshly prepared and energy minimised ligand, and compared the binding site of the preexisting co-crystallised ligand and that of the freshly prepared ligand. The reliability of the docking procedure was indicated by the low RMSD value (0.79 Å) obtained between the bound ligand in the crystal structure and computationally freshly prepared docked ligand (Figure 5).

Table 7. Results of the process and model randomisation tests[a].

	Confidence level	Equation No.	R^2	R^2_r	R^2_p
Process randomisation	90%	6	0.790	0.335	0.533
	90%	7	0.818	0.335	**0.568**
	90%	8	0.751	0.477	0.393
Model randomisation	99%	6	0.790	0.112	0.650
	99%	7	0.818	0.110	**0.689**
	99%	8	0.751	0.008	0.648

[a] The best values of R^2_p (process randomisation and model randomisation) are shown in bold.

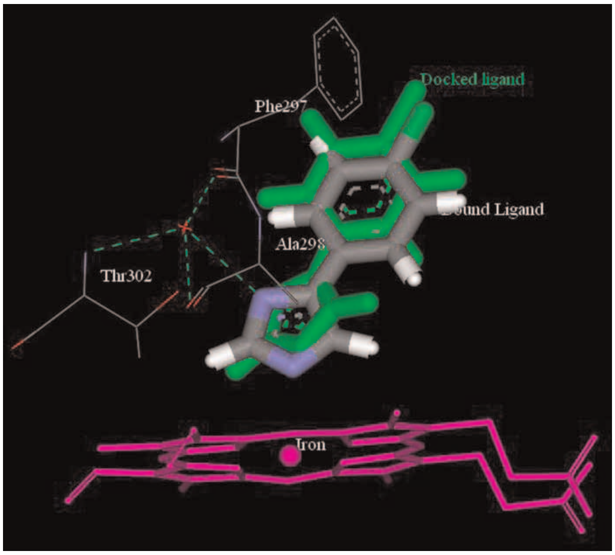


Figure 5. Superimposition of docked ligand and bound ligand (4-(4-chlorophenyl)imidazole) in the active site of human CYP2B6 enzyme.

The important amino acid residues in the active site cavity (within 4 Å) are Phe206, Phe297, Thr302, Glu301, Ala298, Ile114, Ile101, Phe115, Leu363, Val367, Pro368, Val477,Gly366 and Gly487, and this observation is in agreement with the crystal structure information [31]. Although most of the residues in the binding site of CYP2B6 are HYD, there are two residues with polar side chains, Glu301 and Thr302.

Considering the most active compound in the data-set, compound **1** is stabilised in the active site by the non-polar amino acid residues such as Phe206, Phe297, Thr302, Ala298, Ile114, Ile101, Phe115, Leu363, Val367, Pro368 and Val477 (Figure 6) by HYD interactions. The electron-rich centre of compound **1**, i.e. pyridine nitrogen, is in close proximity (2.487 Å) with iron moiety and it coordinates with haeme moiety which is one of the very essential properties for any CYP inhibitors [61]. Similar observations were observed for compounds **2** and **7**. The only difference is the distance between iron and pyridine nitrogen which is 2.573 Å for compound **2** and 2.531 Å for compound **7**. This is in agreement with the *S_aaN* term contribution in QSAR Equations (6) and (7) [*vide supra*]. The term *Hbondacceptor* in QSAR Equation (8) [*vide supra*] appears to indicate contribution of an electron-rich centre capable of formation of coordinate bond with iron of cytochrome instead of formation of a hydrogen bond. In case of compound **3** (Figure 7) containing three HBA groups, the nitro group is close to haeme moiety and this group changes the orientation of the molecule in such a

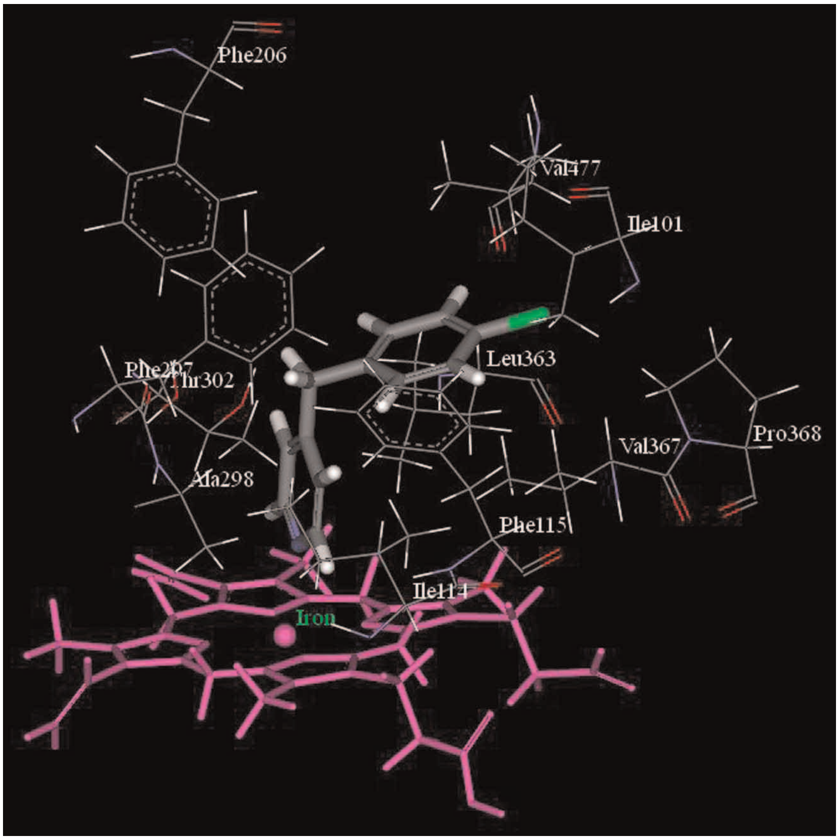


Figure 6. Docked conformation of compound **1** along with the important amino acid residues of human CYP2B6 enzyme.

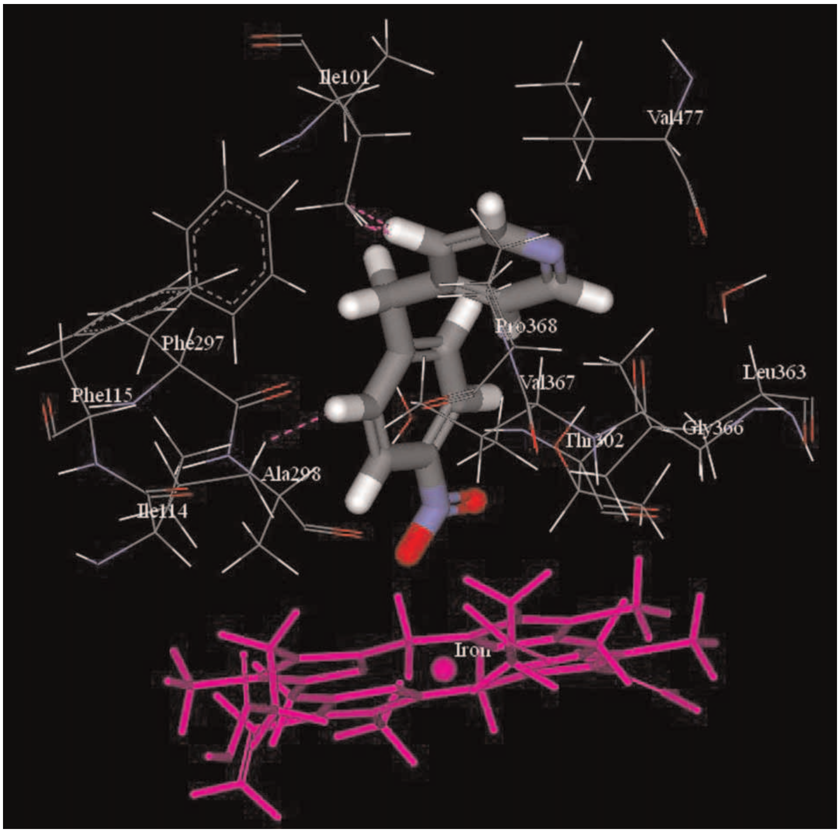


Figure 7. Docked conformation of compound **3** along with the important amino acid residues of human CYP2B6 enzyme.

way which facilitates unfavourable interactions with amino acid residues (Ile101 and Ile114).

For the least active congeners in the data-set (compound **41**), the important amino acids in the active site are Phe206, Val367, Thr302, Ala298, Ile114, Phe297, Cys436 and Leu363 (Figure 8). According to pharmacophoric hypotheses, this compound lacks the RA feature. The docking results show that this molecule is far away from the iron moiety thus making weak binding with the haeme moiety. There is also formation of a bump with Phe297 leading to poor inhibitory activity. The docking study with another least active compound **38** shows appearance of a number of bumps with the keto group of molecule with the haeme moiety as well as a single intramolecular bump (Figure 9).

To justify our QSAR results, we have performed docking study of the molecules with more than 2 HBA groups (electron-rich centres). The keto group of compound **30** produces lots of unfavourable interactions with the haeme moiety and shows poor inhibitory activity (Figure 10). Another compound **45** produces lots of bumps with amino acid residues (Ile114, Thr302 and Leu363) as well as some intramolecular bumps due to its unfavourable orientation in the active site (Figure 11).

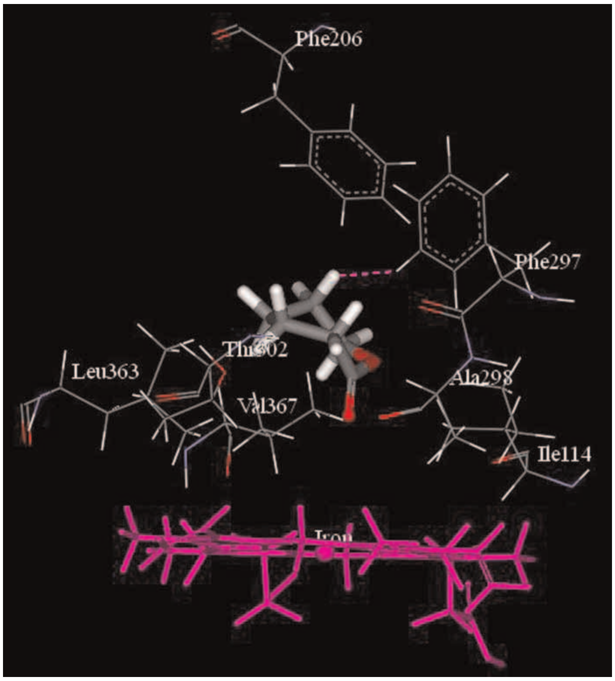


Figure 8. Docked conformation of compound **41** along with the important amino acid residues of human CYP2B6 enzyme.

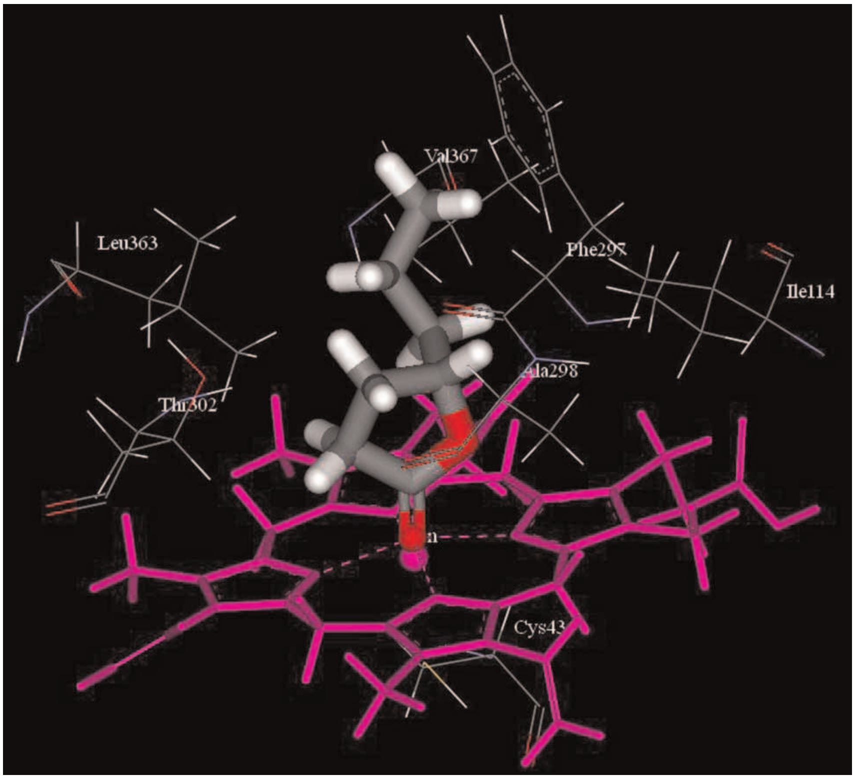


Figure 9. Docked conformation of compound **38** along with the important amino acid residues of human CYP2B6 enzyme.

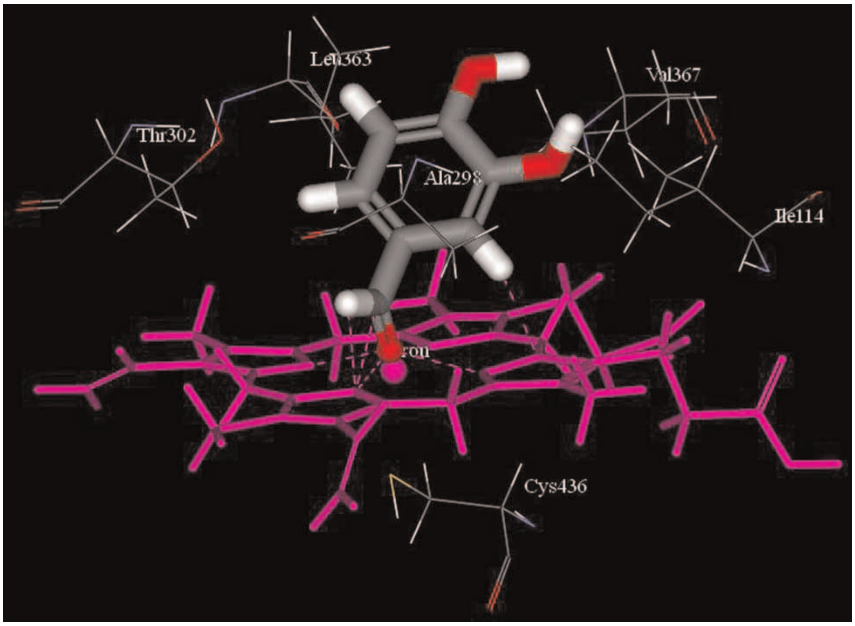


Figure 10. Docked conformation of compound **30** along with the important amino acid residues of human CYP2B6 enzyme.

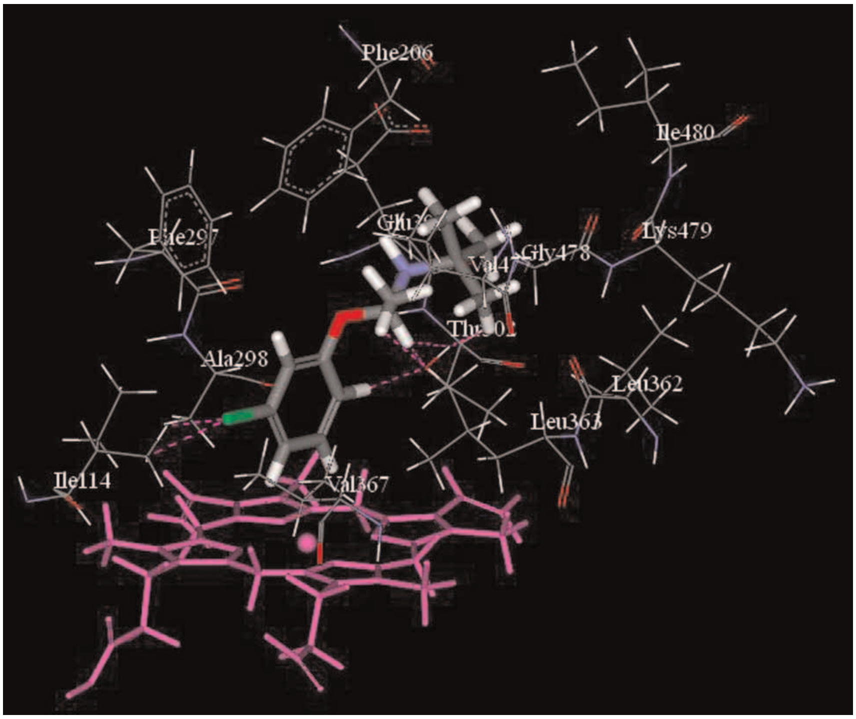


Figure 11. Docked conformation of compound **45** along with the important amino acid residues of human CYP2B6 enzyme.

4. Overview and conclusion

Pharmacophore mapping and QSAR studies were carried out for a structurally diverse set of 48 compounds as CYP2B6 inhibitors. For the two types of analyses, training and test sets were selected very carefully. For pharmacophore development, 22 compounds were selected as the training set and the remaining 26 compounds as the test set. For QSAR study, the training ($n = 36$) and test ($n = 12$) sets were selected by k-means clustering techniques. The reason is that for generation of pharmacophore, we have to select the active molecules in the training set otherwise steric clashes and other unfavourable features of the inactive molecules would decrease the quality of the developed pharmacophore. On the other hand, for QSAR study, the test compounds should be within the domain of training set compounds for good prediction. The generated best hypotheses of pharmacophores from the three methods of conformer generation (FAST, BEST and CAESAR) indicate the importance of two features, namely HBA (electron-rich centre) and RA. The distance between the two centres of features for ideal inhibitors varied from 5.820 to 6.028 Å.

We have also performed docking study for the compounds. It was observed from the docking study that the active site of the enzyme CYP2B6 is almost completely HYD. The only polar residues in the active site are Thr302 and Glu301. From the docking study it can be concluded that the molecules should contain an electron-rich centre capable of formation of a co-ordinate bond with the haeme moiety of the cytochrome enzyme. Furthermore, stabilisation of binding of the molecules in the active site occurs through the HYD interactions and possible pi–pi interactions with aromatic amino acids using the RA feature as observed in the case of the pharmacophore study. Compounds with more than two HBA groups (electron-rich centres) produce unfavourable interactions in the active site cavity leading to poor inhibitory activity.

For the QSAR analysis, models were generated from training set compounds and the predictive ability of the models was judged from the prediction of the CYP2B6 inhibition activity of the test set compounds. A comparison of statistical quality of different models is given in Table 6. The developed QSAR models indicate the importance of different *Jurs* (*Jurs*_FPSA_3, *Jurs*_DPSA_3) structural (Hbond acceptors), thermodynamic (*Molref*), topological branching index ($^0\chi^v, {}^3\chi^v_c, {}^3\kappa_a$) and E-state index for different fragments (*S_aaN*, *S_dsCH*). The properties appearing in the QSAR model such as *S_aaN* (which indicates the presence of aromatic nitrogen) and HBA (electron-rich centre) support the observation of developed pharmacophore. Though we got HBA as an important feature in the pharmacophore study, the docking study suggests that basically an electron-rich centre capable of formation of co-ordinate bond with the iron of cytochrome is important. Overall, the analysis indicates that the different features such as the presence of HBA (electron rich) groups, size (*Molref*) of the molecules, impact of branching and ring system, and distribution of charges are important and these observations are in agreement with those of the source paper [30]. All the three models have Q^2 and R^2_{pred} values greater than 0.5. The GFA-derived models are superior to the G/PLS-derived models. For CYP2B6 inhibition, the GFA model with spline option (Equation (7)) was found to be the best model based on internal validation ($Q^2 = 0.772$) whereas the best predictive model (external validation) was the GFA model with spline option (Equation 6; $R^2_{\text{pred}} = 0.843$). On the basis of $r^2_{\text{m(overall)}}$ criterion, the best model among the reported three models (Table 6) was the GFA model with spline option (Equation (7); $r^2_{\text{m(overall)}} = 0.774$).

Acknowledgements

One of the authors (PPR) thanks the University Grants Commission (UGC), New Delhi for a fellowship.

References

[1] D.R. Nelson, L. Koymans, T. Kamataki, J.J. Stegeman, R. Feyereisen, D.J. Waxman, M.R. Waterman, O. Gotoh, M.J. Coon, R.W. Estabrook, I.C. Gunsalus, and D.W. Nebert, *P450 superfamily: Update on new sequences, gene mapping, accession numbers and nomenclature*, Pharmacogenetics 6 (1996), pp. 1–42.

[2] D.R. Nelson, D.C. Zeldin, S.M. Hoffman, L.J. Maltais, H.M. Wain, and D.W. Nebert, *Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants*, Pharmacogenetics 14 (2004), pp. 1–18.

[3] D.R. Nelson, *Gene nomenclature by default, or BLASTing to babel*, Hum. Genomics 2 (2005), pp. 196–201.

[4] S. Ekins, M. Vandenbranden, B.J. Ring, J.S. Gillespie, T.J. Yang, H.V. Gelboin, and S.A. Wrighton, *Further characterization of the expression in liver and catalytic activity of CYP2B6*, J. Pharmacol. Exp. Ther. 286 (1998), pp. 1253–1259.

[5] L. Gervot, B. Rochat, J.C. Gautier, F. Bohnenstengel, H. Kroemer, V. de Berardinis, H. Martin, P. Beaune, and I. de Waziers, *Human CYP2B6: Expression, inducibility and catalytic activities*, Pharmacogenetics 9 (1999), pp. 295–306.

[6] T. Shimada, H. Yamazaki, M. Mimura, Y. Inui, and F.P. Guengerich, *Interindividual variations in human liver cytochrome P-450 enzymes involved in the oxidation of drugs, carcinogens and toxic chemicals: Studies with liver microsomes of 30 Japanese and 30 caucasians*, J. Pharmacol. Exp. Ther. 270 (1994), pp. 414–423.

[7] M. Mimura, T. Baba, H. Yamazaki, S. Ohmori, Y. Inui, F.J. Gonzalez, F.P. Guengerich, and T. Shimada, *Characterization of cytochrome P450 2B6 in human liver microsomes*, Drug Metab. Dispos. 21 (1993), pp. 1048–1056.

[8] D.M. Stresser and D. Kupfer, *Monospecific antipeptide antibody to cytochrome P-450 2B6*, Drug Metab. Dispos. 27 (1999), pp. 517–525.

[9] E.L. Code, C.L. Crespi, B.W. Penman, F.J. Gonzalez, T.K. Chang, and D.J. Waxman, *Human cytochrome P4502B6: Interindividual hepatic expression, substrate specificity, and role in procarcinogen activation*, Drug Metab. Dispos. 2510 (1997), pp. 985–993.

[10] C. Bathelt, R.D. Schmid, and J. Pleiss, *Regioselectivity of CYP2B6: Homology modeling, molecular dynamics simulation, docking*, J. Mol. Model 8 (2002), pp. 327–335.

[11] H. Hanna, J.R. Reed, F.P. Guengerich, and P.F. Hollenberg, *Expression of human cytochrome P450 2B6 in* E. coli*: Characterization of catalytic activity and expression levels in human liver*, Arch. Biochem. Biophys. 376 (2000), pp. 206–216.

[12] L.M. Hesse, K. Venkatakrishnan, M.H. Court, L.L. von Moltke, S.X. Duan, R.I. Shader, and D.J. Greenblatt, *CYP2B6 mediates the* in vitro *hydroxylation of bupropion: Potential drug interactions with other antidepressants*, Drug Metab. Dispos. 28 (2000), pp. 1176–1183.

[13] S. Miksys, C. Lerman, P.G. Shields, D.C. Mash, and R.F. Tyndale, *Smoking, alcoholism and genetic polymorphisms alter CYP2B6 levels in human brain*, Neuropharmacology 45 (2003), pp. 122–132.

[14] D. Nolan, E. Phillips, and S. Mallal, *Efavirenz and CYP2B6 polymorphism: Implications for drug toxicity and resistance*, Clin. Infect. Dis. 42 (2006), pp. 408–410.

[15] S.R. Faucette, H. Wang, G.A. Hamilton, S.L. Jolley, D. Gilbert, C. Lindley, B. Yan, M. Negishi, and E. LeCluyse, *Regulation of CYP2B6 in primary human hepatocytes by prototypical inducers*, Drug Metab. Dispos. 32 (2004), pp. 348–358.

[16] R.L. Walsky, A.V. Astuccio, and R.S. Obach, *Evaluation of 227 drugs for* in vitro *inhibition of cytochrome P450 2B6*, J. Clin. Pharmacol. 46 (2006), pp. 1426–1438.

[17] M. Turpeinen, H. Raunio, and O. Pelkonen, *The functional role of CYP2B6 in human drug metabolism: Substrates and inhibitors* in vitro, in vivo *and in silico*, Curr. Drug Metab. 7 (2006), pp. 705–714.

[18] U.S. Svensson and M. Ashton, *Identification of the human cytochrome P450 enzymes involved in the* in vitro *metabolism of artemisinin*, Br. J. Clin. Pharmacol. 48 (1999), pp. 528–535.

[19] M. Rotger, S. Colombo, H. Furrer, G. Bleiber, T. Buclin, B.L. Lee, O. Keiser, J. Biollaz, L. Decosterd, and A. Telenti, *Influence of CYP2B6 polymorphism on plasma and intracellular concentrations and toxicity of efavirenz and nevirapine in HIV-infected patients*, Pharmacogenet. Genomics 15 (2005), pp. 1–5.

[20] J.K. Coller, N. Krebsfaenger, K. Klein, K. Endrizzi, R. Wolbold, T. Lang, A. Nussler, P. Neuhaus, U.M. Zanger, M. Eichelbaum, and T.E. Murdter, *The influence of CYP2B6, CYP2C9 and CYP2D6 genotypes on the formation of the potent antioestrogen Z-4-hydroxytamoxifen in human liver*, Br. J. Clin. Pharmacol. 54 (2002), pp. 157–167.

[21] T.K. Chang, G.F. Weber, C.L. Crespi, and D.J. Waxman, *Differential activation of cyclophosphamide and ifosphamide by cytochromes P-450 2B and 3A in human liver microsomes*, Cancer Res. 53 (1993), pp. 5629–5637.

[22] S.L. Mo, Y.H. Liu, W. Duan, M.Q. Wei, J.R. Kanwar, and S.F. Zhou, *Substrate specificity, regulation, and polymorphism of human cytochrome P450 2B6*, Curr. Drug Metab. (2009).

[23] H. Wang and L.M. Tompkins, *CYP2B6: New insights into a historically overlooked cytochrome P450 isozyme*, Curr. Drug Metab. 9 (2008), pp. 598–610.

[24] E.D. Kharasch, D. Mitchell, and R. Coles, *Stereoselective bupropion hydroxylation as an* in vivo *phenotypic probe for cytochrome P4502B6 (CYP2B6) activity*, J. Clin. Pharmacol. 48 (2008), pp. 464–474.

[25] R.L. Walsky, A.V. Astuccio, and R.S. Obach, *Evaluation of 227 drugs for* in vitro *inhibition of cytochrome P450 2B6*, J. Clin. Pharmacol. 46 (2006), pp. 1426–1438.

[26] J. Kumagai, T. Fujimura, S. Takahashi, T. Urano, T. Ogushi, K. Horie-Inoue, Y. Ouchi, T. Kitamura, M. Muramatsu, B. Blumberg, and S. Inoue, *Cytochrome P450 2B6 is a growth-inhibitory and prognostic factor for prostate cancer*, Prostate 67 (2007), pp. 1029–1037.

[27] S. Ekins, G. Bravi, B.J. Ring, T.A. Gillespie, J.S. Gillespie, M. Vandenbranden, S.A. Wrighton, and J.H. Wikel, *Three-dimensional quantitative structure–activity relationship analyses of substrates for CYP2B6*, J. Pharmacol. Exp. Ther. 288 (1999), pp. 21–29.

[28] Q. Wang and J.R. Halpert, *Combined 3D quantitative structure–activity relationship analysis of cytochrome P450 2B6 substrates and protein homology modeling*, Drug Metab. Dispos. 30 (2002), pp. 86–95.

[29] D.F. Lewis, B.G. Lake, Y. Ito, and P. Anzenbacher, *Quantitative structure–activity relationships (QSARs) within cytochromes P450 2B (CYP2B) subfamily enzymes: The importance of lipophilicity for binding and metabolism*, Drug Metabol. Drug Interact. 21 (2006), pp. 213–231.

[30] L.E. Korhonen, M. Turpeinen, M. Rahnasto, C. Wittekindt, A. Poso, O. Pelkonen, H. Raunio, and R.O. Juvonen, *New potent and*

selective cytochrome P450 2B6 (CYP2B6) inhibitors based on Three-dimensional and quantitative structure–activity relationship (3D-QSAR) analysis, Br. J. Pharmacol. 150 (2007), pp. 932–942.

[31] S.C. Gay, M.B. Shah, J.C. Talakad, K. Maekawa, A.G. Roberts, P.R. Wilderman, L. Sun, J.Y. Yang, S.C. Huelga, W.X. Hong, Q. Zhang, C.D. Stout, and J.R. Halpert, *Crystal structure of a cytochrome P450 2B6 genetic variant in complex with the inhibitor 4-(4-Chlorophenyl)imidazole at 2.0Å resolution*, Mol. Pharmacol. 77 (2010), pp. 529–538.

[32] Discovery Studio 2.1 is a product of Accelrys Inc, San Diego, CA, USA.

[33] Y. Kurogi and O.F. Guner, *Pharmacophore modeling and Three-dimensional database searching for drug design using catalyst*, Curr. Med. Chem. 8 (2001), pp. 1035–1055.

[34] A.K. Debnath, *Pharmacophore mapping of a series of 2,4-diamino-5-deazapteridine inhibitors of mycobacterium avium complex dihydrofolate reductase*, J. Med. Chem. 45 (2002), pp. 41–53.

[35] P. Kahnberg, M.H. Howard, T. Liljefors, M. Nielsen, E.O. Nielsen, O. Sterner, and I. Pettersson, *The use of a pharmacophore model for identification of novel ligands for the benzodiazepine binding site of the GABAA receptor*, J Mol Graph Model 23 (2004), pp. 253–261.

[36] J. Faragalla, J. Bremner, D. Brown, R. Griffith, and A. Heaton, *Comparative pharmacophore development for inhibitors of human and rat 5-α-reductase*, J. Mol. Graph. Model 22 (2003), pp. 83–92.

[37] S. Ekins, G. Bravi, J.H. Wikel, and S.A. Wrighton, *Three-dimensional-quantitative structure–activity relationship analysis of cytochrome P-450 3A4 substrates*, J. Pharmacol. Exp. Ther. 291 (1999), pp. 424–433.

[38] A. Smellie, S.L. Teig, and P. Towbin, *Poling: Promoting conformational variation*, J. Comp. Chem. 16 (1995), pp. 171–187.

[39] R. Kristam, V.J. Gillet, R.A. Lewis, and D. Thorner, *Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst*, J. Chem. Inf. Model. 45 (2005), pp. 461–476.

[40] J. Li, T. Ehlers, J. Sutter, S. Varma-O'Brien, and J. Kirchmair, *CAESAR: A new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration*, J. Chem. Inf. Model. 47 (2007), pp. 1923–1932.

[41] Y.H. Hung and Y.S. Liao, *Applying PCA and fixed size LS-SVM method for large scale classification problems*, Inf. Technol. J. 7 (2008), pp. 890–896.

[42] T. Fawcett, *An introduction to ROC analysis*, Pattern Recognit. Lett. 27 (2006), pp. 861–874.

[43] Cerius2 version 4.10 is a product of Accelrys, Inc., San Diego, USA.

[44] J.T. Leonard and K. Roy, *On selection of training and test sets for the development of predictive QSAR models*, QSAR Comb. Sci. 25 (2006), pp. 235–251.

[45] K. Roy and A.S. Mandal, *Development of linear and nonlinear predictive QSAR models and their external validation using molecular similarity principle for anti-HIV indolyl aryl sulfones*, J. Enzyme Inhib. Med. Chem. 23 (2008), pp. 980–995.

[46] Y. Fan, L.M. Shi, K.W. Kohn, Y. Pommier, and J.N. Weinstein, *Quantitative structure-antitumor activity relationships of camptothecinanalogues: Cluster analysis and genetic algorithm-based studies*, J. Med. Chem. 44 (2001), pp. 3254–3263.

[47] D. Rogers and A.J. Hopfinger, *Application of genetic function approximation to quantitative structure–activity relationship and quantitative structure–property relationship*, J. Chem. Inf. Comput. Sci. 34 (1994), pp. 854–866.

[48] W.J. Dunn III and D. Rogers, *Genetic Partial Least Squares in QSAR,* in *Genetic Algorithms in Molecular Modeling*, J. Devillers, ed., Academic Press, London, 1996, pp. 109–130.

[49] K. Hasegawa, Y. Miyashita, and K. Funatsu, *GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists*, J. Chem. Inf. Comput. Sci. 37 (1997), pp. 306–310.

[50] G.W. Snedecor and W.G. Cochran, *Statistical Methods*, Oxford & IBH Publishing, New Delhi, 1967.

[51] S. Wold, *PLS for Multivariate Linear Modeling,* in *Chemometric methods in molecular design*, H. van de Waterbeemd, ed., VCH, Weinheim, 1995, pp. 195–218.

[52] A.K. Debnath, *Quantitative structure–activity relationship (QSAR): A versatile tool in drug design,* in *Combinatorial Library Design and Evaluation: Principles, Software Tools, and Applications in Drug Discovery*, A.K. Ghose and V.N. Viswanadhan, eds., Marcel Dekker, New York, 2001, pp. 73–129.

[53] K. Roy, *On some aspects of validation of predictive QSAR models*, Expert. Opin. Drug. Discov. 2 (2007), pp. 1567–1577.

[54] P.P. Roy and K. Roy, *On some aspects of variable selection for partial least squares regression models*, QSAR Comb. Sci. 27 (2008), pp. 302–313.

[55] K. Roy and P.P. Roy, *Comparative QSAR studies of CYP1A2 inhibitor flavonoids using 2D and 3D descriptors,* Chem. Biol. Drug Des. 72 (2008), pp. 370–382.

[56] I. Mitra, P.P. Roy, S. Kar, P. Ojha, and K. Roy, *On further application of* $\mathrm{r}_{\mathrm{m}}^2$ *as a metric for validation of QSAR models*, J. Chemometrics 24 (2010), pp. 22–33.

[57] K. Roy and S. Paul, *Exploring 2D and 3D QSARs of 2, 4-diphenyl-1, 3-oxazolines for ovicidal activity against tetranychus urticae*, QSAR Comb. Sci. 28(4) (2008), pp. 406–425.

[58] S. Ekins, M. Iyer, M.D. Krasowski, and E.D. Kharasch, *Molecular characterization of cyp2b6 substrates*, Curr. Drug Metab. 9 (2008), pp. 363–373.

[59] P.P. Roy, J.T. Leonard, and K. Roy, *Exploring the impact of the size of training sets for the development of predictive QSAR models*, Chemom. Intell. Lab. Sys. 90 (2008), pp. 31–42.

[60] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, and P. Gramatica, *Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs*, Environ. Health Perspect. 111 (2003), pp. 1361–1375.

[61] D. Itokawa, T. Nishioka, J. Fukushima, T. Yasuda, A. Yamauchi, and H. Chuman, *Quantitative structure–activity relationship study of binding affinity of azole compounds with CYP2B and CYP3A*, QSAR Comb. Sci. 26 (2007), pp. 828–836.