European Journal of Medicinal Chemistry 44 (2009) 1941–1951

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Exploring QSAR and QAAR for inhibitors of cytochrome P450 2A6 and 2A5 enzymes using GFA and G/PLS techniques

Kunal Roy*, Partha Pratim Roy

Drug Theoretics and Cheminformatics Lab, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Raja SC Mullick Road, Kolkata 700 032, West Bengal, India

ARTICLE INFO

Article history:
Received 24 September 2008
Received in revised form
5 November 2008
Accepted 19 November 2008
Available online 3 December 2008

Keywords:
QSAR
QAAR
Cytochrome P450 enzymes
GFA
G/PLS

ABSTRACT

A series of naphthalene and non-naphthalene derivatives ($n = 42$) having cytochrome P450 2A6 and 2A5 inhibitory activities, reported by Rahnasto et al., were subjected to QSAR and QAAR studies. The analyses were performed using electronic, spatial, shape and thermodynamic descriptors to develop quantitative models for prediction of the inhibitory activities and to explore importance of different descriptors for the responses. The data set was divided into training and test sets (with test set size being approximately 25% of the full data set size) based on *K*-means clustering applied on the standardized descriptor matrix. Genetic function approximation (GFA) and genetic partial least-squares (G/PLS) were used as chemometric tools for modeling, and the derived equations were of acceptable statistical and prediction (both internal and external) qualities although different equations varied in quality in a wide range (R^2: 0.561–0.898, R_a^2: 0.508–0.870, Q^2: 0.495–0.814, R_{pred}^2: 0.615–0.914, r^2: 0.679–0.905, r_0^2: 0.639–0.904, r_m^2: 0.494–0.876). In the case of CYP2A5 inhibition, the GFA derived QSAR model is better than the G/PLS derived model considering both internal and external validations. In the case of CYP2A6 inhibitory potency data, the GFA derived QSAR model is better than the G/PLS model considering internal validation whereas the latter is better in external validation (which is more important) than the former. The model development process was subjected to randomization test at 90% confidence level by taking into account the whole pool of descriptors, while the developed models were also subjected to randomization test (99% confidence level) for validation. Based on the randomization test results, GFA models are found to be superior to the G/PLS models. Among the parameters, which were found important in modeling both the responses, were different Jurs descriptors, electronic descriptors (like Sr, Apol), steric descriptors (like shadow indices, Molref), shape descriptors (like COSV, Fo) and lipophilicity descriptors. This indicates that the CYP2A5 and CYP2A6 inhibition of these compounds is related to charge distribution, surface area, electronic, hydrophobic and spatial properties of the molecules.

1. Introduction

Cytochrome P450 (CYP) enzymes are predominantly involved in Phase 1 metabolism of xenobiotics. CYP is a form of large superfamily of heme enzymes. There are over 50 mammalian cytochrome P450 genes in at least 17 families. Microsomal CYP enzymes catalyze the specific steps in the biosynthesis of steroid hormones, cholesterol, prostanoids and bile acids, participate in the catabolism of endogenous compounds, including fatty acids and steroids, and are involved in the degradation of exogenous compounds, including a wide variety of structurally diverse drugs and carcinogens [1–5]. One such enzyme is cytochrome P450 2A6 (CYP2A6), first identified as the human coumarin 7-hydroxylase [6–9]. Coumarin is specifically hydroxylated to 7-hydroxycoumarin in mice by CYP2A5 and in humans by CYP2A6. These enzymes share an 82% similarity in their amino acid sequences. Most of the known CYP2A5 ligands include a lactone moiety [10]. The active site of P450 2A6 is six times smaller than that of human P450 2C8 [11,12], even smaller than that of bacterial P450 101A1 [13]. It was reported by Poso et al. that the 2A5 binding site is larger than that of 2A6 [14].

Crystal structures of CYP2A6 in complexes with coumarin (substrate) and methoxsalen (inhibitor) have recently been published [11,15]. The characteristics of its substrates somewhat resemble CYP1A2 substrates, small planar molecules. The structures of the complexes indicate that Asn297, the only polar residue in the active site, is the hydrogen-bonding residue in the

* Corresponding author. Tel.: +91 98315 94140; fax: +91 33 2837 1078.
E-mail address: kunalroy_in@yahoo.com (K. Roy).
URL: http://www.geocities.com/kunalroy_in

0223-5234/$ – see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2008.11.010

active site interacting with carbonyl oxygen of the ligands. The P450 2A6 structures were solved with the alternative substrate, coumarin, and with the inhibitor, methoxsalen, bound in the active site of the enzyme, adjacent to its iron containing heme group. The compact, hydrophobic active site contains one hydrogen bond donor, Asn297, which orients coumarin for regioselective oxidation [11]. The human cytochrome P450 2A6 is principally involved in the break down of nicotine in the bloodstream as it circulates through the liver. Oxidation of nicotine by this P450 not only leads to detoxication, but the enzyme also activates tobacco-specific procarcinogens to mutagenic products [1,16]. In addition to nicotine, the human CYP2A6 and mouse cytochrome P450 2A5 (CYP2A5) enzymes metabolize several other xenobiotics, including indole and carcinogenic nitrosamines as well as many other tobacco-specific nitrosamines and toxic compounds [17].

Smoking has enormous negative health consequences worldwide, and the use of tobacco is still rising globally [18]. Nicotine is routinely used in smoking cessation therapy in different dosage form as chewing gum, as transdermal patches, or via inhalation [19]. Tobacco is an insidious chemical package because among its numerous toxic, mutagenic, and carcinogenic compounds it contains nicotine, which is responsible for causing the dependency associated with tobacco smoking. According to the World Health Organization (WHO), use of tobacco is responsible for causing disease burden measured in disability adjusted life years in developed countries and one of the top 10 health risk factors even in the poorest developing regions. Individuals having deficient CYP2A6 enzyme function display a decreased capacity for nicotine metabolism, and these individuals may be less likely to become smokers than those having the active enzyme functions [20–26]. An effective inhibitor of P450 2A6 could be used to diminish smoking and tobacco-related cancers.

Poso and coworkers in their recent work with a series of naphthalene (which is structurally close to coumarin) and quinoline derivatives have developed CoMFA models for coumarin 7-hydroxylation inhibition [14]. Comparative modeling of CYP2A6 has shown that inhibitors of the CYP2A enzymes are generally planar molecules with two hydrogen bond acceptors [27]. Structure–activity relationship studies on lactone and non-lactone compounds have yielded information on required structural features for interactions with the CYP2A5 and CYP2A6 enzymes [28]. Comparative quantitative structure–activity relationships (QSARs) of CYP2A5 and CYP2A6 inhibitors can unfold further information on the desired electronic, spatial and shape properties of the ligands for binding with the active site of the enzymes. In this background, we are attempting here to explore the QSAR models of series of naphthalene and non-naphthalene derivatives reported by Rahnasto et al. [18] using different electronic, spatial, shadow, shape and thermodynamic indices to find out the structural information and different properties relevant to the inhibition of CYP2A6 and CYP2A5 enzymes. Naphthalene derivatives are important for the study of CYP2A5 and CYP2A6 inhibitors as naphthalene is structurally similar to coumarin and a relatively potent inhibitor of both enzymes. The applicability of the developed models will be for naphthalene derivatives and related aromatic and heteroaromatic compounds. Attempt was also made to develop quantitative activity–activity relationship (QAAR) models [29] taking one response as the dependent variable and the other as one of the independent variables. All the QSAR and QAAR models were also validated using appropriate strategies. Though Rahnasto et al. [18] have developed CoMFA models for this series of compounds and showed color contour maps of charges for SAR, we have tried to develop statistical models (equations) that readily predict CYP2A5 and CYP2A6 inhibition potencies of the compounds.

2. Material and methods

2.1. The data set and the descriptors

The inhibitory potencies of 26 naphthalene and 16 non-naphthalene compounds (Table 1) against human CYP2A6 and mouse CYP2A5 enzymes reported by Rahnasto et al. [18] have been used as the model data set for the present study. The inhibitory potencies of the naphthalene and non-naphthalene derivatives [IC_{50} (μM)] have been converted to the logarithmic scale [pIC_{50} (mM)] and then used for subsequent QSAR analyses as the response variable. The data set includes naphthalene ($n = 26$), quinoline ($n = 7$), tetralone ($n = 3$) and six non-planar compounds listed in Fig. 1. The analyses were performed using electronic (Apol, Dipole, HOMO, LUMO and Sr), spatial (radius of gyration, Jurs descriptors, Shadow indices, Area, PMI mag, Density, Vm), shape (DIFFV, Fo, NCOSV, COSV, ShapeRMS, SrVol) and thermodynamic (A log P, A log P98, Molref) descriptors. The definitions of the descriptors are given in Table 2. All the descriptors were calculated using descriptor + module of the Cerius2 version 4.10 software [30] under QSAR + environment on a Silicon Graphics O2 workstation running under the IRIX 6.5 operating system.

Table 1
Observed values of CYP2A5 and CYP2A6 inhibition potency of naphthalene and non-naphthalene compounds.

S. No.	SMILES	pIC_{50}(2A5) obs[a]	pIC_{50}(2A6) obs[b]
Training set			
2	Cc1cccc2ccccc12	4.620	4.469
3	Clc1cccc2ccccc12	4.854	4.745
4	Oc1cccc2ccccc12	3.854	3.886
5	Cc2ccc1ccccc1c2	5.155	5.620
6	CCc2ccc1ccccc1c2	5.319	4.921
7	Fc2ccc1ccccc1c2	5.137	6.174
8	Clc2ccc1ccccc1c2	4.921	5.268
9	Brc2ccc1ccccc1c2	5.602	6.26
10	Oc2ccc1ccccc1c2	4.108	3.854
11	COc2ccc1ccccc1c2	5.337	4.208
12	Cc2ccc1ccccc1c2C	4.620	4.585
13	Clc2ccc1ccccc1c2Cl	5.260	4.824
14	Cc2cc(C)c1ccccc1c2	5.004	5.081
16	Clc1cccc2c(Cl)cccc12	5.620	5.602
17	Cc1cccc2c(C)cccc12	4.469	4.509
20	Clc2cc1ccccc1cc2Cl	4.222	–
22	Cc2ccc1ccc(C)cc1c2	5.854	5.77
24	CC(=O)c2ccc1ccc(C)cc1c2	5.387	5.174
25	Clc2cc1cc(Cl)c(Cl)c(Cl)c1cc2Cl	3.523	–
26	c1ccc3c(c1)ccc2ccccc23	4.585	4.013
27	Cc2ccc1ccccc1n2	4.097	3.721
28	Cc2cnc1ccccc1c2	4.000	3.699
30	Cc2ccc1nc(C)ccc1c2	4.398	3.553
32	Cc1nccc2ccccc12	3.770	4.222
33	Cc2cc1ccccc1cn2	5.319	4.602
34	O=C1CCCc2ccccc12	4.854	4.284
36	COc2cc1CCC(=O)Cc1cc2OC	3.678	2.699
37	Clc1ccc(Cl)cc1	6.000	4.770
38	Brc1ccc(Br)cc1	5.658	4.229
39	Clc1ccccc1c2ccccc2	4.886	4.444
41	CN1CCCC1c2cccnc2	3.796	3.237
Test set			
1	c2ccc1ccccc1c2	4.131	4.602
15	Clc1ccc(Cl)c2ccccc12	5.260	5.602
18	Cc2ccc1c(C)cccc1c2	4.495	3.886
19	Cc2ccc1cccc(C)c1c2	4.854	4.507
21	Cc2ccc1cc(C)ccc1c2	5.770	5.000
23	Clc2ccc1ccc(Cl)cc1c2	5.357	–
29	Cc2cc(C)c1ccccc1n2	3.959	3.081
31	Cc2ccc1ccc(C)nc1c2	3.745	3.398
35	O=C2CCc1ccccc1C2	5.013	4.678
40	Clc2ccc(c1ccccc1)cc2	4.886	3.824
42	CN1C(=O)CCC1c2cccnc2	2.824	1.456

[a] Observed potency for CYP2A5 inhibition.
[b] Observed potency for CYP2A6 inhibition [18].

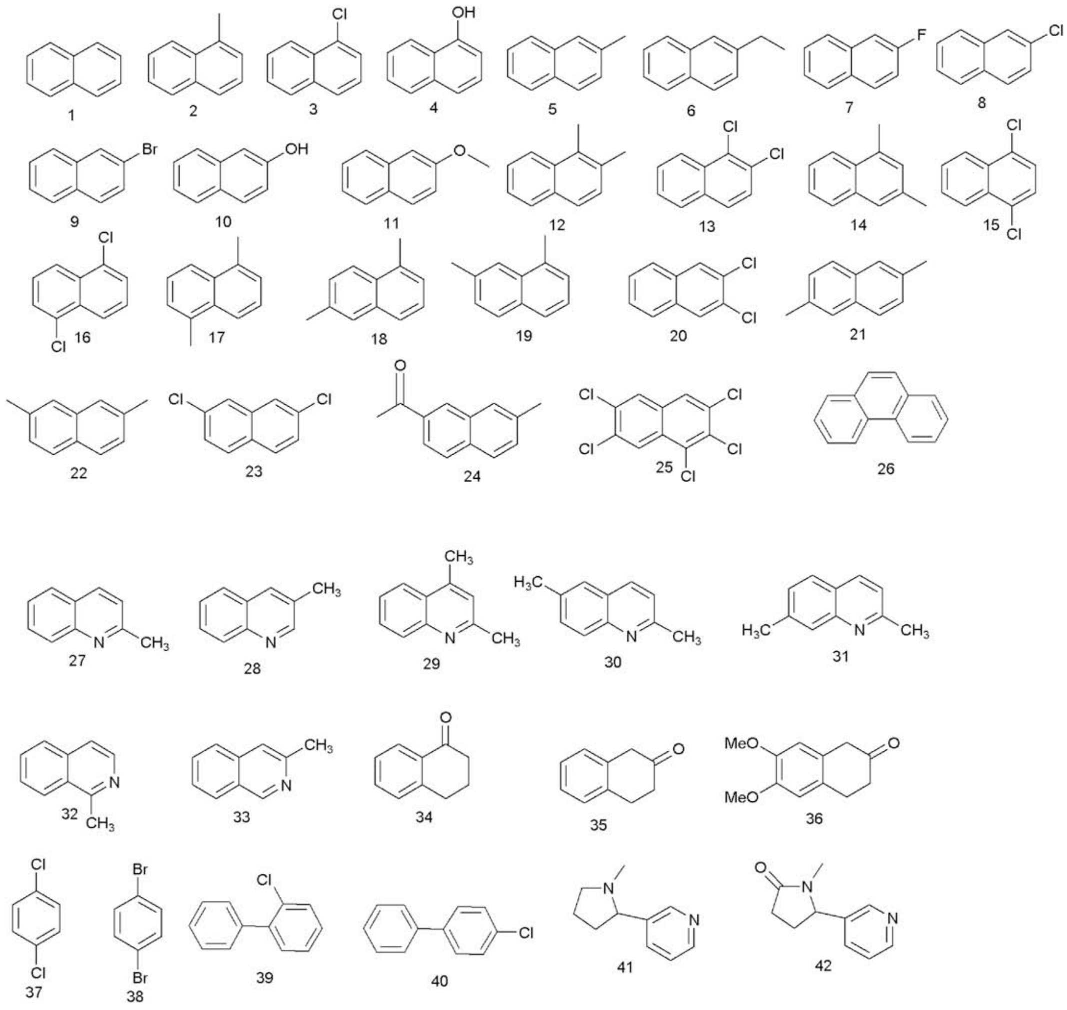

Fig. 1. Structures of naphthalene and non-naphthalene compounds (**1**–**42**).

2.2. Model development

To begin with the model development process, the data set was first classified into clusters by using *K*-means cluster based on standardized descriptor matrix (values 0 to 1) [31]. This approach (clustering) ensures that the similarity principle can be employed for the activity prediction of the test set [32]. *K*-means clustering is a non-hierarchical classification method, which expresses the final cluster membership for each case only [33]. The numbers of compounds for the training sets (in case of QSAR models) were 31 and 29 for CYP2A5 and CYP2A6, respectively, and test sets were composed of 11 and 10 compounds for CYP2A5 and CYP2A6, respectively. In addition to the development of the respective QSAR models, we have also developed QAAR models taking into account one response as the dependent variable and the other as one of the independent variables so that one inhibition potency can be calculated when the other is known. In the case of QAAR models, the numbers of compounds for the training and test sets were 29 and 10, respectively. For the development of the QSAR/QAAR models the statistical techniques used were GFA (genetic function approximation) and G/PLS (genetic partial least squares).

For the computation of shape analysis descriptors, the major steps are (1) generation of conformers and energy minimization; (2) hypothesizing an active conformer (global minimum of the most active compound, though we must acknowledge that minimum energy conformation of an isolated molecule may not be same as that of the molecule bound to the target site); (3) selecting a candidate shape reference compound (based on active conformation); (4) performing pairwise molecular superimposition using the maximum common subgroup (MCSG) method; (5) measuring molecular shape commonality using MSA descriptors; (6) determination of other molecular features by calculating spatial, electronic, and conformational parameters; (7) selection of conformers; and (8) generation of QSAR equations by genetic function algorithm (GFA). However, in the present case, most of the compounds are having rigid aromatic nucleus with limited substituents and thus the question of multiple conformations did not arise in case of many compounds. Optimal search was used as a conformational search method. Each conformer was subjected to an energy minimization procedure using a smart minimizer under the open force field (OFF) to generate the lowest energy conformation for each structure. The conformer of the most active compound (compounds **37** and **9** in the case of inhibitors of CYP2A5 and CYP2A6, respectively) was selected as a shape reference to which all the structures in the study compounds were aligned through pairwise superpositioning. The method used for performing the alignment was a maximum common subgroup (MCSG) [30,34]. This method looks at molecules as points and lines and uses the techniques of graph theory to identify patterns. It finds the largest subset of atoms in the shape reference compound that is shared by all the structures in the study table and uses this subset for alignment. A rigid fit of atom pairings was performed to superimpose each structure so that it overlays the

Table 2
Definition of different variables.

Type of descriptor	Descriptor name	Definition	Comment, if any
Electronic	Apol	Sum of atomic polarizabilities	
	Dipole	Dipole moment	
	HOMO	Highest occupied molecular orbital energy	It measures nucleophilicity of a molecule
	LUMO	Lowest unoccupied molecular orbital energy	It measures electrophilicity of a molecule
	Sr	Superdelocalizability	It may be used to predict relative reactivity in a series of molecules
Shape	DIFFV	The difference between the volume of the individual molecule and the volume of the shape reference compound	
	COSV	The common volume between each individual molecule and the molecule selected as the reference compound	It measures how similar in steric shape the analogs are to the shape reference
	Fo	The common overlap steric volume descriptor (COSV, see above) divided by the volume of the individual molecule	
	NCOSV	The difference between the volume of the individual molecule and the common overlap steric volume (COSV)	
	ShapeRMS	Root mean square (RMS) deviation between the individual molecule and the shape reference compound	
	SRVolume	The volume of the shape reference compound	
Spatial	RadOfGyration	$\sqrt{\frac{(x_i^2+y_i^2+z_i^2)}{N}}$	*N* is the number of atoms and *x*, *y*, *z* are the atomic coordinates relative to the center of mass
	Density	The ratio of molecular weight to molecular volume	It reflects the types of atoms and how tightly they are packed in a molecule
	PMI-mag	It calculates the principal moments of inertia about the principal axes of a molecule	
	Vm	Molecular volume inside the contact surface	
	Area	van der Waals area of a molecule	
	Jurs descriptors	These are calculated by mapping atomic partial charges on solvent-accessible surface areas of individual atoms	This set of descriptors combines shape and electronic information to characterize the molecules
	Shadow indices	This set of geometric descriptors helps to characterize the shape of the molecules	These are calculated by projecting the molecular surface on three mutually perpendicular planes, *XY*, *YZ*, and *XZ*
Thermodynamic	A log P	Log of the partition coefficient	A.K. Ghose, G.M. Crippen, J. Comput. Chem. 1986, 7, 565–77
	Molref	Molar refractivity	A.K. Ghose, G.M. Crippen, J. Comput. Chem. 1986, 7, 565–77
	A log P98	Log of partition coefficient	A. Ghose, V.N. Viswanadhan, J.J. Wendoloski, J. Phys. Chem., 1998, 102, 3762–72

shape reference compound. Finally additional electronic, spatial and thermodynamic descriptors were also calculated.

Genetic function approximation (GFA) technique [35,36] 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 (MARS) 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 penalizes models with too many features. It also inspires the use of splines as a powerful tool for non-linear modeling. 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 genetic partial least squares (G/PLS) algorithm [37,38] may be used as an alternative to a GFA calculation. G/PLS is derived from two QSAR calculation methods: GFA and partial least squares (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.

2.3. Statistical qualities and model validation

The statistical qualities of the equations were judged by the parameters such as *explained variance* (R_a^2), *squared correlation coefficient* (R^2) and *variance ratio* (F) at specified *degrees of freedom* (*df*) [39]. For G/PLS equations, least-squares error (LSE) was taken as a statistical measure, while lack-of-fit (LOF) was noted for the GFA derived equations. After the development of the models, the generated QSAR equations were validated by leave-one-out *cross-validation* R^2 (Q^2) and *predicted residual sum of squares* (*PRESS*) [40–42] and then were used for the prediction of enzyme inhibition potency values of the test set compounds and the prediction qualities of the models were judged by statistical parameters like predictive R^2 (R_{pred}^2), squared correlation coefficient between observed and predicted values with (r^2) and without (r_0^2) intercept. The predictive R^2 value is calculated as follows:

$$R_{\text{pred}}^2 = 1 - \frac{\sum\left(Y_{\text{pred(Test)}} - Y_{\text{(Test)}}\right)^2}{\sum\left(Y_{\text{(Test)}} - \overline{Y}_{\text{Training}}\right)^2}$$

In the above equation, $Y_{\text{pred (Test)}}$ and $Y_{\text{(Test)}}$ indicate predicted and observed activity values, respectively, of the test set compounds and $\overline{Y}_{\text{Training}}$ indicates mean activity value of the training set. It was previously shown that the use of R_{pred}^2 and r^2 might not be sufficient to indicate the external validation characteristics [43]. Thus an additional parameter r_{m}^2 (defined as $r^2 \times (1 - \sqrt{r^2 - r_0^2}))$ [43], which penalizes a model for large differences between the observed and the predicted values, was also calculated. Finally the developed models were subjected to a randomization test for validation purpose. The standard errors for the model fit (S_{fit}), crossvalidation (S_{CV}) and prediction (S_{Pred})

were also reported. As an additional tool for validation, randomization test was applied on the model development process.

2.4. Softwares

The calculation of descriptors and genetic analysis (GFA and G/PLS) were performed using Cerius2 version 4.10 software [30], while cluster analysis was performed using SPSS software [44].

3. Results and discussion

3.1. QSAR

3.1.1. Modeling CYP2A5 inhibitory potency

The view of the aligned training set molecules is shown in Fig. 2. The values of the important descriptors used in the derived equations are given in Tables A1 and A2 of Supplementary material. Though, the development of models was tried at different levels of iterations, models corresponding to the best iteration levels are reported. In general, G/PLS equations require less number of iterations to attain predictive models than GFA equations. In each case, 100 equations were obtained from a single GFA or G/PLS run ranked according to lack-of-fit (for GFA) or least-squares error (for G/PLS) score. The equations show multiple occurrences of Jurs descriptors [45] suggesting importance of charge distribution and surface areas.

Eqs. (1) and (2) were among the best ones obtained from the genetic function approximation (50 000 iterations, no fixed length of the equations) and genetic partial least squares (5000 crossovers, linear terms, scaled variables, and other default settings), respectively.

3.1.1.1. GFA.

$$\begin{aligned}\mathrm{pIC}_{50(2A5)} = {} & 2.682(\pm 0.858) - 0.316(\pm 0.134)\mathrm{Sr} \\ & + 0.728(\pm 0.105)\mathrm{Shadow\_Xlength} \\ & - 7.569(\pm 0.915)\mathrm{FPSA\_2} \\ & - 0.067(\pm 0.010)\mathrm{WNSA\_1} \\ & - 2.497(\pm 0.619)\mathrm{RNCG} - 0.141(\pm 0.030)\mathrm{PNSA\_3}\end{aligned} \tag{1}$$

Fig. 2. Aligned geometry of the training set molecules ($n = 31$, CYP2A5 inhibitory potency as the response variable).

$$\begin{aligned} n_{\mathrm{Training}} & = 31,\ \mathrm{LOF} = 0.253,\ R^2 = 0.796,\ R_a^2 = 0.745, \\ F & = 15.64(\mathrm{df}\ 6, 24),\ Q^2 = 0.686,\ \mathrm{PRESS} = 4.533, \\ n_{\mathrm{Test}} & = 11,\ R_{\mathrm{pred}}^2 = 0.707,\ r^2 = 0.734,\ r_0^2 = 0.732, \\ r_{\mathrm{m}}^2 & = 0.701 \end{aligned}$$

The standard errors of regression coefficient are given within parentheses. The relative importance of the variables appearing in the GFA equation (based on their standardized regression coefficients) is in the following order: Jurs_WNSA_1 > Jurs_FPSA_2 > Jurs_PNSA-3 > Shadow_Xlength > Jurs_RNCG > Sr. Jurs_WNSA_1 is defined as the surface weighted charged partial negative surface area as derived from the following equation:

$$\mathrm{WNSA\_1} = \frac{\mathrm{PNSA}_1 \cdot \mathrm{SASA}}{1000}$$

In the above equation, PNSA_1 is the sum of the solvent-accessible surface areas of all negatively charged atoms ($\mathrm{PNSA}_1 = \sum_{a^-} \mathrm{SA}_a^-$).

The negative coefficient of WNSA_1 indicates that compounds with high values of WNSA_1 (like compounds **20**and **25**) have lower CYP2A5 inhibitory activity than compounds with higher values of WNSA_1 (like compounds **33**and **34**).

Jurs_FPSA_2 (fractional charged partial positive surface area) is obtained by dividing total charge weighted positive surface area (PPSA_2) with the total molecular solvent-accessible surface area (SASA) as follows:

$$\mathrm{FPSA\_2} = \frac{\mathrm{PPSA}_2}{\mathrm{SASA}}$$

In the above equation, PPSA_2 is the partial positive solvent-accessible surface area multiplied by the total positive charge Q^+ ($\mathrm{PPSA}_2 = Q^+ \cdot \sum_{a^+} \mathrm{SA}_a^+$).

Compounds with higher Jurs_FPSA_2 values have lower CYP2A5 inhibitory potency. Compounds **37** and **38** have small values of FPSA_2 and relatively high values of CYP2A5 binding affinity. Compound **36**, a tetralone compound having a high value of Jurs_FPSA_2 has less CYP2A5 binding affinity.

The negative coefficient of Jurs PNSA_3 indicating atomic charge weighted negative surface area (PNSA) [which is the sum of the product of solvent-accessible surface area multiplied by partial charge for all negatively charged atoms ($\mathrm{PNSA}_3 = \sum_{a^-} q_a^- \cdot \mathrm{SA}_a^-$)] is detrimental for the inhibitory potency. Compounds **25** and **36** with lower values of PNSA_3 have poor CYP2A5 inhibitory activity due to high values of WNSA_1 and FPSA_2 parameters which contribute negatively towards the inhibitory activity.

The positive coefficient of Shadow_Xlength indicates that the increase in the length of the molecule in X (Lx) dimension (for example, compound **24**, a naphthalene derivative with –CH_3 group at 2 position and a –COCH_3 group at 7 position) is conducive for inhibitory potency.

The negative coefficient of JursRNCG indicates that the relative negative charge (RNCG), which is defined as the charge of the most negative atom divided by the total negative charge, is detrimental for the inhibitory potency. RNCG is derived from the following equation:

$$\mathrm{RNCG} = \frac{Q_{\max}^-}{Q^-}$$

$Q_{\max}^-$ = charge of the most negative atom
Q^- = total negative charge

Compounds with lower values of RNCG (like compounds **22** and **26**) have higher inhibitory activity than compounds with high values of RNCG (like compound **10**). Compound **11** having a high value of RNCG shows good CYP2A5 inhibitory activity because of low Sr value as well as high Shadow_Xlength value.

The negative coefficient of Sr (superdelocalizability) in Eq. (1) indicates that the increase in electrophilic property is detrimental for the CYP2A5 inhibitory potency. A low value of Sr (like compounds **11** and **34**) favors the binding affinity and a high value of the parameter (like compounds **27** and**28**) reduces the binding affinity.

Eq. (1) could explain 74.5% of the variance (adjusted coefficient of variation) while it could predict 68.6% of the variance (leave-one-out predicted variance). The difference between R^2 and Q^2 values is not very high (less than 0.3) [46]. When the equation was used to predict the CYP2A5 inhibition potency of the test set compounds, the predicted R^2 (R^2_{pred}) value was found to be 0.707. Simple r^2 (squared correlation coefficient) between the observed and predicted values of the test set compounds was 0.734. This value was not significantly changed when the intercept was set to zero ($r_0^2 = 0.732$). This indicates that there are no significant numerical differences between the observed and the predicted values of the test set compounds. Thus, r_m^2 value, calculated according to Ref. [43] is also acceptable ($r_m^2 = 0.701$). The test set size was about 25% of the full data set size [47]. The standard errors of model fit, cross-validation and test set prediction for Eq. (1) are given in Table 3. The intercorrelation matrix for Eq. (1) is given in Table 4.

3.1.1.2. G/PLS.

$$\begin{aligned} pIC_{50(2A5)} &= 2.225 - 0.780RNCG + 0.647Shadow\_Xlength \\ &\quad - 0.0003Apol - 0.439Sr - 51.603FPSA\_3 \\ &\quad - 0.005PMI + 0.014RPCS + 3.131FNSA\_1 \end{aligned} \tag{2}$$

$$\begin{aligned} n_{\text{Training}} &= 31,\ \text{LSE} = 0.091,\ R^2 = 0.779,\ R_a^2 = 0.726, \\ F &= 22.88(\text{df } 4, 26),\ Q^2 = 0.575,\ \text{PRESS} = 6.140, \\ n_{\text{Test}} &= 11,\ R^2_{\text{pred}} = 0.615,\ r^2 = 0.679,\ r_0^2 = 0.639, \\ r_m^2 &= 0.543 \end{aligned}$$

According to the standardized values of the regression coefficients, the relative importance of the variables in the G/PLS equation is in the following order: Shadow_Xlength > PMI_mag > Jurs_FNSA-1 > Apol > Jurs_FPSA_3 > Sr > Jurs_RNCG > Jurs_RPCS. Increase in the length of the molecule in the X dimension (Lx) is conducive (e.g., in case of 2-acetyl-7-methylnaphthalene) for the inhibitory potency. Atomic polarizability (Apol), superdelocalizability (Sr) (electrophlic property) and principal moments of inertia about the principal axes of a molecule (PMI) are detrimental for the inhibitory potency. The positive coefficient of the Jurs FNSA_1 indicates that compounds with higher FNSA_1 values (like polychlorinated naphthalenes) are more active than the alkyl substituted naphthalenes having lower Jurs FNSA_1 values. Compounds with higher values of FNSA_1 (like compounds **37** and **38**) have higher inhibitory activity than compounds with lower values of FNSA_1 (like compounds **36**and **41**). Compound **25** shows poor CYP2A5 inhibitory activity in spite of high FNSA_1 and low FPSA_3 values because of high PMI_mag and Apol values. The FNSA_1 values are obtained by dividing the product of partial negative solvent-accessible surface area and the total negative charge by the total molecular solvent-accessible surface area as shown in the following equation:

Table 3
Comparison of standard errors of model fit, crossvalidation and test set prediction (S_{fit}, S_{CV} and S_{Pred}) for different models.

Eq. No.	Statistical technique	S_{fit}	S_{CV}	S_{Pred}
(1)	GFA	0.313	0.389	0.474
(2)	G/PLS	0.452	0.569	0.542
(3)	GFA	0.530	0.596	0.758
(4)	G/PLS	0.560	0.603	0.702
(5)	GFA	0.213	0.288	0.271
(6)	G/PLS	0.250	0.250	0.289
(7)	GFA	0.417	0.501	0.476
(8)	G/PLS	0.479	0.528	0.479

Table 4
Intercorrelation (r) matrix for Eq. (1).

	Sr	Shadow_Xlength	FPSA_2	WNSA_1	RNCG	PNSA_3
Sr	1.000	0.219	−0.219	0.242	−0.265	0.034
Shadow_Xlength	0.219	1.000	0.197	0.267	−0.147	−0.126
FPSA_2	−0.219	0.197	1.000	−0.608	0.450	−0.017
WNSA_1	0.242	0.267	−0.608	1.000	−0.390	−0.622
RNCG	−0.265	−0.147	0.450	−0.390	1.000	−0.220
PNSA_3	0.034	−0.126	−0.017	−0.622	−0.220	1.000

$$FNSA\_1 = \frac{PNSA_1}{SASA},$$

where $PNSA_1$ is the sum of the solvent-accessible surface areas of all negatively charged atoms.

Jurs_FPSA_3 is derived from the following equation:

$$FPSA\_3 = \frac{PPSA_3}{SASA},$$

where $PPSA_3$ (atomic charge weighted positive surface area) is the sum of the products of atomic solvent-accessible surface areas and partial charges q_a^+ over all positively charge atoms ($PPSA_3 = \sum_{a^+} q_a^+ \cdot SA_a^+$).

Compounds with higher Jurs_FPSA_3 (fractional charged partial positive surface area) values (like 2-hydroxynaphthalene) have less inhibitory potency than those with lower JursFPSA_3 values (polychlorinated naphthalenes). Compounds with low values of Jurs_FPSA_3 (like compounds **37**and **38**) show good CYP2A5 inhibitory activity. Compound **25** shows poor inhibitory activity in spite of low FPSA_3 value as mentioned earlier.

The negative coefficient of Jurs_RNCG in Eq. (2) indicates that the relative negative charge (RNCG) (the charge of the most negative atom divided by the total negative charge) is detrimental for the inhibitory potency. The compounds with a high RNCG value (e.g., 2-hydroxynaphthalene) are less active than chloro-substituted naphthalene congeners. JursRPCS (relative positive charge) is the solvent-accessible surface area of the most positive charge divided by the relative positive charge (RPCG), i.e., RPCS $= SA^+_{\max}/\text{RPCG}$, where RPCG is the partial charge of the most positive atom divided by the total positive charge (RPCG $= Q^+_{\max}/Q^+$).

The positive coefficient of JursRPCS indicates that the relative positive charge (RPCS) is conducive (e.g., 1,4-dichlorobenzene is more active than *ortho*-chloro-substituted biphenyl) for inhibitory activity. Compounds **37**and **38** with high values of the parameter have good inhibitory activities than compound **36** (with low RPCS value). Due to high RNCG values, compounds **4** and **10** show poor inhibitory activity although they have high RPCS values.

Eq. (2) could explain and predict 72.6% and 57.5%, respectively, of the variance. The predictive R^2 value for the test set was found to be 0.615. The squared correlation coefficient (r^2) between the observed and predicted values of the test set compounds with (r^2) and without (r_0^2) intercept are 0.679 and 0.639, respectively. This difference indicates that there are some differences between the observed and predicted values as reflected in the lower r_m^2 value (0.543). The standard errors of model fit, crossvalidation and test set prediction for Eq. (2) are given in Table 3.

3.1.2. Modeling CYP2A6 inhibitory potency

The view of the aligned training set molecules is shown in Fig. 3. The values of the important descriptors used in MSA derived equations are given in Tables A1 and A2 of Supplementary material. Eqs. (3) and (4) were among the best ones obtained from the genetic function approximation (10 000 iterations) and genetic partial least squares (500 crossovers, linear terms, scaled variables and other default settings), respectively.

3.1.2.1. GFA.

$$\begin{aligned}\mathrm{pIC}_{50(2\mathrm{A6})} = & -0.234(\pm 1.666) - 3.380(\pm 0.847)\mathrm{FPSA\_2} \\ & + 0.036(\pm 0.010)\mathrm{COSV} \\ & + 0.082(\pm 0.041)\mathrm{Shadow\_YZ} \qquad (3)\end{aligned}$$

$$\begin{aligned}n_{\mathrm{Training}} &= 29,\ \mathrm{LOF} = 0.431,\ R^2 = 0.607,\ R_a^2 = 0.559, \\ F &= 12.84(\mathrm{df}\ 3, 25),\ Q^2 = 0.502,\ \mathrm{PRESS} = 9.950, \\ n_{\mathrm{Test}} &= 10,\ R_{\mathrm{pred}}^2 = 0.667,\ r^2 = 0.785,\ r_0^2 = 0.648, \\ r_{\mathrm{m}}^2 &= 0.494\end{aligned}$$

The standard errors of the regression coefficients are given within parentheses. The relative importance of the variables appearing in the GFA equation (based on their standardized regression coefficients) is in the following order: Jurs_F-PSA_2 > COSV > Shadow_YZ. The negative coefficient of FPSA_2 in Eq. (3) indicates that the compounds with higher JursFPSA_2 values (fractional charged partial surface area) are detrimental for inhibitory activity. A low value of FPSA_2 (like compounds **7**, **8**, **9** and **16**) favors and a high value of FPSA_2 (like compounds **36** and **41**) reduces the CYP2A6 binding affinity. Compounds **37** and **38** show poor CYP2A6 inhibitory activity because of low Shadow_YZ values in spite of low FPSA_2 value. Common overlap steric volume (COSV) is conducive for the activity, i.e., molecules which are very similar in shape to the shape reference molecule (compound **9**) have more inhibitory potency. The positive coefficient of shadow in the *YZ* plane indicates that an increase in the area of the molecular shadow in the *YZ* plane (S_{YZ}) (for example, 1,3-dimethylnaphthalene) is conducive for the inhibitory potency.

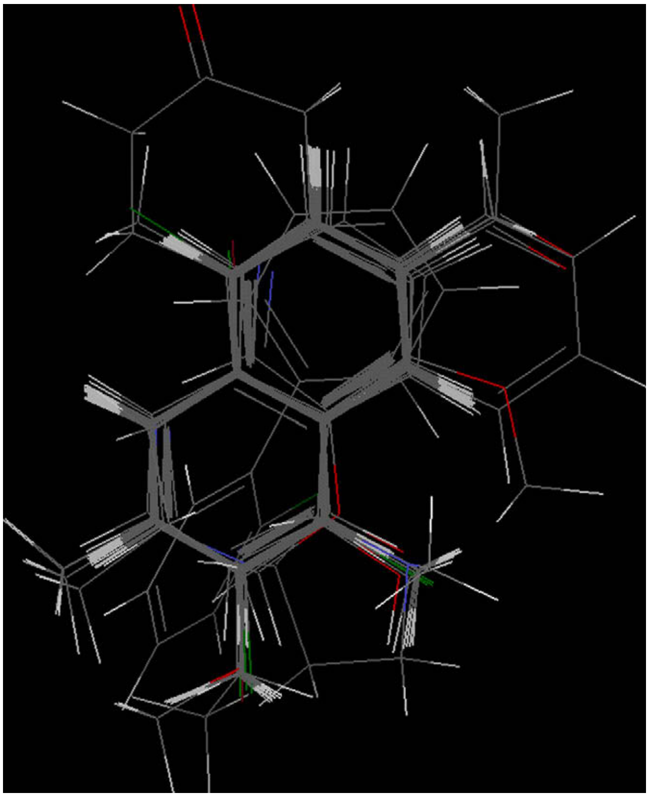

Fig. 3. Aligned geometry of the training set molecules ($n = 29$, CYP2A6 inhibitory potency as the response variable).

Eq. (3) could predict 50.2% of the variance while explaining 55.9% of the variance. The predictive R^2 value for the test set was found to be 0.667. Though the squared correlation coefficient between the observed and the predicted values of the test set compounds was quite high ($r^2 = 0.785$), the value reduces substantially when the intercept was set to zero ($r_0^2 = 0.648$). Accordingly, the value of r_{m}^2 is low ($r_{\mathrm{m}}^2 = 0.494$) and this further confirms that R_{pred}^2 and simple r^2 between the observed and the predicted values of the test set compounds cannot be taken as the only metrics for determining the quality of external validation [43]. The standard errors of model fit, crossvalidation and test set prediction for Eq. (3) are given in Table 3. The intercorrelation matrix for Eq. (3) is given in Table 5.

3.1.2.2. G/PLS.

$$\begin{aligned}\mathrm{pIC}_{50(2\mathrm{A6})} = & \ 1.698 + 0.021\mathrm{COSV} + 0.305\mathrm{A}\log\mathrm{P} \\ & - 0.004\mathrm{PPSA\_2} \qquad (4)\end{aligned}$$

$$\begin{aligned}n_{\mathrm{Training}} &= 29,\ \mathrm{LSE} = 0.282,\ R^2 = 0.561,\ R_a^2 = 0.508, \\ F &= 34.39(\mathrm{df}\ 1, 27),\ Q^2 = 0.489,\ \mathrm{PRESS} = 10.196, \\ n_{\mathrm{Test}} &= 10,\ R_{\mathrm{pred}}^2 = 0.714,\ r^2 = 0.778,\ r_0^2 = 0.721, \\ r_{\mathrm{m}}^2 &= 0.592\end{aligned}$$

Similar to Eq. (3), Eq. (4) also indicates that common overlap steric volume (COSV) is conducive for the inhibitory potency. The positive coefficient of A log P indicates that a hydrophobic group (for example, 1,2-dichloronaphthalene) is conducive for the inhibitory potency. The negative coefficient of Jurs PPSA_2 (total charge weighted positive surface area) indicates that compounds with higher PPSA_2 values (for example, compounds **36** and **41**) are less active. Compounds like **9**, **16**, **37**, and **38** have low values of PPSA_2 but compounds **37**and **38** have poorer CYP2A6 inhibitory activity than compounds **9** and **16** due to less COSV and A log P values. On the other hand, compounds **36** and **41** have high values of PPSA_2 and poor inhibitory activity.

Eq. (4) could explain 50.8% of the variance. The internal validation parameter Q^2 (0.489) of Eq. (4) is not very encouraging, but the external validation parameters are all acceptable ($R_{\mathrm{pred}}^2 = 0.714$, $r_{\mathrm{m}}^2 = 0.592$). According to the standardized values of the regression coefficients the relative importance of the variables is in the following order: A log - P > COSV > Jurs_PPSA_2. The standard errors of model fit, cross-validation and test set prediction for Eq. (4) are given in Table 3.

3.2. QAAR

In addition to the development of the respective QSAR models, we have also developed QAAR models taking one response as the

Table 5
Intercorrelation (r) matrix for Eq. (3).

	FPSA_2	COSV	Shadow_YZ
FPSA_2	1.000	−0.352	0.556
COSV	−0.352	1.000	−0.475
Shadow_YZ	0.556	−0.475	1.000

Table 6
Intercorrelation (r) matrix for Eq. (5).

	$pIC_{50(2A6)}$	Shadow_Xlength	Sr	A Log P98	Molref	FPSA_3
$pIC_{50(2A6)}$	1.000	0.004	0.133	0.596	−0.036	−0.422
Shadow_Xlength	0.004	1.000	0.293	0.182	0.712	−0.020
Sr	0.133	0.293	1.000	0.359	0.303	−0.374
A Log P98	0.596	0.182	0.359	1.000	0.330	−0.511
Molref	−0.036	0.712	0.303	0.330	1.000	0.044
FPSA_3	−0.422	−0.020	−0.374	−0.511	0.044	1.000

dependent variable and the other as one of the independent variables so that one inhibition potency can be calculated when the other is known. It is interesting that the GFA or G/PLS process automatically chooses CYP2A5 or CYP2A6 binding affinity as a variable when either of them is taken as a descriptor.

3.2.1. *CYP2A5 as response variable*

Eqs. (5) and (6) were among the best ones obtained from the genetic function approximation (50 000 iterations) and genetic partial least squares (1000 crossovers, linear terms, scaled variables and other default settings), respectively.

3.2.1.1. *GFA.*

$$\begin{aligned} pIC_{50(2A5)} = {} & 2.989(\pm 0.727) + 0.334(\pm 0.071)pIC_{50(2A6)} \\ & + 0.422(\pm 0.076)\text{Shadow\_Xlength} \\ & - 0.367(\pm 0.118)\text{Sr} - 43.219(\pm 8.814)\text{FPSA\_3} \\ & - 0.058(\pm 0.014)\text{Molref} + 0.236(\pm 0.095) \\ & \text{A log P98} \end{aligned} \tag{5}$$

$$\begin{aligned} n_{\text{Training}} &= 29,\ \text{LOF} = 0.128,\ R^2 = 0.898,\ R_a^2 = 0.870, \\ F &= 32.36(\text{df } 6,22),\ Q^2 = 0.814,\ \text{PRESS} = 2.321, \\ n_{\text{Test}} &= 10,\ R_{\text{pred}}^2 = 0.914,\ r^2 = 0.905,\ r_0^2 = 0.904, \\ r_{\text{m}}^2 &= 0.876 \end{aligned}$$

The standard errors of regression coefficients are given within parentheses. The inhibition of CYP2A5 is favored by the parameters like Shadow_Xlength (increase in length of the molecules in X dimension) and A log P98 and reduced by the parameters like Sr (superdelocalizability), FPSA_3 (fractional charged partial positive surface area) and Molref (molar refractivity). As an example, compound **38** having high inhibition for CYP2A5 has higher values of Shadow_Xlength and A log P98 and lower values of Molref, Sr and FPSA_3. The relative importance of the descriptors appearing in the GFA equation (based on their standardized regression coefficients) is in the following order: Shadow_Xlength > Molref > Jurs_FPSA_3 > $pIC_{50(2A6)}$ > A log P98 > Sr.

Eq. (5) could explain and predict 87.0% and 81.4%, respectively, of the variance. The external validation statistics of Eq. (5) are very good. The predictive R^2, r^2 and r_{m}^2 values for the test set were found to be 0.914, 0.905 and 0.876, respectively. These indicate that the predicted values are in good agreement with the observed ones. The standard errors of model fit, crossvalidation and test set prediction for Eq. (5) are given in Table 3. The intercorrelation matrix for Eq. (5) is given in Table 6.

Table 7
Intercorrelation (r) matrix for Eq. (7).

	pIC50$_{(2A5)}$	Shadow_Ylength	Density	Fo
pIC50$_{(2A5)}$	1.000	0.278	−0.075	−0.361
Shadow_Ylength	0.278	1.000	−0.294	−0.295
Density	−0.075	−0.294	1.000	0.396
Fo	−0.361	−0.295	0.396	1.000

3.2.1.2. *G/PLS.*

$$\begin{aligned} pIC_{50(2A5)} = {} & -1.994 + 0.427pIC_{50(2A6)} \\ & + 0.166\text{Shadow\_Zlength} + 3.850\text{FNSA\_1} \\ & + 0.394\text{Shadow\_Xlength} + 0.012\text{PNSA\_2} \\ & - 0.355\text{Sr} \end{aligned} \tag{6}$$

$$\begin{aligned} n_{\text{Training}} &= 29,\ \text{LSE} = 0.061,\ R^2 = 0.859,\ R_a^2 = 0.794, \\ F &= 36.67(\text{df } 4,24),\ Q^2 = 0.746,\ \text{PRESS} = 3.168, \\ n_{\text{Test}} &= 10,\ R_{\text{pred}}^2 = 0.902,\ r^2 = 0.903,\ r_2^0 = 0.881, \\ r_{\text{m}}^2 &= 0.769 \end{aligned}$$

The inhibition of CYP2A5 is favored by the parameters Shadow_Xlength, Shadow_Zlength, PNSA_2 (total charge weighted negative surface area which is obtained by multiplying partial negative solvent-accessible surface area by the total negative charge, i.e., $\text{PNSA}_2 = Q^- \cdot \sum_{a^-} \text{SA}_a^-$) and FNSA_1, and reduced by superdelocalizability (Sr).

Eq. (6) could explain and predict 79.4% and 74.6%, respectively, of the variance. The predictive R^2 and r^2 values for the test set were found to be 0.902 and 0.903, respectively. But the r_{m}^2 value was lower (0.746) which indicates that considering external validation, this G/PLS equation is less predictive than the GFA Eq. (5). According to the standardized values of the regression coefficients the relative importance of the variables is in the following order: Jurs_FNSA_1 > $pIC_{50(2A6)}$ > Shadow_Xlength > Jurs_PNSA_2 > Sr. The standard errors of model fit, crossvalidation and test set prediction for Eq. (6) are given in Table 3.

3.2.2. *CYP2A6 as response variable*

Eqs. (7) and (8) were among the best ones obtained from the genetic function approximation (50 000 iterations) and genetic partial least squares (1000 crossovers, linear terms, scaled variables and other default settings), respectively.

3.2.2.1. *GFA.*

$$\begin{aligned} pIC_{50(2A6)} = {} & -4.941(\pm 1.685) + 0.908(\pm 0.148)pIC_{50(2A5)} \\ & + 0.500(\pm 0.167)\text{Shadow\_Ylength} \\ & - 0.893(\pm 0.498)\text{Density} + 2.742(\pm 0.830)\text{Fo} \end{aligned} \tag{7}$$

$$\begin{aligned} n_{\text{Training}} &= 29,\ \text{LOF} = 0.320,\ R^2 = 0.757,\ R_a^2 = 0.687, \\ F &= 18.66(\text{df } 4,24),\ Q^2 = 0.648,\ \text{PRESS} = 7.039, \\ n_{\text{Test}} &= 10,\ R_{\text{pred}}^2 = 0.869,\ r^2 = 0.842,\ r_0^2 = 0.842, \\ r_{\text{m}}^2 &= 0.842 \end{aligned}$$

The standard errors of the regression coefficient are given within parentheses. The inhibition of CYP2A6 increases with increase in Shadow_Ylength, Density and Fo values. Shadow_Ylength indicates the length of the molecule in Y (Ly) dimension whereas Fo is the common overlap volume ratio obtained by dividing the common overlap steric volume by the volume of individual molecule.

Eq. (7) could explain and predict 68.7% and 64.8%, respectively, of the variance. The predictive R^2 value for the test set was found to be 0.869. The r_{m}^2 value was also very good (0.842) which indicates good agreement of the predicted values for the test set with the observed ones. The relative importance of the descriptors appearing in the GFA equation (based on their standardized regression coefficients) is in the following order: $pIC_{50(2A5)}$ > Shadow_Ylength > Fo > Density. The standard errors of model fit, crossvalidation and test set

Table 8
Results of randomization test of the model development process.

Eq. No.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Modeling technique	GFA	G/PLS	GFA	G/PLS	GFA	G/PLS	GFA	G/PLS
R from non-random model	0.892	0.882	0.779	0.748	0.947	0.927	0.869	0.823
Confidence level	90%	90%	90%	90%	90%	90%	90%	90%
Mean value of *R* from random trials ± standard deviation	0.371[a] ± 0.087	0.743 ± 0.068	0.539[a] ± 0.150	0.685 ± 0.077	0.569[a] ± 0.146	0.694 ± 0.057	0.519[a] ± 0.187	0.728 ± 0.069

[a] In case of the GFA models, mean values of *R* of random models are significantly lower than those of the corresponding non-random models.

Table 9
Results of randomization test of the developed models.

Eq. No.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Modeling technique	GFA	G/PLS	GFA	G/PLS	GFA	G/PLS	GFA	G/PLS
R from non-random model	0.892	0.882	0.779	0.748	0.947	0.927	0.869	0.823
Confidence level	99%	99%	99%	99%	99%	99%	99%	99%
Mean value of *R* from random trials ± standard deviation	0.423 ± 0.117	0.491 ± 0.114	0.324 ± 0.118	0.257 ± 0.109	0.457 ± 0.123	0.448 ± 0.115	0.350 ± 0.118	0.336 ± 0.116

prediction for Eq. (7) are given in Table 3. The intercorrelation matrix for Eq. (7) is given in Table 7.

3.2.2.2. G/PLS.

$$\begin{aligned}\mathrm{pIC}_{50(2A6)} &= 0.052 + 0.536\mathrm{pIC}_{50(2A5)} + 0.328\ \mathrm{A\ log\ P} \\ &\quad - 0.011\mathrm{NCOSV} + 0.158\mathrm{Shadow\_Ylength} \qquad (8)\end{aligned}$$

$$\begin{aligned} n_{\mathrm{Training}} &= 29,\ \mathrm{LSE} = 0.192,\ R^2 = 0.679,\ R_a^2 = 0.630, \\ F &= 57.07(\mathrm{df}\ 1,27),\ Q^2 = 0.610,\ \mathrm{PRESS} = 7.796, \\ n_{\mathrm{Test}} &= 10,\ R_{\mathrm{pred}}^2 = 0.867,\ r^2 = 0.883,\ r_0^2 = 0.866, \\ r_m^2 &= 0.768\end{aligned}$$

The inhibition of CYP2A6 values increases with Shadow_Ylength and A log P values and decreases with non-common overlap steric volume (NCOSV).

Eq. (8) could explain and predict 63.0% and 61.0%, respectively, of the variance. The predictive R^2 value for the test set was found to be 0.867 which was comparable to the R_{pred}^2 value of the corresponding GFA Eq. (7). However, the r_m^2 of the G/PLS equation was inferior to that of the GFA Eq. (7). According to the standardized values of the regression coefficients the relative importance of the variables is in the following order: $\mathrm{pIC}_{50(2A5)} > \mathrm{A\ log\ P} > \mathrm{NCOSV} > \mathrm{Shadow\_Ylength}$. The standard errors of model fit, crossvalidation and test set prediction for Eq. (8) are given in Table 3.

In each case of the GFA or G/PLS models, model development process was subjected to randomization test at 90% confidence level taking into account the whole pool of descriptors (Table 8). This is different from model randomization in that for process randomization the selection of descriptors is repeated with the randomized response values at different runs while in case of model randomization, the descriptors are kept fixed. In case of the GFA models, the mean value of correlation coefficients (*R*) of random models in each case was much lower than that of the corresponding non-random model while in case of the G/PLS models the mean values of correlation coefficients (*R*) of random models are found to be high values, though lower than those of the corresponding non-random models. This suggests that the GFA models developed in this paper are superior to the corresponding G/PLS models. This also shows the capability of flexible modeling techniques and warns one to use the commercial modeling packages with adequate validation strategies. The final models (Eqs. (1)–(8)) were also subjected to a randomization test with 99% confidence level (Table 9). In no case, correlation coefficient *R* of a random model is superior to that of the corresponding non-random model. The comparative chart of statistical qualities of the above models is presented in Table 10. An attempt was also made to develop PLS models taking into account the whole pool of descriptors which results in much inferior models. This corroborates previous observation [43] of poor performance of PLS models containing noisy variables and justifies the need of variable selection process for PLS. The calculated values of CYP2A5 and CYP2A6 binding affinity values according to different equations are given in Table A3 of Supplementary material. The scatter plots (with bisecting line) of the observed vs. predicted values of CYP2A5 and CYP2A6 binding affinity values of the test set compounds according to different equations are given in Figs. 4 and 5, respectively.

3.3. Further test on external validation

The models were also subjected to the test for criteria of external validation as suggested by Golbraikh and Tropsha [48]. These authors [48] have recommended that in addition to a high value of crossvalidated R^2 (Q^2), the correlation coefficient *r* between the observed and the predicted activities of compounds from an external test set should be close to 1. At least one (but better both) of the correlation coefficients for regressions through the origin (observed vs. predicted activities, or predicted vs. observed activities), i.e., r_0^2 or $r_0'^2$ should be close to r^2. Furthermore, at least one slope of regression lines (*k* or *k*′) through the origin should be close

Table 10
Comparison of statistical qualities of different models.

Model	R	R^2	Q^2	R_{pred}^2	r^2	r_0^2	r_m^2
QSAR of CYP2A5 inhibition potency							
GFA (50 000 iterations)							
1	0.892	0.796	0.686	0.707	0.734	0.732	0.701
G/PLS (5000 iterations)							
2	0.882	0.779	0.575	0.615	0.679	0.639	0.543
QSAR of CYP2A6 inhibition potency							
GFA (10 000 iterations)							
3	0.779	0.607	0.502	0.667	0.785	0.648	0.494
G/PLS (500 iterations)							
4	0.748	0.561	0.489	0.714	0.778	0.721	0.592
QAAR with CYP2A5 as response variable							
GFA (50 000 iterations)							
5	0.947	0.898	0.814	0.914	0.905	0.904	0.876
G/PLS (1000 iterations)							
6	0.927	0.859	0.746	0.902	0.903	0.881	0.769
QAAR with CYP2A6 as response variable							
GFA (50 000 iterations)							
7	0.869	0.757	0.648	0.869	0.842	0.842	0.842
G/PLS (1000 iterations)							
8	0.823	0.679	0.610	0.867	0.883	0.866	0.768

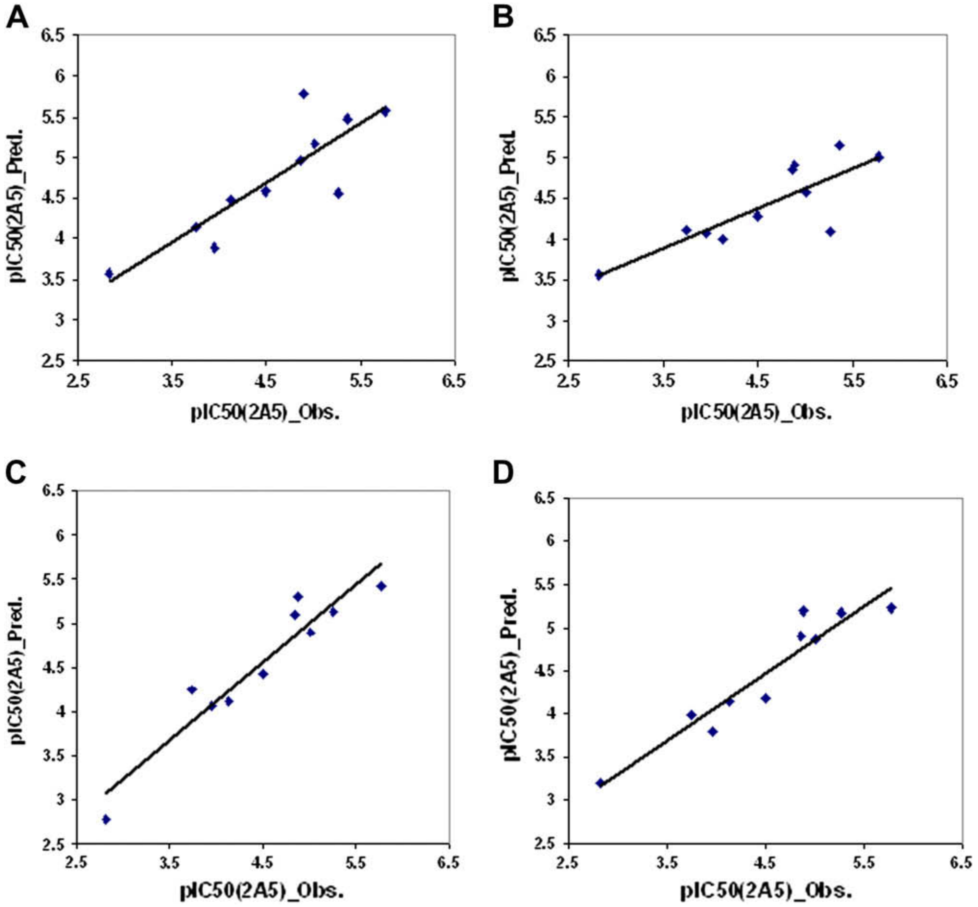

Fig. 4. Scatter plots for observed vs. predicted CYP2A5 inhibitory potency of the test set compounds according to (A) Eq. (1); (B) Eq. (2); (C) Eq. (5); (D) Eq. (6).

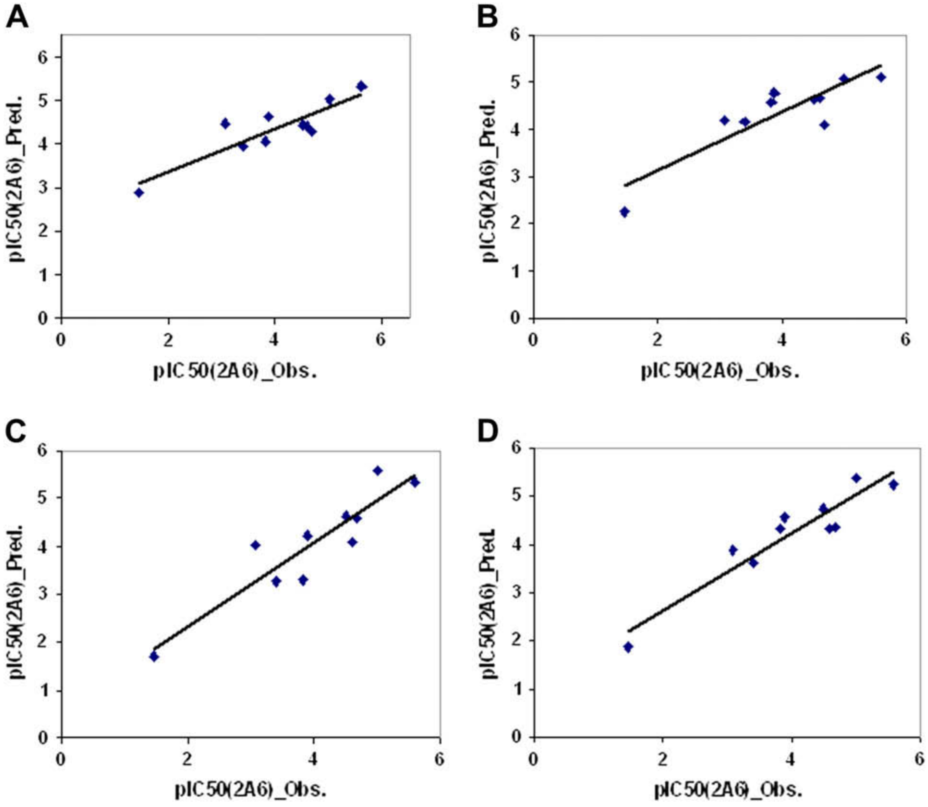

Fig. 5. Scatter plots for observed vs. predicted CYP2A6 inhibitory potency of the test set compounds according to (A) Eq. (3); (B) Eq. (4); (C) Eq. (7); (D) Eq. (8).

Table 11
External validation characteristics of different models according to Golbraikh and Tropsha [48].

Sl. No.	Statistical parameters	Model number							
		1	2	3	4	5	6	7	8
1	r^2	0.734	0.679	0.785	0.778	0.905	0.903	0.842	0.883
2	r_0^2	0.732	0.639	0.648	0.721	0.904	0.881	0.842	0.866
3	r'^2_0	0.600	0.083	−0.266	0.349	0.888	0.839	0.823	0.797
4	$(r^2 - r_0^2)/r^2$	0.003	0.059	0.175	0.073	0.001	0.024	0.000	0.019
5	$(r^2 - r'^2_0)/r^2$	0.183	0.878	1.339	0.551	0.019	0.071	0.023	0.097
6	Minimum of 4 and 5	0.003	0.059	0.175	0.073	0.001	0.024	0.000	0.019
7	k	0.965	1.039	0.935	0.929	0.988	1.009	0.979	0.954
8	k'	1.028	0.952	1.043	1.055	1.009	0.987	1.010	1.039

to 1. Models are considered acceptable, if they satisfy all of the following conditions: (i) $Q^2 > 0.5$, (ii) $r^2 > 0.6$, (iii) r_0^2 or r'^2_0 is close to r^2, such that $[(r^2 - r_0^2)/r^2]$ or $[(r^2 - r'^2_0)/r^2] < 0.1$ and $0.85 \le k \le 1.15$ or $0.85 \le k' \le 1.15$. A list of values of different parameters for different models as recommended by Golbraikh and Tropsha [48] is given in Table 11. Except model 3, all models pass the criteria recommended by Golbraikh and Tropsha [48].

4. Overview

The applicability of the developed models is naphthalene derivatives and related aromatic and heteroaromatic compounds. In the case of CYP2A5 inhibition, the GFA derived model is better than the G/PLS derived model considering both internal and external validations. In the case of the CYP2A6 inhibitory potency data, the GFA derived model is better than the G/PLS model considering internal validation whereas the latter is better in external validation (which is more important [31]) than the former. Considering internal validation, QAAR models with CYP2A5 as response variable are better than those with CYP2A6 as response variable while considering external validation GFA and G/PLS derived QAAR models give similar results for both types of response variables though GFA models are better than the G/PLS ones. Considering randomization test applied on the model development process, GFA models are found to be superior to the corresponding G/PLS models. Among the parameters, which were found important in modeling both the responses, were different Jurs descriptors, electronic descriptors (like Sr, Apol), steric descriptors (like shadow indices, Molref), shape descriptors (like COSV, Fo) and lipophilicity descriptors. This indicates that the CYP2A5 and CYP2A6 inhibition of these compounds is related to charge distribution, surface area, electronic, hydrophobic and spatial properties of the molecules. These observations are in agreement with the CoMFA results of the previous study [18] which showed the importance of charge distribution for the binding affinity. As the CYP2A5 and CYP2A6 binding affinity values are dependent on multiple factors, medicinal chemists should design novel compounds in such a way where the factors contributing positively to the binding affinity are enhanced and detrimental factors are reduced. The developed equations may be helpful to predict CYP2A5 and CYP2A6 inhibition potency of naphthalene derivatives and related aromatic and heteroaromatic compounds.

Acknowledgements

This research work is supported by a Major Research Grant of the University Grants Commission (UGC), New Delhi. One of the authors (P.P. Roy) thanks the UGC, New Delhi for a fellowship.

Appendix. Supplementary material

Supplementary data associated with this article can be found in the online version, at 10.1016/j.ejmech.2008.11.010.

References

[1] R. Arimoto, Curr. Top. Med. Chem. 6 (2006) 1609.
[2] E.M. Sellers, H.L. Kaplan, Clin. Pharmacol. Ther. 68 (2000) 35.
[3] E.M. Sellers, Y. Ramamoorthy, M.V. Zeman, Nicotine Tob. Res. 5 (2003) 891.
[4] C.J. Patten, T.J. Smith, M.J. Friesen, R.E. Tynes, C.S. Yang, S.E. Murphy, Carcinogenesis 18 (1997) 1623.
[5] K. Fujita, T. Kamataki, Environ. Mol. Mutagen. 38 (2001) 339.
[6] J.S. Miles, A.W. McLaren, L.M. Forrester, M.J. Glancey, M.A. Lang, C.R. Wolf, Biochem. J. 267 (1990) 365.
[7] S. Yamano, J. Tatsuno, F.J. Gonzalez, Biochemistry 29 (1990) 1322.
[8] C.H. Yun, T. Shimada, F.P. Guengerich, Mol. Pharmacol. 40 (1991) 679.
[9] M. Oscarson, Drug Metab. Dispos. 29 (2001) 91.
[10] O. Pelkonen, H. Raunio, A. Rautio, J. Mäenpää, M.A. Lang, J. Irish. Coll. Phys. Surg. 22 (1993) 24.
[11] J.K. Yano, M.H. Hsu, K.J. Griffin, C.D. Stout, E.F. Johnson, Nat. Struct. Mol. Biol. 12 (2005) 822.
[12] G.A. Schoch, J.K. Yano, M.R. Wester, K.J. Griffin, C.D. Stout, E.F. Johnson, J. Biol. Chem. 279 (2004) 9497.
[13] T.L. Poulos, B.C. Finzel, I.C. Gunsalus, G.C. Wagner, J. Kraut, J. Biol. Chem. 260 (1985) 16122.
[14] A. Poso, J. Gynther, R. Juvonen, J. Comput. Aided Mol. Des. 15 (2001) 195.
[15] <http://www.patentstorm.us/patents/7148046.html>.
[16] <http://www-ssrl.slac.stanford.edu/research/highlights_archive/cyp2a6.pdf>.
[17] D. Kim, Z.L. Wu, F.P. Guengerich, J. Biol. Chem. 280 (2005) 40319.
[18] M. Rahnasto, H. Raunio, A. Poso, C. Wittekindt, R.O. Juvonen, J. Med. Chem. 48 (2005) 440.
[19] S. Khurana, V. Batra, A.A. Patkar, F.T. Leone, Respir. Med. 97 (2003) 295.
[20] World Health Organization, The World Health Report 2002: Reducing Risks, Promoting Healthy Life, World Health Organization, Geneva, 2002.
[21] J. Hukkanen, P. Jacob 3rd, N.L. Benowitz, Pharmacol. Rev. 57 (2005) 79.
[22] M. Ezzati, S.V. Hoorn, A. Rodgers, A.D. Lopez, C.D. Mathers, C.J. Murray, Lancet 362 (2003) 271.
[23] K. Kitagawa, N. Kunugita, T. Katoh, M. Yang, T. Kawamoto, Biochem. Biophys. Res. Commun. 262 (1999) 146.
[24] K. Inoue, H. Yamazaki, T. Shimada, Arch. Toxicol. 73 (2000) 532.
[25] M. Nakajima, S. Yamagishi, H. Yamamoto, T. Yamamoto, Y. Kuroiwa, T. Yokoi, Clin. Pharmacol. Ther. 67 (2000) 57.
[26] R.F. Tyndale, E.M. Sellers, Ther. Drug Monit. 24 (2002) 163.
[27] D.F. Lewis, M. Dickins, B.G. Lake, P.J. Eddershaw, M.H. Tarbit, P.S. Goldfarb, Toxicology 133 (1999) 1.
[28] R.O. Juvonen, J. Gynther, M. Pasanen, E. Alhava, A. Poso, Xenobiotica 30 (2000) 81.
[29] K. Roy, D.K. Pal, A.U. De, C. Sengupta, Drug Des. Discov. 17 (2001) 315.
[30] Cerius2 version 4.10 is a product of Accelrys, Inc., San Diego, USA, <http://www.accelrys.com/cerius2>.
[31] J.T. Leonard, K. Roy, QSAR Comb. Sci. 25 (2006) 235.
[32] K. Roy, A.S. Mandal, J. Enzyme Inhib. Med. Chem. 23 (2008) 980.
[33] B.S. Everitt, S. Landau, M. Leese, Cluster Analysis, Edward Arnold, London, 2001.
[34] A.J. Hopfinger, J.S. Tokarsi, in: P.S. Charifson (Ed.), Practical Applications of Computer-Aided Drug Design, Marcel Dekker, New York, 1997, pp. 105–164.
[35] Y.L. Fan, M. Shi, K.W. Kohn, Y. Pommier, J.N. Weinstein, J. Med. Chem. 44 (2001) 3254.
[36] D. Rogers, A.J. Hopfinger, J. Chem. Inf. Comput. Sci. 34 (1994) 854.
[37] W.J. Dunn III, D. Rogers, in: J. Devillers (Ed.), Genetic Algorithms in Molecular Modeling, Academic Press, London, 1996, pp. 109–130.
[38] K. Hasegawa, Y. Miyashita, K. Funatsu, J. Chem. Inf. Comput. Sci. 37 (1997) 306.
[39] G.W. Snedecor, W.G. Cochran, Statistical Methods, Oxford & IBH Publishing Co. Pvt. Ltd., New Delhi, 1967.
[40] S. Wold, L. Eriksson, in: H. van de Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, Germany, 1995, pp. 312–317.
[41] A.K. Debnath, in: A.K. Ghose, V.N. Viswanadhan (Eds.), Combinatorial Library Design and Evaluation, Marcel Dekker, Inc., New York, 2001, pp. 73–129.
[42] K. Roy, Expert Opin. Drug Discov. 2 (2007) 1567.
[43] P.P. Roy, K. Roy, QSAR Comb. Sci. 27 (2008) 302.
[44] SPSS is a software of SPSS, Inc., USA.
[45] D.T. Stanton, P.C. Jurs, Anal. Chem. 62 (1990) 2323.
[46] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, P. Gramatica, Environ. Health Perspect. 111 (2003) 1361.
[47] P.P. Roy, J.T. Leonard, K. Roy, Chemom. Intell. Lab. Syst. 90 (2008) 31.
[48] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269.