J Comput Aided Mol Des (2007) 21:559–573
DOI 10.1007/s10822-007-9139-6

Generation of in-silico cytochrome P450 1A2, 2C9, 2C19, 2D6, and 3A4 inhibition QSAR models

M. Paul Gleeson · Andrew M. Davis · Kamaldeep K. Chohan · Stuart W. Paine · Scott Boyer · Claire L. Gavaghan · Catrin Hasselgren Arnby · Cecilia Kankkonen · Nan Albertson

Received: 17 July 2007 / Accepted: 4 October 2007 / Published online: 22 November 2007

Abstract In-silico models were generated to predict the extent of inhibition of cytochrome P450 isoenzymes using a set of relatively interpretable descriptors in conjunction with partial least squares (PLS) and regression trees (RT). The former was chosen due to the conservative nature of the resultant models built and the latter to more effectively account for any non-linearity between dependent and independent variables. All models are statistically significant and agree with the known SAR and they could be used as a guide to P450 liability through a classification based on the continuous pIC50 prediction given by the model. A compound is classified as having either a high or low P450 liability if the predicted pIC_{50} is at least one root mean square error (RMSE) from the high/low pIC_{50} cut-off of 5. If predicted within an RMSE of the cut-off we cannot be confident a compound will be experimentally low or high so an indeterminate classification is given. Hybrid models using bulk descriptors and fragmental descriptors do significantly better in modeling CYP450 inhibition, than bulk property QSAR descriptors alone.

Keywords Cytochrome P450 1A2 · 2C9 · 2C19 · 2D6 · 3A4 · In-silico QSAR · ADMET model · Partial least squares regression · PLS · Regression trees

M. P. Gleeson · A. M. Davis · K. K. Chohan · S. W. Paine
Department of Physical & Metabolic Sciences, AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, UK

A. M. Davis
e-mail: andy.davis@astrazeneca.com

M. P. Gleeson (✉)
Computational and Structural Chemistry, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, UK
e-mail: paul.x.gleeson@gsk.com

S. Boyer · C. L. Gavaghan · C. H. Arnby
Computational Toxicology, Safety Assessment, AstraZeneca R&D, Molndal, Sweden

C. Kankkonen
HTS Screening, AstraZeneca R&D Mölndal, 431 83 Molndal, Sweden

N. Albertson
Discovery DMPK & BAC, AstraZeneca R&D Mölndal, Molndal, Sweden

Introduction

In an increasingly tough commercial and regulatory environment, a current focus for the pharmaceutical industry is speed, quality, efficiency, and attrition through drug discovery and development. A truism that is impacting on the whole process is that much of the future success or failure of a clinical candidate is embodied within its the chemical structure. Hence drug discovery has embraced increasingly complex screening cascades to triage potential weaklings from the point of view of absorption, distribution, metabolism, and elimination and toxicology (ADMET) concerns. Increasing the complexity of the screening cascade would certainly increase the quality of potential clinical candidates, but necessarily will add cost and potentially slow down the process of identifying the best clinical candidate. Speed and quality can be achieved by making better compounds early, and deselecting compounds from synthesis having predictable undesirable properties. Hence the application of predictive methods such as quantitative structure–activity relationships (QSAR) and structure-based design to ADMET has become a very active area.

QSAR methods were introduced in the 1960s to model changes in activity within homologous chemical series. Each biological target would be unique to each project (and each chemotype liable to be also), and hence information was not immediately useful to other projects. ADMET problems differ in that all projects and hence all chemical classes face the same problems in-vivo, and hence common screens can be used generically across all projects. Hence it is not surprising that the compiled data from ADMET screens have become the data-sources of so-called "global" predictive QSAR models. The term global in this context means that the models are based on datasets that are compiled from across a number of projects, and ideally across all projects in which the company is currently and historically has been synthetically active.

While the QSAR models of the 1960s were used to predict the next few compounds in a series, and then rebuilt by QSAR specialists, modern global ADMET QSAR models are used in a different way. Once the ADMET QSAR models are generated, they are often mounted on corporate intranets, and made available to QSAR non-specialists to predict forthcoming chemistries. This takes QSAR into a new area, very large "global" QSAR models, which perhaps are more diffuse in prediction than the smaller series specific models, but may have the power to predict cross-chemistries.

Cytochromes P450s are a superfamily of isoenzymes that catalyze the metabolism of a large number of compounds of both exogenous and endogenous origin [1–3] and are implicated in toxicological events [4]. Metabolism-based drug interactions comprise one of the major concerns during drug development. If two or more drugs that compete for the same P450 are administered concomitantly, and one has a significant affinity for the P450 [4], whether as a substrate or inhibitor, the metabolism of the victim drug may be inhibited causing its plasma levels to rise which may lead to undesirable toxic effects. Early identification of compound-P450 CYP interaction is critical as it enables us to design out this liability early in the lead optimization process. At AstraZeneca (AZ), we focus our efforts on five of the major human CYP isoforms, 1A2, 2C9, 2C19, 2D6, and 3A4 since these make up approximately 60% of the total hepatic CYPs [5, 6] and account for a significant proportion of the P450-mediated metabolism of therapeutic drugs [7]. Focusing on these should allow us minimize the likelihood of drug–drug interactions of our compounds in the clinic [8–10].

This paper describes our attempts to build global QSAR models for cytochrome P450 CYP inhibition, to determine the structure features governing P450 inhibition and how they may be applied to guide chemical synthesis. The dilemma faced when building a QSAR is whether to follow more closely the original QSAR concepts laid down by Hanch and Fujita [11], and a more modern reinterpretation of this, the linear Free energy relationship (LFER) model [12, 13], or to rely on the Black Box approach using modern computational power coupled with complex descriptors and statistical methods. The advantage of the former QSAR approach is that it is considered to offer greater design opportunities through simpler more interpretable models, and the latter giving more predictive albeit complex, uninterpretable models.

The main objective of this study was to demonstrate whether or not it was possible to obtain a set of predictive in-silico models for P450 inhibition, using relatively simple statistical techniques and descriptors familiar to chemists, which have been demonstrated to be predictive of other ADMET parameters including volume of distribution [14] and plasma protein binding [15], and crucially, allow the key learning from the models to used to inspire chemistry ideas with relative ease. To this end we employed two statistical methods PLS regression and regression trees (RT), in conjunction with easily interpretable descriptors, which are described elsewhere [16]. We also consider whether fragmental descriptors derived from either Leadscope or Algorithm Builder, and consensus modeling, proved beneficial in the modeling exercise.

A further impetus for this study was that although a number of molecular modeling and QSAR studies have been reported on cytochrome P450 inhibitors and substrates in the literature these have almost exclusively been limited to relatively small datasets [17–29, 32], often from multiple sources, making it difficult to generalize any findings. By extracting knowledge from a large set of diverse measurements, derived from a single assay, we can provide an increased insight into the SAR as well as to help confirm or reject any of the postulated SAR at present.

Theoretical procedures

Experimental data

An experimental campaign was undertaken at AstraZeneca to screen approximately 1,500 compounds though five different in-house developed cytochrome P450 inhibition assays. A major goal of this study was to determine the physical characteristics required for the inhibition of cytochrome P450s 1A2, 2C9, 2C19, 2D6, and 3A4, as well as the extent of the overlap among the isoforms for different chemotypes. Such an analysis on a truly diverse dataset of compounds should prove considerably more useful than the often questionable assumptions derived from models that are generated on limited literature datasets.

A high throughput assay was developed within AstraZeneca using a standard DMPK fluorescence assay as a

starting point. The compounds selected for screening consisted of 384 oral drugs and 1,152 compounds selected from AstraZeneca compound collection. These 1,536 compounds represent a diverse set of compounds spanning many chemotypes.

For each compound a 20 mM DMSO stock solution was made, and diluted in a solution of 50 mM cyclodextrin in DMSO and seven serial dilutions were performed in cyclodextrin/DMSO. The 7 stock solutions were subsequently added in triplicate to a 384 well plate where the DMSO was allowed to evaporate. The CYP enzyme, buffer and fluorescent substrates were added to each well and left to pre-incubate for 10 min. NADPH/NADP was added and the plate was allowed to incubate for a further 20–30 min. The end point was determined based on the disappearance of the fluorometric substrate (Table 1).

After excluding experiments where solubility meant a satisfactory dose-response curve could not be generated, or where the compounds themselves had fluorescent properties, measurements on at least 1 isoform for ~1000 of the 1,536 were available. The subset of data with quantitative pIC_{50} values was split randomly into a training and test set (70%/30%) and the remaining compounds with non-quantitative values ($<$ /~/ $>$) were used to assess the model in classification only (Table 2).

Bulk property descriptors

A set of key bulk property descriptors used in AZ, which we have described elsewhere [16], were calculated for all compounds. Compounds that reported an error value for calculated descriptors were automatically excluded from the analysis. Although PLS can tolerate some missing values [30] other statistical techniques cannot, and we wished to take a consensus approach.

Molecular fragments

Five wholly fragmental QSAR models were built using fragments extracted from Algorithm Builder 1.0 [31]. Algorithm Builder splits molecules automatically into user-defined fragments such as atoms and cores/chains etc which can then be used to build QSARs. We fragmented the 5 datasets using; (a) atom based fragmentation, (b) chains between 3 and 6 atoms in length, and (c) scaffold/core based fragmentation scheme with substitution patterns specified. The latter could be particularly important for compounds containing a pyridine ring for example as these can potentially bind to the heme. Substitution adjacent to nitrogen atom of pyridine would hinder heme binding and hence attenuate CYP inhibition [32].

Table 1 CYP Isoform studied and the corresponding fluorescent substrate

CYP Isoform	Substrate
1A2	7-Ethoxy-3-cyanocoumarin (CEC)
2C9	7-Methoxy-4-trifluoromethyl coumarin (MFC)
2C19	7-Methoxy-4-trifluoromethyl coumarin (MFC)
2D6	7-Methoxy-4-aminomethyl-coumarin (MAMC)
3A4	7-Benzyloxy-4-trifluoromethylcoumarin (BFC)

Table 2 Number of results in each CYP assay classified as either quantitative, non-quantitative, or no-value and number of compounds used in the model building and validation process

Number of observations	1A2	2C9	2C19	2D6	3A4
Training set	301	457	369	170	463
Test set (quantitative)	126	202	168	89	187
Test set (non-quantitative)	154	25	19	32	85
Total	581	684	556	291	735

The fragment matrix was exported from Algorithm Builder and analyzed in Excel. Fragments were only considered if they were present in more than 10 compounds and if compounds with such a fragment had a mean pIC_{50} significantly different from the mean of the total dataset at the 95% confidence level using a *t*-test.

Leadscope [33] is a tool that allows users to identify whether a particular sub-structural feature from a set of molecules confers a net effect on activity significantly different from those without it. The CYP data was imported into Leadscope and the molecular features that had a significant net effect on pIC_{50} were extracted. The features were coded as SMARTS strings and used as indicator variables alongside the standard in-house molecular descriptors. We refer to the models built using leadscope derived fragments as the LS models.

The numbers of observations in the algorithm builder (AB) and leadscope (LS) datasets differ to those of the whole molecular descriptor dataset due to (a) observation exclusion during the PLS data pre-processing step as a result of low variance in the descriptor matrix and (b) molecule import failures during descriptor generation. As we wanted to compare both RT and partial least squares (PLS) directly for these datasets, only observations that passed the PLS preprocessing were using for RT model building. For the final comparison between the different methods and descriptor sets only compounds common to all methods are used.

Statistical models

Two distinctly different statistical methods have been used in our analysis: PLS [34–36] and RT [37]. Although these techniques employ different basic modeling assumptions,

we would nevertheless expect to obtain complimentary results in prediction [14, 38]. Furthermore, making use of a number of techniques provides the opportunity to explore consensus prediction (average predictions from two or more independent modeling techniques), which may have advantages over any individual model in prediction. Consensus models have already proved useful in a number fields including QSAR [14, 38], protein-ligand docking and scoring [39, 40].

PLS regression model details

Initial PLS models were built using (a) bulk property descriptors, (b) AB fragments, and (c) bulk property descriptors and LS fragments in SIMCA [41, 42]. The standard data preprocessing step in SIMCA was used to exclude descriptors and observations of low variance that were not suitable for use in the model building process. These models were assessed in both fit and prediction of the training and test sets, respectively. Refined PLS models were built in GOLPE [43] with variable reduction achieved using 1 round of D-optimal design (remove 30% of redundant variables) followed by fractional factorial selection to find the key variables that describe inhibition. The number of components fitted was determined automatically in SIMCA using a leave many out procedure to assess their individual significance (leave 1/7 of the out, rebuilding the model 20 times).

Regression tree model details

The regression tree (RT) methodology as implemented in CART 4.0 [44] involves building a set of RT, with each been pruned back using 10-fold cross validation. This has been found to be more effective than using stopping rules. Another characteristic of CART's implementation is the utilization of consensus predictions from a number of different trees. Models were built using the same descriptor combinations used in the PLS models, with each model consisting of a consensus of 15 individual trees. Note that unlike the Random Forests adaptation of CART, all the descriptors are available during the training of each of the individual 15 models, whereas in the latter only a random subset is available in each model building step. It should also be noted that no additional effort was spent optimizing the RT parameters beyond the program default.

The performance of the models were assessed in fit and prediction using the training and test sets. The descriptors used in the model were refined using the average variable importance (VIP) derived from the ensemble. The CART 4.0 VIP is a relative scale in which variables that are involved in a primary split have greater importance than those used further down the tree. We often find that removing those variables with a low variable importance can increase the predictivity of the model [38]. However, an artefact of this process is that the fit statistics can improve considerably compared to prediction statistics. Although this suggests an overfitted model, the fact that they prove more predictive on the independent test set suggests this refinement process is beneficial. This is analogous to the removal of redundant variables in COMFA as performed in GOLPE [43].

PCA model details

We can assess the similarities between the different isoforms based on the overlap in the molecules that are inhibitors and this approach is complimentary to sequence alignment methods. These results complement the individual PLS models in that the latter shows how the activity of each isoform varies with different descriptors while the former is a measure of their similarity. These two results will show independently how likely it is that an inhibitor of one isoform will inhibit another.

A PCA model was built on compounds that had quantitative measurements in at least 4 of the different CYP assays as a portion of missing values can be tolerated in the NIPALS algorithm employed in SIMCA. The number of components fitted was determined automatically in SIMCA using a leave many out procedure to assess the significance.

Assessing the predictive ability of QSAR models

A number of statistical measures may be used to evaluate such models [14, 15]. We use the square of Pearson's correlation coefficient (r^2), the root mean square error in prediction (RMSE) to determine the utility of a model. The r^2 tell us the percentage of variability in the observed value that is predicted by the model, the root mean square error in prediction measures the uncertainly in our predictions in the units of the measurement and the mean absolute error indicates if there is any systematic bias in the predictions. The r^2 and RMSE are calculated using Eqs. 1 and 2, respectively, where x and y are predicted and observed values of the biological response, n is the number of compounds and $\bar{x}$ and $\bar{y}$ are the mean of the predictions and observations respectively.

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum (y-x)^2}{n}} \quad (2)$$

Often the fitted line to the observed versus predicted data has a slope >1 indicating the standard deviation in

predictions is less than the standard deviation in the observations. We therefore also report the r_0^2 value, which is the %variance explained in prediction to the line $x = y$ and is directly related to the RMSE (Eq. 3). The r^2 is most important parameter if one wishes to rank the data only, and an accurate prediction is not imperative.

$$r_0^2 = 1 - \frac{n}{n-1}\left(\frac{\text{RMSE}^2}{\sigma^2_{\text{measured}}}\right) \quad (3)$$

The primary assessment of whether a model is truly predictive is undertaken using the test set, which is a better guide to the model's performance than the training set. We determine if a model is predictive, and how significant that is, using the criteria below. We do this because a large r^2 (or low RMSE) is not necessarily enough to assess a model in prediction in all circumstances.

The following criteria have been used here;

- If the $r_0^2 > 0$, and assuming the result is statistically significant above the 95% confidence level according to an *F*-test, then the model is significant, and may be useful in prediction.
  - The size of the r_0^2 value determines how useful a model is, and determines what environment the model is used (in continuous prediction of a pIC50 or in classification).
  - If the RMSE test set < RMSE of the fitted model then the model predicts the test set better than it fits the training set. This could occur for a particular project series within a global model.
- If the $r_0^2 \approx 0$ and SD $Y_{\text{obs}} >$ RMSE of the fitted model, then the model is not predictive.
- If the $r_0^2 \approx 0$, but the $r^2 > 0$, then the model might be useful to rank compounds, rather than for absolute predictions.
- If the $r_0^2 \approx 0$, but the SD $Y_{\text{obs}} \leq$ RMSE in fit, then it may be very difficult to decide if a model is truly predictive. This is because the data to be predicted lies within the error of the fitted model.

Results

The main objective of this study was to demonstrate whether or not it was possible to obtain a set of predictive in-silico models for P450 inhibition, using simple statistical techniques and descriptors familiar to chemists, which we have been demonstrated to be predictive of other ADMET parameters. In-silico models have been built by us on a range of ADMET parameters, following a similar, methodology, obtaining useful predictive models. Examples which have been reported in the literature include, plasma protein binding [45], volume of distribution [46], and hERG inhibition [47], each of which showed good predictive performance, as evidenced by RMSEs of between 2.5 and 4.0 fold, with r^2s ranging from between 0.4 and 0.7. These models were also easily interpretable making them, in our experience, valuable tools in drug discovery.

We now discuss whether this previously successful methodology will lead to equivalently predictive models for P450s inhibition, given the knowledge that they are more heavily influenced by molecular recognition.

Bulk property QSAR models

Training set results

Comparison of fitted r^2 values between these models and to other models in the literature can be very misleading. First, the degree of over-fit or underfit of models is difficult to assess based on only fit statistics. The P450 RT models in fit have considerably higher r^2s and lower RMSEs than the PLS models suggesting they provide a better description of the inhibition process, but the quality of the models appears more similar when comparing the test set prediction statistics (Table 3). Second, the magnitude of r^2 value will depend on the variance of the biological data relative to the experimental error in the dataset, and so on its own it is not useful in assessing the comparative quality of the models where the ranges of the biological data is significantly different [15]

Test set results

The r^2 and RMSE of the test set gives a more realistic guide to the predictive power of the P450 CYP models (Table 3). It is a difficult and often arbitrary decision when QSAR model training should cease, and this appears particularly true of non-linear models. While both PLS and RT models were built using recommended methods, this is confirmed plotting the fitted r^2 of the training set versus the predicted r^2 of the test sets for both models (Fig. 1). The graph identifies significant overtraining of the RT models while also showing that the predictive ability of the model in the training set for PLS (as given by the r^2) is maintained in the test set. The r^2 of the RT models, on the other hand deteriorates significantly from the training set to the test set even though the RT RMSEs in predictions of the independent test sets are somewhat comparable to those of the PLS models.

By inspection of the r^2, r_0^2, and RMSE, it can be seen that the predictive ability of the different CYP models varies considerably, with 3A4 being most predictive and 2D6 the least. The most predictive QSAR models were

Table 3 PLS and RT results for Training and test set results generated using bulk property models

	Training set results						Quantitative test set results				
PLS	r^2	r_0^2 (q^2)	RMSE	N	σ_Y	Desc./Comp.	r^2	r_0^2	RMSE	N	σ_Y
1A2	0.30	0.30 (0.20)	0.57	301	0.68	(17/3)	0.18	0.17	0.68	126	0.75
2C9	0.31	0.31 (0.31)	0.55	457	0.67	(11/1)	0.29	0.30	0.56	202	0.29
2C19	0.15	0.15 (0.15)	0.56	369	0.61	(15/1)	0.12	0.11	0.55	168	0.12
2D6	0.15	0.15 (0.13)	0.64	170	0.70	(12/1)	0.09	0.07	0.58	89	0.09
3A4	0.36	0.36 (0.30)	0.59	463	0.74	(13/1)	0.36	0.37	0.58	187	0.36
	Training set results						Quantitative test set results				
RT	r^2	r_0^2	RMSE	N	σ_Y	Desc./Trees	r^2	r_0^2	RMSE	N	σ_Y
1A2	0.76	0.72	0.36	301	0.68	25/15	0.15	0.14	0.69	126	0.75
2C9	0.82	0.78	0.31	457	0.67	24/15	0.23	0.23	0.59	202	0.29
2C19	0.77	0.72	0.32	369	0.61	27/15	0.15	0.13	0.55	168	0.12
2D6	0.73	0.67	0.40	170	0.70	32/15	0.13	0.11	0.56	89	0.09
3A4	0.81	0.76	0.36	463	0.74	28/15	0.32	0.33	0.60	187	0.36
	Training set results						Quantitative test set results				
Consensus	r^2	r_0^2	RMSE	N	σ_Y	Models	r^2	r_0^2	RMSE	N	σ_Y
1A2	–	–	–	–	–	RT + PLS	0.21	0.21	0.66	126	0.75
2C9	–	–	–	–	–	RT + PLS	0.30	0.31	0.56	202	0.29
2C19	–	–	–	–	–	RT + PLS	0.17	0.17	0.54	168	0.12
2D6	–	–	–	–	–	RT + PLS	0.16	0.15	0.55	89	0.09
3A4	–	–	–	–	–	RT + PLS	0.41	0.41	0.56	187	0.36

PLS-RT consensus result for the test set are also given. r_0^2 is the correlation coefficient to the line of unity, q^2 is the cross validated r_0^2 from a leave many out procedure and RMSE is the root-mean-square error in prediction: r^2 is Pearson's correlation coefficient squared derived from the line of best fit, N is the number of observations and σ_Y is the standard deviation

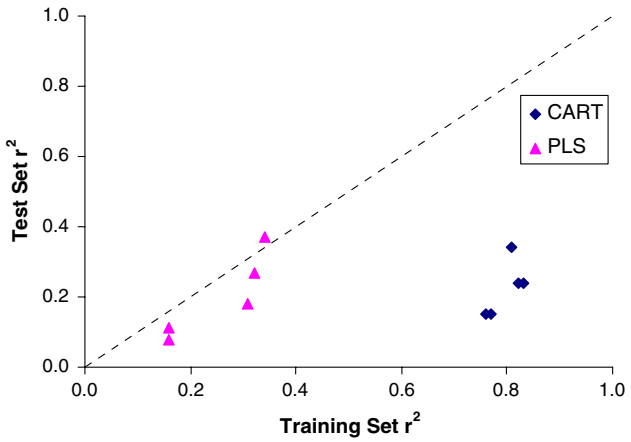

Fig. 1 PLS models are more conservatively fitted than RT models as a results of cross validation and d-optimal/fractional factorial selection of descriptors

produced for 1A2, 2C9, and 3A4 using PLS and the most predictive models for 2C19 and 2D6 were produced using RTs. This might suggest that 2C19 and 2D6 have a much stronger non-linear relationship with molecular structure as represented by our descriptors.

The 3A4 (Fig. 2) and 2C9 models are more predictive than the other 3 isoforms with r^2s of 0.36 and 0.29, respectively and RMSEs in prediction of 0.58 and 0.56 log units, respectively. Although the RMSE of the 2C19 model is lower at 0.54, suggesting this model predicts with lower error, this is a result of the test set observations having the smallest standard deviation. The RMSE of 2C19 model approaches the standard deviation of the observed data (i.e., a random prediction).

In an effort to improve the models further we assessed the benefit of a consensus prediction.

Consensus models

A consensus prediction involving the average of the prediction for each dataset from the RT and PLS models was investigated (Table 3) as it is often found that two different statistical models built on the same dataset can be complementary in prediction. We find that the consensus RT-PLS predictions are better than any of the individual RT or PLS models for all CYP isoforms. These results suggest a consensus prediction should be employed rather than any

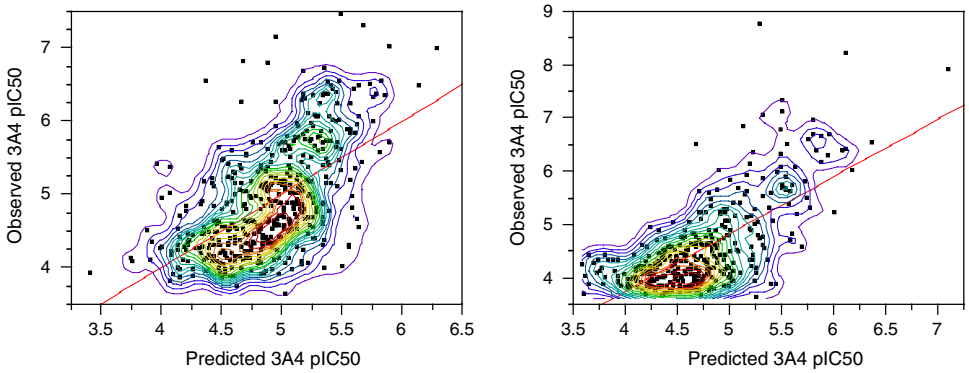

Fig. 2 CYP3A4 PLS training set and test set (right). $r^2 = 0.34$, RMSE = 0.59. and: $r^2 = 0.51$ RMSE = 0.52, respectively. Although relatively weak the model allows a clear differentiation between "high" and "low" compounds

single model should an in-silico P450 screen be implemented.

Performance in binary classification using fuzzy classification

The QSAR models could report a predicted pIC_{50} as well as a classification of the result as either high or low, based on a pIC_{50} cut-off. An IC_{50} for inhibition of a particular cytochrome P450 of 10 μM ($pIC_{50} = 5.0$) is often used as an appropriate flag for concern in drug discovery projects. However, if a compound is predicted close to the pIC_{50} cut-off of 5 we can have little confidence that a compound is truly </> 5 given the intrinsic error associated with the model. The root mean squared error in prediction is ~0.60 for these models suggesting any compound predicted 0.60 units either side of the pIC_{50} cut-off of 5 could potentially be either high or low, since this level of resolution is within the error of the model.

Assuming normal distribution of the error statistics, by excluding predictions within an RMSE of the cut-off, we can thus be ~84% sure that predicted low compounds ($pIC_{50} < 4.4$) will not be observed high, or predicted high compounds ($pIC_{50} > 5.6$) will not be observed low. While this means a number of compounds will be given an "indeterminate" prediction we will now have greater confidence in the high and low predictions. Furthermore, with more predictive models (higher r^2 and lower RMSE's), the uncertainty region will be narrower, and hence higher resolution may be achieved meaning fewer compounds would be unclassified.

The prediction results in classification using these "fuzzy" limits are given in Table 4 and in pictorial form in Fig. 3. This confirms that compounds predicted at the extreme ends of the pIC_{50} range are much more likely to be correctly classified than those close to the cut-off. In 1A2 for example, 91% of compounds that are predicted with a $pIC_{50} < 4.8$ are observed low while 100% of compounds predicted <4.4 are observed low. At pIC_{50}s > 5.6 these models are generally weakest, however, 100% of compounds predicted with a $pIC_{50} > 6$ are potent inhibitors of 1A2 as defined. While a fuzzy classification is considerably more reliable and allows even relatively weak models to be used with greater confidence, a drawback to this classification method is the large proportion of data is not given

Table 4 Results for the fuzzy classifying of the combined quantitative & non-quantitative test data

pIC_{50}	Classification	1A2 Model		2C9 Model		2C19 Model		2D6 Model		3A4 Model	
		Classification success (%)	N_{obs}	Classification success (%)	N_{obs}	Classification success (%)	N_{obs}	Classification success (%)	N_{obs}	Classification success (%)	N_{obs}
<4.0	Inactive	100	14	–	0	100	5	–	0	100	21
4.0–4.4	Inactive	100	23	100	9	92	26	100	24	100	71
4.4–4.8	Borderline inactive	91	92	93	164	87	117	80	61	93	110
4.8–5.2	No classification	–	127	–	178	–	96	–	39	–	81
5.2–5.6	Borderline active	31	114	59	46	52	25	17	12	55	47
5.6–6.0	Active	46	28	86	7	58	12	100	24	88	16
>6.0	Active	100	9	100	1	50	2	–	0	100	7
% of data classified		69		56		66		76		77	
% Success if classified		64		86		81		82		89	

Percentage of total compound (*N*) correctly predicted each bin. No classification (nc) is given for compounds which lie in the bin closest to the cut-off. Model RMSE ~0.60 so compounds predicted with a pIC_{50} <4.4 or >5

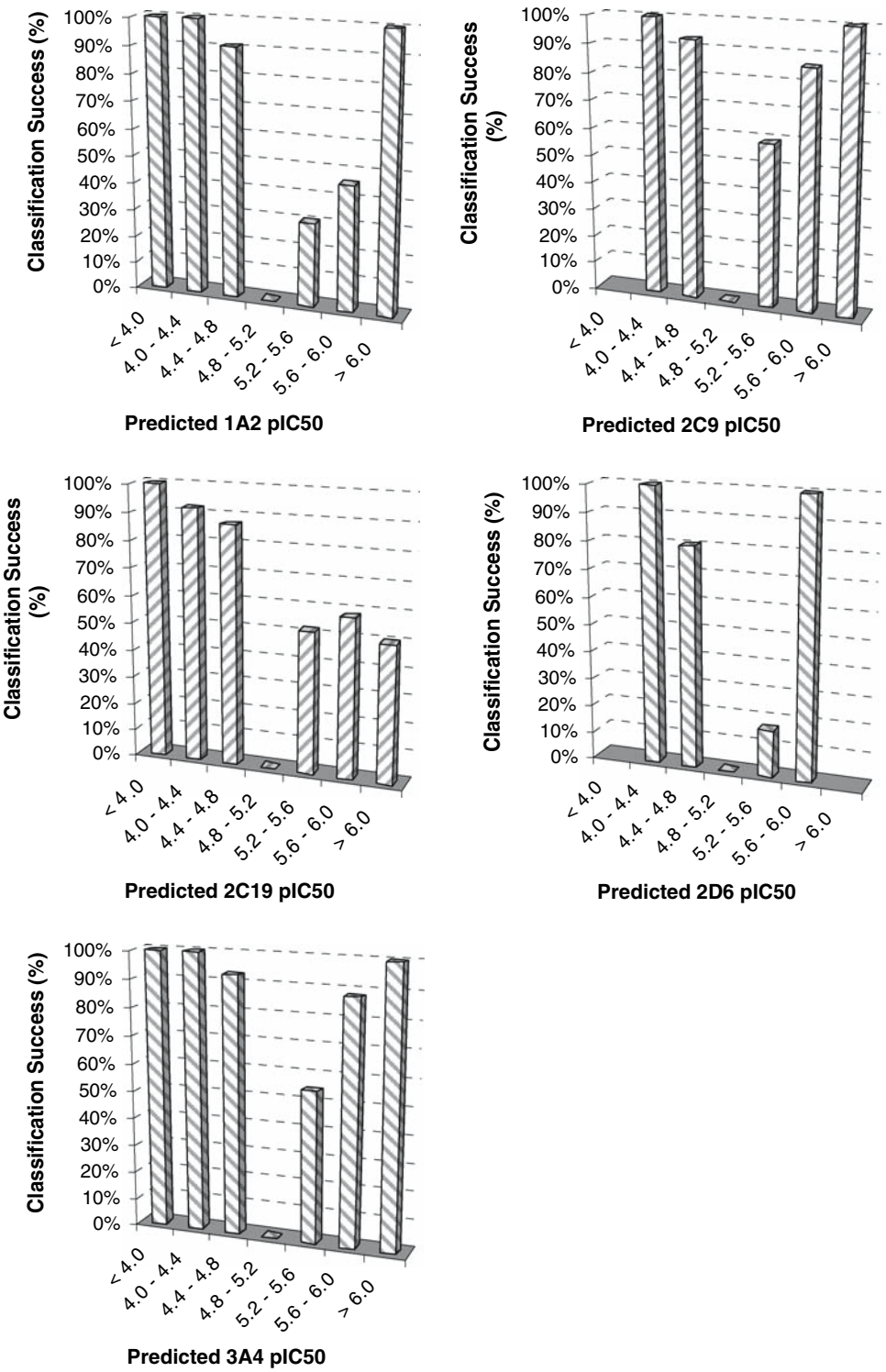

Fig. 3 Fuzzy classification of the combined quantitative & non-quantitative test data. Plot of the classification success rate for each bin against the binned pIC50 prediction. The number of observations in each prediction category can be found in Table 4

any classification. However this may be a more realistic use of such QSAR models, preventing low resolution models being used to make high resolution decisions (model over interpretation).

The key results of using these models in fuzzy classification of the test sets follows Cytochrome P450 1A2

1. 43% of the 1A2 quantitative test set observed "low" ($pIC_{50} < 5$).
2. Model performs better when predictions fall in the extreme low or high regions.
3. Classify compounds as "low" with a high success rate (>91%) if predicted $pIC_{50} < 4.8$.

4. Classify compounds as "high" with a high success rate (100%) if predicted $pIC_{50} > 6.0$.

Cytochrome P450 2C9

1. 58% of the 2C9 quantitative test set observed "low" ($pIC_{50} < 5$).
2. Model performs better when predictions fall in the extreme low or high regions.
3. Classify compounds as "low" with high success rate (>93%) if predicted $pIC_{50} < 4.8$.
4. Classify compounds as "high" with high success rate (>86%) if predicted $pIC_{50} > 5.6$.

Cytochrome P450 2C19

1. 70% of the 2C19 quantitative test set observed "low" ($pIC_{50} < 5$).
2. Model performs better when predictions fall in the extreme low region.
3. Classify compounds as "low" with a high success rate (>87%) if predicted $pIC_{50} < 4.8$.
4. Classify compounds as "high" with only a low success rate (>50%) if predicted $pIC_{50} > 5.2$

Cytochrome P450 2D6

1. 57% of the 2D6 quantitative test set observed "low" ($pIC_{50} < 5$).
2. Model performs better when predictions fall in the extreme low region.
3. Classify compounds as "low" with a high success rate (>80%) if predicted $pIC_{50} < 4.8$.
4. Classify compounds as "high" with a high success rate (100%) if predicted $pIC_{50} > 5.6$.

Cytochrome P450 3A4

1. 66% of the 3A4 quantitative test set observed "low" ($pIC_{50} < 5$).
2. Model performs better when predictions fall in the extreme low or high regions.
3. Classify compounds as "low" with a high success rate (>93%) if predicted $pIC_{50} < 4.8$.
4. Classify compounds as "high" with a high success rate (>88%) if predicted $pIC_{50} > 5.6$.

Key descriptors associated with CYP inhibition

There is significant overlap between the important descriptors identified using the RT and PLS modeling techniques used herein. To simplify matters we only report the PLS descriptor (mean centered & scaled) coefficients as they are more interpretable than the VIP output from CART 4.0 (Table 5). Identification of the key molecular features driving each of the QSAR models should allow us to derive a set of general rules that may be used to reduce the liability of a compound for a given isoform (Table 6). It should be noted that an alteration in one molecule, such as adding a basic group to a neutral molecule, may lead to a problem with another isoform given the interdependency on molecular structure. Furthermore, these guidelines were determined using relatively small datasets (170–463), and although many times larger than comparable literature studies, the conclusions should be taken as a general guide only.

If we take CYP2C9 as an example, it is apparent from Table 5 that as lipophilic character increases so too will 2C9 potency. This is because the model identifies descriptors such as AROM (aromaticity, dependent on number of SP2 carbon atoms), PIAT (number of aromatic, pi atoms, CLOGP/ACDLogD (true lipophilicity measures), and Lipinski failures (i.e., failures have increased MW, reduced polarity etc.) with a net positive coefficient. Pos-charge (indicator variable for positive charge) and charges (indicator variable for any charge type) are identified as having a negative effect on potency suggesting 2C9 disfavors basic compounds. However, a negative charge indicator is not identified explicitly, yet we do find MMPWCT (proportional to max negative charge on molecule). This descriptor has a net negative value for negatively charge molecules such as acids, so a negative coefficient means it has a positive effect on potency. Thus from this QSAR it can be seen that 2C9 generally favors lipophilic, neutral, and acidic molecules, but not bases, which is in line with reports elsewhere [48].

In general: The results show that increasing lipophilicity will lead to an overall increase in potency for 1A2, 2C9, 2C19 and 3A4 but not 2D6. A lipophilicity variable is not explicitly used in the 3A4 model but many surrogates are present. 2D6 does not show a positive or negative dependence on lipophilicity. This may be a manifestation of the fact that 2D6 almost exclusively favors basic compounds, and that recognition is much more important than bulk properties when assessed over a range of chemistries. We have already highlighted that this may be the reason that the 2D6 model is the poorest fitted based on bulk properties.

Neutral molecules are all expected to bind to each of the receptors, however, each distinct P450 isoform will have different preferences for different ionization states due to the considerable differences in their amino-acids sequences, particularly those within, or near to their active sites [49–52]. Thus, we can conclude, based on the presence of particular descriptors in the QSAR models, that 1A2 and 2C9 can preferentially accommodate acids while 2D6 and 3A4 can accommodate bases.

Table 5 Scaled and centered coefficients from the 5 PLS regression models

CYP1A2	Coefficient	CYP2C9	Coefficient	CYP2C19	Coefficient	CYP2D6	Coefficient	CYP3A4	Coefficient
CLOGP	0.2300	PIAT	0.1427	PIAT	0.0776	POSCH	0.0601	MM_VDW_EPP_AREA	0.1494
Negioniz	0.1619	AROM	0.1331	CLOGP	0.0694	POSCHARGED	0.0582	AROM	0.1394
SPEC_FLEX_BD	0.1354	CLOGP	0.1147	AROM	0.0684	CHARGES	0.0576	MM_VDW_EPP_SUM	0.1350
SPEC_SAS_NONPOL_AREA	0.1272	Lipinski	0.1080	NBX	0.0672	CHARGED	0.0557	PIAT	0.1318
MaxRing2	0.0680	ACDLogD65	0.1002	HAROM	0.0601	Posioniz	0.0468	Posioniz	0.1085
MaxRing1	0.0135	ACDLogD74	0.0964	ACDLogD74	0.0597	HMO_LUMO_energy	0.0445	HAROM	0.0922
Aver_neg_charge_GHMO	−0.0983	POSCH	−0.0492	ACDLogD65	0.0578	MM_QC	0.0433	Kier_Chi5c	0.0893
MineV2	−0.1211	CHARGES	−0.0504	Clorine_count	0.0547	Lipinski	0.0419	FLEX_BD	0.0868
MM_PCWT	−0.1238	CHARGED	−0.0534	MMQO	−0.0183	Min_eV2	−0.0447	MM_SAS_EPN_VAR	−0.0738
Kier_Chi3p	−0.1247	POSCHARGED	−0.0561	NBO	−0.0205	MM_RNCS	−0.0450	Polar-count/MW	−0.1036
NB_OH	−0.1252	MM_PCWT	−0.0979	Oxygen_count	−0.0205	MM_MAXNEG	−0.0635	CHARGES	−0.1066
Kier_Chi4p	−0.1310			NEGCH	−0.0280	MM_QMIN	−0.0635	MM_VDW_EPP_VAR	−0.1089
Kier_Chi4c	−0.1343			MM_VDW_EPN_VAR	−0.0347			NEGCH	−0.1172
Pos_ioniz	−0.1448			MM_PCWT	−0.0445				
HMO-LUMO_energy	−0.1465			HMO_pi_energy	−0.0572				
M3M	−0.1848								
Min_eV1	−0.2200								

The descriptors have described elsewhere [16]

Table 6 Typical characteristic of inhibitors of P450s isoforms and the number of observations used in the training set of each QSAR model

P450 Isoform	Inhibitor characteristics	*N*
1A2	Aromatic, lipophilic, neutral, and acidic compounds	301
2C9	Aromatic, lipophilic, neutral, and acidic compounds	457
2C19	Aromatic, lipophilic, and neutral compounds	369
2D6	Aromatic, basic compounds	170
3A4	Aromatic, lipophilic, neutral, and basic compounds	463

All the 5 P450 isoforms contain descriptors that point to the presence of pi-stacking interactions in the active site. Reducing the number of aromatic moieties on compounds should in general reduce exposure to all P450s.

Accounting for molecular recognition in QSAR models

The moderate predictive ability of the bulk property QSAR models is likely to be a result of the complicated nature of P450 inhibition which consists of recognition events, direct binding to the heme and potentially inhibition by a metabolites formed. Given these complexities it is not unsurprising that we cannot accurately predict this phenomenon in a global sense when compared to other ADME properties such as solubility, permeability, CNS penetration etc.

We now discuss our efforts to account for these factors using substructural fragments which have been shown to be useful in ADMET applications elsewhere [53]. The choice of fragments over other geometric, electrotopological or pharmacophoric descriptors is that they are inherently simple and interpretable when used in conjunction with regression based techniques in that the coefficient associated with the fragment estimates the intrinsic liability of that fragment above the other properties in the molecule in question. For example, independent of other parameters it is expected that a naked pyrindine, imidazole or other heterocycles will have greater P450 liability for a given set of molecular properties because or their ability to bind to the heme iron. The most facile, albeit inelegant way to ensure one of the descriptors can encode this piece of "known" SAR is to include a fragmental descriptor containing this structural feature.

Fragment QSAR models

Algorithm builder fragments were used to generate PLS and RT models to investigate if a purely fragmental approach would provide as good or a better description of inhibition compared to those derived from standard bulk-property descriptors. Models were generated outside Algorithm Builder using only the key fragments as described in the computational details sections. The differences in the number of descriptors used in the RT and PLS models arises as SIMCA automatically excludes variables with low variance before building. This occurs for some variables (fragments) because they may be present in the test set in large numbers. While RT does not automatically exclude these descriptors, they cannot play a significant role in the model building process as they have low variance.

The PLS-fragment models are generally over-fitted compared to standard descriptor models as can be seen from a comparison of the r^2/q^2 ratio in the training sets (Tables 3, 7). In all cases the PLS models seem more overfitted than their standard continuous descriptor counterpart as evidenced by the poorer performance in cross validation. It is likely that PLS cross validation procedures are not sufficiently rigorous to lead to a conservative model using categorical descriptor data, in contrast to continuous molecular descriptors. This is likely to be a result of the significant number of fragment descriptors used and the fact that in a few cases relatively small numbers of molecules contain any one fragment, resulting in sparsely occupied descriptor matrices (many zero's few ones). This effect has also been reported in COMFA studies where many descriptors of a similar nature are used in conjunction with PLS [54].

The algorithm builder fragments used in conjunction with RT are generally more predictive on the independent test set, based on r^2 and RMSEs, than the PLS-AB models or our standard (STD) PLS or RT models (Table 7). Apart from one isoform where the performance, on the test set is poorer, the fragmental method demonstrates an improved ranking capability, with r^2s of between 24 and 80% over the best PLS or RT STD descriptor model. The only model not better predicted using a wholly fragmental model is 1A2 where we find the PLS-STD model does better. 2C19 and 2D6 appear to have a greater reliance on molecular recognition, since the variance explained in prediction using a fragmental approach almost doubles compared to that with standard descriptors. 3A4 is also significantly better predicted using fragments in conjunction with RT with an r^2 approaching 0.50.

Apart from 1A2, these results suggest fragmental methods can offer a performance advantage. However, a problem arises using fragmental descriptors when new compound are synthesized containing fragments not represented in the training set. The fragmental approach might therefore be expected to deteriorate more significantly in temporal prediction.

Table 7 PLS and RT training and test set results generated using AB fragmental descriptors only

	Training set results						Quantitative test set results				
PLS (AB)	r^2	r_0^2 (q^2)	RMSE	N	σ_Y	Desc./Comp.	r^2	r_0^2	RMSE	N	σ_Y
1A2	0.66	0.66 (0.20)	0.40	281	0.69	284/1	0.15	0.10	0.72	117	0.77
2C9	0.72	0.72 (0.35)	0.35	447	0.61	308/3	0.32	0.31	0.57	157	0.60
2C19	0.63	0.63 (0.28)	0.37	355	0.67	466 / 4	0.24	0.17	0.54	197	0.69
2D6	0.77	0.77 (0.32)	0.34	164	0.70	210/3	0.19	0.05	0.58	89	0.60
3A4	0.68	0.68 (0.31)	0.41	440	0.73	409/3	0.35	0.33	0.59	177	0.73
	Training set results						Quantitative test set results				
RT (AB)	r^2	r_0^2	RMSE	N	σ_Y	Desc./Trees	r^2	r_0^2	RMSE	N	σ_Y
1A2	0.76	0.70	0.38	281	0.69	330	0.15	0.14	0.71	117	0.77
2C9	0.77	0.75	0.34	447	0.61	359	0.36	0.36	0.55	157	0.60
2C19	0.76	0.69	0.34	355	0.67	533	0.22	0.21	0.53	197	0.69
2D6	0.72	0.68	0.40	164	0.70	236	0.24	0.23	0.52	89	0.60
3A4	0.85	0.81	0.32	440	0.73	461	0.49	0.47	0.53	177	0.73
	Training set results						Quantitative test set results				
Consensus	r^2	r_0^2	RMSE	N	σ_Y	Models	r^2	r_0^2	RMSE	N	σ_Y
1A2	–	–	–	–	–	2	0.20	0.19	0.69	117	0.77
2C9	–	–	–	–	–	2	0.29	0.28	0.58	157	0.60
2C19	–	–	–	–	–	2	0.13	0.11	0.56	197	0.69
2D6	–	–	–	–	–	2	0.09	0.07	0.58	89	0.60
3A4	–	–	–	–	–	2	0.34	0.33	0.59	177	0.73

PLS-RT consensus result for the test set are also given. r_0^2 is the correlation coefficient to the line of unity, q^2 is the cross validated r_0^2 from a leave many out procedure and RMSE is the root-mean-square error in prediction: r^2 is Pearson's correlation coefficient squared derived from the line of best fit, N is the number of observations and σ_Y is the standard deviation

Furthermore, it is apparent that a consensus prediction involving fragment models built using RT and PLS does not improve our overall ability to predict P450 inhibition, with as many models performing worse as better. This is in contrast to the models built using bulk properties.

Hybrid fragment-molecular descriptor QSAR models

Leadscope was used to determine if compounds with particular structural fragments had a different effect on potency compared to the mean as previously described. The identified fragments, heterocycles such as imidazole and pyridine for example, were then used to build new QSAR models in conjunction with the standard molecular descriptors in the hope that a better description of CYP inhibition would be afforded. An added benefit of this hybrid approach is that the absence of a fragment in a temporal test set would not be as significant a problem as the molecular descriptors used in the model would allow us to make a prediction based solely on its overall bulk molecular characteristics.

The results on the test set (Table 8) show that a combination of fragments and traditional descriptors has a noticeable increase in the rank ordering ability (r^2) above those built using our standard descriptors (Table 3). In all cases the hybrid fragment-STD descriptor models show an improvement on the test sets of between 6 and 110% when compared to the best PLS or RT STD descriptor model. The RT-LS-models are generally the most predictive based on r_0^2, apart from 3A4 and 2C19 where the PLS-LS is more accurate. We compare the models using the r_0^2 here, rather than r^2, as there is a significant discrepancy between the two for the RT 2C9, 2C19, and 3A4 models. Analysis of the mean errors reveals this is a result of the models significantly overpredicting the data (-0.14 to -0.18 log units on average).

A comparison of the models built thus far using three different sets of descriptors is somewhat complicated by the fact that some compounds are missing from the LS-test set as they failed to calculate all descriptors. We can make a comparison between the different sized test sets using the RMSE, rather than the r^2/r_0^2 as the standard deviations of the observed values can differ quite significantly. The collective results suggest that 1A2 is best predicted using

Table 8 PLS and RT training and test set results generated using bulk property descriptors and leadscope derived fragmental descriptors

	Training set results						Quantitative test set results				
PLS (LS)	r^2	r_0^2 (q^2)	RMSE	N	σ_Y	Desc./Comp.	r^2	r_0^2	RMSE	N	σ_Y
1A2	0.234	0.23 (0.11)	0.59	283	0.675	28/3	0.21	0.21	0.66	118	0.74
2C9	0.329	0.33 (0.29)	0.538	342	0.612	30/2	0.31	0.29	0.56	154	0.59
2C19	0.16	0.16 (0.11)	0.538	342	0.612	30/2	0.09	0.06	0.57	187	0.67
2D6	0.287	0.29 (0.24)	0.584	160	0.694	12/1	0.14	0.09	0.58	87	0.61
3A4	0.368	0.37 (0.29)	0.58	418	0.73	37/3	0.43	0.43	0.56	168	0.74
	Training set results						Quantitative test set results				
RT (LS)	r^2	r_0^2	RMSE	N	σ_Y	Desc./Trees	r^2	r_0^2	RMSE	N	σ_Y
1A2	0.791	0.688	0.376	283	0.675	214/15	0.26	0.25	0.64	118	0.74
2C9	0.689	0.593	0.402	407	0.657	215/15	0.25	0.20	0.60	154	0.59
2C19	0.681	0.567	0.419	342	0.612	208/15	0.17	0.11	0.56	187	0.67
2D6	0.797	0.715	0.37	160	0.694	196/15	0.27	0.27	0.52	87	0.61
3A4	0.697	0.618	0.451	418	0.73	208/15	0.44	0.36	0.59	168	0.74
	Training set results						Quantitative test set results				
Consensus	r^2	r_0^2	RMSE	N	σ_Y	Models	r^2	r_0^2	RMSE	N	σ_Y
1A2	–	–	–	–	–	2	0.19	0.18	0.67	118	0.74
2C9	–	–	–	–	–	2	0.27	0.26	0.57	154	0.59
2C19	–	–	–	–	–	2	0.12	0.10	0.56	187	0.67
2D6	–	–	–	–	–	2	0.09	0.07	0.58	87	0.61
3A4	–	–	–	–	–	2	0.43	0.43	0.56	168	0.74

PLS-RT consensus result for the test set are also given. r_0^2 is the correlation coefficient to the line of unity, q^2 is the cross validated r_0^2 from a leave many out procedure and RMSE is the root-mean-square error in prediction: r^2 is Pearson's correlation coefficient squared derived from the line of best fit, N is the number of observations and σ_Y is the standard deviation

the PLS-STD model, 2D6 by the RT-LS model and 2C9, 2C19, and 3A4 by the RT-AB model.

Consensus models

A consensus of all the model predictions generated here was evaluated to see if this had any impact on our overall ability to discriminate between potent and non-potent compounds. We report the top models derived from a consensus of the RT and PLS models generated using the different descriptor sets. The PLS-STD model is also included for the purpose of comparison (Table 9). The results show that the PLS-STD model is generally the least accurate model of those reported. The STD descriptor models can be improved somewhat by taking a consensus of the RT and PLS predictions, however these models are still inferior to the individual or consensus models derived from models that contain descriptors that allow an element of recognition to be taken into account.

The benefits of consensus predictions are clearly apparent since the most predictive models we have produced all rely on such an approach (An improvement of between 30 and 120% rank ordering of the test set compared to the PLS STD descriptor model). Furthermore, the best performing individual models all encode elements of molecular recognition using fragments, 1A2 and 2C19 best described by the PLS-LS models and 2C9, 2D6, and 3A4 by the RT-AB models.

The best available literature comparison to these models can be found in the work by Byvatov et al. [55] who built an inhibition model for the 2C9 isoform using support vector machines in conjunction with pharmacophoric fingerprints. While one cannot accurately compare r^2s between datasets where the dataset activity variances are different, the r^2 of 0.55–0.63 obtained by Byvatov et al. on a diverse test set, or 0.36–0.45, on what was defined as a "GPCR" based set (the exact values differ depending on the model), appear more predictive than the models reported here ($r^2 \sim 0.28 - 0.40$). However, they remain less intuitive due to the complexity of the statistical method and the descriptors which may not be preferable in all situations.

In a second modeling exercise, O Brien et al. [56], reported a strongly performing classification based inhibition model on a large 2D6 dataset using pharmacophoric fingerprints and molecular properties. This model proved

Table 9 Consensus model results for the overlapping compounds in the quantitative test set for the best performing models

	r^2
1A2 Model ($N = 110$, SD $Y = 0.756$)	
RT-PLS Consensus (AB+LS)	0.31
PLS (LS)	0.29
RT-PLS Consensus (std)	0.22
PLS (std)	0.23
2C9 Model ($N = 177$, SD $Y = 0.606$)	
RT-PLS Consensus (AB+LS)	0.41
RT (AB)	0.34
RT-PLS Consensus (std)	0.29
PLS (std)	0.28
2C19 Model ($N = 143$, SD $Y = 0.600$)	
RT-PLS Consensus (AB+LS)	0.30
PLS-AB	0.26
RT-PLS Consensus (std)	0.17
PLS (std)	0.16
2D6 Model ($N = 86$, SD $Y = 0.721$)	
RT-PLS Consensus (AB+LS)	0.29
RT (LS)	0.27
RT-PLS Consensus (std)	0.15
PLS (std)	0.13
3A4 Model ($N = 158$, SD $Y = 0.666$)	
RT-PLS Consensus (AB+LS)	0.51
RT (AB)	0.49
RT-PLS Consensus (std)	0.42
PLS (std)	0.40

PLS STD is reported in all cases for the purpose of comparison

Table 10 CYP Isoform loadings for the two component, 5Y PCA model

P450 Isoform	*N*	Component 1	Component 2
1A2	170	0.125	0.527
2C9	229	0.577	0.064
2C19	225	0.605	0.072
2D6	145	0.069	− 0.817
3A4	230	0.53	− 0.214

$r^2X = 0.727$. The variance explained by components one and two is 0.471 and 0.256, respectively

predictive in a class based sense but was not reported to have a ranking capability, nor were details of the descriptor dependencies, making it difficult to make a meaningful comparisons between the models.

SAR similarities between the five CYP isoforms

A PCA model was built on the reported pIC_{50}'s for all compounds that had at least 4 quantitatively measured

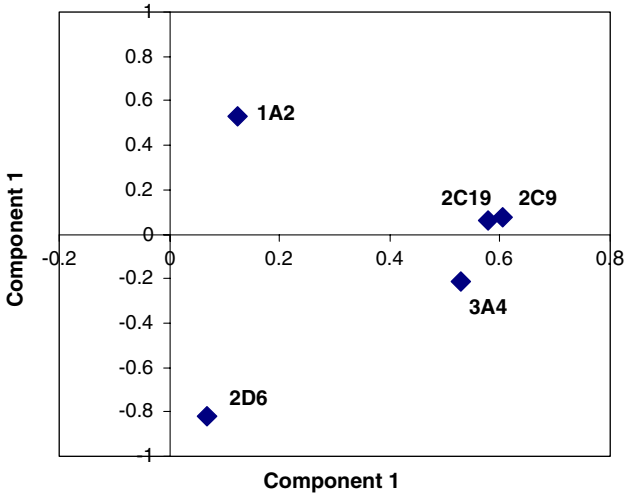

Fig. 4 PCA loadings plot for 5 CYP isoforms built using 237 observations. Isoforms that are strongly related are found close in both dimensions (i.e., 2C19, 2C9, and 3A4). In total, 213 compounds are common to 2C9, 2C19, and 3A4 and an analysis shows r^2 of 0.36 for 2C9–2C19, 0.22 for 2C9–3A4 and 0.16 for 2C19–3A4

values in each of the 5 P450 assays. Such an analysis allows us to compare how strongly each isoforms correlates with each other given that the inhibitors in the dataset are common to all. If for example compounds are active in more than one isoform quite frequently the PCA methodology will rank the isoforms as being similar and give them a similar loading on the extracted components (Table 10)

The optimal PCA model required two component to describe 73% of the variance in the 5 variables of the 237 observations. The first component describes 47% of the variance and the second 26%. Analysis of the loadings plot shows that 2C19, 2C9, and 3A4 are very similar with respect to component 1 but 3A4 is split somewhat from the two more similar isoforms 2C9 and 2C19 based on the second component (Fig. 4). The other two isoforms, 1A2 and 2D6 are found to have little commonality to each other or 2C9, 2C19, and 3A4. This is because both have a near zero loading on the first component and are split from each other, and 2C9, 2C19, and 3A4 based on the second component.

These results suggest that if a compound is active in 2C9 it may also have some activity at 2C19 and 3A4. The results also suggest that activity at 2D6 or 1A2 is unlikely to be accompanied by significant activity at any of the other cytochrome P450s.

Conclusions

Development of QSAR models for CYP inhibition was undertaken using both linear and non-linear statistical methods and a set of easily interpretable descriptors. The

models range from low to moderate predictivity, but all perform considerably better than random and thus could prove useful in assessing the P450 liability of molecules for a particular isoform. The 3A4 in-silico model is considerably more predictive than other isoforms, possibly a result of its more open active site, giving rise to less stringent molecular recognition requirements. In contrast, only poorer models could be produced for 2D6 and 2C19 using bulk molecular descriptors alone suggesting molecular recognition may play a more important role for these.

This study highlights that our ability to globally describe ADMET parameters that require a degree of molecular recognition is limited, at least where a clear understanding is also required. This is because we are using traditional QSAR descriptors that do not encode direct information about the active site(s) in question. We also find that combining predictions from a number of different models, built on different descriptors sets and/or using different statistical methodologies can be beneficial. Fragment based descriptors have been identified here as having utility, either in conjunction with traditional descriptors or as part of a consensus scheme.

References

1. Rendic S, Di Carlo FJD (1997) Drug Metab Rev 29:413
2. Guengerich FP (2002) Drug Metab Rev 4:7
3. Lin JH, Lu AYH (1998) Clin Pharmacokinet 35:361
4. Bertz RJ, Granneman GR (1997) Clin Pharmacokinet 32:210
5. Rendic S, Di Carlo J (1997) Drug Metab Rev 29:413
6. Shimada T, Yamazaki H, Mimura M, Inui Y, Guengerich FPJ (1994) Pharmacol Exp Ther 270:414
7. Masimirembwa CM, Thompson R, Andersson TB (2001) Comb Chem High Throughput Screen 4:245
8. Soars MG, Gelboin HV, Krausz KW, Riley RJ (2003) Br J Clin Pharmacol 35:175
9. McGinnity DF, Griffin SJ, Moody GC, Voice M, Hanlon S, Friedberg T, Riley RJ (1999) Drug Metab Dispos 27:1017
10. Riley RJ, Martin IJ, Cooper AE (2002) Curr Drug Metab 3:527
11. Hanch C, Fugita TJ (1964) J Am Chem Soc 86:1616
12. Abraham MH, Le J (1999) J Pharm Sci 88:868
13. Platts JA, Abraham MH, Zhao YH, Hersey A, Ijaz L, Butina D (2001) Eur J Med Chem 36:719
14. Gleeson MP, Waters, Paine SW, Davis AM (2006) J Med Chem 49:1953
15. Gleeson MP (2007) J Med Chem 50:101
16. Bruneau P (2001) J Chem Inf Comput Sci 41:1605
17. Riley RJ, Parker AJ, Trigg S, Manners CN (2001) Pharm Res 18:652
18. Lewis DFV (2004) In Vitro 18:89
19. Lewis DFV, Modi S, Dickins M (2002) Drug Metab Rev 34:69
20. Susnow RG, Dixon SL (2003) J Chem Inf Comput Sci 43:1308
21. Ekins S, De Groot MJ, Jones JP (2001) Drug Metab Dispos 29:936
22. Vaid TP, Lewis NS (2000) Bioorg Med Chem 8:795
23. Rao S, Aoyama R, Schrag M, Trager WF, Rettie A, Jones JP (2000) J Med Chem 43:2789
24. Smith DA, Ackland MJ, Jones BC (1997) DDT 2:479
25. Lewis DFV, Jacobs MN, Dickins M (2004) DDT 9:530
26. Hatch FT, Lightstone FC, Colvin ME (2000) Environ Mol Mutagen 35:279
27. Lesigiarska I, Pajeva I, Yanev S (2002) Xenobiotica 32:1063
28. Jones JP, He M, Trager WF, Rettie AE (1996) Drug Metab Dispos 24:1
29. De Groot MJ, Ackland MJ, Horne VA, Alex AA, Jones BCJ (1999) Med Chem 42:4062
30. Wold H (1975) Quantitative sociology: international perspectives on mathematical and statistical model building. Academic Press, New York, pp 307–357
31. Japertas P, Didziapetris R, Petrauskas A (2002) QSAR 21:23
32. Afzelius L, Zamora I., Masimirembwa CM, Karlen A, Andersson TB, Mecucci S, Baroni M, Cruciani GJ (2004) Med Chem 47:907
33. Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE Jr (2000) J Chem Inf Comput Sci 40:1302
34. Höskuldsson A (1996) Prediction methods in science and technology. Thor Publishing, Copenhagen, Denmark
35. Wold S, Albano C, Dunn WJ, Edlund U, Esbensen K, Geladi P, Hellberg S, Johansson E, Lindberg W, Sjöström M (1984) In: BR Kowalski BR, Chemometrics: mathematics and statistics in chemistry. D. Reidel Publishing Company, Dordrecht, Holland
36. Wold S, Eriksson L, Sjöström M (2000) PLS in chemistry, encyclopedia of computational chemistry. Wiley, New York
37. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Chapman and Hall, New York
38. Chohan KK, Paine SW, Mistry J, Barton P, Davis AM (2005) J Med Chem 48:5154
39. Asikainen AH, Ruuskanen J, Tuppurainen K (2004) SAR QSAR in Environ Res 15:19
40. Baurin N, Mozziconacci JC, Arnoult E, Chavatte P, Marot C, Morin-Allory L (2004) J Chem Inf Comput Sci 44:276
41. Jackson JE (1991) A user's guide to principal components. John Wiley, New York
42. Wold S, Geladi P, Esbensen K, Öhman J (1987) J Chemom 1:41
43. GOLPE: Multivariate Infometric Analysis Srl., Viale dei Castagni 16, Perugia, Italy. http://www.miasrl.co
44. CART 4.0. Salford Systems, 8880 Rio San Diego Dr., STE. 1045, San Diego, California, 92108. http://www.salford-systems.com
45. Rodgers SL, Davis AM, van de Waterbeemd H (2007) QSAR Comb. Sci 26:511
46. Gleeson MP, Waters NJ, Paine SW, Davis AM (2006) J Med Chem 49:1953
47. Gavaghan CL, Arnby CH, Blomberg N, Strandlund G, Boyer SJ (2007) Comp Aided Mol Des 21:189
48. Lewis FV, Modi S, Dickins M (2001) Drug Metabol Drug Rev 18:18, 221
49. Williams PA, Cosme J, Ward A, Angove HC, Vinkovic DM, Jhoti H (2003) Nature 424:464
50. Rowland P, Blaney FE, Smyth MG, Jones JJ, Leydon VR, Oxbrow AK, Lewis CJ, Tennant MG, Modi S, Eggleston DS, Chenery RJ, Bridges AM (2006) J Biol Chem 281:7614
51. Williams PA, Cosme J, Vinkovic DM, Ward A, Angove HC, Day PJ, Vonrhein C, Tickle IJ, Jhoti H (2004) Science 30:683
52. Yano JK, Wester MR, Schoch GA, Griffin KJ, Stout CD, Johnson EF (2004) J Biol Chem 279:38091
53. Leach AG, Jones HD, Cosgrove DA, Kenny PW, Ruston L, MacFaul P, Wood JM, Colclough N, Law B (2006) J Med Chem 49:6672
54. Golbraikh A, Tropsha A (2002) J Mol Graph Model 20:269
55. Byvatov E, Baringhaus K, Schneider G, Matter H (2006) QSAR Comb Sci 26:618
56. O Brien SE, de Groot MJ (2005) J Med Chem 48:1287