J Comput Aided Mol Des (2007) 21:189–206
DOI 10.1007/s10822-006-9095-6

ORIGINAL PAPER

Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data

Claire L. Gavaghan · Catrin Hasselgren Arnby · Niklas Blomberg · Gert Strandlund · Scott Boyer

Received: 31 August 2006 / Accepted: 2 December 2006 / Published online: 24 March 2007

Abstract A 'global' model of hERG K^+ channel was built to satisfy three basic criteria for QSAR models in drug discovery: (1) assessment of the applicability domain, (2) assuring that model decisions can be interpreted by medicinal chemists and (3) assessment of model performance after the model was built. A combination of D-optimal onion design and hierarchical partial least squares modelling was applied to construct a global model of hERG blockade in order to maximize the applicability domain of the model and to enhance its interpretability. Additionally, easily interpretable hERG specific fragment-based descriptors were developed. Model performance was monitored, throughout a time period of 15 months, after model implementation. It was found that after this time duration a greater proportion of molecules were outside the model's applicability domain and that these compounds had a markedly higher average prediction error than those from molecules within the model's applicability domain. The model's predictive performance deteriorated within 4 months after building, illustrating the necessity of regular updating of global models within a drug discovery environment.

C. L. Gavaghan (✉) · C. H. Arnby · S. Boyer
Computational Toxicology, Safety Assessment,
AstraZeneca R&D, 431 83 Molndal, Sweden
e-mail: claire.gavaghan@astrazeneca.com

N. Blomberg
Global DECS Computational Chemistry, AstraZeneca
R&D, 431 83 Molndal, Sweden

G. Strandlund
Lead Generation, AstraZeneca R&D, 431 83 Molndal,
Sweden

Keywords hERG · Hierarchical PLS modelling · Onion design · QSAR

Introduction

It is now recognized that some drug-induced sudden deaths are secondary to the development of an arrhythmia called Torsades de Pointes (TdP). Recent advances in the understanding of this issue indicate that the primary event is likely to be inhibition of the rapid component of the delayed rectifying potassium current (I_{Kr}) by such agents. These compounds bind to the pore-forming α sub-units of the channel protein carrying this current—sub-units that are encoded by the human ether-a-go-go-related gene (hERG). Since I_{Kr} plays a key role in repolarization of the cardiac action potential, its inhibition slows repolarization and this is manifested as a prolongation of the QT interval. Whilst QT interval prolongation is not a safety concern per se, in a small percentage of people it is associated with TdP and degeneration into ventricular fibrillation [1–6]. This safety issue has led to the withdrawal of a number of drugs from the market (e.g. terfenadine, cisapride and grepafloxacin) and to specific regulatory guidance (ICH S7B and ICH E14). Consequently, screening of compounds for hERG activity has commonly become part of the early drug discovery process [7]. However, experimental assessment of hERG inhibition is time consuming and costly, hence there is a need for reliable in silico predictive tools to aid this screening process in the selection and optimization of drug candidates.

There are a variety of different approaches for modelling and predicting potential hERG inhibitors that have been reported and these have been recently reviewed by Aronov [8]. The pharmacophore models generated by Cavalli et al. [9], Ekins et al. [10], Perlstein et al. [11] and Aronov [12] have provided information based on a range of QT prolonging drugs. A number of homology models have also emerged derived from the KcsA, MthK and KvAP crystal structures [11, 13, 14], a variety of methods have been investigated (neural networks, genetic programming, support vector machines, step-wise and multivariate regression methods) and a range of different descriptors have been utilized (including fragment-based, 2D and 3D molecular property descriptors) to generate predictive QSAR models for hERG inhibition [12, 15–19, 11, 20, 21].

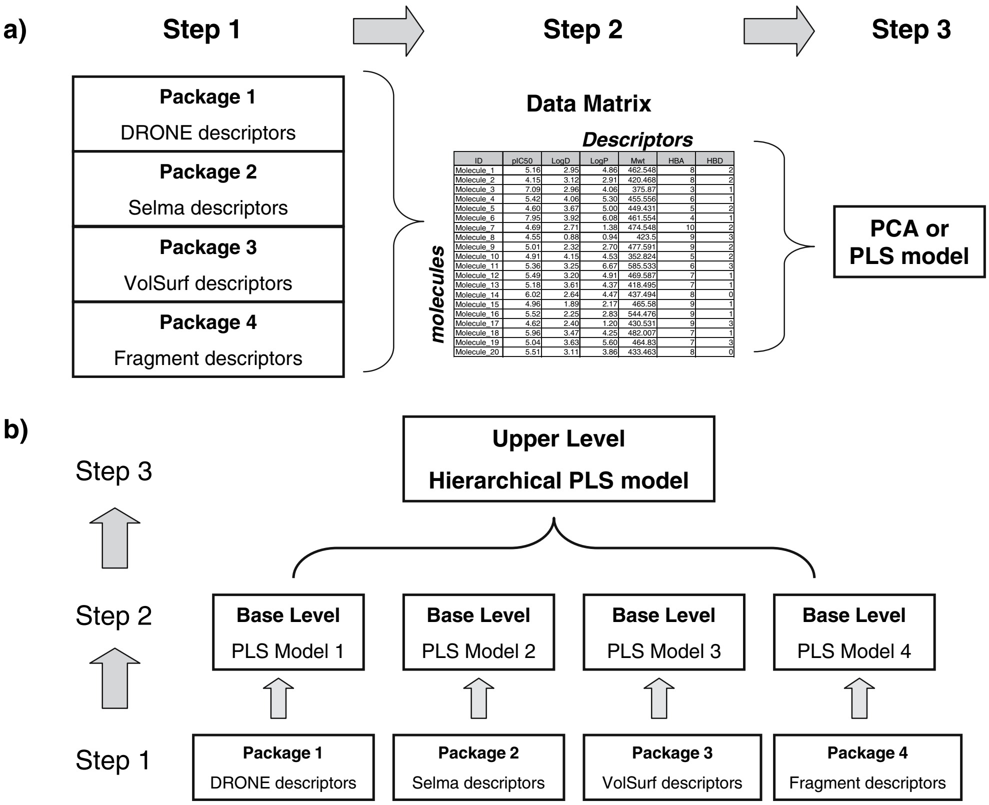
Many of the published models of hERG inhibition have provided valuable information regarding possible binding modes into the hERG channel via pharmacophore models and an appreciation for the general physicochemical properties that drive hERG blockade. Most of the published models are also proposed as "global" models for hERG blockade. These models were "global" in that their predictive range was not restricted to a particular chemical class or structural series but instead were developed to predict molecules with broad structural and chemical diversity. In order to successfully apply a global model in a drug discovery setting it is necessary to demonstrate, with confidence, the model's ability to accurately predict a diverse range of molecules and assure the model's applicability domain and quality by continued revalidation. Furthermore, to qualify as a useful prediction tool in drug discovery, the model and its predictions should be easily interpretable to the medicinal chemist.

In this report, we report one approach for using a global hERG model in a real drug discovery context. The model was "global" in that the compounds included in model building spanned across structural series and represented a diverse range of molecular properties. The model was developed for supporting projects prioritizing both compounds to be synthesized and which compounds may have potential hERG liabilities and should be assayed. Three aspects of this problem are addressed: measuring and maximizing the model's applicability domain; QSAR model interpretation (both for explaining model predictions and diagnosing model weaknesses) and monitoring of model performance over time (temporal evaluation), as new projects with new chemical series are added.

A model's applicability domain is determined by the training data used in its development. Within a pharmaceutical company, predictions are required for molecules that may have shifted from the original applicability domain of the QSAR model and this is a particular concern when developing global models. Ideally, a global model needs to adequately represent the range of chemical properties and structural diversity a medicinal chemist may pursue when designing new molecules but this aim is difficult to achieve in a dynamic field, such as drug design. Therefore, it is necessary to have adequate measures that are capable of alerting the medicinal chemist when a new query molecule is outside of the model's applicability domain. In the current work, we have evaluated a diagnostic parameter commonly used in PLS, distance to model (DModX), as an indicator of applicability domain. The original training dataset contained 1,312 molecules originating from projects in many therapy areas. In order to maximize the applicability domain of the resulting model, a designed training set, utilizing an onion design coupled with D-optimal design, was selected. The size of the model's applicability domain was evaluated with reference to the domain coverage of models built using randomly selected training sets.

To enhance the interpretability of our model, hierarchical modelling was used to construct the global PLS model. Hierarchical modelling is a useful approach when modelling QSAR data containing many descriptors by PLS compared to "regular" PLS modelling and the two approaches are summarized (Fig. 1) [22]. Like most large drug companies, we have access to a number of different molecular descriptor packages, each of which includes different combinations of 2D, 3D, charge and structural fragment-based descriptors. A hierarchical approach was used to facilitate the PLS model building process as it was not known which of these packages, if any, would yield the optimal predictive model or whether the packages could provide complimentary information in the PLS model. The benefits of using hierarchical modelling, including hierarchical PLS, have been demonstrated in modelling other pharmaceutical problems including HIV-1 protease inhibitors [23], carcinogenicity [24] and mutagenicity [25]. The benefits of coupling onion and D-optimal designs for selecting a training set when modelling global datasets containing diverse chemical structures, combined with hierarchical PLS modelling was recently illustrated by Eriksson et al. [26] who used this approach to develop a global dataset of P450 3A4 inhibition and we have adopted the method described therein.

In addition to using a hierarchical modelling for simplifying model interpretation, we also developed new hERG specific structural fragment-based descrip-



ID	pIC50	LogD	LogP	Mwt	HBA	HBD
Molecule_1	5.16	2.95	4.86	462.548	8	2
Molecule_2	4.15	3.12	2.91	420.468	8	2
Molecule_3	7.09	2.96	4.06	375.87	3	1
Molecule_4	5.42	4.06	5.30	455.556	6	1
Molecule_5	4.60	3.67	5.00	449.431	5	2
Molecule_6	7.95	3.92	6.08	461.554	4	1
Molecule_7	4.69	2.71	1.38	474.548	10	2
Molecule_8	4.55	0.88	0.94	423.5	9	3
Molecule_9	5.01	2.32	2.70	477.591	9	2
Molecule_10	4.91	4.15	4.53	352.824	5	2
Molecule_11	5.36	3.25	6.67	585.533	6	3
Molecule_12	5.49	3.20	4.91	469.587	7	1
Molecule_13	5.18	3.61	4.37	418.495	7	1
Molecule_14	6.02	2.64	4.47	437.494	8	0
Molecule_15	4.96	1.89	2.17	465.58	9	1
Molecule_16	5.52	2.25	2.83	544.476	9	1
Molecule_17	4.62	2.40	1.20	430.531	9	3
Molecule_18	5.96	3.47	4.25	482.007	7	1
Molecule_19	5.04	3.63	5.60	464.83	7	3
Molecule_20	5.51	3.11	3.86	433.463	8	0

Fig. 1 Schematic illustrating the differences between the PLS and hierarchical PLS modelling approaches used. (**a**) PLS model: The molecular descriptors are calculated for each molecule in the training set (Step 1) and these are assembled into a matrix (step 2). The PLS model is calculated using all of the molecular descriptors. (**b**) Hierarchical PLS model: The molecular descriptors are calculated for each molecule in the training set (step 1) and separate base level PLS models are calculated for each descriptors package (step 2). The scores are taken from the base level models, combined and used as descriptors to generate an upper level PLS model

tors and examined the benefits and pitfalls of using these as interpretable molecular descriptors in our global hierarchical PLS model of hERG blockade. The application of fragment-based descriptors in QSAR models of hERG blockade have previously been presented by Bains et al. [15] and more recently by Song and Clark [27].

The majority of the published models of hERG inhibition have proven predictive power by measuring prediction accuracy of a test dataset (external validation). However, the test sets have typically been small in number because they have relied upon available published hERG inhibition data, comprising of data originating from a variety of mammalian cell lines [28]. Therefore, although many promising models have been published, it is difficult to gauge how reliably these models would predict in a drug discovery environment. The QSAR model presented here utilized a large dataset for model training and validation, comprising in total of 8,832 unique molecules with measured hERG IC50 values. The biological data generated were high quality (IC50 measurements were averaged from repeated measurements) and the data were obtained from a functional assessment of hERG activity using a single method developed using the the IonWorks$^{\mathrm{TM}}$ HT high throughput electrophysiology assay [29].

We were interested in investigating whether internal and external validation methods do serve as an adequate measure of a model's predictive power in a real drug discovery setting. Internal validation methods are an attractive alternative for determining whether a fitted model is predictive when external validation is not possible and we have attempted to investigate the reliability of two of these approaches (cross-validation and response permutation testing) as indicators of future model predictions and compared them to the external validation of the original test set. We then monitored predictive performance over a period of 15 months using test sets that were generated after the initial model development and validation and compared this to our initial estimates.

Clearly, previous published models of hERG blockade appear, in some cases, to be predictive and most have provided valuable information about the properties that drive hERG binding. Here, we present a novel approach for building a more interpretable QSAR model for hERG and demonstrate a process for evaluating model performance that can give end-users more confidence in the resulting predictions.

Results and discussion

Initial investigation by PCA

Initial investigation of the data by PCA indicated a large overlap between compounds that measured as active (pIC50 > 6), moderate (pIC50 between 5 and 6) and non-active (pIC50 < 5) inhibitors in the hERG IonWorks electrophysiology assay. A plot of the first two principal components showed that although some sub-clusters corresponding to research projects were present, the plot also showed some separation of highly active inhibitors (pIC50 > 6) from non-active (pIC50 < 5) inhibitors (Fig. 2). Consequently, an onion design, coupled with D-optimal design for each layer, was used to select the training and test sets in order to ensure maximum coverage of the structural diversity in the dataset and the feasibility of a global PLS model was investigated.

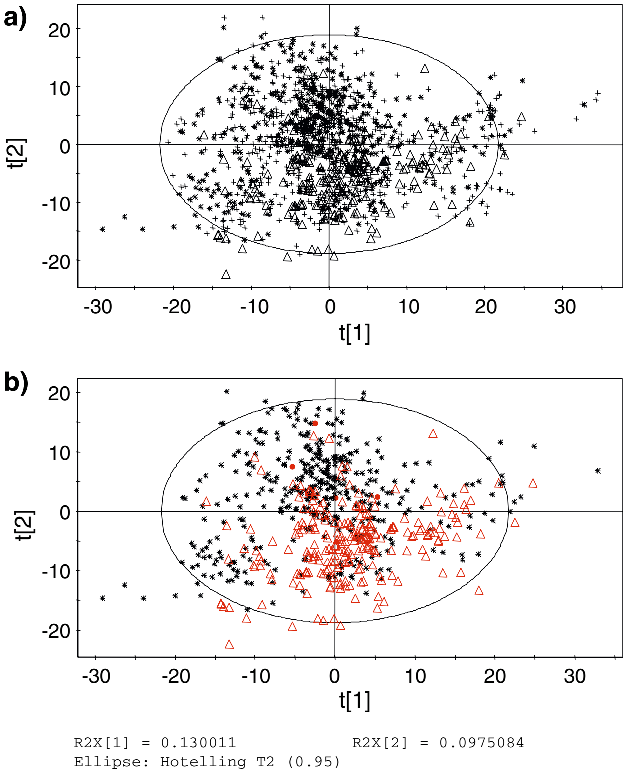


Fig. 2 Scores plot of reference PCA model. Plot (b) is the same plot as plot (a) except that the observations belonging to the moderate class have been removed to show the abundance of high pIC50 compounds in the lower half of the plot (indicated by the dotted line). *Key*: Observations are labelled according to their experimental pIC50 class. Star = pIC50 < 5; Cross = pIC50 between 5 and 6; open triangle = pIC50 ≥ 6

Assessment of training set selection using onion and D-optimal designs

A thorough comparison of training set selection methods was not performed, as this was not the main modelling objective. However, the training set selected from the statistical molecular design method, namely a coupling of onion and D-optimal designs, was evaluated by comparison to reference models generated from randomly selected training sets (Table 1). In total, six different reference models were generated from randomly selected training sets and their predictive performance was compared to a reference model derived from both training and test datasets (M1) and reference PLS and hierarchical PLS models derived from the designed training set molecules (M2 and M3). The initial training and test set RMSE results were similar for all of the models. This was also observed in the RMSE results for the temporal test sets that were collected up to 15 months after the models were generated. All models predicted the temporal datasets within a similar RMSE of 0.5. However, the number of molecules predicted within this RMSE range and the model boundary (distance to model) was notably different between the different models. Both of the models derived from the designed training set (M2 and M3) predicted the highest number of molecules within the model DModX boundary (4,982 and 4,902 respectively, representing 94 and 92% of the validation set). Neither the reference model (M1, Table 1), which used both training and test data to build the PLS model, nor the models generated from randomly selected training sets (M4–M9, Table 1) performed as satisfactorily (with the range of molecules predicted within the DModX boundary ranging between 3,869 and 4,795 respectively, representing 73 and 90% of the validation set). Consequently, more molecules were predicted by the "designed" models within an RMSE of 0.5 and the number of poorly predicted molecules was dramatically reduced.

The comparison also showed that random models are prone to producing models with varying prediction accuracies, even with the reasonably large dataset (n = 1,312) available in this study. Out of the six models generated, four (M4, M5, M7 and M9) produced reasonable results whilst two (M6 and M8), did not perform adequately with respect to the number of molecules predicted within the model distance. This indicated that these training sets consisted of molecules with different ranges of molecular properties and produced different representations of a hERG model.

Although the differences were not analysed to check for significance (as this was not the aim of this study)

Table 1 Comparison of training set selection approaches; onion and D-optimal designs versus random design

									RMSE of Temporal Test Set Data						
Model	Training set data included in PLS model	N	K	A	R^2X	R^2Y	Q^2	June 2004 Training Set RMSE (n=436)	Dataset 1 August 2004 (n=144)	Dataset 2 Oct 2004 (n=1,535)	Dataset 3 May 2005 (n=1,720)	Dataset 4 Sept 2005 (n=1,414)	Mean (Datasets 1-4)	Data within model boundary (number)	Data outside model boundary (number)
M1	All June 2004 data	1,312	585	6	0.35	0.64	0.53	0.40	0.39	0.54	0.55	0.50	0.50	0.48 (3,892)	0.64 (1,418)
M2	Onion & D-optimal Training set	436	585	4	0.30	0.62	0.54	0.43	0.42	0.54	0.53	0.50	0.50	0.50 (4,982)	0.80 (327)
M3	Onion & D-optimal hierarchical	436	13	2	0.35	0.61	0.59	0.44	0.45	0.54	0.54	0.52	0.51	0.52 (4,902)	0.71 (503)
M4	Random training set 1	656	585	4	0.30	0.62	0.52	0.42	0.42	0.53	0.55	0.51	0.50	0.51 (4,795)	0.67 (524)
M5	Random training set 2	656	585	5	0.34	0.63	0.49	0.40	0.43	0.54	0.56	0.51	0.51	0.51 (4,703)	0.71 (606)
M6	Random training set 3	656	585	4	0.30	0.63	0.46	0.42	0.42	0.53	0.55	0.53	0.51	0.51 (3,869)	0.58 (1,140)
M7	Random training set 4	656	585	4	0.30	0.60	0.42	0.43	0.45	0.54	0.53	0.54	0.52	0.51 (4,740)	0.69 (569)
M8	Random training set 5	656	585	4	0.30	0.62	0.53	0.42	0.44	0.54	0.55	0.54	0.52	0.52 (3,894)	0.60 (1,415)
M9	Random training set 6	656	585	4	0.30	0.58	0.48	0.43	0.44	0.53	0.53	0.52	0.51	0.50 (4,806)	0.70 (503)

Key: Step 1, The molecular descriptors for each of the training set molecules are calculated; Step 2, Separate base level PLS models are calculated for the different descriptor packages; Step 3, An upper level hiearchical PLS model is calculated from base level models' scores

these results show that, although the alternative approaches investigated for training set selection (random splitting and inclusion of all available data) can produce comparable predictive models, their performance shows more variability than a model built using statistical molecular design to select a ‘‘balanced'' training set.

Assessment of a hierarchical modelling approach

The main purpose of using a hierarchical approach to model this global dataset was to improve interpretation of the model whilst retaining predictive performance. Model performance statistics were compared between the individual PLS models calculated—the reference model including all descriptors and variables (M1, Table 2; the base level models of the each descriptor package (M2–M5 and M7–M10, Table 2) and the upper level PLS models (M6 and M11, Table 2). Additional data, referred to as the temporal data sets, were provided after the original models were calculated over a time period of 15 months. A comparison of the internal (Q^2 and $\mathrm{RMSE}_{\mathrm{training\ set}}$) and external ($\mathrm{RMSE}_{\mathrm{test\ set}}$ and $\mathrm{RMSE}_{\mathrm{temporal\ data\ sets\ 1–4}}$) statistics revealed that the hierarchical models' results were similar to the reference model and that the hierarchical models—which incorporated information from all descriptors - gave improved predictions relative to the individual PLS models based on only DRONE, Selma, VolSurf or fragment-based descriptors (Table 2).

To interpret ‘‘standard'' base level PLS models, it is common procedure to examine plots of the scores (information relating to the observations), loadings, weights and PLS coefficients to decipher the relationship between the descriptor variables (X matrix) and output variable(s) (Y matrix)—in this case, the relationship between the calculated chemical property and structural descriptors with hERG pIC50. Other diagnostics used for model interpretation include variable importance plots, which rank the descriptors according to their importance in the model with respect to both the descriptor (X) and response (Y) variables, and contribution plots, which allow for a comparison of differences between two or more observations.

The problem of interpreting the reference PLS model (M1, Table 2) becomes clear immediately

Table 2 Comparison of QSAR model performance statistics

								Original Dataset June 2004		Temporal Test Sets				
Model	Descriptors included in PLS model	N	K	A	R^2X	R^2Y	Q^2	Training Set RMSE (n=436)	Test Set RMSE (n= 876)	Dataset 1 RMSE August 2004 (n=144)	Dataset 2 RMSE Oct 2004 (n=1,535)	Dataset 3 RMSE May 2005 (n=1,720)	Dataset 4 RMSE Sept 2005 (n=1,414)	Mean RMSE
M1	All (reference model)	1,312	606	6	0.35	0.64	0.54	0.40	N.A.	0.38	0.54	0.55	0.50	0.49
M2	DRONE	1,312	127	7	0.69	0.44	0.36	0.50	N.A.	0.48	0.56	0.58	0.53	0.54
M3	SELMA	1,312	84	6	0.61	0.51	0.44	0.47	N.A.	0.46	0.53	0.58	0.56	0.53
M4	VOLSURF	1,312	78	3	0.31	0.37	0.33	0.53	N.A.	0.52	0.57	0.60	0.55	0.56
M5	Fragments	1,312	317	4	0.20	0.57	0.48	0.44	N.A.	0.46	0.57	0.59	0.50	0.53
M6	Scores from M2 to M5 (Hierarchical)	1,312	20	3	0.33	0.60	0.57	0.43	N.A.	0.42	0.52	0.53	0.48	0.49
M7	DRONE	436	127	4	0.57	0.37	0.23	0.56	0.53	0.51	0.55	0.57	0.53	0.54
M8	SELMA	436	84	4	0.52	0.49	0.37	0.51	0.50	0.55	0.58	0.63	0.64	0.60
M9	VOLSURF	436	78	2	0.24	0.31	0.23	0.59	0.53	0.57	0.58	0.62	0.56	0.58
M10	Fragments	436	307	3	0.16	0.59	0.41	0.45	0.48	0.49	0.55	0.55	0.48	0.55
M11	Scores from M7 to M8 (Hierarchical)	436	13	2	0.35	0.61	0.59	0.44	0.46	0.45	0.54	0.54	0.52	0.51

Key: *N*, number of compounds in the training set; *K*, number of descriptors; *A*, number of PLS components; R^2X, percentage of explained *X*-variation; R^2Y, percentage of explained *Y*-variation; Q^2, percentage of predicted *Y*-variation with estimated leave-one-out cross-validation

if these plots are examined. This model includes 606 chemical property and structural descriptors originating from four different sources, many of which are different representations (i.e., calculated by different methods) of the same properties, for example size, charge, polarity and lipophilicity. An advantage of PLS is that it can cope with including many descriptors in the model without affecting the model's predictive performance, as exemplified by the reference model, M1. However, it is both a time consuming and often misleading task to interpret this type of PLS model. By adopting a hierarchical PLS approach, interpretation is simplified, as illustrated in Fig. 3. This example, which uses a training set compound, Pimozide, illustrates how to decipher the structural and chemical properties that distinguish a single compound from the PLS model's average position in principal property space (i.e., how the combination of molecular properties for the query molecule differ to the average combination of properties represented by the training set). A scores plot (Fig. 3a) of the first two PLS components from the training model (M11, Table 2) showed separation of the non-active and active inhibitors of hERG (the $pIC50_{obs}$ for inhibiting the hERG ion channel increases from the left to right hand side of the plot). The plot indicates that Pimozide is an inhibitor for the hERG channel. Examination of the upper level model's score contribution plot for Pimozide (Fig. 3b) indicated that the scores from the lower level fragment-based model (M10, Table 2) were the most influential descriptors in the hierarchical PLS model. As this molecule was included in the model's training set, these fragment based descriptors are the main contributions to the observed pIC50 for this hERG inhibitor ($pIC50_{obs}$ = 7.9). This procedure can be applied to identify the main contributing descriptors for any new query molecules.

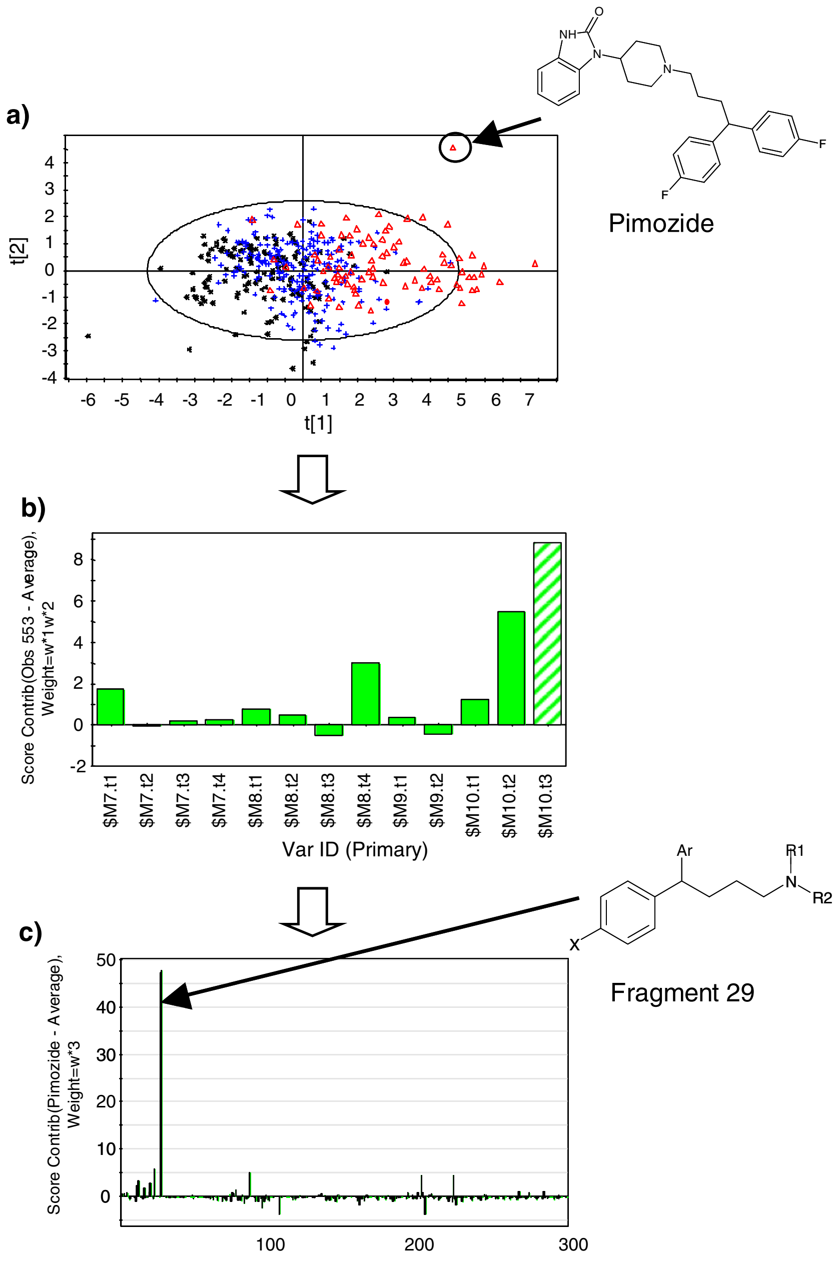


Fig. 3 Schematic illustrating the interpretation of a hierarchical PLS model. (**a**) The upper level scores plot shows separation of compounds with low (pIC50 < 5), moderate (pIC50 between 5 and 6) and high (pIC50 ≥ 6) inhibition in the hERG IonWorks assay. Pimozide is selected for investigation. (**b**) An upper level scores contribution plot for Pimozide indicates which of the underlying descriptor packages contributes to this observation's scores. (**c**) The base level contributions for each descriptor package can be examined for each of the upper level scores. For Pimozide, fragment descriptors were the main contributions to this compounds inhibition at hERG, with Fragment 29 ranking as the most influential descriptor

Although using a hierarchical approach simplified the interpretation of which descriptor set was most influential in the upper level model, it is still necessary to examine which of the original descriptors contribute in new predictions. It is possible to do this by examining the contributions from the base level models and this procedure is illustrated using Pimozide as an example (Fig. 3). Thus, to understand which structural fragment(s) contributed to the prediction of Pimozide as a potential inhibitor, the base level models' score contribution plots (as illustrated in Fig. 3b) were examined. This revealed that Pimozide contained a higher number of "hits" for fragment 29 relative to the rest of the training set. The global dataset (n = 8,832) was examined for the appearance of this fragment. This identified sixteen compounds containing fragment 29, with pIC50$_{obs}$ ranging between 4.2 and 7.9 (mean pIC50$_{obs}$ = 5.9) and the presence of a para-substituted phenyl connected to a tertiary amine via four linker atoms and to another phenyl, as appears in Pimozide, was identified in only six out of these sixteen compounds, with pIC50$_{obs}$ ranging between 5.5 and 7.9. Although the number of examples is small, these compounds did not originate from an analogous series, project or therapy area. Additionally, only two out of the sixteen compounds containing fragment 29 were non-inhibitors in the hERG IonWorks electrophysiology assay (pIC50$_{obs}$ < 5), suggesting that this fragment contributes to increased hERG inhibition. One of these non-inhibitors was not examined further because the molecule was outside of the model's applicability domain (as defined by the critical distance to model boundary) and therefore could not be confidently considered to be a reliable prediction. The other non-inhibitor, which was the marketed drug citirizine, was determined experimentally to be a non-active inhibitor ($pIC50_{obs} < 5$) by the hERG IonWorks assay although an accurate $pIC50_{obs}$ value was not obtained. The hierarchical PLS model calculated $pIC50_{pred}$ as 4.98 and as the prediction was within the model distance boundary, it was considered to be reliable. The prediction was examined by comparing how citirizine's predicted score contributions differed to the average score position of the training set compounds (which had a mean $pIC50_{obs} = 5.4$) using the approach illustrated in Fig. 3. This investigation revealed that although citirizine contains fragment 29, the low $pIC50_{pred}$ value could be rationalized by large score contributions from fragment descriptors containing a carboxylic acid group and Selma and DRONE descriptors describing the presence of a negative charge and a low calculated log*P*. This supports the existing knowledge of non-active properties thought to reduce a molecules blockade of the hERG channel, namely that introducing a carboxylic acid moiety and reducing log*P* have been found to be good strategies for reducing hERG inhibition.

Using this hierarchical approach simplifies the process of interpreting the underlying chemical and structural properties contributing to the predicted hERG activities of new molecules. The added advantage of including fragment-based descriptors into the QSAR model is that the descriptors are more interpretable than indirect structural descriptors, such as connectivity indices, for generating ideas regarding which structural modifications should be made to a molecule to reduce hERG inhibition. This approach can also be used to obtain information relating to ''outlier'' predictions. These predictions can originate from two sources—either they arise from extrapolations outside the model's applicability domain or they are erroneous predictions that are within the model boundary. Both of these types of outlier predictions are worth investigation. Molecules that are outside of the model's applicability domain may indicate a radical change in chemistry and although the model cannot be used for a reliable prediction, these outlier molecules may represent new desirable chemical series. Contrasting this, investigating erroneous predictions for molecules within the model's boundary can help to identify weaknesses in the model itself. Such analysis is an obvious extension to the current work and will be presented separately.

Development of structural fragment-based descriptors relevant to hERG activity

Fifteen fragments that were negatively correlated to hERG $pIC50_{obs}$ were selected from the original dataset ($n = 1{,}312$) using LeadScope®. Caution should be applied when using this singular, unsupervised approach for identification of fragments relating to biological effect, as the analysis is highly dataset dependent. Hence, the importance of these fragments was assessed by interpreting their importance in the hierarchical PLS model. The contribution of these fragments to hERG $pIC50_{obs}$ in the hierarchical PLS model was investigated using the procedure described previously (Fig. 3) by comparing the contributions from the score positions of the non-active group cluster to the model average (average $pIC50_{obs} = 5.2$). This was done for both training set compounds and the predicted scores of the temporal test sets. This analysis identified eight of these hERG specific fragments as relating to a low hERG $pIC50_{obs}$ value (Table 3) suggesting that the inclusion of these fragments in a molecule's structure may contribute towards decreasing the likelihood of hERG inhibition.

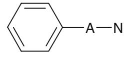
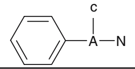
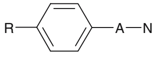
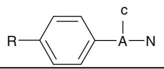
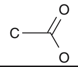
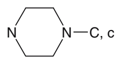
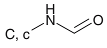
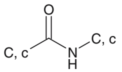
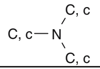
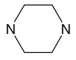
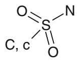
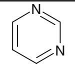
A posteriori knowledge, developed in-house, indicated the existence of a ''hERG pattern'' (Table 3). Molecules with structural features that matched this pattern were found to be more potent at inhibiting hERG. Variations of this pattern were coded as SMARTS and used as descriptors. Interpretation of the hierarchical PLS model identified these fragment-based descriptors as the main contributions to hERG pIC50.

Temporal monitoring of a global hERG PLS model

Data ($n = 7{,}520$), collected after the original models were built, were used to evaluate both the predictive performance of the PLS models over time and the usefulness of distance to model (DModX) as a diagnostic measure for poorly predicted molecules. This measure indicates when a prediction is an extrapolation because the molecule's combination of molecular properties, as they are represented in the model, are outside of the model's applicability domain.

The RMSE of the continuous temporal data setś predictions degenerated during the 15 month time period directly after the models were built (Table 2). Plots of the predicted scores and predicted DModX were examined to assess whether the new temporal data lay outside of the domain covered by the hierarchical models (Fig. 4). The first scores plot shows the position of the designed training set and the ori-

Table 3 Fragments proposed as contributing to a molecules propensity to inhibit hERG

Contribution to hERG pIC50	Fragment
Positive (generated from a priori knowledge)	
	
	
	
Negative (generated from LeadScope analysis)	
	
	
	
	
	
	
	

These fragments were determined to relate either negatively or positively with hERG pIC50. This was done by generating hERG specific fragments by a LeadScope analysis of the original IonWorks assay data (n = 1,311) and then identifying the most influential fragments in the hierarchical PLS training model

Key: A, any 2–7 atom linker; c, any aromatic atom; C, any aliphatic atom; N, amine; R, NO_2, $NHSO_2Me$, Halogen or CN functional groups

The positive fragments were expanded further to define ''N'' as either a primary, secondary or tertiary amine with ''A'' varying between 2 and 7 atom lengths. This resulted in an additional 64 hERG specific fragments that were included in the descriptor set when building the PLS model

ginal test data set (Fig. 4a). Both of these data sets were positioned within the ellipse shown, which indicates a confidence interval based on Hotelling's T^2. The second scores plot shows the scores positions of the temporal test data sets collected between August 2004 and September 2005 (Fig. 4b). This plot reveals that many of the molecules screened after August 2004 were positioned outside of the Hotelling's T^2 ellipse and were therefore outside of the prediction confidence interval (which was calculated at 99% confidence). The number of molecules outside of the model boundary (Fig. 4c) and the predicted distance to model plot (Fig. 4d) shows distance to model information for all data screened between June 2004 and September 2005. Molecules that are positioned outside of the critical model distance have different molecular and structural properties than those found in the earlier June 2004 training and test datasets and are therefore not represented by the model. Both of these plots indicate that these differences begin to occur as early as October 2004 with molecules outside of the model boundary (that have predicted distance to model values >5) appearing from May 2005. These plots illustrate how quickly the domain of applicability of a QSAR model can become outdated in the dynamic setting, such as that found within a pharmaceutical company.

Compound predictions that were outside of the model's DModX were excluded from the temporal prediction dataset and it was expected that the RMSE would be reduced to that obtained from the original test set predictions. However, exclusion of dissimilar molecules, detected by the distance to model plot, did not reduce the RMSE value of $pIC50_{obs}$ versus $pIC50_{pred}$, suggesting the existence of mechanisms for hERG blockade that are not represented or described adequately by the model. Furthermore, a paired t-test comparison of $pIC50_{obs}$ and $pIC50_{pred}$ revealed a bias between $pIC50_{obs}$ and $pIC50_{pred}$ for the October 2004 and September 2005 temporal test sets and many of the potent hERG inhibitors were under-predicted. It cannot be definitively concluded from these results whether the currently used descriptors do not adequately represent a single mechanism of hERG binding or if multiple mechanisms of binding do exist in the hERG ion channel, however, other evidence has been published supporting the possibility of multiple binding modes in the hERG channel [30–32].

To facilitate assessment of model accuracy in predicting the categorical IonWorks measurements (n = 2,707, $pIC50_{obs} < 5$), a confusion matrix was compiled of both continuous and categorical $pIC50_{obs}$ data by categorising $pIC50_{pred}$ values into four classes: pIC50 < 5; pIC50 between 5 and 5.5; pIC50 between 5.5 and 6 and pIC50 ≥ 6 (Table 4). Both predictions within and outside of the model boundary (DModX) were evaluated but results of all predictions are reported, as this filter resulted in only marginal differences in the accuracy of class predictions. The hierarchical model, M11, did not accurately classify $pIC50_{pred}$ into these four categories but this was expected considering that the RMSE of the continuous temporal dataset spans the class boundaries. The majority of predictions were classified into the

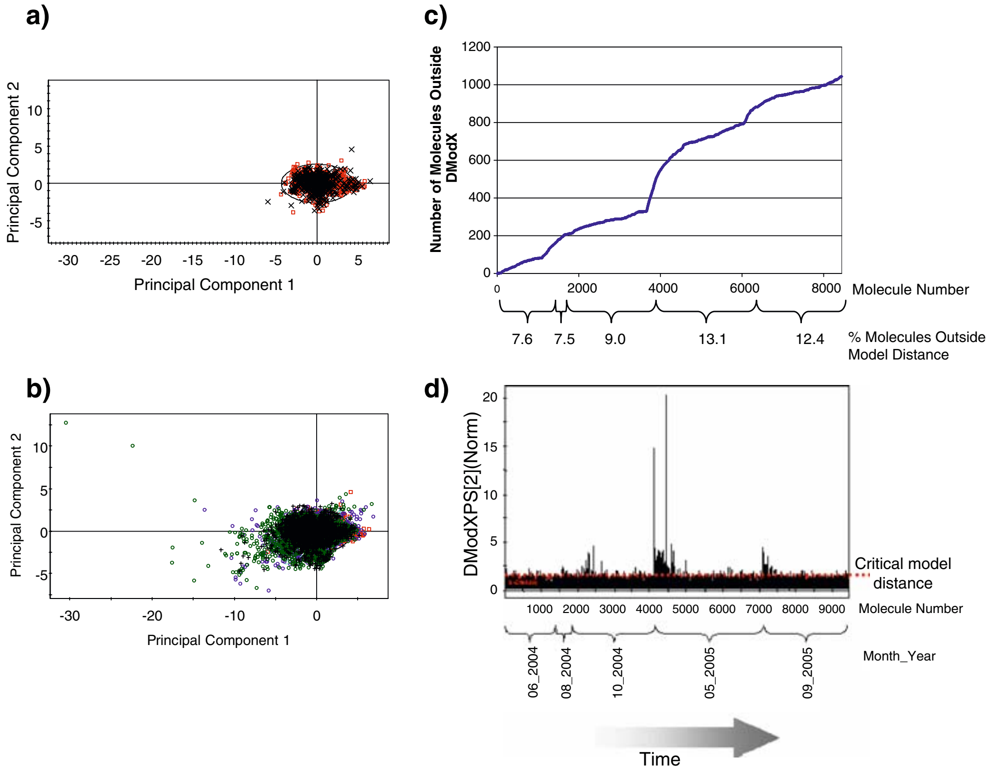


Fig. 4 PLS diagnostics used for identifying outlier pIC50 predictions (**a**) Plot of predicted scores (component 1 versus component 2) for the training and test set data available in June 2004. The ellipse denotes a Hotelling's T^2 associated with the predicted pIC50; therefore, compounds positioned outside of the ellipse are predicted with low confidence. *Key*: Training set = black diagonal crosses; test set = red open squares (**b**) Plot of predicted scores (component 1 versus component 2) for the temporal test set data. The ellipse denotes a Hotelling's T^2 associated with the predicted pIC50; therefore, compounds positioned outside of the ellipse are predicted with low confidence. *Key*: June 2004 dataset = red open squares; August 2004 dataset = light blue open triangles; October 2004 dataset = blue open circles; May 2005 dataset = green open circles; September 2005 dataset = black crosses. (**c**) Plot showing the number of molecules predicted outside of the model distance. Percentage of molecules indicated as outside of model distance was calculated by calculating the percentage of molecules outside DModX at the end of each time period, e.g. the percentage for June 2004 was calculated by dividing the total number of molecules outside of the model distance up until and including June 2004/total number of molecules in dataset up until and including June 2005 = 70/923 = 7.6%; the percentage for August 2004 = 82/1095 = 7.5%. (**d**) Plot of predicted Distance to Model. The model distance boundary was calculated as 1.39. Compounds positioned outside this boundary (i.e., at a greater model distance) are outside of the model's principal property space (i.e., have a significantly different combination of calculated properties than the model's training set molecules). Predictions for molecules outside of the model distance are predicted with low confidence

$pIC50_{obs}$ class ± the adjacent class, for example 91% of the non-active compounds (pIC50obs < 5) are classified into either the (1) pIC50 < 5 or (2) $pIC50_{pred}$ between 5 and 5.5 classes.

These models with their caveats must be utilized in drug discovery programs to reduce the burden on screening programs and prioritise synthesis of new molecules. A recent survey of international pharmaceutical companies reported that twice as many ion channel investigations (via non-clinical electrophysiology assays) are conducted in discovery compared to development, concluding that this was likely due to a need for early identification and culling of compounds posing a risk for developing TdP [33]. It is clear that discovery projects are now facing stringent challenges to deliver suitable candidate drugs that show minimal risk for developing QT interval prolongation/TdP.

Currently, many organisations rely upon the assumption that prolongation in either ventricular repolarisation or the action potential duration of the hERG ion channel are reliable markers for the risk of developing TdP and assess hERG blockade by using non-clinical electrophysiology measurements of these markers [34]. However, experimental screens cannot suggest new areas of chemistry or help the medicinal chemist generate hypotheses to reduce potency at hERG when optimising a promising lead series that exhibits moderate hERG inhibition. To realize

Table 4 Confusion matrix of pIC50$_{obs}$ and pIC50$_{pred}$ for temporal test dataset

		Number of compounds classified into pIC50$_{pred}$ class (percentage correct)				***N***	% Correctly classified (including pIC50$_{pred}$ values^{+}/-1 class)
		<5	5 to 5.5	5.5 to 6	>6		
PIC50$_{obs}$	<5	2,603(57%)	1,594	345	66	4,608	91
	5 to 5.5	340	682(47%)	353	84	1,459	94
	5.5 to 6	100	258	329(39%)	161	848	88
	>6	30	66	130	379(63%)	605	84
N		3,073	2,600	1,157	690	7,520	
Predictivity (including pIC50$_{pred}$ values $^{+}$/-1 class)		96%	97%	70%	78%		

Predictions for both non-continuous (pIC50$_{obs}$ < 5) and continuous pIC50$_{obs}$ IonWorks measurements are included in the matrix. Individual classes were poorly predicted but this was improved by accounting for the model's RMSE when interpreting classification results. This was achieved by including compounds predicted as belonging both to a class and its adjacent class (i.e. $^{+}$/-1 class) in the calculations for predictivity and percentage correct. For example, the predictivity of the non-active inhibitor class, pIC50$_{obs}$ < 5, was calculated by (2603+340)/3073 and the predicitivity of pIC50$_{obs}$ 5–5.5 class was calculated by (1594+682+258)/2600

these challenges, a predictive model is required that can accurately distinguish whether a new series or potential compound is within the recommended safety margin for hERG inhibition (typically recommended pIC50$_{obs}$ < 5) or whether it is a potent inhibitor (pIC50$_{obs}$ ≥ 6). The confusion matrix results (Table 4) were examined further to establish whether pIC50$_{pred}$ results could be classified differently to realize the challenge of distinguishing non-active and highly active inhibitors of hERG. Accordingly, confusion matrices were compiled for the temporal test sets' pIC50$_{pred}$ values using classification cut-offs for pIC50$_{pred}$ of 5, 5.5 and 6. These thresholds yielded overall model accuracies of 67, 82 and 93% respectively, although none of these classification cut-offs achieved accuracies >80% for both non-active and active inhibitor class predictions. The reality in Discovery projects is that many compound series show moderate hERG blockade (pIC50$_{obs}$ between 5 and 6) as measured by an in vitro electrophysiology assay and a method is required that can distinguish these moderately potent inhibitors from either highly or non-active inhibitors. The false positive and negative rates obtained for classification cut-offs at 5 and 5.5 were not considered sufficiently accurate to satisfy this requirement and furthermore, the temporal validation indicated that with the accuracies obtained using these classification cut-offs, model predictions risk either excluding suitable series (due to a high false positive rate) or wasting resources by synthesising potential hERG blockers (due to false negatives).

However, using a classification cut-off at pIC50$_{pred}$ value of 6 yielded accurate prediction of non-actives (96%) with a resultant low number of false positives (4%). This classification scheme is considered to be valuable in aiding discovery projects to prioritize sending predicted inhibitors for experimental assay (for confirmation of potential hERG blockade issues) or selection or optimisation of a lead series. In conjunction with this, predictions from the global model presented (M11), can be classified into four classes (Table 5) and misclassification of pIC50$_{pred}$ estimated according to the RMSE (by including predictions from the adjacent class), as discussed earlier. By this treatment, it is possible to prioritise the non-active (pIC50$_{obs}$ < 5) and active (pIC50$_{obs}$ > 6) inhibitors with sufficient accuracy for target series selection and help generate ideas for possible structural modifications and creating hypothesis for reducing hERG.

Application of the hERG hierarchical PLS model in a drug discovery environment

The global hERG model developed has two uses in early drug discovery. First, it can be used in screening of chemical libraries, by classifying the predicted pIC50 values into two categories, as detailed in Table 4, namely "non-active" molecules, which are included in library designs, have pIC50$_{pred}$ < 5.5 and

Table 5 Classification results for pIC50$_{pred}$ of temporal test set predictions

PIC50 classification cut-off	5	5.5	6
Number of hERG inhibitors (actives)	2,912	1,453	605
Number of non-actives	4,608	6,067	6,915
Sensitivity (% correct actives)	56	69	63
Specificity (% correct non-active)	84	86	96
% False actives	44	14	4
% False non-actives	16	31	38
Overall model accuracy	67	83	93

Thresholds of pIC50$_{obs}$ of 5, 5.5 and 6 were used to classify pIC50$_{pred}$ values as belonging to either non-active or active classes for hERG inhibition ‘‘active’’ molecules, which are excluded from libraries, have $\text{pIC50}_{\text{pred}} > 5.5$. The second application of the model is to provide interpretive information for a query molecule, derived from the model, to the medicinal chemist prior to synthesis, with the intention to suggest ideas for structural modifications that could result in reducing hERG activity (as detailed in Fig. 3). This approach is exemplified in Fig. 5 using Terfenadine to illustrate a query molecule. Terfenidine is predicted to be an active hERG inhibitor ($\text{pIC50}_{\text{pred}} = 6.4$). Examination of the score contributions for Terfenidine relative to the model average (which corresponded to a $\text{pIC50}_{\text{obs}} = 5.4$) indicated that hERG activity could be reduced by incorporating hydrogen bond acceptors, including hydroxyl and carboxylic acid moieties and an aliphatic ether into the molecule. Additionally, the hERG pattern found in Terfenidine (namely a phenyl ring coupled to a tertiary amine through four aliphatic carbons) was found to contribute to its high predicted hERG activity. It is already known that two successful structural modifications of Terfenadine did include these structural modifications: namely Fexofenadine, which includes a carboxlic acid moiety onto the Terfenadine backbone and Cetrizine, which replaces the hERG pattern found in Terfenadine and includes an aliphatic ether coupled to a carboxylic acid. This example, using three marketed compounds, confirms and illustrates the usefulness of this approach for suggesting structural modifications to a medicinal chemist faced with the challenge of reducing hERG activity in a potent hERG inhibitor, prior to any synthesis

Prediction of marketed drugs using the Hierarchical PLS model

A dataset of 34 marketed compounds was measured by IonWorks for hERG inhibition and the prediction results are presented (Table 6) to further illustrate to the reader, with publicly known examples, the performance and weaknesses of the hierarchical model (M11) developed. In total, 21 compounds were within the model DModX and the accuracy of these predictions was examined. A comparison of $\text{pIC50}_{\text{obs}}$ and $\text{pIC50}_{\text{pred}}$ supports the prediction results observed in the temporal test sets. A lower RMSE was obtained for predictions arising from query molecules within the model boundary than those outside it. Additionally, predictions for non-active and moderately active compounds ($\text{pIC50}_{\text{obs}}$ between 4 and 6) are well predicted (the RMSE for these molecules is calculated to be 0.46) but $\text{pIC50}_{\text{pred}}$ is not accurate for highly potent inhibitors of hERG. [expand on the DModX filter].

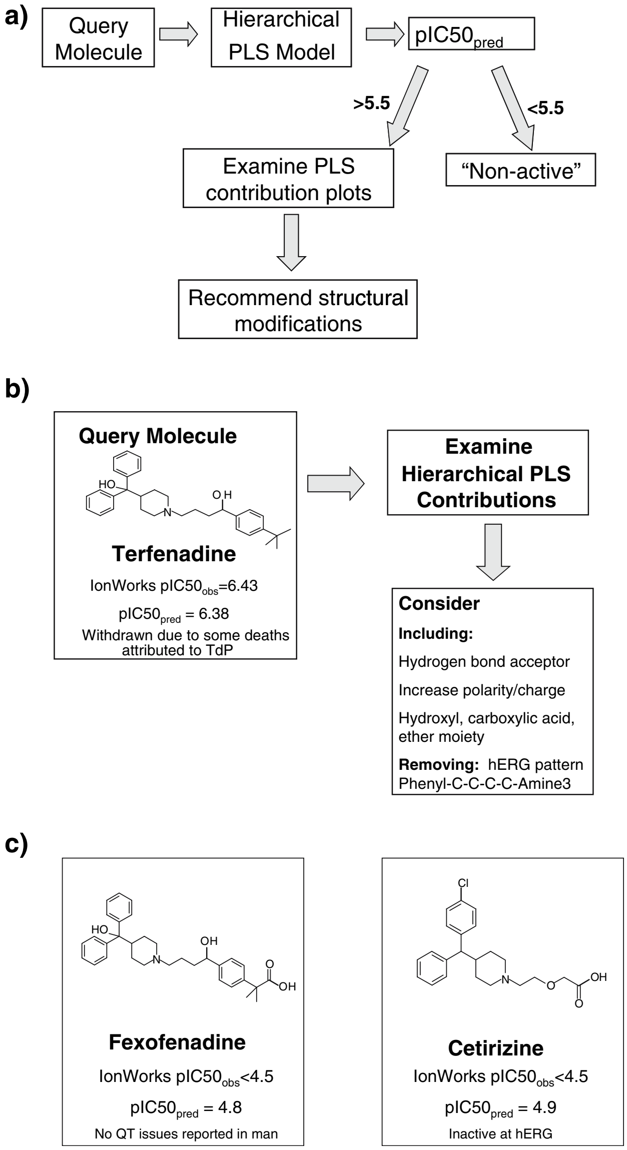

a)
Query Molecule
Hierarchical PLS Model
pIC50$_{pred}$
>5.5
<5.5
Examine PLS contribution plots
“Non-active”
Recommend structural modifications
b)
Query Molecule
HO
OH
N
Terfenadine
IonWorks pIC50$_{obs}$=6.43
pIC50$_{pred}$ = 6.38
Withdrawn due to some deaths attributed to TdP
Examine Hierarchical PLS Contributions
Consider
Including:
Hydrogen bond acceptor
Increase polarity/charge
Hydroxyl, carboxylic acid, ether moiety
Removing: hERG pattern Phenyl-C-C-C-Amine3
c)
HO
OH
N
O
OH
Fexofenadine
IonWorks pIC50$_{obs}$<4.5
pIC50$_{pred}$ = 4.8
No QT issues reported in man
Cl
N
O
OH
O
Cetirizine
IonWorks pIC50$_{obs}$<4.5
pIC50$_{pred}$ = 4.9
Inactive at hERG

Fig. 5 Application of the hERG hierarchical PLS model in a drug discovery environment. (**a**) A query molecule is submitted to the model. If it is predicted as ‘‘non-active’’ it is used for either library design or synthesized. If a query molecule is predicted as ‘‘active’’, the model is interpreted to provide information relating to why it is predicted active and, if possible, recommend structural modifications. (**b**) Prediction of Terfenidine indicates that this is a hERG active molecule and subsequent examination of the PLS predicted scores contributions suggests structural modifications to reduce hERG activity. (**c**) Known structural modifications of Terfenadine, (resulting in Fexofenadine and Cetrizine) confirm that the model and this approach can produce recommendations for structural modifications that are valid and can be useful to the medicinal chemist

Whilst some inhibitors are correctly identified as potent inhibitors of hERG (pimozide, astemizole, haloperidol and terfenidine) others are badly under-predicted to be only moderately potent inhibitors (quinidine, cisapride, E-4031 and dofetilide).

Table 6 Summary of global hierarchical PLS model's predictions for QT prolonging drugs

Name	$pIC50_{obs}$	$pIC50_{pred}$	Error ($pIC50_{pred}$–$pIC50_{obs}$)	Prediction within DModX?	RMSE
Astemizole	7.9	6.6	–1.3	Yes	0.7
Pimozide	7.9	7.2	–0.7		
Haloperidol	7	6.2	–0.8		
E-4031	6.8	5.1	–1.7		
Cisapride	6.5	5.5	–1		
Terfenadine	6.4	6.1	–0.3		
Quinidine	6.2	5.2	–1		
Risperidone	5.8	5.6	–0.2		
Donepezil	5.7	5.9	0.2		
Fluoxetine	5.7	5.7	0		
Buspirone	5.4	4.8	–0.6		
Clozapine	5.3	5.8	0.5		
Ropinirole	5.3	5.2	–0.1		
Loratadine	5.2	5.4	0.2		
Celecoxib	5	5.2	0.2		
Quetiapine	5	5.5	0.5		
Olanzapine	4.7	5.4	0.7		
Zafirlukast	4.6	5	0.4		
Cetirizine	4.5	4.9	0.4		
Bupivacaine	4.4	4.9	0.5		
Phenytoin	4	4.8	0.8		
Dofetilide	7.1	5.2	–1.9	No	1.1
Citalopram	5.4	6.6	1.2		
Clomipramine	5.4	6	0.6		
Sertraline	5.1	5.5	0.4		
Sparfloxacin	4.9	3.2	–1.7		
Fluvoxamine	4.8	5.1	0.3		
Sibutramine	4.8	5.7	0.9		
Propranolol	4.6	5.3	0.7		
Diazepam	4.5	5.1	0.6		
Procaine	4.4	5	0.6		
Triazolam	4.3	5	0.7		
Moclobemide	4.1	5.4	1.3		
Bupropion	4.1	5.2	1.1		

Key: $pIC50_{obs}$ = mean $pIC50_{obs}$ measurement in IonWorks electrophysiology screening assay; $pIC50_{pred}$ = prediction obtained from the hierarchical training set PLS model, M11

Conclusions

A predictive global QSAR model has been constructed by relating calculated molecular descriptors with hERG inhibition data, as measured by the IonWorks HT electrophysiology screening assay. Adopting a hierarchical PLS modelling approach allowed inclusion of four different descriptor packages enabling simple interpretation of the underlying model and retaining the predictive performance of a reference model that included all descriptors. New hERG specific fragment-based descriptors were developed describing both positive and negative correlations with hERG $pIC50_{obs}$ and these descriptors were identified as the most influential descriptor package in the hierarchical PLS model. By using a combined onion and D-optimal design to select the training and test sets, a global model was calculated that covered the known chemical property and structural variation in the dataset, whilst allocating a substantially sized test set for model evaluation purposes. This method for selecting the model's training set was also shown to be superior to random selection with respect to the size of the model's domain of applicability.

Original model evaluation indicated that the global model was both interpretable and predictive with comparable prediction accuracy to other published models. However, we have also included a validation of the model's predictive performance over an extended time period, which has not been done previously, and it was found to deteriorate within a 15 month period. Additionally, the model was unable to distinguish moderate inhibitors ($pIC50_{obs}$ between 5 and 6) from either non-active ($pIC50_{obs} < 5$) or potent ($pIC50_{obs} > 6$) inhibitors of hERG, due to the RMSE of the $pIC50_{pred}$ values. Although the overall RMSE of the temporal data set predictions was reasonable and the model could correctly predict many of the potent inhibitors, PLS diagnostics failed to identify potent inhibitors that were poorly under-predicted (in some cases the error of the $pIC50_{pred}$ values exceeded the model RMSE).

Evaluation of distance to model as a measure of prediction accuracy revealed that although this did identify erroneous predictions it was not a guarantee that predictions within the model's critical distance boundary were predicted within the RMSE calculated for the overall model. This is indicative of deficiencies in the representation of hERG inhibition in the global model, which could be due to either inadequate descriptors or a failure to capture multiple binding modes that may be possible in the ion channel.

Producing a global model that adequately describes any biological system is a challenging problem. Local models (i.e., models based upon an analogous structural series) may yield more accurate predictions and prove more useful in guiding discovery projects during the lead optimisation process. However, it may not always be practical to develop a local model, due to limited availability of experimental data. In this scenario a project must either rely upon a global model or submit new compounds to an experimental screen. However, discovery projects dealing with a target series positioned in the "moderate inhibitor" zone (i.e., $\mathrm{pIC50_{obs}}$ between 5 and 6) cannot always rely upon the experimental assay results to differentiate the effects resulting from structural modifications because the experimental assay error for the IonWorks and traditional "gold standard" Patch Clamp assays ranges between 0.1 and 0.5 log units for replicate $\mathrm{pIC50_{obs}}$ measurements. This limitation extends to the accuracy attainable for any QSAR model based upon these experimental data and it is always necessary to compare the experimental error with the RMSE from a temporal validation.

We conclude that the classification of the $\mathrm{pIC50_{pred}}$ values indicate that the model predictions are sufficiently accurate to aid discovery projects in predicting non-active from potent inhibitors of hERG. Additionally, the interpretability of the global model, both with respect to the use of meaningful fragment-based descriptors and the ease with which hierarchical PLS model can be evaluated, lends itself as a useful tool that can used to estimate how new compounds will inhibit hERG in vitro and enable comparisons between proposed molecules and those in our substantially sized global dataset in terms of both structural and chemical properties relating to hERG inhibition.

Methods

Experimental data

A dataset, comprised of potential drug-like molecules from different disease areas (n = 1,304) and marketed compounds (n = 8), were screened for hERG blockade using the IonWorks^{TM} HT (high throughput) electrophysiology assay [35]. Measured pIC50 values (ranging from 4 to 8.2) were obtained for hERG blockade of the K^+ channel (experimental variation of replicate measurements ranged between 0.1 and 0.5 log units). Mean $\mathrm{pIC50_{obs}}$ values were calculated from replicate measurements (ranging between 2 and 6 replicates). Additional data (n = 7,520) became available to externally test model performance after the initial training and test validation results were obtained. These data comprised of mean IC50 measurements (n = 4,813) and non-active compounds ($\mathrm{pIC50_{obs}} < 5$) for which accurate IC50 values were not obtained (n = 2,707).

Molecular descriptor data

Molecular property descriptors from three packages (Selma [36], DRONE [37] and VolSurf [38]) and structural fragment-based descriptors (generated in-house) were calculated to describe the chemical properties and structural features of the compounds.

The in-house Selma program generates 2D descriptors related to size (e.g., molecular weight, ring structure (e.g., sizes of 3 largest rings), flexibility (e.g., number of rotatable bonds), hydrogen bonds, polarity (e.g., polarizability [39], the connectivity indices [40], electronic environment, partial atom charge [41] and lipophilicity.

The in-house Drone program generates 2D, 3D and charge dependent descriptors. The 2D descriptors describe topological and various atom and sub-structural features. The 3D coordinates are used to generate geometrical descriptors (e.g., gaussian volume, van der Waals surface areas) and the atomic charges are used to compute a series of charge dependent descriptors (e.g., statistics on the distribution of charges and areas of positive or negative electrostatic potential on the van der Waals or solvent accessible surfaces).

The VolSurf approach is described elsewhere [42, 38]. The procedure compresses relevant information present in 3D maps into a few simple, quantitative and interpretable descriptors that can be related to biological activity.

In addition to the traditional generic descriptors, hERG specific structural fragment-based descriptors were generated from two sources: by using LeadScope® and from *a posteriori* knowledge from medicinal chemists within the company regarding structural requirements that induce hERG inhibition. Fragment-based descriptors generated from both of these sources were also compared to the available structural knowledge of the hERG pharmacophore in the literature and they were found to incorporate the features of the previously proposed pharmacophores [9, 10].

From a posteriori knowledge and literature sources we were able to encode fragment-based descriptors that we proposed should contribute to a molecule's propensity to inhibit hERG, however our descriptor set lacked fragment-based descriptors that potentially could reduce a molecule's propensity to inhibit hERG. To overcome this limitation, we analyzed the original dataset (n = 1,312) using LeadScope®.

The algorithm used for fragment identification in LeadScope® has been described previously [43]. The LeadScope analysis involved generating scaffolds and common substructures and sorting all structural fragments (~27,000) included in the LeadScope package according to the pIC50$_{obs}$ values and Z-scores. The Z-score describes how the fragment groups mean pIC50$_{obs}$ value varies compared to the mean of the rest of the dataset by calculating the standard deviation between them. It can therefore be used to evaluate whether a fragment is positively or negatively correlated with pIC50. A standard deviation difference of ± 3 was considered to reflect a difference between the groups and the fragment was considered to be significant if it was present in a minimum of 50 molecules. Although the LeadScope® analysis indicated the presence of fragments similar to the fragment pattern indicated from a posteriori knowledge, it did not identify these patterns exactly.

A collection of general fragments (generated in house) was also included in this descriptor set. Fragments were then coded as SMARTS and SMARTS subgraph pattern matching was performed on the dataset to determine the number of hits in each structure [44]. Examples of fragments included in the model are tabulated in Table 1. In total, 73 hERG specific and 233 generic functional group descriptors were generated for each compound.

QSAR modelling

The molecular descriptor and pIC50$_{obs}$ data were imported into the SIMCA-P+, version 11 software package (www.umetrics.com). Initially, reference models (including all observations, n = 1,312) were calculated (M1–M6, Table 2). Reference PCA and PLS models, including all molecular property and structural descriptors were calculated to check for outliers and any activity related clustering or separation indicated in the scores plots (Fig. 4a). Four separate reference base level PLS models were calculated relating Selma, DRONE, VOLSURF and Fragment-based descriptors (X matrices) to the measured pIC50 hERG activity (Y matrix) and the base level scores (t = 20) were used to calculate an upper level PLS model (Fig. 4b). A three component PLS model was calculated (R^2X = 0.33, R^2Y = 0.60, Q^2 = 0.57) and the scores and Hotelling's T^2 values were extracted to select the training and test sets by D-optimal onion design.

Statistical molecular design is increasingly being used as a method for selecting a representative and diverse range of molecules for the training set when calculating QSAR models. Alternative approaches (e.g., random selection) may introduce bias into a QSAR model if the selected training set is not balanced with respect to the biological activities or chemical properties represented. There are many different types of statistical designs that can be utilized according to the modelling objective. In statistical molecular design, the principles used in experimental design are applied to score vectors calculated by either principal component analysis or partial least squares. Selecting a model training set by using a statistical molecular design ensures that the chosen subset of molecules is diverse, informative and representative of the available dataset.

By coupling onion and D-optimal designs a training set is selected that is reasonably sized and is representative of the molecular structures throughout the chemical property space. D-optimal designs maximize the determinant of the variance-covariance matrix, $X'X$, for a given PCA or PLS model, in the case of QSAR models this matrix consists of the chemical properties that are used to describe the molecules. This aims to select molecules in a way that those selected are spread in a maximum volume of the property space. However, D-optimal designs can suffer from inadequate sampling of the inner part of the chemical property space (as it is described by the PCA or PLS scores). By coupling D-optimal design with an onion design this problem is addressed. Onion designs split the dataset into different layers (or subsets) enabling the selection of training set compounds from each layer. This ensures that the representative molecules are selected from the inner layers of the onion design and that the training set contains a diverse range of chemical structures in addition to covering the range of chemical properties. Coupling D-optimal and onion designs also increases the number of molecules selected for the model's training set, which is desirable when modelling global datasets that contain a diverse range of chemical structures.

In order to achieve a reasonable sized test set (n = 436) that adequately represented the structural and physicochemical variation present within the data, an onion design with forty layers was used and at least 11 compounds were selected from each layer. The layers were assigned using the Hotelling's T^2 values and for each layer a D-optimal quadratic design was calculated from the upper level PLS scores using the MODDE, version 7, software package (www.umetrics.com).

The training set ($n = 436$) was used to calculate a hierarchical PLS model (M11, Table 2) by combining the scores ($n = 13$) derived from four base level PLS models calculated for each of the molecular descriptor packages (M7 = SELMA, M8 = DRONE, M9 = VOLSURF and M10 = fragment-based descriptors). The model statistics and prediction performance of all base level and upper level models was assessed (Table 2).

The use of onion and D-optimal design for training set selection was compared to a random selection approach. The approaches were evaluated by constructing six reference models (containing all descriptors) based upon different randomly selected training sets and comparing the model statistics, prediction performance and the models' coverage of the molecular descriptor space with those obtained from a reference model based upon the onion and D-optimal designed training set (Table 1).

Establishing the predictive power of the PLS models

The models predictive power was determined using both internal and external validation methods. Internal validation was done by using cross-validation [45] and response permutation testing [46]. Cross-validation is known to be misleading when the training and test data have been selected by statistical experimental design. This is because each training set observation is important for generating the designed model and if any are removed, the underlying design often degenerates, resulting in misleading under estimates of predictive power [47]. However, as we had access to a large dataset of high quality data, we were interested in using this opportunity to examine whether cross-validation statistics could serve as a reliable indicator of predictive power when creating models for use in a real drug discovery environment. In this setting, models are often required before adequate data becomes available and it is impossible to validate with a sufficiently representative test set. Hence, there is a need to rely upon the results from internal validation methods when it is not possible to perform a rigorous external validation. As we used statistical molecular design and could not compare the cross-validation results with the predictive performance of the original test set, we evaluated the models' cross-validation statistics using test set data collected after the initial training and test sets were selected. These results were also compared with those generated from the random models.

Cross-validation involves dividing the data into a number of groups and developing a number of parallel models from these reduced datasets. The excluded data is then used as a test set for confirming and validating the initial analysis. The errors between the observed and predicted values from all of the parallel models are evaluated and used as a measure of the predictive ability of the model. In PLS the errors are used to calculate the predictive residual sum of squares (PRESS) and this is often expressed as Q^2Y. It is recommended using the R^2Y and Q^2Y parameters generated from PLS as guidelines for evaluating the reliability of a PLS model [48]. Thus, a value of $Q^2Y > 0.5$ is obtained by a good model. Additionally, this value should be evaluated alongside R^2Y, which is a measure of the fraction of variance explained, and differences between the values of R^2Y and Q^2Y exceeding 0.2 or 0.3 can be indicative of model over-fitting.

Response permutation testing gives a statistical significance to the estimated predictive power [46]. In this method, the numerical values of the end-point or response, for example hERG $\text{pIC50}_{\text{obs}}$, remain the same but are randomly shuffled. As the number of data points remains intact, this method is not affected if designed training sets are used. A number of parallel models are then developed fitting to the randomly ordered data. The resultant ‘‘scrambled'' models can then be evaluated for predictive power, for example by cross-validation and the results compared to the original model. This procedure is repeated many times (for example 100 times) and the collection of permuted models create a reference distribution that can be compared to the original model. The original QSAR is determined to be valid if the comparison determines that its predictive power is outside that of the reference distribution.

The external validation was expected to be a reliable indicator as the training set had been selected by statistical molecular design. Consequently the test set should be within the model's applicability domain. The training and test set scores and predicted scores were plotted to examine whether the test set adequately covered the range of the model's applicability domain. The range of the response variable, hERG pIC50, was deemed to be acceptable as the ranges were similar in the training set ($\text{pIC50}_{\text{obs}}$ between 4.0 and 7.9) and the test set ($\text{pIC50}_{\text{obs}}$ between 4.0 and 8.2).

Temporal assessment of the hierarchical PLS model

Additional IonWorks^{TM} HT electrophysiology assay data was provided 2, 4, 9 and 15 months after the original data were used to build the hierarchical PLS model. These data were used to assess the PLS model's predictive performance by calculating the RMSE of $pIC50_{obs}$ and $pIC50_{pred}$ for the datasets from each time-point.

The reliability of the predictions with respect to identifying non-active, moderately and highly potent hERG inhibitors was also evaluated using the temporal dataset by classifying the $pIC50_{pred}$ values. The predicted values were classified taking the model RMSE into account (0.5). Thus, four classes were allocated according to pIC50 values of <5, 5–5.5, 5.5–6 and >6 (Table 4).

Classification models have been published that use different criteria for defining non-active and active hERG inhibitors and this has ranged between $pIC50_{obs} > 5$ and $pIC50_{obs} > 6$ as the cut-off defining active inhibitors. Realistically, it has been our experience that using a $pIC50_{obs}$ value of 6 as a classification cut-off for predicting active versus non-active inhibitors is not helpful to a medicinal chemist, as many compounds screened have $pIC50_{obs}$ values between 5 and 6. Therefore, in addition to classifying predictions with pIC50 cut-offs of 5 and 6 (to allow comparison to published models) we have also classified the predictions using a pIC50 value of 5.5 (Table 5).

Prediction of marketed drugs by the hierarchical PLS model

A dataset of 34 marketed drugs were measured by the IonWorks^{TM} HT electrophysiology assay and these predictions were evaluated (Table 6).

Acknowledgements The authors thank the following from Safety Pharmacology, Astrazeneca Pharmaceuticals, Department of Safety Assessment, Alderley Park, Macclesfield, Cheshire SKN 4TG, U.K.: B.G. Small, M.H. Bridgeland-Taylor, A.J. Woods and A. Harmer for generating the IonWorks^{TM} HT Data that form the basis of this work and C.E. Pollard for discussions around the biological aspects of this manuscripts

References

1. Brown AM (2004) Cell Calcium 35:543
2. Finlayson K, Witchel HJ, McCulloch J, Sharkey J (2004) Eur J Pharmacol 500:129
3. Guth BD, Germeyer S, Kolb W, Markert M (2004) J Pharmacol Toxicol Meth 49:159
4. Recanatini M, Poluzzi E, Masetti M, Cavalli A, De Ponti F (2005) Med Res Rev 25:133
5. Redfern WS, Carlsson L, Davis AS, Lynch WG, MacKenzie I, Palethorpe S, Siegl PK, Strang I, Sullivan AT, Wallis R, Camm AJ, Hammond TG (2003) Cardiovasc Res 58:32
6. Sanguinetti MCM, John S (2005) Trends Pharmacol Sci 26:119
7. Hammond TPC (2005) Toxicol Appl Pharmacol 207:446
8. Aronov AM (2005) Drug Discov Today 10:149
9. Cavalli A, Poluzzi E, De Ponti F, Recanatini M (2002) J Med Chem 45:3844
10. Ekins S, Crumb WJ, Sarazan RD, Wikel JH, Wrighton SA (2002) J Pharmacol Exp Ther 301:427
11. Pearlstein RA, Vaz RJ, Kang J, Chen X-L, Preobrazhenskaya M, Shchekotikhin AE, Korolev AM, Lysenkova LN, Miroshnikova OV, Hendrix J, Rampe D (2003) Bioorg Med Chem Lett 13:1835
12. Aronov AM, Goldman BB (2004) Bioorg Med Chem 12:2315
13. Osterberg F, Aqvist J (2005) FEBS Lett 579:2944
14. Rajamani R, Tounge BA, Li J, Reynolds CH (2005) Bioorg Med Chem Lett 15:1741
15. Bains W, Basman A, White C (2004) Prog Biophys Mol Biol 86:205
16. Bains W, Basman A, White C (2004) Prog Biophys Mol Biol 86:233
17. Cianchetta G, Li Y, Kang J, Rampe D, Fravolini A, Cruciani G, Vaz RJ (2005) Bioorg Med Chem Lett 15:3642
18. Keseru GM (2003) Bioorg Med Chem Lett 13:2775
19. Olivier Roche GT, Zuegge J, Pflimlin P, Alanine A, Schneider G (2002) ChemBioChem 3:455
20. Roche O, Trube G, Zuegge J, Pflimlin P, Alanine A, Schneider G (2002) Chembiochem 3:455
21. Tobita M, Nishikawa T, Nagashima R (2005) Bioorg Med Chem Lett 15:2890
22. Eriksson L, Johansson E, Lindgren F, Sjostrom M, Wold S (2002) J Comput Aided Mol Des 16:711
23. Kiralj R, Ferreira MMC (2003) J Mol Graph 21:448
24. Purdy R (1996) Environ Health Perspect 104:1085
25. Gute BD, Basak SC (1997) SAR QSAR Environ Res 7:117
26. Eriksson L, Arnhold T, Beck B, Fox T, Johansson E, Kriegl JM (2004) J Chemom 18:188
27. Song M, Clark M (2006) J Chem Inf Model 46:392
28. Fenichel, R.R., http://www.fenichel.net/pages/Professional/subpages/QT/Tables/pbydrug.htm
29. Bridgland-Taylor MH, Hargreaves AC, Easter A, Orme A, Henthorn DC, Ding M, Davis AM, Small BG, Heapy CG, Abi-Gerges N (2006) J Pharmacol Toxicol Meth 54:189
30. Bett G, Rasmusson R (2003) Cell Biochem Biophys 39:183
31. Milnes J, Crociani O, Arcangeli A, Hancox J, Witchel H (2003) Br J Pharmacol 139:887
32. Sanguinetti MC, Mitcheson JS (2005) Trends Pharmacol Sci 26:124
33. Friedrichs GS, Patmore L, Bass A (2005) J Pharmacol Toxicol Meth 52:11
34. Lawrence CL, Pollard CE, Hammond TG, Valentin J-P (2005) J Pharmacol Toxicol Meth 52:59
35. Schroeder K, Neagle B, Trezise DJ, Worley J (2003) J Biomol Screen 8:50
36. Selma is an in-house AstraZeneca program. For further information contact T. Olsson, V.S., Synthesis and Structure Administration (SaSA), Astrazeneca R&D Mölndal, Sweden
37. Bruneau P (2001) J Chem Inf Model 41:1605
38. Cruciani G, Pastor M, Guba W (2000) Eur J Pharm Sci 11:S39
39. Glen RC (1994) J Comput Aided Mol Des 8:457

40. Kier LB, Hall LH, (1986) Research Studies Press. John Wiley and Sons, Letchwort
41. Gasteiger J, Marselli M (1980) Tetrahedron 36:3219
42. Cruciani C, Crivori P, Carrupt PA, Testa B (2000) Theochem-J Mol Struct 503:17
43. Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE (2000) J Chem Inf Comp Sci 40:1302
44. www.daylight.com (Subgraph matching is part of the SMARTS Toolkit)
45. Wold S (1978) Technometrics 20:397
46. van der Voet H (1994) Chemometr Intell Lab Syst 25:323
47. Hawkins DM, Basak SC, Mills D (2003) J Chem Inf Model 43:579
48. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S (2001) Multivariate and megavariate data analysis - principles and applications. Umetrics Academy, pp 65–67