Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2008) 1315–1326
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

The importance of the domain of applicability in QSAR modeling

Shane Weaver, M. Paul Gleeson *

Computational & Structural Chemistry, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, United Kingdom

Received 21 November 2007; received in revised form 11 January 2008; accepted 11 January 2008
Available online 18 January 2008

Abstract

The domain of applicability is an important concept in quantitative structure activity relationships (QSAR) that allows one to estimate the uncertainty in the prediction of a particular molecule based on how similar it is to the compounds used to build the model. In this paper we discuss this important concept, providing details of the development and application of a method to compute the domain of applicability within model descriptor space and structural space as defined by daylight fingerprints.

The importance of the domain of applicability is illustrated using five QSAR models generated on plasma protein binding and P450 inhibition datasets. Such methodologies will be shown to offer us a means to monitor the performance of QSARs over time, providing us both with a way to estimate the accuracy of a given prediction and identify when a model needs to be rebuilt.
© 2008 Elsevier Inc. All rights reserved.

Keywords: Domain of applicability; QSAR; Cytochrome P450; Plasma protein binding; PLS; Neural network; ADMET

1. Introduction

A consideration of the developability characteristics of new chemical entities (NCEs) has become increasingly important in drug discovery in the last two decades. This is driven by the fact that ~60% of drugs fail [1–3] for different ADMET reasons (absorption, distribution, metabolism, excretion and toxicity) leading to an increasing demand for in vivo, in vitro and in silico methods to screen lead compounds much earlier on in the drug discovery process.

Quantitative structure activity relationships (QSARs) [4–7] have become an important component in the compound design and progression process since they represent a much cheaper, rapid alternative to the medium throughput in vitro and low throughput in vivo assays which are generally restricted to later in the discovery cascade. A QSAR is essentially a mathematical equation that is determined from a set of molecules with known activities using computational approaches. The exact form of the relationship between structure and activity can be determined using a variety of statistical methods and computed molecular descriptors and this equation is then used to predict the activity of new molecules.

Early QSARs pioneered by Hanch and Fugita [8] consisted of relatively small number of molecules of a given chemotype being used to derive a simple linear equation to predict the next molecule in the series to be synthesised. The advantage of this approach was that the terms in the equation were generally simple and easily interpretable, while the kinds of molecules being predicted were generally very similar to those that were already synthesised, giving the user greater confidence in the model predictions. In contrast, over the past decade an increasing number of QSARs have been reported based on large, diverse datasets, commonly termed global models, which are considered more reliable at predicting diverse structures than QSARs built on small datasets of low diversity [9–13]. These models are often built using complex statistical methods, and large numbers of often sparsely populated geometrical and electrotopological descriptors [14–17], and while this may allow for a more versatile description of molecular structure and a reliable way to relate structure to activity, the multi-dimensional space defined by such a model will become increasingly complex and fragmented.

Within the pharmaceutical industry the chemotypes being synthesised at any given time depends on several factors such as the biological targets being pursued and the hits identified from

* Corresponding author. Tel.: +44 1438 768682; fax: +44 1438 763352.
E-mail address: paul.x.gleeson@gsk.com (M.P. Gleeson).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.01.002

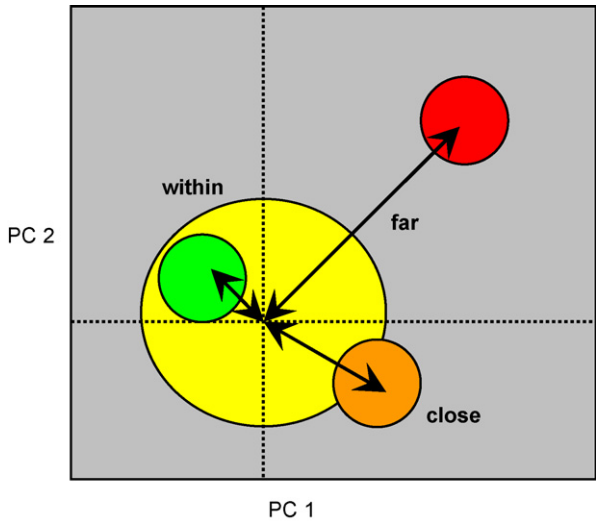

Fig. 1. A graphical illustration of the domain of applicability in principal component (PC) space. The QSAR model training set is represented by the yellow circle. Query molecules are coloured as follows: within the training space (green), close to model space (orange) and far (red). Query compounds predicted further from training model space would be expected to be less reliably predicted.

screening. This has implications for the prediction of new chemotypes not present in a QSAR model training set since these may occupy an area of model space that is not well represented (Fig. 1). Corporate collections are constantly moving further from historical chemical space meaning predictions from QSAR models developed on older, increasingly less relevant datasets will become extrapolations rather than interpolation.

Recognition of this problem in the field of QSAR can be found from the increased number of publications discussing this topic [18–26]. The methods all involve computing the similarity of the query molecules to the model training set using a variety of descriptors and distances (i.e. the so called domain of applicability), and relating this quantity to the prediction error. Readers are referred to references [19,21,23] for an introduction to the concept.

We add to the existing debate by reporting the development and application of a method to compute the domain of applicability of QSAR models, illustrating how it can be used to provide significant additional value in both local and global modeling applications. With examples derived from plasma protein binding and P450 3A4 inhibition datasets, we highlight the way in which such methods can be used to provide extra confidence in QSAR predictions.

2. Results

To illustrate the implications of the domain of applicability in QSAR modeling we have used the following methodology. Five separate QSAR models were built by splitting the respective datasets by date into three different sets as described in the experimental procedures. The earliest dated set was split into training and test sets at random, according to one of the standard practices in QSAR validation, meaning the two datasets are essentially mirror images of each other. The performance of the model on the test set represents a best-case scenario and deterioration in performance over time, and evolving chemistry might be expected. To quantify the deterioration in predictive performance we use the remaining molecules synthesised and tested over the course of at least 1 year following the completion of the model building process (validation set 1 and validation set 2).

2.1. Global plasma protein binding QSAR model

In this first example we study the effect of the domain of applicability on a QSAR model built on plasma protein binding data using a linear statistical technique called PLS regression, combined with relatively simple and interpretable 1D and 2D descriptors. Before modeling, the %bound values were transformed into the more appropriate log K scale (log(%bound/%free)). With the exception of the newly obtained validation set 2, this model has been discussed in detail elsewhere so only a brief description is given here to facilitate a discussion of the domain of applicability calculations [24]. The QSAR, or quantitative structure property relationship (QSPR) model to be more precise, has a moderate r_0^2 of 0.56 (correlation to the line of unity—explaining 56% of the total variation), and an equivalent r^2 (regression line correlation) since this is a fitted relationship based on the 685 training set compounds with a slope of 1 and intercept of 0 (Table 1). The cross-validated q^2 is of similar magnitude at 0.54 suggesting the model is internally consistent. Note a random model prediction would have a root mean square (RMSE) $\geq$ standard deviation (S.D.).

The good model performance on the training set is no guarantee that a model will be predictive on future datasets [27]. We have therefore employed the three additional datasets discussed above (test, validation 1 and validation 2), to assess the utility of the QSPR model, each of which representing an increasingly difficult test for the model due to the increase in time, and structural diversity. Additionally, a fourth literature-derived set was available to assess the protein binding model which is also discussed.

The test set, randomly selected from the training set, is reasonably well predicted by the model. The prediction error, as given by the RMSE. Mean or median errors are comparable with those of the training set, however, the r_0^2 is considerably lower at 0.48. This is because the line of best fit though the data has a slope 0.95 and an intercept of 0.11. The Pearson's correlation coefficient (r^2) is comparable at 0.58 indicating the model has a good ranking capability. Validation set 1, consisting of data measured up to 6 months after the model was built, is less well predicted by the model again ($r_0^2 = 0.50$). The prediction error has increased while the ranking ability of the model has also decreased. Similarly validation set 2, consisting of data measured between 6 months and 1 year after the model was built, displays a further reduced r_0^2 at 0.40. The error as given by the RMSE consistent with the training and test set however, the mean and median are the largest of all the sets indicating the errors are not normally distributed making statistics requiring such normality less reliable. The final

Table 1
Model statistics for the plasma protein binding QSPR model

Set	r^2 (r_0^2)	RMSE (ME)	Mean (median) absolute error	Slope (intercept)	S.D.	N	Mean (median) distance to 5 nearest neighbours				
							Euclidean[a]	Mahalnobis[b]	Euclidean/PLS component[c]	Euclidean/ PLS coefficient[d]	Tanimoto/ daylight fingerprint[e]
Training set	0.56 (0.56)	0.55 (0.00)	0.42 (0.33)	1.00 (0.00)	0.83	685	0.38 (0.35)	0.41 (0.39)	0.34 (0.29)	0.37 (0.34)	0.35 (0.27)
Test set	0.58 (0.48)	0.54 (−0.02)	0.41 (0.31)	0.95 (0.11)	0.83	210	0.36 (0.35)	0.41 (0.37)	0.31 (0.30)	0.35 (0.35)	0.33 (0.25)
Validation set 1	0.51 (0.50)	0.57 (0.07)	0.45 (0.37)	0.97 (−0.02)	0.81	385	0.58 (0.56)	0.62 (0.56)	0.46 (0.40)	0.55 (0.54)	0.57 (0.52)
Validation set 2	0.44 (0.40)	0.53 (−0.03)	0.45 (0.42)	0.78 (0.31)	0.69	132	0.70 (0.72)	0.72 (0.64)	0.57 (0.51)	0.63 (0.63)	0.84 (0.89)
Literature data	0.34 (0.03)	1.05 (0.60)	0.82 (0.68)	0.81 (−0.41)	1.07	324	1.12 (1.09)	1.01 (0.92)	1.18 (1.07)	1.03 (1.01)	1.07 (1.08)

r^2 is Pearson's correlation coefficient, r_0^2 is the correlation coefficient to the line of unity, RMSE is the root mean square error in prediction and ME is the mean error in prediction. Also reported are the mean absolute and median absolute errors, the slope and intercept of the line of best fit, the standard deviation of the response variable (S.D.), the number of observations in each dataset (N) and the mean and median distance to model for each dataset as described in the text.

[a] Domain of applicability calculated using, mean centred and scaled model descriptors.
[b] Domain of applicability calculated using, model descriptors.
[c] Domain of applicability calculated using, PLS model component space.
[d] Domain of applicability calculated using, descriptors weighted by the PLS model coefficients.
[e] Domain of applicability calculated using, standard 1024 bit daylight fingerprint.

dataset used in this study, the literature set, primarily consists of oral drug molecules. Analysis of Table 1 shows the literature set is poorly predicted by the model, with an r_0^2 of 0.03. The line of best fit has an r^2 of 0.34, with a slope of 0.81 and an intercept of −0.41 indicating the model under predicts the set as a whole, although it does display some ranking ability.

In an effort to rationalize why the QSAR model deteriorated over time we next assessed the similarity of the test and validation molecules to those that were used to build the model.

2.1.1. Domain of applicability methods

The similarity of a query compound to the training set was assessed using: (a) the 17 plasma protein binding QSAR model descriptor or (b) daylight fingerprints [28] and the distance between individual molecules computed using a number of common distance types. Calculations within descriptor space were initially assessed using the Euclidean distance with all descriptors being mean centred and scaled. The average distance to 1, 3, 5, 10, 30, 40 and all near neighbours (NN) in the training set were determined, akin to the study reported by Xu et al. [18].

To assess the effect of correlation within the descriptor matrix we calculated average distance to the 5 nearest neighbours using the Mahalanobis distance in a similar way to that reported by Bruneau and McElroy [22], and in an alternate way by calculating the average Euclidean distance to 5NN using the 4 PLS model components. Finally, we assessed the structural relationship between the error and the distance to model using the average Tanimoto coefficient to 5NN in daylight fingerprint space, and the importance of the individual model descriptors by assessing the average Euclidean distance to 5NN where each descriptor is weighted by their corresponding PLS model coefficient.

The relationship between the different distance measures is assessed by computing the correlation between the average Euclidean distance to the 5NN in mean centred and scaled model descriptor space, a parameter that we typically use as a default in our QSAR studies. These values are reported in Table 2 and Supplementary Information Figure S1. From Table 2 it is apparent that altering the number of nearest neighbours when computing the average Euclidean distance in descriptor space has only a minimal effect on the overall value (r^2 of between 0.88 and 0.99). Indeed, the relationship between the different distances and the prediction error is comparable, as might be expected, suggesting no method offers any major advantage over another. It is also found that the error–distance relationship is comparable using the Mahalanobis distance, the Euclidean distance in PLS component, weighted descriptor space or even the daylight-derived Tanimoto distance (Table 3). For the complete correlation matrix see Supplementary Information Figure S2.

Furthermore, the daylight fingerprint derived distance, which has no direct relevance to the plasma protein binding

Table 2
Correlation between average Euclidean distance to 5 nearest neighbours (NN) to other distance types

Distance	NN	Model space	Correlation with Euclidean distance to 5NN/ descriptors		
			r^2	Slope	Intercept
Euclidean	1	MCS descriptors[a]	0.92	1.09	−0.04
Euclidean	3	MCS descriptors[a]	0.99	1.02	−0.01
Euclidean	10	MCS descriptors[a]	0.98	0.98	0.02
Euclidean	30	MCS descriptors[a]	0.91	0.88	0.07
Euclidean	40	MCS descriptors[a]	0.88	0.87	0.08
Euclidean	5	PLS coefficient* descriptors[b]	0.93	0.90	0.03
Mahalanobis	5	Descriptors[c]	0.78	0.82	0.12
Euclidean	5	PLS components[d]	0.76	1.04	−0.07
Tanimoto	5	Daylight fingerprints[e]	0.54	0.70	0.16

[a] Mean centred and scaled model descriptors.
[b] Descriptors weighted by the PLS model coefficients (MCS).
[c] Model descriptors.
[d] PLS model components space.
[e] Standard 1024 bit daylight fingerprint.

Table 3
Plasma protein binding model: mean, median, standard error of the mean and number of observations for the training, test and combined validation/literature set broken down by distance bin

Dataset	Distance bin	Mean (median) absolute error	Standard error mean	*N*
All	–	0.5 (0.39)	0.01	1736
Training set	Bin 1	0.35 (0.28)	0.02	347
Training set	Bin 2	0.47 (0.37)	0.02	296
Training set	Bin 3	0.59 (0.60)	0.07	32
Training set	Bin 4	0.87 (0.90)	0.17	10
Training set	Bin 5	– (–)	–	0
Test set	Bin 1	0.36 (0.29)	0.03	111
Test set	Bin 2	0.44 (0.34)	0.04	87
Test set	Bin 3	0.62 (0.50)	0.16	12
Test set	Bin 4	– (–)	–	0
Test set	Bin 5	– (–)	–	0
Validation set	Bin 1	0.40 (0.29)	0.04	58
Validation set	Bin 2	0.44 (0.38)	0.02	416
Validation set	Bin 3	0.65 (0.54)	0.04	204
Validation set	Bin 4	0.96 (0.92)	0.07	116
Validation set	Bin 5	0.99 (0.88)	0.09	47

QSAR model itself, shows a reasonable correlation with comparable model descriptor-derived parameters ($r^2 = 0.54$ against Euclidean-5NN), and shows a relationship with the prediction error. This suggests the domain of applicability need not be defined using the descriptors used in the model in line with reports by Sheridan et al. [19]. Given the similarity between the different distance metrics we limit our discussion here to the average Euclidean distance to the 5NN in mean centred and scaled model descriptor space for all models discussed. This parameter will henceforth be referred to simply as the ''distance to model''.

2.1.2. Relationship between the distance to model and the prediction error

The plasma protein binding dataset was binned based on the distance to model into five equally populated bins. The distribution is displayed in Fig. 2a where the percentage of each dataset found in each of the five bins can be seen. The training (i.e. similarity to itself using a leave the closest one out procedure) and test set have a similar distribution, as might be expected since they are randomly selected subsets of the same dataset. As time and thus chemistry progresses, here represented by the two subsequently synthesised and measured validations sets, the overlap with the training set decreases. The newer validation set 2 lies further from the training set than validation set 1, but is found to be closer than the diverse literature-based set.

To assess whether there is a relationship between the distance to model and the error, we plot the r^2, median error and median distance for each dataset in Fig. 3a. From this we can see that as the median distance to model of the individual datasets increases so to does the median error, while the r^2 decreases, indicating that the domain of applicability as defined does show a relationship with the prediction error. However, while this may allow us to monitor gross changes in the predictive ability of the model over time, it says little about our ability to estimate the reliability of prediction of new observations. This we can do by comparing the mean (or median) absolute errors for observation in each of the 5 distance bins defined earlier, broken down according to the training, test, validation and literature sets. The validation and literature sets were combined to allow for more statistically meaningful tests to be performed since the combined dataset spans a less restricted number of distance bins. The relationship between the absolute error and the distance is illustrated in Fig. 4a and given in tabular form (Table 4). The relationships illustrated are statistically significant above the commonly used 95% confidence level, according to the Kruskal Wallis ANOVA test, both collectively and when broken down by each set.

To use the error–distance relationship to estimate the prediction error of new observations one must first be confident that the relationship in the training set is reproduced in the test and validations sets. Here we find that the mean absolute error of the training, test and combined validation/literature sets are comparable for bin 1 at 0.35, 0.36, 0.40, respectively. Moving further away from training space to distance bin 2, the mean absolute error has increased compared to bin 1 in each of the three set, and also of comparable magnitude (0.47, 0.44, 0.44, respectively). In distance bin 3 the values have again increased in magnitude and of comparable size for each set (0.59, 0.62, 0.65). In bin 4, the training and combined validation set mean absolute errors are larger than the comparable values in bin 3, at 0.87 and 0.96, respectively. No test set observation is found outside of distance bin 3, while distance bin 5 is only occupied by validation data. The latter 47 observations have a mean absolute error of 0.99 in line with an extrapolation based on the training set distance–error relationship.

Collectively, these results indicate the domain of applicability method used for the plasma protein binding model could be used to estimate the prediction error of new observations. It should be emphasized that observations found far from model space may still have a low prediction error, as can be seen from the error bars in Fig. 4a, however, the probability is much lower than for those close to model space.

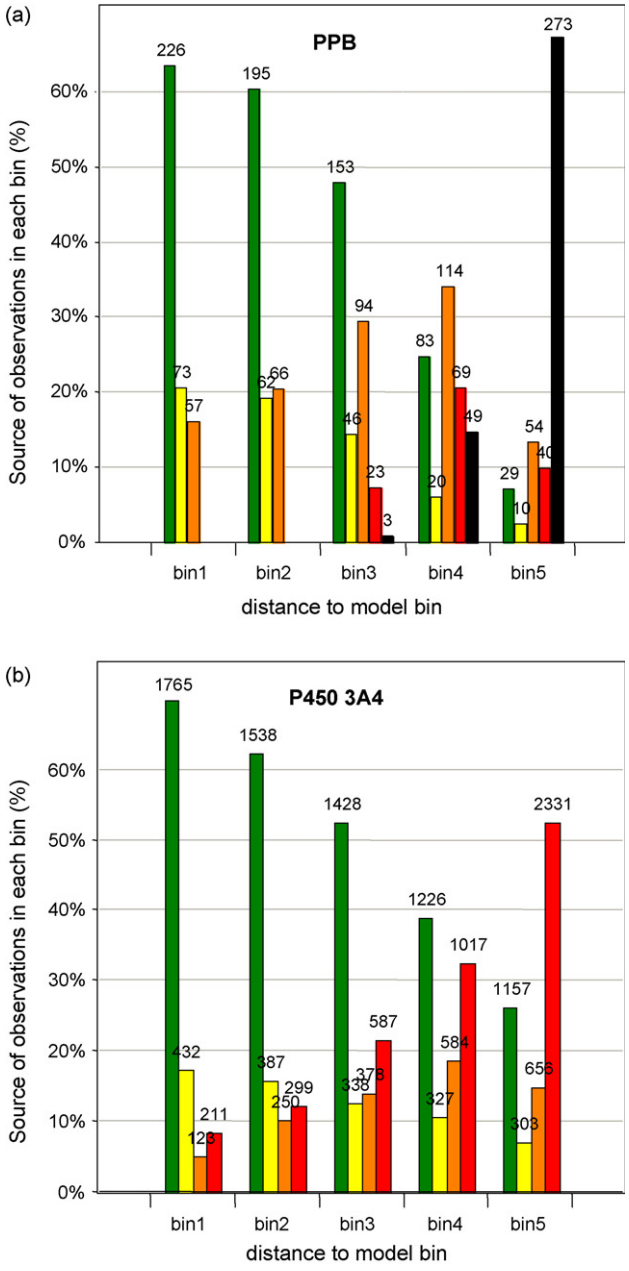

Fig. 2. (a and b) Distributions of distance to model (average Euclidean distance to the 5NN in MCS model descriptors space) for the global plasma protein binding (top) and global P450 3A4 (bottom) QSAR models. In each bin, from left, to right are the training set (green), test set (yellow), validation set 1 (orange) and validation set 2 (red). Also included is the literature-based set (black) for the PPB model. Each model dataset was binned into five equally populated bins. The source of the observations in each bin is given as a percentage on the *Y*-axis and the absolute number above each column.

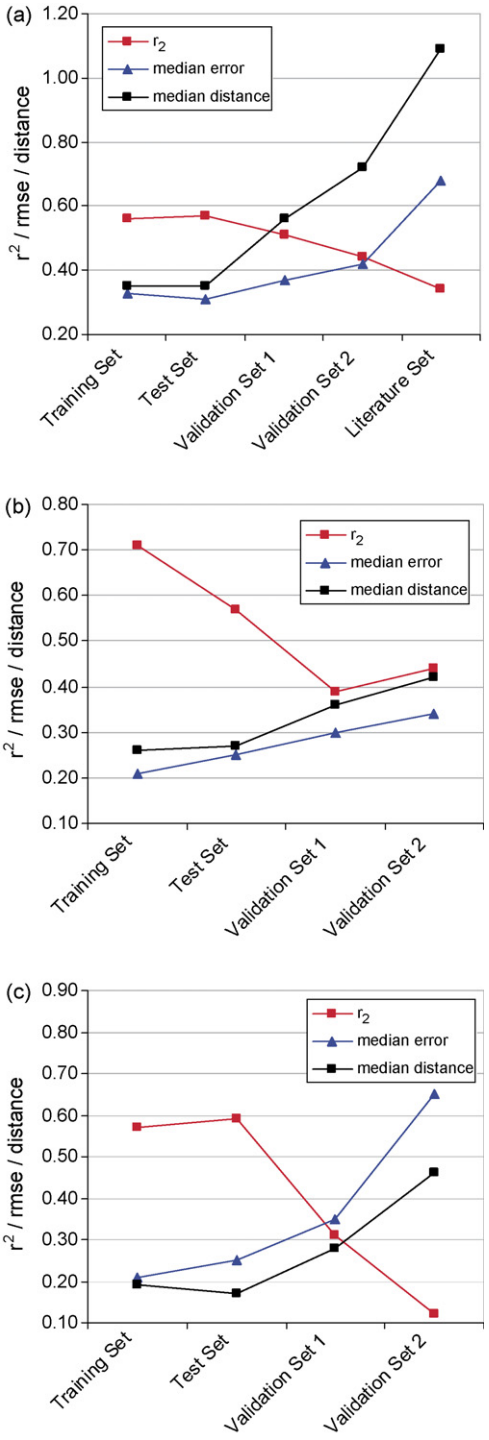

Fig. 3. (a–c) Plot of the r^2, median error and distance against the corresponding dataset for global plasma protein binding QSPR (top), the global P450 3A4 QSAR model (middle) and the local P450 3A4 QSAR model for program A (bottom).

2.2. Global P450 3A4 inhibition QSAR model

To demonstrate the generalizability of this methodology we now report its use on a second, phenomenologically dissimilar example. In the second example a non-linear statistical technique has been used to build a "black box" model on a P450 inhibition dataset using a collection of 1D and 2D molecular descriptors, fragment counts and electrotopological descriptors. The model itself consists of a consensus of 10 individual radian basis function (RBF) neural networks (NNs), thus preventing any meaningful interpretation of the QSAR, which may also have implications for the applicability domain.

The QSAR displays a large r_0^2 of 0.71 with a pIC_{50} prediction error of 0.36 log units, or 2.3-fold in terms of the IC_{50}. This value is close to the experimental variability of the assay (~2-fold) suggesting such a model could prove very useful were its

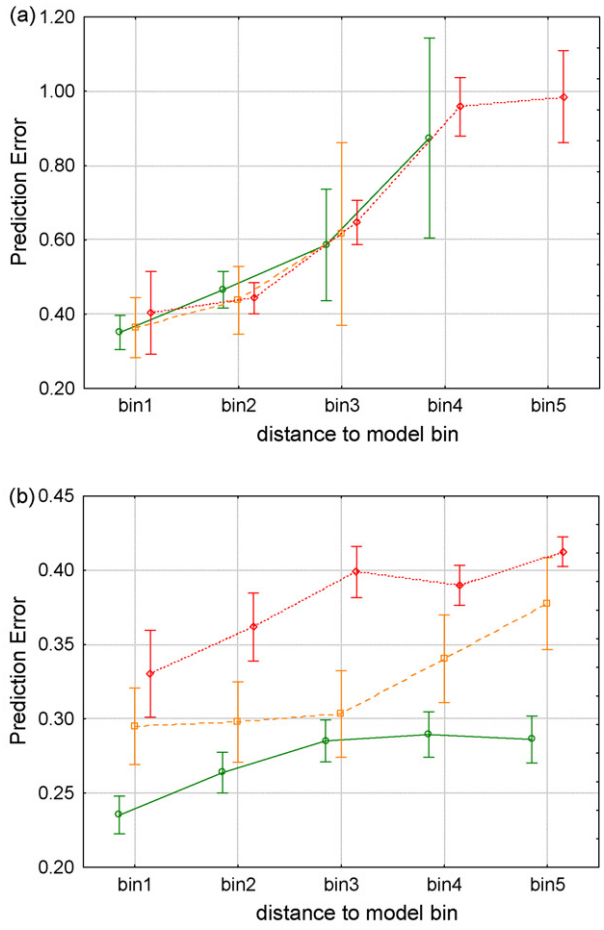

Fig. 4. (a and b) Plot of the absolute error in prediction vs. the binned distance to model for the global plasma protein binding (top) and P450 3A4 (bottom). The binning scheme used here is the same as in Fig. 2, the colour scheme corresponds to the following: training set (green-solid line), test set (orange-dashed line) and the remaining sets combined (red-dotted line). The error bars are proportional to the square root of the standard deviation in the prediction errors of each bin and inversely proportional to the number of observations. Few observations in the training and test set lie in bins 4 and 5 giving rise to much larger error bars.

predictive performance to maintain itself on the test and validation sets (Table 5). On the test set we find the performance has dropped somewhat with an r_0^2 of 0.56, yet the prediction error of 2.6-fold is still reasonable for a purely in silico approach. Such a drop could simply be a reflection of (a) the domain of applicability, (b) model over fitting or (c) a result of both.

Validation set 1, is a more difficult test for the model since it consists of data measured up to 6 months after the model was built. The model displays an $r_0^2 = 0.36$ and 3-fold prediction error. Surprisingly the overall variation of the data in validation set 2, measured between 6 months and 1 year after the model was built, is better ranked by the model ($r_0^2 = 0.43$), however, the prediction error is larger at 3.4-fold meaning the greater ranking ability is simply a result of a greater range in the data.

2.2.1. Relationship between the distance to model and the prediction error

We now discuss the domain of applicability of the QSAR model, to try and rationalize why the prediction error increases by between 0.04 and 0.06 log units moving to each subsequent dataset. To this end, the dataset was binned based on the distance to model into five bins of approximately the same size. The distribution of the data is illustrated in Fig. 2b and this follows the same trends as found in the PPB dataset. The training set distribution unsurprisingly is the tightest and closest to zero (based on a leave the closest one out procedure) and this is similar to that of the randomly selected test set. Data obtained 6 months later, namely the validation set 1, lies further from model space than the test set, but is noticeably closer than the even newer validation set 2 (Fig. 2b). The modest relationship between the distance to model and the error can be demonstrated in Fig. 3b. It is evident that as the median distance increases so to does the median error, while the r^2 decreases. It should be noted that the larger than expected r^2 observed for validation set 2 is a result of the larger variance in that dataset, not because the model performs more effectively. This highlights the problem of using the r^2 alone to assess the

Table 4
P450 3A4 inhibition model (global): mean, median, standard error of the mean and number of observations for the training, test and combined validation/literature set broken down by distance bin

Dataset	Distance bin	*N*	Mean (median) absolute error	Standard error mean
All	–	15,336	0.33 (0.26)	0.28
Training set	Bin 1	1,765	0.24 (0.18)	0.20
Training set	Bin 2	1,537	0.26 (0.20)	0.24
Training set	Bin 3	1,428	0.29 (0.23)	0.24
Training set	Bin 4	1,226	0.29 (0.23)	0.24
Training set	Bin 5	1,157	0.29 (0.23)	0.24
Test set	Bin 1	432	0.29 (0.24)	0.26
Test set	Bin 2	387	0.30 (0.24)	0.25
Test set	Bin 3	338	0.30 (0.23)	0.26
Test set	Bin 4	327	0.34 (0.29)	0.28
Test set	Bin 5	303	0.38 (0.30)	0.32
Validation set	Bin 1	334	0.33 (0.27)	0.27
Validation set	Bin 2	549	0.36 (0.29)	0.29
Validation set	Bin 3	965	0.40 (0.32)	0.31
Validation set	Bin 4	1,601	0.39 (0.33)	0.31
Validation set	Bin 5	2,987	0.41 (0.34)	0.33

Table 5
Model statistics for the 3A4 cytochrome P450 inhibition QSAR models

	Date (MM/YY)	r^2 (r_0^2)	RMSE (ME)	Mean (median) absolute error	Mean (median) absolute distance	S.D.	N
Global dataset							
Training set	06/04–09/05	0.71 (0.71)	0.36 (0.00)	0.27 (0.21)	0.26 (0.25)	0.66	7113
Test set	06/04–09/05	0.57 (0.56)	0.42 (0.01)	0.32 (0.25)	0.27 (0.25)	0.64	1787
Validation set 1	09/05–03/06	0.39 (0.36)	0.48 (0.01)	0.37 (0.30)	0.36 (0.33)	0.60	1991
Validation set 2	03/06–09/06	0.44 (0.43)	0.52 (−0.06)	0.41 (0.34)	0.42 (0.42)	0.51	4445
Program A							
Training set	06/04–07/05	0.57 (0.57)	0.31 (0.00)	0.23 (0.19)	0.25 (0.21)	0.47	431
Test set	06/04–07/05	0.59 (0.58)	0.31 (−0.05)	0.23 (0.17)	0.28 (0.25)	0.46	108
Validation set 1	08/05–11/05	0.31 (0.02)	0.46 (−0.19)	0.35 (0.28)	0.40 (0.35)	0.47	530
Validation set 2	12/05–09/06	0.12 (0.00)	0.61 (−0.36)	0.51 (0.46)	0.68 (0.65)	0.52	529
Program B							
Training set	02/05–10/05	0.38 (0.38)	0.40 (0.00)	0.31 (0.26)	0.30 (0.26)	0.52	182
Test set	02/05–10/05	0.32 (0.30)	0.39 (−0.05)	0.29 (0.21)	0.26 (0.23)	0.47	47
Validation set 1	10/05–04/06	0.45 (0.42)	0.40 (0.01)	0.31 (0.26)	0.41 (0.37)	0.52	229
Validation set 2	04/06–09/06	0.34 (0.01)	0.51 (0.29)	0.42 (0.36)	0.46 (0.41)	0.51	202
Program C							
Training set	06/04–05/05	0.40 (0.40)	0.36 (0.00)	0.26 (0.23)	0.33 (0.34)	0.45	111
Test set	06/04–05/05	0.34 (0.30)	0.40 (0.00)	0.31 (0.26)	0.38 (0.36)	0.48	28
Validation set 1	05/05–11/05	0.10 (0.00)	0.42 (−0.15)	0.33 (0.23)	0.24 (0.21)	0.40	136
Validation set 2	11/05–09/06	0.11 (0.00)	0.49 (−0.22)	0.41 (0.38)	0.60 (0.52)	0.44	136

The experimental date range of each dataset is reported along with the Pearson's correlation coefficient r^2, the correlation coefficient to the line of unity, (r_0^2), is the root mean square error (RMSE) and the mean error (ME). Also reported are the mean absolute and median absolute errors, the slope and intercept of the line of best fit, the standard deviation of the response variable (S.D.), and N, the number of observations in each dataset.

predictive ability of QSAR models without regard to the distribution of the data and the prediction error [24].

Next, we assess the utility of the method to estimate the prediction error of new observations predicted by the model. As before we do this by comparing the absolute error of observations found in each of the 5 distance bins, broken down according to their source (the training set, test set and the combined validation sets). The relationship between the distance and the error is illustrated in Fig. 4b and given in tabular form in Table 5. It is found to be statistically significant, when broken down by individual set as well as collectively, however, there are clear differences compared to the plasma protein binding case (Fig. 4a). While it is clear that the mean absolute error increases with increasing distance overall, the absolute errors are markedly different for observations from the 3 different sets found in the same distance bin. For example, in bin 1 the mean error (ME) of training set observations is 0.24, compared to 0.29 for the test set and 0.33 for the combined validation sets. Also, the mean error of test set observations in bin 4 is the same as the mean error of the combined validation set observations in bin 1, illustrating that the method cannot be used to estimate the absolute prediction error of new observations. This is because the increase in prediction error, for what amounts to a negligible distance change, may simply be a reflection of the model being overfitted. That said, the method is able to rank the observations within a given set according to their prediction error.

The above finding is somewhat worrying given the large dataset employed here to train the model (∼7000), the use of a neural network method according to best practice as defined by the software vendor, as well as commonly used descriptor types [29,30]. This result suggests great care should be taken when training QSAR models, especially on relatively small datasets using complex, non-linear statistical methods, coupled with large, often sparsely populated descriptor sets. This result also highlights the utility of the domain of applicability method to identify overfitting.

2.3. Local P450 3A4 inhibition QSAR models

Local QSAR models differ from global models in that they are built on datasets where the breadth of chemical diversity is restricted, typically limited to one or more related chemotype, at least in the earliest stages of a drug discovery program. The structural diversity will evolve over time though the use of bioisosteric replacements of the core templates by medicinal chemists and through the use of computational scaffold hopping techniques to introduce additional lead series. This process however, can pose a problem for local QSAR models in general as larger changes in either the structure or properties of the original chemotypes can move the new molecules outside of what is an already limited area of overall chemical space.

Program specific models were generated using the three most populated programs from the global P450 3A4 inhibition dataset. The complete program datasets were split in the same way as the global datasets, by date, with the earliest set being partitioned randomly into a training and test set and the two subsequent sets being used for the purpose of validation. The domain of applicability has been assessed using the average Euclidean distance to 5NN in MCS model descriptor space

Program A displays an r_0^2 of 0.57, a cross-validated $q^2 = 0.52$ and a prediction error of 0.31 log units which is as good as one could expect based on the experimental variability in the assay (Table 5). The model is equally predictive of the test set with

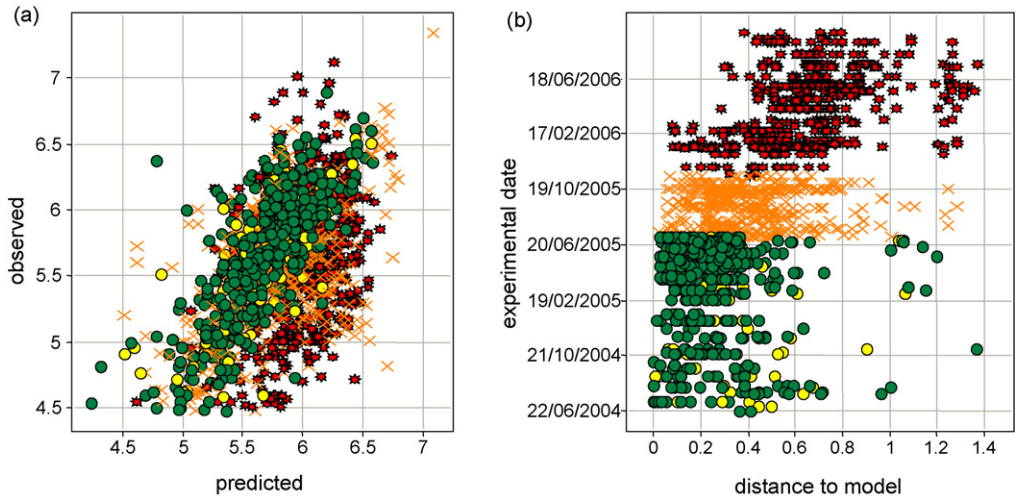

Fig. 5. (a and b) Plot of the observed vs. predicted pIC50s for the 3A4 program A QSAR model (left) and plot of experimental date vs. the distance to model (right). The colour scheme used here is the same as in Fig. 2.

comparable r_0^2 and RMSEs, however, the validation set 1 displays a markedly reduced correlation coefficient ($r_0^2 = 0.02$), while in the validation set 2 it is ($r_0^2 = 0.00$).

The results are given numerically in Table 5 and graphically in Fig. 5a, showing the poor correlation between observed and predicted for the validation set 1 and 2. The main chemotype found in the validation set represents a sizeable change to the original lead scaffold. Worryingly, these molecules are over predicted by between 0.19 and 0.36 log units meaning they could have been excluded from synthesis unnecessarily were a program team to rely exclusively on a model, which apparently appeared predictive.

Plotting the distance to model against the date of experimental measurement shows a clear relationship and this could be used to decide when the model needs to be discarded or regenerated. The relationship between the distance to model and the prediction error can be seen in a more quantitatively fashion in Fig. 3c. As the median distance to model increases across the 4 distinct sets so to does the median error, while the corresponding r^2 decreases.

To assess whether the catastrophic model failure found for program A QSAR was a common feature of QSARs or just a chance effect we built similar models for programs B and C with an assessment being made of their domain of applicability.

Program B displays an r_0^2 of 0.38, a cross-validated $q^2 = 0.36$ and a prediction error of 0.40 log units (Table 5). The model is equally predictive of the test set with comparable r_0^2 and RMSEs, as is the validation set. Although the r_0^2 is larger, the RMSE is the same as the training and test set. This is somewhat surprising given that the mean or median distance to model is considerably larger than that for the test set. Additionally, validation set 2 is predicted with noticeably lower accuracy compared to the other 3 sets, having an RMSE of 0.51 log units. This result is also unexpected since the difference between the median distances in the test and validation set are larger than the difference between the two validation sets. Based on the plot of experimental date versus the distance to model for the dataset it appears the synthetic alterations in the molecules were more limited compared to program A (Fig. 6a). Alternatively, the poor correspondence between the distance and the error could

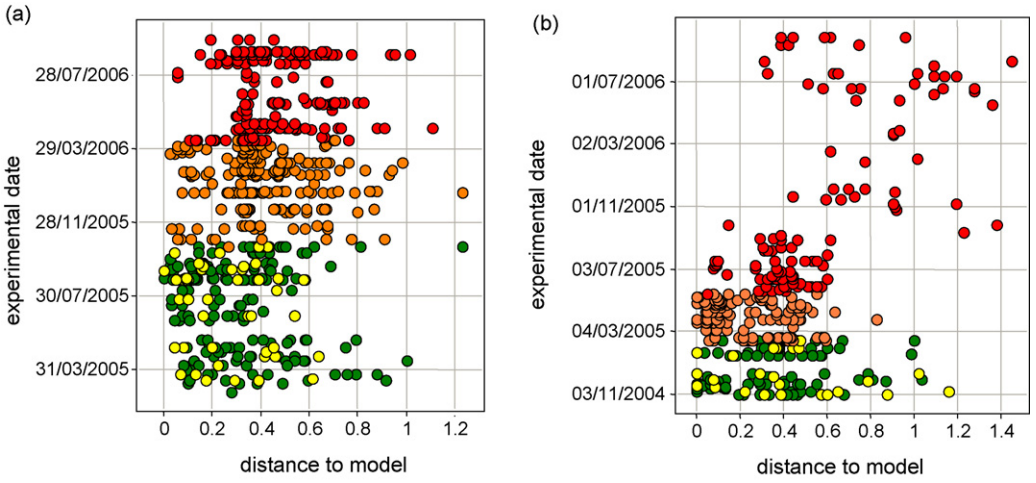

Fig. 6. (a and b) Plot of experimental date vs. the distance to model for program models B (right) and C (left). The colour scheme used here is the same as in Fig. 2.

be an artefact of the relatively limited description of the chemical diversity used here.

Program C displays an r_0^2 of 0.40, a cross-validated $q^2 = 0.29$ and a RMSE of 0.36 log units (Table 5). While the model ranks the data less well compared to program B, the actual prediction error is lower because the data spans a more restricted range. As such it might prove useful in a program as a prioritization tool rather than an accurate in silico predictor. Unfortunately, the prediction of the test set is less effective, displaying a lower r_0^2 of 0.30 and a larger RMSE. This deterioration in performance continues through to both validation sets with each one having an r_0^2 of 0. The model does however display a very limited ranking ability as evidenced by the line of best fit showing r^2 of ~0.10. Surprisingly, from an analysis of the median distance to model values it appears that the validation set 2 is closer to the training set, than itself when employing a leave one out procedure yet it is still poorly predicted (Table 5). In contrast validation set 2 lies far from model space as can be seen from a plot of experimental date versus the distance to model (Fig. 6b).

The results described here on both global and program specific QSAR models collectively indicate that the domain of applicability can prove useful in monitoring the predictive performance. Indeed it could be argued that it is more important to consider this in local QSAR applications because the limited chemical diversity of the model itself means it is relatively easy to stray outside the predictive domain of the model with limited chemistry iterations.It is difficult to define a generic rule of thumb as to when a QSAR model needs to be rebuilt, as this is dependent on how the model is to be applied within drug discovery programs. While we always desire the most predictive model in general, a weak model, for example, that allows one to bias a set the set of building blocks, to lower protein binding, based on the predictions of the virtual compounds before array synthesis, need not be as accurate as one required within a program environment to lower the 3A4 inhibition of a particular chemotype by 2 log units for example. The relationship between model application and accuracy is discussed in more detail elsewhere [32], where it is argued that weak models can still be used to make decisions, albeit at much lower resolution.

3. Conclusions

A methodology to assess the domain of applicability has been described that has proved useful in the interpretation of QSAR models of different type generated in GlaxoSmithKline. A number of distances and descriptor have been implemented in our in house system, as it appears that no method performs best in all situations.

The importance of the domain of applicability has been demonstrated on two global and three local ADMET models, illustrating that one can use the distance to model space to estimate the reliability of a given prediction. Such methodologies also offer us a way to easily monitor the performance of a QSAR over time, providing us with a way to identify when a model needs to be regenerated, as well as helping to identify if a model was originally overfitted or not. The domain of applicability concept is particularly important for local models since it is relatively easy to stray outside the predictive domain of the model with a small number of chemistry iterations.

A number of quite predictive QSAR models have been described herein, showing good performance in cross-validation and initial testing. However, the performance of these models deteriorates with time because the subsequent validation sets lie further outside the applicability domain of the original model. This study highlights the need for more rigorous temporal validation studies [31] and a consideration of the domain of applicability, at least for global models, which often have a shelf life of more than a year. Assessing the robustness of a QSAR model to evolving chemistry over a 6-month period for example, allows one to more effectively quantify the likely deterioration over time that will naturally occur. This however, poses a dilemma since this data should ultimately be incorporated into the model to create the most predictive model with the greatest chemical coverage.

4. Experimental procedures

4.1. Global plasma protein binding QSAR model

The human plasma protein binding model discussed in this study has been described in detail elsewhere [24]. Briefly, the 4 component PLS regression model was built and tested on a dataset of 897 measurements which were split randomly ~75:25 into a training and test set. Approximately forty 1D and 2D physico-chemical descriptors [24] were used to generate the initial model, with descriptor reduction achieved in GOLPE [33] using Fraction factorial selection and D-optimal design. To facilitate greater interpretability, a final refinement was performed in SIMCA P10 [34] by removing descriptors with negligible impact on the first two components (components 1 and 2 describe ~85% of the total variation).

To obtain a more realistic measure of the predictive performance of the model, and assess the effect of the domain of applicability, two subsequent validation sets were defined, the first consisting of data measured up to 6 months after the model was built ($N = 132$), and the second between 6 months and 1½ years after the model was built ($N = 382$).

4.2. Global P450 3A4 inhibition QSAR model

A P450 3A4 inhibition model was built and tested on an in house dataset of ~8900 molecules, which was randomly split 80:20 into a training and test set. Approximately forty 1D and 2D physico-chemical descriptors used in the plasma protein binding model along with ~120 electrotopological (estate) [14] descriptors and a collection of ~150 custom fragmental descriptors known to be important for P450 inhibition were used to build the model. These included fragment counts of heterocycles such as naked pyridines, imidazoles, etc. and is similar to the approach taken by Gepp and Hutter [10] in their modeling studies on hERG inhibition.

A RBF neural network [35] was used to relate the 3A4 activity to the chosen descriptors. 200 models were built using

the standard settings in Statistica 7.0 [36], employing the automatic descriptor pruning option. The final model consisted of a consensus of the 10 best QSARs, chosen based on their performance in bootstrapping. The models used between 44 and 62 descriptors (14 molecular properties, 10 e-states and 38 smarts descriptors), with certain key descriptors such as clogP, molecular weight, certain e-states and fragment descriptor, such as pyridine counts appearing in all networks.

To obtain a more realistic measure of the predictive performance of the model, and to assess the effect of the domain of applicability, two subsequent validation sets were defined. The first validation set consisted of data measured up to 6 months after the model was built ($N = 1991$), and the second between 6 months and 1 year ($N = 4445$).

4.3. Local P450 3A4 inhibition QSAR models

To assess the potential impact of the domain of applicability on local QSAR models, experiments were performed on the larger P450 3A4 dataset. The three most populated program series run over the lifetime of the assay were extracted and approximately forty physico-chemical descriptors used in the other QSAR models were computed. The datasets, each spanning approximately two years, were split into three separate portions, the first consisting of approximately the first year of data, with the remainder being split temporally into two validation sets each of approximately 6 months size.

PLS regression models were built and tested on the first year's worth of data which had been randomly split ~80:20 into a training set and test set. The models were built in SIMCA P10, fitting between two and three components optimally based on the standard fitting routine implemented in the software package.

4.4. QSAR model statistics

We use a variety of statistics to assess the performance of the QSAR models which we have documented elsewhere [12,24]. Briefly, the coefficient of determination (r_0^2) is used to assess the correlation of the predictions to the unity line. In cases where the model predictions deviate from the unity line it is useful to evaluate the Pearson product moment correlation coefficient (r^2), a measure of the overall ranking ability. We also report the RMSE in prediction and the mean error in prediction for each dataset discussed.

For PLS regression models the cross-validated q^2 is reported. This value is computed in SIMCA-P10 by generating 7 different models, each by leaving 1/7 of the data out each time.

4.5. Domain of applicability methods

In our implementation we can compute the distance in descriptor space using the Euclidean distance (Eq. (5)), Mahalanobis distance to N nearest neighbours [22] (Eq. (6)) and the cosine distance (Eq. (7)). The Tanimoto coefficient is used for calculations in daylight fingerprint space as standard (Eq. (8)). The number of nearest neighbours that can be specified ranges from 1 to N where N is the number of training set datapoints.

$$\mathrm{d}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{5}$$

$$\mathrm{d}(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \sum^{-1} (\vec{x}-\vec{y})} \tag{6}$$

$$\mathrm{d}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \tag{7}$$

$$\mathrm{d}(x,y) = 1 - \frac{N_{xy}}{N_x + N_y - N_{xy}} \tag{8}$$

Here, n symbolizes the total number of observations for which a pairwise comparison is performed (x and y), and i is the number of descriptors. $\vec{x}$ and $\vec{y}$ are vectors in the Mahalanobis distance to NN (as opposed to the true Mahalanobis distance) and Σ^{-1} is the covariance matrix. In Eq. (8) N_x refers to the number of bit set on x, N_y refers to the number of bits set on y and N_{xy} the number of common bits set on.

4.5.1. Distance to model process

A schematic representation of the suite of perl and C++ programs written to calculate the domain of applicability can be found in the Supplementary Information (Figure S3).

The first step in the process involves the definition of the input variable and the pretreatment required. Typically pretreatment involves (a) the mean centring and scaling of the data, (b) PLS/PCA component definition or (c) in unscaled descriptor space weighted by the corresponding regression coefficients. The former may prove useful as an alternate means to Mahalanobis distance in removing redundancy due to correlating descriptors, while the latter might prove useful as descriptors with the biggest impact in the model are given the greatest weight.

The resulting scale will differ depending on the distance type and number of descriptors used in the calculation so we subsequently normalize the distance between 0 and 1, where 0 represents v-close to model space and 1 very far. This is done by computing the average distance to N nearest neighbours for each training set molecule to itself, employing a leave the closest one out procedure. The smallest average nearest neighbours distance is assigned a value of 0 and the distance corresponding to the 95% percentile is assigned a value of 1. The 100% percentile is not used as this typically represents a significant outlier, resulting in a heavily skewed distribution. Any query found closer to the training set than the minimum distance from the leave one out procedure is reported a distance of zero.

The domain of applicability computed in daylight fingerprint space does not require any initial normalization of the input bit string, however, the distances are normalized in a similar manner to the descriptor-based distances using the training set as a reference. As such, 1 now represents a molecule far from model space and 0 within.

4.5.2. Domain of applicability statistics

The overall relationship between the distance and the errors was assessed by first binning the complete datasets into five equally sized bins. As the absolute errors of bin are not normally distributed (the absolute errors follow an exponential distribution compared to the normally distributed errors), it is not possible to assess the statistical significance of the mean values of the bins using parametric statistics such as ANOVA. We use the non-parametric Kruskal–Wallis ANOVA test to assess the statistical significance at the 95% confidence level [36].

As with any QSAR, one uses the training set to fit the relationship and the test set to determine the true predictivity of the model. Here, we therefore use the test set to determine the relationship between the distance to model and the error in an analogous manner to Bruneau [25]. In theory one could use the mean error of the 3 nearest neighbours in the training set as the prediction error of the query compound, however, we find this gives very poor results compared to a method that involves the use of an independent test set to assign the prediction error. To generate the distance to model relationship to allow us to predict the likely error of future compounds we first compute the distance of each test set compound to the N nearest neighbours in the training set to allow us to generate the so called model. Since the relationship between the distance and the error is relatively weak we bin the test set compounds into approximately 3–5 distance bins and calculate the median error for each. To then predict the error of a new query compound we first calculate its distance to the N nearest neighbours in the training set and we then, based on its computed distance, assign it the corresponding error of the particular distance bin for the test set.

In this way we can use the test set as a reference so we can quantify the deviation of subsequent data from model space. The median absolute error is used here as standard as it was found to be a slightly more reliable measure of the overall error given that the statistics such as the mean or RMSE require normally distributed errors. In a small number of cases the errors of a set may not be perfectly normal due, to the small n numbers.

Acknowledgments

The authors would like to thank Drs. Anne Hersey, Andrew Leach, Gavin Harper and Nathaniel Woody for their help and useful discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2008.01.002.

References

[1] R.A. Prentis, Y. Lis, S.R. Walker, Pharmaceutical innovation by the seven UK-owned pharmaceutical companies (1964–1985), Br. J. Clin. Pharmacol. 25 (1988) 387–396.
[2] T. Kennedy, Managing the drug discovery/development interface, DDT 2 (1997) 436–444.
[3] R. Frank, F. Hargreaves, Clin. Biomarkers Drug Discov. Dev. Nat. Rev. Drug Discov. 2 (2003) 566–580.
[4] D. Hansch, A. Leo, S.B. Mekapati, A. Kurup, QSAR and ADME, Bioorg. Med. Chem. 12 (2004) 3391–3400.
[5] F. Lombardo, F. Gifford, M.Y. Shalaeva, In silico ADME prediction: data, models, Facts Myths. Mini Rev. Med. Chem. 3 (2003) 861–875.
[6] U. Norinder, C.A.S. Bergstrom, Prediction of ADMET properties, Chem. Med. Chem. 1 (2006) 920–937.
[7] T. Hou, J. Wang, W. Zhang, W. Wang, X. Xu, Recent advances in computational prediction of drug absorption and permeability in drug discovery, Curr. Med. Chem. 13 (2006) 2653–2667.
[8] C. Hanch, T.J. Fugita, p-σ-π analysis. A method for the correlation of biological activity and chemical structure, J. Am. Chem. Soc. 86 (1964) 1616–1626.
[9] E. Byvatov, K. Baringhaus, G. Schneider, H. Matter, A virtual screening filter for identification of cytochrome P450 2C9 (CYP2C9) inhibitors, QSAR Comb. Sci. 26 (2006) 618–628.
[10] M.M. Gepp, M.C. Hutter, Determination of hERG channel blockers using a decision tree, Bioorg. Med. Chem. 14 (2006) 5325–5332.
[11] S. Sciabola, I. Morao, M.J. de Groot, Complex models pharmacophoric fingerprint method (TOPP) for 3D-QSAR modeling: application to CYP2D6 metabolic stability, J. Chem. Inf. Model. 47 (2007) 76–84.
[12] M.P. Gleeson, N.J. Waters, S.W. Paine, A.M. Davis, In silico human and rat Vss quantitative structure-activity relationship models, J. Med. Chem. 49 (2006) 1953–1963.
[13] J.R. Votano, M. Parham, L.M. Hall, L.H. Hall, L.B. Kier, S. Oloff, A. Tropsha, QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation, J. Med. Chem. 49 (2006) 7169–7181.
[14] L.H. Hall, L.B. Kier, Electrotopological state indices for atoms types: a novel combination of electronic. Topological and valence state information, J. Chem. Inf. Comput. Sci. 35 (1995) 1039–1045.
[15] F.R. Burden, Molecular identification number for substructure searches, J. Chem. Inf. Comput. Sci. 29 (1989) 225–227.
[16] R.S. Pearlman, K.M. Smith, Novel software tools for chemical diversity, Perspect. Drug Discov. Des. 9 (1998) 339–353.
[17] M.J. McGregor, S.M. Muskal, Pharmacophore fingerprinting. 1. Application to QSAR and focused library design, J. Chem. Inf. Comput. Sci. 39 (1999) 569–574.
[18] Y. Xu, H. Gao, Dimension related distance and its application in QSAR/QSPR model error estimation, QSAR Comb. Sci. 22 (2003) 422–429.
[19] R.P. Sheridan, B.P. Feuston, V.N. Maiorov, Kearsley similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR, J. Chem. Inf. Comput. Sci. 44 (2004) 1912–1928.
[20] R. Todeschini, V. Consonni, M. Pavan, A distance measure between models: a tool for similarity/diversity analysis of model populations, Chemometr. Int. Lab. Syst. 80 (2004) 55–61.
[21] I.V. Tetko, P. Bruneau, H. Mewes, D.C. Rohrer, G.I. Poda, Can we estimate the accuracy of ADME–Tox predictions? DDT 11 (2006) 700–707.
[22] P. Bruneau, N.R. McElroy, logD7.4 modeling using Bayesian regularized neural networks. Assessment and correction of the errors of prediction, J. Chem. Inf. Model. 46 (2006) 1379–1387.
[23] T.W. Schultz, M. Hewitt, T.I. Netzeva, M.T.D. Cronin, Assessing applicability domains of toxicological QSARs: definition, confidence in predicted values, and the role of mechanisms of action, QSAR Comb. Sci. 26 (2007) 238–254.
[24] M.P. Gleeson, Plasma protein binding affinity and its relationship to molecular structure: an in-silico analysis, J. Med. Chem. 50 (2007) 101–112.
[25] P. Bruneau, Search for predictive generic model of aqueous solubility using Bayesian neural nets, J. Chem. Inf. Comput. Sci. 41 (2001) 1605–1616.
[26] S.L. Rodgers, A.M. Davis, N.P. Tomkinson, H. van de Waterbeemd, QSAR modeling using automatically updating correction libraries: application to a human plasma protein binding model, J. Chem. Inf. Model. 47 (2007) 2401–2407.

[27] T.R. Stouch, J.R. Kenyon, S.R. Johnson, X.Q. Chen, A. Doweyko, Y.J. Li, In silico ADME/Tox: why models fail, J. Comput. Aided Mol. Des. 17 (2003) 83–92.
[28] Daylight Chemical Information Systems, Inc., 120 Vantis–Suite 550–Aliso Viejo, CA, 92656, USA (http://www.daylight.com).
[29] G. Cruciani, M. Pastor, R. Mannhold, Suitability of molecular descriptors for database mining. A comparative analysis, J. Med. Chem. 45 (2002) 2685–2694.
[30] P. Gedeck, B. Rohde, C. Bartels, QSAR—how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets, J. Chem. Inf. Model. 46 (2006) 1924–1936.
[31] S.A. Rodgers, A.M. Davis, H. van de Waterbeemd, Time-series QSAR analysis of human plasma protein binding data, QSAR Comb. Sci. 26 (2006) 511–521.
[32] M.P. Gleeson, A.M. Davis, K.K. Chohan, S.W. Paine, S. Boyer, C.L. Gavaghan, C. Hasselgren-Arnby, C. Kankkonen, N. Albertson, Generation of in-silico cytochrome P450 1A2, 2C9, 2C19, 2D6, and 3A4 inhibition QSAR models, J. Comput. Aided Mol. Des. 21 (2007) 559–573.
[33] GOLPE: Multivariate Infometric Analysis Srl., Viale dei Castagni 16, Perugia, Italy.
[34] SIMCA-P 10, Umetrics, Tvistevägen 48, Box 7960, SE-907 19 Umeå, Sweden.
[35] J. Tetteh, T. Suzuki, E. Metcalfe, S. Howells, Quantitative structure-property relationships for the estimation of boiling point and flash point using a radial basis function neural network, J. Chem. Inf. Comput. Sci. 39 (1999) 491–507.
[36] Statistica System Reference, Statsoft Inc., 2300 East Tulsa, OK74104, USA (http://www.statsoft.com).