Journal of Molecular Graphics and Modelling 29 (2010) 529–537

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

A rapid computational filter for predicting the rate of human renal clearance

Stuart W. Paine[a,*], Patrick Barton[a], James Bird[a], Rebecca Denton[a], Karelle Menochet[c,1], Aaron Smith[a], Nicholas P. Tomkinson[b], Kamaldeep K. Chohan[b,*]

[a] Department of Discovery DMPK, AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, United Kingdom
[b] Department of Chemistry, AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, United Kingdom
[c] School of Pharmacy and Pharmaceutical Science, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, United Kingdom

ARTICLE INFO

Article history:
Received 13 April 2010
Received in revised form 8 October 2010
Accepted 12 October 2010
Available online 20 October 2010

Keywords:
Quantitative structure–activity relationship (QSAR)
Human renal clearance
Partial Least Squares (PLS)
Classification And Regression Trees (CART)
Random Forest (RF)

ABSTRACT

In silico models that predict the rate of human renal clearance for a diverse set of drugs, that exhibit both active secretion and net re-absorption, have been produced using three statistical approaches. Partial Least Squares (PLS) and Random Forests (RF) have been used to produce continuous models whereas Classification And Regression Trees (CART) has only been used for a classification model. The best models generated from either PLS or RF produce significant models that can predict acids/zwitterions, bases and neutrals with approximate average fold errors of 3, 3 and 4, respectively, for an independent test set that covers oral drug-like property space. These models contain additional information on top of any influence arising from plasma protein binding on the rate of renal clearance. Classification And Regression Trees (CART) has been used to generate a classification tree leading to a simple set of Renal Clearance Rules (RCR) that can be applied to man. The rules are influenced by lipophilicity and ion class and can correctly predict 60% of an independent test set. These percentages increase to 71% and 79% for drugs with renal clearances of <0.1 ml/min/kg and >1 ml/min/kg, respectively. As far as the authors are aware these are the first set of models to appear in the literature that predict the rate of human renal clearance and can be used to manipulate molecular properties leading to new drugs that are less likely to fail due to renal clearance.

1. Introduction

The huge cost of pharmaceutical drug development (the current cost of discovering a new therapy is thought to approach US$ 1.3–1.6 billion [1]) and the high attrition of compounds entering clinical development are rightly focusing attention upon every aspect of the efficiency of our industry. While reasons for attrition are varied, including portfolio decisions and lack of clinical efficacy of the biological mechanism, many reasons for compound failure are entirely controlled by the chemical structure. Therefore, there is still much that can be done in the discovery phase, to improve the chances of success of a candidate drug later in development, by the judicious choice of chemical target. Unfavorable absorption, distribution, metabolism, excretion (ADME) properties of new drug candidates cause many of these failures and this has generated intense efforts to identify potential ADME liabilities early in the drug discovery process [2].

The overall clearance rate of a drug from blood is influenced by the rate of metabolism (hepatic or extra-hepatic) and the excretion rate of unchanged drug in the bile or the urine. The half-life of a drug is inversely proportional to blood clearance rate and proportional to the volume of distribution. Therefore, estimating the rate of metabolism and excretion of a drug in man is crucial for an accurate prediction of the pharmacokinetics in man. This is particularly the case for drugs with a low volume of distribution, as they require an especially low overall clearance in order to have an acceptable dosing regimen in man.

A quarter of the top 200 drugs in the U.S. (in 2002) are cleared via excretion of drug into the urine via the kidney, also known as renal clearance [3]. Therefore, the probability of encountering renal clearance in man and the impact it would have on the half-life of a drug requires investigation at an early stage in the discovery process. Renal clearance of xenobiotics involves several major processes, i.e. passive glomerular filtration, active tubular secretion, passive and active re-absorption. Various animal scaling approaches have been used for the prediction of renal clearance in man. The most basic animal scaling technique consists in using pharmacokinetics parameters obtained in preclinical species, i.e. rat or dog, as a prediction for human [4]. Furthermore, simple allometric scaling uses the combination of values measured in several preclinical species to predict man [5]. These methods are labor

* Corresponding authors. Tel.: +44 01509 64 4279; fax: +44 01509 64 5576.
E-mail addresses: Stuart.Paine@astrazeneca.com (S.W. Paine), Kamaldeep.Chohan@astrazeneca.com (K.K. Chohan).
[1] Current address: Department of Discovery DMPK, AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, United Kingdom.

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.10.003

intensive, require the synthesis of significant amounts of test compound and require the use of animals.

Quantitative structure–activity relationships (QSAR) attempt to find relationships between the molecular properties of molecules and the biological responses they elicit when applied to a biological system. The advancements in computer hardware and software now allow the molecular properties of molecules to be easily estimated without the need to synthesize the molecules in question. Thus, the use of predictive computational (in silico) QSAR models allows the biological properties of virtual structures to be predicted, and a more informed choice of target to be selected for synthesis.

The complexity of renal clearance has hindered multivariate computational efforts to date; Doddareddy et al. generated an in silico renal clearance model based upon 130 diverse drugs that predicts the percentage of total clearance that is expected to occur via excretion from the kidneys [6]. However, the model has limited use because it only predicts the likelihood (expressed as a percentage) of the renal clearance but not the absolute rate of renal clearance, which is key to estimating the human pharmacokinetic profile. For instance, a compound may be predicted to have 100% of its total clearance to occur via excretion from the kidneys, however, this tells us nothing about whether this will occur rapidly or slowly. More recently Varma et al. have analysed the physicochemical properties and the human renal clearance for a data set of 391 drugs [7]. Their analysis led to a qualitative set of relationships between renal clearance and physicochemical space, however, no quantitative models were developed that could be used to predict the rate of human renal clearance. The authors are not aware of any publication that proposes a quantitative predictive in silico human renal clearance model.

The aim of this study was to generate quantitative in silico models that can predict the rate of human renal clearance based on chemical structure alone. The models have been developed using a human renal clearance data set of 349 drugs (obtained from various literature sources) that show both active secretion and net re-absorption. Three statistical methods have been used in our analysis of the human renal clearance data—Partial Least Squares (PLS) [8–10], Random Forests (RF) [11,12] and Classification And Regression Trees (CART) [13,14]. Although these techniques employ different basic assumptions in their modeling we would nevertheless expect to obtain complimentary results in prediction. One of the advantages of regression trees is that they are able to model non-linearity in any dataset. On the other hand PLS is less abstract than regression trees, making interpretation more straightforward. In silico models are more than a literature curiosity, and successful ones offer the potential to be used as a virtual screen to filter design targets before synthesis in the drug discovery process.

2. Experimental

2.1. Data sets

A database of 349 compounds together with human renal clearance data was compiled from the literature. The database represents values for the human renal clearance of the parent drug only and does include contributions from biotransformations. The database is comprised of compounds that are actively secreted from the kidney and compounds that undergo net re-absorption. The database includes all charge types [that is, acids (59), bases (124), zwitterions (112) and neutral compounds (54)] and can be found in Supplementary data.

The molecular structure of the 349 compounds (global dataset) in the database was used to calculate 195 descriptors that broadly describe lipophilicity, hydrogen-bonding, size/shape, charge/polarity, topology and drug-ability of molecules. This descriptor calculation process was performed using an in-house descriptor generator engine. These descriptors have been described in detail elsewhere [15–17] and their definitions can be found in Supplementary data. The global dataset which consisted of all charge types was modeled using both parametric (PLS) and non-parametric (RF) methods. Subsequently, the models were validated using two approaches, namely: (a) the hold-out method where the data is randomly split (80:20) into a training and test set and (b) 20-fold cross validation (CV) using all data. Principal component analysis (PCA) within SIMCA-P (version 12.0) applied to the whole dataset showed that there is good overlap in property space between the training and test sets obtained from the hold-out methodology when considering the first six components. These six components describe 74% of the X descriptors. Local charge type models were also built for (i) bases, (ii) neutrals, (iii) acid and zwitterions, (iv) bases and zwitterions and (v) acid, bases and zwitterions. Models for acids or zwitterions alone were not constructed due to the small representation of these charge types in the global dataset.

2.2. QSAR modeling

The log-transformed human renal clearance data was modelled using 2 statistical techniques – PLS and RF. For each modeling method the same descriptor set of 195 descriptors was used.

2.2.1. Partial Least Squares

SIMCA-P (version 12.0) was used to perform PLS modeling on the log-transformed human renal clearance data. Initially, an autofit model was built with the descriptors. Then based on the variable importance plot (VIP), descriptors with a variable importance of less than 1 were removed from the PLS analysis and the model was rebuilt with the remaining descriptors. To test model robustness, Y permutation tests were undertaken in SIMCA-P to see if scrambling the Y data 500 times could lead to a better fitted model. PLS models were built for the global set of compounds (all charge types) and for each of the charge types as described in Section 2.1. For each of the hold-out method models an orthogonal best-line fit r^2_{obf} and root mean square error (RMSE) were computed for the training and test sets. With the 20-fold CV methodology r^2_{obf} was computed for the model fit whereas q^2 was generated for the CV groups and RMSE was calculated for both.

2.2.2. Random Forests

R (version 2.9.0) was used to build RF models using continuous log-transformed human renal clearance data. RF generates multiple un-pruned trees using bootstrap re-sampling of the training set. An ensemble of 100 trees was built and at each node a random selection of a 1/3 of the original set of descriptors was considered (default settings). A single prediction for a compound is achieved by taking the average of the predictions from the individual un-pruned trees. RF models were built for all datasets described previously using both the hold-out and the CV methodologies. As with the PLS models robustness was further evaluated by performing 20 Y permutations.

2.2.3. Classification trees

The software CART (version 5.0) was used to perform tree analysis. The CART methodology employs binary recursive partitioning. A classification tree was built for the global training set, using the hold-out method, where the human renal clearance data was classified as: low (<0.1 ml/min/kg), medium (0.1–1 ml/min/kg) and high (>1 ml/min/kg). The optimal tree (based on relative cost) was selected as a way of generating simple rules for renal clearance

potential of a drug. Trees that were larger than the optimal tree were also analysed and compared with the optimal tree. The larger tree that was selected for analysis had a balance between relative cost and good purity in the terminal nodes. The optimal tree (and the larger tree) was used to predict the global test set and a confusion matrix was generated.

2.3. Evaluation of model predictivity

A number of statistical measures may be used to evaluate the continuous models. We have used the square of the Pearson correlation coefficient that minimizes the distance of the residuals orthogonal to the line of best-fit (r^2_{obf}) and the root-mean square error (RMSE) as a measure of training set model performance. However, the r^2_{obf} obtained for the training set may not be reflective of the model performance on a new dataset. A more realistic measure of model performance is based on how a model performs on compounds that it has not been trained on—the so-called test set compounds. The two approaches that have been applied here are (a) the hold-out method where the data is randomly split (80:20) into a training and test set and (b) 20-fold cross validation (CV). In 20-fold cross-validation, the dataset is randomly partitioned into 20 subsamples. Of the 20 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 19 subsamples are used as training data. The cross-validation process is then repeated 20 times with each of the 20 subsamples used exactly once as the validation data. For test sets we have used the r^2_{obf} [Eq. (1)] and the RMSE [hold-out method; Eq. (2)] or q^2 [CV method; Eq. (3)] as a way of assessing the performance of each model. For Eqs. (1) and (2), x and y are predicted and observed values of the biological response, n is the number of compounds and $\bar{x}$ and $\bar{y}$ are the mean of the predictions and observations, respectively

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

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

$$q^2 = 1 - \frac{\sum(y_{out}-x_{in})^2}{\sum(y_{out}-\bar{y}_{in})^2} \quad (3)$$

where in and out represent compounds in training or left out group, respectively.

For the predictions from the classification tree, the classification performance is normally summarized in a confusion or error matrix that cross tabulates the observed and predicted patterns. A 3-class confusion matrix was generated. For a review of methods for the assessment of prediction errors in classification see Fielding and Bell [18]. A variety of error or accuracy measures can be calculated from a confusion matrix. There are several measures that describe the accuracy of a single class (per class accuracy). For example, in this case the false positive rate is the proportion of low and medium cases that were incorrectly classified as high. Whereas the false negative rate is the proportion of medium and high cases that were incorrectly classified as low. Sensitivity and specificity are usually defined in a 2 class system as the proportion of positive and negative cases that were correctly predicted, respectively. In this 3 class system the terms sensitivity high, medium and low have been used. The positive and negative predictive power are the proportion of high and low predictions that were observed high and low, respectively. A naïve measure of overall accuracy (all classes) is the correct classification rate. The problem with this measure is that it does not account for the fact that some of the apparent classification accuracy could be due to chance. The kappa index of overall agreement for classification was developed by Cohen [19,20] and associates [21] in the context of psychology and psychiatric diagnosis. The kappa index [Eq. (4)] is considered to be superior to using correct classification rate as it assesses the model's improvement in prediction over chance.

$$\mathrm{kappa} = \frac{\text{observed agreement} - \text{chance agreement}}{\text{total observed} - \text{chance agreement}} \quad (4)$$

3. Results and discussion

In order to establish a QSAR model for the log-transformed human renal clearance (ml/min/kg) data, parametric and non-parametric methods were explored. Models have been built for the global set and also for charge-specific sets (local sets). In this context the authors define global as a data set comprising compounds of a wide structural diversity and a good renal clearance range; correspondingly, the models should be able to predict across structural classes. In general, we find that the non-parametric RF method is performing better than the parametric PLS technique. In the following sections we discuss how each model performs by assessing the predictivity of each model as described in Section 2.3. We also discuss simple rules extracted from a CART classification tree. Our intention here is to develop simple guidelines for the human renal clearance of a drug from meaningful and accessible descriptors. Finally, there is a section which shows how significant our models are with respect to the correlation between human renal clearance and the extent of plasma protein binding.

3.1. Partial Least Squares analysis

Model statistics resulting from both the hold-out and the CV methodologies are presented in Tables 1 and 2, respectively. For the hold-out method, global and local PLS models have r^2_{obf} between 0.20 and 0.41 for the training sets. The RMSE of the training sets varied from 0.70 to 0.86 and test set statistics were largely comparable to those of the training sets with r^2_{obf} from 0.28 to 0.60 and the RMSEs from 0.58 to 0.96. The RMSEs for all test sets were better than the variance of the measured data. However, as the RMSEs correspond to approximately 3.8 to 9.1-fold error it is unlikely that

Table 1
PLS model statistics for training and test set.

Model	Training set				Test set			
	n	r^2_{obf}	RMSE	SD	n	r^2_{obf}	RMSE	SD
Global	281	0.37	0.73	0.93	68	0.44	0.72[a]	0.97
Acid and zwitterions	100	0.41	0.71	0.93	13	0.36	0.96	1.18
Bases	105	0.22	0.70	0.80	19	0.38	0.64	0.83
Bases and zwitterions	159	0.20	0.70	0.78	19	0.38	0.64	0.83
Neutrals	85	0.25	0.86	0.99	26	0.60	0.58	0.91
Acid, bases and zwitterions	195	0.28	0.74	0.87	42	0.28	0.88	1.02

[a] The RMSE for the all charge model test set is 0.72. Within this the RMSE for charge types is acids and zwitterions (0.84), bases (0.60), neutrals (0.69), bases and zwitterions (0.61) and acids, bases and zwitterions (0.74).

Table 2
PLS model statistics for training and 20-fold cross-validation.

Model	Training set			20-Fold cross-validation		
	r^2_{obf}	RMSE	Standard deviation of measured data	q^2	RMSE	Standard deviation of measured data
Global	0.39	0.73	0.93	0.28	0.79[a]	0.93
Acid and zwitterions	0.42	0.75	0.99	0.30	0.83	0.99
Bases	0.25	0.69	0.80	0.21	0.71	0.80
Bases and zwitterions	0.23	0.69	0.78	0.20	0.70	0.78
Neutrals	0.30	0.81	0.97	0.29	0.82	0.97
Acid, bases and zwitterions	0.30	0.75	0.90	0.19	0.81	0.90

[a] The RMSE for the global 20-fold cross-validation is 0.79. Within this the RMSE for charge types is acid and zwitterions (0.80), bases (0.71), neutrals (0.86) and bases and zwitterions (0.71).

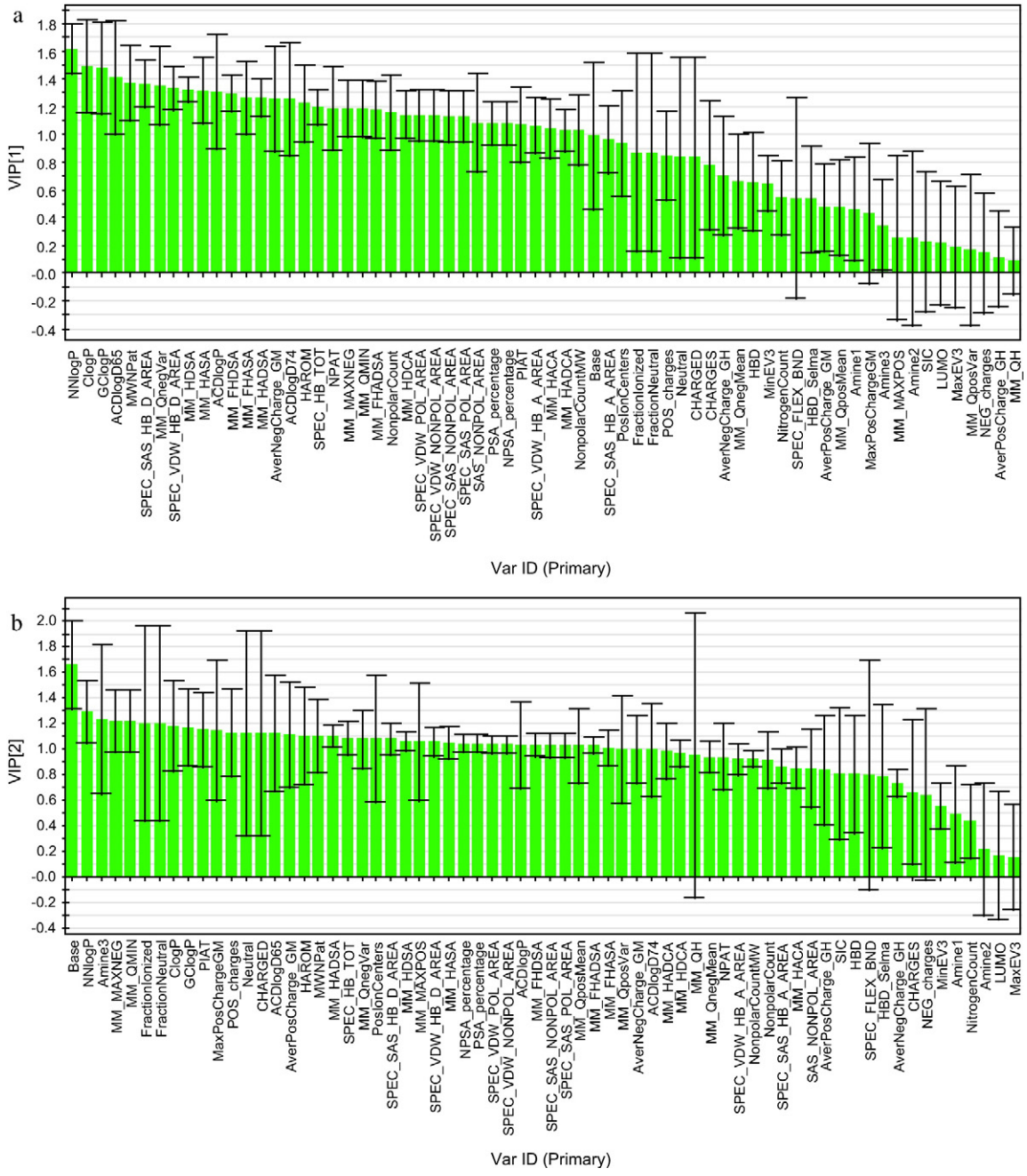

Fig. 1. The variable importance plot for the following hold-out method models: (a) PLS global (component 1), (b) PLS global (component 2), (c) Random Forests global and (d) Random Forests acid and zwitterions.

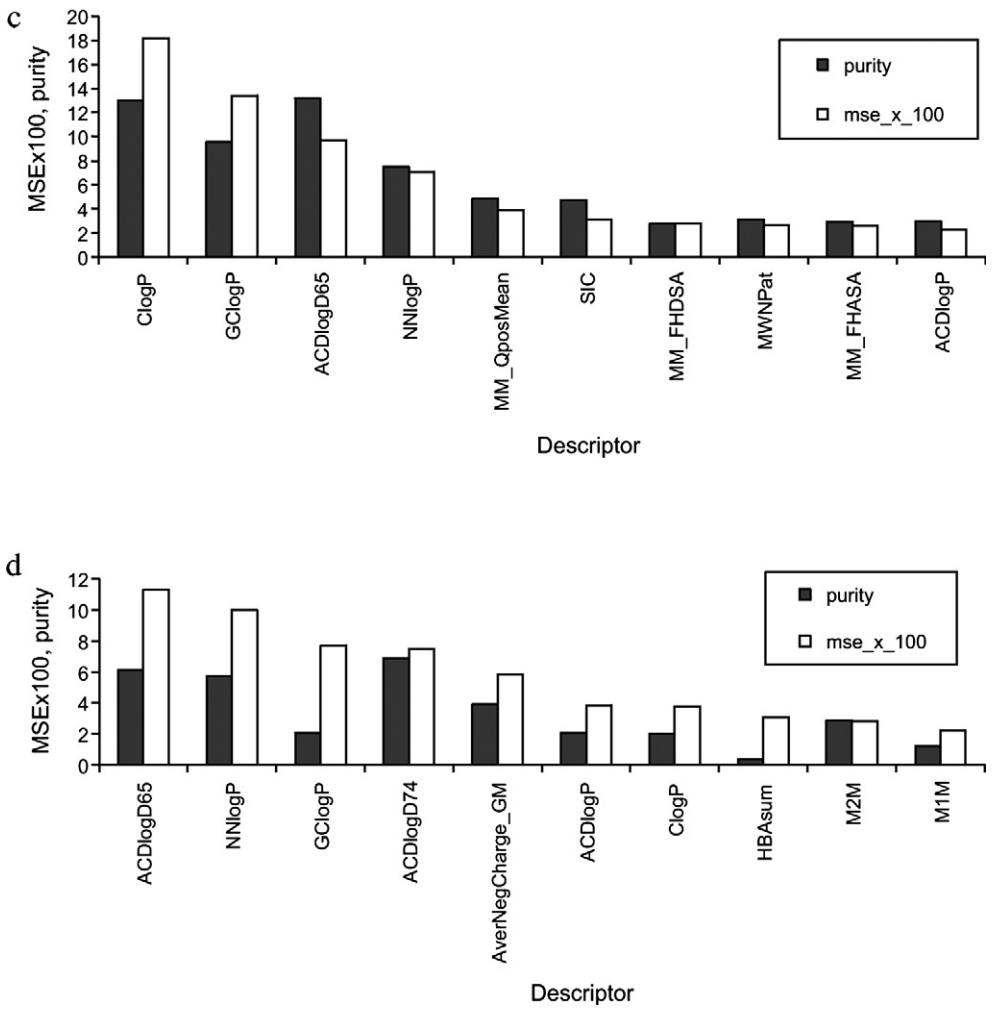

Fig. 1. (*Continued*).

all of these PLS models will yield robust predictions of human renal clearance. All models were significant compared to the 500 Y permutation tests.

Generally, no significant improvements over the global PLS model (test set RMSE = 0.72) were achieved through modeling individual charge types. Attempts at modeling local acid and zwitterions (RMSE = 0.96), bases (RMSE = 0.64), bases and zwitterions (RMSE = 0.64), neutrals (RMSE = 0.58), and acids, bases and neutrals (RMSE = 0.88) showed no significant improvement over the global model when contextualized against the respective training set statistics. Moreover, in test set the local models do no better than the global model's ability to predict the chemotypes with the exception of neutral compounds (Table 1, footnote a).

The descriptors with the largest influence upon each hold-out model were judged by looking at the Variable Importance Plots (VIP) and coefficients plots for individual model components. Component 1 (Fig. 1a) appears to be driven by lipophilicity with lipophilicity descriptors (NNLogP, CLogP, GCLogP, ACDLogD6.5) being most prevalent in the VIP plot. Component 2 (Fig. 1b) had maximum contribution from charge or polarisability descriptors. After interrogation of all the individual charge type models it is apparent that the primary component in each model has the same trend in its coefficient plot. Lipohilicity descriptors have a negative net effect on human renal clearance, i.e. as lipophilicity increases human renal clearance decreases. Charge or polarisability descriptors have a positive effect on human renal clearance. This is what may be expected as free fraction in blood (generally relating to lipohilicity) will influence the amount of drug undergoing glomerular filtration. Furthermore, there is greater potential for passive re-absorption of more lipophilic compounds.

The CV methodology was used in addition to the hold-out method to robustly validate the renal clearance models. For the CV method, global and local PLS models have r^2_{obf} between 0.23 and 0.42 for the training sets. The RMSE of the training sets varied from 0.69 to 0.81 and are very similar to the hold-out methodology which is not surprising as the latter method uses 80% of the CV approach. CV statistics were largely comparable to those of the training sets with q^2 ranging from 0.19 to 0.30 and the RMSEs from 0.70 to 0.83. The RMSEs for CV were better than the variance of the measured data. All models were significant compared to the 500 Y permutation tests. Generally, no significant improvements over the global PLS model (CV RMSE = 0.79) were achieved through modeling individual charge types (Table 2, footnote a).

3.2. Random Forests

Random Forest methodology was used to model the log-transformed human renal clearance in a continuous manner. Model statistics resulting from both the hold-out and the CV methodologies are presented in Tables 3 and 4. For the hold-out method all training set models have high r^2_{obf} (>0.9) and low RMSEs (<0.35) which is relatively common for this methodology. The global model for the hold-out method showed good predictability (RMSE 0.63) when using the independent test set. A summary of all charge

Table 3
Random Forest model statistics for training and test set.

Model	Training set			Test set		
	r^2_{obf}	RMSE	Standard deviation of measured data	r^2_{obf}	RMSE	Standard deviation of measured data
Global	0.93	0.32	0.93	0.63	0.63[a]	0.97
Acid and zwitterions	0.93	0.32	0.95	0.79	0.51	1.06
Bases	0.92	0.30	0.80	0.6	0.54	0.83
Bases and zwitterions	0.92	0.29	0.78	0.61	0.49	0.81
Neutrals	0.93	0.34	0.99	0.43	0.70	0.91
Acid, bases and zwitterions	0.93	0.31	0.87	0.63	0.70	1.02

[a] The RMSE for the global test set is 0.63. Within this the RMSE for charge types is acid and zwitterions (0.63), bases (0.57), neutrals (0.65) and bases and zwitterions (0.52).

Table 4
Random Forest model statistics for training and 20-fold cross-validation.

Model	Training set			20-Fold cross-validation		
	r^2_{obf}	RMSE	Standard deviation of measured data	q^2	RMSE	Standard deviation of measured data
Global	0.93	0.31	0.93	0.38	0.74[a]	0.93
Acid and zwitterions	0.93	0.30	0.99	0.48	0.72	0.99
Bases	0.92	0.29	0.80	0.27	0.68	0.80
Bases and zwitterions	0.92	0.27	0.78	0.34	0.64	0.78
Neutrals	0.93	0.35	0.97	0.24	0.85	0.97
Acid, bases and zwitterions	0.93	0.30	0.90	0.4	0.70	0.90

[a] The RMSE for the global 20-fold cross-validation is 0.74. Within this the RMSE for charge types is acid and zwitterions (0.69), bases (0.69), neutrals (0.83) and bases and zwitterions (0.66).

specific models can be found in Table 3, in which the acids and zwitterions model adds enrichment to the predictability when compared to the global model (Fig. 2). The neutral model with an RMSE of 0.70, however, remains best predicted by the global all charge type model. The test set RMSEs of all models are significantly better than the variance of the measured data. The sub analysis of the data set also allows us to evaluate the behavior of the zwitterions within the data set and in this case shows that they can be successfully modeled in combination with either acids or bases (RMSEs 0.51 and 0.49, respectively). This however, is not the case when all three charge types are combined (RMSE 0.70). All models were significant compared to the 20 Y permutation tests.

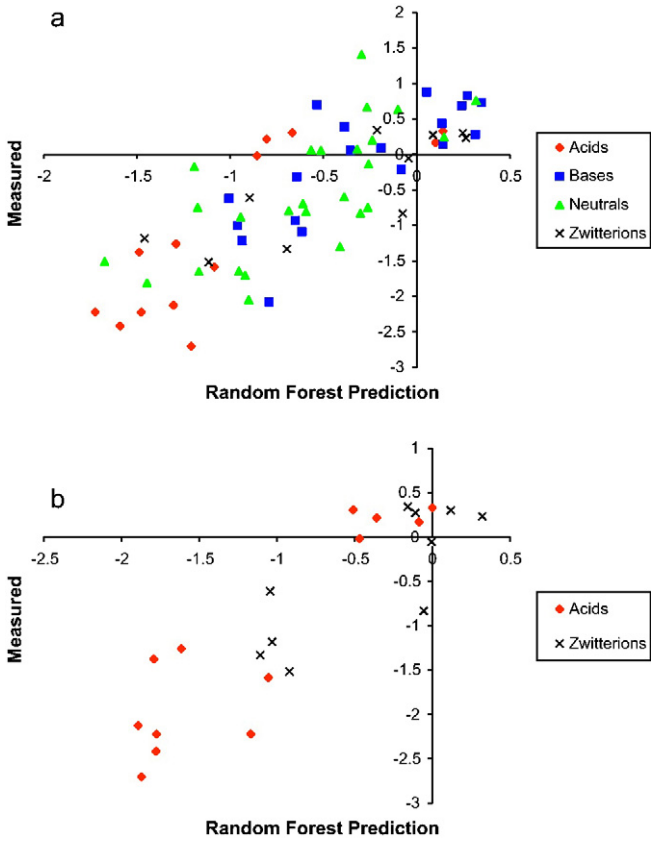

Fig. 2. Measured log-transformed human renal clearance (ml/min/kg) versus the RF prediction for the test set. The plots are for: (a) global (all charges), (b) acid and zwitterion. In these plots the colouring is red for acids, blue for bases, green for neutrals and black for zwitterions.

To fully understand the nature of Random Forests, which in this case built a committee of 100 trees, the underlying importance of the descriptors used in the models must be examined (Fig. 1c and d). Two measures of variable importance are used within RF. The purity measure of the descriptor variables is simply the reduction in impurity that a particular variable creates when it is split on. The second is the average drop in mean square error (MSE) of the predictions (expressed as MSE × 100 in Fig. 1c and d). These two measures do produce different lists but it is good practice to look at the structure of both. The all charge type global model's top descriptors can largely be split into two categories; lipophilicity based descriptors (e.g. ClogP, ACDlogD65) and those relating to the compounds ability to carry a positive charge (e.g. MM_QPOS, MM_FHDSA). Whilst the lipophilicity variables remain important in all models, AverNegCharge_GM and the moment of inertia around the first (M1M) and second plane (M2M) of the molecule have a significant impact on the acid and zwitterions models. In the local model for bases, MWNPat, which describes the proportion of molecular weight accounted for by the excess of non-polar atoms and SAS_NONPOL_AREA, which describes the solvent accessible non-polar surface area make a substantial contribution.

Again, the CV methodology was used in addition to the hold-out method to robustly validate the RF renal clearance models. For the CV method, global and local RF models have $r^2_{obf} > 0.9$ for the training sets. The RMSE of the training sets varied from 0.27 to 0.35 and are very similar to the hold-out methodology. CV statistics were significantly lower compared to the training sets with q^2 ranging from 0.24 to 0.48 and the RMSEs from 0.64 to 0.85 and in the majority of cases is better than the corresponding statistics obtained from PLS CV. The RMSEs for CV were better than the variance of the measured data. Also, all models were significant compared to the 20 Y

permutation tests. Generally, no significant improvements over the global RF model (CV RMSE = 0.74) were achieved through modeling individual charge types (Table 4, footnote a).

3.3. Comparison of methodologies

In this study we have compared the parametric (PLS) and non-parametric (RF) methodologies to build continuous human renal clearance models. For RF the training set statistics are far superior to those obtained from PLS. Also, the statistics generated for the independent test set, using the hold-out method, are superior for RF compared to PLS with the exception of the neutrals model. However, the statistics obtained from 20-fold CV for RF are weaker than those obtained from the hold-out method, but are generally better than those obtained using PLS CV.

There are advantages and disadvantages to using the hold-out and CV methodologies. The hold-out method allows an independent test set to be created which can be used to test model predictivity. This is similar to the situation faced in the drug discovery process where a model is used to predict new chemistry targets. The test set generated within in this work covers the property space of the training set and therefore represents oral drug property space. On one hand this test set may appear to be less challenging than a set of compounds that fall outside the property space of the training set, however, the purpose of these models is to predict the likelihood of renal clearance for oral drug-like compounds. A disadvantage of the hold-out methodology is that the generated test set from a single randomization may lead to predictability which could be artificially high or low. However, the test set generated herein has very good overlap with the property space of the training set and therefore can considered a fair test. The CV method does not suffer from the same issues associated with the hold-out method as all data is ultimately used in the CV test groups. Also, all of the data can be used to build the model, however, in this case the statistics obtained from the training sets are similar between CV and hold-out methodologies. CV affords lower bias than the hold-out method but on the downside may result in large variance [22]. In this study the CV methodology gives a slightly larger, but comparable RMSE to the hold-out method and importantly both give RMSEs lower than the variance in the measured data. The fact that all the models that have been built are better than models generated from the permutation of the Y data adds further confidence to the validation. The best models generated from either PLS or RF produce significant models that can predict acids/zwitterions, bases and neutrals with approximate average fold errors of 3, 3 and 4, respectively, for an independent test set that covers oral drug-like property space.

3.4. Classification trees analysis

Classification trees have been built within CART with the aim of generating simple decision rules for assessing renal clearance risk. By default, based on relative cost, a two-split tree was selected by CART to be the optimal tree. In this optimal tree, the first split is by ACDLogD6.5 (Fig. 3a). If the ACDLogD6.5 ≤ -0.38, the compounds end up in terminal node 1 (TN 1) and are predicted to have high renal clearance. If the ACDLogD6.5 > -0.38, the compounds are then split by ion class. This descriptor essentially splits the compounds according to whether they have a basic centre. Acids and neutral compounds are therefore treated differently to bases and zwitterions. If a compound is an acid or a neutral the optimal tree predicts these compounds to have low renal clearance (terminal node 2; TN2). However, if the compound is basic or a zwitterion these compounds are predicted to have medium renal clearance (terminal node 3; TN3). This decision tree affords simple Renal Clearance Rules (RCR) to flag the potential risk of renal clearance in man of a test drug (Fig. 3b). We have also investigated larger trees and compared these with the optimal tree from CART. One of the larger trees has 11 terminal nodes, two of which are pure nodes. After the ACDLogD6.5 and ion class split the larger tree progresses as follows: for acids and neutrals; if compounds have a polar surface area (PSA) of $\leq$74 Å^2 they end up in a terminal node and will be predicted to have low renal clearance. Acids and neutrals that have a PSA > 74 Å^2 are then split according to the non-polar surface area (NPSA) with those compounds having a NPSA $\leq$ 146 Å^2 being classified as high renal clearance. If the NPSA > 146 Å^2, molecular weight (MW) is the next splitter. Compounds that have a MW $\leq$ 263 are predicted to have low renal clearance. Any acidic or neutral compounds that are left then progress through a lipophilicity descriptor (ACDLogD7.4). If the ACDLogD7.4 $\leq$ 1.7 then the compound is predicted as medium, otherwise the compound is predicted as low. For bases and zwitterions; compounds with ACDLogD6.5 > 2.9 are predicted low by the model. The other compounds progress down the tree and are split according to their ACDLogP. If the ACDLogP is >3.3 then the compounds are predicted to have a medium renal clearance. The next split is ClogP, compounds with a ClogP > 3.1 are predicted as low renal clearance. The final split for compounds with a basic centre is NPSA and if this is $\leq$272 Å^2 then the compounds are predicted with high renal clearance, otherwise they are predicted as medium.

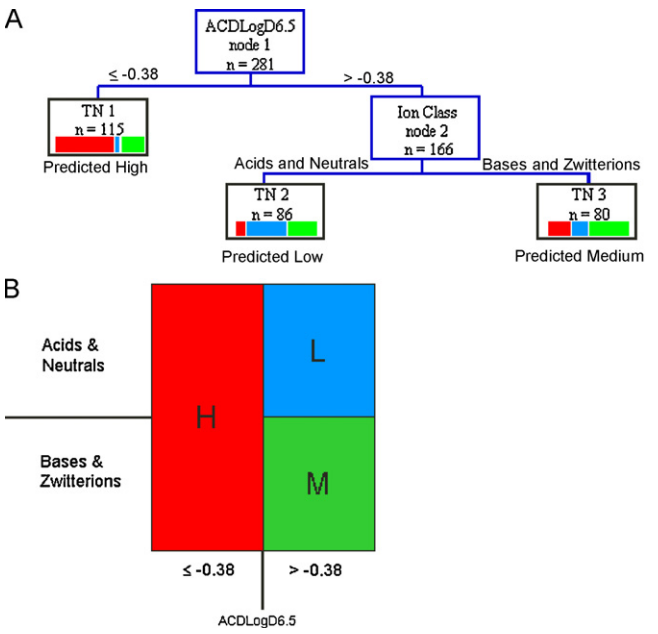

Fig. 3. (a) A single classification tree built on the global training set (281 compounds). In the terminal nodes the colouring blue, green and red refers to low, medium and high, respectively, and (b) a schematic to show the Renal Clearance Rules (RCR).

The training and test set confusion matrix for the optimal tree is shown in Table 5. This optimal tree is able to predict the training and test set with an overall correct classification of 56% and 60%, respectively (Table 6). The overall correct classification appears weak, however, the kappa index of agreement of 0.34 and 0.39 for the training and test set, respectively, implies fair agreement between measured and predicted human renal clearance class. The test set statistics show that it correctly classifies 79% of the high renal clearance class compounds and 71% of the low class compounds, however, the medium class is less well classified (21%) and this leads to the dilution of the overall correct classification. Furthermore, the predictive power of the high renal clearance class (76%) performs better then the low class (56%), and both perform better than the medium class (33%; Table 6). When we compare the opti-

Table 5
A 3-class confusion matrix from the optimal tree: (a) global training set and (b) global test set.

Predicted class	Measured class		
	High	Medium	Low
A			
High	76	30	9
Medium	23	40	17
Low	12	32	42
B			
High	22	6	1
Medium	3	4	5
Low	3	9	15

Table 6
The classification statistics for the global training and test set. All data has been generated based on the confusion matrix and is from the optimal tree.

Measure	Model prediction for training set	Model prediction for test set
Correct classification rate	56%	60%
Sensitivity (high)	68%	79%
Sensitivity (medium)	39%	21%
Sensitivity (low)	62%	71%
False positive rate	23%	18%
False medium rate	22%	16%
False negative rate	21%	26%
Positive predictive power	66%	76%
Medium predictive power	50%	33%
Negative predictive power	49%	56%
Kappa	0.34	0.39

mal tree statistics (RCR) with the larger tree containing 11 terminal nodes as previously described, we find that the training statistics are improved going from the optimal tree to the larger tree. For example, in training the overall correct classification is 56% for optimal tree and 66% for the larger tree, and kappa is 0.34 for optimal tree and 0.49 for larger tree. However, the test set statistics are very similar between the optimal and larger tree. For example, in test set the overall correct classification is 60% for both the optimal and larger tree, and the kappa is 0.39 in both cases. This suggests that larger tree models offer no real advantage and have greater complexity compared to the RCR derived from the optimal tree.

3.5. Effect of plasma protein binding on human renal clearance

The human renal clearance of a drug may be expected to be influenced by the extent of plasma protein binding. Only free drug can be passively filtered through the glomerulus of the kidney and there is much evidence showing the influence of plasma protein binding on drug transport involved in active secretion. Therefore, there is a risk that the models that have been built are essentially describing the plasma protein binding of the drug. The correlation between human renal clearance and log(K), where K is the ratio of bound fraction to free fraction of drug in plasma, is shown in Fig. 4. Although the correlation is statistically significant, the r^2_{obf} value of 0.34 is either comparable or inferior to the statistics obtained from the best models generated using either the hold-out or the CV methodologies. This suggests that the QSAR models contain additional information on top of any influence arising from plasma protein binding.

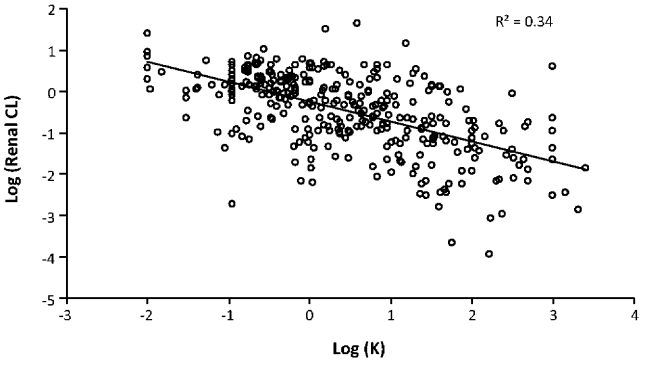

Fig. 4. Correlation between human renal clearance and plasma protein binding (log K).

4. Conclusion

Our investigation of developing robust QSAR models for human renal clearance has resulted in compiling a relatively large database. We have utilized this data to develop in silico models, and as far as we are aware these are the first set of models to appear in the literature that predict the rate of human renal clearance. These models contain additional information on top of any influence arising from plasma protein binding on the rate of renal clearance.

The statistics generated for the independent test set, using the hold-out method, are superior for RF compared to PLS with the exception of the neutrals model. However, the statistics obtained from 20-fold CV for RF are weaker than those obtained from the hold-out method, but are generally better than those obtained from applying CV to PLS. One may expect the RF models to be superior to the linear PLS models as renal clearance is limited by blood flow to the kidneys and therefore may not be expected to follow a linear relationship. All models show general consensus on the important properties governing renal clearance in man where lipophilicity is a key player. A classification analysis afforded a simple set of rules for human renal clearance potential which could be used in the first instance. These Renal Clearance Rules (RCR) use only two simple descriptors namely ACDlogD6.5 and ion class. The RCR give a ball-park approximation of the human renal clearance and can be used in conjunction with the important descriptors from the QSAR regression models to manipulate molecular properties leading to compounds that are less likely to fail due to renal clearance.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2010.10.003.

References

[1] J. Hodgson, ADMET – turning chemicals into drugs, Nat. Biotechnol. 19 (2001) 722–726.
[2] S.A. Roberts, High-throughput screening approaches for investigating drug metabolism and pharmacokinetics, Xenobiotica 31 (2001) 557–589.
[3] J.A. Williams, R. Hyland, B.C. Jones, D.A. Smith, S. Hurst, T.C. Goosen, V. Peterkin, J.R. Koup, S.E. Ball, Drug–drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios, Drug Metab. Dispos. 32 (2004) 1201–1208.
[4] H. Boxenbaum, Comparative pharmacokinetics of benzodiazepines in dog and man, J. Pharmacokinet. Biopharm. 10 (1982) 411–426.
[5] H. Boxenbaum, Interspecies variation in liver weight, hepatic blood flow, and antipyrine intrinsic clearance: extrapolation of data to benzodiazepines and phenytoin, J. Pharmacokinet. Biopharm. 8 (1980) 165–176.
[6] M.R. Doddareddy, Y.S. Cho, H.Y. Koh, D.H. Kim, A.N. Pae, In silico renal clearance model using classical Volsurf approach, J. Chem. Inf. Model. 46 (2006) 1312–1320.
[7] M.V.S. Varma, B. Feng, R.S. Obach, M.D. Troutman, J. Chupka, H.R. Miller, A. El-Kattan, Physicochemical determinants of human renal clearance, J. Med. Chem. 52 (2009) 4844–4852.
[8] A. Höskuldsson, Prediction Methods in Science and Technology, Thor Publishing, Copenhagen, Denmark, 1996.
[9] S. Wold, C. Albano, W.J. Dunn, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, M. Sjöström, Multivariate data analysis in chemistry, in: B.R. Kowalski (Ed.), Chemometrics: Mathematics and

Statistics in Chemistry, D. Reidel Publishing Company, Dordrecht, Holland, 1984.
[10] S. Wold, L. Eriksson, M. Sjöström, PLS in chemistry, in: Encyclopedia of Computational Chemistry, Wiley, 2000.
[11] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[12] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, B.P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comp. Sci. 43 (2003) 1947–1958.
[13] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Pacific Grove, Wadsworth, 1984.
[14] D. Steinberg, P. Colla, CART: Tree-Structured Non-Parametric Data Analysis, Salford Systems, San Diego, CA, 1995.
[15] P. Bruneau, Search for predictive generic model of aqueous solubility using Bayesian neural networks, J. Chem. Inf. Comp. Sci. 41 (2001) 1605–1616.
[16] A. Katritzky, Y. Wang, S. Sild, T. Tamm, M. Kalrelson, QSPR studies on vapor pressure, aqueous solubility, and the prediction of water–air partition coefficient, J. Chem. Inf. Comp. Sci. 38 (1998) 720–725.
[17] Selma is an AstraZeneca in-house software package. For further information contact T. Olsson, V. Sherbukhin, Synthesis and Structure Administration (SaSA), AstraZeneca R&D Mölndal.
[18] A.H. Fielding, J.F. Bell, A review of methods for the assessment of prediction errors in conservation presence/absence models, Environ. Conserv. 24 (1997) 38–49.
[19] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Measur. 20 (1960) 37–46.
[20] J. Cohen, Weighted, Kappa: nominal scale agreement with provision for scaled disagreement or partial credit, Psychol. Bull. 70 (1968) 426–443.
[21] J.L. Fleiss, J. Cohen, B.S. Everitt, Large sample standard errors of kappa and weighted kappa, Psychol. Bull. 72 (1969) 323–327.
[22] L. Breiman, Heuristics of instability and stabilization in model selection, Ann. Stat. 24 (1996) 2350–2383.