Predictivity of Simulated ADME AutoQSAR Models Over Time

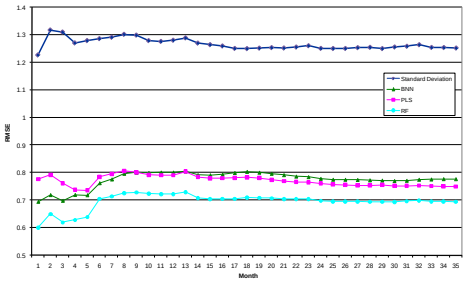
Sarah L. Rodgers[*a], *Andrew M. Davis*[b], *Nick P. Tomkinson*[b] *and Han van de Waterbeemd*[c]

[*] Correspondence author, email address: srodgers@accelrys.com phone: +44(0)1509644560; fax: +44(0)1509644576
[a] Accelrys Ltd, 334 Cambridge Science Park, Cambridge CB4 0WN, UK
[b]AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, UK
[c]14 Rue de la Rasclose, R-66690 Saint Andre, France

An Analysis of the Predictivity of ADME QSAR Models Over Time

Sarah L. Rodgers[*a], *Andrew M. Davis*[b], *Nick P. Tomkinson*[b] *and Han van de Waterbeemd*[c]

Three different global QSAR models, human plasma protein binding, aqueous solubility and log D7.4, were updated on a monthly basis over a period of three years. The effect of updating the models on their predictivity was studied using a series of monthly temporal test sets in addition to a final terminal temporal test set. Partial least squares (PLS), Random Forests (RF) and Bayesian Neural Networks (BNN) models are examined, covering three distinctly different approaches to QSAR modelling. It is demonstrated that the models are able to predict forward in time, but that updating models on a regular basis increases their ability to make predictions for current compounds. The degree of the improvement depends on the property studied and the model building technique used. These results demonstrate the importance of updating models on a regular basis. For both static models predicting forward in time, and regularly updating models it is shown that RF models are the most predictive.



Keywords: Quantitative structure activity relationships (QSAR), autoQSAR, molecular modeling, computational chemistry

Abbreviations: Quantitative structure-activity relationships (QSAR), Partial least squares (PLS), Random Forests (RF), Bayesian Neural Networks (BNN), absorption, distribution, metabolism, elimination and toxicology (ADMET), root mean square error in prediction (RMSE), Mahalanobis Distance (MD)

[*] Correspondence author, email address: srodgers@accelrys.com phone: +44(0)1509644560; fax: +44(0)1509644576
[a] Accelrys Ltd, 334 Cambridge Science Park, Cambridge CB4 0WN, UK
[b]AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, UK
[c]14 Rue de la Rasclose, R-66690 Saint Andre, France

1. Introduction

The current 92% failure rate of candidate drugs in clinical development, risks the long term future of the drug discovery business model [1]. The design of a covalent arrangement of 26 or so atoms that can safely interact with a single biological molecular target to yield a medically beneficial effect, within the 10-100 trillion differentiated cells that comprise the human body, must rank as one of mankind's greatest technical challenges, but we are beginning to learn some simple guidelines that may increase our chances of success. For instance, simple physicochemical properties, such as lipophilicity, charge type and hydrogen bond distribution are strongly associated with causes of attrition, such as solubility, poor permeability and absorption, inappropriate drug distribution, extent of protein binding, high or unwanted metabolism and chances of off-target pharmacology. Effective selection of compounds with favourable absorption, distribution, metabolism, elimination and toxicology (ADMET) properties reduces late stage attrition [2], therefore the measurement of these properties and prediction using empirical quantitative structure-activity (QSAR) models is recognised as crucial in the drug discovery process.

In the quest to improve ADMET QSAR models, focus has been on the statistical techniques applied, the suitability of the physicochemical descriptors employed and the quality of the training set data. Statistical and machine learning methods such as partial least-squares [3], recursive partitioning [4], neural networks [5] and support vector machines [6] have all been applied to the QSAR problem. Descriptor sets such as atom pairs [7], connectivity indices [8] and BCUT descriptors [9] have also been tested. In addition, the quality of the experimental data is very important [10]. Experimental measurements used to train a model need to be consistent and reproducible – ideally originating from the same assay. Models based on literature data tend to be less successful for making predictions on in-house compounds than where in-house training data itself is used [11]. The importance of using the most up-to-date data, or updating models to reflect more recent experimental measurements, has received less attention. This could be due to the difficulties encountered in obtaining accurate and reliable data, particularly in the public domain.

Effective QSAR models are reliant on the data that is used to train them [12], where there is little similarity between the training set compounds and the query compound confidence in predictions will be low. The concept of a distance to model [13] can be used to define the domain of applicability [14] of a model, the region of chemical space represented by the training set where predictions can be made without an extrapolation. As query compounds move away from the property space representing the training compounds, the errors in their predictions are expected to increase. In an attempt to generate QSAR models having the widest domain of applicability, datasets often contain as much data as can be collected, covering as many different chemical series as can be found, in the hope they will show better predictivity with forthcoming chemistries. These so-called generic "global" ADMET models need regular updating to ensure the model space is applicable to new areas of developing chemistry [15]. Expanding the range of chemistry, or the range of the property being modelled, tends to improve model quality [16]. In drug discovery companies, new experimental data is generated every day, yet QSAR models are rarely regularly updated to reflect this new data. Thus the predictions made are based on compounds, and therefore chemical space, that can be months or years out-of-date and likely to be quite different to the current compounds for which we wish to make predictions.

Interest is growing in the automation of QSAR model building [17,18]. Such automated systems will facilitate the regular updating of existing models so that they more closely reflect current chemistry. An alternative is to receive a prediction from a dynamic model generated 'on the fly' as a compound query is made [19,20], these methods are known as *local lazy learning*. A local model is generated using only those compounds similar enough to the query compound, hence this approach is limited by the number of similar neighbours a given query will have. Since the process of collecting experimental

measurements from assays has become more high-throughput and automated, greater automation of the data analysis stage is the next logical step.

In order to determine the benefits of automatically updating models on a regular basis, a series of simulations have been conducted to assess model performance over time. A previous study has examined the effects of updating a human plasma protein binding PLS model over a period of 21 months [21]. Here we extend the previous analysis by modelling three properties using three different model-building techniques over a longer time period.

2. Experimental Methods

The properties selected to model in this QSAR analysis are human plasma protein binding, aqueous solubility measured at pH7.4 and n-octanol-water distribution measured at pH7.4 (logD7.4). These are three commonly measured and predicted physical properties within the pharmaceutical industry.

The degree of plasma protein binding is important in drug discovery, as only free drug is able interact with both on and off-target proteins [22]. Drugs will reversibly bind with plasma proteins, principally serum albumin. While some compounds may have higher affinities for other proteins in plasma, the high concentration of albumin in plasma makes this the dominant protein modulating the free concentration. The amount of available free drug modulates whole blood potency, extent of distribution, metabolism and elimination among other factors. The extent of protein binding needs to be balanced with the intrinsic potency and metabolic stability of the compound to provide a suitable pharmacokinetic profile at an acceptable dose and dose frequency to drive drug efficacy.

Low aqueous solubility can limit the extent of absorption, and can directly lead to toxicity if the compound precipitates in the kidneys. Poorly soluble compounds may lead to variable absorption which can be a risk factor for drugs with low therapeutic margin. This study is concerned with modelling the aqueous solubility of compounds at pH 7.4 (therefore not intrinsic solubility).

The logarithm of the distribution coefficient in octanol/water at pH 7.4 (log D7.4) is used to describe the lipophilicity of compounds. The property log D7.4 takes into account the ionisation state of the molecule, important because many drugs will be ionised at physiological pH. In order to cross membranes by passive diffusion (the most common route), drugs need to display sufficient lipophilicity to be able to penetrate the membrane. Excessive lipophilicity can lead to low solubility, high protein binding [23], and drug disposition in tissues and cells, increasing the risk of an unwanted toxicological consequence. High lipophilicity also correlates with high metabolism, and an increasing likelihood of unwanted off-target pharmacology. Lipophilicity is important in dictating drug-receptor interactions and the pharmacokinetic profile of a drug in addition to toxicological properties.

The three statistical techniques employed for this analysis are partial least-squares (PLS), random forest (RF) and Bayesian neural networks (BNN). These cover three distinctly different approaches to QSAR modelling. PLS is a linear statistical method that defines the y variable in terms of linear combinations of the variability in the x-block. Random forest [24] is a decision tree method, and therefore non-linear in nature; it is currently gaining popularity in the field of QSAR [25]. An ensemble of trees is created and no pruning employed, the prediction is then obtained as the average from all trees. BNNs [26] also model non-linear relationships, they differ from classic neural networks in that a distribution of weights rather than individual weights are employed when attempting to fit the data. Due to the presence of an ensemble of networks, BNNs are thought to be less likely to result in overfitting, a common problem with the neural network approach.

The PLS and RF models were built with the R [27] statistical environment. The PLS package version 2.0.0 and the RF package version 4.5-18 were employed. For each PLS model a 7-fold cross-validation approach was followed, a new component was added each iteration if the r^2 was improved by at least 0.02 and the q^2 was improved from n-1 components. For RF the number of trees was set to 250 and the tuning function used to identify a suitable number for *mtry*, starting at a value of 64 (total number of descriptors/3). The BNN models were built using code written by Neal [28] and developed in-house to include an automated routine for variable selection [11].

An in-house descriptor set [11] was used to characterise the compounds for the QSAR models, these can be grouped into 3 classes: topological (2D), geometrical (3D) and electronic (charge-dependent) descriptors. They briefly include molecular weight, CLOGP, atom and ring counts, hydrogen bond donor and acceptor counts, solvent accessible surface area of donor/acceptor/polar atoms and distribution of charges and areas of positive or negative electrostatic potential. The main focus for these descriptors is the molecular surface since it is considered that molecules interact through their surfaces.

A time period of three years was studied. For each property (human plasma protein binding, solubility and log D7.4) an initial training set was generated using measurements made before the period of study. Measurements were then collected on a monthly basis for a period of 35 months, hence resulting in 36 different training sets (training set 1 = all measurements prior to study, training set 2 = training set 1 + month 1 measurements etc) and 35 test sets (test set 1 = month 1 compounds, test set 2 = month 2 compounds etc). Finally, a terminal temporal test set was formed by collecting all data measured for the 3 months following the 35 month period of study.

The first analysis examines the performance of static models over time. A model was built for each property using the initial training set and PLS, RF and BNN (a total of 9 models). The 35 monthly sets formed a series of temporal test sets, for which predictions were made from each of the models. By calculating the model's ability to predict each test set, the performance of the model over time can be examined.

The second analysis focuses on regularly updating models and how well they are able to predict a terminal temporal test set. After the initial models were built for each property, subsequent models were automatically generated each month using PLS and RF by adding the monthly data set from the following month to the training set. This resulted in a total of 36 PLS and 36 RF models for each property. Due to the extended length of time required to reach convergence, BNN models were not built each month – instead a model was built using data representing each year of this study. Thus, in addition to the initial model described above, 3 further BNN models were built for each property. The first model was built using all monthly data up to month 12, the next one using all data up to month 24 and a final model using all data up to month 35. Each model was used to make predictions for the final terminal temporal test set.

The experimental measurements used to train the models originate from many different sites. However, internal cross-comparison studies have shown that the assays used at the different sites produce data within each other's experimental error and hence are not significantly different (average $r^2 = 0.75$). In prediction no bias was seen between data originating from the different sites.

Validation using an external test set is key for estimating the true predictive power of a QSAR model [29]. Test set selection methods often identify test sets that mirror the compounds represented in the training set. These test set compounds are likely to have near neighbours in the training set thus providing only a weak assessment of model predictivity. The test sets used in this study are all temporal to the models, collected from experimental measurements made subsequent to the training set compounds. This more closely resembles how the models will be used in practice (to predict 'future'

compounds) and hence provides a more robust assessment of the predictive ability of the models. The root mean square error in prediction (*RMSE*) reflects deviations of the predicted from the observed value, it is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (1)$$

where n is the total number of compounds; $\hat{y}_i$ is the predicted dependent value and y_i is the observed dependent value. The RMSE of each test set provides an overview of the predictivity of the model used at that time point (the month over which the test data was measured).

The distance to model is calculated by determining the distance in descriptor space of a query compound to all training set compounds. The average of the 3 closest training set compounds is used as an indication of the distance to model [30]. The distance metric used here is the Mahalanobis distance (MD_{ij}):

$$MD_{ij} = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)} \quad (2)$$

between the query, i, and training set compounds j, where $(x_i - x_j)$ is the column vector of the descriptor value differences between the query x_i and library x_j compounds; $(x_i - x_j)^T$ is the transposed vector of $(x_i - x_j)$; and S^{-1} is the inverse covariance matrix of the descriptors x_j of the library compounds. This distance measure is based on all of the descriptors that are included in the global model, the average distance to the 3 nearest neighbours in the training set is normalised for the number of descriptors:

$$DISTANCE\ TO\ MODEL = \sqrt{\frac{15}{d} av3(MD_{i,j})} \quad (3)$$

where d is the number of descriptors used in the model.

3. Results

3.1. Fixed Model

The predictive ability of the 3 initial models for human plasma protein binding on the individual monthly test sets is shown in Figure 1. All models are able to effectively predict compounds measured up to 3 years in the future from when the model was built, the RMSE in prediction is always lower than the standard deviation of the test set. The standard deviation varies each month with the number and type of compounds that are present. For the majority of months, RF provides the best predictions, however the RMSEs are often not largely dissimilar to those from the PLS and BNN models.

The predictivity of the individual monthly test sets by the initial human plasma protein binding model can also be presented in a cumulative fashion (Figure 2). Here, the first point on the x-axis represents the first test set (month 1); the second point represents both the first and second test sets (months 1 and 2) and so on to the final point which includes all monthly test sets.

Both the RF and BNN lose predictivity over time when all test sets are considered together. However, the PLS predictivity is more stable, the initial and final RMSEs are 0.62. The initial RF model predicts the first test set with an RMSE of 0.53 and the final cumulative test set with an RMSE of 0.57. The BNN model has the greatest reduction in performance over time with an RMSE for the first test set of 0.56 and for the final cumulative test set of 0.64. Initially the BNN model is more predictive than the

PLS model, at approximately month 17 this situation is reversed as the BNN model deteriorates and the PLS model maintains its predictivity.

The RMSEs resulting from the predictions for the individual monthly test sets by the initial solubility model are plotted in Figure 3. The models are predictive of the individual test sets across the time-series. However, there is an anomaly in the data at test set 2 where the standard deviation suddenly increases relative to all of the other data sets. The PLS, BNN and RF models have poor predictivity for this data set. At this point in time (month 2), 2470 compounds were added to the in-house database in just one month (the monthly average number of compounds measured for this dataset is 274). It is likely that a large number of new compounds were added to the database at once, perhaps due to the updating of a specific assay. The initial models are unable to effectively predict this large number of new compounds.

The addition of such a large number of compounds in a single month distorts the shape of the cumulative test set RMSEs plot (Figure 4), the RF models are specifically affected. The RMSE increases sharply with the inclusion of so many extra compounds in the test set. As subsequent test sets, which are all better predicted than the test set from month 2 (evident from Figure 3), are added to the cumulative set a slow reduction in the RMSE results as the post-month 2 compounds account for a greater majority of the cumulative test set.

If the first two test sets (months 1 and 2) from the time-series are excluded from the plot (Figure 5), the pattern of predictivity more closely resembles the initial plasma protein binding model. The PLS model is very stable, but also actually improves over the time-series in terms of its predictivity of the cumulative solubility test sets. The RMSE for prediction of the first test set (month 3) is 0.83 and for the final cumulative test set 0.80. The RF model also improves in predictivity initially, before returning to an RMSE equal to that at the start of the time-series (RMSE = 0.75). As before, the BNN model becomes less predictive over time with an increase in RMSE from 0.78 to 0.82. The BNN model is less predictive than the PLS model after 23 months. In general, the solubility models produce greater errors in prediction than the human plasma protein binding models.

The predictivity of the initial log D model on the individual monthly test sets is plotted in Figure 6. The initial models are all predictive across the time series, the RMSEs each month are significantly lower than the standard deviation of the test sets. The RF model appears to offer the greatest predictivity overall having the lowest RMSE for the majority of months.

Figure 7 presents the predictivity of the initial log D7.4 models where the monthly test sets are considered cumulatively. There is an increase in the RMSEs from the beginning to the end of the time-series with all three initial QSAR models. The addition of the first year's measurements has the greatest effect on the RMSE, after this time it remains relatively stable. The addition of a month's worth of compounds will have a greater effect on the RMSE in the earlier stages as it accounts for a greater overall proportion of the cumulative data set. The PLS model is the least predictive for these test compounds, and RF the most predictive. Initially, the BNN models are similar in predictivity to the RF models, but the predictive ability reduces over the time series such that the BNN models RMSE converges towards the RMSE of the PLS model.

3.2. Updating Model

Having observed the effect of time on the prediction accuracy of the static models, the second analysis involved testing the ability of a series of updated models to predict a terminal temporal test set. The predictivity of the PLS, RF and BNN models over the time-series for human plasma protein binding is shown in Figure 8.

There is a clear reduction in the RMSE across the time-series from the initial model (model = 0) to the final model (model = 35) with PLS, RF and BNN. The initial, 3 year-old, RF model predicts the terminal temporal test set with an RMSE of 0.64, this reduces to 0.56 across the time-series. The PLS model RMSE reduces from 0.67 at the beginning of the time-series to 0.61 with the final model. Only 4 BNN models were built to represent the time-series and show its long-term effect, the initial BNN model predicts the terminal temporal test set with an RMSE of 0.68 and the final model 0.59. The RF models consistently provide the most effective predictions for the terminal temporal test set across the three years studied. Initially, the PLS model is more predictive than the BNN model, but as the models are updated to reflect more recent measurements the BNN model is the most predictive. This occurs approximately 1 year before the end of the study.

The predictivity of the time-series models for solubility is shown in Figure 9.

The initial RMSE for the RF model on the terminal temporal test set is 0.88, this reduces to 0.81 with the final RF model. The predictivity of the PLS model remains stable across the time-series with the RMSE at 0.89 for the three years. The BNN model varies more in predictivity across the time-series with an initial RMSE for the temporal test set of 0.91, in line with the other two models. However, the predictivity of the model deteriorates over time with an increase in RMSE to 0.98 in month 13. This then reduces to a value of 0.89 at the end of the time-series, slightly lower than the PLS model.

The BNN approach used in this study begins by selecting a subset of descriptors from the full set of 193 using Automatic Relevance Determination (ARD) [27]. ARD is able to identify the importance of descriptors and hence can be used to remove the descriptors of little importance. This is achieved by building a neural network with weights associated to each of the input units (descriptors), as the neural network is trained the irrelevant descriptors gain smaller weights and the important descriptors have higher weights. Analysis of the weight associated with each descriptor allows for the removal of those descriptors with small weights.

This process takes place before each model is built (since the training data is different for each model), hence different descriptors are selected each time – even where it is the same property being modelled. The large number of compounds added in month 2 may have skewed the ARD to selecting descriptors relevant to these compounds (and less so to the temporal test) and caused this reduction in predictivity. Over time, as more compounds are added this effect is diluted allowing the next model to be more predictive.

The predictivity of the time-series models for log D7.4 is given in Figure 10.

The initial RF model predicts the log D7.4 terminal temporal test set with an RMSE of 0.67, this reduces to 0.57 with the final model. The PLS models, as with solubility, are more stable with an RMSE from the initial model of 0.69 and the final model 0.68. The BNN models initially produce significantly poorer predictions for the temporal test set than the other two methods (initial RMSE = 0.76), but improve over the time-series with an RMSE for the final model of 0.65.

The training set volume increases dramatically across the time-series for each property considered, doubling in size on average. However, the range of measurements does not change with a shift of only 0.01 or 0.02 log units in the y-values. The measurement space is getting denser with the addition of new compounds to the training set, as is the descriptor space. This increase in information, specifically information (descriptors and measurements) that more closely resemble the compounds we want to predict is improving the predictive ability of the resulting models.

4. Discussion

4.1. Fixed Model

When the compounds collected on a monthly basis are considered separately, there appears to be no relationship between the RMSE in prediction and time since model build. It might be expected that the most recent test sets (i.e. months 32-35) would have greater RMSEs in relation to their standard deviation than the earlier test sets (months 1-4) since the most recently measured compounds are likely to share less similarity with the training set compounds than the older test compounds. The distance to model, calculated as the average Mahalanobis distances of each test compound to its three nearest neighbours in the training set, confirms this reduction in similarity (Table 1). For each property studied, the most recently measured test set (month 35) has a greater distance to model than the earliest test set (month 1). The inter-individual variations in size and spread between the test sets may be too great to show a greater error in prediction over the time-series.

Model	Month 1 Test	Month 35 Test
Human plasma protein binding	2.35	2.95
Solubility	2.02	3.02
Lipophilicity log D7.4	2.46	3.33

Table 1 Average Mahalanobis distance of test compounds to 3 nearest training set compounds from initial model

Where the monthly test sets are considered in a cumulative fashion, there is a small increase in the RMSE over time since model build. This effect is only small, as would be expected since the cumulative test set also includes the earlier compounds (from just one month since model build). However, as the number and proportion of most recent compounds in the test set increases, the RMSE also increases. This reduction in model performance over time has also been demonstrated elsewhere with a human ether-a-go-go (hERG) QSAR model [31]. The RMSE in prediction of a series of temporal test sets (collected over a 15 month period after the model was built) reduced as the time since model build increased.

4.2. Updating Model

The results from the second analysis indicate that, in terms of predicting a future external test set, in general models improve when updated to include more recently measured compounds in the training set. This is in agreement with a preliminary study [21] where the updating of a human plasma protein binding PLS model was tested. These results suggest that some methods are more sensitive to this effect than others. The PLS models tend to be improved less by updating than RF and BNN models, they are more stable across the time-series. This work demonstrates that BNN models are the most sensitive to the lag in time between model build and making predictions. The BNN models exhibited the greatest increase in RMSE with the fixed models and updating test sets. In addition, for the updating models, for the final few months of study there was a greater improvement in model predictivity for the BNN models than the PLS or RF models.

One reason that models produce a large prediction error is the discrepancy between the areas of chemical space represented by the training and test sets [32]. By updating the training set on a regular basis, the domain of applicability is being extended to minimise the distance between the training and test sets. The average distance to model, calculated as the average Mahalanobis distance of the terminal temporal test compounds to the three nearest neighbours in the initial and final training sets, is summarised for each property in Table 2.

	Initial model	Final model
Human plasma protein binding	2.99	2.13

Solubility	3.18	2.09
Lipophilicity log D7.4	3.35	2.43

Table 2 Average Mahalanobis distance of terminal temporal test compounds to 3 nearest training set compounds in initial and final models

In each case, the average distance of the terminal temporal test set compounds to the training sets reduces from the initial to the final model. Not only are there more compounds present in the final training set, but they are more likely to be chemically similar to the terminal temporal test set as they have been measured over a closer time period. This suggests that the reason for the improvement in RMSE across the time-series is a reduction in the distance to model. Compounds measured over the same time period are more likely to share structural similarity and hence the most recently measured compounds are the most valuable in building models to predict future compounds.

4.3. Consensus Predictions

Consensus predictions, obtained by combining the predictions from multiple models, have been found to improve the overall predictivity of models [33]. It is thought that the different modelling techniques will compliment each other and collectively provide the best prediction. Generally, consensus models are likely to be effective because a poor prediction for an individual compound from one model can be compensated for by reasonable predictions from the other models. Hence the consensus value provides an improvement over the poor model, and may not be largely different from the models with reasonable predictions. Consensus models guard against overfitting, but assume that where models agree, they are more likely to be right.

Consensus predictions are only possible in the presence of more than one model. Where models take a long time to build, and also to make predictions from, the benefits provided by the consensus predictions may be significant enough to take account of these other costs.

Consensus predictions were obtained for the initial models created for each property (PLS, RF and BNN models using measurements from before the study period) for the compounds in each of the monthly test sets. The consensus prediction is simply the mean of the predictions from the other models. The predictivity of the consensus model for the human plasma protein binding test sets (using the initial models) are shown in Figure 11. The consensus prediction does not provide any real benefit over the best individual model (which in the majority of cases is RF) for the series of test sets considered.

The consensus predictions from the fixed solubility and lipophilicity logD7.4 models for each of the monthly test sets were also calculated (Figures 12 and 13 respectively).

For solubility the RMSE of the consensus predictions is very similar to that of the best model (RF). For approximately half of the test sets the consensus RMSE is lower than RF and for the other half it is greater. However, the difference between the two in most cases is very small.

For lipophilicity logD7.4, for the majority of test sets the consensus RMSE is approximately equivalent to that from the RF model. However, there are a number of months for which the consensus model appears to provide reasonable improvement over the predictivity of the RF model. For example, for the test sets from months 19 to 23 (excluding month 20 where the PLS model is equivalent to the consensus model in predictivity), there is an average improvement in RMSE of 0.05 log units.

Consensus predictions for the terminal temporal test set compounds using the final human plasma protein binding models from the series were also obtained. The RMSE in prediction was 0.55, only

slightly lower than the best individual model (RF, RMSE = 0.56). There was no difference between the RMSEs for the final solubility consensus model and RF models predicting the terminal temporal test set (RMSE = 0.81). For lipophilicty logD7.4, predicting the terminal temporal test set, the RMSE for the final consensus model is 0.59 (RF, model = 0.58). Hence, as with the monthly test sets, there appears to be little gain in selecting the consensus prediction over the best individual model. This is in agreement with findings from another study [34], where four different data sets were considered and no consistent significant improvement in model predictivity in the consensus model was found. Consensus modelling may provide a good option where there is little confidence in individual methods – the consensus model may be seen as the 'safe' option.

From this large study of three diverse datasets it has been demonstrated that, across the time-series, the RF models produce the best predictions. The general success of the RF approach to QSAR model building mirrors findings published elsewhere [35,36] where many different modelling techniques were tested, including RF, PLS, support vector machines, neural networks and k-nearest neighbours on a series of androgenic, nonandrogenic and solubility assay data. These studies were limited to relatively small random test sets, this study demonstrates that RF is able to effectively predict future compounds.

Random forest is able to model non-linear relationships, which may explain why the random forest models described here are more predictive than the PLS models. Non-linear methods are often prone to over-fitting, this is minimised with random forest since an ensemble of trees (a forest) is built. The prediction returned from the model is an average from all of the trees thus spreading the risk of a poor (over-fit) decision tree.

Preprocessing of the data, such as descriptor selection is not required when building random forest models. The random forest method effectively manages the large number of correlated descriptors that have been used here as it simply ignores irrelevant descriptors (they are not selected as nodes in the trees). PLS models are also resistant to irrelevant descriptors, but BNN models benefit from a descriptor selection step, which introduces bias into the models and also removes information from the training set.

At each node in a random forest tree a random subset of the descriptors (1/3) is considered for splitting the data, this prevents domination of the trees by a small number of descriptors and ensures that the trees all differ from each other. This allows a greater range of descriptors to contribute to the model and may be a real strength of the random forest approach.

The findings from this study, and other studies mentioned above, provide convincing justification for the frequent updating of QSAR models and thus motivation for the development of automatic methods of QSAR model generation. Interest is growing in the area of 'autoQSAR', for example a whole architecture for the automated construction and testing of QSAR models has been presented [37]. and hence it is important that the benefits of such systems are demonstrated. Results so far suggest that a process for the automatic regular updating of QSAR models would improve model predictivity.

This study has focussed on the performance of updating global models over time. A future study will analyse the effects on local models, it is expected that the difference in performance over time may be more significant than that demonstrated here as the addition of compounds to the training set will make up a greater proportion of the total training set size.

5. Conclusion

We have examined the updating of three global models (human plasma protein binding, aqueous solubility and log D7.4) using three different statistical techniques over a three-year period. In each

case, the global models were able to make predictions for future compounds; however, updating the global model by updating the training set to represent more recently measured compounds improved model predictivity. This indicates the need for regular updating of models and demonstrates the possible benefits from the automatic updating of QSAR models. Of the three statistical model building approaches examined, RF consistently produced the most predictive models.

Acknowledgement
The authors wish to thank Dr Pierre Bruneau for his help and support in building the Bayesian neural network models.

References

[1] W. Muster, A. Breidenbach, H. Fischer, S. Kirchner, L. Muller and A. Pahler, Computational toxicology in drug development, Drug Discov. Today 13 (2008), 303-310.
[2] L. Turski in Virtual ADMET Assessment in Target Selection and Maturation (Eds.: B. Testa, L. Turski) IOS Press, Washington, DC, 2006.
[3] W.J. Dunn III, S. Wold, U. Edlund, S. Hellberg, J. Gasteiger, Multivariate Structure-Activity Relationships Between Data from a Battery of Biological Tests and an Ensemble of Structure Descriptors: The PLS Method, Quant. Struct.-Act. Relat. 3 (1984), 131-137.
[4] S.S. Young, D.M. Hawkins, Analysis of a 2^9 Full Factorial Chemistry Library, J. Med. Chem. 38 (1995), 2784-2788.
[5] T. Aoyama, Y. Suzuki, H. Ichikawa, Neural networks applied to pharmaceutical problems. III. Neural networks applied to quantitative structure-activity relationship (QSAR) analysis, J. Med. Chem. 33 (1990), 2583-2590.
[6] R. Burbidge, M. Trotter, B. Buxton, S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, Computers & Chemistry 26 (2001), 5-14.
[7] R.E. Carhart, D.H. Smith, R. Ventkataraghavan, Atom pairs as molecular features in structure-activity studies: Definition and application, J. Chem. Inf. Model. 25 (1985), 64-73.
[8] L.B. Kier, L.H. Hall, Derivation and significance of valence molecular connectivity J. Pharm. Sci. 70 (1981), 583-589.
[9] D.T. Stanton, Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies, J. Chem. Inf. Model. 39 (1999), 11-20.
[10] J. Polanski, A. Bak. R. Gieleciak, T. Magdziarz, Modeling Robust QSAR, J. Chem. Inf. Model. 46 (2006), 2310-2318.
[11] P. Bruneau, Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets, J. Chem. Inf. Model. 41 (2001), 1605-161.
[12] T.R. Stouch, J.R. Kenyon, S.R. Johnson, X.-Q. Chen, A. Doweyko, Y. Li, J., In silico ADME/Tox: why models fail, Comput.-Aided Mol. Des. 17 (2003), 83-92.
[13] Y. Xu, H. Gao, Dimension related distance and its application in QSAR/QSPR model error estimation, QSAR Comb. Sci. 22 (2003), 422-429.
[14] R.W. Stanforth, E. Kolossov, B. Mirkin, A Measure of Domain of Applicability for QSAR Modelling Based on Intelligent K-Means Clustering, QSAR Comb. Sci. 26 (2007), 837-844.
[15] A.M. Davis, R.J. Riley, R.J., Predictive ADMET studies, the challenges and the opportunities, Curr. Opin. Chem. Biol. 8 (2004), 378-386.
[16] 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.
[17] J. Cartmell, D. Enoch, D. Krstajic, D.E. Leahy, Automated QSPR through Competitive Workflow, J. Comput.-Aided Mol. Des. 19 (2005), 821-833.
[18] O. Obrezanova, G. Csányi, J.M.R. Gola, M.D. Segall, Automatic QSAR modelling of ADME properties: blood-brain barrier penetration and aqueous solubility, J. Chem. Inf. Model. 47 (2007), 1847-1857.
[19] R. Guha, D. Dutta, P.C. Jurs, T. Chen, Local Lazy Regression: Making Use of the Neighbourhood to Improve QSAR Predictions, J. Comput.-Aided. Mol. Des. 46 (2006), 1836-1847.
[20] S. Zhang, A. Golbraikh, S. Oloff, H. Kohn, A. Tropsha, A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models, J. Chem. Inf. Model. 46 (2006), 1984-1995.
[21] S.L. Rodgers, A.M. Davis, H. Van De Waterbeemd, Time-Series QSAR Analysis of Human Plasma Protein Binding Data, QSAR Comb. Sci. 26 (2007), 511-521.
[22] H. Van De Waterbeemd, E. Gifford, ADMET in silico modelling: Towards Prediction Paradise? Nat. Rev. Drug Discov. 2 (2003), 192-204.

[23] R. Mannhold in Virtual ADMET Assessment in Target Selection and Maturation (Eds.: B. Testa, L. Turski)Turski) IOS Press, Washington, DC, 2006.
[24] L. Brieman, Random forests, Machine Learning 45 (2001), 5-32.
[25] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, J. Chem. Inf. Model. 43 (2003), 1947-1958.
[26] W.P.W. Ajay, M.A. Murcko, Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules?, J. Med. Chem. 41 (1998), 3314-3324.
[27] R Development Core Team 2006 Version 2.7.1 (Windows). Available from: http://www.r-project.org/index.html
[28] R.M. Neal. Software for Flexible Bayesian Modeling, version of 06-12-1999. http://www.cs.utoronto.ca/~radford
[29] A. Golbraikh, A. Tropsha, Beware of q^2! J. Mol. Graphics Modell. 20 (2002), 269-276.
[30] Bruneau, P. McElroy, N.R., LogD7.4 Modeling Using Bayesian Regularized Neural Networks. Assessment and Correction of the Errors of Prediction, J. Chem. Inf. Model. 46 (2006), 1379-1387.
[31] C.L. Gavaghan, C. Hasselgren Arnby, N. Blomberg, G. Strandlund, S. Boyer, , Interpretation and Temporal Evaluation of a Global QSAR of hERG Electrophysiology Screening Data, J. Comput-Aided Mol. Des. 21 (2007), 189-206.
[32] I.V. Tetko, P. Bruneau, H.-W. Mewes, D.C. Rohrer, G.I. Poda, Can we estimate the accuracy of ADME-Tox predictions? Drug Discov. Today 11 (2006), 700-707.
[33] N. Baurin, J.-C. Mozziconacci, E. Arnoult, P. Chavatte, C. Marot, L. Morin-Allory, 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database, J. Chem. Inf. Model. 44 (2004), 276-285.
[34] Hewitt, M., Cronin, M.T.D., Madden, J.C., Rowe, P.H., Johnson, C., Obi, A. and Enoch, S.J., Consensus QSAR Models: Do the Benefits Outweigh the Complexity? J. Chem. Inf. Model. 47 (2007), 1460-1468.
[35] Palmer, D.S., O'Boyle, N.M., Glen, R.C. and Mitchell, J.B.O., Random Forest Models to Predict Aqueous Solubility, J. Chem. Inf. Model. 47 (2007), 150-158.
[36] Li, Y., Wang, Y. Ding, J., Wang, Y., Chang, Y., Zhang, S., In silico Prediction of Androgenic and Nonandrogenic Compounds Using Random Forest, QSAR Combi. Sci. 4 (2009), 396-405.
[37] Cartmell, J., Krstajic, D. and Leahy, D.E., Competitive Workflow: Novel software architecture for automatic drug design, Curr. Opin. Drug Di. De. 10 (2007), 347-352.