logD$_{7.4}$ Modeling Using Bayesian Regularized Neural Networks. Assessment and Correction of the Errors of Prediction

Pierre Bruneau* and Nathan R. McElroy[†]

AstraZeneca, Parc Industriel Pompelle, BP 1050, 51689 Reims Cedex 2, France

Received September 13, 2005

Bayesian Regularized Neural Networks (BRNNs) employing Automatic Relevance Determination (ARD) are used to construct a predictive model for the distribution coefficient *logD*$_{7.4}$ from an in-house data set of 5000 compounds with experimental endpoints. A method for assessing the accuracy of prediction is established based upon a query compound's distance to the training set. *logD*$_{7.4}$ predictions are also dynamically corrected with an associated library of compounds of continuously updated, experimentally measured *logD*$_{7.4}$ values. A comparison of local models and associated libraries comprising separate ionization class subsets of compounds to compounds of a homogeneous ionization class reveals in this case that local models and libraries have no advantage over global models and libraries.

INTRODUCTION

The pH-dependent distribution coefficient, *logD*, is an important parameter when considering many Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADME/Tox) properties. The ability of a drug to pass through physiological barriers (e.g., biological membranes), reach its intended target area, and act upon a target in the desired manner becomes a major focus of drug design. Depending upon the method of drug administration and final destination, the range of pH environments that a drug may encounter can be quite varied, from stomach acid (pH = 2.0) to blood plasma (pH = 7.4). Knowing more about a compound's aqueous solubility and/or lipophilicity in these pH ranges allows for better drug design through improved property prediction and, ultimately, synthesis of more efficacious leads.

Predictive software is available from commercial and academic resources, but often when it is applied to in-house pharmaceutical databases, the results can be disappointingly inaccurate.[1−4] Methods may include predefined models that are based on a specific training set of data, group contribution, and any number of fixed molecular descriptors applied to linear and nonlinear algorithms. Often the *logD*'s are predicted via calculation from independently predicted parameters, namely *logP* and pK_a.[5−9] To our knowledge, it has rarely been done by modeling *logD* directly. Cerep seems to have followed such an approach for their *logD* pH7.4 model included in their BioPrint package,[10] but full details have not been published elsewhere other than in a conference report.[11] Another approach consists of adjusting *logP* predictions with a library of measured *logD* data.[1,2] While all of these approaches are certainly feasible and have been successfully applied in *logD* prediction for some data sets, they are usually not global in their success and require some modification when applied to a particular company's chemical library.

In this study, data from internal collections with their *logD* measured at pH = 7.4 (*logD*$_{7.4}$) were analyzed with in-house methodology. This system included data pretreatments, molecular descriptor calculations using both commercial and in-house algorithms, model building tools using available Bayesian neural network software, and analyses with both in-house and commercial application.

METHODOLOGY

Database. An initial database of compounds with *logD*$_{7.4}$ values was constructed by extracting compounds from two different AstraZeneca site repositories of physicochemical comprising in-house measurements. These compounds were selected based on homogeneous experimental conditions (solvent, pH, etc.). Any compound with two or more measurements was reported singly using the average of multiple measurements. Compounds for which too great a range between multiple measurements existed were discarded, in this case a range greater than 0.3 log units. This selection process led to 8200 different molecules with associated *logD*$_{7.4}$ measurements that served as the model building data set. This data set was further split into a training set and an 'ex-clusters' validation set, described below.

Later, a second data set was extracted in similar fashion from all AstraZeneca site repositories for use as a general purpose validation set. This set consisted of 20 963 compounds with *logD*$_{7.4}$ endpoints. After data pretreatment, and exclusion of the compounds already included in the training set, a total of 16 325 compounds were available for a global validation set that were applied to the predictive model.

Descriptors. For the initial 8200 compound data set, a diverse set of two- and three-dimensional and charge-dependent molecular descriptors were calculated using an in-house program, Drone.[12] The *logD*$_{7.4}$ values (ACD-LogD74) were calculated from a commercially available suite of programs, ACD Labs,[8] in its UNIX batch version. In total,

* Corresponding author phone: +33 (0)3 26 61 68 52; fax: +33 (0)3 26 61 68 42; e-mail: pierre.bruneau@astrazeneca.com.

[†] Present address: Department of Chemistry, Indiana University of Pennsylvania, 975 Oakland Avenue, Indiana, PA 15705.

10.1021/ci0504014 CCC: $33.50
Published on Web 12/22/2005

Table 1. Distribution of the Number of Compounds of the Initial Set into the Clusters

initial set		no. of compds from each cluster	
no. of compds in clusters	no. of clusters	in training	in ex-clusters validation
1	2770	1	0
2	1539	1	1
3	495	1	2
4	141	1	3
5	41	1	4
6	11	1	5
7	3	1	6
total no. of compds	8189	5000	3189

122 descriptors were calculated and used to establish a predictive model.

Data Pretreatment. The initial data matrix consisting of 8200 compounds (rows) and 122 molecular descriptors (columns) underwent some premodeling treatment to maximize the amount of useful information within. These steps included removing rows for which the ACD program was unable to calculate a $logD_{7.4}$ value, removing columns that had no variation in descriptor values, and descriptor decorrelation. The decorrelation step was performed by calculating the covariance matrix of the descriptors. If two descriptors had a covariance value higher than a predefined cutoff (0.95), then one of those two descriptors was removed from consideration.

Selection of Training Set. To select a diverse subset of compounds for model training, the descriptor matrix was submitted to a hierarchical clustering process. Using Ward's method for standardized data,[13] a complete cluster tree was constructed and 5000 clusters were selected. A training set of 5000 compounds (i.e., one compound per cluster) was formed, and the remaining compounds were placed in an 'ex-cluster' validation set. Table 1 outlines the selection of compounds based on this clustering procedure.

Bayesian Regularized Neural Networks (BRNNs). The implementation of BRNNs in Neal's Flexible Bayesian Modeling (FBM)[14] was used for nonlinear regression. BRNNs offer several advantages for predictive modeling including: the ability to model any function without the need to predefine it; the ability to easily handle real and categorical molecular descriptors; and insensitivity to overfitting and overtraining.[15,16] Use of BRNNs in our methodology has been described.[12] In this former study, we initially trained the BRNNs using up to several thousand iterations in order to reach equilibrium before choosing the final number of networks to average for property prediction. In this case, we defined network equilibrium as a minimization of change in network error, which we found to be related to a minimization of change in the weight values associated with molecular descriptors in the network. In many cases, this led to very long and usually unnecessary training cycles. It was noted that the majority of models were reaching equilibrium at a much lower number of iterations that previously hypothesized. An algorithm was designed to continuously control the convergence of the BRNNs to our defined equilibrium state. After each cycle of 50 training iterations, the model was assessed by its root-mean-square error (rmse) and then allowed to continue training from that endpoint. After a set number of cycles (e.g., 4) in which the rmse did not improve or significantly change, training was halted, and these final cycles of model iterations were used for final property value calculations. Normally, the final 200 to 400 individual neural networks with distinct sets of weight and bias terms were used to calculate a property value mean (y_{pred}) with associated standard deviation ($\mathrm{std}_{\mathrm{pred}}$).

Automatic Relevance Determination (ARD). The principle of parsimony calls for using predictive models that contain only the information that is exactly needed and nothing more.[17,18] In this study, starting with 120 molecular descriptors followed by pretreatment still left more information than was likely required. In FBM, it is possible to implement ARD to allow the selection of only the most relevant molecular descriptors.[14] When the ARD is used in BRNNs, a hyperparameter is added to control the prior probability distributions for each level of the network. Parameters for each network level control distribution parameters for each layer neuron that in turn control the magnitude of weight and bias terms used for each network connection. As training proceeds and probability distributions are adjusted, the weight values associated with irrelevant descriptors diminish, while weights associated with more important descriptors increase in magnitude.

In this implementation, the process starts with the most complete model (i.e., all molecular descriptors after pretreatment). The network is trained until the equilibrium point described above is reached, and the final sets of weight and bias terms are analyzed. For each descriptor, the sum total of squared weight and bias terms for the final number of individual networks are compared. Those descriptors whose sums are declared negligible (e.g., less than 1% the value of the maximum weighted descriptor) are then removed from consideration in the next round of network training. The process is repeated until it is not possible to remove further descriptors without significantly degrading the overall performance of the network.

Although not absolutely necessary in a strictly Bayesian point of view, ARD has been used in other works[19-22] to define the relative importance of descriptors using one model. We used the ARD to continuously remove descriptors, deemed of little importance by the guideline above, retrain the model using a smaller number of descriptors, and repeat the process. This routine led to a series of predictive models with ever decreasing numbers of molecular descriptors, each with an associated overall model error. Our goal was to build a predictive model with the fewest number of descriptors without jeopardizing the quality of the model.

Distance to the Model. No model of a physicochemical property that is developed from a finite set of compounds can predict the same property for the whole chemical universe with a reliable and unique likely error (with the notable exception of molecular weight). It is obvious that the likely errors in the property calculation of the training and validation sets are obtained by applying a model on those compounds and calculating their errors. These calculated errors do not give any information about the likely error associated with the property prediction of a compound that is not a member of either of these sets. To have a way to evaluate the likely error of a prediction, or even a way to say that a model is not applicable to a specific compound due to its dissimilarity to those used to construct the model, a distance metric has been defined. Similar approaches have been published.[23-25] Sheridan et al.[25] discussed the issues arising from applying models in other chemical spaces from which they have been trained. The calculation of the distance between a specific compound and an aggregate value of compounds used to represent a model or simply 'distance to model' concerns only training set compounds. The distance of one compound i to the model (M) is DM_i.

The first step in calculating DM_i is finding the Mahalanobis distances[26] (MD_{ij}) of compound i to each training compound j according to eq 1, where S^{-1} is the

$$\mathrm{MD}_{i,j} = \sqrt{(y_i - y_j)^{\mathrm{T}} S^{-1} (y_i - y_j)} \tag{1}$$

inverse covariance matrix of the descriptors y_j of the training set; $(y_i - y_j)$ is the column vector of the descriptor value differences between compound i and the training set compounds; and $(y_i - y_j)^{\mathrm{T}}$ is the transposed row vector of $(y_i - y_j)$. For our implementation, an average of the three smallest distances is calculated, av3(MD_{ij}). The average of the three smallest distances has been retained after some development work which has shown that it has a good relationship with the likely errors.[27] Finally, a normalized value DM_i is retained as the distance of compound i from the model M as shown in eq 2. Here, d is the number of descriptors used in M.

$$\mathrm{DM}_i = \sqrt{\frac{15}{d}\mathrm{av3}(\mathrm{MD}_{i,j})} \tag{2}$$

Likely Errors. The distribution of DM_i values of the whole validation set was binned into quartiles in order to evaluate the dependencies of the errors of prediction toward the distance to the model. This method allowed an even partition of compounds into four categories, namely: 1st, 2nd, 3rd, and 4th quartiles. These quartiles were in turn categorized into ionization classes as defined later. The performance criteria of models were evaluated for the different categories. Similarly, the standard deviations of individual compound predictions ($\mathrm{std}_{\mathrm{pred}}$) were binned into quartiles, and the rmse of the different categories were calculated.

Classification into Ionization Classes. Because $logD_{7.4}$ has a pK_a component, it is probable that predictions of compounds ionized at pH = 7.4 are more difficult to realize than predictions of neutral compounds for which $logD_{7.4} = logP$. To evaluate the performance of models more effectively for various ionization categories, the following classifications were created: acids, bases, zwitterions, and neutrals. Each classification was defined by two criteria. The first criterion relied on proprietary SMARTS[28] definitions of common preclassified substructures that were likely to be ionized at pH = 7.4. The second criterion used the relative values of calculated $logP$, $logD_{7.4}$, and $logD_{6.5}$ values that were calculated with the ACD Labs[8] program. A compound was then determined to be an acid ($logD_{7.4} - logD_{6.5} < -0.1$), a base ($logD_{7.4} - logD_{6.5} > -0.1$), or a neutral ($logP - logD_{7.4} \leq 0.1$ or $logP - logD_{6.5} \leq 0.1$). All other compounds were classified as zwitterions at pH = 7.4. The compound was retained in its respective class if and only if the two criteria were identical. Otherwise, the compound was classified as doubtful and excluded from analysis. When a compound could not be classified by one of the above methods (e.g., an error in software calculations), it was classified as unknown and also excluded from analysis.

Associated Libraries. The concept of associated libraries was developed by I. Tetko et al.[29,30] By using known errors of prediction on measured data of compounds that may or may not have been used in model training, it is possible to correct the prediction errors of similar compounds for which there are no known measurements. Tetko et al. use eq 3 to obtain a prediction

$$\bar{z}'_i = \bar{z}_i + \frac{\sum_j (y_j - \bar{z}_j) F(\xi_{ij})}{\sum_j F(\xi_{ij})} \tag{3}$$

where $\bar{z}_i'$ is the corrected prediction of the compound i; $\bar{z}_i$ is the prediction 'as is' of compound i without correction; $(y_j - \bar{z}_j)$ is the error observed with the prediction of the nearest neighbors j; $F(\xi_{ij})$ is a function that evaluates the proximity between compound i and its nearest neighbors j; and $\sum_j F(\xi_{ij})$ is a normalizing factor. $F(\xi_{ij})$ must modulate the correction according to the proximity between i, the compound to be predicted and j, its nearest neighbor. Values range from a full correction, $F(\xi_{ij}) = 1$ when compounds i and j are identical or very similar, to no correction, $F(\xi_{ij}) = 0$ when compounds are very dissimilar. It was observed that if errors $(y_j - \bar{z}_j)$ are identical for three nearest neighbors, the correction factors $F(\xi_{ij})$ do not play a role in weighting, and the correction is fully applied even if the nearest neighbors are very dissimilar with the compound to be predicted. In this work, eq 3 has been modified to eq 4.

$$\bar{z}'_i = \bar{z}_i + \frac{\sum_j (y_j - \bar{z}_j) F(\xi_{ij})^2}{\sum_j F(\xi_{ij})} \tag{4}$$

Tetko has explored various functions for $F(\xi_{ij})$, including the Pearson linear coefficient of the predictions of associative neural networks and different measures of distances, such as Euclidean distance and Spearman or Kendall rank correlations. Relying on our experience of Mahalanobis distances, our $F(\xi_{ij})$ is based on MD_{ij} included in eq 5, to decrease the correction factor as compounds

$$\left| \begin{array}{l} \mathit{if}\ \mathrm{MD}_{ij} \leq \mathrm{pl}\ \mathrm{then} F(\xi_{ij}) = 1 \\ \mathrm{else}\ F(\xi_{ij}) = e\dfrac{-(\mathrm{MD}_{ij} - \mathrm{pl})^2}{2\sigma^2} \end{array} \right. \tag{5}$$

become more dissimilar from each other. Here, pl is a parameter that defines a plateau for the function during which the correction factor is constant and equal to 1. The correction factor decreases to zero according to a Gaussian function defined by the two parameters e and σ. As an example, Figure 1 shows the shape of the function $F(\xi_{ij})$ for pl = 0.8, $e = 2.72$, and $\sigma = 0.75$.

Development work showed that the efficiency of the correction increases slightly with $j = 1$ to 3 where it reaches a plateau. For this reason $j = 3$ has been used in this work. The normal use of a library is to correct a prediction according to the knowledge accumulated in the library. The predicted compounds and the compounds which make up the library are usually different. For developmental purposes,

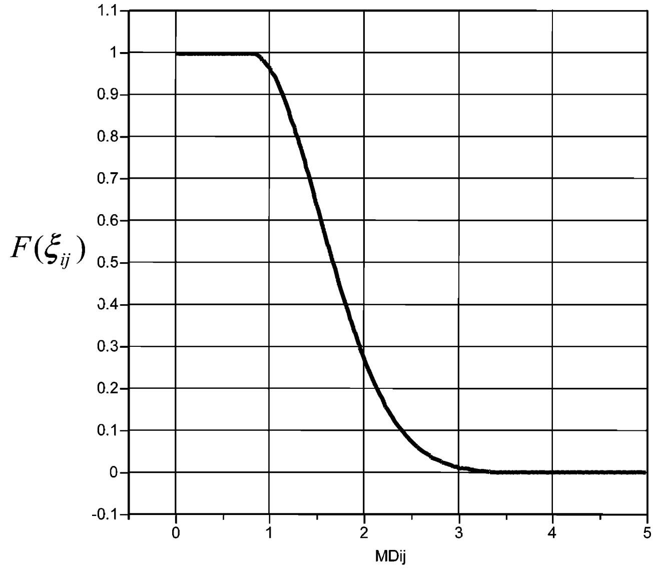

Figure 1. Associated library correction factor, $F(\xi_{ij})$, as a function of Mahalanobis distance (MD_{ij}) from library set data. As compound distance increases, the correction approaches zero in a Gaussian fashion.

the validation set and the library set here were the same. To determine the correction of the validation set with the library, the nearest compound in the library is ignored (i.e., not comparing the compound to itself). This leave-one-out (LOO) type procedure allows us to correct the largest validation set with the greatest number of compounds available. Tetko et al.[1] have shown this procedure to be similar to when the validation set and the library set are entirely different.

Performance Measures. The performance of the model is evaluated by the rmse and by percent **I**mprovement **O**ver the **N**aïve model (ION) as defined by Tiño et al.[31] The rmse for the prediction of a given set is defined by eqs 6 and 7

$$\mathrm{mse}_{\mathrm{set}} = \frac{1}{N_{\mathrm{set}}}\sum_{1}^{N_{\mathrm{set}}}(y_{\mathrm{pred}} - y_{\mathrm{meas}})^2 \qquad (6)$$

$$\mathrm{rmse}_{\mathrm{set}} = \sqrt{\mathrm{mse}_{\mathrm{set}}} \qquad (7)$$

where y_{meas} is the measure value of a compound and y_{pred} is its corresponding predicted value. The naïve model used to define ION is a model where the prediction for all compounds is given by the mean of measured values of the training set ($\bar{y}_{\mathrm{train}}$). Thus, using the naïve model, the corresponding $\mathrm{mse}_{\mathrm{naive}}$ for a given set is defined in eq 8

$$[\mathrm{mse}_{\mathrm{naive}}]_{\mathrm{set}} = \frac{1}{N_{\mathrm{set}}}\sum_{1}^{N_{\mathrm{set}}}(\bar{y}_{\mathrm{trn}} - [y_{\mathrm{meas}}]_{\mathrm{set}})^2 \qquad (8)$$

Tiño et al.[31] define $[\mathrm{ION}_{\mathrm{model}}]_{\mathrm{set}}$, or ION for a model applied on a particular set (eq 9)

$$[\mathrm{ION}_{\mathrm{model}}]_{\mathrm{set}} = \frac{[\mathrm{mse}_{\mathrm{naive}}]_{\mathrm{set}} - [\mathrm{mse}_{\mathrm{model}}]_{\mathrm{set}}}{[\mathrm{mse}_{\mathrm{naive}}]_{\mathrm{set}}}100\% \qquad (9)$$

where the set may be either the training or any validation set. It is noted that when applied to the training set, $[\mathrm{ION}_{\mathrm{model}}]_{\mathrm{set}}$ is not different from 100 times the squared correlation coefficient r^2 between measured and predicted values.

Table 2. Distribution of the Number of Compounds of the Global Validation Set into Ionization Classes

	in-training			validation		
category	N	$\bar{y}$	sd	N	$\bar{y}$	sd
all	*4658*	2.32	1.34	*16325*	2.33	1.28
ACIDS	*633*	1.32	1.36	*1983*	1.12	1.28
BASES	*2067*	2.07	1.07	*6648*	2.14	1.23
NEUTRALS	*1538*	2.96	0.92	*5961*	2.86	1.04
ZWITTERIONS	*81*	1.32	1.34	*237*	1.17	1.08

RESULTS AND DISCUSSION

Training Set. In all, 122 molecular descriptors were calculated for 8189 compounds of the initial data set of 8200 compounds. The covariance coefficients of these descriptors were calculated, and when any pair of descriptors had a coefficient value exceeding 0.95, then one of those two descriptor vectors was removed randomly. This produced a matrix of 98 descriptors and 8189 compounds that was submitted to the hierarchical clustering process to select 5000 maximally diverse compounds to comprise the training set.

Table 1 summarizes the selection process. The 2770 singletons are all in the training set. All of the compounds in the validation set have at least one near neighbor in the training set, but 1539 compounds are singletons. Therefore, they have no near neighbors in the library when the validation set is used as a library in the LOO-like procedure above. These properties do not qualify a validation set selected on this basis as a good validation set. To mimic the ‘real life’ of the use of a model, a different validation set was chosen.

Global Validation Set. Measurements of *logD* from all AstraZeneca sites were collected and treated in an identical manner as those for the initial two-site data set. This database comprised 20 983 compounds that included 4658 compounds of the training set and 16 325 compounds of the validation set above and temporal data that were measured after the establishment of the original model. The characteristics of the in-training set, validation set, and the classified subsets used to evaluate the model performance are shown in Table 2.

Training and Selection of the Model. The ARD process ran for 25 loops and realized a final choice of 56 descriptors. The 5000 compounds by 56 descriptors matrix was submitted to the training of a BRNNs with 43 hidden nodes ($\rho = 2$) (see Bruneau[12] for the basic details of our BRNNs methodology) to give the final model, labeled AZLogD74. For the final training, the controlled convergence process stopped after 450 cycles, leaving a final model whose results were based upon the mean prediction value of the last 200 individual neural nets. This model, when applied to the training set, had an rmse = 0.44 log units and $r^2 = 0.89$ as shown in Table 3. As a comparison, after excluding the compounds which have a range of measurements higher than 2 log units (indicating a likely error of one of the measurement or in the database[32]), the mean of the standard deviations of 307 compounds with triplicate or more measurements is 0.27 log units. This estimation of the experimental error compares well with the results obtained with the model.

Table 3. Results for Model Building Data Set

set	*N*	r^2	rmse
training	5000	0.89	0.44
ex-clusters validation set 'as is'	3153		0.54
ex-clusters corrected validation sub set 'as is'	1799		0.43
ex-clusters corrected validation sub set 'corrected'	1799		0.37

Ex-Clusters Validation Set. The model was applied to the 3189 compounds comprising the ex-clusters validation set and gave a prediction for 3153 compounds (see Table 3). The *logD*$_{7.4}$ value of the remaining 36 compounds (1% of the total number of compounds) could not be predicted because certain model descriptors could not be calculated. An rmse = 0.54 log units for this set represents a moderate decrease in predictability compared to the training set. When the validation set is used as a library to correct these predictions, there are only 1799 compounds from that set to have a near neighbor in the library. Because there is almost no correction observed when distance to the compound in the library exceeds 2.0 (see Figure 1), we observed here that a near neighbor is considered that which has a distance to library less than 2.0. For this qualifying subset of 1799 compounds, the predictions of the model 'as is' (i.e., without library correction) have an rmse = 0.43 log units. This is almost identical to the error of the training set. This result, while puzzling, can be tentatively explained by the mode of selection of the validation and library subsets. Clustering to 3000 groups shows that the majority of the singletons in the 5000 cluster selection remain singletons or doublets in the 3000 cluster selection. Therefore, a compound selected from a populated cluster has a better chance to have a near neighbor in another populated cluster than does a singleton. As a result, the model is more defined near the compounds coming from populated clusters than near compounds from singleton clusters.

In the 1799 subset, the rmse of 0.43 log units is reduced to 0.37 log units via library correction. This is an excellent result because it is comparable to the experimental error of the *logD* measurements. However, as the results obtained with the ex-cluster validation set are biased by the mode of selection of the validation set, it is more meaningful to study the results of the global validation set.

Assessment of the Errors of Predictions. To evaluate the results obtained from the global validation set, the normalized averaged distance of compounds to the model DM_i must be taken into consideration as shown in Table 4. Those compounds with the smallest $MD_{ij} = 0$ are defined as the 'in-training' set, in this case representing 4658 compounds from multiple company sites or global validation set. This subset of compounds has an rmse = 0.43 log units, which is comparable to the rmse of the training set (0.44 log units).

The entire global validation set has an rmse = 0.63 log units. This is higher than the ex-clusters validation set (0.54 log units), thus the differential between training and validation sets increases from 0.10 log units for the ex-clusters validation set to 0.20 log units for the global validation set. This difference may be seen as an indication of an 'over fit' model.[13] Below, a discussion of correcting errors via a compound library will dismiss this explanation.

To evaluate the influence of the distance to the model on the error of predictions of the validation set, DM_i has been binned into quartiles. The means DM_i are 1.7, 2.36, 2.84, and 3.74 for the 1st, 2nd, 3rd, and 4th quartiles, respectively. It is observed that the rmse increases smoothly as the distance to the model increases: from rmse = 0.45 log units in the first quartile nearest the training set to rmse = 0.85 in the fourth quartile farthest from the training set. An estimation of the error of the prediction is therefore available according to the dissimilarity of a compound to be predicted compared to the compounds from the training set.

The standard deviation of prediction is also a good indicator of the error of predictions. As shown in Figure 2, the distances to the model DM_i are correlated with the standard deviations of the predictions, std_{pred}. As a consequence, the rmse of the different compound categories after binning their std_{pred} values into quartiles are remarkably similar to the rmse obtained by binning by their DM_i values into quartiles, as seen in Table 5. The relationship between std_{pred} and the accuracy of the prediction is specific to the BRNNs method. It can be explained by the fact that the final result for a prediction is made by averaging the results of many individual networks which are more tightly fitted together in the area where the training compounds are, compared to an area where there are not training compounds. Explanations of this behavior can be found in R. Neal's work.[14]

When the compounds are split into classes according to their likely ionization state at pH = 7.4, there is no difference in the rmse values of the various clusters for the 'in-training' set. In other words, the model fit equally acidic, basic, neutral, and zwitterionic compounds with rmse = 0.42, 0.43, 0.44, and 0.39 log units, respectively (see Table 4). This is not the case for the validation compounds at greater distances from the model, where rmse values are 0.58 log units for BASES, 0.62 log units for NEUTRALS, 0.69 log units for ZWITTERIONS, and 0.72 log units for the ACIDS. The differences in rmse values for categories ACIDS, BASES, and NEUTRALS may be explained in part by the difference of distribution of their respective distances from the model as shown in Figure 3a. The BASES category comprises compounds found more toward the lower end of the distribution compared to ACIDS and NEUTRALS compounds. But comparing results from within the same quartile bins shows that ACIDS are consistently more difficult to predict than BASES and NEUTRALS at the same distances to the model as shown in Figure 3b.

Corrections with an Associated Library. For a correction of prediction of a compound using an associated library, it is only necessary here to consider those compounds whose distance to nearest neighbor in that library (MD_{ij}) is less than 2.0, as those compounds are the only ones whose values are affected. The rmse of 2956 such compounds in the 'in-training' subset ($MD_{ij} = 0$) was 0.39 log units 'as is' (i.e., without library correction), which was slightly improved by library correction to rmse = 0.37 log units (see Table 6), with a significant difference according to a student paired t-test at 95% confidence (t ratio = 1.7832, DF = 2955). An 'overfit' model in this case would have fit the noise of the training set, and the correction with the library would have reintroduced that noise into the results, thus increasing the rmse of the corrected predictions. It is not the case with this model, and we feel this is a good indication that the model has not over fit the training data.

Table 4. Global Validation Set Errors Relative to Compound Distance-to-Model

			validation set				
			distance from training set (quartiles)				
category		in-training	first	second	third	fourth	all
all	*N*	4658	4065	4077	4092	4091	16325
	rmse	**0.43**	**0.45**	**0.52**	**0.63**	**0.85**	**0.63**
ACIDS	*N*	*633*	*505*	*491*	*409*	*578*	*1983*
	rmse	**0.42**	**0.52**	**0.57**	**0.72**	**0.92**	**0.72**
BASES	*N*	*2067*	*2072*	*1774*	*1577*	*1225*	*6648*
	rmse	**0.43**	**0.45**	**0.54**	**0.60**	**0.79**	**0.58**
NEUTRALS	*N*	*1538*	*1112*	*1434*	*1706*	*1709*	*5961*
	rmse	**0.44**	**0.40**	**0.47**	**0.61**	**0.82**	**0.62**
ZWITTERIONS	*N*	*81*	*61*	*51*	*46*	*79*	*237*
	rmse	**0.39**	**0.47**	**0.59**	**0.86**	**0.92**	**0.69**

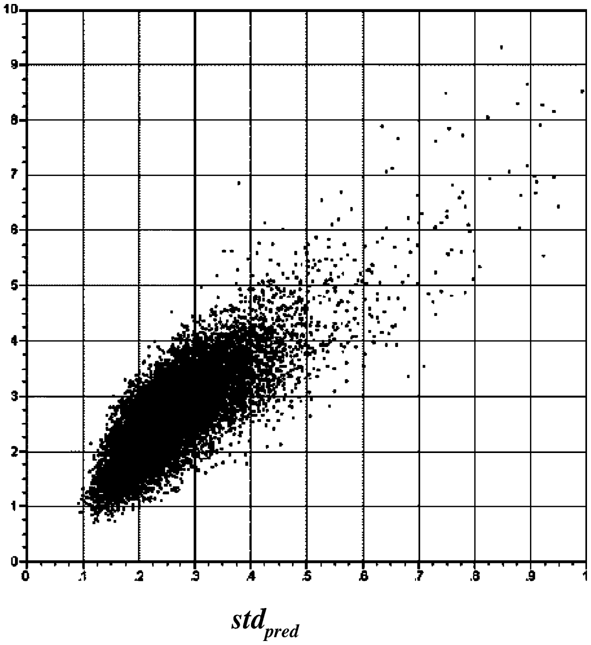

Figure 2. Relationship between DM_i the normalized average of the Mahalanobis distance from the 3 nearest compounds in the training set and std_{pred} the standard deviation of the prediction

Table 5. Comparison of Binned Distance-to-Model with Binned std_{pred}

std predictions (quartiles)		distance from training set (quartiles)				
		first	second	third	fourth	all
first	*N*	*2713*	*988*	*322*	*23*	*4046*
	rmse	**0.40**	**0.48**	**0.55**	**0.55**	**0.43**
second	*N*	*1019*	*1640*	*1127*	*260*	*4048*
	rmse	**0.49**	**0.50**	**0.59**	**0.64**	**0.53**
third	*N*	*277*	*1110*	*1624*	*1042*	*4053*
	rmse	**0.63**	**0.55**	**0.61**	**0.70**	**0.62**
fourth	*N*	*55*	*332*	*997*	*2663*	*4047*
	rmse	**0.83**	**0.66**	**0.70**	**0.92**	**0.85**

Table 6 lists the results of those compounds with $MD_{ij} < 2.0$. As with the ex-clusters validation set, better 'as is' performance is noted considering compounds in the global validation set that have nearest neighbors in the library (11 461 compounds), even in the noncorrected results. Such compounds are likely to also have a nearest neighbor in the training set, and, therefore, with the relationship between the distance from the training set and the performance of the model, the rmse of those particular compounds is lower than that for the whole set of compounds (16 325 compounds). This phenomenon is also noted in the different ionization classes. Using the library correction improves upon the 'as is' results. The rmse for the entire global set with $MD_{ij} < 2.0$ decreased from 0.58 log units to 0.45 log units, while rmse's for ACIDS, BASES, NEUTRALS, and ZWITTERIONS improved to 0.53, 0.43, 0.43, and 0.48 log units,

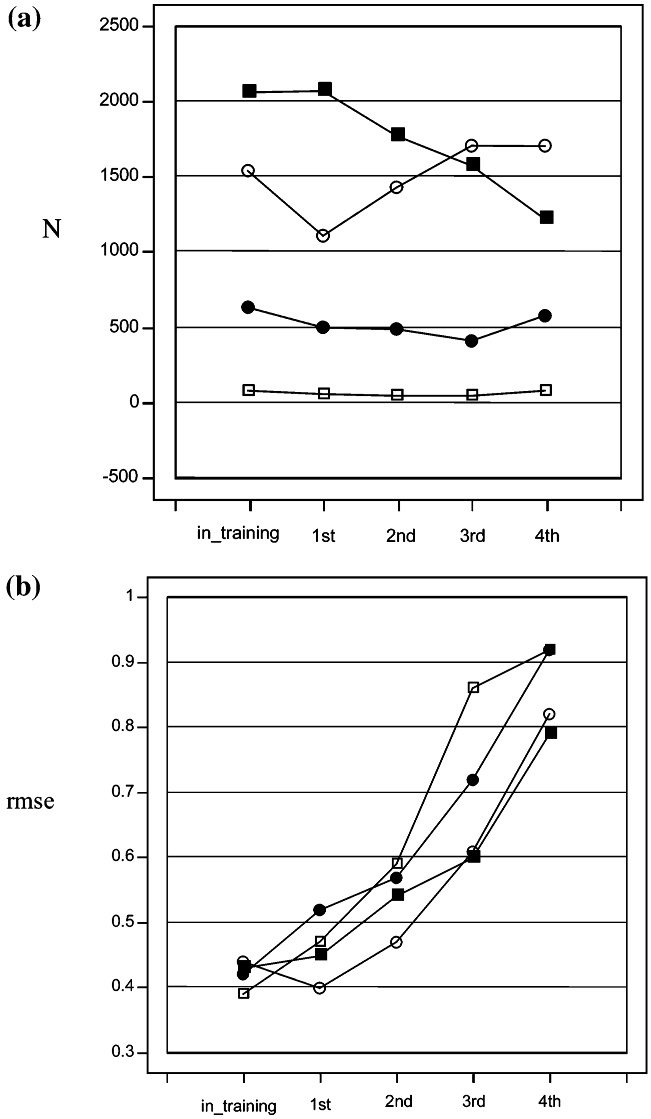

Figure 3. (a) Number of compounds in each categories according to the distance to the model. Solid circle: ACIDS; solid square: BASES; open circle: NEUTRALS; open square: ZWITTERIONS. (b) rmse of the model 'as is' for each category according to the distance to the model. Solid circle: ACIDS; solid square: BASES; open circle: NEUTRALS; open square: ZWITTERIONS

Table 6. Results of Associated Library Correction on Global Validation Set

			validation set				
			distance from training set (quartiles)				
category		in-training	first	second	third	fourth	all
all	*N*	*2956*	*4052*	*3334*	*2410*	*1665*	*11461*
	rmse (as is)	**0.37** (0.39)	**0.40**	**0.42**	**0.47**	**0.57**	**0.45** (0.58)
ACIDS	*N*	*361*	*505*	*388*	*195*	*202*	*1292*
	rmse (as is)	**0.42** (0.40)	**0.48**	**0.48**	**0.60**	**0.66**	**0.53** (0.60)
BASES	*N*	*1441*	*2068*	*1484*	*1010*	*455*	*5017*
	rmse (as is)	**0.38** (0.40)	**0.39**	**0.42**	**0.44**	**0.58**	**0.43** (0.54)
NEUTRALS	*N*	*896*	*1113*	*1162*	*995*	*809*	*4079*
	rmse (as is)	**0.34** (0.36)	**0.37**	**0.40**	**0.44**	**0.53**	**0.43** (0.59)
ZWITTERIONS	*N*	*54*	*62*	*48*	*19*	*19*	*148*
	rmse (as is)	**0.40** (0.41)	**0.47**	**0.55**	**0.36**	**0.39**	**0.48** (0.56)

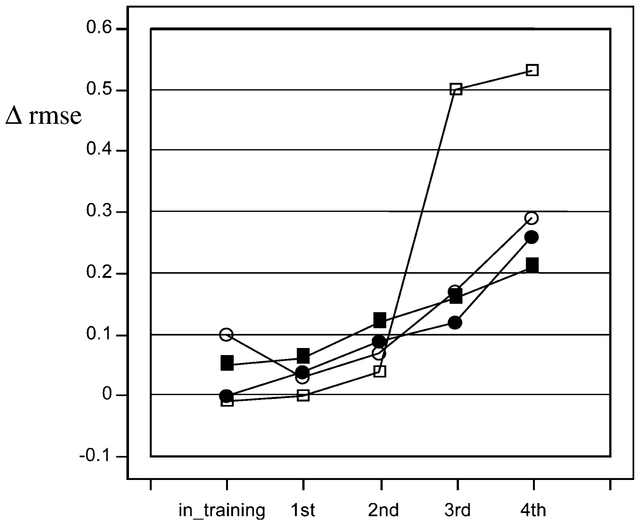

Figure 4. Difference in rmse between the model 'as is' and the model 'corrected', per categories according to the distance to the model. Solid circle: ACIDS; solid square: BASES; open circle: NEUTRALS; open square: ZWITTERION

Table 7. Percentage of Library Compounds in Various Ionization Class Subsets

	subsets of nearest in library			
subsets	same	doubtful	not same	unknown
ACIDS	98.4%	1.2%	0.1%	0.3%
BASES	99.0%	0.7%	0.0%	0.3%
NEUTRALS	97.7%	2.1%	0.0%	0.2%
ZWITTERIONS	89.6%	9.9%	0.0%	0.5%

respectively. The improvement of the rmse's via library correction can be seen in Figure 4 for the different distance categories and ionization classes. We noted that improvement is modest for compounds in or near the training set and highest for those compounds farthest from the training set. This observation is interesting because it allows one to correct the predictions of new synthesized series of compounds that may be different from the series used in model training but without the need to retrain the model.

The correction of a compound belonging to a specific ionization class is made using compounds with $MD_{ij} < 2.0$, irrespective of their ionization class. This means that a basic compound may correct an acidic compound if the two have $MD_{ij} < 2.0$. Tetko et al. have shown that setting up libraries of homogeneous ionization classes for error correction is beneficial.[1] It must be emphasized that, as explained above, both the formula used to correct predictions and the evaluation of distance between compounds are different in this study compared to those used by Tetko et al.[29,30] To check the usefulness of distinct libraries, the ionization class of the nearest compound in the library was checked. Table 7 lists the ionization classes of the nearest compound in the library for $MD_{ij} < 2.0$. In almost all cases, the nearest compound in the library to the test compound is of the same ionization class. A notable exception is the ZWITTERIONS, where 9.9% of test compounds have a nearest neighbor in the DOUBTFUL class. The DOUBTFUL classification was obtained when the two criteria used to classify a compound as acid, base, or neutral were not the same. In other words, a zwitterion needed to meet the two 'acid' and two 'base' criteria in ordered to be classified as ZWITTERIONS. This stipulation increased the chances for zwitterions to be classified as DOUBTFUL if one out of four criteria was not met. Because the criteria are not 100% perfect, it is probable that a doubtful acid is indeed an acidic compound, and most likely the correction of a compound by a member of the corresponding DOUBTFUL class is correct. It is much more worrisome if a compound is corrected with a member of the 'NOT SAME' class (e.g., an acid corrected by a member of BASES or NEUTRALS). This situation occurred only for 0.1% of the ACIDS class and never for the other classes. These results indicated that the global library could be used to correct the predictions without the need to separate data sets by ionization classes. The other indirect consequence was that one set of molecular descriptors was able to differentiate all classes.

Table 8 displays the effect of library correction using Improvement Over Naïve (ION) calculations. Again, considering $MD_{ij} < 2.0$ for compounds compared to the library in the validation set and $MD_{ij} = 0.0$ for those in the 'in-training' set. By the definitions listed above (eqs 6−9), ION gives a relative measure of improvement in error for the ionization class subsets, for 'as is' models, library corrected models, and a *logD*$_{7.4}$ calculation using just ACD calculations. Those subsets with compounds in the 'in-training' model have the greatest ION values overall which is not surprising considering their proximity to the model. Most important are the results of ION values for the library corrected subsets in the validation set. The overall model realized an 88% ION with library correction, and all four subsets saw ION in the 81−88% range. ION values for a straight ACD *logD*$_{7.4}$ calculation saw little or no improvement over naïve models.

Table 8. Improvement over Naïve Results for Global Validation Set

	in-training			validation set						
	naive	model		naive	'as is'		corrected		ACD	
subsets	mse	mse	ION	mse	mse	ION	mse	ION	mse	ION
all	1.78	0.18	**90%**	1.64	0.40	**76%**	0.20	**88%**	1.57	**4%**
ACIDS	1.57	0.18	**89%**	1.67	0.52	**69%**	0.28	**83%**	2.05	**≪0**
BASES	1.86	0.18	**90%**	1.52	0.34	**78%**	0.18	**88%**	1.09	**28%**
NEUTRALS	1.15	0.19	**83%**	1.10	0.638	**65%**	0.18	**83%**	1.27	**<0**
ZWITTERIONS	0.85	0.15	**82%**	1.19	0.47	**60%**	0.23	**81%**	2.89	**≪0**

Table 9. Results for 'Local' Modeling

		in-training	validation
ACIDS	*N*	*631*	*1983*
	rmse (ION)	**0.54 (81%)**	**0.84 (58%)**
BASES	*N*	*2039*	*6648*
	rmse (ION)	**0.34 (94%)**	**0.61 (76%)**
NEUTRALS	*N*	*1531*	*5961*
	rmse (ION)	**0.41 (85%)**	**0.60 (67%)**
ZWITTERIONS	*N*	*93*	*236*
	rmse (ION)	**0.45 (76%)**	**0.83 (42%)**

Local Models. To determine the effectiveness of this model as a general global model in comparison to class specific, or 'local' models, separate training sets of compounds comprising each ionization class were formed. Using the same compounds used in the training of the global model, the entire process of descriptor selection via ARD and final model training of those most important descriptors was performed on the four ionization class training sets. Similarly, the compounds used for the validations of the global model were used for validation of the local models. The results obtained on these models are summarized in Table 9.

The BASES model gives a better training set rmse compared to the global model, an improvement to 0.34 log units from 0.43 log units (compare with Table 4). However, this improvement is not realized in the validation set, where rmse = 0.61 log units; a mild decrease in the predictive ability of the global model (rmse = 0.58 log units). The NEUTRALS model also fared better in the training set with an rmse decrease from 0.44 log units in the global model to 0.41 log units in the local model. A slight improvement in predictive ability was noticed, with a rmse = 0.60 log units in the local model compared to 0.62 log units in the global model. Both the ACIDS and ZWITTERIONS local models showed poorer performance in both training and validation sets. It is a common feeling that 'local' models predict better the compounds of their own chemical space that 'global' models. It is not the case in this work. It is probable that, because of the reduced number of examples in the subsets, some chemical features which are present in the validation set of acids are not present in the acids and zwitterions training sets but are present in the global training set.

Interpretation of the Models. Apart from its numerous advantages, BRNNs share with other NNs a major drawback: it is not possible to evaluate the global influence of the descriptors on the property. Contrary to linear models, the descriptors (or subsets of descriptors) have various influences according to the various chemical spaces covered by the model. Due to the highly nonlinear nature of the BRNNs, it is possible that a descriptor may have a positive influence on a subset of compounds while having a negative influence on another subset. The ARD process allows the removal of descriptors which have no influence at all on the whole chemical space of the model, but a nonzero value of the sum of squared weights related to a descriptor does not indicate that a descriptor has more influence that another on the whole chemical space. In addition, a perfectly valid BRNN model could be obtained with the complete initial descriptor set. It means that the descriptors are not necessarily orthogonal either initially or after reduction by the ARD process. As a consequence, their relative importance even in the 'local' chemical space of a prediction cannot be easily established. This is a complex problem which is currently under study. During the final phase of writing this article, a paper has been published on this subject.[33]

CONCLUSIONS

The application of BRNNs with ARD toward $logD_{7.4}$ was described, and overall results suggest this to be a viable and successful modeling method. One key advantage to BRNNs was the use of ensemble neural networks to calculate a mean predicted value. The associated std_{pred} was related to distance-to-model measurements. In other words, the higher the spread of single predicted values for a compound, the farther away from the model that compound tended to be. A similar trend was noted in error calculations, whereas the farther a compound tended to be from the model, the greater its chance to have a higher error in predicted value. These trends are useful to note when applying new compounds to a predictive model, both to realize how well a compound may be predicted when not similar to the model and how well a library correction may work for that unknown compound.

Several improvements to the postmodeling results were realized by using distance-to-model analyses and prediction corrections with an associated library of compounds. The best improvements were seen using the library correction on a single global model. Further analyses to consider ionization class subsets revealed some differences in success of prediction, but library correction was also successful for these subsets. For this particular study, we noted that building separate training subsets for specific ionization classes, or 'local models', was not justified when compared to our global model.

REFERENCES AND NOTES

(1) Tetko, I. V.; Bruneau, P. Application of ALOGPS to Predict 1-Octanol/Water Distribution Coefficients, logP, and logD of AstraZeneca In-house Database. *J. Pharm. Sci.* **2004**, *93*, 3103−3110.

(2) Tetko, I. V.; Poda, G. I. Application of ALOGPS 2.1 to Predict logD Distribution Coefficient for Pfizer Proprietary Compounds. *J. Med. Chem.* **2004**, *47*, 5601−5604.

(3) Morris, J. J.; Bruneau, P. Prediction of Physical Properties. In *Virtual Screening for Bioactive Molecules*; Bohm, H. J., Schneider, G., Eds.; Wiley-VCH: Chichester, 2000; pp 33−58

(4) Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Xue-Qing; Doweyko, A.; Li Y. In Silico ADME/Tox: Why Models Fail. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 83−92.

(5) Dewitte, R. S.; Kolovanov, E. D. Predicting Molecular Physical Properties. *Biotechnol. Pharm. Aspects* **2004**, *1*, 27−52.

(6) Livingstone, D. J. Theoretical Property Predictions. *Curr. Topics Med. Chem.* **2003**, *3*, 1171−1192.

(7) PrologD. CompuDrug Chemistry Ltd. (www.compudrug.com).

(8) Advanced Chemical Development Inc., 133 Richmond Sreet West, Suite 605, Toronto, Ontario, Canada M5H 2L3 (www.acdlabs.com).

(9) Xing, L.; Glen, R. C. Novel Methods for the Prediction of logP, pK_a, and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796−805.

(10) Krejsa, C. M.; Horwath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F.; Migeon, J. C. Predicting ADME Properties and Side Effects: The BioPrint Approach. *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 4.

(11) http://www.lib.uchicago.edu/SCI/SCIpharm2004/2.2FrederiqueBarbosa.pdf.

(12) Bruneau, P. Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605−1616.

(13) JMP version 5.1.1. Distributed by SAS Institute Inc. http://www.JMP.com.

(14) Neal, R. N. Software for Flexible Bayesian Modeling, version 99-12-06. (www.cs.toronto.ca/∼radford).

(15) Winkler, D. A.; Burden, F. R. Bayesian Neural Nets for Modeling in Drug Discovery. *DDT:Biosilico* **2004**, *2*, 104−111.

(16) Burden, F. R.; Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, *42*, 3183−3187.

(17) Hawkins D. M. The problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12

(18) Hoffmann, R.; Minkin, V. I.; Carpenter B. K. Occam's Razor and Chemistry. *Bull. Soc. Chim. Fr.* **1996**, *133*, 117−130.

(19) Winkler, D. A.; Burden, F. R. Modelling Blood-Brain Barrier Partitioning Using Bayesian Neural Nets. *J. Mol. Graphics Modell.* **2004**, *22*, 499−505.

(20) Li, Y.; Campbell, C.; Tipping, M. Bayesian Automatic Relevance Determination Algorithms for Classifying Genetic Expression Data. *Bioinformatics* **2002**, *18*, 1332−1339.

(21) Enot, D. P.; Gautier, R.; Le Marouillle, J. Y. Gaussian Process: An Efficient Technique to Solve Quantitative Structure−Property Relationship Problems. *SAR QSAR. Environ. Res.* **2001**, *12*, 461−469.

(22) Burden, F. R.; Ford, M. G.; Witley, D. C.; Winkler, D. A. Use of Automatic Relevance Determination in QSAR Studies Using Bayesian Neural Nets. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423−1430.

(23) Todeschini, R.; Consonni, V.; Pavan, M. A Distance Measure Between Models: A Tool for Similarity/Diversity Analysis of Model Populations. *Chemom. Intell. Lab. Syst.* **2004**, *70*, 55−61.

(24) Xu, Y.; Gao, H. Dimension Related Distance and its Application in QSAR/QSPR Model Error Estimation. *QSAR Comb. Sci.* **2003**, *22*, 422−429.

(25) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928

(26) Strictly speaking, the Mahalanobis distance is the measure of the distance between each point i of a multidimensional cloud of points and the centroid of the cloud. It is given by $D_i = \sqrt{(y_i - \bar{y})^T S^{-1} (y_i - \bar{y})}$ where y is a vector of values for a particular point, $\bar{y}$ is the vector of means of each variable, and S is the covariance matrix of the variables.

(27) Not published.

(28) Daylight Inc., Mission Viejo, California, USA.http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

(29) Tetko, I. V. Neural Network studies. 4. Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717−728.

(30) Tetko, I. V.; Tanchuk V. Y. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136−1145.

(31) Tiño, P.; Nabney, I. T.; Williams, B. S.; Lösel, J.; Sun, Y. Nonlinear Prediction of Quantitative Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1647−1653.

(32) A measurement is more likely repeated if an apparently abnormal result is obtained. Thus, high ranges in duplicated measurements do not indicate a global variability of the experimental methodology.

(33) Yang, L.; Wang, P.; Jiang, Y.; Chen, J. Studying the Explanatory Capacity of Artificial Neural Networks for Understanding Environmental Chemical Quantitative Structure−Activity Relationship Models. *J. Chem. Inf. Comput. Sci.* ASAP, Web release October 13, 2005.

CI0504014