Random Forest Models To Predict Aqueous Solubility

David S. Palmer, Noel M. O'Boyle,[†] Robert C. Glen, and John B. O. Mitchell*

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge,
Lensfield Road, Cambridge CB2 1EW, United Kingdom

Received May 5, 2006

Random Forest regression (RF), Partial-Least-Squares (PLS) regression, Support Vector Machines (SVM), and Artificial Neural Networks (ANN) were used to develop QSPR models for the prediction of aqueous solubility, based on experimental data for 988 organic molecules. The Random Forest regression model predicted aqueous solubility more accurately than those created by PLS, SVM, and ANN and offered methods for automatic descriptor selection, an assessment of descriptor importance, and an in-parallel measure of predictive ability, all of which serve to recommend its use. The prediction of log molar solubility for an external test set of 330 molecules that are solid at 25 °C gave an $r^2 = 0.89$ and RMSE = 0.69 log S units. For a standard data set selected from the literature, the model performed well with respect to other documented methods. Finally, the diversity of the training and test sets are compared to the chemical space occupied by molecules in the MDL drug data report, on the basis of molecular descriptors selected by the regression analysis.

INTRODUCTION

The majority of drugs that enter development never reach the marketplace. The cost of research and development is such that by 2000 the annual expenditure of U.S. pharmaceutical companies had reached 26 billion dollars. Of the costs incurred in research and development, approximately 75% have been attributed to failures of potential drug molecules.[1] Therefore there has been much interest in industry in the development of in silico tools to guide drug discovery.

The factors which influence the transit of a drug in the body are termed the pharmacokinetic phase and are commonly described by the acronym ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity). In silico screens have been developed for many of the ADMET properties. Many of these are based upon Quantitative-Structure−Property-Relationships (QSPR), which attempt to relate the physical property of interest to descriptors calculable from a computer representation of the molecule. By their nature, most of these in silico screens are statistical models whose validity is dependent upon the data from which they are derived and the methods used and assumptions made during their construction. Therefore they are frequently improved as new data and methods become available.

In this study, we develop models to predict aqueous solubility. The solubility of a potential drug candidate is an important determinant of its bioavailability. From the solubility the rate of dissolution may also be calculated by models such as those of Hamlin et al.[2] Solubility is defined as the concentration of solute in a saturated solution under equilibrium conditions. However, the solubility of an organic molecule in aqueous media may depend on temperature, pH, counterions, impurities, and the polymorphic form of the solute. We develop a QSPR model for the prediction of (thermodynamic) aqueous solubility at 25 °C.

Of the QSPR models documented in the literature, many are derived from data sets that contain both molecules that are solids and liquids at room temperature. The process of solvation of a crystalline organic molecule can be decomposed into three steps: (1) breakdown of the crystal lattice; (2) cavity formation in the solvent; and (3) solvation of the molecule. However, for the dissolution of a liquid the first step corresponds to overcoming the liquid−liquid interactions, whereas for dissolution of a crystalline solid this corresponds to overcoming the lattice interactions.[3] Therefore in this study we have investigated the prediction of solubility for a data set of molecules which are all solid at room temperature. The use of a data set which contains solids makes it necessary in principle to consider polymorphic form, as many molecules can crystallize to form different crystal structures. For organic molecules, these polymorphs are normally found to have similar lattice energies, rarely differing by more than 10 kJ mol^{-1}. The polymorphic form is not considered explicitly in this work because no information about crystal form was available for the molecules in the data set. It is hoped that the effect of lattice energy on solubility may be accounted for indirectly during model building. This idea might be supported by the work of Ouvrard and Mitchell,[4] who demonstrated that it is possible to predict the lattice energy from simple molecular descriptors. Examples from the literature where the intermolecular interaction energy in the solid phase has been explicitly considered in modeling solubility are the work of Yalkowsky[5] and of Abraham.[6]

The methods employed in deriving solubility models can be grouped by the molecular descriptors selected, the method of regression, and the diversity and constitution of the data set. The last property is difficult to assess because many

* Corresponding author phone: +44-1223-762983; fax: +44-1223-763076; e-mail: jbom1@cam.ac.uk.
† Current address: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, U.K.

RANDOM FOREST MODELS

J. Chem. Inf. Model., Vol. 47, No. 1, 2007 **151**

QSPR papers have provided little information about the data set they employ. One exception is the work of Bergstrom et al.,[7] who used the Chemography method[8] to assess the diversity of their data set. The methods of regression used for solubility modeling can be separated into linear and nonlinear methods. The question of which is more appropriate has been addressed, and for a single data set linear methods were selected.[9] However, the trend in the literature is for nonlinear methods, especially Neural Networks, to give better validation statistics. Lind et al. demonstrated the use of a Support Vector Machine with a Tanimoto similarity kernel.[10] The problem with both Support Vector Machines and Neural Networks is that they are often 'black boxes'; it can be difficult to interpret the importance of individual descriptors to the model. In addition, both methods can be time-consuming to train, which may be a problem as a common occurrence in QSPR is the need to retrain the model as new or more diverse data become available. Finally, the models can be separated into those that employ either 2D descriptors or both 2D and 3D descriptors. Perhaps surprisingly, the best models based on 2D information often outperform those that also incorporate 3D descriptors. It is possible that the bulk properties encoded by 2D descriptors are more important to solubility. Another possibility is that 3D descriptors may suffer from inaccuracies related to the selection of a single 3D molecular conformer. The most significant descriptor in solubility prediction is usually the calculated logarithm of the octanol−water partition coefficient, which is calculated from a 2D representation of a molecule. Three recent review articles have discussed the prediction of solubility from structure.[11−13]

In this paper, we develop Random Forest models for aqueous solubility from a large data set of molecules that are solid at room temperature. We also investigate models developed with Partial-Least-Squares, a Support Vector Machine with a radial-basis-function kernel, and Artificial Neural Networks. Sheridan et al. have demonstrated that dissimilarity between training and test sets correlates with experimental error.[14] To assess the relative diversities of our training and test sets, we compare these to the chemical space occupied by the molecules in the MDL Drug Data Report (MDDR).[15]

THEORY

Random Forest. Random Forest is a method for classification and regression which was introduced by Breiman and Cutler.[16] Recent studies have suggested that Random Forest offers features which make it very attractive for QSPR studies.[17] These include relatively high accuracy of prediction, built-in descriptor selection, and a method for assessing the importance of each descriptor to the model. The theory of Random Forest regression is discussed in the papers of Svetnik et al.[17] and on the Web site of Breiman and Cutler.[18] The method is based upon an ensemble of decision trees, from which the prediction of a continuous variable is provided as the average of the predictions of all trees.

In RF regression, an ensemble of regression trees is grown from separate bootstrap samples of the training data using the CART algorithm.[19] The branches in each tree continue to be subdivided while the minimum number of observations in each leaf is greater than a predetermined value. Unlike

regression trees, the branches are not pruned back. Furthermore, the descriptor selected for branch splitting at any fork in any tree is not selected from the full set of possible descriptors but from a randomly selected subset of predetermined size. There are three possible training parameters for Random Forest: **ntree** − the number of trees in the Forest; **mtry** − the number of different descriptors tried at each split; and **nodesize** − the minimum node size below which leaves are not further subdivided.

The bootstrap sample used during tree growth is a random selection with replacement from the molecules in the data set. The molecules that are not used for tree growth are termed the 'out-of-bag sample'. Each tree provides a prediction for its out-of-bag sample, and the average of these results for all trees provides an in situ cross-validation called the 'out-of-bag validation'.

Random Forest includes a method for assessing the importance of descriptors to the model. When each descriptor is replaced in turn by random noise, then the resulting deterioration in model quality is a measure of descriptor importance. The deterioration in model quality can be assessed by the change in mean-square-error for the out-of-bag validation.

Software for Random Forest regression is available from the Web site of Breiman and Cutler or as part of the Random Forest package in the statistical computing environment, R.[20]

Support Vector Machines. A Support Vector Machine is a kernel-based method for classification, regression, and function approximation. A thorough discussion of the theory of Support Vector Machines is provided by Cristianini and Shawe-Taylor.[21] Lind et al. have previously described the use of a Support Vector Machine with a Tanimoto kernel for the prediction of aqueous solubility.[10] Here we discuss the use of a Support Vector Machine with a radial basis function kernel.

Ant Colony Optimization Algorithm. An Ant Colony Optimization (ACO) algorithm is employed to select a subset of variables for regression by Partial-Least-Squares, Support Vector Machines, and Artificial Neural Networks. Variable selection in QSPR modeling is the problem of selecting an optimum subset from a large number of descriptors so that the best model is formed. The brute force method of selection would be to build every possible model and select the best one. As this is not feasible, it is common practice to use an algorithm to guide the search. In this paper, an Ant Colony Optimization algorithm is used because recent evidence has suggested that they may converge on the optimum solution in a shorter period of time than genetic algorithms.[22]

Ant Colony Optimization algorithms are methods for solving discrete optimization problems which were introduced by Dorigo.[23] The algorithm is modeled upon the behavior of real ant colonies whose members exhibit stigmergy—a form of indirect communication mediated by modifications of the environment. A common example of stigmergy is the sight of ants walking in lines. During the search for food, each ant lays a trail of pheromone between nest and food source that is later followed by other ants with a probability dependent upon the amount of pheromone laid down. As each ant lays additional pheromone upon the trail, the process becomes autocatalytic so that the Ant Colony exhibits self-organizing behavior.

We implemented an Ant Colony Optimization algorithm in R, based on the paper of Shen et al.[22] It was observed that for the fitness function of Shen et al. the term which penalizes the inclusion of extra descriptors swamps the term which measures the quality of the fit. Therefore their algorithm tends to select a very small model. For this reason, the Shen fitness function was replaced by the RMSE for 10-fold cross-validation.

MATERIALS AND METHODS

Data Set. The data set was selected from the Handbook of Aqueous Solubility[24] and two sources from the literature.[25,26] Molecules were selected on the basis that they were organic compounds, structurally diverse and solid at room temperature. For some of the molecules in the Handbook of Aqueous Solubility, more than one experimental measurement is provided. In these cases, the molecule was ignored if the reported values were considered to be inconsistent. Otherwise the molecule was accepted, and a mean value for the molar aqueous solubility was taken. The data set was randomly partitioned into a training set of 658 molecules and a test set of 330 molecules. The range of log molar solubility values (moles of solute per liter) in the training set was from −10.41 to 1.58 with a mean value of −3.42 and a standard deviation of 2.13. The distribution of molecular weights was from 59.1 to 511.6 amu, with a mean of 244.0 and a standard deviation of 90.1. For the test set, the range of log molar solubility values was from −9.16 to 1.09 with a mean value of −3.11 and a standard deviation of 2.06. The distribution of molecular weights was from 60.1 to 588.6 amu with a mean of 235.0 and a standard deviation of 89.9. A diversity analysis is provided for the training and test sets at the end of the Experimental Section.

Statistical Testing. For each model that was tested, three statistics are reported; these are the squared correlation coefficient (r^2), the Root-Mean-Square-Error (RMSE), and the bias, and they are defined in eqs 1−3. A parenthesis nomenclature is adopted to indicate whether the relevant statistic refers to the training set (tr), the 10-fold cross-validation inside the training set (CV), the Random Forest out-of-bag validation inside the training set (oob), or the prediction of the test set (te).

$$r^2 = 1 - \sum_{i=1}^{n}(y_i^{obs} - y_i^{pred})^2 / \sum_{i=1}^{n}(y_i^{obs} - y^{obs,mean})^2 \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{obs} - y_i^{pred})^2} \quad (2)$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(y_i^{obs} - y_i^{pred}) \quad (3)$$

All models were derived by regression against the molecules in the training set. The test set was not used in model building. The 10-fold cross-validation results were selected as the standard by which to compare Random Forest results with other models. The out-of-bag validation is convenient for Random Forest models, due to the bootstrap method of data selection which they utilize. It is not appropriate for our PLS, SVM, or ANN models, which are not based on bootstrap sampling. Although, in principle, a similar data selection and cross-validation technique could be combined with PLS, SVM, or ANN models, this is very rarely done in practice.

Structure Optimization and Descriptors. The data set was converted from SMILES to 3D structures using CON-CORD in SYBYL6.9.[27] All further calculations were carried out in the Molecular Operating Environment (MOE).[28] First, explicit hydrogen atoms were added using the MOE Wash function. The structures were then optimized using the MMFF94 force field before the force field partial charges were replaced by PM3 partial charges. All descriptors were calculated from this single conformer.

The program MOE supports the calculation of more than 200 molecular descriptors (excluding fingerprints), of which 126 2D descriptors and 36 3D descriptors were calculated here. This corresponds to all descriptors that do not require semiempirical Quantum Mechanical calculations or an external alignment step. The 2D descriptors included calculated physical properties (logP, molar refractivity), charged-surface properties (from Gasteiger−Marsili PEOE charge distributions on VDW surfaces), constitutional descriptors (counts of atoms and functional groups), connectivity and topological indices (including the chi, Kier-Hall, Kappa, Wiener, and Balaban indices), and pharmacophore feature counts (numbers of hydrogen bond donors and acceptors). The 3D descriptors included energy terms (total potential energy and contributions of angle bend, electrostatic, out-of-plane, and solvation terms to the molecular mechanics force-field energy), molecular shape descriptors (water-accessible surface areas), volume descriptors, and surface area descriptors. The models were first selected from 2D descriptors only and then from a combined set of both 2D and 3D descriptors.

Random Forest. Random Forests were trained using the randomForest[29] library in the statistical computing environment, R. The Random Forest was trained upon all the calculated 2D descriptors. Optimization of training parameters was performed using R scripts which iteratively changed each parameter one-by-one and regenerated the regression model. The optimum value of each parameter was selected from the following ranges: **mtry** − from 1 to 126; **ntree** − from 1 to 5000; **nodesize** − from 1 to 50. To assess the quality of the model, the fit to the training data and the out-of-bag validation statistics were considered. The fit to the training data was high in all models with $r^2 \sim 0.98$ and RMSE(tr) ~ 0.30 log S units, which demonstrates the ability of the model to learn the information in the training set. However this is not an indication of predictive ability, and so the out-of-bag cross-validation results were used to guide the model building process. The Random Forest was observed to be reasonably insensitive to training parameters, so that variation of **mtry** between 40 and 126, of **ntree** from 250 upward, and of **nodesize** in the region 5−10 had little effect on the cross-validation results. When **mtry** becomes very small, not enough descriptors are considered at each split, and hence the predictive quality of each tree decreases. The exact value of **mtry** below which a decrease in predictive error is observed will depend on the number and relative importance of descriptors present in the data set. As the **nodesize** is increased above the optimum value, the size and range of solubilities in the terminal nodes of the trees
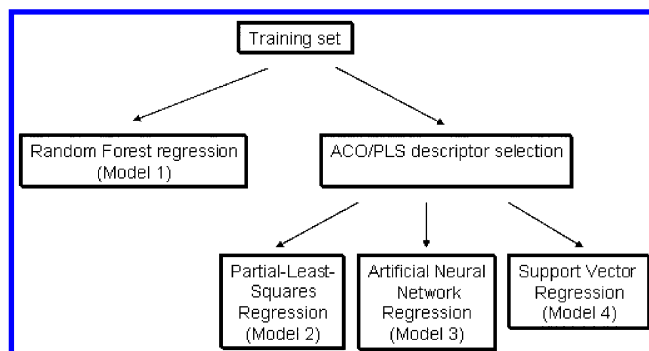
RANDOM FOREST MODELS

J. Chem. Inf. Model., Vol. 47, No. 1, 2007 **153**

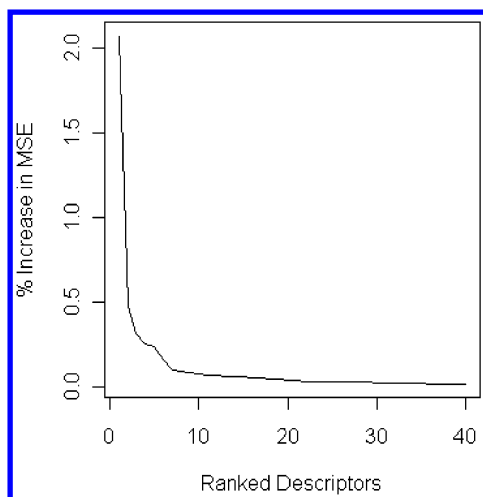**Figure 1.** Derivation of four different regression models.

Figure 2. Increase in mean-square-error when a single descriptor in the Random Forest model is replaced by random noise; shown for the top 40 descriptors as ranked in order of importance.

increases and the predictive accuracy decreases. When the value of **ntree** is decreased too far, the results deteriorate significantly; if **ntree** reaches one, the Random Forest becomes a single unpruned Regression Tree. If the number of trees in the forest is increased above the optimum, there is a general increase in computational expense, but the results do not improve significantly. For this data set an optimum value is reached at 500 trees, though a somewhat smaller value around 250 carries only a very minor cost in accuracy and would be more appropriate for work with larger data sets where CPU time is a limiting factor. Based upon this analysis the parameters were selected with values of **mtry** = 42, **nodesize** = 5, and **ntree** =500.

Descriptor Selection for Random Forest. Random Forest incorporates a method for descriptor selection. However the selection process means that irrelevant descriptors may be incorporated into a very small proportion of the trees generated (as a consequence of the sampling procedure) or may be included in the Forest but without being used in the final model. Therefore efforts were made to remove irrelevant descriptors. Descriptor importance was assessed by replacing each descriptor in turn by random noise and observing the increase in the Mean-Square-Error (MSE) for the out-of-bag validation. When the descriptors were then sorted by relative importance and plotted against the increase in MSE, a graph similar in appearance to a scree plot was obtained (Figure 2). Random Forests were retrained upon smaller subsets of descriptors by pruning away the least important descriptors.

Partial-Least Squares, Support Vector Machines, and Artificial Neural Networks. The most important descriptors for the Random Forest model may not be the best subset for other regression models (which will be discussed). Indeed it was found that better results were obtained for all three methods when they were trained upon a subset of descriptors selected by an Ant Colony Optimization algorithm and Partial-Least-Squares regression. Linear-PLS is widely used as the regression method to select descriptors for nonlinear predictors.[9,30] It may, however, miss descriptors which have strongly nonlinear relationships with the variable to be predicted. Some attempts were made to use the Ant Colony Algorithm for direct selection of descriptors using a SVM and an ANN as the regression method. However, as no improvement was observed, these results are not reported. The ACO/PLS procedure is similar to the widely used GA/PLS method.[30,31] Recent studies suggest that an ACO-based method may converge on optimum solutions in less time.[22]

The ACO/PLS selection was made from all 126 2D descriptors with 50 ants, a pheromone evaporation parameter of 0.7, and an RMSE(CV) fitness function. The selected descriptors were then used as input for Partial-Least-Squares regression and two different machine learning methods (i) a Support Vector Machine for epsilon regression and (ii) a multilayer perceptron feed-forward backward-propagation neural network. The method for descriptor selection for all models is summarized in Figure 1.

The SVM training parameters (epsilon, cost, gamma) were selected by a systematic search, using the *tune* function in the e1071 library in R, so as to minimize the mean-square-error of prediction for an internal hold-out test set. The search was carried out in the following ranges: epsilon − 0.01 to 1; cost − 1 to 20; and gamma − 0.05 to 0.1, with a variable step size reduced from coarse to fine near the global minimum. The SVM with the optimum training parameters was then used to carry out 10-fold cross-validation for all molecules in the training data set, so as to provide an internal estimate of prediction accuracy, before being retrained against all of the training data in order to provide a prediction for molecules in the test set. The optimum training parameters were epsilon = 0.1, cost = 3, and gamma = 0.08.

An artificial neural network was trained by a systematic search method so as to optimize the prediction for an internal hold-out test set. Multilayer feed-forward backward-propagation neural networks were used, and both single and double hidden layer architectures were investigated. The systematic search was carried out in the following ranges: number of neurons in each hidden layer, from 3 to 12; number of iterations, from 50 to 300; learning rate, from 0.004 to 0.016; and momentum, from 0.1 to 1.1. The search was repeated for both a least-mean-squares (LMS) and a least-mean-logarithm-squared (LMLS) error criterion. The best network had a 12−6−1 architecture and was trained over 150 epochs with a learning rate of 0.01, momentum of 0.5, and a LMLS error criterion.

Diversity Analysis. QSPR's are empirical models and as such will be most successful in making predictions for molecules which are similar to the training set. To investigate the diversity of our solid-only data set, the MDDR was selected as an example of the chemical space occupied by druglike organic molecules.

154 *J. Chem. Inf. Model., Vol. 47, No. 1, 2007*

PALMER ET AL.

Table 1. Twelve Descriptors Selected by the ACO/PLS Procedure

molecular descriptor	
SlogP	octanol−water partition coefficient
SMR	molar refractivity
PEOE_RPC-1	relative negative partial charge: the smallest negative atomic partial charge divided by the sum of the negative atomic partial charges
PEOE_VSA_FPOL	fractional polar van der Waals surface area. The sum of the van der Waals surface area of atoms which have an absolute partial charge greater than 0.2 divided by the total surface area.
PEOE_VSA_FNEG	fractional negative van der Waals surface area
TPSA	total polar surface area
a_acc	number of hydrogen bond acceptors
a_don	number of hydrogen bond donors
weinerPol	Weiner polarity index
a_aro	number of aromatic atoms
b_rotR	fraction of rotatable bonds
chi1v_C	carbon valence connectivity index (order 1)

Figure 3. The 12 most important descriptors for a particular Random Forest, as measured by increase in MSE on replacement with random noise. Slight permutations occur in the order of descriptors below vsa_hyd when different Forests are trained with the same data.

The MDDR contains some molecules that cannot be processed by our software. Therfore, prior to the diversity analysis, some standard filters were applied to the MDDR so as to remove those entries that contained counterions, one of each pair of duplicated molecules and all entries that contained elements other than C, H, N, O, S, P, F, Cl, Br, and I. The final MDDR data set contained 108 298 molecules.

For these 108 298 molecules, the 2D descriptors selected for the regression analysis (Table 1) were calculated. A Principal Component Analysis (PCA) was carried out on all scaled and centered descriptors, and the rotation matrix was then used to predict the scores for the MDDR data set and for the QSPR training and test data sets. The distributions of the significant PC scores for the MDDR data set and the QSPR training and test sets were compared. The method is similar to laying down a map (the MDDR) and then locating places (the molecules in the QSPR data sets) within it. In this respect it is similar to the Chemography method of Oprea.[8] A benefit of using the MDDR in conjunction with diversity analysis is that if a region of chemical space is identified which is not represented in the QSPR data sets, then it is trivial to identify molecules from the MDDR which do occupy this region.

The MDDR analysis illustrates the diversity of our data set with respect to a larger region of chemical space. To provide a clear example of the relative similarity between structures in the training and test sets, a separate PCA was carried out based upon the molecules in the training data set. Figure 6 illustrates the factor scores for both training and test data sets.

RESULTS

Random Forest Regression. The first Random Forest model was trained upon all 126 2D descriptors, and the out-of-bag cross-validation results were used to select the optimum training parameters (**mtry** = 42, **ntree** = 500, and **nodesize** = 5). However, the optimum value of **mtry** will depend on the total number of descriptors. Therefore it is more easily expressed as the fraction of the total number of descriptors (in this case, the total number of descriptors divided by three). For the training set the Random Forest
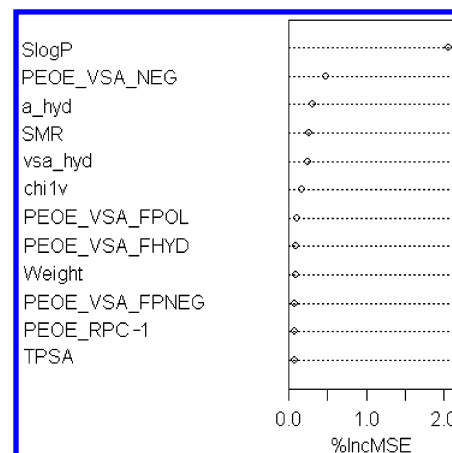
was able to explain a large proportion of the variance in the log molar solubility values giving r^2(tr) = 0.98, RMSE(tr) = 0.28, and bias(tr) = 0.007. The fit of the model to the training data is not an estimate of the predictive ability of the model, and so the out-of-bag validation results were examined, for which r^2(oob) = 0.89, RMSE(oob) = 0.69, and bias(oob) = 0.017. The 10-fold cross-validation results were also calculated so as to provide a comparison with those for PLS, SVM, and ANN models, the results were r^2(CV) = 0.896, RMSE(CV) = 0.685, and bias(CV) = 0.010.

The importance of each descriptor to the model was assessed. Figure 2 is a plot of the descriptors ranked by their importance (only the first 40 descriptors are shown). Figure 2 demonstrates that few of the descriptors contribute significantly to the model. The Random Forest was iteratively retrained, and each time the five least important descriptors were removed. However, no improvement to the prediction error could be made. For example, a Random Forest trained upon the 40 most important descriptors gave a fit to the training data set which was identical to that trained upon all descriptors (r^2(tr) = 0.98, RMSE(tr) = 0.29, and bias(tr) = 0.000 and r^2(oob) = 0.89, RMSE(oob) = 0.695, and bias(oob) = 0.005). In fact, there are a number of reasons why it is unnecessary to remove irrelevant descriptors from RF models. First, Random Forest contains a method for descriptor selection. Second, neither the training of the Forest nor the calculation of the 2D descriptors is computationally expensive, and hence there is no need to define a subset in order to reduce computing time. Third, due to intercorrelations between descriptors the exact order of importance of descriptors may change, and hence the subset selected by our method may not be the unique solution. For this reason, we select the Random Forest model for which all 2D descriptors are available.

When Random Forest training was repeated 10 times, the five most important descriptors were the same in each model and were (in order of importance) SlogP, PEOE_VSA_NEG, a_hyd, SMR, and vsa_hyd. The order of the following descriptors varied slightly between each model; however, the positions 6−12 were always occupied by a subset of PEOE_VSA_FHYD, chi1v, PEOE_VSA_FPOL, PEOE_VSA_FPNEG, Weight, TPSA, PEOE_RPC-1, and
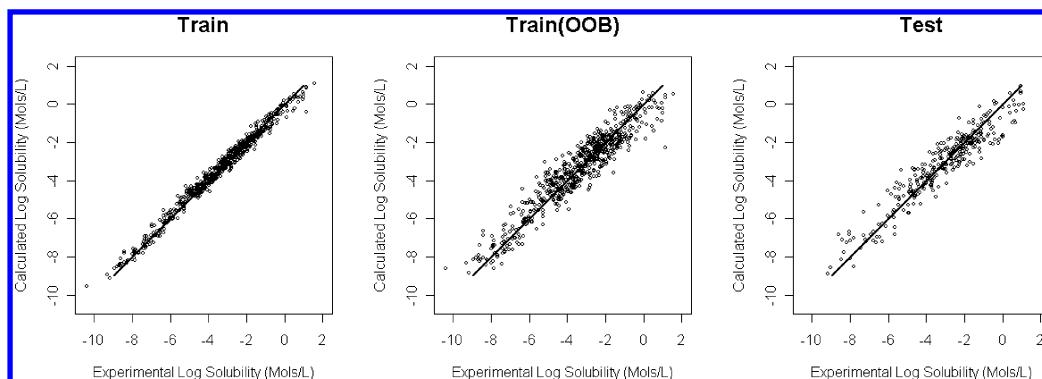
RANDOM FOREST MODELS

J. Chem. Inf. Model., Vol. 47, No. 1, 2007 **155**

Figure 4. Calculated versus experimental log molar solubility for the best Random Forest model for the training and test sets.
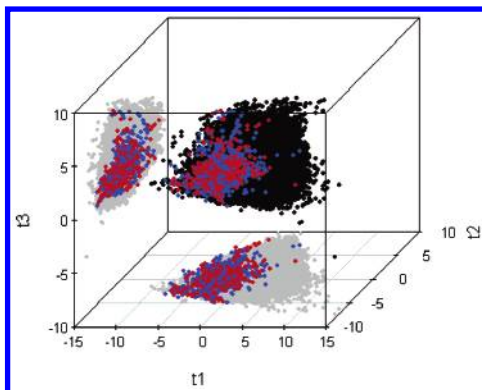

Figure 5. The first three principal component scores for the MDDR data set (black), the training data set (blue), and the test data set (red).

chi0v, and the top 25 ranked descriptors always belonged to a subset of 32 descriptors.

Partial-Least-Squares. The descriptors selected by the ACO/PLS procedure are given in Table 1.

When the PLS model was trained upon the complete training set, the results were as follows: r^2(tr) = 0.873, RMSE(tr) = 0.760, bias(tr) = 0.000, r^2(CV) = 0.856, RMSE(CV) = 0.787, and bias(CV) = 0.001.

Support Vector Machines. The descriptors selected by the ACO/PLS procedure were used as input for an epsilon regression Support Vector Machine with a radial basis function kernel. The statistics for 10-fold cross-validation inside the training set were r^2(CV) = 0.880, RMSE(CV) = 0.726, and bias(CV) = 0.001. The SVM was trained based upon 478 support vectors, and all descriptors were scaled and centered prior to calculation.

Neural Networks. The best network had a 12−6−1 architecture and was trained over 150 epochs of adaptive gradient descent with a least-mean-logarithm-squared error criterion, a learning rate of 0.01, and a momentum of 0.5. For 10-fold cross-validation inside the training set, r^2(CV) = 0.864, RMSE(CV) = 0.742, bias(CV) = −0.020.

External Validation. Table 2 contains the results for all four different methods. Based upon the cross-validation results for all four models, it can be established that the Random Forest performs better than the SVM, ANN, and PLS models (in order of increasing predictive error). The same is true for the prediction of the external test set. Figure 4 shows the correlation between calculated and experimental log solubility values for the RF model. The Random Forest model was able to predict the log molar solubility values for the molecules in the test set with r^2(te) = 0.89,

RMSE(te) = 0.69, and bias(te) = 0.05. The ability of the Random Forest to predict values for molecules not contained within the training set, in conjunction with the 10-fold and out-of-bag cross-validation statistics, suggests that the model is useful for the prediction of the aqueous solubility of as yet unsynthesized molecules.

3D Descriptors. It was hoped that the inclusion of 3D descriptors would further improve the model. Of particular interest were the solvent-accessible-surface-area (SASA) descriptors, which are expected to have a strong correlation with the energy for cavity formation in the solvent (as exemplified by the use of solvent-accessible-surface areas for the empirical prediction of cavity energy in molecular modeling programs[32]). The Random Forest was retrained upon all 2D and 3D descriptors with the training parameters of **mtry** = 61, **nodesize** = 5, and **ntree** = 500. The validation statistics were found to be similar to those for the 2D model. For the out-of-bag validation r^2(oob) = 0.89, RMSE(oob) = 0.694, and bias(oob) = 0.01. When the descriptor importance was assessed, it was found that few 3D descriptors were selected within the top 40 most important descriptors. Although SASA was among the higher ranked descriptors, the failure to improve the regression model demonstrates that similar information is encoded in the 2D descriptors. For this reason, the 2D model was selected in preference to that which included 3D descriptors.

Comparison between Different Models. The Random Forest and ACO selected models contain different subsets of descriptors. The observation is not surprising given that many of the descriptors from which the models are selected have high pairwise correlations. There is therefore some redundancy in the descriptor set and hence a variety of solutions of similar merit.

Eight of the twelve descriptors selected by the ACO/PLS procedure are also within the top 25 ranked descriptors in the Random Forest model (SlogP, SMR, PEOE_RPC-1, PEOE_VSA_FPOL, PEOE_VSA_FNEG, TPSA, a_acc, and weinerPol), but the remaining four (a_don, a_aro, b_rotR, and chi1v_C) do not contribute significantly to the model. The corollary of this is that there are many descriptors (PEOE_VSA_NEG, a_hyd, vsa_hyd, PEOE_VSA_FHYD, Weight, etc.) which appear to be important to Random Forest but which were not selected by the ACO algorithm. The observation that the subset of descriptors that are important for Random Forest is larger than that for the other methods (see Figure 2) is misleading. The predictive ability of Random Forest is not affected by the presence of correlated descriptors, and therefore none were removed prior to
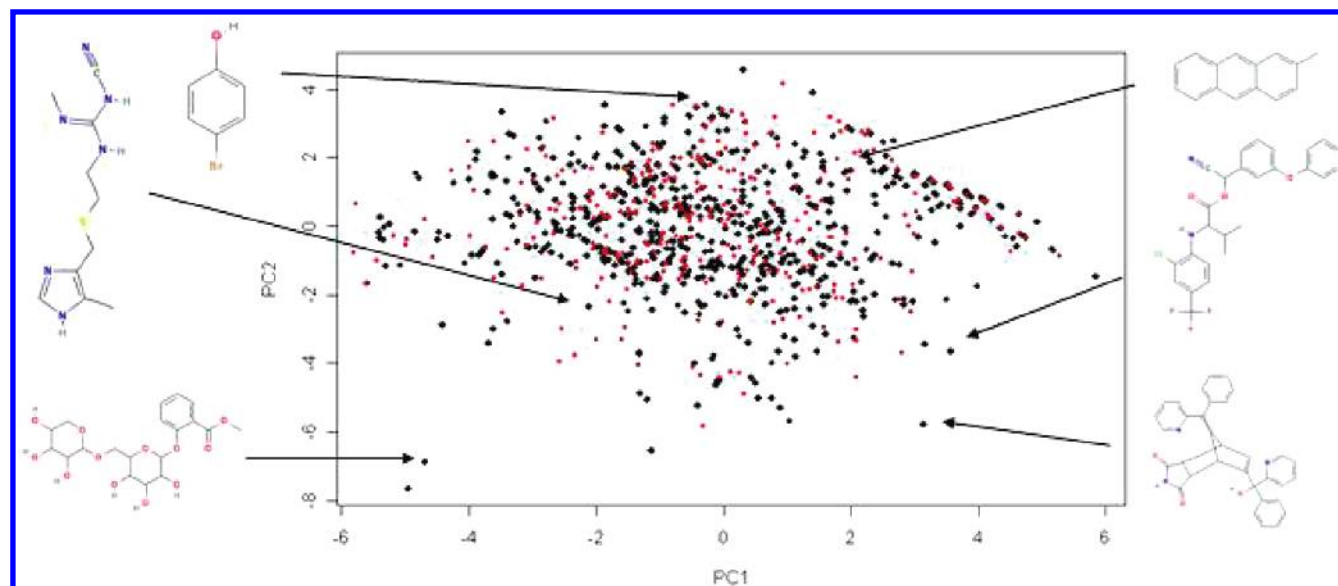
Figure 6. Factor score plot for molecules in the training set (black) and test set (red).

Table 2. Results for 10-Fold Cross-Validation Inside the Training Set and for Prediction of the External Test Set

model	descriptor selection	r^2(CV)	RMSE (CV)	bias(CV)	r^2(te)	RMSE(te)	bias(te)
PLS	ACO/PLS	0.856	0.787	0.001	0.859	0.773	−0.337
ANN	ACO/PLS	0.864	0.742	−0.020	0.866	0.751	0.081
SVM	ACO/PLS	0.880	0.726	0.001	0.878	0.720	−0.038
RF		0.896	0.685	0.010	0.890	0.690	0.050

analysis. However, there will be an effect on the number of descriptors that are assessed as being important. An artificial example of this is provided if the most important descriptor is duplicated in the data set. As the descriptors are identical they will have (nearly) identical measures of importance and will both occur in the list of important descriptors. The list will therefore be one descriptor longer, but the predictive ability does not change.

Diversity Analysis. The diversity of the sample selected for the QSPR training (and test) data sets will affect how well the model generalizes to unseen data. An ideal QSPR data set would be an evenly distributed sample of the population (organic chemical space).

The principal component scores for the MDDR data set were compared to those predicted for the QSPR training and test data sets. The scores derived from the six principal components with the highest eigenvalues were considered; these explain 92.5% of the variance in the MDDR data set. The largest difference between the chemical space occupied by the MDDR and that occupied by the training and test sets is observed for the first PC score and is shown in Figure 5. There are a large number of molecules in the MDDR with a PC1 score greater than 2.5 which are not represented by the QSPR data set. Analysis of the MDDR revealed that there are 12406 molecules in the MDDR which fall into this region. However, there is a strong correlation between PC1 score and molecular size, and it is found that the mean molecular weight for these molecules is 585.54. Molecules from the QSPR data set that do occupy this region of chemical space are norbormide (Figure 6, lower right-hand side), a calcium channel entry blocker, which is used as a rat poison, and etoposide, an antitumor agent which inhibits the enzyme topoisomerase II. By contrast, some of the molecules in the training and test data sets appear to be on

the fringes of the chemical space occupied by the MDDR. Examples include glucose and fructose.

With the exceptions mentioned above, the data set is not localized within a small volume of chemical space but is structurally diverse. Furthermore, the training and test sets occupy similar regions of chemical space; therefore, the external validation is generally an interpolative prediction. However it should be noted that the data set is not a perfect sample of druglike space. In addition to some druglike molecules the data set also contains some agrochemicals and some nondruglike molecules.

Comparison with Other Studies. The direct comparison of QSPR models that use different data sets is difficult as the reported statistics will depend on the size, diversity, and constitution of the test set. A comparison between our model and others in the literature is further complicated by the fact that it is not always clear which studies have used data sets that contain molecules which are liquid at room temperature. As an example, Lind et al. used a SVM equipped with a Tanimoto kernel in order to predict the solubility of an external test set of 412 molecules with r^2(te) = 0.89 and RMSE(te) = 0.68.[10] However this test data set was almost identical to the test set of Huuskonen et al.[25] and is known to contain both liquids and solids. Cheng et al. used multilinear regression based upon 2D molecular descriptors to derive a model which was able to predict log molar solubility in the range of 0.7−1 log S units for four different test sets.[1] However it is not possible to assess whether these test sets contained molecules which were liquids at room temperature.

The prediction of solubility from structure has been the subject of three recent reviews.[11−13] More than 90 different models are discussed in these reviews, and no best model can easily be identified. However, it is clear that for the

RANDOM FOREST MODELS

J. Chem. Inf. Model., Vol. 47, No. 1, 2007 **157**

Table 3. Comparison between Our Random Forest Model and Five Different Models from the Literature for a Data Set That Contains Both Liquids and Solids

models	size of training set	size of test set	method of regresssion	r^2(te)	SD
Random Forest	1033	258	RF	0.92	0.58
Random Forest	884	413	RF	0.92	0.59
Random Forest	797	496	RF	0.91	0.61
Liu[33]	1033	258	ANN	0.86	0.71
QMPR+[33]	1033	258	ANN		0.93
Huuskonen[25]	884	413	ANN	0.92	0.60
Lind[10]	884	413	SVM	0.89	0.68
Tetko[35]	879	412	ANN	0.92	0.60
Yan[34]	797	496	ANN	0.92	0.59

prediction of solubility for druglike molecules a minimum standard error of prediction in the range of 0.5–1 log S unit should be expected, which is comparable to that demonstrated by the Random Forest model.

In the Introduction we made the assumption that the molecular descriptors might be able to account for some of the influence of the lattice energy on solubility during the process of model building. In the same way, the descriptors might account for the cohesive forces present in liquids should these molecules be present in the training set. Furthermore, we were interested in comparing our method to those models which had employed mixed phase data sets. For this reason the Huuskonen data set was selected from the literature. The Random Forest model was retrained upon the data set of Huuskonen et al. Table 3 provides a comparison of the results of this study with six other methods which have used the Huuskonen data set. Unfortunately since each study used different permutations of training and test sets, it was only possible to replicate exactly the data sets selected by Liu et al.[33] (from which the results for QMPR+ are also taken). To approximate the data sets of Yan et al.,[34] Tetko et al.,[35] Huuskonen et al., and Lind et al., a variant of k-fold cross-validation was used. Model building was repeated five times for randomly selected training and test sets of the correct size, and the reported statistics are the mean r^2(te) and the mean standard deviation.

Solid or Liquid. The Huuskonen data set contains both molecules that are liquid and molecules that are solid at room temperature, and therefore it was interesting that the model performed well in the tests shown in Table 3. To investigate this further, the Huuskonen data set was partitioned into a data set of 491 molecules which were known to be liquids at room temperature and 744 molecules known to be solids. Random Forest was used to analyze the data sets. For the group of liquids the model was able to explain most of the variance in the data set with r^2(tr) = 0.990, RMSE(tr) = 0.161, and bias(tr) = 0.002. For out-of-bag validation, r^2(oob) = 0.931, RMSE(oob) = 0.405, and bias = 0.005. This was a better result than obtained for the solid only portion of the Huuskonen data set, for which the best model reported r^2(tr) = 0.985, RMSE(tr) = 0.264, bias(tr) = 0.000, r^2(oob) = 0.910, RMSE(oob) = 0.653, and bias = 0.002.

DISCUSSION

Model Building. The QSPR methods for predicting solubility which are documented in the literature can be roughly divided into those which employ linear regression methods (MLR, PLS, and PCR) and those which employ

machine learning methods (ANN and SVM). With one notable exception,[9] the majority of studies that use machine learning methods have reported better validation statistics than those that use linear methods. In those papers, the most commonly used machine learning method is the Artificial Neural Network, which has been used in commercial as well as academic models. Here we have demonstrated that Random Forest regression has performed better than Neural Networks for this data set. We have also developed a model from a Support Vector Machine with a radial basis function kernel which has proven to be useful for the prediction of aqueous solubility. There are other reasons why Random Forest might be of greater use than SVMs or ANNs. First, Random Forest is easier to train as it includes a descriptor selection procedure and is not strongly dependent upon training parameters. Random Forests are immune to the problems of overfitting common to ANNs and SVMs. Second, descriptor importance can be assessed in Random Forest, which aids in model interpretation.

The inclusion of 3D descriptors did not improve the models. A possible explanation is that strong correlations were observed between some 2D and 3D descriptors. The Pearson correlation coefficient (r) between the 2D Molar Refractivity and the 3D Water-Accessible-Surface Area was 0.95 and that between the 2D and 3D calculated VDW volume was 1.00 (2 decimal places). However this does not explain why 3D descriptors did not replace 2D descriptors in some models. It is also possible that the use of a single conformation to generate 3D descriptors limits their usefulness.

Solids or Liquids. The models developed in the literature have often focused upon mixed-phase data sets which contain many simple organic molecules. However, prediction of the solubility of larger solid-phase molecules which contain many functional groups is often of more interest; an example would be the importance of solubility to the pharmaceutical industry. Therefore we have focused upon developing models for a large data set of molecules that are solid at room temperature. However, we have also provided a comparison with the literature for a data set which contains both solids and liquids. Taking the RMSE(oob) as a guide of the predictive accuracy of the models, the models prepared for the liquid-only data set show higher predictive accuracy than those for data sets containing solids. Some tentative conclusions can be drawn from this observation. The simplest conclusion would be that models that contain liquids in the training and test sets will report better validation statistics; an important point when comparing different QSPR models. A survey of the literature would support this conclusion. However we caution against concluding from these data that the reason solids cannot be modeled as accurately as liquids is that the molecular descriptors do not account for the lattice energy of the crystal. The liquid only data set contains a large number of simple organic molecules and is therefore less diverse than the data set of solids. Furthermore, it is reasonable to propose that experimental errors may be higher for solubility measurements made for solid (crystalline) materials which may exhibit polymorphism or which may form solvates. In this work, the effect of the crystal form on solubility was not considered explicitly, but rather it was hoped that it would be accounted for indirectly during the modeling process.

CONCLUSION

A method for the prediction of aqueous solubility has been developed from a structurally diverse data set based upon 2D molecular descriptors. The method has been shown to perform comparably to other studies in the literature, including some that require 3D structure calculation. Our method is unique in that it uses Random Forest regression, which we have shown performs better than models built by Support Vector Machines or Artificial Neural Networks for this data set. Furthermore, the results confirm that the predictive statistics for solubility QSPR models will depend on the phase adopted by the molecules in the data set at the experimental temperature.

FUTURE WORK

The solubility data were selected from the literature. Although care was taken in excluding inconsistent data, the experimental error will be similar to other studies which use literature data. Clark et al.[36] have estimated this experimental error to be in the region of 0.6 log S units. Therefore to improve the models described here, new experimental data are required. Work has begun in our laboratory to make new measurements of solubility for druglike molecules. For each molecule we are also measuring pK_a's and melting points and characterizing the crystal structure by X-ray crystallography.

ACKNOWLEDGMENT

Supporting Information Available: Training and test data sets. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Cheng, A.; Merz, K. M., Jr. Prediction of Aqueous Solubility of a Diverse Set of Compounds Using Quantitative Structure−Property Relationships. *J. Med. Chem.* **2003**, *46*, 3572−3580.

(2) Hamlin, W. E.; Northam, J. I.; Wagner, J. G. Relationship between in vitro dissolution rates and solubilities of numerous compounds representative of various chemical species. *J. Pharm. Sci.* **1965**, *54*, 1651−1653.

(3) Delgado, E. J.; Alderete, J. B.; Matamala, A. R.; Jaña, G. A. On the Aggregation State and QSPR Models. The Solubility of Herbicides as a Case Study. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 958−963.

(4) Ouvrard, C.; Mitchell, J. B. O. Can we predict lattice energy from molecular structure? *Acta Crystallogr.* **2003**, B*59*, 676−685.

(5) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Structure by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.

(6) Abraham, M. H.; Le, J. The Correlation and Prediction of the Solubility of Compounds in Water using an Amended Solvation Energy Relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.

(7) Bergstrom, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Arturssen, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1477−1488.

(8) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3*, 157−166.

(9) Catana, C.; Gao, H.; Orrenius, C.; Stouten, P. F. W. Linear and Nonlinear Methods in Modeling the Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Model.* **2005**, *45*, 170−176.

(10) Lind, P.; Maltseva, T. Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855−1859.

(11) Dearden, J. C. In silico prediction of aqueous solubility. *Expert Opinion on Drug Discovery* **2006**, *1*, 31−52.

(12) Delaney, J. S. Predicting aqueous solubility from structure. *Drug Discovery Today* **2005**, *10*, 289−295.

(13) Johnson, S. R.; Zheng, W. Recent Progress in the Computational Prediction of Aqueous Solubility and Absorption. *AAPS J.* **2006**, *8*, E27−E40. See http://www.aapsj.org (accessed July 21, 2006).

(14) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. 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.

(15) MDL Drug Data Report database. Available from MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, U.S.A. http://www.mdli.com (accessed July 21, 2006).

(16) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(17) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modelling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(18) Random Forest website. See http://www.stat.berkeley.edu/∼breiman/RandomForests/cc_home.htm (accessed July 21, 2006).

(19) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman & Hall/CRC: Boca Raton, 1984.

(20) R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0, URL http://www.R-project.org (accessed July 21, 2006).

(21) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: 2000.

(22) Shen, Q.; Jiang, J.-H.; Tao, J.-c.; Shen, G.-l.; Yu, R.-Q. Modified Ant Colony Optimization Algorithm for Variable Selection in QSAR Modeling: QSAR Studies of Cyclooxygenase Inhibitors. *J. Chem. Inf. Model.* **2005**, *45*, 1024−1029.

(23) Dorigo, M.; Stützle, T. *Ant Colony Optimisation*; MIT Press: Cambridge, MA, 2004.

(24) Yalkowsky, S. H.; He, Y. *The Handbook of Aqueous Solubility Data*; CRC Press LLC: Boca Raton, 2003.

(25) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.

(26) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.

(27) SYBYL6.9 Tripos Inc., 1699 Hanley Road, St. Louis, MO 63144. http://www.tripos.com (accessed July 21, 2006).

(28) MOE. Chemical Computing Group Inc., Montreal, Quebec, Canada, 2002. http: //www.chemcomp.com (accessed July 21, 2006).

(29) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, 2, 18−22.

(30) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Prediction of the Isoelectric Point of an Amino Acid Based on GA-PLS and SVMs. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 161−167.

(31) Rogers, D. R.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(32) Hasel, W.; Hedrickson, T. F.; Still, W. C.; A Rapid Approximation to the Solvent Accessible Surface Areas of Atoms. *Tetrahedron Comput. Method.* **1988**, *1*, 103−116.

(33) Liu, R.; So, S.-S. Development of Quantitative Structure−Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633−1639.

(34) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429−434.

(35) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.

(36) Clark, T. *Does quantum chemistry have a place in cheminformatics?* *Molecular Informatics: Confronting Complexity*; Hicks, G. M., Kettner, C., Eds.; Proceedings of the Beilstein-Institut Workshop, 2002. See: http://www.beilstein-institut.de/bozen2002/proceedings/ (accessed July 21, 2006).

CI060164K