Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P

Laura D. Hughes,† David S. Palmer, Florian Nigsch, 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 August 16, 2007

This paper attempts to elucidate differences in QSPR models of aqueous solubility (Log S), melting point (T_m), and octanol−water partition coefficient (Log P), three properties of pharmaceutical interest. For all three properties, Support Vector Machine models using 2D and 3D descriptors calculated in the Molecular Operating Environment were the best models. Octanol−water partition coefficient was the easiest property to predict, as indicated by the RMSE of the external test set and the coefficient of determination (RMSE = 0.73, $r^2 = 0.87$). Melting point prediction, on the other hand, was the most difficult (RMSE = 52.8 °C, $r^2 = 0.46$), and Log S statistics were intermediate between melting point and Log P prediction (RMSE = 0.900, $r^2 = 0.79$). The data imply that for all three properties the lack of measured values at the extremes is a significant source of error. This source, however, does not entirely explain the poor melting point prediction, and we suggest that deficiencies in descriptors used in melting point prediction contribute significantly to the prediction errors.

INTRODUCTION

In the past 20 years, the pharmaceutical industry has seen the number of candidate drug molecules dramatically increase, led by developments in combinatorial chemistry, high throughput screening, and robotics. With the vast number of possible lead molecules, medicinal chemists have a greater need to prioritize which compounds to pursue. Consequently, pharmacokinetic information is being used at a much earlier stage, as drug efficacy becomes inconsequential if the drug is unable to reach the intended site of action. To that end, the prediction of pharmacokinetic properties from structure alone remains an important goal. Quantitative structure−property relationships (QSPR) offer one method for the prediction of properties related to pharmacokinetics.

Despite development over the past century, however, the quality of QSPR models varies substantially, based on the property to be predicted. The comparison between prediction of intrinsic aqueous solubility (Log S), melting point (T_m), and octanol−water partition coefficient (Log P), three properties of pharmaceutical importance, provides an interesting exercise. While Log P values can be predicted very accurately, melting point values have yet to be predicted with comparable error. Aqueous solubility prediction lies somewhere between Log P and melting point prediction; it has been predicted better than T_m but not as well as Log P. The relative ease of predicting Log S compared to T_m is somewhat surprising, given that melting point can be measured accurately, whereas reported solubility measurements are known to be laden with errors. The question then becomes: why are there differences in the predictive ability of models for different properties? Furthermore, how are these properties interrelated, and can we attribute the errors in the models to limitations in the data, the methods, or the descriptors?

The differences in the predictive ability of models for different properties could be due to three factors. First, the differences could be attributed to the accuracy of the data; since the model learns information from the data, the experimental error provides a limit on the accuracy of the empirical model. Additionally, if the data set lacks chemical diversity, the model will be unable to predict molecules outside of the region covered by the training data. We will refer to these issues as a failure of the *data*. Second, the descriptors could insufficiently describe the properties of the molecules. For instance, Clark and others have hypothesized that the failure of melting point prediction models can be attributed to the lack of crystal packing information in descriptors.[1−3] We will call this a failure of the *descriptors*. Third, the method of modeling, partial least-squares (PLS), Random Forest (RF), *k*-nearest neighbor (kNN), Support Vector Machine (SVM), or others, could incompletely learn information from the data. Alternatively, the methods could learn too much from the data, leading to overfitted models that have limited applicability to molecules outside the training set. This becomes a failure of the *methods*.

This study seeks to answer these questions by developing models for Log S, T_m, and Log P using drug and druglike molecules. By using a single data set with experimental values for each of these three properties, we attempt to elucidate the relationships between solubility, melting point, and octanol−water partition coefficient. Moreover, we endeavor to explain the differences in QSPR models of these properties.

* Corresponding author phone: +44 (0)1223 762983; fax: +44 (0)-1223 763076; e-mail: jbom1@cam.ac.uk.

† Present address: Department of Chemistry, Stanford University, 333 Campus Drive, Mudd Building, Room 121, Stanford, CA 94305-5080.

10.1021/ci700307p CCC: $40.75
Published on Web 01/11/2008

BACKGROUND

Aqueous Solubility Prediction. Aqueous solubility represents, to a first approximation, the absorption of a drug molecule into the body. Poorly soluble drugs are associated with low oral bioavailability; however, molecules that are too soluble (i.e., too polar) often have difficulty crossing lipid bilayers to reach their intended site of action. Consequently, most drugs have Log S values between −5 and −1,[4] which represents a balance between aqueous solubility and membrane permeability.

The prediction of aqueous solubility tends to use four types of descriptors: methods using melting point and Log P, atom or group contributions, physicochemical and quantum chemical descriptors, and topological indices. Numerous models of Log S have been presented in the recent literature. This brief summary will focus on the prediction of drug and druglike molecules; Dearden presents a more thorough review of recent models.[5]

Methods using melting point and Log P are best exemplified by the general solubility equation (GSE; eq 1).

$$\text{Log S} = 0.5 - 0.01(T_m - 25) - \text{Log P} \qquad (1)$$

The GSE incorporates solution-phase interactions via the octanol−water partition coefficient and implicit solid-state interactions through the melting point term. While the GSE provides a simple way to calculate solubility, it relies on empirical melting point data, limiting its utility in early drug development. Despite this fundamental limitation, however, it has been used as a first approximation for the solubility of drugs with known melting points. Yalkowsky et al. report absolute average errors (AAE) of 0.4−0.6 log molar solubility units for test sets containing environmentally and pharmaceutically relevant compounds.[6−9] However, others have estimated higher errors, with root mean squared errors (RMSE) approaching 1 log molar solubility unit for predominately druglike sets.[10] In particular, Delaney noticed that the AAE almost doubles when the GSE is used to predict the solubility of a class of high molecular weight, druglike molecules (MW = 300−400), compared to prediction of a low MW set.[11]

Physicochemical and quantum chemical descriptors are more commonly used for Log S prediction. Calculated Log P descriptors are frequently included in models of aqueous solubility, often as one of the most important descriptors.[12−14] Bergström et al. have also found hydrophobicity, hydrophilicity, flexibility, electron distribution, and charge to play important roles in prediction.[13] Most methods use linear methods such as multiple linear regression or PLS or nonlinear methods such as artificial neural networks (ANN). In general, nonlinear methods appear to provide better predictions,[12,15,16] although Catana et al. have found that PLS outperformed several nonlinear methods for their data set.[17] RMSEs for models based on physicochemical and quantum chemical descriptors tend to range from approximately 0.7 log units[12] to 1 log unit[13] for the prediction of an independent validation set containing a large number of drugs. Since the standard deviation in experimental solubility data has been estimated to be about 0.6 log molar solubility units,[18] reported errors below this value are likely to be the result of overfitted models.

Melting Point Prediction. Melting point is typically first used in preliminary compound identification and purity analysis. Additionally, melting point has been shown to play a role in aqueous solubility,[8] boiling point,[19] and eutectic composition[20] prediction. Even though melting point can be measured accurately, its prediction has been a notoriously difficult problem. Reasonably accurate models have been built for small subgroups of compounds;[21] however, relatively few models exist for the prediction of druglike molecules. Past studies have emphasized the role of flexibility (or, conversely, rigidity), size, and intra- and intermolecular interactions in determining melting point. Bergström et al. predicted the melting points of a set of 92 drugs, trained on 185 compounds, using electrotopology descriptors and PLS and report a RMSE of 49.8 °C with an r^2 value of 0.54.[1] Karthikeyan et al. used the Bergström data set as a validation set for their ANN model, trained on a selection of compounds from the Molecular Diversity Preservation International (MDPI) database.[2] On a selection of molecules from the MDPI not used in training, they report a RMSE of 50.4 °C with an r^2 value of 0.645. Surprisingly, their prediction of the Bergström data set was much better (RMSE = 41.4 °C, r^2 = 0.662). The improvement in statistics with the ANN model may reflect the advantage of nonlinear methods over linear methods for T_m modeling of drugs. Nigsch et al. in our laboratory used the kNN method to predict the melting points of a set of diverse organic molecules and a set of drug molecules.[22] For the prediction of 80 drug molecules based on a training set of 197 drugs, the RMSE was 46.3 °C with a squared correlation coefficient of 0.30. The predicted values in the model were significantly worse at the extremities; compounds with high melting points were typically underpredicted, and compounds with low melting points were overpredicted. These errors are probably due to the limited number of compounds in the training set with high or low melting points.

Octanol−Water Partition Coefficient Prediction. Octanol−water partition coefficient is a key physicochemical property for the prediction of absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of drug candidates. It is strongly correlated with experimental aqueous solubility[23] and has been associated with blood brain barrier penetration,[24] as only small lipophilic molecules are able to pass into the central nervous system.[25] Similarly, Log P values have been used to predict both dermal and ocular partitioning of drugs.[24] Log P values are widely used to predict pharmacokinetics in Lipinski's Rule of 5, as compounds with Log P values greater than 5 tend to have poor absorption or permeation.[26] Moreover, Log P values can be measured quickly using high throughput screening,[27] increasing their utility in early ADMET prediction. Hansch et al. pioneered work on Log P prediction in the 1960s, proposing that Log P values could be segmented into contributions from molecular fragments.[23] As such, each molecular fragment would have a characteristic contribution to the Log P value, and the overall Log P value could be found by summing the contributions of the individual fragments. This additive-constitutive hypothesis led Chou and Jurs to develop ClogP, a widely used Log P prediction program.[28] To correct for group interactions, ClogP uses a number of correction factors based on experimental data. As such, bonding, branching, multiple halogenation, and proximity effects are corrected

to account for "surprise interactions",[28] allowing more accurate prediction of structural isomers and druglike molecules. These and other Log P prediction programs have been used to predict druglike molecules with standard errors which are as small as 0.3−0.4 log units in favorable cases,[29,30] though values in the approximate range 0.6−1.2 have been found for more challenging data sets.[31,32] In general, the prediction of simple organic molecules is easier than the prediction of drugs, and the methods often fail if the compound to be predicted has fragments with low representation in the training set of the model.

Prediction of Multiple Properties. Several previous studies have incorporated prediction of multiple properties.[3,33−35] While these studies use identical methods to model the properties, they do not use identical data sets. As such, comparisons between the different properties become complicated; differences between models of the properties could be due to differences in the size or composition of the data set. Our work is better equipped to elucidate the relationship between Log S, T_m, and Log P by using a consistent data set and uniform modeling methods. Furthermore, our models allow us to draw several conclusions about the prediction of Log S, T_m, and Log P for drug and druglike molecules.

Modeling Methods. In this work, we determine the best model for each property based on four different modeling methods: partial least-squares, Random Forest, *k*-nearest neighbor, and Support Vector Machine. PLS, also known as projection to latent structures, generates orthogonal latent variables from the input descriptors and attempts to maximize the correlation between these latent variables and the response variable; a regression model is then built from the latent variables. Random Forest builds an ensemble of regression trees using the CART (**Classification and Regression Trees**) algorithm. kNN models use distance matrices to compare unknown molecules to a training set; the unknown is then predicted to have an experimental value of the average of the most similar compounds, i.e., its "nearest neighbors". SVM uses a kernel function to map the descriptors from a lower dimensional space to a higher dimensional one, using a linear or nonlinear target function. A linear learning function determines a relationship between the modified descriptors and the response variable. Since we calculated a large number of descriptors for this work, a subset of these descriptors was selected using an Ant Colony Optimization (ACO) algorithm, as recent work suggests that ACOs efficiently select descriptors.[12,36]

METHODS

Compilation of Data Set. Experimental aqueous solubility data were taken from data sets compiled by Bergström,[13,14] Rytting,[37] and Wassvik.[10] These data sets were selected for their high drug and druglike content and for the quality of the Log S measurements. If provided, melting point data were also taken from these sources. Additional T_m data[8,29,38−41] and Log P data[8,29,30,41−43] were taken from the literature; multiple values for an experimental property were averaged. If a melting range was given, the highest value was used, since an observed range, in contrast to a sharp melting point, may indicate a small quantity of impurity, which would lower the observed melting point such that the top of the range is below the melting point of the pure substance. Additionally, aqueous solubility measured after 72 h was used if solubility at both 24 and 72 h was assayed. Compounds in the original data set that decomposed upon melting or lacked either melting point or Log P data were removed from the data set. The final data set includes 287 molecules (see the Supporting Information for structures and experimental data).

Calculation of Descriptors. Structures of the molecules were downloaded from the PubChem database in SDF format, verified using SciFinder Scholar,[40] and imported into the Molecular Operating Environment (MOE, 2007.05 release).[44] The built-in MOE function was used to add hydrogen atoms, and chirality was explicitly defined according to the PubChem structures. Structures were energy minimized in MOE with MMFF94x using the "Rebuild 3D" option, preserving chirality and calculating forcefield partial charges. A total of 168 2D MOE descriptors were calculated, including atom and bond counts, Gasteiger−Marsili partial charge descriptors,[45] connectivity indices (including Kier−Hall and Balaban), adjacency and distance matrices, calculated physical properties (including octanol−water partition coefficients, aqueous solubility, and molar refractivity), and pharmacophore features (including acidic and basic residues, hydrogen bond donors and acceptors, and polar surface area). Additionally, 53 3D MOE descriptors were calculated, including potential energy, semiempirical quantum mechanical energy, conformation-dependent charge, and surface area, volume, and shape descriptors. An additional 346 functional group counts, atom-centered fragments, and geometrical descriptors were calculated based on SYBYL MOL2 structures using Dragon (version 5).[46] The experimental data and the 567 descriptors were then imported into the statistical package R (version 2.4.1).[47]

Evaluation of Models. Models were evaluated on their RMSE (eq 2), AAE (eq 3), r^2 (eq 4), and bias (eq 5)

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_{i,\mathrm{pred}})^2}{n}} \tag{2}$$

$$\mathrm{AAE} = \frac{\sum_{i=1}^{n} |y_i - y_{i,\mathrm{pred}}|}{n} \tag{3}$$

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - y_{i,\mathrm{pred}})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{4}$$

$$\mathrm{bias} = \frac{\sum_{i=1}^{n} (y_i - y_{i,\mathrm{pred}})}{n} \tag{5}$$

where n is the number of molecules in the prediction set, y_i is the experimental value, $y_{i,\mathrm{pred}}$ is the predicted value, and $\bar{y}$ is the average experimental value. Four different RMSEs will be reported in the text: $\mathrm{RMSE}_{\mathrm{train}}$, the RMSE of the training

set; $RMSE_{CV}$, the RMSE of a 10-fold cross-validation of the training set; $RMSE_{test}$, the RMSE of the internal test set; and $RMSE_{ext}$, the RMSE of the external test set. In the entire analysis, no compounds were removed as outliers, as removing any compound would compromise the integrity of the data set. Apparent outliers, defined as any compound with an absolute prediction error greater than twice the $RMSE_{ext}$, were examined to determine where the model failed.

Descriptor Processing and Selection. Figure 1 outlines the processing steps and descriptor selection. The data set was first randomly divided into three sets: a training set (150 molecules), an internal test set (50 molecules), and an external test set (87 molecules). The external test set was not used in any model building. Consequently, this set serves as an independent test set used to evaluate the generality of the predictive power of each model. All descriptors were then mean-centered and scaled to unit variance based on the mean and variance of the training set.

MOE-calculated Log P descriptors were removed for models predicting Log P, and the MOE-calculated Log S descriptor was removed for models predicting Log S. Descriptors with the same value for more than 90% of the molecules in the data set were removed (Figure 1a). Finally, we removed correlated descriptors, as they are redundant and can degrade model performance for some methods (Figure 1b). First, pairs of descriptors that were more than 95% correlated with each other were identified. For each pair, two PLS regressions were built on the entire set of descriptors with each of the correlated descriptors removed in turn. We then retained the descriptor whose removal led to the greater loss in predictive power. In sum, 126 2D MOE descriptors, 44 3D MOE descriptors, and 86 Dragon descriptors formed the five sets of descriptors.

Descriptor selection was performed using an ACO algorithm (Figure 1c). The algorithm was originally implemented in our laboratory in the R programming language by Dr. Noel O'Boyle,[12] based on a paper by Shen et al.[36] For each set of descriptors and each experimental property, two duplicate ACO models were run. The $RMSE_{CV}$ using PLS was used as the fitness function to guide the ACO toward the global minimum. Each ACO used 25 ants, 250 iterations (N), and a pheromone evaporation rate of 0.7. The descriptors selected by each ant were recorded for the first iteration, every other iteration from $N = 10$ to $N = 200$, and the last iteration. With 25 ants per iteration, a total of 2450 combinations of descriptors were recorded. Since variable selection through ACO tends to produce overfitted models, the internal test set was used to evaluate the 2450 different combinations of descriptors. The three sets of descriptors that produced the three lowest $RMSE_{test}$ were selected for further model building.

Model Building. In this study, four parameters were varied: the predicted property, the method of descriptor calculation, the descriptor selection method, and the regression method (Table 1). Models using 3D MOE descriptors alone were not built, based on previous work that suggests that 3D MOE descriptors alone do not perform well.[12,13,22] For each combination of parameters, PLS, RF, kNN, and SVM models were built.

PLS models were generated in R using the **pls** package.[48] The optimal number of latent variables (**ncomp**) included in each model was determined by varying **ncomp** from 1 to 20. The number of components that yielded the lowest $RMSE_{test}$ was used in the final model that predicted the values of the external test set.

The **randomForest** package in R was used to build the RF models.[49] Since Random Forest models have been previously shown to be insensitive to changes in the parameters,[12,50,51] the default settings for **ntree** and $\mathbf{m_{try}}$ were used (**ntree** $= 500$, $\mathbf{m_{try}} = {}^1/_3 \times$ number of descriptors).

The kNN models for this work were built according to a protocol by Nigsch.[22] Briefly, Euclidean distances were calculated between molecules in the training set. The number of nearest neighbors was varied, with k taking the values 1, 5, 10, or 15. Additionally, the contributions from nearest neighbors were weighted by an arithmetic average, geometric average, inverse distance average, or an exponential scheme. The optimal value for k and the best weighting scheme were determined by the RMSE of the combined training and internal test sets. For kNN, models were built using only ACO-selected descriptors. Preliminary results indicated that kNN models built using all descriptors performed significantly worse than models using ACO-selected descriptors, and thus we did not further pursue development of these models.

SVM models were trained using the **e1071** package in R.[52] For all SVM models, we selected the Radial Basis Function kernel, as Palmer et al. have previously shown this kernel to be useful for modeling aqueous solubility.[12] The *gamma*, *epsilon*, and *cost* parameters were first coarsely optimized and then refined using the **tune.svm** function in R. In the coarse optimization step, *gamma* was varied between 0 and 1 in increments of 0.05; *epsilon* was varied between 0 and 2 in increments of 0.05; and *cost* was varied between 0 and 10 in steps of 0.1. The best set of parameters corresponded to the lowest $RMSE_{test}$. Each model was then refined by sampling the area around the best parameters from the coarse optimization: coarse *gamma* ± 0.05, in steps of 0.005; coarse *epsilon* ± 0.05, in steps of 0.005; and coarse *cost* ± 0.1, in steps of 0.01. The parameters used to build the final SVM models were those resulting in the lowest $RMSE_{test}$.

RESULTS

Solubility Prediction. The best model for aqueous solubility was a SVM model using a combination of 41 MOE 2D and 3D descriptors (see the Supporting Information). For brevity, only the best model for each property will be discussed; comparisons between the regression methods are presented in the Discussion. The SVM model had an $RMSE_{ext}$ of 0.900 log units, an r^2 value of 0.79, and a bias of -0.065 (Table 2). As illustrated by r^2, this model represents a significant improvement over prediction based on the mean. "Prediction based on the mean" refers to a rudimentary model, where all the values in a given test set are predicted to be the same value, namely the mean of the training set. As such, prediction based on the mean provides a measure of the improvement through modeling. For prediction based on the mean, the $RMSE_{ext}$ is 2.02; thus, the SVM model represents a 55.4% reduction in the $RMSE_{ext}$. A plot comparing the predicted Log S values to the experimental values is given in Figure 2.

Even though the $RMSE_{ext}$ is within an acceptable margin (under 1 log unit), there were several compounds whose

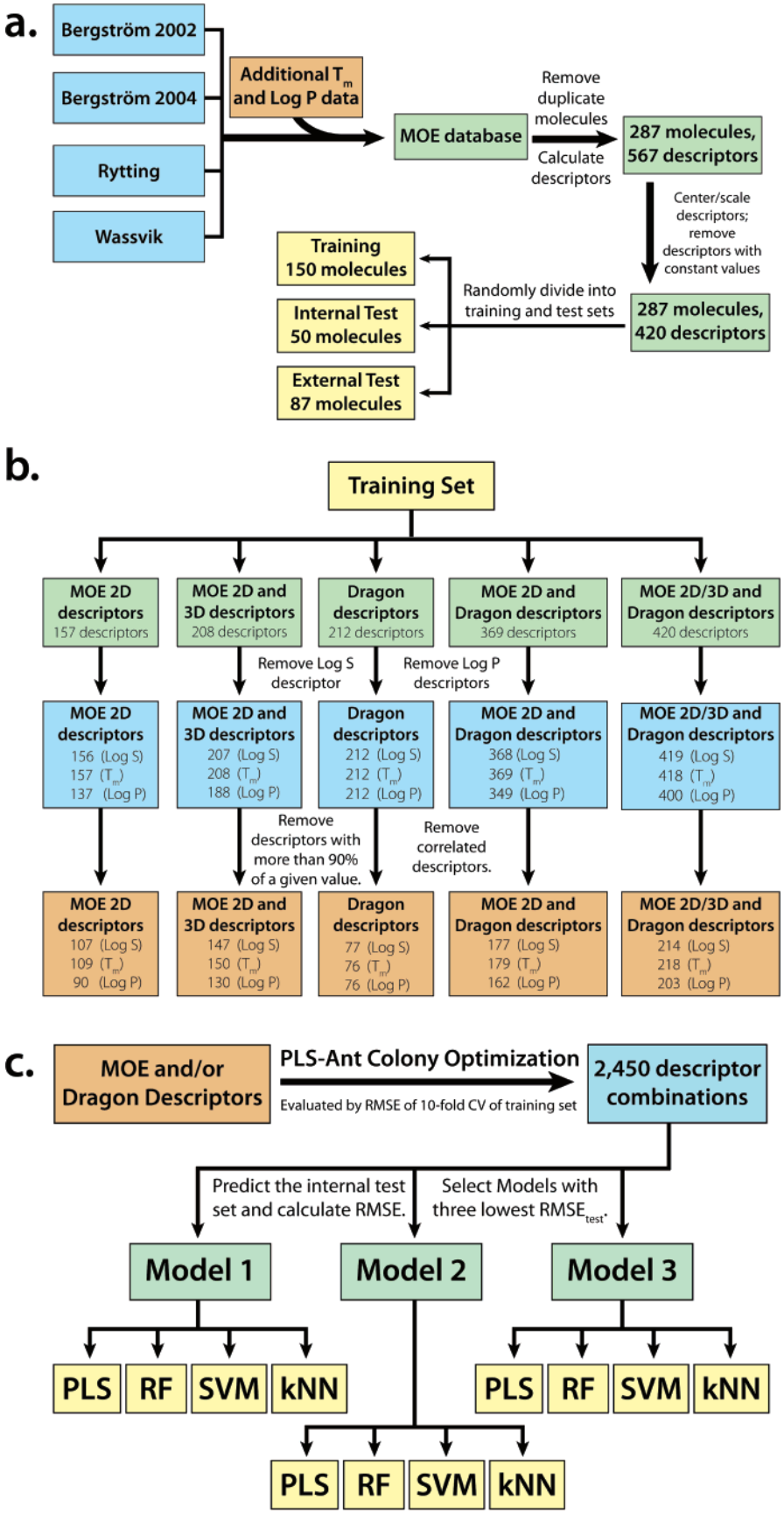

Figure 1. Descriptor processing and selection. a. Experimental data were taken from four primary literature sources and supplemented with literature data of melting point and octanol−water partition coefficient values. 2D and 3D descriptors were calculated in MOE, and additional fragment and geometry descriptors were calculated in Dragon. The data set was then randomly partitioned into training, internal test, and external test sets. b. Redundant and nondescriptive descriptors were eliminated from the data set. c. Descriptors were selected using an Ant Colony Optimization algorithm. The algorithm was guided using the $RMSE_{CV}$ of a PLS model of the training set. To pick the three best sets of descriptors for each variable, the $RMSE_{test}$ was calculated to avoid overfitting.

Table 1. Summary of Sets of Models Constructed[a]

predicted variable	descriptors	descriptor selection	regression method
• Log S • T_m • Log P	• 2D MOE • 2D and 3D MOE • Dragon • 2D MOE and Dragon • 2D/3D MOE and Dragon	• none • ACO (lowest $RMSE_{test}$) • ACO (second lowest $RMSE_{test}$) • ACO (third lowest $RMSE_{test}$)	• PLS • RF • kNN • SVM

[a] For each set of descriptors, PLS, RF, kNN, and SVM models were built.

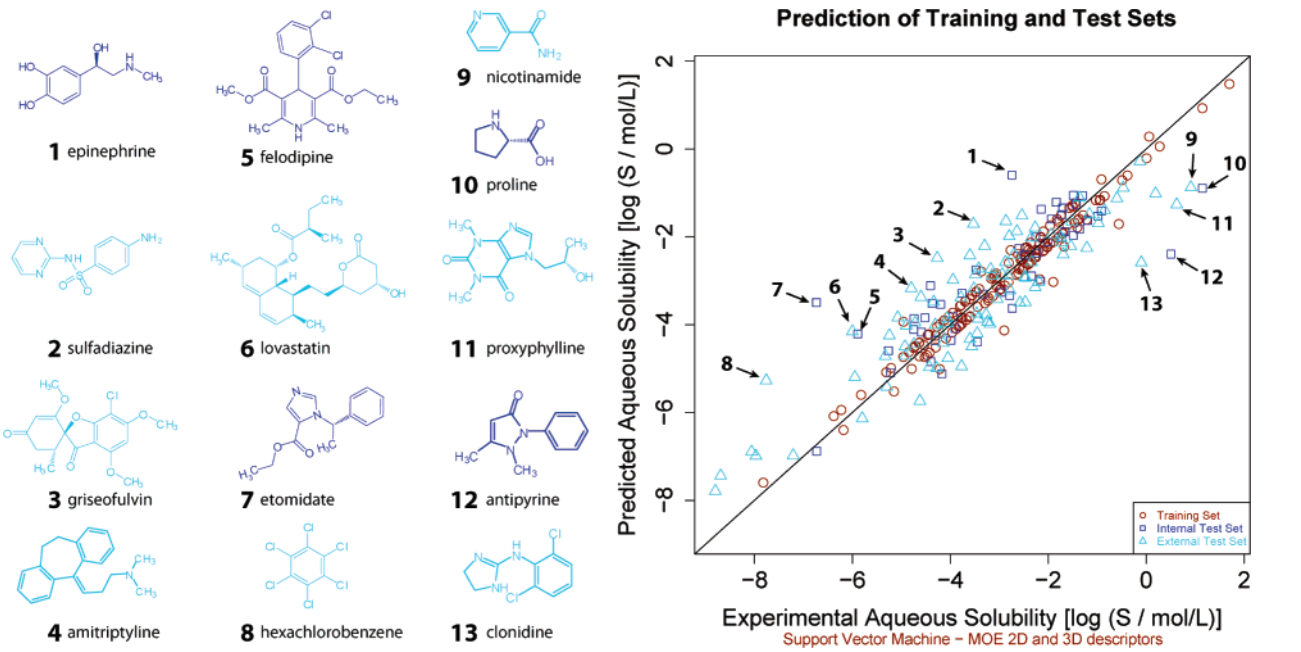

Figure 2. Experimental versus predicted Log S values. Brown circles represent training set data points, blue squares represent internal test set points, and turquoise triangles depict external test set points.

Table 2. Statistics for Best Models of Aqueous Solubility, Melting Point, and Octanol−Water Partition Coefficient

property	no. of descriptors	method	RMSE train	RMSE test	RMSE ext	AAE train	AAE test	AAE ext	r^2 train	r^2 test	r^2 ext	bias train	bias test	bias ext
Log S	41	SVM	0.299	0.936	0.900	0.239	0.657	0.732	0.96	0.68	0.79	0.041	−0.082	−0.065
T_m	35	SVM	27.5	48.1	52.8	23.6	33.8	40.7	0.84	0.50	0.46	−2.0	9.1	6.7
Log P	49	SVM	0.35	0.82	0.73	0.13	0.56	0.52	0.96	0.84	0.87	−0.02	−0.13	0.07

predicted values differed considerably from their experimental values. In particular, the SVM model had difficulty predicting substituted aromatic compounds with very low solubility (Log S less than −6 log units), which typically represent poor candidates for drug molecules. Three of the compounds with the highest errors also had three of the lowest experimental solubility values in the internal and external test sets (hexachlorobenzene, etomidate, and lovastatin). On the other hand, unsubstituted aromatic compounds at the lower end of the aqueous solubility level, such as naphthacene and chrysene, showed smaller differences between the observed and predicted values (errors = 0.230−1.252 log units).

The errors in low solubility compounds are difficult to attribute. While some have observed that QSPR models tend to fail at the extremes,[22] measurements at the high and low solubility limits are equally error-prone. Solubility measurements of low solubility compounds are often performed in a mixture of water and a nonpolar substance to enhance solubility, and the pure aqueous solubility is then extrapolated.[13] As such, experimental errors associated with these compounds tend to be higher than with moderately soluble compounds. The difference in the prediction of low solubility compounds presents an interesting dichotomy. While the series of unsubstituted aromatic compounds and the series of druglike substituted compounds both have a similar range of experimental solubility values (−8.8 to −6 log units), the errors for the unsubstituted compounds were significantly lower. This difference could suggest that it is easier to predict simple compounds rather than more complex, druglike molecules. However, the training set includes three unsubstituted aromatic compounds with Log S values less than −6 but only one substituted aromatic compound with such a low solubility. These data imply that the lack of diverse compounds with low solubility in the training set contributes to large errors in prediction, increasing the average error. We anticipate that increasing the diversity of the data set to include more polyfunctional molecules with low solubility would improve these predictions.

At the other end of the spectrum, compounds with high solubility were also associated with large errors (errors greater than 1.5 log units). Of the six compounds in the internal and external test sets with the largest Log S values (proline, nicotinamide, proxyphylline, antipyrine, pteridine,

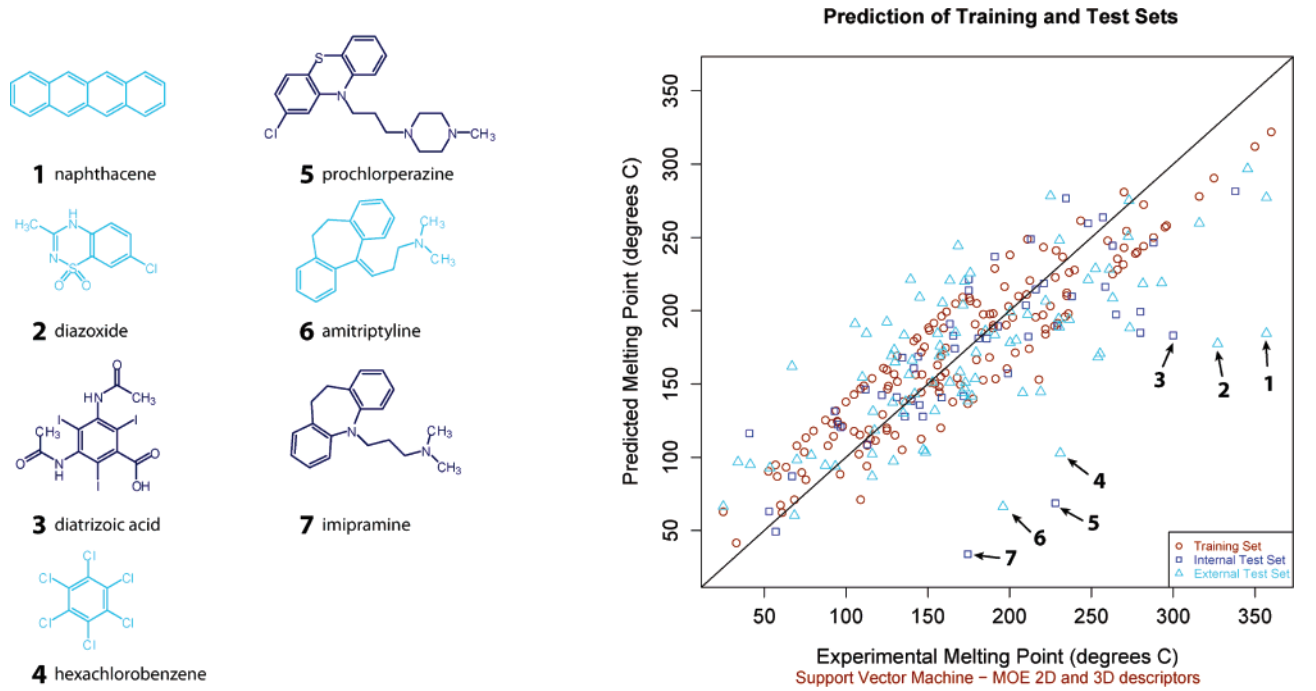

Figure 3. Experimental versus predicted T_m values. Brown circles represent training set data points, blue squares represent internal test set points, and turquoise triangles depict external test set points.

and clonidine), only pteridine had an absolute error less than 1.5 log units. Again, these large errors could be attributed to the fact that fewer compounds with high Log S values were included in the training set. We also note that some compounds may in fact undergo hydrolysis.

One major downside of using SVM modeling is that the relationships between the descriptors and the models are hidden. Consequently, descriptor importance in the model cannot be assigned. More important to a chemist, the influence of each descriptor, or each property of the compound, on the solubility cannot be deduced in a straightforward manner. While this SVM model predicts solubility more accurately than any other model in this study, its utility in suggesting how compounds could be altered to increase or decrease their solubility remains limited. However, with abstract descriptors such as Kier−Hall indices, interpretability is difficult regardless of the modeling method.

Melting Point Prediction. Melting point prediction was not as accurate as the prediction of aqueous solubility, based on r^2 statistics and comparison to prediction based on the mean. As with Log S prediction, the best model for melting point was a SVM model using a combination of MOE 2D and MOE 3D descriptors (see the Supporting Information). The $RMSE_{ext}$ was 52.8 °C, the r^2 value was 0.46, and the bias was 6.7 °C (Table 2). Additionally, the SVM model improved the $RMSE_{ext}$ by 26.8% compared to prediction based on the mean of the training set. While this $RMSE_{ext}$ is higher than has been reported for other models, the data set contains more druglike compounds than other models which report better statistics.

Interestingly, there were fewer outliers in the melting point model than in the Log S model, although, in general, predictions of melting point were less accurate (Figure 3). The decreased number of outliers in the melting point prediction could be a reflection of the relative ease of measuring T_m, compared to measuring Log S. The majority of compounds, especially the outliers, tended to be underpredicted, as indicated by the positive bias in the internal and external test set prediction (9.1 °C and 6.7 °C for the internal and external test sets, respectively). Of the seven apparent outliers, all were underpredicted by more than 105.6 °C (twice the $RMSE_{ext}$). Three of these compounds (naphthacene, diazoxide, and diatrizoic acid) had experimental melting point values concentrated in the high range of melting point values, although none of the experimental values were outside the range of the training set. Thus, similar to our Log S predictions and previous T_m models,[22] compounds at the extremes are associated with large errors. However, the proportion of outliers at the extremes is much lower for T_m: only three of the seven outliers had melting points above the 90th percentile of experimental values. The training set also included four compounds with experimental T_m values in this range, whereas the training set for Log S had only one complex molecule with low experimental solubility. Additionally, four compounds with experimental melting points greater than 300 °C in the internal and external test sets were predicted with absolute errors less than 100 °C. Moreover, none of the compounds in the lowest 10% of experimental values had absolute errors of prediction greater than 100 °C. Taken as a whole, these data suggest that prediction at extreme values does not play as large a role in the errors in melting point prediction as it does in solubility prediction. It is also possible that some outlier melting point data are caused by thermal decomposition.

Octanol−Water Partition Coefficient Prediction. Octanol−water partition coefficient proved to be the easiest of the three properties to predict. Similarly to both Log S and T_m, the best model used MOE 2D and 3D descriptors in a SVM model (see the Supporting Information). The r^2 value for the external test set was the highest for all the predicted properties, with a value of 0.87 (Table 2). The $RMSE_{ext}$ was also lower than that of the Log S prediction (0.73 log units

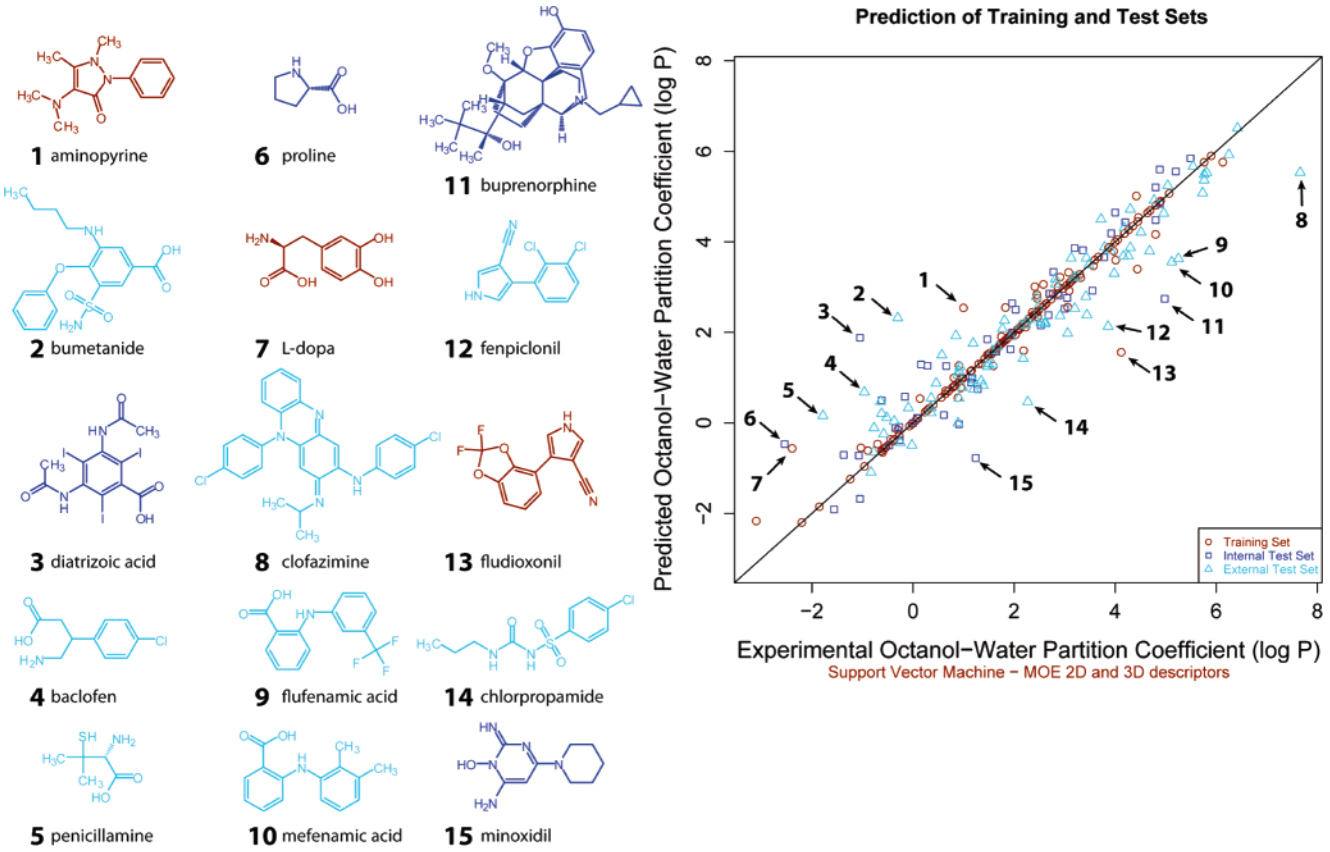

Figure 4. Experimental versus predicted Log P values. Brown circles represent training set data points, blue squares represent internal test set points, and turquoise triangles depict external test set points.

for Log P, 0.90 log units for Log S). In addition, the SVM model improved the $RMSE_{ext}$ by 65.4%, compared to a prediction based on the mean. This improvement of the $RMSE_{ext}$ through modeling was most pronounced for Log P, compared to both Log S and T_m, suggesting that Log P is easier to predict than both Log S and T_m. This conclusion agrees with prior work on all three properties through the use of a uniform data set and consistent modeling methods.

Figure 4 shows the agreement between the predicted Log P values and their experimental values. Similar to both Log S and T_m prediction, Log P prediction was associated with outliers at the upper and lower limits of the data range. Of the seven compounds that were significantly overpredicted, five had experimental values in the lowest 10% of experimental values. Similarly, half of the compounds that were underpredicted occurred near the upper limit of Log P values. For the compounds with the highest experimental Log P values, simple unsubstituted aromatic compounds were predicted more successfully than substituted ones. This result corresponds to the predictions of Log S for poorly soluble compounds. Additionally, Log P was the only property to have apparent outliers in the training set: fludioxonil, L-dopa, and aminopyrine were all predicted with absolute errors greater than 1.5 log units. The poor prediction of compounds at the upper and lower limits of Log P values implicates either a lack of data in these ranges and/or large experimental errors as major sources of error in Log P prediction, similar to Log S prediction.

Table 3. Inclusion of 3D Descriptors in Model[a]

model	no. of descriptors	method	external test set RMSE	r^2	bias
	Log S				
MOE 2D	37	SVM	0.946	0.77	−0.019
MOE 2D and MOE 3D	41	SVM	0.900	0.79	−0.065
MOE 2D and Dragon	54	PLS	0.944	0.77	−0.096
MOE 2D/3D and Dragon	214	PLS	0.945	0.77	−0.112
	T_m				
MOE 2D	32	SVM	54.0	0.43	6.2
MOE 2D and MOE 3D	35	SVM	52.8	0.46	6.7
MOE 2D and Dragon	44	RF	54.1	0.43	5.1
MOE 2D/3D and Dragon	66	RF	53.1	0.45	3.9
	Log P				
MOE 2D	41	PLS	0.89	0.81	0.11
MOE 2D and MOE 3D	49	SVM	0.73	0.87	0.07
MOE 2D and Dragon	72	PLS	0.78	0.86	0.09
MOE 2D/3D and Dragon	82	PLS	0.80	0.85	0.06

[a] Comparing the best model (based on $RMSE_{ext}$) using only 2D descriptors to the best model using 2D and 3D descriptors, adding 3D descriptors improves the models.

DISCUSSION

Comparison of Descriptors. In all cases, selecting a subset of descriptors using the ACO algorithm improved the model statistics. We will not explicitly discuss the inclusion of Dragon descriptors, due to the lack of model improvement. We should note, however, that we only used a subset of all the available Dragon descriptors; the entire set of Dragon descriptors, totaling over 1500 descriptors, could provide different results.

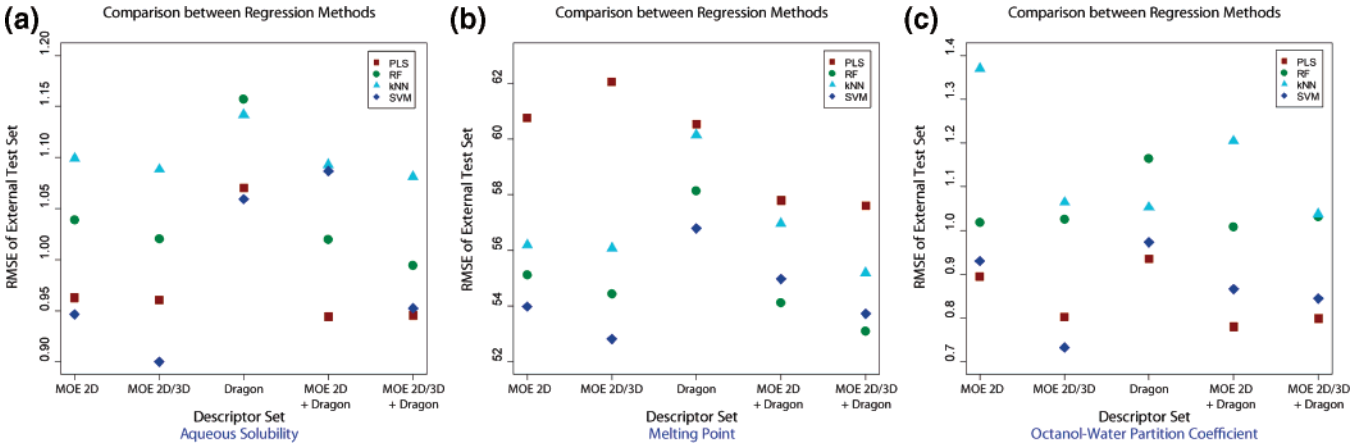

Figure 5. Best models for each regression method, using five sets of descriptors. Methods are compared based on the $RMSE_{ext}$: a. aqueous solubility, b. melting point, and c. octanol−water partition coefficient.

The incorporation of 3D descriptors with MOE 2D descriptors did provide new information, as the RMSE, r^2, and bias statistics improved for all three variables (Table 3). Indeed, the best overall models of aqueous solubility, melting point, and octanol−water partition coefficient all included between seven and 16 3D MOE descriptors. However, we saw no significant difference between Log S and Log P models when MOE 3D descriptors were included with MOE 2D and Dragon descriptors. While there was a slight improvement in statistics for melting point models when 3D MOE descriptors were present in addition to MOE 2D and Dragon descriptors, the improvement was less than when MOE 3D descriptors were used with MOE 2D descriptors alone. Interestingly, the improvement when MOE 3D descriptors were included with MOE 2D descriptors was largest for Log P prediction. The $RMSE_{ext}$ decreased by 18% for Log P, compared to 2.2% for T_m, and 4.9% for Log S when 3D descriptors were included with 2D descriptors. It is likely that the inclusion of 3D descriptors replaced some of the 2D descriptors, since many 3D descriptors are correlated with 2D ones. The three-dimensional descriptors therefore could provide more accurate calculation of these correlated properties, which could explain why their inclusion improves the quality of the models.

It should be noted that the direct comparison of the models presented in Table 3 could be misleading. Since the number and identity of the 2D descriptors differ between the pairs of models, the changes in performance could be attributed to differences in the 2D descriptors, rather than an improvement due to the addition of 3D descriptors.

Comparison of Regression Methods. The regression methods (PLS, RF, kNN, and SVM) performed differently based on the property being predicted (Figure 5). In all cases, the overall best model was a SVM model. However, low $RMSE_{ext}$ values were also observed for PLS models of Log S and Log P. The ability of PLS models to describe octanol−water partition coefficients agrees with Hansch's supposition that Log P should be a linearly additive property.[53] In contrast, most studies on Log S that compare nonlinear and linear modeling methods conclude that nonlinear methods more accurately represent solubility.[12,15,16] Consequently, it is interesting that linear PLS methods perform almost as well as nonlinear methods for modeling Log S, similar to results by Catana et al.[17] The ability of both linear and nonlinear methods to adequately predict solubility values could be due to the relationship between Log S and Log P. In the 1960s, Hansch observed that Log P values were correlated linearly with Log S values.[23] This observation led Yalkowsky to develop the GSE for the dissolution of solids in water;[8] indeed, the majority of models of solubility include some sort of octanol−water partition coefficient term, often as one of the most important descriptors.[12−14] Therefore, if Log S does depend linearly on Log P, and if most models rely heavily on Log P descriptors, linear methods such as PLS could model solubility reasonably well. However, we observed that, at high and low solubility values, the experimental Log P values did not correlate with the experimental Log S values as strongly as for midrange values; compounds with low and high solubility deviate from a linear regression equation of best fit. This observation suggests nonlinear behavior for solubility values as the values approach the upper and lower limits of solubility. SVM models could, therefore, be able to replicate the nonlinearity at the extremes of the data, whereas PLS models perform well on the middle portion of the data.

The best SVM model predicted compounds at the upper and lower limits of Log S better than PLS. In a PLS model of Log S using the descriptors from the best SVM model, the 20 compounds with the highest and lowest Log S values in the internal and external test sets were predicted with a RMSE 8% worse than the SVM model (Table SI1). Only one-quarter of the 20 values were predicted more accurately using PLS than with SVM. Additionally, the middle third of the internal and external training set Log S values were better predicted using the PLS method than using SVM. The RMSE for these 47 compounds improved by 11% when PLS is used instead of SVM. However, the RMSE for the 20 compounds with the most median experimental Log S values did not differ significantly between PLS and SVM models. One very poor prediction (erythromycin) strongly impacted the RMSE of the PLS predictions. The AAE for these 20 compounds, on the other hand, was 12.6% lower using PLS than SVM, indicating that, on average, the PLS method was associated with lower errors. Taken as a whole, these data suggest that the PLS model using MOE 2D and 3D descriptors modeled the moderate values of Log S better than the SVM model, but the SVM model accommodated extreme values better, lowering the overall RMSE.

Random Forest models performed slightly worse or slightly better than SVM models for melting point, depending

Table 4. Prediction Using the General Solubility Equation[a]

	RMSE			AAE			r^2			bias		
	train	test	ext	train	test	ext	train	test	ext	train	test	ext
GSE: experimental T_m, experimental Log P	0.848	1.065	0.974	0.593	0.746	0.708	0.67	0.59	0.75	−0.037	−0.146	−0.032
GSE: predicted T_m, experimental Log P	0.895	1.143	1.057	0.655	0.806	0.803	0.63	0.53	0.71	−0.017	−0.237	−0.098
GSE: predicted Log P, experimental T_m	0.854	1.161	0.901	0.600	0.831	0.725	0.66	0.51	0.79	−0.019	−0.013	−0.104
GSE: predicted Log P, predicted T_m	0.889	1.153	1.031	0.656	0.819	0.819	0.63	0.52	0.72	0.001	−0.105	−0.171
SVM model	0.299	0.936	0.900	0.239	0.657	0.732	0.96	0.68	0.79	0.041	−0.082	−0.065

[a] GSE indicates the statistics using experimental T_m and Log P values. Other predictions were made using different combinations of predicted and experimental T_m and Log P values. The statistics from the SVM prediction of aqueous solubility are provided for reference.

on the descriptor set (Figure 5b). The success of these two methods supports the conclusion that nonlinear methods more effectively model melting point. A RF model using MOE 2D/3D and Dragon descriptors performed nearly as well as the SVM model using MOE 2D and 3D descriptors ($RMSE_{ext}$ 53.1 °C and 52.8 °C, respectively). RF models can therefore be used in place of SVM models with little sacrifice in model performance; furthermore, RF models do not require time-consuming parameter optimization, and the importance of the descriptors can be easily assessed. For the RF model using MOE 2D/3D and Dragon descriptors, the ten most important descriptors are b_rotR, a_ICM, opr_nrot, E_sol, lip_don, PEOE_VSA-4, PM3_LUMO, a_don, DISPv, and J3D (see the Supporting Information for an explanation of the descriptors). This model suggests that rotatable bonds, size, and hydrogen bond donation all play a key role in melting point. The selected descriptors are similar to those in the PLS model by Bergström et al., where the number of rings, flexibility, partial charges, LUMO energy, and polar surface area were the most important descriptors.[1]

Comparison to Literature Models. Both the Log S and T_m predictions performed similarly to literature models, while Log P predictions were somewhat worse than past models. Both of the Bergström models of aqueous solubility perform similarly, with $RMSE_{ext}$ of 1.01 with some extrapolation (2004 model)[7] and 0.90 (2002 model).[27] Our SVM model of Log S therefore represents a slight improvement in prediction. Our $RMSE_{ext}$ is better than the 2004 model and the same as the 2002 model, which predicted the solubility of only five compounds. For melting point prediction, our $RMSE_{test}$ and $RMSE_{ext}$ are similar to the RMSE values reported by Bergström,[1] Karthikeyan,[2] and Clark.[3] These three studies cite RMSE of test sets between 48.9 and 50.4 °C, while our model has an $RMSE_{test}$ of 48.1 °C and an $RMSE_{ext}$ of 52.8 °C. The prediction of Log P values deviates the most from literature values, where commercial software tends to perform better than our SVM for the prediction of druglike molecules.[29,30] The difference between these models could suggest that fragment-based descriptors better describe Log P values, or it could reflect the significantly larger size of the training sets used to build commercial Log P prediction software. For all three properties, the similar statistics with respect to literature models emphasizes the difficulty of predicting a set of diverse druglike compounds.

Comparison to General Solubility Equation. The GSE (eq 1) was used to predict the aqueous solubility of the compounds in this study for comparison to our models. The GSE predicted the training, internal test, and external test sets similarly well, with the RMSE of 0.848 for the training set, 0.974 for the external test set, and 1.065 for the internal test set (Table 4). The internal test set performed worst of the three sets, consistent with findings in our work. Overall, the GSE performed worse than our prediction of aqueous solubility; moreover, our method does not rely on experimental melting point values.

Additionally, the combination of using predicted melting point and predicted octanol−water partition coefficient in the GSE produces models of satisfactory quality ($RMSE_{ext}$ ∼ 1 log unit), only slightly worse than predictions using experimental values ($RMSE_{ext}$ = 0.974, Table 4). This supports the validity of the GSE in predicting aqueous solubility and tentatively suggests that the melting point prediction presented in this work is sufficient for use in predicting solubility. However, the success of the GSE with predicted T_m values could merely reflect the small size of our data set. Unsurprisingly, using experimental rather than predicted melting point values with predicted Log P values improves the quality of the prediction. Since the largest errors of prediction were found for models of melting point, replacing these predicted values with experimental values should improve the model.

Comparison between Models of Log S, T_m, and Log P. Models of octanol−water partition coefficient performed best overall, with the largest coefficient of determination (r^2 = 0.87). Log P models also had the largest improvement in $RMSE_{ext}$ based on the mean (65.4%). Log S models followed Log P models in performance, with a coefficient of determination of 0.79 and an improvement in the $RMSE_{ext}$ of 55.4% compared to prediction based on the mean. Melting point models performed significantly worse than either Log P or Log S models, with a coefficient of determination of 0.46 and an improvement over mean-based prediction of 26.8%. In summary, Log P was the easiest property to predict and melting point was the most difficult.

The errors in prediction tend to be easier to understand for Log P and Log S models. In both cases, the majority of the poor predictions come from compounds in the internal and external test sets with values near the upper and lower bounds of the experimental range. These compounds tend to be underrepresented in the training set, making prediction of similar compounds difficult. In Log S and Log P prediction, 69.2 and 60.0%, respectively, of the outliers had experimental values in the lowest or highest 10% of the experimental values. Moreover, for both Log S and Log P models, simple unfunctionalized aromatic compounds were predicted more accurately than polyfunctional molecules. The T_m model had fewer outliers in the extreme values. Only 43% of the outliers had experimental melting points in the highest 10% of experimental values; additionally, no compounds were strongly overpredicted (no error was greater

than +105.6 °C). These data imply that melting point prediction suffers fewer errors from the prediction of extremes than Log S or Log P; this results from a combination of the prediction methods and the distribution of the data. As such, it is more likely that deficiencies in the descriptors, rather than the data set, are responsible for the poor prediction of T_m.

Overall, there are multiple ways of generating models with similar statistics. Different sets of descriptors (MOE 2D, MOE 3D, and Dragon), different selected descriptors within those sets, and different modeling methods can all be combined in ways to produce models with similar performance. Despite these variations in model performance, several commonalities emerge. Out of all the combinations of MOE 2D, MOE 3D, and Dragon descriptors, a combination of MOE 2D and 3D descriptors performed best for predicting aqueous solubility, melting point, and octanol−water partition coefficient. Additionally, for all three experimental properties, SVM regressions performed best, despite the fact that the descriptors were selected using an ACO/PLS algorithm. Attempts to build a set of common descriptors for the three properties proved fruitless. The models built on the consensus descriptors had an $RMSE_{ext}$ 8.4−143.4% higher than models built on individual sets of descriptors.

Comparing the models of the different experimental properties, Log P models performed the best, followed by Log S, and then T_m. Log S and Log P predictions were also strongly interrelated: lipophilic compounds were found to have large errors in both cases. Additionally, the best model of Log S included a large number of calculated subdivided surface area contributions to Log P as well as two MOE calculations of Log P, and the best model of Log P included the one calculated Log S descriptor (see the Supporting Information). When we plot the experimental Log S values against their Log P counterparts, we find a significant correlation, with $r^2 = 0.54$, 0.63 and 0.62 for the training, internal test, and external test sets, respectively. In contrast, the experimental Log S and T_m data are uncorrelated, with $r^2 =$ close to zero for these three data sets. These data imply that aqueous solubility is more closely related to octanol−water partition coefficient than melting point; or, in other words, solution-phase interactions appear to dominate solubility. Past literature reports corroborate these conclusions. The question then is: why do QSPR models consistently perform significantly worse with regard to melting point?

In the Introduction, we proposed three reasons for the failure of QSPR models: problems with the data, the descriptors, or the modeling methods. We find issues with the data unlikely to be the only source of error in Log S, T_m, and Log P predictions. Although the accuracy of the data provides a fundamental limit on the quality of a QSPR model, we attempted to minimize its influence by selecting consistent, high quality data. The Log S data all measured intrinsic solubility; moreover, the Bergström and Wassvik data came from the same research group, which used similar criteria to select reliable solubility measurements. With regards to the accuracy of T_m and Log P data, both properties are associated with smaller errors than Log S measurement. Moreover, the melting point model performed the worst, yet it is by far the most straightforward property to measure. For all three predictions, the $RMSE_{ext}$ is larger than estimates of experimental errors, which suggests that there are additional sources of error besides the accuracy of the data. Moreover, preliminary work in our laboratory suggests that using consistent, high quality solubility data does not significantly improve predictions. Two models of the same set of molecules, one using Log S values measured by one individual using the same method and the other using literature Log S values, performed similarly well. Taken together, this work and ongoing research in our laboratory imply that other sources of error are responsible for the poor prediction of melting point and solubility.

Despite the minimal influence of data quality on the errors, the lack of data at the extremes clearly plays a role for the outliers. Approximately 40−70% of the most poorly predicted compounds in all three models occurred in the lowest or highest 10% of experimental values. This source of error was more pronounced for Log P and Log S predictions; in Log S predictions, in particular, the training set lacked any complex molecules with low solubility. The comparatively low percentage of T_m outliers at the extremes (43% compared to 69.2% for Log S), in conjunction with the lack of compounds considerably overpredicted, suggests another significant source of error for melting point prediction.

The second source of errors, a failure of the methods, seems unlikely to explain the discrepancy between prediction of melting point and prediction of the other two properties. We used the same modeling methods for all three properties: the descriptors were selected in the same manner, and the models were built in the same way. Both linear and nonlinear methods were tried, and for all three properties, SVM performed the best. The modeling methods for melting point performed more equally poorly, while there was much larger variation observed for Log S and Log P. The average deviation from the lowest $RMSE_{ext}$ for all the models built was 9.2% for T_m, compared to 16.0% for Log S, and 34.6% for Log P. The limited difference in melting point models suggests two conclusions: either the selected methods are better suited for Log S and Log P prediction, or the descriptors incompletely explain melting point.

The latter source of error, the descriptors, seems likely to be a significant source of error in melting point prediction. While current descriptors sufficiently describe the properties of an isolated gas-phase molecule, they incompletely describe the condensed phase intermolecular interactions, as they do not account for crystal packing information. For melting point, these interactions are essential, as they describe the interactions that must be overcome for the crystal lattice to break down and the compound to melt. Although liquid-phase interactions are important, we believe new descriptors that describe crystal packing need to be developed for melting point prediction to advance. While the most accurate descriptors would include the explicit crystal structure, a plausible hypothetical structure might be sufficient to generate models with errors less than 40 °C, since differences in melting point for polymorphs are substantially less than the errors in current models.

CONCLUSIONS

The combination of MOE 2D and 3D descriptors performed best for predicting each of aqueous solubility, melting

point, and octanol−water partition coefficient. For all three experimental properties, SVM regression was the most effective modeling method.

Our results suggest that Log P is the easiest property to predict, followed by Log S, and that T_m is the hardest. We suggest that the failure of existing chemoinformatics descriptors adequately to describe interactions in the crystalline solid phase may be a significant cause of error in melting point prediction.

ACKNOWLEDGMENT

L.D.H. thanks the Gates Cambridge Trust for support. D.S.P. is supported by the Pfizer Institute for Pharmaceutical Materials Science, and F.N. is supported by Unilever. We would also like to acknowledge Unilever plc for their support of the Unilever Centre for Molecular Science Informatics.

Supporting Information Available: Structures, experimental values, and predicted values for the 287 compounds in the data set; list of descriptors included in the best Log S, T_m, and Log P models; experimental and predicted values for the data set; and predictions of the 20 molecules with the lowest and highest experimental Log S values (Table SI1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bergstrom, C. A.; Norinder, U.; Luthman, K.; Artursson, P. Molecular descriptors influencing melting point and their role in classification of solid drugs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1177−1185.

(2) Karthikeyan, M.; Glen, R. C.; Bender, A. General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Model.* **2005**, *45*, 581−590.

(3) Clark, M. Generalized fragment-substructure based property prediction method. *J. Chem. Inf. Model.* **2005**, *45*, 30−38.

(4) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355−366.

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

(6) Ran, Y.; He, Y.; Yang, G.; Johnson, J. L.; Yalkowsky, S. H. Estimation of aqueous solubility of organic compounds by using the general solubility equation. *Chemosphere* **2002**, *48*, 487−509.

(7) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208−1217.

(8) Ran, Y.; Yalkowsky, S. H. Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.

(9) Yang, G.; Ran, Y.; Yalkowsky, S. H. Prediction of the aqueous solubility: comparison of the general solubility equation and the method using an amended solvation energy relationship. *J. Pharm. Sci.* **2002**, *91*, 517−533.

(10) Wassvik, C. M.; Holmen, A. G.; Bergstrom, C. A.; Zamora, I.; Artursson, P. Contribution of solid-state properties to the aqueous solubility of drugs. *Eur. J. Pharm. Sci.* **2006**, *29*, 294−305.

(11) Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.

(12) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest models to predict aqueous solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150−158.

(13) Bergstrom, C. A.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and local computational models for aqueous solubility prediction of druglike molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1477−1488.

(14) Bergstrom, C. A.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for prediction of aqueous drug solubility. *Pharm. Res.* **2002**, *19*, 182−188.

(15) 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.

(16) Huuskonen, J.; Rantanen, J.; Livingstone, D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.* **2000**, *35*, 1081−1088.

(17) Catana, C.; Gao, H.; Orrenius, C.; Stouten, P. F. Linear and nonlinear methods in modeling the aqueous solubility of organic compounds. *J. Chem. Inf. Model.* **2005**, *45*, 170−176.

(18) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. QSPR studies on vapor pressure, aqueous solubility, and the prediction of water-air partition coefficients. *J. Chem. Inf. Model.* **1998**, *38*, 720−725.

(19) Walters, A. E.; Myrdal, P. B.; Yalkowsky, S. H. A method for estimating the boiling points of organic compounds from their melting points. *Chemosphere* **1995**, *31*, 3001−3008.

(20) Law, D.; Wang, W.; Schmitt, E. A.; Long, M. A. Prediction of poly(ethylene) glycol-drug eutectic compositions using an index based on the van't Hoff equation. *Pharm. Res.* **2002**, *19*, 315−321.

(21) Katritzky, A. R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; Karelson, M. Perspective on the relationship between melting points and chemical structure. *Cryst. Growth Des.* **2001**, *1*, 261−265.

(22) Nigsch, F.; Bender, A.; van Buuren, B.; Tissen, J.; Nigsch, E.; Mitchell, J. B. O. Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. *J. Chem. Inf. Model.* **2006**, *46*, 2412−2422.

(23) Hansch, C.; Quinlan, J. E.; Lawrence, G. L. Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.* **1968**, *33*, 347−350.

(24) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.

(25) Silverman, R. B. *The Organic Chemistry of Drug Design and Drug Action*, 2nd ed.; 2004.

(26) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(27) Faller, B.; Grimm, H. P.; Loeuillet-Ritzler, F.; Arnold, S.; Briand, X. High-throughput lipophilicity measurement with immobilized artificial membranes. *J. Med. Chem.* **2005**, *48*, 2571−2576.

(28) Chou, J. T.; Jurs, P. C. Computer-assisted computation of partition coefficients from molecular structures using fragment constants. *J. Chem. Inf. Model.* **1979**, *19*, 172−178.

(29) Machatha, S. G.; Yalkowsky, S. H. Comparison of the octanol/water partition coefficients calculated by ClogP, ACDlogP and KowWin to experimentally determined values. *Int. J. Pharm.* **2005**, *294*, 185−192.

(30) Mannhold, R.; Petrauskas, A. Substructure versus whole-molecule approaches for calculating Log P. *QSAR Comb. Sci.* **2003**, *22*, 466−475.

(31) Eros, D.; Kövesdi, I.; Orfi, L.; Takács-Novák, K.; Acsády, G.; Kéri, G. Reliability of logP predictions based on calculated molecular descriptors: a critical review. *Curr. Med. Chem.* **2002**, *9*, 1819−1829.

(32) 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.

(33) Raevsky, O. A.; Trepalin, S. V.; Trepalina, H. P.; Gerasimenko, V. A.; Raevskaja, O. E. SLIPPER-2001 - Software for predicting molecular properties on the basis of physicochemical descriptors and structural similarity. *J. Chem. Inf. Model.* **2002**, *42*, 540−549.

(34) Sun, H. A universal molecular descriptor system for prediction of LogP, LogS, LogBB, and absorption. *J. Chem. Inf. Model.* **2004**, *44*, 748−757.

(35) Wegner, J. K.; Zell, A. Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077−1084.

(36) 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.

(37) Rytting, E.; Lentz, K. A.; Chen, X. Q.; Qian, F.; Vakatesh, S. Aqueous and cosolvent solubility data for drug-like organic compounds. *AAPS J.* **2005**, *7*, E78−E105.

(38) Lide, D. R. *CRC Handbook of Chemistry and Physics*, 84th ed.; CRC Press: 2003.

(39) O'Neil, M. J.; Hechelman, P. E.; Koch, C. B. *The Merck Index*, 14th ed.; Merck: Whitehouse Station, NJ, 2006.

(40) SciFinder Scholar, Chemical Abstracts Service, P.O. Box 3012, Columbus, OH 43210, U.S.A. http://www.cas.org/SCIFINDER/ (accessed August 16, 2007).

(41) *EPI Suite*; EPA: 2005. http://www.syrres.com/esc/epi.htm (accessed April 19, 2007).
(42) Salminen, T.; Pulli, A.; Taskinen, J. Relationship between immobilised artificial membrane chromatographic retention and the brain penetration of structurally diverse drugs. *J. Pharm. Biomed. Anal.* **1997**, *15*, 469−477.
(43) Zissimos, A. M.; Abraham, M. H.; Du, C. M.; Valko, K.; Bevan, C.; Reynolds, D.; Wood, J.; Tam, K. Y. Calculation of Abraham descriptors from experimental data from seven HPLC systems; evaluation of five different methods of calculation. *J. Chem. Soc., Perkin Trans. 2* **2002**, 2001−2010.
(44) *MOE*; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2007.
(45) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity−a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(46) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *Dragon Professional*, 5; Milano, Italy, 2004.
(47) R Development Core Team. In *R: A language and environment for statistical computing;* R Foundation for Statistical Computing: Vienna, Austria, 2006.
(48) Wehrens, R.; Mevik, B.-H. *pls: Partial Least Squares Regression* (*PLSR*) *and Principal Component Regression* (*PCR*), *R version 2.0-0*; 2006.
(49) Liaw, A.; Wiener, M. Classification and regression by Random Forest. *R News* **2002**, *2*, 18−22.
(50) 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 modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(51) Cannon, E. O.; Bender, A.; Palmer, D. S.; Mitchell, J. B. O. Chemoinformatics-based classification of prohibited substances employed for doping in sport. *J. Chem. Inf. Model.* **2006**, *46*, 2369−2380.
(52) Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, A. *e1071: Misc Functions of the Department of Statistics* (*e1071*), *R package version 1.5-16*; 2006.
(53) Iwasa, J.; Fujita, T.; Hansch, C. Substituent constants for aliphatic functions obtained from partition coefficients. *J. Med. Chem.* **1965**, *8*, 150−153.

CI700307P