Support Vector Machines for the Estimation of Aqueous Solubility

Peter Lind* and Tatiana Maltseva

Medivir AB, Lunastigen 7, 141 44 Huddinge, Sweden

Received May 28, 2003

Support Vector Machines (SVMs) are used to estimate aqueous solubility of organic compounds. A SVM equipped with a Tanimoto similarity kernel estimates solubility with accuracy comparable to results from other reported methods where the same data sets have been studied. Complete cross-validation on a diverse data set resulted in a root-mean-squared error = 0.62 and $R^2 = 0.88$. The data input to the machine is in the form of molecular fingerprints. No physical parameters are explicitly involved in calculations.

INTRODUCTION

The aqueous solubility of compounds is one of the most important factors determining their usefulness as drugs. Insufficient solubility is a common reason for poor bioavailability of drug candidates. Computational methods are often applied to virtual libraries early in the drug development process to screen out sublibraries based on estimated physical properties. These methods must be computationally efficient and reliable. Several methods have been proposed for the estimation of aqueous solubility, and most of them fall into two categories: Methods of the first type calculate logS as a sum of contributions from functional groups or fragments.[1−6] These procedures often employ various correction terms to account for pairwise group interactions.[2−6] Methods of the second type use regression of experimental or nonexperimental molecular parameters.[7−11] Some nonexperimental parameters can be derived from the molecular structure with low cost, such as topological indexes or count of hydrogen bond donors. These are often called 2D-descriptors or fast descriptors. Others parameters require more calculation, such as those from ab initio calculations. Commonly used regression methods are partial least squares, multiple linear regression, and artificial neural networks. Methods which include experimental molecular parameters are of no interest for virtual screening.

Support Vector Machines (SVMs) are computer programs for regression and classification. These were first applied to classification problems;[12] the methodology for regression was developed later. In recent years, SVM methods have successfully been applied to a range of pattern-recognition problems. The potential of SVMs for use in QSAR and QSPR has been discussed.[13,14]

In this paper, we report the application of Support Vector Regression for the estimation of aqueous solubility. We will briefly outline the theory underlying Support Vector Regression. For a full description with historical background, see monographs of Vapnik,[15] Christianini,[16] Herbrich,[17] and Schölkopf.[18]

THEORETICAL BASIS AND DEFINITIONS

The general learning problem is to find a relationship between objects $x \in X$ and targets $y \in Y$ based solely on a sample $z = (x,y) = ((x_1,y_1),...,(x_m,y_m)) \in (X \times Y)^m$ of size $m \in N$. This relationship is called a *hypothesis*. If the output space Y contains a finite number of elements $Y = (y_1,...y_n) \in Y^n$, then the task is called *classification*. If the output space is a set of n real targets $Y \in R^n$, then the task is called *regression*. In this paper, we deal with regression and with input spaces X which are sets of molecules.

A function $\Phi:X \rightarrow R$ that maps each object $x \in X$ to a real value $\Phi_i(x)$ is called a *feature*. Combining n features $\Phi_1,...,\Phi_n$ results in a *feature mapping* $\Phi:X \rightarrow F$ where the space F is called a *feature space*. The number of features may be infinite, corresponding to an infinite dimensional feature space.

A *kernel* is an inner product function $k:X \times X \rightarrow R$ in F, so that for all $x_i,x_j \in X$ and $\Phi:X$ we have

$$k(\mathbf{x}_i,\mathbf{x}_j) \equiv \langle \Phi(\mathbf{x}_i),\Phi(\mathbf{x}_j) \rangle = \langle \mathbf{x}_i,\mathbf{x}_j \rangle = \mathbf{K}_{ij} \qquad (1)$$

where $\mathbf{K}$ is a *Gram matrix* for a given kernel. The Gram matrix and the feature space are also called the *kernel matrix* and the *kernel space*, respectively.

Support Vector Regression (SVR). A support vector machine is first trained on a sample with objects having known target values. After training, the machine is used to predict or estimate target values for objects where these values are unknown. A kernel-induced feature space with function $k(x_i,x)$ is used for the mapping of objects onto target values. Thus a nonlinear feature mapping will allow the treatment of nonlinear problems in a linear space. The prediction or approximation function used by a basic SVM is

$$f(x) = \sum_{i=1}^{l} \alpha_i k(x_i,x) + b \qquad (2)$$

where α_i is some real value, x_i is a feature vector corresponding to a training object, and $k(x_i,x)$ is a kernel function. The components of vector α and the constant b represent the hypothesis and are optimized during training. It may be useful to think of the kernel, $k(x_i,x)$ as comparing patterns, or as evaluating the proximity of objects in their feature space. Thus a test point is evaluated by comparing it to all training points. Training points with nonzero weight α_i are called the *support vectors*.

* Corresponding author fax: +46 8 6083199; e-mail: Peter.lind@medivir.se.

10.1021/ci034107s CCC: $25.00
Published on Web 10/17/2003

Kernel Requirements. The function $k(x,z)$ is a kernel corresponding to the feature mapping Φ if the kernel matrix K is symmetric, and there is an orthogonal matrix Λ such that $K = V\Lambda V^T$, where Λ is a diagonal matrix containing the eigenvalues λ_t of K with corresponding eigenvectors $v_t = (v_{ti})_{i=1}^{n}$ and $\Phi(x_i) = \sqrt{\Lambda}v_i$, so that

$$K_{ij} = \langle\Phi(x_i),\Phi(x_j)\rangle = \sum_{t=1}^{n}\lambda_t v_{ti}v_{tj} = (V\Lambda V^T)_{ij} = K(x_i,x_j) \quad (3)$$

It is required that the eigenvalues of K are nonnegative (K is positive semidefinite), because a negative eigenvalue λ_s with eigenvector v_s would give rise to a point in the feature space with norm squared $||\Phi(x_s)||^2 = v_s^T\Lambda v_s = \lambda_s < 0$, contradicting the geometry of an inner product space (see also Mercer's theorem[19]).

Some commonly used kernels[15−18] are

$$\text{Linear: } k(x_i,x_j) = [x_i,x_j]$$

$$\text{Polynomial: } k(x_i,x_j) = ([x_i,x_j] + 1)^d$$

$$\text{Radial Basis Function, RBF: } k(x_i,x_j) = \exp(-(x_i - x_j)^2/(2\sigma^2))$$

It may seem necessary to first design a mapping function specifically for the particular problem at hand and then work out the corresponding kernel, but this is seldom done in practice. A key idea of kernel-based learning methods is that a kernel may often be chosen directly, without first examining the corresponding feature space.

Tanimoto Kernel. In chemistry, the Tanimoto, or Jaccard, index[20] (T_{ab}) is often used to measure the similarity between molecules. T_{ab} is implemented in a number of commercial computational chemistry toolkits and is typically applied to bit vectors known as molecular fingerprints (see below). In the case of binary valued vectors, the Tanimoto index may be defined as

$$T_{ab} \equiv N_{ab}/(N_a + N_b - N_{ab}) = \langle x_a,x_b\rangle/(\langle x_a,x_a\rangle + \langle x_b,x_b\rangle - \langle x_a,x_b\rangle) \quad (4)$$

where N_{ab} is the number of bits that patterns a and b have in common, and N_a and N_b are the numbers of bits set in a and b, respectively.

An important feature of this index is that it disregards information of nonoccurrences common to the compared objects. For example, the similarity between two molecules differing by one atom will depend on molecule size, so that similarity is high between a pair of large molecules but low between a pair of small ones.

The Tanimoto index is a symmetric ($T_{ab} = T_{ba}$) and positive semidefinite function as shown by Gower,[21] thus it fulfills the necessary conditions of a kernel function. It has been used as an SVM kernel function in the context of image analysis.[22] Applications of the Tanimoto function in chemistry include similarity searching of chemical databases[23] and clustering.[24] Two comparative studies on 22 binary similarity indexes[25,26] applied to similarity searches concludes that the Tanimoto index may be the best single measure of similarity and that no combinations with other indexes results in consistent improvement over its use.

SVM Training. During training, an expression for the cost of errors called the *loss function*, $L(y,f(x,\alpha))$, is minimized. In this work, we use a so-called linear ϵ-insensitive machine,[15] for which

$$L(y,f(x,\alpha)) = L(|y - f(x,\alpha)|_\epsilon) \quad (5)$$

here

$$|y - f(x,\alpha)|_\epsilon = \begin{cases} 0, \text{ if } |y - f(x,\alpha)| \le \epsilon \\ |y - f(x,\alpha)| - \epsilon, \text{ otherwise} \end{cases}$$

is minimized. The parameter ϵ, which is chosen a priori, defines a band of width 2ϵ around the output function. The cost of errors for points lying inside that band is zero. Points lying outside the band defined by ϵ are support vectors and will give rise to nonzero components in the optimized α vectors. The purpose of ϵ is to protect against overfitting. For a detailed discussion of the solution to this minimization problem, see ref 15. For a description of the algorithm which is used in this work to solve this problem numerically, see ref 27. A parameter called C defines the maximum value of all α_i and is called the cost, or regularization parameter**.** The purpose of the constraint C is to limit the influence of outliers.

Molecular Fingerprints. Historically, molecular fingerprints have evolved from *structural keys*, which are used for searching chemical databases. Generation of structural keys employs a predefined dictionary of substructures and lets each bit in a bit vector correspond to the presence or absence of a particular substructure. Fingerprints does not make explicit use of substructures; the common approach is to enumerate all subpatterns of bonds and atoms in all possible paths of the molecular graph up to some predefined length. Then, some representation of each pattern seeds a pseudorandom number generator, the output of which indexes the bit vector where bits are turned on as corresponding patterns are encountered. Thus, a particular bit may be set by unrelated features, and the process does not strictly guarantee that different molecules give distinct fingerprints; however, a substructure will always turn on some common set of bits in the fingerprints of all its superstructures. Typical fingerprints are of size 1024 to 4096 bits. Software for the generation of molecular fingerprints is available from a number of companies[28−33] and as open source.[34] The generation of fingerprints is in itself a feature mapping, which may be seen as separate from the mapping from fingerprints to kernel space.

Fingerprint Kernels. The SVM kernel used in this work evaluates the Tanimoto similarity of molecular fingerprints. We use the SynChemistrySimilarityMatchEx[35] function in the Accord Software Development Kit,[28] which generates sets of attributes for a pair of molecules and evaluates their Tanimoto distance. The exact algorithm used by this function is not disclosed by Accelrys, and the underlying attributes may be different from those coded by traditional fingerprints. No comparisons with other types of similarity functions were made.

SOFTWARE

The software which we have used for SVM calculations is based on the shareware program Libsvm[36] v. 2.33 of Chih-

Chung Chang and Chih-Jen Lin. Libsvm is available as Java and C++ code. The basic training algorithm of libsvm is a simplification of both Sequential Minimal Optimization (SMO)[27] by Platt and SVMLight[37] by Joachims. The Java version was rewritten in VB.NET and modified for use with chemistry. The chemical functionality of the Accord Software Development Kit (SDK) version 5.2[28] was incorporated in the program as separate classes. This choice of software components was based on considerations of cost, development time, and run-time efficiency. All functionality was compiled into a single program, where SVM and chemistry functions run in the same process. Although various separate fingerprint and similarity functions are availible in the Accord SDK, only the SynChemistrySimilarityMatchEx function was tried in this work. Classes for statistical analysis and batch run control were added. A graphical interface for control of parameters was also constructed. This modified program takes a text file with SMILES[38] strings as input and uses Accord SDK functions to instantiate corresponding chemical objects.

Values for training parameters ϵ and C were chosen after running partial (20×20) cross-validations on the training data. The optimization stopping parameter of libsvm was 0.1 in all experiments. Root-mean-squared errors were calculated as

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

and the squared correlation coefficient as

$$R^2 = 1 - \sum_{i=1}^{n}(y_i^{obs} - y_i^{pred})^2 / \sum_{i=1}^{n}(y_i^{obs} - y^{average})^2$$

where y^{obs} are experimental and y^{pred} are predicted values.

DATA SETS

Three data sets were used in this study: set A ($n = 883$) is essentially[39] the training data set of Huuskonen,[7] compiled from the AQUASOL[40] and PHYSPROP[41] databases. This set, which has been used in several studies,[9,42−44] contains a large proportion of drugs and pesticides. LogS values span from −11.62 to 1.58, with standard deviation of 2.01. Set B ($n = 412$) is the data referred to as the testing set in Huuskonens work.[7,39] This set is similar in character to set A. Set C ($n = 411$) is the data used in the work of Katritzky et al.,[8] in which regression models based on electrostatic, quantum chemical, and topological parameters are studied. This set contains no drugs and consists mostly of haloalkanes and monofunctional compounds. The logS values of all data sets are based on solubilities given as moles per liter. The data sets of Huuskonen are available as supplementary data of ref 7. The data set of Katritzky et al. is available from ref 8.

RESULTS

On set A, a complete cross-validation using the Tanimoto fingerprint kernel gave rmse = 0.62 and $R^2 = 0.88$. Figure 1 is a graph showing the results of this cross-validation. On the smaller set B, a similar cross-validation resulted in an rmse = 0.77 and $R^2 = 0.86$.

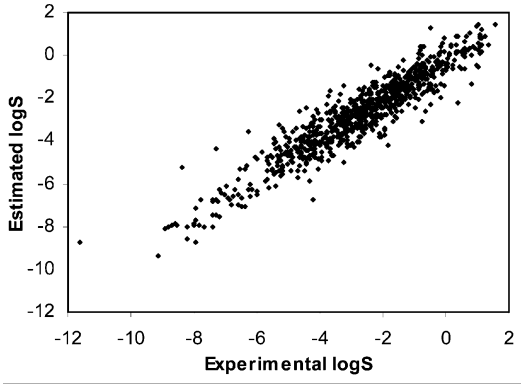

Figure 1. Complete cross-validation on data set A.

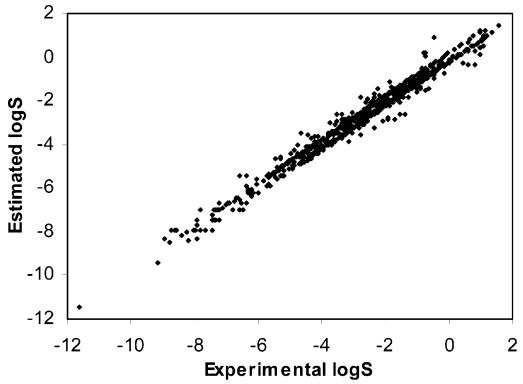

Figure 2. Fitting of solubility data for set A.

A machine was trained on set A and used to predict logS values in set B. The effect of the ϵ parameter on the fit to set A training data can be seen in Figure 2, where a band of width 2ϵ is apparent. The measures of fit on the training data are rmse = 0.29 and $R^2_{\mathrm{train}} = 0.98$. A number of 630 compounds (71%) were selected as support vectors by the machine in this experiment. This trained machine predicts the log S values of set B with a standard deviation of prediction, rmse = 0.68, and with $R^2_{\mathrm{pred}} = 0.89$. Parameters used for experiments on sets A and B were $\epsilon = 0.2$ and $C = 0.15$.

A complete cross-validation on set C resulted in a root-mean-square error (rmse) = 0.57 and a squared correlation coefficient (R^2) = 0.88. Parameters for this run were $\epsilon = 0.1$ and $C = 20$.

Dependency on Similarity with Training Set. An important question is to what degree the successful prediction requires the presence of compounds in the training set which are similar to the compounds to be predicted. In other words, can some measure of similarity between test and training data be used to grade the confidence of prediction of individual points? In an attempt to answer this question, compounds in set A were partitioned into two dissimilar subsets, A1 (629 compounds) and A2 (254 compounds). A standard clustering algorithm based on mean pairwise Tanimoto similarities was used for this division. Compounds in A1 had mean internal Tanimoto similarity = 0.53. The average similarity between members in set A1 and A2 was 0.32. The mean internal similarity in A2 was 0.51. A SVM trained on A1 predicted logS values in set B with rmse = 1.02. Thus the predictive ability was lowered as compared

Table 1. Compounds in Test Set B Binned by Absolute Error of Prediction

bin no.	logS absolute error	*n*	av Tanimoto similarity of bin compds with compds in A1	av no. of compds in training set being more similar than: 0.5	0.6	0.7	0.8	0.9
1	<0.1	41	0.50	296	152	64	19	5
2	0.1−0.25	57	0.49	276	144	65	21	4
3	0.25−0.5	82	0.49	297	163	70	22	4
4	0.5−0.7	65	0.47	273	143	54	14	3
5	0.7−1.0	61	0.47	255	129	47	14	2
6	1.0−2.0	80	0.41	187	98	38	12	1
7	>2.0	26	0.27	30	11	2	0	0

to machines trained on the complete set A. This may be due to the smaller size of the training set, or to the exclusion of a class of compounds in the training set which are similar to some proportion of compounds in B, or to both factors.

The compounds in set B were binned into groups based on their absolute error of prediction, using an SVM trained on set A1. The average Tanimoto similarity with all compounds in the training set was calculated. For each test compound, compounds in the training set more similar than 0.5, 0.6, 0.7, 0.8, and 0.9 were counted. Table 1 summarizes the results. Compounds in the bin with the highest error of prediction have a significantly lower number of similar compounds in the test set as compared to compounds in other bins. Test compounds having at least one compound in the training set more similar than 0.6 ($n = 332$) have a root-mean-squared error of prediction = 0.76, significantly lower than the overall value. This suggests that some measure of overall similarity to test set members may be used to grade the level of confidence for individual predictions, although there is not enough evidence to state any quantitative rule.

Comparison with Other Methods. The Tanimoto kernel method was compared to several other reported methods where large training sets have been used. Two of these methods use the same data set as the present work, making a direct comparison possible in these cases. Katritzky et al.[8] uses a five-descriptor model and multilinear regression to estimate solubilities of compounds in the set C. The method makes use of molecular mechanics, AM1 geometry optimization, and quantum-chemical descriptors. A complete cross-validation was run to assess the quality of the model. The results of this cross-validation is rmse = 0.57, $R^2 = 0.87$. The method we present here gives a model of comparable quality (rmse = 0.57 and $R^2 = 0.88$) on the same data set.

Huuskonen[7] uses multiple linear regression (MLR) and an artificial neural network (ANN) to fit E-state[45] and other topological indices to logS data. The molecules in set B[39] were used for testing. The fit on test data was rmse = 0.71 and $R^2 = 0.88$ for the best MLR model. The best ANN model gave rmse = 0.60 and $R^2 = 0.92$. The corresponding values in the present work are rmse = 0.68 and $R^2 = 0.89$.

Similar work of Tetko et al.[42] resulted in a model which had rmse = 0.81 for a MLR model, and rmse = 0.60 for a ANN model when tested on compounds from set B.

A fragment-based method of Klopman et al.[1] was developed by fitting data from 1168 compounds. The fit on training data was rmse = 0.50, $R^2 = 0.95$. The estimation of logS on a 120 compound test set resulted in rmse = 0.79.

A method of Cheng and Merz[46] uses descriptors derived from molecular graphs. The fit on a 784 compound training set is rmse = 0.87 and $R^2 = 0.84$. The fit on a test set was rmse = 0.79 and $R^2 = 0.88$.

Gao and Shanmugasundaram reports a method[47] which is based on molecular descriptors calculated with MOE software.[32] The descriptors were selected by means of principal components regression and a genetic algorithm. Testing of the model on a 249 compound diverse test set resulted in rmse = 0.40 and $R^2 = 0.91$.

Comparison with Simple Similarity Method. We compared the SVM method to a method in which test compounds were given similar values as similar compounds in the training data set. The logS values were calculated as

$$y = \sum_{i=1}^{m} k(x,x_i)^n y_i / \sum_{i=1}^{m} k(x,x_i)^n$$

where the exponent n is an integer. For both sets A and C the lowest observed rmse was 1.02. This value was found for set A with $n = 30$ and for set C with $n = 36$. These values of n give a strong weighting of contributions from the very closest neighbors.

Computational Efficiency. The time for training on set A was 140 s on a standard 1000 MHz PC. Training time was found to scale with slightly less than the square of training set size, as expected for this algorithm.[27] Testing performance was in the order of 7 compounds per second for a machine trained on set A having 630 support vectors. Most of the CPU time during testing was spent on the instantiation of Accord chemical objects from SMILES and on evaluating the Accord similarity function, processes which include generation of the underlying fingerprints. It follows that the testing throughput of this application type is in practice limited by the rate to which test compound fingerprints can be generated and fed to the SVM.

DISCUSSION

We use support vector regression to learn solubility directly from experimental data and general-purpose molecular fingerprints. There is no manual selection of descriptors and no construction of a fragment library. No prior knowledge about the physical phenomena underlying solubility is used.

Fixed Training Process. The training process is fixed once the training parameters and fingerprint type are chosen. This means that the machine can be trained on new or expanded data sets without any need for manual steps or expert guidance. This can be useful in a drug development scenario where experimental logS values for representatives of a new class of compounds become available after synthesis. The binning analysis described above and the experiments with data sets A1 and A2 suggest that adding training compounds similar to the compounds to be tested will improve the accuracy of prediction, although it is not clear to what extent.

Interpretability. By the nature of molecular fingerprints, there is no direct way to express a learned hypothesis in human-understandable terms. This is a disadvantage of the method.

CONCLUSIONS

A support vector machine equipped with a Tanimoto similarity kernel acting on general-purpose molecular fin-

gerprints is able to estimate aqueous solubility with good accuracy. It may be possible to estimate a confidence of prediction for individual points by using a measure of similarity between the test point and the training points. Once implemented, the learning system can be applied to new data sets without expert guidance.

REFERENCES AND NOTES

(1) Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*(2), 439−445.

(2) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.

(3) Wakita, K.; Yoshimoto, M.; Miyamoto, S.; Watanabe, H. A method for calculation of the aqueous solubility of organic molecules by using new fragment solubility constant. *Chem. Pharm. Bull. (Tokyo)* **1986**, *34*, 4663−4681.

(4) Suzuki, T. Development of an automatic calculation system for both the partition coefficient and aqueous solubility. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 149−166.

(5) Kuhne, R.; Ebert, R.-U.; Kleint F.; Schmidt, G.; Schüürmann, G. Group Contribution Method to Estimate Water Solubility of Organic Chemicals. *Chemosphere* **1995**, *30*, 2061−2077.

(6) Lee, Y.; Myrdal, P. B.; Yalkowski, S. H. Aqueous Functional Group Activity Coefficients (AQUAFAC) 4: Applications to Complex Organic Compounds. *Chemosphere* **1996**, *33*, 2129−2144.

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

(8) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. QSPR Studies on Vapour Pressure, Aqueous Solubility, and the Prediction of Water−Air Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720−725.

(9) 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.

(10) Butina, D.; Gola, J. M. R. Modeling Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, xxxx−xxxx.

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

(12) Boser, B.; Guyon, I.; Vapnik, V. N. A training algorithm for optimal margin classifiers. *Fifth Annual Workshop on Computational Learning Theory*; ACM: Pittsburgh, 1992; pp 144−152.

(13) Burbidge, R.; Trotter, M.; Buxton, B.; Holden S. Drug Design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.

(14) Czerminski, R.; Yasri, A.; Hartsourgh, D. Use of Support Vector Machine in pattern classification: Application to QSAR studies. *Quant. Struct.-Act. Relat.* **2001**, *20*, 227−240.

(15) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: 1995.

(16) Christianini, N.; Shawe-Taylor, J. *An introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: 2000.

(17) Herbrich, R. *Learning Kernel Classifiers: theory and algorithms*; MIT Press: 2002.

(18) Schölkopf, B.; Smola A. J. *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*; The MIT Press: Cambridge, MA; London, England.

(19) Mercer, J. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. R. Soc. London, Ser. A* **1909**, *209*, 415−446.

(20) Tanimoto, T. *An elementary mathematical theory of classification and prediction*; IBM Report; 1958.

(21) Gower, J. C.; Legendre P. Metric and Euclidean properties of dissimilarity coefficients. *J. Classification* **1986**, *3*, 5−48.

(22) Pękalska, E.; Paclík, P.; Duin, R. P. W. A Generalized Kernel Approach to Dissimilarity-based Classification. *J. Machine Learning Res.* **2001**, *2*, 175−211

(23) Willet P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(24) Wild, D. J.; Blankley, C. D. Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155−162.

(25) Salim, N.; Holliday J.; Willet P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−442.

(26) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of coefficients for the calculation of intermolecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screening* **2002**, *5*(*2*), 155−166.

(27) Platt, J. *Fast Training of Support Vector Machines using Sequential Minimal Optimization*; Microsoft Research Technical Report MSR-TR-98-14; 1998.

(28) Software and documentation from Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121-3752, U.S.A., www.accelrys.com.

(29) BCI Software and documentation from Barnard Chemical Information Ltd., www.bci.gb.com.

(30) Software and documentation from Daylight Chemical Information Systems, Inc., 27401 Los Altos − Suite 360, Mission Viejo, CA 92691 U.S.A., info@daylight.com.

(31) Software and documentation from MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, info@mdl.com.

(32) Software and documentation from Chemical Computing Group, Inc., 1010 Sherbrooke Street West, Suite 910 Montreal, Quebec, Canada H3A 2R7, info@chemcomp.com.

(33) Software and documentation from Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144-2913, U.S.A., www.tripos.com.

(34) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500. See also Web site on http://cdk.sourceforge.net/.

(35) This is not clear from the official v.5.2 documentation, but Accelrys has confirmed that a Tanimoto function is used for this similarity measure.

(36) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

(37) Joachims, T. 11 In *Making large-Scale SVM Learning Practical. Advances in Kernel Methods − Support Vector Learning*; Schölkopf, B., Burges, C., Smola, A., Ed.; MIT Press: 1999.

(38) Weiniger, D. Smiles, a Chemical Language and Information-System. 1 Introduction to methodology and encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(39) Compounds 216 and 1214 in the supporting information of ref 7 were excluded from the data set. The compound 216 is permanently charged, and we were unable to identify the structure of compound 1214.

(40) Yalkowsky, S. H.; Dannelfelser R. M. *The Arizona Database of Aqueous Solubility*; College of Pharmacy, University of Arizona, Tucson, AZ, 1990.

(41) Syracuse Research Corporation. *Physical*/*Chemical Property Database* (*PHYSPROP*); SRC Environmental Science Center, Syracuse, NY, 1994.

(42) 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.

(43) 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.

(44) Liu, R.; Sun, H.; 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.

(45) Kier L. B.; Hall L. B. *Molecular Structure Description − The Electrotopological State*; Academic Press: 1999.

(46) Cheng, A.; Merz, K. Patent WO 0300699, Pharmacopeia, Inc.

(47) Gao, H.; Shanmugasundaram, V.; Lee, P. Estimation of aqueous solubility of organic compounds with QSPR approach. *Pharm. Res.* **2002**, *19*(*4*), 497−503.

CI034107S