Comparative Analysis of Machine Learning Techniques for the Prediction of LogP

Edward W. Lowe, Jr., Mariusz Butkiewicz, Matthew Spellings, Albert Omlor, Jens Meiler

Abstract—**Several machine learning techniques were evaluated for the prediction of logP. The algorithms used include artificial neural networks (ANN), support vector machines (SVM) with the extension for regression, and kappa nearest neighbor (k-NN). Molecules were described using optimized feature sets derived from a series of scalar, two- and three-dimensional descriptors including 2-D and 3-D autocorrelation, and radial distribution function. Feature optimization was performed as a sequential forward feature selection. The data set contained over 25,000 molecules with experimentally determined logP values collected from the Reaxys and MDDR databases, as well as data mining through SciFinder. LogP, the logarithm of the equilibrium octanol-water partition coefficient for a given substance is a metric of the hydrophobicity. This property is an important metric for drug absorption, distribution, metabolism, and excretion (ADME). In this work, models were built by systematically optimizing feature sets and algorithmic parameters that predict logP with a root mean square deviation (rmsd) of 0.86 for compounds in an independent test set. This result presents a substantial improvement over XlogP, an incremental system that achieves a rmsd of 1.41 over the same dataset. The final models were 5-fold cross-validated. These fully *in silico* models can be useful in guiding early stages of drug discovery, such as virtual library screening and analogue prioritization prior to synthesis and biological testing. These models are freely available for academic use.**

I. Introduction

The process of modern drug design involves eliminating compounds with undesirable properties from the available chemical space while optimizing efficacy. The ability to predict properties which influence absorption, distribution, metabolism, and excretion of compounds prior to synthesis, such as the octanol-water partition coefficient (logP), could drastically reduce both the cost and time involved in drug discovery. Computational models can quickly assess the properties of large sets of compounds *in silico*. LogP, a measure of hydrophobicity or hydrophilicity of a molecule indicates whether a compound reaches a target protein as it influences the ability to cross the blood/brain barrier [1, 2]. It plays further a key role in the binding of a ligand to a target in aqueous solution [3].

This work is supported by 1R21MH082254 and 1R01MH090192 to Jens Meiler. Edward W. Lowe, Jr. acknowledges NIH support through the Integrative Training in Therapeutic Discovery training grant (T90DA022873; PI Larry Marnett).

Edward W. Lowe, Jr., PhD, is a post-doctoral fellow at the Center for Structural Biology at Vanderbilt University, Nashville TN, 37232;

Mariusz Butkiewicz is a graduate student in Chemistry at Vanderbilt University, Nashville TN, 37232;

Matthew Spellings is an undergraduate student in Chemical Engineering at Vanderbilt University; Nashville TN, 37232;

Albert Omlor is an undergraduate student in Chemistry at Saarland University, Germany;

Corresponding author: Jens Meiler, PhD, Center for Structural Biology, Vanderbilt University, Nashville, TN 37232.

978-1-4244-9897-0/11/$26.00 ©2011 IEEE

Formally, logP is the logarithm of the equilibrium ratio of concentrations of a compound in the organic and aqueous phases of an octanol-water system [4]. LogP is a widely used, well-defined property with experimental values available for large numbers of compounds, which makes it ideal for prediction by machine learning methods.

A well-established method for prediction of logP is XlogP [5], which assigns each atom in the molecule an empirically-determined contribution depending on its type and then sums these contributions for the logP estimation of the entire molecule. This incremental method resembles a multiple linear regression model. We test the hypothesis that logP has a nonlinear dependence on composition, charge distribution, and shape of the molecule. Therefore, we expect non-linear models to improve prediction accuracy.

Machine learning techniques have been successful in approximating nonlinear separable data in Quantitative Structure Property Relationship studies [6-9]. Here, we present several predictive models for logP using machine learning techniques including artificial neural networks (ANN) [10], support vector machines with the extension for regression estimation (SVR) [11], and kappa nearest neighbors (k-NN) [12].

II. Machine Learning Techniques

A. Artificial Neural Networks

The utility of ANNs for classification is well-known in chemistry and biology [13-16]. ANNs model the human brain and, thus, consist of layers of neurons linked by weighted connections w_{ji}. The input data x_i are summed according to their weights, activation function applied, and output used as the input to the j-th neuron of the next layer. For a three-layer feed forward ANN, such a training iteration would proceed as:

$$y_j = f(net_j) = f(\sum_{i=1}^{d} x_i w_{ji}) \quad 1 \le j \le n_H \tag{1}$$

$$z_k = f(net_k) = f(\sum_{j=1}^{n_H} y_j w_{kj}) \quad 1 \le k \le c \tag{2}$$

where $f(x)$ is the activation function, d is the number of features, n_H is the number of hidden neurons, and c is the number of outputs. For supervised training, the difference between the calculated output z_k, and the target value t_k, determines the errors for back-propagation:

$$\Delta w_{kj} = \eta(t_k - z_k) f'(net_k) y_j \tag{3}$$

$$\Delta w_{ji} = \eta\left[\sum_{k=1}^{c} w_{kj} (t_k - z_k) f'(net_k)\right] f'(net_j) x_i \tag{4}$$

The ANN training iterations produce weight changes that minimize the rmsd between the predicted and target values,

$$rmsd = \sqrt{\frac{\sum_{i=1}^{n}(exp_i - pred_i)^2}{n}} \quad (5)$$

which in this case is predicted and experimental logP values, respectively.

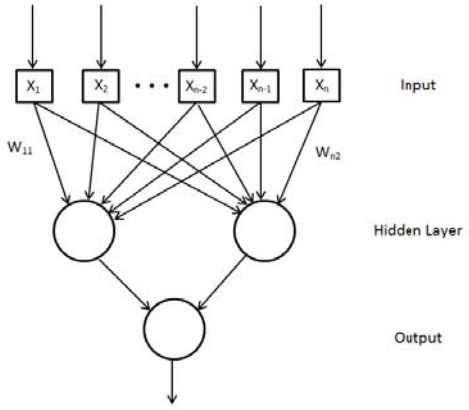

Fig 1: Schematic view of an ANN: Up to 1,142 descriptors are fed into the input layer. The weighted sum of the input data is modified by the activation function and serves as input to the next layer. The output describes the predicted logP value of the molecule.

In this study, the ANNs have up to 1142 inputs, 8 hidden neurons, and one output (logP). The activation function of the neurons is the sigmoid function:

$$g(x) = \frac{1}{1+e^{-x}} \quad (6)$$

B. Support Vector Machines

The second machine learning approach applied in this study is SVM learning with extension for regression estimation [17, 18]. Linear functions defined in high-dimensional feature space [19], risk minimization according to Vapnik's ε - intensive loss function, and structural risk minimization [20] which minimizes the risk function consisting of the empirical error and the regularized term are the core principles integrated in SVM regression estimation.

The training data is defined by $(x_i \in X \subseteq R^n, y_i \in Y \subseteq R)$ with $i = 1, \dots, l$ where l is the total number of available input data pairs consisting of molecular descriptor data and experimental logP value. The following function defines a linear plane in a high-dimensional space for SVM estimation:

$$f(x, w) = w * \phi(x) + b \quad (7)$$

where $\phi(x)$ describes a nonlinear transformation function as a distance measure in an input space X. The parameter w describes a normal vector perpendicular to the separating hyperplane whereas b is a bias parameter. Both parameters are optimized by estimating the minimum of Vapnik's linear loss function as a measure of the error of approximation:

$$|y - f(x, w)| = \begin{cases} 0, & if\ |y - f(x, w)| \le \varepsilon \\ |y - f(x, w)| - \varepsilon, & otherwise \end{cases} \quad (8)$$

The error is zero if the difference between the measured value y and the predicted value $f(x, w)$is less than a given threshold ε. Thus, Vapnik's insensitivity loss function defines an ε - tube (Fig 2).

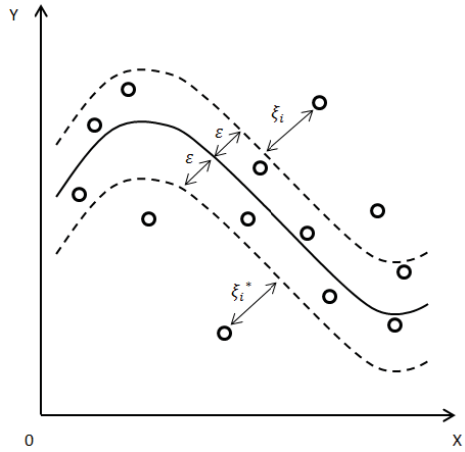

Fig 2: Schematic view of a Support Vector ε - tube : data points in ε – tube are not penalized, while points outside the tube get a penalty according to their distance from the tube edge.

Predicted values positioned within the ε - tube have an error of zero. In contrast, data points outside the tube are panelized by the magnitude of the difference between the predicted value and the outer rim of the tube. The regression problem is solved by minimizing function L:

$$L_{w,\xi,\xi^*} = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{l}\xi_i + \sum_{i=1}^{l}\xi_i^*\right) \quad (9)$$

under constraints:

$$\begin{aligned} &y_i - g(x, w) \le \varepsilon + \xi_i \\ &g(x, w) - y_i \le \varepsilon + \xi_i^* \\ &\xi_i^{(*)} \ge 0 \qquad i = 1, \dots, l \end{aligned} \quad (10)$$

where the slack variables ξ_i and $\xi_i^*$are shown in Fig 2 for measurements above and below an ε - tube, respectively. Both slack variables are positive values and their magnitude can be controlled by penalty parameter C. To estimate a numerical solution the optimization problem is converted into the given dual problem by:

$$f(x) = \sum_{i=1}^{N_{SV}} (\alpha_i - \alpha_i^*) * K(x_i, x) + b$$
$$0 \le \alpha_i \le C,$$
$$0 \le \alpha_i^* \le C \quad (11)$$

where α_i and α_i^* define Lagrange multipliers associated with ξ_i and ξ_i^*, N_{SV} shows the number of support vectors SV defining the SVM and $K(x_i, x_j)$ denotes the kernel function. In this study the Radial Basis Function kernel

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\gamma^2}} \quad (12)$$

was used to train the Support Vector Machine. The influence of the approximation error and the weight vector $\|w\|$ norm is balanced by the penalty constant C. It is optimized along with kernel parameter γ by using a grid search approach on a monitoring dataset.

C. Kappa Nearest Neighbors

The third machine learning approach utilized in this research is the k-NN [12, 21-24]. k-NNs are considered an unsupervised learning algorithm. This method uses a distance function to calculate pair-wise distances between query points and reference points, where query points are those to be classified (Fig 3). The predicted value of a query point is then that of the weighted average of its *kappa* nearest reference points. In this research, the distance measure was the Euclidean distance between feature vectors:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (13)$$

The reference activities were weighted as $\frac{1}{\text{distance}}$, and the value of *kappa* was 5.

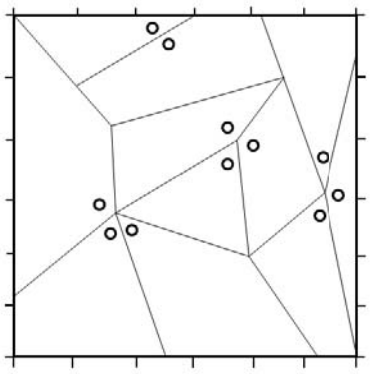

Fig 3: Schematic view of k-NN cluster centers with its determined nearest neighbor environments

III. Training Data

The octanol-water partition coefficient, or partition constant, is a measure of differential solubility of a substance. Specifically, it is the ratio of concentrations of a substance in the two phases of a mixture of the two immiscible solvents. These partition constants are useful in approximating the distribution of a substance within the body in medicinal settings. In pharmacology, logP is an indicator of how easily a substance can reach its intended target, and influences absorption, distribution, metabolism, and excretion (ADME) properties of the substance. Thus, certain ranges of logP are desirable depending on the intended target and interaction of the substance making logP an important property in drug discovery. The ability to accurately predict this property is beneficial in the early stages of therapeutic design, such as during virtual high-throughput screening efforts and in analogue prioritization and optimization.

The training data for this investigation was obtained through data mining of the MDL Drug Data Report (MDDR) [25] and Reaxys [26] databases, as well as through literature searches using SciFinder [27]. Data mining resulted in ~26000 compounds with experimentally determined values of logP. Of the compounds retrieved, the values of logP ranged from approximately -14 to 14. From this range, 13% of the compounds were removed from the training set from both extrema reaching a range of -5 to 8 in order to eliminate possible outliers at the limits of logP determination.

The remaining molecules in the training data set were numerically encoded using a series of transformation-invariant descriptors which serve as unique fingerprints. The descriptors (Table I) were calculated using in-house code.

IV. Implementation / Method

All machine learning algorithms, and descriptor calculations used for this study were implemented in our in-house C++ class library, the BioChemistryLibrary (BCL). A third-party 3D conformation generator, CORINA [28], was used to generate 3D coordinates for the molecules prior to descriptor calculation.

A. Dataset Generation

The data set used in this study was obtained through data mining and filtering which resulted in a final data set of 22,582 molecules. During the training of the models, 10% of the data set was used for monitoring and 10% were used for independent testing of the trained models, leaving 80% for the training data set.

B. Quality Measure

The machine learning methods are evaluated by calculating the rmsd (eq. 5) using the cross-validated models.

The average rmsd of the cross-validated models for a feature set with n features is used to determine the predictive ability of the model. Additionally, correlation plots and the calculation of r^2 are also used to evaluate the trained models.

$$r^2 = \left(\frac{n\sum(exp*pred) - \sum exp \sum pred}{\sqrt{[n\sum(exp^2) - (\sum exp)^2][n\sum(pred^2) - (\sum pred)^2]}} \right)^2 \quad (14)$$

C. Feature Selection

1142 descriptors in 54 categories were generated using the BCL. The 54 categories consisted of scalar descriptors, as well as 2D and 3D autocorrelation functions, radial distribution functions, and van der Waals surface area

weighted variations of each of the non-scalar descriptors (see Table I).

TABLE I
THE ORIGINAL MOLECULAR DESCRIPTORS BY CATEGORY

	Descriptor Name	Description
Scalar descriptors	Weight	Molecular weight of compound
	HDon	Number of hydrogen bonding acceptors derived from the sum of nitrogen and oxygen atoms in the molecule
	HAcc	Number of hydrogen bonding donors derived from the sum of N-H and O-H groups in the molecule
	TPSA	Topological polar surface area in [Å^2] of the molecule derived from polar 2D fragments
Vector descriptors	Ident	weighted by atom identities
2D Autocorrelation	SigChg	weighted by σ atom charges
(11 descriptors) /	PiChg	weighted by π atom charges
3D Autocorrelation	TotChg	weighted by sum of σ and π charges
(12 descriptors) /	SigEN	weighted by σ atom electronegativities
Radial Distribution	PiEN	weighted by π atom electronegativities
Function	LpEN	weighted by lone pair electronegativities
(48 descriptors)	Polariz	weighted by effective atom polarizabilities

Every Vector descriptor available with and without van der Waals surface area weighting

Sequential forward feature selection [29] was used for feature optimization for each machine learning technique individually. Additionally, each feature set was trained with 5-fold cross-validation. To cope with the computational expense associated with this thorough feature selection process, a subset of the training data was used taking randomly only 8000 of the 22582 molecules. The number of models generated during this process for each training method was $\sum_{i=1}^{54} 5(n-1)$. Upon identification of the optimized feature set for each algorithm, any algorithm-specific parameters were optimized using the entire training data set and using 5-fold cross-validation.

V. RESULTS

ANNs were trained using 7500 iterations of simple propagation evaluating the rmsd every 100 steps during the feature optimization process. For the training of the final model with the optimized feature set, 100,000 iterations were performed evaluating the rmsd every 500 steps.The weight matrices were randomly initialized with values in the range of [-0.1, 0.1] if the rmsd of the monitoring data set had not improved in the last 10000 iterations. The ANN algorithm runs on graphics processing units (GPUs) using OpenCL implemented within the BCL. The training time was 28 minutes per final network on a C2050 NVidia GPU on a Dell T3500 with 8-core Xeon 3.2 GHz microprocessor running CentOS 5 64-bit. For the best model, an rmsd of 1.20 for the independent data set was achieved.

SVMs were trained using a C of 1.0 and γ of 0.1 during the feature optimization process. Upon identification of the optimal feature set, the cost and γ parameters were optimized to 0.1 and 0.1, respectively. The training time for the final model was 12 minutes using 6 cores. Using these optimized parameters, the cross-validated model achieved an rmsd for the independent data set of 1.21.

The k-NN algorithm was used to predict the logP values of the training, monitoring, and independent data sets. The value of kappa, the number of neighbors to consider, was optimized with the full data set using the optimized feature set determined during the feature selection process. The prediction time for the final model was 0.75 minutes on 8 cores. Using an optimal kappa of 5, the relative rmsd achieved for the independent data set was 1.03.

TABLE II
MODEL STATISTICS FOR BEST PREDICTORS

Method	RMSD	R^2
ANN	1.20	0.70
SVM	1.21	0.67
k-NN	1.03	0.72
XlogP	1.41	0.56

In order to further evaluate the resulting models, cross correlation plots were created on the independent data sets of each model and compared with that of the XlogP algorithm [5].

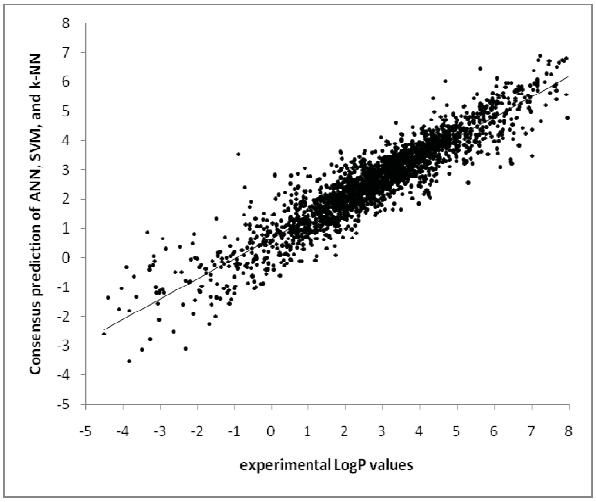

Fig 4 : Correlation plot for best consensus prediction using ANN/SVM/k-NN compared with Reaxys/MDDR experimental values

Consensus predictors were also created by averaging the predictions using different combinations of models. This yields better results with the best rmsd of 0.86 achieved using ANN/SVM/k-NN models (Table III).

TABLE III
CONSENSUS PREDICTORS

Method	RMSD	R^2
ANN/SVM	1.04	0.8
ANN/k-NN	1.06	0.81
SVM/k-NN	0.99	0.77
ANN/SVM/k-NN	0.86	0.86

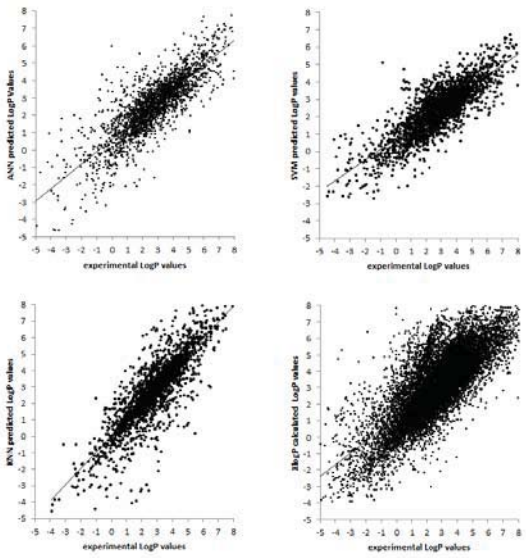

Fig 5 : ANN, SVM, k-NN, and XlogP compared with Reaxys/MDDR experimental values.

VI. CONCLUSIONS

Here, we present the utility of a series of machine learning techniques for the construction of predictive models capable of predicting logP, the water-octanol partition coefficient, with high accuracy compared to XlogP. The XlogP algorithm is a standard and is used by the NIH-funded PubChem database as well as the Chemical Abstracts Service. We have shown that the three models using artificial neural networks, support vector machines, and kappa nearest neighbors outperform this method as do all of the consensus models.

The best individual model was found to be that of the k-NN. This is likely due to the k-NNs ability to predict only around the cluster space of kappa neighbors while the other machine learning methods take the entire sample space into account when training. This all-inclusive training allows the SVM and ANN to over-train on the mean due to underrepresentation of the extrema. This is reflected by the slope of the best-fit linear regression lines in Fig 5. The best performing method was the consensus prediction of ANN/SVM/k-NN models which achieved an rmsd of 0.86. This is likely due to the ANNs and SVMs ability to predict the core range of the logP space more accurately while the k-NN was more accurate at the extrema. This consensus predictor will be made freely available through a web-interface accessible through www.meilerlab.org .

VII. REFERENCES

[1] J. Kai, K. Nakamura, T. Masuda, I. Ueda, and H. Fujiwara, "Thermodynamic aspects of hydrophobicity and the blood-brain barrier permeability studied with a gel filtration chromatography," *J Med Chem,* vol. 39, pp. 2621-4, Jun 21 1996.

[2] M. H. Abraham, W. E. Acree, Jr., A. J. Leo, D. Hoekman, and J. E. Cavanaugh, "Water-solvent partition coefficients and Delta Log P values as predictors for blood-brain distribution; application of the Akaike information criterion," *J Pharm Sci,* vol. 99, pp. 2492-501, May.

[3] S. Miyamoto and P. A. Kollman, "What determines the strength of noncovalent association of ligands to proteins in aqueous solution?," *Proc Natl Acad Sci U S A,* vol. 90, pp. 8402-6, Sep 15 1993.

[4] A. Leo, C. Hansch, and D. Elkins, "Partition coefficients and their uses," *Chemical Reviews,* vol. 71, pp. 525-616, 1971.

[5] R. Wang, Y. Fu, and L. Lai, "A New Atom-Additive Method for Calculating Partition Coefficients," *Journal of Chemical Information and Computer Sciences,* vol. 37, pp. 615-621, 1997.

[6] V. V. Zernov, K. V. Balakin, A. A. Ivaschenko, N. P. Savchuk, and I. V. Pletnev, "Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions," *J Chem Inf Comput Sci,* vol. 43, pp. 2048-56, Nov-Dec 2003.

[7] A. Bleckmann and J. Meiler, "Epothilones: Quantitative Structure Activity Relations Studied by Support Vector Machines and Artificial Neural Networks," *QSAR Comb. Sci.,* vol. 22, pp. 719-721, 2003.

[8] R. Mueller, A. L. Rodriguez, E. S. Dawson, M. Butkiewicz, T. T. Nguyen, S. Oleszkiewicz, A. Bleckmann, C. D. Weaver, C. W. Lindsley, P. J. Conn, and J. Meiler, "Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening," *ACS Chemical Neuroscience*.

[9] M. Butkiewicz, R. Mueller, D. Selic, E. Dawson, and J. Meiler, "Application of Machine Learning Approaches on Quantitative Structure Activity Relationships," in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, K. C. Wiese, Ed. Nashville, 2009.

[10] D. Winkler, "Neural networks as robust tools in drug lead discovery and development," *Molecular Biotechnology,* vol. 27, pp. 139-167, 2004.

[11] B. Schoelkopf, "SVM and Kernel Methods," *www,* 2001.

[12] M. Shen, Y. Xiao, A. Golbraikh, V. K. Gombar, and A. Tropsha, "Development and Validation of k-Nearest-Neighbor QSPR Models of Metabolic Stability of Drug Candidates," *Journal of Medicinal Chemistry,* vol. 46, pp. 3013-3020, 2003.

[13] T. Fox and J. M. Kriegl, "Machine learning techniques for in silico modeling of drug metabolism," *Curr Top Med Chem,* vol. 6, pp. 1579-91, 2006.

[14] J. Meiler, "PROSHIFT: Protein Chemical Shift Prediction Using Artificial Neural Networks," *J. Biomol. NMR,* vol. 26, pp. 25-37, 2003.

[15] W. P. Walters and M. A. Murcko, "Prediction of 'drug-likeness'," *Adv Drug Deliv Rev,* vol. 54, pp. 255-71, Mar 31 2002.

[16] I. V. Tetko, V. V. Kovalishyn, and D. J. Livingstone, "Volume Learning Algorithm Artificial Neural Networks for 3D QSAR Studies," *Journal of Medicinal Chemistry,* vol. 44, pp. 2411-2420, 2001.

[17] J. S. Alex and S. Bernhard, *A tutorial on support vector regression*: Kluwer Academic Publishers, 2004.

[18] B. C. Drucker H and V. V, "Support vector regression machines."

[19] B. Schoelkopf and A. J. Smola, *Learning with Kernels*. Cambridge, Massachusetts: The MIT Press, 2002.
[20] V. Vapnik, *The Nature of Statistical Learning Theory (Information Science and Statistics)*: Springer, 1999.
[21] C. Yan, J. Hu, and Y. Wang, "Discrimination of outer membrane proteins using a K-nearest neighbor method," *Amino Acids,* vol. 35, pp. 65-73, Jun 2008.
[22] B. F. Jensen, C. Vind, S. B. Padkjaer, P. B. Brockhoff, and H. H. Refsgaard, "In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors," *J Med Chem,* vol. 50, pp. 501-11, Feb 8 2007.
[23] F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch, and J. B. Mitchell, "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization," *J Chem Inf Model,* vol. 46, pp. 2412-22, Nov-Dec 2006.
[24] S. Ajmani, K. Jadhav, and S. A. Kulkarni, "Three-dimensional QSAR using the k-nearest neighbor method and its interpretation," *J Chem Inf Model,* vol. 46, pp. 24-31, Jan-Feb 2006.
[25] "MDDR."
[26] "Reaxys."
[27] A. B. Wagner, "SciFinder Scholar 2006: an empirical analysis of research topic query processing," *J Chem Inf Model,* vol. 46, pp. 767-74, Mar-Apr 2006.
[28] J. Gasteiger, C. Rudolph, and J. Sadowski, "Automatic Generation of 3D-Atomic Coordinates for Organic Molecules," *Tetrahedron Comput. Method.,* vol. 3, pp. 537-547, 1992.
[29] K. Z. Mao, "Orthogonal forward selection and backward elimination algorithms for feature subset selection," *IEEE Trans Syst Man Cybern B Cybern,* vol. 34, pp. 629-34, Feb 2004.