Comparative Analysis of Machine Learning Techniques for the Prediction of the DMPK Parameters Intrinsic Clearance and Plasma Protein Binding

Edward W. Lowe, Jr., Mariusz Butkiewicz, Zollie White III, Matthew Spellings, Albert Omlor, Jens Meiler
Department of Chemistry
Vanderbilt University
Nashville, TN 37235 USA
jens@jens-meiler.de

Abstract

Several machine learning techniques were evaluated for the prediction of parameters relevant in pharmacology and drug discovery including rat and human microsomal intrinsic clearance as well as plasma protein binding represented as the fraction of unbound compound. The algorithms assessed in this study include artificial neural networks (ANN), support vector machines (SVM) with the extension for regression, kappa nearest neighbor (KNN), and Kohonen Networks. The data sets, obtained through literature data mining, were described through a series of scalar, two- and three-dimensional descriptors including 2-D and 3-D autocorrelation, and radial distribution function. The feature sets were optimized for each data set individually for each machine learning technique using sequential forward feature selection. The data sets range from 400 to 600 compounds with experimentally determined values. Intrinsic clearance (CL_{int}) is a measure of metabolism by cytochrome P-450 enzymes primarily in the vesicles of the smooth endoplasmic reticulum. These important enzymes contribute to the metabolism of an estimated 75% of the most frequently prescribed drugs in the U.S. The fraction of unbound compound (*fu)* greatly influences pharmacokinetics, efficacy, and toxicology. In this study, machine learning models were constructed by systematically optimizing feature sets and algorithmic parameters to calculate these parameters of interest with cross validated correlation/RMSD values reaching 9.53 over the normalized data set. These fully *in silico* models are useful in guiding early stages of drug discovery, such as analogue prioritization prior to synthesis and biological testing while reducing costs associated with the *in vitro* determination of these parameters. These models are made freely available for academic use.

1 Introduction

During the 1990's, poor pharmacokinetic and bioavailability properties accounted for approximately 40% of drug candidate attrition during human trials[1]. A decade later, these properties accounted for approximately 10% of attrition during human trials due to the implementation of early determination (pre-clinical) of drug metabolism and pharmacokinetics (DMPK) properties in the drug discovery workflow through both *in vitro* and *in vivo* studies. While many proposed new chemical entities (NCE) are now eliminated in earlier stages, these preliminary studies are time consuming and add to the mounting real and opportunity costs which now approach an estimated \$1.30 – \$1.76 billion [2, 3]. Thus, the use of *in silico* models for the prediction of these DMPK properties trained on existing data would increase the efficiency of the drug discovery process while mitigating the costs[4]. Indeed, computational models can quickly assess large data sets of proposed molecules for DMPK parameters[5].

Two important DMPK properties which are routinely determined in drug discovery are microsomal intrinsic clearance (CL_{int}) and plasma protein binding as the fraction of unbound compound (*fu*). CL_{int} is a measure of metabolism primarily by cytochrome P-450 (CYP) enzymes in the vesicles of the smooth endoplasmic reticulum of hepatocytes. CYP enzymes contribute to the metabolism of approximately 75% of the top 200 most prescribed drugs in the United States[6]. *fu* is an indication of the extent to which a compound binds to plasma proteins which influences to a large degree pharmacokinetics, efficacy, and toxicology *in vivo* [7, 8].

Previous work has proven machine learning techniques useful in the approximation of nonlinear separable data in Quantitative Structure Property Relationship (QSPR) studies [5, 9-12]. Here, we present several predictive models based on machine learning techniques for human and rat CL_{int} as well as human and rat *fu*. The machine learning techniques used include artificial neural networks[13], support vector machine with the extension for regression[14], kappa nearest neighbor[15], and Kohonen networks[16].

2 Methods

All descriptors calculated and machine learning algorithms used in this study are implemented in the in-house C++ class library, the BioChemistry Library (BCL). CORINA, a 3rd-party 3D conformation generator, was used for the generation of 3D coordinates prior to descriptor calculations [17].

2.1 Machine Learning Techniques

2.1.a Artificial Neural Network

The utility of artificial neural networks (ANN) for classification is proven in chemistry and biology [18-21]. ANNs model the human brain, consisting of layers of nodes linked by weighted connections w_{ji}. Input data x_i are summed by their weights, followed by the application of an activation function, and the output used as the input to the j-th neuron of the next layer.

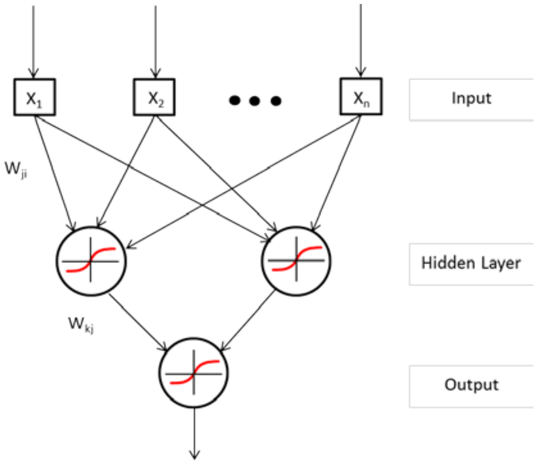

Figure 1: Schematic view of an ANN: Up to 1,284 descriptors are used as training input. The activation function is applied to the weighted sum of the input data which serves as input to the next layer. The output describes the predicted value of that feature.

For a common feed forward ANN with a single hidden layer, the training iteration would proceed as:

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

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

where $f(x)$ is the activation function (commonly sigmoid), d is the number of features, n_H is the number of hidden neurons, and c is the number of outputs. The difference between the calculated output z_k, and the target value t_k, provides the errors for back-propagation through the network:

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

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

The weight changes produced attempt to minimize the objective function (RMSD) between the predicted and target values,

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

In this study, the ANNs have up to 1284 inputs, 8 hidden neurons, and one output (DMPK parameter of interest). The activation function of the neurons is the sigmoid function:

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

2.1.b Support Vector Machine with extension for regression estimation

SVM learning with extension for regression estimation [14] represents a supervised machine learning approach successfully applied in the past [5, 11, 12]. The core principles in SVR lay in linear functions defined in high-dimensional feature space [22], risk minimization according to Vapnik's ε - intensive loss function, and structural risk minimization [23] of a risk function consisting of the empirical error and the regularized term.

The training data is described 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 the experimental DMPK property.

Given a defined error threshold ε, SVR seeks to approximates Vapnik's insensitivity loss function through the definition of an ε – tube incorporating all data points of the given problem (see Fig X). The error is zero if the difference between the experimentally measured value and the predicted value is less than ε.

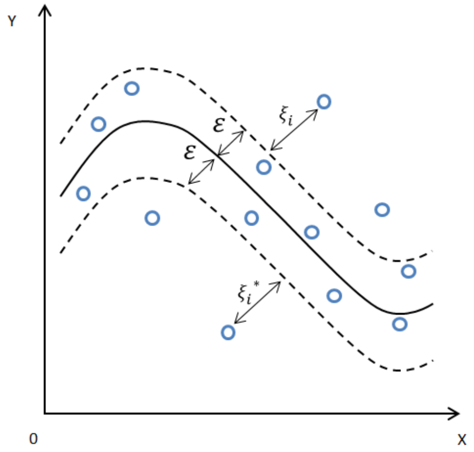

Figure 2: schematic depiction of a support vector epsilon tube. A function is approximated to incorporate as many data points as possible in an ε - wide tube. Outliers are penalized depending on the distance to the edge of the tube.

Predicted values positioned within the ε - tube have an assigned error value of zero. On the opposite, data points outside the tube are panelized by the distance of the predicted value from the edge of the tube. The solution to the regression problem is obtained by minimizing the following function L:

$$L_{w,\xi,\xi^*} = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{l} \xi_i + \sum_{i=1}^{l} \xi_i{}^*\right) \qquad (7)$$

under constraints:

$$y_i - g(x,w) \le \varepsilon + \xi_i \quad , \quad g(x,w) - y_i \le \varepsilon + \xi_i{}^*$$

$$\text{and} \quad \xi_i{}^{(*)} \ge 0 \qquad i = 1, \dots, l$$

where the parameter w describes a normal vector perpendicular to the separating hyperplane in the higher dimensional space. The slack variables ξ_i and $\xi_i^*$are shown in Fig X for measurements above and below an ε - tube, respectively. Both slack variables are positive values and their magnitude can be controlled by the penalty parameter C. In this study the Radial Basis Function kernel

$$K(x_i, x_j) = e^{-\frac{\left\|x_i - x_j\right\|^2}{2\gamma^2}} \qquad (8)$$

was applied as distance measure. The penalty constant C determined the influence of the approximation error penalty. A grid search approach was conducted to optimize both parameters γ and C using a monitoring dataset.

2.1.c Kappa Nearest Neighbor

Kappa nearest neighbor (KNN) was also utilized in this study[15, 24-27]. KNNs are an unsupervised learning algorithm using a distance function to calculate pair-wise distances between query points and reference points. Query points are those to be classified (Fig 3). The query point is classified through a weighted average of the known output of its *kappa* nearest reference points. The distance measure used in this study was the Euclidean distance measure between feature vectors:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (9)$$

The reference activities were weighted as $\frac{1}{\mathrm{d(x,y)}}$, and the value of *kappa* was optimized for each data set.

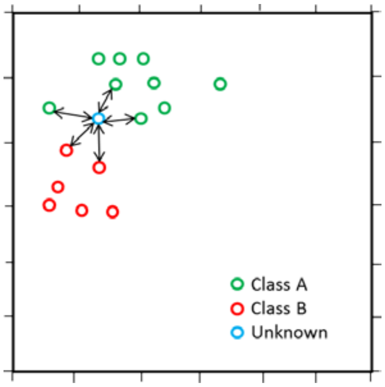

Figure 3: Schematic view of KNN classification with k = 5 nearest neighbors

2.1.d Kohonen Network

The kohonen network represents an unsupervised learning algorithm. It is conceptually derived from artificial neural networks consisting of one input layer connected by weighted connections with a two dimensional grid of neurons, the kohonen network [28].

The training data defined by pairs of numerical molecular descriptor data x_i and the respective experimental DMPK parameter y_i $(x_i \in X \subseteq R^n, y_i \in Y \subseteq R)$ with $i = 1, \dots, l$ where l is the total number of available input data pairs.

For every training data point, a node most similar to the data point is determined for placement in the grid. Weight vectors are updated using the Gaussian kernel as a neighborhood function. A radius of four neighboring nodes is considered.

To determine the classification result of an unknown compound, the most similar node is determined and the average prediction values of all neighboring nodes are computed.

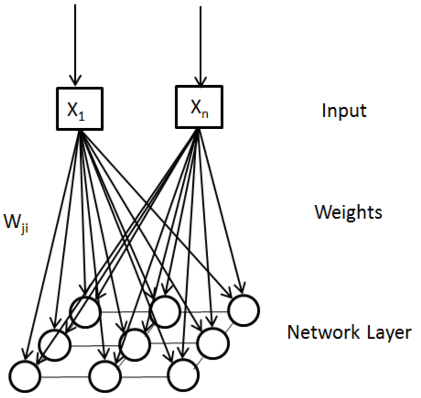

Figure 4: Schematic view of a kohonen network showing the connectivity and involved network layer.

2.2 Data Set Generation

The data sets used in this study were obtained through literature search and data mining using the reaxys and pubmed. The data sets ranged from 386 (human *fu*) to 601 (rat $\mathrm{CL_{int}}$) as seen in Table I.

TABLE I
DATA SET COMPOSITION

Data Set	Number Molecules
Rat *fu*	388
Human *fu*	386
Rat $\mathrm{CL_{int}}$	601
Human $\mathrm{CL_{int}}$	576

Three-dimensional conformations for the molecules were generated. The molecules in the data sets were then numerically encoded using transformation-invariant descriptors (Table II) which represent features for the machine learning techniques.

2.3 Quality Measures

The calculated RMSD (eq. 5) is used to evaluate the predictive power of the machine learning models. Specifically, the average RMSD of the cross-validated models for a feature set with **n** features is used. Additionally, the Pearson (r_p) and Spearman (r_s) correlation coefficients, are computed.

$$r_{p/s} = \frac{n\sum(exp*pred) - \sum exp \sum pred}{\sqrt{\left[n\sum(exp^2) - (\sum exp)^2\right]\left[n\sum(pred^2) - (\sum pred)^2\right]}} \quad (10)$$

In contrast to the Pearson correlation coefficient, the Spearman correlation coefficient takes the ranking of two dependent variables into account rather than their actual value.

TABLE II
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
	VCharge	weighted by VCharge atom charges
(12 descriptors) /	SigEN	weighted by σ atom electronegativities
Radial Distribution	PiEN	weighted by π atom electronegativities
Function (48 descriptors)	LpEN	weighted by lone pair electronegativities
	Polariz	weighted by effective atom polarizabilities

All molecular fingerprints are considered with and without van der Waals surface area weighting

2.4 Feature Selection

The BCL was used to generate 1284 descriptors in 60 categories. The 60 categories consist of scalar, 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 II).

Sequential forward feature selection [29] was used for feature optimization for each machine learning technique individually. Each feature set was trained with 5-fold cross-validation. The number of models generated during this process for each training method was $cv * \frac{n(n+1)}{2}$ where n is the number of feature categories and cv is the number of cross validations. Thus, 9150 models were trained for each machine learning algorithm on each data set during feature selection. Upon identification of the optimized feature set for each algorithm, algorithm-specific parameters were then optimized using a grid search with 5-fold cross-validation.

3 Results

During feature selection, ANNs were trained for 100 epochs of simple back-propagation using η = 0.1 and α = 0.5 with weight updates and the evaluation of RMSD every step using 5-fold cross validation. Weight matrices were initialized randomly with values in the range [-0.1, 0.1]. A grid search was performed to optimize eta and alpha parameters using the optimized feature set and trained 100 epochs using 5-fold cross validation (Table III). Eighty percent of each data set was used as the training set while 10% was used for the monitoring data set and 10% for the test data set.

TABLE III
OPTIMIZED PARAMETERS

Data Set	**ANN (η/α)**	**SVM (*C*/γ)**	**KNN (*k*)**
Rat *fu*	0.25/0.015625	2.0/0.25	5
Human *fu*	0.125/0.5	2.0/0.03125	24
Rat CL$_{int}$	0.03125/0.03125	0.25/0.25	14
Human CL$_{int}$	0.03125/0.0625	1.0/0.03125	16

SVMs were trained using a *C* of 0.1 and γ of 0.5 during the feature optimization process. Upon identification of the optimal feature set, the cost and γ parameters were optimized using a grid search approach (Table III). The SVMs were trained 100 iterations and used 5-fold cross validation. Each iteration step accumulated up to 200 support vectors.

TABLE IV
MODEL CORRELATION RESULTS $\frac{r_S}{RMSD}$ ($\frac{r_p}{RMSD}$) FOR INDEPENDENT VALIDATION

Machine Learning Method(s)	**Human *fu***	**Rat *fu***	**Human CL$_{int}$**	**Rat CL$_{int}$**
ANN	8.87 (8.47)	9.2 (9.52)	1.34 (1.27)	1.46 (1.37)
ANN/KNN	**10.47** (9.74)	**9.67** (9.41)	1.55 (1.48)	1.73 (1.67)
ANN/KNN/Kohonen	9.69 (9.24)	9.49 (9.26)	1.65 (1.58)	**1.79**(1.71)
ANN/KNN/Kohonen/SVM	9.47 (9.48)	9.16 (8.92)	1.63 (1.57)	1.6 (1.53)
ANN/KNN/SVM	9.6 (9.81)	9.03 (8.80)	1.57 (1.51)	1.49 (1.43)
ANN/Kohonen	8.65 (8.41)	9.1 (9.33)	1.58 (1.52)	1.69 (1.6)
ANN/Kohonen/SVM	8.36 (8.80)	8.61 (8.74)	1.57 (1.51)	1.44 (1.37)
ANN/SVM	7.75 (8.72)	8.04 (8.31)	1.44 (1.39)	1.2 (1.12)
KNN	10.31 (9.62)	8.74 (8.28)	1.45 (1.38)	1.57 (1.54)
KNN/Kohonen	9.29 (9.03)	8.91 (8.59)	1.67 (1.6)	1.69 (1.64)

KNN/Kohonen/SVM	9.08 (9.37)	8.46 (8.18)	**1.67**(1.61)	1.48 (1.44)
KNN/SVM	8.9 (9.63)	7.84 (7.47)	1.61 (1.53)	1.29 (1.27)
Kohonen	7.42 (7.56)	7.79 (8.22)	1.62 (1.57)	1.51 (1.45)
Kohonen/SVM	7.24 (8.28)	7.1 (7.49)	1.61 (1.56)	1.22 (1.16)
SVM	5.04 (6.64)	3.35 (2.48)	1.46 (1.41)	0.03 (0.04)

KNNs were trained by optimizing kappa, the number of neighbors to consider, during feature selection from k=1 to k=25 using 5-fold cross validation (Table III). Eighty percent of each data set was used as reference features while 10% were queried as the monitoring data set. The remaining 10% was then used as a test set.

Kohonen networks were trained using a grid of 10x10 nodes. An ensemble model approach was then investigated by using all combinations of optimized models to arrive at single best predictors for each data set. This approach provided the best models in terms of correlation. The resulting $\frac{r_P}{RMSD}$ and $\frac{r_S}{RMSD}$ values are listed in Table IV with correlations plots shown for each of the best predicting models in Figure 5.

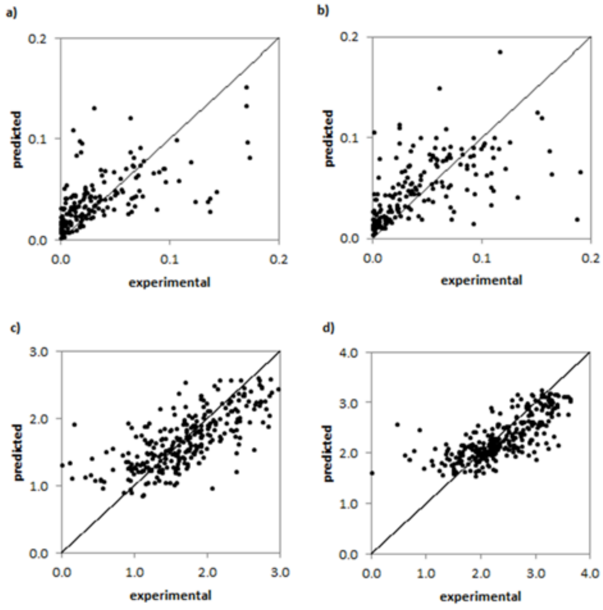

Figure 5: Correlation plots are shown for the machine learning models with the highest predictive power as determined by $\frac{r_p}{RMSD}$ for a) human *fu (ANN/KNN)*, b) rat *fu (ANN/KNN)*, *c)* human CL$_{int}$ (KNN/Kohonen/SVM), and d) rat CL$_{int}$ (ANN/KNN/Kohonen) , respectively. The selected models are shown in bold in Table IV.

4 Conclusion

Here, we present ensemble models based on machine learning techniques capable of predicting several parameters relevant to drug discovery. We have shown that ensemble models are in some cases capable of outperforming single algorithms using artificial neural networks, support vector machines, kappa nearest neighbors, and kohonen networks.

KNN algorithm consistently performs very well for all 4 data sets examined. The predictors constructed during this study compare favorably against recent studies[30] and are of great utility in early drug discovery. The top scoring predictors for human and rat *fu*, and human and rat CL$_{int}$ are ANN/KNN, ANN/KNN, KNN/ Kohonen/SVM, and ANN/KNN/Kohonen, respectively. These predictors will be made freely available through a web-interface accessible through www.meilerlab.org.

5 References

1. Kola, I. and J. Landis, *Can the pharmaceutical industry reduce attrition rates?* Nat Rev Drug Discov, 2004. 3(8): p. 711-716.
2. DiMasi, J.A., R.W. Hansen, and H.G. Grabowski, *The price of innovation: new estimates of drug development costs.* Journal of health economics, 2003. 22(2): p. 151-85.
3. Paul, S.M., et al., *How to improve R&D productivity: the pharmaceutical industry's grand challenge.* Nature reviews. Drug discovery, 2010. 9(3): p. 203-14.
4. Richon, A.B., *Current status and future direction of the molecular modeling industry.* Drug Discovery Today, 2008. 13(15-16): p. 665-9.
5. Lowe, E.W., Jr., et al., *Comparative Analysis of Machine Learning Techniques for the Prediction of LogP (Accepted)*, in *SSCI 2011 CIBCB - 2011 Symposium on Computational Intelligence in Bioinformatics and Computational Biology*2011: Paris, France.
6. Williams, J.A., et al., *Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios.* Drug metabolism and disposition: the biological fate of chemicals, 2004. 32(11): p. 1201-8.
7. Schmidt, S., D. Gonzalez, and H. Derendorf, *Significance of protein binding in pharmacokinetics and pharmacodynamics.* Journal of pharmaceutical sciences, 2010. 99(3): p. 1107-22.
8. Berezhkovskiy, L.M., *On the influence of protein binding on pharmacological activity of drugs.* Journal of pharmaceutical sciences, 2010. 99(4): p. 2153-65.
9. Zernov, V.V., et al., *Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions.* J Chem Inf Comput Sci, 2003. 43(6): p. 2048-56.

10. Bleckmann, A. and J. Meiler, *Epothilones: Quantitative Structure Activity Relations Studied by Support Vector Machines and Artificial Neural Networks.* QSAR Comb. Sci., 2003. 22(7): p. 719-721.

11. Butkiewicz, M., et al., *Application of Machine Learning Approaches on Quantitative Structure Activity Relationships*, in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, K.C. Wiese, Editor 2009: Nashville.

12. Mueller, R., et al., *Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening.* ACS Chemical Neuroscience, 2010. 1(4): p. 288-305.

13. Winkler, D.A., *Neural networks as robust tools in drug lead discovery and development.* Molecular biotechnology, 2004. 27(2): p. 139-68.

14. Schoelkopf, B., *SVM and Kernel Methods.* www, 2001.

15. Shen, M., et al., *Development and Validation of k-Nearest-Neighbor QSPR Models of Metabolic Stability of Drug Candidates.* Journal of Medicinal Chemistry, 2003. 46(14): p. 3013-3020.

16. Korolev, D., et al., *Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach.* J Med Chem, 2003. 46(17): p. 3631-43.

17. Gasteiger, J., C. Rudolph, and J. Sadowski, *Automatic Generation of 3D-Atomic Coordinates for Organic Molecules.* Tetrahedron Comput. Method., 1992. 3: p. 537-547.

18. Fox, T. and J.M. Kriegl, *Machine learning techniques for in silico modeling of drug metabolism.* Curr Top Med Chem, 2006. 6(15): p. 1579-91.

19. Meiler, J., *PROSHIFT: Protein Chemical Shift Prediction Using Artificial Neural Networks.* J. Biomol. NMR, 2003. 26: p. 25-37.

20. Walters, W.P. and M.A. Murcko, *Prediction of 'drug-likeness'.* Adv Drug Deliv Rev, 2002. 54(3): p. 255-71.

21. Tetko, I.V., V.V. Kovalishyn, and D.J. Livingstone, *Volume Learning Algorithm Artificial Neural Networks for 3D QSAR Studies.* Journal of Medicinal Chemistry, 2001. 44(15): p. 2411-2420.

22. Schoelkopf, B. and A.J. Smola, *Learning with Kernels*. Adaptive Computation and Machine Learning, ed. T. Dietterich2002, Cambridge, Massachusetts: The MIT Press.

23. He, X.Y., et al., *Type 10 17beta-hydroxysteroid dehydrogenase catalyzing the oxidation of steroid modulators of gamma-aminobutyric acid type A receptors.* Mol Cell Endocrinol, 2005. 229(1-2): p. 111-7.

24. Yan, C., J. Hu, and Y. Wang, *Discrimination of outer membrane proteins using a K-nearest neighbor method.* Amino Acids, 2008. 35(1): p. 65-73.

25. Jensen, B.F., et al., *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, 2007. 50(3): p. 501-11.

26. Nigsch, F., et al., *Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization.* J Chem Inf Model, 2006. 46(6): p. 2412-22.

27. Ajmani, S., K. Jadhav, and S.A. Kulkarni, *Three-dimensional QSAR using the k-nearest neighbor method and its interpretation.* J Chem Inf Model, 2006. 46(1): p. 24-31.

28. Kohonen, T. and P. Somervuo, *Self-organizing maps of symbol strings.* Neurocomputing, 1998. 21(1-3): p. 19-30.

29. Mao, K.Z., *Orthogonal forward selection and backward elimination algorithms for feature subset selection.* IEEE Trans Syst Man Cybern B Cybern, 2004. 34(1): p. 629-34.

30. Paixao, P., L.F. Gouveia, and J.A. Morais, *Prediction of the in vitro intrinsic clearance determined in suspensions of human hepatocytes by using artificial neural networks.* European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences, 2010. 39(5): p. 310-21.