Regression Methods for Developing QSAR and QSPR Models to Predict Compounds of Specific Pharmacodynamic, Pharmacokinetic and Toxicological Properties

C.W. Yap[1,2], H. Li[1,2], Z.L. Ji[3] and Y.Z. Chen[1,2,*]

[1]Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Block S4, 18 Science Drive 4, Singapore 117543; [2]Centre for Computational Science and Engineering, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543; [3]Department of Biomedical Sciences, School of Life Science, #114, Block of Biology II, Xiamen University, P.R., China

Abstract: Quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) models have been extensively used for predicting compounds of specific pharmacodynamic, pharmacokinetic, or toxicological property from structure-derived physicochemical and structural features. These models can be developed by using various regression methods including conventional approaches (multiple linear regression and partial least squares) and more recently explored genetic (genetic function approximation) and machine learning (k-nearest neighbour, neural networks, and support vector regression) approaches. This article describes the algorithms of these methods, evaluates their advantages and disadvantages, and discusses the application potential of the recently explored methods. Freely available online and commercial software for these regression methods and the areas of their applications are also presented.

Key Words: ADME, ADMET, compound, drug, pharmacodynamics, pharmacokinetics, toxicity, QSAR, QSPR, statistical learning methods, structure-activity relationship.

INTRODUCTION

As part of the effort for accelerating and reducing the cost of drug discovery processes, computational methods have been explored for predicting compounds that possess such pharmaceutically-relevant properties as a specific pharmacodynamic, pharmacokinetic or toxicological property [1-4]. In particular, quantitative structure activity relationship (QSAR) and quantitative structure-property relationship (QSPR) models has been instrumental for facilitating the prediction of biological and pharmaceutically-relevant activities of drug candidates. QSAR/QSPR is a method to model and predict the biological or pharmaceutically-relevant activity of a compound from a selected set of structure-derived physicochemical and structural features by using a statistically derived mathematical equation [5]. It is based on a similarity principle which assumes that compounds with similar physicochemical properties or structural frameworks tend to exhibit similar biological and pharmaceutically-relevant activities.

The process of developing a QSAR/QSPR model starts with the collection of high quality activity data and the elimination of low quality ones that are likely to affect the accuracy of the model. The next step is the selection of representative compounds into a training set and a validation set to calibrate and evaluate the QSAR/QSPR model respectively. Molecular descriptors are then computed for representing the physicochemical and structural properties of the compounds studied and those that are redundant or contain little information are removed prior to the modelling process. A regression method, such as multiple linear regression and neural networks, is then used to develop a model that relates the investigated activities of the compounds to their physicochemical and structural properties. During the modelling process, optimization of the essential parameters of the regression methods and the selection of relevant descriptor subsets are conducted simultaneously. The optimal set of parameters and descriptor subset are used to construct a final QSAR/QSPR model, which is subsequently subjected to evaluation by one or more of the various validation methods to ensure that the constructed model is valid and useful.

This article describes the algorithms, advantages, disadvantages and application potential of various regression methods that are commonly used for developing QSAR/QSPR models of specific pharmacodynamic, pharmacokinetic or toxicological property. It is to be noted that the selection of the algorithms and their descriptions are neither exhaustive nor comprehensive because of the limitation of a mini review and the reader is requested to refer to other resources for more detailed descriptions of the algorithms. The process of data collection, data pre-processing, computation and selection of molecular descriptors, and model validation have been extensively reviewed elsewhere [6-15] and they are thus not described here. Freely available online software and commercial software available for these regression methods are also discussed.

MOLECULAR DESCRIPTORS FOR REPRESENTING COMPOUNDS

Molecular descriptors are used for representing physicochemical and structural properties of compounds from their

[*]Address correspondence to this author at Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Block S4, 18 Science Drive 4, Singapore 117543; Tel: 65-6516-6877; Fax: 65-6779-1554; E-mail: phacyz@nus.edu.sg

1389-5575/07 $50.00+.00

1D, 2D or 3D structure. The most popularly used computer programs for deriving molecular descriptors are CoMFA [16], DRAGON [17], GRID [18], HyperChem [19], JOELib [20], MOE [21], Molconn-Z [22], VolSurf [23] and Xue descriptor set [24]. Web-servers such as MODEL (http://jing.cz3.nus.edu.sg/cgi-bin/model/model.cgi) have also emerged for facilitating the computation of molecular descriptors. Over 3,000 molecular descriptors can be derived from these programs, which range from constitutional descriptors to more complex 2D and 3D descriptors representing different geometric, connectivity, and physicochemical properties.

The commonly used descriptors can be divided into 18 classes. These include constitutional descriptors such as molecular weight, geometrical descriptors such as volume and surface areas, topological descriptors such as the number of rings and rotatable bonds, RDF descriptors representing inter-atomic distances in the entire molecule and other useful information such as bond distances, ring types, planar and non-planar systems, atom types and molecular weight [25], molecular walk counts [26], 3D-MoRSE descriptors describing features such as molecular weight, van der Waals volume, electronegativities and polarizabilities [27], BCUT descriptors representing connectivity information and atomic properties relevant to intermolecular interaction [28], WHIM descriptors describing size, shape, symmetry, atom distribution and polarizability of a molecule [29], Galvez topological charge indices and charge descriptors [30], GETAWAY descriptors [31], 2D autocorrelations, functional groups, atom-centred descriptors, aromaticity indices [32], Randic molecular profiles [33], electrotopological state descriptors [34], linear solvation energy relationship descriptors [35], and other empirical and molecular properties.

METHODS FOR DEVELOPING REGRESSION MODELS

Various regression methods have been applied to QSAR and QSPR model construction. The most widely used methods include such conventional approaches as multiple linear regression and partial least squares and such recently explored approaches as genetic function approximation and machine learning methods (k-nearest neighbour, neural networks, and support vector regression).

Multiple Linear Regression (MLR)

MLR [36] is one of the most commonly used and simplest methods for constructing QSAR/QSPR models. A MLR model is constructed under the assumption that a linear relationship exists between a set of molecular descriptors of a compound (which is represented by a feature vector **x** with each descriptor as its component) and a specific activity (which is represented by a quantity *y*). A MLR model can be described using the following equation $\hat{y} = \beta 0 + \beta 1 X1 +$, $\beta 2 X2 + \ldots + \beta k Xk$ where $\{X1,\ \ldots,\ Xk\}$ are molecular descriptors, $\beta 0$ is the regression model constant, $\beta 1$ to βk are the coefficients for individual descriptor *X1* to *Xk*. The values for $\beta 0$ to βk are chosen by minimizing the sum of squares of the residuals between the observed and predicted values defined by the equation so as to give the best prediction of *y* from **x**.

The advantage of MLR is its simplistic form and easily interpretable mathematical expression. The sign of the coefficients $\beta 1$ to βk indicates whether each molecular descriptor contributes positively or negatively to a specific activity and their magnitudes indicates the relative importance of each descriptor to that activity. However, MLR works well only when the structure-activity relationship is linear in nature, the set of molecular descriptors are mathematically independent (orthogonal) of each another, and the number of compounds in the training set exceeds the number of molecular descriptors by at least a factor of five [37]. It has been found that, when collinear descriptors are used, the derived coefficients $\beta 1$ to βk tend to be larger than the real values and sometimes have opposite signs [15]. Therefore, the assumption of a linear relationship between a set of molecular descriptors and a specific activity may not always be appropriate, especially in the cases involving multiple mechanisms.

Partial Least Squares (PLS)

PLS [38] constructs a QSAR or QSPR model by creating latent variables, $\mathbf{x}_{\text{new}}$, which are molecular descriptors carrying the same information as the original molecular descriptors $\mathbf{x}_{\text{original}}$, but fewer in numbers. PLS differentiates from principal component analysis (PCA) in that PLS finds latent variables from **x** that are also relevant for *y*. Specifically, PLS regression searches for a set of latent variables by performing a simultaneous decomposition of **x** and *y* with the constraint that these latent variables explain the maximum covariance between **x** and *y*. This is different from PCA, which creates its scores by finding the linear combination of the explanatory variables **x** that have the maximum variance. These scores are not chosen optimally for a regression analysis because the principal component scores have been chosen without even considering the activity variable *y*. An important consideration in PLS is the selection of an appropriate number of latent variables for constructing a QSAR model. This is usually determined by using such cross-validation methods as 5-fold cross-validation and leave-one-out.

Genetic Function Approximation (GFA)

GFA is a method which combines genetic algorithm (GA) [39] and multivariate adaptive regression splines (MARS) [40] algorithm to produce multiple QSAR/QSPR models [41]. An initial population of models are first created by randomly selecting some descriptors for building basis functions, which are functions of one or more descriptors, and developing QSAR/QSPR models from these basis functions. These models are scored by using Friedman's "lack of fit' (LOF) measure, which is resistant to over-fitting problems better than the traditional least-squares error (LSE) measure. LOF is given by $LOF = LSE / [1-(c+dp) \quad / M]^2$ where *c* is the number of basis functions, *d* is a smoothing parameter, *p* is the total number of descriptors contained in all basis functions and *M* is the total number of compounds in the training set. GA is then used to incrementally improve the LOF score of the models by selecting more relevant basis functions for building the models. At the end of the GA process, the model with the lowest LOF score can be selected as the final QSAR/QSPR model. Alternatively, a consensus model can be built from the population of models.

GFA has several advantages over standard regression analysis. GFA is able to automatically select descriptors and combination of basis functions that are important for the models. In addition, it can provide additional information such as preferred model length and useful partitions of the datasets.

k-Nearest Neighbour (kNN)

kNN is a basic instance-based method [42], which measures the Euclidean distance between a given compound (represented by a vector **x**) and each of the near neighbor compounds in a training set (represented by a group of vectors {**x**$_i$}) [42, 43]. The Euclidean distance between vector **x** and **x**$_i$ is computed by using the formula $D=\sqrt{\|\mathbf{x}-\mathbf{x}_i\|^2}$. The activity of the studied compound is determined by the average of the activity values of a total of k number of training compounds nearest to that compound $\hat{y}=(\sum_{i=1}^{k} y_i\}/k$ ·

Feedforward Back Propagation Neural Network (FFBPNN)

FFBPNN is a form of artificial neural network which has two distinct phases: forward propagation of activation and backward propagation of error [44]. It is composed of an input layer, a variable number of hidden layers and an output layer. The input and output layers contain neurons representing the molecular descriptors and activity value of a studied compound respectively. In a fully connected FFBPNN, each neuron in the input layer sends its value to all neurons in the first hidden layer. Each neuron in the hidden layers receives inputs from all neurons in the previous layer and computes a weighted sum of the inputs. The neuron output is determined by passing the weighted sum through a transfer function, which is usually a linear or sigmoidal function. The single neuron in the output layer determines the predicted activity value of a compound by computing a weighted sum of the outputs of all neurons in the last hidden layer. Weights for the connections between neurons in adjacent layers are initially randomly assigned. These weights are then refined *via* a backward propagation of error process during training of the FFBPNN. In backpropagation learning, every time an input vector of a training sample is presented, the output vector is compared to the desired value by evaluating the squared difference $Err=(V_{desired}-V_{output})^2$. The goal of backpropagation is to gradually minimize the sum of *Err*, $Minimize\sum Err=(V_{desired}-V_{output})^2$, for all the training samples to ensure that the network behave in the most desired way.

A difficulty in using FFBPNN is the construction of an optimal architecture for a given problem. An undersized network is not capable of generating an optimal QSAR or QSPR model, while an oversized network may lead to an over-fitted model. Moreover, the physicochemical basis of the connection weights of FFBPNN are not easily interpreted, which makes it difficult for medicinal chemists to rationally optimize the structures of active compounds based on a FFPBNN model.

General Regression Neural Network

GRNN [45] is a form of neural network designed for regression through the use of Bayes' optimal decision rule. In GRNN, the activity value of a studied compound is derived from the most probable value sampled over the activity of all of the compounds in a training set, which is given by $\hat{y}=[\int_{-\infty}^{\infty} yf(x,y)dy]/[\int_{-\infty}^{\infty} f(x,y)dy]$ where *f*(**x**,*y)* is the joint density which can be estimated straightforwardly by using Parzen's nonparametric estimator [46] $g(x)=[\sum_{i=1}^{n} W((x-x_i)/\sigma)]/n\sigma$ where n is the sample size, σ is a scaling parameter which defines the width of the bell curve that surrounds each compound, *W(d)* is a weight function which has its largest value at $d = 0$ and $(x - x_i)$ is the distance between a given compound and a compound in the training set. The Parzen's nonparametric estimator was later expanded by Cacoullos [47] for the multivariate case, $g(x_1,\ldots,x_p)=.$ $\frac{1}{n\sigma_1\ldots\sigma_p}\sum_{i=1}^{n} W(\frac{x_1-x_{1,i}}{\sigma_1},\ldots,\frac{x_p-x_{p,i}}{\sigma_p})$ The Gaussian function is frequently used as the weight function because it is well behaved, easily calculated and satisfies the conditions required by Parzen's estimator. Thus the probability density function for the multivariate case becomes $g(\mathbf{x})=.\ \frac{1}{n}$ $\sum_{i=1}^{n}\exp(-\sum_{j=1}^{p}\left(\frac{x_j-x_{j,i}}{\sigma_j}\right)^2)$ Substituting Parzen's nonparametric estimator for *f*(**x**,*y)* and performing the integrations leads to the fundamental equation of GRNN $\hat{y}=,\ [\sum_{i=1}^{n} y_i\ \exp(-D(x,x_i))]/[\sum_{i=1}^{n}\exp(-D(x,x_i))]$ where $D(x,x_i)=\sum_{j=1}^{p}\left[(x_j-x_{ji})/\sigma_j\right]^2$ ·

Support Vector Regression (SVR)

SVR is an extension of support vector machine (SVM) to solve nonlinear regression problems by introducing an ε-insensitive loss function [48-50]. A kernel function (in the form of a polynomial, gaussian, or sigmoidal function) is used to map the input vectors into a higher dimensional feature space and then a linear regression model is conducted in this feature space. The quality of estimation is measured by the ε-insensitive loss function $L(y,f(x,\omega))=0$ if $|y-f(x,\omega)|\leq\varepsilon$ otherwise $L(y,f(x,\omega))=|y-f(x,\omega)|\ -\varepsilon$. The optimal regression function can be represented by $\hat{y}$ = $\sum_{i=1}^{nsv}(\alpha_i-\alpha_i^*)K(x_i,x)+b$ under the conditions $0\leq\alpha_i,\ \alpha_i^*\leq C$ and $\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)=0$ · Where $\hat{y}$ represents the predicted activity value of a specific property, *nsv* is the number of support vectors, constant C determines the trade off between the flatness of function *f* and the amount up to which deviations larger than ε are tolerated and *K* is the kernel function, normally Gaussian kernel function $K(x_i, x_j) = e^{-(\|x_j - x_i\|^2)/(2\sigma^2)}$ is used.

Similar with other multivariate regression models, the generalization capability of SVR depends on proper selection of parameters *C*, ε, the kernel type and its parameters. Some strategies are needed for optimizing these factors to achieve best generalization capability. For instance, a too small *C* value may lead to under-fitting and a too large *C* value may lead to over-fitting of the training data. A larger ε value may lead to a higher number of support vectors. The selection of the kernel function and corresponding parameters is very important because they define the distribution of the training set samples in the high dimensional feature space [51].

CURRENT APPLICATIONS OF REGRESSION METHODS

Table **1** summarises the performance of the commonly used regression methods for predicting compounds of various pharmacodynamic, pharmacokinetic and toxicological properties. The majority of the QSAR/QSPR models have primarily been developed by using conventional regression methods such as MLR or PLS. It is highly likely that these regression methods have been used because of their simplicity and because the derived models can be easily interpreted. The performance of these studies is primarily measured by the r^2 value, which measures the explained variance between the computed activities and experimentally estimated activities. Moreover, q^2 values and RMSE values are also frequently computed to further evaluate the predictive capability of these studies. The number of compounds in many of the studies listed in Table **1** is in the range of tens to hundreds of compounds, which is significantly lower than the hundreds to thousands of compounds typically used in classification studies [52].

The computed r^2 values are in the range of 0.30 to 0.99 with the majority concentrated in the range of 061 to 0.91. These results suggest that the regression methods surveyed here have certain level of capability for predicting the activity of compounds of different pharmacodynamic, pharmacokinetic and toxicological properties. In these studies, the r^2 of models developed by using more recently explored regression methods such as FFBPNN, GRNN or SVR appear to be higher than the corresponding values of models developed by using conventional regression methods. One likely reason for the higher r^2 derived from these more recently explored regression methods is that these methods do not rely on the existence of a fixed relationship between a specific activity and the molecular descriptors of the studied compounds. This makes it possible to model compounds of complex relationships and thus improve the prediction capabilities of the developed models. However, such added flexibility makes the more recently explored regression methods more susceptible to overfitting problems than the conventional regression methods [53, 54]. Overfitted models appear to show good prediction performance for the training set but exhibit poor performance for compounds not in the training set. Hence proper validation of QSAR/QSPR models is important to ensure that the models are valid and has reasonably good generalization capability.

FREELY AVAILABLE ONLINE AND COMMERCIAL SOFTWARE FOR QSAR/QSPR MODELING

A number of commercial and free software are available for facilitating the development of QSAR/QSPR models. A particularly useful source for such software is the public web-servers: QSAR and Modelling Society (http://www.qsar.org) and Cheminformatics (http://www.cheminformatics.org). Some of the relevant software is based on a particular data analysis method while others include a number of data analysis methods. Moreover, some of the relevant software do not include molecular descriptor computing module, in which case such software or web-servers as DRAGON [17], Molconn-Z [22], MODEL [55], or VolSurf [23], can be used for deriving the needed molecular descriptors.

CONCLUDING REMARKS

Evaluation of literature reported performances of commonly used regression methods in QSAR/QSPR studies shows that these methods consistently exhibit promising capability for predicting the activity of compounds of diverse ranges of structures and of a wide variety of pharmacodynamic, pharmacokinetic, and toxicological properties. A summary of the characteristics of these methods is given in Table **2**. Regression methods can be used for quantitative prediction of the activity levels of new compounds in cases that the activity data are available for a sufficient number of known compounds. These methods have the capacity for estimating the contribution of specific structural and physicochemical features of the selected compounds to a particular property [56]. This capacity may be explored for probing the mechanism of action for a specific group of compounds that possess a particular property.

Development of new regression methods and exploration of those developed in other fields is highly useful for further advancement of QSAR/QSPR research. Several methods have recently been developed in other fields, which include kernel partial least squares (K-PLS) [57], hierarchical PLS (Hi-PLS) [58], orthogonal PLS (OPLS) [59], robust continuum regression [60], and deepest regression [61]. These methods have been shown to be useful for the prediction of a wide variety of properties including the levels of moisture, oil, protein and starch in corn [57], output of a polymer processing plant [57], chaotic Mackey-Glass time-series [57], human signal detection performance monitoring [57], binding strength of ligand-protein complexes [62], wood chip dry content [59], X-ray analysis of hydrometallugical solutions [60], and Michaelis–Menten model of enzyme kinetics [63]. It is of interest to explore these and other new regression methods for developing QSAR/QSPR models that can cover a more diverse spectrum of compounds and are capable of describing a more extensive range of pharmacodynamic, pharmacokinetic and toxicological properties.

Table 1. QSAR Models for Pharmacodynamic, Pharmacokinetic and Toxicological Agents

Property	Method and Reference of Reported Study	Number of Compounds	Reported Prediction Statistics
Antitrichomonal agents	LDA [64]	196	r^2 = 0.749 - 0.845
Carbonic anhydrase inhibitors	MLR and NN [65]	142	r^2 = 0.921 - 0.943 (MLR) r^2 = 0.971 - 0.992 (NN)
COX-2 inhibitors	MLR and NN [66]	273	r^2 = 0.666 - 0.669 (MLR) r^2 = 0.719 - 0.883 (NN)
	SVR [67]	53	r^2 = 0.869
	MLR [68]	16	r^2= 0.9839
	PLS, NN [69]	322	q^2=0.52 (PLS) q^2=0.53 (NN)
1,4-dihydropyridine calcium channel antagonists	LSSVM [70]	45	r^2 = 0.8696
	GEP and HM [71]	45	r^2= 0.88 – 0.93 (GEP) r^2= 0.86 – 0.91 (HM)
Angiotensin-converting enzyme (ACE) inhibitors	MLR [68]	58	r^2 = 0.9398
	PLS, NN [69]	114	q^2=0.72 (PLS), q^2=0.72 (NN)
Acetylcholinesterase (AChE) inhibitors	PLS, NN [69]	111	q^2=0.30 (PLS), q^2=0.45 (NN)
Benzodiazepine receptor binders	PLS, NN [69]	163	q^2=0.34 (PLS), q^2=0.35 (NN)
Dihydrofolate reductase inhibitors (DHFR)	PLS, NN [69]	397	q^2=0.52 (PLS), q^2=0.61 (NN)
Glycogen phosphorylase b (GPB) inhibitors	PLS, NN [69]	66	q^2=0.42 (PLS), q^2=0.48 (NN)
Thermolysin inhibitors (THER)	PLS, NN [69]	76	q^2=0.65 (PLS), q^2=0.64 (NN)
Thrombin inhibitors	PLS, NN [69]	88	q^2=0.45 (PLS), q^2=0.64 (NN)
Protein Tyrosine Phosphatase 1B Inhibitors	MLR [72]	128	r^2 =0.859

Na+/H+ antiporter inhibitors	MLR and NN [73]	113	RMSE= 0.473 - 0.546 (MLR)
			RMSE= 0.228 - 0.296 (NN)

Human type 1 5alpha-reductase inhibitors	NN [74]	93	r^2= 0.89 - 0.97
Murine and human soluble epoxide hydrolase inhibition by urea-like compounds	NN [75]	348	r^2= 0.61 - 0.66
HIV-1 protease inhibitors	PLS [76]	48	r^2=0.91, q^2=0.84
	MLR, PLS [77]	35	r^2=0.763 – 0.798, q^2=0.703 -0.741 (MLR) r^2=0.865 (PLS)
CCR5 receptor binders	MLR [78]	93	r^2=0.951
	MLR [79]	79	r^2=0.834
	MLR [80]	52	r^2=0.837
Glycogen synthase kinase-3 inhibitors	MLR and ANN [81]	277	r^2= 0.507 - 0.896 (MLR) r^2= 0.679 - 0.782 (ANN)
Platelet-derived growth factor inhibitors	MLR and ANN [82]	78 - 123	r^2= 0.684 - 0.698 (MLR) r^2= 0.71 - 0.81 (ANN)

(Table 1. Contd….)

Property	Method and Reference of Reported Study	Number of Compounds	Reported Prediction Statistics
Artemisinin analogues	NN [83]	179	r^2= 0.88 – 0.96
PDGFR inhibition	NN [84]	79	r^2= 0.61 - 0.93
	LS-SVM [70]	45	r^2=0.817
hERG channel inhibitors	BPNN [85]	439	r^2= 0.76
	SVR, PLS and RF [86]	90	r^2= 0.849 – 0.912 (SVR) r^2= 0.753 – 0.882 (PLS) r^2= 0.785 – 0.889 (RF)
	MLR [87]	104	r^2= 0.636 – 0.704
	RP [88]	134	r^2= 0.83
	PLS [89]	348-544	r^2=0.76-0.77
Estrogen Receptor binders	KNN [90]	61-68	r^2 =0.69-0.79
	PLS [91]	40-44	r^2 =0.45-0.96

Human Androgen Receptor binders	MLR, RBFNN, SVR [92]	146	RMS = 0.76 (MLR)
			RMS = 0.69 (RBFNN)
			RMS = 0.55 (SVR)

	PLS [93]	70	r^2 =0.66
Calcium Channel Antagonists	LS-SVM [70]	45	r^2 =0.82
	ANN [94]	110	r^2=0.93 – 0.94
	PCANN, MLR [95]	46	r^2=0.55 (MLR) r^2=0.73 (PCANN)
Potassium channel openers	PLS [96]	27	r^2=0.94
Dopamine Antagonists	PLS, kNN [97]	29	r^2=0.73 (PLS) r^2=0.79 (kNN)
Platelet aggregation inhibitor	MLR [98]	35	r^2=0.74
Skin permeation	MLR, ANN [99]	143	r^2=0.804 – 0.919(MLR) r^2=0.72 – 0.813 (ANN)
Human and rat steady-state volume of distribution	BNN, CART and PLS [100]	199 – 2086	Human: r^2=0.560 – 0.794 (BNN), r^2=0.573 – 0.876 (CART), r^2=0.587 – 0.641 (PLS) Rat: r^2=0.527 – 0.767 (BNN), r^2=0.470 – 0.846 (CART), r^2=0.463 – 0.519 (PLS)
Human oral absorption	SVR [101]	169	r^2=0.70 – 0.86
Human intestine absorption	MLR [102, 103]	169 – 467	r^2=0.79 – 0.82
	Sigmoidal [104]	20	r^2=0.94
	PLS [105, 106]	79 – 169	r^2=0.55 – 0.921
	ANN [107-109]	77 – 581	r^2=0.80 – 0.92
	GRNN [110]	77	RMSE=6.5
	CART [111]	899	AAE=0.120 – 0.200
	PLS [112, 113]	20	r^2=0.903
	SVR [114]	20	r^2=0.779 – 0.877

(Table 1. Contd….)

Property	Method and Reference of Reported Study	Number of Compounds	Reported Prediction Statistics
Bioavailability	ML [115]	591	r^2=0.71
	MLR [116]	169	r^2=0.72
	ANN [117]	152	r^2=0.736
	NN [118]	28	q^2=0.90
Blood Brain Barrier penetration	MLR [119-137]	20 - 150	r^2=0.56 - 0.95
	PLS [138]	86	r^2=0.89
	PCR [139, 140]	75 - 100	r^2=0.576 - 0.83
	PLS [106, 113, 141-145]	56 - 97	r^2=0.617 - 0.910
	NN [118]	36	q^2=0.88
	BNN [146]	106	r^2=0.76
	GRNN [147]	159	r^2=0.701
	SVR [114]	59	r^2=0.82 – 0.85
HSA binding	MLR [148, 149]	94	r^2=0.68 - 0.88
	GRNN [147]	93	r^2=0.851
	SVR [150]	94	r^2=0.89 - 0.94
Milk-plasma ratio	ANN [151]	123	r^2=0.61
	GRNN [147]	122	r^2=0.677
Total clearance	kNN [152]	38	r^2=0.94
	ANN [153]	6	r^2=0.731
	GRNN [154]	23	r^2=0.775
P-gp inhibitor	PLS [155]	100	r^2=0.731
Genotoxicity	NN [156]	82	r^2= 0.871
	MLR [157]	95	r^2=0.70
	MLR [158]	29	r^2=0.44
Toxicity to *Vibrio fischeri*	MLR [159]	56	r^2= 0.820 – 0.865
Hepatotoxicity	LR, MLR [160]	15-28	r^2=0.801 (LR) r^2=0.561 – 0.892 (MLR)
Aquatic toxicology	MLR [161]	92	r^2= 0.738 – 0.885
	CPNN [162]	282	r^2=0.79
Cytotoxicity	MLR [163]	42	r^2= 0.823 – 0.831
	MLR [164-166]	7-64	r^2=0.59 – 0.95
	MLR [167]	29	r^2= 0.87 – 0.89
Tetrahymena pyriformis toxicity	MLR [168]	40	r^2= 0.821 – 0.831
	PLS [169]	476	r^2= 0.801-0.826
	NN,GAM,MARS,PPR [170]	203	r^2= 0.73 (NN) r^2= 0.61 – 0.75 (GAM) r^2= 0.73 – 0.74 (MARS) r^2= 0.71 – 0.80 (PPR)
	PNN [171]	1084	r^2= 0.8033 - 0.8989

(Table 1. Contd….)

Property	Method and Reference of Reported Study	Number of Compounds	Reported Prediction Statistics
Toxicity in fish fathead minnow	CPNN [172]	541	r^2= 0.861 – 0.93
	MLR [173]	408	r^2= 0.803– 0.922
Toxicity of chlorophenols	CoMFA [174]	10	r^2= 0.727 – 0.968
Toxicity to rainbow trout *Onchorhyncus mykiss Walbaum*	MLR, PLS [175]	75	q^2= 0.75 -0.78 (MLR) q^2= 0.80 (PLS)
General toxic chemicals	MLR, RBFNN, SVR [176]	76	r^2=0.85 (MLR) r^2=0.88 (RBFNN) r^2=0.95 (SVR)

Abbreviations:

hERG: human ether-a-go-go-related gene; **MLR**: multiple linear regressions; **PLS**: partial least squares; **kNN**: k nearest neighbors; **PCA**: principal component analysis; **SVR**: support vector regression; **LS-SVM:** least square support vector machine; **ANN**: artificial neural network; **BPNN**: back-propagation neural network; **BNN**: Bayesian neural networks; **RBFNN**: radial basis function neural network; **GRNN**: generalized regression neural network; **CPNN**: counter propagation neural network model; **HCA**: hierarchical cluster analysis; **PCRA**: principal component regression analysis; **GEP**: gene expression programming; **HM**: heuristic method; **GA-MLR**: multiple linear regressions combined with genetic algorithm; **RF**: random forests; **RP**: recursive partitioning; **PPR**: projection pursuit regression; **GAM**: generalized additive model; **MARS**: multivariate adaptive regression splines; **CART**: classification and regression trees; **CoMFA**: comparative molecular field analysis.

Table 2. Characteristics of the Various Regression Methods

Regression Methods	Dataset	Descriptors	Model
MLR	- Single mechanism of action - One target property	- Should not have intercorrelation - Total number must not exceed one-fifth of the number of compounds in training set	- No optimizable parameter - Fast training speed - Fast prediction speed - Easy to interpret - Low risk of overfitting
PLS	- Can have multiple mechanism of action (if non-linear extensions such as quadratic PLS, spline PLS, GIFI-PLS, are used) - Can have multiple target properties	- Can have intercorrelation - No restriction on the total number used	- One optimizable parameter - Fast training speed - Fast prediction speed - Low risk of overfitting
GFA	- Can have multiple mechanism of action - One target property	- Can have intercorrelation - Total number used restricted by LOF statistics	- Multiple optimizable parameters - Slow training speed - Fast prediction speed - Easy to interpret - Low risk of overfitting
kNN	- Can have multiple mechanism of action - Can have multiple target properties	- Can have intercorrelation - No restriction on the total number used	- One optimizable parameter - Fast training speed - Prediction speed may be slow with large training sets - Difficult to interpret - Risk of overfitting
FFBPNN	- Can have multiple mechanism of action - Can have multiple target properties	- Can have intercorrelation - No restriction on the total number used	- Multiple optimizable parameters - Slow training speed - Fast prediction speed - Difficult to interpret - Risk of overfitting - Non-uniqueness

(Table 2. Contd….)

Regression Methods	Dataset	Descriptors	Model
GRNN	- Can have multiple mechanism of action - Can have multiple target properties	- Can have intercorrelation - No restriction on the total number used	- One optimizable parameter - Fast training speed - Prediction speed may be slow with large training sets - Difficult to interpret - Risk of overfitting
SVR	- Can have multiple mechanism of action - One target property	- Can have intercorrelation - No restriction on the total number used	- Multiple optimizable parameters - Training speed may be slow with large training sets - Prediction speed may be slow - Difficult to interpret - Low risk of overfitting

REFERENCES

[1] Hansch, C.; Leo, A.; Mekapati, S. B.; Kurup, A. *Bioorg. Med. Chem.,* **2004**, *12*, 3391.

[2] Katritzky, A. R.; Karelson, M.; Lobanov, V. *Pure Appl. Chem.,* **1997**, *69*, 245.

[3] Manallack, D. T.; Livingstone, D. J. *Eur. J. Med. Chem.,* **1999**, *34*, 195.

[4] van de Waterbeemd, H.; Gifford, E. *Nat. Rev. Drug Discov.,* **2003**, *2*, 192.

[5] Johnson, M. A.; Maggiora, G. M., *Concepts and Applications of Molecular Similarity*. Wiley: New York, **1990**.

[6] Cronin, M. T. D.; Schultz, T. W. *THEOCHEM,* **2003**, *622*, 39.

[7] Susnow, R. G.; Dixon, S. L. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 1308.

[8] Wold, S.; Eriksson, L. In *Chemometric Methods in Molecular Design*, van de Waterbeemd, H. Ed. VCH: Weinheim; New York; Basel; Cambridge; Tokyo, **1995**.

[9] Gramatica, P.; Pilutti, P.; Papa, E. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 1794.

[10] Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. D. *SAR QSAR Environ. Res.,* **2003**, *14*, 59.

[11] Rajer-Kanduc, K.; Zupan, J. M., N. *Chemom. Intell. Lab. Sys.,* **2003**, *65*, 221.

[12] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*. Wiley-VCH: Weinheim, **2000**.

[13] Livingstone, D. J. *Data Analysis for Chemists: Applications to QSAR and Chemical Product Design*. Oxford University Press: Oxford, **1995**.

[14] Guyon, I.; Elisseeff, A. *J. Mach. Learn. Res.,* **2003**, *3*, 1157.

[15] Eriksson, L.; Jaworska, J.; Cronin, M.; Worth, A.; Gramatica, P.; McDowell, R. *Environ. Health Perspect.,* **2003**, *111*, 1361.

[16] Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *Prog. Clin. Biol. Res.,* **1989**, *291*, 161.

[17] Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *DRAGON*, Version 5.3; Talete SRL: Milan, **2005**.

[18] Pastor, M.; Cruciani, G.; Watson, K. A. *J. Med. Chem.,* **1997**, *40*, 4089.

[19] Hypercube *HyperChem 7.5*, **2006**.

[20] Wegner, J. K. *JOELib/JOELib2*, **2005**.

[21] *Molecular Operating Environment (MOE)*, Chemical Computing Group: **2006**.

[22] Hall, L. H.; Kellogg, G. E.; Haney, D. N. *Molconn-Z*, Version 4.05+; eduSoft, LC: **2002**.

[23] Cruciani, G.; Pastor, M.; Guba, W. *Eur. J. Pharm. Sci.,* **2000**, *11*, S29.

[24] Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 1630.

[25] Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *Vib. Spectrosc.,* **1999**, *19*, 151.

[26] Rücker, G.; Rücker, C. *J. Chem. Inf. Comput. Sci.,* **1993**, *33*, 683.

[27] Schuur, J. H.; Setzer, P.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.,* **1996**, *36*, 334.

[28] Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.,* **1999**, *39*, 28.

[29] Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. *J. Comput. Aided Mol. Des.,* **1997**, *11*, 79.

[30] Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. *J. Chem. Inf. Comput. Sci.,* **1994**, *34*, 520.

[31] Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 682.

[32] Randic, M. *Tetrahedron,* **1975**, *31*, 1477.

[33] Randic, M. *New J. Chem.,* **1995**, *19*, 781.

[34] Kier, L. B.; Hall, L. H., *Molecular Structure Description: The Electrotopological State*. Academic Press: San Diego, **1999**.

[35] Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. *J. Chem. Inf. Comput. Sci.,* **1999**, *39*, 835.

[36] Allison, P. *Multiple Regression*. Pine Forge Press: Thousand Oaks, CA, **1999**.

[37] Topliss, J. G.; Edwards, R. P. *J. Med. Chem.,* **1979**, *22*, 1238.

[38] Abdi, H. In *Encyclopedia of Measurement and Statistics.*, Salkind, N. J. Ed. Sage: Thousand Oaks, CA, **2007**.

[39] Glodberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Pub. Co.: Boston, MA, **1988**.

[40] Friedman, J. H.; Roosen, C. B. *Stat. Methods Med. Res.,* **1995**, *4*, 197.

[41] Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.,* **1994**, *34*, 854.

[42] Fix, E.; Hodges, J. L. *Discriminatory Analysis: Non-Parametric Discrimination: Consistency Properties*; 4; USAF School of Aviation Medicine, Randolph Field: Texas, **1951**.

[43] Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*. Prentice Hall: Englewood Cliffs, NJ, **1982**.

[44] Wythoff, B. J. *Chemom. Intell. Lab. Sys.,* **1993**, *18*, 115.

[45] Specht, D. F. *IEEE T Neural Netw.,* **1991**, *2*, 568.

[46] Parzen, E. *Ann. Math. Stat.,* **1962**, *33*, 1065.

[47] Cacoullos, T. *Ann. I. Stat. Math.,* **1966**, *18*, 179.

[48] Smola, A. J.; Scholkopf, B. In *A Tutorial on Support Vector Regression*, NeuroCOLT2 Technical Report NC2-TR-1998-030.

[49] Yuan, Z.; Huang, B. X. *Proteins,* **2004**, *57*, 558.

[50] Vapnik, V. N. *The Nature of Statistical Learning Theory*. Springer: New York, **1995**.

[51] Yao, X. J.; Panaye, A.; Doucet, J. P.; Zhang, R. S.; Chen, H. F.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 1257.

[52] Yap, C. W.; Xue, Y.; Li, H.; Li, Z. R.; Ung, C. Y.; Han, L. Y.; Zheng, C. J.; Cao, Z. W.; Chen, Y. Z. *Mini Rev. Med. Chem.,* **2006**, *6*, 449.

[53] Hawkins, D. M. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 1.

[54] Babyak, M. A. *Psychosom. Med.,* **2004**, *66*, 411.

[55] Li, Z. R.; Han, L. Y.; Xue, Y.; Yap, C. W.; Li, H.; Jiang, L.; Chen, Y. Z. *Biotechnol. Bioeng*., **2007**, *97*, 389.

[56] Stanton, D. T. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 1423.

[57] Rosipal, R.; Trejo, L. J. *J. Mach. Learn. Res.,* **2001**, *2*, 97.

[58] Eriksson, L.; Johansson, E.; Lindgren, F.; Sjostrom, M.; Wold, S. *J. Comput. Aided Mol. Des.,* **2002**, *16*, 711.

[59] Trygg, J.; Wold, S. *J. Chemometr.,* **2002**, *16*, 119.

[60] Serneels, S.; Filzmoser, P.; Croux, C.; Van Espen, P. J. *Chemom. Intell. Lab. Sys.,* **2005**, *76*, 197.

[61] Rousseeuw, P. J.; Hubert, M. *J. Am. Statist. Ass.,* **1999**, *94*, 388.

[62] Lindström, A.; Pettersson, F.; Almqvist, F.; Berglund, A.; Kihlberg, J.; Linusson, A. *J. Chem. Inf. Model.,* **2006**, *46*, 1154.

[63] Van, A. S.; Rousseeuw, P. J.; Hubert, M.; Struyf, A. *J. Multivariate Analysis,* **2002**, *81*, 138.

[64] Marrero-Ponce, Y.; Meneses-Marcel, A.; Castillo-Garit, J. A.; Machado-Tugores, Y.; Escario, J. A.; Barrio, A. G.; Pereira, D. M.; Nogal-Ruiz, J. J.; Aran, V. J.; Martinez-Fernandez, A. R.; Torrens, F.; Rotondo, R.; Ibarra-Velarde, F.; Alvarado, Y. J. *Bioorg. Med. Chem.,* **2006**, *14*, 6502.

[65] Mattioni, B. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 94.

[66] Kauffman, G. W.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.,* **2001**, *41*, 1553.

[67] Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Comput. Aided Mol. Des.,* **2004**, *18*, 389.

[68] Liu, S. S.; Yin, C. S.; Wang, L. S. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 749.

[69] Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. *J. Med. Chem.,* **2004**, *47*, 5541.

[70] Yao, X.; Liu, H.; Zhang, R.; Liu, M.; Hu, Z.; Panaye, A.; Doucet, J. P.; Fan, B. *Mol. Pharm.,* **2005**, *2*, 348.

[71] Si, H. Z.; Wang, T.; Zhang, K. J.; Hu, Z. D.; Fan, B. T. *Bioorg. Med. Chem.,* **2006**, *14*, 4834.

[72] Patankar, S. J.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 885.

[73] Kauffman, G. W.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.,* **2000**, *40*, 753.

[74] Bakken, G. A.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.,* **2001**, *41*, 1255.

[75] McElroy, N. R.; Jurs, P. C.; Morisseau, C.; Hammock, B. D. *J. Med. Chem.,* **2003**, *46*, 1066.

[76] Kiralj, R.; Ferreira, M. M. *J. Mol. Graph. Model.,* **2003**, *21*, 435.

[77] Leonard, J. T.; Roy, K. *Bioorg. Med. Chem.,* **2006**, *14*, 1039.

[78] Roy, K.; Leonard, J. T. *J. Chem. Inf. Comput. Sci.,* **2005**, *45*, 1352.

[79] Thomas Leonard, J.; Roy, K. *Bioorg. Med. Chem. Lett.,* **2006**, *16*, 4467.

[80] Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. *J. Comput. Aided Mol. Des.,* **2006**, *20*, 83.

[81] Katritzky, A. R.; Pacureanu, L. M.; Dobchev, D. A.; Fara, D. C.; Duchowicz, P. R.; Karelson, M. *Bioorg. Med. Chem.,* **2006**, *14*, 4987.

[82] Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; M., K. *Bioorg. Med. Chem.,* **2005**, *13*, 6598.

[83] Guha, R.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 1440.

[84] Guha, R.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 2179.

[85] Seierstad, M.; Agrafiotis, D. K. *Chem. Biol. Drug Des.,* **2006**, *67*, 284.

[86] Song, M.; Clark, M. *J. Chem. Inf. Model.,* **2006**, *46*, 392.

[87] Yoshida, K.; Niwa, T. *J. Chem. Inf. Model.,* **2006**, *46*, 1371.

[88] Ekins, S.; Balakin, K. V.; Savchuk, N.; Ivanenkov, Y. *J. Med. Chem.,* **2006**, *49*, 5059.

[89] Cianchetta, G.; Li, Y.; Kang, J.; Rampe, D.; Fravolini, A.; Cruciani, G.; Vaz, R. J. *Bioorg. Med. Chem. Lett.,* **2005**, *15*, 3637.

[90] Asikainen, A. H.; Ruuskanen, J.; Tuppurainen, K. A. *Environ. Sci. Technol.,* **2004**, *38*, 6724.

[91] Yu, S. J.; Keenan, S. M.; Tong, W.; Welsh, W. J. *Chem. Res. Toxicol.,* **2002**, *15*, 1229.

[92] Zhao, C. Y.; Zhang, R. S.; Zhang, H. X.; Xue, C. X.; Liu, H. X.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *SAR QSAR Environ. Res.,* **2005**, *16*, 349.

[93] Soderholm, A. A.; Lehtovuori, P. T.; Nyronen, T. H. *J. Med. Chem.,* **2005**, *48*, 917.

[94] Hemmateenejad, B.; Akhond, M.; Miri, R.; Shamsipur, M. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 1328.

[95] Viswanadhan, V. N.; Mueller, G. A.; Basak, S. C.; Weinstein, J. N. *J. Chem. Inf. Comput. Sci.,* **2001**, *41*, 505.

[96] Carosati, E.; Lemoine, H.; Spogli, R.; Grittner, D.; Mannhold, R.; Tabarrini, O.; Sabatini, S.; Cecchetti, V. *Bioorg. Med. Chem.,* **2005**, *13*, 5581.

[97] Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. *J. Med. Chem.,* **1999**, *42*, 3217.

[98] de Candia, M.; Summo, L.; Carrieri, A.; Altomare, C.; Nardecchia, A.; Cellamare, S.; Carotti, A. *Bioorg. Med. Chem.,* **2003**, *11*, 1439.

[99] Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Hur, E.; Tamm, K.; Kuruncz, i. L.; Karelson, M.; Varnek, A.; Solov'ev, V. P. *J. Med. Chem.,* **2006**, *49*, 3305.

[100] Gleeson, M. P.; Waters, N. J.; Paine, S. W.; Davis, A. M. *J. Med. Chem.,* **2006**, *49*, 1953.

[101] Liu, H. X.; Hu, R. J.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Comput. Aided Mol. Des.,* **2005**, *19*, 33.

[102] Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. *J. Pharm. Sci.,* **2001**, *90*, 749.

[103] Klopman, G.; Stefan, L. R.; Saiakhov, R. D. *Eur. J. Pharm. Sci.,* **2002**, *17*, 253.

[104] Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. *Pharm. Res.,* **1997**, *14*, 568.

[105] Oprea, T. I.; Gottfries, J. *J. Mol. Graph. Mod.,* **1999**, *17*, 261.

[106] Sun, H. M. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 748.

[107] Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. *J. Chem. Inf. Comput. Sci.,* **1998**, *38*, 726.

[108] Agatonovic-Kustrin, S.; Beresfordb, R.; Pauzi, A.; Yusof, M. *J. Pharm. Biomed. Anal.,* **2001**, *25*, 227.

[109] Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B. *Mol. Divers.,* **2004**, *8*, 379.

[110] Niwa, T. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 113.

[111] Bai, J. P. F.; Utis, A.; Crippen, G.; He, H.-D.; Fischer, V.; Tullman, R.; Yin, H.-Q.; Hsu, C.-P.; Jiang, L.; Hwang, K.-K. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 2061.

[112] Norinder, U.; Österberg, T.; Artursson, P. *Eur. J. Pharm. Sci.,* **1999**, *8*, 49.

[113] Norinder, U.; Österberg, T. *J. Pharm. Sci.,* **2001**, *90*, 1076.

[114] Norinder, U. *Neurocomputing,* **2003**, *55*, 337.

[115] Andrews, C. W.; Bennett, L.; Yu, L. X. *Pharm. Res.,* **2000**, *17*, 639.

[116] Turner, J. V.; Glass, B. D.; Agatonovic-Kustrin, S. *Anal. Chim. Acta,* **2003**, *485*, 89.

[117] Turner, J. V.; Maddalena, D. J.; Agatonovic-Kustrin, S. *Pharm. Res.,* **2004**, *21*, 68.

[118] Dorronsoro, I.; Chana, A.; Abasolo, M. I.; Castro, A.; Gil, C.; Stud, M.; Martinez, A. *Quant. Struct.-Act. Relat.,* **2004**, *23*, 89.

[119] Young, R. C.; Mitchell, R. C.; Brown, T. H.; Ganellin, C. R.; Griffith, R.; Jones, M.; Rana, K. K.; Saundesr, D.; Smith, I. R.; Sore, N. E.; Wilks, T. J. *J. Med. Chem.,* **1988**, *31*, 656.

[120] Abraham, M. H.; Chadha, H. S.; Mitchell, R. *J. Pharm. Sci.,* **1994**, *83*, 1257.

[121] Lombardo, F.; Blake, J. F.; Curatolo, W. J. *J. Med. Chem.,* **1996**, *39*, 4750.

[122] Kaliszan, R.; Markuszewski, M. *Int. J. Pharm.,* **1996**, *145*, 9.

[123] Segarra, V.; Lopez, M.; Ryder, H.; Palacios, J. M. *Quant. Struct.-Act. Relat.,* **1999**, *18*, 474.

[124] Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J. P. *Pharm. Res.,* **1999**, *16*, 1514.

[125] Clark, D. E. *J. Pharm. Sci.,* **1999**, *88*, 815.

[126] Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.,* **2000**, *43*, 3714.

[127] Keserü, G. M.; Molnár, L. *J. Chem. Inf. Comput. Sci.,* **2001**, *41*, 120.

[128] Liu, R.; Sun, H.; So, S. S. *J. Chem. Inf. Comput. Sci.,* **2001**, *41*, 1623.

[129] Platts, J. A.; Abraham, M. H.; Zhao, Y. H.; Hersey, A.; Ijaz, L.; Butina, D. *Eur. J. Med. Chem.,* **2001**, *36*, 719.

[130] Kaznessis, Y. N.; Snow, M. E.; Blankley, C. J. *J. Comput. Aided Mol. Des.,* **2001**, *15*, 697.

[131] Hou, T. J.; Xu, X. J. *J. Mol. Model.,* **2002**, *8*, 337.

[132] Rose, K.; Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 651.

[133] Iyer, M.; Mishru, R.; Han, Y.; Hopfinger, A. J. *Pharm. Res.,* **2002**, *19*, 1611.

[134] Lobell, M.; Molnár, L.; Keserü, G. M. *J. Pharm. Sci.,* **2003**, *92*, 360.

[135] Hou, T. J.; Xu, X. J. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 2137.

[136] Pan, D.; Iyer, M.; Liu, J.; Li, Y.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 2083.

[137] Narayanan, R.; Gunturi, S. B. *Bioorg. Med. Chem.,* **2005**, *13*, 3017.

[138] Cheng, A.; Diller, D. J.; Dixon, S. L.; Egan, W. J.; Lauri, G.; Merz, K. M. J. *J. Comput. Chem.,* **2002**, *23*, 172.
[139] Feher, M.; Sourial, E.; Schmidt, J. M. *Int. J. Pharm.,* **2000**, *201*, 239.
[140] Labute, P. *J. Mol. Graph. Model.,* **2000**, *18*, 464.
[141] Norinder, U.; Sjöberg, P.; Österberg, T. *J. Pharm. Sci.,* **1998**, *87*, 952.
[142] Luco, J. M. *J. Chem. Inf. Comput. Sci.,* **1999**, *39*, 396.
[143] Osterberg, T.; Norinder, U. *Eur. J. Pharm. Sci.,* **2001**, *12*, 327.
[144] Ooms, F.; Weber, P.; Carrupt, P. A.; Testa, B. *Biochim. Biophys. Acta,* **2002**, *1587*, 118.
[145] Subramanian, G.; Kitchen, D. B. *J. Comput. Aided Mol. Des.,* **2003**, *17*, 643.
[146] Winkler, D. A.; Burden, F. R. *J. Mol. Graph. Model.,* **2004**, *22*, 499.
[147] Yap, C. W.; Chen, Y. Z. *J. Pharm. Sci.,* **2005**, *94*, 153.
[148] Hall, L. M.; Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.,* **2004**, *43*, 2120.
[149] Colmenarejo, G.; Alvarez-Pedraglio, A.; Lavandera, J. L. *J. Med. Chem.,* **2001**, *44*, 4370.
[150] Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 1693.
[151] Agatonovic-Kustrin, S.; Ling, L. H.; Tham, S. Y.; Alany, R. G. *J. Pharm. Biomed. Anal.,* **2002**, *29*, 103.
[152] Ng, C.; Xiao, Y. D.; Putnam, W.; Lum, B.; Tropsha, A. *J. Pharm. Sci.,* **2004**, *93*, 2535.
[153] Turner, J. V.; Maddalena, D. J.; Cutler, D. J. *Int. J. Pharm.,* **2004**, *270*, 209.
[154] Karalis, V.; Tsantili-Kakoulidou, A.; Macheras, P. *Eur. J. Pharm. Sci.,* **2003**, *20*, 115.
[155] Klein, C.; Kaiser, D.; Kopp, S.; Chiba, P.; Ecker, G. F. *J. Comput. Aided Mol. Des.,* **2002**, *16*, 785.
[156] Shoji, R.; Kawakami, M. *Mol. Divers.,* **2006**, *10*, 101.
[157] Cash, G. G. *Mutat. Res.,* **2001**, *491*, 31.
[158] Cash, G.; Anderson, B.; Mayo, K.; Bogaczyk, S.; Tunkel, J. *Mutat. Res.,* **2005**, *585*, 170.
[159] Roy, K.; Ghosh, G. *Bioorg. Med. Chem.,* **2005**, *13*, 1185.
[160] Moridani, M. Y.; Siraki, A.; O'Brien, P. J. *Chem. Biol. Interact.,* **2003**, *145*, 213.
[161] Roy, K.; Ghosh, G. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 559.
[162] Mazzatorta, P.; Smiesko, M.; Lo Piparo, E.; Benfenati, E. *J. Chem. Inf. Model.,* **2005**, *45*, 1767.
[163] Roy, K.; Leonard, J. T. *Bioorg. Med. Chem.,* **2005**, *13*, 2967.
[164] Verma, R. P.; Hansch, C. *Mol. Pharmacol.,* **2006**, *3*, 441.
[165] Selassie, C. D.; Kapur, S.; Verma, R. P.; Rosario, M. *J. Med. Chem.,* **2005**, *48*, 7234.
[166] Saiz-Urra, L.; Gonzalez, M. P.; Teijeira, M. *Bioorg. Med. Chem.,* **2006**, *14*, 7347.
[167] Sovadinova, I.; Blaha, L.; Janosek, J.; Hilscherova, K.; Giesy, J. P.; Jones, P. D.; Holoubek, I. *Environ. Toxicol. Chem.,* **2006**, *25*, 1291.
[168] Aptula, A. O.; Roberts, D. W.; Cronin, M. T.; Schultz, T. W. *Chem. Res. Toxicol.,* **2005**, *18*, 844.
[169] Netzeva, T. I.; Schultz, T. W.; Aptula, A. O.; Cronin, M. T. *SAR QSAR Environ. Res.,* **2003**, *14*, 265.
[170] Ren, S. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 1679.
[171] Kaiser, K. L.; Niculescu, S. P.; Schultz, T. W. *SAR QSAR Environ. Res.,* **2002**, *13*, 57.
[172] Vracko, M.; Bandelj, V.; Barbieri, P.; Benfenati, E.; Chaudhry, Q.; Cronin, M.; Devillers, J.; Gallegos, A.; Gini, G.; Gramatica, P.; Helma, C.; Mazzatorta, P.; Neagu, D.; Netzeva, T.; Pavan, M.; Patlewicz, G.; Randic, M.; Tsakovska, I.; Worth, A. *SAR QSAR Environ. Res.,* **2006**, *17*, 265.
[173] Pavan, M.; Netzeva, T. I.; Worth, A. P. *SAR QSAR Environ. Res.,* **2006**, *17*, 147.
[174] Liu, X.; Chen, J.; Yu, H.; Zhao, J.; Giesy, J. P.; Wang, X. *Chemosphere,* **2006**, *64*, 1619.
[175] Bermudez-Saldana, J. M.; Cronin, M. T. *Pest. Manag. Sci.,* **2006**, *62*, 819.
[176] Zhao, C. Y.; Zhang, H. X.; Zhang, X. Y.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *Toxicology,* **2006**, *217*, 105.

Received: 11 December, 2006 **Revised: 14 May, 2007** **Accepted: 15 May, 2007**