Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Molecular Graphics and Modelling 24 (2006) 383–395
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Quantitative structure–pharmacokinetic relationships for drug clearance by using statistical learning methods

C.W. Yap [a], Z.R. Li [a,b], Y.Z. Chen [a,c,*]

[a] *Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore*
[b] *College of Chemistry, Sichuan University, Chengdu 610064, PR China*
[c] *Shanghai Center for Bioinformation Technology, Shanghai 201203, PR China*

Received 27 July 2005; accepted 4 October 2005
Available online 14 November 2005

Abstract

Quantitative structure–pharmacokinetic relationships (QSPkR) have increasingly been used for the prediction of the pharmacokinetic properties of drug leads. Several QSPkR models have been developed to predict the total clearance (CL_{tot}) of a compound. These models give good prediction accuracy but they are primarily based on a limited number of related compounds which are significantly lesser in number and diversity than the 503 compounds with known CL_{tot} described in the literature. It is desirable to examine whether these and other statistical learning methods can be used for predicting the CL_{tot} of a more diverse set of compounds. In this work, three statistical learning methods, general regression neural network (GRNN), support vector regression (SVR) and k-nearest neighbour (KNN) were explored for modeling the CL_{tot} of all of the 503 known compounds. Six different sets of molecular descriptors, DS-MIXED, DS-3DMoRSE, DS-ATS, DS-GETAWAY, DS-RDF and DS-WHIM, were evaluated for their usefulness in the prediction of CL_{tot}. GRNN-, SVR- and KNN-developed models have average-fold errors in the range of 1.63 to 1.96, 1.66–1.95 and 1.90–2.23, respectively. For the best GRNN-, SVR- and KNN-developed models, the percentage of compounds with predicted CL_{tot} within two-fold error of actual values are in the range of 61.9–74.3% and are comparable or slightly better than those of earlier studies. QSPkR models developed by using DS-MIXED, which is a collection of constitutional, geometrical, topological and electrotopological descriptors, generally give better prediction accuracies than those developed by using other descriptor sets. These results suggest that GRNN, SVR, and their consensus model are potentially useful for predicting QSPkR properties of drug leads.

Keywords: Clearance; Consensus models; Computational ADME; General regression neural network; k-nearest neighbour; QSAR; Support vector regression

1. Introduction

Drug clearance is measured by a quantity, total clearance (CL_{tot}), which is a proportionality constant describing the relationship between a substance's rate of transfer, in amount per unit time, and its concentration, in an appropriate reference fluid [1]. Drug clearance occurs by perfusion of blood to the organs of extraction, which are generally the liver and the kidney [2]. The CL_{tot} value of a drug is an important pharmacokinetic parameter because it is directly related to bioavailability and drug elimination and can be used to determine the dosing rate and steady-state concentration of a drug [3]. Thus, it is important to predict the CL_{tot} value of drug leads during drug discovery so that compounds with acceptable metabolic stability can be identified and those with poor bioavailability can be eliminated.

Traditionally, the CL_{tot} value of a drug candidate is obtained via in vivo and in vitro studies [4–7], which tends to be time-consuming and costly. Therefore, an in silico method, quantitative structure–pharmacokinetic relationship (QSPkR) modeling, has recently been explored for predicting the CL_{tot} value of drug candidates [8–12] in an effort to eliminate undesirable agents in a fast and cost-effective manner. An initial PLS study conducted by Karalis et al. [8] using 272 structurally unrelated compounds failed to find any correlation between CL_{tot} and a large variety of molecular descriptors used in that study. Karalis et al. [9] then developed a partial least square (PLS) model and non-linear regression model for CL_{tot} by using 23 cephalosporins. The r^2 and q^2 values of the PLS-developed model are 0.775 and 0.731, while the r^2 value of the

* Corresponding author. Tel.: +65 6874 6877; fax: +65 6774 6756.
E-mail address: yzchen@cz3.nus.edu.sg (Y.Z. Chen).

1093-3263/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2005.10.004

non-linear regression model is 0.804. These two studies suggest that multiple mechanisms may be involved in CL_{tot} and thus linear methods may not always be suitable for constructing QSPkR models for CL_{tot}. Another study for the prediction of CL_{tot} was done by Turner et al. [10] who used artificial neural network (ANN), which gives a r^2 value of 0.982 for a training set of 16 cephalosporins and a r^2 value of 0.998 for a validation set of four cephalosporins. Subsequently, Turner et al. [11] used a larger training set of 56 compounds to develop an ANN-based QSPkR model, which gives a r^2 value of 0.731 for a validation set of six compounds. These results suggest that non-linear methods may be useful for developing models for CL_{tot} prediction of structurally unrelated compounds. Two QSPkR models for CL_{tot} were developed by Ng et al. [12] by using k-nearest neighbour (KNN) and PLS. The KNN-developed QSPkR model gives a q^2 value of 0.77 for a training set of 38 antimicrobial agents and a r^2 value of 0.94 for a validation set of six antimicrobial agents. There are 68% of the 44 compounds having predicted CL_{tot} within two-fold of actual values. For the PLS-developed QSPkR model, there are only 50% of the 44 compounds having predicted CL_{tot} within two-fold of actual values and the q^2 value of this model is 0.09 for the training set and its r^2 value is 0.35 for the validation set. These results are consistent with the study of Turner [11] and further confirm the usefulness of non-linear methods for developing QSPkR models for predicting CL_{tot}. All of the previous QSPkR models for predicting CL_{tot} have primarily been developed and tested by using a relatively small number of compounds (<70), which is significantly smaller in number and diversity than the number of compounds with known CL_{tot} data. Thus, it is of interest to evaluate the prediction capabilities of QSPkR models that are developed by using much larger and more diverse datasets.

Recently, non-linear statistical learning methods such as KNN [12], general regression neural network (GRNN) [13] and support vector regression (SVR) [14] have shown promising potential for predicting compounds of various pharmacokinetic and pharmacodynamic properties. GRNN has been explored for QSPkR modeling of drug distribution properties [13] and human intestinal absorption [15]. SVR has been applied to blood brain barrier penetration [14] and human intestinal absorption [14]. KNN has been used for the prediction of CL_{tot} [12] as well as metabolic stability of drug candidates [16]. It is of interest to evaluate the usefulness of these methods and other non-linear statistical learning methods for the prediction of CL_{tot}.

This work is intended to evaluate the capability of several statistical learning methods for predicting CL_{tot} by using 503 compounds found from a comprehensive literature search, which is substantially larger in number and more diverse in structure than those used in earlier studies. The methods used include GRNN, SVR and KNN. Different descriptor sets, which encode different combination of the structural and physiochemical properties of a compound, were also compared for their usefulness for constructing QSPkR models to predict CL_{tot}. Consensus modeling strategy has been introduced for developing prediction systems based on multiple models [17,18]. In this work, this strategy was also applied to the development of consensus QSPkR (cQSPkR) models for the prediction of CL_{tot} by using QSPkR models generated from different statistical learning methods.

2. Method

2.1. Dataset

Compounds with known human CL_{tot} values were selected from several sources including *Micromedex* [19], a classic pharmacology textbook [20] and a number of publications [5,6,10–12,21,22]. In order to ensure that experimental variations in determining CL_{tot} do not significantly affect the quality of our data sets, only CL_{tot} values obtained from healthy adult males and from intravenous administration were used for constructing the dataset. In addition, a number of compounds were excluded because they are known to possess certain molecular characteristics which do not permit reliable calculations of the molecular descriptors used in this study [8]. Examples of these compounds are quarternary ammonium compounds, molecules with complex chemical structures like amphotericin-B, aminoglycosides, vancomycin, and compounds containing one or more metal atoms. A total of 503 compounds were selected from this process and these were used as the dataset for this work. The CL_{tot} value for each of these compounds was log-transformed ($\log CL_{tot}$) to normalize the data and to reduce unequal error variances [23].

Representative training set and validation set were constructed from our dataset according to their distribution in the chemical space by using a method used in several studies [24–26]. Here, chemical space is defined by the structural and chemical descriptors used to represent a compound. Each compound occupies a particular location in this chemical space. All possible pairs of these compounds were generated and a similarity score was computed for each pair. These pairs were then ranked in terms of their similarity scores, based on which compounds of similar structural and chemical features were evenly assigned into separate datasets. For those compounds without enough structurally and chemically similar counterparts, they were assigned to the training set. After the dataset separation procedure, the training set and validation set contain 398 and 105 compounds, respectively. The list of compounds with their CL_{tot} values and their allocation into training and validation sets is provided in the supplementary material.

Prediction capability of QSPkR models is known to be strongly affected by the diversity of samples used in the training set [27,28]. Independent validation sets have frequently been used for evaluating the predictive performance of these QSPkR models, and these need also to be sufficiently diverse and representative of the samples studied in order to accurately assess the capabilities of the QSPkR models [27,28]. The diversity of a dataset can be estimated by a diversity index (DI) which is the average value of the similarity between all of the pairs of compounds in that dataset [29]:

$$\mathrm{DI} = \frac{\sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} \mathrm{sim}(i, j)}{n(n-1)} \quad (1)$$

where sim(i,j) is a measure of the similarity between compound i and j, and n is the number of compounds in a dataset. The diversity of a dataset increases with decreasing DI. The similarity between two compound i and j is commonly described by the Tanimoto coefficient [30–32]:

$$\mathrm{sim}(i,j) = \frac{\sum_{d=1}^{p} x_{di}x_{dj}}{\sum_{d=1}^{p}(x_{di})^2 + \sum_{d=1}^{p}(x_{dj})^2 - \sum_{d=1}^{p} x_{di}x_{dj}} \quad (2)$$

where p is the number of descriptors of the compounds in the dataset. Similarly, the level of representativity of a validation set can be estimated by a representativity index (RI) which is the mean Tanimoto coefficient between the compounds in a validation set and those in the training set. The validation set is considered to be more representative of the training set if its RI value is high.

2.2. Molecular structures and descriptors

The 2D structure of each of the compounds studied was generated by using DS ViewerPro 5.0 [33], which was subsequently converted into 3D structure by using CONCORD [34]. The 3D structure of each compound was manually inspected to ensure that the chirality of each chiral agent is properly represented. All salts and elements, such as sodium or calcium, were removed prior to descriptor calculation. Six different sets of descriptors were used to describe the structural and physico-chemical properties of the compounds. The first set (DS-MIXED) contains a number of commonly used descriptors, including 21 constitutional descriptors, six geometrical descriptors, 72 topological descriptors and 108 electrotopological state descriptors [35]. The second set (DS-3DMoRSE) includes 224 3D-MoRSE descriptors [36], which are representations of the 3D structure of a molecule and encode features such as molecular weight, van der Waals volume, electronegativities and polarizabilities. The third set (DS-ATS) is composed of 209 Moreau-Broto topological autocorrelation (ATS) descriptors [37], which describes how molecular properties such as polarizability, charge, electronegativity, are distributed along the topological structure. The fourth set (DS-GETAWAY) consists of 340 GETAWAY descriptors [38], which encodes both molecular structure and chemical information such as atomic mass, polarizability, van der Waals volume and electronegativity. The fifth set (DS-RDF) contains 203 RDF descriptors [39], which provides information about interatomic distances in the entire molecule and also other useful information such as bond distances, ring types, planar and non-planar systems, atom types and molecular weight. The last set (DS-WHIM) includes 126 WHIM descriptors [40], which encodes information about the size, shape, symmetry, atom distribution and polarizability of a molecule. All of the descriptors were computed from the 3D structure of each compound using our own designed molecular descriptor computing program. Our program has been tested on a number of compounds with readily available descriptor values [41] to ensure the accuracy of our computed descriptors.

Objective feature selection is applied to all of the six sets of descriptors to remove descriptors irrelevant or redundant to the CL_{tot} of the compounds, so as to improve computation speed, performance and interpretability of predictive models. The first step involves the removal of all irrelevant descriptors such as constant descriptors. Redundant descriptors were then eliminated by removing one of the two descriptors with pairwise correlation coefficient of greater than 0.90 [42,43]. The final number of descriptors for each descriptor set is 84, 109, 142, 155, 111 and 44 for DS-MIXED, DS-3DMoRSE, DS-ATS, DS-GETAWAY, DS-RDF, and DS-WHIM, respectively. All of the remaining descriptors in each descriptor set were autoscaled to a mean value of zero and a variance of one to ensure that all descriptors have equal potential to affect the QSPkR model [44].

2.3. Statistical learning methods

2.3.1. GRNN algorithm

GRNN was introduced by Specht in 1991 [45] and is a form of neural network designed for regression through the use of Bayes' optimal decision rule. For GRNN, the predicted value of the response is the most probable value $E[y|\boldsymbol{x}]$, which is given by

$$\hat{y}(\boldsymbol{x}) = E[y|\boldsymbol{x}] = \frac{\int_{-\infty}^{\infty} y f(\boldsymbol{x},y)\,\mathrm{d}y}{\int_{-\infty}^{\infty} f(\boldsymbol{x},y)\,\mathrm{d}y} \quad (3)$$

where $f(\boldsymbol{x},y)$ is the joint density and can be estimated by using Parzen's nonparametric estimator [46]:

$$g(\boldsymbol{x}) = \frac{1}{n\sigma}\sum_{i=1}^{n} W\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{\sigma}\right) \quad (4)$$

where n is the training set size, σ is a scaling parameter which defines the width of the bell curve that surrounds each sample point, $W(d)$ is a weight function which has its largest value at $d = 0$ and $(\boldsymbol{x} - \boldsymbol{x}_i)$ is the distance between a to-be-predicted vector $\boldsymbol{x}$ and a individual vector $\boldsymbol{x}_i$ in the training set. Substituting Parzen's nonparametric estimator for $f(\boldsymbol{x},y)$ and performing the integrations leads to the fundamental equation of GRNN.

$$\hat{y}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} y_i \exp\left(-D(\boldsymbol{x},\boldsymbol{x}_i)\right)}{\sum_{i=1}^{n} \exp\left(-D(\boldsymbol{x},\boldsymbol{x}_i)\right)} \quad (5)$$

where

$$D(\boldsymbol{x},\boldsymbol{x}_i) = \sum_{j=1}^{p}\left(\frac{x_j - x_{ij}}{\sigma}\right)^2 \quad (6)$$

2.3.2. SVR algorithm

The theory of SVR has been extensively described [47,48]. Thus, only a brief description is given here. SVR is based on the structural risk minimization principle from statistical learning theory [49]. Each instance is represented by a vector $\boldsymbol{x}$ with molecular descriptors as its components. A kernel function, $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, is used to map the vectors into a higher dimensional

feature space and linear regression is then conducted in this space. The optimal regression function can be represented by:

$$\hat{y}(\boldsymbol{x}) = \sum_{i=1}^{l}(\alpha^* - \alpha)K(\boldsymbol{x}, \boldsymbol{x}_i) + b \tag{7}$$

where l is the number of support vectors and the coefficients α, α^* and bias b are determined by maximizing the following Langrangian expression:

$$-\varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}(\alpha_i + \alpha_i^*)y_i - \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i + \alpha_i^*)(\alpha_j + \alpha_j^*) \times (x_i, x_j) \tag{8}$$

under the following conditions:

$$0 \leq \alpha_i, \alpha_i^* \leq C \tag{9}$$

$$\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) = 0 \tag{10}$$

where n is the training set size and C is a penalty for training errors.

2.3.3. KNN algorithm

KNN measures the Euclidean distance between a to-be-predicted vector $\boldsymbol{x}$ and each individual vector $\boldsymbol{x}_i$ in the training set [50,51]. A total of k number of vectors nearest to the vector $\boldsymbol{x}$ are then used to determine its response value:

$$\hat{y}(\boldsymbol{x}) = \frac{\sum_{i=1}^{k} y_i}{k} \tag{11}$$

2.3.4. Optimization of the parameters of GRNN, SVR and KNN

In GRNN, the only parameter to be optimized is the scaling parameter, σ. SVR was trained by using a Gaussian kernel function with an adjustable parameter σ. For KNN, the optimum number of nearest neighbours, k, needs to be derived for each training set. Optimization of the parameter for each of these statistical learning methods was conducted by scanning the parameter value through a range from 1 to 30. The predictive capability of the QSPkR model developed from a particular parameter value can be determined by using cross-validation methods, such as five-fold cross validation, 10-fold cross-validation and modeling testing set. An earlier study has shown that the use of a modeling testing set gives the best performance for assessing the predictive capability of a model [24]. Thus, this validation method was used to select the optimum parameter for each statistical learning method in this study. The following function was used to measure the predictive capability of a QSPkR model [42,43]:

$$F = \mathrm{MAE}_{\mathrm{train}} + |\mathrm{MAE}_{\mathrm{train}} - \mathrm{MAE}_{\mathrm{test}}| \tag{12}$$

where $\mathrm{MAE}_{\mathrm{train}}$ and $\mathrm{MAE}_{\mathrm{test}}$ are the mean absolute error of the modeling training set and modeling testing set, respectively. The modeling testing set was derived by dividing the original training set into a modeling training set and modeling testing set of 303 and 95 compounds, respectively, by using the same procedure for dividing a dataset into the training set and validation set described in the previous section.

2.4. cQSPkR method

In this work, cQSPkR models were developed by combining QSPkR models generated from different statistical learning methods. cQSPkR models compute the predicted $\mathrm{CL}_{\mathrm{tot}}$ of a compound by averaging the predicted $\mathrm{CL}_{\mathrm{tot}}$ of that compound from the different QSPkR models [52].

2.5. Evaluation of QSPkR models

The validation set, not used in the derivation of the QSPkR models, was used to estimate the prediction capability of the QSPkR models. LOO and 10-fold cross-validation were not used in this work because there are reports of the lack of correlation between cross-validation methods and the prediction capability of a QSPkR model [53–56]. Moreover, cross-validation methods have a tendency of underestimating the prediction capability of a QSPkR model, especially if important molecular features are present in only a minority of the compounds in the training set [42,57]. Thus, a model having low cross-validation results can still be quite predictive [42].

The fold-error for each compound was determined as follows [6]:

$$\text{fold-error} = \begin{cases} \dfrac{\hat{y}(\boldsymbol{x})}{y(\boldsymbol{x})} & \text{if } \hat{y}(\boldsymbol{x}) > y(\boldsymbol{x}) \\ \dfrac{y(\boldsymbol{x})}{\hat{y}(\boldsymbol{x})} & \text{all others} \end{cases} \tag{13}$$

and the percentage of compounds in the validation set where the fold-error is less than two or three were calculated. The predictive capability of the QSPkR models can be measured by the Spearman rank correlation coefficient (R_{s}) and average-fold error [58]. R_{s} is used to assess the ability of the QSPkR models to rank compounds based on their $\mathrm{CL}_{\mathrm{tot}}$. The average-fold error is the geometric mean of the ratio of predicted and actual values, and can be computed by:

$$\text{Average-fold error} = 10^{\sum(|\log \hat{y}(\boldsymbol{x})/y(\boldsymbol{x})|)/n} \tag{14}$$

The average-fold error avoids the cases in which poor over-predictions are cancelled by equally poor underpredictions. A QSPkR model that predicts $\mathrm{CL}_{\mathrm{tot}}$ perfectly gives a value of 1 and a model with an average-fold error of less than 2 is considered to be a successful one [58]. The predicted log $\mathrm{CL}_{\mathrm{tot}}$ values of the compounds were converted back to $\mathrm{CL}_{\mathrm{tot}}$ prior to the calculation of fold-errors and average-fold errors.

2.6. Functional dependence study of QSPkR models

Descriptors in models developed by using non-linear statistical learning methods are related to $\mathrm{CL}_{\mathrm{tot}}$ in the form of a non-linear relationship which can potentially provide more

information about the relationships between the descriptors and CL_{tot} than those in models developed by linear methods. QSPkR models usually contain descriptors correlated with each other which make it difficult to determine the relationship between a specific molecular characteristic and CL_{tot}. A principal component analysis (PCA)-based method was used in this study to overcome this difficulty. This method has been described in detail elsewhere [13], thus only a brief description is given here. PCA was used to extract dominant patterns in the descriptor subsets and to group similar descriptors under a single principal component (PC). Different PCs encode different molecular characteristics and the orthogonality among the PCs can be exploited to determine the correlation between a molecular characteristic and CL_{tot} without the influence of other molecular characteristics. Artificial testing sets containing artificial compounds with varying PC values were created and their log CL_{tot} were predicted by using the models developed from each of the three statistical learning methods. The RIs of the artificial testing sets were computed to ensure that the artificial testing sets were representative of the training set and thus were suitable for functional dependence study. Plots of log CL_{tot} against the PCs can then be used to find the trends between various molecular characteristics and CL_{tot}.

3. Results and discussion

3.1. Dataset analysis

The DI of the training set and validation set used in this study and those of several reference datasets are given in Table 1. It is found that the DI values of the training set and validation set is very small, as low as 0.067, which is at the level of those of highly diverse datasets. For comparison, the DI values of datasets containing congeneric compounds are typically greater than 0.452, and those of the compounds used in QSAR and QSPR studies are typically in the range of 0.274–0.584. This suggests that the training set and validation set are sufficiently diverse. The RI value between the training set and validation set is 0.881, which suggests that the validation set is representative of the training set and thus is suitable for assessing the predictive capability of the QSPkR models developed in this work.

Table 1
Diversity indices of the datasets used in this and other studies

	Dataset	Number of compounds	Diversity index
Datasets used in this work	Training set	398	0.067
	Validation set	105	0.068
Highly diverse datasets	Satellite structures [70]	8	0.076
	FDA approved drugs	1121	0.069
	NCI diversity set [71]	1804	0.124
Congeneric datasets	Penicillins	59	0.452
	Cephalosporins	73	0.568
	Fluoroquinolones	39	0.579
QSAR, QSPR datasets	Estrogen receptor ligands [72]	1009	0.274
	Benzodiazepine receptor ligands [72]	405	0.314
	Dihydrofolate reductase (DHFR) inhibitors [72]	756	0.384
	Cyclooxygenase 2 (COX2) inhibitors [72]	467	0.584

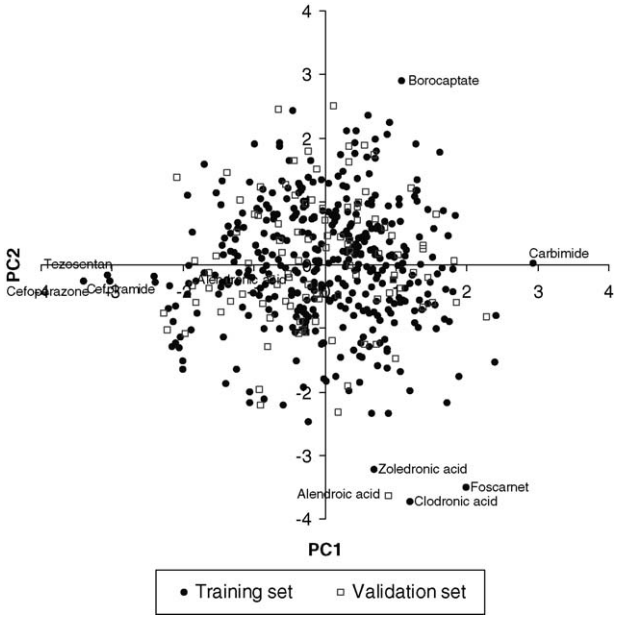

Fig. 1. Score plot of the first two principal components for training set and validation set.

PCA [59] was performed by using the dataset of 503 compounds to identify outliers and clusters. Two principal components were derived which is able to explain 73.2% of the total variance in the descriptors. Component one and two are able to explain 60.9 and 12.3% of the variance, respectively. Fig. 1 shows a score plot of the compounds in the training set and validation set by using the first two principal components. There are no distinct clusters in the training set and validation set. The validation set is evenly distributed throughout the score space of the training set, confirming the representativeness of the validation set. Four compounds, alendronic acid, clodronic acid, foscarnet and zoledronic acid, were found to be farther away from the majority of compounds and are located at the bottom right of the score space. Other compounds that are farther away from the majority of compounds are cefoperazone, cefpiramide and tezosentan, which are located at the left of the score space, and carbimide and borocaptate, which are located at the right and top right of the score space, respectively. There seems to be no evidence to suggest that these compounds are outliers. Thus, they are retained in the training set and validation set.

3.2. Analysis of descriptor sets

The computed R_s values and average-fold errors of the QSPkR models developed by using different descriptor sets are

Table 2
Average-fold errors of QSPkR models developed by using different statistical learning methods and different descriptors sets

Statistical learning methods	Model	Descriptor set	Optimum parameter	R_s	Average-fold error
GRNN	G-MIXED	DS-MIXED	2	0.636	1.73
	G-3DMoRSE	DS-3DMoRSE	3	0.540	1.75
	G-ATS	DS-ATS	3	0.448	1.86
	G-GETAWAY	DS-GETAWAY	3	0.520	1.80
	G-RDF	DS-RDF	3	0.558	1.80
	G-WHIM	DS-WHIM	2	0.302	1.96
	G-ALL	DS-ALL	7	0.633	1.63
SVR	S-MIXED	DS-MIXED	3	0.558	1.73
	S-3DMoRSE	DS-3DMoRSE	4	0.518	1.81
	S-ATS	DS-ATS	7	0.548	1.74
	S-GETAWAY	DS-GETAWAY	8	0.564	1.78
	S-RDF	DS-RDF	4	0.607	1.76
	S-WHIM	DS-WHIM	5	0.346	1.95
	S-ALL	DS-ALL	13	0.643	1.66
KNN	K-MIXED	DS-MIXED	2	0.523	2.00
	K-3DMoRSE	DS-3DMoRSE	2	0.360	2.23
	K-ATS	DS-ATS	3	0.406	2.03
	K-GETAWAY	DS-GETAWAY	2	0.522	2.00
	K-RDF	DS-RDF	3	0.447	1.98
	K-WHIM	DS-WHIM	3	0.392	2.01
	K-ALL	DS-ALL	2	0.513	1.90
PLS	P-MIXED	DS-MIXED	17	0.528	1.89
	P-3DMoRSE	DS-3DMoRSE	8	0.377	2.26
	P-ATS	DS-ATS	7	0.562	2.09
	P-GETAWAY	DS-GETAWAY	10	0.474	1.92
	P-RDF	DS-RDF	6	0.468	1.99
	P-WHIM	DS-WHIM	28	0.282	2.10
	P-ALL	DS-ALL	5	0.559	1.96

The average-fold errors were assessed by using the validation set.

shown in Table 2. Comparison of the QSPkR models based on the six descriptor sets shows that models based on the DS-MIXED descriptor set generally give higher R_s values and lower average-fold errors than those based on other descriptors sets. This suggests that models based on the DS-MIXED descriptor set are more useful and it may be advantageous to use a variety of descriptors for prediction of pharmacokinetic properties than to use a specialized descriptor set which may partially neglect some important features.

The descriptors in the six descriptor sets were combined to form a new descriptor set (DS-ALL). The G-ALL, S-ALL and K-ALL models developed by using DS-ALL have higher predictive capabilities compared to models developed by using individual descriptor sets. This suggests that all of the three statistical learning methods are able to extract useful information from the different descriptor sets and to effectively combine them to develop more predictive QSPkR models.

3.3. Predictive capability of QSPkR and cQSPkR models

Table 2 shows the predictive capabilities of the QSPkR models developed by using GRNN, SVR and KNN. PLS was used as a reference QSPkR method for comparison of the predictive capabilities of the different models. The results for the corresponding PLS-developed QSPkR models are also given in Table 2. All of the GRNN- and SVR-developed QSPkR models have average-fold errors less than 2 while KNN-developed models have average-fold errors near 2, which are similar to those of PLS-developed models. GRNN- and SVR-developed QSPkR models were also found to generally give higher R_s values than the corresponding KNN- and PLS-developed models. This suggests that both GRNN and SVR are more useful than either KNN or PLS for developing QSPkR models of drug clearance [58].

To assess the performance of the three statistical learning methods for CL_{tot} prediction of a more diverse set of compounds, it is useful to examine whether the predictive capability of these methods is at a similar level as those derived from the use of a significantly smaller set of compounds. It is noted that, a direct comparison with results from previous studies is inappropriate because of the differences in the dataset, molecular descriptors, and computing algorithms used. Although desirable, it is impossible to conduct a separate comparison using results directly from other studies without full information about the algorithms of molecular descriptors and modeling methods used in each study. Nonetheless, a tentative comparison may provide some crude estimate regarding the approximate level of predictive capability of the QSPkR models studied in this work.

Table 3 gives the prediction results of the G-ALL, S-ALL, K-ALL and P-ALL models from this work along with those derived from previous studies. The percentage of compounds in

Table 3
Number of compounds with the predicted CL_{tot} within two-fold error of the actual CL_{tot} from this work and other studies

Model	Number of compounds	Number (%) of compounds with fold-errors <2
G-ALL (this work)	105	73 (69.5)
S-ALL (this work)	105	78 (74.3)
K-ALL (this work)	105	65 (61.9)
P-ALL (this work)	105	63 (60.0)
Multiple linear regression (Wajima et al. [6])	68	44 (64.7)
KNN (Ng et al. [12])	44	30 (68.2)
PLS (Ng et al. [12])	44	22 (50.0)
Parallel tube (Obach [22])	29	16 (55.2)

the validation set with predicted CL_{tot} within two-fold error of actual values of G-ALL and S-ALL models are comparable and in some cases slightly better than those of earlier studies that were tested by using a much smaller number of compounds. This suggests that statistical learning methods, particularly GRNN and SVR, are useful for prediction of CL_{tot} of a broad range of compounds. A possible reason for the better performance of GRNN and SVR is that multiple mechanisms are involved in determining CL_{tot}. A variety of factors may interact in complex ways to affect the CL_{tot} of a compound. Therefore, methods based only on linear relationships, such as PLS, may not be the most efficient approach for constructing a QSPkR model for predicting CL_{tot}. Thus, nonlinear methods, such as GRNN and SVR, which do not require prior knowledge about the molecular mechanism or structure-activity relationship of a particular drug property may be more suitable.

Plots of the predicted CL_{tot} against the actual values for the G-ALL and S-ALL models are shown in Fig. 2a and b. These plots show that both models tend to under-predict the CL_{tot} value of compounds rather than over-predicting the CL_{tot}. Under-prediction of CL_{tot} is more desirable than over-prediction of CL_{tot} during drug development because over-prediction results in more frequent dosing of a drug candidate during clinical trials which may lead to higher rates of adverse drug reactions. For compounds with fold-errors greater than 2, the G-ALL model underpredicted 22 and overpredicted 10 of these compounds, respectively. The corresponding values for the S-ALL model are 18 and 11, respectively. There are seven compounds having fold-errors greater than 3 for both models and their chemical structures are shown in Fig. 3. A possible reason for the high fold-errors of some of these compounds is that the descriptors used in this study may be inadequate to properly describe these compounds. Examples of these compounds are chlorphenamine and fendiline, which contain two aromatic rings separated by an atom, allopurinol, which contains a complex two rings system with multiple heteroatoms, carbidopa, which has a hydrazine group that is highly reactive and reducing, and raltitrexed, which has two carboxylic acid groups that makes it highly charged at physiological pH. Studies have suggested that compounds containing these structural features may not be adequately represented by currently available descriptors [24–26]. Thus, by using the currently available algorithm, these compounds are misrepresented and incorrectly positioned in the chemical space, leading to inaccurate prediction of their CL_{tot} values.

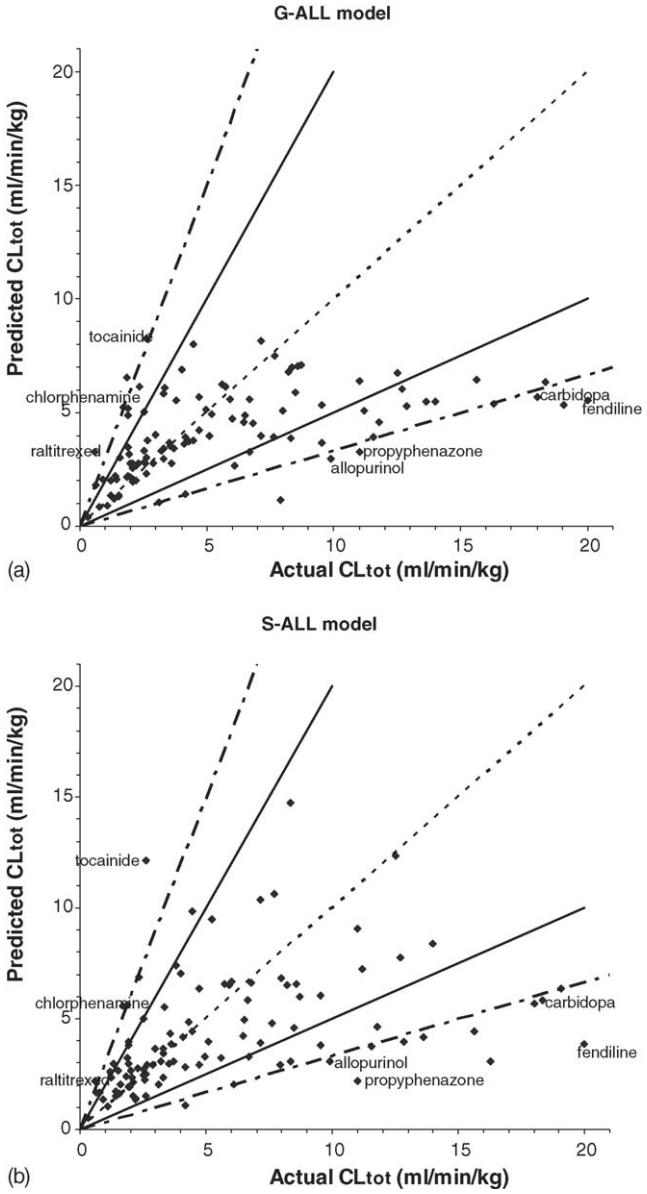

Fig. 2. (a) Plot of predicted CL_{tot} vs. actual CL_{tot} for the G-ALL model. The dotted line represents line of unity. The area between the two solid lines and between the two dotted-dash lines represents an area between two-fold and three-fold error, respectively. Compounds in validation set with fold-error greater than 3 for both G-ALL and S-ALL models are identified. (b) Plot of predicted CL_{tot} vs. actual CL_{tot} for the S-ALL model. The dotted line represents line of unity. The area between the two solid lines and between the two dotted-dash lines represents an area between two-fold and three-fold error, respectively. Compounds in validation set with fold-error greater than three for both G-ALL and S-ALL models are identified.

A cQSPkR model was developed by using G-ALL and S-ALL models. The K-ALL model was not used because its prediction capability is significantly lower than those of the G-ALL and S-ALL models and hence may reduce the prediction capability of the cQSPkR model. The cQSPkR model has an average-fold error of 1.61. Thus, the cQSPkR model had slightly better prediction capability than either the G-ALL

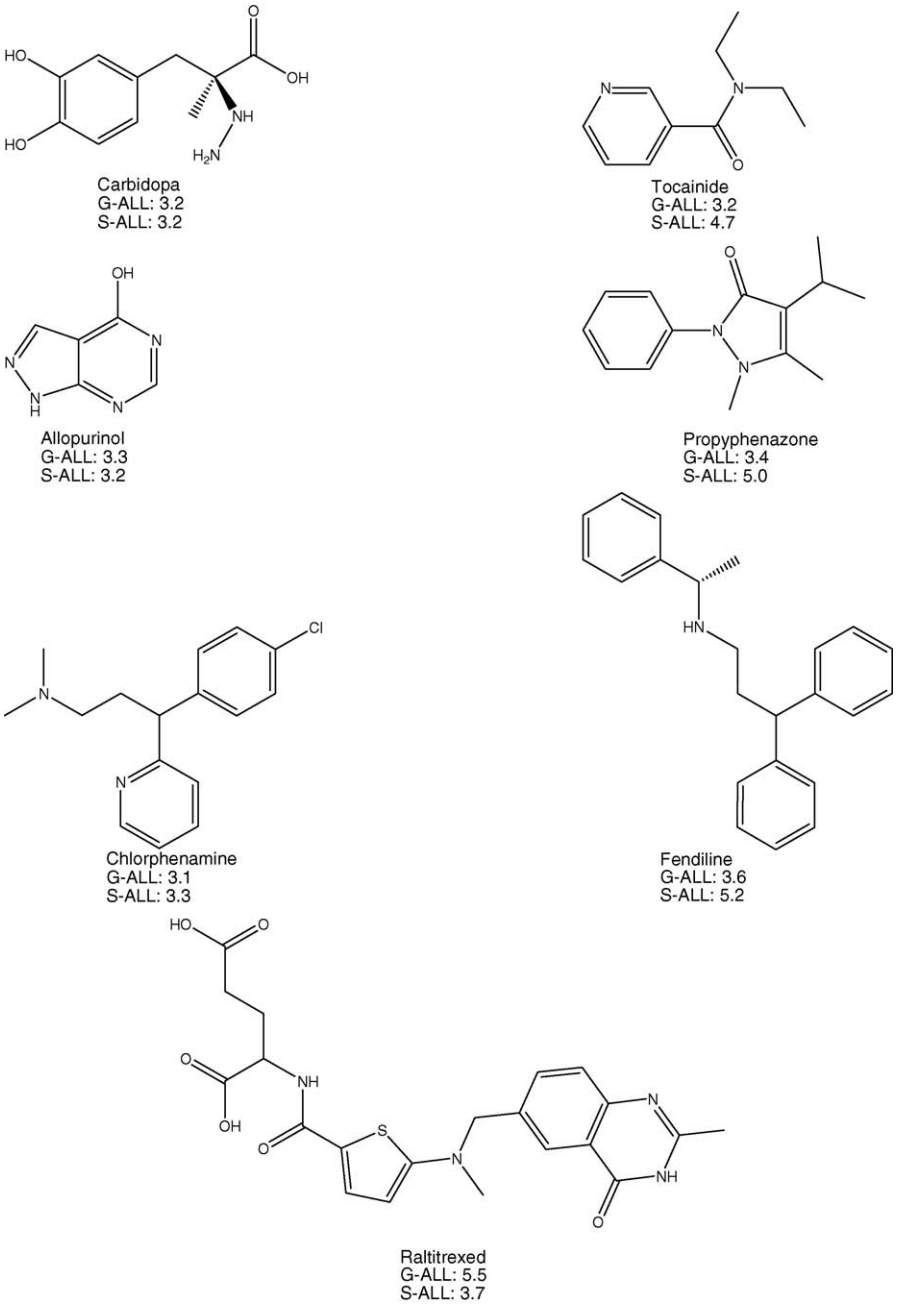

Fig. 3. Chemical structures of compounds in validation set with fold-errors greater than three for both G-ALL and S-ALL models. The numbers represent the fold-errors of each compound for both G-ALL and S-ALL models.

or S-ALL model. The relatively small average-fold error suggests that the model is useful for the prediction of CL_{tot}. The cQSPkR model correctly predicted 77 (73.3%) compounds in validation set to be within two-fold error of actual CL_{tot}. For compounds with fold-errors greater than 2, the cQSPkR model under-predicted 19 and over-predicted 9 of these compounds, respectively. None of the under-predicted or over-predicted compounds have fold-errors greater than 4.5. This is significantly improved over that of the G-ALL and S-ALL models which have two and four compounds with fold-errors greater than 4.5, respectively. The cQSPkR model gives an R_s value of 0.652 which suggests that the model may be useful for ranking compounds according to their CL_{tot} in large chemical libraries.

3.4. Functional dependence analysis

Multiple elimination processes are involved in drug clearances. Thus, it is difficult to determine which molecular characteristics are important in affecting CL_{tot}. Nonetheless, it is possible to infer some information from a functional dependence study of the QSPkR models. It is noted that the results of a functional dependence study may vary with respect to different QSPkR models. Thus, the following interpretation

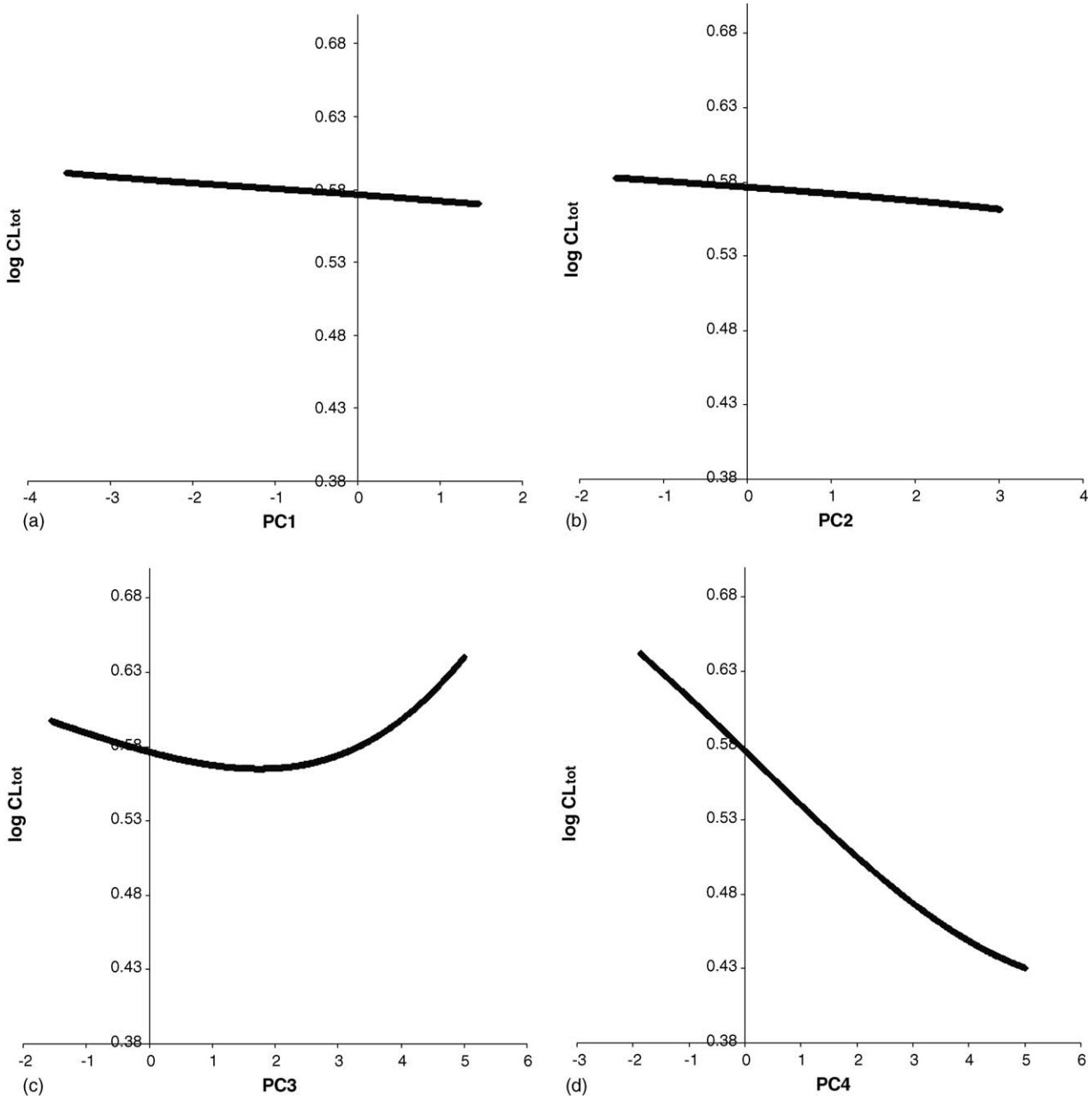

Fig. 4. (a) Plot of log CL_{tot} against first PC for G-ALL model. Increasing values of PC1 denotes increasing sphericity of a compound. (b) Plot of log CL_{tot} against second PC for G-ALL model. Increasing values of PC2 denotes decreasing lipophilicity of a compound. (c) Plot of log CL_{tot} against third PC for G-ALL model. Increasing values of PC3 denotes decreasing flexibility of a compound. (d) Plot of log CL_{tot} against fourth PC for G-ALL model. Increasing values of PC4 denotes increasing molecular size of a compound. (e) Plot of log CL_{tot} against fifth PC for G-ALL model. (f) Plot of log CL_{tot} against sixth PC for G-ALL model. Increasing values of PC6 denotes increasing hydrogen bond accepting ability of a compound. (g) Plot of log CL_{tot} against seventh PC for G-ALL model. Increasing values of PC7 denotes increasing hydrogen bond donoring ability of a compound.

of the descriptors must be taken in light of the absolute predictive ability of the QSPkR models. Fig. 4a–g shows the prediction results of the first seven PCs of the G-MIXED model by using artificial testing sets. The first seven PCs are able to explain approximately 60% of the total variance of the descriptors. Plots of log CL_{tot} against the PCs for the S-MIXED model are similar and thus are not given here. The DS-MIXED descriptor set was used to determine the relationship between a specific molecular characteristic and CL_{tot} because models developed by using this descriptor set have higher predictive capabilities than those developed by using other descriptor sets. In addition, it is relatively easier to assign the descriptors in the DS-MIXED descriptor set to specific molecular characteristics. Table 4 gives the list of the dominant descriptors and the corresponding molecular characteristic in different PCs.

Analysis of the variations of the descriptors shows that the first PC is primarily determined by topological descriptors. These include 3D-Wiener index, which decreases with increasing sphericity of a structure, and $^{2}\chi^{v}$, which is the valence molecular connectivity Chi index for path order 2 and

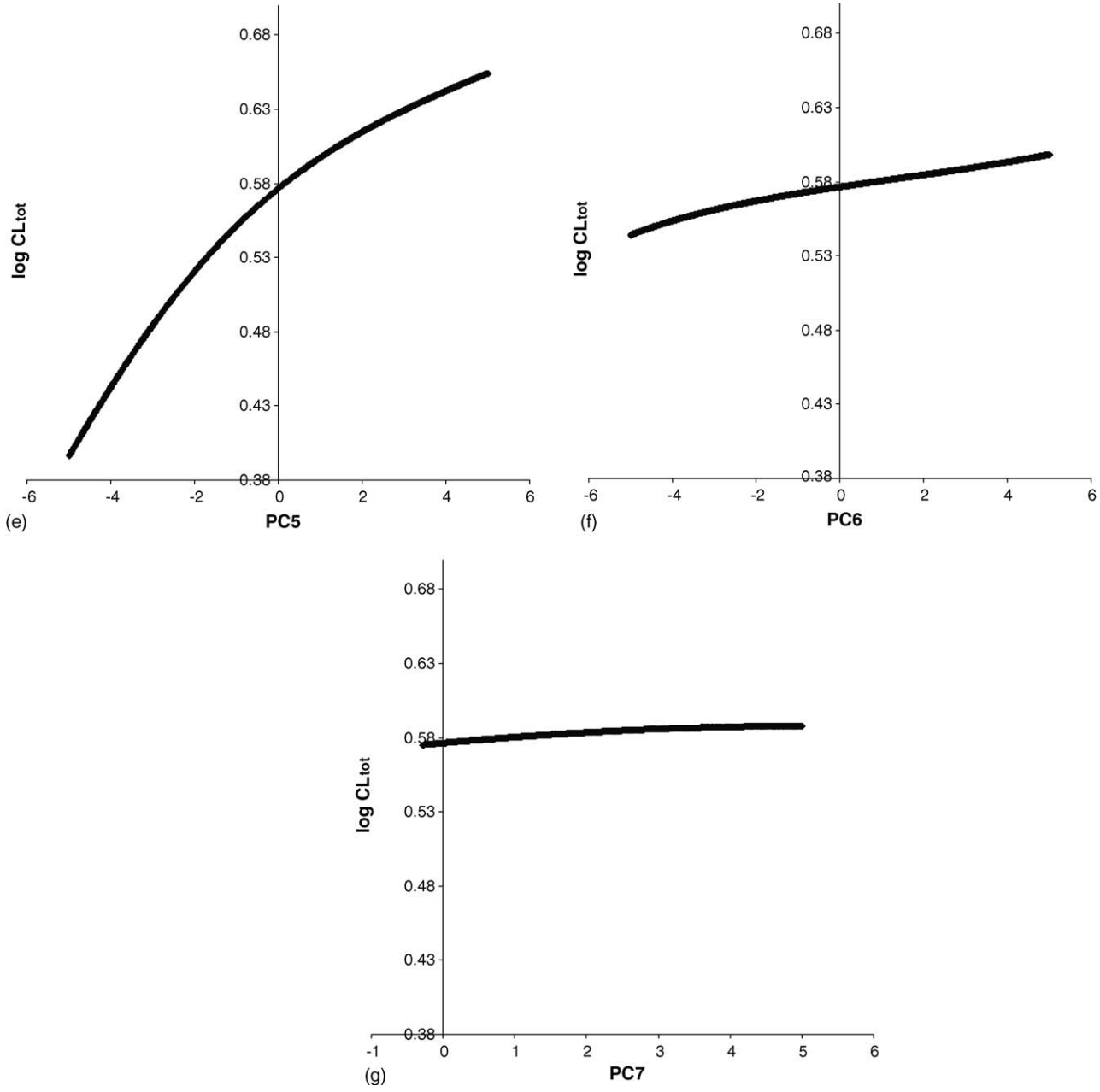

Fig. 4. (*Continued*).

encodes the relative degree of branching in a compound [60]. Prediction results by using the artificial testing set show that CL_{tot} generally decreases with increasing value of the 3D-Wiender index and $^{2}\chi^{v}$ (Fig. 4a). This suggests that spherically-shaped molecules with fewer side chain branching tend to have higher CL_{tot} than that of aspherical molecules with multiple branches.

Electrotopological state descriptors like Estate_aaCH and Estate_aasC, which describe electrotopological properties of carbons in aromatic rings, and AlogP [61], which measures the partition coefficient of a compound, are the main contributor to the second PC. Thus, it is likely that the second PC is a measure of the lipophilicity of a compound. Results of the artificial testing set suggest CL_{tot} increases with increasing lipophilicity of a compound.

The third PC is determined primarily by KierFlexibilityIndex, which is related to the flexibility of a molecule. The complex role of molecular flexibility in membrane permeation has been found by two studies. One found a positive correlation between flexibility and permeation [62] while the other found a negative correlation [63]. Using the artificial testing set, it was found that compounds with low or high flexibility have higher CL_{tot} than those with moderate flexibility. This may partially explain the apparent contradiction between the two earlier studies.

The fourth PC is formed mainly by AMW, which is the average molecular weight and the Gravitational3DIndex. These are related to the volume of a molecule and the distribution of atomic masses within the molecular space. The contribution of these two descriptors to the fourth PC suggests that the fourth

Table 4
The dominant descriptors and the corresponding molecular characteristic in different principal components

PC	Dominant descriptors	Corresponding molecular characteristic
First	3D-Wiener index [73] Valence molecular connectivity Chi index for path order 2 [60]	Molecular shape
Second	Atom-type Estate sum for:CH: (sp^2, aromatic) [35] Atom-type Estate sum for:C: [35] AlogP [61]	Lipophilicity
Third	Kier flexibility index [74]	Flexibility
Fourth	Average molecular weight Gravitational 3D index [43]	Molecular size
Fifth	Atom-type Estate sum for = S=< [35] Solvation molecular connectivity Chi index for path order 2 [75] Mean topological charge index for path order 1 [76]	Charge and molecular solvation
Sixth	Number of H-bond acceptors	Hydrogen bond accepting capability
Seventh	Number of H-bond donors	Hydrogen bond onating capability

PC is a measure of molecular size. The artificial testing sets show that CL_{tot} generally increases with decreasing molecular size. This is consistent with the findings that small molecular size is necessary for good membrane penetration [64].

The main contributors to the fifth PC are Estate_ddssS, which is the electrotopological descriptor for sulfur atoms, $^{2}\chi^{s}$, which is the solvation molecular connectivity Chi index for path order 2, and Mean$^{1}G^{c}$, which is the mean topological charge index for path order 1. It is difficult to attribute these descriptors to a single molecular characteristic. Nonetheless, studies have consistently shown that charge and molecular solvation are important in determining the metabolism [65,66] and renal clearance [11,67] of a molecule.

The sixth and seventh PCs are determined primarily by descriptors encoding the hydrogen bond acceptor and donor properties of a compound, respectively. Fig. 4f and g shows that CL_{tot} increases with increasing hydrogen bonding capability of a compound. Studies have found that binding affinity to human serum albumin generally decreases with increasing hydrogen bonding capability of these compounds [13,68]. Many compounds bind to serum albumin and the albumin-bound fraction is not available for hepatic metabolism or renal clearance [69]. Thus, factors which decrease serum albumin binding are expected to increase the CL_{tot} of a compound.

The rate of change in CL_{tot} per unit change in the PC values can provide a useful hint about the contribution of a molecular characteristic to the clearance of a compound. The plots in Fig. 4a–g show the contribution of each PC in the following order: PC5 > PC4 > PC3 > PC6 > PC1 ≈ PC2 > PC7. Thus, charge, molecular solvation, molecular size and flexibility are the most important molecular properties which influence clearance of a compound.

4. Conclusion

Our study suggests that both GRNN and SVR are potentially useful for developing QSPkR models to predict drug clearance from a large diverse set of compound data. QSPkR models developed by using GRNN, SVR and KNN were tested and compared with those developed by using a linear method, PLS. All of the GRNN- and SVR-developed models show better prediction capability than the corresponding KNN- or PLS-developed models. The predictive capabilities of the QSPkR models developed in this study are comparable to those of previous studies and can be further improved by using consensus modeling methods.

A collection of constitutional, geometrical, topological and electrotopological descriptors seems to be more useful for modeling drug clearance than specialized descriptor sets such as 3DMoRSE, ATS, GETAWAY, RDF and WHIM. An individual descriptor set tends to partially neglect some important features and thus the use of all the available descriptors may help to alleviate such type of feature bias. The three statistical learning methods used in this work appears to be capable of combining the information encoded in the different descriptor sets effectively to develop more predictive QSPkR models.

Acknowledgement

This work was supported in part by grants from Singapore ARF R-151-000-031-112, Shanghai Commission for Science and Technology (04DZ19850), and the '973' National Key Basic Research Program of China (2004CB720103).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2005.10.004.

References

[1] G.R. Wilkinson, Clearance approaches in pharmacology, Pharmacol. Rev. 39 (1981) 1–47.
[2] D.A. Smith, H. van de Waterbeemd, D.K. Walker, Pharmacokinetics and Metabolism in Drug Design, Wiley-VCH, Weinheim, Chichester, 2001.
[3] P.L. Toutain, A. Bousquet-Melou, Plasma clearance, J. Vet. Pharmacol. Ther. 27 (2004) 415–425.
[4] J. Zuegge, G. Schneider, P. Coassolo, T. Lave, Prediction of hepatic metabolic clearance: comparison and assessment of prediction models, Clin. Pharmacokinet. 40 (2001) 553–563.
[5] Y. Naritomi, S. Terashita, S. Kimura, A. Suzuki, A. Kagayama, Y. Sugiyama, Prediction of human hepatic clearance from in vivo animal experiments and in vitro metabolic studies with liver microsomes from animals and humans, Drug Metab. Dispos. 29 (2001) 1316–1324.
[6] T. Wajima, K. Fukumura, Y. Yano, T. Oguma, Prediction of human clearance from animal data and molecular structural parameters using multivariate regression analysis, J. Pharm. Sci. 91 (2003) 2489–2499.
[7] T. Wajima, K. Fukumura, Y. Yano, T. Oguma, Prediction of human pharmacokinetics from animal data and molecular structural parameters

using multivariate regression analysis: oral clearance, J. Pharm. Sci. 92 (2003) 2427–2440.
[8] V. Karalis, A. Tsantili-Kakoulidou, P. Macheras, Multivariate statistics of disposition pharmacokinetic parameters for structurally unrelated drugs used in therapeutics, Pharm. Res. 19 (2002) 1827–1834.
[9] V. Karalis, A. Tsantili-Kakoulidou, P. Macheras, Quantitative structure–pharmacokinetic relationships for disposition parameters of cephalosporins, Eur. J. Pharm. Sci. 20 (2003) 115–123.
[10] J.V. Turner, D.J. Maddalena, D.J. Cutler, S. Agatonovic-Kustrin, Multiple pharmacokinetic parameter prediction for a series of cephalosporins, J. Pharm. Sci. 92 (2003) 552–559.
[11] J.V. Turner, D.J. Maddalena, D.J. Cutler, Pharmacokinetic parameter prediction from drug structure using artificial neural networks, Int. J. Pharm. 270 (2004) 209–219.
[12] C. Ng, Y.D. Xiao, W. Putnam, B. Lum, A. Tropsha, Quantitative structure–pharmacokinetic parameters relationships (QSPKR) analysis of antimicrobial agents in humans using simulated annealing k-nearest-neighbor and partial least-square analysis methods, J. Pharm. Sci. 93 (2004) 2535–2544.
[13] C.W. Yap, Y.Z. Chen, Quantitative structure–pharmacokinetic relationships for drug distribution properties by using general regression neural network, J. Pharm. Sci. 94 (2005) 153–168.
[14] U. Norinder, Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimisations and variable selection, Neurocomputing 55 (2003) 337–346.
[15] T. Niwa, Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures, J. Chem. Inf. Comput. Sci. 43 (2003) 113–119.
[16] M. Shen, Y.D. Xiao, A. Golbraikh, V.K. Gombar, A. Tropsha, Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates, J. Med. Chem. 46 (2003) 3013–3020.
[17] P.D. Mosier, P.C. Jurs, L.L. Custer, S.K. Durham, G.M. Pearl, Predicting the genotoxicity of thiophene derivatives from molecular structure, Chem. Res. Toxicol. 16 (2003) 721–732.
[18] A.H. Asikainen, J. Ruuskanen, K.A. Tuppurainen, Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds, SAR QSAR Environ. Res. 15 (2004) 19–32.
[19] MICROMEDEX *MICROMEDEX*, Edition expires 12/2003, MICROMEDEX, Greenwood Village, Colorado.
[20] J.G. Hardman, L.E. Limbird, A. Goodman Gilman, Goodman and Gilman's the Pharmacological Basis of Therapeutics, McGraw-Hill, New York, 2002.
[21] K. Ito, T. Iwatsubo, S. Kanamitsu, Y. Nakajima, Y. Sugiyama, Quantitative prediction of in vivo drug clearance and drug interactions from in vitro data on metabolism, together with binding and transport, Annu. Rev. Pharmacol. Toxicol. 38 (1998) 461–499.
[22] R.S. Obach, Prediction of human clearance of twenty-nine drugs from hepatic microsomal intrinsic clearance data: an examination of in vitro half-life approach and nonspecific binding to microsomes, Drug Metab. Dispos. 27 (1999) 1350–1359.
[23] J. Neter, M.H. Kutner, C.J. Nachtsheim, W. Wasserman, Diagnostics and remedial measures, in: J. Neter, M.H. Kutner, C.J. Nachtsheim, W. Wasserman (Eds.), Applied Linear Statistical Models, Irwin, Chicago, 1996, pp. 95–151.
[24] C.W. Yap, Y.Z. Chen, Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines, J. Chem. Inf. Model. 45 (2005) 982–992.
[25] H. Li, Y. Xue, C.Y. Ung, C.W. Yap, Z.R. Li, Y.Z. Chen, Prediction of genotoxicity of chemical compounds by statistical learning methods, Chem. Res. Toxicol. 18 (2005) 1071–1080.
[26] Y. Xue, C.W. Yap, L.Z. Sun, Z.W. Cao, J.F. Wang, Y.Z. Chen, Prediction of *p*-glycoprotein substrates by support vector machine approach, J. Chem. Inf. Comput. Sci. 44 (2004) 1497–1505.
[27] T.W. Schultz, T.I. Netzeva, M.T.D. Cronin, Selection of data sets for QSARs: analyses of Tetrahymena toxicity from aromatic compounds, SAR QSAR Environ. Res. 14 (2003) 59–81.
[28] K. Rajer-Kanduc, J.M.N. Zupan, Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment, Chemom. Intell. Lab. Sys. 65 (2003) 221–229.
[29] J.J. Perez, Managing molecular diversity, Chem. Soc. Rev. 34 (2005) 143–152.
[30] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.
[31] L. Molnar, G.M. Keseru, A neural network based virtual screening of cytochrome P450 3A4 inhibitors, Bioorg. Med. Chem. Lett. 12 (2002) 419–421.
[32] T. Potter, H. Matter, Random or rational design? Evaluation of diverse compound subsets from chemical structure databases, J. Med. Chem. 41 (1998) 478–488.
[33] B. Hollas, An analysis of the autocorrelation descriptor for molecules, J. Math. Chem. 33 (2003) 91–101.
[34] R.S. Pearlman, CONCORD User's Manual, Tripos Inc., St. Louis, MO.
[35] L.B. Kier, L.H. Hall, Molecular Structure Description: The Electrotopological State, Academic Press, San Diego, 1999.
[36] J.H. Schuur, P. Setzer, J. Gasteiger, The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure–spectra correlations and studies of biological activity, J. Chem. Inf. Comput. Sci. 36 (1996) 334–344.
[37] G. Moreau, P. Broto, The autocorrelation of a topological structure: a new molecular descriptor, Nouveau J. de Chimie 4 (1980) 359–360.
[38] V. Consonni, R. Todeschini, M. Pavan, Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors, J. Chem. Inf. Comput. Sci. 42 (2002) 682–692.
[39] M.C. Hemmer, V. Steinhauer, J. Gasteiger, Deriving the 3D structure of organic molecules from their infrared spectra, Vibrat. Spectrosc. 19 (1999) 151–164.
[40] G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, A. Zaliani, MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids, J. Comput. Aided Mol. Des. 11 (1997) 79–92.
[41] I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D.J. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, V.V. Prokopenko, Virtual computational chemistry laboratory—design and description, J. Comput. Aided Mol. Des. 19 (2005) 453–463.
[42] P.D. Mosier, P.C. Jurs, QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks, J. Chem. Inf. Comput. Sci. 42 (2002) 1460–1470.
[43] M.D. Wessel, P.C. Jurs, J.W. Tolan, S.M. Muskal, Prediction of human intestinal absorption of drug compounds from molecular structure, J. Chem. Inf. Comput. Sci. 38 (1998) 726–735.
[44] D.J. Livingstone, Data pre-treatment<, in: D.J. Livingstone (Ed.), Data Analysis for Chemists: Applications to QSAR and Chemical Product Design, Oxford University Press, Oxford, 1995, pp. 48–64.
[45] D.F. Specht, A general regression neural network, IEEE Trans. Neural Netw. 2 (1991) 568–576.
[46] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (1962) 1065–1076.
[47] A.J. Smola B. Scholkopf, A tutorial on support vector regression, NeuroCOLT2 Technical Report Series.
[48] Z. Yuan, B.X. Huang, Prediction of protein accessible surface areas by support vector regression, Proteins 57 (2004) 558–564.
[49] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[50] E. Fix, J.L. Hodges, Discriminatory Analysis: Non-parametric Discrimination: Consistency Properties, USAF School of Aviation Medicine, Randolph Field, Texas, 1951, pp. 261–279.
[51] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, Englewood Cliffs, NJ, 1982.
[52] J.J. Sutherland, D.F. Weaver, Development of quantitative structure-activity relationships and classification models for anticonvulsant activity of hydantoin analogues, J. Chem. Inf. Comput. Sci. 43 (2003) 1028–1036.

[53] A. Golbraikh, A. Tropsha, Beware of q2! J. Mol. Graph. Mod. 20 (2002) 269–276.
[54] A. Kozak, R. Kozak, Does cross validation provide additional information in the evaluation of regression models? Can. J. For. Res. 33 (2003) 976–987.
[55] J. Reunanen, Overfitting in making comparisons between variable selection methods, J. Machine Learn. Res. 3 (2003) 1371–1382.
[56] I.-M. Olsson, J. Gottfries, S. Wold, D-optimal onion designs in statistical molecular design, Chemom. Intell. Lab. Sys. 73 (2004) 37–46.
[57] D.M. Hawkins, S.C. Basak, D. Mills, Assessing model fit by cross-validation, J. Chem. Inf. Comput. Sci. 43 (2004) 579–586.
[58] R.S. Obach, J.G. Baxter, T.E. Liston, B.M. Silber, B.C. Jones, F. Macintyre, D.J. Rance, P. Wastall, The prediction of human pharmacokinetic parameters from preclinical and in vitro metabolism data, J. Pharmacol. Exp. Ther. 283 (1997) 46–58.
[59] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Sys. 2 (1987) 37–52.
[60] L.B. Kier, L.H. Hall, Molecular connectivity in structure-activity analysis, Research Studies Press, Wiley, Letchworth, Hertfordshire, England, New York, 1986.
[61] V.N. Viswanadhan, M.R. Reddy, R.J. Bacquet, M.D. Erion, Assessment of methods used for predicting lipophilicity: application to nucleosides and nucleoside bases, J. Comput. Chem. 14 (1993) 1019–1026.
[62] M. Iyer, R. Mishru, Y. Han, A.J. Hopfinger, Predicting blood–brain barrier partitioning of organic molecules using membrane-interaction QSAR analysis, Pharm. Res. 19 (2002) 1611–1621.
[63] D.F. Veber, S.R. Johnson, H.Y. Cheng, B.R. Smith, K.W. Ward, K. Kopple, Molecular properties that influence the oral bioavailability of drug candidates, J. Med. Chem. 45 (2002) 2615–2623.
[64] W.M. Pardridge, CNS drug design based on principles of blood–brain barrier transport, J. Neurochem. 70 (1998) 1781–1792.
[65] D.A. Smith, M.J. Acklanda, B.C. Jones, Properties of cytochrome P450 isoenzymes and their substrates. Part 2. Properties of cytochrome P450 substrates, Drug Discov. Today 2 (1997) 479–486.
[66] M.J. de Groot, S. Ekins, Pharmacophore modeling of cytochromes P450, Adv. Drug Deliv. Rev. 54 (2002) 367–383.
[67] D. Venturoli, B. Rippe, Ficoll and dextran vs. globular proteins as probes for testing glomerular permselectivity: effects of molecular size, shape, charge, and deformability, Am. J. Physiol. Renal Physiol. 288 (2005) F605–F613.
[68] L.M. Hall, L.H. Hall, L.B. Kier, QSAR modeling of beta-lactam binding to human serum proteins, J. Comput. Aided Mol. Des. 17 (2003) 103–118.
[69] G. Colmenarejo, In silico prediction of drug-binding strengths to human serum albumin, Med. Res. Rev. 23 (2003) 275–301.
[70] T.I. Oprea, J. Gottfries, Chemography: the art of navigating in chemical space, J. Comb. Chem. 3 (2001) 157–166.
[71] NCI/NIH Developmental therapeutics program.
[72] J.J. Sutherland, L.A. O'Brien, D.F. Weaver, Spline-fitting with a genetic algorithm: a method for developing classification structure-activity relationships, J. Chem. Inf. Comput. Sci. 43 (2003) 1906–1915.
[73] S.C. Basak, B.D. Gute, S. Ghatak, Prediction of complement-inhibitory activity of benzamidines using topological and geometric parameters, J. Chem. Inf. Comput. Sci. 39 (1999) 255–260.
[74] L.B. Kier, Indexes of molecular shape from chemical graphs, in: L.B. Kier (Ed.), Computational Chemical Graph Theory, Nova Science Publishers, New York, 1990, pp. 151–174.
[75] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
[76] J. Galvez, R. Garcia, M.T. Salabert, R. Soler, Charge indexes. New topological descriptors, J. Chem. Inf. Comput. Sci. 34 (1994) 520–525.