Quantitative and qualitative prediction of corneal permeability for drug-like compounds

Mehdi Ghorbanzad'e [a,b,*], Mohammad H. Fatemi [b], Masoumeh Karimpour [a,b], Patrik L. Andersson [a]

[a] Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden
[b] Chemometrics Laboratory, Faculty of Chemistry, University of Mazandaran, Babolsar, Iran

ABSTRACT

A set of 69 drug-like compounds with corneal permeability was studied using quantitative and qualitative modeling techniques. Multiple linear regression (MLR) and multilayer perceptron neural network (MLP-NN) were used to develop quantitative relationships between the corneal permeability and seven molecular descriptors selected by stepwise MLR and sensitivity analysis methods. In order to evaluate the models, a leave many out cross-validation test was performed, which produced the statistic $Q^2 = 0.584$ and SPRESS = 0.378 for MLR and $Q^2 = 0.774$ and SPRESS = 0.087 for MLP-NN. The obtained results revealed the suitability of MLP-NN for the prediction of corneal permeability. The contribution of each descriptor to MLP-NN model was evaluated. It indicated the importance of the molecular volume and weight. The pattern recognition methods principal component analysis (PCA) and hierarchical clustering analysis (HCA) have been employed in order to investigate the possible qualitative relationships between the molecular descriptors and the corneal permeability. The PCA and HCA results showed that, the data set contains two groups. Then, the same descriptors used in quantitative modeling were considered as inputs of counter propagation neural network (CPNN) to classify the compounds into low permeable (LP) and very low permeable (VLP) categories in supervised manner. The overall classification non error rate was 95.7% and 95.4% for the training and prediction test sets, respectively. The results revealed the ability of CPNN to correctly recognize the compounds belonging to the categories. The proposed models can be successfully used to predict the corneal permeability values and to classify the compounds into LP and VLP ones.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

The cornea acts as a protective barrier to invasion of foreign substances and also as a barrier to drug transport [1,2]. It contains three primary layers: endothelium, stroma and epithelium [3]. The corneal endothelium is a monolayer of hexagonal cells found at the internal base of the cornea. It is considered to have minor significance in drug permeation through the cornea. The epithelium and stroma are more significant for drug delivery. The stroma is a fibrous tissue that forms the bulk of the cornea and is made up of primarily large collagen fibers embedded in a proteoglycan matrix. The epithelium is a multilayer of cells found at the external surface of the cornea. The major barrier for lipophilic drugs is aqueous stroma, while for hydrophilic drugs the lipophilic corneal epithelium is the rate-limiting barrier [4–6]. The major route for drug absorption from eye drops is permeation through the cornea [7]. Experimental determination of corneal permeability of drugs is costly, laborious, time consuming and needs sufficient quantity of the pure compounds and fresh tissue samples [5,8–11]. Thus, the ability to predict the corneal permeability of drug candidates from their physicochemical properties using some theoretical methods such as quantitative structure property relationship (QSPR) would be a powerful tool in the development of new drugs. With such a tool, the number of animal experiments would be reduced and the permeability of newly synthesized drugs can be assessed without stepping into the laboratory.

QSPR analysis is a well-established technique to correlate physicochemical properties of a compound with its molecular structure, through a variety of descriptors. QSPR studies have been applied by researchers to predict many physicochemical properties of molecules [12–15]. In QSPR studies many statistical techniques such as multiple linear regression (MLR), partial least squares regression (PLS), support vector machine (SVM), and various types of artificial neural networks (ANNs) can be used

* Corresponding author at: Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden. Tel.: +98 112 5342350/+46 90 7869241; fax: +98 112 5342350.
E-mail addresses: mehdi.mojaveri@chem.umu.se, m.ghorbanzade@umz.ac.ir, ghorbanzade@gmail.com (M. Ghorbanzad'e), mhfatemi@umz.ac.ir (M.H. Fatemi), masoumeh.karimpour@chem.umu.se (M. Karimpour), patrik.andersson@chem.umu.se (P.L. Andersson).

to derive correlation models between descriptors of molecular structures and properties [16–18]. There have been reported some inexpensive and less time-consuming *in silico* QSPR models for prediction of corneal permeability in the literature. For example, Agatonovic-Kustrin et al. [6] used ANN mapping technique along with calculated molecular descriptors to build a model for predicting corneal permeability of 45 structurally different drugs. The best model obtained included four input descriptors and 12 hidden layers. They found that polar surface area, global shape index unweighted, global shape index weighted by atomic Sanderson electro negativities and oxygen atom of the carbonyl group give the best predictive performance. Yoshida and Topliss [19] used distribution coefficient (log D) and the difference between the octanol–water partition coefficient and the alkane–water partition coefficient ($\Delta \log P$) to develop a model for predicting the corneal permeability of various compounds. Some other models have been developed to predict corneal permeability as a function of the partition coefficient or the distribution coefficient of the drug [9,20]. In a recent work, Kidron et al. [21] have used PLS regression technique to build corneal permeability models based on a structurally diverse set of 69 drug-like compounds. They selected 10 descriptors affecting membrane permeability. They concluded that total number of hydrogen bonds and the logarithm of the octanol–water partition coefficient at pH 7.4 and 8.0 ($\log D_{7.4}$ and $\log D_{8.0}$) are the most important descriptors for predicting corneal permeability. Using these descriptors, they constructed models with Q^2 and R^2 values ranging from 0.77 to 0.79. In the present study, we developed new QSPR models and compared these with the PLS models obtained by Kidron et al. We hypothesized that nonlinear relationship exists for the structure–corneal permeability relationship and that multilayer perceptron neural network (MLP-NN) model as a nonlinear mapping technique can reveal that phenomena. Furthermore, we aimed at classifying the 69 assessed drug-like compounds into low permeable (LP) and very low permeable (VLP) categories based on their permeability through the cornea using counter propagation neural network (CPNN).

2. Materials and methods

2.1. Experimental data

The *in vitro* experimental values of corneal permeability for 69 drug-like compounds were taken from the recent study by Kidro et al. [21]. The reported permeability of the compounds has been measured in similar condition in the rabbit cornea. The values of permeability range from 1.7×10^{-7} to 7.9×10^{-5} cm/s. The permeability values were transformed into logarithmic unit and the relationship between logarithm of corneal permeability (log permeability) and molecular descriptors was examined. The compounds in the data set were randomly divided into the training, internal, and external test sets, consisting of 47, 12, and 10 members, respectively. The training set and internal test set participated in the generation of the MLP-NN model and adjusting its parameters. The external test set was used to validate the predictive performance of the model. In the case of MLR and CPNN analysis both internal and external test sets were considered as validation set. For classification analysis, the training set included 23 drug-like compounds of LP category (the compounds permeating faster than 9×10^{-6} cm/s) and 24 drug-like compounds of VLP category (the compounds permeating slower than 9×10^{-6} cm/s) and validation set contained 22 compounds (9 LPs and 13 VLPs). The experimental values of corneal permeability for all molecules studied in this work are listed in Table 1.

2.2. Molecular descriptors calculation, screening and selection

Molecular descriptors are the simple mathematical representation of a molecule and are used to encode significant structural features of molecules. In order to calculate descriptors, the Hyperchem program (ver. 7) [22] was applied to construct all molecular structures. Molecular geometry was optimized with the Austin Model 1 (AM1) semi empirical method. A more precise optimization of molecular structures was then done with MOPAC (version 6.0) [23]. Then Dragon [24] and CODESSA [25] were used for calculation of a wide variety of molecular descriptors. Some of these descriptors encoded similar molecular information, therefore, it was necessary to test descriptors and eliminate those with low variation and those which encoded similar information (descriptors with the absolute value of Pearson correlation coefficient above 0.8). In a next step the most significant descriptors were selected from the pool of molecular descriptors in order to reach predictive models. Stepwise MLR, which combines the forward and backward procedures, was used for selection of the descriptors [26]. Due to the complexity of intercorrelations, the variance explained by certain variables will change when new variables enter the equation. Sometimes a variable, that is qualified to enter the model, loses some of its predictive validity when other variables enter. If this happens, the stepwise method will remove the weakened variable. A final set of models was then tested for stability and validity through a variety of statistical methods. The best MLR model is the one with high multiple correlation coefficient (R), high F-statistic, low standard error (SE), least number of descriptors and high ability for prediction. Based on these criteria eight descriptors were selected using stepwise MLR technique. Out of eight descriptors, four are Dragon based descriptors, i.e. hydrophilic factor, A total size of index/unweighted, Borto–Moreau autocorrelation of lag 1 weighted by Van der Waals volume and property potential (mass weighted). The other four descriptors are maximum partial charge on an H atom ($\delta_H(\max)$), count of H acceptor sites, molecular volume and average information content order 1 which were calculated using CODESSA. The eight selected descriptors were used as inputs for MLP-NN modeling. Then, ten descriptors which had been used in previous work [21] were added to improve the performance of MLP-NN model and to investigate the nonlinear relationship between these descriptors and corneal permeability. The descriptors selected by Kidron et al. were; molecular weight (MW), molecular volume (MV), polar surface area (PSA), number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), total number of putative hydrogen bonds, i.e. HBD + HBA (HBtot), the logarithm of the octanol-water partition coefficient of the neutral form ($\log P$) and at pH 7.0, 7.4 and 8.0 ($\log D_{7.0}$, $\log D_{7.4}$ and $\log D_{8.0}$, respectively). These molecular descriptors had been calculated with the ACDLabs software package version 6.0 [27]. In the next step, the correlations of descriptors with each other were examined once more, and eleven descriptors (from eighteen descriptors) were selected by stepwise MLR and used as new inputs for MLP-NN. Finally, sensitivity analysis was used to discard the insignificant portions of data and to find the most important descriptors that best correlate with the investigated property. The inputs that produced low sensitivity values were considered insignificant and were removed from the network. Following a sensitivity analysis the number of inputs was reduced to seven descriptors. Table 2 represents the correlation matrix among these seven descriptors.

2.3. Diversity analysis

In order to make sure the structures of the training and test sets represent those of the whole data set, diversity analysis [28,29] was done on the data set. In this way, the mean distance of one

Table 1
Calculated and observed log permeability values for compounds used in this study.

No.	Structure name	Permeability (cm/s)	Log permeability	Pred. MLR	Pred. MLP-NN
1[et]	Acetazolamide	9.10E−07	−6.04	−5.83	−6.19
2	Alprenolol	2.90E−05	−4.54	−4.13	−4.46
3	Atenolol	6.70E−07	−6.17	−5.79	−6.22
4	Benzolamide	1.40E−07	−6.85	−6.33	−6.80
5	Bevantolol	5.40E−05	−4.27	−4.17	−4.27
6	Bromoacetazolamide	3.80E−07	−6.42	−5.95	−6.25
7[et]	Bufuralol	5.70E−05	−4.24	−3.98	−4.35
8[it]	Chlorzolamide	1.80E−05	−4.74	−4.86	−4.71
9	Clonidine	3.10E−05	−4.51	−4.54	−4.37
10[it]	Corynanthine	1.10E−05	−4.96	−4.81	−4.75
11	Cyclophosphamide	1.10E−05	−4.96	−4.80	−4.94
12	Ethoxazolamide	5.60E−05	−4.25	−4.52	−4.38
13	2-Benzothiazole-sulfonamide	3.60E−05	−4.44	−4.70	−4.40
14[it]	6-Hydroxy-2-benzo-thiazolesulfonamide	5.60E−06	−5.25	−5.26	−5.39
15	6-Chloro-2-benzo-thiazolesulfonamide	4.30E−05	−4.37	−4.53	−4.35
16	4.6-Dichloro-2-benzo-thiazolesulfonamide	3.90E−05	−4.41	−4.46	−4.38
17	6-Amino-2-benzo-thiazolesulfonamide	6.70E−06	−5.17	−5.47	−5.13
18	6-Nitro-2-benzo-thiazolesulfonamide	6.60E−06	−5.18	−5.11	−5.09
19	6-Hydroxyethoxy-2-benzothiazolesulfonamide	1.50E−06	−5.82	−5.54	−5.78
20	6-Benzyloxy-2-benzo-thiazolesulfonamide	4.70E−05	−4.33	−4.22	−4.30
21[it]	6-Acetamido-2-benzo-thiazolesulfonamide	4.70E−06	−5.33	−5.32	−5.48
22[it]	Levobunolol	2.00E−05	−4.70	−4.34	−4.41
23[et,*]	Labetalol	1.40E−05	−4.85	−5.25	−5.20
24	Methazolamide	2.60E−06	−5.59	−5.58	−5.75
25[et]	5-Imino-4-methyl-1.3.4-thiadiazoline-2-sulfonamide	7.80E−07	−6.11	−6.56	−6.35
26[it]	Metoprolol	2.50E−05	−4.60	−4.65	−4.61
27	Nadolol	1.00E−06	−6.00	−5.45	−5.95
28	Oxprenolol	2.90E−05	−4.54	−4.47	−4.50
29[et]	Penbutolol	4.50E−05	−4.35	−3.68	−4.29
30	Phenylephrine	9.40E−07	−6.03	−5.50	−6.03
31[et]	Pindolol	1.00E−05	−5.00	−4.88	−5.26
32	Propranolol	4.80E−05	−4.32	−4.14	−4.46
33	Rauwolfine	9.20E−06	−5.04	−4.79	−5.01
34[it]	SKF86607	7.90E−05	−4.10	−4.26	−4.23
35[it]	SKF86466	7.10E−05	−4.15	−3.15	−4.19
36	SKF72223	4.90E−05	−4.31	−4.33	−4.38
37	Sotalol	1.60E−06	−5.80	−5.66	−5.82
38	1[(N-methylamino)sulfonyl]-4-chloro-benzene	6.50E−05	−4.19	−4.07	−4.22
39	4-Chlorobenzene sulfonamide	5.50E−05	−4.26	−4.69	−4.32
40[it]	6-Sulfonamido-3-substituted-3H-1.3.4.-thiadiozolo [2.3-C]-1.2.4-thiadiazole	7.90E−06	−5.10	−5.23	−5.26
41[*]	3-Chloro der. of 40	1.30E−05	−4.89	−5.00	−5.01
42	4-Chloro der. of 40	8.30E−06	−5.08	−4.95	−4.96
43	3-Methoxy der. of 40	4.50E−06	−5.35	−5.30	−5.39
44[it]	4-Methoxy der. of 40	5.20E−06	−5.28	−5.20	−5.18
45[et]	4-Hydroxy der. of 40	3.50E−07	−6.46	−5.92	−6.76
46	3-Fluoro der. of 40	6.40E−06	−5.19	−5.11	−5.14
47	4-Fluoro der. of 40	4.10E−06	−5.39	−5.20	−5.24
48	4-Dimethylamino der. of 40	5.80E−06	−5.24	−5.38	−5.24
49	Timolol	1.20E−05	−4.92	−5.72	−4.97
50	Trichlormethazolamide	1.10E−05	−4.96	−5.20	−5.29
51	Trifluormethazolamide	3.90E−06	−5.41	−5.20	−5.33
52	Yohimbine	1.80E−05	−4.74	−4.81	−4.75
53	Cinoxacin	1.50E−06	−5.82	−6.40	−5.87
54	Nalidixic acid	1.70E−05	−4.77	−5.12	−4.71
55	Enoxacin	1.70E−06	−5.77	−5.57	−5.82
56[it]	Norfloxacin	1.40E−06	−5.85	−5.46	−5.77
57	Ofloxacin	1.90E−06	−5.72	−5.53	−5.63
58	Ciprofloxacin	1.30E−06	−5.89	−5.64	−5.93
59	Moxifloxacin	9.10E−06	−5.04	−5.63	−5.16
60	Levofloxacin	2.90E−06	−5.54	−5.53	−5.63
61[*]	Gatifloxacin	2.50E−06	−5.60	−5.65	−5.40
62[et]	Lomefloxacin	3.50E−06	−5.46	−5.14	−5.51
63[et]	Buspirone	6.65E−05	−4.18	−3.62	−4.13
64	Apraclonidine	3.80E−05	−5.42	−5.78	−5.43
65[et]	Fluorescein	1.60E−05	−4.80	−4.51	−4.81
66	Pilocarpine	9.80E−06	−5.01	−4.56	−4.99
67	Nepafenac	7.40E−05	−4.13	−4.87	−4.26
68	Betaxolol	6.00E−05	−4.22	−4.37	−4.30
69[it]	Dexamethasone	9.20E−06	−5.04	−4.95	−5.25

[it] and [et] refer to compounds used as internal test and external test set, respectively and [*]refers to the misclassified compounds.

Table 2
Correlation matrix of descriptors used in MLP-NN and CPNN modeling.

	MW	MV	δ_H (max)	HBD	HBA	Log P	Log $D_{7.4}$
MW	1						
MV	0.616	1					
δ_H (max)	0.400	0.482	1				
HBD	−0.004	−0.033	0.253	1			
HBA	0.624	−0.054	0.212	0.074	1		
Log P	0.450	0.451	0.001	−0.251	−0.105	1	
Log $D_{7.4}$	0.174	0.154	−0.349	−0.210	−0.255	0.764	1

chemical to the remaining ones ($\bar{d}_\iota$) was computed from descriptor space matrix as follows:

$$\bar{d}_\iota = \frac{\sum_{j=1}^n d_{ij}}{n-1} \quad i = 1, 2, \ldots, n \tag{1}$$

where d_{ij} is a distance score for two different compounds, which can be measured by the Euclidean distance norm based on the compounds descriptors (x_{ik} and x_{jk}):

$$d_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2} \tag{2}$$

Then the mean distances were normalized within the interval [0,1]. The closer to 1 the distance is, the more diverse each compound is from the others. For the used data set, the mean distances of samples versus log permeability values are plotted in Fig. 1, which illustrates the diversity of the molecules in the training and test sets. As can be seen from this figure, the structures of compounds are diverse in all subsets. It warrants model stability and that the external test set is suitable to assess the predictive performance of the developed model.

2.4. Neural networks

Artificial neural networks are capable of mapping complex patterns of non-linearity. The flexibility of ANNs enables them to discover non-linearity between selected descriptors and corneal permeability. In this work two supervised neural networks (MLP-NN and CPNN) were used to investigate non-linear relationship between selected molecular descriptors and corneal permeability of drugs. Generally, each neural network is built from several layers: one input layer, one or more hidden layers, and one output layer. The node in each layer is connected to the nodes of the next layer by weights. During training, these weights and biases

are iteratively adjusted to minimize the network errors. A detailed description of the theory behind neural networks is described elsewhere [30,31]. In MLP-NN, the inputs are fully connected to the hidden layer and hidden layer neurons are fully connected to the outputs. In the learning phase, the sequenced input patterns, presented to the network, are propagated in the forward direction layer by layer until the final layer output is computed. The error is computed as the difference between the calculated output and the target value. These calculated errors are taken as input pattern for feedback connections from which the synaptic weights are adjusted in backward direction layer by layer [32]. In order to obtain an optimal MLP-NN model, different networks with three logistic, exponential and tangent hyperbolic functions in each hidden and output layer were trained using Quasi–Newton algorithm. Comparison among the results based on root mean square error (RMSE) of training, internal and external test sets revealed that using exponential and logistic transfer function, respectively in hidden and output layer leads to the lowest RMSE in each set. In addition to the type of transfer function, the number of neurons in hidden layer can affect the performance of model. In order to find the optimal number of neurons in hidden layer, different networks with 4–10 neurons in hidden layer were trained. According to RMSE values for training, internal and external test sets, the optimal number of neurons in hidden layer was set at six. The trained network was used as an analytical tool to predict log permeability of the compounds. Detailed description of this type of ANN modeling has been published [33–35].

Counter propagation neural network was used to classify drug-like compounds into LP and VLP categories. Counter propagation neural network [36,37] is briefly explained here. It is built up from two layers of neurons arranged in two-dimensional rectangular matrices and the combination of these motifs allows for the supervised learning of CPNN. Both layers of neurons are placed exactly one above the other. Thus, the Kohonen and output neurons are in one-to-one correspondence. The input or Kohonen layer contains information on input values (descriptors) while the output layer is associated with the output values (log permeability). Network training is performed during learning epochs. One epoch consists of presenting every object exactly once to the network. In the first phase of each training step, the winning neuron (the neuron with the closest weights to the input values) in the Kohonen layer is selected, and the weights of each neuron in the Kohonen layer are updated on the basis of the difference between their old value and the values of the input vector. Then this correction is scaled according to the topological distance from the winner. On the other hand, the weights of the output layer are updated in a supervised manner, but considering the response component of the target, instead of the input vector. Additionally, the learning rate constant is also changing during the training phase up to a minimum value, which is reached when the number of training epochs equals a predefined maximum value. CPNN modeling technique has been successfully used for resolving several classification and regression problems [38–43]. In order to build a CPNN classification model with predictive ability, some parameters should be optimized. The optimization of CPNN parameters was made on the
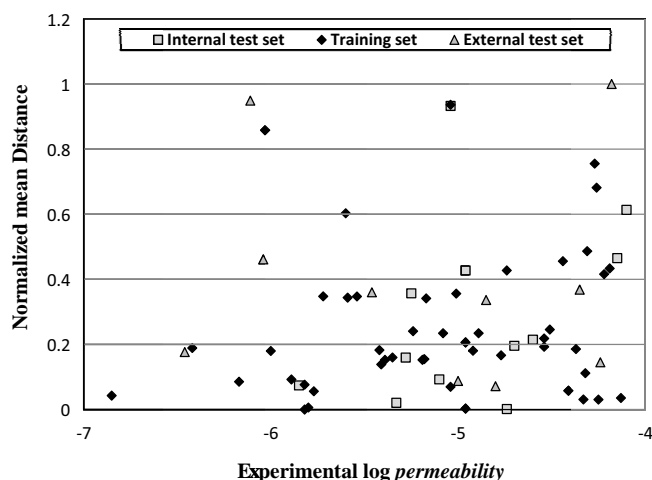
Fig. 1. Scatter plot of normalized mean distance of samples versus experimental log permeability.

basis of minimum error rate on the training as well as prediction set. The dimension of the Kohonen layer and number of epochs were varied from 4×4 to 10×10 and 50 to 250, respectively. The optimal model parameters were found to be an 8×8 Kohonen grid, trained for 100 epochs with a hexagonally toroidal neighborhood function.

3. Results and discussion

In this work, stepwise MLR technique was used to select eight descriptors from descriptors calculated by the Dragon and CODESSA softwares. Then, ten descriptors computed by ACDLabs software were added to eight previously selected descriptors and stepwise MLR was again used for selecting the most relevant descriptors. In the next step, sensitivity analysis was performed to refine the model and the number of descriptors reduced to seven descriptors. Finally, MLR and MLP-NN were used to build linear and nonlinear QSPR models, respectively, to predict the corneal permeability of the drug-like compounds and CPNN was applied to discriminate between LP and VLP compounds in the data set.

3.1. Quantitative models for corneal permeability prediction

3.1.1. MLR modeling

The main goal of MLR modeling was to select a set of suitable descriptors that can be used as inputs for building models. Due to the co-linearity problem in the MLR models, prior to models generation, the number of significant descriptors was reduced according to the criteria mentioned in Section 2.2. The obtained MLR model using descriptors, selected by combining stepwise MLR and sensitivity analysis, has the following specifications:

$$\log \text{perm} = -0.005(\pm 0.003)\text{MW} + 0.005(\pm 0.002)\text{MW}$$
$$- 0.166(\pm 0.061)\text{HBD} - 0.081(\pm 0.075)\text{HBA}$$
$$+ 0.164(\pm 0.092)\log P + 0.209(\pm 0.075)\log D_{7.4}$$
$$- 5.994(\pm 4.483)\delta_H(\text{max}) - 3.740(\pm 0.364) \quad n = 47,$$
$$R = 0.879, \text{SE} = 0.349, F = 18.870 \tag{3}$$

The regression coefficients obtained from training of MLR model were used to calculate the corneal permeability for validation set. The calculated values of log permeability for training and validation sets are shown in Table 1. Also, the leave many-out cross-validation test was performed to validate the MLR model. The statistical parameters of this calculation are shown in Table 3.

3.1.2. Neural network modeling

A three-layer network with an exponential transfer function in the hidden layer and a logistic transfer function in the output layer was designed and the number of neurons in the hidden layer was optimized to be six. Then, the optimized 7-6-1 network was used to predict the log permeability for external test set as well as training and internal test sets. In neural network modeling, due to non-linearity, it is difficult to interpret the effects of individual descriptors. However, nonlinear behavior of descriptors can be assessed by visual inspection of three dimensional surface plots. The network output was plotted against two input descriptors to generate a functional dependence surface. This gives the idea of how the network output alters in response to the two selected input variables [44]. Fig. 2 shows functional dependence surfaces of the four most important descriptors identified by sensitivity analysis. Non-linearity of inputs is clearly evident suggesting the complex relationship between the input descriptors and log permeability value. In order to compute the relative importance of
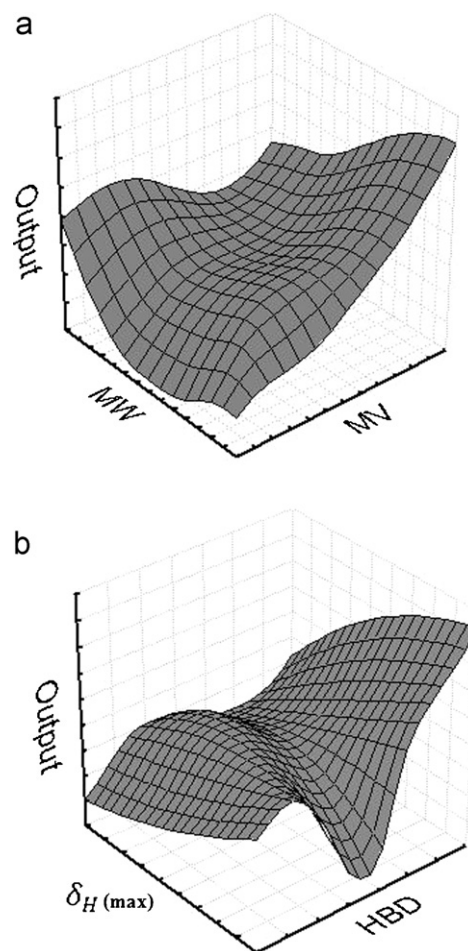
Fig. 2. Functional dependence surfaces for the four most important descriptors. The descriptors are (a) MV and MW; (b) HBD and δ_H.

each descriptor in the MLP-NN model, sensitivity analysis approach was used to assess the importance of the input variables in the fitted model. In this approach, the sum of squares residuals for the model is computed when the respective descriptor is eliminated from the MLP-NN, then, ratios of the full model versus the reduced model are also calculated and the descriptors can be sorted by their importance for the particular neural net [45]. According to the sensitivity analysis, the order of descriptor importance was; $\text{MV} \geq \text{HBD} > \delta_H(\text{max}) > \log D_{7.4} > \log P > \text{HBA}$. Based on this result, the most important descriptors are the molecular volume and molecular weight which is in accordance with some previous studies [46,47]. Kidron et al. concluded that total number of putative hydrogen bonds and lipophilicity variables resulted in the simplest model and they used two-descriptor models for predicting the corneal permeability. In contrast, we concluded that molecular volume and molecular weight as well as max partial charge on hydrogens are more important than lipophilicity in drug diffusion across cornea. In order to evaluate the predictive power and generalization of the obtained MLP-NN model, the trained network was used to predict the log permeability of the compounds included in the internal and external test sets, respectively. The MLP-NN calculated values of log permeability for the training, internal, and external test sets are presented in Table 1. Also, the statistical parameters obtained for the MLP-NN model are shown in Table 3. Comparison among the values in this table implies that the MLP-NN model produces better statistical results in terms of R, RMSE and Fisher statistic value over the MLR model. This strongly suggested a nonlinear relationship between the selected descriptors

Table 3
Statistical parameters of MLR and MLP-NN models.

Model	Training		Internal test		External test		Validation		Q^2	SPRESS
	R	RMSE	R	RMSE	R	RMSE	R	RMSE		
MLR	0.879	0.318					0.909	0.376	0.584	0.378
MLP-NN	0.990	0.096	0.953	0.151	0.991	0.195			0.774	0.087

and log permeability of drug-like compounds. To further evaluate the MLP-NN model, the leave-many-out cross-validation test was performed and the values of the cross-validation correlation coefficient (Q^2) and standard deviation based on predicted residual sum of square (SPRESS) were calculated for the constructed MLP-NN model. The obtained values of Q^2 and SPRESS for the whole data set are also shown in Table 3. Inspecting these values reveals the successful predictive power of the neural network model. Golbraikh and Trospsha [48] and Roy and Roy [49] suggested a number of criteria to further assess the predictive ability of a QSPR model;

$$q^2 > 0.5 \tag{4}$$

$$R^2 > 0.6 \tag{5}$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \quad \text{or} \quad \frac{(R^2 - R_0'^2)}{R^2} < 0.1 \tag{6}$$

$$0.85 \ll k \ll 1.15 \quad \text{or} \quad 0.85 \ll k' \ll 1.15 \tag{7}$$

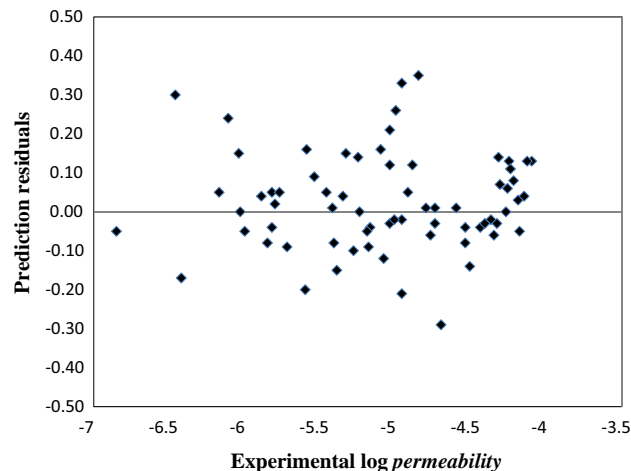$$R_m^2 = R^2 \left(1 - \left| \sqrt{R^2 - R_0^2} \right| \right) > 0.5 \tag{8}$$

Definitions of above parameters are presented obviously in [50] and are not written again here for shortness. The statistical values for the external validation set used in MLP-NN modeling were; $q^2 = 0.938$, $R^2 = 0.982$, $k = 0.972$, $k' = 1.0281$, $(R^2 - R_0^2)/R^2 = 0.012$, $(R^2 - R_0'^2)/R^2 = 0.017$ and $R_m^2 = 0.876$. The obtained values of the model are in good agreement with the limits described above; demonstrating once again the high predictive ability of the MLP-NN model. Fig. 3 shows the plot of the predicted values of log permeability against the experimental values. As can be seen the agreement between measured and predicted results across the entire range of values is excellent. The residuals of the predicted values are plotted against the experimental values in Fig. 4 and the propagation of the residuals on both sides of the zero line indicates that no systematic error exists in the MLP-NN model. The better prediction results of MLP-NN method supports our guess that a nonlinear relationship may exist between selected descriptors and log permeability values of drug-like compounds.
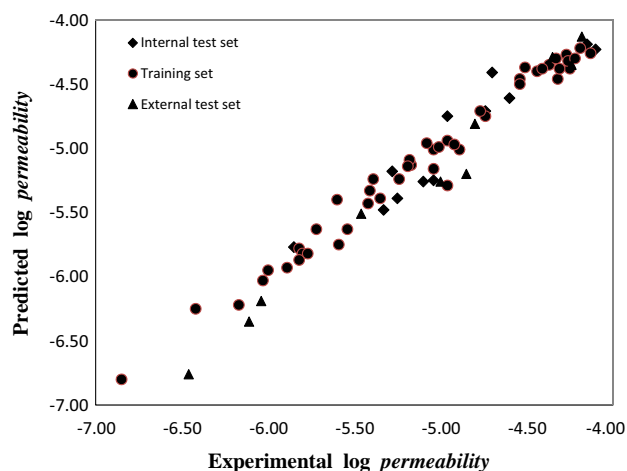
3.2. Qualitative models for corneal permeability prediction

In order to study the chemical variation among the 69 analyzed chemicals and to investigate if these could be separated in groups, hierarchical clustering was applied on a data set including eighteen descriptors (eight descriptors from Dragon and CODESSA and ten descriptors from ACDLabs). Fig. 5 shows the dendrogram of the data. In the dendrogram, objects are merged together according to the Euclidean distance as a similarity measure. The objects, which are least distant from each other, are the most similar ones, and those objects are linked at the lowest levels of the dendrogram. The clusters are found by cutting the highest link(s). There are several methods to join the objects to each other and the one used here is Ward's method [51]. Two clusters (A and B) are found by cutting the highest link of the dendrogram shown in Fig. 5. Also, principal component analysis (PCA) was applied on the same data set using cross-validation as a validation method. 55.7% of the total variance was explained by first two principal components. The PCA grouping of samples is shown in Fig. 6, where each compound is presented as a point. The Figs. 5 and 6 suggest classification (clustering) of drug-like compounds into two different groups of compounds, group A (32 samples; open triangles in Fig. 6) and group B (37 samples; diamonds in Fig. 6). A careful look at the compounds belonging to group A indicated that, 31 compounds (96.87%) permeate across the cornea with the rate of higher than 9×10^{-6} cm/s and only one compound (compound 18) has the permeability value lower than 9×10^{-6} cm/s. On the other hand, about 84% (31 out of 37) of compounds in group B permeate slower than 9×10^{-6} cm/s and six compounds (23, 41, 49, 50, 54 and 59) permeate faster than 9×10^{-6} cm/s. According to this, the corneal permeability of 9×10^{-6} cm/s was selected as a cut off value to classify the desired compounds into low permeable (9×10^{-6} cm/s) and very low permeable (9×10^{-6} cm/s) ones using supervised CPNN classification technique. In order to classify the drug-like compounds, the seven most important descriptors used in MLP-NN modeling (Table 3), were considered as inputs for CPNN. The optimal CPNN model was evaluated with cross-validation test. Classification non-error rate

Fig. 3. Plot of MLP-NN calculated versus experimental log permeability values for all molecules in the data set.

Fig. 4. Plot of prediction residuals versus experimental values of log permeability for all molecules in the data set.
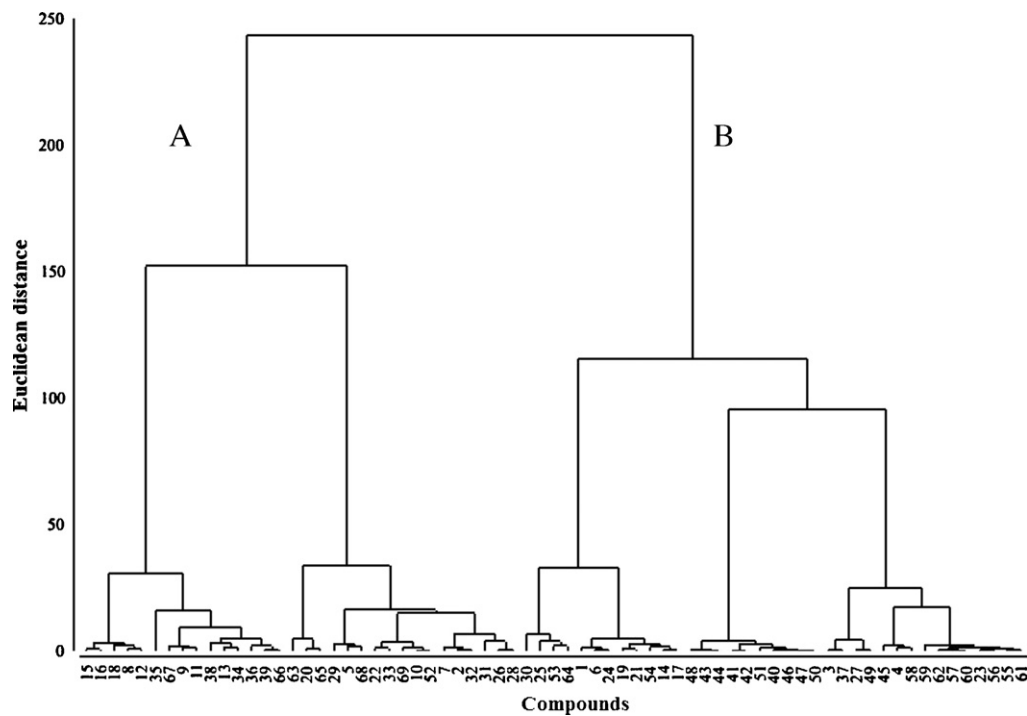
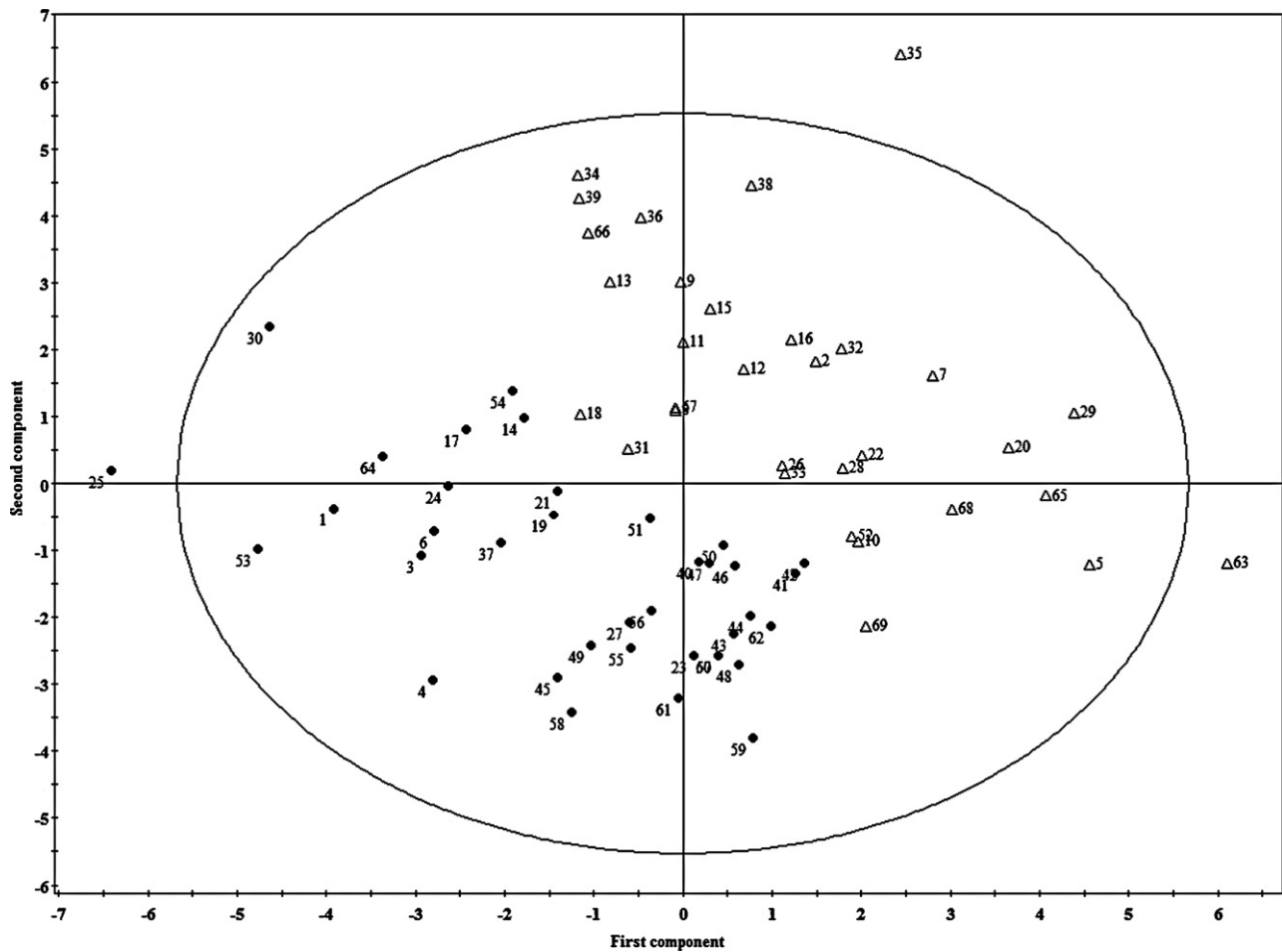**Fig. 5.** Dendrogram of cluster analysis obtained for the 69 compounds.

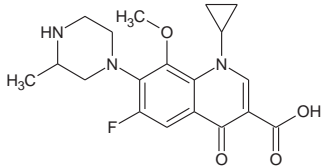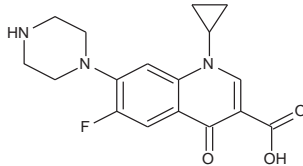**Fig. 6.** PCA score plot (PC1 versus PC2) for the 69 compounds.

Table 4
Confusion matrix and classification parameters for CPNN model.

		LP	VLP	Specificity	Sensitivity	Precision
Training	LP	22	1	0.96	0.96	0.96
	VLP	1	23	0.96	0.96	0.96
Validation	LP	9	0	0.92	1.00	0.90
	VLP	12	1	1.00	0.92	1.00

Table 5
Structural similarity between gatifloxacin and ciperofloxacin.

Name	Gatifloxacin	Ciprofloxacin
Structure		
Defined category	VLP	VLP
Predicted category	LP	VLP

for cross-validation test was 0.89. Also, classification parameters of precision, sensitivity and specificity for LP category were; 0.82, 1.00 and 0.79, and for VLP category were; 1.00, 0.79 and 0.89, respectively. In order to validate the predictive ability of the model, the trained network was used to predict the category of compounds in external validation set. The results of the classification are shown in Table 4. This model provided good performances both in modeling and in prediction. The overall non error rate of this model was 95.7% and 95.4% for the training and prediction sets, respectively. The confusion matrices of this model show that only three drugs (compounds marked by asterisk in Table 1) in all compounds were misclassified. Gatifloxacin (compound 61) in VLP category is included in training set and cannot be assigned to true class by the model. The most similar compound to gatifloxacin is ciprofloxacin (compound 58) which is correctly assigned to VLP category. The structures of these chemicals are shown in Table 5. Inspection of the structures revealed that the difference between these two drugs is methoxy and methyl groups in gatifloxacin. Introduction of these groups in the molecule makes it more lipophilic than ciperofloxacin, therefore gatifloxacin can permeate across the cornea easier and faster than ciperofloxacin. Also, labetalol (compound 23) in validation set and 3-chloroderavative of compound 40 (compound 41) in training set were misclassified. These compounds are among six compounds categorized as B in dendrogaram but permeate faster than 9×10^{-6} cm/s. These compounds which belong to LP group, categorized as VLP because the chemistry of these chemicals are similar to ones belonging to VLP category. These results revealed that the compounds can be categorized into two groups based on the cut off value of 9×10^{-6} cm/s and also revealed the ability of CPNN model to correctly recognize objects belonging to the LP and VLP categories in training and prediction sets.

3.3. Applicability domain

The evaluation of applicability domain is useful to determine if a QSPR model is capable of predicting the property of new compounds with unavailable experimental data. The applicability domain of a QSPR model is defined as the response and chemical structure spaces in which the model makes predictions with a given reliability [52]. In this work, leverage approach and Williams plot were used for the evaluation of the structural applicability domain [53,54]. The leverage or hat value is defined as:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad i = 1, \ldots, n \tag{9}$$

where h_i is the leverage or hat value of the compound (i) in the descriptor space, x_i is the descriptor raw-vector of the query compound, and X is the descriptor matrix. The superscript T refers to the transpose of the matrix and vector. The observation that a chemical has a leverage value greater than the warning leverage (h^*) indicates that the chemical falls outside the applicability domain. The warning leverage is calculated as follows:

$$h^* = \frac{3(p)}{n} \tag{10}$$

where p is the number of model variable plus one, and n is the number of training compounds. As can be seen from Fig. 7, none of the compounds exceed the warning value of h^*. On the other hand, the hat value less than $h^* = 0.511$ does not necessarily guarantee that the chemicals fall within the applicability domain because compounds may be located outside the applicability domain owing to large values of standardized residuals (values more than three times of standard deviation unit). In order to visualize the applicability domain of MLP-NN based on both leverage and standardized residual, the Williams plot is used. It is a plot of standardized residuals versus leverage values with two boundary lines, the first for outliers and the second for high leverage compounds (Fig. 7). As can be seen, no significantly outlier and high leverage compound is
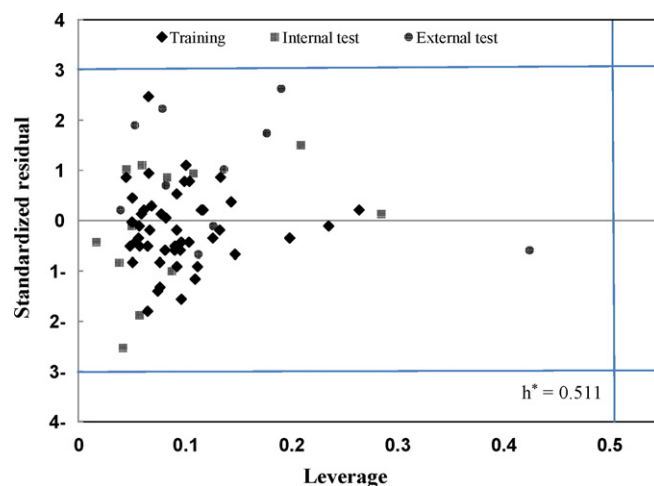
Fig. 7. William plot; plot of standardized residuals versus hat values, with a warning leverage of 0.511.

detected by analyzing the William plot. Therefore, the models can be applied to predict corneal permeability of new drug-like compounds if the calculated leverage values for them are not greater than the warning value of $h^* = 0.511$.

4. Conclusion

In conclusion, the selected descriptors can encode different aspects of molecular structure affecting the permeation of drug candidates across the cornea. Using these descriptors as inputs, an MLP-NN model with good predictive performance for estimating corneal permeability of drug-like compounds has been developed. The MLP-NN model was found to be more successful than MLR equation, indicating that the relationship between descriptors and permeability of the compounds is nonlinear. Cluster analysis and PCA are capable of categorizing the data set in two groups (low permeable and very low permeable compounds) and CPNN, as a supervised nonlinear classification technique, can be used to extract actual information and knowledge from the data set and finally, to discriminate between two categories of drugs.

References

[1] I. Ahmed, R.D. Gokhale, M.V. Shah, T.F. Patton, J. Pharm. Sci. 76 (1987) 583–586.
[2] P. Ashton, S.K. Podder, V.H.L. Lee, Pharm. Res. 8 (1991) 1166–1174.
[3] H. Van de Waterbeemd, G. Camenish, G. Folkers, O.A. Raevsky, Quant. Struct. Acta Relat. 15 (1996) 480–490.
[4] H.S. Huang, R.D. Schoenwald, J.L. Lach, J. Pharm. Sci. 72 (1983) 1272–1279.
[5] A. Edwards, M.R. Prausnitz, Pharm. Res. 18 (2001) 1497–1508.
[6] S. Agatonovic-Kustrin, A. Evans, R.G. Alany, Pharmazie 58 (2003) 725–729.
[7] D.M. Maurice, S. Mishima, in: M.L. Sears (Ed.), Ocular Pharmacokinetics, Springer Verlag, Berlin-Heidelberg, 1984, pp. 16–119.
[8] M.G. Ghosn, V.V. Tuchin, K.V. Larin, Invest. Ophthalmol. Vis. Sci. 48 (2007) 2726–2733.
[9] R.D. Schoenwald, R.L. Ward, J. Pharm. Sci. 67 (1978) 786–788.
[10] E. Toropainen, V.P. Ranta, A. Talvitie, P. Suhonen, A. Urtti, Invest. Ophthalmol. Vis. Sci. 42 (2001) 2942–2948.
[11] E. Toropainen, V.P. Ranta, K.S. Vellonen, J. Palmgren, A. Talvitie, M. Laavola, P. Suhonen, K.M. Hämäläinen, S. Auriola, A. Urtti, Eur. J. Pharm. Sci. 20 (2003) 99–106.
[12] E.A. Tehrany, F. Fournier, S. Desobry, J. Food Eng. 64 (2004) 315–320.
[13] S. Yang, W. Lu, N. Chena, Q. Hu, J. Mol. Struct. 719 (2005) 119–127.
[14] W. Lu, Y. Chen, M. Liu, X. Chen, Z. Hu, Chemosphere 69 (2007) 469–478.
[15] W. Ma, F. Luana, H. Zhang, X. Zhang, M. Liu, Z. Hu, B. Fan, J. Chromatogr. A 1113 (2006) 140–147.
[16] Y. Pan, J. Jiang, R. Wang, H. Cao, Chemom Intell. Lab. Syst. 32 (1996) 177–191.
[17] D. Sola, A. Ferri, M. vBanchero, S. Sicardi, Fluid Phase Equilb. 263 (2008) 33–42.
[18] M.H. Fatemi, J. Chromatogr. A 955 (2002) 273–280.
[19] F. Yoshida, J.G. Topliss, J. Pharm. Sci. 85 (1996) 819–823.
[20] R.D. Schoenwald, H.R. Huang, J. Pharm. Sci. 72 (1983) 1266–1272.
[21] H. Kidron, K. Vellonen, E.M. del Amo, A. Tissari, A. Urtti, J. Pharm. Sci. 27 (2010) 1398–1407.
[22] HyperChem Release 7.0 for windows, Hypercube, Inc., 2002.
[23] J.P.P. Stewart, Version 6.0, Quantum Chemistry Program Exchange, QCPE, No. 455, India University, 1989.
[24] http://www.disat.unimib.it/chem.
[25] A. R. Katritzky, V. S. Lobanov, M. Karelson, Version 2.0, 1994.
[26] N.H. Bingham, J.M. Fry, Regression Linear Models in Statistics, Springer-Verlag, London, 2010.
[27] Advanced Chemistry Developement, Toronto, Canada.
[28] D.H. Rouvray, J. Chem. Inf. Comput. Sci. 32 (1992) 580–586.
[29] A.G. Maldonado, J.P. Doucet, M. Petitjean, B.T. Fan, Mol. Divers. 10 (2006) 39–79.
[30] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, Wiley-VCH, Weinheim, 1999.
[31] N.K. Bose, P. Liang, Neural Network Fundamentals, McGraw-Hill, New York, 1996.
[32] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice-Hal, NJ, 1999.
[33] M. Egmont-Petersen, J.L. Talmo, A. Hasman, Neural Networks 11 (1998) 623–635.
[34] G. Bologna, C. Pellegrini, Phys. Med. 13 (1997) 183–187.
[35] T.F. Rathbun, S.K. Rogers, M.P. DeSimio, Neurocomputing 17 (1997) 195–216.
[36] J. Zupan, M. Novic, J. Gasteiger, Chemom Intell. Lab. Syst. 27 (1995) 175–187.
[37] J. Zupan, M. Novic, I. Ruisánchez, Chemom Intell. Lab. Syst. 38 (1997) 1–23.
[38] M.H. Fatemi, M. Ghorbanzad'e, Eur. J. Med. Chem. 45 (2010) 5051–5055.
[39] D. Ballabio, R. Kokkinofta, R. Todeschini, C.R. Theocharis, Chemom Intell. Lab. Syst. 87 (2007) 52–58.
[40] D. Ballabio, M. Vasighi, V. Consonni, M. Kompany-Zareh, Chemom. Intell. Lab. Syst. 105 (2011) 56–64.
[41] D. Ballabio, V. Consonni, R. Todeschini, Chemom. Intell. Lab. Syst. 98 (2009) 115–122.
[42] V.K. Gupta, H. Khani, B. Ahmadi-Roudi, S. Mirakhorli, E. Fereyduni, S. Agarwal, Talanta 83 (2011) 1014–1022.
[43] N. Fjodorov, M. Vracko, M. Tusar, A. Jezierska, M. Novic, R. Kühne, G. Schüürmann, Mol. Divers. 14 (2010) 581–594.
[44] S. Agatonovic-Kustrin, I.G. Tucker, M. Zecevic, L.J. Zivanovic, Anal. Chim. Acta 418 (2000) 181–195.
[45] D.S. Yeung, I. Cloete, D. Shi, W.W.Y. Ng, Sensitivity Analysis for Neural Networks, Springer-Verlag, Berlin, 2010.
[46] X.C. Fu, W.Q. Liang, Int. J. Pharm. 232 (2002) 193–197.
[47] A.P. Worth, M.T.D. Cronin, ATLA 28 (2000) 403–413.
[48] A. Golbraikh, A. Tropsha, J. Mol, Graph Model. 20 (2002) 269–276.
[49] P.P. Roy, K. Roy, QSAR Comb. Sci. 27 (2008) 302–313.
[50] A. Tropsha, P. Gramatica, V.K. Gombar, QSAR Comb. Sci. 22 (2003) 69–77.
[51] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, Florida, 1989.
[52] T.N. Netzeva, A.P. Worth, T. Aldenberg, A. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D. Roberts, T.W. Schultz, D. Stanton, J.J. van de Sandt, W. Tong, G. Veith, C. Yang, ATLA 33 (2005) 1–19.
[53] A. Tropsha, Mol. Inf. 29 (2010) 476–488.
[54] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, ATLA 33 (2005) 445–459.