Prediction Acidity Constant of Various Benzoic Acids and Phenols in Water Using Linear and Nonlinear QSPR Models

Aziz Habibi-Yangjeh,* Mohammad Danandeh-Jenagharad, and Mahdi Nooshyar

Department of Chemistry, Faculty of Science, University of Mohaghegh Ardebili, P.O. Box 179, Ardebil, Iran
**E-mail: habibiyangjeh@yahoo.com; ahabibi@uma.ac.ir*
Received April 18, 2005

An artificial neural network (ANN) is successfully presented for prediction acidity constant (p*K*a) of various benzoic acids and phenols with diverse chemical structures using a nonlinear quantitative structure-property relationship. A three-layered feed forward ANN with back-propagation of error was generated using six molecular descriptors appearing in the multi-parameter linear regression (MLR) model. The polarizability term (π_1), most positive charge of acidic hydrogen atom (q^+), molecular weight (MW), most negative charge of the acidic oxygen atom (q^-), the hydrogen-bond accepting ability (ε_B) and partial charge weighted topological electronic (PCWTE) descriptors are inputs and its output is p*K*a. It was found that properly selected and trained neural network with 205 compounds could fairly represent dependence of the acidity constant on molecular descriptors. For evaluation of the predictive power of the generated ANN, an optimized network was applied for prediction p*K*a values of 37 compounds in the prediction set, which were not used in the optimization procedure. Squared correlation coefficient (R^2) and root mean square error (RMSE) of 0.9147 and 0.9388 for prediction set by the MLR model should be compared with the values of 0.9939 and 0.2575 by the ANN model. These improvements are due to the fact that acidity constant of benzoic acids and phenols in water shows nonlinear correlations with the molecular descriptors.

Key Words : Quantitative structure-property relationship, Artificial neural networks, Acidity constant, Phenols, Benzoic acids

Introduction

The macroscopic (bulk) activities/properties of chemical compounds clearly depend on their microscopic (structural) characteristics. Development of quantitative structure-property/activity relationships (QSPR/QSAR) on theoretical descriptors is a powerful tool not only for prediction of the chemical, physical and biological properties/activities of compounds, but also for deeper understanding of the detailed mechanisms of interactions in complex systems that predetermine these properties/activities.[1-10] QSPR/QSAR models are essentially calibration models in which the independent variables are molecular descriptors that describe the structure of the molecules and the dependent variable is the property or activity of interest. Since these theoretical descriptors are determined solely from computational methods, a priori predictions of the properties/activities of compounds are possible, no laboratory measurements are needed thus saving time, space, materials, equipment and alleviating safety (toxicity) and disposal concerns. An enormous number of descriptors have been used by researchers to increase the ability to correlate biological, chemical and physical properties. To obtain a significant correlation, it is crucial that appropriate descriptors be employed.[11,12]

Various methods for constructing QSPR/QSAR models have been used including multi-parameter linear regression (MLR), principal component analysis (PCA) and partial least-squares regression (PLS).[13-16] In some cases, it is more convenient that a linear relationship between property/activity and descriptors is considered. If there is not a well-defined linear relationship, the discussed method cannot give a perfect QSPR/QSAR model. Artificial neural networks (ANNs) are capable of recognizing highly nonlinear relationships.[17-20] ANNs are biologically inspired computer programs designed to simulate the way in which the human brain processes information. ANNs gather their knowledge by detecting the patterns and relationships in data and learned (or trained) through experience, not from programming. There are many types of neural networks designed by now and new ones are invented every week.[20] The behavior of a neural network is determined by transfer functions of its neurons, by learning rule, and by the architecture itself. An ANN is formed from artificial neuron or processing elements (PE), connected with coefficients (weights), which constitute the neural structure and are organized in layers. The first layer is termed the input layer, and the last layer is the output layer. The layers of neurons between the input and output layers are called hidden layers. Neural networks do not need on explicit formulation of the mathematical or physical relationships of the handled problem. These give ANNs an advantage over traditional fitting methods for some chemical application. For these reason in recent years, ANNs have been used to a wide variety of chemical problems such as simulation of mass spectra, ion interaction chromatography, aqueous solubility and partition coefficient, simulation of nuclear magnetic resonance spectra, prediction of bioconcentration factor, solvent effects on reaction rate and prediction of normalized polarity parameter

in mixed solvent systems.[21-36]

It has been shown that the acid-base properties affect the toxicity, chromatographic retention behavior and pharmaceutical properties of organic acids and bases.[37,38] On the other hand, interpretation and prediction of p*K*a values for chemical compounds are of general importance and usefulness for chemists.[39] Although in the last years several theoretical studies have been performed for correlation of p*K*a values with molecular parameters, but in these studies linear equations have been used.[38-46]

The main aim of present work is to develop a linear and nonlinear QSPR models based on molecular descriptors for prediction p*K*a values of various benzoic acids and phenols with diverse chemical structures (including 242 compounds).

Theory

A detailed description of theory behind a neural network has been adequately described by different researchers.[17-19] There are many types of neural network architectures, but the type that has been most useful for QSAR/QSPR studies is the multilayer feed - forward network with back-propagation (BP) learning rule.[20] The number of neurons in the input and output layers are defined by system's properties. The number of neurons in the hidden layer could be considered as an adjustable parameter, which should be optimized. The input layer receives the experimental or theoretical information. The output layer produces the calculated values of dependent variable. The use of ANNs consists of two steps: "training" and "prediction". In the training phase the optimum structure, weight coefficients and biases are searched for. These parameters are found from training and validation data sets. After the training phase, the trained network can be used to predict (or calculate) the outputs from a set of inputs. ANNs allow one to estimate relationships between input variables and one or several output dependent variables. The ANN reads the input and target values in the training data set and changes the values of the weighted links to reduce the difference between the calculated output and target values. The error between output and target values is minimized across many training cycles until network reaches specified level of accuracy. If a network is left to train for too long, however, it will overtrain and will lose the ability to generalize.[22-36]

Experimental Section

Descriptor generation. The derivation of theoretical molecular descriptors proceeds from the chemical structure of the compounds. In order to calculate the theoretical descriptors, the z-matrices (molecular models) were constructed with the aid of HyperChem 7.0 and molecular structures were optimized using AM1 algorithm.[47] In order to calculate some of theoretical descriptors, the molecular geometries of molecules were further optimized with the same algorithm in MOPAC program version 6.0. The other molecular electronic descriptors were calculated by *Dragon* package version 2.1.[48] For this propose the output of the HyperChem software for each compound feed into the *Dragon* program and the descriptors were calculated. As a result, a total of 18 theoretical descriptors were calculated for each compound in the data sets (242 compounds).

Linear correlations. Acidity constant of benzoic acids and phenols are literature values at 25 ºC.[49] MLR model was developed for prediction of p*K*a values by molecular descriptors. The method of stepwise multi-parameter linear regression was used to select the most important descriptors and to calculate the coefficients relating the p*K*a to the descriptors. The MLR models were generated using spss/pc software package release 9.0.

Neural network generation. The specification of a typical neural network model requires the choice of the type of inputs, the number of hidden layers, the number of neurons in each hidden layer and the connection structure between the inputs and the output layers. The number of input nodes in the ANNs was equal to the number of molecular descriptors in the MLR model. A three-layer network with a sigmoidal transfer function was designed. The initial weights were randomly selected between 0 and 1. Before training, the input and output values were normalized between 0.1 and 0.9. The optimization of the weights and biases was carried out according to the resilient back-propagation algorithm.[50] The data set was randomly divided into three groups: a training set, a validation set and a prediction set consisting of 168, 37 and 37 molecules, respectively. The training and validation sets were used for the model generation and the prediction set was used for evaluation of the generated model, because a prediction set is a better estimator of the ANN generalization ability than a validation (monitoring) set.[51]

The performances of training, validation and prediction of ANNs are evaluated by the mean percentage deviation (MPD) and root-mean square error (RMSE), which are defined as follows:

$$\mathrm{MPD} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{(P_i^{\exp} - P_i^{cal})}{P_i^{\exp}}\right| \quad (1)$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N}\frac{(P_i^{\exp} - P_i^{cal})^2}{N}} \quad (2)$$

where $P_i^{\exp}$ and P_i^{cal} are experimental and calculated values of p*K*a with the models and N denote the number of data points.

Individual percent deviation (IPD) is defined as follows:

$$\mathrm{IPD} = 100 \times \left(\frac{P_i^{calc} - P_i^{\exp}}{P_i^{\exp}}\right) \quad (3)$$

The processing of the data was carried using Matlab 6.5.[52] The neural networks were implemented using Neural Network Toolbox Ver. 4.0 for Matlab.[50]

Results and Discussion

A major challenge in the development of MLR equations is connected with the possible multicollinearity of molecular descriptors. In order to decrease the redundancy existed in the descriptors data matrix, the correlation of descriptors with each other and with p*K*a of the compounds was examined and collinear descriptors were detected ($r > 0.85$). Among the collinear descriptors, one with the lowest correlation with the property was removed from the data matrix. Table 1 demonstrates that all of the descriptors are strongly orthogonal which reflects the statistical reliability of the model.

Multi-parameter linear correlation of p*K*a values of 168 benzoic acids and phenols *versus* the molecular descriptors in the training set gives the results in Table 2. It can be seen from this table that six descriptors are appeared in the MLR model. These descriptors are: polarizability index (π_I), most positive charge of acidic hydrogen atom (q^+), molecular weight (MW), most negative charge of acidic oxygen atom (q^-), the hydrogen-bond accepting ability (ε_B) and partial charge weighted topological electronic (PCWTE) descriptors.

The negative coefficient for π_I, q^+, q^- and MW descriptors indicate that with increasing these descriptors, acidity constant (*K*a) increases. With increasing q^+ and q^- of the compounds, interactions of water molecules with acidic hydrogen and oxygen of the compounds increase, then acidic hydrogen can be easily removed from the compounds. Polarizability and then the dipole-induced dipole interactions increase with increasing π_I and MW, as a result acidity of the compounds increases with increasing these descriptors.[53] Acidity constant of the compounds decrease with increasing ε_B and PCWTE descriptors, because basicity of phenolic oxygen atom increases with increasing these descriptors. Effects of π_I, q^+ and MW on p*K*a are higher than that of the other descriptors, because standardized coefficients of π_I, q^+ and MW are higher than those of the other descriptors.

The calculated values of p*K*a for the compounds in training, validation and prediction sets using the MLR model have been plotted *versus* the experimental values of it in Figure 1.

The next step in this work was generation of the ANN model. There are no rigorous theoretical principles for choosing the proper network topology; so different structures were tested in order to obtain the optimal hidden neurons and training cycles.[36] Before training the network, the number of nodes in the hidden layer was optimized. In order to optimize the number of nodes in the hidden layer, several training sessions were conducted with different numbers of hidden nodes. The root mean squared error of training (RMSET) and validation (RMSEV) sets were obtained at various iterations for different number of neu-

Table 1. Correlation coefficients between various theoretical descriptors that have been used in the multi-parameter linear regression (MLR) and artificial neural network (ANN) models

Descriptor	π_I	q^+	q^-	ε_B	MW	PCWTE
π_I	1	0.530	0.042	0.150	0.329	0.285
q^+	0.530	1	0.642	0.546	0.368	0.070
q^-	0.042	0.642	1	0.236	0.155	0.018
ε_B	0.150	0.546	0.236	1	0.248	0.038
MW	0.329	0.368	0.155	0.248	1	0.237
PCWTE	0.285	0.070	0.018	0.038	0.237	1

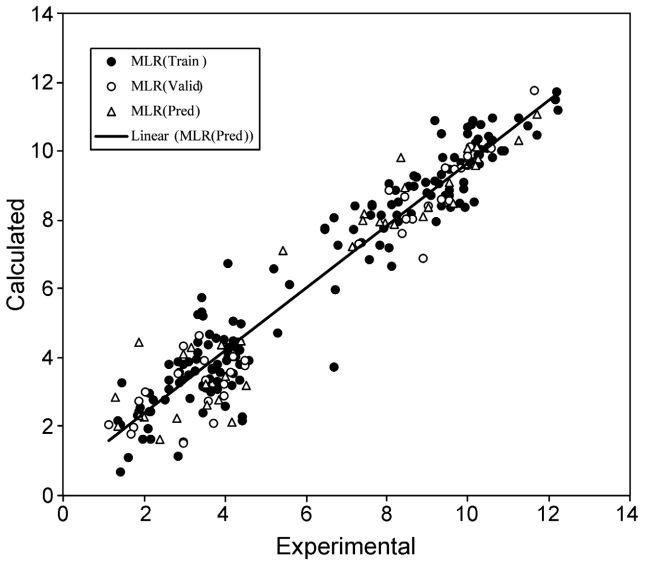

Figure 1. Plot of the calculated values of p*K*a from the MLR model *versus* the experimental values of it for training, validation and prediction sets.

Table 2. Descriptors, symbols and results of the multi-parameter linear regression (MLR) model[a]

No.	Descriptor	Symbol	Coefficient	β
1	polarizability term	π_I	−8.361	0.080
2	most positive charge of acidic hydrogen atom	q^+	−110.471	0.521
3	molecular weight	MW	−0.0051	0.074
4	most negative charge of the phenolic oxygen atom	q−	−26.394	0.321
5	the hydrogen-bond accepting ability	ε_B	34.445	0.080
6	partial charge weighted topological electronic	PCWTE	0.0902	0.101
7	constant		42.278	

[a] The β is standardized coefficient of descriptors. The polarizability term (π_I) is obtained by dividing the polarizability volume by the molecular volume. The ε_B is equal $0.3–0.01(E_{lw}–E_h)$, in which E_{lw} and E_h are referring to the LUMO energy for water and HOMO energy for the compound, respectively.

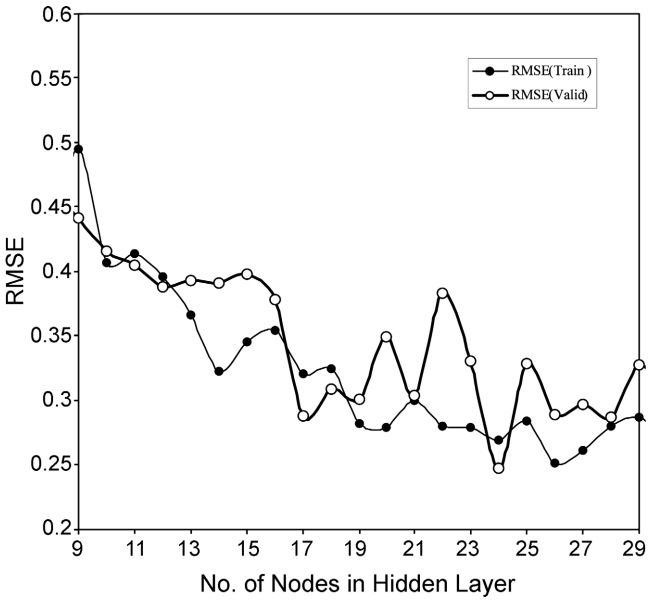

Figure 2. Plot of RMSE for training and validation sets *versus* the number of nodes in hidden layer.

rons at the hidden layer and the minimum value of RMSEV was recorded as the optimum value. Plot of RMSET and RMSEV *versus* the number of nodes in the hidden layer has been shown in Figure 2. It is clear that the twenty-four nodes in hidden layer is optimum value.

The six descriptors appearing in the MLR model (including π_I, q^+, MW, q^-, ε_B, and PCWTE descriptors) were considered as inputs for developing the ANN. Then an ANN with architecture 6-24-1 was generated. It is note worthy that training of the network was stopped when the RMSEV started to increases *i.e.* when overtraining begins. The overtraining causes the ANN to loose its prediction power.[34,36] Therefore, during training of the networks, it is desirable that

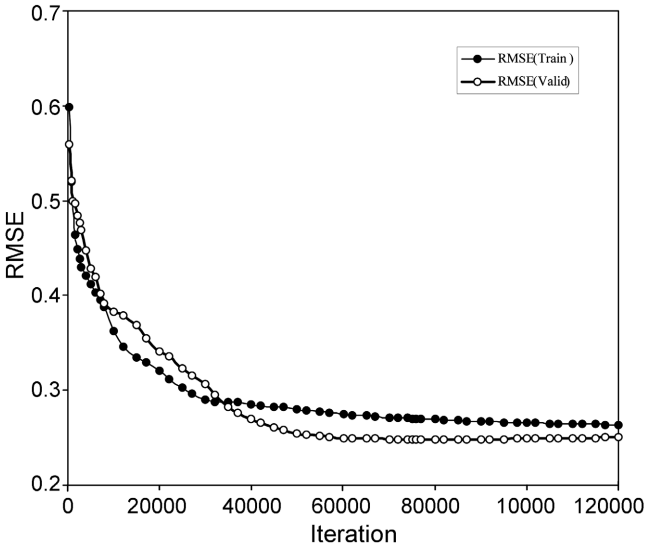

Figure 3. Plot of RMSE for training and validation sets *versus* the number of iterations.

iterations are stopped when overtraining begins. To control the overtraining of the network during the training procedure, the values of RMSET and RMSEV were calculated and recorded to monitor the extent of the learning in various iterations. Results obtained showed that after 77000 iterations the value of RMSEV started to increase very little and overfitting slightly began (Figure 3).

The generated ANN was then trained using the training and validation sets for the optimization of the weights and biases. For the evaluation of the predictive power of the generated ANN, an optimized network was applied for prediction of the p*K*a values of the compounds in the prediction set, which were not used in the modeling procedure (Table 3). The calculated values of p*K*a for the compounds in training, validation and prediction sets using the ANN model have been plotted *versus* the experimental values of it in Figure 4.

As expected, the calculated values of p*K*a are in good agreement with those of the experimental values. The correlation equation for all of the calculated values of p*K*a from the ANN model and the experimental values is as follows:

$$pKa(cal) = 0.99299\ pKa(exp) + 0.04454 \quad (4)$$

($R^2 = 0.9931$; MPD = 4.5044; RMSE = 0.2648; F = 34295.94)

Similarly, the correlation of p*K*a (cal) *versus* p*K*a (exp) values in the prediction set gives equation (5):

$$pKa(cal) = 1.01212pKa(exp) - 0.08200 \quad (5)$$

($R^2 = 0.9939$; MPD = 5.0361; RMSE = 0.2575; F = 5718.11)

Plot of IPD for p*K*a values in the prediction set *versus* the experimental values of it has been illustrated in Figure 5. As can be seen, the model did not show proportional and systematic error, because the slope (a = 1.01212) and intercept (b = −0.08200) of the correlation equation are not significantly different from unity and zero, respectively and the propagation of errors in both sides of zero is random (Figure 5).

Table 4 compares the results obtained using the MLR and ANN models. The squared correlation coefficient (R^2) and RMSE of the models for total, training, validation and prediction sets show potential of the ANN model for prediction of p*K*a values of various benzoic acids and phenols in water with one model.

As a result, it was found that properly selected and trained neural network could fairly represent dependence of the acidity constant of benzoic acids and phenols in water on the molecular descriptors. Then the optimized neural network could simulate the complicated nonlinear relationship between p*K*a values and the molecular descriptors. It can be seen from Table 4 that although the parameters appearing in the MLR model are used as inputs for the generated ANN, the statistics is shown a large improvement. These improvements are due to the fact that p*K*a values of the compounds show nonlinear correlations with the molecular descriptors.

Table 3. Experimental and calculated values of p*K*a for various benzoic acids and phenols in water at 25 °C for training, validation and prediction sets by multi-parameter linear regression (MLR) and artificial neural network (ANN) models along with individual percent deviation (IPD)[a]

No.	Compound	Exp.	MLR	IPD_{MLR}	ANN	IPD_{ANN}
	Training set					
1	2-acetylphenol	9.19	10.867	18.25	9.272	0.89
2	4-acetylphenol	8.05	9.05	12.42	8.794	9.24
3	2-allylphenol	10.28	9.637	−6.25	9.972	−3.00
4	4-bromophenol	9.34	8.42	−9.85	9.126	−2.29
5	2,6-di-*tert*-butyl-4-bromophenol	10.83	10.011	−7.56	10.975	1.34
6	2,6-di-*tert*-butyl-4-methylphenol	12.23	11.201	−8.41	11.983	−2.02
7	2,6-di-*tert*-butyl-4-methoxyphenol	12.15	11.484	−5.48	11.936	−1.76
8	2-*tert*-butylphenol	11.24	10.96	−2.49	10.752	−4.34
9	3-*tert*-butylphenol	10.1	10.773	6.66	10.504	4.00
10	4-*tert*-butylphenol	10.31	10.768	4.44	10.788	4.64
11	1-chloro-2,6-dimethyl-4-hydroxybenzene	9.549	9.476	−0.76	9.944	4.14
12	4-chloro-2-nitrophenol	6.48	7.724	19.20	6.475	−0.08
13	2-chlorophenol	8.55	8.974	4.96	8.117	−5.06
14	3-chlorophenol	9.10	8.72	−4.18	8.971	−1.42
15	4-chlorophenol	9.43	8.725	−7.48	9.03	−4.24
16	*o*-cresol	10.26	9.875	−3.75	10.174	−0.84
17	4-cyano-2,6-dimethylphenol	8.27	7.934	−4.06	7.872	−4.81
18	4-cyano-3,5-dimethylphenol	8.21	8.869	8.03	8.033	−2.16
19	3-cyanophenol	8.61	8.168	−5.13	8.2	−4.76
20	3,5-dibromophenol	8.056	7.186	−10.80	8.024	−0.40
21	2,4-dichlorophenol	7.85	8.141	3.71	7.96	1.40
22	2,6-dichlorophenol	6.78	7.264	7.14	6.827	0.69
23	3,5-diethoxyphenol	9.370	9.813	4.73	9.529	1.70
24	3-(diethoxyphosphinyl)phenol	8.68	9.267	6.76	8.628	−0.60
25	4-(diethoxyphosphinyl)phenol	8.28	8.517	2.86	8.276	−0.05
26	3,4-dihydroxybenzaldehyde	7.55	6.84	−9.40	7.623	0.97
27	1,2-dihydroxybenzene	9.356	10.487	12.09	9.456	1.07
28	1,4-dihydroxy-2,6-dinitrobenzene	4.42	2.184	−50.59	4.425	0.11
29	1,3-dihydroxy-2-methylbenzene	10.05	9.685	−3.63	9.603	−4.45
30	1,2-dihydroxy-3-nitrobenzene	6.68	3.728	−44.19	6.68	0.00
31	1,2-dihydroxy-4-nitrobenzene	6.701	8.082	20.61	6.824	1.84
32	3,5-diiodophenol	8.103	6.653	−17.89	8.126	0.28
33	3,5-dimethoxyphenol	9.345	9.32	−0.27	9.497	1.63
34	2,6-dimethyl-4-nitrophenol	7.190	7.736	7.59	7.439	3.46
35	3,5-dimethyl-4-nitrophenol	8.245	8.131	−1.38	8.299	0.65
36	2,3-dimethylphenol	10.50	10.117	−3.65	10.246	−2.42
37	2,6-dimethylphenol	10.59	9.824	−7.23	10.264	−3.08
38	3,4-dimethylphenol	10.32	10.055	−2.57	10.298	−0.21
39	3,5-dimethylphenol	10.15	10.113	−0.36	10.068	−0.81
40	2,4-dinitrophenol	4.08	6.75	65.44	4.081	0.02
41	2,5-dinitrophenol	5.216	6.568	25.92	5.222	0.12
42	3,5-dinitrophenol	6.732	5.961	−11.45	6.658	−1.10
43	2-ethoxyphenol	10.109	10.886	7.69	10.117	0.08
44	3-ethoxyphenol	9.655	9.801	1.51	9.617	−0.39
45	2-ethylphenol	10.2	10.221	0.21	10.28	0.78
46	2-fluorophenol	8.73	9.247	5.92	9.112	4.38
47	3-fluorophenol	9.29	9.072	−2.35	9.16	−1.40
48	4-fluorophenol	9.89	9.078	−8.21	9.992	1.03
49	2'-hydroxyacetophenone	9.90	8.906	−10.04	9.232	−6.75
50	3'-hydroxyacetophenone	9.19	9.114	−0.83	9.46	2.94
51	3-hydroxybenzaldehyde	9.00	8.799	−2.23	9.25	2.78
52	4-hydroxybenzaldehyde	7.620	8.423	10.54	7.96	4.46

Table 3. Continued

No.	Compound	Exp.	MLR	IPD_{MLR}	ANN	IPD_{ANN}
53	2-hydroxybenzyl alcohol	9.92	8.355	−15.78	10.146	2.28
54	3-hydroxybenzyl alcohol	9.83	9.663	−1.70	9.866	0.37
55	1-hydroxy-2,4-dihydroxymethylbenzene	9.79	8.488	−13.30	9.615	−1.79
56	2-hydroxy-3-methoxybenzaldehyde	7.912	7.76	−1.92	7.974	0.78
57	(2-hydroxy-5-methylbenzene)-methanol	10.15	8.525	−16.01	9.876	−2.70
58	1-hydroxy-2-propylbenzene	10.50	10.439	−0.58	10.586	0.82
59	4-hydroxy-α,α,α-trifluorotoluene	8.675	8.982	3.54	8.559	−1.34
60	1-hydroxy-2,4,6-trihydroxymethylbenzene	9.56	8.357	−12.58	9.623	0.66
61	4-indanol	10.32	10.033	−2.78	10.289	−0.30
62	4-iodophenol	9.200	7.939	−13.71	9.088	−1.22
63	2,6-di-iodo-4-nitrophenol	3.32	4.131	24.43	3.304	−0.48
64	2-methoxyphenol	9.99	10.489	4.99	9.427	−5.64
65	2-methoxy-4-(2-propenyl)phenol	10.0	10.686	6.86	10.093	0.93
66	6-methyl-2-butylphenol	11.72	10.478	−10.60	11.065	−5.59
67	2-methyl-4-*tert*-butylphenol	10.59	10.963	3.52	11.007	3.94
68	2,2'-methylenebis(4-chlorophenol)	7.6	8.15	7.24	7.566	−0.45
69	2,2'-methylenebis(4,6-dichlorophenol)	5.6	6.129	9.45	5.637	0.66
70	4-methylsulfonyl-3,5-dimethylphenol	8.13	8.435	3.75	8.094	−0.44
71	3-(s-methylthio)phenol	9.53	8.572	−10.05	9.636	1.11
72	4-(s-methylthio)phenol	9.53	8.717	-8.53	9.743	2.24
73	2-nitrohydroquinone	7.63	8.45	10.75	7.633	0.04
74	2-nitrophenol	7.222	8.396	16.26	7.166	−0.78
75	4-nitrosophenol	6.48	7.768	19.88	6.693	3.29
76	phenol	9.99	9.578	−4.12	10.346	3.56
77	2-phenylphenol	9.55	8.847	−7.36	9.367	−1.92
78	5,6,7,8-tetrahydro-1-naphthol	10.28	9.883	−3.86	10.473	1.88
79	5,6,7,8-tetrahydro-2-naphthol	10.48	10.088	−3.74	10.639	1.52
80	2,4,6-tri-*tert*-butylphenol	12.19	11.724	−3.82	12.342	1.25
81	2,4,5-trichlorophenol	7.37	7.345	−0.34	7.396	0.35
82	3,4,5-trichlorophenol	7.839	7.275	−7.19	7.771	−0.87
83	3-trifluoromethylphenol	8.950	9.098	1.65	9.242	3.26
84	2,3,4-trimethylphenol	10.59	10.322	−2.53	10.648	0.55
85	2,4,5-trimethylphenol	10.57	10.325	−2.32	10.684	1.08
86	3,4,5-trimethylphenol	10.25	10.358	1.05	10.498	2.42
87	2,4,6-trimethylphenol	10.88	10.002	−8.07	10.58	−2.76
88	2,4,6-tripropylphenol	11.47	10.741	−6.36	11.165	−2.66
89	2-acetamidobenzoic acid	3.63	4.692	29.26	3.641	0.30
90	3-acetamidobenzoic acid	4.07	4.289	5.38	4.212	3.49
91	4-acetamidobenzoic acid	4.28	4.035	−5.72	3.87	−9.58
92	4-acetoxybenzoic acid	4.38	3.351	−23.49	3.979	−9.16
93	2-acetylbenzoic acid	4.13	4.025	−2.54	4.023	−2.59
94	3-acetylbenzoic acid	3.83	3.79	−1.04	3.597	−6.08
95	4-acetylbenzoic acid	3.70	3.654	−1.24	3.87	4.59
96	3-amino-1-naphthoic acid	2.61	3.817	46.25	2.834	8.58
97	anthracene-9-carboxylic acid	3.65	2.991	−18.05	3.562	−2.41
98	1,3-benzenedicarboxylic acid	3.62	3.374	−6.80	3.571	−1.35
99	1,4-benzenedicarboxylic acid	3.54	3.266	−7.74	3.975	12.29
100	1,2,4,5-benzenetetracarboxylic acid	1.92	2.563	33.49	2.002	4.27
101	1,2,3-benzenetricarboxylic acid	2.88	3.582	24.38	2.771	−3.78
102	1,2,4-benzenetricarboxylic acid	2.52	2.79	10.71	2.375	−5.75
103	1,3,5-benzenetricarboxylic acid	2.12	2.971	40.14	2.34	10.38
104	benzilic acid	3.09	3.872	25.31	3.27	5.83
105	benzylamine-4-carboxylic acid	3.59	4.358	21.39	4.14	15.32
106	2-biphenylcarboxylic acid	3.46	3.194	−7.69	3.665	5.92
107	2-bromobezoic acid	2.85	3.896	36.70	2.918	2.39

Table 3. Continued

No.	Compound	Exp.	MLR	IPD_{MLR}	ANN	IPD_{ANN}
108	3-bromobezoic acid	3.810	3.065	−19.55	3.915	2.76
109	3-*tert*-butylbenzoic acid	4.199	5.079	20.96	4.418	5.22
110	4-*tert*-butylbenzoic acid	4.389	4.985	13.58	4.104	−6.49
111	2-chlorobenzoic acid	2.877	3.287	14.25	3.086	7.26
112	3-chlorobenzoic acid	3.83	3.319	−13.34	3.698	−3.45
113	2-chloro-4-nitrobenzoic acid	1.96	1.619	−17.40	1.868	−4.69
114	2-chloro-5-nitrobenzoic acid	2.17	1.654	−23.78	1.929	−11.11
115	2-chloro-6-nitrobenzoic acid	1.342	2.153	60.43	1.417	5.59
116	2-cyanobenzoic acid	3.14	2.809	−10.54	2.962	−5.67
117	3,5-diaminobenzoic acid	5.30	4.731	−10.74	5.225	−1.42
118	3,6-dichlorophthalic acid	1.46	3.287	125.14	1.271	−12.95
119	2,4-dihydroxybenzoic acid	3.29	3.956	20.24	3.639	10.61
120	2,5-dihydroxybenzoic acid	2.97	3.802	28.01	3.408	14.75
121	3,5-dihydroxybenzoic acid	4.04	3.911	−3.19	3.805	−5.82
122	2,6-dimethoxybenzoic acid	3.44	5.312	54.42	3.447	0.20
123	2,3-dimethylbenzoic acid	3.771	4.55	20.66	3.668	−2.73
124	2,4-dimethylbenzoic acid	4.217	4.499	6.69	4.019	−4.70
125	2,5-dimethylbenzoic acid	3.990	4.522	13.33	3.799	−4.79
126	3,5-dimethylbenzoic acid	4.302	4.467	3.84	4.189	−2.63
127	2,3-dimethylnaphthalene-1-carboxylic acid	3.33	4.439	33.30	3.532	6.07
128	2,3-dinitrobenzoic acid	1.85	2.331	26.00	2.258	22.05
129	2,4-dinitrobenzoic acid	1.43	0.688	−51.89	1.587	10.98
130	2,5-dinitrobenzoic acid	1.62	1.12	−30.86	1.905	17.59
131	3,5-dinitrobenzoic acid	2.85	1.127	−60.46	2.681	−5.93
132	2-ethylbenzoic acid	3.79	4.575	20.71	3.757	−0.87
133	4-ethylbenzoic acid	4.35	4.451	2.32	4.2	−3.45
134	2-fluorobenzoic acid	3.27	3.629	10.98	3.67	12.23
135	3-fluorobenzoic acid	3.865	3.586	−7.22	3.584	−7.27
136	3-hydroxybenzoic acid	4.076	3.993	−2.04	4.496	10.30
137	4-hydroxybenzoic acid	4.582	3.912	−14.62	4.627	0.98
138	2-hydroxy-5-bromobenzoic acid	2.61	3.096	18.62	2.722	4.29
139	2-hydroxy-5-chlorobenzoic acid	2.63	3.342	27.07	3.394	29.05
140	4-hydroxy-3-methoxybenzoic acid	4.355	3.803	−12.68	4.022	−7.65
141	2-hydroxy-5-methylbenzoic acid	4.08	4.275	4.78	3.365	−17.52
142	2-hydroxy-6-methylbenzoic acid	3.32	5.268	58.67	3.231	−2.68
143	2-hydroxy-3-nitrobenzoic acid	1.87	2.427	29.79	1.93	3.21
144	2-hydroxy-5-nitrobenzoic acid	2.12	2.435	14.86	2.059	−2.88
145	2-hydroxy-6-nitrobenzoic acid	2.24	2.774	23.84	2.723	21.56
146	4-iodobenzoic acid	4.00	2.59	−35.25	4.162	4.05
147	mesitylenic acid	4.32	4.467	3.40	4.189	−3.03
148	2-methoxybenzoic acid	4.09	4.146	1.37	3.963	−3.11
149	3-methoxybenzoic acid	4.08	4.006	−1.81	4.305	5.51
150	3-methylbenzoic acid	4.269	4.282	0.30	4.303	0.80
151	4-methylbenzoic acid	4.362	4.214	−3.39	4.541	4.10
152	2-methyl-3,5-dinitrobenzoic acid	2.97	1.552	−47.74	2.982	0.40
153	2-methyl-1-naphthoic acid	3.11	3.509	12.83	3.135	0.80
154	3-methylsulfonylbenzoic acid	3.52	3.113	−11.56	3.581	1.73
155	4-methylsulfonylbenzoic acid	3.64	3.062	−15.88	3.323	−8.71
156	1-naphthalenecarboxylic acid	3.695	3.68	−0.41	3.938	6.58
157	2-naphthalenecarboxylic acid	4.161	3.214	−22.76	3.693	−11.25
158	4-nitrobenzene-1,2-dicarboxylic acid	2.11	1.933	−8.39	1.917	−9.15
159	2-nitrobenzoic acid	2.18	2.444	12.11	2.429	11.42
160	3-nitrobenzoic acid	3.46	2.38	−31.21	3.196	−7.63
161	4-nitrobenzoic acid	4.441	2.295	−48.32	3.286	−26.01
162	*o*-phthalic acid	2.950	3.361	13.93	2.555	−13.39

Table 3. Continued

No.	Compound	Exp.	MLR	IPD_{MLR}	ANN	IPD_{ANN}
163	3-sulfamylbenzoic acid	3.54	3.209	−9.35	3.73	5.37
164	4-sulfamylbenzoic acid	3.47	3.159	−8.96	3.324	−4.21
165	2,3,5,6-tetramethylbenzoic acid	3.415	5.746	68.26	3.401	−0.41
166	2,4,6-tribromobenzoic acid	1.41	2.063	46.31	1.408	−0.14
167	3,4,5-trihydroxybenzoic acid	4.19	3.528	−15.80	3.824	−8.74
168	2,4,6-trimethylbenzoic acid	3.448	5.211	51.13	3.641	5.60
	Validation set					
169	2-bromophenol	8.452	8.691	2.83	8.673	2.61
170	2,4-di-*tert*-butylphenol	11.64	11.741	0.87	11.887	2.12
171	4-chloro-2,6-dinitrophenol	2.97	1.531	−48.45	2.959	−0.37
172	*m*-cresol	10.00	9.835	−1.65	10.017	0.17
173	1,3-dichloro-2,5-dihydroxybenzene	7.30	7.295	−0.07	7.066	−3.21
174	3,4-dichlorophenol	8.630	8.014	−7.14	8.343	−3.33
175	1,3-dihydroxybenzene	9.44	9.497	0.60	9.415	−0.26
176	2,4-dimethylphenol	10.58	10.069	−4.83	10.367	−2.01
177	2,6-dinitrophenol	3.713	2.105	−43.31	3.718	0.13
178	3-ethylphenol	10.07	10.127	0.57	10.217	1.46
179	4'-hydroxyacetophenone	8.05	8.846	9.89	8.155	1.30
180	4-hydroxybenzyl alcohol	9.82	9.53	−2.95	9.829	0.09
181	3-hydroxy-4-methoxybenzaldehyde	8.889	6.87	−22.71	8.744	−1.63
182	2-iodophenol	8.464	8.042	−4.99	8.505	0.48
183	3-methoxyphenol	9.652	9.475	−1.83	9.698	0.48
184	3-methylsulfonylphenol	9.33	8.579	−8.05	9.219	−1.19
185	3-nitrophenol	8.360	7.627	−8.77	8.185	−2.09
186	4-phenylphenol	9.55	8.554	−10.43	9.647	1.02
187	1,2,3-trihydroxybenzene	9.03	8.412	−6.84	8.938	−1.02
188	2-acetoxybenzoic acid	3.48	3.934	13.05	3.743	7.56
189	4-amino-2-naphthoic acid	2.89	3.547	22.73	3.028	4.78
190	1,2,3,4-benzenetetracarboxylic acid	2.05	3.004	46.54	2.064	0.68
191	benzoic acid	4.204	4.015	−4.50	4.297	2.21
192	4-bromobezoic acid	3.99	2.879	−27.84	4.057	1.68
193	4-chlorobenzoic acid	3.986	3.234	−18.87	3.88	−2.66
194	3-cyanobenzoic acid	3.60	2.725	−24.31	3.381	−6.08
195	3,4-dihydroxybenzoic acid	4.48	3.758	−16.12	4.153	−7.30
196	2,6-dimethylbenzoic acid	3.362	4.656	38.49	3.486	3.69
197	2,6-dinitrobenzoic acid	1.14	2.051	79.91	1.103	−3.25
198	4-fluorobenzoic acid	4.14	3.573	−13.70	3.854	−6.91
199	2-hydroxy-3-methylbenzoic acid	2.99	4.336	45.02	3.376	12.91
200	2-iodobenzoic acid	2.86	3.547	24.02	2.922	2.17
201	4-methoxybenzoic acid	4.49	3.91	−12.92	4.456	−0.76
202	2-methyl-4-nitrobenzoic acid	1.86	2.746	47.63	2.968	59.57
203	2-nitrobenzene-1,4-dicarboxylic acid	1.73	1.981	14.51	1.953	12.89
204	2-phenoxybenzoic acid	3.53	3.36	−4.82	3.758	6.46
205	2,4,6-trihydroxybenzoic acid	1.68	1.793	6.73	1.793	6.73
	Prediction set					
206	3-bromophenol	9.031	8.378	−7.23	9.17	1.54
207	2,6-di-*tert*-butylphenol	11.7	11.053	−5.53	11.895	1.67
208	4-chloro-3-methylphenol	9.549	9.111	−4.59	9.593	0.46
209	*p*-cresol	10.26	9.774	−4.74	10.216	−0.43
210	2,3-dichlorophenol	7.44	8.196	10.16	7.827	5.20
211	3,5-dichlorophenol	8.179	7.873	−3.74	8.198	0.23
212	1,4-dihydroxybenzene	9.91	9.613	−3.00	10.153	2.45
213	1,4-dihydroxy-2,3,5,6-tetramethylbenzene	11.25	10.3	−8.44	10.723	−4.68

Table 3. Continued

No.	Compound	Exp.	MLR	IPD_{MLR}	ANN	IPD_{ANN}
214	2,5-dimethylphenol	10.22	10.115	−1.03	10.31	0.88
215	3,4-dinitrophenol	5.424	7.121	31.29	5.319	−1.94
216	4-ethylphenol	10.0	10.064	0.64	10.293	2.93
217	2-hydroxybenzaldehyde	8.34	9.833	17.90	8.155	−2.22
218	4-hydroxybenzonitrile	7.95	7.911	−0.49	8.166	2.72
219	4-hydroxy-3-methoxybenzaldehyde	7.396	7.974	7.82	7.896	6.76
220	3-iodophenol	8.879	8.099	−8.78	8.921	0.47
221	4-methoxyphenol	10.20	9.587	−6.01	10.282	0.80
222	4-methylsulfonylphenol	7.83	7.936	1.35	7.647	−2.34
223	4-nitrophenol	7.150	7.232	1.15	7.219	0.97
224	3-phenylphenol	9.63	8.485	−11.89	9.671	0.43
225	1,3,5-trihydroxybenzene	8.45	8.929	5.67	8.107	−4.06
226	3-acetoxybenzoic acid	4.00	3.47	−13.25	3.822	−4.45
227	anthracene-2-carboxylic acid	4.18	2.148	−48.61	4.186	0.14
228	1,2,3,5-benzenetetracarboxylic acid	2.38	1.625	−31.72	2.379	−0.04
229	2-benzoylbenzoic acid	3.54	3.223	−8.95	3.185	−10.03
230	2-bromo-6-nitrobenzoic acid	1.37	2.004	46.28	0.957	−30.15
231	2-chloro-3-nitrobenzoic acid	2.02	2.266	12.18	2.536	25.54
232	4-cyanobenzoic acid	3.55	2.619	−26.23	3.873	9.10
233	2,6-dihydroxybenzoic acid	1.30	2.864	120.31	1.084	−16.62
234	3,4-dimethylbenzoic acid	4.41	4.471	1.38	4.255	−3.51
235	3,4-dinitrobenzoic acid	2.82	2.251	−20.18	2.738	−2.91
236	2-hydroxybenzoic acid	2.98	4.091	37.28	3.313	11.17
237	2-hydroxy-4-methylbenzoic acid	3.17	4.308	35.90	3.128	−1.32
238	3-iodobenzoic acid	3.86	2.771	−28.21	3.529	−8.58
239	2-methylbenzoic acid	3.90	4.357	11.72	3.749	−3.87
240	2-methyl-6-nitrobenzoic acid	1.87	4.44	137.43	1.939	3.69
241	3-nitrobenzene-1,2-dicarboxylic acid	1.88	2.334	24.15	1.872	−0.43
242	4-phenoxybenzoic acid	4.52	3.194	−29.34	3.993	−11.66

[a]Exp. refers to the experimental values of p*K*a, MLR and ANN refer to multi-parameter linear regression and artificial neural network calculated values of p*K*a, respectively.

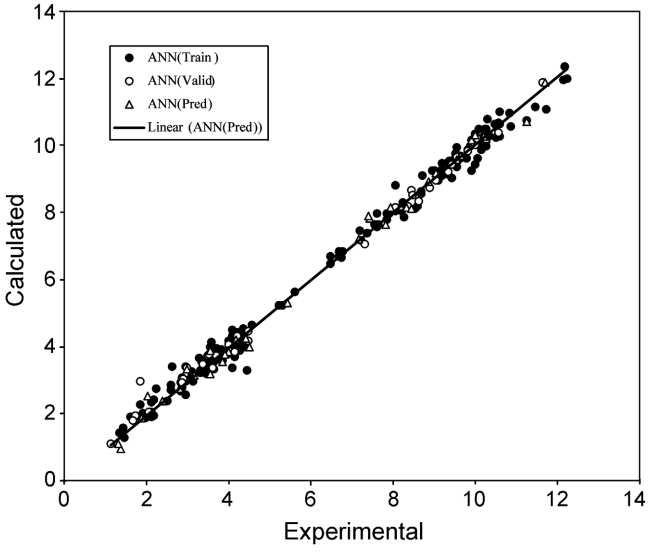

Figure 4. Plot of the calculated values of p*K*a from the ANN model *versus* the experimental values of it for training, validation and prediction sets.

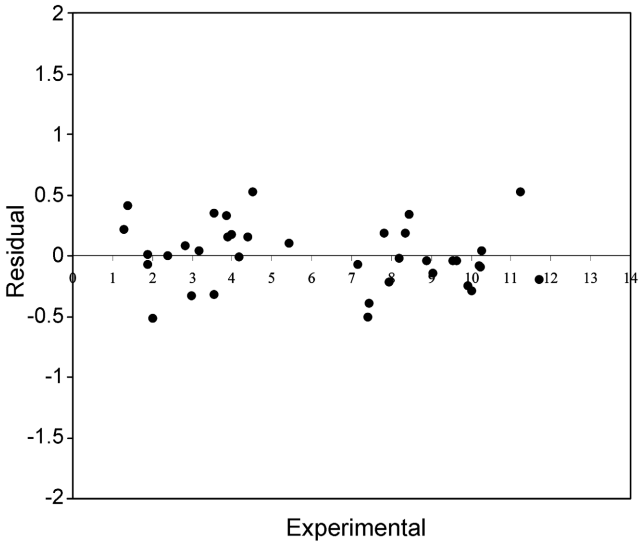

Figure 5. Plot of the residual for calculated values of p*K*a from the ANN model *versus* the experimental values of it for prediction set.

Table 4. Comparsion of statistical parameters obtained by the MLR and ANN models for correlation acidity constant of phenols and benzoic acids with the molecular descriptors[a]

Model	R^2_{tot}	R^2_{train}	R^2_{valid}	R^2_{pred}	$RMSE_{tot}$	$RMSE_{train}$	$RMSE_{valid}$	$RMSE_{pred}$
MLR	0.9266	0.9268	0.9400	0.9147	0.8610	0.8553	0.8034	0.9388
ANN	0.9931	0.9926	0.9943	0.9939	0.2648	0.2700	0.2479	0.2575

[a]Subscript train is referring to the training set, valid is referring to the validation set and pred is referring to the prediction set, tot is refering to the total data set and R is the correlation coefficient.

Conclusions

A linear and non-linear QSPR models have been developed for prediction of acidity constant (p*K*a) for various benzoic acids and phenols in water. Comparison of the values of RMSE for training, validation and prediction sets (and other statistical parameters in Table 4) for the MLR and ANN models show superiority of the nonlinear model over the regression model. Root-mean square error of 0.9388 for the prediction set by the MLR model should be compared with the value of 0.25751 for the ANN model. Since the improvement of the results obtained using nonlinear model (ANN) is considerable, it can be concluded that the nonlinear characteristics of the molecular descriptors on the p*K*a values of the compounds in water is serious.

Acknowledgements. The Authors wish to acknowledge the vice-presidency of research, university of Mohaghegh Ardebili, for financial support of this work.

References

1. Katritzky, A. R.; Karelson, M.; Lobanov, V. S. *Pure Appl. Chem.* **1997**, *69*, 245.
2. Balaban, A. T. *J. Chem. Inf. Comut. Sci.* **1997**, *37*, 645.
3. Benfenati, E.; Gini, G. *Toxicology* **1997**, *119*, 213.
4. Cronce, D. T.; Famini, G. R.; Soto, J. A. D.; Wilson, L. Y. *J. Chem. Soc., Perkin Trans. 2* **1998**, 1293.
5. Engberts, J. B. F. N.; Famini, G. R.; Perjessy, A.; Wilson, L. Y. *J. Phys. Org. Chem.* **1998**, *11*, 261.
6. Hiob, R.; Karelson, M. *J. Chem. Inf. Comut. Sci.* **2000**, *40*, 1062.
7. Habibi-Yangjeh, A. *Indian J. Chem.* **2003**, *42B*, 1478.
8. Habibi-Yangjeh, A. *Indian J. Chem.* **2004,** *43B*, 1504.
9. Nikolic, S.; Milicevic, A.; Trinajstic, N.; Juric, A. *Molecules* **2004**, *9*, 1208.
10. Devillers, J. *SAR and QSAR Environ. Res.* **2004**, *15*, 501.
11. Karelson, M.; Lobanov, V. S. *Chem. Rev.* **1996**, *96*, 1027.
12. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
13. Kramer, R. *Chemometric Techniques for Quantitative Analysis*; Marcel Dekker: New York, 1998.
14. Wold, S.; Sjöström, M. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 3.
15. Barros, A. S.; Rutledge, D. N. *Chemomet. Intell. Lab. Syst.* **1998**, *40*, 65.
16. Garkani-Nejad, Z.; Karlovits, M.; Demuth, W.; Stimpfl, T.; Vycudilik, W.; Jalali-Heravi, M.; Varmuza, K. *J. Chromatogr. A* **2004**, *1028*, 287.
17. Patterson, D. W. *Artificial Neural Networks: Theory and Applications*; Simon and Schuster: New York, 1996; Part III, Ch. 6.
18. Bose, N. K.; Liang, P. *Neural Network Fundamentals*; McGraw-Hill: New York, 1996.
19. Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinhein, 1999.
20. Agatonovic-Kustrin, S.; Beresford, R. *J. Pharm. Biomed. Anal.* **2000**, *22*, 717.
21. Fatemi, M. H. *J. Chromatogr. A* **2002**, *955*, 273.
22. Xing, W. L.; He, X. W. *Anal. Chim. Acta* **1997**, *349*, 283.
23. Bunz, A. P.; Braun, B.; Janowsky, R. *Fluid Phase Equilib.* **1999**, *158*, 367.
24. Homer, J.; Generalis, S. C.; Robson, J. H. *Phys. Chem. Chem. Phys.* **1999**, *1*, 4075.
25. Goll, E. S.; Jurs, P. C. *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 974.
26. Vendrame, R.; Braga, R. S.; Takahata, Y.; Galvao, D. S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1094.
27. Gaspelin, M.; Tusar, L.; Smid-Korbar, J.; Zupan, J.; Kristl, J. *Int. J. Pharm.* **2000**, *196*, 37.
28. Gini, G.; Cracium, M. V.; Konig, C.; Benfenati, E. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1897.
29. Urata, S.; Takada, A.; Uchimaru, T.; Chandra, A. K.; Sekiya, A. *J. Fluorine Chem.* **2002**, *116*, 163.
30. Koziol, J. *Internet Electron J. Mol. Des.* **2003**, *2*, 315.
31. Wegner, J. K.; Zell, A. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077.
32. Valkova, I.; Vracko, M.; Basak, S. C. *Anal. Chim. Acta* **2004**, *509*, 179.
33. Sebastiao, R. C. O.; Braga, J. P.; Yoshida, M. I. *Thermochimica Acta* **2004**, *412*, 107.
34. Jalali-Heravi, M.; Masoum, S.; Shahbazikhah, P. *J. Magn. Reson.* **2004**, *171*, 176.
35. Habibi-Yangjeh, A.; Nooshyar, M. *Bull. Korean Chem. Soc.* **2005**, *26*, 139.
36. Habibi-Yangjeh, A.; Nooshyar, M. *Physics and Chemistry of Liquids* **2005**, *43*, 239.
37. Selassie, C. D.; DeSoyza, T. V.; Rosario, M.; Gao, H.; Hansch, C. *Chemico-Biological Interaction* **1998**, *113*, 175.
38. Zhao, Y.-H.; Yuan, L.-H.; Wang, L.-S. *Bull. Environ. Contam. Toxicol.* **1996**, *57*, 242.
39. Hemmateenejad, B.; Sharghi, H.; Akhond, M.; Shamsipur, M. *J. Solution Chem.* **2003**, *32*, 215.
40. Gruber, C.; Buss, V. *Chemosphere* **1989**, *19*, 1595.
41. Citra, M. J. *Chemosphere* **1999**, *38*, 191.
42. Schuurmann, G. *Quant. Struct. Act. Relat.* **1996**, *15*, 121.
43. Gross, K. C.; Seybold, P. G. *Int. J. Quant. Chem.* **2001**, *85*, 569.
44. Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. C. *J. Am. Chem. Soc.* **2002**, *124*, 6421.
45. Hanai, T.; Koizumi, K.; Kinoshita, T. *J. Liq. Chromatogr. Relat. Technol.* **2000**, *23*, 363.
46. Ma, Y.; Gross, K. C.; Hollingsworth, C. A.; Seybold, P. G.; Murray, J. S. *J. Mol. Model* **2004**, *10*, 235.
47. HyperChem, Release 7.0 for Windows, *Molecular Modeling System*; Hypercube Inc.: 2002.
48. Todeschini, R.; Consonni, V.; Pavan, M. *Dragon Software*, Version 2.1; 2002.
49. Dean, J. A. *Lange's Handbook of Chemistry*, 15th Ed.; McGraw-Hill, Inc.: 1999.
50. Demuth, H.; Beale, M. *Neural Network Toolbox*; Mathworks: Natick, MA, 2000.
51. Despagne, F.; Massart, D. L. *Analyst* **1998**, *123*, 157R.
52. *Matlab 6.5*; Mathworks: 1984-2002.
53. Famini, G. R.; Wilson, L. Y. *J. Phys. Org. Chem.* **1999**, *12*, 645.