J Incl Phenom Macrocycl Chem (2012) 72:423–435
DOI 10.1007/s10847-011-0006-7

ORIGINAL ARTICLE

QSPR probing of Na^+ complexation with 15-crown-5 ethers derivatives using artificial neural network and multiple linear regression

Hiua Daraei · Mohsen Irandoust · Jahan B. Ghasemi · Ali Reza Kurdian

Received: 4 May 2011 / Accepted: 18 June 2011 / Published online: 7 July 2011

Abstract A quantitative structure–property relationship (QSPR) study is performed to develop a model, relating to Na^+ complex stability constant (log *K*) and the structure of 74 derivatives of 1,4,7,10,13-pentaoxacyclo-pentadecane ethers (15C5). Stepwise Multiple Linear Regression (SMLR) and Artificial Neural Network (ANN) methods have been exploited as linear and nonlinear methods, respectively to build the QSPR model. MOPAC software embedded in ChemOffice 2004 package was used for the minimizing energy using semi-empirical AM1 method. The optimum structures have been applied to generate more than 50 descriptors using available servers in ChemOffice 2004. The five most important constitutional, steric, electronic, thermodynamic and molecular descriptors were selected using the common preselection method combined by SMLR method. SMLR and ANN models were constructed based on the five selected descriptors. Both proposed models efficiently predict log *K* of 15C5 complexes. However, the results of ANN were more effective respect to SMLR model. This phenomenon reveals that log *K* of 15C5 complexes have a deviation from linear behavior related to the selected descriptors.

Keywords Quantitative structure–property relationship · 1,4,7,10,13-pentaoxacyclo-pentadecane ethers · Sodium ion · Stability constant · Artificial neural network

Introduction

Crown ethers are compounds with multiple oxygen heteroatoms (three or more) incorporated in a monocyclic carbon backbone. They were first synthesized by Pedersen in 1967 [1, 2]. Their generic name originates from their molecular shape, reminiscent of a royal crown [3]. Because of their selective complex formation with hard metal ions as well as their negligible water solubility, crown ethers have been extensively used as suitable ion-carriers in solvent–solvent and solid phase extractions [4–7], ion-transport [8–10], ion-selective and PVC membrane ion-selective electrode studies [11–13], and the crown ether complexes were applied as a nano-switch recently [14–16].

The complexation of these molecules with suitable well-tailored ions may be considered as trigger step accounting for this widely applications. Among plenty studied ions, alkali metal ions are under attention [17]. Therefore, predicting the log *K* values of the complexation reactions as the most important complexation property without expend the time and cost in laboratory is a motive to use the quantitative structure property relationship (QSPR) in this branch of chemistry [3].

Quantitative structure activity and structure–property relationship (QSAR/QSPR) studies are unquestionably of great importance in modern chemistry and biochemistry [18]. Currently, these methods are increasingly employed

H. Daraei
Environmental Health Research Center, Kurdistan University of Medical Sciences, Sanandaj, Iran

M. Irandoust (✉)
Department of Chemistry, Razi University, Kermanshah, Iran
e-mail: irandoust1341@yahoo.com

J. B. Ghasemi
Department of Chemistry, K. N. Toosi University of Technology, Tehran, Iran

A. R. Kurdian
Department of Chemical Engineering, Razi University, Kermanshah, Iran

in the prediction of chemical and physical properties or bioactivities of different types of molecules [3, 19–27]. In the field of complexation, these methods have potential applications to predict log K that has been developed with the help of QSPR [3].

The successful applications have inspired us to perform more exhaustive study in order to validate the applicability of traditional QSPR. Furthermore, the analog complexation reactions have been experimentally studied recently [28–30], due to its advantages authors have seriously attempted to develop QSPR in this filed [3].

Materials and methods

Dataset

The 74 studied 15C5 derivatives of chemical structures and their experimental complex stability constants values with Na^+ ion taken from the literature [17] are presented in Table 1. Since both the temperature and solvent affect the log K, we used data obtained at standard temperature (25 °C) and just in methanol solution [3].

The dataset was split into training and testing sets for SMLR study. The same train set was used to construct the ANN model. Nevertheless SMLR test dataset was split into validation and test sets for ANN. The training set of 53 complexes was used to adjust the parameters of the models. The test set of 21 complexes was used to evaluate SMLR model prediction ability. The validation set of 7 complexes was used to prevent overtrain and the test set of 14 complexes was used to evaluate ANN model prediction ability [31]. Members of each set were assigned randomly [27].

Molecular modeling, molecular optimization and descriptor generation

The structures were drawn in ChemDraw Ultra8 and exported into a file with compatible format with MOPAC program. The minimization of energy was performed on a PC computer with Intel (R) Pentium (R) Dual CPU with windows XP operating system with the semi-empirical quantum method Austin Model 1 (AM1) [32] embedded in the MOPAC program. The gradient norm criterion 0.001 kcal/mol in presence of precise keyword was applied in order to the minimize energy for all structures.

MOPAC output files were used by the ChemPropPro, ChemPropStd, CLogP, MM2, MOPAC and Topology Indices servers embedded in ChemOffice 2004 program to compute more than 40 steric, electronic, and thermodynamic descriptors for the all 74 optimized 15C5 structures. The generation of the descriptors is carried out without taking into account of the solvation of the ligands molecules. It means that the generated descriptors are carried out using the gas-phase geometry calculation of AM1.

Selection of molecular descriptors

A preselection of descriptors was carried out by eliminating those descriptors that are not available for each structure, descriptors having a small variation in magnitude for all structures and descriptors which exhibit a very small correlation with log K values combines with a SMLR method to more reduce in pool of descriptors to receive a minimum number of most important descriptors [27].

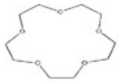
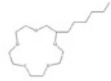
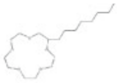
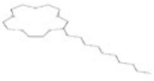
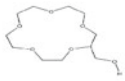
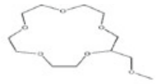
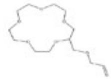
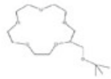
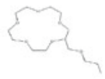
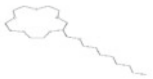
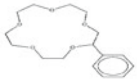
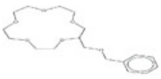
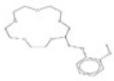
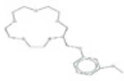
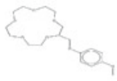
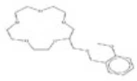
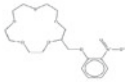
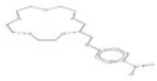
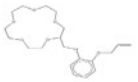
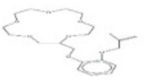
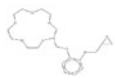
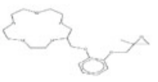
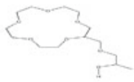
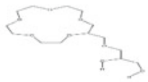
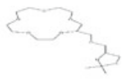
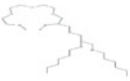
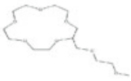
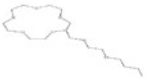
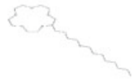
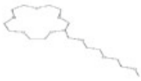
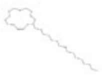
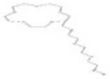
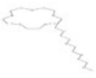
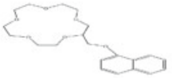
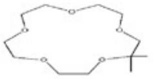
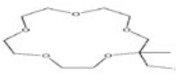
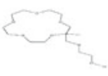
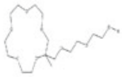
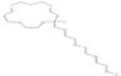
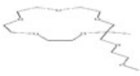
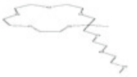
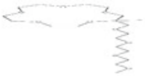
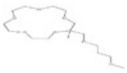
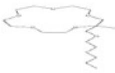
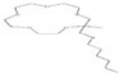
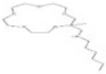
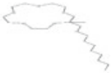
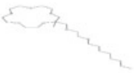
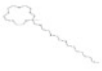
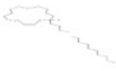
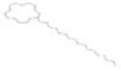
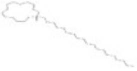
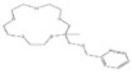
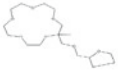
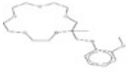
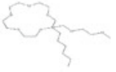
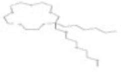
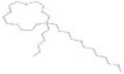
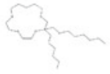
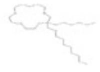
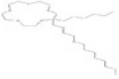
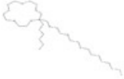
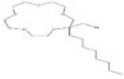
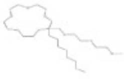
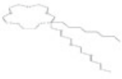
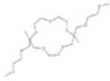
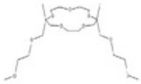
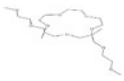
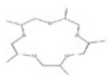
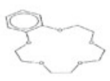
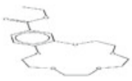
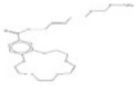
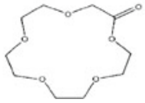
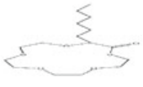
Stepwise, forward and backward SMLR are commonly used regression methods which are proposed to evaluate only a small number of subsets by either adding or deleting variables one at a time according to a specific criterion [25, 27, 33]. The forward selection method adds variables to the model one at a time. The first variable included in the model is the one which has the highest correlation with the independent variable log K. The variable that enters the model as the second variable is one which has the highest correlation with log K, after log K has been adjusted for the effect of the first variable. This process terminated when the last variable entering the model has insignificant regression coefficients or all the variables are included in the model [34]. In contrast to forward selection, backward elimination begins with the full model and successively eliminates one at a time. The first variable deleted is the one with the smallest contribution to the reduction of predictive error sum of squares (PRESS). Assuming that there are more variables that are insignificant, the process operates by eliminating the next most insignificant variable. The process is terminated when all the variables are significant or all but one variable has been deleted.

In stepwise procedure a variable that entered the model in the earlier stages of selection may be deleted at the later stages. The calculations made for inclusion and elimination of variables are the same as forward selection and backward procedures. That is, the stepwise method is essentially a forward selection procedure, but at any stages the possibility of deleting a variable, as in backward elimination, is considered. The number of variables retained in the model is based on the levels of significance assumed for inclusion and exclusion of variables from the model that is assumed 0.05 and 0.1, respectively here.

Methodology of modeling

The selected descriptors for the 53 15C5 derivatives and correspond Na^+ complex log K values were correlated by SMLR and nonlinear ANN models. The SMLR analysis, a

Table 1 chemical structure of 74 studied 15C5

 1	 2	 3	 4	 5	 6
 7	 8	 9	 10	 11	 12
 13	 14	 15	 16	 17	 18
 19	 20	 21	 22	 23	 24
 25	 26	 27	 28	 29	 30
 31	 32	 33	 34	 35	 36
 37	 38	 39	 40	 41	 42
 43	 44	 45	 46	 47	 48
 49	 50	 51	 52	 53	 54
 55	 56	 57	 58	 59	 60
 61	 62	 63	 64	 65	 66
 67	 68	 69	 70	 71	 72
 73	 74				

commonly used method in QSPR study, was employed to establish the quantitative regression models [3, 25, 35]. Equation 1 gives the mathematical representation of the linear equation that should correlate the best log K with a certain number (n) of molecular descriptors (d_i) weighted by the regression coefficients b_i:

$$\log K = b_0 + \Sigma b_i d_i \qquad I = 1, 2, \ldots, n. \tag{1}$$

Then the same data set and descriptors was used to build the nonlinear model using ANNs.

ANNs are parallel computational devices consisting of groups of highly interconnected processing elements called neurons. Neural networks are characterized by topology, computational characteristics of their elements, and training rules. Customary neural network have neurons arranged in a series of layers consist of an input, a hidden, and an output layer. The first layer is input layer that does not process the information; it only distributes the input vectors to the hidden layer. The last layer is the output layer, and its neurons handle the output from the network. The layers of neurons between the input and output layers are called hidden layers. Each layer may make its independent computations and may pass the results yet to another layer. In feed-forward neural networks the connections among neurons are directed upwards, i.e. connections are not allowed among the neurons of the same layer or the preceding layer. Networks where neurons are connected to themselves, with neurons in the same layer or neurons from a preceding layer, are termed feedback or recurrent networks [36].

Feed forward backpropagation network have been used in this study because it is very fast, easy to use and some other advantages [31]. This algorithm is multilayer feed-forward neural networks trained by backpropagation of errors (traditionally) using gradient descent with momentum weight and bias learning function. The weights of the connections between neurons being adjusted in order to decrease the mean squared error (MSE) between calculated and expected values for all train molecules in the database [37].

Data set was randomly divided into three parts. The training set was used to adjust the parameters of the models, and testing set used to evaluate its prediction ability. An important problem of neural networks is overtraining probability. An overtrained network has usually learned the pattern it has seen (training set) perfectly but cannot give accurate predictions for unseen compounds, and it is no longer able to generalize. There are several methods for avoiding this problem. One of the superior methods is to use a test set to validate the prediction power of the network during its training [22, 31].

Results and discussion

SMLR modeling

The SMLR method was used to develop the linear model for the prediction of log K using all the descriptors which remain after preselection step for training data set. The number of descriptors reduced from more than 40–30 using preselection method. The plot of the number of descriptors involved in the obtained models versus square correlation coefficient (R^2) and the cross-validated square correlation coefficient (R^2 adjusted) corresponding to those models is shown in Fig. 1. The model corresponding to the break point shows the optimum number of descriptors to be used in linear modeling that presented in Eq. 2.

$$\log K = 12.78152 - 0.08124 d_1 + 0.008912 d_2 - 0.01433 d_3 + 2.19 \times 10^{-7} d_4 + 0.000734 d_5 \tag{2}$$

As seen from Eq. 2, five descriptors model were selected in SMLR process. The log K is assumed to be highly dependent upon the stretch energy, d_1, freezing point, d_2, critical temperature, d_3, d_4, and heat of formation, d_5. The values of these five descriptors for 74 15C5 are listed in Table 2.

The linear QSPR model has been generated using a training set of 53 crown ethers. The test set of 21 crown ethers was used to assess the predictive ability of the QSPR model produced in the regression. The linear model statistical parameters have been presented in Table 3.

where b_i, standard error and t-test are the regression coefficients, standard errors of the regression coefficients and t significance, respectively.

The SMLR predicted values of log K are presented in Table 4. The stability and validity of the model was tested by prediction of the response values for the test set.

The plots of predicted log K versus experimental log K are presented in Fig. 2.

ANN modeling

After the linear model establishing, ANN was then used to develop a nonlinear model based on the SMLR same subset [27] just the test set that was used in SMLR study, split to a validation and test set in order to prevent of overtrain and to evaluate the prediction ability of the models correspondingly. The input and output data have been normalized before using to construct and to test the ANN models [36]. The inputs were normalized by the Eq. 3.

$$p_{\mathrm{st}} = \frac{p - \overline{p}}{\mathrm{std}_p} \tag{3}$$

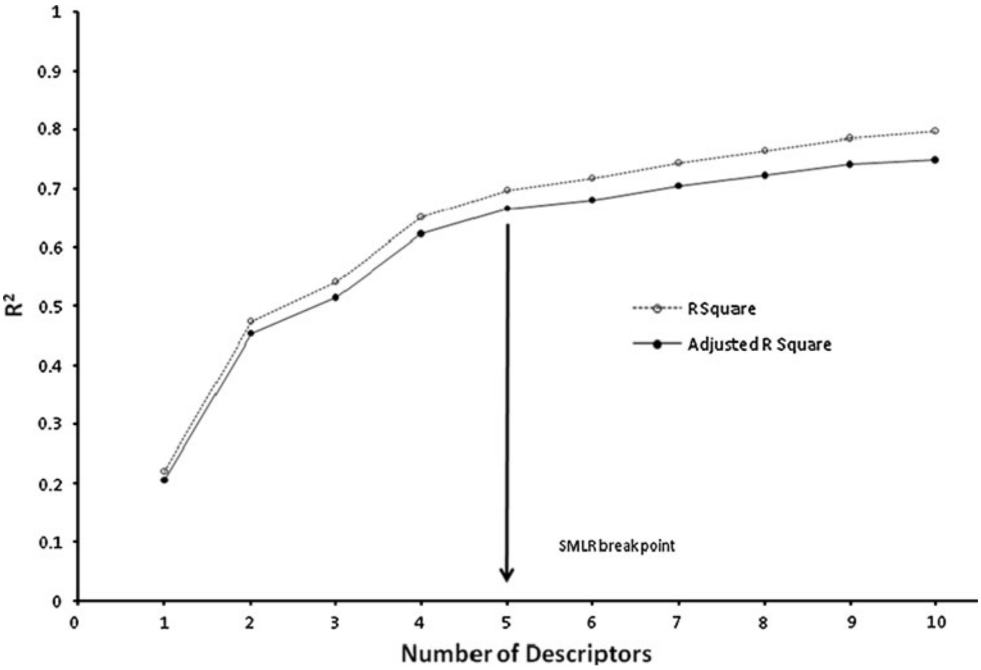


Fig. 1 Correlation and cross-validated correlation coefficients (R^2 and R^2_{CV}) versus number of descriptors

where p, $\overline{p}$, std_p and p_{st} are the input vector, mean value of the input vector, standard deviation value of the input vector and standardized value of p values, respectively. Also the output values were normalized by the Eq. 4.

$$t_{\mathrm{st}} = \frac{2(t - t_{\min})}{t_{\max} - t_{\min}} - 1 \qquad (4)$$

where t, $t_{\min}$, $t_{\max}$ and t_{st} are the output vector, minimum value of the output vector, maximum value of the output vector and standardized value of t values, respectively.

In order to obtain better results, the parameters that influence the performance of ANN models were optimized. The selection of the optimal number of hidden layer, number of hidden layer neurons, transfer function types and learning rate value for ANN was performed by systemically changing their values and types in the training step [27].

Once the network has been trained, the weights of each neuron are saved in the ANN model and could be used to predict log K for unknown compound. The parameters which determine the ANN have been listed in Table 5. The best ANN model constructed using 1 hidden layer with 13 neuron, tansig as a transfer function for both hidden layer and output layer, and 0.11 learning rate.

The tansig transfer function defines as Eq. 5.

$$a = \mathrm{tansig}(n) = \frac{2}{(1 + \exp(-2n))} - 1 \qquad (5)$$

The architecture of best obtained (5-13-1) ANN model has been showed in Fig. 3.

The predicted results of the ANN models are shown in Table 4 and Fig. 4.

Comparison between ANN and SMLR models

In this study, our goal was set to measure the predictive ability of the ANN model by comparison with SMLR method. On the basis of this test and all the other information presented here, it appears that the ANN model described here is very superior for predicting log K of complexation reaction related compounds.

The differences in the results of predictions obtained by using different models are as function of the modeling approach employed, the descriptors used and the data set of compounds [20]. The ANN model presented here obviously gives the better statistical results. A summary of the comparisons of ANN with SMLR is given in Table 4. In addition, the consistency of the ANN model as compared with SMLR method was revealed by test quantified with predictive Q^2.

The Q^2 values measure the goodness of the predictions of the held out cases exactly in the same way as R^2 does with the cases included in the modeling phase. But Q^2 is always lower and may be even negative if the predictions are worse than just using the average value of the response. The Q^2 value should be at least 0.3–0.4 in order to assess that the model has statistically significant prediction ability [20, 21, 38]. The Q^2 values of the models are calculated by the Eq. 6.

Table 2 Values of five selected descriptors

Compound	Stretch energy	Freezing point	Critical temp.	Balaban index	Heat of formation
1	1.77235	314.75	820.92	94500	−890.51
2	2.7335	378.13	846.7	483643	−1034.69
3	3.26059	400.67	859.68	781248	−1075.97
4	3.09364	423.21	875.3	1217626	−1117.25
5	1.75404	382.6	835.54	166689	−1083.72
6	2.20897	355.28	834.02	220116	−1084.35
7	1.9161	376.06	846.66	375302	−1000.2
8	4.07653	391.51	844.52	461983	−1155.02
9	1.79053	377.82	842.92	375302	−1125.63
10	1.93018	434.17	876.55	1217626	−1228.83
11	3.9742	404.55	896.7	308788	−798.16
12	2.31834	449.32	910.91	636211	−971.66
13	3.18001	484.07	917.42	736537	−1115.35
14	3.18	484.07	917.42	750817	−1115.35
15	2.97444	484.07	917.42	764818	−1115.35
16	6.77167	495.34	923.66	916940	−1135.99
17	15.0882	582.46	937.96	878633	−1123.6
18	12.692	582.46	937.96	924779	−1123.6
19	3.18	504.85	933.05	1070472	−1031.2
20	3.18002	502.16	938.61	1273034	−1061.63
21	4.68517	551.12	952.19	987340	−1215.83
22	4.68517	586.29	956.72	1160217	−1221.23
23	1.89434	423.64	852.59	476256	−1283.14
24	1.79679	484.46	884.28	601146	−1435.37
25	3.01805	495.33	898.87	770230	−1396.17
26	4.58027	542.52	973.16	5104652	−1643.03
27	1.51449	400.05	849.96	483643	−1257.85
28	1.4405	433.86	869.86	979551	−1319.77
29	1.37765	478.94	905.76	2229644	−1402.33
30	1.79368	444.82	871.22	979551	−1431.35
31	3.86445	523.71	940.85	3834334	−1575.83
32	1.85368	516.91	913.17	1501140	−1604.22
33	1.84343	489.59	897.96	1836255	−1604.85
34	7.78241	528.35	974.34	810072	−853.98
35	2.28189	356.95	827.08	162785	−936.89
36	1.45871	416.75	868.28	211351	−910.56
37	1.85848	462.54	860.57	453433	−1262.62
38	2.11662	507.31	886.93	915618	−1436.12
39	2.37519	552.08	918.94	1724095	−1609.62
40	1.67062	435.22	854.08	577553	−1263.25
41	1.92842	479.99	875.89	1139332	−1436.75
42	2.04053	524.76	903.21	2097827	−1610.25
43	1.85555	446.49	860.24	730062	−1283.89
44	2.85841	446.8	865.64	915618	−1192.95
45	3.02263	458.97	894.74	915618	−1018.86
46	2.72739	477.23	873	915618	−1007.26
47	3.30509	469.34	881.47	1406796	−1234.23
48	3.09135	514.11	911.28	2535125	−1407.73

Table 2 continued

Compound	Stretch energy	Freezing point	Critical temp.	Balaban index	Heat of formation
49	3.95961	558.88	947.05	4308159	−1581.23
50	3.47221	514.42	921.34	3043675	−1316.79
51	5.52459	559.19	959.56	5082408	−1490.29
52	4.18454	581.73	988.53	6965416	−1531.57
53	3.86589	544.76	921.78	739891	−923.63
54	0.224494	473	892	605383	−1243.83
55	3.18001	519.24	921.58	851615	−1120.75
56	0.00013	491.57	891.54	1473272	−1366.45
57	3.13313	536.34	923.11	2544206	−1539.95
58	3.44727	581.11	960.82	4246320	−1713.45
59	4.31655	503.15	910.31	2127769	−1296.15
60	4.91771	525.69	933.1	3030374	−1337.43
61	2.54798	570.46	973.62	4994873	−1510.93
62	3.7133	615.23	1021.34	7929252	−1684.43
63	3.47788	495.64	905.08	1079111	−1055.04
64	2.23428	558.88	947.05	3509669	−1581.23
65	2.98462	603.65	989.4	5645460	−1754.73
66	2.27967	555.69	904.98	2193684	−1635.99
67	2.22751	555.69	904.98	2193684	−1635.99
68	2.27884	555.69	904.98	2072861	−1635.99
69	2.48569	349.9	829.07	335358	−1095.41
70	3.92495	405.81	877.87	187912	−735.69
71	9.89642	494.47	907.47	571901	−1126.36
72	17.4701	696.4	1082.5	7720075	−1770.7
73	2.44724	337.55	880.95	125610	−1063.47
74	4.18525	400.93	903.41	580491	−1207.65

Table 3 Statistical parameters of the best MLR model

i	b_i	St. error	*t*-Test	Descriptor
0	12.78152	1.88343	6.786298	Intercept
1	−0.08124	0.015184	−5.34987	Stretch energy
2	0.008912	0.001342	6.643167	Freezing point
3	−0.01433	0.002488	−5.75913	Critical temperature
4	2.19E−07	4.52E−08	4.857182	Balaban index
5	0.000734	0.000272	2.697069	Heat of formation

$$Q^2 = 1 - \frac{\sum_{i=1}^{N} \left(K_{\mathrm{exp}}^{i} - K_{\mathrm{pre}}^{i}\right)^2}{\sum_{i=1}^{N} \left(K_{\mathrm{exp}}^{i} - K_{\mathrm{exp}}^{\mathrm{mean}}\right)^2} \tag{6}$$

where K_{exp} and K_{pre} are the experimental and the predicted K, respectively, $K_{\mathrm{exp}}^{\mathrm{mean}}$ is average of experimental K. Another validation analysis of the comparison of ANN with other conventional methods is RMSE (Root-Mean-Square Error) as an indicator of reliability or accuracy of the models. RMSE is computed on the basis that the data fit the model, and that all misfits in the data are merely a reflection of the stochastic nature of the model [20]. RMSE values of the models are calculated by the Eq. 7.

$$\mathrm{RMSE} = \left[\frac{\sum_{i=1}^{N} \left(K_{\mathrm{exp}}^{i} - K_{\mathrm{pre}}^{i}\right)^2}{N}\right]^{0.5} \tag{7}$$

Table 4 Comparison of ANN and MLR models

Data set	Compound	Exp. K_f	Predicted values	
			ANN	MLR
Train	1	3.23	3.21	3.05
	2	3.20	3.22	3.14
	3	3.18	3.17	3.15
	6	2.99	3.12	3.07
	8	2.95	2.86	3.09
	10	3.18	3.25	3.30
	11	3.30	3.28	2.70
	14	2.89	3.11	3.04
	16	3.04	2.96	2.78
	17	2.83	2.82	2.68
	18	2.72	2.74	2.88
	19	3.07	3.11	3.13
	20	3.04	2.87	3.05
	21	3.03	2.99	2.99
	22	3.02	3.06	3.28
	23	3.14	3.34	3.35
	24	3.00	3.13	3.36
	25	3.03	3.09	3.22
	26	2.97	3.00	3.21
	28	3.09	3.19	3.31
	30	3.16	2.94	3.28
	31	3.23	3.18	3.34
	32	3.04	3.03	3.30
	33	3.09	3.15	3.35
	34	2.74	2.75	2.45
	38	3.88	3.74	3.57
	39	3.73	3.68	3.54
	40	3.87	3.66	3.49
	45	3.08	2.97	3.26
	46	3.08	3.06	3.77
	47	3.54	3.47	3.47
	48	3.75	3.88	3.58
	49	3.88	3.92	3.66
	50	3.42	3.39	3.58
	51	3.75	3.75	3.59
	52	3.89	3.81	3.87
	53	3.58	3.49	3.60
	54	4.02	3.77	3.42
	56	3.90	3.97	3.71
	59	3.56	3.59	3.39
	60	3.39	3.40	3.38
	61	3.62	3.65	3.69
	62	3.75	3.79	3.83
	63	2.79	3.04	3.41
	65	3.75	3.72	3.69
	66	3.89	4.09	3.86
	67	4.22	4.09	3.87
	68	4.15	4.08	3.84
	69	3.34	3.29	3.09
	70	3.05	3.09	3.00
	71	2.49	2.51	2.68
	72	2.47	2.51	2.45
	73	1.98	1.96	2.22
Validation	9	3.05	3.10	3.18
	15	3.00	3.14	3.06
	27	3.05	2.93	3.23
	36	2.86	2.79	3.31
	41	3.89	3.81	3.55
	57	3.91	3.81	3.51
	58	3.71	3.68	3.59
Test	4	3.18	3.14	3.21
	5	2.94	3.09	3.32
	7	3.12	3.10	3.19
	12	2.97	2.96	2.97
	13	3.24	3.12	3.03
	29	3.22	3.17	3.42
	35	2.99	3.06	3.27
	37	3.88	3.87	3.59
	42	3.87	3.85	3.63
	43	3.48	3.72	3.50
	44	3.57	3.40	3.45
	55	3.79	3.49	3.31
	64	3.86	3.58	3.62
	74	1.48	1.64	2.31
Train		Q^2	0.948	0.697
		RMSE	0.1033	0.2516
Test		Q^2	0.933	0.684
		RMSE	0.1539	0.3080
All data		Q^2	0.945	0.692
		RMSE	0.1136	0.2689

Between the two approaches, ANN outperformed SMLR significantly according to Table 4.

The statistical significance of the relationship between the K_f and chemical structure descriptors was further demonstrated by y-randomization procedure. Y-randomization is a tool used in validation of QSPR/QSAR models, whereby the performance of the original model in data description (R^2) is compared to that of models built for permuted (randomly shuffled) response, based on the original descriptor pool and the original model building procedure. The test was done by (1) repeatedly permuting the K_f values of the data set, (2) using the permuted values to generate QSPR models and (3) comparing the resulting scores with the score of the original QSPR model generated from non-randomized K_f values. If the original QSPR model is statistically significant, its score should be significantly better than those from permuted data. The mean of R^2 values for 100 trials based on permuted data is smaller than 0.08 that significantly is different from the R^2 of the models by both SMLR and ANN.

Furthermore, in order to avoid uncertainties related to a selection of single external test set, a more severe 74

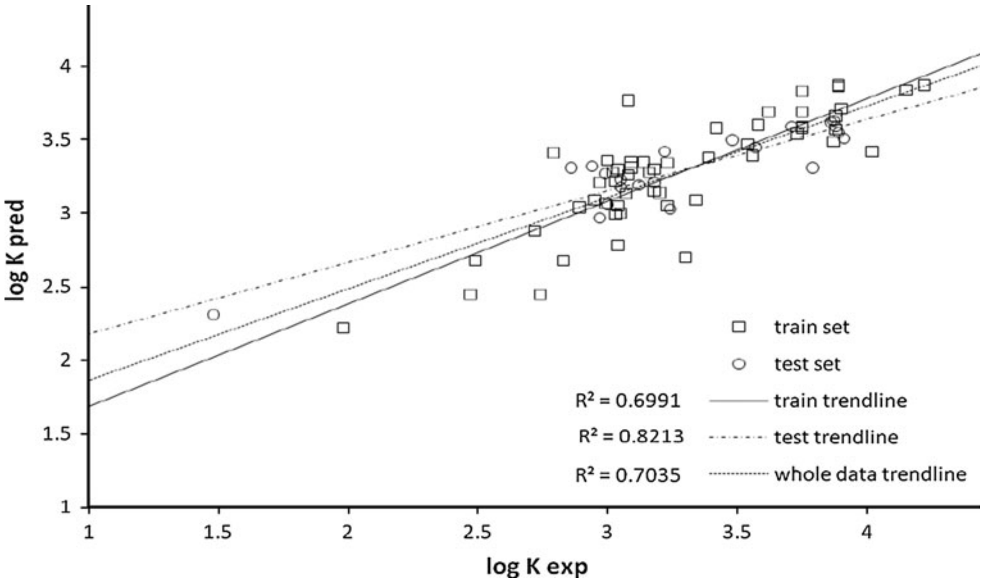


Fig. 2 Scatter plot of the MLR predicted values versus experimental log *K* values

Table 5 The (5-13-1) ANN parameters

Hidden layer parameters							Output layer parameters	
							b^2 (bias)	−0.25392
Neuron	IW (weights of hidden layer)					b^1 (bias)	W (weights of output layer)	
	i_1	i_2	i_3	i_4	i_5		Neuron	Weight
1	−0.85325	1.5273	−0.76228	−1.2556	−0.42823	2.3551	1	−0.29184
2	1.3401	0.23753	1.4117	−0.83827	−0.6729	−2.0565	2	−0.16412
3	1.6861	−1.4562	1.2276	−0.95571	0.60211	−0.36929	3	0.99278
4	0.15772	1.9317	−2.1238	0.93127	2.3632	0.80242	4	1.7665
5	−1.3769	−0.43189	0.28179	−1.7616	0.26133	0.70157	5	−0.21736
6	0.54349	1.1665	0.083291	−1.2267	1.3664	−0.1965	6	−2.112
7	−2.608	−0.64044	1.3225	−0.61568	−1.4887	0.51709	7	1.0057
8	0.15037	−2.3493	−0.58666	2.2313	1.9384	0.86537	8	−1.3618
9	−0.42002	−1.1255	−0.58772	−0.80233	2.7154	1.2855	9	−1.0982
10	0.83224	−0.54429	−1.265	−1.7138	0.74057	0.3036	10	1.4517
11	−0.74453	−1.9348	0.75185	−0.98055	−0.97021	−1.8301	11	−1.0545
12	0.13032	1.1722	−0.60931	−1.5931	−1.133	1.8688	12	−0.56647
13	2.161	0.76981	0.74113	0.35639	−2.2778	2.8179	13	−0.88038

(Leave-one-out), 5 (leave 20% out) and 3 (leave 33% out) -fold cross validation have been used to verify the ANN models predictability.

In (*n*)-fold cross-validation, the entire dataset is randomly split into *n* approximately equal size subsets. The model will then be trained and tested *n* times. At each time, one of the *n* subsets is used as the test set and the other ($n - 1$) subsets are put together to form a training set. The benefit *n*-fold cross validation is that it is not important how the data are divided. Every data point appears in a test set only once, and appears in a training set ($n - 1$) times. The overall accuracy of the built model is then just the simple average of the *n* individual accuracy measures [19]. The averages of Q^2 for 74, 5 and 3 fold cross validation are 0.977, 0.735 and 0.682, respectively. Based on Q^2 definition presented in "Comparison between ANN and SMLR models" section these results clearly proved the good predictability of ANN model like external test.

The selected descriptors

The descriptors involved in the QSPR model are: (i) stretch energy (d_1), (ii) freezing point (d_2), (iii) critical temperature (d_3), (iv) Balaban index (d_4), (v) heat of formation (d_5).

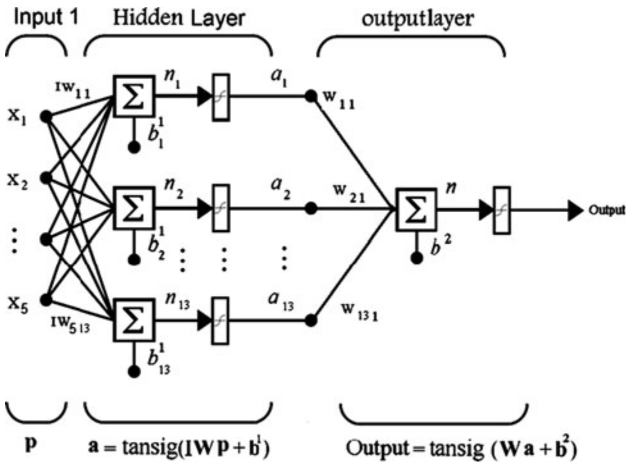


Fig. 3 The architecture of (5-13-1) backpropagation ANN

The first selected significant descriptor (d_1) involved in the Eq. 2 is stretch energy. It represents the energy associated with distorting bonds from their optimal length. Defined as in Eq. 8:

$$E_{\text{stretch}} = 71.94 \sum_{\text{Bonds}} K_{\text{s}}(r - r_0)^2 \tag{8}$$

The bond stretching energy equation is based on Hooke's law. The K_{s} parameter controls the stiffness of the spring's stretching (bond stretching force constant), while r_0 defines its equilibrium length. Unique K_{s} and r_0 parameters are assigned to each pair of bonded atoms based on their atom types (C–C, C–H, and O–C). The parameters are stored in the Bond Stretching parameter table. The constant, 71.94, is a conversion factor to obtain the final units as kcal/mole. The result of this equation is the energy contribution associated with the deformation of a bond from its equilibrium bond length. In addition, this descriptor has important role in some previous QSPR models [39].

The second and third significant descriptor (d_2 and d_3) involved in the Eq. 2 are freezing point and critical temperature [40]. Freezing points are commonly assumed to be the phase transition when the pressure is 1 atm. A more exact terminology for these temperatures might be the ''normal'' freezing points.

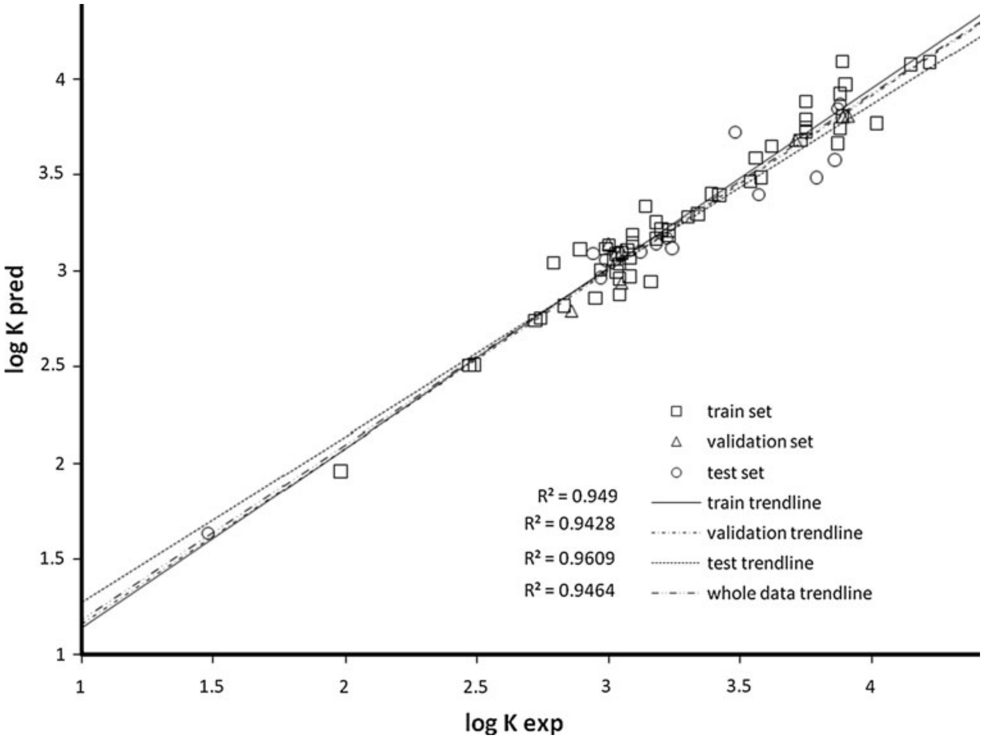
In addition, Vapor–liquid critical temperature (Tc), pressure (Pc), and volume (Vc), are the purecomponent constants of greatest interest. Number of methods to estimate the normal boiling point and critical properties has been proposed in the thermodynamic references [41].

The fourth descriptor is Balaban index that is rather unknown descriptor among five selected descriptors [42, 43]. It's a topological descriptor that defined as in Eq. 9.

$$\text{Balaban index} = \frac{q}{\mu + 1} \sum_{\text{edges}ij} (S_i S_j)^{-0.5} \tag{9}$$

where q is number of edges in the molecular graph, $\mu = (q - n + 1)$ is the cyclomatic number of the molecular graph, n is number of atoms in the molecular graph and S_i is distance sums calculated as the sums over the rows or columns of the topological distance matrix of the molecule.

Fig. 4 Scatter plot of the ANN predicted values versus experimental log K values



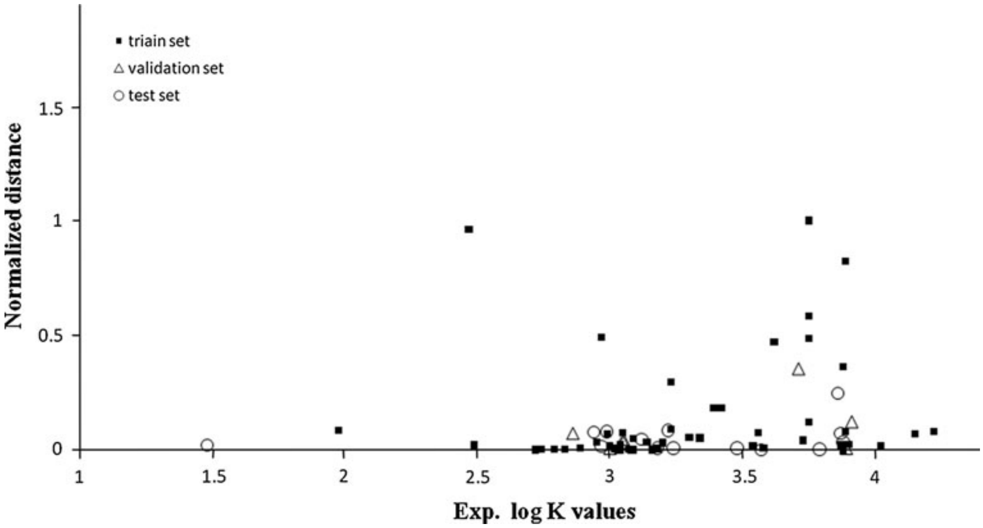


Fig. 5 Normalized diversity values versus experimental log *K* values

The final descriptor is heat of formation, d_5. This energy value represents the heat of formation for a molecule [44]. The heat of formation in MOPAC is the gas-phase heat of formation at 298 K of one mole of a compound from its elements in their standard state. The heat of formation is composed of the following terms $\Delta H_f = E_{\mathrm{elec}} + E_{\mathrm{nucl}} + E_{\mathrm{isol}} + E_{\mathrm{atoms}}$where E_{elec} is calculated from the SCF calculation, E_{nucl} is the core–core repulsion based on the nuclei in the molecule, E_{isol} and E_{atoms} are parameters supplied by the potential function for the elements within your molecule.

Molecular diversity validation

Two fundamental research themes in chemical database analysis are similarity and diversity sampling [27]. The diversity problem involves defining a diverse subset of 'representative' compounds so that researchers can scan only a subset of the huge database each time. In this study, diversity analysis was performed for the data set to make sure the structures of the training or test cases can represent those of the whole ones.

We consider a database of n compounds generated from m highly correlated chemical descriptors. Each compound X_i is represented as a vector:

$$X_i = (x_{i1}, x_{i2,}x_{i3}, \ldots, x_{im})^T \quad \text{for compound with } i = 1, 2, \ldots, n.$$

where x_{ij} denotes the value of descriptor j belongs compound X_i. The collective database X is represented by the $n \times m$ matrix X:

$$X = (X_1, X_2, X_3, \ldots, X_n)^T;$$

Here the superscript T denotes the vector/matrix transpose. A distance score for two different compounds X_1 and X_2 can be measured by the Euclidean the mean distances of one sample to the remaining ones were computed as follow:

$$d_{12} = X_1 - X_2 = \sqrt{\sum_{k=1}^{m} (x_{1k} - x_{2k})^2}$$

Distance norm based on the compound descriptors:

$$\bar{d}_1 = \frac{\sum_{i=1}^{n} d_{1i}}{n-1}$$

And then the mean distances of all compounds were normalized within the interval [0, 1]. The closer to one the distance is the more diverse to each other the compound is. For the data sets, the normalized mean distances of samples versus experimental log K are shown in Fig. 5, which illuminates the diversity of the molecules in the training and test/validation sets. As can be seen from the Fig. 5, the structures of the compounds are diverse in both sets. The training set with a broad representation of the chemistry space was adequate to ensure the model's stability and the diversity of test set can prove the predictive capability of the model.

References

1. Pedersen, C.J.: Cyclic polyethers and their complexes with metal salts. J. Am. Chem. Soc. **89**, 2495–2496 (1967)
2. Pedersen, C.J.: Cyclic polyethers and their complexes with metal salts. J. Am. Chem. Soc. **89**, 7017–7036 (1967)
3. Ghasemi, J.B., Saaidpour, S.: QSPR modeling of stability constants of diverse 15-crown-5 ethers complexes using best multiple linear regression. J. Inclusion Phenom.Macrocyclic Chem **60**, 339–351 (2008)
4. Lee, M., Oh, S.Y., Pathak, T.S., Paeng, I.R., Cho, B.Y., Paeng, K.J.: Selective solid-phase extraction of catecholamines by the

chemically modified polymeric adsorbents with crown ether. J. Chromatogr. A **1160**, 340–344 (2007)

5. Costero, A.M., Sanchis, J., Peransi, S., Gil, S., Sanz, V., Domenech, A.: Bis(crown ethers) derived from biphenyl: extraction and electrochemical properties. Tetrahedron **60**, 4683–4691 (2004)
6. Kijak, A.M., James, A.: Self-assembled monolayers of crown ethers for solid phase extraction in flow-injection analysis. Anal. Chim. Acta **489**, 13–19 (2003)
7. Yun, L.: High extraction efficiency solid-phase microextraction fibers coated with open crown ether stationary phase using sol–gel technique. Anal. Chim. Acta **486**, 63–72 (2003)
8. Gherrou, A., Kerdjoudj, H.: Specific membrane transport of silver and copper as $Ag(CN)_3^{2-}$ and $Cu(CN)_4^{3-}$ ions through a supported liquid membrane using K^+-crown ether as a carrier. Desalination **151**, 87–94 (2002)
9. Chauhan, B.S., Boudjouk, P.: New neutral carrier-type ion sensors. Crown ether derivatives of poly(methylhydrosiloxane). Tetrahedron Lett. **40**, 4123–4126 (1999)
10. Aghaie, H., Giahi, M., Monajjemi, M., Arvand, M., Nafissi, G.H., Aghaie, M.: Tin(II)-selective membrane potentiometric sensor using a crown ether as neutral carrier. Sens. Actuators B **107**, 756–761 (2005)
11. Mahajan, R.K., Kumar, M., Sharma (nee Bhalla), V.: Erratum to ''Cesium ion selective electrode based on calix[4]crown ether/ester''. Talanta **58**, 445–450 (2002)
12. Su, C.C., Chang, M.C., Liu, L.K.: New Ag^+ and Pb2C-selective electrodes with lariat crown ethers as ionophores. Anal. Chim. Acta **432**, 261–267 (2001)
13. Gupta, V.K., Pal, M.K., Singh, A.K.: Comparative study of Ag(I) selective poly(vinyl chloride) membrane sensors based on newly developed Schiff-base lariat ethers derived from 4,13-diaza-18-crown-6. Anal. Chim. Acta **631**, 161–169 (2009)
14. Gromov, S., Alfimov, M.: Supramolecular organic photochemistry of crown-ether-containing styryl dyes. Russ. Chem. Bull. **46**, 611–636 (1997)
15. Takeshita, M., Soong, C., Irie, M.: Alkali metal ion effect on the photochromism of 1,2-bis(2,4-dimethylthien-3-yl)-perfluorocyclopentene having benzo-15-crown-5 moieties. Tetrahedron Lett. **39**, 7717–7720 (1998)
16. Kawai, S.: Photochromic bis(monoaza-crown ether)s. Alkali-metal cation complexing properties of novel diarylethenes. Tetrahedron Lett. **39**, 4445–4448 (1998)
17. Izalt, R.M., Pawlak, K., Bradshaw, J.S.: Thermodynamic and kinetic data for macrocycle interaction with cations and anions. Chem. Rev. **91**, 1721–2085 (1991)
18. Roberts, D.W., Marshall, S.J.: Application of hydrophobicity parameters to prediction of the acute toxicity of commercial surfactant mixtures. SAR QSAR Environ. Res. **4**, 167–176 (1995)
19. Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P.A., Markopoulos, J., Igglessi-Markopoulou, O.: Prediction of intrinsic viscosity in polymer–solvent combinations using a QSPR model. Polymer **47**, 3240–3248 (2006)
20. Buyukbingol, E., Sisman, A., Akyildiz, M., Alparslan, F.N., Adejare, A.: Adaptive neuro-fuzzy inference system (ANFIS): a new approach to predictive modeling in QSAR applications: A study ofbneuro-fuzzy modeling of PCP-based NMDA receptor antagonists. Bioorg. Med. Chem. **15**, 4265–4282 (2007)
21. Fassihi, A., Abedi, D., Saghaie, L., Sabet, R., Fazeli, H., Bostaki, G., Deilami, O., Sadinpour, H.: Synthesis, antimicrobial evaluation and QSAR study of some 3-hydroxypyridine-4-one and 3-hydroxypyran-4-one derivatives. Eur. J. Med. Chem. **44**, 2145–2157 (2009)
22. Yao, X., Wang, Y., Zhang, X., Zhang, R., Liu, M., Hua, Z., Fan, B.: Radial basis function neural network-based QSPR for the prediction of critical temperature. Chemom. Intell. Lab. Syst. **62**, 217–225 (2002)
23. Kardanpour, Z., Hemmateenejad, B., Khayamian, T.: Wavelet neural network-based QSPR for prediction of critical micelle concentration of Gemini surfactants. Anal. Chim. Acta **531**, 285–291 (2005)
24. Turner, J.V., Glass, B.D., Agatonovic-Kustrin, S.: Prediction of drug bioavailability based on molecular structure. Anal. Chim. Acta **485**, 89–102 (2003)
25. Xu, J., Liang, H., Chen, B., Xu, W., Shen, X., Liu, H.: Linear and nonlinear QSPR models to predict refractive indices of polymers from cyclic dimer structures. Chemom. Intell. Lab. Syst. **92**, 152–156 (2008)
26. Mercader, A.G., Duchowicz, P.R., Sanservino, M.A., Fernández, F.M., Castro, E.A.: QSPR analysis of fluorophilicity for organic compounds. J. Fluorine Chem **128**, 484–492 (2007)
27. Luan, F., Liu, H., Gao, Y., Li, Q., Zhang, X., Guo, Y.: Prediction of hydrophile–lipophile balance values of anionic surfactants using a quantitative structure–property relationship. J. Colloid Interface Sci. **336**, 773–779 (2009)
28. Irandoust, M., Shamsipur, M., Daraei, H.: Proton NMR study of the stoichiometry, stability and thermodynamics of complexation of Rb^+ ion with 18-crown-6 in binary dimethylsulfoxide–nitrobenzene mixtures. J. Inclusion Phenom.Macrocyclic Chem **66**, 365–370 (2010)
29. Shamsipur, M., Irandoust, M., Alizadeh, K., Lippolis, V.: Proton NMR study of the stoichiometry, stability and thermodynamics of complexation of Ag^+ ion with octathia-24-crown-8 in binary dimethylsulfoxide–nitrobenzene mixtures. J. Incl. Phenom. Macrocycl. Chem. **59**, 203–209 (2007)
30. Shamsipur, M., Irandoust, M.: A proton NMR study of the stoichiometry and stability of 18-crown-6 complexes with K^+, Rb^+ and Tl^+ ions in binary dimethyl sulfoxide-nitrobenzene mixtures. J. Solution Chem **37**, 657–664 (2008)
31. Niculescu, S.P.: Artificial neural networks and genetic algorithms in QSAR. J. Mol. Struct. THEOCHEM **622**, 71–83 (2003)
32. Dewar, M.J.S.: J. Am. Chem. Soc. **107**, 3902–3909 (1985)
33. Moon, T., Chi, M.W., Choi, M.J., Yoon, C.N.: Quantitative structure–polarization relationships (QSPR) study of BTEX tracers for the formation of antibody–BTEX–EDF complex. Bioorg. Med. Chem. Lett. **14**, 3461–3466 (2004)
34. Chatterjee, S., Price, B.: Regression analysis by example. Wiley, New York (1977)
35. Ghasemi, J., Saaidpour, S.: Quantitative structure–property relationship study of n-octanol–water partition coefficients of some of diverse drugs using multiple linear regression. Anal. Chim. Acta **604**, 99–106 (2007)
36. Hagan, T., Demuth, H.B.: Neural network design. PWS Publishing Company, Boston, MA (1996)
37. Jorjani, E., Chelgani, S.C., Mesroghli, S.: Application of artificial neural networks to predict chemical desulfurization of Tabas coal. Fuel **87**, 2727–2734 (2008)
38. Jalali-Heravi, M., Fatemi, M.H.: Prediction of thermal conductivity detection response factors using an artificial neural network. J. Chromatogr. A **897**, 227–235 (2000)
39. Chung, W.K., Hou, Y., Holstein, M., Freed, A., Makhatadze, G.I., Cramer, S.M.: Investigation of protein binding affinity in multimodal chromatographic systems using a homologous protein library. J. Chromatogr. A **1217**, 191–198 (2010)
40. Zaier, I., Shu, C., Ouarda, T.B.M.J., Seidou, O., Chebana, F.: Estimation of ice thickness on lakes using artificial neural network ensembles. J. Hydrol. **383**, 330–340 (2010)

41. Poling, B.E., Prausnitz, J.M., O'Connell, J.P.: The properties of gases and liquids, 5th edn. The McGraw-Hill Companies, New York (1997)
42. Balaban, A.T.: Chem. Phys. Lett. **89**, 399–404 (1982)
43. Balaban, A.T.: Pure Appl. Chem. **55**, 199–206 (1983)
44. Yang, P., Chen, J., Chen, S., Yuan, X., Schramm, K., Kettrup, A.: QSPR models for physicochemical properties of polychlorinated diphenyl ethers. Sci. Total Environ. **305**, 65–76 (2003)