

UDC 539.19; 541.66

**PREDICTION OF WATER-TO-POLYDIMETHYLSILOXANE PARTITION COEFFICIENT FOR SOME ORGANIC COMPOUNDS USING QSPR APPROACHES**© 2010 **H. Golmohammadi**<sup>1\*</sup>, **Z. Dashtbozorgi**<sup>2</sup><sup>1</sup>*Department of Chemistry, Mazandaran University, Babolsar, Iran*<sup>2</sup>*Department of Chemistry, Islamic Azad University, Science and Research Branch, Tehran, Iran**Received December, 11, 2009*

A Quantitative Structure — Property Relationship (QSPR) model based on Genetic Algorithm (GA), Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) techniques was developed for the prediction of water-to-polydimethylsiloxane partition coefficients ( $\log K_{\text{PDMS-water}}$ ) of 139 organic compounds. A suitable set of molecular descriptors was calculated and important descriptors were selected by genetic algorithm and stepwise multiple regression. These descriptors were: Minimum Atomic Orbital Electronic Population ( $P_{\mu\mu}$ ), Kier Shape Index (order 3) ( $^3\kappa$ ), Polarity Parameter / Square Distance (PP), and Complementary Information Content (order 2) ( $^2\text{CIC}$ ). In order to find a better way to depict the nonlinear nature of the relationships, these descriptors were used as inputs for a generated ANN. The root mean square errors for the neural network calculated  $\log K_{\text{PDMS-water}}$  of training, test, and validation sets were 0.116, 0.179, and 0.183, respectively, which are smaller than those obtained by MLR model (0.422, 0.425, and 0.480, respectively). The results obtained showed the ability of developed artificial neural network to predict water-to-polydimethylsiloxane partition coefficients of various organic compounds. Also, the results revealed the superiority of the artificial neural network over the multiple linear regression model.

**Keywords:** quantitative structure-property relationship, water-to-polydimethylsiloxane partition coefficient, artificial neural network, multiple linear regression, genetic algorithm.

**INTRODUCTION**

Partition coefficients of organic compounds are of remarkable importance because their bioavailability as well as residual concentrations in atmosphere, water and soil strictly depends on partition properties [ 1, 2 ]. When an apolar substance (polydimethylsiloxane, PDMS) is selected as extracting phase vs. water, the partition coefficient ( $K_{\text{PDMS-water}}$ ) of an organic compound can be defined as its index of hydrophobicity. Recently, solid-phase microextraction (SPME) method was applied to measure water-to-polydimethylsiloxane partition coefficients. SPME is a versatile analytical technique developed by Pawliszyn and coworkers [ 3, 4 ] that combines sampling and sample preparation into a single step. Successful implementation of a feasible SPME-based method is strongly dependent upon an accurate determination of  $K_{\text{PDMS-water}}$  values between the SPME sorbent phase and sample matrix. When an aqueous sample of volume  $V_W$  is equilibrated with a volume  $V_{\text{PDMS}}$  of extraction phase, continually renewed each time, the following formula can be applied:

$$K_{\text{PDMS-water}} = \left[ \left( \frac{C_i}{C_{i+1}} \right) - 1 \right] \cdot \frac{V_W}{V_{\text{PDMS}}} \text{ (isothermal),} \quad (1)$$

where  $C_i$  and  $C_{i+1}$  are the analyte equilibrium concentrations in the aqueous phase for two consecutive

---

\* E-mail: hassan.gol@gmail.com

extractions performed on the same (unchanged) sample aliquot. The integer value  $i$  indicates how many identical batch equilibrium extractions have been carried out. Of course,  $C_i$  is always greater than  $C_{i+1}$ , and in theory their ratio is a constant.

There are some experimental methods to determine air/water-to-polydimethylsiloxane partition coefficients of organic compounds [ 5—10 ] but these methods are time-consuming and require high-purity samples and skilled operators, so the development of an alternative method such as Quantitative Structure — Property Relationship (QSPR) would be useful for the theoretical calculation of  $\log K_{\text{PDMS-water}}$  values. QSPR methods represent an attempt to correlate structural descriptors of molecules with their desired chemical properties and/or activities. The advantages of this approach lie in the fact that it requires only the knowledge of chemical structure and is not dependent on any experiment properties. QSPR studies can be used for the selection of principal structural characteristics (descriptors), finding their relation to property values and the derivation of mathematical models that involve these multivariate data in order to be applied for predictive purposes in every chemical system.

A number of attempts to model the relationship between partition coefficients and the property of organic compounds have been performed. Klopman et al. [ 11 ] developed a relationship between the structure of diverse organic compounds and gas-to-olive oil partition coefficient. Lu et al. [ 12 ] predicted octanol—water partition coefficients of 133 polychlorinated biphenyls using heuristic method (HM) implemented in CODESSA. Puzyn and Falandysz [ 13 ] estimated octanol—water and octanol—air partition coefficients of 75 chloronaphthalene congeners by means of six chemometrics approaches. Ohlenbusch et al. [ 14 ] investigated the sorption of various phenols to Aldrich-HA and BSA by a QSAR study based on a linear free energy relationship (LFER) model. Metivier-Pignon et al. [ 15 ] used QSPRs to determine the structural features that influence most adsorption processes of 22 commercial dyes onto activated carbon cloths using multiple linear regressions method. Sprunger et al. [ 16 ] predicted water-to-PDMS and gas-to-PDMS partition coefficients for various gaseous and organic solutes. They also correlated the partition coefficients for solute transfer to 1,2-dichloroethane from both water gas phase using Abraham solvation parameter model [ 17 ]. Hierleman et al. [ 18 ] examined the performance of the Abraham linear free energy relationship to describe the sorption coefficients of organic vapors on thickness-shear-mode resonators coated with different polymers.

Recently Artificial Neural Networks (ANNs) have been used for investigation of a wide variety of chemical problems such as spectral analysis [ 19 ], prediction of dielectric constants [ 20 ] and mass spectral search [ 21 ]. ANNs have been applied to QSPR analysis since the late 1980s due to their flexibility in the modeling of nonlinear problem, mainly in response to increased accuracy demands. They have been widely used to predict many physicochemical properties [ 22—26 ]. In this investigation, the descriptors calculated from structures were utilized as the only source to predict the water-to-PDMS partition coefficients of 139 organic compounds using the ANN and QSPR methods.

## METHODS

**Data set.** The water-to-polydimethylsiloxane partition coefficients ( $\log K_{\text{PDMS-water}}$ ) of 139 organic compounds taken from [ 16 ] were used as a data set. The list of molecules of the data set including hydrocarbons, alkylhalides, alcohols, ethers, esters, ketones, nitriles, halobenzenes, polycyclic aromatic hydrocarbons, heterocyclic compounds and benzene derivatives are given in Table 1. The partition coefficients fall in the range between 0.12 and 7.48 values for phenethyl alcohol and tetradecane, respectively. The data set was randomly divided into three separate sections: the training, test, and external validation sets, consisting of 93, 23, and 23 members, respectively. The training and test sets were used to build and optimize the QSPR model and the external validation set was used to evaluate the prediction power of the obtained model.

**Molecular descriptors generation.** Molecular descriptors are mathematical values that describe the structure or shape of molecules and help to predict the activity and properties of molecules in complex experiments [ 27 ]. A wide variety of descriptors have been reported for using in QSPR/QSAR analysis [ 28—34 ]. Due to the diversity of the molecules studied, various descriptors were calculated. The calculation procedure for obtaining the molecular descriptors was as follows. First, all molecules

Table 1

The data set and the corresponding experimental, and MLR and ANN predicted values of  $\log K_{\text{PDMS-water}}$

<i>N</i>	Name	Exp	MLR	ANN	<i>N</i>	Name	Exp	MLR	ANN
1	2	3	4	5	6	7	8	9	10
Training set									
1	Methane	1.16	1.42	1.24	45	1,2,4-Trichlorobenzene	3.48	3.07	3.32
2	Propane	2.32	2.37	2.29	46	1,2,3,4-Tetrachlorobenzene	3.90	3.69	3.79
3	2-Methylpropane	2.88	2.79	2.69	47	1,2,4,5-Tetrachlorobenzene	4.09	3.61	4.16
4	2,2-Dimethylpropane	3.23	3.48	3.41	48	Bromobenzene	2.51	2.34	2.33
5	Octane	4.70	4.79	4.95	49	Iodobenzene	2.73	2.51	2.64
6	Nonane	5.40	5.29	5.28	50	Phenyl methyl ether	1.71	1.76	1.61
7	Decane	5.82	5.80	5.92	51	4-Chloroanisole	2.37	2.41	2.23
8	Undecane	6.27	6.31	6.46	52	2-Chloroaniline	1.04	1.12	0.98
9	Tridecane	7.27	7.34	7.18	53	4-Chloroaniline	0.84	1.48	0.92
10	Tetradecane	7.48	7.86	7.29	54	2,4-Dichloroaniline	1.69	1.56	1.59
11	Cyclopropane	1.43	1.33	1.49	55	3,5-Dimethylphenol	0.42	0.68	0.46
12	Ethene	1.34	1.42	1.46	56	4-Ethylphenol	0.60	0.63	0.55
13	Propene	1.80	1.73	1.62	57	2-Chlorophenol	0.56	0.26	0.58
14	1-Butene	2.31	2.16	2.22	58	Pentachlorophenol	2.65	2.29	2.64
15	2-Methyl-1-propene	2.16	2.32	2.18	59	Naphthalene	2.83	2.86	2.81
16	Trichloromethane	1.71	1.87	1.84	60	1-Methylnaphthalene	3.26	3.18	3.11
17	Tetrachloromethane	2.84	2.98	2.92	61	2-Methylnaphthalene	3.17	3.21	3.15
18	1,1,1-Trichloroethane	2.75	2.51	2.59	62	2,6-Dimethylnaphthalene	3.59	3.66	3.66
19	1,2-Dichloropropane	2.10	2.90	2.25	63	Fluorene	3.72	3.52	3.54
20	Trichloroethylene	2.24	2.60	2.31	64	Anthracene	3.84	3.78	3.78
21	Tetrachloroethylene	3.27	3.41	3.21	65	Fluoranthene	4.26	4.09	4.11
22	Dibromochloromethane	2.16	1.73	2.00	66	Benz[a]anthracene	4.77	4.75	4.70
23	Pentan-2-one	0.41	0.52	0.46	67	Chrysene	4.69	4.78	4.72
24	Hexan-2-one	0.86	0.97	0.92	68	Benzo[b]fluoranthene	5.16	5.04	5.25
25	Hexan-3-one	0.98	1.11	1.02	69	Perylene	4.98	4.95	5.02
26	Heptan-2-one	1.35	2.55	1.41	70	Benzonitrile	1.04	2.18	1.12
27	Acetophenone	1.04	0.76	1.12	71	Tetrafluoromethane	1.57	0.25	1.39
28	4-Chloroacetophenone	1.64	1.27	1.43	72	Sulfur hexafluoride	2.10	2.03	1.99
29	Isobutyl acetate	1.66	1.41	1.62	73	Phenethyl alcohol	0.12	0.36	0.14
30	Phenyl acetate	0.86	1.44	0.95	74	3-Methylbenzyl alcohol	0.17	0.54	0.20
31	Methyl benzoate	1.65	1.36	1.46	75	2-Chlorobiphenyl	3.97	3.74	3.91
32	Benzene	2.10	2.04	2.21	76	2,4,4'-Trichlorobiphenyl	4.70	4.79	4.81
33	Toluene	2.24	2.38	2.36	77	2,2',4,5,5'-Pentachlorobiphenyl	5.71	5.77	5.79
34	Ethylbenzene	2.71	2.79	2.74	78	2,2',5,5'-Tetrachlorobiphenyl	5.30	5.71	5.44
35	1,2-Dimethylbenzene	2.50	2.87	2.68	79	Limonene	4.14	3.32	4.00
36	1,3-Dimethylbenzene	2.95	2.82	2.83	80	3,4-Dichloroaniline	1.39	2.58	1.45
37	1,4-Dimethylbenzene	2.76	2.89	2.93	81	2,3,3',4,4'-Pentachlorobiphenyl	5.89	5.88	5.88
38	Propylbenzene	3.14	3.22	3.18	82	2,2',3,4,4',5-Hexachlorobiphenyl	6.20	6.29	6.07
39	Isopropylbenzene	3.25	3.28	3.29	83	2,3',4,4',5-Pentachlorobiphenyl	5.87	5.79	5.80
40	1,3,5-Trimethylbenzene	3.25	3.37	3.28	84	Bromoform	1.87	2.09	1.96
41	Styrene	2.86	2.66	2.65	85	2,2',4,5,5'-Pentachlorobiphenyl	5.71	5.70	5.72
42	Chlorobenzene	2.40	2.05	2.21	86	2,2',4,4',5,5'-Hexachlorobiphenyl	6.16	6.54	6.28
43	1,2-Dichlorobenzene	2.87	2.67	2.65	88	1,2-Dichloroethane	1.16	2.41	1.19
44	1,4-Dichlorobenzene	2.93	2.75	2.70	89	Benzonitrile	0.86	2.00	0.91

Continued Table 1

1	2	3	4	5	6	7	8	9	10
Training set									
90	Diethyl ether	0.66	1.84	0.65	92	2-Chlorotoluene	3.07	2.45	3.04
91	2,3,5,6-Tetrachlorobiphenyl	5.34	5.20	5.34	93	2,4,5-Trichloroaniline	2.08	2.93	2.18
Test set									
94	Ethane	1.71	2.01	1.82	106	3,4-Dimethylaniline	1.07	1.70	1.14
95	Pentane	3.47	3.30	3.26	107	4-Chlorotoluene	2.87	2.46	2.61
96	Heptane	4.61	4.29	4.41	108	Camphor	1.48	1.39	1.34
97	Cyclohexane	3.52	3.12	3.37	109	Biphenyl	3.37	3.71	3.16
98	1,1,1,2-Tetrachloroethane	2.66	3.23	2.89	110	Acenaphthene	3.63	3.11	3.40
99	Trifluoromethane	0.60	0.41	0.56	111	Pyrene	4.32	4.04	4.12
100	Ethyl acetate	0.27	0.49	0.31	112	Benz[a]pyrene	5.24	4.95	5.02
101	Ethyl benzoate	2.12	1.57	2.01	113	2,2',3,4,4',5-Hexachlorobiphenyl	6.20	5.68	6.05
102	1,2,4-Trimethylbenzene	2.94	3.32	3.12	114	4,4'-Dichlorobiphenyl	4.59	4.50	4.41
103	1,3-Dichlorobenzene	3.29	2.60	3.48	115	2,2',3,4,4',5,5'-Heptachloro-	6.40	6.81	6.21
104	1,2,3,5-Tetrachlorobenzene	4.18	3.58	4.40		biphenyl			
105	Hexachlorobenzene	5.01	5.42	5.27	116	Thiophene	1.75	1.24	1.61
Validation set									
117	Butane	2.93	2.85	2.70	129	Phenanthrene	4.00	3.78	3.81
118	Hexane	4.04	3.80	3.88	130	Benzo[k]fluoranthene	5.33	5.05	5.13
119	Dodecane	6.82	6.82	6.99	131	1-Methylphenanthrene	4.50	4.10	4.32
120	1,3-Butadiene	1.78	2.05	1.96	132	2,4',6'-Trichlorobiphenyl	5.00	4.72	4.80
121	1,1,2,2-Tetrachloroethane	2.17	3.13	2.03	133	Hexafluoroethane	2.40	2.18	2.21
122	Methyl 2-methylbenzoate	2.15	1.63	2.00	134	2,2',4,4',6,6'-Hexachlorobiphenyl	6.03	6.36	6.26
123	1,3,5-Trichlorobenzene	3.64	3.11	3.42	135	Ethanethiol	1.12	1.66	1.24
124	Pentachlorobenzene	4.62	4.09	4.41	136	1,2,3-Trichlorobenzene	3.45	3.13	3.26
125	3-Chlorophenol	0.31	0.24	0.34	137	2,4,5-Trichlorotoluene	4.17	3.50	4.36
126	Nitrobenzene	1.21	2.45	1.44	138	Acridine	3.17	3.39	3.37
127	3-Bromophenol	0.46	0.32	0.41	139	Benzo[ghi]perylene	5.50	5.31	5.29
128	1,2-Dimethylnaphthalene	3.47	3.21	3.25					

were drawn with Hyperchem (Version 7) [ 35 ] and then they were pre-optimized using MM<sup>+</sup> molecular mechanics force field. A more precise optimization was then done with the semiempirical AM1 method in Mopac (Version 6) [ 36 ]. All calculations were carried out at a restricted Hartree-Fock level with no configuration interaction. As a next step, the Mopac output files were used by the CODESSA program [ 37, 38 ] to calculate five classes of descriptors including constitutional, geometrical, topological, electrostatic, and quantum-chemical descriptors. The software CODESSA, developed by Karitzky group, enables the calculation of a large number of quantitative descriptors based only on the molecular structure information and codes the chemical information into a mathematical form [ 37, 38 ]. CODESSA combines diverse methods for quantifying the structural information about the molecule with advanced statistical analysis to establish quantitative structure-property relationship.

**Variable Selection Using Genetic Algorithm.** Genetic Algorithms (GAs) are adaptive heuristic search algorithms that can be applied when the dimension of the data space is too large for an exhaustive search. They have been proved to be an efficient method in the feature selection problems [ 39, 40 ]. GAs have several advantages in comparison with other optimization algorithms. They have the ability to move from local optima present on the response surface. They require no knowledge or gradient information about the response surface and can be employed for a wide variety of optimization problems [ 41 ]. The major drawbacks of GAs are potential difficulties in finding the exact global

optimum, which requires a large number of response (fitness) function evaluations and configuring the problem is not straightforward [42]. There are some basic steps in genetic algorithms as follow: (1) a chromosome is represented by a binary bit string and an initial population of chromosomes is created in a random way; (2) a value for the fitness function of each chromosome is evaluated; (3) according to the values of fitness function, the chromosomes of the next generation are reproduced by selection, crossover and mutation operations. In this paper, GA program was written with MATLAB 7.0 [43] and based on Leardi's method [44] with a few minor modifications in our laboratory. The size of population was 30, the probability of cross over was 0.5, the probability of mutation was 0.01 and the number of evaluations was 200. For each set of data 100 runs were performed. Here, we try to use varieties of fitness functions which are proportional to the residual error of the training set, test set and the number of selected variables according to the following equation:

$$\text{fitness} = \frac{1}{\text{SEC} + \text{SEP} + (m)^w} \quad (2)$$

In this equation, SEC and SEP are standard error of calibration (training) and test set, respectively;  $m$  is the number of variables in the represented model and  $w$  is a numerical value that implies the weights of  $m$  in the value of fitness. In fact, the value of  $w$  determines the number of variables presented in a selected chromosome. Some experiments were done using different values of  $w$ . Acquired results showed that for a small value of  $w$ , the number of variables in the fittest individual was high and on the other hand if the value of  $w$  was to be high, the number of variables in the best chromosome was small. Hence, after some experiments the value of  $w$  was set to be 0.3. It is worth noting that the parameter of  $w$  was determined in a preliminary study, before the overall genetic algorithm optimization had been carried out.

**Multiple Linear Regression (MLR).** Multiple linear regression is a common method used in QSPR study. The equation linking the structural features to the  $\log K_{\text{PDMS-water}}$  is developed in the form:

$$\log K_{\text{PDMS-water}} = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n, \quad (3)$$

where  $a_0$  is the intercept and  $a_1, a_2, \dots, a_n$  are the regression coefficients of the descriptors. The descriptors ( $x_1, x_2, \dots, x_n$ ) included in the equation are used to describe chemical structure of compounds and  $n$  is the number of the descriptors to find the best regression model. The main goal of the generation of the MLR model was to choose a set of suitable descriptors that could be used as inputs for the generation of the ANN model. From pairs of variables with  $R > 0.90$ , only one of them was used in the modeling and the variables that more than in 90 % cases were equal to zero were eliminated. Remaining descriptors were used to generate the models using the SPSS/PC software package [45]. A stepwise procedure was used for selection of the descriptors. This method combines the forward and backward procedures. Due to the complexity of inter-correlations, the variance caused by certain variables will change when new variables enter the equation. Sometimes a variable that is qualified to enter loses some of its predictive validity when other variables enter. If this takes place, the stepwise method will remove the weakened variable. A final set of selected equations was then tested for stability and validity through a variety of statistical methods. The choice for an equation suitable for further consideration was made by using four criteria, namely, multiple correlation coefficients ( $R$ ), standard error (SE),  $F$ -statistics and the number of descriptors in the model. The best Multiple Linear Regression (MLR) model is the one that has high  $R$  and  $F$ -values, low standard error, least number of descriptors and high ability for prediction. The best model selected in this work is presented in Table 2.

**Artificial Neural Network.** ANNs are basically a data-driven black-box model capable of solving highly non-linear complex problems. They have the ability to capture the relationship between input and output variables from given patterns (historical data or measured data on input and output variables of the system of the concern) and this enables them to solve large-scale complex problems. The network learns basically by finding the optimal network-connection-weights that would generate an output vector as close as possible to the target values of the output vector, with a desired accuracy. A detailed description of the theory behind a neural network has been adequately described elsewhere [46–48]. Therefore, only the points relevant to this work are described here. A fundamental process-

Table 2

## Specification of best multiple linear regression models

Descriptor	Notation	Coefficient	Mean effect
Minimum Atomic Orbital Electronic Population	$P_{\mu\mu}$	10.065±1.276	8.480
Kier Shape Index (order 3)	${}^3\kappa$	0.411±0.019	2.969
Polarity Parameter / Square Distance	PP	-6.383±0.802	-0.289
Complementary Information Content (order 2)	${}^2\text{CIC}$	0.007±0.002	0.207
Constant		-8.355±1.129	

sion element of an ANN is a node. Each node has a series of weighted inputs,  $W_{ij}$ , and acts as a summing point of weighted input signals. The summed signals pass through a transfer function that may be in sigmoidal form. The output of node  $j$ ,  $O_j$ , is given by

$$O_j = 1/[1 + \exp(-X)], \quad (4)$$

where  $X$  is defined by the following equation:

$$X = \sum W_{ji} O_i + B_j. \quad (5)$$

In Eq. (5),  $B_j$  is a bias term,  $O_i$  is the output of the node of the previous layer and  $W_{ji}$  represents the weight between the nodes of  $i$  and  $j$ .

A feedforward neural network consists of three layers. The first layer (input layer) consists of nodes and acts as an input buffer for the data. Signals introduced to the network, with one node per element in the sample data vector, pass through the input layer to the layer called the hidden layer. Each node in this layer sums the inputs and forwards them through a transfer function to the output layer. These signals are weighted and then pass to the output layer. In the output layer the processes of summing and transferring are repeated. The output of this layer now represents the calculated value for the node  $k$  of the network.

The training of back-propagation neural network requires the comparison of the network output with an expected value. This comparison may be presented in an iterative fashion to the network with a weighted adjustment after each run. The differences between the output and the expected value back-propagated to the network and followed by adjustment of the weights and biases. The adjusted weights and biases can be calculated according to

$$\Delta W_{kj}(n) = \eta \delta_{pk} O_{pj} + \alpha \Delta W_{kj}(n-1), \quad (6)$$

$$\Delta B_{kj}(n) = \gamma \delta_{pk} O_{pj}. \quad (7)$$

In these equations,  $\Delta W_{kj}$  and  $\Delta B_{kj}$  are the changes in the weights and biases between the node  $j$  in the hidden layer and the node  $k$  in the output layer, respectively;  $\delta_{pk}$  is the error term obtained from the differences between the output and the expected value. The parameters  $\eta$  and  $\gamma$  are learning rate of the weight and bias, respectively;  $\alpha$  represents the momentum and  $n$  and  $n-1$  refer to the present and the previous iterations, respectively.

Equations similar to the Eqs. (6) and (7) were used to adjust weights and biases connecting the hidden layers to the input one. The criterion for stopping the iteration during the training process could be a predefined number of iterations ( $p$ ) or a desired difference between the output and its expected value. In order to obtain a parsimonious model, the network architecture was modified and tested. The number of hidden layer nodes, learning rates and momentum parameters were optimized.

In the present work, an ANN program was written with MATLAB 7. This network was feed-forward fully connected with three layers with sigmoidal transfer function. Descriptors selected by MLR methods were used as inputs of network and its output signal represented the water-to-polydimethylsiloxane partition coefficients for the compounds of interest. Thus this network has four nodes in input layer and one node in output layer. The values of each input were divided by their mean value to bring them into dynamic range of the sigmoidal transfer function of the network. The initial values of weights were randomly selected from a uniform distribution that ranged between -0.3 to

+0.3 and the initial values of biases were set to be unity. These values were optimized during the network training. The back-propagation algorithm was used for the training of the network. During the training, the network parameters would be optimized. These parameters were: the number of nodes in the hidden layer, weights and biases of learning rates and the momentum. Procedures for the optimization of these descriptors were reported elsewhere [49, 50]. Then the optimized network was trained using a training set for adjustment of weights and biases values. To maintain the predictive power of the network at a desirable level, training was stopped when the value of error for the test set started to increase. Since the test error is not a good estimation of the generalization error, the prediction potential of the model was evaluated on a third set of data, named validation set. Compounds in the validation set were not used during the training process and were reserved to evaluate the predictive power of the generated ANN.

#### EVALUATION OF THE PREDICTABILITY OF THE QSPR MODEL

For the optimized QSPR model, several parameters were selected to test the prediction ability of the model. A real QSPR model may have a high predictive ability when it is close to ideal. This may imply that the correlation coefficient  $R$  between the experimental  $y$  and predicted  $\tilde{y}$  properties must be close to 1 and regression of  $y$  against  $\tilde{y}$  or  $\tilde{y}$  against  $y$  through the origin, i.e.  $y^{r0} = k\tilde{y}$  and  $\tilde{y}^{r0} = k'y$ , respectively, should be characterized by at least either  $k$  or  $k'$  close to 1 [51]. Slopes  $k$  and  $k'$  were calculated as follows:

$$k = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2}, \quad (8)$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2}. \quad (9)$$

The criteria formulated above may not be sufficient for a QSPR model to be truly predictive. Regression lines through the origin defined by  $y^{r0} = k\tilde{y}$  and  $\tilde{y}^{r0} = k'y$  (with the intercept set to unity) should be close to optimum regression lines  $y^r = a\tilde{y} + b$  and  $\tilde{y}^r = a'y + b'$  ( $b$  and  $b'$  are intercepts). Correlation coefficients for these lines  $R_0^2$  and  $R_0'^2$  are calculated as follows:

$$R_0^2 = 1 - \frac{\sum (\tilde{y}_i - y_i^{r0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2}, \quad (10)$$

$$R_0'^2 = 1 - \frac{\sum (y_i - \tilde{y}_i^{r0})^2}{\sum (y_i - \bar{y})^2}, \quad (11)$$

where  $\bar{y}$  and  $\bar{\tilde{y}}$  are the average values of the observed and predicted properties, respectively and the summations are over all  $n$  compounds in the validation set.

A difference between  $R^2$  and  $R_0^2$  values ( $R_m^2$ ) needs to be studied to explore the prediction potential of a model [52]. This term was defined in the following manner:

$$R_m^2 = R^2 (1 - |\sqrt{R^2 - R_0^2}|). \quad (12)$$

Finally, the following criteria for evaluation of the predictive ability of QSPR models should be considered:

1. High value of cross-validated  $R^2$  ( $q^2 > 0.5$ ).
2. Correlation coefficient  $R$  between the predicted and actual properties from an external test set close to 1.  $R_0^2$  or  $R_0'^2$  should be close to  $R^2$ .
3. At least one slope of regression lines ( $k$  or  $k'$ ) through the origin should be close to 1.
4.  $R_m^2$  should be greater than 0.5.

The predictive power of the ANN models developed on the selected training sets was estimated from the predictions of validation set chemicals, by calculating the  $q^2$  that is defined as follows:

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}, \quad (13)$$

where  $y_i$  and  $\hat{y}_i$  are the experimental and predicted values of the dependent variable (water-to-polydimethylsiloxane partition coefficient), respectively,  $\bar{y}$  is the averaged value of dependent variable of the training set and the summations cover all the compounds.

## RESULTS AND DISCUSSION

Table 1 shows the data set used and the corresponding observed, and predicted MLR and ANN values of water-to-polydimethylsiloxane partition coefficients for all the molecules studied in this work. It can be seen from Table 2 that four descriptors appeared in the MLR model. These descriptors were: Minimum Atomic Orbital Electronic Population ( $P_{\mu\mu}$ ), Kier Shape Index (order 3) ( ${}^3\kappa$ ), Polarity Parameter / Square Distance (PP), and Complementary Information Content (order 2) ( ${}^2\text{CIC}$ ). The numerical values of these descriptors are listed in Table 3. Table 4 represents the correlation matrix for these descriptors.

By interpreting the descriptors in the models, it is possible to gain some insight into the factors that are likely related to the water-to-PDMS partition coefficients for the organic compounds. For the inspection of the relative importance and contribution of each descriptor in the model, the value of mean effect (ME) was calculated for each descriptor using the following equation:

$$\text{ME}_j = \frac{\beta_j \sum_{i=1}^n d_{ij}}{\sum_j \beta_j \sum_i d_{ij}}, \quad (14)$$

where  $\text{ME}_j$  is the mean effect for a considered descriptor  $j$ ,  $\beta_j$  is the coefficient of the descriptor  $j$  and  $d_{ij}$  is the value of interested descriptors for each molecule, and  $m$  is the number of descriptors in the model. The calculated values of MEs are represented in the last column of Table 2 and are also plotted in Fig. 1. The value and sign of the mean effect shows the relative contribution and direction of influence of each descriptor on the partition coefficient. As shown in Table 2 the most relevant descriptor based on its mean effect is Minimum Atomic Orbital Electronic Population ( $P_{\mu\mu}$ ), a quantum-chemical descriptor. This descriptor describes the nucleophilicity of the molecule. Molecules with high  $P_{\mu\mu}$  are more able to donate their electrons and hence are relatively reactive compared to molecules with low  $P_{\mu\mu}$ . A molecule with higher reactivity will be adsorbed stronger into the polymer phase. Thus, an increase in the descriptor value leads to an increase in the dispersion forces during the sorption process.

The second descriptor in this model, Kier Shape Index (order 3) ( ${}^3\kappa$ ) [53, 54], is defined as:

$${}^3\kappa = (N_{SA} + \alpha - 1)(N_{SA} + \alpha - 3)^2 ({}^3P + \alpha)^2 \text{ if } N_{SA} \text{ is odd,} \quad (15)$$

$${}^3\kappa = (N_{SA} + \alpha - 3)(N_{SA} + \alpha - 2)^2 ({}^3P + \alpha)^2 \text{ if } N_{SA} \text{ is even,} \quad (16)$$

where  $N_{SA}$  is the number of non-hydrogen atoms in the molecule,  ${}^3P$  is the number of paths of the length 3 in the molecular graph and  $\alpha$  is the sum of the  $a_i$  parameters for all skeletal atoms minus 1. This descriptor shows the effect of molecule's shape in sorption process and, as it can be seen in Table 2, it has positive sign. Water molecules have strong hydrogen-bonding ability and considerable parts of them are combined with one another to form three-dimensional networks. Due to network formation,

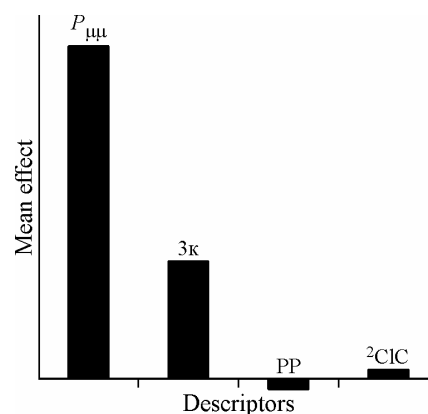


Fig. 1. Plot of descriptor's mean effects



Table 3

*The values of the descriptors that were used in this work*

$N^*$	$P_{\mu\mu}$	${}^3\kappa$	PP	${}^2\text{CIC}$	$N^*$	$P_{\mu\mu}$	${}^3\kappa$	PP	${}^2\text{CIC}$
1	2	3	4	5	6	7	8	9	10
1	0.9335	1.0000	0.0389	8.0000	45	0.8489	5.9282	0.0194	16.0000
2	0.9236	3.0000	0.0101	19.5098	46	0.8476	5.9282	0.0067	20.0000
3	0.9164	4.0000	0.0102	33.2842	47	0.8341	7.2000	0.0096	13.5098
4	0.9274	5.0000	0.0037	51.0195	48	0.8403	8.4763	0.0098	16.0000
5	0.9213	8.0000	0.0018	70.5293	49	0.8302	8.4763	0.0072	20.0000
6	0.9213	9.0000	0.0018	84.4224	50	0.8540	4.8493	0.0108	18.3645
7	0.9213	10.0000	0.0011	99.0195	51	0.8564	5.0939	0.0034	18.3645
8	0.9212	11.0000	0.0011	114.2199	52	0.8020	5.3193	0.0485	23.1194
9	0.9212	13.0000	0.0008	146.1466	53	0.7983	6.5881	0.0125	16.7549
10	0.9212	14.0000	0.0006	162.7682	54	0.7993	5.6039	0.1652	12.0000
11	0.8924	1.3333	0.0085	20.2647	55	0.8316	5.6039	0.1690	14.0000
12	0.8910	1.7400	0.0095	10.0000	56	0.7934	6.8742	0.1653	8.7549
13	0.8834	2.7400	0.0101	6.7549	57	0.7834	6.3023	0.2767	26.2647
14	0.8816	3.7400	0.0064	8.7549	58	0.7832	6.3023	0.2765	18.7549
15	0.8888	3.7400	0.0038	19.5098	59	0.7789	5.6039	0.2725	10.0000
16	0.8422	4.8700	0.0562	4.7549	60	0.7615	10.7092	0.2649	18.3645
17	0.8422	4.8700	0.0562	4.7549	61	0.8680	5.4822	0.0016	26.0000
18	0.8932	6.1600	0.0633	8.0000	62	0.8675	6.4127	0.0076	21.5098
19	0.8573	5.8700	0.0541	9.5098	63	0.8679	6.4127	0.0067	23.5098
20	0.8820	5.5800	0.0108	6.7549	64	0.8683	7.3545	0.0013	27.0196
21	0.8655	5.6100	0.0260	2.0000	65	0.8651	6.9029	0.0020	42.0000
22	0.8255	5.6100	0.0260	2.0000	66	0.8659	7.5714	0.0017	37.5098
23	0.8837	6.9000	0.0248	10.0000	67	0.8638	8.1164	0.0012	55.2193
24	0.8152	5.2500	0.0529	2.0000	68	0.8655	9.6658	0.0017	54.7549
25	0.7137	5.6700	0.1050	13.5098	69	0.8664	9.6658	0.0012	57.0196
26	0.7138	6.6700	0.1050	19.5098	70	0.8625	10.2190	0.0025	70.1390
27	0.7225	6.6700	0.1053	25.5098	71	0.8659	10.2190	0.0008	47.5489
28	0.8116	7.6700	0.1050	27.0196	72	0.7859	8.2190	0.1608	27.5489
29	0.7182	6.0167	0.1062	23.1194	73	0.7599	4.7200	0.1701	8.0000
30	0.7196	7.2889	0.1066	16.7549	74	0.7891	6.9300	0.0849	15.5098
31	0.7253	7.6300	0.1270	24.2647	75	0.7604	6.3023	0.2828	22.3645
32	0.7516	6.9630	0.1238	23.1194	76	0.7788	6.3023	0.2817	16.7549
33	0.7455	6.9630	0.1253	23.1194	77	0.8530	8.3338	0.0462	44.8939
34	0.8699	3.4116	0.0025	31.0196	78	0.8369	10.8155	0.0029	35.9161
35	0.8699	4.3808	0.0066	23.1194	79	0.8324	13.3247	0.0025	33.2193
36	0.8673	5.3585	0.0030	25.1194	80	0.8712	12.0674	0.0048	39.5098
37	0.8704	5.3585	0.0022	31.5098	81	0.8289	7.5855	0.0014	29.5098
38	0.8698	5.3585	0.0067	29.5098	82	0.8060	6.8742	0.0122	10.7549
39	0.8702	5.3585	0.0023	35.5098	83	0.8422	13.3247	0.0023	31.9742
40	0.8676	6.3417	0.0021	27.1194	84	0.8345	14.5862	0.0025	27.5098
41	0.8688	6.3417	0.0030	35.8743	85	0.8358	13.3247	0.0025	29.2193
42	0.8700	6.3417	0.0017	47.5489	86	0.8277	5.4400	0.0315	4.7549
43	0.8676	5.1037	0.0029	20.3645	87	0.8295	13.3247	0.0038	27.2193
44	0.8534	4.6636	0.0462	18.3645	88	0.8516	14.5862	0.0052	39.5098

Continued Table 3

1	2	3	4	5	6	7	8	9	10
89	0.8296	14.5862	0.0023	27.5098	116	0.8511	9.5703	0.0017	46.0000
90	0.8763	4.5800	0.0172	12.0000	117	0.8310	15.8513	0.0054	35.1613
91	0.8520	4.8590	0.0672	18.3645	118	0.8305	3.0370	0.0133	8.0000
92	0.8213	4.9600	0.0503	27.5098	119	0.9225	4.0000	0.0063	27.5098
93	0.8305	12.0674	0.0074	34.2647	120	0.9215	6.0000	0.0028	45.5098
94	0.8542	5.6432	0.0461	14.7549	121	0.9212	12.0000	0.0008	129.9483
95	0.7884	8.1495	0.0094	10.7549	122	0.8789	3.4800	0.0062	14.0000
96	0.9294	2.0000	0.0099	17.5098	123	0.8468	7.1600	0.0214	12.0000
97	0.9214	5.0000	0.0063	35.0196	124	0.7338	7.9515	0.1245	19.5098
98	0.9213	7.0000	0.0028	57.4839	125	0.8344	7.2000	0.0088	19.0196
99	0.9220	4.1667	0.0028	58.5293	126	0.8275	9.7557	0.0101	18.3645
100	0.8620	7.1600	0.0280	6.7549	127	0.7795	5.6039	0.2753	8.0000
101	0.7992	3.7900	0.1483	4.7549	128	0.8290	5.9085	0.0192	18.3645
102	0.7247	5.6300	0.1273	11.5098	129	0.7797	5.7905	0.2756	8.0000
103	0.7283	7.9515	0.1290	25.1194	130	0.8679	6.4127	0.0067	23.5098
104	0.8704	6.3417	0.0067	42.0391	131	0.8669	7.5714	0.0011	34.2647
105	0.8382	5.9282	0.0087	14.0000	132	0.8626	10.2190	0.0017	71.7744
106	0.8310	8.4763	0.0080	14.0000	133	0.8662	8.4853	0.0079	31.7744
107	0.8894	11.0372	0.0078	31.0196	134	0.8345	10.8155	0.0089	35.1613
108	0.8194	6.3022	0.1693	26.2647	135	0.7591	7.5800	0.0490	17.5098
109	0.8539	5.6432	0.0462	16.7549	136	0.8325	14.5862	0.0029	43.5098
110	0.7167	7.3261	0.1077	40.5293	137	0.8775	3.3500	0.0560	6.7549
111	0.8653	7.1087	0.0016	58.7291	138	0.8452	7.2000	0.0213	13.5098
112	0.8664	6.2419	0.0013	17.5098	139	0.8349	8.1891	0.0096	13.5098
113	0.8664	8.1164	0.0011	42.0000	140	0.8490	7.5079	0.0167	23.6096
114	0.8649	10.2190	0.0024	51.0195	141	0.8653	10.8079	0.0006	66.7290
115	0.8656	11.7622	0.0008	62.7291					

The definitions of the descriptors are given in Table 2.

\* The numbers refer to the numbers of the molecules given in Table 1.

molecules which are large in size are often difficult to dissolve in water, so the sorption process into PDMS can perform easily for these molecules. Thus as the value of Kier shape index increases, the water-to-polydimethylsiloxane partition coefficient increases.

The third descriptor was PP [ 55 ] which defines as:

$$PP = \frac{Q_{\max} - Q_{\min}}{R_{\text{mm}}^2}, \quad (17)$$

where  $Q_{\max}$  is the most positive atomic partial charge in the molecule,  $Q_{\min}$  is the most negative atomic partial charge in the molecule and  $R_{\text{mm}}$  is the distance between the most positive and the most negative atomic partial charges in the molecule. As it can be seen from Table 2, the polarity parameter has a negative effect in the model proposed. There is an old principle "*like dissolves like*", so there will be a strong interaction between molecules with high polarity parameter and water as a polar solvent. Thus as polarity parameter of molecules increases, the tendency of sorption into PDMS will decrease.

The last descriptor was Complementary Information Content (order 2) ( ${}^2\text{CIC}$ ) [ 56 ] which was calculated as follows:

$${}^k\text{CIC} = \log_2 n - {}^k\text{IC}, \quad (18)$$

Table 4

Correlation matrix between selected descriptors

	$P_{\mu\mu}$	${}^3\kappa$	PP	${}^2\text{CIC}$
$P_{\mu\mu}$	1	-0.053	0.634	0.407
${}^3\kappa$		1	-0.182	0.522
PP			1	-0.299
${}^2\text{CIC}$				1

Table 5

Architecture and specifications of optimized ANN model

Number of nodes in the input layer	4
Number of nodes in the hidden layer	5
Number of nodes in the output layer	1
Weights learning rate	0.2
Biases learning rate	0.1
Momentum	0.5
Transfer function	Sigmoid

where  $n$  is the total number of atoms in the molecule,  $k$  is the number of atomic layers in the coordination sphere around a given atom and the information content (IC) is equal to average information content multiplied by the total number of atoms. This descriptor describes the connectivity and branching in a molecule and can be related to molecular shape and symmetry. The relative number of rings in the fragments can also be related to molecular shape. The positive mean effect for  ${}^2\text{CIC}$  reflects the fact that molecules with lower symmetry have weaker sorption ability that leads to lower  $K_{\text{PDMS-water}}$ .

From the above discussion, it can be seen that all descriptors involved in the QSPR model have physical meaning, and these descriptors can account for structural features that affect the water-to-polydimethylsiloxane partition coefficients of the compounds studied.

The next step was the construction of an artificial neural network. During the training of the ANNs, the parameters of the network including the number of nodes in the hidden layer, weights and biases learning rates and momentum values were optimized. Table 5 shows the architecture and specifications of the optimized network. After the optimization of the network parameters, the network was trained by using the training set for adjustment of the weights and biases values by back-propagation algorithm. It is known that neural network can become overtrained. An overtrained network has usually learned perfectly the stimulus pattern it has seen but cannot give accurate prediction for unseen stimuli. There are several methods to overcome this problem. One method is to use a test set to evaluate the prediction power of the network during its training. In this method after each 1000 training iterations the network was used to calculate  $\log K_{\text{PDMS-water}}$  of molecules included in the test set. To maintain the predictive power of the network at a desirable level, training was stopped when the value of errors for the test set started to increase. The results obtained showed that overtraining began after 23 000 iterations. Table 1 represents the experimental, and MLR and ANN calculated values of water-to-polydimethylsiloxane partition coefficients for the training, test and validation sets. The statistical parameters obtained by the ANN and MLR models for these sets are shown in Table 6. The standard errors of training, test and validation sets for the MLR model were 0.422, 0.425 and 0.480, respectively. These values could be compared with the values of 0.116, 0.179, and 0.183, respectively, for the ANN model. The comparison between these values and other statistical parameters in Table 6 reveals the superiority of the ANN model over MLR. The key strength of neural networks, unlike MLR

Table 6

Statistical parameters obtained using the ANN and MLR models\*

Model	$\text{SE}_c$	$\text{SE}_t$	$\text{SE}_v$	$R_c$	$R_t$	$R_v$	$F_c$	$F_t$	$F_v$
ANN	0.116	0.179	0.183	0.998	0.995	0.995	21122	1919	2030
MLR	0.422	0.425	0.480	0.971	0.969	0.964	1512	321	277

\* The index c refers to the calibration (training) set; t refers to the test set; v refers to the validation set;  $R$  is the correlation coefficient; SE is the standard error and  $F$  is the statistical  $F$  value.

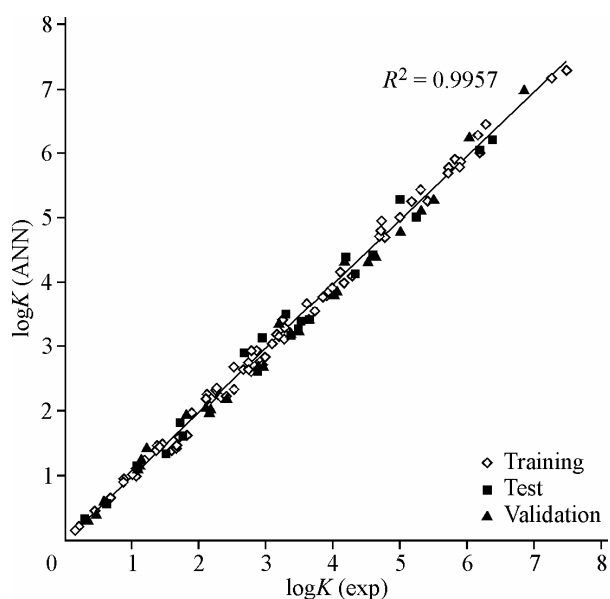


Fig. 2. Plot of ANN calculated water-to-polydimethylsiloxane partition coefficient against experimental values

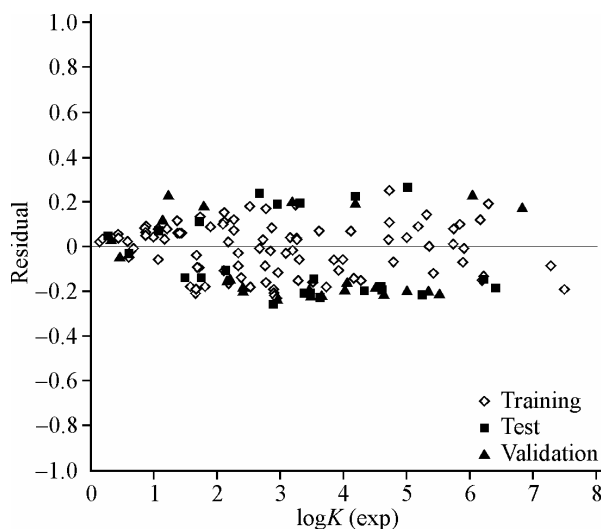


Fig. 3. Plot of residual of ANN calculated versus experimental values of water-to-polydimethylsiloxane partition coefficient

analysis, is their ability for flexible mapping of the selected features by manipulating their functional dependence implicitly.

The statistical values of the validation set for the ANN model was characterized by  $q^2 = 0.989$ ,  $R^2 = 0.990$  ( $R = 0.995$ ),  $R_0^2 = 0.988$ ,  $R_m^2 = 0.956$  and  $k = 1.015$ . These values and other statistical parameters shown in Table 6 reveal a high predictive ability of the model. Fig. 2 shows the plot of the ANN predicted versus experimental values for water-to-polydimethylsiloxane partition coefficients for all of the molecules in data set. The residuals of the ANN calculated values of the water-to-polydimethylsiloxane partition coefficients are plotted against the experimental values in Fig. 3. The propagation of the residuals on both sides of zero line indicates that no systematic error exists in the constructed QSPR model.

## CONCLUSION

In the present work GA as a feature selection tool and MLR and ANN as feature mapping techniques were used for prediction of the water-to-polydimethylsiloxane partition coefficients of 139 organic compounds. The optimized 4-5-1 ANN model showed a remarkable improvement over the linear model. The GA-based MLR approach is especially useful for modeling large variable data sets. The physical meaning of the selected descriptors, which are, according to the GA method, are the most predictive and informative, was identified. The water-to-polydimethylsiloxane partition mechanisms of investigated compounds were interpreted rationally with these four descriptors. The result obtained indicate that while the GA and MLR method could be more powerful in precise selecting of important parameters and assuming the significance of each of the descriptors, the introduction of a neural network gives a significant improvement in prediction quality.

## REFERENCES

1. Tanford C. The Hydrophobic Effect. Formation of Micelles, Biological Membranes, 2nd ed. – N.Y.: Wiley—Interscience, 1980.
2. Braumann T. // J. Chromatogr. A. – 1986. – 373. – P. 191.
3. Belardi R.P., Pawliszyn J. // Water Pollut. Res. J. Can. – 1989. – 24. – P. 179.
4. Pawliszyn J. // Solid Phase Microextraction: Theory, Practice. – New York: Wiley, 1997.
5. De Coensel N., Desmet K., Gorecki T., Sandra P. // J. Chromatogr. A. – 2007. – 1150. – P. 183.

6. Yang Z.Y., Greenstein D., Zeng E.Y., Maruya K.A. // J. Chromatogr. A. – 2007. – **1148**. – P. 23.
7. Kwon J.H., Wuethrich T., Mayer P., Escher B.I. // Anal. Chem. – 2007. – **79**. – P. 6816.
8. Isidorov V.A., Vinogorova V.T. // J. Chromatogr. A. – 2005. – **1077**. – P. 195.
9. Zeng E.Y., Tsukada D., Noblet J.A., Peng J. // J. Chromatogr. A. – 2005. – **1066**. – P. 165.
10. Makrodimitri Z.A., Dohrn R., Economou I.G. // Macromolecules. – 2007. – **40**. – P. 1720.
11. Klopman G., Ding C., Macina O.T. // J. Chem. Inf. Comput. Sci. – 1997. – **37**. – P. 569.
12. Lu W., Chen Y., Liu M. et al. // Chemosphere. – 2007. – **69**. – P. 469.
13. Puzyn T., Falandysz J. // Phys. Chem. Ref. Data. – 2007. – **36**. – P. 203.
14. Ohlenbusch G., Frimmel F.H. // Chemosphere. – 2001. – **45**. – P. 323.
15. Metivier-Pignon H., Faur C., Le Cloirec P. // Chemosphere. – 2007. – **66**. – P. 887.
16. Sprunger L., Proctor A., Acree W.E. Jr., Abraham M.H. // J. Chromatogr. A. – 2007. – **1175**. – P. 162.
17. Sprunger L., Gibbs J., Acree W.E. Jr., Abraham M.H. // Fluid Phase Equilibria. – 2008. – **273**. – P. 78.
18. Hierleman A., Zellers E.T., Ricco A.J. // Anal. Chem. – 2001. – **73**. – P. 3458.
19. Vegas J.M., Zufiria P.J. // Neural Networks. – 2004. – **17**. – P. 233.
20. Schweitzer R.C., Morris J.B. // Anal. Chem. Acta. – 1999. – **384**. – P. 285.
21. Tong C.S., Cheng K.C. // Chemometrics, Intelligent Laboratory Systems. – 1999. – **49**. – P. 135.
22. Golmohammadi H., Fatemi M.H. // Electrophoresis. – 2005. – **26**. – P. 3438.
23. Baher E., Fatemi M.H., Konož E., Golmohammadi H. // Microchim. Acta. – 2007. – **158**. – P. 117.
24. Konož E., Golmohammadi H. // Anal. Chem. Acta. – 2008. – **619**. – P. 157.
25. Golmohammadi H. // J. Comput. Chem. – 2009. – **30**. – P. 2455.
26. Golmohammadi H., Konož E., Dashtbozorgi Z. // Anal. Chem. – 2009. – **25**. – P. 1137.
27. Ohlenbusch G., Frimmel F.H. // Chemosphere. – 2001. – **45**. – P. 323.
28. Todeschini R., Consonni V. Handbook of Molecular Descriptors. – Weinheim: Wiley-VCH, 2000.
29. Fatemi M.H., Jalali-Heravi M., Konuze E. // Anal. Chem. Acta. – 2003. – **486**. – P. 101.
30. Kier L.B., Hall L.H. Molecular Connectivity in Structure-Activity Analysis. – Chichester, UK: RSP-Wiley, 1986.
31. Kostantinova E.V. // J. Chem. Inf. Comp. Sci. – 1997. – **36**. – P. 54.
32. Rucker G., Rucker C. // J. Chem. Inf. Comp. Sci. – 1993. – **33**. – P. 683.
33. Galvez J., Garcia R., Salabert M.T., Soler R. // J. Chem. Inf. Comp. Sci. – 1994. – **34**. – P. 520.
34. Broto P., Moreau G., Vandicke C. // Eur. J. Med. Chem. – 1984. – **19**. – P. 66.
35. Hyperchem for Windows, Release 4, Autodesk, Sansalito, CA, 1995.
36. Stewart J.J.P. Semiempirical Molecular Orbital Program; QCPE, 445, 1983; Version 6, 1990.
37. Katritzky A.R., Labadov V.S., Carelson M. CODESSA Training Manual, University of Florida, Gainesville, 1995.
38. Katritzky A.R., Labadov V.S., Carelson M. CODESSA Version 1 Reference Manual, University of Florida, Gainesville, Florida, 1994.
39. Leardi R., Boggia R., Terrile M. // J. Chemometr. – 1992. – **6**. – P. 267.
40. Leardi R., Gonzalez A.L. // Chemometr. Intell. Lab. Syst. – 1998. – **41**. – P. 195.
41. Chambers L. Practical Handbook of Genetic Algorithms. Lewis Publishing, 1995.
42. Hibbert D.B. // Chemometr. Intell. Lab. Syst. – 1993. – **19**. – P. 277.
43. MATLAB 7.0, The Mathworks Inc., Natick, MA, USA, <http://www.mathworks.com>.
44. Blank T.B., Brown S.T. // Anal. Chem. – 1993. – **65**. – P. 3081.
45. Zupan J., Gasteiger J. // Neural Network in Chemistry, Drug Design. – Weinheim: Wiley-VCH, 1999.
46. SPSS/PC, Statistical Package for IBMPC, Quiad Software, Ontario, 1986.
47. Beal T.M., Hagan H.B., Demuth M. Neural Network Design; PWS, Boston, 1996.
48. Zupan J., Gasteiger J. Neural Networks for Chemists: an Introduction. – Weinheim: VCH, 1993.
49. Blank T.B., Brown S.T. // Anal. Chem. – 1993. – **65**. – P. 3081.
50. Jalali-Heravi M., Fatemi M.H. // J. Chromatogr. A. – 2001. – **915**. – P. 177.
51. Golbraikha A., Tropsha A. // J. Mol. Graphics Model. – 2002. – **20**. – P. 269.
52. Roy P.P., Roy K. // QSAR Comb. Sci. – 2008. – **27**. – P. 302.
53. Kier L.B. // Quant. Struct.-Act. Relat. – 1985. – **4**. – P. 109.
54. Kier L.B. in: Computational Chemical Graph Theory / Ed. D.H. Rouvray – New York: Nova Science Publishers, 1990.
55. Osmialowski K., Halkiewicz J., Radecki A., Kaliszan R. // J. Chromatogr. – 1985. – **346**. – P. 53.
56. Basak S.C., Harriss D.K., Magnuson V.R. // J. Pharm. Sci. – 1984. – **73**. – P. 429.