ЖУРНАЛ СТРУКТУРНОЙ ХИМИИ

*2012. Том 53, № 2* *Март – апрель* *С. 268 – 277*

UDC 548.737

# PREDICTION OF GAS TO WATER PARTITION COEFFICIENT OF SOME ORGANIC COMPOUNDS USING THEORETICALLY DERIVED MOLECULAR DESCRIPTORS

**Z. Dashtbozorgi[1], H. Golmohammadi[2]**

[1]*Young Researchers Club, Central Tehran Branch, Islamic Azad University, Tehran, Iran,*
e-mail: z.dashtbozorgi@gmail.com
[2]*Department of Chemistry, Mazandaran University, Babolsar, Iran*

*Received June, 11, 2010*

An artificial neural network (ANN) is constructed and trained for the prediction of gas to water partition coefficients of various organic compounds. The inputs of this neural network are theoretically derived from molecular descriptors that were chosen by the genetic algorithm-partial least squares (GA-PLS) feature selection technique. These descriptors are: area-weighted surface charge of hydrogen bonding donor atoms (HDCA-2), average bond order of a C atom ($P_C$), Kier flexibility index (Φ), atomic charge weighted partial positively charged surface area (PPSA-3), and difference between atomic charge weighted partial positive and negative surface areas (DPSA-3). By comparing the results obtained from PLS and ANN models, one can see that statistical parameters (Fisher ratio, correlation coefficient, and standard error) of the ANN model are better than those of the PLS model, which indicates that a nonlinear model can simulate more accurately the relationship between the structural descriptors and the partition coefficients of the investigated molecules.

**K e y w o r d s:** artificial neural network, gas to water partition coefficient, genetic algorithm, partial least squares.

## INTRODUCTION

The distribution of a solute between two phases has been an essential topic for theoretical and experimental studies for many years [ 1—3 ]. The ratio of the concentration of solutes distributed between two phases is called the partition coefficient. Partition coefficients have since been used repeatedly to assess the biological activity of chemicals in all areas of science. Their use today is ordinary in the subjects such as medicinal chemistry, agrochemical and pesticide research, formulation science and prediction of environmental effects of chemicals [ 4, 5 ]. The gas to water partition coefficient ($K_W$) is defined as the ratio of a compound concentration in water to its concentration in gas after the partition between two phases reaches equilibrium at a specified temperature. According to this definition, the $K_W$ or log $K_W$ value for a chemical can be calculated as follows:

$$K_W = \frac{C_W}{C_G}, \quad (1)$$

where $C_W$ and $C_G$ are the concentrations of a test chemical in water and gas phases respectively. If the concentration units are the same in water and in the gas phase, for example $mol \cdot dm^{-3}$, then $K_W$ is dimensionless and entirely equivalent to L, the Ostwald solubility coefficient.

There are some analytical techniques to determine partition coefficients [ 6, 7 ]. However, these methodologies are laborious, expensive, or time-consuming and require an adequate amount of pure

compounds, hence not suitable for high-throughput screening of several compounds. Therefore, the theoretical and computational methodologies such as the quantitative structure property relationship (QSPR) would be supportive in the evaluation of the partition coefficients. The advantage of this approach over other methods lies in the fact that the descriptors used can be calculated from the structure alone and are not dependent on any experiment properties. Once the structure of a compound is known, any descriptor can be calculated. Thus, once a reliable model is established, we can use this model to predict the property of a compound, whether it was acquired or not.

A number of attempts to model the relationship between the partition coefficients and the property of organic compounds have been performed. Chen et al. [ 8 ] developed a QSPR model based on nine quantum chemical descriptors computed by PM3 Hamiltonian for the prediction of the logarithm of octanol—air partition coefficients ($\log K_{OA}$) of polychlorinated biphenyls (PCBs). Patel et al. [ 9 ] predicted the dimyristoyl phosphatidyl choline (DMPC)—water partition coefficient ($\log k_{DMPC—W}$) of 49 organic compounds using the QSAR approach and physicochemical descriptors. Toropov et al. [ 10 ] constructed a model to correlate lipid-water partition coefficients of two sets of diverse functional aliphatic and aromatic compounds to their structures. They found that the statistical characteristics of the lipid-water partition coefficient model based on the weighting of the kinds of atoms together with NNC values are better than those based on the weighting of the kinds of atoms together with the Morgan extended connectivity values. Raska et al. [ 11 ] estimated the octanol/water partition coefficient of vitamins and various organic compounds that are not vitamins. An analysis of the correlation weights of their models indicates that vitamins are most likely to interact with water and octanol by more complex mechanisms than various compounds under consideration, which are not vitamins. Fatemi and Karimian [ 12 ] studied the micelle—water partition coefficients of 81 organic compounds in an SDS solution by the quantitative structure—property relationship method. Chen et al. [ 13 ] correlated 209 molecular structure patterns of polychlorinated diphenyl ethers (PCDEs) with their *n*-octanol/water partition coefficient ($\lg K_{ow}$) and sub-cooled liquid water solubilities ($-\log S_{w,l}$) using the stepwise multiple regression (SMR). Finally, Abraham et al. [ 14 ] have developed a LSER method to predict the gas to water partition coefficients for 374 compounds at 310 K through the relatively simple linear equations.

Recently, artificial neural networks (ANNs) have been used for the investigation of a broad variety of chemical problems such as spectral analysis [ 15 ], prediction of dielectric constants [ 16 ], and mass spectral search [ 17 ]. Artificial neural networks have been applied to the QSPR analysis since the late 1980s due to its suppleness in modeling the nonlinear problem, mainly in response to increase accuracy demands. They have been commonly used to predict many physicochemical properties [ 18—20 ]. The main aim of the present work is the development of a QSPR model using a genetic algorithm and ANN to predict the gas to water partition coefficients of various organic compounds.

## METHODOLOGY

**Data set.** The data set of gas to water partition coefficients was taken from the values reported by Sprunger et al. [ 21 ]. The molecules in the data set, including alkyl halides, alcohols, phenols, ethers, esters, ketones, amides, amines, acids, polycyclic hydrocarbons, heterocyclic compounds, and benzene derivatives, are shown in Table 1. The partition coefficients of all molecules included in the data set were obtained under the same conditions. All $\log K_W$ values are dimensionless and refer to a temperature of 298 K. The partition coefficients fall in the range of 2.55 to 11.32 for tetrahydrofuran and phenobarbital respectively. The data set was randomly divided into three separate sections (the training, test, and validation sets) consisting of 57, 12, and 12 members respectively. The training set was used to regulate the parameters of the models; the test set was used to prevent the network from overfitting; and the external validation set was used to assess the prediction ability of constructed models.

**Calculation of descriptors.** The calculation process of the molecular descriptors was described as follows: the molecules were drawn with Hyperchem and then pre-optimized using the MM+ molecular mechanics force field [ 22 ]. A more accurate optimization is then done with the semiempirical PM6 method in MOPAC 6.0 [ 23 ]. All calculations are carried out at a restricted Hartree—Fock level with no configuration interaction. The molecular structures were optimized using the Polak—Ribiere

Table 1

*Data set and the corresponding experimental*, PLS *and* ANN *predicted values of the gas to water partition coefficient*

| Number | Name | log$K_W$ (EXP) | log$K_W$ (PLS) | log$K_W$ (ANN) | Number | Name | log$K_W$ (EXP) | log$K_W$ (PLS) | log$K_W$ (ANN) |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Training set | | | | | |
| 1 | Methanol | 3.74 | 4.10 | 3.62 | 30 | 2-Hydroxybenzoicacid | 5.39 | 6.90 | 5.61 |
| 2 | Ethanol | 3.67 | 3.70 | 3.84 | 31 | 4-Hydroxybenzoicacid | 6.78 | 6.90 | 6.65 |
| 3 | 1-Pentanol | 3.35 | 3.37 | 3.29 | 32 | 2,4-Dihydroxybenzoicacid | 8.39 | 9.49 | 8.57 |
| 4 | 1-Heptanol | 3.09 | 3.25 | 3.36 | 33 | 2-Methyl-1-propanol | 3.30 | 3.59 | 3.48 |
| 5 | Phenol | 4.85 | 4.69 | 4.60 | 34 | 2-Butanol | 3.39 | 3.86 | 3.36 |
| 6 | 2-Nitrophenol | 3.36 | 4.76 | 3.24 | 35 | Pentobarbital | 8.63 | 9.54 | 8.63 |
| 7 | 4-Nitrophenol | 7.81 | 6.73 | 7.62 | 36 | Phenobarbital | 11.32 | 10.68 | 11.04 |
| 8 | 2-Fluorophenol | 3.88 | 4.59 | 4.08 | 37 | Aniline | 4.30 | 4.67 | 4.50 |
| 9 | 4-Fluorophenol | 4.54 | 4.51 | 4.71 | 38 | 4-Methylaniline | 4.09 | 4.69 | 4.37 |
| 10 | 2-Chlorophenol | 3.34 | 4.28 | 3.48 | 39 | 4-Hexylaniline | 3.69 | 4.57 | 3.51 |
| 11 | 4-Chlorophenol | 5.16 | 4.32 | 5.15 | 40 | Pyridine | 3.44 | 3.01 | 3.49 |
| 12 | 2-Bromophenol | 3.71 | 4.70 | 3.89 | 41 | Ephedrine | 6.92 | 4.72 | 6.95 |
| 13 | 4-Bromophenol | 5.23 | 4.65 | 5.17 | 42 | Tetrahydrofuran | 2.55 | 2.53 | 2.42 |
| 14 | 2-Iodophenol | 4.55 | 5.32 | 4.86 | 43 | 4-Aminobenzoicacid | 9.43 | 7.48 | 9.62 |
| 15 | 4-Iodophenol | 5.58 | 5.27 | 5.51 | 44 | 2,4-Dichlorophenol | 3.65 | 3.49 | 3.83 |
| 16 | 2-Methylphenol | 4.31 | 4.78 | 4.01 | 45 | 2,4,6-Trichlorophenol | 3.63 | 3.58 | 3.57 |
| 17 | 4-Methylphenol | 4.50 | 4.71 | 4.38 | 46 | 1-Naphthylamine | 5.35 | 5.30 | 5.16 |
| 18 | 2-Ethylphenol | 4.42 | 3.94 | 4.15 | 47 | 2-Naphthylamine | 5.39 | 5.63 | 5.43 |
| 19 | 2-Methoxyphenol | 4.09 | 4.45 | 4.33 | 48 | 2-Aminophenol | 7.45 | 6.61 | 7.50 |
| 20 | 4-Methoxyphenol | 6.15 | 4.68 | 6.29 | 49 | 2-Hydroxybenzaldehyde | 3.48 | 6.18 | 3.74 |
| 21 | 2,4-Dimethylphenol | 4.41 | 4.32 | 4.55 | 50 | Methyl4-hydroxybenzoate | 6.86 | 4.93 | 6.90 |
| 22 | 2,5-Dimethylphenol | 4.34 | 4.87 | 4.23 | 51 | 1-Nitroso-2-naphthol | 5.36 | 5.78 | 5.51 |
| 23 | 3,4-Dimethylphenol | 4.77 | 4.60 | 4.46 | 52 | 4-Nitrosophenol | 5.51 | 5.55 | 5.36 |
| 24 | 4-Propylphenol | 4.45 | 4.64 | 4.25 | 53 | 2-Hydroxy-3-methoxy- | 5.05 | 5.81 | 4.97 |
| 25 | 1-Naphthol | 5.87 | 5.49 | 5.78 | | benzaldehyde | | | |
| 26 | 2-Naphthol | 5.95 | 5.57 | 6.11 | 54 | Catechol | 7.20 | 6.50 | 7.26 |
| 27 | Aceticacid | 4.88 | 5.44 | 4.78 | 55 | Hydroquinone | 8.87 | 6.66 | 8.69 |
| 28 | Propanoicacid | 4.73 | 5.12 | 4.92 | 56 | Vanillin | 6.42 | 6.33 | 6.22 |
| 29 | Heptanoicacid | 4.20 | 4.02 | 3.98 | 57 | Peracetic acid | 5.49 | 4.62 | 5.45 |
| | | | | Test set | | | | | |
| 58 | 1-Propanol | 3.56 | 3.51 | 3.39 | 64 | 3,5-Dimethylphenol | 4.60 | 4.74 | 4.47 |
| 59 | 1-Hexanol | 3.23 | 3.13 | 3.45 | 65 | Pentanoicacid | 4.45 | 3.86 | 4.61 |
| 60 | 3-Fluorophenol | 4.62 | 4.79 | 4.41 | 66 | 4-Methoxybenozicacid | 5.51 | 5.12 | 5.69 |
| 61 | 3-Bromophenol | 5.11 | 4.19 | 4.97 | 67 | 3-Methyl-1-butanol | 3.24 | 3.99 | 3.36 |
| 62 | 3-Methylphenol | 4.60 | 4.67 | 4.40 | 68 | Hexobarbital | 9.45 | 7.42 | 9.62 |
| 63 | 3-Methoxyphenol | 5.80 | 5.08 | 5.52 | 69 | 2-Nitroso-1-naphthol | 6.44 | 4.19 | 6.28 |
| | | | | Validation set | | | | | |
| 70 | 1-Butanol | 3.46 | 3.49 | 3.45 | 76 | Butanoicacid | 4.63 | 3.83 | 4.91 |
| 71 | 3-Nitrophenol | 7.06 | 6.03 | 6.83 | 77 | 3-Hydroxybenzoicacid | 7.00 | 6.37 | 7.26 |
| 72 | 3-Chlorophenol | 4.80 | 4.21 | 4.68 | 78 | 4-Pentylaniline | 3.82 | 4.62 | 4.01 |
| 73 | 3-Iodophenol | 5.68 | 5.30 | 5.44 | 79 | 2,6-Dichlorophenol | 3.37 | 3.94 | 3.56 |
| 74 | 4-Ethylphenol | 4.50 | 4.66 | 4.33 | 80 | Methyl2-hydroxybenzoate | 2.97 | 4.75 | 3.16 |
| 75 | 2,6-Dimethylphenol | 3.86 | 4.37 | 4.02 | 81 | Resorcinol | 8.35 | 7.37 | 8.65 |

algorithm until the root-mean-square gradient reached 0.001. The resulting geometry was transferred into the CODESSA software that can calculate constitutional, topological, electrostatic, and quantum chemical descriptors. The CODESSA software, developed by the Kartitzky group, enables the calculation of a large number of quantitative descriptors based lonely on the molecular structure information and codes the chemical information into the mathematical form [ 24, 25 ]. CODESSA combines various methods for quantifying the structural information about the molecule with an advanced statistical analysis to establish a quantitative structure-property relationship.

**GA-PLS based variable selection.** GA-PLS is a refined hybrid approach that combines GA [ 26 ] as a commanding optimization method with PLS [ 27 ] as a forceful statistical method for variable selection. GA is inspired by the biological concept of natural selection and evolution, just as the most fit organisms are most likely to survive and be reproduced by crossover together with random mutations of chromosomes in the surviving ones. In GA-PLS, the chromosome and its fitness in the species correspond to a set of variables and internal prediction of the derived PLS model respectively.

In QSPR studies, it is important to obtain a model containing as few variables as possible because this will guide to a simple and interpretable model. Therefore, the quality of a chromosome is determined by both the internal predictivity it gives and the number of variables it uses. In order to enhance the quality of chromosomes in the population, an extra rule is added to GA-PLS following the idea of Leardi et al. [ 28 ]: the best chromosome using the same number of variables is protected unless a chromosome with a lower number of variables gives better internal predictivity. The protected chromosomes in the final population of GA can be considered as the significant combinations of variables.

In this paper, GA-PLS followed Leardi′s method [ 29 ]. Because each GA gives a faintly different model, at least each run is repeated five times to verify the robustness of the predictive ability and importance of the selected model. If some variables (descriptors) are present only in one model, it can be concluded that they have been selected by chance and therefore, they can be ignored in the final model.

**Partial least squares (PLS).** The partial least squares (PLS) regression is a modern technique that generalizes and combines features from the principal component analysis and multiple regression. It is particularly helpful when we need to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors). The PLS regression has acquired a famous position in chemometrics [ 30 ]. One reason for this is that it overcomes the deficiencies of the ordinary least squares (OLS) regression in the case of highly collinear data. Moreover, PLS allows an analysis of the data in terms of independent latent variables or components. These PLS components span a subspace of the regressors (columns of **X**), which is relevant for describing both **X** and response **Y**. Ardent proponents of PLS consider it superior to other biased regression methods [ 31 ]. However, it is unlikely that there is a single superior technique for predictive modeling.

It is assumed that $X$ ($n\times N$) contains the descriptors that can be used for predicting the activities $Y$ ($n\times M$). It is distinguished that PLS decomposes the data matrices $X$ and $Y$ into a two matrix product plus residual in a single process. The matrices $E$ and $F$ contain residuals for $X$ and $Y$ respectively

$$X = TP' + E, \tag{2}$$

$$Y = UQ' + F, \tag{3}$$

where $T$ and $U$ are the score matrices and $P'$ and $Q'$ are the loading matrices for $X$ and $Y$ respectively. These two equations can be written as a multiple regression model

$$Y = XB + G \tag{4}$$

where the matrix $B$ contains the PLS regression coefficients [ 32 ].

The PLS algorithm used in this study was singular value decomposition (SVD)-based PLS. This algorithm was proposed by Lobert et al. in 1987 [ 42 ]. A brief discussion of the SVD-based PLS algorithm can be found in the literature [ 33, 34 ]. The program of PLS modeling based on SVD was written with MATLAB 7 in our laboratory [ 35 ].

**ANN generation.** Artificial Neural Networks (ANNs) are collections of mathematical models that emulate the real neural structure of the brain. In general, ANN is made of individual interrelated simple processing elements called neurons, arranged in a layered structure to form a network that is

capable of performing massively parallel computation. A detailed description of the theory behind a neural network has been adequately described elsewhere [ 36 ].

In the present work, an ANN program was written with MATLAB 7. This network was feed-forward fully connected with the sigmoidal transfer function that has three layers. Descriptors selected by PLS methods were used as inputs of the network, and its output signal represents the gas to water partition coefficients of interested compounds. Thus, this network has five nodes in the input layer and one node in the output layer. The value of each input was divided into its mean value to bring them into a 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 one. These values were optimized during the network training. The back-propagation algorithm was used for the training of the network. Before training, the network parameters would be optimized. These parameters are: number of nodes in the hidden layer, weights and biases learning rates, and the momentum. Procedures for the optimization of these descriptors were reported elsewhere [ 37, 38 ]. Then the optimized network was trained using the training set for the adjustment of weights and biases. 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 the 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 predictive ability of a 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, if it is close to the ideal one. This may imply that the correlation coefficient $R$ between the experimental (actual) $y$ and predicted $\tilde{y}$ properties must be close to 1 and the regression of $y$ against $\tilde{y}$ through the origin, i.e. $y^{r0} = k\tilde{y}$ should be characterized by $k$ close to 1 [ 39 ]. The slope $k$ is calculated as follows:

$$k = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2}. \tag{5}$$

The criteria formulated above may not be sufficient for a QSPR model to be truly predictive. The regression line through the origin defined by $y^{r0} = k\tilde{y}$ (with the intercept set to one) should be close to the optimum regression line $y^r = a\tilde{y} + b$ ($b$ is the intercept). The correlation coefficient for this line $R_0^2$ is calculated as follows:

$$R_0^2 = 1 - \frac{\sum (\tilde{y}_i - y_i^{r0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2}, \tag{6}$$

where $\bar{\tilde{y}}$ is the average value of the observed property and the summation is 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 examine the prediction potential of a model [ 40 ]. This term was defined in the following manner:

$$R_m^2 = R^2(1 - |\sqrt{R^2 - R_0^2}\ |). \tag{7}$$

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

1. A high value of cross-validated $R^2$ ($q^2 > 0.5$).

2. The correlation coefficient $R$ between the predicted and actual properties from an external test set close to 1. $R_0^2$ should be close to $R^2$.

3. The slope of the regression line ($k$) through the origin should be close to 1.

4. $R_m^2$ should be greater than 0.5.

## RESULTS AND DISCUSSION

**Molecular diversity validation.** The basic research themes in chemical database analysis are the diversity of sampling [ 41 ]. The diversity problem involves defining a diverse subset of representative compounds. In this study, the diversity analysis was performed on the data set to make sure that the structures of the training, test or validation sets can represent those of the whole ones. We consider a database of n compounds generated from m highly correlated chemical descriptors $\{X_J\}_{j=1}^{m}$. Each compound $X_i$ is represented as the following vector (eq. 8):

$$X_i = (x_{i1}, x_{i2}, x_{i3}, \ldots x_{im}) \qquad \text{for } i = 1, 2, \ldots, n, \tag{8}$$

where $x_{ij}$ denotes the value of the descriptor $j$ of the compound $X_i$. The collective database $X = \{X_i\}_{i=1}^{N}$ is represented an $n \times m$ matrix of $X$ as follows (Eq. 9):

$$X = (X_1, X_2, \ldots X_N)^T = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix}, \tag{9}$$

where the superscript $T$ denotes the vector/matrix transpose. A distance score $d_{ij}$ for two different compounds $X_i$ and $X_j$ can be measured by the Euclidean distance norm based on the compound descriptors (Eq. 10)

$$d_{ij} = \| X_i - X_j \| = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}. \tag{10}$$

The mean distances of one sample to the remaining ones were computed as follows (Eq. 11):

$$\bar{d}_i = \frac{\sum_{j=1}^{n} d_{ij}}{n-1} \qquad i = 1, 2, \ldots, n. \tag{11}$$

Then the mean distances were normalized within the interval of zero to one. In order to calculate the values of mean distances according to the Eqs. (10) and (11), a MATLAB program was written in our laboratory. This program combines maximum dissimilarity search algorithms and general multidimensional measurements of chemical similarity based on different molecular descriptors. The closer to one the distance, the more diverse to each other the compound is. The mean distances of samples were plotted against log$K_W$ (exp) and are shown in Fig. 1, which illuminates the diversity of the molecules in the training, test, and validation sets. As can be seen from this figure, the structures of the compounds are diverse in all sets, and the training set with a broad representation of the chemistry space was adequate to ensure the model stability, and the diversity of test and validation sets can prove the predictive capability of the model.

**PLS modeling.** The data set and the corresponding observed PLS and ANN predicted values of the gas to water partition coefficients of all molecules studied in this work are shown in Table 1. Table 2 shows the specifications of best PLS model. It can be seen from this table that five descriptors appeared in this model. These descriptors are: area-weighted surface charge of hydrogen bonding donor atoms (HDCA-2), average bond order of a C atom ($P_C$), Kier flexibility index (Φ), atomic charge weighted partial

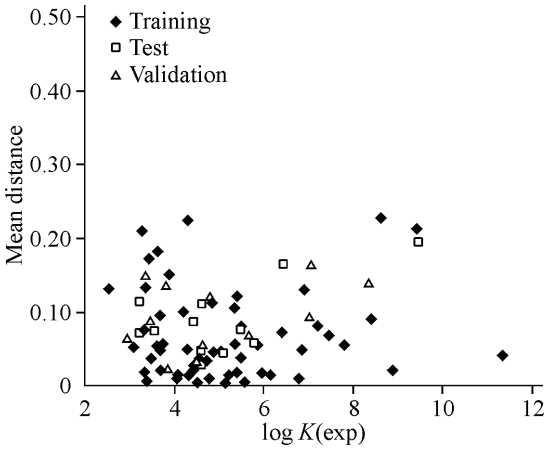


*Fig. 1.* Scatter plot of samples for the training, test, and validation sets

Table 2

*Partial least squares regression coefficients*

| Descriptor | Notation | Coefficient | Mean Effect |
|---|---|---|---|
| Area-weighted surface charge of hydrogen bonding donor atoms | HDCA-2 | 1.5965 | 1.9695 |
| Average bond order of a C atom | $P_C$ | 3.3069 | 3.3911 |
| Kier flexibility index | Φ | –0.2958 | –0.5827 |
| Atomic charge weighted partial positively charged surface area | PPSA-3 | 0.0650 | 1.6813 |
| Difference between atomic charge weighted partial positive and negative surface areas | DPSA-3 | 0.0140 | 0.6886 |
| Constant | | –1.9833 | |

positively charged surface area (PPSA-3), and difference between atomic charge weighted partial positive and negative surface areas (DPSA-3). Each of these descriptors encodes different aspects of the molecular structure. By interpreting the descriptors in the models, it is possible to gain some insight into factors that are likely to relate to the gas to water partition coefficients of the organic compounds.

For the examination of the relative significance and contribution of each descriptor in the model, the mean effect (ME) value was calculated for each descriptor by the following equation:

$$ME_j = \frac{\beta_j \sum_{i=1}^{n} d_{ij}}{\sum_{j}^{m} \beta_j \sum_{i}^{n} d_{ij}}, \tag{12}$$

where $ME_j$ is the mean effect for the considered descriptor $j$, $\beta_j$ is the coefficient of the descriptor $j$, $d_{ij}$ 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. The value and sign of MEs show the relative contribution and the direction of influence of each descriptor on the partition coefficient respectively. As shown in Table 2, the most relevant descriptor based on its mean effect is $P_C$: a charge distribution-related descriptor. This descriptor represents or depends directly on the quantum chemically calculated charge distribution in the molecules, and therefore describes the polar interactions between molecules or their chemical reactivity. The positive coefficient of this descriptor means that as the value of this descriptor increases, the values of log$K_W$ increase. The second and forth relevant descriptors, according to the ME value, are charged partial surface area (CSPA) descriptors (PPSA-3 and DPSA-3). These descriptors are calculated as

$$\mathrm{DPSA3} = \mathrm{PPSA3} - \mathrm{PNSA3}, \tag{13}$$

where PPSA3 and PNSA3 were calculated as follows:

$$\mathrm{PPSA3} = \sum (+SA_i)(Q_i^+), \tag{14}$$

$$\mathrm{PNSA3} = \sum (-SA_i)(Q_i^-), \tag{15}$$

where $(+SA_i)$ and $(-SA_i)$ are the surface area contribution of the $i$th positive or negative atom in the molecule and $Q_i^+$ and $Q_i^-$ are the partial atomic charges for the $i$th positive and negative atoms. These descriptors have been invented by Jurs et al. [ 42 ] in terms of the whole surface area of the molecule and in terms of functional group portions. The descriptors encode the features responsible for polar interactions between molecules. They can be taken as an indication of the solute—solvent interaction that arises through the presence of the positive coefficient that implies the polar compounds have high tendency to dissolve in the aqueous system. The third descriptor is HDCA-2, a quantum-chemical descriptor. This descriptor can be defined as follows:

$$\mathrm{HDCA2} = \sum_{D} \frac{q_D \sqrt{S_D}}{\sqrt{S_{\mathrm{tot}}}} \qquad D \in H_{\text{H-donor}}, \tag{16}$$

where $S_D$ is the solvent-accessible surface area of H-bonding donor H atoms, selected by the threshold charge; $q_D$ is the partial charge on H-bonding donor H atoms, selected by the threshold charge; and $S_{tot}$

is the total solvent-accessible molecular surface area. The positive sign of this descriptor shows that an increase in the value of this descriptor causes an increase in log$K_W$ values. The last descriptor that is present here is Φ, a topological descriptor. This descriptor is calculated as follows:

$$\Phi = \frac{^1\kappa\, ^2\kappa}{N_{SA}}, \quad (17)$$

where $^1\kappa$ and $^2\kappa$ are the Kier shape index, order 1 and the Kier shape index, order 2 respectively and $N_{SA}$ is the number of the skeleton atom. This descriptor describes the atomic connectivity in the molecule [ 43 ] and characterizes the size and shape of the molecule. As we know, the size, shape, and symmetry of molecules play a key role in the processes of the distribution of molecules from one solvent to another. The quantification of the molecular shape and size helps understanding the cavity effect that is the endoergic effect of disrupting solvent—solvent bonds. The negative coefficient of this descriptor means that as the value of this descriptor increases, the values of log$K_W$ decrease.

From the above discussion, it can be seen that all descriptors involved in the QSPR model have physical meaning, and they can account for the structural features that affect the partition coefficients of the interested molecules.

**Neural network modeling.** The next step was the generation of ANN. Before training ANNs, the network parameters, including the number of nodes in the hidden layer, weights and biases learning rates, and momentum values, were optimized. After the optimization of the network parameters, the network was trained using the training set for the adjustment of the weights and biases by the back-propagation algorithm. It is known that a neural network can become over-trained. An over-trained network has usually learned perfectly the stimulus pattern it has seen, but cannot give an accurate prediction for unseen stimuli, and it is no longer able to generalize. There are numerous methods for overcoming 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$K_W$ 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. Results obtained showed overtraining began after 35 000 iterations.

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

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}, \quad (18)$$

where $y_i$ and $\hat{y}_i$, respectively are the measured and predicted values of the dependent variable (gas to water partition coefficient), $\bar{y}$ is the averaged value of the dependent variable of the training set and the summations cover all the compounds. The calculated value of $q^2$ was 0.984.

Table 1 represents the experimental, PLS and ANN calculated values of gas to water partition coefficients for the training, test, and validation sets. The statistical parameters obtained by ANN and PLS models for these sets are shown in Table 3. The standard errors of training, test, and validation sets for the PLS model are 0.907, 0.828, and 0.799 respectively, which would be compared with the

Table 3

*Statistical parameters obtained using the* ANN *and* PLS *models* [a]

| Model | $SE_c$ | $SE_t$ | $SE_v$ | $R_c$ | $R_t$ | $R_v$ | $F_c$ | $F_t$ | $F_v$ |
|---|---|---|---|---|---|---|---|---|---|
| ANN | 0.175 | 0.194 | 0.216 | 0.995 | 0.994 | 0.993 | 5751 | 833 | 678 |
| PLS | 0.907 | 0.828 | 0.799 | 0.864 | 0.886 | 0.895 | 162 | 36 | 40 |

[a] 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.

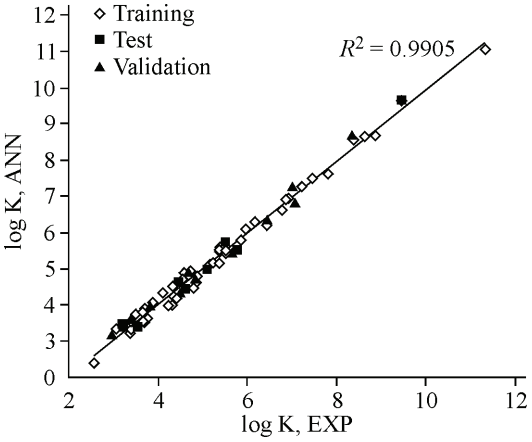


*Fig. 2.* Plot of the ANN calculated gas to water partition coefficient against experimental values

values of 0.175, 0.194, and 0.216 respectively for the ANN model. A comparison between these values and other statistical parameters in Table 3 reveals the superiority of the ANN model over PLS ones. The key strength of neural networks, unlike regression analysis, is their ability to flexible mapping of the selected features by manipulating their functional dependence implicitly.

The statistical values of validation set for the ANN model was characterized by $q^2 = 0.984$, $R^2 = 0.986$ ($R = 0.993$), $R_0^2 = 0.984$, $R_m^2 = 0.956$, and $k = 0.987$. These values and other statistical parameters, which are shown in Table 3, reveal the high predictive ability of the model. Fig. 2 shows the plot of the ANN predicted versus experimental values for the gas to water partition coefficients of all molecules in the data set.

## CONCLUSIONS

In this research, the QSPR models for the prediction of gas to water partition coefficients of 81 organic compounds were constructed by PLS and ANN methods. For each compound a variety of descriptors in 5 classes using the CODESSA software were calculated. The best set of calculated descriptors was selected by the PLS method. Good agreement between the experimental results and the predicted values confirms the validity of the obtained models. The calculated statistical parameters of these models reveal the superiority of ANN over the PLS model. The result shows that the ANN model can describe accurately the relationship between the structural parameters and gas to water partition coefficients of organic compounds.

*REFERENCES*

1. *Leo A.* // Chem. Rev. – 1993. – **1281**. – P. 93.
2. *Leo A., Hansch C., Elkins D.* // Chem. Rev. – 1971. – **525**. – P. 71.
3. *Nasal A., Szmtawska M., Bucinski A., Salesman R.J.* // J. Chromatogr. A. – 1995. – **83**. – P. 692.
4. *Taylor P.J.* in: C.A. Ramsden. (Ed.) Comprehensive Medicinal Chemistry, vol. 4. – Oxford: Pergamon Press, 1990.
5. *Livingstone D.J.* // J. Chem. Inf. Comput. Sci. – 2000. – **40**. – P. 195.
6. *Sato A., Nakajima T.* // Toxicol. Appl. Pharmacol. – 1979. – **47**. – P. 41.
7. *Gargas M.L., Burgess R.J., Voisard D.E., Cason G.H., Andersen M.E.* // Toxicol. Appl. Pharmacol. – 1989. – **98**. – P. 87.
8. *Chen J., Xue X., Schramm K.W. et al.* // Chemosphere. – 2002. – **48**. – P. 535.
9. *Patel H., Schultz T.W., Cronin M.T.D.* // J. Mol. Struct. (Theochem). – 2002. – **593**. – P. 9.
10. *Toropov A.A., Roy K.* // J. Chem. Inf. Comput. Sci. – 2004. – **44**. – P. 179.
11. *Raska I. Jr., Toropov A.* // Eur. J. Medicin. Chem. – 2006. – **41**. – P. 1271.
12. *Fatemi M.H., Karimian F.* // J. Colloid, Interface Science. – 2007. – **314**. – P. 665.
13. *Chen J., Zeng X.L., Wang Z.Y. et al.* // Science of the Total Environment. – 2007. – **382**. – P. 59.
14. *Abrahama M.H., Ibrahim A., Acree W.E. Jr.* // Fluid Phase Equilibria. – 2007. – **251**. – P. 93.
15. *Xing W.L., He X.W.* // Anal. Chim. Acta. – 1997. – **349**. – P. 283.
16. *Jalali-Heravi M., Fatemi M.H.* // J. Chromatogr. A. – 2001. – **915**. – P. 177.
17. *Fatemi M.H., Jalali-Heravi M., Konoz E.* // Anal. Chim. Acta. – 2003. – **486**. – P. 101.
18. *Dashtbozorgi Z., Golmohammadi H.* // Eur. J. of Medicin. Chem. – 2010. – **45**. – P. 2182.
19. *Golmohammadi H., Safdari M.* // Microchem. J. – 2010. – **95**. – P. 140.
20. *Golmohammadi H.* // J. Comput. Chem. – 2009. – **30**. – P. 2455.
21. *Sprunger L.M., Proctor A., Acree W.E. Jr. et al.* // Fluid Phase Equilibria. – 2008. – **270**. – P. 30.
22. Hyperchem, re. 4. for Windows, Autodesk, Sansalito, CA, 1995.
23. *Stewart J.J.P.* Semiempirical Molecular Orbital Program; QCPE, 445, 1983, Version 6, 1990.

24. *Katritzky A.R., Labadov V.S., Carelson M.* CODESSA Training Manual, University of Florida, Gainesvill, 1995.
25. *Katritzky A.R., Labadov V.S., Carelson M.* CODESSA Version 1 Reference Manual, University of Florida, Gainesville, Florida, 1994.
26. *Goldberg D.E.* Genetic Algorithms in Search, Optimization, Machine learning, Addison-Wesley, New York, 1989.
27. *Hoskuldsson A.* Prediction Methods in Science, Technology Vol. 1: Basic Theory, Thur Publishing, Denmark, 1996.
28. *Leardi R., Boggia R., Terrile M.* // J. Chemom. – 1992. – **6**. – P. 267.
29. *Leardi R., Gonzalez A.L.* // Chemom. Intell. Lab. Sys. – 1998. – **41**. – P. 195.
30. *Martens H., Næs T.* // Multivariate Calibration, Wiley: Chichester, 1989.
31. *Hoskuldsson A.* // Chemom. Intell. Lab. Syst. – 1992. – **14**. – P. 139.
32. *Wold S., Sjostorm M., Eriksson L.* // Chemom. Intell. Lab. Syst. – 2001. – **58**. – P. 109.
33. *Lorber A., Wangen L., Kowalsky B.R.* // J. Chemom. – 1987. – **1**. – P. 19.
34. *Hoskuldsson A.* // Chemom. Intell. Lab. Syst. – 2001. – **55**. – P. 23.
35. MATLAB 7.0, The Mathworks Inc., Natick, MA, USA, http://www.mathworks.com
36. *Zupan J., Gasteiger J.* // Neural Network in Chemistry, Drug Design; Wiley-VCH. Weinheim, 1999.
37. *Blank T.B., Brown S.T.* // Anal. Chem. – 1993. – **65**. – P. 3081.
38. *Jalali-Heravi M., Fatemi M.H.* // J. Chromatogr. A. – 2001. – **915**. – P. 177.
39. *Golbraikh A., Tropsha A.* // J. Mol. Graphics Modell. – 2002. – **20**. – P. 269.
40. *Roy P.P., Roy K.* // QSAR Comb. Sci. – 2008. – **27**. – P. 302.
41. *Maldonado A.G., Doucet J.P., Petitjean M. et al.* // Mol. Divers. – 2006. – **10**. – P. 39.
42. *Stanton D.T., Jurs P.C.* // Anal. Chem. – 1990. – **62**. – P. 2323.
43. *Kier L.B., Hall L.* Molecular connectivity in structure—activity analysis. – Letchworth: Research Studies Press, 1986.