EurJChem

European Journal of Chemistry

Journal homepage: www.eurjchem.com

EurJChem

Prediction of *n*-octanol-water partition coefficient for polychlorinated biphenyls from theoretical molecular descriptors

Majid Safdari[a] and Hassan Golmohammadi[b,*]

[a] *Department of Chemistry, Esfahan University of Technology, Esfahan, IR-84156-8311, Iran*
[b] *Department of Chemistry, Mazandaran University, Babolsar, IR-47415, Iran*

**Corresponding author at: Department of Chemistry, Mazandaran University, Babolsar, IR-47415, Iran. Tel.: +98.21.77635178; fax: +98.21.77635178.*
E-mail address: hassan.gol@gmail.com (H. Golmohammadi).

ARTICLE INFORMATION

Received: 12 April 2010
Received in revised form: 24 July 2010
Accepted: 28 July 2010
Online: 31 December 2010

KEYWORDS

Quantitative Structure-Property Relationship
n-Octanol–water partition coefficient
Artificial neural network
Partial least squares
Genetic algorithm

ABSTRACT

A quantitative structure–property relationship (QSPR) study was performed to develop models that relate the structures of 133 polychlorinated biphenyls to their *n*-octanol–water partition coefficients (log K_{ow}). Molecular descriptors were derived solely from 3D structures of the molecules. The genetic algorithm-partial least squares (GA-PLS) method was applied as a variable selection tool. The partial least square (PLS) method was used to select the best descriptors and the selected descriptors were used as input neurons in neural network model. These descriptors are: Balabane index (J), XY Shadow (SXY), Kier shape index (order 3) (3κ), Wiener index (W) and Maximum valency of C atom (VmaxC). The use of descriptors calculated only from molecular structure eliminates the need for experimental determination of properties for use in the correlation and allows for the estimation of log Kow for molecules not yet synthesized. The root mean square errors for ANN predicted partition coefficients of training, test and external validation sets were 0.063, 0.112 and 0.126, respectively, while these values are 0.230, 0.164 and 0.297 for the PLS model, respectively. Comparison between these values and other statistical parameters for these two models revealed the superiority of the ANN over the PLS model.

1. Introduction

Polychlorinated biphenyls (PCBs) are persistent organic contaminants and widespread environmental pollutants that are found at an appreciable concentration in the polar regions presumably, as a result of long-range atmospheric transport [1]. PCBS are a family of 209 congeners each of which consists of two benzene rings and 1–10 chlorine atoms, are ubiquitous in the global environment because of their biological and chemical stability and their historical widespread use in the power-generation industry [2,3]. Toxicological effects of exposure to PCBs include hepatotoxicity, immunotoxicity, and reproductive problems, as well as respiratory, mutagenic and carcinogenic effects [3,4]. Although the manufacture and use of PCBs have been banned in many countries, the compounds remain serious environmental contaminants due to unceasing release from hazardous waste sites. Risk assessment of PCBs involves their behavior in environment, so it is important to understand their physico-chemical properties.

The octanol-water partition coefficient expressed as log K_{ow} is an important property for various applications in pharmacology, toxicology and medicinal chemistry [5]. Log K_{ow} is used to model partitioning of chemicals between the lipophilic membrane and the relative hydrophobic cellular cytoplasmic material. Log K_{ow} quantities hydrophobicity of chemicals and is important both for predicting pharmacokinetics and pharmacodynamics of drugs and toxicants [6]. Lipophilicity is traditionally measured in the octanol-water system. Log K_{ow} values have been shown to be generally satisfactory for modeling protein binding and lipophilic interactions with biological membranes consisting largely of protein [7]. The K_{ow} or log K_{ow} is defined as the ratio of a compound's concentration in octanol to its concentration in water after the partition between two phases reaches equilibrium at a specified temperature. According to this definition, log K_{ow} value for a chemical can be calculated as follows:

$$\text{Log } K_{ow} = \text{Log} \frac{C_{org}}{C_{aq}} \qquad (1)$$

C_{org} and C_{aq} are the concentrations of test chemical in organic and aqueous phases, respectively. Log K_{ow} can be used to explain certain bioconcentration factors (BCFs) and bioaccumulation factors (BAFs) [8]. Although other partition coefficients such as octanol–air partition coefficient were discovered to influence the BCFs of organic chemicals recently, log K_{ow} still plays an important role in governing PCBs. Several methods have been described in the literature for the estimation of the octanol/water partition coefficient [9,10].

In view of fact that experimental determination of the partition coefficients of a large set of compounds is time-consuming and required high-purity samples and skilled operators, the development of an alternative method such as quantitative structure-property relationship (QSPR) would be useful for the theoretical calculation of log K_{ow} values. The QSPR method, which establishes the correlation by expressing a specific physicochemical property of target compounds using appropriate molecular descriptors, is applicable for modeling of log K_{ow} of PCBs. QSPR models can provide insight into the

ISSN 2153-2249 (Print) / ISSN 2153-2257 (Online) © 2010 EURJCHEM
DOI:10.5155/eurjchem.1.4.266-275.59

major molecular structural factors that affect the specific physical-chemical property of chemicals [11,12]. There have been numerous reports on QSPR studies of octanol-water partition coefficient. Lu *et al.* [13] predicted octanol-water partition coefficients of 133 polychlorinated biphenyls using heuristic method (HM) implemented in CODESSA. In order to indicate the influence of different molecular descriptors on log K_{ow} values and well understand the important structural factors affecting the experimental values, they built three multivariable linear models derived from three groups of different molecular descriptors. Padmanabhan *et al.* [14] developed a QSPR model for predicting the lipophilic behaviour (log K_{ow}) of the data set containing various polychlorinated biphenyl (PCB) congeners using the conceptual density functional theory based global reactivity parameter such as electrophilicity index (ω) along with energy of lowest unoccupied molecular orbital (E_{LUMO}) and number of chlorine substituents (N_{Cl}) as descriptors. Puzyn and Falandysz [15] estimated octanol-water and octanol-air partition coefficients of 75 chloronaphtalene congeners by means of six chemometrics approaches. Li *et al.* [16] estimated octanol–water partition coefficients of polybrominated diphenyl ethers (PBDEs) using the partial least-squares regression method.

Recently artificial neural networks (ANNs) have been used for investigation of wide variety of chemical problems such as spectral analysis [17], prediction of dielectric constants [18] and mass spectral search [19]. ANNs have been applied to QSPR analysis since the late 1980s due to their flexibility in modeling of nonlinear problem, mainly in response to increase accuracy demands. They have been widely used to predict many physicochemical properties [20-23]. In this investigation, the calculated descriptors from structures were used lonely to predict the octanol-water partition coefficients of 133 Polychlorinated biphenyls (PCBs) using the ANN and QSPR methods.

2. Methodology

2.1. Data set

The data set in this investigation was extracted from the values reported by Padmanabhan *et al.* [14]. The names of molecules in data set including 133 PCBs are shown in Table 1. The octanol-water partition coefficients fall in the range of 4.63 to 7.94 for 4-chlorobiphenyl and 2,2’,3,3’,4,4’,5,5’,6-nonachloro biphenyl, respectively. The data set was randomly divided into three groups including training, test and external validation set, which consists of 83, 25 and 25 molecules, respectively. The training set was used to adjust the parameters of models; the test set was used for monitoring the extent of overtraining and external validation set was used for evaluation of the prediction power of obtained model.

2.2. Descriptor calculation

One important step in QSPR investigation is the numerical representation of the chemical structure (often called molecular descriptors). The built model’s performance and accuracy of the results obtained are strongly dependent on the way that descriptors were performed. Due to diversity of the molecules studied, different descriptors were calculated. The calculation process of the molecular descriptors was described as follows: molecules were drawn with Hyperchem package (Version 7) [24] and then pre-optimized using MM^+ molecular mechanics force field. The final geometries of the minimum energy conformation were obtained by more precise optimization with the semi-empirical AM1 method, applying a gradient limit of 0.01 kcal/Å as a stopping criterion for optimized structures. A more precise optimization is then done with the semiempirical AM1 method in Mopac (Version 6) [25]. All calculations are 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 [26,27] to calculate five classes of descriptors including constitutional; geometrical; topological; electrostatic and quantum-chemical descriptors. The software CODESSA, developed by Kartitzky group, enables the calculation of a large number of quantitative descriptors based lonely on the molecular structure information and codes chemical information into mathematical form [26,27]. CODESSA combines diverse methods for quantifying the structural information about the molecule with advanced statistical analysis to establish quantitative structure-property relationship.

Some of the descriptors generated for each compound, encode similar information about the molecule of interest, therefore it was desirable to test each descriptor and eliminate those that show high correlation (R>0.95) with each other. A total of 123 out of 476 descriptors showed high correlation and were removed from the next generation. Subsequently genetic algorithm-partial least squares (GA-PLS) variable subset selection method was used for selection of important descriptors. Since the number of descriptors considered is large, a suitable feature selection method should be combined with a proper feature mapping technique. In the present work we have considered GA-PLS as a feature selection tool and PLS and ANN were employed for feature mapping.

2.3. GA–PLS based variable selection

GA-PLS is a sophisticated hybrid approach that combines GA [28] as a powerful optimization method with PLS [29] as a robust 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 [30].

GA-PLS consists of three basic steps. (1) A chromosome is presented by a binary bit string and initial population of chromosomes is created in random way. (2) A value for the fitness function of each chromosome is evaluated by the internal predictivity of PLS. (3) According to the values of the fitness function, the chromosomes of the next generation are reproduced by selection, cross over and mutation operations.

In QSPR studies, it is important to obtain a model containing as few variables as possible because this will lead 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 increase quality of chromosomes in the population, an extra rule is added to GA-PLS following the idea of Leardi *et al.* [31]: 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 regarded as the important combinations of variables.

In this paper, GA-PLS followed Leardi’s method [32]. The size of population is 30, the probability of cross over is 0.5, the probability of mutation is 0.01 and the number of evaluation is 200. For each set of data 100 runs were performed. Because each GA gives a slightly different model, at least each run repeat 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 selected by chance and therefore, they can be disregarded in the final model.

Table 1. Data set and corresponding observed and predicted values of *n*-octanol-water partition coefficient [a].

Number	Name	log K_{ow} (EXP)	log K_{ow} (PLS)	log K_{ow} (ANN)
Training set				
1	3-Chlorobiphenyl	4.66	4.80	4.63
2	4-Chlorobiphenyl	4.63	4.90	4.68
3	2,2’-Dichlorobiphenyl	4.72	4.73	4.78
4	2,4-Dichlorobiphenyl	5.15	5.05	5.09
5	3,3’-Dichlorobiphenyl	5.27	5.35	5.19
6	4,4’-Dichlorobiphenyl	5.23	5.57	5.28
7	2,2’,3-Trichlorobiphenyl	5.12	5.05	5.20
8	2,2’,5-Trichlorobiphenyl	5.33	5.23	5.38
9	2,2’,6-Trichlorobiphenyl	5.04	4.95	4.98
10	2,3,4-Trichlorobiphenyl	5.68	5.29	5.59
11	2,3’,4-Trichlorobiphenyl	5.54	5.75	5.63
12	2,4,4’-Trichlorobiphenyl	5.71	5.86	5.76
13	2,4’,6-Trichlorobiphenyl	5.24	5.38	5.32
14	2,3’,4’-Trichlorobiphenyl	5.71	5.62	5.63
15	2,3’,5’-Trichlorobiphenyl	5.71	5.61	5.65
16	2,2’,3,3’-Tetrachlorobiphenyl	5.67	5.57	5.62
17	2,2’,3,5’-Tetrachlorobiphenyl	5.73	5.69	5.80
18	2,2’,3,6-Tetrachlorobiphenyl	4.84	5.39	4.85
19	2,2’,3,6’-Tetrachlorobiphenyl	4.84	5.40	4.91
20	2,2’,4,4’-Tetrachlorobiphenyl	5.94	6.02	5.99
21	2,2’,4,6-Tetrachlorobiphenyl	5.75	5.58	5.68
22	2,2’,5,5’-Tetrachlorobiphenyl	5.79	5.81	5.88
23	2,2’,6,6-Tetrachlorobiphenyl	5.24	5.33	5.29
24	2,3,3’,4-Tetrachlorobiphenyl	6.10	5.92	6.01
25	2,3,4,4’-Tetrachlorobiphenyl	6.24	6.01	6.19
26	2,3,4’,6-Tetrachlorobiphenyl	5.76	5.81	5.81
27	2,3’,4,4’-Tetrachlorobiphenyl	5.98	6.22	6.02
28	2,3’,4,5-Tetrachlorobiphenyl	6.32	6.06	6.28
29	2,3’,4’,6-Tetrachlorobiphenyl	5.76	5.95	5.83
30	2,4,4’,6-Tetrachlorobiphenyl	6.03	6.18	6.11
31	2,2’,3,4,4’-Pentachlorobiphenyl	6.18	6.37	6.24
32	2,2’,3,4,6-Pentachlorobiphenyl	6.50	5.89	6.42
33	2,2’,3,4,6’-Pentachlorobiphenyl	5.60	5.95	5.69
34	2,2’,3,5,5’-Pentachlorobiphenyl	6.32	6.16	6.28
35	2,2’,3,5,6-Pentachlorobiphenyl	6.06	5.82	6.07
36	2,2’,3,5’,6-Pentachlorobiphenyl	5.92	5.95	5.91
37	2,2’,3,4’,5’-Pentachlorobiphenyl	6.30	6.15	6.22
38	2,2’,4,4’,5-Pentachlorobiphenyl	6.41	6.49	6.41
39	2,2’,4,5’,6-Pentachlorobiphenyl	6.11	6.19	6.20
40	2,3,3’,4,4’-Pentachlorobiphenyl	6.79	6.64	6.77
41	2,3,3’,4,5-Pentachlorobiphenyl	6.92	6.29	6.88
42	2,3,3’,5’,6-Pentachlorobiphenyl	6.45	6.24	6.41
43	2,3,4,4’,5-Pentachlorobiphenyl	6.71	6.47	6.68
44	2,3,4,4’,6-Pentachlorobiphenyl	6.44	6.34	6.51
45	2,3’,4,4’,5-Pentachlorobiphenyl	6.57	6.61	6.49
46	2,3’,4,5,5’-Pentachlorobiphenyl	6.30	6.69	6.37
47	2,3’,4,5’,6-Pentachlorobiphenyl	6.42	6.47	6.35
48	2,2’,3,3’,4,5’-Hexachlorobiphenyl	7.30	6.71	7.29
49	2,2’,3,3’,4,6-Hexachlorobiphenyl	6.78	6.41	6.69
50	2,2’,3,3’,5,6-Hexachlorobiphenyl	6.20	6.33	6.25
51	2,2’,3,3’,5,6’-Hexachlorobiphenyl	6.32	6.46	6.42
52	2,2’,3,4,4’,5-Hexachlorobiphenyl	6.82	6.76	6.91
53	2,2’,3,4,4’,6’-Hexachlorobiphenyl	6.58	6.67	6.67
54	2,2’,3,4,5,5’-Hexachlorobiphenyl	6.75	6.69	6.81
55	2,2’,3,4’,5,5’-Hexachlorobiphenyl	6.85	6.81	6.91
56	2,2’,3,4’,5’,6-Hexachlorobiphenyl	6.41	6.52	6.50
57	2,2’,4,4’5,5’-Hexachlorobiphenyl	6.80	6.95	6.88
58	2,2’,4,4’,6,6’-Hexachlorobiphenyl	6.54	6.72	6.60
59	2,3,3’,4,4’,5-Hexachlorobiphenyl	7.44	6.92	7.38
60	2,3,3’,4’,5,6-Hexachlorobiphenyl	6.78	6.71	6.85
61	2,3,3’,5,5’,6-Hexachlorobiphenyl	7.00	6.62	6.95
62	2,3’,4,4’,5,5’-Hexachlorobiphenyl	7.29	7.10	7.27
63	3,3’,4,4’,5,5’-Hexachlorobiphenyl	7.55	7.35	7.52
64	2,2’,3,3’,4,5,6’-Heptachlorobiphenyl	6.85	6.91	6.90
65	2,2’,3,3’,4,5’,6-Heptachlorobiphenyl	6.92	7.05	6.88
66	2,2’,3,3’,4,5’,6’-Heptachlorobiphenyl	6.73	6.98	6.77
67	2,2’,3,3’,5,5’,6-Heptachlorobiphenyl	6.85	6.98	6.92
68	2,2’,3,4,4’,5,5’-Heptachlorobiphenyl	7.21	7.37	7.24
69	2,2’,3,4,4’,5,6-Heptachlorobiphenyl	7.13	7.04	7.09
70	2,2’,3,4,4’,5,6’-Heptachlorobiphenyl	6.92	7.15	6.99
71	2,2’,3,4,4’,5’,6-Heptachlorobiphenyl	7.04	7.18	6.99
72	2,2’,3,4,5,5’,6-Heptachlorobiphenyl	6.99	6.90	6.91
73	2,2’,3,4’,5,6,6’-Heptachlorobiphenyl	6.78	6.94	6.72
74	2,3,3’,4,4’,5,5’-Heptachlorobiphenyl	7.72	7.47	7.68
75	2,3,3’,4,4’,5’,6-Heptachlorobiphenyl	7.21	7.36	7.30
76	2,3,3’,4’,5,5’,6-Heptachlorobiphenyl	7.21	7.23	7.20
77	2,2’,3,3’,4,4’,5,5’-Octachlorobiphenyl	7.62	7.78	7.66
78	2,2’,3,3’,4,4’,5,6-Octachlorobiphenyl	7.35	7.47	7.42
79	2,2’,3,4,4’,5,5’,6-Octachlorobiphenyl	7.49	7.59	7.52
80	2,2’,3,4,4’,5,6,6’-Octachlorobiphenyl	7.48	7.45	7.42
81	2,3,3’,4,4’,5,5’,6-Octachlorobiphenyl	7.62	7.75	7.70

Table 1. *(Continued).*

Number	Name	log K_{ow} (EXP)	log K_{ow} (PLS)	log K_{ow} (ANN)
82	2,2',3,3',4,4',5,5',6-Nonachlorobiphenyl	7.94	8.05	7.88
83	2,2',3,3',4,4',5,6,6'-Nonachlorobiphenyl	7.88	7.90	7.79
Test set				
84	2,3'-Dichlorobiphenyl	4.84	4.96	4.90
85	3,4'-Dichlorobiphenyl	5.15	5.47	5.27
86	2,2',4-Trichlorobiphenyl	5.39	5.32	5.45
87	2,3,4'-Trichlorobiphenyl	5.29	5.53	5.42
88	2,3',5-Trichlorobiphenyl	5.65	5.48	5.50
89	2,2',3,4-Tetrachlorobiphenyl	5.79	5.58	5.67
90	2,2',4,5-Tetrachlorobiphenyl	5.69	5.71	5.82
91	2,2',4,6'-Tetrachlorobiphenyl	5.51	5.63	5.64
92	2,3,4',5-Tetrachlorobiphenyl	6.10	6.06	6.02
93	2,3',4,6-Tetrachlorobiphenyl	6.03	5.83	5.99
94	2,3',4',5'-Tetrachlorobiphenyl	5.98	5.99	5.96
95	2,2',3,4,5-Pentachlorobiphenyl	6.38	6.04	6.24
96	2,2',3,4',5-Pentachlorobiphenyl	6.32	6.32	6.29
97	2,2',3,4',6'-Pentachlorobiphenyl	6.04	6.08	6.12
98	2,3,3',4',6-Pentachlorobiphenyl	6.20	6.27	6.38
99	2,3,4',5,6-Pentachlorobiphenyl	6.39	6.27	6.47
100	2,3',4,4',5'-Pentachlorobiphenyl	6.64	6.62	6.51
101	2,2',3,3',4,6'-Hexachlorobiphenyl	6.20	6.45	6.34
102	2,2',3,4,5,6'-Hexachlorobiphenyl	6.56	6.37	6.44
103	2,2',3,5,5',6-Hexachlorobiphenyl	6.42	6.43	6.37
104	2,3,3',4,4',6-Hexachlorobiphenyl	6.78	6.80	6.95
105	2,2',3,3',4,4',5-Heptachlorobiphenyl	7.08	7.22	7.21
106	2,2',3,3',4,6,6'-Heptachlorobiphenyl	6.55	6.78	6.71
107	2,3,3',4,4',5,6-Heptachlorobiphenyl	7.08	7.19	7.19
108	2,2',3,3',4,4',5,6'-Octachlorobiphenyl	7.43	7.57	7.53
Validation set				
109	2,3-Dichlorobiphenyl	4.99	4.96	4.85
110	3,4-Dichlorobiphenyl	5.23	5.47	5.13
111	2,4'-Dichlorobiphenyl	5.09	5.32	5.18
112	2,3,3'-Trichlorobiphenyl	5.60	5.53	5.43
113	2,3,6-Trichlorobiphenyl	5.44	5.48	5.28
114	2,4',5-Trichlorobiphenyl	5.68	5.58	5.57
115	2,2',3,4'-Tetrachlorobiphenyl	5.72	5.71	5.87
116	2,2',4,5'-Tetrachlorobiphenyl	5.87	5.63	5.98
117	2,2',5,6'-Tetrachlorobiphenyl	5.55	6.06	5.36
118	2,3,5,6-Tetrachlorobiphenyl	5.96	5.83	5.80
119	2,4,4',5-Tetrachlorobiphenyl	6.10	5.99	6.23
120	2,2',3,3',6-Pentachlorobiphenyl	5.60	6.04	5.75
121	2,2',3,4,5'-Pentachlorobiphenyl	6.23	6.32	6.34
122	2,2',3,4',6-Pentachlorobiphenyl	5.87	6.08	6.02
123	2,2',4,4',6-Pentachlorobiphenyl	6.23	6.27	6.37
124	2,3,3',5,6-Pentachlorobiphenyl	6.41	6.27	6.23
125	3',4,4',6-Pentachlorobiphenyl	6.40	6.62	6.55
126	2,2',3,3',4,5-Hexachlorobiphenyl	6.76	6.45	6.89
127	2,2',3,4,4',5'-Hexachlorobiphenyl	6.73	6.37	6.84
128	2,2',3,4,5',6-Hexachlorobiphenyl	6.45	6.43	6.38
129	2,2',4,4',5,6'-Hexachlorobiphenyl	6.65	6.80	6.77
130	2,3,3',4',5',6-Hexachlorobiphenyl	6.63	7.22	6.75
131	2,2',3,3',4,5,5'-Heptachlorobiphenyl	7.21	6.78	7.13
132	2,2',3,3',5,6,6'-Heptachlorobiphenyl	6.41	7.19	6.53
133	2,3,3',4,5,5',6-Heptachlorobiphenyl	7.21	7.57	7.10

[a] EXP refers to experimental; PLS refers to partial least squares; ANN refers to artificial neural network.

2.4. Partial least squares (PLS)

The PLS method takes into account information of dependent variables during the decomposition of the independent variables data matrix. Assume that X represents independent variables (X is a matrix) and Y represents dependent variables (Y is a vector). Then a brief description of computations is given as follows.

$$X = TP^{T} + E \qquad (2)$$

$$Y = QS^{T} + F \qquad (3)$$

The matrices E and F contain residual for X and Y, respectively. T and P are score and loading matrices associated with the X, Q and S are the score and loading of Y and superscript T indicates the transposed matrix.

The relationship between scores and dependent variable is obtained from:

$$Y = TBQ^{T} + F \qquad (4)$$

Where B is the matrix of the regression coefficient obtained by a least squares procedure.

The PLS algorithm used in this study was the singular value decomposition (SVD)-based PLS. This algorithm was proposed by Lobert *et al.* in 1987 [33]. A brief discussion of the SVD-based PLS algorithm can be found in the literature [34-36]. The program of PLS modeling based on SVD was written with MATLAB 7 in our laboratory [37]. The PLS regression was run on the data matrices containing the descriptors selected by GA.

2.5. Neural network construction

Artificial neural networks (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 the selected accuracy. A detailed description of the theory behind a neural network has been adequately described elsewhere [38-40]. Therefore, only the points relevant to this work are described here. A fundamental procession 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 Eq. (5):

$$O_j = 1/[1+\exp(-X)] \tag{5}$$

where X is defined by the following equation:

$$X = \sum W_{ji}O_i + B_j \tag{6}$$

In Eq. 6, 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 feed-forward 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.

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 Eqs. (7) and (8).

$$\Delta W_{kj}(n) = \eta\delta_{pk}O_{pj} + \alpha\,\Delta W_{kj}(n-1) \tag{7}$$

$$\Delta B_{kj}(n) = \gamma\,\delta_{pk}O_{pj} \tag{8}$$

In these equations, ΔW_{kj} and ΔB_{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; δ_{pk} is the error term obtained from the differences between the output and the expected value. The parameters η and γ are learning rate of the weight and bias, respectively; α represents the momentum and *n* and *n*-1 refer to the present and the previous iterations, respectively.

Equations similar to the Eqs. (7) and (8) were used to adjust weights and biases connecting the hidden layers to the input one. The criterion for the stopping of 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 that has three layers with sigmoidal transfer function. Descriptors selected by PLS methods were used as inputs of network and its output signal represent the *n*-octanol-water partition coefficients of interested compounds. Thus this network has five nodes in input layer and one node in output layer. The value of each input was divided into its 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 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 [41,42]. Then the optimized network was trained using 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.

2.6. Estimation of the predictive ability of a QSPR model

For the optimized QSPR model several parameters were selected to test prediction ability of the model. A real QSPR model may have a high predictive ability, if it is close to 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 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 [43]. Slopes k and k' are calculated as follows:

$$k = \frac{\sum y_i\tilde{y}_i}{\sum \tilde{y}_i^2} \tag{9}$$

$$k' = \frac{\sum y_i\tilde{y}_i}{\sum y_i^2} \tag{10}$$

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 one) 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'^2_0 are calculated as follows:

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

$$R'^2_0 = 1 - \frac{\sum(y_i - \tilde{y}_i^{r0})^2}{\sum(y_i - \bar{y})^2} \tag{12}$$

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.

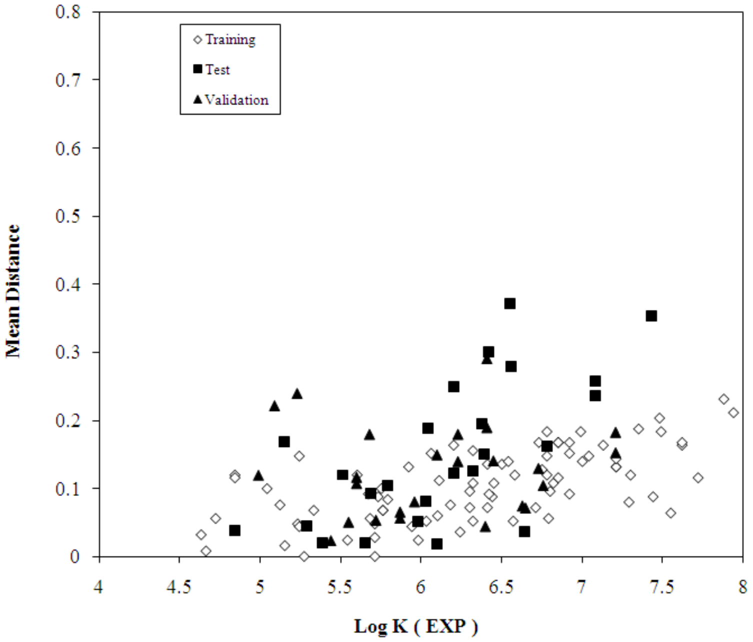

Figure 1. Scatter plot of samples for training, test and validation sets.

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 [44]. This term was defined in the following manner:

$$R_m^2 = R^2(1 - \left|\sqrt{R^2 - R_0^2}\right|) \quad (13)$$

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'^2_0 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.

3. Results and discussion

3.1. Molecular diversity validation

The fundamental research themes in chemical database analysis are diversity of sampling [45]. The diversity problem involves defining a different subset of representative compounds. In this study, 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 following vector (eq. 14):

$$X_i = (x_{i1}, x_{i2}, x_{i3}, \ldots x_{im}) \quad \text{for } i = 1,2,\ldots,n \quad (14)$$

where x_{ij} denotes the value of descriptor j of compound X_i. The collective database $X = \{X_i\}_{i=1}^N$ is represented a n×m matrix of X as follows (Eq. 15):

$$X = (X_1, X_2, \ldots X_N)^T = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad (15)$$

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. 16):

$$d_{ij} = \|X_i - X_j\| = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \quad (16)$$

The mean distances of one sample to the remaining ones were computed as follows (Eq. 17):

$$\bar{d}_i = \frac{\sum_{j=1}^{n} d_{ij}}{n-1} \quad i = 1,2,\ldots,n \quad (17)$$

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. (16) and (17) a MATLAB program was written in our laboratory. This program combines maximum dissimilarity search algorithms and general multi-dimensional measurements of chemical similarity based on different molecular descriptors. The closer to one the distance is the more diverse to each other the compound is. The mean distances of samples were plotted against log K_{ow} (EXP) and was shown in Figure 1. Inspections to this figure illuminate 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's stability and the diversity of test and validation sets can prove the predictive capability of the model.

3.2. PLS modeling

The data set and corresponding observed PLS and ANN predicted values of n-octanol-water partition coefficients of all molecules studied in this work are shown in Table 1. Using GA-PLS variable selection, 12 descriptors were selected. These descriptors with short description are given in Table 2. After variable selection with GA-PLS, obtained descriptors were used to build PLS model. Among these 12 descriptors, 5 descriptors were chosen by this model. Specifications of finally selected descriptors by PLS are given in Table 3. These descriptors are: Balabane index (J), XY Shadow (S_{XY}), Kier shape index (order3) ($^3\kappa$), Wiener index (W) and Maximum valency of C atom (V_{max} C). The numerical values of these descriptors are shown in Table 4. Table 5 represents the correlation matrix for these descriptors.

Table 2. Selected descriptors by GA-PLS.

Notation	Description
J	Balaban index
RNDB	Relative number of double bonds
S_{XY}	XY Shadow
3κ	Kier shape index (order 3)
RNBR	Relative number of benzene rings
W	Wiener index
TMSA	Total molecular surface area
Q_{max} Cl	Maximum net atomic charge for a Cl atom
V_{max} C	Maximum valency of a C atom
FPSA-2	Fractional total charge weighted partial positive surface area
HBSA	H-bonding surface area
Q_{max} C	Maximum partial charge for a C atom

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

$$ME_j = \frac{\beta_j \sum_{i=1}^{n} d_{ij}}{\sum_{j}^{m} \beta_j \sum_{i}^{n} d_{ij}} \qquad (18)$$

where, ME_j is the mean effect for considered descriptor j, β_j is the coefficient of 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 3 and are also plotted in Figure 2. The value and sign of mean effect shows the relative contribution and direction of influence of each descriptor on the partition coefficient.

Table 3. The partial least squares regression coefficients.

Descriptor	Notation	Coefficient	Mean effect
Balabane index	J	-1.6504	-4.4486
XY Shadow	S_{XY}	0.1559	1.1185
Kier shape index (order3)	$^3\kappa$	0.5982	1.5999
Wiener index	W	0.0052	2.5621
Max. valency of C atom	V_{max} C	-5.7974	-22.7372
Constant	-	28.3132	-

As shown in Table 3 the most relevant descriptor based on its mean effect is V_{max} C, a quantum-chemical descriptor. This descriptor relates to the strength of intramolecular bonding interactions and characterizes the stability of the molecules, their conformational flexibility and other valency-related properties [46]. Molecule with higher value of V_{max} C is more hydrophile, therefore, its tendency to water phase increase, hence the coefficient of this descriptor has negative sign. The second relevant descriptor according to the mean effect value is a topological descriptor, Balabane index (J). This descriptor is defined by the following formula:

$$J = \left(\frac{q}{\mu+1}\right) \sum_{i,j}^{q} (S_i S_j)^{-\frac{1}{2}} \qquad (19)$$

where q is the number of edges in the molecular graph, μ is the cyclometric number and S_i and S_j are the distance sums (or distance degrees), obtained by summation the row i and column i (or row j and column j, respectively) of the distance matrix between atoms in the molecule. The negative coefficient of this descriptor means as the value of this descriptor increase, the values of log K_{ow} increase. The third relevant descriptor according to the mean effect value is Wiener index. The Wiener index [47] can be expressed in the terms of the distance matrix. The distance matrix is a square matrix (NSA x NSA), and the entries d_{ij} correspond to the number of bonds in the shortest path connecting the pair of atoms i and j. The Wiener index W equals to the half-sum of all distance matrix entries:

$$W = \frac{1}{2} \sum_{(i,j)}^{N_{SA}} d_{ij} \qquad (20)$$

The next descriptor is Kier shape index (order3) (3κ). The shape of molecule depends on the number of skeletal atoms, the molecular branching and the special parameter a_i which is calculated as the ratio of the atomic radius (r_i) and the radius of the carbon atom in the sp^3 hybridization state (r_0) [48]:

$$^{3}K = (N_{SA} + \alpha - 1)(N_{SA} + \alpha - 3)^2 (^{3}P + \alpha)^2 \qquad (21)$$

(if N_{SA} is odd)

where ^3P is the number of paths of the length n in the molecular skeleton, and α is the sum of the a_i parameters for all skeletal atoms minus 1.

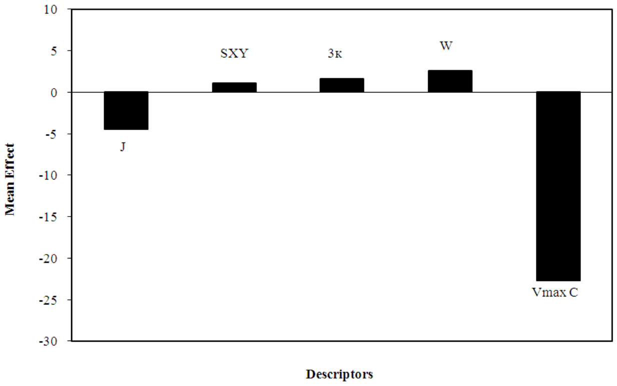

Figure 2. Plot of descriptor's mean effects.

These two Topological descriptors (also called topological indices) describe the atomic connectivity in the molecule [49-51]. These molecular descriptors have positive signs for their mean effect, which reveal that a larger molecule with flexible conformation is more likely to be found in organic phase. The last descriptor that is presented here is XY Shadow (S_{XY}), a geometrical descriptor. The shadow areas are calculated by applying 2D square grid on the molecular projection and by summation of the areas of squares overlapped with a

Table 4. The values of the descriptors that were used in this work [a].

Number[b]	J	S_{XY}	3κ	W	V_{max} C
1	2.475	6.726	1.923	246	3.953
2	2.424	6.642	1.923	252	3.954
3	2.609	6.168	1.944	287	3.951
4	2.528	6.428	2.092	298	3.952
5	2.492	7.172	2.296	301	3.953
6	2.397	7.098	2.296	315	3.954
7	2.647	5.940	2.124	344	3.950
8	2.635	6.318	2.298	346	3.951
9	2.693	5.998	2.124	338	3.950
10	2.603	6.782	2.124	352	3.951
11	2.541	7.578	2.460	360	3.952
12	2.494	7.546	2.460	368	3.953
13	2.584	6.478	2.298	354	3.952
14	2.551	7.550	2.298	358	3.953
15	2.590	7.536	2.460	352	3.953
16	2.681	6.728	2.334	408	3.950
17	2.671	6.810	2.475	410	3.951
18	2.739	6.384	2.334	400	3.949
19	2.726	6.342	2.334	401	3.950
20	2.586	6.780	2.660	426	3.951
21	2.699	6.444	2.475	407	3.949
22	2.662	6.680	2.660	412	3.951
23	2.798	6.908	2.334	391	3.948
24	2.611	7.336	2.475	421	3.952
25	2.565	7.110	2.475	430	3.952
26	2.635	6.946	2.475	418	3.951
27	2.543	7.442	2.660	433	3.953
28	2.601	7.344	2.660	423	3.952
29	2.631	7.908	2.475	417	3.953
30	2.597	7.932	2.660	425	3.951
31	2.650	7.302	2.681	494	3.951
32	2.769	6.808	2.518	472	3.949
33	2.738	6.778	2.518	476	3.950
34	2.717	6.520	2.834	480	3.950
35	2.800	6.814	2.518	466	3.948
36	2.761	6.502	2.681	472	3.950
37	2.686	6.528	2.681	486	3.951
38	2.641	7.362	2.834	496	3.951
39	2.722	6.786	2.834	480	3.950
40	2.607	8.428	2.681	502	3.953
41	2.683	7.342	2.681	488	3.951
42	2.716	7.102	2.834	480	3.952
43	2.638	7.702	2.681	498	3.951
44	2.671	7.404	2.681	492	3.951
45	2.599	7.454	2.834	504	3.953
46	2.636	7.888	3.028	496	3.953
47	2.678	7.142	3.028	488	3.952
48	2.737	7.424	2.864	562	3.950
49	2.796	7.010	2.726	550	3.949
50	2.828	7.040	2.726	543	3.948
51	2.810	7.072	2.864	546	3.950
52	2.717	7.312	2.864	568	3.951
53	2.741	7.170	2.864	562	3.950
54	2.754	7.558	2.864	559	3.951
55	2.729	7.264	3.036	564	3.951
56	2.771	6.768	2.864	555	3.950
57	2.692	7.438	3.036	573	3.951
58	2.776	7.388	3.036	555	3.948
59	2.675	7.644	2.864	577	3.952
60	2.738	7.396	2.864	563	3.951
61	2.776	6.884	3.036	554	3.951
62	2.658	7.888	3.036	580	3.953
63	2.625	8.934	3.036	587	3.953
64	2.840	7.294	2.913	632	3.951
65	2.834	7.390	3.069	634	3.950
66	2.834	7.488	2.913	634	3.949
67	2.834	7.488	2.913	634	3.949
68	2.762	8.124	3.069	652	3.951
69	2.828	7.692	2.913	638	3.949
70	2.803	7.452	3.069	642	3.951
71	2.796	7.514	3.069	644	3.950
72	2.866	7.558	2.913	628	3.949
73	2.870	7.290	3.069	626	3.948
74	2.728	8.178	3.069	660	3.953
75	2.761	8.056	3.069	652	3.952
76	2.791	7.838	3.069	644	3.952
77	2.826	8.322	3.120	738	3.950
78	2.875	7.752	2.977	726	3.949
79	2.869	7.836	3.120	728	3.949
80	2.911	7.696	3.120	717	3.948
81	2.834	8.310	3.120	737	3.952
82	2.928	8.138	3.183	820	3.950
83	2.964	7.816	3.183	810	3.948
84	2.549	6.226	2.092	294	3.952
85	2.444	7.156	2.296	308	3.953
86	2.596	6.264	2.298	352	3.951
87	2.542	6.832	2.298	360	3.953
88	2.579	6.424	2.460	354	3.952
89	2.664	6.484	2.334	412	3.951
90	2.653	6.618	2.475	414	3.951
91	2.676	6.432	2.475	410	3.950
92	2.593	7.180	2.660	425	3.952
93	2.644	6.518	2.660	416	3.951
94	2.620	7.962	2.475	419	3.954
95	2.734	7.240	2.518	478	3.951
96	2.679	6.890	2.834	488	3.951
97	2.730	6.752	2.681	478	3.949
98	2.677	7.224	2.681	488	3.952
99	2.700	7.424	2.681	486	3.950
100	2.608	7.714	2.834	502	3.953
101	2.779	6.982	2.726	553	3.950
102	2.803	6.940	2.726	548	3.951
103	2.820	6.988	2.864	545	3.949
104	2.708	7.382	2.864	570	3.949
105	2.769	7.900	2.913	650	3.951
106	2.877	7.006	2.913	624	3.948
107	2.785	7.930	2.913	648	3.951
108	2.861	7.656	3.120	729	3.950
109	2.580	5.944	1.944	291	3.952
110	2.484	7.170	2.092	303	3.953
111	2.498	7.116	2.092	301	3.953
112	2.590	6.670	2.298	352	3.952
113	2.676	6.432	2.124	342	3.950
114	2.531	6.590	2.460	362	3.952
115	2.633	6.528	2.475	417	3.951
116	2.623	6.538	2.660	419	3.951
117	2.716	6.210	2.475	403	3.950
118	2.741	6.962	2.334	402	3.949
119	2.555	8.028	2.660	432	3.952
120	2.769	6.624	2.518	470	3.949
121	2.687	6.874	2.681	486	3.950
122	2.722	6.790	2.681	480	3.950
123	2.683	6.978	2.834	488	3.950
124	2.747	6.802	2.681	476	3.950
125	2.640	8.404	2.834	496	3.952
126	2.761	7.612	2.726	557	3.951
127	2.699	7.716	2.864	571	3.951
128	2.789	7.098	2.864	552	3.949
129	2.733	7.184	3.036	564	3.951
130	2.735	7.458	2.864	562	3.953
131	2.799	7.916	3.069	642	3.951
132	2.909	6.944	2.913	616	3.948
133	2.823	7.184	3.069	638	3.951

[a] The definitions of the descriptors are given in Table 2.

[b] The numbers refer to the numbers of the molecules given in Table 1.

projection. Those indices therefore reflect the size (natural shadow indices) and geometrical shape (normalized shadow indices) of the molecule. As we know, for a solute to enter into aqueous solution, a cavity must be formed in the solvent for the solute molecule to occupy. Increasing shadow area, leads to increasing cavity information energy in solvent (water), the larger the solute, the grater the energy demand to make cavity and lower the solubility in water. Consequently, increasing of XY Shadow, increasing the K_{ow}.

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

Table 5. Correlation matrix of the five descriptors used in this work[a]

	J	S_{XY}	$^3\kappa$	W	V_{max} C
J	1.000	0.230	0.697	0.836	-0.774
S_{XY}		1.000	0.644	0.630	0.197
$^3\kappa$			1.000	0.820	-0.298
W				1.000	-0.432
V_{max} C					1.000

[a] The definitions of the descriptors are given in Table 2.

3.3. Neural network modeling

The next step was the construction of an artificial neural network. Before training the ANNs, the parameters of network including the number of nodes in the hidden layer, weights and biases learning rates and momentum values were optimized. Table 6 shows the architecture and specification of the optimized network. After optimization of the network parameters, the network was trained by using training set for adjustment of the weights and biases values by back-propagation algorithm. It is known that neural network can become over–trained. An over-trained network has usually learned perfectly the stimulus pattern it has seen but cannot give accurate prediction for unseen stimuli, and it no longer able to generalize. There are several 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 iteration the network was used to calculate log K_{ow} 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 65000 iterations.

Table 6. Architecture and specification of the generated ANN.

Parameter	Value
No. of nodes in the input layer	5
No. of nodes in the hidden layer	5
No. of nodes in the output layer	1
Weights learning rate	0.2
Bias learning rate	0.6
Momentum	0.3
Transfer function	Sigmoid

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 follow:

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{22}$$

Where y_i and $\hat{y}_i$, respectively are the measured and predicted values of the dependent variable(*n*-octanol-water partition coefficient), $\bar{y}$ is the averaged value of dependent variable of the training set and the summations cover all the compounds. The calculated value of q^2 was 0.969.

Table 1 represents the experimental, PLS and ANN calculated values of *n*-octanol-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 7. The standard errors of training, test and validation sets for the PLS model are 0.230, 0.164 and 0.297, respectively which would be compared with the values of 0.063, 0.112 and 0.126, respectively, for the ANN model. Comparison between these values and other statistical parameters in Table 7 reveals the superiority of the ANN model over PLS one. The key strength of neural networks, unlike PLS 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 =0.969, R^2 = 0.958 (R=0.979), $R_0^2 = 0.966$, $R_m^2 = 0.872$ and k= 0.999. These values and other statistical parameters which are shown in Table 7 reveal the high predictive ability of the model. Figure 3 shows the plot of the ANN predicted versus experimental values for *n*-octanol-water partition coefficients of all of the molecules in data set. The residuals of the ANN calculated values of the *n*-octanol-water partition coefficients are plotted against the experimental values in Figure 4. The propagation of the residuals in both sides of zero line indicates that no systematic error exists in the constructed QSPR model.

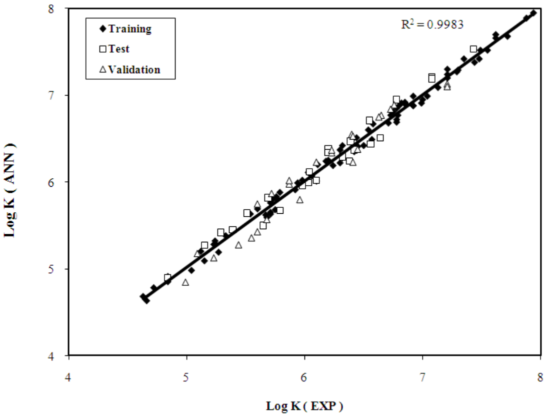

Figure 3. Plot of ANN calculated *n*-octanol-water partition coefficients against experimental values

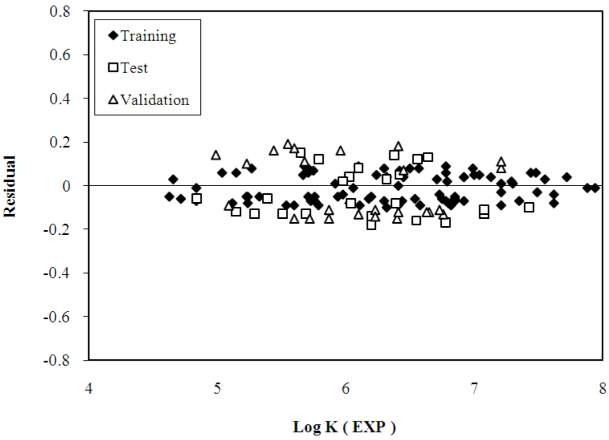

Figure 4. Plot of residual versus experimental values of *n*-octanol-water partition coefficient

Lu *et al.* [13] reported a QSPR model for the prediction of n-octanol-water partition coefficients of 133 polychlorinated biphenyls by the Heuristic method of CODESSA (comprehensive descriptors for structural and statistical analysis) technique. They developed three QSPR models. The best model they obtained has the squared correlation coefficients (R^2) of 0.9263 for the training set and 0.9336 for the test set. They also used radial basis function neural network (RBFNN) to build nonlinear prediction model for further discussing of the correlation between the molecular structure and the n-octanol-water partition coefficients values of PCBs based on the same subset of three descriptors. The model gave correlation coefficients (R^2) of 0.9393 for the training set and 0.9023 for the test set, respectively. Comparison between results obtained by Lu *et al.* and present study indicated that the model demonstrated in this work performs substantially better than the former models in predicting of partition coefficients.

Padmanabhan *et al.* [14] developed a QSPR model for estimation of the lipophilic behaviour (log K_{ow}) of the data set containing 133 polychlorinated biphenyl (PCB) congeners using the conceptual density functional theory based global reactivity parameter such as electrophilicity index (x) along with energy of lowest unoccupied molecular orbital (E_{LUMO}) and number of chlorine substituents (N_{Cl}) as descriptors. They

Table 7. 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	0.063	0.112	0.126	0.997	0.985	0.979	13959	749	543
PLS	0.230	0.164	0.297	0.960	0.967	0.887	961	335	85

[a] c refers to the calibration (training) set; t refers to test set; v refers to validation set; R is the correlation coefficient; SE is standard error and F is the statistical F value.

have performed linear/multilinear regression method using experimental log K_{ow} as dependent variable and various combinations of the selected descriptors as independent variables. The correlation coefficients (R^2) of training and test set of their model were 0.914 and 0.909, respectively. Comparison between results attained by Padmanabhan *et al*. and this study revealed the superiority of our model.

4. Conclusion

Results of this study reveal that ANN can be used successfully in development of a QSRR model to predict the *n*-octanol-water partition coefficients of polychlorinated biphenyls. Descriptors appear in these QSPR model provide some information related to different molecular properties, which can participate in the intermolecular interactions that affected on the *n*-octanol-water partition coefficient. The good agreement between experimental results and predicted values confirm the validity of obtained models. The calculated statistical parameters of these models reveal the superiority of ANN over PLS model. The result shows that ANN model can describe accurately the relationship between the structural parameters and *n*-octanol-water partition coefficient of compounds.

References

[1]. Wania, F.; Mackay, D. *Ambio*. **1993**, *22*, 10-18.

[2]. Giesy, J. P.; Kannan, K. *Crit. Rev. Toxicol.* **1998**, *28*, 511- 569.

[3]. King, C. M.; King, R. B.; Bhattacharyya, N. K.; Newton, M. G. *Organomet. Chem.* **2000**, *600*, 63-70.

[4]. Sanchez, E.; Fernandez, S. M.; Lopez-Aparicio, P.; Recio, M. N.; Perez-Albarsanz, M. A. *Chem. Biol. Interact.* **2000**, *125*, 117-131.

[5]. Leo, A. *J. Chem. Rev.* **1993**, *93*, 1281-1306.

[6]. Klopman, G.; Li, J. Y.; Wang, S. J. *Chem. Inf. Comput. Sci.* **1994**, *34*, 752-781.

[7]. Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835-845.

[8]. Woodrow, B. N.; Dorsey, J. G. *Environ. Sci. Technol.* **1997**, *31*, 2812-2820.

[9]. Platts, J. A.; Abraham, M. H.; Butina, D.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 71-80.

[10]. Khadikar, P. V.; Singh, S.; Shrivastava, A. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1125-1128.

[11]. Chen, J. W.; Xue, X. Y.; Schramm, K. W.; Xie, Q.; Yang, F. L.; Kettrup, A. *Chemosphere* **2002**, *48*, 535-544.

[12]. Niu, J. F.; Yang, Z. F.; Shen, Z. Y.; Long, X. X.; Yu, G. *Chemosphere* **2006**, *64*, 658-665.

[13]. Lu, W.; Chen, Y.; Liu, M.; Chen, X.; Hu, Z. *Chemosphere* **2007**, *69*,469-478.

[14]. Padmanabhan, J.; Parthasarathi, R.; Subramaniana, V.; Chattaraj, P. K. *Bioorganic & Medicinal Chemistry* **2006**, *14*, 1021-1028.

[15]. Puzyn, T.; Falandysz, *J. Phys. Chem. Ref. Data* **2007**, *36*, 203-214.

[16]. Li, L.; Xie, S.; Cai, H.; Bai, X.; Xue, Z. *Chemosphere* **2008**, *72*, 1602-1606.

[17]. Vegas, J. M.; Zufiria, P. J. *Neural Networks* **2004**, *17*, 233-245.

[18]. Schweitzer, R. C.; Morris, J. B. *Anal. Chem. Act.* **1999**, *384*, 285-303.

[19]. Tong, C. S.; Cheng, K. C. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 135-150.

[20]. Golmohammadi, H.; Fatemi, M. H. *Electrophoresis* **2005**, *26*, 3438-3444.

[21]. Baher, E.; Fatemi, M. H.; Konoz, E.; Golmohammadi, H. *Microchim Acta* 2007, *158*, 117-122.

[22]. Konoz, E.; Golmohammadi, H. *Anal. Chem. Act.* **2008**, *619,* 157-164.

[23]. Lui, F.; Liang, Y.; Cao, C. *Chemom. Intell. Lab. Syst.* **2006**, *81*, 120-126.

[24]. Hyperchem, Rel. 4. for Windows, Autodesk, Sansalito, CA, 1995.

[25]. Stewart, J. J. P. Semiempirical Molecular Orbital Program; QCPE, 445 (1983), Version 6, 1990.

[26]. Katritzky, A. R.; Labadov, V. S.; Carelson, M. CODESSA Training Manual, University of Florida, Gainesville, 1995.

[27]. Katritzky, A. R.; Labadov, V. S.; Carelson, M. CODESSA Version 1 Reference Manual, University of Florida, Gainesville, Florida, 1994.

[28]. Goldberg, D. E. Genetic Algorithms in Search, Optimization and Machine learning, Addison-Wesley, New York, 1989.

[29]. Hoskuldsson, A. Prediction Methods in Science and Technology Vol. 1: Basic Theory, Thur Publishing, Denmark, 1996.

[30]. Hasegawa, K.; Kimura, T.; Funatsu, K. *Quant. Struct. Acta. Relat.* **1999**, *18*, 262-272.

[31]. Leardi, R.; Boggia, R. Terrile, M. *J. Chemom.* **1992**, *6*, 267-281.

[32]. Leardi, R.; Gonzalez, A. L. *Chemom. Intell. Lab. Sys.* **1998**, *41*, 195-207.

[33]. Lorber, A.; Wangen, L.; Kowalsky, B. R. *J. Chemom.* **1987**, *1*, 19-31.

[34]. Khayamian, T.; Ensafi, A. A.; Hemmateenejad, B. *Talanta* **1999**, *49*, 587-596.

[35]. Shamsipur, M.; Hemmateenejad, B.; Akhond, M.; Sharghi, H. *Talanta* **2001**, *54*, 1113-1120.

[36]. Hoskuldsson, A. *Chemom. Intell. Lab. Syst.* **2001**, *55*, 23-38.

[37]. MATLAB 7. 0, The Mathworks Inc., Natick, MA, USA, http://www. mathworks. com.

[38]. Zupan, J.; Gasteiger, J. Neural Network in Chemistry and Drug Design; Wiley-VCH. Weinheim, 1999.

[39]. Beal, T. M.; Hagan, H. B.; Demuth, M., Neural Network Design; PWS, Boston, 1996.

[40]. Zupan, J.; Gasteiger, J. Neural Networks for Chemists: an Introduction; VCH. Weinheim, 1993.

[41]. Blank, T. B.; Brown, S. T. *Anal. Chem.* **1993**, *65*, 3081-3089.

[42]. Jalali-Heravi, M.; Fatemi, M. H. *J. Chromatogr. A* **2001**, *915*, 177-183.

[43]. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269-276.

[44]. Roy, P. P.; Roy, K. *QSAR Comb. Sci.* **2008**, *27*, 302-313.

[45]. Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B. T. *Mol. Divers.* **2006**, *10*, 39-79.

[46]. Sannigrahi, A. B. *Adv. Quant. Chem.* **1992**, *23*, 301-351.

[47]. Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17-20.

[48]. Kier, L. B. Computational Chemical Graph Theory Rouvray, D. H. (editor), Nova Science Publishers, New York, 1990, pp. 151-174.

[49]. Kowalski, B. R. (Ed.), Chemometrics, Reidel, Dordrecht, 1984.

[50]. Stankevich, M. I.; Stankevich, I. V.; Zefirov, N. S.; *Russ. Chem. Rev.* **1988**, *57*, 191-208.

[51]. El-Basil, S.; Randic, M. *Adv. Quant. Chem.* **1992**, *24*, 239-290.