Journal of Biostatistics and Biometric Applications
Volume 9 | Issue 1
ISSN: 2455-765X

Research Article Open Access

Prediction of the Soil-Water Distribution Coefficient from Theoretical Derived Molecular Descriptors

Maedeh Hosseinnia, Mohammad H Fatemi*

Department of Chemistry, University of Mazandaran, Babolsar, Iran

* **Corresponding Author:** Mohammad H Fatemi, Department of Chemistry, University of Mazandaran, Babolsar, Iran, Tel: + 98 112 5242002, Fax.: + 98 112 5242002, E-mail: mhfatemi@umz.ac.ir

Citation: Maedeh Hosseinnia, Mohammad H Fatemi (2024) Prediction of the Soil-Water Distribution Coefficient from Theoretical Derived Molecular Descriptors, J Biostat Biometric App 9(1): 103

Received Date: June 04, 2024 **Accepted Date:** July 04, 2024 **Published Date:** July 08, 2024

Abstract

In this work an artificial neural network (ANN) was constructed and trained for the prediction of the soil organic carbon to water distribution coefficients of some aliphatic and aromatic hydrocarbons and benzene derivatives based on quantitative structure- property relationships approaches. The inputs of this network are theoretically derived descriptors, which were chosen by the stepwise multiple linear regression (MLR) variables selection techniques. These descriptors are: R-maximal index/weighted by atomic Sanderson electronegativities, R-maximal autocorrelation of lag5/weighted by atomic Sanderson electronegativities, R-autocorrelation of lag3/weighted by atomic Sanderson electronegativities, 3D-Morse-signal 01/weighted by atomic van der Waals volumes and H autocorrelation of lag3/weighted by atomic polarizabilities which encode topological and electronic aspects of molecules. The standard errors of the obtained ANN model are 0.16, 0.19 and 0.27 for training, test and prediction sets respectively. Comparison between statistical results calculated from the MLR and ANN models reveals that all statistics have been improved considerably in the case of the ANN model. Also the results obtained in this paper demonstrate that it is possible to generate robust neural network, capable in estimation of the soil organic carbon to water distribution coefficient using theoretically calculated molecular descriptors.

Keywords: Soil-water distribution coefficient; artificial neural network; quantitative structure-property relationship; molecular descriptors; multiple linear regressions

Introduction

The term of sorption is used frequently in environmental studies to denote the uptake of a solute by a solid (soil or sediment or component of soil) without refering to specific mechanism. Sorption play an important role in determining the environmental fate and impact of organic chemicals. This process affects on a variety of specific fate phenomena of organic compounds like volatilization [1], bioavailability [2], photolysis [3], and hydrolysis [4]. Soil organic carbon to water distribution coefficient (koc), quantitavely describe the extent to which an organic chemical is distributed between an environmental solid and an aqueous phase, which is defined by the following equation5:

$$k_{oc=}C_{soil}/C_w \quad (1)$$

Where C_{soil} is the concentration of solute (gram of carbon per gram of soil) in the soil, and C_w denotes the concentration of solute in the aqueous phase. The distribution of a molecule between soil and water was a result of competitive interactions of the solute between these phases. Sorption occurs when the free energy of the interactions between an environmental solid and an organic chemical sorbet is negative. Relatively weak intermolecular forces called Van der Waals interactions drive this process. These types of interactions refers to all attractive and repulsive interactions between chemically nonbonding molecules namely; electrostatic, induction, and dispersion interactions. The molecular structure and chemical properties of the interested molecule determine type and extent of these interactions and therefore affected on the value of k_{oc}.

The value of k_{oc} plays an important role in determining the pollution potential of the nonionic organic compounds such as pesticides. The experimental determination of soil organic carbon to water distribution coefficient is time consuming, difficult task and requiring a number of measured values for different types of soils over a reasonable concentration range [6-8]. Because of experimental complications, and the needs to access these values for new synthesized materials, there is considerable interest to use an estimation method based on surrogate solute properties that are more amenable to experimental determination [9-14]. One of the most successful approaches in the prediction of chemical properties starting only from their molecular structural informations is quantitative structure-activity/property relationships (QSAR/QSPR). In these studies, a correlation between chemical structural parameters and chosen property is defined. Such studies consist of two main stages. In the first step the chemical compounds are translated into a computer readable form and afterwards the quantitative correlation between chemical structure and its property can be obtained using different statistical and learning procedures, like multiple linear regression (MLR), artificial neural network (ANN) and partial least square (PLS). There are some reports about the applications of QSPR approaches to predict the soil organic carbon distribution coefficients in some soil-solvent systems [15, 16]. For example Delgado, Alderete and Jana predict the soil sorption coefficients of 80 polar and nonpolar organic compounds using QSPR approach [17]. This is a 5-parameter MLR model based on molecular descriptors. These parameters are; number of benzene rings, molecular weights, number of N atoms, number of O atoms and number of S atoms. Also Liu and Yu developed a QSPR approaches for the prediction of the soil sorption coefficient of some polar organic chemicals [18]. Since the commonly used MLR will fail to develop an appropriate QSPR model when the nonlinear phenomenon is significant to some extent within the investigated data; therefore nonlinear modeling techniques such as artificial neural networks were necessary for building an accurate and reliable QSPR model [19-21]. Todays application of artificial intelligence in modeling of environmental risk assessment of chemicals is very important and necessary. Computational toxicology is a rapidly growing field that utilizes artificial intelligence and machine learning to predict the toxicity of chemical compounds. Computational toxicology is an important tool for assessing the risks associated with the exposure of soil, water and air to environmental contaminants. By providing insights into the behavior and effects of these compounds, computational models can help to inform management decisions and protect the health of aquatic ecosystems and the humans who depend on them for food and recreation. The main aim of the present work was to development of a QSPR model using artificial neural network for the prediction of soil organic carbon to water distribution coefficient of some aliphatic and aromatic hydrocarbons, heterocyclic compounds and benzene derivatives.

Methods

Data Set

The soil organic carbon to water distribution coefficients of 57 neutral organic compounds weretaken from Ref. [22], which were be used as data set. The molecules inthe dataset consist of aliphaticandaromatic hydrocarbons, heterocyclic compoundsandbenzene derivatives (Table 1). The soil-water distribution coefficients of these compoundswere obtained in the same conditions andfalls in the range of 1.3 to 4.85 for 2-methylphenol and pyrene, respectively. Data sets were randomly divided into three separate sections: the training, test, and external validation sets, consisting of 37, 10, and 10 members, respectively. The training set was used to adjust the parameters of models, the test set was used to prevent the network from overfitting, and the external validation set was used to evaluate the prediction ability of constructed models.

Table 1: Data set and corresponding observed ANN and MLR predicted values of $\log K_{oc}$.

No.	Name	$(\log K_{oc})_{EXP}$	$(\log K_{OC})_{ANN}$	$(\log K_{OC})_{MLR}$	E_{abs}	E_r
	Training set					
1	Benzene	1.90	1.96	2.17	0.06	0.03
2	1,3-Dimethylbenzene	2.50	2.33	2.37	-0.17	-0.07
3	1,4-Dimethylbenzene	2.72	2.41	2.57	-0.31	-0.11
4	n-Buthylbenzene	3.53	3.49	3.29	-0.04	-0.01
5	Naphtalene	2.97	3.12	3.12	0.15	0.05
6	Pyrene	4.85	4.73	5.12	-0.12	-0.02
7	Chlorobenzene	2.34	2.30	2.39	-0.04	-0.02
8	2-Methylphenol	1.30	1.42	1.31	0.12	0.09
9	1,3-Dichlorobenzene	2.48	2.68	2.30	0.20	0.08
10	Pyridine	2.48	2.34	2.30	-0.14	-0.06
11	1,2,3,4,-Tetrachlorobenzene	3.96	3.80	3.79	-0.16	-0.04
12	1,3,5-Trichlorobenzene	3.09	2.82	2.86	-0.27	-0.08
13	1,2,3,5-Tetrachlorobenzene	3.38	3.86	3.48	0.48	0.14
14	Bromobenzene	2.55	2.49	2.46	-0.06	-0.02
15	Iodobenzene	3.10	2.72	2.93	-0.38	-0.12
16	Dichloromethane	1.56	1.59	1.51	0.03	0.02
17	Tribromomethane	2.10	2.08	1.87	-0.02	-0.01
18	4-Methylphenol	1.56	1.72	1.91	0.16	0.10
19	1,1,2,2,-Tetrachloroethane	1.92	1.89	1.68	-0.03	-0.02
20	4-Nitroaniline	1.88	2.27	2.41	0.39	0.20
21	Tetrachloroethene	2.38	2.10	2.03	-0.28	-0.12
22	Trichloroethene	2.04	1.78	1.80	-0.26	-0.13
23	Benzylalcohol	1.43	1.75	1.04	0.32	0.22
24	3-Nitroaniline	1.73	2.21	2.29	0.48	0.28

25	Nitrobenzene	2.20	2	2.07	-0.20	-0.09
26	Anisol	1.54	1.56	1.71	0.02	0.01
27	Acetophenone	1.57	1.68	1.81	0.11	0.07
28	4-Methylaniline	1.90	1.87	1.95	-0.03	-0.02
29	Quinoline	2.84	2.86	2.83	0.02	0.01
30	Phenol	1.55	1.76	1.82	0.21	0.14
31	4-Bromophenol	2.41	1.83	2.21	-0.58	-0.24
32	2-Nitrophenol	2.01	1.65	1.54	-0.36	-0.18
33	2-Chlorophenol	1.67	1.65	2.07	-0.02	-0.01
34	1-Aminonaphtalene	3.50	3.48	3.12	-0.02	-0.01
35	2-Methoxyphenol	1.56	1.64	1.47	0.08	0.05
36	3-Methoxyphenol	1.50	1.73	1.72	0.23	0.15
37	Naphtol	3.07	2.97	2.82	-0.10	-0.03
	Test set					
38	Toluene	2.06	1.44	1.67	-0.62	-0.30
39	1,3,5-Trimethylbenzene	3	3.25	3.18	0.25	0.08
40	Phenanthrene	4.35	4.38	4.22	0.03	0.01
41	1,2,3-Trichlorobenzene	3.29	3.42	3.45	0.13	0.04
42	Trichloromethane	1.81	1.62	1.53	-0.19	-0.10
43	1,1,2-Trichloroethane	1.99	1.80	1.62	-0.19	-0.09
44	3-Nitrophenol	1.68	1.82	1.93	0.14	0.08
45	4-Chlorophenol	1.80	1.78	2.04	-0.02	-0.01
46	1,4-Dichlorobenzene	2.80	2.50	2.41	-0.30	-0.11
47	1,2-Dichloropropane	1.68	2.05	2.15	0.37	0.22
	Prediction set					
48	Ethylbenzene	2.22	1.97	2.10	-0.25	-0.11
49	1,2,3-Trimethylbenzene	3.04	2.66	2.65	-0.38	-0.13
50	Anthracene	4.38	4.53	4.23	0.15	0.03
51	1,2,4-Trichlorobenzene	3.35	3.11	2.83	-0.24	-0.07
52	Tetrachloromethane	1.53	1.86	1.67	0.33	0.22
53	1,1,1-Trichloroethane	1.82	1.56	1.42	-0.26	-0.14
54	4-Nitrophenol	1.70	1.98	2.14	0.28	0.16
55	3-Chlorophenol	1.78	1.81	1.94	0.03	0.02
56	1,2-Dichloroethane	1.62	1.78	1.74	0.16	0.10
57	1,2-Dichlorobenzene	2.50	2.44	2.71	-0.06	-0.02

Molecular Descriptors

One important subject in QSPR studies is the numerical representation of the chemical structure of the interested molecule, which often called molecular descriptors. The built model performance and the accuracy of results are strongly depends on the way that the structural representation was performed. In this project the structures of the molecules were drawn with HyperChem 4.0 package[23]and exported in a file format suitable for MOPAC program [24]. The geometry optimization was performed with the semiempirical quantum method AM1 [25] using the MOPAC 6.0. All geometries had been fully optimized without symmetry restrictions. In all cases frequency calculations have been performed in order to ensure that all the calculated geometries correspond to true minima. Also the HyperChem output files were used by the Dragon 3.0 package [26]. Dragon is a new, freely available software (by Milano Chemometrics and QSAR Research Group) for the calculation of more than 1000 molecular descriptors. The generated numerical descriptors were responsible for encoding important features of the structures of molecules. Some descriptors, which generated for each compound encoded similar information about the molecule of interest. Therefore, it was desirable to test each descriptor and eliminate those that show high correlation (R>0.90) with each other. Subsequently, the method of stepwise multiple linear regression was used for the selection of important descriptors and MLR model construction. The main goal of the generation of the MLR model was to choose a set of suitable descriptors that can be used as inputs for generation of the ANN model. The specifications of selected MLR model is presented in Table2 and also the correlation matrix between selected descriptors is shown in Table 3.

Table 2: Specification of multiple linear regression models*

Descriptor	Notation	Coefficient	SE
R maximal index	$RT^{+}(e)$	-2.507	0.539
R maximal autocorrelation of lag5	$R_{5}^{+}(e)$	-7.485	1.136
P autocorrelation of lag3	$R_{3}(e)$	-0.258	0.104
3D-Morse-signal 01	Mor01v	0.009	0.002
H autocorrelation of lag3	$H_{3}(P)$	2.539	0.323
Constant		2.627	0.221

* SE refers to standard error of coefficients.

Table 3: Correlation matrix for descriptors applying in this work

	$R_{3}(e)$	**Mor01v**	**$RT^{+}(e)$**	**$R_{5}^{+}(e)$**	**$H_{3}(P)$**
$R_{3}(e)$	1				
Mor01v	0.187	1			
$RT^{+}(e)$	-0.324	-0.539	1		
$R_{5}^{+}(e)$	0.160	0.216	-0.098	1	

Artificial Neural Network

An ANN is a biologically inspired computer program desired to learn from data in a manner of emulating the learning pattern in the brain. Most ANN systems are very complex and high-dimension processing systems. Details descriptions of the theory behind neural networks have been adequately described elsewhere [27-29]. The relevant principle of supervised learning in an ANN is that it takes numerical inputs (the training data) and transfers them into desired outputs. The inputs and outputs nodes may be connected to the ‘external world’ and to other nodes within the network. The way in which each node transforms its input depends on the so-called ‘connection weights’ or ‘connection strength’ and bias of the node, which are modifiable. The output value of each node depends on the weight strength and bias values. For the present purposes, the great power of Ann’s stems from the fact that it is possible to train them. Continually presenting the network with known inputs and outputs and modifying the connection weights and biases between the individual nodes do training. This process is continued until the output nodes of the network match to the desired outputs in a stated degree of accuracy. In the present work an ANN program was written in FORTRAN 77 in our laboratory. This network was feed-forward fully connected that has three layers with sigmoid transfer function. Descriptors appearing in the MLR models were used as inputs of network and signal of the output node represent the distribution coefficient of interested compound. Thus, this network has five nodes in input layer and one node in output layer. The value of each input was divided into its maximum value to bring them into dynamic range of the sigmoid transfer function of the network. The initial values of weights were randomly selected from a uniform distribution that ranged between -0.3 and +0.3. The initial values of biases were set to be one. These values were optimized during the network training. Before training, the network’s parameters would be optimized. These parameters are: number of nodes in the hidden layer, weights and biases learning rates, and the momentum values. Then the optimized network was trained using training set for the adjustment of weights and biases values. It is known that a neural network can become over-trained. An over-trained network has usually learned the stimulus pattern it has seen perfectly, but cannot give an accurate prediction for unseen stimuli, and it is 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 cycle, the network was used to calculate the distribution coefficient of molecules included in the test set. To maintain the prediction power of the network at a desirable level, training was stopped when the standard error of prediction for the test set started to increase. In order to evaluate the performance of the ANN, standard error of calibration and standard error of prediction were used [30]. For the evaluation of the predictive power of the networks, the trained ANN was used to calculate log k_{oc} of the molecules included in the prediction set.

Results and Discussion

The data set and corresponding observed and predicted values of the $\log K_{oc}$ of all molecules studied in this work are shown in Table1. The selected MLR model is presented in Table 2. It can be seen from this table that five descriptors appeared in the MLR model. These descriptors are: R-maximal index/weighted by atomic sanderson electronegativities ($RT^+(e)$), R- maximal autocorrelation of lag5/weighted by atomic sanderson electronegativities ($R^+_5(e)$), R autocorrelation of lag3/weighted by atomic sanderson electronegativities($R_3(e)$), 3D-Morse-signal 01/weighted by atomic van der Waals volumes (Mor01v) and H autocorrelation of lag3/weighted by atomic polarizabilities ($H_3(P)$). A detailed description of the theory of these descriptors has been described in Ref. [31]. According to Eq.1, the value of K_{OC} determined by the ratio of concentration of solute in soil phase to it's concentration in aqueous phase. As mentioned in introduction section, the value of K_{OC} affected by two compatative process; which are sorption and solubility. Sorption process is driven by relatively weak van der Waals intermolecular interactions. The charge distribution in a molecule can affected on molecular polarizability and hence on the van der Waals interactions and leading on the sorption process. On the other hand the molecular solubility is depends on the intermolecular forces that exist between the solute and solvent molecules. Therefore the solute-solvent interactions, such as hydrogen bonding and dipole-dipole interactions are highly responsible in the solubility of molecules. The parameters of $RT^+(e)$, $R^+_5(e)$, $R_3(e)$ and $H_3(P)$ which appeared in the model are geometry, topology and atomic weight assembly (GETAWAY) descriptors that were presented by Consonni et al. [31]. They encodes geometrical information that given from influence matrix, topological information given by molecular graph and chemical information from selected atomic properties. The Mor01v is a new type of descriptors for encoding of the molecular structure and named 3D molecular representation of structure based on electron diffraction code (3D--MORSE code) [32, 33]. They can be calculated by the summing of atomic properties viewed by different angular scattering functions which weighted by some atomic properties, like atomic number, atomic mass, partial atomic charge, atomic electronegativity and atomic polarizability. The possibility for choosing an appropriate atomic property give great flexibility to these descriptors for adapting them to the problem under investigations. This fact that these topological descriptors wighted by electronic properties of molecule reveals that they can also encode electronic distribution (or polarizability) of the interested molecules which has an important role in the interactions of molecules with soil and water and therefore affected on soil-water distribution coefficient of organic compounds. Other descriptors that appeared in model is $H_3(P)$ which, encode information about the charge distribution in molecule. This descriptor is particularly important in modeling some solution properties of compounds which depend on solute/solvent interactions and sometimes frequently used to represent polarizability term in linear solvation energy relationships.[31] The appearance of these descriptors ($RT^+(e)$, $R^+_5(e)$, $R_3(e)$, $H_3(P)$ and Mor01v) in the model reveals the role of electronic and charge distribution as well as size, shape and steric parameters of a molecule to it's K_{OC}. The correlations matrix between these descriptors are shown in Table 3. As can be seen from this table there are not significant correlations between selected descriptors. The next step was to construction of the ANN model. The selected descriptors were used as inputs for generated ANN. Before training the network, the parameters of the number of nodes in the hidden layer, weights and biases learning rate and momentum values would be optimized. Procedures for the optimization of these parameters were reported in our previous papers [34-36]. Table 4 shows the architecture and specification of the optimized network. After optimization of the ANN parameters, the network was trained for the adjustment of weights and biases values. The ANN calculated values of distribution coefficient for training, test, and prediction sets are shown in Table 1. To compare the applied chemometric methods of MLR and ANN in predicting the k_{oc} values, some statistics for these models calculated and shown in Table 5. The ANN standard error values of 0.16, 0.19 and 0.27 were obtained for the training, test and prediction sets respectively, should be compared with corresponding values of 0.25, 0.34 and 0.32 for MLR model. It can be seen from these values and also those shown in Table 5 that all statistics have improved considerably in the case of ANN model.

Table 4: Architecture and specification of the generated ANNs

No. of nodes in the input layer	5
No. of nodes in the hidden layer	3
No. of nodes in the output layer	1
Weights learning rate	0.2
Bias learning rate	0.3
Momentum	0.8
Transfer function	Sigmoid

Table 5: Comparison of the statistical parameters obtained using the ANN and MLR models*

*F is the statistical F value, R is the correlation coefficient and SE is the standard error of model

	MLR			ANN		
	F	SE	R	F	SE	R
Training	334.4	0.25	0.95	849.67	0.16	0.98
Test	52.17	0.34	0.93	184.3	0.19	0.97
Prediction	64.09	0.32	0.94	97.79	0.27	0.96

Figure 1 shows a plot of the ANN calculated versus the experimental values of log k_{oc} for data set molecules. Correlation coefficient of this plot is 0.955, for a 5-parameter ANN model. The residual of the ANN calculated values of the log koc are plotted against their experimental values in Figure. 2. The propagation of the residuals on both sides of zero line indicates that no systematic error exists in the development of the neural network. Any residuals can be raised to some experimental uncertainty in determination of log koc and/or existence of molecules in the out of the applicability domain of the model. The choice of an optimal metric space that represents bitterly the structural diversity of a compound population is determinant in the efficiency of the model. Molecular diversity analysis explores the way of molecules to cover a determined structural space and underlies many approaches for compound selection and design of combinatorial libraries. A correct diversity/similarity space should permit us to place molecules in good position relative to others in a well-parameterized way. The measures of molecular similarity and/or diversity involve in general three main components the descriptors, the coefficients and the weighting scheme. It is usual to define a "chemistry space" as a space with N descriptors that define the position of a molecule in an N-dimensional chemistry space; the diversity or similarity of compounds is intuitively related to their inter-molecular distance as measured in that space.

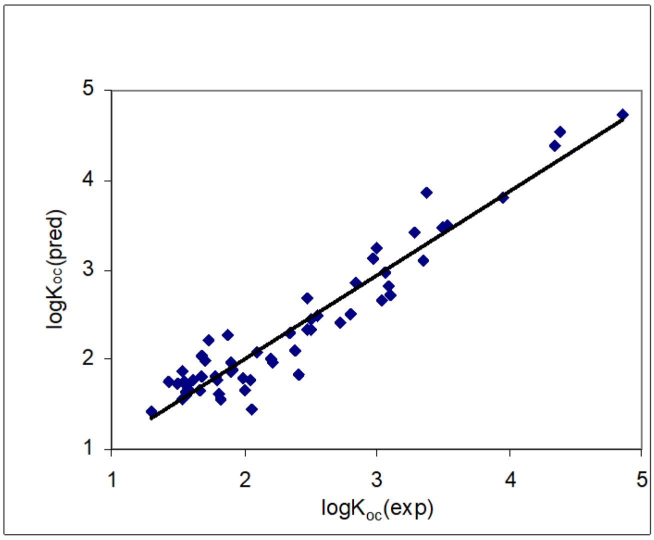

Figure 1: Plot of the ANN predicted distribution coefficients against the experimental values.

Figure 1 M.Hossein nia, M.H.Fatemi*, Prediction of the Soil-Water Distribution Coefficient from Theoretical Dervived Molecular Descriptors

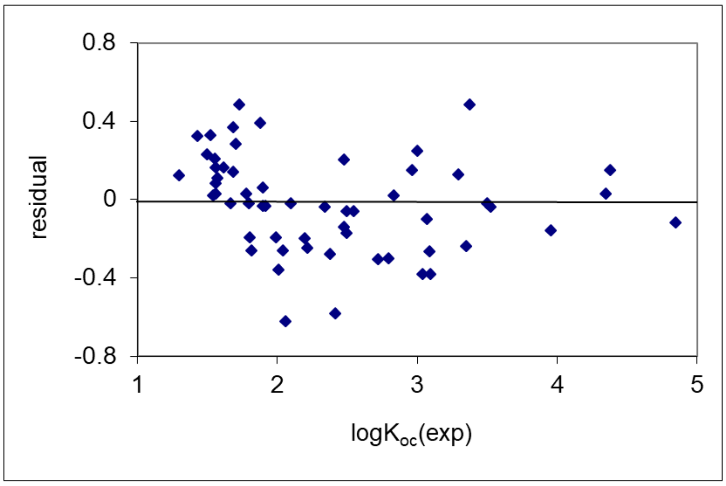

Figure 2: Plot of the residual versus experimental values of distribution coefficients

Figure 2 M.Hossein nia, M.H.Fatemi*, Prediction of the Soil-Water Distribution Coefficient from Theoretical Dervived Molecular Descriptors

Conclusion

The results of this study demonstrate that QSPR method using the artifical neural networks techniques can generate suitable model for the prediction of soil-water distribution coefficients of aliphatic and aromatic hydrocarbons and benzene derivatives. The key strength of the neural networks is their ability to allow for flexible mapping of the selected features by manipulating their functional dependence implicity, unlike regression analysis. Neural network handles both linear and nonlinear relationships without adding complexity to model. This capability offset the large computing time required and the complexity of the ANN method with respect to MLR. Also descriptors that appear in the MLR models and included in ANN, provide information related to the different molecular properties that can participate in the physicochemical process that occure in distribution of soil organic carbon to water.

References

1. Chiou CT, Shoup TD (1985) Soil sorption of organic vapors and effects of humidity on sorptive mechanism and capacity, Environ Sci Technol, 19: 1196.

2. Ditoro DM, Zarba CS, Hasen DJ, Betty WJ, Swuartz RC et al. (1991) Technical Basis for Establishing Sediment Quality Criteria for Nonionic Organic Chemicals Using Equilibrium Partitioning, Environ Toxicol Chem, 10: 1541.

3. Tao S, Piao HS, Dawson R, Lu XX, Hu HY (1999) Estimation of Organic Carbon Normalized Sorption Coefficient (KOC) for Soils Using the Fragment Constant Method Environ Sci Technol, 33: 2719.

4. Deeley GM, Reinhard M, Stearns SM (1991) J Transformation and sorption of 1, 2-dibromo-3-chloropropane in subsurface samples collected at Fresno, California Environ Qual, 20: 547.

5. Delgado EJ, Alderete JB, Jana GAJ (2003) A simple QSPR model for predicting soil sorption coefficients of polar and nonpolar organic compounds from molecular formula Chem Inf Comput Sci, 43: 1928.

6. Von Oepen B, Kordel W, Klein W (1989) Soil preparation for the estimation of adsorption coefficients (koc) of organic chemicals Chemosphere, 18: 1495.

7. Von Oepen B, Kordel W, Klein W (1991) Sorption of nonpolar and polar compounds to soils: processes, measurements and experience with the applicability of the modified OECD-Guideline 106 Chemosphere, 22: 285.

8. American Society for Testing and Materials, Standard Test Method for Determining a Sorption Constant (Koc) for an Organic Chemical in soil and Sediments; ASTM; West Conshohocken (1966).

9. Gerstl ZJ Contam (1990) QSAR MODELLING OF PREDICTING SOIL SORPTION COEFFICIENTS FOR ORGANIC CHEMICALS Hydrol, 6: 357.

10. Gaewlik RM, Sotiriou N, Feicht EA, Schulte-Hostede S, Kettrup A (1997) Alternatives for the determination of the soil adsorption coefficient, KOC, of non-ionicorganic compounds—a review Chemosphere, 34: 2525.

11. Lyman WJ, Reehl WF, Rosenblatt DH (1990) Hand book of chemical property Estimation Methods; American Chemical Society; Washington DC.

12. Von Oepen B, Kordel W, Klein W, Schuurmann G (1991) Predictive QSPR models for estimating soil sorption coefficients:

potential and limitations based on dominating processes Sci Total Environ, 343: 109-10.

13. Hermens (1995) Prediction of environmental toxicity based on structure-activity relationships using mechanistic information J Sci Total Environ, 171: 235.

14. Sabljic A, Gusten H, Verhaar H, Hermans (1995) QSAR modelling of soil sorption. Improvements and systematics of log Koc vs. log Kow correlations J Chemosphere, 31: 4489.

15. Delgado EJ, Alderete JB, Jana GA (2003) A simple QSPR model for predicting soil sorption coefficients of polar and nonpolar organic compounds from molecular formula J Chem Inf Comput Sci, 43: 1928.

16. Gramatica P, Corradi M, Consonni V (2000) Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors Chemosphere, 41: 763.

17. Delgado EJ, Alderete JB, Jana GAJ (2003) A simple QSPR model for predicting soil sorption coefficients of polar and nonpolar organic compounds from molecular formula Chem Inf Comput Sci, 43: 1928.

18. Liu G, Yu J (2005) QSAR analysis of soil sorption coefficients for polar organic chemicals: Substituted anilines and phenols Water Research, 39: 2048.

19. Zupan J, Gasteiger J (1991) Neural networks: A new method for solving chemical problems or just a passing phase? Anal Chim Acta, 248: 1.

20. Fatemi MH (2000) Prediction of the electrophoretic mobilities of some carboxylic acids from theoretically derived descriptors J Chromatogr A, 1002: 221.

21. Marengo E, Gianotti V, Angio S, Gennaro MC (2004) Optimization by experimental design and artificial neural networks of the ion-interaction reversed-phase liquid chromatographic separation of twenty cosmetic J Chromatogr A, 1029: 57.

22. Poole SK, Poole CF (1999) Chromatographic models for the sorption of neutral organic compounds by soil from water and air J Chromatogr A, 845: 381.

23. Hyperchem 4.0 (1994) Hypercube.

24. Stewart JJP (1989) MOPAC 6.0, Quantum Chemistry Program Exchange; QCPE, No. 455; India University, Bloomington.

25. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) Extension of MNDO to d Orbitals: Parameters and Results for the Second-Row Elements and for the Zinc Group J Am Chem Soc, 107: 902.

26. http://www.disat.unimib.it/chm/ (2001).

27. Zupan J, Gasteiger J (1999) Neural Networks in Chemistry and Drug Design; Wiley-VCH; Weinheim.

28. Haykin S (1994) Neural Network; Prentice-Hall; Englewood Cliffs, NJ.

29. Bose NK, Liang P (1996) Neural Network, Fundamentals; McGraw-Hill; New York.

30. Blank TB, Brown, ST (1993) Prediction of Retention Factors in Supercritical Fluid Chromatography Using Artificial Neural Network Anal Chem, 65: 3084.

31. Todeschini R, Consonni V (2000) Handbook of Molecular Descriptors; Wily-VCH; Weinheim.

32. Schuur JH, Selzer P, Gasteiger J (1996) The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity J Chem Inf Comput Sci, 36: 36.

33. Schuur JH, Gasteiger (1997) Infrared spectra simulation of substituted benzene derivatives on the basis of a 3D structure representation J Anal Chem, 69: 2398.

34. Jalali-Heravi M, Fatemi MH (2000) Simulation of mass spectra of noncyclic alkanes and alkenes using artificial neural network Anal Chim Acta, 415: 95.

35. Jalali-Heravi M, Fatemi MH (2000) Prediction of thermal conductivity detection response factors using an artificial neural network J Chromatogr A, 897: 161.

36. Fatemi MH, Baher E (2005) Prediction of retention factors in supercritical fluid chromatography using artificial neural network J Anal Chem, 60: 969.