J. Iran. Chem. Soc., Vol. 8, No. 1, March 2011, pp. 176-192.

JOURNAL OF THE
Iranian
Chemical Society

Prediction of Gas/Particle Partitioning Coefficients of Semi Volatile Organic Compounds *via* QSPR Methods: PC-ANN and PLS Analysis

O. Deeb[a,*], P.V. Khadikar[b] and M. Goodarzi[c]

[a]*Faculty of Pharmacy, Al-Quds University, P.O. Box 20002 Jerusalem, Palestine*

[b]*Research Division, Laxmi Fumigation and Pest Control Pvt. Ltd., 3, Khatipura, Indore-452 007, India*

[c]*Department of Chemistry, Faculty of Sciences, Islamic Azad University, Arak Branch, P.O. Box 38135-567 Arak, Markazi, Iran*

(Received 4 July 2009, Accepted 24 June 2010)

Linear and non-linear quantitative structure property relationship (QSPR) models for predicting the gas/particle partitioning coefficients of semivolatile organic compounds were developed based on partial least squares (PLS) and artificial neural network (ANN) to identify a set of structurally based numerical descriptors. Multilinear regression (MLR) was used to build the linear QSPR models using combination of the compounds structural descriptors and topological indices related to environmental conditions such as temperature, pressure and particle size. The prediction results for PLS and ANN models give very good coefficient of determination (0.97). In consistent with experimental studies, it was shown that linear and non-linear regression analyses are useful tools to predict the relationship between the calculated descriptors and gas/particle partitioning coefficient.

Keywords: Gas/particle partitioning coefficient, Semivolatile organic compounds, QSPR, PLS, PC-ANN

INTRODUCTION

Semivolatile organic compounds (SVOCs) are distributed throughout the environment. An example of SVOCs includes polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs) and others. These compounds have been implicated in causing a range of health problems in the immune, endocrine, nervous and reproductive systems of animals and humans [1]. Polychlorinated biphenyls (PCBs) were prohibited in Europe and the USA in the 1970s [2,3] because they are toxic on living organisms and bio-accumulative characteristics. PCB sources contributing the atmospheric concentration levels are burning of PCB containing materials, air-water/soil exchange, vaporization from waste disposal areas sludge dewatering beds, electronic devices containing PCBs [4,5]. Polycyclic aromatic hydrocarbons (PAHs) are organic compounds composed of at least two aromatic rings combined together. PAHs are produced during incomplete combustion of organic materials from sources such as residential heating (coal, wood, and oil), vehicle exhausts, aluminum production, cement manufacture, production of coal tar, coke and asphalt, and petroleum catalytic cracking. Many of PAH compounds have been classified as potential human carcinogens. As a result, they have received widespread interest in recent decades in air pollution studies [6,7]. PAHs are categorized as priority pollutants in the United States Environmental Protection Agency (US EPA) and the European Environment Agency (EEA). Polychlorinated dibenzo-*p*-dioxins (PCDD) are well-known organic pollutants [8,9]. These fine particles play an essential role in the transport of toxic compounds and related increased health risk to humans (US EPA, 1996).

*Corresponding author. E-mail: deeb2000il@yahoo.com

SVOCs concentrations are generally higher in metropolitan areas than in rustic areas [10-13] and show seasonal disparities. SVOCs in atmosphere are realized to partition between gas and particle phase. Gas/particle partitioning of SVOCs is an important topic in understanding the atmospheric behavior of SVOCs. To fully understand atmospheric phenomena and possible health effects, it is necessary to undertake a partitioning study that considers site specific characteristics (particle size, surface area, and mass), source region and meteorological conditions. The gas/particle partitioning coefficient, expressed as K_p, is a necessary input parameter for mathematical models that attempt to interpret SVOCs' transportation and vapor exchange across the air-water interface [8,14,15]. It is also an indispensable parameter to estimate the concentrations of SVOCs [16] in different environmental matrixes. Therefore, information about the K_p values for SVOCs is needed.

However, the experimental procedures determining the values of K_p are always time-consuming, cost-expensive, and difficult to accurately distinguish congener species with similar physicochemical properties [15-17]. A different but highly effective tool depending on QSPR can be utilized to predict K_p values for those compounds with no literature values.

QSPR studies are very active in different fields including chemistry, biochemistry and environmental problems. In QSPR, a correlation between physicochemical properties and molecular descriptors are generated [17]. Once a QSPR model is constructed, the whole congeners can be predicted.

Recently, the direct prediction of gas/particle partitioning for SVOCs using QSPR methods were reported [18] for 209 polychlorinated biphenyl compounds by partial least squaresmethod. In continuation to previous studies [19-23], this study aims to evaluate multivariate statistical models for predication of the gas/particle partitioning coefficients of SVOCs. For this purpose, the relationship between molecular descriptors, related to the factors found experimentally to affect the gas/particle partitioning of SVOCs [6,7] and the logarithm of K_p were comprehensively explored and then QSPR models were constructed using partial least squares (PLS) and artificial neural network (ANN) method. Finally, the performances of the QSPR models were validated.

A total of 70 semivolatile organic compounds were collected. These compounds represent several chemical classes including PAHs, PCBs, Polychlorinated Naphthalenes (PCNs), PBDEs and PCDD/Fs which are pollutants widely identified in atmosphere. The K_p values (25 °C) for PAHs and PCBs are cited from the literature by Finizio *et al*. [24] PCNs, PBDEs and PCDD/Fs are cited from Harner *et al*. [25], Chen *et al*. [26] and Kadowaki *et al*. [27], respectively, according to the method of Finizio *et al*. [24]. These SVOCs together with observed logarithms of their gas/particle partitioning coefficient ($logK_p$) values listed in Table 1.

EXPERIMENTAL

Software

Geometry Optimization was performed by HyperChem

Table 1. Non-Ionic Organic Compounds Used in this Study and Their Experimental $logK_p$ Values

No.	Compound name	$logK_p$	No.	Compound name	$logK_p$
1[b]	Fluorene	-4.48	36[a]	1,2,4,5,8-Petachlorinated naphthalene	-3.39
2[d]	Phenanthrene	-4.20	37[b]	1,2,3,4,6,7-Hexachlorinated napthalene	-3.14
3[a]	Anthracene	-4.27	38[a]	1,2,3,4,5,7-Hexachlorinated napthlene	-3.09
4[a]	Pyrene	-3.27	39[a]	1,2,3,5,7,8-Hexachlorinated napthalene	-3.09
5c	Fluoranthene	-3.37	40a	1,2,4,5,6,8-Hexachlorinated napthalene	-3.04
6a	Chrycene	-2.16	41a	1,2,3,4,5,6-Hexachlorinated napthalene	-2.96

Table 1. Continued

7[c]	Benz[a]anthracene	-2.13	42[a]	1,2,3,4,5,8-Hexachlorinated napthalene	-2.92
8[a]	Benzo[a]pyrene	-1.08	43[a]	1,2,3,4,5,6,7-Heptachlorinated napthalene	-2.60
9[c]	Benz[k]fluoranthene	-1.32	44[b]	1,2,3,4,5,6,8-Heptachlorinated napthalene	-2.56
10[a]	2,4'-Dichlorinated biphenyls	-4.60	45[b]	1,2,3,4,5,6,7,8-Octachlorinated napthalene	-2.16
11[a]	2,3',4'-Trichlorinated biphenyls	-4.03	46[b]	2,4,4'-Tribrominated biphenyle ether	-3.49
12[a]	2,2',3,3'-Tetrachlorinated biphenyls	-3.72	47[c]	2,2,4,4'-Tera brominated biphenyle ether	-2.72
13[a]	2,2'4,5'-Terachlorinated biphenyls	-3.89	48[a]	2,3',4,4'-Tetrabrominated biphenyle ether	-2.59
14[b]	2,2',5,5'-Tetrachlorinated biphenyls	-3.80	49[a]	2,2,'3,4,4'-Pentabrominated biphenyle ether	-1.80
15[c]	2,3',4,4'-Terachlorinated biphenyls	-3.66	50[a]	2,2'4,4',5-Pentabrominated biphenyle ether	-1.98
16[a]	2,3',4,'5-Tetrachlorinated biphenyls	-3.61	51[a]	2,2',4,4',6-Pentabrominated biphenyle ether	-2.14
17[a]	2,4,4',5-Tetrachlorinated biphenyls	-3.63	52[a]	2,2,'3,4,4',5'-Hexabrominated biphenyle ether	-1.28
18[c]	2,2',3,4',5'-Penta chlorinated biphenyls	-3.29	53[c]	2,2',4,4'5,5'-Hexabrominated biphenyle ether	-1.32
19[a]	2,2',4,4',5-Pentachlorinated biphenyls	-3.35	54[b]	2,2',4,4',5,6'-Hexabrominated biphenyle ether	-1.50
20[a]	2,2'4,5,5'-Pentachlorinated biphenyls	-3.39	55[c]	2,2,'3,4,4',5',6-Heptabrominated biphenyle ether	-0.91
21[a]	2,2,3,3',4,4'-Hexachlorinated biphenyls	-2.67	56[c]	1,2,3,4-Tetrachlorinateddibenzodioxine	-2.36
22[b]	2,2',3,4,4',5'-Hexachlorinated biphenyls	-2.79	57[a]	1,2,3,7-Tetrachlorinateddibenzodioxine	-2.41
23[c]	2,2,'3,3'4,4',5-Heptachlorinated biphenyls	-2.88	58[a]	1,3,6,8-Tetrachlorinateddibenzodioxine	-3.01
24[b]	2,2',3,3',4,4',5-Heptachlorinated biphenyls	-2.23	59[c]	2,3,7,8-Terachlorinateddibenzodioxine	-3.32
25[a]	2,2',3,4,4',5,5'-heptachlorinated biphenyls	-2.36	60[b]	1,2,3,4,7-Pentachlorinateddibenzodioxine	-2.17
26[b]	1,3,5-Trichlorinated napthalene	-4.46	61[c]	1,2,3,4,7,8-Hexachlorinateddibenzodioxine	-1.23
27[a]	1,4,6-Trichlorinated napthalene	-4.46	62[a]	1,2,3,4,6,7,8-Heptachlorinateddibenzodioxine	-0.60
28[b]	1,3,5,7-Tetrachlorinated napthalene	-4.15	63[a]	Octachlorinateddibenzodioxine	-0.57
29[b]	1,4,6,7-Tetrachlorinated napthalene	-4.03	64[a]	2,3,7,8-Tetrachlorinated dibenzofuran	-2.40
30[a]	1,2,3,5-Tetrachlorinated napthalene	-3.93	65[a]	2,3,4,7,8-Pentachlorinated dibenzofuran	-2.26
31[c]	1,2,5,8-Tetrachlorinated napthalene	-3.86	66[a]	1,2,3,4,7,8-Hexachlorinated dibenzofuran	-1.91
32[c]	1,2,4,5,7-Pentachlorinated naphthalene	-3.57	67[a]	1,2,3,6,7,8-Hexachlorinated dibenzofuran	-1.91
33[a]	1,2,4,5,6-Pentachlorinated napthalene	-3.46	68[a]	1,2,3,4,6,7,8-Heptachlorinated dibenzofuran	-0.89
34[a]	1,2,4,7,8-Pentachlorinated napthalene	-3.46	69[a]	1,2,3,4,7,8,9-Heptachlorinated dibenzofuran	-0.58
35[a]	1,2,3,5,8-Pentachlorinated napthalene	-3.45	70[b]	Octachlorinated dibenzofuran	-0.44

[a]Compound belongs to the training set. [b]Compound belongs to the test set. [c]Compound belongs to the validation set as applied in the ANN analysis. Note that in PLS analysis, [a]and [c]compounds are belonging to the same set (training set). [d]Utlier.

(Version 7.0 Hypercube, Inc.) at the Austin model 1 (AM1), semi-empirical method level. An AM1 optimization was chosen since it was developed and parameterized for common organic structures. Descriptors were calculated using HyperChem and Dragon software (Milano Chemometrics and QSPR Group, http://www.disat.unimib.it/chm/). SPSS Software (version 13.0, SPSS, Inc.) was used for the simple MLR analysis. PLS, PCA and ANN regression were performed in the MATLAB (Version 7.0.1 (R14), Mathworks, Inc.) environment.

Chemical Data and Descriptors

Compounds name and their logarithms of the gas/particle partitioning coefficient ($\log K_p$) are included in Table 1. Chemical structure of these compounds was obtained from HyperChem software and optimized on AM1 semi-empirical level. The Optimization was preceded by the Polak-Rebiere algorithm to reach 0.01 root mean square gradient. In this study, 19 molecular descriptors including combination of structural descriptors and topological indices were calculated using HyperChem and Dragon software, these descriptors are J, $Jhet_Z$, $Jhet_m$, $Jhet_v$, $Jhet_e$, $Jhet_p$, BAC, $^0\chi$, $^1\chi$, 2χ, $^0\chi^v$, $^1\chi^v$, $^2\chi^v$, MR, MV, PC, Ir, ST, Pl (See Appendix 1).

Multiple Linear Regression (MLR) Analysis

MLR analysis using the method of maximum-R^2 with stepwise selection and elimination of variables [28] was employed to model the logarithms of the gas/particle partitioning coefficient ($\log K_p$) relationships with different set of structural descriptors and topological indices to select initial input models for the artificial neural networks algorithm (ANN).

Principal Component-Artificial Neural Network (PC-ANN)

In contrast to MLR, the artificial neural networks (ANN) are capable of recognizing highly nonlinear relationships. The flexibility of ANN enables it to discover more complex relationships in experimental data, when it is compared with the traditional statistical models. The principal component-artificial neural network (PC-ANN) was proposed by Gemperline *et al*. [29], to improve training speed and decrease the overall calibration error.

In this method [29], as a preliminary treatment, the input data (*i.e.*, molecular descriptors) was normalized so as to have zero mean and unity variance, and then were subjected to principal component analysis (PCA) before being introduced into the neural network. The most significant principal components (PCs), which explain most of the variances in the original data (>95%), were selected, ranked according to decreasing Eigen-value and then used as ANN input. It should be noted for each MLR resulted model separate PC-ANN models were developed so that the input's descriptors were the subsets selected by the stepwise MLR methods.

In the case of each MLR model, a feed-forward neural network with back-propagation of error algorithm was constructed to model the property structure relationships between the extracted PCs of the descriptors in one hand and the logarithm of gas/particle partitioning coefficient data of the semivolatile organic compounds in the second hand. More details about the model development in PC-ANN and the network architecture are explained in references [19,20,22]. Over-fitting problem or poor generalization capability happens when a neural network over learns during the training period.

A too well-trained model may not perform well on unseen data set due to its lack of generalization capability. An approach to overcome this problem is the early stopping method in which the training process is concluded as soon as the overtraining signal appears. This approach requires the data set to be divided into three subsets: training, test and validation sets. The training and the validation sets are the norm in all model training processes. The test set is used to test the trend of the prediction accuracy of the model trained at some point of the training process. At later training stages, the validation error increases. This is the point when the model should cease to be trained to overcome the over-fitting problem. To achieve this purpose, the extracted PCs for each MLR model were classified into training set (60%), validation set (20%) and external test set (20%). Then, the training and validation sets were used to optimize the network performance. The regression between the network output and the property was calculated for the three sets individually. The training function "trainscg" in MATLAB was used to train the network. To find models with lower errors, the ANN algorithm was run many times, each time run with different geometry and/or initial weights.

Partial Least Squares (PLS) Analysis

PLS is a method for building regression models on the latent variable (LV) decomposition relating two blocks, matrices **X** and **Y**, which contain the independent, x, and dependent, y, variables, respectively. In this procedure, it is necessary to find the best number of latent variables, which is normally performed by using cross-validation, based on determination of minimum prediction error. Leave-one-out cross validation was carried out using the NIPALS algorithm. Applications of PLS have been discussed by several workers [21,30,31]. The data was divided into 80% training set and 20% test set. To have comparable data with that used in the ANN analysis, the outliers and test set compounds are kept the same as in the PC-ANN analysis.

RESULTS AND DISCUSSION

MLR Analysis

Table 1 shows the SVOCs listed together with observed $\log K_p$ values. Table 2 records the regression models suggested from MLR analyses on $\log K_p$ and a set of 19 molecular de scriptors including combination of structural descriptors and

Table 2. Correlation Coefficient for MLR, PLS and ANN Models **3-16** and Cross Validation Parameters for PLS and ANN Models

	Descriptors	MLR		PC-ANN					PLS				
M#[a]		R	SE	#PCs	R^c	$R^2_{CV}{}^c$	R^p	PE^p	LV	R^c	$R^2_{CV}{}^c$	R^p	PE^p
3	$^2\chi$, MR, ST	0.972	0.260	2	0.978	0.951	0.988	7.263	2	0.968	0.933	0.984	6.396
4	$^2\chi$, MR, ST, $^0\chi$	0.974	0.253	3	0.978	0.944	0.989	6.47	3	0.971	0.938	0.984	6.546
5	$^2\chi$, MR, ST, $^0\chi$, MV	0.975	0.250	2	0.975	0.904	0.984	10.227	3	0.969	0.936	0.984	6.658
6	$^2\chi$, MR, ST, $^0\chi$, MV, Ir	0.979	0.232	3	0.982	0.958	0.993	6.138	2	0.970	0.936	0.984	6.607
7	$^2\chi$, MR, ST, $^0\chi$, MV Ir, $^1\chi^v$	0.98	0.225	3	0.986	0.966	0.993	6.366	2	0.966	0.958	0.984	6.542
8	$^2\chi$, MR, ST, $^0\chi$, MV, Ir, $^1\chi^v$, $^1\chi$	0.981	0.223	3	0.969	0.9	0.986	9.315	3	0.970	0.938	0.983	6.844
9	$^2\chi$, MR, ST, $^0\chi$, MV, Ir, $^1\chi^v$, $^1\chi$, BAC	0.984	0.207	3	0.972	0.936	0.985	10.306	8	0.980	0.958	0.994	5.365
10	2χ, MR, ST, 0χ, MV, Ir, 1χv, 1χ, BAC, 0χv	0.985	0.203	3	0.974	0.936	0.985	6.536	3	0.966	0.929	0.988	5.669

Table 2. Continued

11	$^{2}\chi$, MR, ST, χ_0, MV, Ir, $^{1}\chi^{v}$, $^{1}\chi$, BAC, $^{0}\chi^{v}$, $Jhet_p$	0.985	0.204	3	0.977	0.946	0.984	7.174	3	0.970	0.937	0.988	6.219
12	$^{2}\chi$, MR, ST, $^{0}\chi$, MV, Ir, $^{1}\chi^{v}$, $^{1}\chi$, BAC, $^{0}\chi^{v}$, Jhetp, PC	0.985	0.206	3	0.972	0.938	0.973	9.546	4	0.972	0.941	0.987	6.423
13	$^{2}\chi$, MR, ST, $^{0}\chi$, $^{0}\chi$v, Ir, $^{1}\chi^{v}$, $^{1}\chi$, BAC, $^{0}\chi^{v}$, $Jhet_p$, PC, $^{2}\chi^{v}$	0.985	0.207	3	0.98	0.947	0.985	8.272	3	0.970	0.937	0.988	6.043
14	$^{2}\chi$, MR, ST, $^{0}\chi$, MV, Ir, $^{1}\chi^{v}$, $^{1}\chi$, BAC, $^{0}\chi^{v}$, $Jhet_p$, PC, $^{2}\chi^{v}$, $Jhet_m$	0.985	0.209	3	0.981	0.96	0.991	6.223	3	0.969	0.935	0.988	5.999
15	$^{2}\chi$, MR, ST, $^{0}\chi$, MV, Ir, $^{1}\chi^{v}$, $^{1}\chi$, BAC, $^{0}\chi^{v}$, V, $Jhet_p$, PC, $^{2}\chi^{v}$, $Jhet_m$, $Jhet_e$	0.985	0.211	3	0.981	0.957	0.986	7.02	3	0.968	0.934	0.987	6.021
16	$^{2}\chi$, MR, ST, $^{0}\chi$, MV, Ir, $^{1}\chi^{v}$, $^{1}\chi$, BAC, $^{0}\chi^{v}$, $Jhet_p$, PC, $^{2}\chi^{v}$, $Jhet_m$, $Jhet_e$, Pl	0.985	0.212	3	0.975	0.939	0.987	7.409	3	0.970	0.937	0.985	6.747

[a]M# is model number, [c]refers to the training (calibration) set and [p]refers to the external test (prediction) set, R is correlation coefficient, $R^2_{CV}(Q^2)$ is cross-validated coefficient of determination, SE is standard error of the estimate, EP is the relative standard error of prediction.

topological indices (See Appendix 1). The number of descriptors in these models is varied between 3 and 16.

The highest MLR correlation coefficient (R) obtained is 0.985 for a regression model with 10-16 descriptors (models **10-16**). Table 2 shows that R is constant for models 10-16 pointing up that adding more than 9 or 10 variables will not improve the correlation. However, model **8** has close coefficient of determination (R^2) to that of models **6** and **7** (~0.96) which implies that 6-8 descriptors are enough to describe the relation with $\log K_p$. This number is less than what is recommended by the rule of the thumb [32]. Artificial neural networks algorithm (ANN) was used seeking to investigate the obtained regression models.

PC-ANN

The inputs of the ANN were the subset of the descriptors used in different MLR models (Table 2). The correlation data matrix for these descriptors is represented in Table 3. As it is observed, some descriptors represent high degree of collinearity. Collinear descriptors add redundancy to the input data matrix and therefore the performances of the models obtained by using these descriptors will be degraded. Principal component analysis (PCA) and more specifically factor analysis (FA) groups together variables that are collinear to form a composite indicator capable of capturing as much of common information of those indicators as possible. Each factor reveals the set of variables having the highest association with it. The idea under this approach is to account for the highest possible variation in the indicators set using the smallest possible number of factors. Therefore, the index no longer depends upon the dimensionality of the dataset but it is rather based on the "statistical" dimensions of the data. Application of PCA on a descriptor data matrix results in a loading matrix containing factors or principal components, which are orthogonal and therefore do not correlate with each other. We used these factors as the inputs of ANN instead of the original descriptors.

Since, the information contents of some extracted features (PC's) may not be in the same direction of the property data, the main problem arises from all of the PCA-based algorithms, is how many and which PC's constitute a good subset for predictive purposes. Different methods have been addressed to select the significant PC's for calibration [32-38]. The simplest and most common one is a top-down variable selection where the factors are ranked in the order of decreasing eigenvalues (eigenvalue ranking, ER). The factors with the highest eigenvalue are considered as the most significant ones, and, subsequently, the factors are introduced into the calibration model until no further improvement of the calibration model is obtained.

Firstly, PCA was used to classify the molecules into training, validation and prediction sets. PCA was performed on the whole data of 70 compounds and 19 descriptors and the first principal was plotted versus the second and third ones. Figure 1 shows the distribution of the data in the space of the first and second PCs (Fig. 1a) as well as their distribution in the space of the first and third PCs (Fig. 1b). Figure 1a shows that compound **2** is an outlier, *i.e.* molecule **2** behaves differently from other molecules with respect to both molecular structure (descriptors) and the logarithm of gas/particle partitioning coefficients ($\log K_p$). Therefore, this molecule was not used in the future analysis. The space of the first and second PCs was not enough to describe the distribution of all the data, hence, the space of the first and third PCs was considered to have clearer picture of the data set distribution. According to the pattern of the distribution of the data in factor spaces (Fig. 1) the training, validation and prediction molecules were selected homogenously, so that molecules in different zones of Figs. 1a and 1b included to all three subsets. After removing the outliers and subjecting the data for the remaining 69 compounds to the preliminary treatment mentioned above, the classified data was used as an input for the ANN.

In this study, a three-layered feed-forward ANN model with back-propagation learning algorithm [39] was employed. First, the nonlinear relationship between the subset of descriptors selected by stepwise selection-based MLR (Table 2) and $\log K_p$ was proceeded by PC-ANN models with similar structure. The number of hidden layer's nodes was set 7 for all models and the number of nodes in the input layer was the number of PCs extracted for each subset of descriptors. The results of PC-ANN modeling for MLR models number **3-16** are given in Table 2.

This table shows that, from a statistical viewpoint, models **6**, **7** and **14** are comparable. This table shows that the mentioned models have almost the highest correlation

Table 3. Correlation Matrix for the Variables Used in ANN Models

	$logK_p$	$Jhet_m$	$Jhet_e$	$Jhet_p$	BAC	$^0\chi$	$^1\chi$	$^2\chi$	$^0\chi^v$
$logK_p$	1								
$Jhet_m$	-0.080	1							
$Jhet_e$	-0.221	0.946	1						
$Jhet_p$	-0.484	0.665	0.746	1					
BAC	0.581	0.614	0.467	0.089	1				
$^0\chi$	0.885	0.008	-0.160	-0.440	0.714	1			
$^1\chi$	0.913	-0.314	-0.453	-0.601	0.460	0.913	1		
$^2\chi$	0.925	-0.152	-0.296	-0.536	0.594	0.950	0.977	1	
$^0\chi^v$	0.726	0.185	0.000	-0.472	0.694	0.800	0.663	0.725	1
$^1\chi^v$	0.782	0.051	-0.131	-0.547	0.614	0.827	0.748	0.783	0.983
$^2X^v$	0.688	0.186	0.015	-0.427	0.621	0.727	0.610	0.668	0.935
MR	0.899	-0.061	-0.219	-0.483	0.532	0.834	0.831	0.834	0.825
Mv	0.738	0.067	-0.120	-0.443	0.653	0.834	0.715	0.755	0.887
PC	0.777	0.087	-0.091	-0.432	0.649	0.841	0.733	0.786	0.854
Ir	0.409	-0.313	-0.267	-0.125	-0.235	0.067	0.329	0.239	-0.080
ST	0.777	-0.107	-0.159	-0.412	0.308	0.559	0.687	0.697	0.388
Pl	0.881	-0.060	-0.216	-0.464	0.533	0.829	0.822	0.824	0.814

	$^1\chi^v$	$^2\chi^v$	MR	MV	PC	Ir	St	Pl
$^1\chi^v$	1							
$^2\chi^v$	0.927	1						
MR	0.869	0.781	1					
MV	0.878	0.805	0.897	1				
PC	0.848	0.778	0.878	0.929	1			
Ir	0.039	-0.006	0.309	-0.139	-0.034	1		
ST	0.454	0.400	0.572	0.259	0.398	0.709	1	
Pl	0.856	0.768	0.987	0.896	0.872	0.286	0.544	1

coefficient for the external test set (0.993, 0.993 and 0.991 for models **6**, **7** and **14**, respectively) which reflects their high predictive power. The PRESS/SST ratio, where PRESS is the predictive residual sum of squares and SST is the regression sum of squares, calculated as

$$(PRESS / SST = \sum_{i=1}^{n} (y_{\exp} - y_{pred})^2 / \sum_{i=1}^{n} (y_{pred} - \bar{y})^2)$$

is an indicator of how reasonable the models are.

These models have a relatively low PRESS/SST ratio (0.042, 0.034 and 0.040 for models **6**, **7** and **14**, respectively) compared with other models which make them the most reasonable models among all. Furthermore, the training set in models **6**, **7** and **14** has a correlation coefficient of 0.982 and 0.986 and 0.981, respectively. The cross-validation coefficients of determination (R^2_{CV} or Q^2) for model **6** are

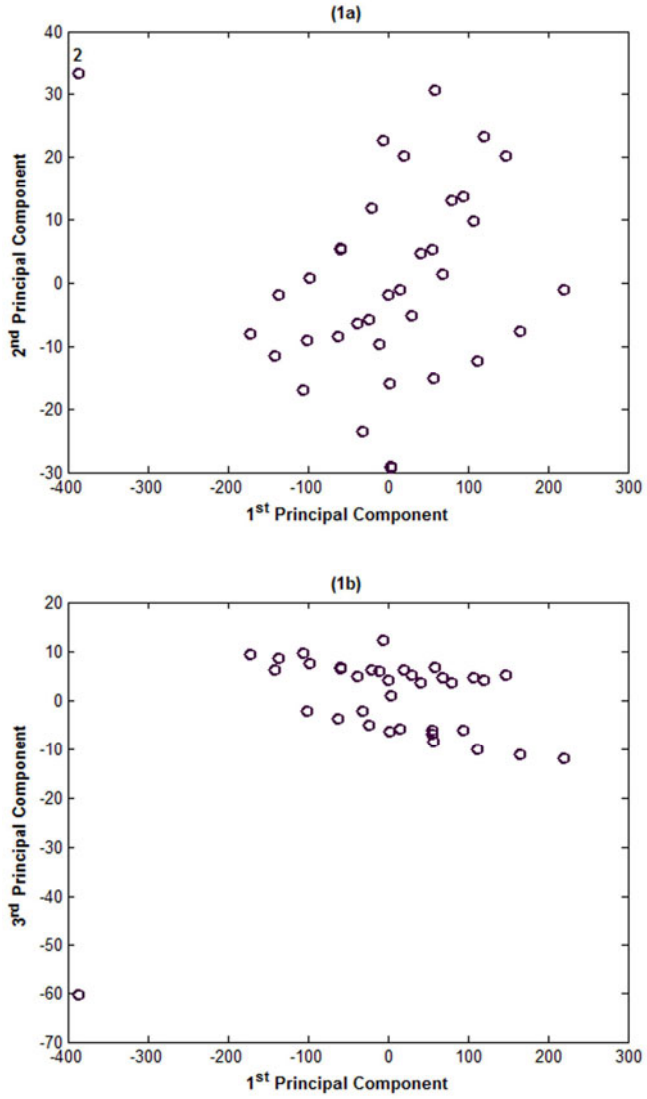

Fig. 1. Correlation of 1st principal component with 2nd principal component (1a) and 1st and 3rd principal components (1b) for the factor spaces of the descriptors and the logarithms of their gas/article partitioning coefficient ($logK_p$).

0.958 and 0.973 for calibration and prediction, respectively. In other words, the three PCs selected by eigenvalue ranking procedure can explain at least 95.8% and 97.3% variances in $logK_p$ for the calibration and prediction, respectively. Model **7**, which models 3 PCs as well, has comparable prediction R^2_{CV} value (0.976) and higher calibration R^2_{CV} (0.966). Model **14**, which models also 3 PCs, has a comparable calibration and prediction cross-validation coefficients of determinations (0.960 and 0.972, respectively) to those of model **6**. Figure 2a shows plots of PRESS against ANN model number**s** (**3**-**16**) for training and test sets. This figure demonstrates that the values of PRESS for ANN models **6**, **7** and **14** are the minima for training and test sets simultaneously which allocates them as nominees for carrying outing the feature analysis. Accordingly, the number of hidden nodes for models **6**, **7** and **14** was optimized and evaluated.

In order to optimize the performance of the suggested ANN models, we trained the ANN using different number of hidden nodes starting from 1 hidden node to 20 hidden nodes. Figure 2b shows plots of PRESS against number of hidden nodes for training and test sets for ANN model **6**. This plot shows that the minima on PRESS curves for the test set is observed when using 7 hidden nodes while that for the training set is detected when using 15 hidden nodes. However, the PRESS value for the training set when using 15 hidden nodes is slightly lower than that when using 7 hidden nodes while the PRESS value for the test set when using 7 hidden nodes is slightly lower than that when using 15 hidden nodes. Needless to say, the difference between the PRESS values for the training or test sets when using 7 or 15 hidden nodes is quite small. Furthermore, it is noticed that the PRESS values for the external test set, in general, are lower than that for the training set.

Table 4 shows regression and cross validation parameters obtained from optimizing the number of hidden nodes for model **6**. This table shows that using 7 and 15 hidden nodes gives the best cross validation parameters. However, recognizing that large numbers of hidden nodes often draw attention to the risk of overfitting [40] implies that using 7 hidden nodes is preferable on using 15.

Figure 2c shows plots of PRESS against number of hidden nodes for training and test sets of ANN model **7**. This plot shows that the minima of PRESS curves for both the training and test sets take place when using 7 hidden nodes. Although the models obtained using 15 and 16 hidden nodes are analogous to that obtained using 7 hidden nodes, the later number of hidden nodes gives a prediction error (PE) of 0.472% while using 15 or 16 hidden nodes give a prediction

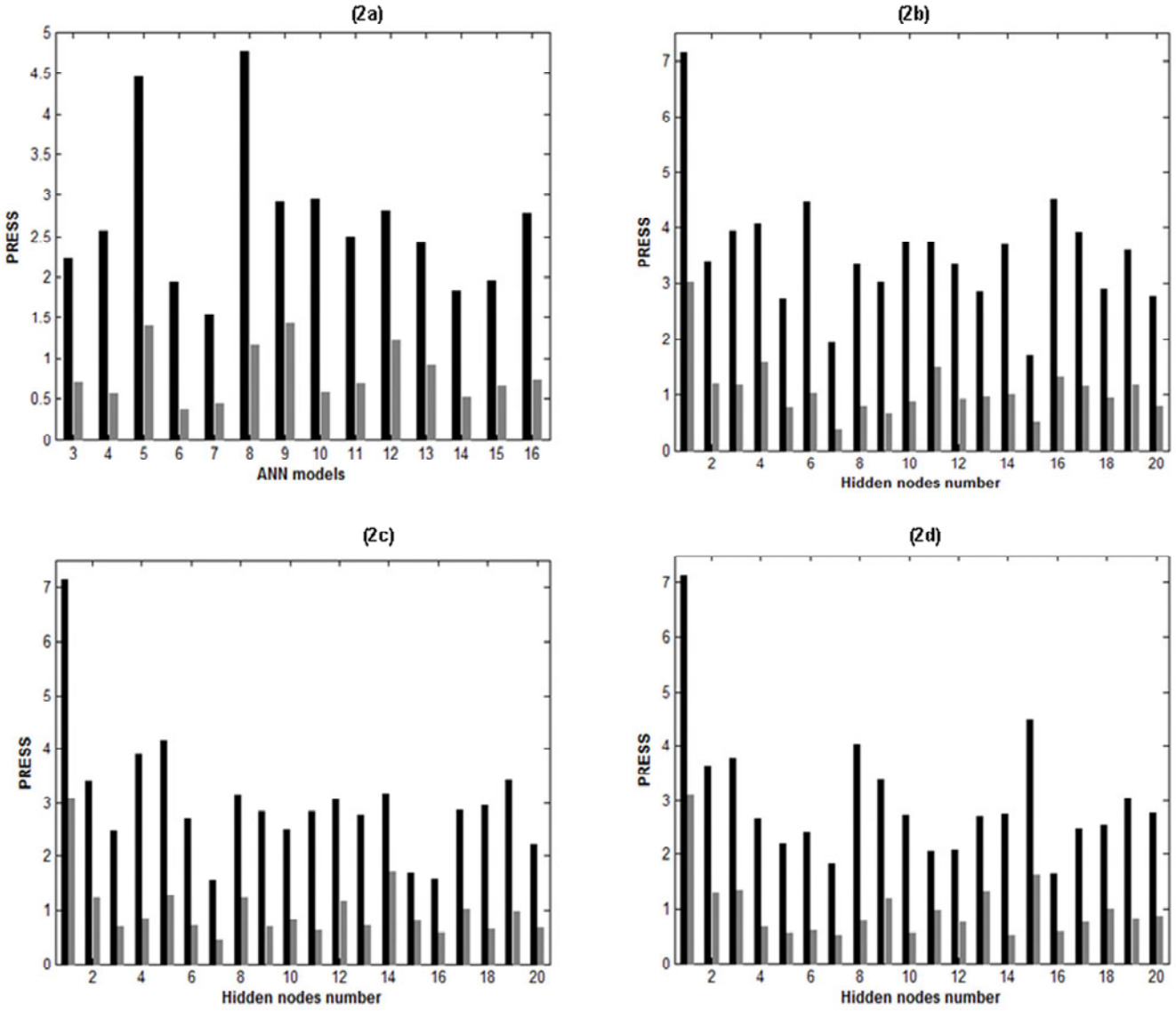

Fig. 2. Correlation of PRESS with: (2a) ANN models (**3-16**), (2b) different numbers of hidden nodes for model **6**, (2c) different number of hidden nodes for model **7** and (2d) different number of hidden nodes for model **14**. Black and grey columns represent PRESS values for the training and test sets, respectively.

error of 0.631% and 0.545%, respectively. Therefore, using 7 hidden nodes is considered the optimal number for this model given that the evaluation of the predictive ability of a multivariate calibration model is based on determination of minimum prediction error [41].

Table 4 shows the results for optimizing the number of hidden nodes for model 14. An ANN with 7 hidden nodes gives a reasonable model with low PRESS/SST ratio (0.034) and high correlation coefficient for both the training set (0.986) and the prediction set (0.993). This model gives and R^2_{CV} of 0.966. Figure 2d shows plots of PRESS against number of hidden nodes for training and test sets for this

Table 4. Correlation Coefficients and Cross Validation Parameters for ANN Models **6**, **7** and **14** Using Different Numbers of Hidden Nodes

	Model 6				Model 7				Model 14			
hn#[a]	R^c	$R^2_{CV}{}^c$	R^p	PE (%)^p	R^c	$R^2_{CV}{}^c$	R^p	PE (%)^p	R^c	$R^2_{CV}{}^c$	R^p	PE (%)^p
1	0.958	0.843	0.974	15.054	0.956	0.843	0.974	15.176	0.954	0.844	0.969	15.212
2	0.970	0.926	0.983	9.430	0.971	0.926	0.983	9.527	0.969	0.921	0.978	9.918
3	0.965	0.914	0.981	9.405	0.976	0.946	0.987	7.136	0.965	0.918	0.981	10.040
4	0.970	0.911	0.985	10.869	0.961	0.916	0.977	7.978	0.971	0.942	0.984	7.027
5	0.970	0.941	0.984	7.588	0.959	0.909	0.982	9.820	0.976	0.951	0.987	6.417
6	0.969	0.906	0.982	8.686	0.973	0.941	0.988	7.283	0.976	0.948	0.988	6.746
7	0.982	0.958	0.993	6.136	0.986	0.966	0.993	5.758	0.981	0.96	0.991	6.221
8	0.963	0.927	0.985	7.698	0.968	0.931	0.975	9.588	0.968	0.914	0.982	7.637
9	0.968	0.934	0.983	6.966	0.969	0.938	0.982	7.222	0.964	0.926	0.985	9.357
10	0.961	0.918	0.979	8.112	0.973	0.946	0.982	7.832	0.977	0.941	0.987	6.404
11	0.971	0.918	0.986	10.564	0.974	0.938	0.985	6.795	0.98	0.955	0.986	8.503
12	0.963	0. 297	0.98	8.295	0.973	0.933	0.985	9.308	0.979	0.955	0.985	7.515
13	0.972	0.938	0.983	8.466	0.971	0.939	0.982	7.295	0.979	0.941	0.985	9.930
14	0.959	0.919	0.986	8.661	0.972	0.931	0.984	11.308	0.982	0.941	0.988	6.270
15	0.985	0.963	0.988	6.087	0.983	0.963	0.983	7.698	0.979	0.902	0.986	10.991
16	0.953	0.902	0.969	9.954	0.983	0.965	0.984	6.648	0.984	0.964	0.985	6.612
17	0.962	0.915	0.977	9.308	0.970	0.938	0.979	8.734	0.975	0.946	0.983	7.478
18	0.969	0.937	0.980	8.393	0.972	0.936	0.983	6.978	0.980	0.945	0.983	8.686
19	0.967	0.921	0.985	9.344	0.968	0.925	0.984	8.588	0.971	0.934	0.987	7.881
20	0.970	0.940	0.983	7.673	0.976	0.951	0.984	7.051	0.971	0.939	0.979	8.088

[a]hn# refers to hidden nods number.

model.

This figure shows that the minima of PRESS curves for the training set is obtained using 7 or 16 hidden nodes (1.824 and 1.628, respectively) while those for the test set are 0.518 and 0.585, respectively. However, following the same argument used for model **6**, the preference is for the model obtained using the smaller number of hidden nodes (7 in this case) to avoid overfitting risk. Table 4 shows regression and cross validation parameters for optimizing the number of hidden nodes for model **7**. An ANN with 7 hidden nodes gives the most reasonable model with low PRESS/SST ratio (0.040) and a very high calibration correlation coefficient of 0.981 along with the highest prediction correlation coefficient (0.991). This model gives an R^2_{CV} of 0.960 and prediction error of 0.510%.

Comparing the cross-validation parameters obtained for the three models mentioned above, it can be noticed that these models are in close proximity from a statistical point of view. Again, as stated in [41], choosing the best model is normally achieved by using cross-validation, based on determination of minimum prediction error and for the optimal number of hidden nodes for model 7 (7 hidden nodes) has the lowest prediction error (0.472%) compared with those for models **6** and **14** (0.503% and 0.510%, respectively), model **7** obtained using 7 hidden nodes is regarded as the optimal one. This model contains the following seven descriptors: $^2\chi$, MR, ST, $^0\chi$, MV Ir, $^1\chi^v$ (see Appendix 1) represented by 3 PCs.

Table 5 shows regression and cross validation parameters for randomization test that is performed to investigate the probability of chance correlation for model 7 obtained using 7

Table 5. Correlation Coefficient and Cross Validation Parameters for Chance Correlation Investigation for Model **7** Using 7 Hidden Nodes

Trial	R^{tr}	$PRESS^{tr}$	$R^2_{CV}{}^{tr}$	R^{ts}	$PRESS^{ts}$	$R^2_{CV}{}^{ts}$
1	-0.140	52.606	-0.152	0.011	20.094	-0.040
2	0.517	35.484	0.225	0.649	14.279	0.255
3	-0.282	71.015	-0.544	-0.511	38.343	-0.702
4	0.234	45.521	0.003	0.040	20.943	-0.095
5	-0.911	116.847	-1.510	-0.963	52.538	-1.400
6	-0.124	50.768	-0.099	0.165	18.294	0.026
7	-0.479	63.063	-0.365	-0.582	26.993	-0.286
8	0.421	39.472	0.167	0.773	11.133	0.405
9	-0.442	77.593	-0.460	-0.633	28.445	-0.467
10	-0.576	67.892	-0.475	-0.546	27.260	-0.321

hidden nodes. This table shows that the correlation coefficients obtained by chance are low while PRESS and PRESS/SST ratio are high indicating that the proposed optimal PCA-ANN model is superior to that obtained by chance.

Figure 3 shows regression between observed and predicted $\log K_p$ as well as their residuals for training, validation and test sets of model **7** obtained using 7 hidden nodes.

However, the correlation coefficients obtained from MLR and ANN analysis are close which lead to the conclusion that the linear and nonlinear regressions are equal for this data set. Hence, we have carried out PLS analysis to furthermore examination of these models.

PLS

More on inspection the obtained models, a PLS analysis with cross validation was carried out. Model validation was achieved through leave-one-out cross-validation (LOO CV) and external validation (for a test set), and the predictive ability was statistically evaluated through the root mean square errors of calibration and validation. The calibration and prediction qualities were quantified with R^2 (training set) and R^2_{CV} (leave one out cross-validation on training set), select the LV when the R^2_{CV} has a high number, or determine it by computing the prediction error sum of squares (PRESS) for cross-validated models. PRESS is a standard index to measure the accuracy of a modeling method based on the cross-validation technique.

The cross-validation method employed was to eliminate only one sample at a time and then PLS calibrate the remaining standard descriptor. By using this calibration the $\log K_p$ of the sample, left out was predicted. This process was repeated until each standard had been left out once. Figure 4a and 4b show the associated PRESS, prediction error (PE%) and R^2_{CV} values for each model. Table 2 shows that the minimum prediction error (5.365) occurs for model 9. The cross validation coefficient of determination for this model is the highest (0.958). This model has the lowest PRESS values for the training and test sets at the same time and has the lowest PE%. While other models have higher R^2_{CV} values but also have higher PRESS values for the training set. Accordingly model 9 is the best model according to PLS analysis. This model has a regression coefficient of 0.980 and 0.994 for the training and tests sets, respectively. The PRESS/SST value for this model (0.034) shows that this is an excellent model.

Table 6 shows regression and cross validation parameters for randomization test that is performed to investigate the probability of chance correlation for model 9 using PLS analysis. This table shows that the proposed optimal PLS model is superior to that obtained by chance.

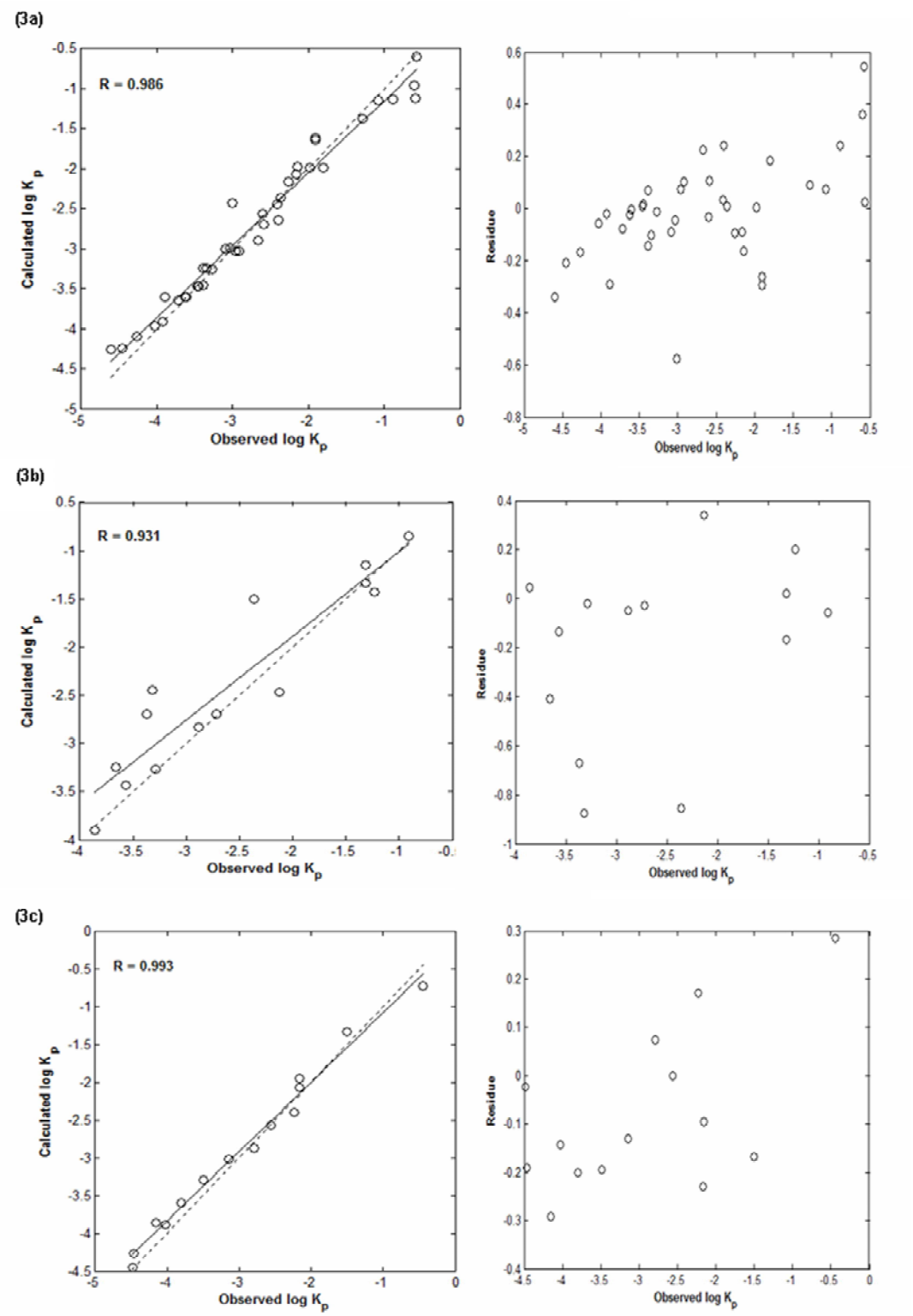

Fig. 3. Correlation of the predicted logK_p against observed one as well as their residues for (3a) training set, (3b) validation set and (3c) external test set of ANN model **7** using 7 hidden nodes.

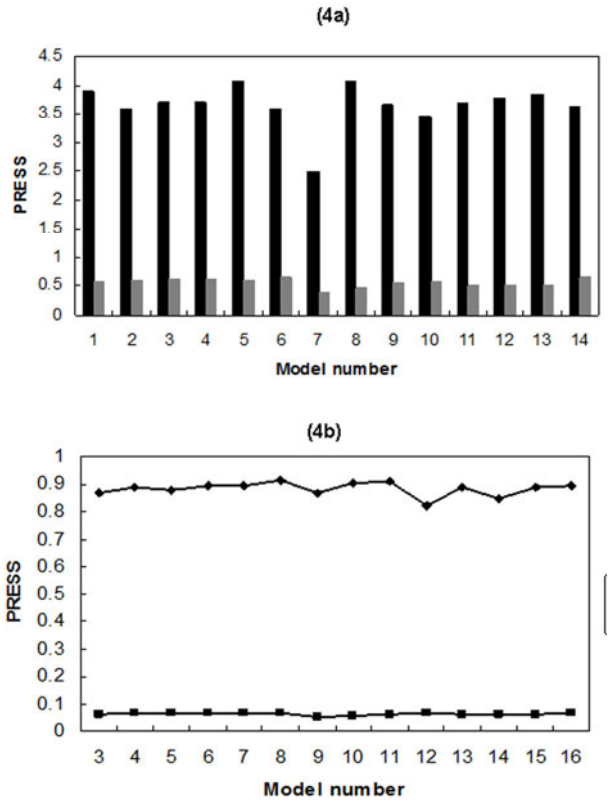

Fig. 4. (4a) Correlation of PRESS with PLS models (**3-16**). Black and grey columns represent PRESS values for the training and test sets, respectively. (4b) Correlation of R^2_{CV} and PE% with PLS models (**3-16**): (■) PE%, (◆) R_2CV.

Figure 5 shows regression between observed and predicted $logK_p$ as well as their residuals for training and test sets of model **9** using PLS analysis. This model contains the following nine descriptors which are represented by 8 latent variables (LV): $^2\chi$, MR, ST, $^0\chi$, MV, Ir, $^1\chi^v$, $^1\chi$, BAC (see Appendix 1). However the optimal models obtained from the PLS and ANN analysis are close to each other.

The following conditions proposed by Golbraikh and Tropsha [42] were applied to conclude that the QSAR model has acceptable prediction power if: (1) $R^2_{CV} > 0.5$; (2) $R^2 > 0.6$; (3) $(R^2 - R^2_0)/R^2 < 0.1$ and $0.85 < k < 1.15$ or $(R^2 - R'^2_0)/R^2 < 0.1$ and $0.85 < k' < 1.15$, where R^2_0 and R'^2_0 are the coefficients of determination characterizing linear regression with Y-intercept set at zero, the first associated with observed *vs*. predicted values, the second related to predicted *vs*. observed values; k and k' are the slopes of the regression lines forced through zero, relating observed *vs*. predicted and predicted *vs*. observed values. Alternatively, the parameter R^2_m, where $R^2_m = R^2 * (1 - (R^2-R^2_0)^{1/2})$, can be used [43]. This parameter, which penalizes a model for large differences between observed and predicted values, was also calculated. R^2_m should be larger than 0.5 for a good external prediction, which is the case for model **7** from the ANN analysis (R^2_m = 0.934) and model **9** from the PLS analysis (R^2_m = 0.911). If a model shows good statistical performance for all these criteria, on both the training and the test sets, its reliability and robustness are high as it is achieved in this study.

The descriptors used in these models are in consistence

Table 6. Correlation Coefficient and Cross Validation Parameters for PLS Chance Correlation Investigation for Models 3-16 after 100 Trials

Statistics		Mod3	Mod4	Mod5	Mod6	Mod7	Mod8	Mod9
R^2 for original Y		0.9371	0.9419	0.9398	0.9401	0.9339	0.9416	0.9598
R^2 for randomized-Y	Avg	0.0980	0.1110	0.1260	0.1490	0.0760	0.1010	0.0850
	SD	0.0070	0.0130	0.0080	0.0210	0.0180	0.0110	0.0070
R^2_{cv} for randomized-Y	Avg	0.0950	0.0980	0.1180	0.1420	0.0730	0.1000	0.0790
	SD	0.0080	0.0110	0.0120	0.0170	0.0160	0.0090	0.0050
		Mod10	Mod11	Mod12	Mod13	Mod14	Mod15	Mod16
R^2 for original Y		0.9340	0.9407	0.9441	0.9403	0.9387	0.9377	0.9409
R^2 for randomized-Y	Avg	0.1440	0.1000	0.1220	0.1480	0.1310	0.1150	0.0880
	SD	0.0070	0.0120	0.0070	0.0040	0.0210	0.0080	0.0140
R^2_{cv} for randomized-Y	Avg	0.1390	0.0960	0.1150	0.1440	0.1290	0.1080	0.0810
	SD	0.0130	0.0100	0.0070	0.0210	0.0170	0.0150	0.0100

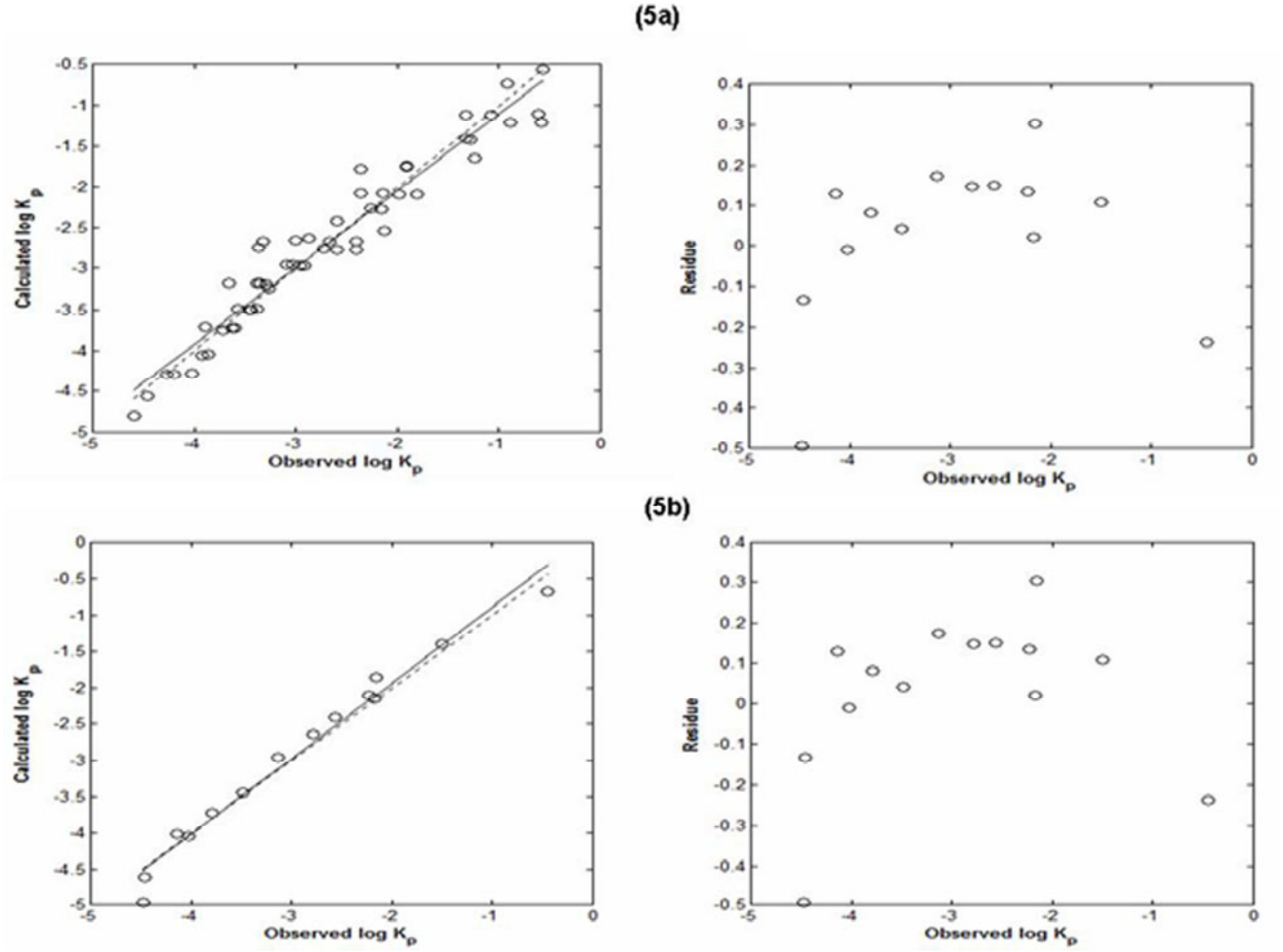

Fig. 5. Correlation of the predicted logK_p against observed one as well as their residues for (5a) training set, (5b) external test set of model **9** as observed from PLS analysis.

with the suggested experimental factors to affect the gas/particle partitioning. Most of these descriptors depend on the pressure and temperature or both of them. The refractivity and polarizability can be an indicator of the gas/particle partitioning. On the other hand, the surface area depends on the particle size. The connectivity indices depend on atoms connectivity and racimification of the molecules, which correlates with its volume of course.

Table 2 shows that the PC-ANN and PLS models are very comparable. However, the prediction error for the PLS optimal model (5.365) is less than that for the ANN optimal model (6.366) while the correlation coefficient of the prediction set is almost the same for both models (0.994 and 0.993 for PLS and ANN, respectively). On the other hand, the correlation coefficient of the training set for the ANN optimal model (0.986) is slightly higher than that for the PLS model (0.980) while the cross validation coefficient of determination for the ANN optimal model (0.966) is higher than that for the PLS optimal model (0.958). The above discussion indicates that the ANN optimal model has a slightly improved generalization power while the PLS optimal model has a slightly improved prediction power. However, considering the number of

descriptors used in each model, the ANN optimal model with the less number of variables (7 descriptors represented by 3 PC's) is preferred over the PLS optimal model (9 descriptors represented by 8 LV's).

Comparison with other Studies

Wei *et al*. [18] have performed a study on some PCBs using Partial Least-Squares Regression method (PLS). The highest R^2 and R^2_{CV} they obtained are 0.987 and 0.985, respectively for a set of 20 compounds without using an external test set while we obtained R^2 around 0.97 and R^2_{CV} of 0.966 and 0.958 and using an external test set of 14 compounds from the ANN and PLS analyses, respectively. R^2 and R^2_{CV} obtained in this study are lower than those obtained by Wei *et al*. [18] for a smaller size training set where such small size of the data set may be a possible sign of an overfitted model simply leading to predictions that are far beyond the range of the training. However, their model represents $\log K_p$ of one type of compounds while our model represents $\log K_p$ of more than one group of compounds which is expected to lower its predictivity power numerically and on the other hand, increasing its reliability.

The model of Wei *et al*. [18] contains the descriptors Σq^2_C (sum of squared atom electron densities on carbon atoms in a given molecule) and Σq^2_H (sum of squared atom electron densities on hydrogen atoms in a given molecule) while the optimal model suggested in this study contains the following nine descriptors: $^2\chi$, MR, ST, $^0\chi$, MV, Ir, $^1\chi^v$, $^1\chi$, BAC (see Appendix 1) which are consistent with the experimental results [6,7].

CONCLUSIONS

A quantitative-structural relationship analysis has been performed on the logarithm of gas/particle partitioning coefficient ($\log K_p$) for 70 different semivolatile organic compounds by using the PC-ANN modeling method, with application of eigenvalue ranking factor selection procedure. The PC-ANN gives good regression models with good prediction ability using a relatively low number of PCs. The optimal models obtained by PC-ANN and PLS analyses are in close proximity from the statistical point of view. The results obtained offers excellent regression models that hold good prediction ability using a relatively low number of PCs compared with other studies on the same data set of compounds. A coefficient of determination around 0.97 was obtained using PC-ANN and PLS analysis.

Appendix 1

J is Balaban index

$Jhet_Z$ is Balaban-type index from Z weighted distance matrix (Barysz matrix)

$Jhet_m$ is Balaban-type index from mass weighted distance matrix

$Jhet_v$ is Balaban-type index from van der Waals weighted distance matrix

$Jhet_e$ is Balaban-type index from electronegativity weighted distance matrix

$Jhet_p$ is Balaban-type index from polarizability weighted distance matrix

BAC is Balaban centric index

$^0\chi$ is connectivity index chi-0

$^1\chi$ is connectivity index chi-1 (Randic connectivity index)

2χ is connectivity index chi-2

$^0\chi^v$: valence connectivity index chi-0

$^1\chi^v$ is valence connectivity index chi-1

$^2\chi^v$ is valence connectivity index chi-2

MR is Ghose-Crippen molar refractivity

MV is mean atomic van der Waals volume (scaled on Carbon atom)

PC is Parachore

Ir is index of refraction

ST is surface tension

Pl is polarzability

REFERENCES

[1] H.W. Vallack, D.J. Bakker, I. Brandt, E. Brostrom-Lunden, A. Brouwer, K.R. Bull, C. Gough, R. Guardans, I. Holoubek, B. Jansson, et al. Environ. Toxicol. Phar. 6 (1998) 143.

[2] M.D. Erickson, Analytical Chemistry of PCBs, CRC Press LLC, Boca Raton, FL, 1997.

[3] L. Manodori, A. Gambaro, I. Moret, G. Capodaglio, W.R.L. Cairns, P. Cescon, Chemosphere 62 (2006) 449.

[4] M. Biterna, D. Voutsa, Environ. Int. 31 (2005) 671.

[5] Y. Tasdemir, M. Odabasi, T.M. Holsen, Atmos. Environ. 39 (2005) 885.
[6] S. Lee et al. Environ. Pollut. 153 (2008) 215.
[7] N. Vardar, et al. Environ. Pollut. 155 (2008) 298.
[8] T.F. Bidleman, Environ. Sci. Technol. 22 (1988) 361.
[9] R. Lohmann, K.C. Jones, Sci. Total Environ. 219 (1998) 53.
[10] R. Lohmann, G.L. Northcott, K.C. Jones, Atmosphere. Environ. Sci. Technol. 34 (2000) 2892.
[11] A. Gambaro, L. Manodori, I. Moret, G. Capodaglio, P. Cescon, ABC 378 (2004) 1806.
[12] Y. Tasdemir, N. Vardar, M. Odabasi, T.M. Holsen, Environ. Pollut.131 (2004) 35.
[13] S. Cindoruk, Y. Tasdemir, Environ. Pollut. 148 (2007) 325.
[14] R.M. Hoff, et al., Atmos. Environ. 30 (1996) 3505.
[15] D. Mackay, S. Paterson, Envir. Sci. Technol. 20 (1986) 810.
[16] D. Mackay, S. Paterson, Envir. Sci. Technol. 25 (1991) 427.
[17] M. Karelson, V.S. Lobanov, A.R. Katritzky, Chem. Rev. 96 (1996)1027.
[18] B. Wei, *et al.*, Chemosphere 66 (2007) 1807.
[19] B. Hemmateenejad, M. Shamsipur, Internet Electron J. Mol. Des. 3 (2004) 316.
[20] B. Hemmateenejad, M.A. Safarpour, R. Miri, N. Nesari, J. Chem. Inf. Model 45 (2005) 190.
[21] O. Deeb, B. Hemmateenejad, A. Jaber, R. Garduno-Juarez, R. Miri, Chemosphere 67 (2007) 2122.
[22] O. Deeb, B. Hemmateenejad, Chem. Biol. Drug Des. 70 (2007) 19.
[23] M. Goodarzi, M.P. Freitas, Chemom. Int. Lab. Sys. 96 (2009) 59.
[24] A. Finizio, et al., Atmos. Environ. 3 (1997) 2289.
[25] T. Harner, T.F. Bidleman, Environ. Sci. Technol. 32 (1998) 1494.
[26] L.G. Chen, et al., Environ. Sci. Technol. 40 (2006) 1190.
[27] S. Kadowaki, H. Naitoh, Chemosphere 59 (2005) 1439.
[28] Chaterjee, S. Hadi, A.S. Price, B. Regression Analysis by Examples, 3rd ed., Wiley, New York, 2000.
[29] P.J. Gemperline, J.R.V. Long, G. Gregoriou, Anal. Chem. 63 (1991) 2313.
[30] M. Goodarzi, M.P. Freitas, J. Phys. Chem. A 112 (2008) 11263.
[31] M. Goodarzi, M.P. Freitas, R. Jensen, J. Chem. Inf. Model. 49 (2009) 824.
[32] M.S. Tute, in: N.J. Harter, A.B. Simmord (Eds.), History and Objectives of Quantitative Drug Design in Advances in Drug Research, Vol. 6, Academic Press, London, 1971.
[33] Y.L. Xie, J.H. Kalivas, Anal. Chim. Acta 348 (1997) 19.
[34] J.M. Sutter, J.H. Kalivas, J. Chemom. 6 (1992) 217.
[35] J. Sun, J. Chemom. 9 (1995) 21.
[36] U. Depczynski, V.J. Frost, K. Molt, Anal. Chim. Acta 420 (2000) 217.
[37] A.S. Barros, D.N. Rutledge, Chemom. Intell. Lab. Syst. 40 (1997) 65.
[38] J.D. Verdu-Andres, L. Massart, Appl. Spectrosc. 52 (1998)1425.
[39] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Nature 323 (1986) 533.
[40] E.P.P.A Derks, L.M.C. Buydens, Chemom. Intell. Lab. Syst. 41 (1998) 171.
[41] H. Martens, T. Naes, Multivariate Calibration. John Wiley, Chichester, 1989.
[42] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269.
[43] K. Roy, P. Roy, Europ. J. Med. Chem. 44 (2009) 2913.