Quantitative Structure-Retention Relationship Study of Some Phenol Derivatives in Gas Chromatography

Zahra Garkani-Nejad

Chemistry department, Faculty of Science, Vali-e-Asr University, Rafsanjan, Iran

Abstract

A quantitative structure-retention relationship (QSRR) model has been developed for the gas chromatographic retention times of 37 phenolic derivatives in a DB-5 non-polar column (95% dimethyl and 5% diphenyl-polysiloxane). As a first step, multiple linear regression (MLR) was employed to gain informative descriptors that can predict the retention times of these compounds. Descriptors appearing in the MLR model are categorized as topological and geometric parameters that comply with the applied column. Furthermore, each molecular descriptor in this model was examined to unfold the relationship between molecular structures and their retention times. Then, a 4-4-1 neural network was developed using the descriptors selected by the MLR model. The comparison of the standard errors and correlation coefficients reveals the superiority of artificial neural networks (ANN) over the MLR model. This refers to the fact that the retention behaviors of molecules display non-linear characteristics. The consistency and reliability of ANN model was investigated using the L4O cross-validation technique. The obtained results are closely in compliance with the experiment. Moreover, the mean effect of descriptors shows that Kier symmetry index is the most important factor affecting the retention behavior of molecules.

Introduction

Gas chromatography (GC) is mainly used as a criterion of purity degree in organic compounds. This technique is mostly utilized to measure the efficiency of the purification processes. Theoretically, the retention times can be used for the identification of compounds. This compound identification is often completed by comparing the GC peak with standard samples of the suspected material. However, obtaining samples of pure standard materials is not always possible. Thus, it seems essential to develop a theoretical model for estimating the retention times.

Quantitative structure-retention relationships (QSRRs) represent a powerful tool in chromatography. The principal aim of QSRR is to predict retention data from the molecular structure. One of the crucial problems is how to represent molecular structure for QSRR. Generally, the descriptors encoding the molecular structure are classified as physicochemical, quantum-chemical, topological, etc., descriptors. A key to building successful QSRR models is a proper feature selection (i.e., selection of the most relevant descriptors from a large number of inputs) (1–2). Thus, it is essential to select the method that has been described the most, and in most cases, the multiple linear regression (MLR) technique has been used for this purpose (1–8).

Also, the use of artificial neural networks (ANN) in the modeling of retention behavior and optimization of conditions in chromatography has been studied (1,2,9–15). Compared to MLR, ANN is a more flexible modeling methodology because both linear and non-linear functions can be used (or combined) in the processing units. This allows the description of more complex relationships between a high-dimensional descriptor space and the given retention data, which may lead to better predictive power of the resulting ANN model compared to MLR.

In this study, the retention times of phenol derivatives have been predicted using ANN as modeling tool, and usefulness of the neural network was compared with the MLR technique.

Theory

ANN

ANNs are mathematical systems that simulate biological neural networks (16–18). They consist of processing elements (nodes, neurons) organized in layers. Back-propagation neural networks (BPNNs) are most often used in analytical applications. The BPNN receives a set of inputs, which are multiplied by each node and then a nonlinear transfer function is applied. The goal of training the network is to change the weights between the layers in a direction to minimize the output errors. The changes in the values of the weights can be obtained using Eq. 1:

$$\Delta W_{ij}(n) = \eta \delta_i O_j + \alpha W_{ij}(n-1) \qquad \text{Eq. 1}$$

where ΔW_{ij} is the change in the weight factor for each network node, δ_i is the actual error of node i, and O_j is output of node j. The coefficients η and α are the learning rate and the momentum factor, respectively. The goal of training process is to find the optimum weights, and the process starts with random connection weights. The computed output (O_{pm}) is compared to target value (T_{pm}) (i.e., experimental retention times in this work), and an error term $(T_{pm} - O_{pm})^2$ is determined. The mean

* Author to whom correspondence should be addressed: E-mail garakani@mail.vru.ac.ir.

square error (MSE) is used as a criterion for finalizing the learning process and computed using the following equation:

$$MSE = \frac{1}{P \times M} \sum_{p=1}^{P} \sum_{m=1}^{M} (T_{pm} - O_{pm})^2 \qquad \text{Eq. 2}$$

where M is the number of neurons in output layer and P denotes the number of patterns (i.e., the number of experimental retention time data employed in the training process of the network). The number of neurons in the hidden layer and epochs has been optimized by minimizing MSE term.

Cross-validation technique

The consistency and reliability of a method can be explored using the cross-validation technique (19). Two different strategies of leave-one-out (LOO) and leave-multiple-out (LMO) can be carried out in this method. In the LOO strategy, by deleting each time one object from the training set, a number of models will be produced. Obviously, the number of models produced by the LOO procedure is equal to the number of available examples *n* (*n* = 37). Prediction error sum of squares (PRESS) is a standard index to measure the accuracy of a modeling method based on the cross-validation technique. Based on the PRESS and SSY (sum of squares of deviations of the experimental values from their mean) statistics, the Q^2_{LOO} can be easily calculated by Eq. 3:

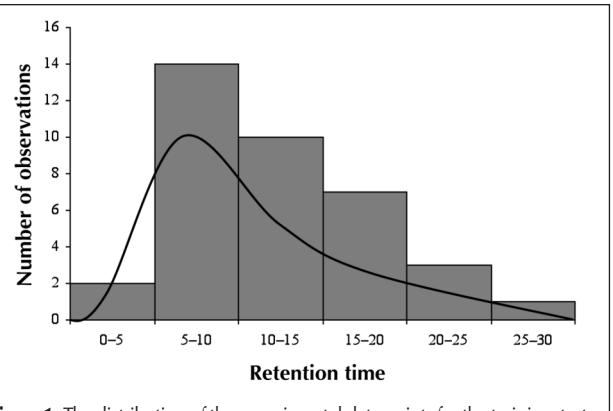

Figure 1. The distribution of the experimental data points for the training, test, and prediction sets.

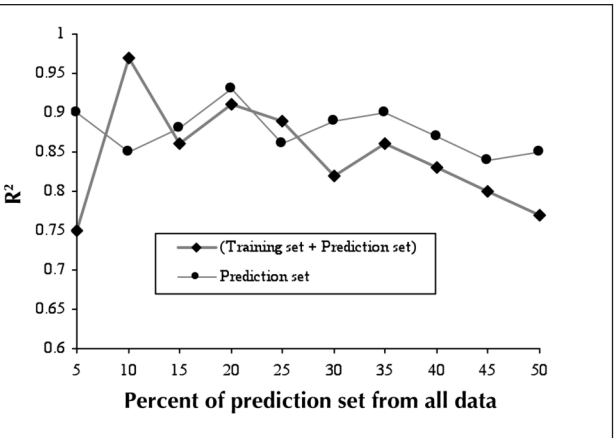

Figure 2. Effect of percent of prediction set on the squared correlation coefficient of (training set + prediction set) and prediction set in MLR.

$$Q^2_{LOO} \frac{PRESS}{SSY} = 1 - \frac{\sum_{i=1}^{n} (y_{exp} - y_{pred})^2}{\sum_{i=1}^{n} (y_{exp} - \bar{y})^2} \qquad \text{Eq. 3}$$

In the case of LMO, M represents a group of randomly selected data points that would be left out at the beginning and be predicted by the model, which was developed using the remaining data points. So, M molecules are considered as prediction set. The R^2_{LMO} can be calculated:

$$R^2_{LMO} \frac{PRESS}{SSY} = -1 \frac{\sum_{i=1}^{test} (y_{exp} - y_{pred})^2}{\sum_{i=1}^{test} (y_{exp} - \bar{y}_{train})^2} \qquad \text{Eq. 4}$$

It is common to choose 10–15% of the total number of molecules to be left out. Therefore, in the present work, calculation of R^2_{LMO} was based on 60 random selections of groups of four samples. The higher the Q^2_{LOO} or R^2_{LMO} the higher the predictive power of the model. The detailed description of this method can be found elsewhere (19).

Experimental

Data set

The retention times of a series of phenol derivatives consisting of 37 molecules were taken from the literature (20) as the data set. The applied approach in this reference involves the use of fused-silica, wide-bore, and open-tubular columns with different polarities. The column used in this work is a nonpolar DB-5 (30 m × 0.53 mm i.d., cross-linked and chemically bonded with 95% dimethyl and 5% diphenyl-polysiloxane, 0.83-µm or 1.5-µm film thickness). This column is connected to an injection tee and an electron capture detection system (ECD). Chromatographic conditions were: column temperatures programmed from 150 to 275°C at a rate of 3°C/min. Injector and detector temperatures were 250°C and 320°C, respectively. Helium and nitrogen were used as a carrier and makeup gases, and the flow rate of these gases were 6 mL/min and 20 mL/min, respectively. The interaction between the phenol derivatives and the stationary phase of the columns makes the derivatives separate.

In the present work, the relationship between the structure and the retention times of these compounds has been studied in the aforementioned DB-5 column. Experimental retention times of these compounds are illustrated in Table I. The distribution of experimental data and the retention time values for the full set of 37 derivatives of phenol is shown in Figure 1.

Descriptor generation

The selection and calculation of the structural descriptors as numerical encoded parameters reflecting the chemical structures is regarded as an essential step in every quantitative structure-activity relationship (QSAR) and QSRR study. To calculate the molecular descriptors, the three-dimensional structures of the studied molecules were generated and optimized using semi-empirical, quantum-chemical methods of AM1 Hamiltonian implemented in a Hyperchem package (21). In the present study, 12 molecular descriptors were produced by applying the Hyperchem package after improving their structures. Then

Dragon software (version web 5) was also utilized for producing additional descriptors (22). Dragon software has been widely used for calculating chemical descriptors in many QSRR and QSAR studies (23–26). The calculated descriptors using this software were divided into 20 groups as shown in Table II. A total of 1661 descriptors were calculated for each molecule using this software. Finally, a total of 301 (289 from Dragon + 12 from Hyperchem) descriptors were calculated for each compound. These descriptors were obtained from a large number of descriptors after removing the parameters, which have more than 10% constant or zero values.

Regression analysis

Generation models were formed by the use of a step-wise multiple linear regression procedure. Only one of the variables from pairs with $R > 0.95$ was used in modeling. From 301 original descriptors, 100 descriptors were eliminated, and the remaining ones were used for the generation of models by applying the SPSS/PC software package (27). For regression analysis, data set was randomly divided into two groups: training and prediction sets. For performing this work, we select randomly 20% of this data set and put it in prediction set, and all others were put in training set. The effect of the percent of a prediction set on the data set was considered in this work. The results of this study are shown in Figure 2. As can be seen in Figure 2, if the percent of the prediction set is very low, there will be an uncertainty in the squared correlation coefficient of prediction set in comparison to the main data set; also, if the percent of the prediction set is very high, the obtained model cannot be as complete and as oral so as to predict the retention time accurately. Really, there is an optimum range for the percent of prediction set. As can be seen in Figure 2, it can be said if the percent of prediction set from main data set is between 15–40%, the model constructed by the training set can predict the prediction set as well as training set. Of course because the nature of selection of prediction set from main data set is random, Figure 2 cannot exactly be repeated with the exact same range, but the behavior shown in Figure 2 is quite fixed.

Table I. Experimental, ANN and MLR Calculated Values of Retention Times Together with the Values of the Descriptors Appearing in the Model*

No	Component	Descriptor				Retention Time (min)		
Training set		**S0k**	**Tie**	**R5v+**	**Ats4m**	**RT_{ANN}**	**RT_{MLR}**	**RT_{EXP}**
1	Phenol	15.398	19.669	0.009	2.187	5.741	4.595	4.69
2	2-Methylphenol	19	26.828	0.015	4.196	6.137	6.14	5.68
3	4-Methylphenol	18	25.679	0.011	3.804	6.156	5.991	6.21
4	2,6-Dimethylphenol	17.759	34.933	0.01	7.044	6.787	6.407	7.08
5	2,4-Dimethylphenol	21.549	33.703	0.011	6.652	7.684	8.246	7.34
6	2,3-Dimethylphenol	21.549	33.85	0.011	7.419	7.72	8.319	7.96
7	2-Chlorophenol	22.459	44.696	0.015	5.778	7.968	8.373	7.34
8	3-Chlorophenol	22.459	32.782	0.018	9.116	7.593	8.187	7.86
9	3,4-Dimethylphenol	22.496	32.557	0.01	7.027	8.264	8.872	8.46
10	2-Chloro-5-methylphenol	24.946	52.412	0.024	8.539	9.482	9.084	9.12
11	2,6-Dichlorophenol	23.612	26.38	0.014	17.593	8.696	9.924	9.73
12	4-Chloro-2-methylphenol	24.946	39.905	0.022	10.173	9.251	9.271	9.73
13	3,5-Dichlorophenol	23.612	58.023	0.017	24.269	10.066	10.68	11.02
14	2,4-Dichlorophenol	27.765	84.116	0.019	17.352	11.873	12.391	11.02
15	2,4,6-Trichlorophenol	28.731	47.629	0.019	37.389	13.864	14.239	12.85
16	2,3-Dichlorophenol	27.765	148.271	0.017	12.708	14.06	13.104	12.01
17	3,4-Dichlorophenol	27.765	56.69	0.019	12.466	11.424	11.539	12.51
18	2,3,6-Trichlorophenol	33.347	65.985	0.045	24.522	14.613	12.939	13.93
19	2-Nitrophenol	29.996	40.753	0.015	12.646	13.205	12.935	12.51
20	2,3,5-Trichlorophenol	33.347	63.073	0.046	27.86	14.312	13.1	15.02
21	2,3,5,6-Tetrachlorophenol	32.243	284.615	0.049	39.674	17.257	16.502	17.71
22	2,3,4,6-Tetrachlorophenol	39.013	153.543	0.045	44.318	18.249	19.042	17.96
23	2,3,4-Trichlorophenol	33.347	84.79	0.019	24.281	16.153	15.994	16.81
24	4-Nitrophenol	27.329	121.262	0.01	9.229	14.096	12.913	15.69
25	Pentachlorophenol	37.021	473.865	0.047	59.47	22.936	23.824	22.96
26	2,5-Dinitrophenol	43.961	39.263	0.013	21.038	20.404	21.284	20.51
27	2,5-Dibromotoluene*	25.856	24.989	0.085	25.184	3.406	4.138	3.16
28	2,2',5,5'-Tetrabromobiphenyl*	48.435	62.564	0.053	80.908	24.776	25.244	25.16
29	2,4-Dibromophenol†	27.765	39.443	0.026	61.235	16.023	15.075	16.02
Prediction set								
30	3-Methylphenol‡	19	25.747	0.014	4.948	6.203	6.303	6.05
31	2,5-Dimethylphenol§	22.496	33.861	0.015	6.956	7.845	8.345	7.08
32	4-Chlorophenol‡	19.998	30.726	0.02	5.537	6.362	6.309	8.19
33	4-Chloro-3-methylphenol‡	24.946	39.086	0.017	9.021	9.375	9.691	10.18
34	2,5-Dichlorophenol§	27.765	125.794	0.052	12.708	10.76	9.003	10.71
35	2,4,5-Trichlorophenol‡	33.347	52.937	0.047	24.281	14.72	12.512	15.02
36	3-Nitrophenol§	29.996	28.721	0.013	10.58	13.302	12.784	13.69
37	2,3,4,5-Tetrachlorophenol§	39.013	312.665	0.049	39.433	21.021	20.46	20.51

* Internal Standard used in GC researches. † Surrogate used in GC researches.
‡ and † refer to test and prediction sets in ANN model, respectively.

The applied step-wise procedure for the selection of descriptors combines forward and backward procedures. When new variables enter the equation, complexity of intercorrelations will cause a change in the amount of the variance, which is explained by certain variables. When new variables enter, a variable sometimes loses some parts of its predictive validity, and in this case the stepwise method will automatically remove the weakened variable. The mathematical equation describing the model was constructed by using four criteria: multiple correlation coefficient (R), standard error (SE), F statistic, and the number of descriptors in the model. High R and F values, low standard error, the least number of descriptors, and high ability for prediction are the noticeable features of the best MLR model. The features of the best equation are illustrated in Table III.

The break-point MLR (BMLR) algorithm was used in order to avoid over-correlation of the regression equations (28). This is done through monitoring the increase of R^2 in the equations with a successive number of descriptors involved. This procedure, known as the break-point technique, shows the break-point (the change in the slope) in the plot of R^2 versus the number of descriptors added (Figure 3). The procedure was stopped when the difference between R^2 of the two consequent regression equations was less than or equal to 0.02. Figure 3 reveals the notion that increasing the number of parameters only up to four has a large influence on

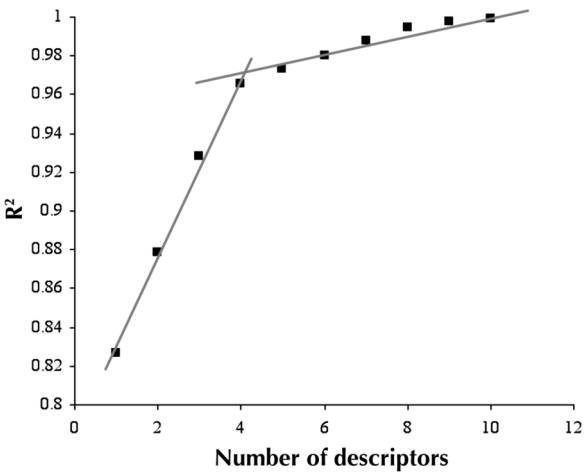

Figure 3. Influence of number of descriptors on R^2 of MLR model.

Table II. Groups of the 1661 Molecular Descriptors Generated by Software Dragon

Group name	Dimensionality	No. of descriptors	No. of descriptors in model
Constitutional descriptors	0	48	5
Topological descriptors	2	119	14
Molecular walk counts	2	47	1
Connectivity indices	2	33	6
Information indices	2	47	10
2D autocorrelation descriptors	2	96	31
Edge adjacency indices	2	107	0
BCUT descriptors	2	64	16
Galvez topological charge indices	2	21	5
Eigenvalue-based indices	2	44	2
Randic molecular profiles	3	41	2
Geometrical descriptors	3	74	8
RDF descriptors	3	150	20
3D-MoRSE descriptors	3	160	82
WHIM descriptors	3	99	28
GETAWAY descriptors	3	197	58
Functional groups	1	152	0
Atom-centered fragments	1	120	0
Charge descriptors	1	14	0
Molecular properties	1	28	1
Sum		1661	289

Table III. Selected Descriptors of Multiple Linear Regression

Descriptor	Type of descriptor	Notation	Coefficient	Mean effect
Kier symmetry index	Topological	S0k	0.528 (± 0.041)	14.592
E-state topological parameter	Topological	Tie	0.01448 (± 0.003)	1.125
Getaway (R maximal autocorrelation of lag 5 / weighted by atomic van der waals volumes)	Geometric	R5v+	–107.857 (± 14.570)	–2.656
2D autocorrelations (Broto-Moreau utocorrelation of a topological structure-lag4 / weighted by atomic masses	Topological	Ats4m	0.09311 (± 0.018)	2.004
Constant			–3.053(± 0.915)	

$R^2_{training}$ = 0.966, $R^2_{prediction}$ = 0.932, $SE_{training}$ = 0.989, $SE_{prediction}$ = 1.391, F = 171.605

improving correlation. Therefore, we have chosen four descriptors as the optimum number of parameters. The descriptors appearing in this model are Kier symmetry index (S0k), E-state topological parameter (Tie), getaway-weighted by atomic van der Waals volumes (R5v+), and 2D autocorrelations-weighted by atomic masses (Ats4m), whose definitions are given in Table III. As it can be seen from the correlation matrix (Table IV), there is no significant correlation between the selected descriptors.

Neural network generation

In order to generate and educate the neural network in this study, the Matlab 7.1 package was used (29). For ANN generation, data set was divided into three groups: training, test, and prediction sets. The training set, comprising 29 molecules, was used for the model generation. However, the test set, comprising four molecules, was used to maintain the overtraining. The prediction set, comprising four molecules, was used to evaluate the generated model. It is worth noting that the molecules in the test and prediction sets were just the same as those selected as prediction set in MLR model. The effect of the percent of prediction set and test set from the main data set on accuracy of results has been considered in Figure 4. Any simple neural network can fit any data set with any complexity. The power of prediction in application of neural networks for nonlinear pattern recognition is a very important factor. When the percent of the test set is increased as well as percent of prediction set, the quality of fitting of neural network, which has been constructed by means of training set, are slowly reduced. As shown in Figure 2 for MLR, there is an optimum percent range for selecting the prediction and test set in a neural network construction. This optimum range ideally is the range in which the accuracy of constructed network for training set, prediction set, and test set are equal. But in practice the word "equal" is replaced by "quite equal". As can be seen in Figure 4, the best range we can select is about 10% for test set and 10% for prediction set. If the percent of the prediction set and test set is very low, the constructed neural network cannot predict the test set well; and if the percent of the prediction set and test set is very high, the pattern recognition for neural network will be difficult. Thus, in these two forms, the obtained accuracy of the constructed neural network for these three sets by constructed network will not be "quite equal." Of course, in this work we used equal percentages for prediction set and test set. Results of Figure 4 can be repeated, but because of the random nature of selecting the prediction and test sets, the values will be changed partially, but the oral behavior will be constant.

Descriptors that appeared in the MLR model were used to generate the network as its inputs. A three-layer network with a tangent sigmoidal transfer function was designed. Before training, the input and the output values were normalized between –1 and +1. The appropriate number of nodes in the hidden layer was identified through training the network with a diverse number of nodes in the hidden layer. Learning rate and momentum values and type of transfer function were opti-

mized in a similar way. Architecture and specifications of ANN model are given in Table V.

Results and Discussion

Regression analysis

The relationship between the structure of phenol derivatives and their retention times in GC is the main objective of this work. For this purpose, linear and non-linear models were tested. The linear model (i.e., MLR) has been developed for two purposes. First, step-wise multiple linear regression procedure was used to select the suitable variables. It can be seen from Table III that four descriptors of Kier symmetry index (S0k), E-state topological parameter (Tie), getaway-weighted by atomic van der Waals volumes (R5v+), and 2D autocorrelations-weighted by atomic masses (Ats4m) were chosen out of 200 descriptors. These descriptors can be classified as topological (S0k, Tie, and Ats4m) and geometric (R5v+) descriptors. Because the DB-5 is a non-polar column, the lack of electronic descriptors in the model puts emphasis on the fact that polar interactions have no important effect on the retention behavior of the molecules. Thus, shape and symmetry of molecule, molecular mass or volume, atomic distances in the molecule, and substructure of molecule are effective parameters to mesh the molecules in the stationary phase and identify their retention times. This fact can be demonstrated through geometric and topological descriptors apparent in the MLR model. Also, in order to obtain the extent of each descriptor's contribution in the prediction of retention behavior of molecules, the mean effect of each parameter was calculated. The mean effect of each descriptor can be regarded as a measure of its part in retention behavior of molecules. These descriptors may have negative and/or positive roles. The mean effect for each descriptor is given in Table III, which illustrates that S0k is the most essential parameter influencing the retention behavior of the molecules. The comparison of compounds 7, 8, and 32 demonstrates the effect of a shift in the symmetric degree of molecules with similar substituents on the retention behavior of aforementioned molecules. This comparison reveals the fact that the higher symmetry a molecule has, the easier it can enter stationary phase holes and display a longer retention time. The S0k descriptor illustrates clearly the previously mentioned molecular behavior. The Tie descriptor is very similar to distance connectivity indices, and it is computed through electro-topological indices of adjacent atoms in H-depleted molecular graphs and generally indicates the complexity of molecular substructure. According to Table I, the more that substituents on phenyl ring takes place, the more complex the molecular substructure will be, and the chance of its interference with stationary phase and thus its meshing considerably increases. The Tie descriptor well-defines the mentioned process. Figure 5 properly elucidates the role of Tie descriptor. The difference between the retention times of three substituted compounds 20 and 23 may be due to their molecular volume, which in turn is controlled by atomic van der Waals volumes. This suggests that the more voluminous a compound becomes, the smaller longitudinal diffusion (B/U) will be due to existence of less diffusion coefficient DM in mobility phase. As a result, a short retention time is assessed. Furthermore, the more voluminous a molecule is, the more difficult it is to enter the stationary phase holes. This process has been illustrated through the negative role of R5v+ descriptor in the model. The comparison of compounds with similar symmetries 3, 24, and 32 proves the fact that the heavier a molecule is, the smaller diffusion coefficient DS in film it exhibits; and consequently, it holds a less bias for separation from stationary phase than its lighter molecules. This process is considered one of the most common ones in chromatography. Thus, the appearance of Ats4m descriptor in model will be easily justifiable. The detailed description of these descriptors is given in literature (30). The second purpose of developing this model was to assess the linear relationship between these descriptors and the retention times of phenol derivatives. In the MLR model, a R^2 value for prediction set 0.932 was obtained,

Table IV. Correlation Matrix for the Four Selected Descriptors

	S0k	Tie	R5v+	Ats4m
S0k	1			
Tie	0.467	1		
R5v+	0.54	0.429	1	
Ats4m	0.76	0.507	0.597	1

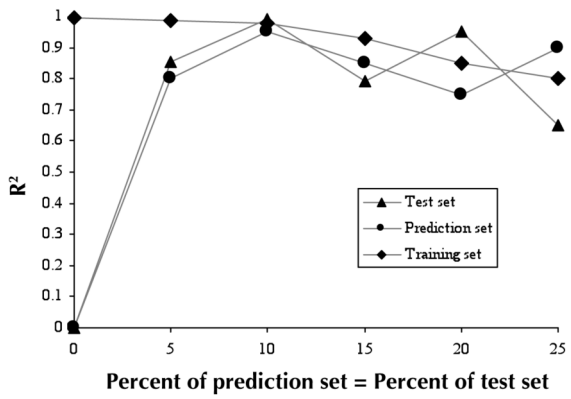

Figure 4. Effect of percent of testing set on the squared correlation coefficient of BPANN results. In this section the percent of validation set assumed equal to percent of testing set.

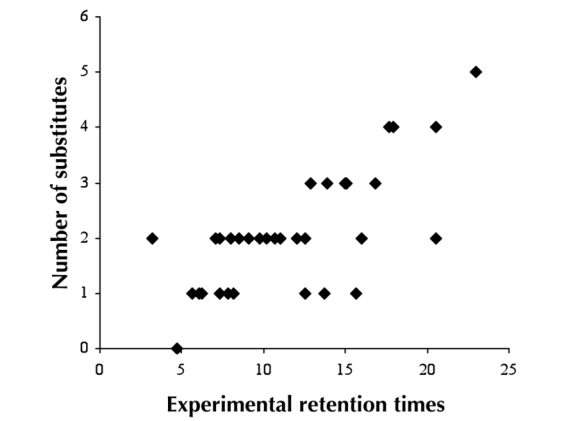

Figure 5. Experimental retention times versus number of substituents plot.

which demonstrated this model's ability to distinguish between the retention behaviors of molecules.

Neural network generation

As a second step, the non-linear characteristics of the descriptors were investigated. Therefore, a back propagation artificial neural network was developed using the descriptors appearing in the MLR model as its inputs. It is a common practice to optimize the parameters of number of nodes in the hidden layer, learning rate, and momentum in developing a reliable network. Different numbers of neurons in the hidden layer were tested at an arbitrary learning rate and momentum, epochs, and transfer function. The number of neurons in the hidden layer with the minimum value of SE was selected as the optimum number. Then, learning rate, momentum, epochs, and transfer function were optimized in a similar way. The experimental and calculated values of the phenol derivatives' retention times using MLR and ANN methods as well as the values of the descriptors appearing in the selected MLR model are given in Table I. The specifications together with R^2 and SE for the training, test, and prediction sets are given in Table VI. Comparison of the results in Table VI reveals superiority for ANN model over the MLR model. It can be seen from Table VI that the ANN model shows a SE of 1.013, which is much lower than those of the MLR model. Also, the R^2 value of 0.955 for the ANN should be compared with a value of 0.932 for the MLR model in the prediction set. For both models, several inconsistencies exist between the sequences of the real and predicted retention times. A summary of these inconsistencies is shown in Figure 6. In this figure, components with predicted values that have a diversion higher than 1 compared to their real ones in two different directions are displayed with heavier lines. It can be understood from the figure that diversions and their gradients are smaller in ANN model than MLR model; and this fact, in turn, demonstrates the prediction power of ANN model. Compounds 1, 11, 15, 16, 17, 24, and 32 show an incorrect order for the ANN model, while the order of retention times for compounds 7, 14, 15, 16, 20, 21, 22, 24, 31, 32, 34, and 35 is not correct for the MLR model. Large deviation of –2.77 for the MLR calculated value of the compound 24 should be compared with the large deviation of –2.05 for the ANN calculated value of the compound 16 counterpart.

Because of the small number of molecules included in the data set, the cross-validation method was used to evaluate the ability of the constructed ANN model. In this method, four species were removed randomly from the data set each time, and the model was generated with the remaining molecules (leave-4-out procedure) (19). Then the retention time of the removed molecules was predicted using the generated model. This procedure was continued until each analyte was predicted once. As a result, seven rounds of runs were needed for cross-validation of the ANN model. The values obtained using the cross-validation method for different groups of compounds are given in Table VII. As can be seen from this table, the results do not depend on the molecules in training and prediction set.

Figure 7 shows the calculated retention times versus experimental retention times for training, test, and prediction sets, and a value of 0.978 for R^2 was obtained from this plot. Figure 8 shows the plot of residuals compared to experimental values of retention times for the ANN model. The propagation of the residuals in both sides of zero indicates that no systematic error exists in the development of the ANNs.

Table V. Architecture and Specifications of the BPANN model

Number of nodes in the input layer	4
Number of nodes in the hidden layer	4
Number of nodes in the output layer	1
Number of iterations	36800
Learning rate	0.1
Momentum	0.85
Transfer function	Tangent sigmoidal

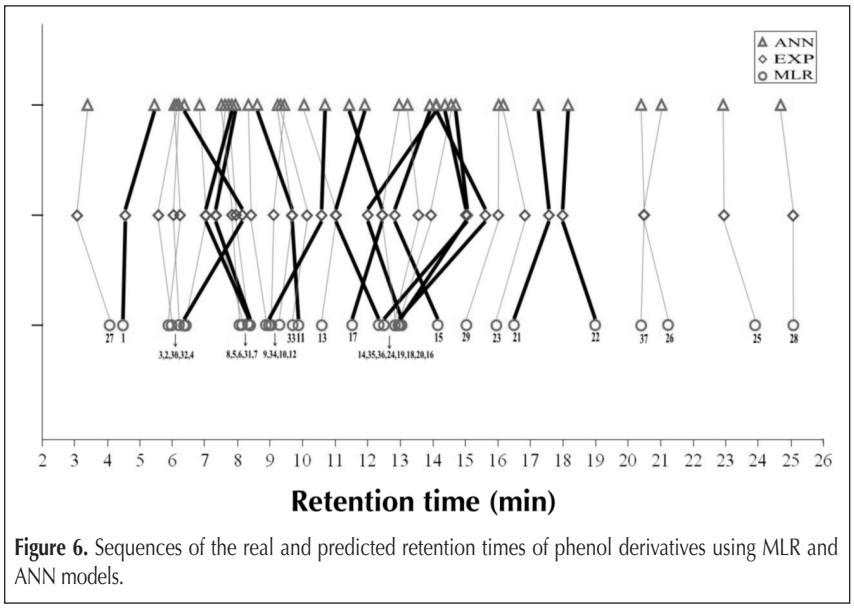

Figure 6. Sequences of the real and predicted retention times of phenol derivatives using MLR and ANN models.

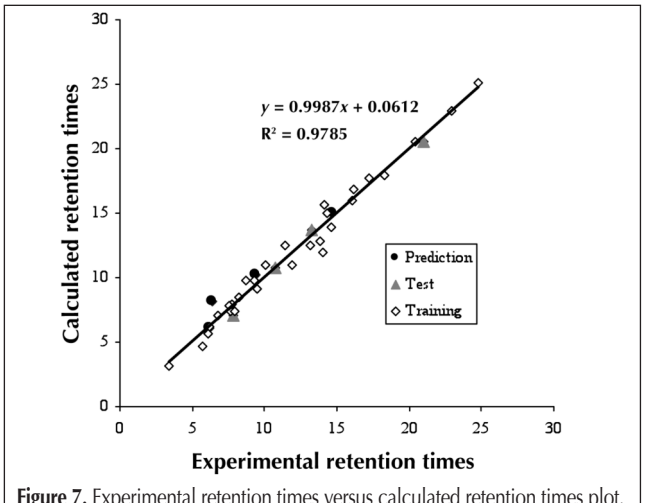

Figure 7. Experimental retention times versus calculated retention times plot.

Conclusions

Two common methods of MLR and ANN were used to predict the retention times of 37 derivatives of phenols. Both methods seem invaluable, but the comparison of these methods shows the superiority of ANN over that of the regression model. Also, the ANN is able to predict the trend of variation in the values of retention times for different derivatives, while the MLR has a lower predictive ability in this regard. The superiority of ANN over that of MLR reveals the fact that the retention times of substituted phenols manifest some nonlinear characteristics. On the other hand, because the descriptors appearing in the MLR model were used as inputs for the ANN, it can be included that the former method is a suitable technique for choosing the inputs for the neural networks. Moreover, the conformity between selected descriptors and stationary phase of column is a vigorous demonstration of results obtained.

Table VI. Statistical Results of ANN Model Compared to MLR Model

	$R^2_{training}$	$SE_{training}$	R^2_{test}	SE_{test}	$R^2_{prediction}$	$SE_{prediction}$
ANN model*	0.981	0.754	0.992	0.5	0.955	1.013
MLR model†	0.966	0.989	–	–	0.932	1.391

* Transfer function = Tangant sigmoidal, momentum = 0.85, learning rate = 0.1, epochs = 36,800, initial weights were between –2 and + 2.
† Feature selection was based on stepwise method.

Table VII. The Results of Cross-Validation Test

Test model	SE(T)	SE(P)	R^2_{cv}
I	0.682	1.025	0.949
II	0.589	0.997	0.968
III	0.823	1.045	0.935
IV	0.796	1.103	0.958
V	0.741	0.996	0.951
VI	0.696	0.985	0.962
VII	0.775	1.121	0.947

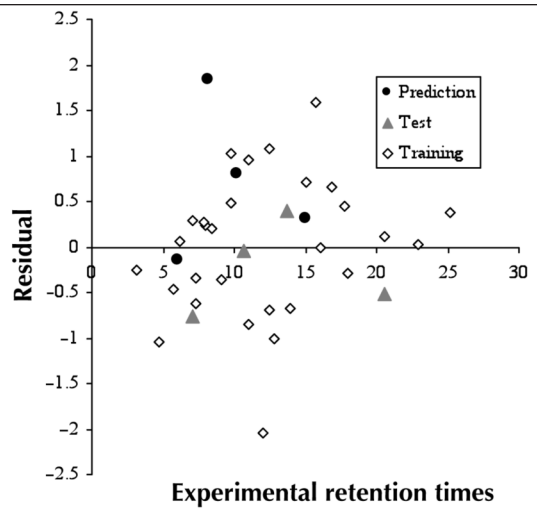

Figure 8. Experimental retention times versus residual plot.

References

1. K. Heberger. Quantitative structure–(chromatographic) retention relationships. *J. Chromatogr. A* **1158:** 273–305 (2007).
2. R. Put and Y. Vander Heyden. Review on modelling aspects in reversed-phase liquid chromatographic quantitative structure–retention relationships. *Anal. Chim. Acta* **602:** 164–172 (2007).
3. Y. Rena, H. Liua, X. Yaoa, and M. Liu. Three-dimensional topographic index applied to the prediction of acyclic C5–C8 alkenes Kováts retention indices on polydimethylsiloxane and squalane columns. *J. Chromatogr. A* **1155:** 105–111 (2007).
4. F. Liu, Y. Liang, C. Cao, and N. Zhou. QSPR study of GC retention indices for saturated esters on seven stationary phases based on novel topological indices. *Talanta* **72:** 1307–1315 (2007).
5. T.B. Czek and R. Kaliszan. Combination of linear solvent strength model and quantitative structure–retention relationships as a comprehensive procedure of approximate prediction of retention in gradient liquid chromatography. *J. Chromatogr. A* **962:** 41–55 (2002).
6. V. Andrisano, C. Bertucci, V. Cavrini, M. Recanatini, A. Cavalli, L. Varoli, G. Felix, and I.W. Wainer. Stereoselective binding of 2,3-substituted 3-hydroxypropionic acids on an immobilised human serum albumin chiral stationary phase: stereochemical characterisation and quantitative structure–retention relationship study. *J. Chromatogr. A* **876:** 75–86 (2000).
7. J. Olivero and K. Kannan. Quantitative structure–retention relationships of polychlorinated naphthalenes in gas chromatography. *J Chromatogr A* **849:** 621–627 (1999).
8. C. Lu, W. Guo, and C. Yin. Quantitative structure-retention relationship study of the gas chromatographic retention indices of saturated esters on different stationary phases using novel topological indices. *Anal. Chim. Acta* **561:** 96–102 (2006).
9. M. Jalali-Heravi and Z. Garkani-Nejad. Prediction of gas chromatographic retention indices of some benzene derivatives. *J. Chromatogr. A* **648:** 389–393 (1993).
10. J. Acevedo-Martínez, J.C. Escalona-Arranz, A. Villar-Rojas, F. Téllez-Palmero, R. Pérez-Rosés, L. González, and R. Carrasco-Velar. Quantitative study of the structure–retention index relationship in the imine family. *J. Chromatogr. A* **1102:** 238–244 (2006).
11. G. Carlucci, A.A. D'Archivio, M.A. Maggi, P. Mazzeo, and F. Ruggieri. Investigation of retention behaviour of non-steroidal anti-inflammatory drugs in high-performance liquid chromatography by using quantitative structure–retention relationships. *Anal. Chim. Acta* **601:** 68–76 (2007).
12. F. Ruggieri, A.A. D'Archivio, G. Carlucci, and P. Mazzeo. Application of artificial neural networks for prediction of retention factors of triazine herbicides in reversed-phase liquid chromatography. *J. Chromatogr. A* **1076:** 163–169 (2005).
13. B. Škrbic and A. Onjia. Prediction of the Lee retention indices of polycyclic aromatic hydrocarbons by artificial neural network. *J. Chromatogr. A* **1108:** 279–284 (2006).
14. M. Jalali-Heravi and M.H. Fatemi. Artificial neural network modeling of Kováts retention indices for noncyclic and monocyclic terpenes. *J. Chromatogr. A* **915:** 177–183 (2001).
15. W. Guo, Y. Lu, and X.M. Zheng. The predicting study for chromatographic retention index of saturated alcohols by MLR and ANN. *Talanta* **51:** 479–488 (2000).
16. D.W. Patterson. Artificial Neural Networks: Theory and Applications, Simon and Schuster, New York, 1996.
17. J. Zupan and J. Gasteiger. Neural networks: A new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta* **248:** 1–30 (1991).
18. N.K. Bose and P. Liang. Neural Network-Fundamentals, McGraw-Hill, New York, 1996.
19. D.W. Osten. Selection of optimal regression models via cross-validation. *J. Chemom.* **2:** 39–48 (1998).
20. http://www.epa.gov/sw-846/pdfs/8041a.pdf 21. Hyperchem, Molecular Modeling System, Hyper Cube, Inc. and Auto Desk, Inc, 1993, Developed by Hyper Cube, Inc.
22. R. Todeschini, V. Consonni, A. Mauri, and M. Pavan. Software Dragon: Calculation of Molecular Descriptors, Department of Environmental Sciences, University of Milano-Bicocca, and Talete, srl., Milan, Italy, (2003).
23. Z. Garkani-Nejad, M. Karlovits, W. Demuth, T. Stimpfl, W. Vycudilik, M. Jalali-Heravi, and K. Varmuza. Prediction of gas chromatographic retention indices of a diverse set of toxicologically relevant compounds. *J. Chromatogr. A* **1028:** 287–295 (2004).
24. F. Liu, Y. Liang, and C. Cao. QSPR modeling of thermal conductivity detection response factors for diverse organic compound. *Chemom. Intell. Lab. Syst.* **81:** 120–126 (2006).
25. A. Khalafi-Nezhad, M.N. Soltani Rad, H. Mohabatkar, Z. Asrari, and B. Hemmateenejad. Design, synthesis, antibacterial and QSAR studies of benzimidazole and imidazole chloroaryloxyalkyl derivatives. *Bioorg. Med. Chem.* **13:** 1931–1938 (2005).
26. J. Shen, Y. Du, Y. Zhao, G. Liu, and Y. Tang. In Silico Prediction of Blood-Brain Partitioning Using a Chemometric Method Called Genetic Algorithm Based Variable Selection. *QSAR Comb. Sci.* **27:** 704–717 (2008).
27. SPSS/PC, Statistical Package for IBMPC, Quiad software, Ontario, 1986
28. A.R. Katritzky, E.S. Ignatchenko, R.A. Barcock, V.S. Lobanov, M. Karelson. Prediction of Gas Chromatogrphic Retention Times and Response Factors Using a General Quantitative Structure-Property Relationship Treatment. *Anal. Chem.* **66:** 1799 (1994).
29. MATLAB for Windows, The Language of Technical Computing, Ver.7.1.0.450 Release 12.1, The MathWorks, Inc., 2001.
30. R. Todeschini and V. Consonni. Handbook of Molecular Descriptors, Wiley/VCH, Weinheim, Germany, 2000.

Manuscript received September 28, 2008;
revision received December 20, 2008.