ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Quantitative Structure/Retention Relationship Study of Benzene Derivatives.

Karima DJELLOUL MOKRANI[1,2]* , Hamza HADDAG[1], and Djelloul MESSADI[1].

[1]Environmental and Food Safety Laboratory, Badji Mokhtar - Annaba University, Algeria
[2]Functional and Evolutionary Ecology Laboratory, University Chadli Bendjedid El Tarf. Algeria

ABSTRACT

Retention indices of 38 benzene derivatives, separated by gas chromatography were correlated with 2 connectivity indices, using PM3 semi-empirical calculation method and hybrid genetic algorithms/ multiple linear regression approach. For the sake of external validation, the available set of chemicals was separated using Kennard and Stone algorithm into training set of 28 compounds and an external set of 10 compounds. The proposed hybrid model was validated using different criterions, and its predictive capability meets the conditions defined by Golbraikh et al. In comparison to the previously published model, our model exhibits a large enhancement and its mechanistic interpretation was attempted to connect the selected variables to the retention phenomenon.

Keywords: Benzene derivatives- Kováts index- QSRR- Internal and external predictivity validation-Chemical applicability domain.

**Corresponding author*

INTRODUCTION

The mono-aromatic hydrocarbons, which are often present in the urban environments, constitute an important source of pollution and health hazard[1].

Hence, the development of reliable structural identification and quantification of these substances is imperative. Gas chromatography coupled with the mass spectrometry or infrared analysis by Fourier transform is largely used to this aim. Nevertheless, the measurement of their retention indices constitutes, even nowadays, a simple means of effective, sensitive and affordable identification.

Advantageously, any parameter of retention can be derived a priori from the molecular structure of the considered compound. The prediction of the retention indices of a set of 38 benzene derivatives separated by isothermal gas chromatography was obtained by Jalali-Heravi and Garkani-Nejad [2] who adopted a QSRR approach [3](Quantitative Structure-Retention/Relationship). Based on a training set of 32 randomly selected compounds, linear models were developed following a stepwise involving successive additions of 58 variables (topological, geometrical and electronic) along with previously characterized physical properties. Although widely used, the disadvantage of this approach is that it cannot account for combined effects since each variable is considered separately.

Genetic algorithms [4,5], based on the stochastic search, constitute an alternative method of choice for the selection of variables subsets (VSS: Variable Subset Selection).

The optimization of the molecules geometry, necessary to the calculation of certain descriptors was conducted by applying the MNDO (Modified Neglect of Diatomic Overlap) semi-empirical method [6] while MNDO is known to be ineffective when calculating the molecular structures and the heats of formation of the molecules containing fluorine [7].

The model includes four descriptors (XV0: valence connectivity index of order zero; NOCH$_3$: number of methyl groups in the molecule; VOL: Van der Waals volume of the molecule; DIMO: Dipole moment of the molecule) [2] is validated using the following set of parameters: the coefficient of determination R^2, Fisher parameter F and the standard deviation S,while the application domain of this model is not defined.

In addition, for models including more than two descriptors, low coefficients of correlation cannot positively ensure the complete independence of the descriptors. This aspect was not assessed by Jalali-Heravi and Garkani-Nejad [2]. Finally, the predictive capability of the proposed model was tested by calculating the retention indices of the six compounds not retained for its construction.

In this work, we proposed a statistical linear model using the same database by calculating the molecular descriptors with the software Dragon [8]. This statistical linear model is justified using different criteria and its prediction capability is assessed following Golbraikh et al.'s conditions [9,10].

Finally, the applicability domain (AD)is discussed using the Williams plot [11,12] that represents the standardized residual of predictions versus the leverage values (h_{ii}).

The semi-empirical method PM3 (Parametric Method 3) [13] was useful for optimization of the geometry of the molecules. It consists in re-parameterization of AM1 method (Austin Model1) [14] that is itself an improved version of the MNDO method.

It is important to define rationally the training set during the construction of the model and, for its assessment an external test set comprising 15 to 40% of the available data. The available set of chemicals was preliminary separated using Kennard and Stone algorithm [15].The hybrid approach Genetic Algorithm Multiple Linear Regression(GA-MLR) was adopted in our work.

MATERIALS AND METHODS

Database:

The Kováts indices of the 38 benzene derivatives (table1) were extracted from reference [2]that also provides a detailed description of the conditions of the chromatographic separation.

The extreme values are 664.1 and 1287.7 index units (iu) with an average of 965.2 iu.

Table1: Retention indices (experimental and calculated) and values of the used descriptors

N°	Compounds	CAS Number	$RI_{Experimental}$	$RI_{Calculated}$	$X0Av$	$X1sol2$
01	Benzene	71-43-2	681.3	689.94	0.577	9.0000
02	Fluorobenzene	462-06-6	664.1	678.31	0.538	9.0000
03	Chlorobenzene	108-90-7	877.9	863.89	0.646	13.5645
04	Bromobenzene	108-86-1	979.6	973.14	0.764	15.7688
05	Toluene	108-88-3	788.2	789.50	0.627	11.5192
06	Anisole	100-66-3	923.6	913.61	0.599	15.4606
07	p-Chloroanisole	623-12-1	1131.7	1124.67	0.650	21.2890
08	p-Xylene	106-42-3	889.2	895.62	0.664	14.3489
09	p-Fluorotoluene	352-32-9	777.7	777.28	0.586	11.5192
10	p-Bromotoluene	106-38-7	1096.3	1089.48	0.784	19.0532
11	p-Bromofluorobenzene	460-00-4	940.9	955.86	0.706	15.7688
12	p-Chlorobromobenzene	106-39-8	1174.4	1182.14	0.801	21.6597
13	m-Chloroanisole	2845-89-8	1126.0	1124.67	0.650	21.2890
14	m-Methylanisole	100-84-5	1029.6	1033.68	0.635	18.7143
15	m-Xylene	108-38-3	892.0	895.62	0.664	14.3489
16	m-	108-37-2	1179.0	1182.14	0.801	21.6597
17	m-Bromotoluene	591-17-3	1100.0	1089.48	0.784	19.0532
18	m-Fluorotoluene	352-70-5	778.0	777.28	0.586	11.5192
19	m-Dibromobenzene	108-36-1	1287.7	1306.01	0.905	24.4234
20	o-Methylanisole	578-58-5	1013.5	1038.63	0.635	18.8616
21	o-Chloroanisole	766-51-8	1135.6	1129.96	0.650	21.4462
22	o-Bromofluorobenzene	1072-85-1	959.6	955.86	0.706	15.7688
23	o-Xylene	95-47-6	916.2	899.96	0.664	14.4780
24	o-Bromochlorobenzene	694-80-4	1197.6	1187.46	0.801	21.8182
25	p-Methylanisole	104-93-8	1029.5	1033.68	0.635	18.7143
26	o-Bromotoluene	95-46-5	1095.7	1094.47	0.784	19.2019
27	m-Fluoroanisole	456-49-5	908.5	903.77	0.566	15.4606
28	p-Chlorotoluene	106-43-4	989.2	976.50	0.680	16.6138
29	p-Fluoroanisole	459-60-9	910.6	903.77	0.566	15.4606
30	m-Chlorotoluene	108-41-8	990.9	976.50	0.680	16.6138
31	m-Chlorofluorobenzene	625-98-9	835.4	851.08	0.603	13.5645
32	m-Dichlorobenzene	541-73-1	1060.5	1063.55	0.697	19.0532
33	o-Fluorotoluene	95-52-3	777.4	777.28	0.586	11.5192
34	o-Fluoroanisole	321-28-8	919.7	903.77	0.566	15.4606
35	o-Chlorofluorobenzene	348-51-6	862.0	851.08	0.603	13.5645
36	p-Chlorofluorobenzene	352-33-0	840.5	851.08	0.603	13.5645
37	o-Chlorotoluene	95-49-8	986.3	981.17	0.680	16.7526
38	m-Bromofluorobenzene	1073-06-9	932.8	955.86	0.706	15.7688

Descriptors calculation:

We have used the Hyperchem[16] to represent each molecule, whose geometry is initially pre-optimized by molecular mechanics calculation. Then for each molecule we have determined its (x,y,z) atomic coordinates corresponding to the conformation of lowest energy determined by the PM3 method. All calculations were carried in the frame of the Hartree Fock formalism with spin constraint (or RHF: for Restricted Hartree-Fock) without confirmation interaction.

The molecular structures were optimized; according to the Polack-Ribiere algorithm adopting a stopping criterion corresponding to a mean square root of the gradient of 0.001 kcal/mol. Following this optimization, the molecule geometries were transferred to Dragon software[8] for the calculation of 1664 descriptors belongingto20 different classes. The descriptors of the same group exhibiting constant values(standard deviation lower than 0.0001)provide no information and thus, are removed from subsequent analysis. Similarly, two highly correlated descriptors $r \geq 0,92$ conveying redundant information automatically exclude one that is correlated with the greatest number of descriptors. Consequently, the initial pool of 1664 descriptors was reduced to 203 elements.

Kennard and Stone algorithm [15]:

It is a sequential technique that maximizes the Euclidean distances between new selected samples and previous analyzed samples. It starts by locating the two most distant samples, which are removed from the original data set and assigned to the training set.

For each sample (sample i) not selected previously, the algorithm calculates its distance to each sample; and assigns to (sample i) the smallest distances.

The sample (sample i) associated with the greatest distance is the furthest of all the samples already selected. The procedure is repeated until the target number of training samples is reached.

This technique has two significant advantages. Selecting the most distant samples from each other introduces diversity across the training set. Obtaining a uniform distribution is another advantage of this technique.

As a result, using the algorithm of Kennard and Stone, the complete data set was divided into two subsets: the training subset containing 28 compounds and validation subset including the 10 remaining compounds.

Model validation development:

The variable subsets selection (VSS) is realized using the genetic algorithm (GA-VSS) by maximizing the prediction coefficient Q^2_{LOO}.

Genetic algorithms are optimization algorithms based on technique derived from genetic and natural evolution mechanisms: i.e, crossing (or crossover) and mutation that are responsible for the generation of new individuals.

In the MobyDigs software [17] such processes are controlled by a user-defined parameter T varying between zero and one, defining the relative extent of crossing and mutation.

In the terminology of genetic algorithms, the binary vector I, called chromosome, is a vector of dimension p where each position (a gene) corresponds to a variable (1: if it appears in the model; zero 0: otherwise. Each chromosome is a model with a subset of variables [4,5].

The genetic algorithm parameters have been defined as follow:

- Model population: Pop = 100

- Maximum number of variables in the model: L = 5, so as to associate a minimum of five compounds to each descriptor; the minimum number is arbitrarily 1
- T value: chosen equal to 0.5 to balance the effects of crossover and mutation.

To avoid models with co-linearity lacking high prediction capabilities, we have applied the QUIK (Q Under Influence of K) rule [18] based on multivariable correlation index [19] defined as follow:

$$K = \frac{\sum_J \left| \frac{\lambda_j}{\sum_j \lambda_j} - \frac{1}{p} \right|}{2(p-1)\big/ p}; j = 1, \ldots, p \quad \text{et} \ 0 \le k \le 1 \qquad (1)$$

λ_j :are the eigenvalues of the correlation matrix of the data set $(n \times p)$.

n : The number of objects and p the number of variables.

This rule derives from the assumption that the total correlation in the set formed by predictors **X** of the model, and the response **Y (** K_{xy}) must always be greater than the correlation measured with the set of predictors (K_{xx}) taken separately.

To calculated K_{xy} we have considered the response Y as an x variable and determined the corresponding correlation matrix. Generally [20], models that do not verify the relationship are rejected:

$$D(K) = K_{xy} - K_{xx} > 0{,}05 \qquad (2)$$

A model with a minimum number of explanatory variables is sought (rule of parsimony). Its variables can be related to the retention phenomena in the apolar stationary phase Apiezon MH used in the analyses and could be easly interpreted.

The model will be justified by means of different statistical parameters (R^2 , R_{aj}^2 , Fisher parameter F, standard error S) and by considering the Leave Many Out (LMO) cross- validation, the randomization test of Y, as well as the bootstrap technique.

The adjusted R^2 (R_{aj}^2) calculated using the formula:

$$R_{aj}^2 = 1 - \left[\frac{n-1}{n-p-1} \left(1 - R^2\right) \right] \qquad (3)$$

Is a better measure of the percentage of the total variation explained by the model than the coefficient of determination R^2

The Fisher parameter F is defined by the ratio of the average of the squares due to regression to the mean of the squares of the residuals, which to compare the variance explained by the model to the residual variance: a high value of R^2 is proof of the reliability of this model.

The cross validation consists in re-computing the model considering only (n-q) objects and using this new model to predict the dependent variable value of the q excluded compounds. The process is repeated for the n objects of the training set.

If $q = 1$ the technique is called LOO (Leave One Out), otherwise it is LMO (Leave Many Out). A prediction coefficient LOO or LMO designated by Q^2 or R_{CV}^2 respectively, is calculated considering the dispersion of the estimation [21]:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i/i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \qquad (4)$$

$\hat{y}_{i/i}$: corresponds to the response of the ith object using a model obtained without involving this object; y_i, $\bar{y}$: represent respectively, the value of the ith observation and the average value of the n observations; the summation covers all of the compounds in training.

In order to establish a nonrandom model, we have applied the randomization test of **Y**(Y-scrambling) [21]. The test consists in generating a vector of the studied propriety by a random permutation of the components of the real vector. Then, we calculate the result QSRR model vector according to the usual method. This process is repeated 100 times in this study. If a high score is reached, the original model is not acceptable.

In the Bootstrop validation technique, we simulate new samples of size (n), by random pooling with reduction. As such, the training set that maintains its initial size (n), is composed of generally, repeated objects, since the set of evaluation includes the removed objects [22,23].

The model is calculated both on the training set and on the predicted responses set combined. This construction procedure of the training and evaluation sets is repeated 3000 times in this study, and an average prediction capacity is calculated Q^2_{BOOT} [23].

The validation of the model has been completed using a test set. Equation (5) details the calculation of Q^2_{EXT} for the test set.

$$Q^2_{EXT} = 1 - \frac{\sum_{i=1}^{n_{EXT}}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n_{EXT}}\left(y_i - \bar{y}_{cal}\right)} \qquad (5)$$

y_i, $\hat{y}_i$: are respectively the observed and the predicted values, and $\bar{y}_{cal}$ is the average of the observed values of the training set. The sum considers all the samples of the test set.

According to Golbraikh et al, [9,24] a QSRR model can provide an acceptable prediction if it verifies the following conditions:

$$Q^2_{EXT} > 0{,}5 \quad (6-a) \quad ; \quad r^2 > 0{,}6 \quad (6-b)$$

$$\left(r^2 - r_0^2\right)/r^2 < 0{,}1 \quad \text{or} \quad \left(r^2 - r_0'^2\right)/\mathrm{r}^2 < 0{,}1 \quad (6-c)$$

$$0{,}85 \le k \le 1{,}15 \quad \text{or} \quad 0{,}85 \le k' \le 1{,}15 \quad (6-d)$$

r is the correlation coefficient between the calculated and experimental values in the test set ;
r_0^2 (Calculated versus observed values) and $r_0'^2$ (observed versus calculated values) are the coefficients of determination; k , k' are slopes of the regression lines through the origin of calculated versus observed and observed versus respectively.

Applicability domain:

The applicability domain (AD) is a theoretical region of space defined by the descriptors of the model and the modeled response, for which a given QSRR model is expected to lead to reliable predictions. This region, which depends on the nature of the compounds of the training set, can be characterized in different ways. In this work the structure of AD has been determined by the leverages approach, defined by the diagonal elements of the **H** matrix that allows, by simple multiplication to associate the vector $\mathbf{y}$ to the vector $\hat{\mathbf{y}}$ .The diagonal h_{ii} element is defined by:

$$h_{ii} = x_i \left(X^T X \right)^{-1} x_i^T \qquad (7)$$

x_i is the line-vector of the compound descriptors i, and **X** the matrix of the model deduced from the descriptors values of all the training set; the exponent T denoting the transposition vector (or matrix).

The h_{ii} element determines the influence of observation i on estimators obtained by the least squares method. A leverage point is an observation that significantly influences the estimators. In practice, an observation i is considered as a point of leverage if:

$$h_{ii} > h^* = 3\left(\sum_i h_{ii} \right) \Big/ n = 3\left(p+1 \right)/n \qquad (8)$$

The Williams plot displaying the standardized residual of predictions against the leverage values h_{ii} was used with the aim of detecting both **X** outliers (leverage points) and **Y** outliers in (standard residuals higher in absolute values than 3 standard deviation units: $3s$).

RESULTS AND DISCUSSION

Model development and validation:

Table 2 shows that the retention index is linearly correlated to the descriptor $X1sol$, or better to its square $X1sol2$

Table 2: Comparison of the statistical parameters of different models.

n^a	Descriptors	R^2	R^2_{aj}	Q^2_{LOO}	$Q^2_{L(5)O}$	Q^2_{BOOT}	F	S	DK
28	$X1sol$	98.00	97.7	97.64	97.86	97.64	1247.37	23.16	-
28	$X1sol2$	98.19	98.12	97.95	98.14	97.76	1409.32	21.82	-
28	$X0Av \,.\, X1sol2$	99.59	99.56	99.47	99.52	99.4	3068.88	10.53	0.123
32^b	$NOCH_3 \,.\, XV0 \,.\, VOL \,.$ $DIMO$	99.63	99.57	99.46	99.47	99.36	1814.17	10.12	0.126

[a] Training set compounds, [b] Jalali-Heravi et al. model.

We have adopted the model with 2 descriptors, $X0AV$, $X1sol2$. The corresponding equation, calculated using the centered reduced values is given by:

$$IR = 0.168\, X0Av + 0.872\, X1sol2 \qquad (9)$$

Where $X0Av$ denotes the valence connectivity index of zero order, and $X1sol$ the Solvation connectivity index of the first order [25-26].

The combination of these two descriptors provides an improvement of all statistical parameters as detailed and compared in the table 2. In particular, the standard error is divided by a factor greater than two

(21.82 to 10.58) and close to the value (10.12) obtained with the four descriptor model, published by Jalali-Heravi et al. Also note DK (= 0.123) is higher than the prescribed limit of 0.05.

The obtained statistical parameters provide substantial ground that the proposed model (equation 9) establishes a strong correlation between the 2 selected variables and the studied property, characterized by an excellent coefficient of determination R^2 =99.59%that explains about 99.60% of data variation. In addition, the very high value of the Fisher parameter F (= 3068.88), indicates the excellent capability of the model in the prediction of RI values, with an acceptable standard error (s = 10.53). Equation (9) presents a R_{aj}^2 = 99.56 indicating excellent agreement between correlation and variation of the data.

The minor difference between Q_{LOO}^2 and $Q_{L(5)O}^2$ informs about the robustness of the model. The cross-validation prediction coefficient provides indication of the reliability of the model when addressing the sensitivity against the elimination of any chosen five data. The value of $Q_{Boot}^2(=99.4)$ confirms both the internal predictability and stability of the proposed model. Figure 1 plots the graph of statistical coefficients Q^2 and R^2 which allows comparing the results for randomized models (circles) to the initial model (rhombus). It appears clearly that retention indices statistics obtained for the modified vectors are lower than those of the real QSRR model. This observation ensures that a real structure/retention relationship has been established.

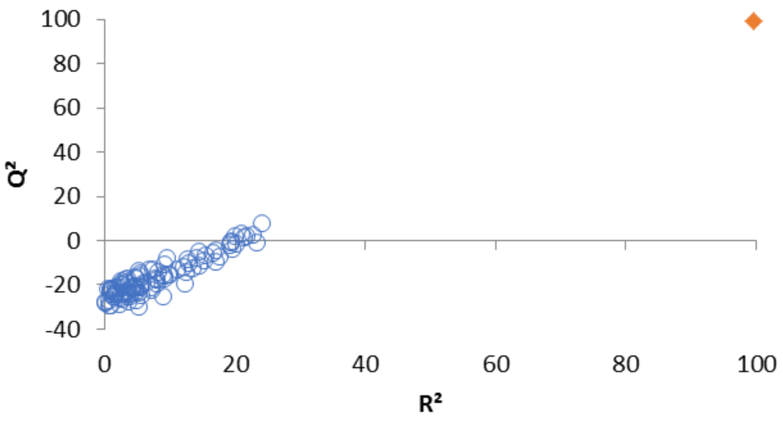

Figure 1: Graphical representation of the randomization test.

The following statistical parameters obtained for the external tests set verify the well-accepted conditions (6-a to 6-d), which reinforces the predictive capabilities of the present model.

$$Q_{EXT}^2 = 0.9869 > 0.5 \; r^2 = 0.9765 > 0.6$$

$$\left(r^2 - r_0^2\right)/r^2 = \left(0.9765 - 1.000\right)/0.9765 = -0.0240 < 0.1$$

$$\text{or} \;\; \left(r^2 - r_0^{'2}\right)/r^2 = \left(0.9765 - 1.000\right)/0.9765 = -0.0240 < 0.1$$

$$0.85 < k = 1.0002 < 1.15 \quad \text{or} \quad 0.85 < k^{'} = 0.9996 < 1.15$$

Application domain:

Figure 2 compares Williams plots derived either from our2-descriptors model, and the 4- descriptors model [2].In both cases, the leverage values of all training and test compounds, are lower than the corresponding critical values $h*$ (respectively 0.321 and 0.468) and, in both cases, none of the compounds is found influential.

Furthermore, for 2-descriptors model (figure 2-A) all training and test compounds exhibit standard residuals values lower, in absolute value to 3 units of standard deviation (3s), which confirms that the relevance of data set and the removal of insignificant data.
However, two outlier data points are found with the 4-descriptors model (figure 2-B), one of the training set (compound 30: o-xylene) and the other from the test set (compound 38: m-Bromofluorobenzene).

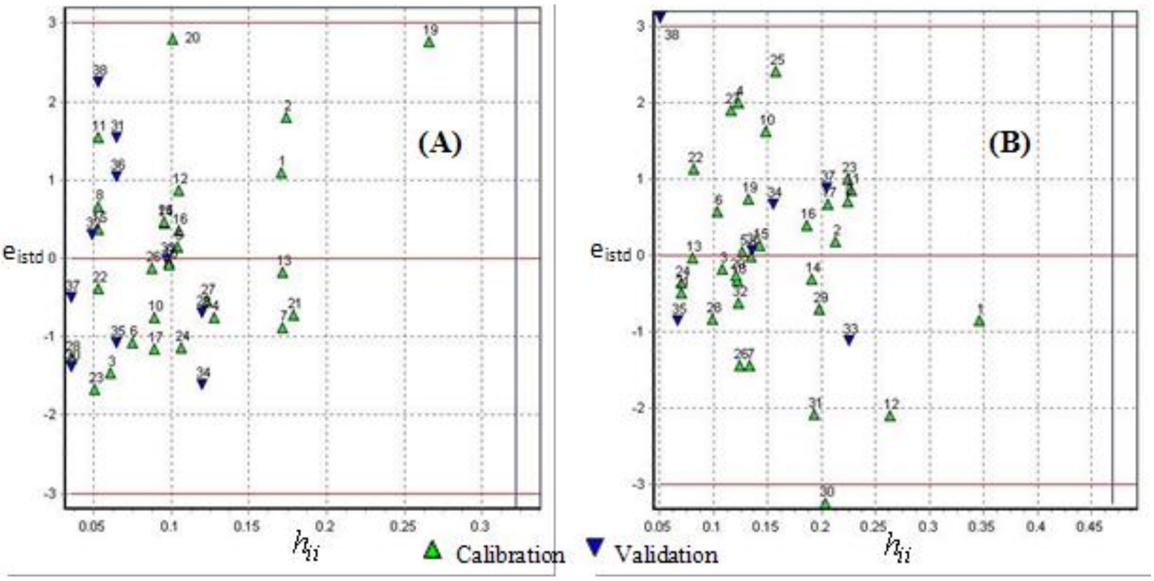

Figure 2: Williams plot for the 2-descriptors model (A) and the 4-descriptors model (B).

Interpretation of the model:

The descriptor X1sol2 which is highly correlated with RI significantly governs the model response as shown by the values of the coefficients of the 2- descriptors model, and the associated Student t values equal to 48.33 for X1sol2 and 9.31 for X0Av respectively.

Solvation connectivity indices are defined [27] for a H-depleted molecular graph, where fluorine atoms as well as hydrogen's are not taken into account, their dimension being very close to that of the hydrogen atom.

The Solvation connectivity index of the first order is derived from the equation:

$$X\ 1sol = \frac{1}{4}\sum_{b=1}^{B}\frac{\left(L_i . L_j\right)_b}{\left(\delta_i \delta_j\right)_b^{0,5}} \qquad (10)$$

Where b runs to the number of bonds B, L_i and L_j are the principal quantum numbers of 2 vertices (atoms) incidents to the considered bond; δ_i and δ_j represent the degrees (valences) of the corresponding vertices. The solvation connectivity indices make it possible to model solvation entropy and describe the interactions of dispersion in solution which play a decisive role in the retention phenomenon.

The average valence connectivity index of order zero ($X\,0Av$) is obtained by dividing the valence connectivity index of order zero ($X\,0V$) by the number of edges B (bonds) of the H-depleted molecular graph. $X\,0V$ is defined [28,29] by:

$$X\ 0Av = \sum_{i=1}^{N}\left(\delta_i^v\right)^{-0,5} \qquad (11)$$

N is the number of graph vertices, the number of atoms in the molecule other than hydrogen.

δ_i^v is calculated for the atom i, from the expression:

$$\delta_i^V = \frac{Z_i^V - H_i}{Z_i - Z_i^V - 1} \qquad (12)$$

Z_i, Z_i^δ respresent, respectively, the atomic number and the number valence electron of atom i.

Whereas H_i designates the number of hydrogen atoms bounded to the considered atom.

Besides the fact that it introduces relative corrections to the differences between halogen types contained in a given molecule, the descriptor $X\ 0Av$ is related to the size and degree of ramification of the molecules that may have a significant role in the distribution process of the solute between the two chromatographic phases (mobile/stationary)

CONCLUSION

A bi-parametric model was developed for the retention of 38 benzene derivatives separated by gas chromatography on apolar Apiezon MH column.

Solvation connectivity index describes the interactions of dispersion in solution. Whose role is crucial in the phenomenon of retention, while the average valence connectivity index of order zero plays a significant role in the distribution process of the solute between the two chromatographic phases (mobile/stationary).

The selection of these explanatory variables was carried out by genetic algorithm based software MobyDigs among 203 descriptors calculated using the Dragon software.

This optimal model was validated by different statistical approaches using the training set and the external validation set, defined rationally by adopting the Kennard and Stone technique.

The obtained statistical parameters show that the model with two descriptors established a strong correlation between the two selected variables and the studied property, which indicates the excellence of the model in the prediction of retention indices of benzene derivatives.

REFERENCES

[1] Bertinetto C, Duce C, Solaro R, Tiné MR.MATH Commun Math Comput Chem 2013; 70:1005-1021

[2] Jalali-Heravi M, Garkani-Nejad Z.J Chromatogr1993; 648: 389-393.

[3] Kaliszan R.Chem. Rev2007;107: 3212-3246.

[4] Leardi R, Boggia R, Terrile M. JChemom1992; 6: 267-281.

[5] Clark DE .Evolutionary Algorithms in Molecular Design. Wiley-VCH, Weinheim, 2000, 288p.

[6] Dewar M J S, Thiel W. J Am Chem Soc1977; 99: 4899-4907.

[7] Dewar M J S, Rzepa HS.J Am Chem Soc1978;100: 778-784.

[8] R. Todeschini, V. Consonni, M. Pavan ; 2005. DRAGON software for the calculation of molecular descriptors. Release 5.3 for Windows, Milano.

[9] Golbraikh A, Tropsha A.J Comput Aided Mol Des2002; 16: 357-369.

[10] Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A.J Comput Aided Mol Des 2003; 17: 241-253.

[11] Eriksson L, Jaworska J, Worth AP, Cronin MTD, Mc Dowell RM, Gramatica P.Environ HealthPerspect2003; 111: 1361-1375.

[12] Tropsha A, Gramatica P, Grombar VR. QSAR Comb Sci2003, 22: 69-76.

[13] Stewart JJP. JComput Chem1989, 10: 109-221.

[14] Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP.J Am Chem Soc1985; 107: 3902-3909.

[15] Kennard RW, Stone LA.Technometrics1969;11: 137-148.

[16] Hyperchem TM, 2000. Release 6.03 for windows, molecular modeling System.

[17] Todeschini R., Ballabio D., Consonni V., Mauri A., Pavan M., 2009. MOBYDIGS software for multilinear regression analysis and variable subset selection by genetic algorithm. Release for windows, Milano.

[18] Todeschini R, Consonni V, Maiocchi A.Chemom Intell Lab Syst 1998;46: 13-29.

[19] Todeschini R. Anal Chim Acta 1997;348: 419-430.

[20] Todeschini R, Consonni V, Mauri A, Pavan M.Anal Chim Acta 2004; 515: 199-208.

[21] Wold S, Eriksson L. Statistical validation of QSAR results. In: H. Van de Waterbeemd ed. Chemometrics methods in molecular design. VCH, New York, 1995, Vol. 2, pp. 309-318.

[22] Efron B, Tibshirani RJ.An introduction to the bootstrap, Chapman and Hall,1993, 456p.

[23] Wehrens R, Putter H, Buydens LMC.Chemom Intell Lab Syst2000;54: 35-52.

[24] Golbraikh A, Tropsha A. J Mol Graph Model 2002; 20: 269-276

[25] Todeschini R, Consonni V. Handbook of molecular descriptors. Edited by Mannhold R, Kubinyi H, Timmerman H, Wiley-VCH Verlag GmbH, Weinheim, 2000, 688p.

[26] Todeschini R, Consonni V.Molecular descriptors for chemoinformatics. Second, Revised and Enlarged Edition.Vol. I: Alphabetical listing, Series Editors: Mannhold R, Kubinyi H, Folkers G, Wiley-VCH Verlag GmbH CO. KGaA, 2009, 967 p.

[27] Zefirov NS, Polyulin VA.J Chem Inf Comput Sci2001; 41: 1022-1027.

[28] Kier LB, Hall LH.J Pharm Sci1981; 70: 583-589.

[29] Kier LB, Hall LH.J Pharm Sci1983;72: 1170-1173.