IJournals: International Journal of Software & Hardware Research in Engineering
ISSN-2347-4890
Volume 4 Issue 2 February, 2016

Modeling of the Retention indices of a set of polycyclic aromatic hydrocarbons using a hybrid approach

Khadidja Amirat[1] ;Fatiha Mebarki [2]; Nadia Ziani [3];Djelloul Messadi [4]

Environmental and food Security laboratory, Badji Mokhtar University, BP 12, 23000, Annaba, Algeria

1:khadidja_amirat@yahoo.fr

2:fatiha_mebarki@yahoo.fr

3:Ziani_nadia84@yahoo.fr

4: D_messadi@yahoo.fr

Abstract

A structure/retention indices relationship was searched for 93 PAHs while promoting the hybrid genetic algorithm/multilinear regression approach, the structural parameters being calculated with the software Spartan and DRAGON. Among about a hundred of 2 regressor models gotten, we selected the one that present best values of the prediction parameter (Q^2) and of the determination coefficient (R^2).the reliability of the proposed model was further illustrated using various evaluation techniques :leave-many-out, cross-validation procedure, randomization test, and validation through the test set.

Keywords: structure/ retention indices; PAHs; software; molecular descriptors.

1.INTRODUCTION

Polycyclic aromatic hydrocarbons (PAHs) are important classes of organic compounds, which usually have two to six fused benzene rings, with occasional incorporation of cyclopentene rings. A wide variation of alkyl substituents gives rise to thousands of different PAHs and many have been identified in environmental samples. PAHs are generally highly toxic and carcinogenic compounds[1] and ubiquitous contaminants of aquatic and atmospheric ecosystems, where they are present as a result of natural processes such as forest fires, volcanic emissions, but the predominant PAH sources in the environment are related to human activities such as oil spills, burning fossil fuels and domestic wastes,, transport emissions.

In past decades, Quantitative Structure-Activity properties Relationships / (QSAR / QSPR/ QSRR/) or QSXR has become a powerful theoretical tool, alternative to quantum mechanics, for the description and prediction of properties of complex molecular systems in different environments. The approach QSXR proceeds from the assumption of a one correspondence between any physical property, chemical affinity or biological activity of a chemical compound and its molecular structure [2]. The latter can be represented by the chemical composition, the connectivity of the atoms, the potential energy surface, and the electron wave function of a compound. Different physicochemical molecular descriptors reflecting the structure can be determined empirically or by using theoretical and computational methods of different complexities. It is emphasized that the knowledge of the exact chemical constitution and / or the three-dimensional molecular structure of the studied compounds is a prerequisite to the application of QSXR approach. The success of the approach QSXR depends critically on the precise definition and the appropriate use of molecular descriptors. We distinguish arbitrarily empirical molecular descriptors theoretical molecular descriptors. The empirical descriptors can be divided into two general classes, the first reflects the intramolecular electronic interactions (structural descriptors) while the second takes into account the intermolecular interactions in condensed media such as liquids and solutions (solvation descriptors).The objective of this work is to develop a robust QSRR model that could predict the retention indices for a diverse set of PAHs, using the general molecular descriptors and to seek the important features related to the retention indices value.

2. Materials and methods

2.1. Dataset:

The experimental retention indices values for 93 PAHs were taken from the article published by Jujun Kang *et al* [3] .A complete list of the compounds name and their corresponding retention indices of the 93 PAHs is shown in Table 1. The data set was randomly divided into two subsets: a training set of 70 compounds and a test set of 23 compounds.

2.2. Descriptor Generation:

The chemical structure of each compound was sketched on a PC using Spartan 10 [4]program and optimized using PM6 semi empirical method .the resulted geometry was transferred into the soft ware Dragon version 5.3[5], to calculate 1600

descriptors of the type geometrical and Getaway(Geometry, Topology and Atom Weighted Assembly).descriptors with constant or near constant values inside each group were discarded .For each pair of correlated descriptors (with correlation coefficient r≥0.95),the one showing the highest pair correlation with the other descriptors was excluded. The GA (Genetic Algorithm) [6] has been considered superior to other methods of variable selection techniques .So, variable selection was performed on the training set, using GA in the MobyDigs version of Todeschini [7] by maximizing the cross-validated explained variance Q^2_{LOO}.

The chemical structure of each compound was sketched on a PC using Spartan10[4] program and optimized using PM6 semi empirical method .the resulted geometry was transferred into the soft ware Dragon version 5.3[5], to calculate 1600 descriptors of the type geometrical and Getaway(Geometry, Topology and Atom Weighted Assembly).descriptors with constant or near constant values inside each group were discarded .For each pair of correlated descriptors (with correlation coefficient r≥0.95),the one showing the highest pair correlation with the other descriptors was excluded. The GA (Genetic Algorithm) [6] has been considered superior to other methods of variable selection techniques .So, variable selection was performed on the training set, using GA in the MobyDigs version of Todeschini [7] by maximizing the cross-validated explained variance Q^2_{LOO}.

2.4. Model Development and Validation:

Multiple linear regression analysis and variable selection were performed by package MobyDigs for windows/PC [7][31] using the Ordinary Least Square regression (OLS) method.

The goodness of fit of the calculated models were assessed by means of the multiple determination coefficients, R^2, and the standard deviation error in calculation (SDEC).

$$SDEC = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \qquad (1)$$

Cross validation techniques allow the assessment of internal predictivity (Q^2_{LMO} cross validation; bootstrap) in addition to the robustness of model (Q^2_{LOO} cross validation).

Cross validation methods consist in leaving out a given number of compounds from the training set and rebuilding the model, which is then used to predict the compounds left out. This procedure is repeated for all compounds of the training set, obtaining a prediction for every one. If each compound is taken away one at a time the cross validation procedure is called leave-one-out technique (LOO technique), otherwise leave-more-out technique (LMO technique). An LOO or LMO correlation coefficient, generally indicated with Q^2, is computed by evaluating the accuracy of these "test" compounds prediction.

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

The "hat" of the variable y, as is the usual statistical notation, indicates that it is a predicted value of the studied property, and the sub index "i/i" indicates that the predicted values come from models built without the predicted compound.

TSS is the total sum of squares.

The predictive residual sum of squares (PRESS) measures the dispersion of the predicted values. It is used to define Q^2 and the standard deviation error in prediction (SDEP).

$$SDEP = \sqrt{PRESS/n} \qquad (3)$$

A value $Q^2 > 0.5$ is generally regarded as a good result and $Q^2 > 0.9$ as excellent [32, 33][8,9].

However, studies [10, 11][34, 35] have indicated that while Q^2 is a necessary condition for high predictive power a model, is not sufficient.

To avoid overestimating the predictive power of the model LMO procedure (repeated 5000 times, with 5 objects left out at each step) was also performed ($Q^2_{L(5)O}$).

In bootstrap validation technique K n-dimensional groups are generated by a randomly repeated selection of n-objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2 is calculated for each model. The bootstrapping was repeated 8000 times for each validated model.

By using the selected model the values of the response for the test objects are calculated and the quality of these predictions is defined in terms of Q^2_{ext}, which is defined as

$$Q^2ext = 1 - \frac{\sum_{i=1}^{next}(\hat{y}_{i/i} - y_i)^2 / n_{ext}}{\sum_{i=1}^{ntr}(y_i - \bar{y}_{tr})^2 / n_{tr}} = 1 - \frac{PRESS / n_{ext}}{TSS / n_{tr}} \qquad (4)$$

Here n_{ext} and n_{tr} are the number of objects in the external set (or left out by bootstrap) and the number of training set objects, respectively.

Other useful parameters are R^2, calculated for the validation chemicals by applying the model developed on the training set, and external standard deviation error of prediction ($SDEP_{ext}$), defined as:

$$SDEP_{ext} = \sqrt{\frac{1}{n_{ext}}\sum_{i=1}^{n_{ext}}\left(yi - \bar{y}\right)^2} \qquad (5)$$

where the sum runs over the test set objects (n_{ext}).

According to Golbraikh and Tropsha[11]. A QSPR model is successful if it satisfies several criteria as follows :

$R^2_{CVext} > 0.5$ (6)
$r^2 > 0.6$ (7)
$(r^2-r^2_0)/r^2 <$ or $(r^2- r^{2\prime}0)/r^2 < 0.1$ (8)
$0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$ (9)

Here:

$$r = \frac{\sum(y_i - \tilde{y}_i)(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum(y_i - \tilde{y})^2 \sum(\tilde{y}_i - \bar{\tilde{y}}_i)^2}} \quad (10)$$

$$r_0^2 = 1 - \frac{\sum(\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum(\tilde{y}_i - \bar{\tilde{y}}_i)^2} \quad (11)$$

$$r_0'^2 = 1 - \frac{\sum(y_i - y_i^{r_0})^2}{\sum(y_i - \bar{y})^2} \quad (12)$$

$$k = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2} \quad (13)$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \quad (14)$$

$$T1 = \frac{(r^2 - r^2_0)}{r^2} \quad (15)$$

$$T2 = \frac{(r^2 - r'^2_0)}{r^2} \quad (16)$$

$$Ab = [r^2_0 - r'^2_0] \quad (17)$$

where r is the correlation coefficient between the calculated and experimental values in the test set; r^2_0 (calculated versus observed values) and r'^2_0 (observed versus calculated values) are the coefficients of determination; k and k' are slopes of regression lines through the origin of calculated versus observed and observed versus calculated, respectively $y_i^{r_0}$, $\tilde{y}_i^{r_0}$; are defined as $y_i^{r_0} = k\tilde{y}_i$ and $\tilde{y}_i^{r_0} = k'y_i$ and the summations runs over the test set.

2.5. QSAR AD (Applicability Domain)

The AD was discussed by the Williams plot [8, 9] of jacknified residuals versus leverages (hat diagonal values (h_i). The jacknifed residuals (or Studentized residuals) are the standardized cross-validated residuals. Each residuals is divided by its standard deviation, which is calculated without the i-th observation. The leverage(h_i). value of a chemical in the original variable space is defined as :

$$h_i = x_i (X^T X)^{-1} x_i^T \quad (i=1,\ldots..,n) \quad (18)$$

Where x_i is the descriptor row-vector of the query compound, and X is the n (p+1) matrix of p model parameter values for n training set compounds. The superscript T refers to the transpose of the matrix/vector. The warning leverage value (h^*) is defined as 3(p +1)/ n. When h value of a compound is lower than h^*, the probability of accordance between predicted and actual values is as high as that for the compounds in the training set. A chemical with $h_i > h^*$ will reinforce the model if the chemical is in the training set. But such a chemical in the validation set and its predicted data may be unreliable. However, this chemical may not appear to be an outlier because its residual may be low. Thus the leverage and the jacknified residual should be combined for the characterization of the AD.

3.Results and Discussion

Application of the GA-VSS led to several good models for the prediction of based on different sets of molecular descriptors. The best biparametric model was constructed using: Molecular weight; salvation energy .All data concerning value of descriptors and the retention indices are summarized in Table 1.

The equation of the optimal model can be written as:

Ri = - 50. ±(4.400) -4.63 ±(0.281)SOLV EN +1.52±(0.031) MW (19)

Here Molecular weight (MW); solvation energy(SOLV EN) are descriptors calculated with Spartan software, All relevant statistical parameters are reported in Table 2.

Values of $R^2$and R^2_{adj} attest the good fitting performances of the model which, moreover, is very highly significant (great value of the Fisher parameter *F*).

The model is robust, the difference between R^2 and Q^2 is small 0.09(%). The model demonstrates a very good stability in internal validation (difference between Q^2_{LOO} and Q^2_{LMO} is about 0.04(%). While bootstrapping confirms the internal predictivity and stability of the model.$SDEP_{ext}$is a little bit different from *SDEP*.Some important statistical parameters (as given in Table 3) were used to evaluate the involved descriptors. The t-value of a descriptor measures the statistical significance of the regression coefficients. The high absolute t-values shown in Table 3express that the regression coefficients of the descriptors involved in the GA/MLR model are significantly larger than the standard deviation. The t-probability of a descriptor can describe the statistical

significance when combined together within an overall collective QSRR model (i.e., descriptors' interactions). Descriptors with t-probability values below 0.05 (95% confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance[12][13] .The smaller t-probability suggests the more significant descriptor.The t-probability values of the tree descriptors are very small, indicating that all of them are highly significant descriptors.The VIF values and the correlation matrix as shown in Table 4 suggest that these descriptors are weakly correlated with each other. The distributions of errors for the entire dataset are given in **Figure1**.As the errors are distributed on both sides of the zero line,one may conclude that there is no systematic error in the model development.The model was also verified by Y-scrambling. **Figure 2** clearly ensures the existence of a linear relationship between Ri and the descriptors Molecular weight, As can be observed the permuted responses yield poor predictive models, all having $Q^2 < 0.2$ On the other hand, the correctly ordered Ri yield good statistical parameters, and therefore it is located isolated in the plot.

The statistical parameters of Tropsha et all reported in Table 5 were obtained for the test set, which obviously satisfy the generally accepted condition and thus demonstrate the predictive power of the present model:

R^2cv_{ext}= >0.5

r^2= >0.6

T1= $(r^2-r^20)/r^2$= -0.0012 <1

Or T2= $(r^2-r^{2'}0)/r^2$= -0.0012 <0.1

0.85≤ k= 1.0101 <1.15 or 0.85 k'= 0.9899 ≤1.15

Table 1: Values of Ri, Molecular weight,solvation energy for a set of 93 PAHs. The last 23 chemicals are the test set.

Chemical	Ri	Solvation energy (Solv En)	Molecular weight (MW)
Naphthalene	200	-11.24	128.174
1-Methylnaphthalene	221.04	-11.61	142.201
2-Ethylnaphthalene	236.08	-9.99	156.228
1-Ethylnaphthalene	236.56	-10.86	156.228
2,7-Dimethylnaphthalene	237.71	-11.75	156.228
1,3-Dimethylnaphthalene	240.25	-12.03	156.228
1,7-Dimethylnaphthalene	240.66	-11.6	156.228
1,6-Dimethylnaphthalene	240.72	-11.97	156.228
1,4-Dimethylnaphthalene	243.57	-11.62	156.228
Acenaphthelene	244.63	-15.77	152.196
1,5-Dimethylnaphthalene	244.98	-11.75	156.228
1,2-Dimethylnaphthalene	246.49	-11.95	156.228
2,3,6-Trimethylnaphthalene	263.31	-11.84	170.255
1-Methylacenaphthelene	265.24	-16.66	166.223
2,3,5-Trimethylnaphthalene	265.9	-12.12	170.255
Phenanthrene	300	-16.21	178.234
1-Phenylnaphthalene	315.19	-11.69	204.272
3-Methylphenanthrene	319.46	-16.56	192.261
2-Methylanthracene	321.57	-15.64	192.261
2-Methylphenanthrene	321.57	-16.3	192.261
4-Methylphenanthrene	323.17	-16.97	192.261
Chemical	**Ri**	**Solvation energy (Solv En)**	**Molecular weight (MW)**
1-Methylanthracene	323.33	-15.67	192.261
1-Methylphenanthrene	323.9	-16.73	192.261
9-Methylanthracene	329.13	-17.14	192.261
9-Ethylphenanthrene	337.05	-15.98	206.288
2-Ethylphenanthrene	337.5	-14.89	206.288
3,6-Dimethylphenanthrene	337.83	-16.89	206.288
2,7-Dimethylphenanthrene	339.23	-16.32	206.288
9-Isopropylphenanthrene	345.78	-13.62	220.315
1,8-Dimethylphenanthrene	346.26	-17.19	206.288
9-n-Propylphenanthrene	350.3	-14.47	220.315
Pyrene	351.22	-20.51	202.256
9-Methyl-10-Ethylphenanthrene	359.91	-15.5	220.315
1-Methyl-7-isoprppylphenanthrene	368.67	-14	234.342
4-Methylpyrene	369.54	-21	216.283
1-Methylpyrene	373.55	-21.3	216.283
9,10-Dimethyl-3-ethylphenanthrene	381.85	-16.26	234.342
1-Ethylpyrene	385.35	-20.48	230.31
2,7-Dimethylpyrene	386.34	-20.64	230.31
Benzo(c)phenanthrene	391.39	-19.54	228.294
9-Phenylanthracene	396.38	-14.51	254.332
Cyclopenta(cd)pyrene	396.54	-24.09	226.278
Benzo(a)anthracene	398.5	-19.93	228.294
Triphenylene	400	-21.24	228.294
9-Phenylphenanthrene	406.9	-15.76	254.332
11-Methylbenzo(a)anthracene	412.72	-20.37	242.321

1-Methylbenzo(a)anthracene	414.37	-20.87	242.321
1-n-Butylpyrene	414.87	-18.24	258.364
1-Methyltriphenylene	416.32	-20.86	242.321
9-Methylbenzo(a)anthracene	416.5	-20.17	242.321
9-Methyl-10-phenylphenanthrene	417.16	-15.77	268.354
8-Methylbenzo(a)anthracene	417.56	-20.35	242.321
6-Methylbenzo(a)anthracene	417.57	-20.39	242.321
3-Methylchrysene	418.1	-21.11	242.321
2-Methylchrysene	418.8	-20.83	242.321
Chemical	**Ri**	**Solvation energy (Solv En)**	**Molecular weight (MW)**
12-Methylbenzo(a)anthracene	419.39	-20.52	242.321
4-Methylbenzo(a)anthracene	419.67	-20.51	242.321
5-Methylchrysene	419.68	-20.35	242.321
4-Methylchrysene	420.83	-20.42	242.321
1-Phenylphenanthrene	421.66	-16.02	254.332
1-Methylchrysene	422.87	-21.24	242.321
7-Methylbenzo(a)anthracene	423.14	-21.88	242.321
1,12-Dimethylbenzo(a)anthracene	436.82	-17.49	256.348
Benzo(j)fluoranthene	440.92	-25.17	252.316
Benzo(b)fluoranthene	441.74	-23.86	252.316
Benzo(k)fluoranthene	442.56	-23.42	252.316
1,6,11-Trimethyltriphenylene	446.24	-21.44	270.375
Benzo(e)pyrene	450.73	-25.11	252.316
Benzo(a)pyrene	453.44	-24.4	252.316
Perylene	456.22	-24.93	252.316
Pentacene	486.81	-22.65	278.354
Dibenzo(a,c)anthracene	495.01	-24.59	278.354
Dibenzo(a,h)anthracene	495.45	-24.35	278.354
Picene	500	-25.02	278.354
Dibenzo(def,mno)chrysene	503.89	-27.94	276.338
2-Methylnaphthalene	218.14	-11.4	142.201
2,6-Dimethylnaphthalene	237.58	-11.34	156.228
2,3-Dimethylnaphthalene	243.55	-11.73	156.228
1,8-Dimethylnaphthalene	249.52	-12.71	156.228
Anthracene	301.69	-15.34	178.234
9-Methylphenanthrene	323.06	-16.71	192.261
2-Phenylnaphthalene	332.59	-13.02	204.272
Fluoranthene	344.01	-20.27	202.256
9,10-Dimethylanthracene	355.49	-18.19	206.288
2-Methylpyrene	370.15	-20.63	216.283
Chrysene	400	-20.72	228.294
2-Methylbenzo(a)anthracene	413.78	-20.26	242.321
3-Methylbenzo(a) anthracene	416.63	-20.04	242.321
5-Methylbenzo(a)anthracene	418.72	-20.54	242.321
Chemical	**Ri**	**Solvation energy (Solv En)**	**Molecular weight (MW)**
6-Methylchrysene	420.61	-21.61	242.321
1,3-Dimethyltriphenylene	432.32	-21.37	256.348
7,12-Dimethylbenzo(a)anthracene	443.38	-21.4	256.348
Benzo(b)chrysene	497.66	-24.29	278.354

Table 2 statistical parameters of a developed model.

n_{tr}	n_{ext}	Q^2_{LOO}	R^2	$Q^2_{LMO/50}$	Q^2_{BOOT}	R^2_{adj}	Q^2
70	*23*	99.18	99.27	99.1324	99.12	99.25	99.17
SDEC	***SDEP***	***SDEP$_{ext}$***	***S***	***F***			
6.108	6.475	6.519	6.2428	4569.847			

Table 3. Characteristics of the selected descriptors in the best GA/ MLR model.

Predictor	Coef	SE Coef	T	P	VIF
Constant	-49.968	4.400	-11.36	0.000	-
SOLV EN	-4.6285	0.2803	-16.51	0.000	2.341
WEIGHT	1.52264	0.03104	49.05	0.000	2.341

Table 4: correlation matrix between retention indices and the selected descriptors.

	Ri	SOLV EN	Molecular WEIGHT
Ri	*1.000*		
SOLV EN	-0.855	*1.000*	
Molecular WEIGHT	0.981	-0.757	*1.000*

Table 5: The statistical parameters of Tropsha *et al.*

R^2cv ext	r^2	k	k'	r^2o
0.9954	0.9969	1.0101	0.9899	0.9980
r^2'o	Ab	T1	T2	
0.9980	0.000	-0.0012	-0.0012	

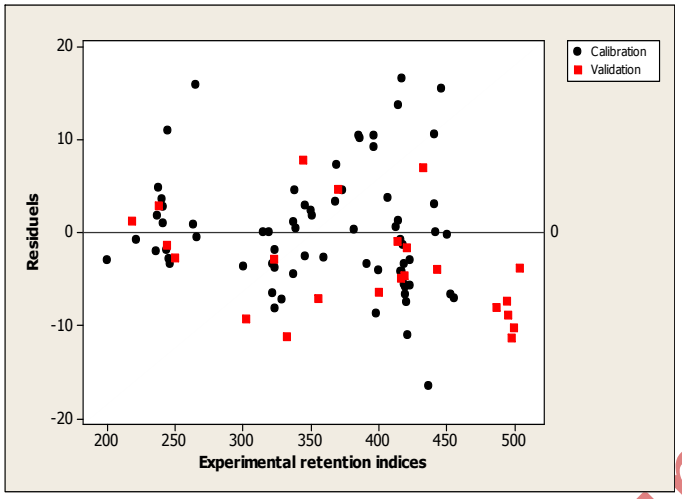

Fig1 :Residuals versus Ri (exp) for the entire dataset.

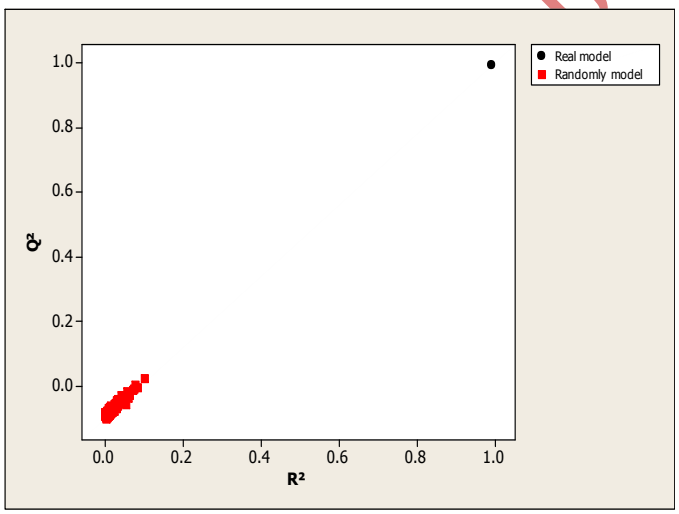

Fig 2 :Randomization test associated to the previous QSRR model. Square represent the randomly orded retention and the circle corresponds to the real retention.

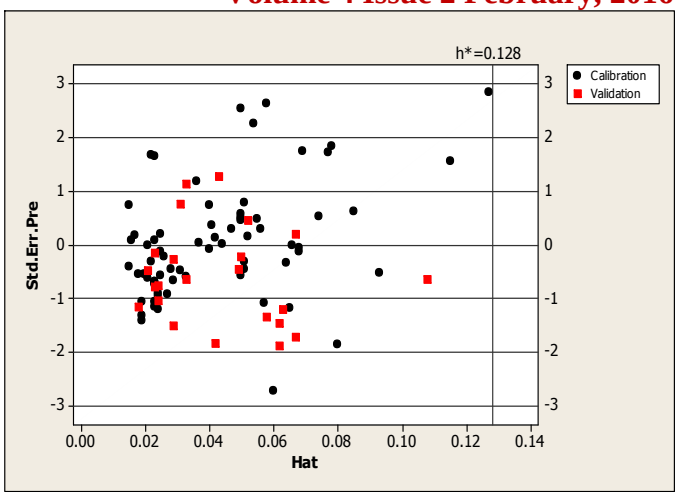

Fig 3: Williams plot of the current QSRR model.

3.3 Applicability Domain

On analyzing the model applicability domain from Williams plot, all residuals were located within the range of three Standard deviations, and there is no influential compound both for training or prediction set (Figure.3), which means that the model has a good external predictivity.

4. Conclusion

A QSPR model for the estimation of the retention indices for 93 PAHs was established in the following six steps: molecular structure input and generation of the files containing the chemical structures;quantum mechanics geometry optimization with a semi –emperical method;structural descriptors computation;structural descriptors selection;structure model generation with a multivariate method and statistical analysis.

According to obtained results it is concluded that the salvation energy and the molecular weight can be used succesfully for modeling retention indices (Ri) of the under study compounds. High correlation coefficient (0. 9927) and low prediction error (SDEP=6.475;SDEPext=6.519) obtained confirm good predictive ability of the model.The QSRR model proposed with the double calculated molecular descriptors can be used to estimate retention indices for new compounds even in the absence of the standard candidates.

References

[1] National Research Council (1979), Committee on the Assessment of Polychlorinated Biphenyls in the Environment. Polychlorinated biphenyls: a report; National Academy of Sciences: Washington, U.S.A.

[2] Angulo Lucena, R., Farouk Allam, M., Serrano Jiménez, S. and Luisa Jodral Villarejo, M. A.(2007), "review of environmental exposure to persistent organochlorine residuals during the last fifty years". Current Drug Safety,Vol. 2 No.2, pp. 163-172.

[3] Roveda, A. M., Veronesi, L., Zoni, R., Colucci, M. E. and Sansebastiano, G. (2006), "Exposure to polychlorinated

biphenyls (PCBs) in food and cancer risk: recent advances",Igiene e Sanita Pubblica,Vol. 62 No.6, pp. 677-696.

[4] Lundqvist, C.,Zuurbier, M., Leijs, M., Johansson, C.,Ceccatelli, S.,Saunders, M.,Schoeters, G., Ten Tusscher, G. and Koppe, J. G. (2006)," The effects of PCBs and dioxins on child health", Acta Paediatrica,Vol.95 No.453,pp.55-64.

[5] Poppenga, R. H. (2000), "Current environmental threats to animal health and productivity", The Veterinary Clinics of North America. Food Animal Practice ,Vol. 16 No.3, pp.545-558.

[6] Bren, U.,Zupan, M., Guengerich, F. P. and Mavri, J.(2006)" Chemical Reactivity as a Tool to Study Carcinogenicity: Reaction between Chloroethylene Oxide and Guanine", The Journal of Organic Chemistry ,Vol. 71 No.11,pp. 4078-4084.

[7] Lebeuf, M., Noël, M.,Trottier, S. and Measures, L. (2007), "Temporal trends (1987-2002) of persistent,bioaccumulative and toxic (PBT) chemicals in beluga whales (Delphinapterus leucas) from the St.Lawrence Estuary, Canada", Sciences of the Total Environment, Vol.383 No. (1-3), pp.216-231.

[8]]Eriksson L., Jaworska J., Worth A., Cronin M Mc., Dowell R.M. & Gramatica P., 2003. Methods for Reliability, uncertainty assessment , and applicability evaluations of regression based and classification QSARs, Environmental Health Perspectives, Vol. 111(10),1361-1375.

[9] Tropscha A., Gramatica P. & Grombar V.K., 2003.The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, QSAR & Combinatorial Science, Vol. 22(1), 69-77

[10] Kubinyi H., Hamprecht F.A. & Mietzner T., 1998. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices, Journal of Medicinal Chemistry, Vol.41(14), 2553-2564.

[11] Golbraikh, A. and Tropsha, A. (2002), "Beware of q^2!",Journal of Molecular Graphics and Modelling, Vol.20 No.4, pp.269-276.

[12] Ramsey, L. F.; Schafer, W. D. The Statistical Sleuth; Wadsworth Publishing Company: USA, 1997.

