Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS
J. Serb. Chem. Soc. *86* (1) 63–75 (2021)
JSCS–5404
Original scientific paper

Predicting retention indices of PAHs in reversed-phase liquid chromatography: Quantitative structure retention relationship approach

NABIL BOUARRA[1,2]*, NAWEL NADJI[1], LOUBNA NOURI[1], AMEL BOUDJEMAA[1], KHALDOUN BACHARI[1] and DEJLLOUL MESSADI[2]

[1]*Centre de Recherche Scientifique et Technique en Analyses Physico–Chimiques, BP 384, Zone Industrielle Bou-Ismail, 42004 Tipaza, Algeria and* [2]*Laboratory of Environmental and Food Safety, Department of Chemistry, Badji Mokhtar– Annaba University, PB 12, 23000, Annaba, Algeria*

(Received 19 February, revised 24 March, accepted 16 April 2020)

Abstract: In this work, the liquid chromatography retention time in monomeric and polymeric stationary phases of PAHs was investigated. Quantitative structure retention relationship approach has been successfully performed. At first, 3224 molecular descriptors were calculated for the optimized PAHs structure using Dragon software. Afterwards, the modelled dataset was divided using the CADEX algorithm into two subsets for internal and external validation. The genetic algorithm-based on a multiple linear regression was used for feature selection of the most significant descriptors and the model development. The selected models with five descriptors: *nCIR*, *GGI3*, *GGI4*, *JGT* and *DP14* were used for the monomeric column and *nR10*, *EEig01x*, *L1m*, *H5v* and *HATS6v* were introduced for the polymeric column. Robustness and predictive performance of the suggested models were verified by both internal and external statistical validation. The good quality of the statistical parameters indicates the stability and predictive power of the suggested models. This study demonstrated the suitability of the established models in the prediction of liquid chromatographic retention indices of PAHs.

Keywords: molecular descriptors; genetic algorithm; multiple linear regression; prediction.

INTRODUCTION

Polycyclic aromatic hydrocarbons (PAHs) establish a huge family of neutral and stable organic compounds, composed of carbon and hydrogen, containing from 2 to 6 fused aromatic rings.[1] PAHs are found in our environment as ubiquitous, persistent and toxic molecules.

* Corresponding author. E-mail: bouarranabil@yahoo.com
https://doi.org/10.2298/JSC200219019B

Indeed, the pyrolysis or the carbonization of organic compounds as coal, oil, and wood is the major source of PAHs.[2] Generally, PAHs are applied as intermediate compounds in pharmaceuticals, photographic products, lubricating materials, agricultural foodstuffs, thermosetting plastics, and other industrial products.[3,1] The emissions of PAHs during the incomplete combustion pose a serious disquiet for the environment.[2] The reversed-phase liquid chromatography (RPLC) is one of the important methods most used, versatile analytical techniques.[4] Usually, the C18 phase is used in RPLC and LC[5] as a stationary phase due to its excellent detection, separation and selectivity to PAHs and their isomers.[6]

The experimental determination of LC retention indices (*RI*) for all possible PAHs compounds is labor-intensive, and thus time-consuming and expensive. Recently, alternative approaches have been extensively investigated in an attempt to enhance the performance of substitute ways to obtain theoretical *RI*.[4]

Nowadays, quantitative structure retention relationship (QSRR) has gained much interest from researchers in the area of separation science. The QSRR is a powerful tool that provides promising methods for the valuation of *RI* based on structural descriptors calculated from the molecular structure.[7] The advantage of the QSRR approach is that the built model permits the estimation of *RI* for an unknown, unmeasured or novel compounds which have a similar structure to those used to build up the model.[8] Moreover, the use of QSRR allows for the fast and easily done input of the compounds' structures that are usually studied. Also, QSRR is used for the calculation, the analysis of descriptors and for the generation and validation of the model equation. Several estimation methods to understand the specific molecular interactions that govern PAHs chromatographic separation have been investigated[9-12] using topology of the molecule and/or quantum chemistry parameters calculated for the optimized molecular structure.

The main purpose of the present work is to establish robust and accurate models, which are capable to estimate the RI for the PAHs using combined method genetic algorithm-multiple linear regression (GA-MLR). The predictive power of the obtained models was validated by a prediction set. Moreover, the applicability domain of the developed models was checked graphically based on Williams plot. In addition, the developed models can help to understand and describe the retention behavior by highlighting the necessary factors to represent the relationship between the *RI* of PAHs and their molecular structure.

EXPERIMENTAL

Dataset

The experimental *RI* were taken from the study of Sander and Wise.[13] The reported values are between 2 and 6 for both polymeric and monomeric columns. The retention data are expressed as logarithms of *RI* (log *I*, Table S-I of the Supplementary material to this paper).

RI of the liquid chromatography of both monomeric and polymeric C18 reversed-phase column were determinated using Vydac 201TP 10μ (polymeric) 4.6 mm×25 cm (the separation group, Hesperia, CA, USA) and Zobrax ODS 6μ(monomeric), 6.4 mm×25 cm (Dupon, Wilington, DE, USA). The mobile phase used was composed of 85 % acetonitrile in water for both columns. Pop *et al.*[14] illustrated that *RI* was similar to Kováts indices in the gas chromatography (GC). The equation used was expressed as:

$$\log I_x = \log I_n + \frac{\log R_x - \log R_n}{\log R_{(n+1)} - \log R_n} \quad (1)$$

where x, n and (n+1) and R represents the solute, the lower, the higher standard and the values are the corresponding corrected retention volumes, respectively.

Descriptors generation

The chemical structures of the studied molecules were downloaded from the WebBook database[15] as SD-file format. The final geometry of the minimum energy conformation was obtained by the B3LYP functional approach[6] in combination with the 6-31G (d) basis set. For more information, the technical details of the geometric optimization of molecules are available at the NIST website.[15] Afterwards, Dragon software V5.5[17] was used for the calculation of 3224 molecular descriptors to describe the chemical's structural diversity. Then, constant descriptors, near-constant descriptors and highly correlation descriptors (R > 0.95) were excluded by using built-in variable exclusion procedure in Dragon to reduce the initial pool of descriptors and eliminate the redundant information that was not used. In the end, 398 molecular descriptors were retained.

Dataset division

In order to generate a strong QSRR model that could estimate the *RI* of PAHs, based on the algorithm of Kennard and Stone (Cadex),[18] the data set was divided into the training and the prediction set. The training set made up of 92 compounds was investigated to make the final model while the remaining 40 compounds (prediction set) were used to validate the built model (see Table S-II of the Supplementary material).

Model development and validation

QSARINS software[19] was used to construct the QSRR model. In order to select the best modelling descriptors, the statistical quality of all combinations of the whole descriptors was explored by using multiple linear regression and the genetic algorithm-variable subset selection (GA/VSS) methods based on ordinary least squares (OLS). The variable subset selection procedure generates a 'population' of models, ranked according to decreasing R^2 values. The optimal models were chosen according to Q^2Leave-One-Out (Q^2_{LOO}) as the optimization value and take into consideration the parsimony principle regarding the complexity of the models, which should be as small as possible. Moreover, the correlation between the descriptors and the response was verified by the Q under influence of K rule[20] to remove models with high predictor collinearity and exclude chance correlation.

A key step in QSRR studies is the validation of the built models, aims to guarantee their goodness of fit, reliability, robustness and ability to provide good predictions for new compounds.[21]

The coefficient (R^2) was calculated to evaluate the goodness-of-fit and to estimate the degrees of overall correlation. A robust model should have an R^2 values greater than 0.7.[22] The R^2 was calculated using:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \quad (2)$$

where y_i, $\hat{y}_i$, $\bar{y}$, and n represents the experimental, the calculated, the mean value of the experimental *RI* and the number of samples in the training set; respectively.

The reliability and consistency of a model can be investigated via the cross-validation technique (CV).[23] Leave-one-out (LOO) and leave-many-out (LMO) strategies can be carried out by calculating the cross-validation coefficient (Q^2_{LOO}) and (Q^2_{LMO}). In the LOO technique, each time one sample from the training set would be removed, consequently, several models will be generated. Whereas, in LMO technique M represents a group of compounds randomly selected that would be removed at the beginning and be predicted by the model, which was developed using the remaining compounds.

According to Chirico and Gramatica,[22,24] a model is statistically reliable if the Q^2_{LOO} value is greater than 0.6. Q^2_{LOO} values close to the Q^2_{LMO} indicated that the model is robust.[25] The Q^2_{LOO} is defined as follows:

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_{i/i} - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \quad (3)$$

where, $\hat{y}_{i/i}$ is the value of log *I* predicted by the model built without the compound *i* according to LOO method.

The predictive power of the built models was evaluated by calculating several external validation metrics such as R^2_{ext}, Q^2_{F1},[26] Q^2_{F2},[27] Q^2_{F3},[28,29] and the concordance correlation coefficient (*CCC*).[22,24,30] For the predictive power of a model to be considered adequate, the values of Q^2_{F1}, Q^2_{F2} and Q^2_{F3} should be greater than 0.7, and the CCC_{ext} value must be greater than 0.85.[22,24] The equations defined the external validation metrics are regrouped in Table S-III of the Supplementary material.

Besides, the root mean square error (*RMSE*)[31] is calculated by squaring individual errors, summing them, dividing the sum by their total number, and then taking the square root of this quantity. Therefore, the *RMSE* summarizes the global error of the model used to measure and compare the accuracy of the predictions in training and prediction set, *i.e.*, the precision of the QSRR and can be applied to predictions (*i.e.*, $RMSE_{Pr}$):

$$RMSE_{tr(pr)} = \sqrt{\frac{1}{n_{tr(pr)}}\sum_{i=1}^{n_{tr(pr)}}\left(y_i - \hat{y}_i\right)^2} \quad (4)$$

Applicability domain (AD)

The applicability domain (AD)[32] of constructed model is a theoretical region in chemical space, defined by the modeled response and model descriptors. In this region, compounds possess similar structure, biological or physicochemical properties to the ones of the training compounds. Williams plot is the typical graphical description of the AD, which represents the standardized residuals versus leverages values (h_i). Leverage values are calculated as the diagonal of the Hat matrix:

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

where x_i is the descriptor vector of the compound and X is the **n**×**p** matrix containing **p** descriptor values and **n** training compounds. Generally, the warning leverage[33] is defined as follows:

$$h^* = \frac{3(k+1)}{m} \tag{2}$$

where m is the number of compounds in the training set and k is the number of descriptors in the developed model.

Williams plot utilized as a graphical detection of both the outliers responses (*i.e.*, chemicals with absolute standardized residuals greater than 3 standard deviation units) and the structurally influential chemicals (*i.e.*, chemicals with leverage values greater than the threshold value $h_i > h^*$).[34]

To ensure that the established model is reliable and robust, the Y-scrambling is one of the most widely used techniques.[34] Indeed, it is not uncommon to obtain fortuitous correlations, *i.e.*, a model with good statistical results (R^2, Q^2) for training, but involving descriptors that in reality are not related to the modeled property. These random models can be detected by the Y-scrambling procedure. It consists of randomly mixing the experimental property for the training set and using the same descriptors, re-training the learning algorithm to try to obtain a model. Normally, the obtained models should have very low performance.[19]

RESULTS AND DISCUSSION

In order to develop the models, each column was treated separately by using the same pool of available descriptor to the regression analysis procedure. For each column type, it was a very important to see which of the descriptors were considered significant in predicting the RI values. The regression analysis procedure generated several reasonable models. The choice of the best model was made by using many criteria such as the size of the model, the multiple correlation coefficients (R^2), leave one out cross-validation coefficient (Q^2_{LOO}), the external validation metrics, *etc*. The best model is one that has a high value of the coefficients cited above. The optimal selected models for each column were deduced to be:

Monomeric column:

$$\log I = 1.40 + 0.0176nCIR + 0.442GGI3 + 0.733 + 1.36JGT + 0.130DP14 \tag{7}$$

$R^2 = 0.9799$, $Q^2_{LOO} = 0.9771$, $Q^2_{LMO} = 0.9770$, $RMSE_{cv} = 0.1155$, $RMSE_{tr} = 0.1083$, $CCC_{tr} = 0.9898$, $RMSE_{ext} = 0.1313$, $R^2_{ext} = 0.9721$, $Q^2_{F1} = 0.9714$, $Q^2_{F2} = 0.9706$, $Q^2_{F3} = 0.9705$, $CCC_{ext} = 0.9849$, $s = 0.1120$, $F = 838.6887$.

Polymeric column:

$$\log I = -7.11 + 0.0970nR10 + 1.95EEig01x + 0.169L1m + \\ + 1.41H5v + 0.905HATS6v \tag{8}$$

$R^2 = 0.9677$, $Q^2_{LOO} = 0.9623$, $Q^2_{LMO} = 0.9617$, $RMSE_{cv} = 0.1263$, $RMSE_{tr} = 0.1083$, $CCC_{tr} = 0.9836$, $RMSE_{ext} = 0.1292$, $R^2_{ext} = 0.9603$, $Q^2_{F1} = 0.9584$, $Q^2_{F2} = 0.9565$, $Q^2_{F3} = 0.9605$, $CCC_{ext} = 0.9770$, $s = 0.1209$, $F = 514.9955$.

The statistical parameters of the developed models prove that the established models are stable, robust and predictive. Thus, the two models were approved, R^2 values greater than 0.7 and CCC_{tr} values greater than 0.85. Additionally, these two models had the smallest $RMSE_{tr}$ values and the greatest CCC_{tr} values, which indicate that these models presented the least error and the smallest differences between the experimental and predicted data. Also, Q^2_{LOO} and Q^2_{LMO} values for models (polymeric and monomeric) are greater than 0.6 and very close to R^2. Additionally, the developed models presented the smallest $RMSE_{cv}$ values, which confirm that these are the best models. The external validation results of the developed models indicate that they have good predictive power, given that all presented R^2_{ext} values greater than 0.7, CCC_{ext} values are also greater than 0.85. The external validation results (Q^2_{F1}, Q^2_{F2} and Q^2_{F3}) for each subset make it clear that the two models were approved, in agreement with the criteria recommended in the literature.[22,24] Then, the built models were used to predict the prediction set data (Table S-I).

Table I shows the statistical and the definition of the selected descriptors. The regression coefficients of the descriptors presented in the models are significantly larger than the standard deviation indicated by the high absolute *t*-values. The values of probability (*P*) are less than 0.05 for each descriptor that means the presence of every descriptor is statistically significant and indicates that the models are not a result of mere chance. As can be seen from Table I, the variance inflation factor (*VIF*) values of all descriptors are less than five.[35] Thus, there is no collinearity between the selected descriptors, and the obtained models are stable.

TABLE I. Names, definitions and coefficients of selected descriptors in the developed models

Descriptor	Descriptor definition	Coeff	SE. Coeff	*T*	*P*	*VIF*
	Polymeric					
Constant		–7.110	0.627	–11.33	0.000	–
nR10	Number of 10-memberd rings	0.097	0.007	12.62	0.000	1.460
EEig01x	Eigen value 01 from edge adj. matrix weighted by edge degrees	1.945	0.136	14.30	0.000	1.789
H5v	H autocorrelation of lag 5 / weighted by atomic van der Waals volumes	0.168	0.007	23.24	0.000	1.951
L1m	1st component size directional WHIM index / weighted by atomic masses	1.410	0.265	5.31	0.000	2.347
HATS6v	Leverage-weighted autocorrelation of lag 6 / weighted by van der Waals volume	0.904	0.262	3.44	0.001	1.496
	Monomeric					
Constant		1.397	0.114	12.160	0.000	–
nCIR	Number of circuits	0.017	0.002	7.210	0.000	3.670
GGI3	Topological charge index of order 3	0.441	0.035	12.590	0.000	1.793
GGI4	Topological charge index of order 4	0.733	0.039	18.640	0.000	2.518
JGT	Global topological charge index	1.362	0.371	3.670	0.000	1.788
DP14	Molecular profile no. 14	0.130	0.007	17.98	0.000	1.544

As shown in Fig. 1, the experimental and predicted values are very close. This model fits well the experimental data (R^2 = 0.96, 0.97, *RMSE* = 0.11, 0.10 for the training set and R^2_{ext} = 0.96, 0.97, *RMSE* = 0.13, 0.11 for the prediction set) for polymeric and monomeric column; respectively. A good internal robustness (Q^2_{LMO} = 0.97, 0.96). So, the proposed models show an excellent agreement between the experimental and predicted values.

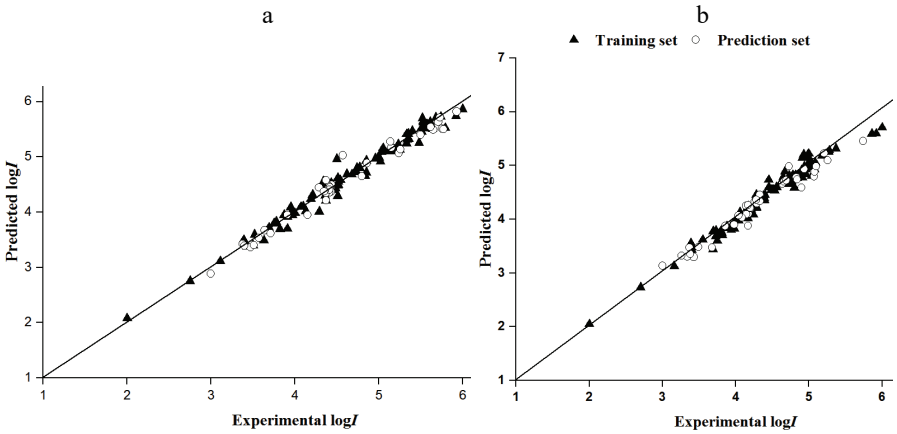

Fig. 1. Predicted versus experimental values of log *I*: a) monomeric column; b) polymeric column.

Fig. 2 shows the residuals of the training and prediction data set. All the residuals, that are distributed on both sides of the zero line uniformly and randomly, indicated the absence of the systematic error in the developed model.

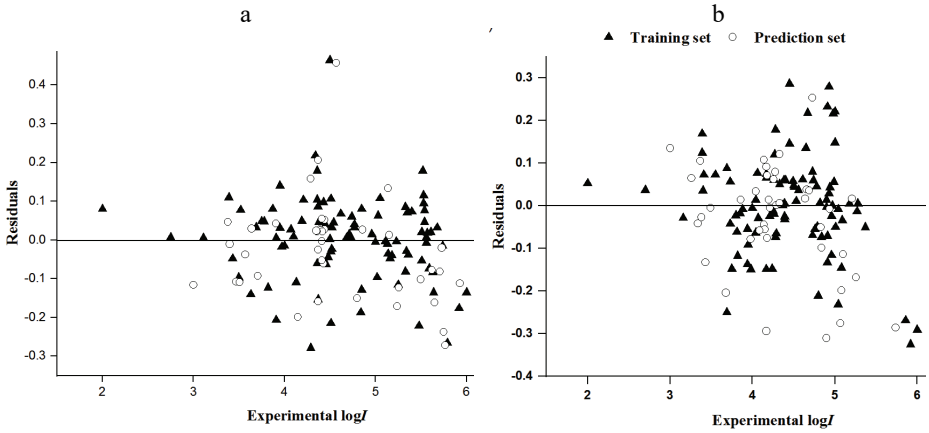

Fig. 2. The residuals *vs.* training and prediction values for the developed models: a) monomeric column; b) polymeric column.

The built models were analyzed using Williams plots. The results show that most of 132 compounds are within the area of the applicability domain of the model and have been well predicted (Fig. 3). The leverage value of Naphthalene for the polymeric column model is greater than the threshold $h^* = 0.195$, while for the monomeric column model four compounds (Indeno[1,2,3-*cd*]fluoranthene, naphthalene, benzo[ghi]perylene and dibenzo[def,mno]chrysene) from the training set exceed the h^* value ($h^* = 0.195$). But their standardized residual values are less than 3s. Thus, these compounds can stabilize the models and make them more precise.[33]

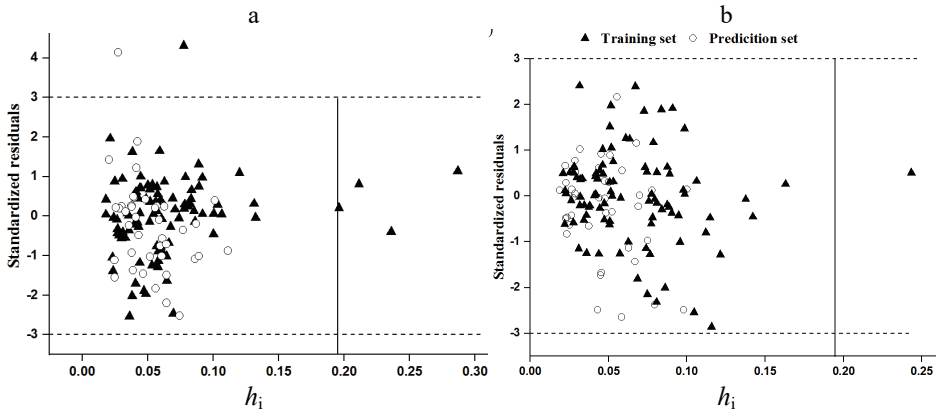

Fig. 3. Williams plot: a) monomeric column; b) polymeric column.

In order to identify outlier compounds, Williams plot shows that the standardized residuals for all molecules in the training and prediction sets are smaller than three standard deviation units (±3s) in absolute value, except for monomeric model only one compound in the training set (Benzo[a]naphthacene) and one in the prediction set (perylene-3-methyl) were wrongly predicted, but have lower leverage values, which means that they belong to the applicability domain of the developed models. When the statistical parameters Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and $CCC_{ext,}$, which are indicators of the predictive ability have high values. Then, the predicted log I by the developed model is reliable. Perhaps, the wrong predictions may have happened due to incorrect experimental values.

In the goal of verifying the robustness of the developed models, Y-randomization test was applied. The dependent variable vector (log RI) was shuffled randomly within the training set by using 200 iterations, knowing that on every iteration a novel model will be generated.

Figure 4 (A-B) which represents the diagram of the statistical coefficients Q^2_{Yscr} and R^2_{Yscr} makes it possible to compare the results obtained for the randomized models (squares) with the developed model (ring). It is clear that R^2

and Q^2 values of the model are very high compared to the values obtained for the randomized models. These results indicate that the robustness and the reliability of the developed models are not due to chance correlation.

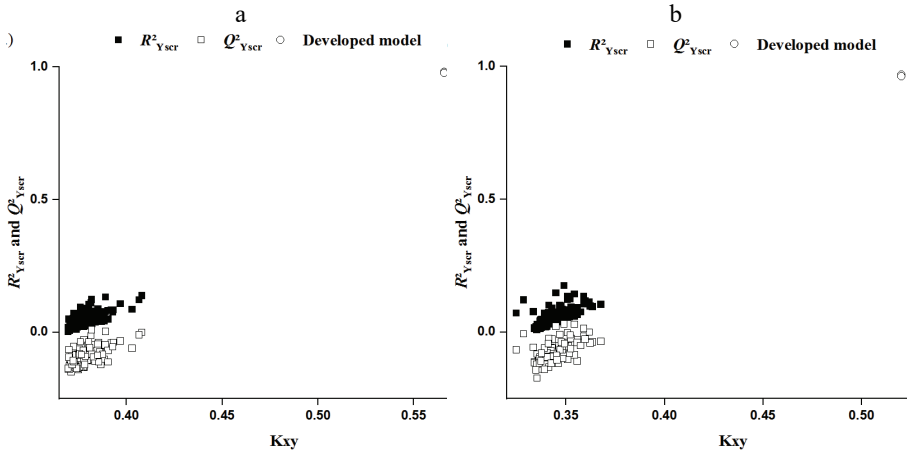

Fig. 4. Y-Scramble plot of R^2 and Q^2 *vs*. *Kxy* for random models (*Kxy*: correlations among the block of the descriptors and the experimental data).

The orders of importance for the selected descriptors in monomeric and polymeric columns are as follow:

For the monomeric column: *GGI4* (27.4862 %) > *DP14* (26.2492 %) > *nCIR* (19.7418 %) > *GGI3* (14.2261 %) > > *JGT* (12.2967 %).

For polymeric column: *L1m* (32.419 %) > *EEig01x* (22.102 %) > *nR10* (19.789 %) > *H5v* (13.444 %) > > *HATS6v* (12.2434 %).

The two descriptors belonged to the GETAWAY-type descriptor[36] are namely *H5v* and *HATS6v*. *H5v* which belongs to the GETAWAY *H*-indices-type descriptors that was calculated by *H* autocorrelation function along with the topological structure and weighted by atomic van der Waals volumes.[36] The diagonal elements of the molecular influence matrix were used to calculate the *HAT6v* descriptor taking into account the relative position of each atom in the three-dimensional molecular space weighted by van der Waals volume.[36] The *HAT6v* values are directly proportional to the ramification of molecules. The positive influence of *H5v* and *HATS6v* on the log *I* indicated that the logarithm of *RI* in the polymeric column will be increased when the volume of a solute is increased.

The WHIM descriptor[37] involved in polymeric model *L1m*, represent different sources of chemical information for the entire three-dimensional molecular structure such as size, shape, symmetry, and atom distribution. They are calculated by carrying out a PCA on a weighted covariance matrix of the centred rectangular coordinates of a molecule obtained from different atoms weighting

schemes.[37] *L1m* encodes the size of the molecule along with the second component weighted *via* atomic masses. In Eq. (8), *L1m* showed a positive effect, indicating that the increase of the molecule size increases the log *I*.

GGI3 and *GGI4* are topological charge indexes of order 3 and 4 respectively, while *JGT* is the global topological charge index which belongs to the Galvez topological charge indices. Topological descriptors facilitate the identification of the molecules through to their size, degree of ramification, flexibility and global shape.[38] Topological charge indices describe the charge transfer between atoms, and therefore the global charge transfer in the molecule as it relates to the topology.[39] These descriptors have a positive sign in the model equation which means log *I* increase with the increasing of these predictors.

The two constitutional descriptors *nR10* and *nCIR* are reported in the built models reflecting the molecular composition of a compound without connectivity and geometry information. The *nR10* descriptor explains the presence of either independent or fused 10-membered rings in molecules which play the main role in the determination of physicochemical properties.[38] While the *nCIR* descriptor is the number of the circuit and includes both rings and circuits. The positive coefficient of *nR10* and *nCIR* suggests that the log *I* increase with the increasing number of *n*-member rings.

The *EEig01x* is an edge adjacency index, that belongs to topological descriptor obtained from the edge adjacency matrix, which encodes the connectivity between graph edges. The positive coefficient of *EEig01x* descriptor indicates that the *RI* increases with the increasing of *EEig01x* values.[38] The Randić molecular profile *DP14*[40] can characterize the three-dimensional structure of each molecule. The Randić molecular profile descriptors are particularly suitable in similarity/diversity analysis since each profile well characterizes a molecule. The positive sign of *DP14* coefficient suggests that the log *I* increase with the molecule branching.

CONCLUSION

QSRR method was utilized to investigate the relationship between the molecular structure of PAHs and their retention indices. The good performance of statistical parameters (R^2, Q^2_{LOO}, $Q^2_{F1\text{-}F3}$ and *CCC*) and low *RMSE* values suggest that the developed models possess a good predictive capacity, which helps to estimate the *RI* of PAHs in cases where *RI* values are not available. The results provide a simple and straightforward way to predict the *RI* just from the molecular structures and gave some insight into the structural features related to *RI* of the PAHs.

Acknowledgement. The authors are grateful to the Directorate General for Scientific Research and Technological Development DGRSDT of Algeria for the financial support to this work.

SUPPLEMENTARY MATERIAL

Additional data are available electronically at the pages of journal website: https://www.shd-pub.org.rs/index.php/JSCS/index, or from the corresponding author on request.

И З В О Д

ПРЕДВИЂАЊЕ РЕТЕНЦИОНИХ ИНДЕКСА ПОЛИЦИКЛИЧНИХ АРОМАТИЧНИХ УГЉОВОДОНИКА У РЕВЕРСНО-ФАЗНОЈ ТЕЧНОЈ ХРОМАТОГРАФИЈИ: ПРИСТУП КВАНТИТАТИВНЕ РЕЛАЦИЈЕ СТРУКТУРЕ И РЕТЕНЦИОНИХ ИНДЕКСА

NABIL BOUARRA[1,2], NADJI NAWEL[1], NOURI LOUBNA[1], AMEL BOUDJEMAA[1], KHALDOUN BACHARI[1] и DEJLLOUL MESSADI[2]

[1]*Centre de Recherche Scientifique et Technique en Analyses Physico–Chimiques, BP 384, Zone Industrielle Bou-Ismail, 42004 Tipaza, Algeria* и [2]*Laboratory of Environmental and Food Safety, Department of Chemistry, Badji Mokhtar– Annaba University, PB 12, 23000, Annaba, Algeria*

У овом раду су истраживана ретенциона времена у мономерним и полимерним стационарним фазама полицикличних ароматичних угљоводоника (ПАУ) у течној хроматографији. Приступ квантитативном релацијом структуре и ретенције успешно је спроведен. Почето је са 3224 молекулска дескриптора израчунатим за оптимизоване структуре ПАУ помоћу Dragon софтвера. Затим је, моделовани скуп података подељен користећи CADEX алгоритам у два подскупа за интерну и екстерну валидацију. Генетички алгоритам заснован на вишеструкој линеарној регресији употребљен је да се изврши избор карактеристика најзначајнијих дескриптора и развој модела. Одабрани модели који укључују пет дескриптора: *nCIR*, *GGI3*, *GGI4*, *JGT* и *DP14* употребљени су за мономерне колоне, а *nR10*, *EEig01x*, *L1m*, *H5v* и *HATS6v* су уведени за полимерну колону. Робустност и ваљаност предвиђања сугерисаних модела проверене су и интерном и екстерном статстичком валидацијом. Висок квалитет статистичких параметара указује на стабилност и способност предвиђања сугерисаних модела. Ова студија демонстрира погодност утврђених модела за предвиђање ретенционих индекса код течне хроматографије ПАУ.

(Примљено 19. фебруара 2019, ревидирано 24. марта, прихваћено 16. априла 2020)

REFERENCES

1. M. Pogorzelec, K. Piekarska, *Sci. Total Environ*. **631** (2018)1431 (https://dx.doi.org/10.1016/j.scitotenv.2018.03.105)
2. H. I. Abdel-Shafy, M. S. M. Mansour, *Egypt. J. Petrol*. **25** (2016) 107 (https://dx.doi.org/10.1016/j.ejpe.2015.03.011)
3. N. E. Kaminski, B. L. Faubert Kaplan, M. P. Holsapple, *Casarett and Doull's Toxicology, the basic science of poisons*, C. D. Klaassen (Ed.), Mc-Graw Hill, Inc., New York, 2008, p. 1280 (ISBN: 978-0071470513)
4. R. Put, Y. Vander Heyden, *Anal. Chim. Acta* **602**(2007) 164 (https://dx.doi.org/10.1016/j.aca.2007.09.014)
5. K. D. Bartle,M. L Lee,S. A. Wise, *Chem. Soc. Rev*. **10** (1981) 113 (https://dx.doi.org/10.1039/CS9811000113).
6. *EPA Test Method, Polynuclear Aromatic Hydrocarbons- Method 610*, US Environmental Protection Agency, Environmental Monitoring and Support Laboratory, 1982 (https://www.epa.gov/sites/production/files/2015-10/documents/method_610_1984.pdf)
7. R. Kaliszan, *Chem. Rev*. **107** (2007) 3212 (https://dx.doi.org/10.1021/cr068412z)
8. N.Goudarzi, D. Shahsavani,F. Emadi-Gandaghi, M. Arab Chamjangali, *J. Chromatogr., A* **1333** (2014) 25 (https://dx.doi.org/10.1016/j.chroma.2014.01.048)

9. M. M. C. Ferreira, *Chemosphere* **44** (2001) 125 (https://dx.doi.org/10.1016/S0045-6535(00)00275-7)
10. F. A. L. Ribeiro, M. M. C. Ferreira, *J. Mol. Struct.: Theochem*. **663** (2003) 109 (https://dx.doi.org/10.1016/j.theochem.2003.08.107)
11. T. Moon, M. W. Chi, S. J. Park, C. N. Yoon, *J. Liq. Chromatogr. Rel. Technol*. **26** (2003) 2987 (https://dx.doi.org/10.1081/JLC-120025413)
12. K. A. Lippa, L. C. Sander, S. A. Wise, *Anal. Bioanal. Chem*. **378** (2004) 365 (https://dx.doi.org/10.1007/s00216-003-2419-7)
13. L. C. Sander, S. A.Wise, *J. Chromatogr. Libr*. **57**(1995) 337 (https://dx.doi.org/10.1016/S0301-4770(08)60622-3)
14. M. Popl, V. Dolansky,J.Mostecky, *J. Chromatogr*. **117** (1976) 117 (https://doi.org/10.1016/S0021-9673(00)81072-9)
15. National institute of standards and technology, https://webbook.nist.gov/chemistry/
16. A. D. Becke, *J. Chem. Phys*. **98** (1993) 5648 (https://dx.doi.org/10.1063/1.464913)
17. *TaleteSrl. Dragon for windows (Software for Molecular Descriptor Calculation), version 5.5*, Milano, 2007 (software available at: http://www.talete.mi.it)
18. R. W. Kennard, L. A. Stone, *Technometrics* **11** (1969) 137 (https://dx.doi.org/10.1080/00401706.1969.10490666)
19. P. Gramatica, N. Chirico, E. Papa, S. Cassani, S. Kovarich, *QSARINS, Software for the Development and validation of QSAR MLR Models* (available on request at http://www.qsar.it)
20. R. Todeschini, A. Maiocchi, V. Consonni, *Chemometr. Intell. Lab. Sys*. **46** (1999) 13 (https://dx.doi.org/10.1016/S0169-7439(98)00124-5)
21. P. Gramatica, *QSAR Comb. Sci*. **26** (2007) 694 (https://dx.doi.org/10.1002/qsar.200610151)
22. N. Chirico, P. Gramatica, *J. Chem. Inf. Model*. **51** (2011) 2320 (https://dx.doi.org/10.1021/ci200211n)
23. D. W. Osten. *J. Chemometr.* **2** (1998) 39 (https://dx.doi.org/10.1002/cem.1180020106)
24. N. Chirico, P. Gramatica, *J. Chem. Inf. Model*. **52** (2012) 2044 (https://dx.doi.org/10.1021/ci300084j)
25. R. Kiralj, M. M. C. Ferreira, *J. Braz. Chem. Soc*. **20** (2009) 770 (https://dx.doi.org/10.1590/S0103-50532009000400021)
26. P. Gramatica, *Mol. Inf*. **33** (2014) 311 (https://dx.doi.org/ 10.1002/minf.201400030)
27. G. Schüürmann, R. Ebert, J. Chen, B. Wang, R. Kühne. *J. Chem. Inf. Model*. **48** (2008) 2140 (https://doi.org/10.1021/ci800253u)
28. V. Consonni, D. Ballabio, R. Todeschini, *J. Chem. Inf. Model*. **49** (2009) 1669 (https://dx.doi.org/10.1021/ci900115y)
29. V. Consonni, D. Ballabio, R. Todeschini, *J. Chemometr.* **24** (2010) 194 (https://dx.doi.org/10.1002/cem.1290)
30. L. I. Lin, *Biometrics* **45** (1989) 255 (https://dx.doi.org/10.2307/2532051)
31. A. O. Aptula, N. G. Jeliazkova, T. W. Schultz, M. T. D. Cronin, *QSAR Comb. Sci*. **24** (2005) 385 (https://dx.doi.org/10.1002/qsar.200430909)
32. A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci*. **22** (2003) 69 (https://dx.doi.org/10.1002/qsar.200390007)
33. L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin,R. M. McDowell, P. Gramatica. *Environ. Health Perspect*. **111** (2003)1361(https://dx.doi.org/10.1289/ehp.5758)
34. S. Kherouf, N. Bouarra,A. Bouakkadia, D. Messadi**,** *J. Serb. Chem. Soc*. **84** (2019) 575 (https://dx.doi.org/10.2298/JSC180820016K)

35. S. Chatterjee, A. Hadi, B. Price, *Regression Analysis by Examples*, Wiley-VCH, New York, 2000, p. 368 (ISBN-13: 978-0471319467)
36. V. Consonni, R. Todeschini, M. Pavan. *J. Chem. Inf. Comput. Sci.* **42** (2002) 682 (https://dx.doi.org/10.1021/ci015504a)
37. R. Todeschini, P. Gramatica, *Quant. Struct.-Act. Relat.* **16** (1997) 113 (https://dx.doi.org/10.1002/qsar.19970160203)
38. R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH, New York, 2009, p.1257 (ISBN-13: 978-3527318520)
39. J. Galvez, R. Garcia-Domenech, J. V. de Julian-Ortiz, R. Soler, *J. Chem. Inf. Comput. Sci*. **35** (1995) 272(https://dx.doi.org/10.1021/ci00024a017)
40. M. Randic, G. Krilov, *Chem. Phys. Lett*. **272** (1997) 115 (https://dx.doi.org/10.1016/S0009-2614(97)00447-8).