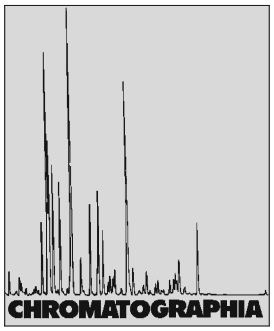
QSRR Models for Kováts' Retention Indices of a Variety of Volatile Organic Compounds on Polar and Apolar GC Stationary Phases Using Molecular Connectivity Indexes



2010, *72*, 893–903

Raouf Ghavami✉, Shadab Faham

Department of Chemistry, Faculty of Science, University of Kurdistan, P.O. Box 416, Sanandaj, Iran;
E-Mail: rghavami2000@yahoo.com; rghavami@uok.ac.ir

Received: 26 April 2010 / Revised: 13 July 2010 / Accepted: 2 August 2010
Online publication: 9 September 2010

Abstract

Quantitative structure-retention relationship (QSRR) approaches, based on molecular connectivity indices are useful to predict the gas chromatography of Kováts relative retention indices (GC-RRIs) of 132 volatile organic compounds (VOCs) on different 12 (4 apolar and 8 polar) stationary phases (C_{67}, C_{103}, C_{78}, C_{∞}, POH, TTF, MTF, PCL, PBR, TMO, PSH and PCN) at 130 °C. Full geometry optimization based on Austin model 1 semi-empirical molecular orbital method was carried out. The sets of 30 molecular descriptors were derived directly from the topological structures of the compounds from DRAGON program. By means of the final variable selection method, which is elimination selection stepwise regression algorithms, three optimal descriptors were selected to develop a QSRR model to predict the RRI of organic compounds on each stationary phase with a correlation coefficient between 0.9378 and 0.9673 and a leave-one-out cross-validation correlation coefficient between 0.9325 and 0.9653. The root mean squares errors over different 12 phases were within the range of 0.0333–0.0458. Furthermore, the accuracy of all developed models was confirmed using procedures of Y-randomization, external validation through an odd–even number and division of the entire dataset into training and test sets. A successful interpretation of the complex relationship between GC RRIs of VOCs and the chemical structures was achieved by QSRR. The three connectivity indexes in the models are also rationally interpreted, which indicated that all organic compounds' RRI was precisely represented by molecular connectivity indexes.

Keywords

Gas chromatography
Quantitative structure–retention relationship
Kováts retention indices
Connectivity indices
Elimination selection stepwise regression algorithms

Introduction

Retention is a phenomenon primarily dependent on the interactions between the solute and the stationary-phase molecules, which included directional force, induction force, dispersion force, hydrogen bond and so on. These forces can be related to the topological structures, geometric and electronic environments of the solute; therefore, it was possible to predict the solute retention from molecular parameters. Quantitative structure–retention relationships

Electronic supplementary material The online version of this article (doi:10.1365/s10337-010-1741-4) contains supplementary material, which is available to authorized users.

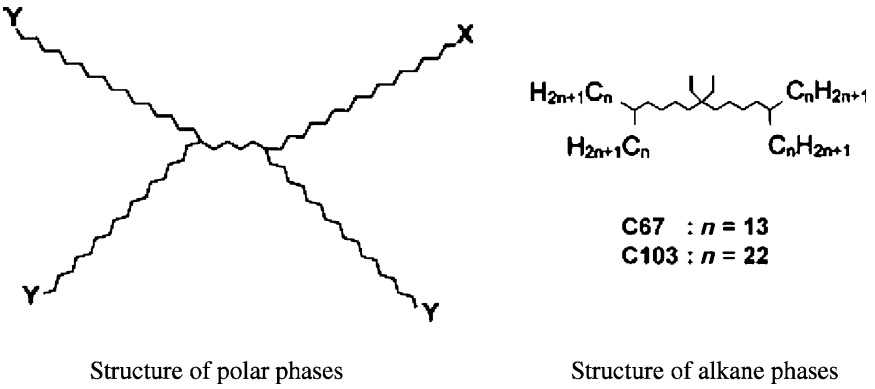


Fig. 1. Schematic representation of the stationary phases synthesized and used by Kováts and co-workers [32–35], and Laffort and co-workers [36]. The polar phases, with X and Y representing different functional groups (see also Table 1) are indicated on the left. The structures of apolar phases are indicated on the right

(QSRRs) have been demonstrated to be a powerful tool for the investigation of the chromatographic parameters. In QSRR approaches, the structural features of solutes encoded by non-empirical numerical descriptors and then the relationship between these descriptors and solute retention were studied [1–5]. The main steps in this method includes: data collection, molecular geometry optimization, molecular descriptors generation, descriptor selection, model development and finally model performance evaluation [5, 6].

The Kováts retention index is the most popular dependent variable in QSRR studies because of its reproducibility and accuracy [7]. The influence of the stationary phase polarity on the correlation and prediction of the retention of a variety of solutes by using different molecular descriptors has been the focus of several publications [8–14]. The development of QSRR studies is required for the suitable mathematical technique to make the models. In this regard, many available classical multivariate calibration techniques could be used to predict gas chromatographic retention index. These include multivariate linear regression (MLR), partial least-squares regression (PLS), and principal components regression (PCR) [13, 15–20]. Junkes et al. [13] applied new topological index, called semi-empirical topological index (I_{ET}) by MLR to predict the chromatographic retention of aliphatic ketones and aldehydes on stationary phases of different polarities. Farkas and Héberger [15] compared ridge regression (RR), PLS, pairwise correlation (PCM), forward selection (FS), and best subset variable selection (BSS) methods for prediction of retention indices for 44 aliphatic alcohols. Initially Farkas et al. calculated 109 descriptors; they then reduced these to 17 by use of principal-component analysis. The 17 descriptors were ranked in five different ways by use of RR, PLS, PCM, FS, and BSS. Models for prediction of RIs were then built with multiple linear regression using the best three and four descriptors selected by RR, PLS, PCM, FS, and BSS. If the standard derivation of the models increase and the F values decrease, the model is unsatisfactory. Héberger et al. [16–18] applied different linear multivariate techniques such as PCA, cluster, PLS and MLR in order to establish correlations between Kováts retention indices and different molecular descriptors for 35 aliphatic ketones and aldehydes, on different stationary phases at four temperatures. Tulasamma et al. [19] developed a unified QSRR with modified valence connectivity index, ${}^{n}\chi^{\mathrm{Vm}}$, based on the summation over inverse geometric mean terms by MLR to predict the retention properties of oxygen containing organic compounds on any new stationary phase. Good QSRR models were obtained to predict the retention index of 98 saturated esters on seven different polar stationary phases by MLR using the chemical descriptors proposed by Wang et al. [20].

Topological indices include valence and nonvalence molecular connectivity indices (MCIs); have been playing an important part in QSRR study for a long time. A large number of studies have demonstrated that many physicochemical and biological properties correlate with the connectivity index [21–24]. The main advantage of the graph theoretical approach to the prediction of properties is that it permits the interpretation of results in terms of structurally related concepts. In spit of that, the most important criticism of the so-called topological indices is concerned with their physical meaning [25].

Molecular connectivity is a method of molecular structure quantification based only on bonding and branching patterns rather than physical or chemical characteristics. Weighted counts of substructure fragments are incorporated into numerical indices and structural features (size, branching, unsaturation, heteroatoms content and cyclicity) are encoded. These indices are related to the number of atoms and how they are connected in a molecule. Only the carbon or heavy atoms are taken into consideration and the connectivity indices are derived from the hydrogen-suppressed graph of the molecule. Each atom is represented by a vertex in the graph, while the bonds become edges. Molecular valence connectivity index [26] uses the same invariant but modifies vertex degrees to account for heteroatoms by using the number of valence electrons in the corresponding atom. The details of their definition and the calculation method can be found elsewhere [21–24]. The general expression for the mth-order molecular valence connectivity index is as follows:

$$ {}^{m}\chi_k^v = \sum_{j=1}^{n_m} \left(\prod_{i=1}^{m+1} \delta_i^v \right)_j^{-1/2} \qquad (1) $$

where m is the order of the molecular valence connectivity index, k denotes a contiguous path type of fragment, which

Table 1. Name and structure details of the stationary phases synthesized and used by Kováts and co-workers[a] and Laffort and co-workers (experimentally[b] and interpolated or extrapolated[c])

Abbreviated name	Formula	Chemical name	Functional group	Structure of C_{16} branches for polar phases	
C_{67}^{b}	$C_{67}H_{136}$	19,19-Diethyl-14,24-ditridecylheptatricontane (67 carbon atoms)	Without		
C_{103}^{b}	$C_{103}H_{208}$	28,28-Diethyl-23,33-dicosylpentapentacontane (103 carbon atoms)	Without		
C_{78}^{c}	$C_{78}H_{158}$	19,24-Dioctadecyldotetracontane (78 carbon atoms)	Without		
C_{∞}^{c}		(infinite carbon atoms)	Without		
POH[a]	$C_{77}H_{156}O$	18,23-Dioctadecyl-1-untetracontanol	Primary alcohol	$X = CH_2OH$	$Y = CH_2CH_3$
TTF[a]	$C_{78}H_{146}F_{12}$	19,24-Bis-(18,18,18-trifluorooctadecyl)-1,1,1,42,42,42-hexafluorodotetracontane	Tetrakistrifluoromethyl	$X = CH_2CF_3$	$Y = CH_2CF_3$
MTF[a]	$C_{78}H_{155}F_3$	1,1,1-Trifluoro-19,24-dioctadecyldotetracontane	Monotrifluoromethyl	$X = CH_2CF_3$	$Y = CH_2CH_3$
PCL[a]	$C_{77}H_{155}Cl$	1-Chloro-18,23-dioctadecyluntetracontane	Primary chloro	$X = CH_2Cl$	$Y = CH_2CH_3$
PBR[a]	$C_{77}H_{115}Br$	1-Bromo-18,23-dioctadecyluntetracontane	Primary bromo	$X = CH_2Br$	$Y = CH_2CH_3$
TMO[a]	$C_{74}H_{150}O$	17,22, Bis-(16-methoxyhexadecyl)-1,38-dimethoxyoctatricontane	Tetramethoxy	$X = OCH_3$	$Y = OCH_3$
PSH[a]	$C_{77}H_{156}S$	18,23-Dioctadecyl-1-untetracontanethiol	Primary thiol	$X = CH_2SH$	$Y = CH_2CH_3$
PCN[a]	$C_{78}H_{155}N$	1-Cyano-18,23-dioctadecyluntetracontane	Primary cyano	$X = CH_2CN$	$Y = CH_2CH_3$

is divided into paths (P), cluster (C), path-clusters (PC), and chains (cycles) (CH). n_m is the number of the relevant paths, and δ_i^v is the atomic valence connectivity index and is defined as:

$$\delta_i^v = \frac{Z_i^v - h_i}{Z_i - Z_i^v - 1} \quad (2)$$

where Z_i is the number of electrons of atom i, Z_i^v the number of valence electrons, and h_i the number of hydrogen connected with atom i.

The study of the relationship between quantitative topological indices and chromatographic RI could be traced back to the 1970s. Many investigators have obtained good correlations between the experimental gas chromatographic RI and structural characteristics of molecules by using different topological indices as structural descriptors [27–31]. Randic [27] first studied the relationship between molecular connection index and gas chromatographic RI of alcohols. To predict chromatographic RIs of alkyl benzenes, Sutter et al. [28] chose six from 182 variables, which were from topological, geometrical, and electronic descriptors. Liu et al. [29] modeled the relationship between novel topological indices, polarizability effect index (PEI), odd–even index (OEI) and steric effect index (SV_{ij}) of 90 saturated esters and their GC RIs on seven GC columns (SE-30, OV-7, DC-710, OV-25, XE-60, OV-225 and Silar-5CP) by the MLR method. The average prediction errors over seven phases are within the range of 0.5–0.7%. Katritzky et al. [30] chose four variables from 129 topological descriptors and built a good model of 178 methyalkanes with squared correlation coefficient of 0.9585 and standard deviation of 5.8, and they particularly interpreted the relationship between variables selected in the model and molecular structures. Recently, QSRR equations have been established to model gas chromatographic retention data of alkyl pyridines on apolar and polar stationary phases by Tulasamma and Reddy [31].

The purpose of the present study was to investigate the relationship between gas chromatographic Kováts RRIs of 132 VOCs having C, H, O, N, and halogen atoms and their valence and non-valence topological molecular descriptors on different 12 (4 apolar and 8 polar) stationary phases at 130 °C using elimination selection stepwise best multiple linear regression (BMLR) analysis. Moreover, molecular descriptors were discussed to explore the influence of structural features on the values of RI. This study provided a simple and straightforward way to predict the RIs of VOCs from their structures and gave some insight into structural features related to the retention of the molecules. It has been found that the QSRR models for each stationary phase have not only high estimation qualities and high stabilities but also good predictive potentials.

Experimental and Methodology

Kováts Retention Indices

The QSRR treatment started with the assembly of the dataset. The chromatographic data used were obtained from scientific resources omitted hexamethyldisiloxane [32–36] and consisted of Kováts gas chromatographic RI of 132 solutes consisting of alkanes, alkens, ethers, amines, alcohols, alkylbenzenes, and alkylhalides on 12 diverse stationary phases with different polarities at 130 °C. The chemical structures of apolar and polar stationary phases are depicted in Fig. 1. The polar phases are all isochoric and isosteric with the C_{78} skeleton (Fig. 1 left side). All the polar phases have 78 heavy atoms (other than hydrogen), with the heavy atoms of the polar groups substituting for methylene or methyl groups in the hydrocarbon skeleton. The alkanes are more branched than the polar phases; hence they have a lower melting point. The alkane's family represented in Fig. 1 (right part), only concerns compounds with an odd number of carbon atoms. The C_{78} and C_{∞}

Table 2. Correlation matrix for the inter-correlation of various connectivity indices of 132 solutes

	$^{1}X_{v}$	$^{3}X_{v}$	$^{4}X_{v}$	$^{1}X_{sol}$	$^{2}X_{sol}$	$^{3}X_{sol}$	$^{4}X_{sol}$
$^{1}X_{v}$	1.0000						
$^{3}X_{v}$	0.5944	1.0000					
$^{4}X_{v}$	0.3565	0.2819	1.0000				
$^{1}X_{sol}$	0.6904	0.4127	0.6249	1.0000			
$^{2}X_{sol}$	0.7254	0.5196	0.3168	0.6241	1.0000		
$^{3}X_{sol}$	0.3942	0.6432	0.6716	0.7232	0.4915	1.0000	
$^{4}X_{sol}$	0.1834	0.1801	0.8298	0.6576	0.3386	0.8074	1.0000

Table 3. The value of mean effect for each descriptor of QSRR models

St. Ph.	Mean effect		
	$^{1}X_{sol}$	$^{1}X_{v}$	$^{4}X_{sol}$
C_{67}	0.400	−0.075	0.038
C_{103}	0.396	−0.076	0.040
C_{78}	0.399	−0.075	0.039
C_{∞}	0.387	−0.077	0.043
POH	0.412	−0.092	0.040
TTF	0.433	−0.098	0.037
MTF	0.408	−0.081	0.039
PCL	0.404	−0.081	0.039
PBR	0.404	−0.081	0.040
TMO	0.432	−0.100	0.038
PSH	0.403	−0.081	0.040
PCN	0.424	−0.096	0.039

alkanes have been derived from C_{67} and C_{103} (Eqs. 2, 3, Ref. [36]). Its detailed experimental conditions were listed in Table 1. Because of the large scale of RIs of dataset, organic compounds (max RI 1400 for tetradecane on all stationary phases), the RRIs of these compounds were recalculated by dividing the RI of a reference compound such as tetradecane as the internal standards. A complete list of the names and corresponding experimental Kováts RRI values of VOCs on each stationary phase has been categorized in Table S1 of the supplementary data.

Computer Hardware and Software

All calculations were run on a Pentium IV personal computer (CPU at 2.6 MB) under Windows XP operating system. The ISIS/Draw version 2.3 software was used for drawing the molecular structures [37]. Molecular modeling and geometry optimization were employed by HyperChem (version 7.1, HyperCube) [38]. Dragon software [39] was employed for calculation of theoretical topological connectivity indices. SPSS software (version 13.0, SPSS) http://www.spss.com/ was used for stepwise MLR analysis and other calculations were performed in the MATLAB (version 7.0, Math Works) environment.

Descriptor Generation and Models Developing

To obtain QSRR models, solutes must be represented using molecular descriptors. Descriptors are generated solely from the molecular structures and aimed to numerically encode meaningful features of each molecule. The calculation process of the molecular descriptors is described as below: all the two-dimensional structures of the molecules were drawn using ISIS/Draw 2.3 program [37]. Then the 3D geometry structures of the molecules were pre-optimized using MM+ molecular mechanics force filed and precisely optimized with semi-empirical Austin Model 1 (AM1) method implemented in HyperChem software package (HyperCube, version 7.1) [38]. All calculations were carried out at restricted Hartree–Fock level with no configuration interaction. The molecular structures were optimized using the Polak-Ribiere algorithm until the root mean square gradient was 0.01 kcal Å mol^{-1} (200 K, gas phase). In order to prevent the structures locating at local minima, geometry optimization was run many times with different stating points for each molecule. The output files exported from HyperChem software were transferred into software Dragon, developed by Todeschini et al. [39], to calculate Randic topological indices as the mostly used topological indices in the literature Randic indices [23, 40]. We computed 30 different MCIs by the Dragon software including a set of 12 connectivity and average connectivity index (6 connectivity indices $^{0}X - {}^{5}X$, 6 average connectivity indices $^{0}X_{A} - {}^{5}X_{A}$), 12 valence and average valence connectivity index (6 valence connectivity indices $^{0}X_{v} - {}^{5}X_{v}$, 6 average valence connectivity indices $^{1}X_{Av} - {}^{5}X_{Av}$), and 6 solvation connectivity index ($^{0}X_{sol} - {}^{5}X_{sol}$).

The calculated molecular descriptors were collected in a data matrix (**D**) whose number of rows and columns were the number of molecules and descriptors, respectively. At the beginning, in order to minimize the information overlap in descriptors and to reduce the number of descriptors required in regression equation, the concept of non-redundant descriptors (NRD) [41] was used in our study. That is, when two descriptors are correlated by a linear correlation coefficient value >0.85, both descriptors are correlated with the dependent variables, the better correlation is used for the actual analysis, leaving out the descriptors showing a lower correlation. This objective-based feature selection left reduced and predictive descriptors for the studied compounds. By using these criteria, 23 out of 30 original descriptors were eliminated. These descriptors are not correlating with each other as revealed from the correlation matrix presented in Table 2. In addition, these descriptors can give

some information on the affecting degree for RI of different descriptors and well understand the correlation between the experimental and calculated values. Therefore, ${}^{1}X_{v}$, ${}^{3}X_{v}$, ${}^{4}X_{v}$, ${}^{1}X_{sol}$, ${}^{2}X_{sol}$, ${}^{3}X_{sol}$ and ${}^{4}X_{sol}$ set of descriptors has been used in the QSRR model for all stationary phases.

In order to select the subset of descriptors that best explain Kováts RI for each stationary phase, we have used elimination selection stepwise regression (ES-SWR) algorithm to select the most appropriate [42–44]. This method can be regarded as a combination of the forward and backward approaches. Stepwise model-building techniques for regression designs with a single dependent variable involve identifying an initial model, repeatedly altering the model from the previous step by adding (forward stepwise) or removing (back stepwise) a predictor variable and terminating the search when stepping does not further improve the model. The forward stepwise method employs a combination of the forward entry of independent variables and backward removal of insignificant variables. The best single predictor, which is the most significant variable, was used for the initial linear regression step. Next, descriptors were added one at a time, always adding the one that most improved the fit, until the fit was not significantly improved. Once all the significant variables were determined, the regression equation was constructed. The number of variables retained in the model is based on the levels of significance assumed for inclusion and exclusion of variables from the model for each stationary phase column.

An MLR model assumes that there is a linear relationship between the molecular descriptors of a compound, which is usually expressed as a feature vector **x** (with each descriptor as a component of this vector), and its target property, y. An MLR model can be described using the following equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon \quad (3)$$

where $\{X_1,\ldots, X_k\}$ are molecular descriptors, β_0 is the regression model constant, β_1 to β_k are the coefficients corresponding to the descriptors X_1 to X_k and y is dependent variable. The values for β_0 to β_k are chosen by minimizing the sum of squares of the vertical distances of the points from the hyperplane so as to give the best prediction of y from X. Regression coefficients represent the independent contributions of each calculated molecular descriptor. In matrix notation, we will write the MLR model is defined in Eq. (4) as:

$$y = \mathbf{X}b + e \quad (4)$$

where $\mathbf{X}_{(n\times k)}$ is of full column rank, including a column of 1 s for the intercept if the intercept is included in the mean function y. We will further assume that we have selected a parameterization for the y so that **X** has full column rank, meaning that the inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ exists; this is not an important limitation on regression models because we can always delete terms from the y, or equivalently delete columns from $\boldsymbol{X}$, until we have full rank. The $k \times 1$ vector b is the unknown parameter vector. The vector e consists of unobservable errors that we assume are equally variable and uncorrelated, unless stated otherwise. In appropriate Eq. (4), the least squares solution estimate b by $\hat{b} = (X^TX)^{-1}X^Ty$, and the fitted values y corresponding to the experimental RIs are then given by:

$$\hat{y} = X\hat{b} = X\left(X^TX\right)^{-1}X^Ty = Hy \quad (5)$$

where **H** is the $n \times n$ matrix defined by $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, which is called the hat matrix because it transforms the vector of experimental responses y into the vector of fitted responses $\hat{y}$. The vector of residuals $\hat{e}$ is defined by $\hat{e} = y - \hat{y} = y - X\hat{b} \quad = y - X(X^TX)^{-1}X^Ty = (I - H)y$.

The advantages of MLR are that it is simple to use and the derived models are easy to interpret. The sign of the coefficients β_0 to β_k shows whether the molecular descriptors contribute positively or negatively to the target property and their magnitudes indicates the relative importance of the descriptors to the target property. However, the molecular descriptors should be mathematically independent (orthogonal) of one another and the number of compounds in the training set should exceed the number of molecular descriptors by at least a factor of 5 [45]. Studies have shown that collinear descriptors may result in the coefficients β_0 to β_k being larger than expected or have the wrong sign [46]. To examine multicollinearity, the variance inflation factor (VIF) was calculated for each variable in the regressions, which is defined as:

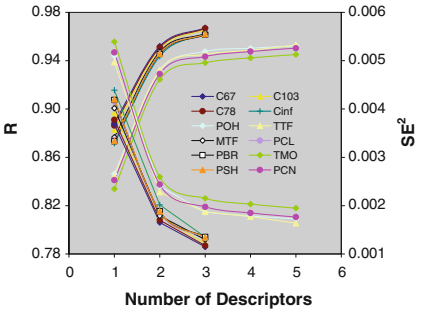


Fig. 2. The influences of the number of descriptors on the correlation coefficient (R) and the standard deviation (SE^2)

$$\mathrm{VIF} = \frac{1}{1 - R_j^2} \quad (6)$$

where R_j^2 is the squared correlation coefficient between the jth coefficient regressed against all the other descriptors in the model [47]. The VIF is uniform and equal to 1.00 if there is no linear correlation between a given variable and rest of the variables in the regression equations. Higher values of VIF indicate a more serious multicollinearity problem (popular cut off value is 10). In addition, for inspection of the relative importance and contribution of each descriptors in the QSRR models, the value of mean effect (MF) was calculated for each descriptors by the following equation and it is shown in Table 3:

$$MF_j = \frac{\beta_j \sum_{i=1}^{n} d_{ij}}{\sum_{j=1}^{m} \beta_j \sum_{i=1}^{n} d_{ij}} \quad (7)$$

where MF_j is the mean effect for considered descriptor j, β_j is the coefficient of descriptor j, d_{ij} denotes the value of descriptor j of molecule i, m is the number of descriptors in the model and n is the number of molecules in the data sets. The value and sign of the mean effect shows the relative contribution and direction of influence of each descriptor on the RI.

Table 4. QSRR models and statistical parameters of GLC-RRI values versus tetradecane for the total sets ($n = 132$) of solutes on 12 column stationary phases

St. Ph.	X_0 (± SE)	$^1X_{sol}$ (± SE)	^1X_v (± SE)	$^4X_{sol}$ (± SE)	R	RMS	REP	F	R_{cv}	RMS_{cv}	R_{max}
C_{67}	0.0969 (±0.0146)	0.1438 (±0.0071)	−0.0294 (±0.0031)	0.0495 (±0.0065)	0.9673	0.0333	5.9569	619.6	0.9653	0.0343	0.0480
C_{103}	0.1004 (±0.0150)	0.1433 (±0.0073)	−0.0300 (±0.0032)	0.0520 (±0.0067)	0.9658	0.0343	6.1027	592.4	0.9637	0.0343	0.0282
C_{78}	0.0982 (±0.0148)	0.1436 (±0.0071)	−0.0297 (±0.0032)	0.0505 (±0.0066)	0.9667	0.0337	6.0145	608.7	0.9647	0.0347	0.0513
C_∞	0.1069 (±0.01158)	0.1423 (±0.0077)	−0.0311 (±0.0034)	0.0567 (±0.0071)	0.9629	0.0362	6.3940	543.4	0.9604	0.0374	0.0540
POH	0.1294 (±0.0186)	0.1453 (±0.0090)	−0.0354 (±0.0040)	0.0511 (±0.0083)	0.9475	0.0425	7.3326	375.0	0.9432	0.0442	0.0453
TTF	0.1345 (±0.0187)	0.1476 (±0.0090)	−0.0368 (±0.0040)	0.0447 (±0.0083)	0.9451	0.0426	7.3211	357.3	0.9402	0.0445	0.0538
MTF	0.1104 (±0.0155)	0.1443 (±0.0075)	−0.0316 (±0.0033)	0.0489 (±0.0069)	0.9624	0.0353	6.2303	536.7	0.9601	0.0361	0.0551
PCL	0.1102 (±0.0158)	0.1443 (±0.0076)	−0.0317 (±0.0034)	0.0503 (±0.0070)	0.9618	0.0360	6.3427	527.0	0.9593	0.0372	0.0503
PBR	0.1107 (±0.0159)	0.1442 (±0.0077)	−0.0318 (±0.0034)	0.0507 (±0.0071)	0.9615	0.0362	6.3721	522.5	0.9590	0.0374	0.0415
TMO	0.1397 (±0.0201)	0.1476 (±0.0097)	−0.0375 (±0.0043)	0.0460 (±0.0089)	0.9378	0.0458	7.8109	311.6	0.9325	0.0477	0.0473
PSH	0.1092 (±0.0158)	0.1444 (±0.0076)	−0.0319 (±0.0034)	0.0511 (±0.0070)	0.9620	0.0361	6.3571	529.1	0.9595	0.0372	0.0487
PCN	0.1344 (±0.0192)	0.1466 (±0.0093)	−0.0365 (±0.0041)	0.0477 (±0.0085)	0.9434	0.0438	7.5112	345.3	0.9385	0.0456	0.0469

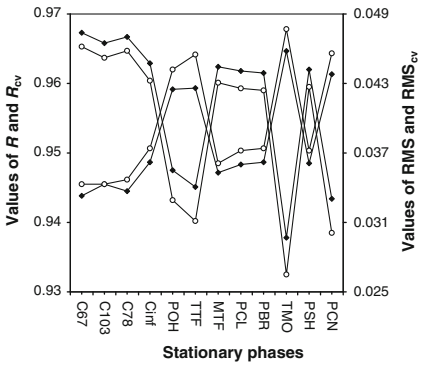


Fig. 3. Plot of the root mean square error (♦ RMS and ○ RMS_{cv}) and correlation coefficient (♦ R and ○ R_{cv}) for the resulted 12 QSRR models versus GC stationary phases

Results and Discussion

Selecting seven molecular theoretical descriptors including 3 valence connectivity index and 4 solvation connectivity index and the GC-RRI values of 132 solutes on 12 column stationary phases as a dependent one, the best combination of variables is selected by the ES-SWR algorithm to build the relationship between the molecular structure and RRIs using BMLR analysis. The influence of the best number of variables for the subset of all solutes was selected by ES-SWR on the calibration correlation coefficient (R) and square standard error (SE^2) for each stationary phase were included in Fig. 2. As can be seen in Fig. 2 for the former 12 GC stationary phases, the R values increases gradually with increasing the number of variables until reaches a plateau while SE^2 declines until it drops to a lower limit value. We used the best correlation equation with three optimal variables including solvation connectivity index of order 1 ($^1X_{sol}$), valence connectivity index of order one (^1X_v) and a solvation connectivity index of order 4 ($^4X_{sol}$) for the analysis of all QSRR models. The descriptors that appear in the BMLR equations for the 12 different stationary phases are identical. A complete list of solutes and the calculated values of the molecular connectivity indexes appearing in the QSRR models are summarized in Table S2 of the supplementary data. These descriptors are related to shape, and the degree of branching of the molecules. This indicates that dispersion interactions and the extent of branching of the molecules affected the retention behavior of organic compounds on the polar and apolar stationary phase columns. The solvation connectivity index ($^1X_{sol}$) shows an average mean effect of 0.409 for all columns, which is the largest among the descriptors appearing in the QSRR models. This descriptor can be considered as entropy of solvation and somehow indicates the dispersion interactions occurring in the solution. $^1X_{sol}$ also is a measure of branching of the molecules. The large contributions of this descriptor in the retention behavior of organic molecules is in agreement with the contribution that one would expect for the interaction of nonpolar stationary phases such as C_{67}, C_{103}, C_∞, and C_{78} with the nonpolar organic molecules. The coefficient of correlation between the GC-RRIs of 132 solutes and $^1X_{sol}$ index was $0.8339 > 0.8411 > 0.8442 > 0.8463 > 0.8716 > 0.8733 > 0.8736 > 0.8743 > 0.8766 > 0.8828 > 0.8863 > 0.8886$, respectively, for the TMO, PCN, TTF, POH, C_∞, PBR, PSH, PCL, MTF, C_{103}, C_{78} and C_{67} stationary phases. The presence of ^1X_v as a connectivity index with the average mean effect −0.084 in all models indicates that the retention indices depend on the presence and the position of the heteroatoms in the organic molecules. The descriptor shows the negative effect on the retention indices, consequently, the RI decreases with increasing of ^1X_v. The solvation connectivity index $^4X_{sol}$ with the average mean effect 0.039 in all models is a measure of branching of the molecules. Thus, the emergence of the $^4X_{sol}$ in all QSRR models reflects the influence of the degree of branching on the values of RIs. The positive sign of the corresponding coefficient indicates that the higher ramifications in the solutes, the smaller is $^4X_{sol}$ and therefore, the bigger is the molecule. This descriptor could also be considered as entropy of solvation and somehow indicates the influence of the dispersion interaction occurring in the stationary phases on the values of RIs. The regression coefficients and the statistical results of the resulted QSRR models for each stationary phase are given in Table 4. The value after the symbol "±" in the parenthesis is the standard deviation related to the regression coefficient. The results in Table 4 indicate that GC-RRIs of 132 solutes on all column stationary phases are strongly dependent on the first order solvation

connectivity index ($^1X_{sol}$) with the maximum mean effect (Table 3) for each stationary phase. The qualities of the models derived from various subsets are evaluated using some statistics, such as the correlation coefficient (R), root mean-square error (RMS), relative error of prediction (REP) and the Fisher's criterion at the 95% level probability (F) are included in Table 4, also the best fitted equations and results of the root mean square error (RMSE) along with correlation coefficient (R) are plotted in Fig. 3. From Table 4 and Fig. 3, the predicted correlation coefficients over 0.9378, the overall F values higher than 310, and the RMS and REP below 0.0458 and 7.8109, respectively, indicated that the BMLR models have good statistical qualities with low prediction error and demonstrated an excellent predictive power of the obtained QSRR models for all stationary phases. From Table 4, the VIF values are lower than 10 (4.32, 2.54, and 2.34) for three variables $^1X_{sol}$, ^1X_v, and $^4X_{sol}$, respectively, indicating that the QSRR models could have some multicollinearity but it was not serious. According to the statement of Mihalic and Trinajstic [48], the models we have constructed represent good QSRR models judging from the statistics.

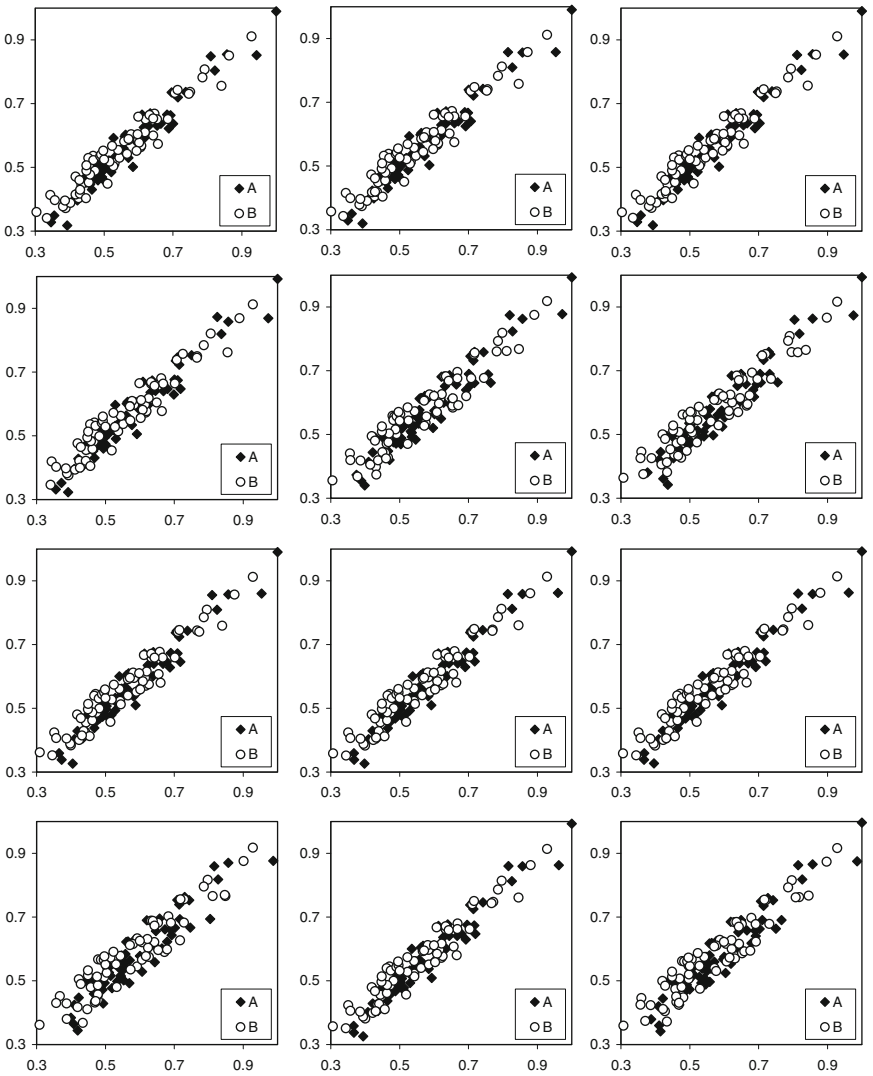


Fig. 4. Plots of the RRIs estimated for the odd set (♦ A) and even set (○ B) samples by holdout model versus that observed RRIs experimentally for all stationary phases of 132 solutes

Model Prediction-Validation

Model validation is a critical component of QSRR development. A number of procedures have been established to determine the quality of QSRR models. Therefore, a leave-one-out cross-validation (LOO-CV), Y-randomization, and external validation (EV) procedures through an odd–even number and division of the entire dataset into training and test sets are used to validate the predictive ability and check the statistical significance of the developed 12 QSRR models.

Cross-Validation

The most popular validation method is cross-validation (CV), known as jackknifing or leave-one-out (LOO). This method systematically removes one data point at a time from the training set, and constructs a model with the reduced dataset. Subsequently, the model is used to predict the data point that has been left out. By repeating the procedure for the entire dataset, a complete set of predicted properties and cross-validated statistics can be obtained. It has been argued that the LOO procedure often overestimates the predictivity of the model and that, subsequently, the QSRR models are overoptimistic [49]. For cross-validated statistics, it has been suggested that prediction residual error sum of squares (PRESS), cross-validated square correlation coefficient (R^2_{cv}) and root mean square error in cross-validation (RMS_{cv}) are good estimates of the real prediction error of a model:

$$\mathrm{PRESS} = \sum_{i=1}^{N} \left(y_{\mathrm{pred},i} - y_{\mathrm{obs},i}\right)^2 \tag{8}$$

$$R^2_{cv} = 1 - \frac{\sum_{i=1}^{N} \left(y_{\mathrm{pred},i} - y_{\mathrm{obs},i}\right)^2}{\sum_{i=1}^{N} \left(y_{\mathrm{obs},i} - \bar{y}_{\mathrm{obs}}\right)^2} = 1 - \frac{\mathrm{PRESS}}{\sum_{i=1}^{N} \left(y_{\mathrm{obs},i} - \bar{y}_{\mathrm{obs}}\right)^2} \tag{9}$$

$$\mathrm{RMS}_{cv} = \sqrt{\frac{\sum_{i=1}^{N} \left(y_{\mathrm{pred},i} - y_{\mathrm{obs},i}\right)^2}{N}} \tag{10}$$

where N is the number of training patterns, $y_{obs,i}$ and $y_{pred,i}$ are the experimental, and predicted RRIs of the left-out compound i, respectively and $\bar{y}_{obs}$ is the average experimental RRI of left-in

Table 5. Statistical parameters of the over-fitting and predictive ability of the models

St. Ph.	Odd samples				Even samples			
	$RMSE_{RS}$	R_{RS}	$RMSE_{HO}$	R_{HO}	$RMSE_{RS}$	R_{RS}	$RMSE_{HO}$	R_{HO}
C_{67}	0.0310	0.9703	0.0319	0.9701	0.0351	0.9647	0.0360	0.9646
C_{103}	0.0318	0.9692	0.0327	0.9690	0.0362	0.9630	0.0371	0.9628
C_{78}	0.0313	0.9698	0.0322	0.9696	0.0355	0.9641	0.0364	0.9639
C_{∞}	0.0333	0.9670	0.0343	0.9666	0.0384	0.9595	0.0393	0.9592
POH	0.0392	0.9525	0.0401	0.9523	0.0452	0.9432	0.0461	0.9428
TTF	0.0393	0.9495	0.0402	0.9494	0.0452	0.9416	0.0461	0.9415
MTF	0.0327	0.9662	0.0336	0.9660	0.0376	0.9594	0.0385	0.9593
PCL	0.0330	0.9662	0.0339	0.9660	0.0384	0.9581	0.0393	0.9580
PBR	0.0331	0.9660	0.0340	0.9658	0.0387	0.9577	0.0395	0.9575
TMO	0.0421	0.9433	0.0433	0.9427	0.0487	0.9334	0.0498	0.9327
PSH	0.0330	0.9664	0.0340	0.9661	0.0386	0.9582	0.0393	0.9580
PCN	0.0400	0.9493	0.0409	0.9489	0.0468	0.9384	0.0478	0.9379

Table 6. QSRR models of GC-RRI values versus tetradecane for the training sets ($n = 82$) of solutes on 12 GC stationary phases

St. Ph.	X_0 (± SE)	$^1X_{sol}$ (± SE)	^1X_v (± SE)	$^4X_{sol}$ (± SE)
C_{67}	0.0927 (±0.0193)	0.1452 (±0.0101)	−0.0311 (±0.0048)	0.0552 (±0.0094)
C_{103}	0.0972 (±0.0199)	0.1445 (±0.0104)	−0.0319 (±0.0050)	0.0583 (±0.0097)
C78	0.0945 (±0.0195)	0.1449 (±0.0103)	−0.0314 (±0.0049)	0.0564 (±0.0095)
C_{∞}	0.1054 (±0.0209)	0.1431 (±0.0110)	−0.0333 (±0.0053)	0.0642 (±0.0102)
POH	0.1290 (±0.0244)	0.1459 (±0.0128)	−0.0382 (±0.0061)	0.0596 (±0.0119)
TTF	0.1363 (±0.0242)	0.1473 (±0.0127)	−0.0398 (±0.0061)	0.0549 (±0.0118)
MTF	0.1085 (±0.0204)	0.1449 (±0.0107)	−0.0335 (±0.0051)	0.0562 (±0.0100)
PCL	0.1085 (±0.0208)	0.1450 (±0.0109)	−0.0339 (±0.0052)	0.0576 (±0.0102)
PBR	0.1093 (±0.0209)	0.1449 (±0.0110)	−0.0340 (±0.0053)	0.0581 (±0.0102)
TMO	0.1386 (±0.0264)	0.1494 (±0.0139)	−0.0414 (±0.0066)	0.0544 (±0.0129)
PSH	0.1072 (±0.0209)	0.1453 (±0.0110)	−0.0341 (±0.0052)	0.0582 (±0.0102)
PCN	0.1351 (±0.0252)	0.1474 (±0.0132)	−0.0399 (±0.0063)	0.0568 (±0.0123)

compounds different from *i*. Values of R^2_{cv} can range from 1 to <0. A value of one indicates a perfect prediction, and a value of 0 means that the QSRR derived has no modeling power. Negative values arise from a situation where the derived QSRR is a poorer description of data than no model at all. The R^2_{cv} values can be considered as a measure of the predictive power of a model: whereas R^2 can always be increased artificially by adding more parameters, R^2_{cv} decreases if a model is over parameterized [50], and is therefore a more meaningful summary statistic for predictive models. The correlation coefficients (R_{cv}) and RMS_{cv} for each subset are presented in Table 4 and the resulted values are plotted in Fig. 3. The cross-validation results shows that the R_{cv} are higher than 0.9325 and RMS_{cv} lower than 0.0477 for all GC stationary phases, respectively. Furthermore, in all cases, the cross-validated R_{cv} values are very close to the corresponding R values and the cross-validated RMS_{cv} values are only slightly larger than the corresponding RMS values. Clearly, the cross-validation demonstrates the final models to be statistically significant.

This method is not a very rigorous model predictivity test and suffers from two other major deficiencies: the time to carry out the cross-validation increases as the square of the size of training set; the method produces n final models (each corresponding to one of the training set molecules being left out) and it is not clear which is the 'best' model. To further check the prediction ability of the resulting QSRR models two better methods are applied here, one by Hawkins [51] namely as odd–even external validation and the other better method is to remove a percentage of the training set into a prediction set [50, 52].

Odd–Even External Validation

To validate and develop a credible QSRR model, it is not enough to build a model for the whole dataset. So, the 132 dataset solutes for all stationary phases were sorted in the ascending order of GC Kováts RRI values and then divided into two sets namely "odd set" and "even set" RRIs [50, 52]. This way of splitting ensures that the distribution of RRI values of the two subsets were very similar. The QSRR models were fitted to the odd set and even set samples separately and the resulted fitness were assessed by applying QSRR models to both samples. To compare the estimation abilities of the models, two statistical parameters namely root mean squares error (RMSE) and R, were calculated. The same dataset (i.e., 'calibration set') that was already used to fit the models was employed to determine resubstitution parameters, i.e. $RMSE_{RS}$ and R_{RS}, also to determine holdout parameters, i.e. $RMSE_{HO}$ and R_{HO} for the other dataset, which was not involved in the fitting. The resubstitution statistical parameters of the samples base their predictions on the regression fitted to those samples and this is while the holdout

statistical parameters base their predictions on the regression fitted to the other samples. The plots of RRIs estimated by odd- and even-set QSRR models (holdout prediction) versus the RRIs observed experimentally are given in Fig. 4, also Table 5 summarizes these statistical parameters achieved by this approach. As can be seen, in the odd and even-set samples, the resubstitution and holdout RMSE are very similar, indicating that the same sample and other sample predictions are equally precise for all stationary phases.

Y-Randomization Test

Another procedure that is easy to perform is a randomization test called *Y*-randomization (randomization of response, i.e., in our case RRI). In this method for each column stationary phase, the output RRIs values of the compounds are shuffled randomly, and the resulting dataset is examined by the QSRR method against real (unscrambled) input descriptors to determine the correlation and predictivity of the resulting "model" [53–56]. The whole procedure is repeated on many different scrambled datasets. The rationale behind this test is that the significance of the real QSRR model would be suspected if there is a strong correlation between the selected descriptors and the randomized response variables. The randomization was repeated ten times. If the statistical qualities of these models are much lower than the original model, it can be considered that the model is reasonable and had not been obtained by chance. The results are shown in last column of Table 4. Very low level of R_{max} (in the interval of 0.0282 for C_{103} and 0.0551 for MTF) indicates good results in our original models and is not due to a chance correlation or structural dependency of the training set for each stationary phase of the GC column.

Calibration and Prediction Sets

In this investigation, for further testing the predictive ability of the models for the external compounds without the models, part of the congeners are picked up from 132 solutes to construct a training set which is used to develop a prediction model and then predict the values of Kováts RRIs in the remaining congeners. How to pick up the compounds in the training set is very important for developing of the predictive QSRR models. In this case, before each training run, all datasets were split randomly into two separate sub-matrices: the training set matrix and external testing set matrix. Out of 132 organic compounds, 82 solutes (62%) were used for the training set and 50 congeners (38%) were used as external validation. The solutes constituting the training and testing sets are clearly presented in Table S1 of the supplementary data. Moreover, the same divisions were repeated with corresponding RRIs values. The test examples are marked as bold font and training set was also used to obtain the best fit equation of MLR with three molecular descriptors. Furthermore, the testing set was used to monitor overfitting the MLR models. The resulted MLR models for training set congeners were the same as those obtained for the entire set of all solute in each subset subject to use descriptors of all solute

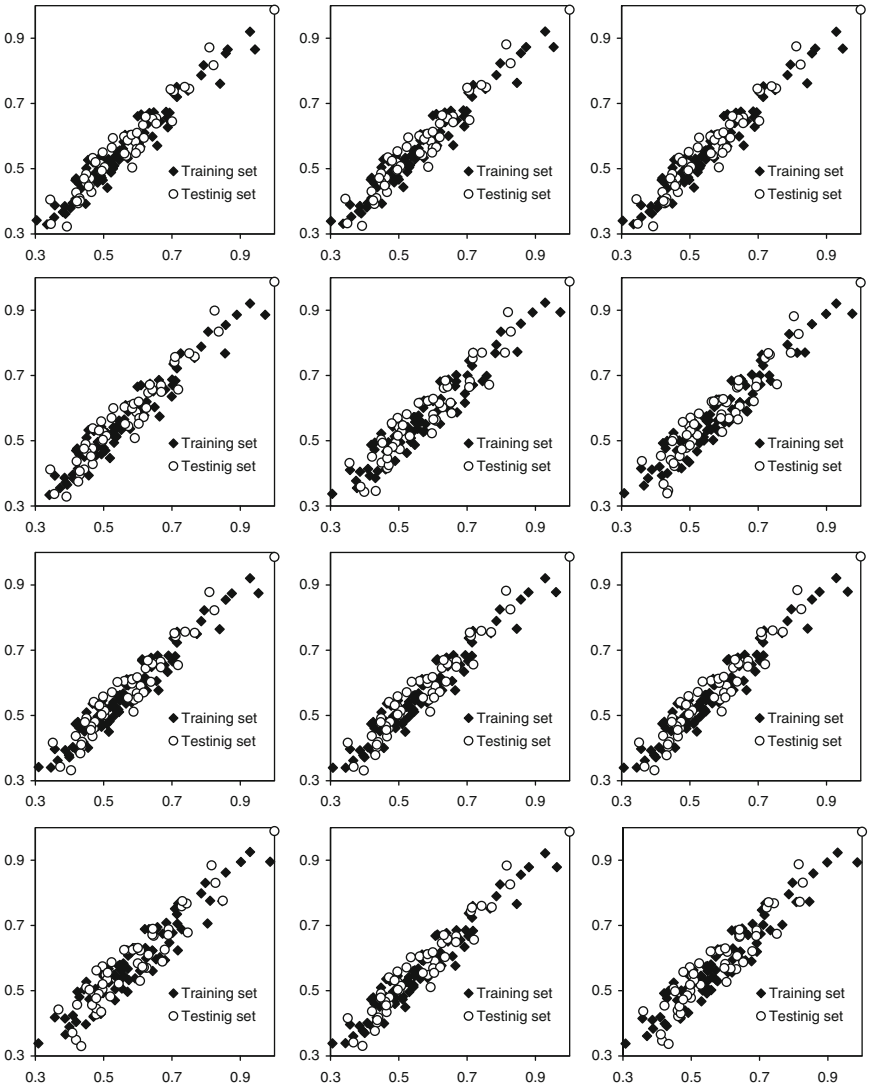


Fig. 5. Plots of the RRIs estimated by the QSPR models in Table 7 versus that observed for 82 training set solutes (♦) and 50 testing set ones (○) for all stationary phases

Table 7. Statistical parameters of the QSRR models obtained using different molecular descriptors

St. Ph.	Training set						Testing set		
	RMS	REP	F	R	R_{cv}	RMS_{cv}	RMS	REP	R
C_{67}	0.0331	5.8983	395.3	0.9687	0.9654	0.0347	0.0346	6.2093	0.9659
C_{103}	0.0340	6.0420	378.2	0.9673	0.9637	0.0358	0.0357	6.3831	0.9643
C_{78}	0.0335	5.956	388.5	0.9681	0.9647	0.0352	0.0350	6.2740	0.9652
C_{∞}	0.0359	6.3248	348.0	0.9646	0.9603	0.0380	0.0380	6.7379	0.9609
POH	0.0417	7.1883	245.5	0.9509	0.9419	0.0454	0.0451	7.7950	0.9418
TTF	0.0414	7.0971	240.2	0.9499	0.9380	0.0461	0.0462	7.9746	0.9366
MTF	0.0349	6.1432	348.8	0.9647	0.9600	0.0371	0.0375	6.6375	0.9594
PCL	0.0357	6.2697	337.9	0.9636	0.9588	0.0380	0.0377	6.6680	0.9594
PBR	0.0359	6.3019	334.7	0.9633	0.9583	0.0382	0.0379	6.7013	0.9591
TMO	0.0452	7.6901	203.5	0.9416	0.9405	0.0496	0.0484	8.2826	0.9310
PSH	0.0358	6.2952	338.2	0.9636	0.9588	0.0381	0.0377	6.6629	0.9598
PCN	0.0431	7.3842	225.6	0.9469	0.9359	0.0473	0.0465	8.0051	0.9372

models supporting sufficient ability for the prediction set of 50 solutes. The resulting regressions equations of the training set for individual GC column stationary phases with the optimal three molecular descriptors are indexed in Table 6, and results obtained are plotted in Fig. 5. Statistical parameters for the best-fitted models are also presented in Table 7. The correlation coefficients (R) of the obtained models are >0.94 for all the stationary phases and the highest one is 0.9687 for stationary phase C_{67}. The RMS and relative error prediction (REP) of estimation ranged from 0.0331, 5.8983 of C_{67} stationary phases to 0.0452, 7.6901 of TMO stationary phase, respectively, also the F statistic values are >203.5. The LOO-CV method was used to examine the stability of QSRR models, and the values of R_{cv} and RMS_{cv} for the models were above 0.9359 and in the range of 0.0347 for C_{67} stationary phase and 0.0496 for TMO stationary phase, respectively. The predicted Kováts RRIs versus the observed Kováts RRIs of the 82 solute training sets are plotted in Fig. 5. As shown in Table 7 and Fig. 5, the QSRR statistical results exhibit good estimation capacity and stability for internal training set solute samples to individual stationary phases. High predictive ability of QSRR models for external examples is another criterion of a good QSRR model. The predicted RRIs of 50 solutes in the external testing set by the models in Table 6 are also demonstrated in Fig. 5 versus the observed RRIs of 12 GC stationary phases. For all 12 GC stationary phases, the regression of the observed and predicted RRIs had a high agreement with the diagonal of each chart. The predicted correlation coefficients (R) over 0.9310 and the RMS REP below 0.0346 and 8.2626, respectively, demonstrated an excellent predictive power of the obtained QSRR models.

Conclusion

In this study, a novel QSRR tool of BMLR is performed to describe the GC Kováts RRI values of 132 solutes on 12 polar and apolar stationary phase columns based on connectivity molecular descriptors. MLR analysis produced more predictive, informative and significantly improved QSRR models. The use of connectivity indices molecular descriptors revealed to be a completely successful strategy. The effectiveness of the stepwise forward and backward elimination algorithm is demonstrated by the selection of the best set of molecular descriptors. All QSRR models provide a reasonably good calibrated correlation coefficient. The validation and predictive ability of the models were examined by the leave one-out cross-validation, Y-randomization, and external validation. The three methods indicated that the resulting multiparametric QSRR models possess high prediction ability and low overfitting.

Open Access

References

1. Kaliszan R (2000) In: Valko K (ed) Recent advances in quantitative structure-retention relationships; separation methods in drug synthesis and purification. Elsevier, Amsterdam
2. Kaliszan R (1997) Structure and retention in chromatography. A chemometric approach. Harwood, Amsterdam
3. Karelson M, Lovanov M, Katritzky AR (1996) Chem Rev 96:1027–1044. doi:10.1021/cr950202r
4. Kaliszan R (2007) Chem Rev 107:3212–3246. doi:10.1021/cr068412z
5. Liu F, Liang Y, Cao C, Zhou N (2007) Anal Chim Acta 594:279–289. doi:10.1016/j.aca.2007.05.023
6. Hu RJ, Liu HX, Zhang RS, Xue CX, Yao XJ, Liu MC, Hu ZD, Fan BT (2005) Talanta 68:31–39. doi:10.1016/j.talanta.2005.04.034
7. Héberger K (2007) J Chromatogr A 1158:273–305. doi:10.1016/j.chroma.2007.03.108

8. Junkes BS, Amboni RDMC, Yunes RA, Heinzen VEF (2003) Anal Chim Acta 477:29–39. doi:10.1016/S0003-2670(02)01413-7
9. Ren B (2003) Chemom Intell Lab Syst 66:29–39. doi:10.1016/S0169-7439(03)00004-2
10. Estrada E, Gutierrez Y (1999) J Chromatogr A 858:187–199. doi:10.1016/S0021-9673(99)00808-0
11. Liu S, Yin C, Cai S, Li Z (2002) Chemom Intell Lab Syst 61:3–15. doi:10.1016/S0169-7439(01)00146-0
12. Tello AM, Lebrón-Aguilar R, Quintanilla-López JE, Santiuste JM (2009) J Chromatogr A 1216:1630–1639. doi:10.1016/j.chroma.2008.10.025
13. Junkes BS, Amboni RDMC, Yunes RA, Heinzen VEF (2004) J Braz Chem Soc 15:183–189. doi:10.1590/S0103-50532004000200005
14. Garkani-Nejad Z, Karlovits M, Demuth M, Stimpfl T, Vycudilik W, Jalali-Heravi M, Varmuza K (2004) J Chromatogr A 1028:287–295. doi:10.1016/j.chroma.2003.12.003
15. Farkas O, Héberger K (2005) J Chem Inf Model 45:339–346. doi:10.1021/ci049827t
16. Héberger K, Görgényi M (1999) J Chromatogr A 845:21–31. doi:10.1016/S0021-9673(99)00323-4
17. Héberger K, Görgényi M, Sjöström M (2000) Chromatographia 51:595–600. doi:10.1007/BF02490818
18. Körtvélyesi T, Görgényi M, Héberger K (2001) Anal Chim Acta 428:73–82. doi:10.1016/S0003-2670(00)01220-4
19. Tulasamma P, Reddy KS (2007) Internet Electron J Mol Des 6:345–362
20. Wang Y, Yao X, Zhang X, Zhang R, Liu M, Hu Z, Fan B (2002) Talanta 57:641–652. doi:10.1016/S0039-9140(02)00078-4
21. Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. Wiley, New York
22. Hall LH, Kier LB (2001) J Mol Graph Model 20:4–18. doi:10.1016/S1093-3263(01)00097-3
23. Randić M (2001) J Mol Graph Model 20:19–35. doi:10.1016/S1093-3263(01)00098-5
24. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim, Germany
25. Mihalic Z, Nikolic S, Trinajstic N (1992) J Chem Inf Comput Sci 32:28–37. doi:10.1021/ci00005a005
26. Estrada E (2001) Chem Phys Lett 336:248–252. doi:10.1016/S0009-2614(01)00127-0
27. Randic M (1978) J Chromatogr 161:1–14. doi:10.1016/S0021-9673(01)85209-2
28. Sutter JM, Peterson TA, Jurs PC (1997) Anal Chim Acta 342:113–122. doi:10.1016/S0003-2670(96)00578-8
29. Liu F, Liang Y, Cao C, Zhou N (2007) Talanta 72:1307–1315. doi:10.1016/j.talanta.2007.01.038
30. Katritzky AR, Cheng K, Maran Y, Carlson DA (2000) Anal Chem 72:101–109. doi:10.1021/ac990800w
31. Tulasamma P, Reddy KS (2006) J Mol Graph Model 25:507–513. doi:10.1016/j.jmgm.2006.04.003
32. Reddy KS, Dutoit JC, Kováts Esz (1992) J Chromatogr 609:229–259. doi:10.1016/0021-9673(92)80167-S
33. Reddy KS, Cloux R, Kováts Esz (1994) J Chromatogr A 673:181–209. doi:10.1016/0021-9673(94)85037-2
34. Defayes G, Reddy KS, Dallos A, Kováts Esz (1995) J Chromatogr A 699:131–154. doi:10.1016/0021-9673(95)00023-G
35. Reddy KS, Cloux R, Kováts Esz (1995) J Chromatogr A 704:387–436. doi:10.1016/0021-9673(95)93206-B
36. Laffort P, Chauvin F, Dallos A, Callegari P, Valentin D (2005) J Chromatogr A 1100:90–107. doi:10.1016/j.chroma.2005.09.022
37. ISIS Draw 2.3 (1990–2000) MDL Information Systems, Inc
38. HyperChem Release 7.1 for windows molecular modeling system program package, HyperCube, 2002
39. Todeschini R, Consonni V (2002) Dragon software version 2.1. http://www.disat.unimib.it/chm/dragon.htm
40. Randic M, Zupan J (2001) J Chem Inf Comput Sci 41:550–560. doi:10.1021/ci000095o
41. Olivero J, Garcia T, Payares P, Vivas R, Diaz D, Daza E, Geerliger P (1997) J Pharm Sci 86:625–630. doi:10.1021/js960196u
42. Brown S, Tauler R, Walczak B (2009) Comprehensives chemometrics. Elsevier, Amsterdam
43. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Markopoulou OI (2006) QSAR Comb Sci 25:928–935. doi:10.1002/qsar.200530208
44. Rencher AC (2002) Methods of multivariate analysis, 2nd edn. Wiley, Canada
45. Topliss JG, Edwards RP (1979) J Med Chem 22:1238–1244. doi:10.1021/jm00196a017
46. Eriksson L, Jaworska J, Cronin M, Worth A, Gramatica P, McDowell R (2003) Environ Health Perspect 111:1361–75 (PMID: 12896860)
47. Chatterjee S, Hadi AS, Price B (2006) Analysis of collinear data. In: Regression analysis by example, 4th edn. Wiley, New York, pp 221–258 (Chap. 9)
48. Mihalic M, Trinajstic N (1992) J Chem Educ 69:701–712. doi:10.1021/ed069p701
49. Shao J (1993) J Am Stat Assoc 88:486–494
50. Hawkins DM, Basak SC, Mills D (2003) J Chem Inf Comput Sci 43:579–586. doi:10.1021/ci025626i
51. Hawkins DM (2004) J Chem Inf Comput Sci 44:1–12. doi:10.1021/ci0342472
52. Tetko IV, Livingstone DJ, Luik AI (1995) J Chem Inf Comput Sci 35:826–833. doi:10.1021/ci00027a006
53. Golbraikh A, Tropsha A (2000) Mol Divers 5:231–243. doi:10.1023/A:1021372108686
54. Melagraki G, Afantitis A, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2007) Bioorg Med Chem 15:7237–7247. doi:10.1016/j.bmc.2007.08.036
55. Ghavami R, Sadehgi F (2009) Chromatographia 70:851–868. doi:10.1365/s10337-009-1233-6
56. Hemmateenejad B, Sanchooli M (2007) J Chemometr 21:96–107. doi:10.1002/cem.1039