Inter*net* Electronic Journal of
Molecular Design

January 2003, Volume 2, Number 1, Pages 33–49

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 5

Guest Editor: Jun–ichi Aihara

Semiempirical Topological Index: A Novel Molecular Descriptor for Quantitative Structure–Retention Relationship Studies

Berenice da Silva Junkes, Renata Dias de Mello Castanho Amboni, Rosendo Augusto Yunes, and Vilma Edite Fonseca Heinzen

Departamento de Química, Universidade Federal de Santa Catarina, Campus Universitário, Trindade, Florianópolis (SC) 88040–900, Brazil

Received: July 18, 2002; Revised: October 2, 2002; Accepted: November 29, 2002; Published: January 31, 2003

Citation of the article:

B. S. Junkes, R. D. M. C. Amboni, R. A. Yunes, and V. E. F. Heinzen, Semiempirical Topological Index: A Novel Molecular Descriptor for Quantitative Structure–Retention Relationship Studies, *Internet Electron. J. Mol. Des.* **2003**, *2*, 33–49, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Semiempirical Topological Index: A Novel Molecular Descriptor for Quantitative Structure–Retention Relationship Studies[#]

Berenice da Silva Junkes,* Renata Dias de Mello Castanho Amboni, Rosendo Augusto Yunes, and Vilma Edite Fonseca Heinzen*

Departamento de Química, Universidade Federal de Santa Catarina, Campus Universitário, Trindade, Florianópolis (SC) 88040–900, Brazil

Received: July 18, 2002; Revised: October 2, 2002; Accepted: November 29, 2002; Published: January 31, 2003

Internet Electron. J. Mol. Des. **2003**, *2* (1), 33–49

Abstract

Motivation. An important property that has been extensively studied in QSPR is the chromatographic retention. Based on new considerations about the chromatographic behavior and experimental data, our group has developed a new topological index designed semi–empirical topological index, I_{ET}. The main goal of the present paper is to generalize the semi–empirical topological index, verifying the predictive–ability of the chromatographic retention for a diverse set of organic compounds (alkanes, alkenes, esters, ketones, aldehydes, and alcohols) and to obtain a general QSRR model. QSRR may be used as an important complementary tool for the elucidation of the molecular structure or for the prediction of the chromatographic retention.

Method. This index is based on the hypothesis that the chromatographic retention is due to the interaction of each atom of the molecule with the stationary phase, and consequently the value of the index is reduced by steric effects from its neighbors. Considering that the complexity involved in the solute–stationary phase interactions cannot be estimated only by theoretical considerations, values were attributed to the atoms of the molecules from the experimental chromatographic retention and theoretical deductions.

Results. The simple linear regression between the chromatographic retention and the index proposed, for all 548 organic compounds, is extremely satisfactory (correlation coefficient, $r = 1.0000$, standard deviation, SD = 7.01, and leave–one–out cross–validation correlation coefficient, $r^2_{\mathrm{CV}} = 0.999$). The predictive quality of the QSRR was tested for an external prediction set of 182 compounds randomly chosen from 548 compounds ($r = 1.0000$ and SD = 7.65).

Conclusions. Statistical analysis shows that the semi–empirical topological index has excellent predictive power using a single descriptor for a large data set of organic compounds.

Keywords. QSRR; quantitative structure–retention relationships; topological index; chromatographic retention; semi–empirical topological index; alkanes; alkenes; esters; aldehydes; ketones; alcohols.

Abbreviations and notations

I_{ET}, semi–empirical topological index	QSRR, quantitative structure–retention relationships
QSAR, quantitative structure–activity relationships	*RI*, retention index
QSPR, quantitative structure–property relationships	TI, topological indices

[#] Dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday.

* Correspondence author; phone: +55–48–331–9219; fax: +55–48–331–9711; E–mail: bsjunkes@hotmail.com and heinzen@qmc.ufsc.br.

1 INTRODUCTION

The use of graph–theoretical topological indices in quantitative structure–property and structure–activity relationships (QSPR/QSAR) studies has received major interest in recent years [1–7]. The topological indices became a powerful tool for predicting numerous physicochemical properties and/or biological activities of compounds as well as for molecular design. One of the most important properties that have been extensively studied is the chromatographic retention [8–15]. Quantitative structure–chromatographic retention relationship (QSRR) studies are widely investigated in gas chromatography (GC) and high–performance liquid chromatography (HPLC) [16].

The identification of the compounds by GC methods is carried out by peak comparison against a standard sample of each compound. The development of the QSRR is important since standards are not always available. It can efficiently help predict retention parameters by using theoretical descriptors from the chemical structure.

Correlations between gas chromatographic retention indices and molecular parameters provide significant information on the effect of molecular structure, on retention time and on the possible mechanism of absorption and elution [17].

Gas chromatographic retention is a very complex process. It involves the interaction of molecules by multiple intermolecular forces, as dispersion (or London forces), orientation (dipole–dipole or Keesom forces), induction (dipole–induced dipole or Debye forces), and electron donor–acceptor complexation, including hydrogen–bonding forces, leading to the partition of the solute between the gas and liquid phases [18–20]. Others factors, such as, steric hindrance of substituent groups within the solute molecule can also affect the chromatographic behavior [21,22].

Topological indices (TI) are numbers obtained via mathematical operations from the corresponding molecular graphs of compounds [23–28] in contrast to physicochemical characterization used by traditional QSAR [29]. One of the main advantages of TI is that they can be easily and rapidly computed for any constitutional formula yielding good correlation abilities. However, important disadvantages can be cited including its difficulty to encode stereo–chemical information, for example, to distinguish between *cis*– and *trans*–isomers, and its lack of physical meaning.

Many topological indices have been proposed since the pioneering works by Wiener [30] and by Kier *et al*. in the use of QSAR studies [23]. The TI recently developed to QSAR/QSRR studies can be illustrated by Estrada's approach to edge weights using quantum chemical parameters [31] and by Ren's atom–type AI topological indices derived from the topological distance sums and vertex degree further [6].

In the last years, our group has published several articles establishing structure–property relationships employing structural descriptors such as topological, geometrical and electronic or physicochemical descriptors for classes of organic compounds with different structural features. Using multiple linear regressions as statistical methods, the best descriptors were selected in the final structure–property model [32–40].

Based on a new hypothesis about the chromatographic behavior, our group recently developed a new topological index called semi–empirical topological index (I_{ET}). This index was initially developed to predict the chromatographic retention for linear and branched alkanes and linear alkenes with the objective of differentiating their *cis*– and *trans*– isomers and to obtain QSRR models [41]. The excellent results obtained stimulated our group to extend this new topological descriptor to other classes of compounds [42–46]. The equation obtained to calculate the I_{ET} was generated from the molecular graph and the values of the carbon atoms and the functional group were attributed observing the experimental chromatographic behavior and supported by theoretical considerations. This was done due to the difficulty to obtain a complete theoretical description of the interaction between the stationary phase and solute. Based only on theoretical equations or hypothesis it is not possible, for example, to estimate how the molecular conformation of the solute affects the intermolecular forces. In view of this, we believe that from experimental behavior we can obtain insights about these facts in order to apply them to other processes involved in QSPR studies. Thus, it can be noted that the semi–empirical topological index has a clear physical meaning.

The main goal of the present paper is to verify the predictive–ability of the chromatographic retention for a large data set (alkanes, alkenes, esters, ketones, aldehydes, and alcohols) using the semi–empirical topological index and to obtain a general QSRR model for these organic compounds.

2 MATERIALS AND METHODS

2.1 Chemical Data

The Kováts experimental retention indices of organic compounds used in the present investigation are taken from literature, as presented in Table 1 (see Supplementary Material). The data set for the chromatographic prediction of this study consists of 153 alkanes [41], 79 linear alkenes [41], 43 branched alkenes [42], 178 methyl–branched alkanes [43], 81 esters [44], 43 ketones [45], 11 aldehydes [45] and 44 alcohols [46] for a total of 632 organic compounds. These compounds represent a heterogeneous set of monofunctional compounds with different structural features. The experimental *RI*s of these compounds were measured on low–polarity stationary phases (squalane, DB–1, HP–1 and OV–1) and at different temperatures, as shown in Table 3.

2.2 Calculation of Semiempirical Topological Indices

Some basic considerations were taken into account in the development of this novel semi–empirical topological index. The representation of molecules was based on the molecular graph theory, where the carbon atoms are considered as the vertexes of the graph and hydrogen are suppressed [47]. As we know, the chromatographic retention of the solute molecules is mainly due to the number and interaction of each atom of the molecule with the stationary phase. This interaction is determined by its electrical properties and by the steric hindrance to this interaction by other atoms attached to it. Thus, values were attributed to the atoms of the molecules, based on the result of the general behavior of the experimental chromatographic retention of the molecules and theoretical assumptions.

2.2.1 Calculation of I_{ET} for alkanes and alkenes

Values were attributed to the carbon atoms (vertices in the molecular graphs) according to the following rules:

(1) According with Kováts convention, the correlation between the retention index and number of carbon atoms is linear for the alkanes [48]. However, branched alkanes do not present such linear relationship with Kováts index, since the retention of the tertiary and quaternary carbon atoms is decreased by the steric effects of their neighboring groups. It is evident that secondary, tertiary and quaternary carbon atoms have a value less than 100 u.i., as previously attributed by Kováts.

(2) Observing the experimental chromatographic behavior, approximate numeric values were attributed: 100 u.i. for the carbon atom in the methyl group in agreement with Kováts, 90 u.i. for the secondary carbon atoms, 80 u.i. for the tertiary and 70 u.i. for the quaternary. All values were divided by 100 to adapt them to the common topological values.

(3) The contribution of these carbon atoms on the chromatographic retention also depends on the neighboring substituent groups due to steric effects. In order to estimate steric effects, it was observed that the values of experimental *RI* decreased as the branch increased, showing a log trend. Therefore, it was necessary to add the value of the logarithm of each adjacent carbon atom. Thus, the new semi–empirical topological index (I_{ET}) is expressed as:

$$I_{ET} = \sum_i (C_i + \delta_i)$$
$$\delta_i = \sum\nolimits_{j\sim i} \log C_j \qquad (1)$$

where C_i is the value attributed to each carbon atom i and to the functional group in the molecule and δ_i is the sum of the logarithm of the value of each adjacent carbon atom (C_1, C_2, C_3 and C_4) and/or the logarithm of the value of the functional group, and ~ means ‘adjacent to’.

(4) For alkenes, the mainly interaction force between the solute and stationary phase is the

dispersive force, that is reduced by neighboring steric effects; however, the electrostatic force is also involved. The influence of conformational effects on the intermolecular forces makes it very difficult to predict these effects based only on theoretical considerations. For this reason, the values attributed to the carbon atom of the double bond for alkenes were calculated by numerical approximation based on the experimental retention indices as described in our previous publication [39].

2.2.2 Calculation of I_{ET} for compounds with oxygen–containing functional groups

The values attributed to the carbon atoms and functional groups (vertex of the molecular graphs) were based on the following rules:

(1) For this group of compounds, the main intermolecular forces that contribute to the chromatographic behavior in low polarity stationary phase are dispersive and inductive forces. The values attributed to functional groups are also based on the experimental retention index.

(2) The –COO– (ester), C=O (ketone or aldehyde) and C–OH (alcohol) groups were considered as a single vertex of the molecular graph of the compounds studied. This was done due to the difficulty and the inconsistency in calculating the individual values of the carbon atoms and the oxygen atoms of these groups. Thus, better numerical approximations were obtained, capable of reflecting the experimental chromatographic behavior of these compounds, when these groups were treated as a single vertex.

(3) The same considerations that were taken into account during the development of the semi–empirical topological method for the prediction of retention indices of alkanes and alkenes [41,42] were employed to develop the I_{ET} for oxo–compounds.

(4) The contribution of the chromatographic retention coming from carbon atoms and functional groups was represented by a single symbol, C_i, as indicated in Eq. (1). The semi–empirical topological index can be expressed by a general equation, for the entire set of compounds included in this work, where C_i is the value attributed to the –COO– (ester), C=O (ketone or aldehyde), C–OH (alcohol) groups and/or to each carbon atom, i, in the molecule, and δ_i is the sum of the logarithm of the values of each adjacent carbon atom (C_1, C_2, C_3, and C_4) and/or the logarithm of the value of the –COO– (ester), C=O (ketone or aldehyde), C–OH (alcohol) groups, and ~ means 'adjacent to'.

(5) In a first step, an approximate I_{ET} (I_{Eta}) was calculated for each compound. This was achieved using the equation previously obtained (RI_{exp} = 123.6871 I_{ETa} – 47.3557) for linear alkanes containing from 3 to 10 carbon atoms and Kováts experimental retention indices of compounds.

(6) Subsequently, the values of C_i for primary and secondary carbon atoms, previously attributed to alkanes [41], and the approximate I_{ET}, calculated above, were used in Eq. 1 in order to calculate

the values of –COO–, C=O and C–OH groups of linear compounds. Thus, values were attributed to each class of functional group in accordance with the position of the group in the carbon chain.

(7) One of the fundamental factors taken into consideration for the development of this topological index was the importance of the steric and other mutual intramolecular interactions between the functional group and atoms nearby. In this way, for branched molecules, different values were attributed to carbon atoms in the α, β, and γ position with respect to the functional groups compared to those previously attributed to alkanes [41], as described in refs. [44–46].

2.3 Regression Analysis

The statistical evaluation of the data was performed by the Origin [49] and Bilin [50] program packages. To test the quality of the regression equation, the correlation coefficient (r), the coefficient of determination (r^2), and the standard deviation (SD) were utilized as statistical parameters. To verify the validity and stability of the model obtained the cross–validation test (r^2_{CV}), using the “leave–one–out” method [51] was performed using the Bilin computer program. A further examination of external stability of the model was carried out by means of a procedure in which the entire data set was randomly divided into a training set of 366 compounds and a test set of 182 compounds.

3 RESULTS AND DISCUSSION

The semi–empirical topological index developed by our group is based on the supposition that the chromatographic behavior of a molecule results predominantly from the number and interaction of the atoms of the molecule with the stationary phase. This interaction is determined by its electrical properties and by steric hindrance to this interaction by other atoms attached to it. In order to obtain topological indices that consider all the complex reality of the molecular interactions, specific values were attributed to the carbon atoms and the functional group of the molecule considering the general chromatographic behavior and supported by theoretical deductions.

It is well known that the Kováts retention indices of non–polar substances on non–polar stationary phases show an almost completely linear dependence on column temperature. On polar stationary phases, this relationship is represented by a hyperbolic curve described by the Antoine–type equation. This curve can depict a significant linear segment, whose length mainly depends on the polarity of the substance examined, on the stationary phase applied, and on their interactions [18,52,53].

Considering the predominance of this linear relationship between *RI* and temperature in non–polar and low polar stationary phases, the semi–empirical topological index can be applied to different temperature ranges of low–polarity stationary phases [41–46]. The influence of the temperature on the quality of the linear regressions between *RI* and I_{ET} can be verified in a recently published article [42]. In this work, the topological descriptor was generated for branched alkenes on squalane at 80 °C. It was subsequently applied to obtain QSRR models ($RI = a + b\ I_{ET}$) on diverse low stationary phases at different temperatures (Table 3 in ref. [42]). The similarity between angular coefficients and the quality of the statistical parameters of QSRR models demonstrated the applicability of I_{ET} for these experimental conditions. This clearly indicates that the semi–empirical topological descriptors can be considered as invariants of the system within some limits such as stationary low–polarity phases and appropriate temperature range.

Analyzing the influence of structural features on the chromatographic behavior of the organic compounds selected in this study, it is possible to verify that the retention mainly depends on the number of carbon atoms, the degree of branching, the presence of the heteroatom and the position of the functional group in the carbon chain. Thus, the following general qualitative statements can be drawn:

(1) The branching of a chain reduces the *RI*s of the compounds due to steric effects, in the following order: CH_3 < $–CH_2–$ < >CH– < >C<. In a general way, the tertiary and quaternary carbon atoms in the α, β and γ position attached to the functional groups exhibit different values from the C_i values previously attributed to the same carbon atom for alkanes [41]. As expected, the steric hindrance on the carbon atoms attached to the functional group generally decreased as the carbon atom moved away from the functional group in the order α, β and γ. These values are specific for each functional group because it is necessary to take into account not only the influence of the steric effect but also other mutual intramolecular interactions between the functional group and neighboring carbon atoms.

(2) As earlier observed for alkanes and alkenes [41], one of the most important factors in the chromatographic retention of esters, aldehydes, ketones and alcohols is the dispersion interaction between the surfaces of solute and the stationary phase, which is related to steric factors, molecular size, and branching. However, it is necessary to consider the permanent dipole moment of oxygen–containing functional groups, which should provoke dipole–induced–dipole interactions with any non–polar stationary phase. This can be observed comparing the contributions to the retention indices of some structural fragments that represents different functional groups having the same topological structure. These fragments are those representing C–OH in 2–heptanol, the C=O group of the 2–heptanone, –COO– of the pentyl acetate and $–CH_2–$ in heptane. The correlation between the C_i values attributed to fragments, C–OH, C=O, –COO–, and $–CH_2–$, and the retention indices of these respective compounds is illustrate in Figure 1. The contributions of these groups to the chromatographic retention are ordered as follows: hydroxyl > carbonyl > carboxyl > methylene, indicating the importance of the nature of the functional group of the given type of compound, which include dispersive interactions, dipole–induced–dipole and hydrogen bonding interactions.

(3) We also observed the influence of the position of functional groups (esters, aldehydes, ketones and alcohols) and the double bond (alkenes) on the chromatographic retention. The retention indices of compounds decrease when the functional group moves towards the center of the carbon backbone. This result can be attributed to the steric hindrance of aliphatic side–chains on the functional group, significantly reducing the contribution of the oxygen atom to the chromatographic retention of these compounds.

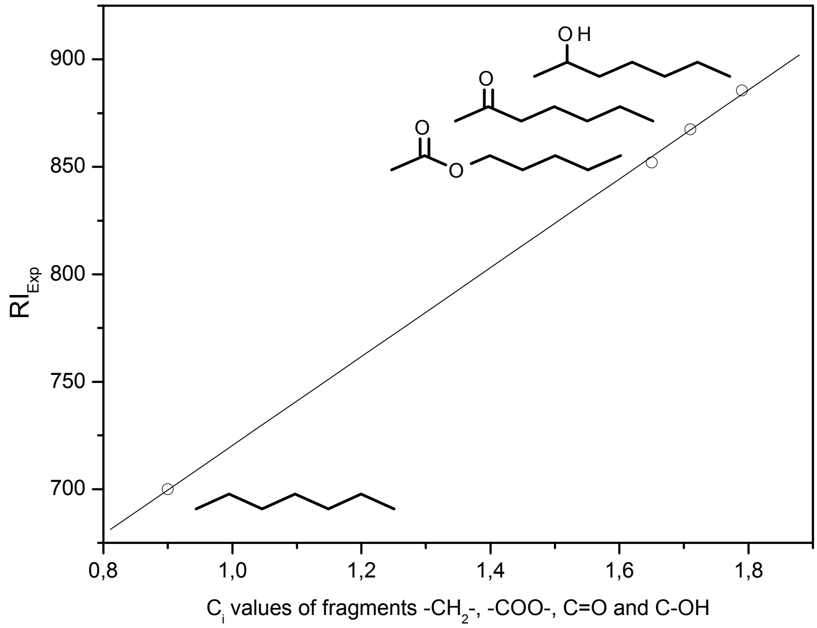


Figure 1. Correlation between the C_i values attributed to fragments $–CH_2–$, –COO–, C=O and C–OH and the retention indices of heptane, pentyl acetate, 2–heptanone and 2–heptanol, respectively.

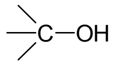
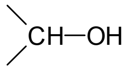
The values of C_i for the carbon atoms and the values attributed to the functional groups of esters, aldehydes, ketones and alcohols are listed in Table 2.

As the starting point, the I_{ET} was developed for alkanes on non– polar stationary phase. These compounds are the simplest ones and their properties almost completely depend on topological features. Subsequently, this novel topological descriptor was extended to different classes of organic compounds with more complex structural features. Due to the influence of the polarity of the stationary phase on the chromatographic behavior, the I_{ET} has been developed on low polarity stationary phases. Recently, our group has investigated the predictive ability of I_{ET} to predict the *RI* of aldehydes and ketones on stationary phases of different polarity and better results were found on stationary phases of low polarity, as it was expected [54]. Thus, in a near future, we intend to improve the I_{ET} in order to apply it on polar stationary phases, reflecting the interactions between polar molecules and polar stationary phases.

Table 2. Values of C_i for the carbon atoms and the values attributed to the functional groups of esters, aldehydes, ketones and alcohols [41–46].

Class of organic compounds	Fragment	Fragment position	C_i
Linear and branched alkanes	$-CH_3$	–	1.0000
	$-CH_2-$	–	0.9000
	$-CH<$	–	0.8000
	$>C<$	–	0.7000
Linear alkenes	$CH_2=$; $-CH=$	1C	0.8975
	$-CH=$ *trans–*	2C	0.8950
	cis–		0.9100
	$-CH=$ *trans–* [a]	3C	0.8750
	cis– [a]		0.8850
	$-CH=$ *trans–* [a]	4C	0.8650
	cis– [a]		0.8700
	$-CH=$ *trans–*	5C	0.8650
	cis–		0.8550
	$-CH=$ *trans–*	6C	0.8600
	cis–		0.8500
	$-CH=$ *trans–*	7C	0.8575
	cis–		0.8450
Branched alkenes	$=CH_2$; $=CH-$	1C/2C/3C	0.8975
	$=C<$	1C/2C/3C	0.8500
	$=C<$ *trans–*	2C/3C	0.8800
	cis–		0.8200
	$=CH-R$ [b]	1C/2C/3C	0.8400
	$=CH-R$ [b]; $=C<^{R}$ [b] *trans–*	2C/3C	0.8100
	cis–	2C/3C	0.7700
	$=CH-R$ [c]	1C/2C/3C	0.7700
	$=CH-R$ [c] *trans–*	1C/2C/3C	0.7700
	$=C<^{R}$ [b]	1C/2C/3C	0.7900
Esters	$-CH_3$	α acid	1.0700
	$-CH_3$	α alcohol	1.0700
	$-CH_2-$	α alcohol	0.8500
	$-CH_2-$	β alcohol	0.8800
	$-CH_2-$	γ alcohol	0.8900
	$-CH_2-$	α acid	0.8950
	$-CH_2-$	β acid	0.8700
	$-CH_2-$	γ acid	0.8970
	$>CH-$	α alcohol	0.6400
	$>CH-$	β alcohol	0.7000
	$>CH-$	γ alcohol	0.7400
	$>CH-$	α acid	0.7000
	$>CH-$	β acid	0.7100
	$>CH-$	γ acid	0.7500
	$>C<$	α alcohol	0.5200
	$R-COO-R'$ [d]	–	1.6500
	$HCOO-R'$ [d]	–	2.1500

Table 2. (Continued).

Class of organic compounds	Fragment	Fragment position	C_i
Aldehydes and ketones	HC=O	aldehyde	2.0940
	C=O	2	1.7100
	C=O	3	1.6900
	C=O	Middle of the chain [e]	1.6000
	>CH–	α	0.7300
	>CH–	β	0.7000
	>CH–	γ	0.7650
	>C<	α	0.6100
	>C<	β	0.6100
Alcohols	$–CH_2–OH$	–	2.6300
		2	1.7900
		3	1.7800
		Middle of the chain [e]	1.6800
		2	1.2600
		3	1.3600
	–CH<	α	0.7500
		β	0.7300
	>C<	α	0.6100
		β	0.6300

[a] Above 10 carbon atoms in the carbonic chain, the values for *cis*– and *trans*– linear alkene isomers should be inverted
[b] R = alkyl group with α or β branching at the double bond
[c] $R = –C(CH_3)_3$ group at the α position
[d] R corresponds to the acid side and R' corresponds to the alcohol side of the molecule
[e] For compounds with more than 6 carbon atoms in the backbone

A summary of the best simple linear regression models ($RI = a + b\,I_{ET}$) and the statistical data for each data set of compounds, obtained in previous QSRR studies, is given in Table 3. The results illustrated in Table 3 clearly indicate that it is possible to generate a general QSRR model for different classes of organic compounds employing a single topological descriptor, I_{ET}. Thus, a new data set with all 632 compounds was used to build the general QSRR model.

The best simple linear regression obtained for the whole data set of 632 compounds using a single descriptor, I_{ET}, is given as follows:

$$\begin{gathered} RI = -\,55.4551 + 123.7183\, I_{ET} \\ n = 632 \quad r = 0.9999 \quad r^2 = 0.9997 \quad \mathrm{SD} = 17.71 \end{gathered} \tag{2}$$

where RI is the retention index and I_{ET} is the semi–empirical topological index. Table 1 (see Supplementary Material) shows the values of experimental retention indices (RI_{exp}), values of calculated retention indices (RI_{calc}) using Eqs. (2) and (3), ΔRI ($RI_{\mathrm{exp}} - RI_{\mathrm{calc}}$) and values of calculated semi–empirical topological index (I_{ET}) for alkanes, alkenes, esters, ketones, aldehydes and alcohols.

A good improvement of this QSRR model was obtained removing the branched alkanes that showed large residuals (≥ 20), considered as outliers (Table 1). This result was expected since these compounds are generally small and have a high degree of branching. For these compounds, the steric effect is probably highlighted due to the conformations of the molecules of the solute that plays an important role in solute–stationary phase interactions. Despite the good results obtained for the correlation between I_{ET} values and experimental retention indices for 157 linear and branched alkanes [41], our topological index was not able to encode some conformational effects, which are probably responsible by the chromatographic behavior of these compounds. However, in a later work, for a specific set of alkanes (methyl–branched alkanes produced by insects) [43], the I_{ET} was improved considering conformational factors in order to distinguish the methyl alkane isomers. The statistical parameters obtained in this QSRR model were excellent (r^2 = 0.9999, SD=4.31). In the near future, we intend to improve our topological index to obtain better results for these outliers branched alkanes and information about the molecular conformations involved in the chromatographic retention processes.

Table 3. Summary of the best simple linear regressions ($RI_{\text{calc}} = a + b\, I_{ET}$) found for different data sets on low–polarity stationary phases.

No	Data Set	Phase	Temperature (ºC)	*a*	*b*	*n*	*r*	SD	Ref.
1	Alkanes	SQ	100	116.8000	–19.0500	157	0.9901	26.20	41
2	Cis–/trans– linear alkenes	SQ	100	122.8446	–41.7054	79	1.0000	2.35	41
3	Branched alkenes	SQ	80	120.4671	–29.0457	59	0.9985	5.76	42
4	Methyl–branched alkanes	DB–1	Programmed [a]	123.1610	–39.5251	178	1.0000	4.31	43
5	Esters	SQ	81	123.7900	–48.1400	81	0.9995	5.79	44
6	Aldehydes and ketones	HP–1 and OV–1	50 and 60	123.4951	–45.6553	54	0.9999	5.01	45
7	Alcohols	OV–1	60	124.1239	–51.3739	44	0.9991	5.70	46

[a] Temperature programmed from 60 to 320 ºC.

A final QSRR model was obtained (removing these outliers) using the present method for 548 compounds, as illustrated bellow:

$$\begin{gathered} RI = –48.0866 + 123.4758\, I_{ET} \\ n = 548 \quad r = 1.0000 \quad r^2 = 1.0000 \quad \text{SD} = 7.01 \quad r^2_{\text{CV}} = 0.999 \end{gathered} \tag{3}$$

As it can be seen, this model explains more than 99 % of the variance in the experimental values of the retention index for this data set of compounds. Good results were obtained considering that the data originated from different sources under different experimental conditions.

The correlation between the experimental (RI_{exp}) and calculated retention index (RI_{calc}) for all compounds in this data set is shown in Figure 2. The residual values were plotted against the calculated ones to check the relationship between them (Figure 3). In this model, 33 compounds (29 branched alkanes) were identified as statistical outliers, showing $\Delta RI \geq 15$ index units. The histogram of residual values is illustrated in Figure 4. It can be seen that 72% of the compounds had absolute values lower than 6 index units and only 6% of the compounds revealed residual values greater than 15 index units.

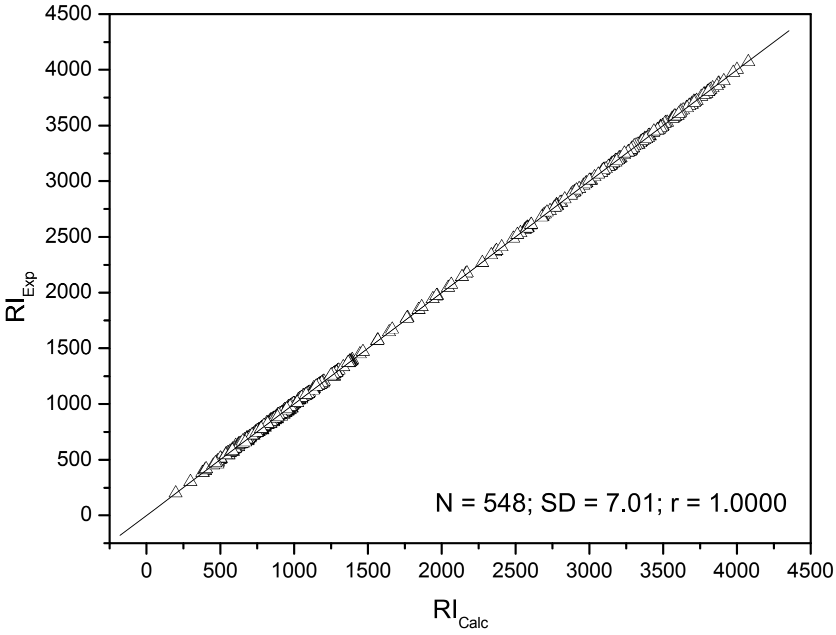


Figure 2. Experimental retention index (RI_{exp}) *vs*. calculated retention index (RI_{calc}) for the data set of 548 organic compounds.

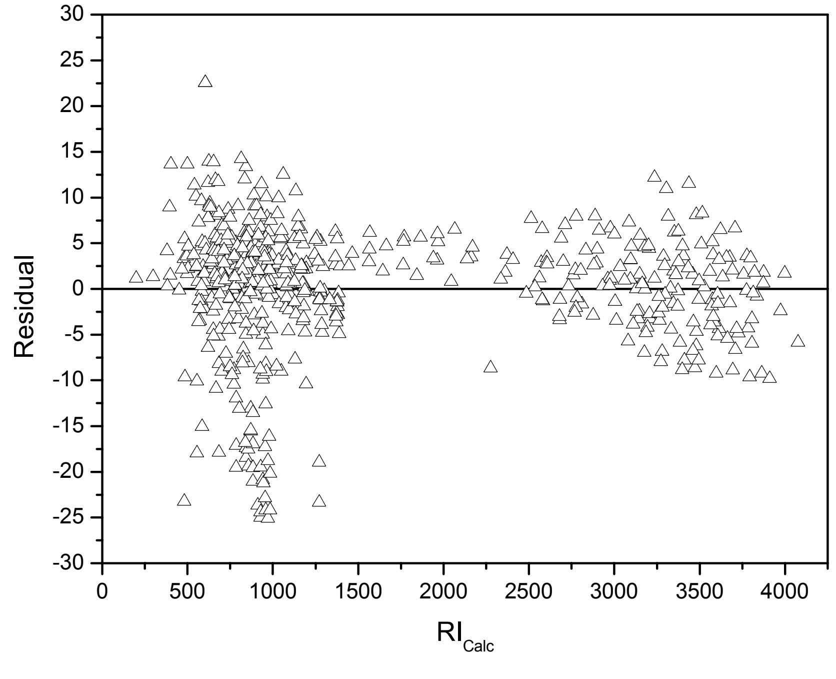


Figure 3. Plot of the residuals vs. calculated retention indices (RI_{calc}) for the data set of 548 organic compounds.

An excellent QSPR model should have not only good estimation ability for any internal sample, but also good prediction ability for an external sample. The most usual method to prove that a model has excellent prediction ability is a cross–validation method (r^2_{CV} = 0.999). In the present work, $n - 1$ samples from a total data set were used to construct a calibration set and to build a QSRR model. The property of the sample is then predicted using the one sample that was left out of the data set. The procedure above is repeated until every sample in the total data set is used for a prediction.

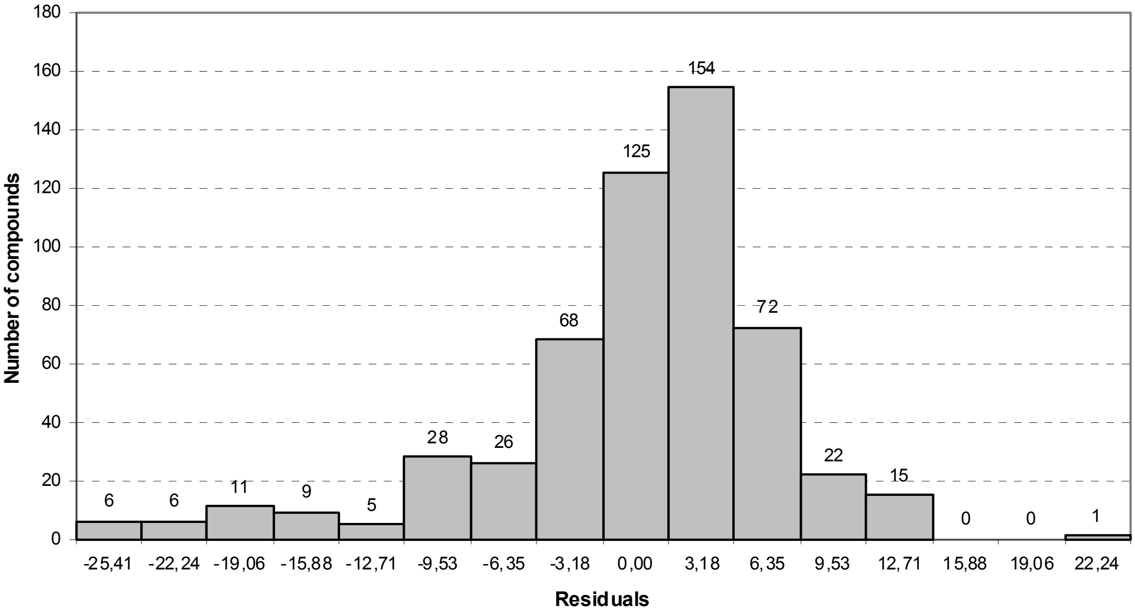


Figure 4 – Histogram of residual values of 548 organic compounds.

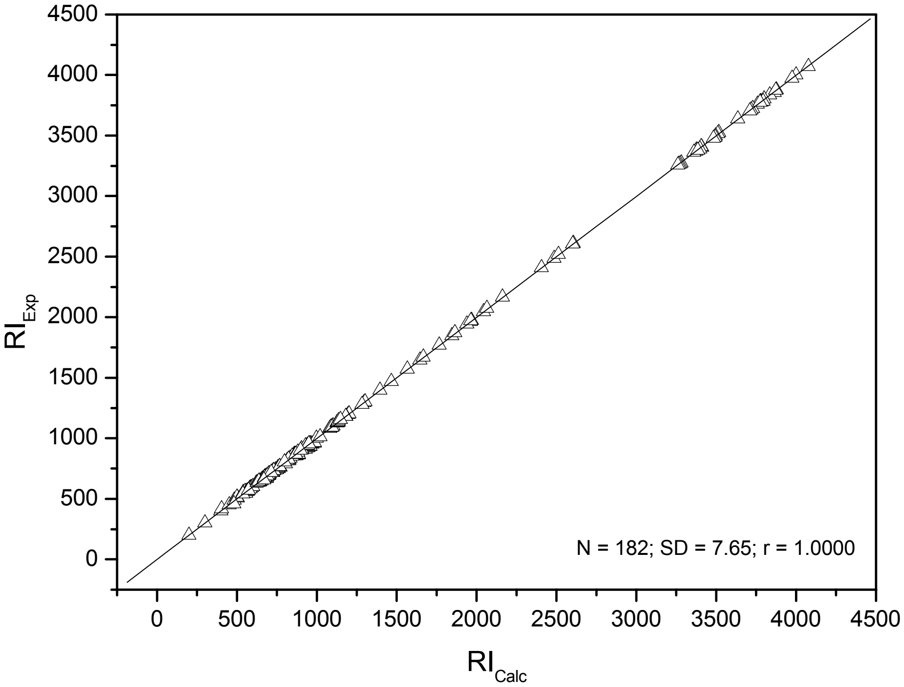


Figure 5– Experimental retention index (RI_{Exp}) vs. calculated retention index (RI_{Calc}) for the external prediction set of 182 organic compounds.

In order to further validate the external stability of the above QSRR model from Eq. (3), a new model was obtained by using 366 compounds in the training set or calibration set randomly chosen from whole 548 compounds. This model ($RI = –48.0866 + 123.4758\ I_{ET}$) was used to predict the chromatographic retention index of 182 remaining compounds, considered as the external prediction set. The graph of predicted *RI* values versus experimental retention indices for these compounds is shown in Figure 5 ($r = 1.0000$, SD = 7.65). These results also indicate a satisfactory predictive ability for the external data of QSRR model generated using our semi–empirical topological index.

The final prediction results from Eq. (3), using our single topological index, showed statistical data comparable to those obtained from similar studies recently reported elsewhere [9,20,55]. In one of the articles [20], the authors developed a method for the prediction of retention indices for a diverse set of compounds (184 organic compounds) using radial basis function networks (RBFNNs) from their physicochemical parameters. The selected compounds included acyclic and cyclic alkanes, alkenes, alcohols, ethers, ketones, and esters. For a test set of 34 compounds, a predictive correlation coefficient $r = 0.9910$ and root mean squared error of 14.1 were obtained by the method.

4 CONCLUSIONS

A novel semi–empirical topological method for the prediction of retention indices for a diverse set of compounds has been presented. Very satisfactory results were obtained with this topological index. Because of its simplicity, it is also suitable for routine work.

As observed in our previous publications using the semi–empirical topological method, the results of this work suggest that the role of the steric factors must be more important in the chromatographic retention than the polar effects on the non–polar and low–polarity stationary phases.

The predictive quality of the QSRR was tested for an external prediction set of 182 compounds randomly chosen from all 548 compounds ($r = 1.0000$, SD = 7.65). Statistical analysis shows that the semi–empirical topological index has good predictive power using a single descriptor for a large data set of organic compounds. The worst QSRR model (SD = 17.71) was observed when the small and highly branched alkanes were included in the data set. This result was expected because of the conformational effects, which were not well encoded by I_{ET}, presumably play an important role in the intermolecular interactions between these molecules and the stationary phase. However, as the I_{ET} was developed on non–polar and low–polarity stationary phases, the applicability of the proposed approach is limited to this type of stationary phases.

Many researchers have developed new topological descriptors where the main objective is to use them in drug design and to be able to predict biological activities. Thus, the good results obtained from the prediction of chromatographic retention using the novel semi–empirical topological index, I_{ET}, can be considered as an initial step towards forthcoming QSAR studies.

Acknowledgment

The authors acknowledge the financial support from CNPq (Brazil) to this research.

Supplementary Material

Table 1. Values of experimental retention indices (RI_{exp}), values of calculated retention indices (RI_{calc}) using Eqs. 2 and 3, ΔRI ($RI_{\mathrm{exp}} - RI_{\mathrm{calc}}$) and values of calculated semi–empirical topological index (I_{ET}) for alkanes, alkenes, esters, ketones, aldehydes and alcohols.

5 REFERENCES

[1] O. Ivanciuc, T. Ivanciuc, and A. T. Balaban, Quantitative structure–property relationship study of normal boiling for halogen–/oxygen–/sulfur–containing organic compounds using the CODESSA program, *Tetrahedron* **1998**, *54*, 9129–9142.

[2] A. Balaban, D. Mills, and S. C. Basak, Correlation between structure and normal boiling points of acyclic carbonyl compounds, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 758–764.

[3] A. R. Katritzky, U. Maran, V. S. Lobanov, and M. Karelson, Strucutre diverse quantitative structure–property relationship correlations of technologically relevant physical properties, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1–18.

[4] M. Randić and S. C. Basak, A new descriptor for structure–property and structure–activity correlations, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 650–656.

[5] S. D. Nelson and P. G. Seybold, Molecular structure–property relationships for alkenes, *J. Mol. Graph. Modell.* **2001**, *20*, 36–53.

[6] B. Ren, Application of novel atom–type AI topological indices in the structure–property correlations, *J. Mol. Struct.* (*Theochem*) **2002**, *586*, 137–148.

[7] O. Ivanciuc, T. Ivanciuc, and A. T. Balaban, Quantitative Structure–Property Relationships for the Normal Boiling Temperatures of Acyclic Carbonyl Compounds, *Internet Electron. J. Mol. Des.* **2002**, *1*, 252–268, http://www.biochempress.com.

[8] A. R. Katritzky, E. S. Ignatchenko, R. A. Barcock, V. S. Lobanov, and M. Karelson, Prediction of gas chromatographic retention times and response factors using a general quantitative structure–property relationship treatment, *Anal. Chem.* **1994**, *66*, 1799–1807.

[9] M. Pompe and M. Novic, Prediction of gas–chromatographic retention indices using topological descriptors, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 59–67.

[10] E. Estrada and Y. Gutierrez, Modeling chromatographic parameters by novel graph theoretical sub–structural approach, *J. Chromatogr. A* **1999**, *858*, 187–199.

[11] B. Ren, A new topological index for QSRR of alkanes, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 139–143.

[12] O. Ivanciuc, T. Ivanciuc, D. Carbol–Bass, and A. T. Balaban, Comparison of Weighting schemes for molecular graph descriptor: Application in quantitative structure–retention relationship models for alkylphenols in gas–liquid chromatography, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 732–743.

[13] A. R. Katritzky, K. Chen, U. Maran, and D. A. Carlson, QSRR correlation and predictions of GC retention indexes for methyl–branched hydrocarbons produced by insects, *Anal. Chem.* **2000**, *72*, 101–109.

[14] B. Ren, Application of novel atom–type AI topological indices to QSRR studies of alkanes, *Comput. Chem.* **2002**, *26*, 357–369.

[15] T. Ivanciuc and O. Ivanciuc, Quantitative Structure–Retention Relationship Study of Gas Chromatographic Retention Indices for Halogenated Compounds, *Internet Electron. J. Mol. Des.* **2002**, *1*, 94–107, http://www.biochempress.com.

[16] M. Markuszenwski and R. Kaliszan, Quantitative structure–retention relationships in affinity high–performance liquid chromatography, *J. Chromatogr. B* **2002**, *768*, 55–66.

[17] T. Körtvélyesi, M. Görgényi, and K. Hérberger, Correlation between retention indices and quantum–chemical descriptor of ketones and aldehydes on stationary phases of different polarity, *Anal. Chim. Acta* **2001**, *428*, 73–82.

[18] R. Kaliszan, Quantitative structure–chromatographic retention relationships, Wiley–Interscience, New York, 1987.

[19] C. T. Peng, Prediction of retention indices V. Influence of electronic effects and column polarity on retention

index, *J. Chromatogr. A* **2000**, *903*, 117–143.
[20] X. Yao, X. Zhang, R. Zhang, M. Liu, Z. Hu, and B. Fan, Prediction of gas chromatographic retention indices by the use of radial basis function neural networks, *Talanta* **2002**, *57*, 297–306.
[21] D. F. Fritz, A. Sahil, and E. Kováts, Study of the adsorption effects on the surface of poly–(ethylene glycol)–coated column packings, *J. Chromatogr.* **1979**, *186*, 63–80.
[22] C. T. Peng, S. F. Ding, R. L. Hua, and Z. C. Yang, Prediction of retention indexes I. Structure–retention index relationship on apolar columns, *J. Chromatogr.* **1988**, *436*, 137–172.
[23] L. B. Kier and L H. Hall, Molecular connectivity chemistry and drug research, Academic Press, New York, 1976.
[24] I. Rios–Santamarina, R. García–Domenech, J. Cortijo, P. Santamaria, E. J. Morcillo, and J. Gálvez, Natural Compounds with Bronchodilator Activity Selected by Molecular Topology, *Internet Electron. J. Mol. Des.* **2002**, *1*, 70–79, http://www.biochempress.com.
[25] A. A. Toropov and A. P. Toropova, QSAR Modeling of Mutagenicity Based on Graphs of Atomic Orbitals, *Internet Electron. J. Mol. Des.* **2002**, *1*, 108–114, http://www.biochempress.com.
[26] D. J. G. Marino, P. J. Peruzzo, E. A. Castro, and A. A. Toropov, QSAR Carcinogenic Study of Methylated Polycyclic Aromatic Hydrocarbons Based on Topological Descriptors Derived from Distance Matrices and Correlation Weights of Local Graph Invariants, *Internet Electron. J. Mol. Des.* **2002**, *1*, 115–133, http://www.biochempress.com.
[27] S.–S. Liu, H.–L. Liu, Y.–Y. Shi, and L.–S. Wang, QSAR of Cyclooxygenase–2 (COX–2) Inhibition by 2,3–Diarylcyclopentenones Based on MEDV–13, *Internet Electron. J. Mol. Des.* **2002**, *1*, 310–318, http://www.biochempress.com.
[28] O. Ivanciuc, T. Ivanciuc, D. Cabrol–Bass, and A. T. Balaban, Optimum Structural Descriptors Derived from the Ivanciuc–Balaban Operator, *Internet Electron. J. Mol. Des.* **2002**, *1*, 319–331, http://www.biochempress.com.
[29] R. García–Domenech, A. Catalá–Gregori, C. Calabuig, G. Antón–Fos, L. del Castillo, and J. Gálvez, Predicting Antifungal Activity: A Computational Screening Using Topological Descriptors, *Internet Electron. J. Mol. Des.* **2002**, *1*, 339–350, http://www.biochempress.com.
[30] H. Wiener, Structural determination of paraffin boiling points, *J. Am. Chem. Soc.* **1947**, 69, 17–20.
[31] E. Estrada, The Structural Interpretation of the Randić Index, *Internet Electron. J. Mol. Des.* **2002**, *1*, 360–366, http://www.biochempress.com.
[32] V. E. F. Heinzen and R. A. Yunes, Relationship between gas chromatographic retention indices and molecular connectivity indices of chlorinated pesticides and structurally related compounds, *J. Chromatogr.* **1992**, 598, 243–250.
[33] V. E. F. Heinzen and R. A. Yunes, Correlation between gas chromatographic retention indices of linear alkylbenzene isomers and molecular connectivity indices, *J. Chromatogr. A* **1993**, *654*, 83–89.
[34] A. C. Arruda, V. E. F. Heinzen, and R. A. Yunes, Relationship between Kováts retention indices and molecular connectivity indices of tetralonas, coumarins and structurally related compounds, *J. Chromatogr. A* **1993**, *630*, 251–256.
[35] V. E. F. Heinzen and R. A. Yunes, Using topological indices in the prediction of chromatographic retention indices of linear Alkylbenzene isomers, *J. Chromatogr. A* **1996**, *719*, 462–467.
[36] V. E. F. Heinzen, V. Cechinel Filho, and R. A. Yunes, Correlation of activity of 2–(X–benzyloxy)–4,6–dimethoxyacetophenones with topological indices and with the Hansch equation, *Il Farmaco* **1999**, *54*, 125–129.
[37] M. F. Soares, F. D. Monache, E. F. Heinzen, and R. A. Yunes, Prediction of gas chromatographic retention indices of coumarins, *J. Braz. Chem. Soc.* **1999**, *10*, 189–196.
[38] M. F. Soares, C. R. Boing, V. E. F. Heinzen, and R. A. Yunes, Aplicação da teoria de QSRR na interpretação da retenção cromatográfica de uma série de acetofenonas, *Anais Assoc. Bras. Quim.* **2000**, *49*, 24–30.
[39] R. D. M. C. Amboni, B. S. Junkes, R. A. Yunes, and V. E. F. Heinzen, Quantitative structure–odor relationships of aliphatic esters using topological indices, *J. Agric. Food Chem.* **2000**, *48*, 3517–3521.
[40] R. A. Yunes, V. E. F. Heinzen, V. Cechinel Filho, and M. Lazzarotto, From the manual method of Topliss to a modified quantitative method, *Arzneim.– Forsch./ Drug Res.* **2002**, *52*, 125–132.
[41] V. E. F. Heinzen, M. F. Soares, and R.A. Yunes, Semi–empirical topological method for prediction of the chromatographic retention of *cis*– and *trans*–alkene isomers and alkanes, *J. Chromatogr.* **1999**, *849*, 495–506.
[42] B. S. Junkes, R. D. M. C. Amboni, V. E. F. Heinzen, and R. A. Yunes, Use of a semi–empirical topological method to predict the chromatographic retention of branched alkenes, *Chromatographia* **2002**, *55*, 75–81.
[43] B. S. Junkes, R. D. M. C. Amboni, V. E. F. Heinzen, and R. A. Yunes, Quantitative structure–retention relationships (QSRR), using the optimum semi–empirical topological index, for methyl–branched alkanes produced by insects *Chromatographia* **2002**, *55*, 707–714.
[44] R. D. M. C. Amboni, B. S. Junkes, R. A. Yunes, and V. E. F. Heinzen, Semi–empirical topological method for prediction of the chromatographic retention of esters, *J. Mol. Struct.* (*Theochem*) **2002**, *579*, 53–62.
[45] R. D. M. C. Amboni, B. S. Junkes, R. A. Yunes, and V. E. F. Heinzen, Quantitative structure–property relationship study of chromatographic retention indices and normal boiling points for oxo compounds using the

semi–empirical topological method, *J. Mol. Struct.* (*Theochem*) **2002**, *586*, 71–80.

[46] B. S. Junkes, R. D. M. C. Amboni, R. A. Yunes, and V. E. F. Heinzen, Prediction of the chromatographic retention of saturated alcohols on stationary phases of different polarity applying the novel semi–empirical topological index, *Anal. Chim. Acta* **2003**, *477*, 29–39.

[47] P. J. Hansen and P. C. Jurs, Chemical applications of graph theory. Part I. Fundamental and topological indices, *J. Chem. Educ.* **1988**, *65*, 574–580.

[48] E. Kováts, Zu Fragen der Polarität. Die Method der LinearKombination der Wechselwirkungskräfte (LKWW), *Chimia* **1968**, *22*, 459–462.

[49] MicroCal Origin version 5.0.

[50] H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches*, Eds. R. Mannhold, P. Krogsgaard–Larsen, H. Timmerman, VCH, Weinheim, 1993.

[51] S. Muresan, C. Bologa, M. Mracec, A. Chiriac, B. Jastorff, Z. Simon, and G. Náray–Szabó, Comparative QSAR study with eletronic and steric parameters for cAMP derivatives with large substituents in positions 2, 6, and 8, *J. Molec. Struct.* (*Theochem*) **1995**, *342*, 161–171.

[52] L. S. Ettre and K. Billed, Consideration on the retention index concept I. Retention index and column temperature, *J. Chromatogr.* **1967**, *30*, 1–11.

[53] M. V. Budahegyi, E.R. Lombosi, T.S. Lombosi, S. Y. Mészáros, Sz. Nyiredy, G. Tarján, I. Timar, and J. M. Takács, Twenty–fifth anniversary of the retention index system in gas–liquid chromatography, *J. Chromatogr.* **1983**, *271*, 213–307.

[54] B. S. Junkes, R. D. M. C. Amboni, V. E. F. Heinzen, and R. A. Yunes, Application of novel semi–empirical topological index in the QSRR of aliphatic ketones and aldehydes on stationary phases of different polarity, *J. Braz. Chem. Soc.* **2002** (submitted).

[55] A. Yan, R. Zhang, M. Liu, Z. Hu, M. A. Hopper, and Z. Zhao, Large artificial neural networks applied to prediction of retention indices of acyclic and cyclic alkanes, alkenes, alcohols, esters, ketones and ethers, *Computers Chem.* **1998**, *22*, 405–412.