Journal of Chromatography A, 1422 (2015) 277–288

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Quantitative structure–property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase

Cristian Rojas[a,b,*], Pablo R. Duchowicz[a], Piercosimo Tripaldi[c], Reinaldo Pis Diez[d]

[a] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (CCT La Plata-CONICET, UNLP), Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina
[b] Decanato General de Investigaciones, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo, Apartado Postal 01.01.981, Cuenca, Ecuador
[c] Laboratorio de Química-Física de Alimentos, Facultad de Ciencia y Tecnología, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo, Apartado Postal 01.01.981, Cuenca, Ecuador
[d] CEQUINOR, Centro de Química Inorgánica (CONICET, UNLP), Departamento de Química, Facultad de Ciencias Exactas, UNLP, C.C. 962, 1900 La Plata, Argentina

ARTICLE INFO

Article history:
Received 28 February 2015
Received in revised form 7 October 2015
Accepted 7 October 2015
Available online 22 October 2015

Keywords:
Fragrance
Carbowax 20M column
QSPR theory
Dragon software
Replacement method
k-Means cluster analysis

ABSTRACT

A quantitative structure–property relationship (QSPR) was developed for modeling the retention index of 1184 flavor and fragrance compounds measured using a Carbowax 20M glass capillary gas chromatography column. The 4885 molecular descriptors were calculated using Dragon software, and then were simultaneously analyzed through multivariable linear regression analysis using the replacement method (RM) variable subset selection technique. We proceeded in three steps, the first one by considering all descriptor blocks, the second one by excluding conformational descriptor blocks, and the last one by analyzing only 3D-descriptor families. The models were validated through an external test set of compounds. Cross-validation methods such as leave-one-out and leave-many-out were applied, together with Y-randomization and applicability domain analysis. The developed model was used to estimate the *I* of a set of 22 molecules. The results clearly suggest that 3D-descriptors do not offer relevant information for modeling the retention index, while a topological index such as the Randić-like index from reciprocal squared distance matrix has a high relevance for this purpose.

1. Introduction

Fragrance and flavor substances [1] are strong-smelling organic compounds. Their major common characteristic is a pleasant odor (fragrance chemicals) or a pleasant taste (flavor chemicals). Fragrance substances are used as components in perfumes or perfumed products, while flavor substances are used as flavoring or to enhance the flavor of beverages and food products. Gas chromatography (GC) techniques are generally used for analyzing the contents and impurities of fragrances and flavors, as well as for quality control and in-process control, in order to provide details of their profiles in few minutes. Since the majority of fragrance and flavor compounds are volatile, the GC technique is commonly used for both separation and quantitative determination [2].

In 1977, three publications appeared for the first time on QSPR theory, or what is currently known as QSRR (quantitative structure–retention relationships) [3]. Subsequently, in 1987 and 1997 two pioneering books were published containing several hundred publications and applications of the QSRR theory [4,5], which is a powerful tool in Chromatography data analysis [6]. The aim of QSRR is to predict retention data for non-synthesized compounds from the knowledge of their molecular structure. The accurate prediction of the retention index (*I*) represents a challenge in QSPR, because this requires quality and precision in the experiments. However, the methodology is useful for chromatographers in order to prepare experimental designs [7] and to optimize the separation of complex mixtures. In addition, reliable QSRR methods have been established to understand the molecular mechanism of retention on diverse stationary phases and, therefore, to rationally design new phases with defined properties [3].

* Corresponding author at: Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (CCT La Plata-CONICET, UNLP), Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina. Tel.: +54 221 4257430; fax: +54 221 4254642.
E-mail address: crojasvilla@gmail.com (C. Rojas).

http://dx.doi.org/10.1016/j.chroma.2015.10.028
0021-9673/

Several QSPR studies have been published in the past years for both the modeling and the prediction of the *I* parameter on the stationary polar phase Carbowax 20M. A large number of studies were performed on alkylbenzene derivatives [8–14], pyrazine derivatives [15–18], and acyclic and monocyclic terpenes [19].

On the other hand, few studies regarding aroma compounds can be found in the literature. In 1990, Anker and Jurs [16] measured the *I* parameter of 115 odor compounds and established a QSPR model after removing four outliers, leading to quality parameters of $R^2 = 0.994$, and $S = 17.1$ as well as $S_{cv} = 24.7$ for internal validation of leave-one-out. Yan et al. [20] performed a QSRR for 434 flavor compounds. The whole set was divided into a training set (330 compounds) and a test set (107 compounds). Subsequently, 195 molecular descriptors were calculated and six were selected by means of the stepwise method. Using this technique, a well-correlated model was achieved both for the training set ($R^2_{train} = 0.923$, $S_{train} = 104.24$, and $R^2_{cv} = 0.922$, $S_{cv} = 104.48$) and for the test set ($R^2_{test} = 0.926$, $S_{test} = 105.48$).

Recently, Rojas et al. [21], used a data set of 1208 aroma compounds and calculated molecular descriptors using the well-known Dragon software in order to develop a predictive QSPR for the non-polar stationary phase OV-101. It was found that the 3D descriptors did not give any relevant information for the prediction of this property. Therefore, the main purpose of the present work is to use the same data set for the development of a predictive QSPR model for the *I* measured on the stationary polar phase Carbowax 20M following the same methodology as in our previous work. If this QSPR model is shown to predict the *I*, it may be a useful tool for people working on aroma and flavor chemistry to understand the retention mechanism of volatile compounds in this polar stationary phase, and in some cases to estimate the *I* property of other compounds not considered in this data set.

2. Materials and methods

2.1. Experimental data set

The chemical domain analyzed in the present study involves 1206 aromatic substances reported by Jennings and Shibamoto [22]. The experimental property reported by these authors is the Kováts retention index in the polar stationary capillary column (0.2 mm × 80 m), which is coated with polyethylene glycol Carbowax 20M, and programmed from 70 to 170 °C at 2 °C min^{-1}.

The data set is composed of 1184 substances, in which the *I* values vary in the range from 500 to 2640. The *I* values for 22 compounds were not reported by Jennings and Shibamoto, and are considered as a prediction set in the current study. The chemical names, SMILES (simplified molecular input line entry system) notations as obtained with the Open Babel software [23] and *I* values are presented in Table 1S. When a molecule had two or more *I* values the average value was used.

2.2. Molecular descriptors

A crucial problem in QSPR studies is to find a convenient structure representation. Generally, researchers use molecular descriptors as structural characterizations. Descriptors are the final result of a logical and mathematical procedure that transforms chemical information encoded within a symbolic representation of a molecule into a numerical quantity or into the result of some standardized experiment [24]. In this study, the compounds were first drawn using HyperChem for Windows [25]. For geometry optimization, the molecular mechanics force field (MM+) was applied, followed by the PM3 semiempirical method to refine the structures. The conjugate gradient algorithm, in the Polak-Ribiere version, was used for the optimizations, and the geometries were considered optimized when the root mean square deviation of the gradient vector became less than 0.01 kcal (Å mol)$^{-1}$. We computed 4885 molecular descriptors (refer to the descriptors.txt file in the supplementary material) implemented in the software Dragon [26]. This well-known descriptors database includes twenty nine descriptor families: 0D-descriptors (constitutional indices), 1D-descriptors (functional group counts, atom-centered fragments, molecular properties), 2D-descriptors (ring descriptors, topological indices, walk and path counts, connectivity indices, information indices, 2D matrix-based descriptors, 2D autocorrelations, Burden eigenvalues, P_VSA-like descriptors, edge adjacency indices, CATS 2D, 2D atom pairs, atom-type E-state indices, ETA indices, drug-like indices), and 3D-descriptors (Randic molecular profiles, geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, charge descriptors, 3D matrix-based descriptors, 3D autocorrelations, 3D atom pairs) [24]. The first data set was composed by descriptors belonging to all the blocks. In the second data set only non-conformational molecular descriptors were considered. Finally, in the third data set we only considered 3D-descriptors. This was done in order to compare whether 3D-molecular descriptors played an important role for modeling the *I* parameter. In all the cases, non-informative molecular descriptors were excluded. For example, molecular descriptors with constant values (descriptors with all values equal, e.g., number of phosphorous atoms "nP") and near-constant values (descriptors with only one value different from the remaining ones, e.g., number of bromine atoms "nBR") were excluded since they were not relevant descriptors for QSPR analysis.

2.3. Model development

2.3.1. Molecular descriptor selection in MLR

Another issue to address in the QSPR theory is the selection of the most useful molecular descriptors from a large number of correlated variables. In the last few years there was an increasing interest in developing methodologies regarding the selection of the best pool of descriptors in QSAR/QSPR studies. Lučić et al. [27] proposed a variable selection technique based on multivariable linear regression (MLR) to predict the Chromatographic Retention Times. This method was based on two steps: (1) A full search of descriptors for MLR from 1 to 4 descriptors, and (2) a stepwise procedure in order to build MLR models from 5 to 10 descriptors. However, a methodology involving a full search procedure to study large data sets is impractical because it is very time consuming, and requires high computational costs. Consequently, we used the replacement method (RM) variable subset selection [28].

RM is an efficient optimization tool that generates MLR models by searching in a set having *D* descriptors for an optimal subset having $d \ll D$ with the smallest *RMSD*. The quality of the results achieved with this technique is close to performing an exact (combinatorial) full search of molecular descriptors, although it requires much less computational work. The RM has been previously applied for chromatographic purposes [21,29]. For further details of this variable selection technique, refer to the RM procedure.docx in the supplementary material.

In addition to the use of RM for optimization, a fundamental step in the development of a QSPR model is the determination of the optimal size of the model, i.e., the number of descriptors to be included in the QSPR equation. In order to avoid overfitting in the models, an external validation set of compounds was considered. In this way, we considered the quality parameter models given by the RM which are: R^2_{train}, R^2_{val}, $RMSD_{train}$, $RMSD_{val}$, and $R^2_{ij\ max}$. The best model combines the highest values for the squared correlation coefficients (R^2), and the lowest values for both the root mean

squared deviations (*RMSD*) and the maximum correlation coefficient among descriptors ($R^2_{ij\ \max}$).

2.3.2. Model validation

The established QSPR models were validated in order to determine their predictive power, by predicting the *I* on compounds not considered during the calibration and comparing such values with the experimental ones. Therefore, the whole set of 1184 compounds was split into training (395), validation (396) and test (393) sets of compounds [30]. The training set was used to calibrate the model and to obtain its parameters; the validation set helps to partially validate the model; the test set includes compounds "never seen" during the calibration and demonstrates the predictive capability. It is also known that such splitting should be done by achieving similar structure–property relationships in the three sets; in other words, the training set molecules should be representative of the validation and test set compounds [31]. There are available in the literature several standard techniques that allow designing a rational partition of a data set, such as principal components analysis (PCA), discriminant analysis (DA), cluster analysis (CA), or methods based on the fuzzy logic theory [32].

In this work, we chose the training validation and test set compounds following the procedure developed in our previous study, one which has been applied elsewhere [21,33]. This methodology is based on the *k*-means cluster analysis (*k*-MCA) method [34]. The essence of *k*-MCA is to create *k*-clusters or groups of compounds, in such a way that compounds in the same cluster are very similar in terms of distance metrics (i.e., Euclidean distance), and compounds in different clusters are very distinct. The procedure applied to the retention index data set involves the following steps:

a. prepare a matrix (**C**) that includes the experimental property and the 1815 non-conformational molecular descriptors, i.e., this is done to account for the structure–property relationship during the classification process. Furthermore, only geometry independent descriptors are used in order to avoid optimization biases. Now the size of **C** is 1184×1815.
b. remove the linearly dependent variables from the previous matrix. The actual size of **C** is 1184×1116.
c. standardize **C** for centering and scaling its matrix elements. This is done for discerning better the matrix elements.
d. create N^0_{train} clusters with the 1184 compounds through the *k*-MCA method, for which the **C** matrix is used together with the Euclidean metrics, and 20 runs for the numerical optimization algorithm of *k*-MCA in order to achieve the best solution. This computes N^0_{train} cluster centroid locations, each centroid of 1×1116 size. $N^0_{train} = N_{train} - N_{\min\max}$, where N_{train} is the number of compounds in train and $N_{\min\max}$ is the number of compounds that have minimum or maximum values for the experimental property.
e. the training set is designed by including one compound per cluster, which is the compound that is nearer to the centroid in each cluster. It also includes the $N_{\min\max}$ compounds.
f. create N_{val} clusters with the remaining $1184 - N_{train}$ compounds through the *k*-MCA method, in the same numerical conditions as described previously. This computes N_{val} cluster centroid locations.
g. the validation set is designed by including one compound per cluster, which is the compound that is nearer to the centroid in each cluster.
h. Finally, the test set includes the remaining $1184 - N_{train} - N_{val}$ compounds.

We carried out the cross-validation technique of leave-one-out (loo) and leave-many-out (ln%o) with n% being the percentage of molecules removed from the training set). The statistical parameters $R_{\ln\%o}$ and $S_{\ln\%o}$ (correlation coefficient and standard deviation of leave-many-out) measure the stability of the QSPR upon inclusion/exclusion of molecules. The number of cases for random data removal analyzed is 50000.

The Y-randomization procedure [35] was applied in order to verify the model robustness and to avoid the development of spurious or false positive correlations. This technique depends on scrambling the experimental property values in such a way that they do not correspond to the respective compounds. After analyzing 10000 cases of Y-randomization, the standard deviation obtained (S^{rand}) has to be a poorer value than the one found by considering the true calibration (*S*).

2.3.3. Applicability domain analysis

The applicability domain (AD) of the QSPR model was also explored, since not even the best predictive model is expected to reliably predict the modeled property for the whole universe of molecules. The AD is a theoretically defined area that depends on the descriptors and the experimental property [36]. Only the molecules falling within this AD are not considered model extrapolations. One possible way to characterize the AD is based on the leverage approach [37], which allows one to verify whether a given compound can be considered as interpolated (with reduced uncertainty, thus a reliable prediction) or extrapolated outside the domain (unreliable prediction). Each compound *i* has a calculated leverage value (h_i) and there exists a warning leverage value (h^*). These parameters are defined in Table 2S. When $h_i > h^*$ for a test set compound, then a warning should be given, which means that the prediction is the result of substantial extrapolation of the model and could not be treated as reliable.

2.3.4. Degree of contribution of selected descriptors

In order to find out the relative importance of the *j*th descriptor in the linear model, we standardized its regression coefficient (b^s_j, see Table 2S). The larger the absolute value of b^s_j, the greater the importance of such a descriptor [38].

2.4. Software

Open Babel [23] was used to obtain the simplified molecular input line entry system (SMILES notations). HyperChem [25] was used for both molecular design and optimization. Molecular descriptors were calculated by means of Dragon version 6 [26]. Partition of the data set based on *k*-MCA, variable selection by means of RM, model fitting and validation were carried out in MatLab [39], by using toolboxes and functions written by the authors.

3. Results and discussion

As a first step, we applied the *k*-MCA clustering-based procedure for splitting the data set of 1184 compounds into $N_{train} = 395$, $N_{val} = 396$, and $N_{test} = 393$ set compounds (refer to Table 1S), thus ensuring a design with balanced sets of compounds. The $N_{train} = 395$ and $N_{val} = 396$ cluster centroid locations, in terms of descriptor values that minimize the squared sum of Euclidean distances of compounds located within them are provided in two matrices, respectively, as the C1.txt and C2.txt files from the Supplementary Material.

The RM variable subset selection method provides a way to explore a pool containing (a) 2895 molecular descriptors of all the blocks, (b) 1815 non-conformational descriptors, and (c) 1080 3D-descriptors. In this way, we intended to recognize whether 3D-descriptors are really important for modeling the *I* parameter measured in the polar Carbowax 20M column. Tables 1 and 2

Table 1
The best QSPR models obtained by considering all descriptor blocks. The chosen result appears in bold.

d	R^2_{train}	$RMSD_{train}$	R^2_{val}	$RMSD_{val}$	$R^2_{ij\ max}$	molecular descriptors
1	0.72	232.3	0.69	222.0	0.00	SCBO
2	0.81	189.9	0.80	175.7	0.11	SCBO, Hy
3	0.85	170.0	0.85	155.4	0.11	SCBO, C-001, Hy
4	0.88	153.2	0.88	135.6	0.87	Sv, RDF010p, CATS2D_02_AA, Hy
5	0.89	146.6	0.90	127.8	0.88	Sp, SM6_B(s), MATS2e, RDF010e, nHDon
6	**0.91**	**134.9**	**0.91**	**119.4**	**0.88**	**Sp, SM5_B(s), RDF010e, nHDon, C-005, CATS2D_04_DA**
7	0.90	138.0	0.91	119.9	0.91	Sp, SM5_B(s), MATS2e, RDF010p, Mor18s, Hy, PDI
8	0.92	127.9	0.92	116.1	0.91	Sp, EE_B(s), MATS2e, RDF010p, C-005, CATS2D_04_DA, Hy, PDI
9	0.91	134.0	0.92	112.2	0.35	SCBO, SM1_Dz(m), ATSC2s, E1s, nR = Cp, nHDon, C-001, NdssC, CATS2D_04_DA
10	0.91	129.0	0.92	114.5	0.38	SCBO, P_VSA_s_3, R1m+, nR = Cp, nArOR, C-001, H-050, O-058, O-060, CATS2D_04_DA

Table 2
The best QSPR models obtained by non-conformational descriptor blocks. The chosen result appears in bold.

d	R^2_{train}	$RMSD_{train}$	R^2_{val}	$RMSD_{val}$	$R^2_{ij\ max}$	molecular descriptors
1	0.72	232.3	0.69	222.0	0.00	SCBO
2	0.81	189.9	0.80	175.7	0.11	SCBO, Hy
3	0.85	170.0	0.85	155.4	0.11	SCBO, C-001, Hy
4	0.86	164.3	0.87	146.1	0.84	Sv, P_VSA_m_1, C-002, Hy
5	0.87	157.8	0.89	133.2	0.94	Se, P_VSA_m_1, NssCH2, CATS2D_04_DA, CATS2D_02_AA
6	0.89	143.7	0.89	131.2	0.86	H_Dz(v), SM3_B(p), SM6_B(s), C-001, SdCH2, Hy
7	**0.90**	**137.6**	**0.91**	**119.8**	**0.85**	**Se, Chi_H2, SM1_Dz(e), WiA_Dz(p), nHDon, SaaO, CATS2D_04_DA**
8	0.91	132.0	0.91	117.4	0.94	piID, XMOD, SpAD_B(m), SpDiam_B(s), C-001, CATS2D_04_DA, B02[O-O], Hy
9	0.91	134.2	0.92	115.0	0.71	SCBO, Xindex, J_Dz(p), P_VSA_m_2, SpMAD_EA(dm), C-001, SdCH2, CATS2D_04_DA, Hy
10	0.92	121.0	0.92	112.0	0.72	DBI, H_Dz(Z), J_Dz(p), SM1_B(s), ATSC2s, SpAD_EA(dm), C-001, C-015, H-050, CATS2D_04_DA

summarize the best MLR models found having 1-10 descriptors considering all the blocks of descriptors and non-conformational ones, respectively. It is appreciated that the $RMSD_{train}$ and $RMSD_{val}$ parameters do not have a significant variation between models of the same size (d). On the other hand, the 3D-molecular descriptors did not produce good models (refer to Table 3S). In fact, when 3D-descriptors were considered, we needed a ten-parameter model in order to achieve similar results with the seven-parameter non-conformational model. The meaning for each descriptor used in the current study is given in Table 4S. In this way, we decided to use the seven-parameter non-conformational model due to its simplicity compared with the six-parameter model that considered all families of descriptors.

This study demonstrates again that 3D-descriptors can be avoided for modeling the retention index on a Carbowax 20M column. The authors consider this to be an important finding since we avoid the loss of predictive capability of the QSPR associated with ambiguities resulting from an incorrect geometry optimization owing to the existence of compounds in various conformational states. In addition, conformational models involve high computational costs and long times for the calculation of the optimum molecular geometry. In fact, Doweyko [40] established three potential problems when 3D-descriptors are used to develop QSAR models. The first one indicates the lack of relationship between the predictive squared correlation coefficient (Q^2) and the predictive capability of the model. The second one explains the different features due to the alignment paradigm chosen. The third potential problem is the use of descriptors that could provide both a model with uninterpretable correlations and the over-description of the system.

For the selection of the optimal model for the two data sets, we considered simultaneously all quality parameter models given by the RM which are: R^2_{train}, R^2_{val}, $RMSD_{train}$, $RMSD_{val}$, $R^2_{ij\ max}$, and d. The best model satisfied the highest values for the squared correlation coefficient (R^2), and the lowest values for both the root mean squared deviation ($RMSD$) and the maximum correlation coefficient among descriptors ($R^2_{ij\ max}$). In addition, the number of descriptors d was kept as small as possible according to the principle of parsimony (Ockham's razor) [41], in order to avoid any possible unwarranted increased correlation between descriptors. Tables 1 and 2 summarize model results for all descriptor blocks and non-conformational descriptors, respectively. Thus, we chose the following seven-descriptor structure–retention relationship that includes non-conformational descriptors:

$$I = -141.9 - 23.6\,Se + 826.5\,Chi\_H2 + 1042.3\,SM1\_Dz(e) - 316.6\,WiA\_Dz(p) + 317.8\,nHDon - 36.1 SaaO - 161.4\,CATS2D\_04\_DA \quad (1)$$

$$N_{train} = 395,\ R^2_{train} = 0.902,\ S_{train} = 137.600,$$

$$F = 507.242,\ R^2_{ij\,max} = 0.852$$

$$o(3S) = 9,\ R^2_{loo} = 0.894,\ S_{loo} = 142.874,\ R^2_{l20\%o} = 0.938,$$

$$S_{l20\%o} = 152.276,\ S^{rand} = 423.006$$

$$N_{val} = 396,\ R^2_{val} = 0.910,\ S_{val} = 119.829$$

$$N_{test} = 393,\ R^2_{test} = 0.904,\ S_{test} = 121.978$$

where, F is the Fisher parameter and $R^2_{ij\ max}$ denotes the maximum squared correlation coefficient between descriptor pairs, $o(3S)$ indicates the number of outlier compounds having a residual (difference between experimental and calculated property) greater than three-times S_{train}.

This model is predictive using the external test set: the percentages of explained variances are $R^2_{train} = 90\%$, $R^2_{val} = 91\%$ and $R^2_{test} = 90\%$. In addition, the root mean square deviations are: $RMSD_{train} = 137.600$, $RMSD_{val} = 119.829$ and $RMSD_{test} = 121.978$. The established QSPR also shows the internal validation process of cross-validation through the exclusion of one molecule at a time and also by excluding 20% of them (79 molecules). The Y-randomization procedure demonstrates that $S_{train} < S^{rand}$ (423.006), and thus a valid structure–property relationship is achieved. We established that Eq. 1 accomplished the validation

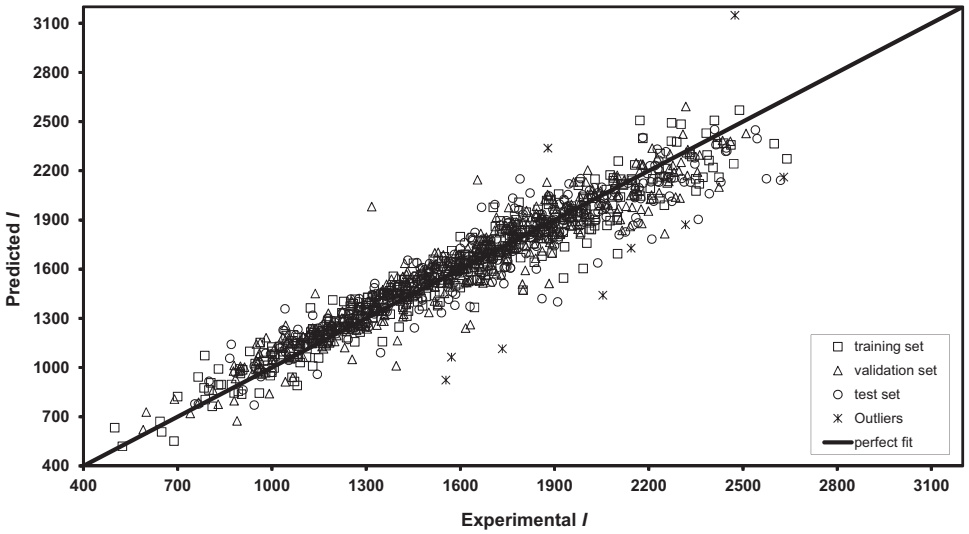


Fig. 1. Experimental versus predicted retention index according to the QSPR model for Carbowax 20M column.

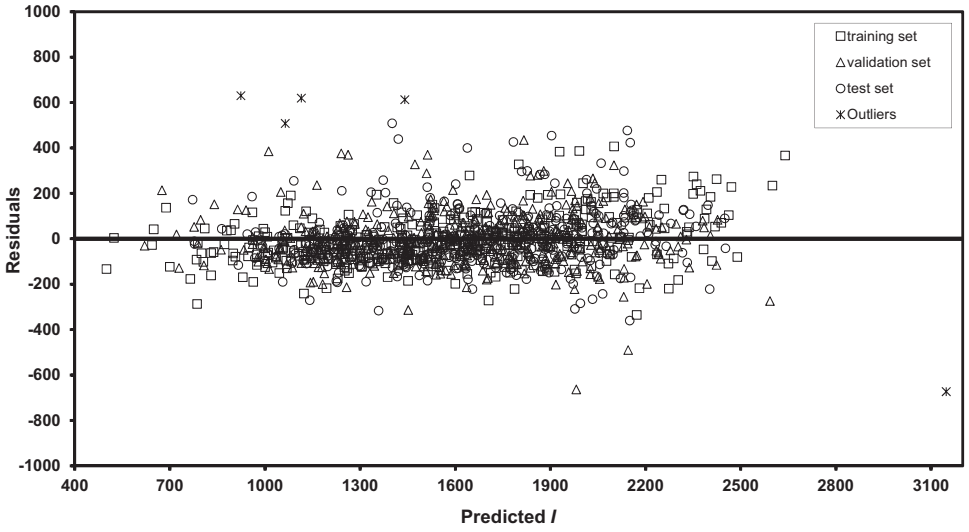


Fig. 2. Dispersion plot of residuals for the QSPR model.

criteria suggested by Golbraikh and Tropsha to avoid the proposal of overoptimistic, erroneously called, "predictive" QSPR model [42]:

$$R^2_{loo} > 0.5 \quad (0.894)$$

$$R^2_{test} > 0.6 \quad (0.904)$$

$$1 - \frac{R^2_0}{R^2_{test}} < 0.1\,(0.000) \quad \text{or} \quad 1 - \frac{R'^2_0}{R^2_{test}} < 0.1\,(0.013)$$

$$0.85 \le k(0.996) \le 1.15 \quad \text{and} \quad 0.85 \le k'(0.998) \le 1.15$$

$$R^2_m > 0.5 \quad (0.898)$$

These parameters are also defined in Table 2S.

Fig. 1 plots the predicted *I* as a function of the experimental values for the training, validation and test sets (numerical data is provided in Table 5S), revealing that the assumption of the MLR is fulfilled, i.e., there exists a tendency for the points to have a straight line trend. On the other hand, the dispersion plot of the residuals, presented in Fig. 2, shows that the points follow a random pattern around the zero line. Nine compounds exist from Eq. 1 that show residues outside the 3S limit value (412.8). These are: **musk xylol** (673.1), **dimethyl sulfoxide** (630.1), **methyl isovalerate** (619.5), **methyl 2-hydroxyisobutyrate** (612.9), **diethyleneglycol diethyl ether** (508.2), **6-methylcoumarin** (468.2), **jasmal** (458.8), **γ-dodecalactone** (445.3), and **δ-decalactone** (415.7). After an exhaustive control analysis of these compounds at the source, we are confident that their molecular structures and *I* values are correct. Hence, we can assume that this irregular behavior may be attributed either to un-controlled analytical aspects during the gas-chromatography technique or the wide structural diversity of the molecules considered in the analyzed data set.

Among the descriptors appearing in the QSPR model, there was one Constitutional index: sum of atomic Sanderson electronegativities (scaled on Carbon atom) (*Se*), one Functional group counts: number of donor atoms for H-bonds (N and O) (*nHDon*), one Atom-type E-state index: Sum of aaO E-states (*SaaO*), one CATS 2D: CATS2D Donor-Acceptor at lag 04 (*CATS2D_04_DA*), while there were three 2D matrix-based descriptors: Randić-like index from reciprocal squared distance matrix (*Chi_H2*), spectral moment of order 1 from Barysz matrix weighted by Sanderson electronegativity (*SM1_Dz(e)*), and average Wiener-like index from Barysz matrix weighted by polarizability (*WiA_DZ(p)*). Such descriptors selected by the RM technique were considered appropriate to study

Table 3
A comparison of various QSPR models taken from the literature for the Carbowax 20M column.

Reference	Compound family	Number of compounds	Number of descriptors	R^2_{train}	$RMSD_{train}$	R^2_{test}	$RMSD_{test}$
[15]	Substituted pyrazine	107	6	0.986	33.72	0.986	51.59
[16]	Odor compounds	111	7	0.994	17.1	–[a]	–[a]
[17]	Mono- and di-substituted pyrazine	74	4	0.990	40.3	–[a]	–[a]
[8]	Hydrocarbon mixture composed of alkylbenzenes, styrene derivatives, and associated aromatic acids and nonaromatic ringed compounds	81	5	0.974	23.3	–[a]	–[a]
	Alkylbenzenes	40	3	0.988	13.0	–[a]	–[a]
[9]	Alkylbenzenes	16	1	0.975	9.6	–[a]	–[a]
[10]	Alkylbenzenes (MLR)	150	6	0.982	18.0	–[a]	21.8
	Alkylbenzenes (ANN)			–[a]	11.7	–[a]	18.9
[11]	Alkylbenzenes (MLR)	165	7	0.875	44.76	–[a]	–[a]
	Alkylbenzenes (NMLR)			0.970	23.08	–[a]	–[a]
	Alkylbenzenes (ANN1)			0.973	21.79	–[a]	–[a]
	Alkylbenzenes (ANN2)			0.979	19.46	–[a]	–[a]
[12]	Alkylbenzenes (MLR)	170	8	0.968	26.08	0.985	18.19
	Alkylbenzenes (NMLR)			0.989	15.53	0.988	16.59
	Alkylbenzenes (ANN)			0.993	12.57	0.993	13.24
[19]	Acyclic and monocyclic terpenes (MLR)	53	6	0.774	–[a]	0.504	–[a]
	Acyclic and monocyclic terpenes (ANN)			0.941	–[a]	0.884	–[a]
[13]	Alkylbenzenes	–[a]	3	0.992	17.60	–[a]	–[a]
[14]	Alkylbenzenes	34	1	0.947	25.1	–[a]	–[a]
[18]	Pyrazines	35	3	0.985	–[a]	0.983	–[a]
[20]	Flavor compounds	434	6	0.923	104.24	0.926	105.48
This work	Flavors and Fragrances	1184	7	0.902	137.60	0.904	121.978

[a] Not available.

the I of these compounds measured on the Carbowax 20M column. The maximum squared correlation coefficient between the *Se* and *Chi_H2* descriptors was $R^2_{ij\ max} = 0.852$ (see correlation matrix in Table 6S). This value reflected a moderate correlation between such descriptors, which indicated that they were not collinear. Each one included different aspects of the molecular structure that succeeded in combining with the remaining variables of Eq. 1 [29]. The numerical values given by the seven descriptors are included in Table 7S.

The relative degree of contribution of each descriptor (b^s_j) revealed that the Randić-like index from reciprocal squared distance matrix had the greatest importance in the equation: *Chi_H2* (1.97) > *WiA_Dz(p)* (0.97) > *Se* (0.51) > *SM1_Dz(e)* (0.38) > *nHDon* (0.30) > *CATS2D_04_DA* (0.13) > *SaaO* (0.10). Furthermore, three descriptors took positive numerical values of the coefficients, indicating that such variables had a synergistic effect on the I value, i.e., higher values for the four descriptors for a given compound would lead to a higher predicted I value. On the other hand, there were four descriptors that had negative values of the coefficients, suggesting that they had an antagonist influence over the prediction of this property, i.e., higher values of these descriptors for a given compound would be reduced to a lower value for this property.

The terms, *WiA_DZ(p)* and *SM1_Dz(e)* are descriptors calculated from the Barysz matrix weighted by the atomic polarizability and the Sanderson electronegativity, respectively; which considered contemporarily the presence of multiple bonds and heteroatoms in a molecule. *Se* is an atomic property calculated as the sum of the relationships between the electronegativity values of each atom with respect to carbon. This descriptor indicates how the bonding electrons will become distributed between these two atoms when they are connected by a chemical bond.

The term *nHDon* is a measure of the hydrogen-bonding capability or hydrogen-bond donors of a molecule expressed as the sum of the hydrogens bonded to electronegative atoms such as nitrogen or oxygen without a negative charge in the molecule.

The term *CATS2D_04_DA* is a descriptor in which the assignment of the atoms is made by the hydrogen-bond donor and hydrogen-bond acceptor pair as a potential pharmacophore point pair at a topological distance of 4. *SaaO* is a molecular descriptor that describes the presence or absence of the aromatic-oxygen-aromatic (*aaO*) E-states as their sum. In addition, the structural information regarding the electron accessibility provided by this atom-type is taken into account.

In terms of the above-mentioned explanation these six molecular descriptors are correlated to the presence/absence of the electric charge in the molecule, i.e., they have an influence on the cohesion energy density of such compounds. The solubility parameter calculated as the square root of cohesive energy density has been widely employed in order to correlate polymer solvent interactions (Hildebrands theory) [43].

On the other hand, the Randić connectivity index was the first one proposed [44] to measure the degree of branching of saturated hydrocarbon skeletons, which was well correlated with chromatographic retention times as well as other physicochemical properties such as boiling point, enthalpies of formation, etc. In fact, molecules containing a high degree of branching exhibit highest values of the mixture entropy between the molecule and the stationary phase. Consequently, due to the mixture stability, there is a need to increase the temperature in order to return the compound to the mobile-phase, and, as a result, the I value

increases. This is in agreement with the modified Flory-Huggins equation [45,46] which represents approximately the entropy of the interaction between the compound and the stationary-phase.

This descriptor was in close agreement to the second Zagreb index proposed by Bollobás and Erdös [47] as a general Randić index. In addition, Randić-like indices were calculated from the reciprocal squared topological distances between any pair atoms from an H-depleted skeleton, i.e., from the reciprocal squared distance matrix (H2) considering a generalization of the classical formula of the Randić connectivity index based on the vertex degree [48].

Table 4
Predicted retention index and leverage values for the 22 flavor and aroma molecules not provided in the data set.

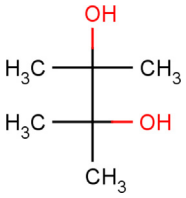
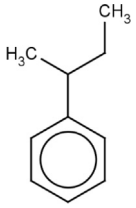
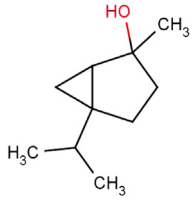
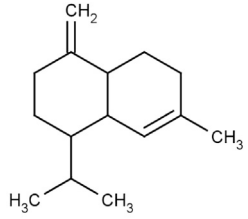
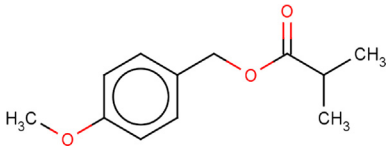
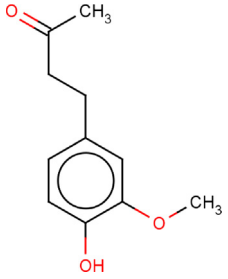
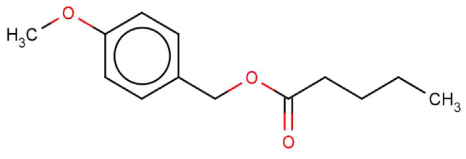
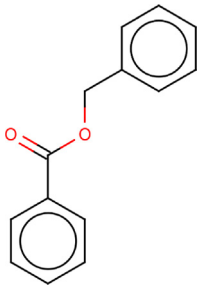
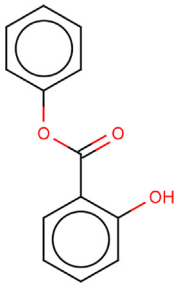
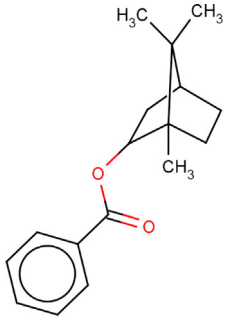
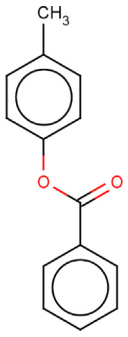
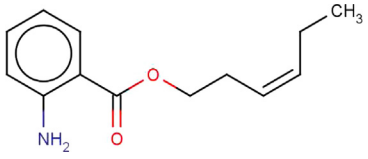
Name	Molecule	h_i	Predicted *I*	Reported *I*
Pinacol		0.0812[a]	1510	1338[b] [60,61]
sec-Butyl benzene		0.0142	1427	1227 and 1248 [61]
trans-Sabinene hydrate		0.0214	1676	1458[c] [53]
				1548.9 [62]
γ-Muurolene		0.0246	1709	1675 [58]
				1681 [53]
				1689.8[d] [62]
Anisyl isobutyrate		0.0096	2016	–[e]
Zingerone		0.0227	2551	–[e]
Anisyl n-valerate		0.0121	2189	2400 [61]

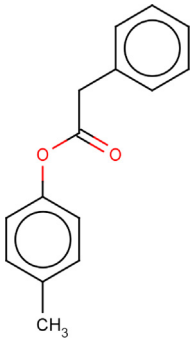
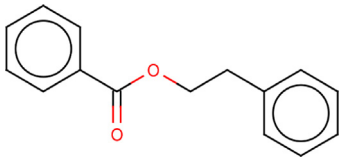
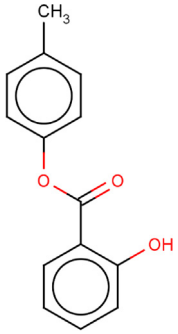
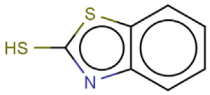
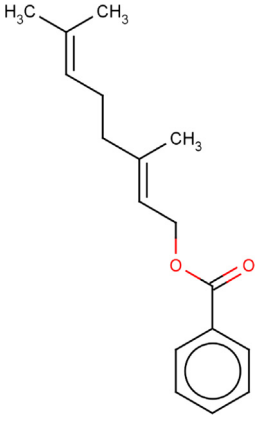
Table 4 (*Continued*)

Name	Molecule	$\mathbf{h_i}$	Predicted *I*	Reported *I*
Benzyl benzoate		0.0383	2474	2071[c] [53] 2612.7[d] [62] 2655 [61]
Phenyl salicylate		0.0976[a]	2593	–[*]
Bornyl benzoate		0.0066	2294	2114[c] [53]
p-Cresyl benzoate		0.0298	2424	–[*]
cis-3-hexenyl anthranilate		0.1084[a]	2583	–[*]

The molecular descriptor information encoded in the adjacency matrix (A) and the reciprocal squared distance Matrix (H2) is very similar when a data set involves very heterogeneous molecules of different sizes. In fact, in a recent study [21], we demonstrated the use of the solvation connectivity index of first order (*X1sol*) to calculate the retention index for the OV-101 stationary phase. There were several other studies that used topological indexes to describe this property [49–51].

Table 3 presents a comparison of the model obtained in this work to similar ones taken from the literature. It is noted that when a QSRR model is built by considering only a group or a few groups of compounds having a similar structures, the results lead to a good correlation with R^2 close to one. The results are poorer when a greater number of compounds are considered having diverse chemical structures, although it is always possible to split the data set to perform an external validation [10,12,15,20].

Table 4 (*Continued*)

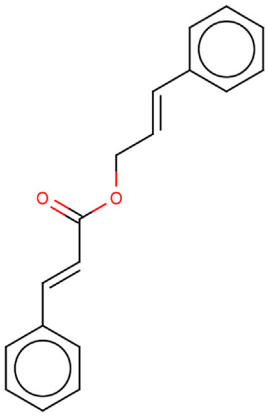
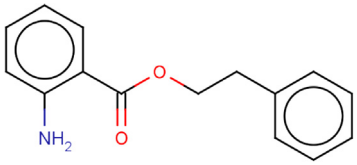
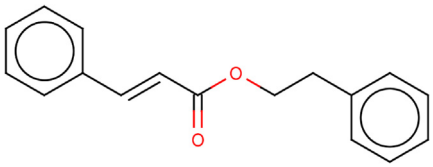
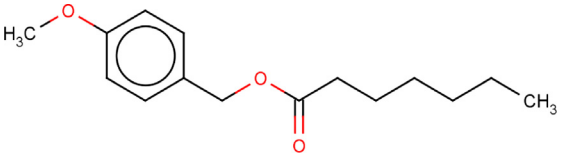
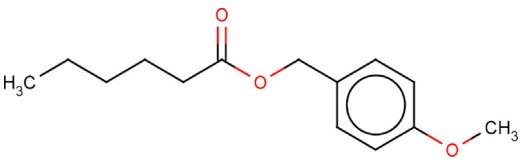
Name	Molecule	$\mathbf{h_i}$	Predicted *I*	Reported *I*
p-Cresyl phenylacetate		0.0311	2505	–[*]
Phenylethyl benzoate		0.0402	2540	2189[c] [53] 2654 [61]
p-Cresyl salicylate		0.0927[a]	2605	–[*]
2-Mercaptobenzothiazole		0.0276	2028	–[*]
Geranyl benzoate		0.0168	2432	2617[e] [61]

In fact, when a large data set is considered the quality parameters are slightly worse, i.e., the R^2 parameter decreases, meanwhile the *RMSD* value increases; however, the model shows a good predictive capability [20].

In our case, the data set was 2.7 times bigger than the largest one previously analyzed [20] but the parameters were very close for the training set and slightly different for the test set. In contrast, seven QSRR studies from Table 3 did not perform external validation [8,9,11,13,14,16,17] or did not present the parameters for the test set of the 81 hydrocarbons [8].

The model developed in the current work has a reasonable number of uncorrelated descriptors. This result is good for both calibration and external validation, and the model can be used for prediction purposes. Due to the wide range of chemical structures considered for building the QSPR model, compounds that can be predicted belong to the volatile families, such as aromatic hydrocarbons, alcohols, acids, ketones, aldehydes, ethers, acid esters, amines, etc., for which the leverage value should lie below the warning leverage value ($h^* = 0.0608$) established in the present model.

Table 4 (*Continued*)

Name	Molecule	$\mathbf{h_i}$	Predicted *I*	Reported *I*
Cinnamyl cinnamate		0.0678[a]	2987	–[*]
Phenylethyl anthranilate		0.1257[a]	2970	–[*]
Phenylethyl cinnamate		0.0541	2809	3100 [61]
Anisyl n-heptanoate		0.0135	2355	2600 [61]
Anisyl n-hexanoate		0.0126	2271	–[*]

[a] molecules with leverage value above the warning leverage $h^* = 0.0608$.
[b] measured on a DB-Wax polar column.
[c] Interpolated retention index.
[d] measured on a Polyethylene glycol (PEG) stationary phase.
[e] measured on a HP-Innowax FSC polar column.
[*] *I* value not found in the literature.

The AD of Eq. 1 reveals that fourteen compounds from the validation set (iso-amyl salicylate, iso-butyl salicylate, n-heneicosane, n-hexyl salicylate, methyl anthranilate, methyl salicylate, n-propyl n-methyl anthranilate, n-propyl salicylate, 1,3-butanediol, 1,4-butanediol, hydroxycitronellol, octyl salicylate, n-propyl anthranilate, n-amyl anthranilate) and twelve compounds from the test set (allyl salicylate, iso-butyl antranilate, ethyl anthranilate, cis-3-hexenyl salicylate, methyl n-methyl anthranilate, isopropyl n-methyl anthranilate, isopropyl salicylate, methyl 3-hydroxybutyrate, m-aminoacetophenone, isoamyl anthranilate, methyl n-propylanthranilate, n-butyl n-methylanthranilate) had leverage values over the warning leverage $h^* = 0.0608$ (refer to Table 8S). The majority of the compounds outside the AD are anthranilate and salicylate derivatives. The principle of AD obliges the researchers to define model limitations with respect to the molecular descriptors and the property space. In other words, any robust, validated and predictive QSPR model is not capable of producing a reliable prediction for the whole set of molecules [36,52]. Moreover, QSRR studies regarding GC responses are usually carried out by considering only data sets of homogeneous families of compounds, reducing their ability to generalize the AD of these models to other kinds of volatile compounds [8–15,17–19].

Using the model developed in this study, we predicted the *I* of 22 volatile compounds for which this property was not provided by Jennings and Shibamoto. Table 4 summarizes both the predicted and reported *I* as well as the leverage values for these compounds. There are six compounds that fall outside the applicability domain of the training set (their leverage values are bigger than the warning leverage of 0.0608), i.e., they are considered an extrapolation of the model. These compounds are: pinacol (0.0812), phenyl salicylate (0.0976), cis-3-hexenyl anthranilate (0.1084), p-cresyl salicylate (0.0927), cinnamyl cinnamate (0.0678), and phenylethyl

anthranilate (0.1257). Therefore, the prediction capability of the present model was used to estimate the *I* for the chemicals inside the applicability domain.

According to the results presented in Table 4, phenylethyl cinnamate presents the largest deviation (291 retention units) between the predicted and reported retention index. Other molecules also exhibit appreciable departures from the reported retention indexes, namely, anisyl n-heptanoate (245 retention units), anisyl n-valerate (211 retention units), sec-butyl benzene (186 retention units), geranyl benzoate (185 retention units), pinacol (172 retention units), trans-sabinene hydrate (127 retention units), and phenylethyl benzoate (114 retention units). The model is more accurate to predict the retention index in molecules such as γ-muurolene, with a deviation of 28 retention units, and benzyl benzoate (95 retention units). Interestingly, benzyl benzoate and phenylethyl benzoate present predicted *I* values that agree better with reported data than values reported in the flavornet database [53]. In contrast, six molecules (benzyl benzoate, phenylethyl benzoate, phenylethyl cinnamate, cinnamyl cinnamate, phenylethyl anthranilate, and phenylethyl cinnamate) show either reported or predicted *I* values above the maximum retention index of the training set (I_{max} = 2640). The relatively low accuracy in predicting retention indexes of Eq. 1 for the molecules reported in Table 4 may be attributed to the low reproducibility of retention index measured in different laboratories.

Low reproducibility in this case may be caused by several factors affecting GC analysis such as the polar interaction between the analyzed compound and the wall interfaces of the stationary phase. In addition, misidentification of the compounds, sensitivity of the GC-equipment, variation of the conditions in the retention index measurement (carrier gas, temperature conditions, detector parameters) and variations in the intrinsic properties of stationary phase (type, thickness of the stationary phase film, column dimensions and column age) may lower the reproducibility of the analysis. Also, properties of the sample (thermal stability and impurities), factors involved during the sample preparation and factors that affect the introduction of the sample in the GC-equipment (selection of the proper injection system, i.e., limited or excessive volume and split of the sample) may be involved. Sources of "spurious" errors (e.g., equipment malfunction or analyst error), error and variability associated with peak integration, as well as the normal experimental errors and data entry errors inherent in all studies may affect the reproducibility of the data [8,54–57].

Additionally, a few *I* values of the present data set have been compared with the corresponding ones reported for the Carbowax 20M [58] (refer to Table 9S), and the Innowax column [59] (refer to Table 10S). In the particular case of the elemol compound, it has the largest deviation (778 retention units) with respect to the corresponding one (see Table 10S). Additionally, in the data set published by Jennings and Shibamoto, compounds exist that exhibit two *I* values, e.g., n-butyl n-hexanoate (1402 and 2435 with a difference of 1033 retention units), and ethyl n-octanoate (1424 and 2460 with a difference of 1037 retention units). These discrepancies clearly reflect the influence of all the factors discussed above in predicting accurately the *I* property. In fact, Gramatica [36] claimed that the erroneous predictions could possibly be due to the incorrect experimental data. A way to ensure the identification and subsequent exclusion of erroneous *I* values for fragrances is to perform a data curation by comparing *I* values with respect to available data (e.g., the NIST standard reference database [60]).

All these factors described above contributed to the prediction of retention indices with relatively low accuracy; and therefore, our model can be considered valid as a complement to other studies to describe the retention index mechanism in the Carbowax 20M stationary phase, as well as to estimate *I* values of other compounds with a predictive ability of 90.4%. Finally, a way to achieve models with good prediction capability is to develop models for each family of compounds as described in Table 3, as well as to build models using an established range along the *I* values.

4. Conclusions

An application of the QSPR theory is presented for the modeling of the gas chromatographic retention index of 1184 flavor and fragrance compounds in the polar stationary capillary column Carbowax 20M. In this study, we developed a model with both acceptable stability in regression and predictive power on the test set, which can be in some cases used to predict the retention index for un-evaluated and un-synthesized flavors or fragrances. In fact, we have predicted the *I* for 22 aroma volatile compounds not previously reported, and a comparison with respect to the published ones is done. The actual QSPR model is not capable of making reliable prediction of the *I* property for anthranilate and salicylate derivatives, as well as some other compounds. The features affecting the prediction ability of the model were categorized into three groups: (1) factors affecting the reproducibility of the GC analytical technique, (2) incorrect experimental data of the retention index property, and (3) intrinsic model limitations defined by the molecular descriptors and the *I* property space. On the other hand, we have also established that the Randić-like index from reciprocal squared distance matrix is strongly correlated to the *I* as a synergistic effect. The model used here produces a more general quantitative structure–retention relationship, and complements previously reported results from the literature. Finally, 3D-molecular descriptors do not significantly improve the quality of the parameters of the QSPR model. According to these results, the conformation-independent QSPR method continues to emerge as an alternative approach for developing models based on constitutional and topological molecular features of compounds.

Acknowledgments

Cristian Rojas is grateful for his PhD Fellowship from the National Secretary of Higher Education, Science, Technology and Innovation (SENESCYT) from the Republic of Ecuador. Pablo R. Duchowicz wishes to thank the National Scientific and Technical Research Council of Argentina (CONICET) for the project grant PIP11220100100151, and the Minister of Science, Technology and Productive Innovation for the use of the electronic library facilities. Pablo R. Duchowicz and Reinaldo Pis Diez are members of the Scientific Researcher Career of CONICET.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.chroma.2015.10.028.

References

[1] K. Bauer, D. Garbe, H. Surburg, Common Fragrance and Flavor Materials. Preparation, Properties and Uses, forth ed., Wiley-VCH, Weinheim, 2001.
[2] T. Cserháti, Chromatography of Aroma Compounds and Fragrances, Springer, 2010.
[3] R. Kaliszan, QSRR: quantitative structure–(chromatographic) retention relationships, Chem. Rev. 107 (2007) 3212–3246.
[4] R. Kaliszan, Quantitative Structure–Chromatographic Retention Relationships, Wiley-Interscience, 1987.
[5] R. Kaliszan, Structure and Retention in Chromatography: A Chemometric Approach, Harwood Academic Publishers, 1997.
[6] K. Héberger, Quantitative structure–(chromatographic) retention relationships, J. Chromatogr. A 1158 (2007) 273–305.
[7] M. Righezza, A. Hassani, B.Y. Meklati, J.R. Chrétien, Quantitative structure–retention relationships (QSRR) of congeneric aromatics series studied on phenyl OV phases in gas chromatography, J. Chromatogr. A 723 (1996) 77–91.

[8] T.F. Woloszyn, P.C. Jurs, Prediction of gas chromatographic retention data for hydrocarbons from naphthas, Anal. Chem. 65 (1993) 582–587.
[9] N. Dimov, A. Osman, O. Mekenyan, D. Papazova, Selection of molecular descriptors used in quantitative structure-gas chromatographic retention relationships: I. Application to alkylbenzenes and naphthalenes, Anal. Chim. Acta 298 (1994) 303–317.
[10] J.M. Sutter, T.A. Peterson, P.C. Jurs, Prediction of gas chromatographic retention indices of alkylbenzenes, Anal. Chim. Acta 342 (1997) 113–122.
[11] A. Yan, G. Jiao, Z. Hu, B.T. Fan, Use of artificial neural networks to predict the gas chromatographic retention index data of alkylbenzenes on Carbowax-20M, Comp. Chem. 24 (2000) 171–179.
[12] A. Yan, Z. Hu, Linear and non-linear modeling for the investigation of gas chromatography retention indices of alkylbenzenes on Cit.A-4, SE-30 and Carbowax 20M, Anal. Chim. Acta 433 (2001) 145–154.
[13] C. Zhou, C. Nie, Modeling quantitative structure property relationships with a semi-empirical topological descriptor and path numbers, Chromatographia 66 (2007) 545–554.
[14] L.C. Porto, É.S. Souza, B. da Silva Junkes, R.A. Yunes, V.E.F. Heinzen, Semi-Empirical Topological Index: development of QSPR/QSRR and optimization for alkylbenzenes, Talanta 76 (2008) 407–412.
[15] D.T. Stanton, P.C. Jurs, Computer-assisted prediction of gas chromatographic retention indexes of pyrazines, Anal. Chem. 61 (1989) 1328–1332.
[16] L.S. Anker, P.C. Jurs, P.A. Edwards, Quantitative structure–retention relationship studies of odor-active aliphatic compounds with oxygen-containing functional groups, Anal. Chem. 62 (1990) 2676–2684.
[17] P.A. Edwards, L.S. Anker, P.C. Jurs, Quantitative structure–property relationship studies of the odor threshold of odor active compounds, Chem. Senses 16 (1991) 447–465.
[18] I. Touhami, K. Mokrani, D. Messadi, Modèles QSRR hybrides algorithme génétique-régression linéaire multiple des indices de rétention de pyrazines en chromatographie gazeuse, Lebanese Sci. J. 13 (2012) 75–88.
[19] M. Jalali-Heravi, M. Fatemi, Artificial neural network modeling of Kovats retention indices for noncyclic and monocyclic terpenes, J. Chromatogr. A 915 (2001) 177–183.
[20] J. Yan, D.-S. Cao, F.-Q. Guo, L.-X. Zhang, M. He, J.-H. Huang, Q.-S. Xu, Y.-Z. Liang, Comparison of quantitative structure–retention relationship models on four stationary phases with different polarity for a diverse set of flavor compounds, J. Chromatogr. A 1223 (2012) 118–125.
[21] C. Rojas, P.R. Duchowicz, P. Tripaldi, R. Pis Diez, QSPR analysis for the retention index of flavors and fragrances on a OV-101 column, Chemom. Intell. Lab. Syst. 140 (2015) 126–132.
[22] W. Jennings, T. Shibamoto, Qualitative Analysis of Flavor and Fragrance Volatiles by Glass Capillary Gas Chromatography, Academic Press, Inc., London, 1980.
[23] Open Babel: The Open Source Chemistry Toolbox, Open Babel, http://openbabel.org/.
[24] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, Weinheim, 2009.
[25] HyperChem, Hypercube, Inc., http://www.hyper.com.
[26] TALETE, srl., Dragon (version 6), Software for Molecular Descriptor Calculation, http://www.talete.mi.it/, 2014.
[27] B. Lučić, N. Trinajstic, S. Sild, M. Karelson, A.R. Katritzky, A new efficient approach for variable selection based on multiregression: prediction of gas chromatographic retention times and response factors, J. Chem. Inf. Comp. Sci. 39 (1999) 610–621.
[28] P.R. Duchowicz, E.A. Castro, F.M. Fernández, M.P. González, A new search algorithm of QSPR/QSAR theories: normal boiling points of some organic molecules, Chem. Phys. Lett. 412 (2005) 376–380.
[29] P.R. Duchowicz, J.J. Marrugo, H.R. Vivas-Reyes, E.A. Castro, QSPR applied on gas chromatography indices of polycyclic aromatic compounds, Int. J. Environ. Sci. (IJES) 1 (2010) 73–77.
[30] A. Miller, Subset Selection in Regression, CRC Press, 2012.
[31] T.M. Martin, P. Harten, D.M. Young, E.N. Muratov, A. Golbraikh, H. Zhu, A. Tropsha, Does rational selection of training and test sets improve the outcome of QSAR modeling? J. Chem. Inf. Model. 52 (2012) 2570–2578.
[32] F. Ros, O. Taboureau, M. Pintore, J.R. Chrétien, Development of predictive models by adaptive fuzzy partitioning. Application to compounds active on the central nervous system, Chemom. Intell. Lab. Syst. 67 (2003) 29–50.
[33] D. Dimić, A.G. Mercader, E.A. Castro, Chalcone derivative cytotoxicity activity against MCF-7 human breast cancer cell QSAR study, Chemom. Intell. Lab. Syst. 146 (2015) 378–384.
[34] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 2005.
[35] C. Rücker, G. Rücker, M. Meringer, Y-Randomization and its variants in QSPR/QSAR, J. Chem. Inf. Model. 47 (2007) 2345–2357.
[36] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701.
[37] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environ. Health Perspect. 111 (2003) 1361–1375.
[38] N.R. Draper, H. Smith, Applied Regression Analysis, 1981, New York.
[39] The MathWorks, Inc., MatLab, Masachussetts, USA, http://www.mathworks.com.
[40] A.M. Doweyko, 3D-QSAR illusions, J. Comp.-Aided Mol. Design 18 (2004) 587–596.
[41] R. Hoffmann, V.I. Minkin, B.K. Carpenter, Ockham's razor and chemistry, Bull. Soc. Chim. France 133 (1996) 117–130.
[42] A. Golbraikh, A. Tropsha, Beware of q2!, J. Mol. Graphics Modell. 20 (2002) 269–276.
[43] D.W. Van Krevelen, K. Te Nijenhuis, Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions, Elsevier, 2009.
[44] M. Randić, Characterization of molecular branching, J. Am. Chem. Soc. 97 (1975) 6609–6615.
[45] C. Qian, S.J. Mumby, B. Eichinger, Phase diagrams of binary polymer solutions and blends, Macromolecules 24 (1991) 1655–1661.
[46] Y. Bae, J. Shim, D. Soane, J. Prausnitz, Representation of vapor–liquid and liquid–liquid equilibria for binary systems containing polymers: applicability of an extended Flory–Huggins equation, J. Appl. Polym. Sci. 47 (1993) 1193–1206.
[47] B. Bollobás, P. Erdös, Graphs of extremal weights, Ars Combinatoria 50 (1998) 225–234.
[48] V. Consonni, R. Todeschini, Multivariate analysis of molecular descriptors, in: M. Dehmer, K. Varmuza, D. Bonchev (Eds.), Statistical Modelling of Molecular Descriptors in QSAR/QSPR, Wiley-Blackwell, 2012, pp. 111–147.
[49] L. Buydens, D.L. Massart, Prediction of gas chromatography retention indexes from linear free energy and topological parameters, Anal. Chem. 53 (1981) 1990–1993.
[50] L. Buydens, D.L. Massart, P. Geerlings, Prediction of gas chromatographic retention indexes with topological, physicochemical, and quantum chemical parameters, Anal. Chem. 55 (1983) 738–744.
[51] M. Pompe, M. Novič, Prediction of gas-chromatographic retention indices using topological descriptors, J. Chem. Inform. Comput. Sci. 39 (1998) 59–67.
[52] S. Dimitrov, G. Dimitrova, T. Pavlov, N. Dimitrova, G. Patlewicz, J. Niemela, O. Mekenyan, A stepwise approach for defining the applicability domain of SAR and QSAR models, J. Chem. Inf. Model. 45 (2005) 839–849.
[53] T. Acree, H. Arn, Flavornet and Human Odor Space, 2004, Available from: http://www.flavornet.
[54] L. Soják, V.G. Berezkin, J. Janák, Effect of adsorption on the reproducibility of retention indices of hydrocarbons in capillary gas–liquid chromatography, J. Chromatogr. A 209 (1981) 15–20.
[55] R. Wittkowski, R. Matissek, Capillary Gas Chromatography in Food Control and Research, Technomic Publishing Co., Inc., 1993.
[56] V.J. Barwick, Sources of uncertainty in gas chromatography and high-performance liquid chromatography, J. Chromatogr. A 849 (1999) 13–33.
[57] V. Babushok, P. Linstrom, J. Reed, I. Zenkevich, R. Brown, W. Mallard, S. Stein, Development of a database of gas chromatographic retention properties of organic compounds, J. Chromatogr. A 1157 (2007) 414–421.
[58] C. Bicchi, P. Rubiolo, E.E. Saranz Camargo, W. Vilegas, J. de Souza Gracioso, A.R. Monteiro Souza Brito, Components of *Turnera diffusa* Willd. var. *afrodisiaca* (Ward) Urb. essential oil, Flavour Fragr. J. 18 (2003) 59–61.
[59] N. Tabanca, B. Demirci, T. Ozek, N. Kirimer, K.H.C. Baser, E. Bedir, I.A. Khan, D.E. Wedge, Gas chromatographic–mass spectrometric analysis of essential oils from *Pimpinella* species gathered from Central and Northern Turkey, J. Chromatogr. A 1117 (2006) 194–205.
[60] P.J. Linstrom, W.G. Mallard, NIST Chemistry WebBook, NIST Standard Reference Database Number 69, National Institute of Standards and Technology, Gaithersburg, MD, 2001.
[61] H. Shiratsuchi, M. Shimoda, K. Imayoshi, K. Noda, Y. Osajima, Off-flavor compounds in spray-dried skim milk powder, J. Agric. Food Chem. 42 (1994) 1323–1327.
[62] V.I. Babushok, P.J. Linstrom, I.G. Zenkevich, Retention indices for frequently reported compounds of plant essential oils, J. Phys. Chem. Ref. Data 40 (2011) 043101.