Chemometrics and Intelligent Laboratory Systems 140 (2015) 126–132

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

QSPR analysis for the retention index of flavors and fragrances on a OV-101 column

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

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

ARTICLE INFO

Article history:
Received 27 May 2014
Received in revised form 17 August 2014
Accepted 20 September 2014
Available online 21 November 2014

Keywords:
Flavors and fragrances
OV-101 column
QSPR theory
Dragon software, Replacement method
k-Means cluster analysis, Total-order ranking, DART

ABSTRACT

A predictive quantitative structure–property relationships (QSPR) is developed for modeling the retention index measured on the OV-101 glass capillary gas chromatography column, in a set of 1208 flavor and fragrance compounds. The 4885 molecular descriptors are calculated using the Dragon software and then are simultaneously analyzed through multivariable linear regression analysis using the replacement method (RM) variable subset selection technique. We proceed in three steps, the first one by considering all descriptor blocks, the second one by excluding conformational descriptors blocks, and the last one by analyzing only 3D-descriptors families. The models are properly validated through an external test set of compounds. Cross-validation methods such as leave-one-out and leave-more-out are applied, together with Y-randomization and applicability domain analysis. The results clearly suggest that 3D-descriptors do not offer relevant information for modeling the retention index, while a topological index such as the solvation connectivity index of first order 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 chemical) or a pleasant taste (flavor chemical). A fragrance substance is used as a component in a perfume or a perfumed product, while a flavor substance is used as a flavoring or to enhance the flavor of beverages and food products. Chromatography techniques are generally used for analyzing the content and impurity of fragrances and flavors, as well as for quality control and in-process control, in order to provide details of their profiles in a few minutes.

In 1977, three publications have appeared for the first time on QSPR theory, or what is currently known as quantitative structure–retention relationships (QSRR) [2]. Subsequently, in 1987, a monograph is published containing several hundred of publications demonstrating that, in fact, QSRR methods represent a powerful tool in Chromatography data analysis [3,4]. 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 (RI) [5] 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 [6] 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 rational design new phases of defined properties [2].

Several QSPR studies have been published in past years for the prediction of the RI parameter on different stationary phase types. In 1989, Staton and Jurs [7] have used the ADAPT software to study a series of 107 substituted pyrazine compounds taken from the literature, with RI values being measured on methyl silicone OV-101 and poly(ethylene glycol) Carbowax 20M columns. Their results are reliable: $R^2 = 0.994$ for OV-101, and $R^2 = 0.986$ for Carbowax 20M. A year later, a similar study from Anker and Jurs [8] measure RI on 115 odor compounds (38 alcohols, 11 aldehydes, 19 ketones, and 47 esters) and establishes a QSPR, leading to $R^2 = 0.998$ for OV-101 and $R^2 = 0.992$ for Carbowax 20M.

In 2000, Héberger et al. [9] have built a quantitative structure–retention relationship for 35 aliphatic ketones and aldehydes, by means of the partial least squares (PLS) technique. The retention index is determined using HP-1, HP-50, DB-210, and HP-INNOWax capillary columns. The authors claim that ketones and aldehydes cannot be separated in two classes solely on the basis of RI data, whereas physical properties such as boiling point, molar volume, molecular weight, molar refraction, and the octanol-water partition coefficient contain the additional information for appropriately separating them into such classes.

In 2002, Liu et al. [10] have proposed the molecular electronegativity distance vector to describe the structure of 209 polycyclic aromatic

[*] Corresponding author. Tel.: +54 221 4257430; fax: +54 221 4254642.
E-mail address: crojasvilla@gmail.com (C. Rojas).

http://dx.doi.org/10.1016/j.chemolab.2014.09.020
0169-7439/© 2014 Elsevier B.V. All rights reserved.

hydrocarbons, in a programmed temperature SE-52 capillary-GC column, and establish a model by means of the multiple linear regression cross-validated technique. In this research, a better QSPR is achieved by removing 37 outliers ($R^2 = 0.987$).

In another study in 2007, Lu et al. [11] have analyzed the GC retention indices of 55 polyhalogenated biphenyl compounds through two 2D-descriptor types including the molecular electronegativity distance vector based on 13 atomic types (MEDV-13) and the atom-type electrotopological state (E-state) index. The resulting equation validated by leave-one-out cross-validation represents a very good model for calculating RI ($R^2 = 0.984$, and $R^2_{\text{loo}} = 0.975$). Goodner [12] in 2008 has developed a high-quality regression model utilizing the boiling point and the logarithm of the octanol–water partition coefficient, for several columns: OV-101 (91 molecules), DB-1 (57 molecules), DB-5 (94 molecules), and DB-Wax (102 compounds). In addition, a combination of several outlier tests are used, such as the Grubb's test, the Dixon's *Q*-test, and a modified “huge” rule.

In 2008, Dua et al. [13] have predicted the retention times on 43 constituents of saffron aroma using the best multi-linear regression (BMLR) and projection pursuit regression (PPR) methods, leading to acceptable results for both methods: $R^2_{\text{train}} = 0.943$, $R^2_{\text{test}} = 0.872$, and $R^2_{\text{train}} = 0.981$, $R^2_{\text{test}} = 0.946$, respectively. Finally, in a recent study of 2012, Yan et al. [14] have performed a QSRR for 1341 flavor compounds using four stationary phases of different polarities (OV-101, DB-5, OV-17, and Carbowax 20M). They select the following descriptors for such columns: ${}^0\chi$ (molecular connectivity index of zero-th order), ${}^1\chi$ (molecular connectivity index of first order), ndonr (number of hydrogen-bond donors), MW (molecular mass), DPSA1 (difference between partial positively and negatively charged surface areas), FPSA1 (fractional partial positive surface area), qhmax (maximum positive charge on hydrogen), μ (dipole moment), *ipc* (information content from the adjacent matrix), and ${}^4\chi_c$ (molecular connectivity index for four clusters). Model's stability and validity are tested by internal and external validation. The squared leave-one-out correlation coefficient (Q^2) show well-correlated models: 0.959, 0.953, 0.959, and 0.922 on stationary phase OV-101 (${}^0\chi$, ${}^1\chi$, ndonr, MW), DB-5 (MW, ${}^0\chi$, ndonr, ${}^1\chi$), OV-17 (DPSA1, FPSA1, qhmax, μ, ipc), and Carbowax 20M (ndonr, FPSA1, qhmax, μ, ${}^4\chi_c$), respectively. All these QSPR predictive models may be useful for the prediction of the RI parameter on flavor compounds when experimental data is unavailable.

Given these premises, the main purpose of the present work is to use a novel data set of 1208 flavor and fragrance volatile compounds having RI values being measured on the stationary phase methyl silicone OV-101 for developing a predictive QSPR model. This data set has not yet been used for QSPR studies. Following this purpose, a large set of molecular descriptors are calculated using the well-known Dragon software; besides that, the replacement method (RM) is used as a variable subset selection technique; then, the total-order ranking methodology implemented in DART software [15] is used for choosing the best model from a pool of models having 1 to 7 descriptors. For validation purposes, the data set is split intro training, validation, and test set following the k-means cluster analysis. Therefore, this research would allow us to obtain a model to predict RI values for both un-evaluated and un-synthesized flavors/fragrances, and thus it would be useful for people who work on aroma and flavor chemical synthesis.

2. Materials and methods

2.1. Experimental Data set

The chemical domain analyzed involves 1208 aromatic substances reported by Jennings and Shibamoto [16]. The experimental property reported by these authors is the Kováts retention index in the non-polar stationary capillary column (0.28 mm × 50 m), which is coated with methyl silicone OV-101, admixed with 1 % Carbowax 20M as an antitailing additive, and programmed from 80 to 200 °C at 2 °C min^{-1}.

The RI values vary in the range from 350 to 2180. The chemical names, simplified molecular input line entry system (SMILES) notations as obtained with the Open Babel software [17], and RI values are presented in Table S1. When a molecule has two or more RI values the average value is 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 characterization. Descriptors are the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into an useful number or the result of some standardized experiment [18]. In this way, the compounds are first drawn using HyperChem for Windows [19]. For geometry optimization, the molecular mechanics force field (MM+) is used, followed by the PM3 semi-empirical method to refine the structures. The conjugate gradient algorithm, in the Polak–Ribiere version, is used for the optimizations. The geometries are considered optimized when the root mean square deviation of the gradient vector becomes less than 0.01 kcal.(Å $\text{mol})^{-1}$. Geometry optimizations are also carried out with HyperChem. We compute 4885 molecular descriptors of all types using Dragon version 6.0 [20]. This well-known descriptor database includes twenty-nine descriptor families: 0D-descriptors (constitutional indices), 1D-descriptors (functional group counts, atom-centred 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) [18]. The first data set is composed of descriptors belonging to all the blocks. In the second data set, only non-conformational molecular descriptors are considered. Finally, in the third data set, we only consider 3D-descriptors. This is done in order to compare whether 3D-molecular descriptors play an important role for modeling the RI parameter. In all the cases, we exclude descriptors with constant and near-constant values and descriptors with at least one missing value.

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. A large number of feature selection methods is available to search for the best descriptors from a pool of variables, and the replacement method (RM) [21,22], employed here, has been successfully applied elsewhere [23–32]. In brief, the RM is an efficient optimization tool which generates multivariable linear regression (MLR) models by searching in a set having *D* descriptors for an optimal subset having $d \ll D$ ones with smallest *RMSD*. The quality of the results achieved with this technique is close to performing an exact (combinatorial) full search of molecular descriptors, although requires much less computational work. RM algorithm is programmed in Matlab [33].

For the selection of the best model among the MLR models, we use a novel methodology, which is called total-order ranking that belongs to multicriteria decision making (MCDM) [34]. This is an useful tool for selecting the best alternative from a pool of potential candidates that have some criteria. One of the simplest approaches for the total-order ranking is the utility function, which is defined by a transformation in a scale between 0 (minimum utility) and 1 (maximum utility) by an arbitrary function (f_k), i.e., linear, inverse linear, normal, etc., that represents the trend of each criterion throughout its lower and highest values as the worst and best conditions, respectively. In this way, each criterion (y_{ki}) is independently transformed into an utility function (u_{ki}). Then the overall utility (U_i) of each candidate is calculated as an average value if each criterion has the same weight. Total-order calculations are carried out using DART software [15].

2.3.2. Model validation

We validated the established QSPR models in order to determine their predictive power, by predicting the RI on compounds not considered during the calibration and comparing such values with the experimental ones. Therefore, the whole set of 1208 compounds is split into training (train), validation (val), and test (test) sets of compounds [35]. The training set is 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 representatives of the validation and test set compounds [36]. 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 [37].

In this work, we choose the training, validation and test set compounds by applying a new procedure developed in our group that is based on the k-means cluster analysis (k-MCA) method [38] implemented in Matlab. 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 a distance metrics (i.e., Euclidean distance), and compounds in different clusters are very distinct. Our 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. This is done to consider 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 1208 × 1816.
b. Remove the linearly dependent variables from the previous matrix. The new size of **C** is 1208 × 1137.
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 1208 compounds through the k-MCA method, for which the **C** matrix is used together with the Euclidean metrics, and 90 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 × 1137 size. $N^0_{\text{train}} = N_{\text{train}} - N_{\text{min max}}$, where N^0_{train} is the number of compounds in train and $N_{\text{min max}}$ 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 $N_{\text{min max}}$ compounds.
f. Create N_{val} clusters with the remaining $1208 - N_{\text{train}}$ 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 $1208 - N_{\text{train}} - N_{\text{val}}$ compounds.

We practice the cross-validation technique of leave-one-out (loo) and leave-more-out (ln%o, with n% being the percentile of molecules removed from the training set). The statistical parameters $R_{\text{ln } \% \text{ o}}$ and $S_{\text{ln } \% \text{ o}}$ (correlation coefficient and standard deviation of leave-more-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 [39] is applied in order to verify the model robustness and to avoid the development of fortuitous correlations. This technique consists 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 is also explored, as not even a 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 [40]. 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 [41], which allows to verify whether a new compound can be considered as interpolated (with reduced uncertainty, 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^*); Table S2 includes the definitions for h_i and h^*. When $h_i > h^*$ for a test set compound, then a warning should be given: it 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 standardize its regression coefficient (b^s_j, see Table S2). The larger the absolute value of b^s_j, the greater the importance of such descriptor [42].

3. Results and discussion

As a first step, we apply the k-MCA clustering-based procedure for splitting the data set of 1208 compounds into $N_{\text{train}} = 400$, $N_{\text{val}} = 405$, and $N_{\text{test}} = 403$ set compounds (refer to Table S1), thus ensuring the design of balanced sets of compounds. The $N_{\text{train}} = 400$ and $N_{\text{val}} = 405$ 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,

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

d	R^2_{train}	RMSD$_{\text{train}}$	R^2_{val}	RMSD$_{\text{val}}$	$R^2_{ij\ \text{max}}$	U	Molecular descriptors
1	0.87	124.38	0.89	107.12	0.00	0.167	*X1sol*
2	0.89	112.46	0.92	91.93	0.00	0.470	*Chi0_EA*, *GATS1p*
3	0.91	101.71	0.93	83.58	0.88	0.568	*nHDon*, *RDF010e*, *Sp*
4	0.92	97.75	0.94	79.65	0.88	0.721	*nHDon*, *DP02*, *RDF010e*, *Sp*
5	**0.92**	**95.44**	**0.94**	**78.31**	**0.32**	**0.784**	***PDI***, ***Hy***, ***ATSC2s***, ***EE_B(s)***, ***X1sol***
6	0.93	91.47	0.95	76.19	0.76	0.682	*H-050*, *R1s+*, *Mor05p*, *RDF010s*, *ATSC2s*, *X1sol*
7	0.93	89.07	0.94	76.94	0.76	0.683	*H-050*, *nCconj*, *R2s+*, *Mor05p*, *RDF010s*, *ATSC2s*, *X1sol*

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

d	R^2_{train}	$\text{RMSD}_{\text{train}}$	R^2_{val}	RMSD_{val}	$R^2_{ij\ \max}$	U	Molecular descriptors
1	0.87	124.38	0.89	107.12	0.00	0.167	*X1sol*
2	0.89	112.46	0.92	91.93	0.00	0.489	*Chi0_EA*, *GATS1p*
3	0.91	105.09	0.93	83.80	0.19	0.699	*PDI*, *Hy*, *X1sol*
4	**0.91**	**100.94**	**0.93**	**82.99**	**0.17**	**0.806**	***PDI*, *H-050*, *SpMax1_Bh*(*s*), *X1sol***
5	0.92	96.43	0.94	82.74	0.38	0.770	*PDI*, *O-058*, *H-050*, *ATSC4s*, *X1sol*
6	0.93	94.00	0.93	85.15	0.76	0.602	*O-058*, *H-050*, *C-044*, *ATSC4e*, *H_Dt*, *X1sol*
7	0.93	91.09	0.94	80.11	0.32	0.763	*PDI*, *Hy*, *C-044*, *C-033*, *ATSC2s*, *EE_B*(*s*), *X1sol*

(b) 1815 non-conformational descriptors, and (c) 1080 3D-descriptors. In this way, we try to identify whether 3D-descriptors are really important for modeling the RI parameter in the OV-101 column. Tables 1–3 summarize the best MLR models found having 1–7 descriptors. The meaning for each involved descriptor is supplied in Table S3. It is appreciated that the $\text{RMSD}_{\text{train}}$ and $\text{RMSD}_{\text{test}}$ parameters do not have a significant variation between models of the same size (d). Therefore, this finding clearly demonstrates that 3D-descriptors can be avoided for modeling the RI property. This can be considered as an important result, as any model that includes quantum chemical descriptors usually involves a relatively difficult calculation of the optimum molecular geometry, involving high computational costs and long times. The exclusion of 3D structural aspects also avoids problems associated with ambiguities, resulting from an incorrect geometry optimization due to the existence of compounds in various conformational states. Such kind of problems may also lead to the loose of predictive capability of the QSPR when applied to the prediction of an external test set of compounds.

For the selection of the optimal model for each data set, we use the total-order ranking as an utility function. We consider the six quality parameter models given by the RM, that is, R^2_{train}, R^2_{val}, $\text{RMSD}_{\text{train}}$, RMSD_{val}, $R^2_{ij\ \max}$, and d. A linear function is used for representing R^2_{train} and R^2_{val}, whereas $\text{RMSD}_{\text{train}}$, RMSD_{val}, and $R^2_{ij\ \max}$ are modulated by an inverse linear function. A normal function is considered for the number of descriptors in the model d. In this way, we keep the model size as small as possible (Ockham's razor), in order to avoid any possible fortuitous correlation or an increased correlation between descriptors. DART results of utility (U) for each model is given in Tables 1–3, and the best one is placed in bold according to the highest utility value. Thus, we choose the following four-descriptor structure–retention relationship that includes non-conformational descriptors:

$$\text{RI} = -1104.8 + 169.3\,X1sol + 26.0\,SpMax1\_Bh(s) + 136.5\,H{-}050 + 1370.2\,\text{PDI} \tag{1}$$

$$N_{\text{train}} = 400,\ d = 4,\ R^2_{\text{train}} = 0.914,\ S_{\text{train}} = 100.9,\ F = 1049,\ R^2_{ij\ \max} = 0.172$$
$$o(3S) = 5,\ R^2_{\text{loo}} = 0.912,\ S_{\text{loo}} = 102.3,\ R^2_{120\%\text{o}} = 0.907,\ S_{120\%\text{o}} = 105.0,\ S^{\text{rand}} = 335.1$$
$$N_{\text{val}} = 405,\ R^2_{\text{val}} = 0.935,\ S_{\text{val}} = 83.0$$
$$N_{\text{test}} = 403,\ R^2_{\text{test}} = 0.927,\ S_{\text{test}} = 78.6$$

Here, F is the Fisher parameter, $R_{ij\ \max^2}$ denotes the maximum squared correlation coefficient between descriptor pairs, and $o(3S)$ 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 $R^2_{\text{train}} = 91\ \%$, $R^2_{\text{val}} = 94\ \%$, and $R^2_{\text{test}} = 93\ \%$. In addition, the root mean square deviations are $\text{RMSD}_{\text{train}} = 100.9$, $\text{RMSD}_{\text{val}} = 83.0$, and $\text{RMSD}_{\text{test}} = 78.6$. 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 (80 molecules). The Y-randomization procedure demonstrates that $S_{\text{train}} < S^{\text{rand}}$ (335.1), and thus a valid structure–property relationship is achieved. We check that Eq. (1) accomplishes with the validation criteria suggested by Golbraikh and Tropsha to assure predictive capability [43]:

$$R^2_{\text{loo}} > 0.5(0.912)$$

$$R^2_{\text{test}} > 0.6(0.927)$$

$$1 - R^2_0/R^2_{\text{test}} < 0.1(0.000)\ \text{or}\ 1 - R'^2_0/R^2_{\text{test}} < 0.1(0.008)$$

$$0.85 \le k(0.99) \le 1.15\ \text{and}\ 0.85 \le k'(1.00) \le 1.15$$

$$R^2_m > 0.5(0.917)$$

The R^2_0, R'^2_0, k, k', and R^2_m parameters are defined in Table S2.

Fig. 1 plots the predicted RI as a function of the experimental values for the training, validation, and test sets (numerical data are provided in Table S4), showing that there exists a tendency for the points to have a straight line trend. The dispersion plot of residuals for the selected model (i.e., residuals as a function of predicted RI) is shown in Fig. 2, which reveals that residuals tend to obey a random pattern around the zero line, suggesting that the assumption of the MLR technique is fulfilled. The five outliers from Eq. (1) are compounds **523**, **961**, **1057**, **1058**, and **1202**. After checking the literature to be certain that their experimental IR values and molecular structures are correct (which they are), we can assume that this irregular behavior may be attributed to the wide structural diversity of the molecules considered in the analyzed data set.

Among the descriptors appearing in the QSPR, there are two 2D-descriptors: a connectivity index ($X1sol$) and a Burden eingenvalue ($SpMax1\_Bh(s)$), while there are two 1D-descriptors: an atom-centred fragment ($H-050$) and a molecular property (PDI). Such descriptors selected by RM are enough to study this data set. The maximum squared correlation coefficient between $X1sol$ and PDI descriptors is $R^2_{ij\ \max} = 0.172$ (see correlation matrix in Table S5). This value reflects a low correlation between such variables, which indicates that they are not

Table 3
The best QSPR models obtained using 3D-descriptor blocks. The chosen result appears in bold.

d	R^2_{train}	$\text{RMSD}_{\text{train}}$	R^2_{val}	RMSD_{val}	$R^2_{ij\ \max}$	U	Molecular descriptors
1	0.84	138.11	0.88	114.04	0.00	0.167	*G2*
2	0.86	129.98	0.90	101.54	0.06	0.404	*Mor11p*, *G2*
3	0.88	119.99	0.91	99.11	0.37	0.536	*Mor03p*, *Mor18v*, *G2*
4	0.89	112.28	0.91	96.21	0.57	0.678	*DP08*, *TDB03p*, *SpMAD_RG*, *G2*
5	**0.91**	**103.07**	**0.93**	**88.61**	**0.81**	**0.714**	***R1e+*, *H2v*, *RDF025p*, *RDF010p*, *Chi_RG***
6	0.91	101.98	0.93	87.37	0.81	0.667	*R1m+*, *H2v*, *RDF010s*, *RDF025p*, *RDF010p*, *Chi_RG*
7	0.92	100.29	0.92	89.79	0.81	0.640	*Q2*, *R2s+*, *L2m*, *RDF010s*, *RDF025p*, *RDF010p*, *Chi_RG*

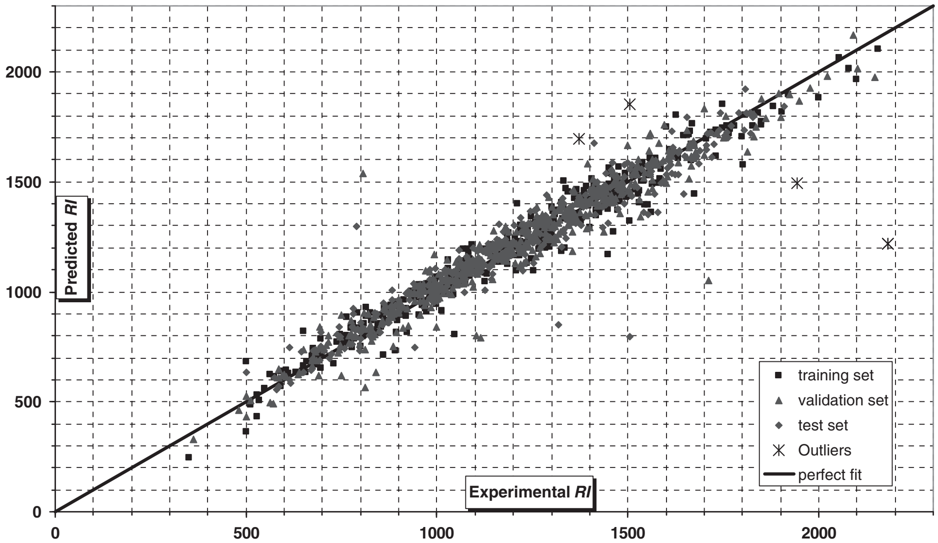

Fig. 1. Experimental versus predicted retention index according to the QSPR model for methyl silicone OV-101 column.

collinear, and each one includes different aspects of the molecular structure that succeed in combining with the remaining variables of Eq. (1) [44]. The numerical values taken by the four descriptors are included in Table S6.

The relative degree of contribution of each descriptor (b_j^s) reveals that $X1sol$ has the greatest importance in the equation: $X1sol$ (0.86) > PDI (0.21) > $H-050$ (0.16) > $SpMax1\_Bh(s)$ (0.09). Moreover, as these descriptors take on positive numerical values, it is concluded that the sign of each regression coefficient in Eq. (1) has a synergistic effect on RI value. Therefore, higher values for the four descriptors for a given compound would lead to a higher predicted RI value.

Analyzing the most relevant descriptor $X1sol$ proposed by Zefirov and Palyulin [45], this descriptor has been defined to model the solvation entropy and to describe dispersion interactions in solution. This connectivity index is used for studying alkyl sulfides, and among several topological indices, it is well correlated to their boiling point. In fact, it is well known that boiling points govern the retention in gas chromatography for apolar stationary phases. Its calculations involve an hydrogen and fluorine-depleted graphs. The product of the principal quantum numbers of the two vertices incident to the considered edge are divided by the squared root of the product of the corresponding vertices degree, then a summation goes over all the considered edges, and finally it is multiplied for a normalization factor that allows coinciding the indices for compounds containing only second-row atoms [18].

Buydens and Massart [46] have also demonstrated that topological indices (i.e., molecular connectivity) and structural parameters are very useful for describing the interactions between molecules of the same family and a stationary phase. Molecular connectivity has a good correlation with both polar and non-polar stationary phases, showing its efficiency for these calculations. In a subsequent study, the same authors [47] use a combination of topological, physicochemical, and quantum chemical descriptors. Their results show that for non-polar stationary phases the topological parameters are able to explain a large part of the total variance; however, for a polar stationary phase using these topological descriptors, an electronic parameter is necessary. Pompe and Novič [48] have used topological descriptors for improving the prediction capabilities of GC-RI. They select 16 molecular descriptors which optimize the R^2 parameter. Comparing their results with ours, Eq. (1) presented in our study is a much simpler model for

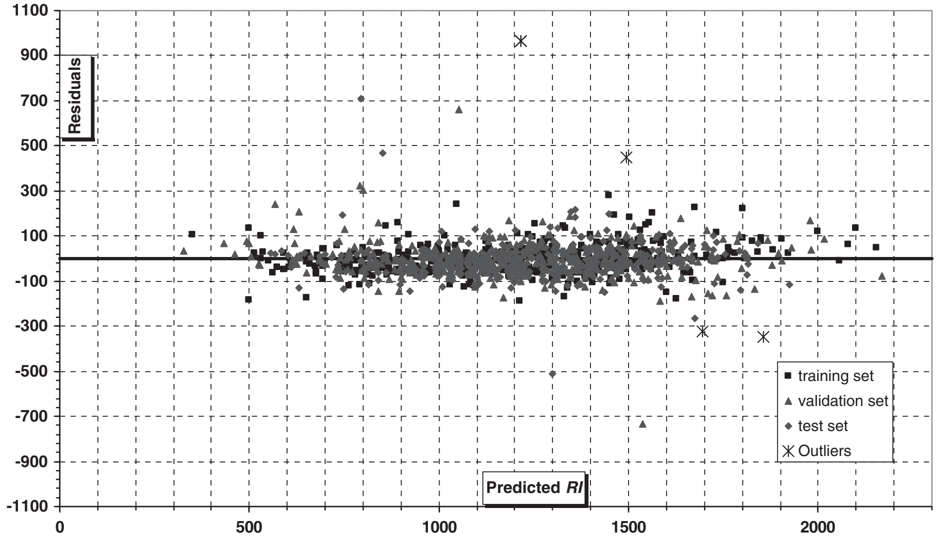

Fig. 2. Dispersion plot of residuals for the QSPR model.

Table 4
A comparison of various QSPR models taken from the literature for the OV-101 column.

Reference	Compound family	Number of compounds	Number of descriptors	R^2_{train}	$RMSD_{train}$	R^2_{test}	$RMSD_{test}$
[7]	Substituted pyrazines	107	6	0.994	21.4	0.992	32.4
[8]	Alcohols, aldehydes, ketones, and esters	115	7	0.998	11.1	–[a]	–
[49]	Alkylbenzenes	39	7	1.000	3.4	0.999	–
	Alkylnaphthalenes	15	5	1.000	1.7	0.999	–
	Alkyl aryl carbamates (Group 1)	27 for the three groups	3	1.000	1.4	–	–
	Alkyl aryl carbamates (Group 2)		4	1.000	0.6	–	–
	Alkyl aryl carbamates (Group 3)		4	0.994	7.0	–	–
[12]	Fragrance compounds	91	5	0.994	–	–	–
[14]	Flavor compounds	297	4	0.961	59.6	0.959	58.0
This work	Flavors and Fragrances	1208	4	0.914	100.9	0.927	78.6

[a] Not available

a non-polar column considering four molecular descriptors; therefore, our QSPR can be easily interpreted and applied for prediction purposes.

Table 4 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. On the other hand, when all three groups of alkyl aryl carbamates are considered, it leads to a low $R^2 = 0.230$. 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 into training and test sets [7,14]. In our case, the data set is 4.1 times bigger than the largest one previously analyzed [14]. In addition, three QSRR studies from Table 4 do not perform external validation [8,12,49]. The model found in the current work has the smallest 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 RI should be in the range 350 to 2180 and the leverage value should lie below the warning leverage value ($h^* = 0.0375$) established in the present model.

The AD of Eq. (1) reveals that nine compounds from the validation set (**129**, **316**, **723**, **854**, **950**, **972**, **1097**, **1102**, **1205**) and five compounds from the test set (**18**, **38**, **772**, **1037**, **1198**) have leverage values (Table S7) over the warning leverage $h^* = 0.0375$. After an exhaustive control analysis of these compounds at the source, we do not find any mistakes. Hence, we assume that this particular behavior is due to the complexity of the data set, i.e., the structural heterogeneity of the molecules considered in this study. In fact, QSRR studies regarding gas chromatography responses are usually carried out by considering only a few families of compounds [3,7–14,27,50–58].

4. Conclusions

A successful application of the QSPR theory is presented for the prediction of the gas chromatographic retention index of 1208 flavor and fragrance compounds in the non-polar stationary capillary column OV-101. We develop a model with acceptable predictive power on the test set, which can be used to predict this property for un-evaluated and un-synthesized flavors or fragrances. Furthermore, the utility function demonstrates to be a useful tool for selecting the best model among a pool of seven ones having six criteria to be considered. We have also demonstrated that the solvation connectivity index of first order is strongly correlated to the RI as a synergistic effect. Our model complements previous reported results from the literature, and it produces a more general and predictive quantitative structure–retention relationship. Finally, 3D-molecular descriptors do not improve the parameters quality of the QSPR model. In this context, the conformation-independent QSPR method continues to emerge as an alternative approach for developing models based on constitutional and topological molecular features of compounds.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2014.09.020.

Acknowledgments

Cristian Rojas is grateful for a 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.

References

[1] K. Bauer, D. Garbe, H. Surburg, Common Fragrance and Flavor Materials. Preparation, Properties and Uses, Fourth ed. WILEY-VCH, Weinheim, 2001.
[2] R. Kaliszan, QSRR: quantitative structure-(chromatographic) retention relationships, Chem. Rev. 107 (2007) 3212–3246.
[3] K. Héberger, Quantitative structure-(chromatographic) retention relationships, J. Chromatogr. A 1158 (2007) 273–305.
[4] Q.S. Wang, L. Zhang, M. Zhang, X.D. Xing, G.Z. Tang, A system for predicting the retentions of *O*-alkyl, *n*-(1-methylthioethylideneamino) phosphoramidates on RP-HPLC, Chromatographia 49 (1999) 444–448.
[5] IUPAC, Compendium of Chemical Terminology—The Gold Book, 2012.
[6] 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.
[7] D.T. Stanton, P.C. Jurs, Computer-assisted prediction of gas chromatographic retention indexes of pyrazines, Anal. Chem. 61 (1989) 1328–1332.
[8] 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.
[9] K. Héberger, M. Görgényi, M. Sjöström, Partial least squares modeling of retention data of oxo compounds in gas chromatography, Chromatographia 51 (2000) 595–600.
[10] S. Liu, C. Yin, S. Cai, Z. Li, Molecular structural vector description and retention index of polycyclic aromatic hydrocarbons, Chemom. Intell. Lab. Syst. 61 (2002) 3–15.
[11] C. Lu, A. Jalbout, L. Adamowicz, Y. Wang, C. Yin, QSRR study for gas and liquid chromatographic retention indices of polyhalogenated biphenyls using two 2D descriptors, Chromatographia 66 (2007) 717–724.
[12] K.L. Goodner, Practical retention index models of OV-101, DB-1, DB-5, and DB-Wax for flavor and fragrance compounds, LWT Food Sci. Technol. 41 (2008) 951–958.
[13] H. Dua, J. Wanga, Z. Hua, X. Yao, Quantitative Structure-Retention Relationship study of the constituents of saffron aroma in SPME-GC-MS based on the Projection Pursuit Regression method, Talanta 77 (2008) 360–365.
[14] 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.
[15] TALETE srl., DART (Decision Analysis by Ranking Techniques), http://www.talete.mi.it/ 2007.
[16] W. Jennings, T. Shibamoto, Qualitative Analysis of Flavor and Fragrance Volatiles by Glass Capillary Gas Chromatography, ACADEMIC PRESS, INC, London, 1980.
[17] Open Babel, Open Babel, http://openbabel.org/wiki/Windows_GUI.
[18] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, WILEY-VCH, Weinheim, 2009.
[19] Hypercube, Inc., HyperChem, http://www.hyper.com.

[20] TALETE, Srl., Dragon (Software for Molecular Descriptor Calculation), http://www.talete.mi.it/ 2014.

[21] P.R. Duchowicz, E.A. Castro, F.M. Fernández, Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies, MATCH Commun. Math. Comput. Chem. 55 (2006) 179–192.

[22] 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.

[23] P.R. Duchowicz, A. Talevi, L.E. Bruno-Blanch, E.A. Castro, New QSPR study for the prediction of aqueous solubility of drug-like compounds, Bioorg. Med. Chem. Lett. 16 (2008) 7944–7955.

[24] P.R. Duchowicz, N.C. Comelli, E.V. Ortiz, E.A. Castro, QSAR study for carcinogenicity in a large set of organic compounds, Curr. Drug Saf. 7 (2012) 282–288.

[25] P.R. Duchowicz, M.A. Giraudo, E.A. Castro, A.B. Pomilio, Amino acid profiles and quantitative structure-property relationship models as markers for Merlot and Torrontés wines, Food Chem. 140 (2013) 210–216.

[26] P.R. Duchowicz, M.P. González, A.M. Helguera, M. Natália Dias Soeiro Cordeiro, E.A. Castro, Application of the replacement method as novel variable selection in QSPR. 2. Soil sorption coefficients, Chemometr. Intell. Lab. Syst. 88 (2007) 197–203.

[27] 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. 1 (2010) 73–77.

[28] J. García, P.R. Duchowicz, M.F. Rozas, J.A. Caram, M.V. Mirífico, F.M. Fernández, E.A. Castro, A comparative QSAR on 1,2,5-thiadiazolidin-3-one 1,1-dioxide compounds as selective inhibitors of human serine proteinases, J. Mol. Graph. Model. 31 (2011) 10–19.

[29] M. Goodarzi, P.R. Duchowicz, C.H. Wu, F.M. Fernández, E.A. Castro, New hybrid genetic based support vector regression as QSAR approach for analyzing flavonoids-GABA(A) complexes, J. Chem. Inf. Model. 49 (2009) 1475–1485.

[30] G. Pasquale, G.P. Romanelli, J.C. Autino, J. García, E.V. Ortiz, P.R. Duchowicz, Quantitative structure-activity relationships on chalcone derivatives: mosquito larvicidal studies, J. Agric. Food Chem. 60 (2012) 692–697.

[31] A.B. Pomilio, P.R. Duchowicz, M.A. Giraudo, E.A. Castro, Amino acid profiles and quantitative structure-property relationships for malts and beers, Food Res. Int. 43 (2010) 965–971.

[32] A.B. Pomilio, M.A. Giraudo, P.R. Duchowicz, E.A. Castro, QSPR analyses for aminograms in food: citrus juices and concentrates, Food Chem. 123 (2010) 917–927.

[33] The MathWorks, Inc., Matlab, Masachussetts, USA, http://www.mathworks.com.

[34] M. Pavan, R. Todeschini, Total order ranking methods, in: M. Pavan, R. Todeschini (Eds.), Scientific Data Ranking Methods: Theory and Applications, Elsevier, 2008, pp. 51–72.

[35] A. Miller, Subset selection in regression, CRC Press, 2012.

[36] 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.

[37] 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.

[38] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 2005.

[39] C. Rücker, G. Rücker, M. Meringer, Y-Randomization and its variants in QSPR/QSAR, J. Chem. Inf. Model. 47 (2007) 2345–2357.

[40] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701.

[41] 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.

[42] N.R. Draper, H. Smith, Applied Regression Analysis, 1981. (New York).

[43] A. Golbraikh, A. Tropsha, Beware of q2! J. Mol. Graph. Model. 20 (2002) 269–276.

[44] P.R. Duchowicz, F.M. Fernández, E.A. Castro, Orthogonalization methods in QSPR-QSAR Studies, in: E.A. Castro (Ed.), QSPR-QSAR Studies on Desired Properties for Drug Design, Research Signpost, Kerala, 2010, pp. 189–203.

[45] N.S. Zefirov, V.A. Palyulin, QSAR for boiling points of “Small” sulfides: are the “high-quality structure–property–activity regressions” the real high quality QSAR models? J. Chem. Inf. Comput. Sci. 41 (2001) 1022–1027.

[46] L. Buydens, D.L. Massart, Prediction of gas chromatography retention indexes from linear free energy and topological parameters, Anal. Chem. 53 (1981) 1990–1993.

[47] 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.

[48] M. Pompe, M. Novič, Prediction of gas-chromatographic retention indices using topological descriptors, J. Chem. Inf. Comput. Sci. 39 (1998) 59–67.

[49] V.A. Gerasimenko, V.M. Nabivach, Relationships between gas chromatographic retention indices and molecular structure of aromatic hydrocarbons, J. Chromatogr. A 498 (1990) 357–366.

[50] J. Bermejo, M.D. Guillen, Prediction of Kovats retention index of saturated alcohols on stationary phases of different polarity, Anal. Chem. 59 (1987) 94–97.

[51] B. da Silva Junkes, R. Dias de Mello Castanho Amboni, R. Augusto Yunes, 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. 477 (2003) 29–39.

[52] O. Farkas, I.G. Zenkevich, F. Stout, J.H. Kalivas, K. Héberger, Prediction of retention indices for identification of fatty acid methyl esters, J. Chromatogr. A 1198–1199 (2008) 188–195.

[53] V.E.F. Heinzen, R.A. Yunes, Using topological indices in the prediction of gas chromatographic retention indices of linear alkylbenzene isomers, J. Chromatogr. A 719 (1996) 462–467.

[54] F. Liu, Y. Liang, C. Cao, N. Zhou, QSPR study of GC retention indices for saturated esters on seven stationary phases based on novel topological indices, Talanta 72 (2007) 1307–1315.

[55] K. Ośmiałowski, J. Halkiewicz, A. Radecki, R. Kaliszan, Quantum chemical parameters in correlation analysis of gas–liquid chromatographic retention indices of amines, J. Chromatogr. A 346 (1985) 53–60.

[56] J. Raymer, D. Wiesler, M. Novotny, Structure-retention studies of model ketones by capillary gas chromatography, J. Chromatogr. A 325 (1985) 13–22.

[57] J.M. Sutter, T.A. Peterson, P.C. Jurs, Prediction of gas chromatographic retention indices of alkylbenzenes, Anal. Chim. Acta. 342 (1997) 113–122.

[58] D. Zakarya, M. Chastrette, M. Tollabi, S. Fkih-Tetouani, Structure-camphor odour relationships using the Generation and Selection of Pertinent Descriptors approach, Chemom. Intell. Lab. Syst. 48 (1999) 35–46.