

# QSRR prediction of gas chromatography retention indices of essential oil components

Yovani Marrero-Ponce<sup>1,2,3</sup> · Stephen J. Barigye<sup>4,5</sup> · María E. Jorge-Rodríguez<sup>6</sup> · Trang Tran-Thi-Thu<sup>6</sup>

Received: 24 April 2017 / Accepted: 22 July 2017  
© Institute of Chemistry, Slovak Academy of Sciences 2017

**Abstract** A comprehensive and largest (to the best of our knowledge) database of 791 essential oil components (EOCs) with corresponding gas chromatographic retention properties has been built. With this data set, Quantitative structure–retention relationship (QSRR) models for the prediction of the Kováts retention indices (RIs) on the non-polar DB-5 stationary phase have been built using the DRAGON molecular descriptors and the regression methods: multiple linear regression (MLR) and artificial neural networks (ANN). The obtained models demonstrate good

performance, evidenced by the satisfactory statistical parameters for the best MLR ( $R^2 = 96.75\%$  and  $Q_{\text{ext}}^2 = 98.0\%$ ) and ANN ( $R^2 = 97.18\%$  and  $Q_{\text{ext}}^2 = 98.4\%$ ) models, respectively. In addition, the built models provide information on the factors that influence the retention of EOCs over the DB-5 stationary phase. Comparisons of the statistical parameters for the QSRR models in the present study with those reported in the literature demonstrate comparable to superior performance for the former. The obtained models constitute valuable tools for the prediction of RIs for new EOCs whose experimental data are undetermined.

**Electronic supplementary material** The online version of this article (doi:10.1007/s11696-017-0257-x) contains supplementary material, which is available to authorized users.

✉ Stephen J. Barigye  
sjbarigye@gmail.com

- <sup>1</sup> Colegio de Ciencias de la Salud (COCSA), Escuela de Medicina, Universidad San Francisco de Quito (USFQ), Grupo de Medicina Molecular y Traslacional (MeM&T), Edificio de Especialidades Médicas, Av. Interoceánica Km 12 ½ -Cumbayá, Quito 170157, Ecuador
- <sup>2</sup> Instituto de Simulación Computacional (ISC-USFQ), Diego de Robles y vía Interoceánica, 17-1200-841, Quito, Ecuador
- <sup>3</sup> Grupo de Investigación Ambiental (GIA), Programas Ambientales, Facultad de Ingenierías, Fundación Universitaria Tecnológico de Comfenalco (COMFENALCO), Programa de Ingeniería de Procesos, Bolívar, Cartagena de Indias, Bolívar, Colombia
- <sup>4</sup> Department of Chemistry, Federal University of Lavras, Lavras, Brazil
- <sup>5</sup> Facultad de Medicina, Universidad de Las Américas, Quito, Pichincha, Ecuador
- <sup>6</sup> Department of Pharmacy, Faculty of Chemistry-Pharmacy, Central University of Las Villas, Villa Clara, 54830 Santa Clara, Cuba

**Keywords** Gas chromatography · Retention index · Essential oil · Quantitative structure–retention relationships · Multiple linear regression · Artificial neural networks

## Introduction

Essential oils are composite mixtures of varying composition (20–60 components) and abundance [ranging from traces (ng/g) to fairly high concentrations (g/100 g)] with the major components belonging to the phenolic mono- and sesqui-terpenoid chemical groups. Essential oils have increasingly found applications in a diverse range of industries including: food processing, perfumery, cosmetics, pharmaceutical production, winery, etc., particularly due to their antioxidant and antimicrobial activity, attributed to the phenyl moieties in their structures (Bajpai et al. 2009). The evaluation of the essential oils' composition profiles is crucial in determining the components responsible for their chemical/biological activity.

**Table 1** QSRR models for the prediction of the RIs for essential oils components using MLR and families of DRAGON MDs based on the apolar stationary phase DB-5

MD type	Models	No.
All DRAGON MDs	RI = 315.4270 ( $\pm 22.8496$ ) + 8.9710 ( $\pm 0.2768$ ) $M_w$ - 64.7769 ( $\pm 3.2362$ ) RBN - 68.5399 ( $\pm 5.0842$ ) X2v - 180.4735 ( $\pm 10.7349$ ) MATS2m - 178.2409 ( $\pm 11.8427$ ) GATS2m - 174.6514 ( $\pm 30.2488$ ) E3u + 60.5358 ( $\pm 6.9529$ ) HTm - 821.4969 ( $\pm 123.1227$ ) HATS3m - 566.4689 ( $\pm 62.1945$ ) R5m + 46.4828 ( $\pm 2.2600$ ) nCs - 47.7686 ( $\pm 3.6273$ ) nCrs	(1)
3D	RI = 175.5397 ( $\pm 44.5928$ ) + 200.0965 ( $\pm 15.8300$ ) DP01 - 16.0241 ( $\pm 2.2008$ ) J3D + 1.0708 ( $\pm 0.1290$ ) H3D - 340.6341 ( $\pm 41.5330$ ) RDF020m + 224.8065 ( $\pm 19.4184$ ) RDF020v - 177.6682 ( $\pm 24.3254$ ) Mor27m - 36.2096 ( $\pm 5.3076$ ) Mor03v - 294.3802 ( $\pm 32.6961$ ) Mor26v + 529.8989 ( $\pm 88.3242$ ) Du - 1779.0363 ( $\pm 153.8564$ ) Dv + 131.4808 ( $\pm 5.7147$ ) HTm	(2)
2D	RI = -536.6519 ( $\pm 105.8288$ ) - 3.1879 ( $\pm 0.2117$ ) ZM2V + 73.0030 ( $\pm 2.5197$ ) Dz + 1045.2636 ( $\pm 158.3998$ ) MSD - 12.9239 ( $\pm 0.9881$ ) Rww - 319.9770 ( $\pm 86.2795$ ) piPC01 + 354.8943 ( $\pm 53.5003$ ) piPC02 + 7.1269 ( $\pm 0.9637$ ) PCD - 2.4443 (0.2821) TIC2 + 56.6151 ( $\pm 12.5004$ ) IC4 - 148.1708 ( $\pm 11.9387$ ) MATS2m - 116.3300 ( $\pm 14.7029$ ) GATS2m	(3)
0D-1D	RI = 196.4010 ( $\pm 17.0234$ ) + 13.5301 ( $\pm 1.8236$ ) Ss + 63.2356 ( $\pm 3.6204$ ) SCBO - 75.0738 ( $\pm 9.6783$ ) nCIC - 36.8603 ( $\pm 4.0298$ ) RBN - 54.6929 ( $\pm 4.8279$ ) nDB + 50.2084 ( $\pm 3.5599$ ) nCs + 8.9209 ( $\pm 3.5054$ ) nCt - 31.3332 ( $\pm 4.6627$ ) nCrs + 29.7728 ( $\pm 3.8676$ ) nR = Cs + 6.3350 ( $\pm 1.0142$ ) H-046-42.8890 ( $\pm 7.3439$ ) MLOGP	(4)

The gas chromatography (GC) technique is one of the most powerful tools in analytical chemistry and has been widely used (almost irreplaceably) in the analysis of essential oils (Adams 2001; Olivero et al. 1997; Zhao et al. 2009). The GC method produces a single parameter (retention index), which may be used for the identification of virtually any compound under well-defined conditions (Acevedo-Martínez et al. 2006). Chromatographic retention is a complex phenomenon in which various types of intermolecular forces are involved and these include: dipole-dipole (or Keesom) forces, dipole-induced dipole forces, London dispersion forces, electron donor-acceptor complexes, as well as hydrogen bonds. These forces collectively determine the partition of the solute between the mobile and stationary phases (Fritz et al. 1979; Ong and Hites 1991; Peng et al. 1988; Yancey 1994). The chromatographic retention profile for molecules can be measured using different parameters which include: retention time, linear-temperature programmed retention index, Lee retention index, boiling point correlation, equivalent chain length, Kováts retention distance, and the most popular one the Kováts retention index (RI) (Babushok 2015; Kováts 1958; von Mühlen and Marriott 2011). The RI is a relative retention parameter normalized with respect to *n*-alkane series as a standard. It thus has an advantage of being independent of individual chromatographic system characteristics, which explains its wide application in QSRR studies (Rohrschneider 1965).

Nowadays, gas chromatography systems hyphenated with mass spectrometry (GC-MS) are considered as standard analytical platforms, with the latter providing complementary information for structural identification.

However, GC-MS data (retention times and mass spectra) do not always provide sufficient evidence for structural profiling and thus prediction models may be useful for ultimate structural verification. In addition, the identification of compounds is often performed by matching the GC peaks with a standard of the suspected chemical. The setback for this approach is that standard samples with the required degree of purity are sometimes unavailable, and in such cases, theoretical model for estimating the RIs may be a useful alternative (Hodjmoammadi et al. 2004). The identification of essential oils components is particularly challenging as many constituent terpenes provide identical MS, owing to the fact that they yield similar fragments upon ionization. Moreover, compounds not registered in the existing MS libraries are rather difficult to identify and often lead to erroneous assignments (false positives). In this sense, the integration of chemical information from theoretical retention indices allows for the elucidation of MS data, and consequently more accurate peak assignment of essential oil components. Indeed, MS libraries are currently incorporating retention index data for all registered compounds (Mondello 2015; NIST 2017).

The correlation between gas chromatographic retention indices and molecular parameters provides a significant information on the effect of the molecular structural characteristics on the retention time and on the possible mechanisms for absorption and elution (Körtvélyesi et al. 2001). Good correlations have been obtained between RIs and theoretically calculated data for molecules with different functional groups: alkanes (Görgényi et al. 1989), alkybenzenes and naphthalenes (Dimov et al. 1994), dialkylhydrazones (Király et al. 1996), phenol derivatives

**Table 2** Principal statistical parameters for the four optimal MLR models

Model	$R^2$	$R^2_{\text{adj}}$	$Q^2$	$Q^2_{\text{boot}}$	$Q^2_{\text{ext}}$	a ( $R^2$ )	a ( $Q^2$ )	$F$	SDEP <sup>a</sup>	SDEC <sup>b</sup>
All MDs (1)	96.8	96.69	96.60	96.54	98.00	-0.008	-0.049	1616.01	54.668	53.447
3D (2)	95.8	95.73	95.63	95.55	98.23	-0.007	-0.048	1239.38	62.011	60.73
2D (3)	95.9	95.83	95.72	95.67	98.13	-0.008	-0.049	1270.35	61.313	60.016
0D-1D (4)	95.2	95.12	94.98	94.89	98.22	-0.007	-0.048	1077.56	66.399	64.927

<sup>a,b</sup> Standard deviation error of prediction and calculation, respectively

(Kaliszan and Hóltje 1982), azo compounds (Kortvelyesi et al. 1995), primary, secondary, and tertiary amines (Osmialowski et al. 1985), polycyclic aromatic hydrocarbons (Rohrbaugh and Jurs 1986), various aromatic compounds (Gautzsch and Zinn 1996), alcohols, esters, ketones, monoterpenes, di- and tricyclimethyl esters, and monocyclic ketones (Duvenbeck and Zinn 1993), odor-active aliphatic compounds with oxygen-containing functional groups (Anker et al. 1990), stimulants and narcotics (Georgakopoulos et al. 1991a), anabolic steroids (Georgakopoulos et al. 1991b), etc., using several models ranging from linear ones (e.g., multiple linear regression, partial least squares, and principal component regression) to non-linear ones (e.g., artificial neural network, support vector machine, etc.) (Albaugh et al. 2009; Qin et al. 2013a, b; Qin et al. 2009; Skrbic and Onjia 2006).

Nonetheless, it is important to bear in mind that the ultimate objectives of constructing QSRR models are for their use in the prediction of properties of newly identified compounds whose chemical profiles are not known and for providing greater knowledge on the general mechanisms involved in solute–solute, solute–mobile phase, and solute–stationary phase interactions, respectively. To achieve this goal, it is imperative that the models are constructed over wide chemical space and have good generalizability. Unfortunately, the majority of the QSRR studies have been performed on small and/or congeneric data sets (Azar et al. 2011; Jalali-Heravi and Ebrahimi-Najafabadi 2011; Noorzadeh and Farmany 2010; Qin et al. 2009, 2013b; Yan et al. 2015), with a few exceptions in the literature (Dossin et al. 2016; Garkani-Nejad et al. 2004; Zhang et al. 2017). In the particular case of essential oils, despite the vast literature on the identification and characterization of novel constituent compounds, QSRR models have been generally built on data sets of sizes ranging from 25 to 169 compounds, with most of these belonging to a single chemical series. Certainly, such models have reduced utility as they do not cover an extensive structural space of known essential oils components. It is thus important that wide and diverse data sets for essential oils be constructed and QSRR models be built thereof to guarantee a wider applicability domain (AD) and thus generalization ability.

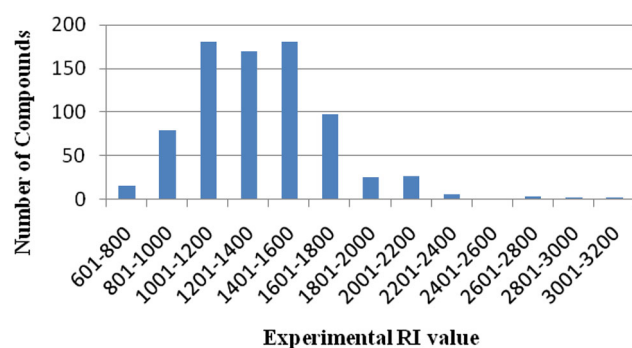
The aim of the present study is to develop a comprehensive data set of constituent components of essential oils and posteriorly develop statistical and artificial intelligence models relate the retention behavior of these components over the apolar stationary phase DB-5 [(5% Phenyl)-methylpolysiloxane] with DRAGON's molecular structure characterizing parameters.

## Experimental

### Construction of the essential oils data set

An extensive literature review on theoretical and experimental studies on essential oils was performed and a data set comprised of 791 chemical structures with their corresponding average RI values was built. To guarantee the homogeneity and thus comparability of the RI values for the constituent components, only target Kováts RI values obtained under similar experimental conditions on standard non-polar 5% phenyl methyl polysiloxane stationary phase (DB-5 or HP-5) of GC–MS system were considered. This is a critical step as it minimizes the possibility of experimental outliers. The distribution of retention indices is shown in Fig. 1, which demonstrates the diversity of the constructed data set.

The compounds and their corresponding retention indices are given as Supplementary Information, SI1 and SI2. For dissimilar RI values obtained under homogeneous conditions for the same compound, the average value was considered. In general, the measurement errors of the GC retention indices are in the range of  $\pm 2$  standard deviations ( $s$ ) of the RI values. Compounds whose RI values presented high standard deviations were not included in during the data set compilation. The molecular structures of data set were sketched using ChemDraw Ultra module of the ChemOffice software (Jaworska 2005). The sketched structures were exported to Chem3D module to create their 3D structures. The geometry optimization was done using semiempirical AM1 (Austin Model) Hamiltonian method and closed shell restricted wave function available in the MOPAC module.



**Fig. 1** Distribution of retention index values for data set

### Descriptor generation

A total of 3224 molecular descriptors (MDs) were computed for the constructed data set using the DRAGON software (Todeschini et al. 2007). Given the high dimensionality of the obtained data matrix ( $791 \times 3224$ ), we applied a simple variable selection procedure based on Spearman's correlation coefficient ( $R$ ), where for each pair of descriptors with  $R \geq 0.95$ , only one is retained. Consequently, a lower dimensionality data matrix comprising of 1476 MDs was obtained.

### Data set splitting and statistical analysis

The earnest predictive power of any model can only be assessed over a set of compounds not used in the model training, also known as the test set. In this sense, the essential oils data set is split into training and test sets, respectively, using the cluster analysis technique implemented in the STATISTICA 8.0 software (Statsoft 2001). First, hierarchical clustering was performed using the Euclidean distance and complete linkage, as the distance measure and linkage rule, respectively. The output dendrogram is comprised of a hierarchical cascade in which the base level represents each compound as belonging to a separate cluster, and for each level uphill, compounds (or clusters) with minimum distances are grouped together. To determine the optimum number of clusters, the distance corresponding to the steepest ascent in the amalgamation schedule is used as the cutoff. Posteriorly,  $k$ -means cluster analysis is performed; with  $k$  representing the number of clusters determined using hierarchical cluster analysis. Finally, the test set (representing approximately 25% of the data set) was selected by rank ordering the chemical compounds in each of the clusters according to the experimental RI values and sampled to span the entire property space. This splitting procedure guarantees that the structural and property spaces are broadly represented in the test for external prediction.

### Multiple linear regression

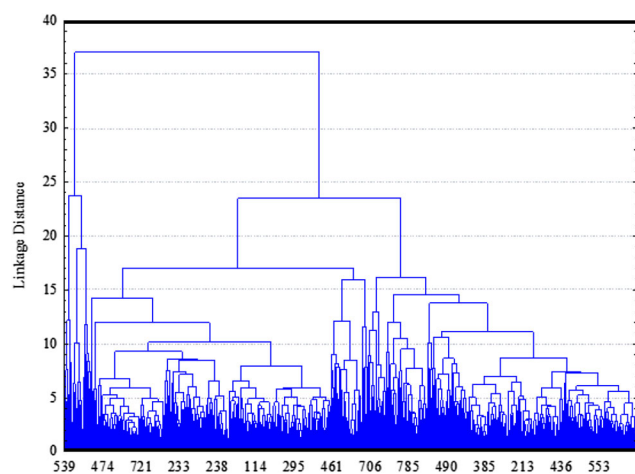
To obtain the linear QSRR models for the gas chromatography Kováts retention index (GC-RI) of essential oils components, multiple linear regression coupled with the Genetic Algorithm (MLR-GA) (Devilleers 1996; Kubinyi 1994; Leardi 1994) was used as the fitting method for the RI and variable selection strategy, respectively. The choice of MLR statistical technique is because of its simplicity, while the key advantage of the GA as a search strategy is that a set of optimum models are obtained with less computational effort, in the sense that a global maximum is achieved without exploring all combinations of variables in the data matrix space. The leave one out cross-validation parameter ( $Q_{100}^2$ ) was used as the objective function. For this study, the MOBY-DIGS program (Todeschini et al. 2004) was employed and the following configurations for the GA were considered: population size, 100; generations, 10 000; probability of mutation, 0.5; number of crossover, 5000.

### Model validation techniques

The obtained models were tested for their robustness using the bootstrapping validation ( $Q_{boot}^2$ ) procedure and Y-randomization [ $a(Q^2)$ ] was used to check for fortuitous correlation. Other statistical parameters considered include: Fisher score ( $F$ ), standard deviation error of prediction (SDEP), and standard deviation error in calculation (SDEC). Therefore, a multi-criteria approach was used to select the best model from the set of models obtained with the MLR-GA method. Posteriorly, the best models were assessed for their predictive power and using the external validation ( $Q_{ext}^2$ ) procedure on the external set compounds. In addition, the Y-randomization test was carried out to check for fortuitous correlation; low intercept values [i.e., a ( $R^2$ ) and a ( $Q^2$ )] are indicative of stability to this phenomenon.

### Model applicability domain (AD)

The AD is a theoretical region in the chemical space defined by the model's independent and response variables, and thus by the nature of the chemicals in the training set as represented by specific MDs (Gramatica 2007). A large and diverse training set contributes to a wide AD, although it is equally important to employ an inclusive structure description method that characterizes (explicitly or implicitly) all relevant structure features. If the structural characteristics of novel compounds are represented in the training data, and also adequately encoded by the model descriptors, it is reasonable to expect that there will be an



**Fig. 2** Dendrogram of the hierarchical analysis  $k$ -MCA

increased probability of accurate property predictions for these compounds. In fact, only the predictions for chemicals that lie within the AD for a given model can be considered as reliable (Jaworska 2005).

Several approaches have been reported in the literature on determining the model's AD, with the most common being the leverage approach (Atkinson 1985). This approach is based on some sort of "distance" metric (also known as the leverage, denoted by  $h$ ) with which the separation of compounds from the model's experimental space (the structural centroid of the training set) is determined as a measure of the influence of chemical structures on the model, in the sense that chemicals close to the centroid are less influential in model building than extreme points. Prediction should be considered unreliable for compounds with  $h$  values greater than the critical value  $h^*$ , as these lie outside the AD of the model, i.e., are structurally distant from the training chemicals ( $h^* = 3p'/n$ ,  $p'$  is the number of model variables plus one, and  $n$  is the number of the objects used to calculate the model). In addition, the models AD should be examined for possible outlying compounds, i.e., poorly fitted data points that cause models to deviate from the actual line of best fit. The criterion for flagging a compound as an outlier involves the computation of the standardized (or studentized) residuals. It follows that compounds with standardized residual values greater than  $\pm 3$  (or  $\pm 2.5$  in the case of studentized residuals) should be analyzed for possible outlying behavior (Alvarez 1995).

### Artificial neural networks

Non-linear methods for multiple regressions such as artificial neural networks, support vector machine, and random forest have increasingly found utility in QSRR studies

following the understanding that the relationship between molecular parameters and corresponding properties may not necessarily be linear. In fact, it has been demonstrated that non-linear models are capable of providing improved predictions of QSRR models. Therefore, to examine the possible non-linear relationship of the RIs and the MDs employed in the present report, artificial neural network (ANN) models were trained using the feedforward back-propagation algorithm. For these models, the same variables contained in the final MLR models were used as inputs and the following parameters were optimized: initial weights, number of nodes in the hidden layer, learning rate, the momentum, and number of iteration cycles. The values corresponding to the minimum prediction error were selected as the optimal parameters. In this study, a three-layered ANN was employed, i.e., comprised of one hidden layer, in addition to the input and output layers, respectively. The early stopping criterion was used to avoid the over-fitting phenomenon. The training and test errors were reported for each 500 cycles and these values used to construct a learning curve with which the network's training was monitored to guard against overtraining and subsequent loss of generalizability of the models. An ascent in the learning curve (corresponding to an increase in the prediction error) was used as a flagging point for stopping the learning process.

## Results and discussion

### Design of the training and validation sets using cluster analysis

Figure 2 shows the output dendrogram for hierarchical cluster analysis. From the amalgamation schedule, 20 clusters were determined and this number ( $k$  value) posteriorly used to perform the  $k$ -means cluster analysis.

The ensuing clusters were then employed to split the data in training and validation sets. Note that 13 structurally atypical compounds were identified with cluster analysis and these were excluded, as they were indicative of outlier behavior, and thus, 778 chemical structures were retained. The resulting data set was split into training and test sets, with the 650 and 128 compounds, respectively.

### Construction of MLR models and determination of optimum dimension

The MLR-based GC-RI models of size 4–15 variables were built over the training set and the best regressions (according to the optimization function) for each model magnitude selected for posterior validation. Figure 3 shows the statistical parameters for the best 4–15 variable models

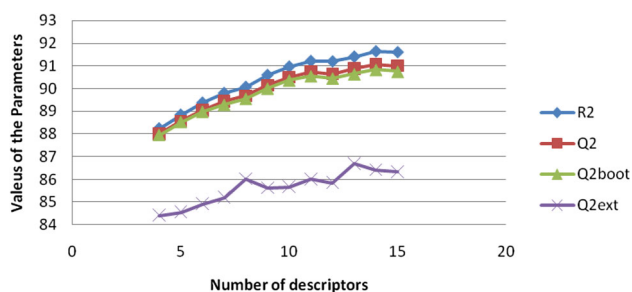
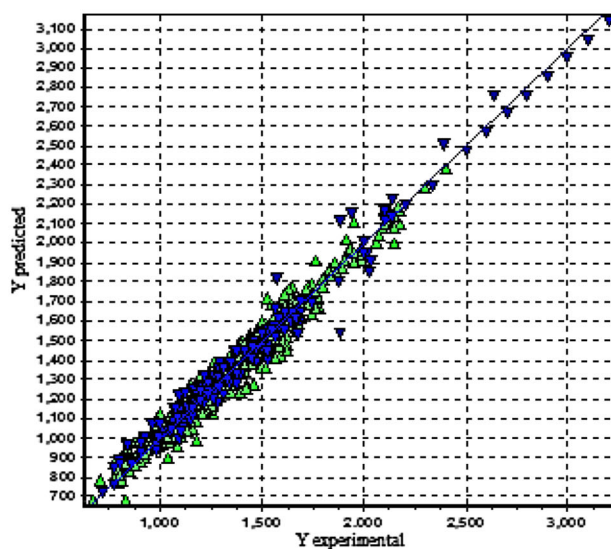


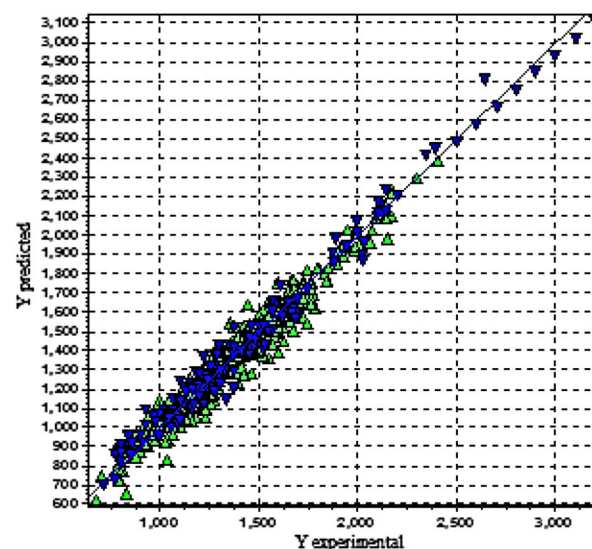
Fig. 3 Statistical parameters for MLR models for different sizes

obtained in the present study. As can be observed, the obtained models generally possess good statistical behavior with correlation coefficients for the internal cross-validation  $Q_{\text{loo}}^2$ ,  $Q_{\text{boot}}^2$  superior to 87%, while the external validation coefficients  $Q_{\text{ext}}^2$  are greater than 84% for all model sizes. It can thus be inferred that generally, the obtained models possess high predictive power and generalizability.

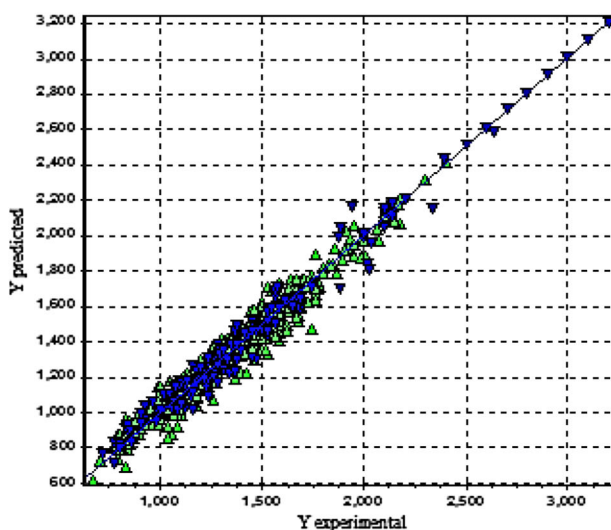
To determine the optimum model size, the statistical parameters  $R^2$ ,  $Q_{\text{loo}}^2$ ,  $Q_{\text{boot}}^2$ , and  $Q_{\text{ext}}^2$  for the different model sizes were compared (see Fig. 3). Although the 14-variable model yields superior statistical parameters, the 11-variable



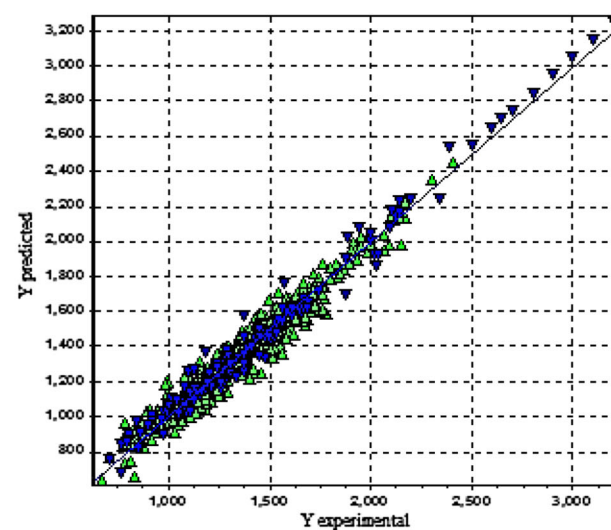
Model 1



Model 2



Model 3



Model 4

Fig. 4 Experimentally predicted RIs for the MLR models. Training set in green and prediction set in blue

**Table 3** Optimization parameters for the four ANN models

Models	Intervals for the initial weight values	Number of hidden nodes ( $x$ )	Number of adjustable parameters	Learning rate	Momentum	Number of optimum iteration cycles
1a	(0.3; 0.3)	8	105	0.1	0.4	24,000
2a	(0.4; 0.4)	15	196	0.2	0.2	1500
3a	(0.5; 0.5)	5	66	0.1	0.5	9500
4a	(0.4; 0.4)	7	92	0.1	0.2	7000

**Table 4** Statistical parameters obtained with the MLR (Eqs. 1–4) and ANN (Eqs. 1a–4a) models

	Model 1 (a)		Model 2 (a)		Model 3 (a)		Model 4 (a)	
	MLR	ANN	MLR	ANN	MLR	ANN	MLR	ANN
$R_t^2$	96.8	97.18	95.8	96.17	95.9	96.21	95.2	96.29
$R_p^2$	98.00	98.35	98.23	98.50	98.13	98.29	98.22	98.24
SDEC	53.447	53.48	60.73	59.98	60.016	63.96	64.927	63.35
SDEP	54.668	67.10	62.011	63.43	61.313	68.46	66.399	69.31

model is empirically chosen as the optimal model size considering the parsimony principal (Occam razor). The correlation matrix of the 11 MDs contained in the selected model as well as the corresponding Pareto diagram are provided as Supplementary Information, SI3 and SI4, respectively.

Subsequently, the selected model's AD was examined for possible outlying compounds. For the best model, 29 statistical outliers were identified, and when these compounds were excluded from the training set, the model's descriptive and predictive ability significantly improved, justifying their ultimate exclusion. The structures of these outlying compounds are provided as Supplementary Information SI5. Therefore, the final data set constituted 762 compounds with the training and test sets, comprised of 609 and 153 compounds, respectively. In parallel, the DRAGON MDs were stratified into three groups, i.e., 0–1D, 2D, 3D, and models built for each case and their performances compared with the model built from the entire set of MDs. Tables 1, 2 show the equations for the best MLR models and the corresponding statistical parameters, respectively, for each of the MD sets (all DRAGON MDs, 3D, 2D, and 0–1D). As can be observed, the results obtained are satisfactory; all the correlation coefficients,  $Q_{\text{loo}}^2$  and  $Q_{\text{boot}}^2$ , are greater than 94%, while the  $Q_{\text{ext}}^2$  values are superior to 98%. In addition, the  $a$  ( $Q^2$ ) and  $a$  ( $R^2$ ) parameters and thus the models are not prone to chance correlation.

Other parameters considered in the analysis of the quality of the obtained models include: the Fisher score ( $F$ ), root-mean-squared errors calculated on the training and test sets, denoted by SDEC and SDEP, respectively. As can be observed from Table 2, the models' parameters are

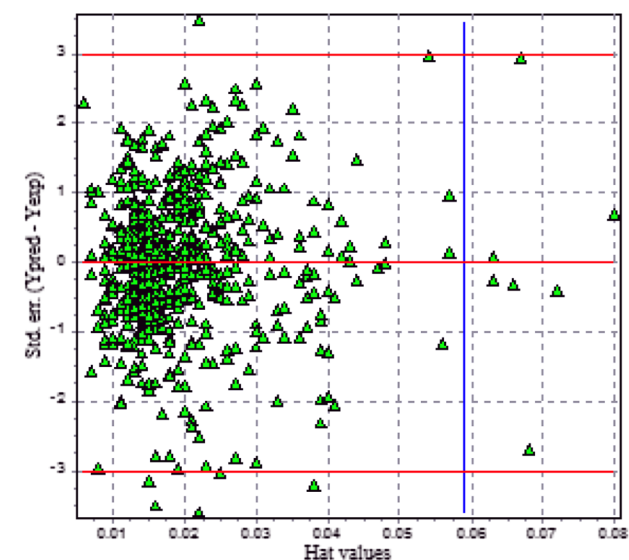
satisfactory. It can, therefore, be deduced that the built models are robust and possess high predictive power.

Figure 4 shows the relationship between the experimental and predicted results of the training and prediction sets for the 4 models (see Supplementary Information SI6 for experimental and predicted values for the models). As can be observed, there exists a close linear association between the experimental and predicted GC–RI endpoints for both the training and test sets, respectively.

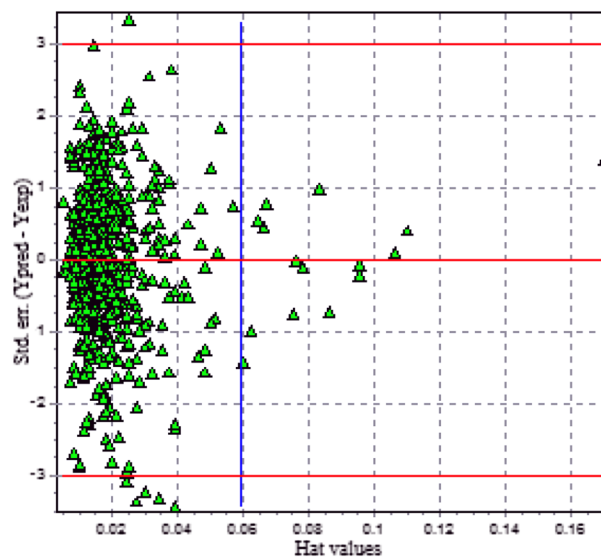
### Model applicability domain (AD)

Figure 5 shows the Williams plots for the obtained models. As can be observed, almost all chemicals lie within the AD, demonstrating the reliability of the models. Some chemicals slightly exceed the critical HAT value (vertical line), but these belong to the training set. Moreover, the removal of these compounds did not significantly alter the corresponding statistical parameters and thus their removal is not justified. On the other hand, a few chemicals are wrongly predicted ( $>3$  s) for each model, but in all the cases belong the models' AD as their HAT values are lower than the cutoffs. This erroneous prediction could probably be attributed to wrong experimental data rather than significant structural differences with respect to compounds within the AD. We presume that the measured GC–RI values are not appropriate and need additional verification. The identities of these compounds as well as their corresponding sources (references) are provided as Supporting Information SII–2, SI6.

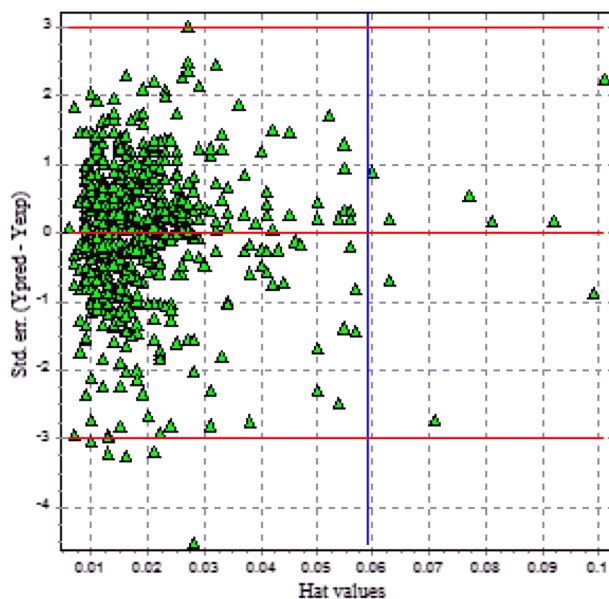
The validation set was designed to include examples spanning all structural moieties in the data set. Therefore, the satisfactory prediction of the validation structures



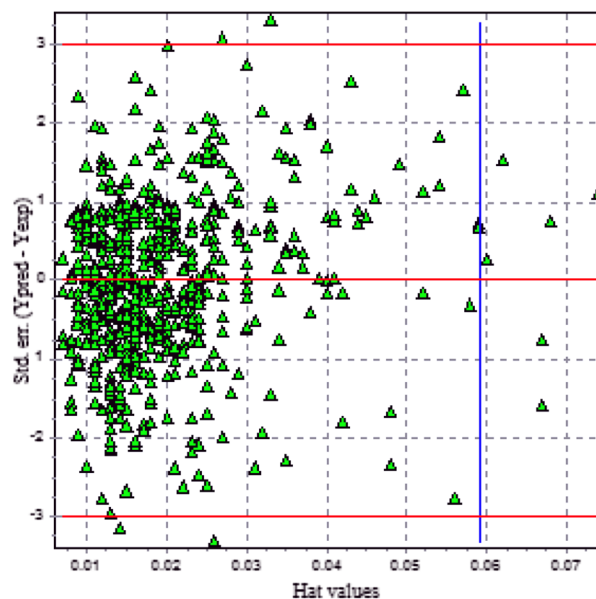
Model 1



Model 2



Model 3



Model 4

**Fig. 5** Williams plot for MLR models on the training set

suggests that the obtained models could be successfully applied in the prediction of GC-RI of a diverse set of compounds provided that they lie within the models' AD (Albaugh et al. 2009). For test set William's plots for the MLR models, see Supplementary Information, SI7. No significant differences have been found between the statistical parameters of four models neither in the training nor in the validation set, although the first model provides the best description for the GC-RI on the independent test set. Nonetheless given that each model has a distinct AD, it is

desirable that the 4 models be jointly used for predicting the GC-RI to enhance the reliability of the modeling procedure. This is known as consensus modeling or ensemble averaging. In fact, it is observed that ensemble averaging of the obtained models provides greater approximation to the experimental GC-RI values. For example, the predicted values for *trans*- $\alpha$ -Bergamotene are 1478.28, 1412.29, 1456.17, and 1462.47 i.u, according to models 1, 2, 3, and 4, respectively, yielding an average value of  $1452.30 \pm 28.25$ , while experimental the GC-RI is

**Table 5** Sensitivity analysis of the MDs for each ANN model

Model 1a										
MW	RBN	X2v	MATS2m	GATS2m	E3u	HTm	HATS3 m	R5m	nCs	nCrs
76.909	36.102	6.285	3.021	2.721	1.277	2.783	1.425	1.826	15.281	3.915
Model 2a										
DP01	J3D	H3D	RDF020 m	RDF020v	Mor27 m	Mor03v	Mor26v	Du	Dv	HTm
5.156	1.368	2.943	1.452	4.035	1.324	1.757	1.691	1.299	2.013	6.043
Model 3a										
ZM2 V	Dz	MSD	Rww	piPC01	piPC02	PCD	TIC2	IC4	MATS2 m	GATS2 m
3.080	59.853	3.476	3.579	2.133	7.101	1.185	1.584	1.307	1.957	1.458
Model 4a										
Ss	SCBO	nCIC	RBN	nDB	nCs	nCt	nCrs	nR_Cs	H_046	MLOGP
3.376	62.38	3.854	4.526	2.592	7.723	1.125	1.826	1.337	2.063	1.556

1436.33 ± 4.50 i.u., which is a good correspondence of averages with overlapped standard deviations. The utility of consensus modeling in enhancing the performance of QSRR models, particularly in the identification of compounds without commercially available references, has been validated in other reports in the literature (Dossin et al. 2016).

### Artificial neural network-based regression models

For the ANN models, the same variables (MDs) and set of training and test compounds used to build the final MLR models (i.e., 609 and 153 compounds, respectively) were employed. The input layer was comprised of 11 neurons corresponding to the models' independent variables/MDs, while the output layer contained a single neuron corresponding to the dependent variable (RI values). Table 3 shows the optimum parameters for the four ANN models obtained for 30,000 cycles (Eqs. 1a–4a).

With these parameters, four ANN models were trained and posteriorly validated over the test set. Table 4 shows the correlation parameters obtained with the ANN models. In addition, a comparison of the results obtained ANN regressions with the MLR models is performed to evaluate the contribution of non-linear relationships in modeling the GC–RI (see Table 4). As can be observed, all models yield minor improvements in correlations with the RI values for both the training and test sets using the ANN compared to the MLR technique. Although these improvements may not be considered as statistically significant, the incorporation of the ANN in consensus modeling should contribute to more robust and reliable predictions.

On the other hand, note that the SDEP values are higher for the ANN than MLR models due to compound 433, which has an extremely high relative error (see Supplementary Information, SI8). It can thus be inferred that this

compound is not adequately predicted by non-linear models but rather linear ones. Nonetheless, this compound is a member of the prediction set, and, therefore, the AD and validity of the ANN models are not affected. Residual analysis of the ANN models to check for possible systematic errors was performed and it was observed that the residual points are randomly propagated over both sides of the zero residue axes and, therefore, the regressions were correctly computed (see Supplementary Information SI9 for residual plots).

Moreover, the sensitivity of the variables (MDs) in the ANN model was assessed to determine their relative importance. This parameter is measured as the difference between SDE values when all MDs are considered as inputs [SDE (*n*)] and when the *i*th MD is excluded [SDE (*n*–1)], with both values computed over the same data set. Greater differences are associated with higher relevance for the excluded MD. Table 5 shows the sensitivity of the MDs in each of the models.

As can be observed, the most relevant MDs are the molecular weight ( $M_w$ ), mass weighted total autocorrelation MD on leverage matrix/H total index ( $H_{Tm}$ ), Pogliani index (Dz), and Sum of Conventional Bond Orders (SCBO) for models 1, 2, 3 and 4, respectively. The  $M_w$  and  $H_{Tm}$  indices are related with the structural bulk of chemicals, which, in turn, possesses a close relationship with the dispersion forces in chromatographic retention. On the other hand, the Dz distinguishes heteroatoms in a compound, while the SCBO characterizes bond types. To understand the relevance of these MDs, an inferential evaluation of the information codified is performed. First, it is known that heteroatomic compounds have (permanent) dipoles, while compounds with unsaturated bond systems (e.g., aromatic systems) are polarizable. Therefore, when the latter interacts with the former, the electric field from the permanent dipole induces a reverse dipole in the

**Table 6** Comparison of obtained results with those reported in the literature

No	Model	Technique	$N$	$N_{tr}$	$N_{pre}$	SDEC	$R^2$	$F$	$Q_{log}^2$	$Q_{ext}^2$	References
1	RI = 315.4270 ( $\pm 22.8496$ ) + 8.9710 ( $\pm 0.2768$ ) $M_w$ – 64.7769 ( $\pm 3.2362$ ) RBN – 68.5399 ( $\pm 5.0842$ ) $X_{2v}$ – 180.4735 ( $\pm 10.7349$ ) MATS2m – 178.2409 ( $\pm 11.8427$ ) GATS2m – 174.6514 ( $\pm 30.2488$ ) E3u + 60.5358 ( $\pm 6.9529$ ) JHTm – 821.4969 ( $\pm 123.1227$ ) HATS3m – 566.4689 ( $\pm 62.1945$ ) R5m + 46.4828 ( $\pm 2.2600$ ) nCs – 47.7686 ( $\pm 3.6273$ ) nCrS ANN architecture 11-8-1	MLR-GA	762	609	153	53.98	96.75	1616	96.54	98	Present study
2	RI = 300.00 + 3.83 ( $\pm 1.41$ ) + 21.7 ( $\pm 1.58$ ) $n_{19}$ + 18.26 ( $\pm 0.93$ ) $n_{28}$ + 18.89 ( $\pm 0.84$ ) $n_{37}$ + 15.28 ( $\pm 0.84$ ) $n_{46}$ + 5.78 ( $\pm 1.02$ ) $n_{or}$ – 0.93 ( $\pm 1.01$ ) $n_{me}$ + 0.43 ( $\pm 1.72$ ) $n_{pa}$	ANN MLR	762 43	609 30	153	53.48 0.86	97.18 0.9974	1073.3	–	98.35	Present study Schade and Andersson (2006)
3	RI = 1563.9 + 161.4 $G_1$ + 188.6 $G_2$ + 181.5 $G_3$ + 195.3 $G_4$ + 372.4 $G_5$ + 310.4 $G_6$ + 349.7 $G_7$ + 337.3 $G_8$ + 401.7 $G_9$ + 386.3 $G_{10}$ + 558.2 $G_{11}$ + 519.7 $G_{12}$ + 515.3 $G_{13}$ + 592.0 $G_{14}$ + 753.3 $G_{15}$ + 90.8 $G_{16}$ – 1.1 $G_{17}$	MLR- Stepwise	135	57	58	6.1	0.9995				Liang et al. (2000)
4	RI = 15.36 $n_{1,9}$ + 16.16 $n_{2,8}$ + 16.17 $n_{3,7}$ + 14.77 $n_{4,6}$ + 0.96 $n_{ortho}$ – 3.16 $n_{meta}$ – 0.9 $n_{para}$ + 2.73	MLR	25	18	7		0.9993				Sielex and Andersson (2000)
5	ANN architecture: 4-6-1 Inputs: Bp, $M_w$ , $Z$ , SF	ANN	96	32 (16*)	30					0.8800	Skrbic and Onjia (2006)
6	RI = 3.637*10 <sup>3</sup> + 8.392 RevDif <sub>x</sub> <sup>l</sup> – 79.799 MaxDP – 1.571*10 <sup>3</sup> RelnH – 1.310*10 <sup>3</sup> MaxQO – 2.204*10 <sup>3</sup> Rev <sub>x</sub> <sup>l</sup>	MLR- Stepwise	37	37		18.8	0.988	253.89	0.963		Hodj Mohammadi et al. (2004)
7	RI = 2205.9941 + 48.3581 ( $\pm 3.3325$ ) OEI – 111.7530 ( $\pm 11.3713$ ) $SX_{1CH}$ – 57.439 ( $\pm 22.4324$ ) $N^{2/3}$ – 155.4130 ( $\pm 9.5636$ ) $X_{eq}PEI$ – 504.5270 ( $\pm 131.4974$ ) $X_{eq}$ + 4.9453 ( $\pm 0.8324$ ) $MPEI_mIMPEI_m$	MLR	106	86	20	15.31	0.9901	653	0.9879	0.992	Liu et al. (2007)
8	RI = 141.068 (52.871) – 238.311 (109.969) RBF + 150.028 (54.587) IC1 – 2.901 (0.487) RDF045e + 41.373(4.699) $\mu$ – 27.025 (12.430) LUMO	GA-MLR	101	80	21	86.74	0.935		0.936	0.860	Noori (2012)
		GA-PLS GA-KPLS	101 101	80 80	21 21	82.51 68.59	0.942 0.967		0.942 0.968	0.871 0.919	Noori (2012) Noori (2012)

\* Validation set. Data sets: 7-essential oils components; 1-Alkylated dibenzothiophenols;

2-Polychlorinated dibenzofurans; 3-

Polychlorinated dibenzothiophenols; 4-

Polycyclic aromatic hydrocarbons; 5-

Phenytol, benzodiazepines, barbiturates; 6-

Oxygenated organic compounds (esters, ketones,

aldehydes, and alcohols)

polarizable system yielding dipole-induced dipole interactions. It can, therefore, be concluded that the Dz and SCBO are related with chromatographic induction forces.

### Comparison with other approaches reported in the literature

Finally, a quantitative comparison of the performance of the models obtained in the present study with those reported in the literature is performed with the aim of assessing the practical contribution of the obtained models in the prediction of the GC-RI of essential oils (see Table 6). As can be observed, the studies reported in the literature are based on much smaller sized data sets relative to the data set in the present study and in most cases congeneric in nature. Even then, similar results are obtained.

It can, therefore, be inferred that for the first time, QSRR models for predicting the RI of essential oil components with a wide AD, good statistical quality, robustness, and high predictive power are obtained. In addition, these models provide knowledge on the factors that influence the chromatographic retention of essential oils components over the DB-5 stationary phase.

### Conclusions

Retention indices have gained an increasingly relevant role in analytical chemistry given their utility in reducing false-positive (or negative) compound identification rates. Indeed, MS database repositories, e.g., NIST, Wiley, and FFNC currently include associated RIs to ensure more accurate identification of confounding molecular structures. In this report, QSRR models were built for Kováts retention indices based on a large and structurally diverse database of 791 essential oils components for the non-polar GC DB-5 column. These models were vigorously validated using both internal and external validation techniques on the training and test sets, respectively, and the corresponding statistical parameters were satisfactory, showing predictive ability of these models. The descriptors included in the prediction models provide information on the different molecular properties and/or interaction forces that influence the chromatographic retention/elution of essential oils components on the DB-5 stationary phase. All together, the obtained models provide valuable tools for the prediction of RIs for new essential oils components within the models' ADs and whose experimental data are undetermined.

**Acknowledgements** YM-P gives thanks to support from USFQ with partial finance of Project ID5400 "Chancellor Grant 2016".

### References

- Acevedo-Martínez J, Escalona-Arranz JC, Villar-Rojas A, Téllez-Palmero F, Pérez-Rosés R, González L et al (2006) Quantitative study of the structure-retention index relationship in the imine family. *J Chromatogr A* 1102:238–244. doi:[10.1016/j.chroma.2005.10.019](https://doi.org/10.1016/j.chroma.2005.10.019)
- Adams RP (2001) Identification of essential oil components by gas chromatography/quadrupole mass spectrometry, 3rd edn. Allured Publishing Corp, Illinois, p 456
- Albaugh DR, Hall LM, Hill DW, Kertesz TM, Parham M, Hall LH (2009) Prediction of HPLC retention index using artificial neural networks and IGroup E-state indices. *J Chem Inform Model* 49:788–799. doi:[10.1021/ci9000162](https://doi.org/10.1021/ci9000162)
- Alvarez R (1995) Estadística Multivariante y no Paramétrica con SPSS: aplicación a las ciencias de la salud. Díaz de Santos edn, Madrid
- Anker LS, Jurs PC, Edwards PA (1990) Quantitative structure-retention relationship studies of odor-active aliphatic compounds with oxygen-containing functional groups. *Anal Chem* 62:2676–2684
- Atkinson AC (1985) Plots, transformations and regression. Clarendon Press edn, Oxford
- Azar AP, Nekoei M, Riahi S, Ganjali MR, Zare K (2011) A quantitative structure-retention relationship for the prediction of retention indices of the essential oils of *Ammoides atlantica*. *J Serb Chem Soc* 76:891–902. doi:[10.2298/JSC100219076A](https://doi.org/10.2298/JSC100219076A)
- Babushok V (2015) Chromatographic retention indices in identification of chemical compounds. *TrAC Trends Anal Chem* 69:98–104. doi:[10.1016/j.trac.2015.04.001](https://doi.org/10.1016/j.trac.2015.04.001)
- Bajpai VK, Al-Reza SM, Choi UK, Lee JH, Kang SC (2009) Chemical composition, antibacterial and antioxidant activities of leaf essential oil and extracts of *Metasequoia glyptostroboides* Miki ex Hu. *Food Chem Toxicol* 47:1876–1883. doi:[10.1016/j.fct.2009.04.043](https://doi.org/10.1016/j.fct.2009.04.043)
- Devillers J (1996) Genetic algorithms in computer-aided molecular design. In: Devillers J (ed) Genetic algorithms in molecular modeling. Academic Press, London, pp 131–157
- Dimov N, Osman A, Mekenyan OV, Papazova D (1994) Selection of molecular descriptors used in quantitative structure-gas chromatographic retention relationships: I. Application to alkylbenzenes and naphthalenes. *Anal Chim Acta* 298:303–317. doi:[10.1016/0003-2670\(94\)00280-0](https://doi.org/10.1016/0003-2670(94)00280-0)
- Dossin E, Martin E, Diana P, Castellon A, Monge A, Pospisil P, Bentley M, Guy PA (2016) Prediction models of retention indices for increased confidence in structural elucidation during complex matrix analysis: application to gas chromatography coupled with high-resolution mass spectrometry. *Anal Chem* 88:7539–7547. doi:[10.1021/acs.analchem.6b00868](https://doi.org/10.1021/acs.analchem.6b00868)
- Duvenbeck C, Zinn P (1993) List operations on chemical graphs. 3. Development of vertex and edge models for fitting retention index data. *J Chem Inform Comput Sci* 33:211–219. doi:[10.1021/ci00012a005](https://doi.org/10.1021/ci00012a005)
- Fritz DF, Sahil A, Kováts E (1979) Determination of hydroxyl groups in poly(ethylene glycols). *Anal Chem* 51:7–12. doi:[10.1021/ac50037a010](https://doi.org/10.1021/ac50037a010)
- Garkani-Nejad Z, Karlovits M, Demuth W, Stimpfl T, Vycudilik W, Jalali-Heravi M, Varmuza K (2004) Prediction of gas chromatographic retention indices of a diverse set of toxicologically relevant compounds. *J Chromatogr A* 1028:287–295. doi:[10.1016/j.chroma.2003.12.003](https://doi.org/10.1016/j.chroma.2003.12.003)
- Gautzsch R, Zinn P (1996) Use of incremental models to estimate the retention indexes of aromatic compounds. *Chromatographia* 43:163–176. doi:[10.1007/BF02292946](https://doi.org/10.1007/BF02292946)

- Georgakopoulos CG, Kiboris JC, Jurs PC (1991a) Prediction of gas chromatographic relative retention times of stimulants and narcotics. *Anal Chem* 63:2021–2024
- Georgakopoulos CG, Tsika OG, Kiboris GC, Jurs PC (1991b) *Anal Chem* 63: 2025
- Görgényi M, Fekete Z, Seres L (1989) Estimation and prediction of the retention indices of selected trans-diazenes. *Chromatographia* 27:581–584. doi:10.1007/BF02258982
- Gramatica P (2007) Principles of QSAR models validation: internal and external. *QSAR Comb Sci* 26:694–701. doi:10.1002/qsar.200610151
- Hodj Mohammadi MR, Ebrahimi P, Pourmorad F (2004) Quantitative structure–retention relationships (QSRR) of some CNS agents studied on DB-5 and DB-17 phases in gas chromatography. *QSAR Comb Sci* 23:295–302. doi:10.1002/qsar.200530869
- Jalali-Heravi M, Ebrahimi-Najafabadi H (2011) Modeling of retention behaviors of most frequent components of essential oils in polar and non-polar stationary phases. *J Sep Sci* 34:1538–1546. doi:10.1002/jssc.201100042
- Jaworska J (2005) QSAR applicability domain estimation by projection of the training set in descriptor space: a review. *ATLA-NOTTINGHAM* 3:445–459
- Kaliszan R, Höltje HD (1982) Gas chromatographic determination of molecular polarity and quantum chemical calculation of dipole moments in a group of substituted phenols. *J Chromatogr A* 234:303–311. doi:10.1016/S0021-9673(00)81868-3
- Kiraly Z, Körtvélyesi T, Seres L, Görgényi M (1996) Structure-retention relationships in the gas chromatography of *N,N*-dialkylhydrazones. *Chromatographia* 42:653–659. doi:10.1007/BF02267697
- Kortvélyesi T, Gorgenyi M, Seres L (1995) Correlation of retention indices with van der Waals' volumes and surface areas: alkanes and azo compounds. *Chromatographia* 41:282–286. doi:10.1007/BF02688041
- Körtvélyesi T, Görgényi M, Héberger K (2001) Correlation between retention indices and quantum-chemical descriptors of ketones and aldehydes on stationary phases of different polarity. *Anal Chim Acta* 428:73–82. doi:10.1016/S0003-2670(00)01220-4
- Kováts ES (1958) Gas chromatographic characterization of organic compounds. I. Retention indexes of aliphatic halides, alcohols, aldehydes, and ketones. *Helv Chim Acta* 41:1915–1932
- Kubinyi H (1994) Variable selection in QSAR studies. I. An evolutionary algorithm. *Quant Struct Act Rel* 13:285–294. doi:10.1002/qsar.19940130306
- Leardi R (1994) Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection. *J Chemom* 8:65–79. doi:10.1002/cem.1180080107
- Liang X, Wang W, Schramm W, Zhang Q, Oxyinos K, Henkelmann B, Kettrup A (2000) A new method of predicting of gas chromatographic retention indices for polychlorinated dibenzofurans (PCDFs). *Chemosphere* 41:1889–1895. doi:10.1016/S0045-6535(00)00052-7
- Liu F, Liang Y, Cao C, Zhou N (2007) Theoretical prediction of the Kováts retention index for oxygen-containing organic compounds using novel topological indices. *Anal Chim Acta* 594:279–289. doi:10.1016/j.aca.2007.05.023
- Mondello L, Wiley FFNSC Library (2015) Mass spectra of flavors and fragrances of natural and synthetic compounds. Wiley, Hoboken
- NIST (2017) Mass Spectral Library, Standard Reference Data Program, ed, National Institute of Standards and Technology, Gaithersburg, Maryland
- Noori H (2012) Linear and nonlinear quantitative structure linear retention indices relationship models for essential oils. *Eurasian J Anal Chem* 8:50–63
- Noorizadeh H, Farmany A (2010) QSRR models to predict retention indices of cyclic compounds of essential oils. *Chromatographia* 72:563–569. doi:10.1365/s10337-010-1660-4
- Olivero J, Gracia T, Payares P, Vivas R, Díaz D, Daza E, Geerlings P (1997) Molecular structure and gas chromatographic retention behavior of the components of ylang–ylang oil. *J Pharm Sci* 86:625–630. doi:10.1021/js960196u
- Ong VS, Hites RS (1991) Relationship between gas chromatographic retention indexes and computer-calculated physical properties of four compound classes. *Anal Chem* 63:2829–2837. doi:10.1021/ac00024a005
- Osmialowski K, Halkiewicz J, Radecki A, Kaliszan R (1985) Quantum chemical parameters in correlation analysis of gas–liquid chromatographic retention indices of amines. *J Chromatogr A* 346:53–60. doi:10.1016/S0021-9673(00)90493-X
- Peng CT, Ding SF, Hua RL, Yang WC (1988) Prediction of retention indexes: I. Structure–retention index relationship on apolar columns. *J Chromatogr A* 436:137–172. doi:10.1016/S0021-9673(00)94575-8
- Qin L-T, Liu S-S, Liu H-L, Tong J (2009) Comparative multiple quantitative structure–retention relationships modeling of gas chromatographic retention time of essential oils using multiple linear regression, principal component regression, and partial least squares techniques. *J Chromatogr A* 1216:5302–5312. doi:10.1016/j.chroma.2009.05.016
- Qin L-T, Liu S-S, Chen F, Xiao Q-F, Wu Q-S (2013a) Chemometric model for predicting retention indices of constituents of essential oils. *Chemosphere* 90:300–305. doi:10.1016/j.chemosphere.2012.07.010
- Qin LT, Liu SS, Chen F, Wu QS (2013b) Development of validated quantitative structure–retention relationship models for retention indices of plant essential oils. *J Sep Sci* 36:1553–1560. doi:10.1002/jssc.201300069
- Rohrbaugh RH, Jurs PC (1986) Prediction of gas chromatographic retention indexes of polycyclic aromatic compounds and nitrated polycyclic aromatic compounds. *Anal Chem* 58:1210–1212. doi:10.1021/ac00297a052
- Rohrschneider L (1965) Die vorausberechnung von gaschromatographischen retentionszeiten aus statistisch ermittelten “Polaritäten”. *J Chromatogr A* 17:1–12. doi:10.1016/S0021-9673(00)99831-5
- Schade T, Andersson TJ (2006) Speciation of alkylated dibenzothiophenes through correlation of structure and gas chromatographic retention indexes. *J Chromatogr A* 1117:206–213. doi:10.1016/j.chroma.2006.03.079
- Sielex K, Andersson J (2000) Prediction of gas chromatographic retention indices of polychlorinated dibenzothiophenes on non-polar columns. *J Chromatogr A* 886:105–120. doi:10.1016/S0021-9673(99)01079-1
- Skrbic B, Onjia A (2006) Prediction of the Lee retention indices of polycyclic aromatic hydrocarbons by artificial neural network. *J Chromatogr A* 1108:279–284. doi:10.1016/j.chroma.2006.01.080
- Statsoft (2001) *Statistica*, 6th edn. Data Analysis Software System, Tulsa
- Todeschini R, Ballabio D, Consonni V, Mauri A, Pavan M (2004) MOBYDIGS computer software. TALETE srl, Milano
- Todeschini R, Consonni V, Mauri A, Pavan M (2007) DRAGON, v. 5.5, Talete srl, Milano
- von Mühlen C, Marriott PJ (2011) Retention indices in comprehensive two-dimensional gas chromatography. *Anal Bioanal Chem* 401:2351–2360
- Yan J, Liu X-B, Zhu W-W, Zhong X, Sun Q, Liang Y-Z (2015) Retention indices for identification of aroma compounds by GC: development and application of a retention index database. *Chromatographia* 78:89–108. doi:10.1007/s10337-014-2801-y

- Yancey JA (1994) Review of liquid phases in gas chromatography, part I: intermolecular forces. *J Chromatogr Sci* 32:349–357. doi:[10.1093/chromsci/32.8.349](https://doi.org/10.1093/chromsci/32.8.349)
- Zhang J, Zheng C-H, Xia Y, Wang B, Chen P (2017) Optimization enhanced genetic algorithm-support vector regression for the prediction of compound retention indices in gas chromatography. *Neurocomputing* 240:183–190. doi:[10.1016/j.neucom.2016.11.070](https://doi.org/10.1016/j.neucom.2016.11.070)
- Zhao C, Zeng Y, Wan M, Li R, Liang Y, Li C et al (2009) Comparative analysis of essential oils from eight herbal medicines with pungent flavor and cool nature by GC–MS and chemometric resolution methods. *J Sep Sci* 32:660–670. doi:[10.1002/jssc.200800484](https://doi.org/10.1002/jssc.200800484)