IRSTI 31.27.21 https://doi.org/10.26577/IJBCh202518215

S. Kherouf[1], Y. Driouche[2], N. Bouarra[3, 4*]

[1]Organic Synthesis Laboratory Modeling and Optimization of Chemical Processes (LOMOP), Badji Mokhtar University, Annaba, Algeria
[2]Environmental Research Center (CRE), Annaba, Algeria
[3]Center of Scientific and Technical Research in Physico-Chemical Analysis, Bou-Ismail, Tipaza, Algeria
[4]Laboratory of Environmental Engineering, Badji Mokhtar University, Annaba, Algeria
*e-mail: bouarranabil@yahoo.com/bouarra.nabil@crapc.dz

(Received 11 November 2025; received in revised form 19 December 2025; accepted 25 December 2025)

Linking molecular structure to chromatographic behavior: a quantitative structure-retention relationship study of *Olea europaea* L. essential oil components

Abstract. A robust quantitative structure-retention relationship (QSRR) model was developed to accurately predict the linear retention indices (LRI) of 51 essential oil compounds. Molecular descriptors were calculated using alvaDesc software, and model construction was achieved through a multiple linear regression (MLR) approach. A rigorous variable selection process identified relevant descriptors, resulting in a statistically significant model with strong predictive performance (R^2 = 0.9533, Q^2_{LOO} = 0.9339, Q^2_{LMO} = 0.9293, $RMSE_{tr}$ = 55.0581, s = 50.0169). External validation further confirmed the model's reliability, demonstrating excellent predictive capability (R^2_{ext} = 0.9381, Q^2_{F1} = 0.9361, Q^2_{F2} = 0.9354, Q^2_{F3} = 0.9646, CCC_{ext} = 0.9663, $RMSE_{ext}$ = 40.3308). The findings highlight the efficiency of the QSRR-MLR model in predicting retention indices, providing valuable insights into molecular properties influencing compound retention. Additionally, the applicability domain assessment ensured reliable predictions within the studied chemical space. This methodology offers a deeper understanding of chromatographic behavior and presents potential for application to other chemical classes for predictive modeling.

Keywords: QSRR, essential oil, volatile chemicals, prediction set, validation.

Introduction

The essential oils are aromatic and volatile chemicals extracted from plants by distillation under steam [1]. They smell like the original plant they were taken from. Essential oils have been used for therapeutic purposes for thousands of years; they are also fascinating and potent plant-based substances. They have been quite significant up till the current day. Essential oils also have various uses in the fields of flavoring, food, perfume, and the cosmetics industry [2, 3]. The olive tree (*Olea europaea* L.), a member of the *Oleaceae* family, is a regional specialty [4, 5]. Antioxidant, antibacterial, antiviral, hypoglycemic, anti-inflammatory, and anticancer bioactivities are only some of the many attributed to olive bioactive components. Because of its potential health advantages, olive bioactive components have attracted much attention [6].

The primary techniques for identifying these plant oils are gas chromatography (GC) and liquid chromatography (LC). In GC analysis, volatile chemicals or molecules with low volatility which may be chemically transformed to more appropriate molecules are often separated and analyzed using this approach [7]. When utilized under controlled analytical circumstances, the GC's single output parameter (retention index) allows for the identification of any volatile chemical [6]. Several intermolecular interactions, including dipole-diploe forces, dipole-induced forces, hydrogen bonds, etc., affect whether a compound elutes or is retained [8]. Looking for a quantitative link between molecular structure and GC retention indices is a fundamental task in chemistry. More theoretical depth may be gained by understanding the relationships between mobile and stationary phases by correlating retention time values and molecular structure. In addition, they may provide crucial details on how chemical structure influences retention behavior and potential absorption and elution processes [9].

One of the aims of this scientific endeavor is to allow accurate prediction of physicochemical properties of compounds based only on their

molecular structure [10]. Many different approaches can be employed in establishing a relationship between the molecular structure of a compound and the physicochemical characteristics; quantitative structure-property relationship (QSPR) is one of the successful methods that has been used for this purpose. Each QSPR study has as its foundation the pursuit of optimal quantitative relationships for use in property prediction from molecular structures [11]. Once a solid relationship has been established, it may be used to predict similar characteristics in other buildings that are still without any kind of measurement or even design. QSRRs are statistical models that allow for the prediction of retention indices of new compounds by quantifying the relationship between a molecule's structure and its chromatographic retention index [12]. These relationships may provide theoretical light on the complex interplay that occurs between chemicals, mobile phases, and stationary phases during chromatographic analyses. They are also able to give useful information on the influence of the structure of the chemicals on the behavior of retention as well as the probable mechanisms behind adsorption and elution processes [13]. Studies using QSRR to estimate retention indices (RI) for a variety of organic compounds have been published in recent years. Choosing the right set of variables is essential for these methods to provide accurate predictions [14, 15].

Recent studies have focused on enhancing QSRR models by incorporating advanced feature selection techniques and expanding datasets to improve prediction accuracy and robustness. Some used multiple linear regression (MLR) and partial least squares (PLS) models with descriptors chosen by a genetic algorithm to predict retention for 80 oils including 20 test compounds [16, 17]. The support vector machine (SVM) nonlinear models were used in other studies employing the same dataset as a means of enhancing prediction performance [16]. In other work, QSRR models have also been published for retention indices of essential oils using GA with MLR or PLS, kernel PLS and Levenberg-Marquardt artificial neural networks [18-20]. Driouche and Messadi [21] developed a QSRR model for the prediction of retention indices of essential oil constituents from *Thymus vulgaris Lamiaceae* and demonstrated its efficiency and reliability through external validation. Navabi et al. [22] developed a GA-MLR and GA-BPANN model for predicting retention indices of essential oil components in Polygonum minus essential oil.

The objective of this study is to develop a robust QSRR model for predicting the retention indices of 51 essential oil components derived from *Olea europaea* L. grown in Mediterranean and arid regions of Algeria. By employing advanced statistical methods and validation techniques, this work aims to provide accurate and reliable predictions of LRIs, contributing to a deeper understanding of the chromatographic behavior of essential oils and their molecular properties.

Materials and methods

Dataset collection

This study used experimental linear retention index (LRI) data sourced from the work of Tlili et al. [23]. The LRI values, presented in Table 2, were determined through GC-MS analysis were performed with a Varian CP-3800 gas-chromatograph equipped with a DB-5 capillary column and a Varian Saturn 2000 ion trap mass detector. The dataset includes 51 volatile compounds identified in *Olea europaea* L. cultivars, providing detailed insight into their chromatographic retention characteristics.

Descriptors generation

The structures of the compounds are produced using MarvinSketch V24.3.0 [24]. Thereafter, PM7 semi-empirical method has been used in MOPAC software (version 2016, Stewart Computational chemistry) [25] to improve the finale geometry. The resulting minimum-energy conformations were used to calculate three-dimensional geometrical descriptors with alvaDesc (version 3.0.4) [26]. These descriptors, derived from the 3D arrangement of atoms, capture molecular shape, size, and spatial orientation. They help distinguish between structurally similar compounds and reflect conformational variability across molecules.

Dataset Division

The current challenge in QSRR model development is building a model that can accurately predict the behavior of new chemical compounds. When external experimental data are not available for validation, a common approach is to divide the available dataset into two parts: a training set used to construct the model, and a prediction set used to assess its predictive performance. The prediction set typically represents 15 to 40% of the total dataset [27].

In this study, the Kennard and Stone algorithm (CADEX) [28], was used to divide the 51 compounds based on their molecular descriptors. This deterministic method ensures an even and representative distribution of samples across the descriptor space. The algorithm starts by selecting the two most dissimilar samples using Euclidean distance, then iteratively adds the next most distant compound from those already selected. This process continues until all compounds are ranked.

The dataset was split into a training set of 35 compounds and a prediction set of 16 compounds. The training set was used to establish the relationship between the descriptors and the linear retention index (LRI), while the prediction set was used to evaluate the model's robustness, goodness-of-fit, and external predictive power.

Model development and validation

The quantitative structure-retention relationship (QSRR) model was developed using the Multiple Linear Regression (MLR) technique, based on the ordinary least squares (OLS) method as implemented in the QSARINS software (version 2.2.4) [29]. This modeling approach was selected for its simplicity, reproducibility, and interpretability. After preparing and splitting the dataset, molecular descriptor selection was performed on the training set by evaluating all possible combinations of up to four descriptors. This step ensured exhaustive coverage of low-dimensional models before considering more complex ones. The optimal model was chosen based on its leave-one-out cross-validation coefficient (Q^2_{LOO}), ensuring a balance between predictive performance and model simplicity.

The final regression equation takes the general form of:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n \quad (1)$$

where $\hat{\boldsymbol{y}}$ is the predicted linear retention index (LRI); $\boldsymbol{x_i}$ are the selected molecular descriptors, and $\boldsymbol{b_i}$ are their respective regression coefficient. Only descriptor with p-values ≤ 0.05 were retained to ensure statistical significance and reduce overfitting.

To assess internal model performance and robustness, two cross-validation techniques were applied: Leave-One-Out (LOO) and Leave-Many-Out (LMO). The Q^2_{LOO} value, indicating how well the model predicts compounds not included during training, was calculated using the following equation:

$$Q^2_{Loo} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_{i/i} - y_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

where $\boldsymbol{y_i}$ is the experimental value, $\hat{\boldsymbol{y}}_{\boldsymbol{i/i}}$ is the predicted value for compound $\boldsymbol{i}$ when it is omitted from the training set, and $\bar{\boldsymbol{y}}$ is the mean of the experimental values in the training set.

For external validation, the dataset of 51 compounds was divided using the Kennard and Stone (CADEX) algorithm into training set of 35 compounds and a prediction set of 16 compounds. This deterministic method ensures a representative and uniform distribution across the descriptor space. The model's predictive performance was assessed using several statistical indicators, such as Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and CCC_{ext} (Eq. 3-8), which have been previously described and discussed in our earlier publications [30-32].

$$Q^2_{F1} = 1 - \frac{PRESS_{EXT}}{SS_{EXT}(\bar{y}_{TR})} \quad (3)$$

$$PRESS_{EXT} = \sum(y_i - \hat{y}_i)^2 \quad (4)$$

$$Q^2_{F2} = 1 - \frac{PRESS_{EXT}}{SS_{EXT}(\bar{y}_{EXT})} \quad (5)$$

$$SS_{EXT}(\bar{y}_{EXT}) = \sum(y_i - \bar{y}_{EXT})^2 \quad (6)$$

$$Q^2_{F3} = 1 - \frac{\left(\frac{PRESS_{EXT}}{n_{EXT}}\right)}{\left(\frac{TSS}{n_{TR}}\right)} \quad (7)$$

$$CCC_{ext} = \frac{2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2 + n\left(\bar{y} - \bar{\hat{y}}\right)^2} \quad (8)$$

where n_{ext} and n_{tr} are the number of compounds in the prediction and training sets, and $\bar{y}_{tr}$ and $\bar{y}_{ext}$ are the means of the training and prediction experimental values.

To verify the statistical robustness and eliminate the possibility of chance correlation, a Y-randomization test (Y-scrambling) was conducted. The dependent variable (LRI) was randomly permuted, and models were rebuilt. A sharp drop in Q^2 values after scrambling confirmed the model's validity.

Applicability domain

The model's applicability domain was assessed using the Williams plot, which shows standardized residuals alongside leverage values. Williams plots are used to identify outliers and establish a confidence interval for the model's predictions [33]. It guarantees that the investigation adheres to the third principal of the OECD [34]. Leverage is calculated as:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (9)$$

where x_i is the descriptor row vector of the query compound, and X is the n*k matrix containing k model descriptor values for n training set compounds. Transpose of matrix/vector denoted by the superscript ''T'' the control leverage h* is fixed at (3k+1)/n, where k is the number of model parameters and n is the number of observations used to compute the model. Compounds with $h_i > h^*$ or residuals outside ±3 standard deviations were considered influential or outlier compounds, respectively.

The QSRR-MLR model developed in this study met all these conditions, demonstrating excellent fit, predictive power, and structural interpretability. These results confirm the model's reliability for predicting LRI values of essential oil constituents and its applicability in chromatographic behavior modeling.

Results and discussion

Modeling of the linear retention index (LRI)

In this study, the linear retention indices (LRI) of essential oil compounds were modeled using a quantitative structure-retention relationship (QSRR) approach based on multiple linear regression (MLR). The modeling was performed using the QSARINS software [29] a robust platform for building, validating, and interpreting QSAR/QSPR models using a genetic algorithm-based variable selection approach and extensive internal and external validation techniques. The regression model is expressed as:

$$\textbf{LRI}=373.90+1.83\ \textbf{BertzCT}+116.71\ \textbf{Hy}+284.87\ \textbf{Chi1\_EA(ed)}-262.53\ \textbf{Mor25v} \quad (10)$$

where: **BertzCT:** Bertz complexity index; **Hy:** Hydrophilic factor; **Chi1_EA(ed)**: Connectivity-like index of order 1 from edge adjacency mat. Weighted by edge degree; **Mor25v**: 3D-MoRSE.signal 25 / weighted by van der Waals volume.

The model's accuracy and predictive performance were evaluated using key statistical metrics summarized in Table 1.

Table 1 – Statistical performance indicators for the training and prediction sets

Statistical Parameters			
Training Set		Prediction Set	
R^2	0.954	R^2_{ext}	0.915
Q^2_{LOO}	0.946	Q^2_{F1}	0.909
$RMSE_{tr}$	25.70	Q^2_{F2}	0.904
s	23.40	Q^2_{F3}	0.910

The QSRR model shows strong statistical performance in both training and prediction sets. In the training phase, the model achieved an R^2 of 0.954, indicating that 95.4% of the variation in experimental LRI values is captured by the selected descriptors. The Q^2_{LOO} value of 0.946, close to R^2, confirms model robustness and lack of overfitting. Low RMSE (25.70) and standard error (s=23.40) indicate accurate predictions within the training data. The high Fisher statistic (F=208.50) supports the significance of the regression.

External validation further confirms model reliability. The prediction set yielded R^2_{ext} = 0.915, with Q^2_{F1}, Q^2_{F2}, and Q^2_{F3} all above 0.90, exceeding the OECD's threshold (>0.5) for valid QSAR models. The external concordance correlation coefficient (CCC_{ext} = 0.957) and $RMSE_{ext}$ = 20.57 affirm high predictive power and low prediction error. Together, these metrics demonstrate that the model is both statistically sound and externally predictive.

Table 2 illustrates the prediction results and the values of the descriptors used in the model developed.

The coefficients and associated statistical parameters are summarized (Table 3).

Table 2 – List of names, experimental, predicted LRI, and descriptors values involved in the model

ID	Name	Status	Exp.LRI	Pred. LRI	BertzCT	Hy	Chi1_EA (ed)	Mor25v
1	hexanal	Tr	802	831.064	41.445	-0.802	1.644	-0.025
2	(E)-2-hexenal	Tr	856	875.783	64.58	-0.802	1.644	-0.034
3	(Z)-3-hexen-1-ol	Tr	857	914.321	48.142	-0.088	1.644	0.022
4	1-hexanol	Tr	869	872.599	27.361	-0.088	1.644	0.036
5	n-nonane	Tr	900	907.511	33.303	-0.954	2.144	0.102
6	heptanal	Tr	903	899.535	50.349	-0.828	1.894	0.036
7	(Z)-2-heptenal	Tr	958	955.484	74.456	-0.828	1.894	-0.009
8	6-methyl-5-hepten-2-one	Tr	987	985.815	118.529	-0.848	1.695	-0.042
9	(E,E)-2,4-heptadienal	Tr	1011	980.633	100.954	-0.828	1.894	0.08
10	limonene	Tr	1032	1064.477	162.877	-0.96	1.779	0.009
11	phenylacetaldehyde	Tr	1045	1074.079	170.744	-0.848	1.919	0.229
12	n-undecane	Tr	1100	1099.823	49.059	-0.965	2.644	0.017
13	linalool	Tr	1101	1110.171	154.256	-0.294	1.733	0.021
14	phenylethyl alcohol	Tr	1111	1111.710	155.406	-0.213	1.919	0.261
15	(E,E)-2,6-nonadienal	Tr	1153	1194.102	122.69	-0.864	2.394	-0.055
16	menthone	Tr	1154	1102.784	148.952	-0.877	1.851	-0.119
17	methyl chavicol	Tr	1197	1191.035	209.257	-0.877	2.093	0.228
18	safranal	Tr	1198	1174.535	231.257	-0.877	1.655	-0.031
19	-cyclocitral	Tr	1222	1146.305	198.918	-0.877	1.655	-0.149
20	1,2-benzisothiazole	Tr	1223	1205.518	245.642	-0.742	1.851	0.224
21	carvone	Tr	1244	1233.890	223.027	-0.877	1.837	-0.117
22	(E)-2-decenal	Tr	1262	1232.220	105.44	-0.877	2.644	-0.055
23	bornyl acetate	Tr	1287	1283.986	269.495	-0.835	2.12	0.342
24	(E,Z)-2,4-decadienal	Tr	1293	1253.442	133.808	-0.877	2.644	0.062
25	theaspirane I	Tr	1298	1341.364	264.409	-0.905	2.157	0.097
26	p-vinylguaiacol	Tr	1314	1251.880	256.736	-0.206	1.862	0.375
27	theaspirane II	Tr	1315	1392.556	264.409	-0.905	2.157	-0.098
28	phenylethyl propionate	Tr	1351	1381.369	245.264	-0.822	2.627	0.358
29	(E)- -damascenone	Tr	1382	1527.077	327.145	-0.905	2.285	-0.034
30	n-tetradecane	Tr	1400	1371.431	74.039	-0.975	3.394	-0.034
31	dihydrodehydro- -ionone	Tr	1422	1457.342	292.357	-0.905	2.32	0.027
32	(E)-geranylacetone	Tr	1454	1449.609	229.62	-0.905	2.685	0.015
33	dihydroactinidiolide	Tr	1536	1404.142	288.465	-0.822	1.998	-0.11
34	caryophyllene oxide	Tr	1581	1557.778	330.327	-0.917	2.521	0.122
35	benzophenone	Tr	1627	1609.632	352.152	-0.905	2.885	0.477
36	benzaldehyde	Pr	963	951.392	158.826	-0.828	1.624	0.302
37	3-ethenyl pyridine	Pr	969	998.288	172.826	-0.828	1.624	0.221
38	(E,Z)-2,4-heptadienal	Pr	997	980.896	100.954	-0.828	1.894	0.079
39	octanal	Pr	1002	1011.323	59.588	-0.848	2.144	-0.063
40	p-cymene	Pr	1028	1059.216	180.078	-0.96	1.779	0.149
41	1-octanol	Pr	1071	1030.862	43.832	-0.213	2.144	0.035
42	nonanal	Pr	1103	1082.127	69.129	-0.864	2.394	-0.002

Continuation of the table

ID	Name	Status	Exp.LRI	Pred. LRI	BertzCT	Hy	Chi1_EA (ed)	Mor25v
43	1,4-Dimethyl-DELTA.-3-tetrahydroacetophenone	Pr	1150	1156.890	202.918	-0.877	1.707	-0.105
44	(E)-2-nonenal	Pr	1163	1136.662	94.901	-0.864	2.394	-0.03
45	2-phenylethyl formate	Pr	1177	1277.730	197.603	-0.79	2.467	0.261
46	n-dodecane	Pr	1200	1156.257	56.439	-0.969	2.894	0.123
47	n-tridecane	Pr	1300	1279.479	66.107	-0.972	3.144	-0.009
48	(E,E)-2,4-decadienal	Pr	1317	1245.304	133.808	-0.877	2.644	0.093
49	eugenol	Pr	1358	1369.930	269.563	-0.244	2.157	0.318
50	(Z)-jasmone	Pr	1395	1340.027	233.389	-0.888	2.306	0.055
51	(E)- -ionone	Pr	1488	1476.244	292.357	-0.905	2.32	-0.045

Table 3 – Characteristic of the descriptors in the optimal MLR model

Descriptor	Coefficient	Std. Coefficient	Std. Error	(+/-) Co. int. 95%	p-value
Intercept	373.90	–	44.53	90.9323	0.000
BertzCT	1.83	0.812	0.10	0.204	0.000
Hy	116.71	0.136	41.57	84.8926	0.0083
Chi1_EA(ed)	284.87	0.561	23.57	48.1442	0.000
Mor25v	-262.53	-0.189	64.93	132.6146	0.0003

The statistical quality of the model is high, as indicated by the very low p-values (< 0.01) for all descriptors, confirming their significant contribution to the prediction of the linear retention index.

Assessing the model's applicability domain is a key part of the validation process. The Williams plot (Fig. 1) shows standardized residuals versus leverage values, which reflect how similar each compound is to those in the training set. All residuals fall within the ±3s range, except for one training compound ((E)- - damascnone), identified as a Y outlier due to a likely error in its experimental value.

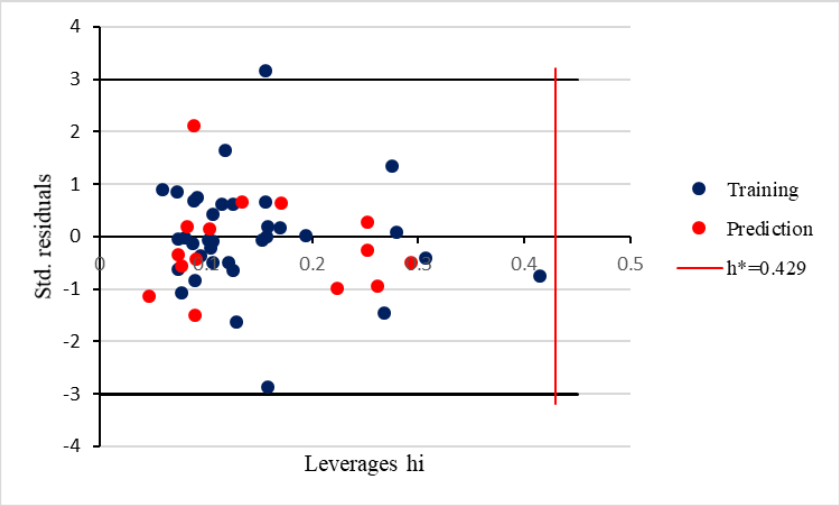

Figure 1 – Williams plot of the developed MLR model showing standardized residuals versus leverage values

All leverage values (h_i) remain below the threshold ($h^* = 0.429$), indicating that no compound exerts excessive influence on the model. This confirms that the MLR model provides reliable linear retention index (LRI) predictions. The model can be applied to screen chemical databases or virtual compounds, using the applicability domain to exclude structurally dissimilar molecules.

Figure 2 shows the correlation between the predicted and experimental linear retention index (LRI) values obtained using the QSRR-MLR model. Red dots represent the training set, while blue dots correspond to the external prediction set. Most points are closely aligned along the diagonal line, indicating a strong agreement between observed and predicted values.

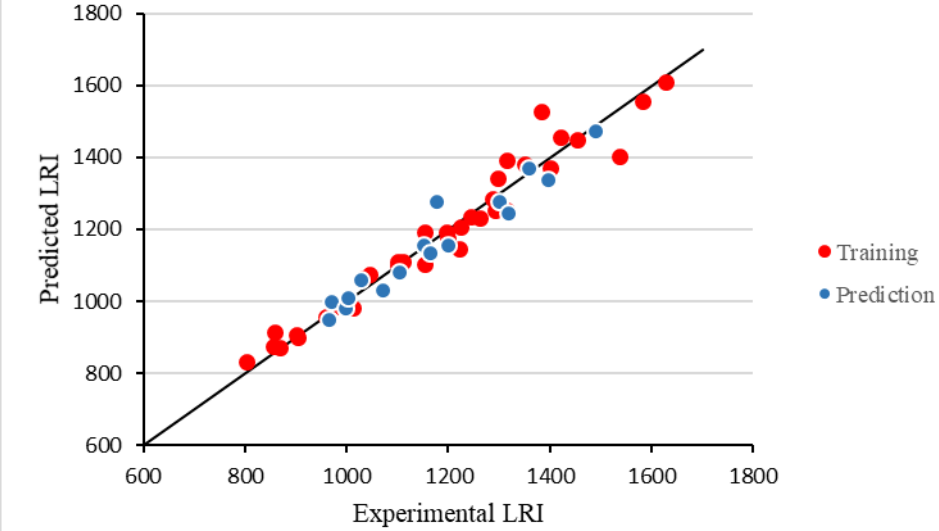

Figure 2 – Plot of predicted vs Experimental LRI values

The determination coefficients (R^2 = 0954 for the training set and R^2_{ext} = 0.915 for the prediction set) confirm that the model performs well both in fitting the training data and in predicting unseen data. The absence of major outliers supports the robustness and reliability of the regression model in capturing retention behavior based on molecular descriptors.

Figure 3 illustrates the results of the Y-randomization (or Y-scrambling) test used to assess the statistical reliability of the QSRR model. Red and yellow dots represent the R^2 and Q^2 values, respectively, for models developed from randomly shuffled response data. The blue dots in the upper right corner correspond to the original model's R^2 and Q^2 values.

The randomized models yield significantly lower R^2 and Q^2 values, mostly clustered near or below zero, clearly separated from the original model's performance. This confirms that the developed QSRR model is not a result of chance correlations. It demonstrates genuine predictive ability based on meaningful structural information encoded in the selected descriptors.

Model descriptors interpretation

The regression model obtained in equation 10, provides a coherent mechanistic view of the chromatographic retention of essential oil compounds on a DB-5 type nonpolar stationary phase. On such a column, the separation mechanisms are mainly governed by London dispersion forces, where retention increases with the size, molecular surface area, and boiling point of the analytes [35]. Analysis of the model coefficients allows us to decipher the quantitative influence of different molecular properties on the retention index.

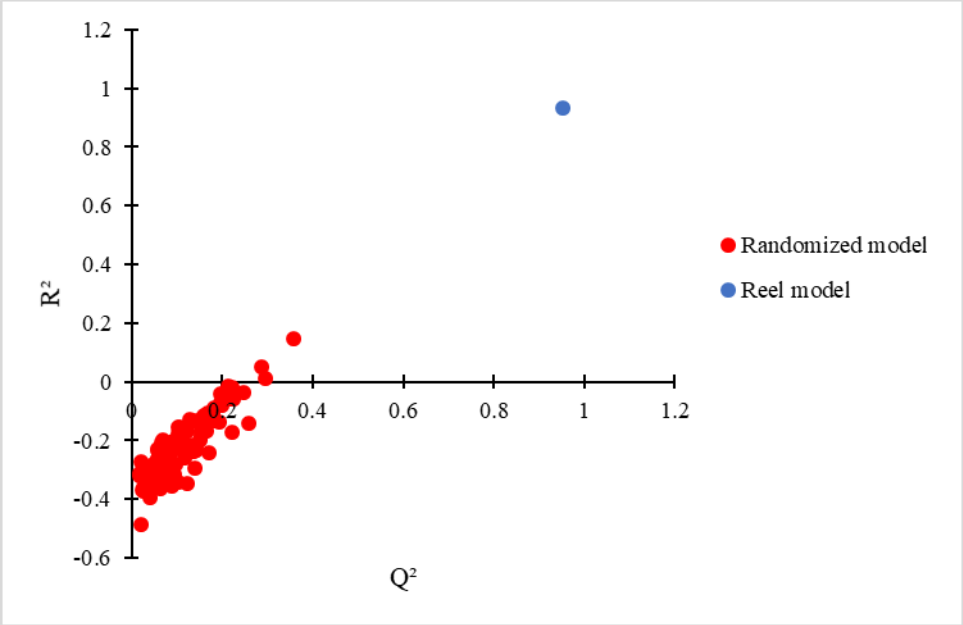

Figure 3 – Y-Randomization test

The topological descriptor Chi1_EA(ed) (connectivity index) and Bertz complexity index (BertzCT) positively influence retention, with respective coefficients of +284.87 and +1.83. The high coefficient of the Chi1_EA(ed) index reflects the electronic and topological connectivity of a molecule: a high value reflects a more rigid, better connected, and often conjugated or aromatic structure [36]. These characteristics promote polarization and π–π or dispersion interactions with the stationary phase, which prolongs the elution time. Similarly, BertzCT quantifies overall structural complexity by taking into account ramifications, cycles, and types of atoms present [37]. More complex and ramified molecules have an increased contact surface area and interact more strongly with the nonpolar DB-5 phase, which delays their elution. This behavior is well documented in QSRR models based on topological descriptors [38].

The Mor25v descriptor, derived from 3D-MoRSE descriptors, has the most strongly negative coefficient (-262.53), acting as the main factor unfavorable to retention. The 3D-MoRSE descriptors encode precise information about the three-dimensional geometry of the molecule based on simulated electron diffraction [39]. The negative coefficient suggests that this descriptor is an indicator of molecular compactness or sphericity. A more compact molecule, for the same molar mass, has a smaller external contact surface, which significantly reduces the possible dispersion forces with the stationary phase [40]. The negative coefficient suggests that a molecular geometry leading to a more dispersed electron distribution (and therefore a smaller effective contact surface with the stationary phase) promotes faster elution.

The Hy (hydrophilic factor) descriptor has a positive coefficient (+116.71), indicating that polarity contributes modestly but significantly to retention. Although the DB-5 phase is essentially nonpolar, it contains about 5% phenyl groups, whose π electrons are polarizable. Thus, analytes with polar groups (hydroxyl, carbonyl, ester) can interact via induced dipole-dipole forces with the aromatic rings of the phase [40]. These interactions, although secondary to dispersion forces, slightly increase retention. In addition, more polar molecules generally have higher boiling points, which reduces their volatility and indirectly increases their retention. This behavior has already been observed in QSRR studies on polar compounds in essential oils [35]. The Hy (hydrophilic factor) descriptor has a positive coefficient (+116.71), indicating that polarity contributes modestly but significantly to retention. Although the DB-5 phase is essentially non-polar, it contains approximately 5% phenyl groups, whose π electrons are polarizable. Thus, analytes with polar groups (hydroxyl, carbonyl, ester) can interact via induced dipole-dipole forces

with the aromatic rings of the phase [40]. These interactions, although secondary to dispersion forces, slightly increase retention. In addition, more polar molecules generally have higher boiling points, which reduces their volatility and indirectly increases their retention. This behaviour has already been observed in QSRR studies on polar compounds in essential oils [35].

Conclusion

The QSRR-MLR model developed in this work provides an effective and interpretable approach for predicting the linear retention indices (LRI) of 51 essential oil components from *Olea europaea* L. The model, built using a rigorously selected set of molecular descriptors, demonstrates strong statistical performance with high internal (R^2 = 0.9533, Q^2_{LOO} = 0.9339) and external validation metrics (R^2_{ext} = 0.9381, $Q^2_{F1\text{-}F3}$ > 0.93, CCC_{ext} = 0.9663). The selected descriptors reflect key structural features-topological complexity, polarity, connectivity, and 3D molecular shape-that significantly influence chromatographic behavior. The model passed all recommended validation steps, including Y-randomization and applicability domain analysis, confirming its reliability and robustness. These results support the use of QSRR-MLR as a practical and accurate tool for retention time prediction in essential oil analysis. The methodology can be extended to other volatile compounds datasets, facilitating compound identification, optimization of chromatographic methods, and molecular-level understanding of retention mechanisms.

Acknowledgments

We thank Prof. Paola Gramatica for the free license of QSARINS software. We are thankful to the Algerian Directorate-General for Scientific Research and Technological Development (DGRSDT) for providing financial assistance for this research.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

1. Hyldgaard M., Mygind T., Meyer R.L. (2012) Essential oils in food preservation: mode of action, synergies, and interactions with food matrix components. *Front Microbiol*., vol. 3, p. 12. https://doi.org/10.3389/fmicb.2012.00012.
2. Baris O., Güllüce M., Sahin F., Ozer H., Kılıc H., Ozkan H., Sökmen M., Ozbek T. (2006) Biological activities of the essential oil and methanol extract of Achillea biebersteinii Afan (Asteraceae). *Turk J Biol*., vol. 30, pp. 65–73.
3. Wei A., Shibamoto T. (2010) Antioxidant/Lipoxygenase inhibitory activities and chemical compositions of selected essential oils. *J Agric Food Chem*., vol. 58, pp. 7218–7225. https://doi.org/10.1021/jf101077s.
4. Gilani A.H., Khan A.U. (2010) Medicinal value of combination of cholinergic and calcium antagonist constituents in olives. In: Preedy V.R., Watson R.R. (Eds.), Olives and Olive Oil in Health and Disease Prevention. Elsevier: Amsterdam, The Netherlands, pp. 835–843. https://doi.org/10.1016/B978-0-12-374420-3.00089-9.
5. Mannina L., Segre A.L. (2010) NMR and olive oils: A geographical characterization. In: Preedy V.R., Watson R.R. (Eds.), Olives and Olive Oil in Health and Disease Prevention. Elsevier: Amsterdam, The Netherlands, pp. 117–124. https://doi.org/10.1016/B978-0-12-374420-3.00014-0.
6. Jiang L., Lu J., Qin Y., Jiang W., Wang Y. (2020) Antitumor effect of guava leaves on lung cancer: A network pharmacology study. *Arab J Chem*., vol. 13, pp. 7773–7797. https://doi.org/10.1016/j.arabjc.2020.09.010.
7. Pavlić B., Teslić N., Kojić P., Pezo L. (2020) Prediction of the GC–MS retention time for terpenoids detected in sage (Salvia officinalis L.) essential oil using QSRR approach. *J Serb Chem Soc*., vol. 85, no. 1, pp. 9–23. https://doi.org/10.2298/JSC190416097P.
8. Azar A.P., Nekoei M., Riahi S., Ganjali M.R., 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*., vol. 76, no. 6, pp. 891–902. https://doi.org/10.2298/JSC100219076A.
9. Pourbasheer E., Riahi S., Ganjali M.R., Norouzi P. (2010) Quantitative structure–retention relationship (QSRR) models for predicting the GC retention times of essential oil components. *Acta Chromatogr*., vol. 22, no. 3, pp. 357–373. https://doi.org/10.1556/achrom.22.2010.3.2.
10. Rojas C., Duchowicz P., Tripaldi P., Diez R.P. (2015) QSPR analysis for the retention index of flavors and fragrances on a OV-101 column. *Chemom Intell Lab Syst*., vol. 140, pp. 126–132. https://doi.org/10.1016/j.chemolab.2014.09.020.
11. Hu R., Liu H., Zhang R., Xue C., Yao X., Liu M., et al. (2005) QSPR prediction of GC retention indices for nitrogen-containing polycyclic aromatic compounds from heuristically computed molecular descriptors. *Talanta*., vol. 68, no. 1, pp. 31–39. https://doi.org/10.1016/j.talanta.2005.04.034.

12. Luan F., Liu H.T., Wen Y., Zhang X. (2008) Quantitative structure–property relationship study for estimation of quantitative calibration factors of some organic compounds in gas chromatography. *Anal Chim Acta*., vol. 612, no. 2, pp. 126–135. https://doi.org/10.1016/j.aca.2008.02.037.

13. Polyakova Y., Jin L., Row K. (2006) Linear regression based QSPR models for the prediction of the retention mechanism of some nitrogen containing heterocycles. *J Liq Chromatogr Relat Technol*., vol. 29, pp. 533–552. https://doi.org/10.1080/10826070500479062.

14. Navabi A., Isfahani T., Ramazani M., Alimoradi M. (2021) QSPR models for predicting retention indices of Polygonum minus Huds. essential oil composition using GA–BWMLR and GA–BPANN methods. *J Essent Oil Bear Plants*., vol. 24, pp. 879–896. https://doi.org/10.1080/0972060X.2021.1976284.

15. Bayat Z., Yazdan Abad M.F. (2011) Quantitative structure–property relationship (QSPR) study of Kovats retention indices of some adamantane derivatives by the genetic algorithm and multiple linear regression (GA–MLR) method. *Pet Coal*., vol. 53, no.2.

16. Riahi S., Pourbasheer E., Ganjali M.R., Norouzi P. (2009) Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine. *J Hazard Mater*., vol. 166, no. 2–3, pp. 853–859. https://doi.org/10.1016/j.jhazmat.2008.11.097.

17. Noorizadeh H., Farmany A. (2010) QSRR models to predict retention indices of cyclic compounds of essential oils. *Chromatographia*., vol. 72, pp. 563–569. https://doi.org/10.1365/s10337-010-1660-4.

18. Noorizadeh H., Farmany A., Noorizadeh M. (2011) Quantitative structure–retention relationships analysis of retention index of essential oils. *Quim Nova*., vol. 34, pp. 242–249. https://doi.org/10.1590/S0100-40422011000200014.

19. Noorizadeh H., Farmany A. (2010) Exploration of linear and nonlinear modeling techniques to predict retention index of essential oils. *J Chin Chem Soc*., vol. 57, pp. 1268–1277. https://doi.org/10.1002/jccs.201000188.

20. Noorizadeh H., Farmany A., Khosravi A. (2013) Investigation of retention behaviors of essential oils by using QSRR. *J Chin Chem Soc*., vol. 57, pp. 982–991. https://doi.org/10.1002/jccs.201000137.

21. Driouche Y., Messadi D. (2019) Quantitative structure–retention relationship model for predicting retention indices of constituents of essential oils of Thymus vulgaris (Lamiaceae). *J Serb Chem Soc*., vol. 84, no. 4, pp. 405–416. https://doi.org/10.2298/JSC180817010D.

22. Navabi A., Isfahani T., Ramazani M., Alimoradi M. (2021) QSPR models for predicting retention indices of Polygonum minus Huds. essential oil composition using GA–BWMLR and GA–BPANN methods. *J Essent Oil Bear Plants*., vol. 24, pp. 879–896. https://doi.org/10.1080/0972060X.2021.1976284.

23. Tlili A., Bouziane M., Flamini G., Hadj-Mahammed M. (2022) Volatiles variation of two major cultivars of Olea europaea L. cultivated in Mediterranean and arid regions of Algeria. *Rec Nat Prod*., vol. 16, no. 1, pp. 34–45. http://doi.org/10.25135/rnp.249.21.02.1989.

24. ChemAxon Ltd. (2024) ChemAxon software suite. Available at: https://chemaxon.com

25. Stewart J.J.P. (2016) MOPAC2016, Stewart Computational Chemistry, Colorado Springs, CO, USA.

26. Alvascience. (2024) alvaDesc (software for molecular descriptors calculation), version 3.0.4. Available at: https://www.alvascience.com

27. Martin T., Harten P., Young D., Muratov E., Golbraikh A., Zhu H., Tropsha A. (2012) Does rational selection of training and test sets improve the outcome of QSAR modeling? *J Chem Inf Model*., vol. 52, no. 10, pp. 2570–2578. https://doi.org/10.1021/ci300338w.

28. Kennard R., Stone L.A. (1969) Computer aided design of experiments. *Technometrics*., vol. 11, pp. 137–148. https://doi.org/10.1080/00401706.1969.10490666.

29. Gramatica P., Chirico N., Papa E., Cassani S., Kovarich S. (2013) QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *J Comput Chem*., vol. 34, pp. 2121–2132. https://doi.org/10.1002/jcc.23361.

30. Kherouf S., Bouarra N., Messadi D. (2019). Quantitative modeling for prediction of boiling points of phenolic compounds. *Int J Chem Technol*., vol. 3, pp. 121–128. http://dx.doi.org/10.32571/ijct.636581.

31. Bouarra N., Nadji N., Nouri L., Boudjemaa A., Bachari K., Messadi D. (2021). Predicting retention indices of PAHs in reversed-phase liquid chromatography: A quantitative structure retention relationship approach. *J Serb Chem Soc*., vol. 86, pp. 63–75. https://doi.org/10.2298/JSC200219019B.

32. Bouarra N., Nadji N., Kherouf S., Nouri L., Boudjemaa A., Bachari K., Messadi D. (2022). QSER modeling of half-wave oxidation potential of indolizines by theoretical descriptors. *J Turk Chem Soc A: Chem.*, vol. 9, pp. 709–720. https://doi.org/10.18596/jotcsa.1065043.

33. Golbraikh A., Tropsha A. (2002) Beware of q²! . *J Mol Graph Model*., vol. 20, no. 4, pp. 269–276. https://doi.org/10.1016/s1093-3263(01)00123-1.

34. OECD. (2007) Principles for the validation, for regulatory purposes, of (quantitative) structure–activity relationship models. Paper presented at 37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology, Paris, France. ENV/JM/MONO(2007)2.

35. Poole C.F. (2012) Packed columns for gas–liquid and gas–solid chromatography. In: *Gas Chromatography*, pp. 97–121.ISBN 978-0-12-385540-4.

36. Bertz S.H. (1981) The first general index of molecular complexity. *J Am Chem Soc*., vol. 103, no. 12, pp. 3599–3601. https://doi.org/10.1021/ja00402a071.

37. Schuur J.H., Selzer P., Gasteiger J. (1996) The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure–spectra correlations and studies of biological activity. *J Chem Inf Comput Sci*., vol. 36, no. 2, pp. 334–344. https://doi.org/10.1021/ci950164c.

38. Kier L.B., Hall L.H. (1986) Molecular Connectivity in Structure–Activity Analysis. Research Studies Press: Letchworth, Hertfordshire, England. ISBN 9780471909835

39. Randić M. (1975) Characterization of molecular branching. *J Am Chem Soc*., vol. 97, no. 23, pp. 6609–6615. https://doi.org/10.1021/ja00856a001.

40. Todeschini R., Consonni V. (2009) Molecular Descriptors for Chemoinformatics. Wiley–VCH: Weinheim. ISBN:9783527318520. https://doi.org/10.1002/9783527628766.

Information about authors:

Soumaya Kherouf – Senior Researcher, Organic Synthesis Laboratory Modeling and Optimization of Chemical Processes (LOMOP), Department of Chemistry, Badji Mokhtar University (Annaba, Algeria, e-mail: soumaya.kherouf@univ-annaba.org).

Youssouf Driouche – Senior Researcher, Environmental Research Center (CRE) (Annaba, Algeria, e-mail: y.driouche@cre.dz).

Nabil Bouarra – Doctor of Science, Senior Researcher, Center of Scientific and Technical Research in Physico-Chemical Analysis, Bou-Ismail, (Tipaza, Algeria), Laboratory of Environmental Engineering, Faculty of Engineering Sciences, Department of Process Engineering, Badji Mokhtar University (Annaba, Algeria, e-mail: bouarranabil@yahoo.com, bouarra.nabil@crapc.dz).