Egypt. J. Chem. **Vol. 65,** No. 7, pp. 1 - 9 (2022)

Egyptian Journal of Chemistry

http://ejchem.journals.ekb.eg/

A QSRR model for predicting gas chromatography retention indices of essential oils using an improved chemical reaction optimization algorithm

Maimoonah Khalid Qasim

Department of General Science, University of Mosul, Mosul, Iraq

E-mail: maimoonah.qasim@uomosul.edu.iq

Abstract

A quantitative structure–retention relationship (QSRR) model with chemical reaction optimization algorithm (CROA) for predicting retention indices (RI) of 169 constituents of essential oils is proposed. In this, the CROA is utilized to select the most informative descriptors with high prediction. The proposed model was internal and external validated based on Q^2_{int}, Q^2_{LGO}, Q^2_{Boot}, $\text{MSE}_{\text{train}}$, Y-randomization test, Q^2_{ext}, MSE_{test}, and the applicability domain (AD). The validation results indicate that the model is robust and not due to chance correlation. In addition, the results indicate that the descriptors selection and prediction performance of the proposed model for training dataset outperforms the other two used modeling methods. The proposed model shows the highest Q^2_{int}, Q^2_{LGO}, and Q^2_{Boot}, and the lowest $\text{MSE}_{\text{train}}$. For the test dataset, proposed model shows higher external validation value (Q^2_{ext} = 0.936), and lower value of MSE_{test} compared with the other methods, indicating its higher predictive ability. In conclusion, the results reveal that the proposed model is an efficient approach for modeling high dimensional QSRRs and useful for the estimation of RI of essential oils that have not been experimentally tested.

Keywords: Chemical reaction optimization algorithm; lasso; descriptor selection; QSRR; Essential oils

1. Introduction

Essential oils are natural compounds extracted from plant secondary metabolism ADDIN EN.CITE [. Essential oils have been widely used in food, medicine, cosmetic, and fragrance industry [1-4]. It was proven that they have a wide range of biological properties such as antimicrobial and antioxidant activity [5-8]. The characteristics and antimicrobial activity of the essential oils depend on their components [9]. However, the ingestion of large quantities or wrong use of essential oils may cause toxic effects to humans [1]. Isolation of essential oils is carried out using several extraction methods such as organic solvent extraction, microwave assisted distillation, high pressure solvent extraction, supercritical CO_2 extraction, ultrasonic extraction, and solvent free microwave extraction [2]. Various procedures using gas chromatography (GC) and gas chromatography with mass spectrum (GC-MS) can be used for qualitative identification and quantitative determination of the essential oils components [1,2]. However, sometimes such as in the case of isomers the identification of the essential oils using GC-MS may not be accurate [1,3].

Due to the time consuming of using experimental methods, quantitative structure–retention relationship (QSRR) procedures are used to predict the GC retention indices. QSRR is a theoretical approach which is used in computational chemistry. The principle of QSRR is to correlate molecular descriptors derived from chemical structures of the studied components quantitatively with their experimental retention indices [1,10]. QSRR procedures are developed and validated to select the most informative descriptors that can define the retention index of the desired essential oil components. QSRR procedures can provide theoretical information about the compounds interactions with the mobile and

*Corresponding author e-mail: *maimoonah.qasim@uomosul.edu.iq*.
Receive Date: 08 June 2020, Revise Date: 22 July 2020, Accept Date: 21 December 2021
DOI: 10.21608/EJCHEM.2021.32116.2682

stationary phases, influence of the molecular structure on retention, and may provide explanation of possible absorption and elution mechanisms [3].

Many studies have been carried out to develop QSRR models for predicting the retention indices of essential oils. Fragkaki, et al. [11] carried out a QSRR study by the correlation of the GC-MS relative retention times of α-, β_1-, and β_2-agonists with their molecular characteristics using multiple linear regression and partial least squares as regression methods. QSRR study was conducted for gas chromatographic retention indices of 90 saturated esters using MLR regression [12]. A QSRR model was proposed to estimate the retention of 83 various drugs using ant colony optimization for variable selection. Multiple linear regression and support vector machines were used as regression methods [13]. Artificial neural networks, principal component analysis and cluster analysis were complementarily applied to construct QSRR models based on retention factors of the 59 esters alkoxyphenylcarbamic acid [14]. QSRR model was developed based on MLR method to predict the retention indices in gas chromatography of 169 essential oils components. The variable selection was done using ordered predictors selection (OPS) algorithm [1]. QSRR models were developed based on the chromatographic retention of amino acid analogues using PLS and MLR regression methods [15]. QSRR relationships were used for predicting retention times of 89 sulfur-containing compounds in two dimensional gas chromatography based on MLR regression method using CODESSA software [16]. Filipic, et al. [17] constructed QSRR models based on retention behavior of 22 imidazoline drugs using PLS and MLR regression.

In chemometrics area, the dimensionality of data becomes larger, where the dimension of data may grow exponentially with the sample size [18]. Such high-dimensional data present simultaneous challenges of statistical accuracy and computational feasibility [19]. Recent studies are also conducted [20-23].

In this paper, the chemical reaction optimization algorithm is utilized to perform descriptor selection and to enhance the constructed QSRP model. The performance is compared with other penalized methods.

2. MATERIALS AND METHODS

2.1. Data set

The gas chromatography retention indices (RI) of 169 constituents of essential oils were obtained from Conforti, et al. [24]. The chemical composition of the studied essential oils were experimentally isolated and characterized by GC and GC-MS [24]. A QSRR study have been conducted for these 169 compounds by Qin, et al. [1]. In the present study, we followed the same procedure that was used by Qin, et al. [1] to split the dataset into training and test datasets. The data were divided into 85 compounds as a training dataset and 84 compounds as a test dataset. The training dataset was used for constructing the QSAR model, and the test dataset was used for the evaluation of the QSAR model performance based on several evaluation criteria.

2.2. Molecular descriptor calculation

The molecular structures of the compounds were sketched using CHEM3D software (CambridgeSoft Corporation, Cambridge, MA,USA) [25]. The structures were optimized using the molecular mechanics (MM2) method implemented in Chem3D software, and then using molecular orbital package (MOPAC) module implemented in the same Chem3D software. DRAGON software (version 6.0) was used to generate 4,885 molecular descriptors based on the optimized structures [26-28]. To include consistent and useful descriptors, preprocessing steps were performed as follows. First, descriptors that had constant or zero values for all compounds, 301 descriptors, were excluded. Second, the remaining descriptors were refined further by removing those in which 70% of their values were zeros (237 descriptors). After that, descriptors with a relative standard deviation of less than 0.001, 174 descriptors, were removed. In addition, the correlation of the remaining descriptors was examined to omit multicollinearity by removing those that were highly correlated ($r_{ij} \geq 0.90$) (108 descriptors). Finally, 4071 descriptors remained for constructing the QSRR model.

2.3. Chemical reaction optimization algorithm

We often encounter optimization problems in scientific and technological research and development. Over the past decades, a number of evolutionary algorithms have been suggested [29,30].

The chemical reaction optimization algorithm (CROA) is evolutionary optimization techniques developed by Lam and Li [31]. CROA is an optimization technique inspired by chemical reaction process. It mimics the interactions of molecules in chemical reaction to reach a low energy stable state. In CROA, a candidate solution for a specific problem is encoded as a molecule. Each molecule represents a point in the search space, and hence a possible solution to the problem. A population consists of a finite number of molecules, each molecule is decided by an evaluating mechanism to obtain its potential energy (PE). Based on this potential energy and undergoing CROA operators, a new molecule(s) is generated.

In a chemical reaction process, a sequence of collisions among molecules occurs. Molecules collide either with each other or with the walls of the container. Collisions under different conditions provoke distinct elementary reactions, each of which

may have a different way of manipulating the energies of the involved molecule(s). The elements of the CROA are as follows:

2.3.1 The manipulated agent

CROA is a multi-agent algorithm and the manipulated agents are molecules. Each molecule has several attributes, some of which are essential to the basic operations of CROA .The essential attributes include (a) the molecular structure (w); (b) the potential energy (PE); and (c) the kinetic energy (KE).The rest depends on the algorithm operators and they are utilized to construct different CROA variants for particular problems provided that their implementations satisfy the characteristics of the elementary reactions. The optional attributes adopted in most of the published CROA variants are (d) the number of hits ($NumHit$); (e) the minimum structure ($MinStruct$): the minimum PE ($MinPE$); and (g) the minimum hit number ($MinHir$). Ilustrations of the attributes mentioned above are listed in the following:

- Molecular structure w captures a solution of the problem. It is not required to be in any specific format: it can be a number, a vector, or even a matrix. For example, if the problem solution space is defined as a set of vectors composed of five real numbers, then o can be any of these vectors
- Potential energy PE is defined as the objective function value of the corresponding solution represented by w. If f denotes the objective function, then we have

$$PE_w = f(\mathrm{w}) \quad \ldots(1)$$

- Kinetic energy KE is a non-negative number and it quantifies the tolerance of the system accepting a worse solution than the existing one. We will elaborate on the concept later in this section.
- Number of hits When a molecule undergoes a collision one of the elementary reactions will be triggered and it may experience a change in its molecular structure. NumHit is a record of the total number of hits (i.e. collisions) a molecule has taken.
- Minimum structure MinStruct is the w with the minimum corresponding PE which a molecule has attained so far. After a molecule experiences a certain number of collisions, it has undergone many transformations of its structure, with different corresponding PE. MinStruct is the one with the lowest PE in its own reaction history.
- Minimum potential energy When a molecule attains its $MinStruct$, $MinPE$ is the corresponding PE.
- Minimum hit number $MinHit$ is the number of hits when a molecule realizes $MinStruct$. It is an abstract notation of time when MinStruct is achieved.

2.3.2. Elementary reactions

There are four types of elementary reactions, each of which takes place in each iteration of CROA. They are employed to manipulate solutions (i.e. explore the solution space) and to redistribute energy among the molecules and the buffer. Note that there is no strict requirements on the mechanisms of the operators and operators designed for other algorithms may also be adopted. However, CROA ensures the conservation of energy when new solutions are generated with the operators.

2.3.2.1. On-wall ineffective collision

An on-wall ineffective collision represents the situation when a molecule collides with a wall of the container and then bounces away remaining in one single unit. In this collision, we only perturb the existing w to w' , 1.e.

This can be done by picking o in the neighborhood of w'o. Let $N(.)$ be any neighborhood search operator, we have $w' = N(\mathrm{w})$ and $\mathrm{PE}_{w'} = f(w')$ Moreover a certain of KE of transformed molecule is withdrawn to the central energy buffer (buffer). Let $KElossRate$ be a parameter of CRO, $0 \leq KElosskate \leq 1$, and $a \neq [KElossRate, 1]$ be a random number, uniformly distributed from $KElossRate$ to 1. We get

$$KE_{w'} = (\mathrm{PE}_w - \mathrm{PE}_{w'} + \mathrm{KE}_{w'}) \times \mathrm{a} \quad \ldots(2)$$

and the remaining energy,

$$(\mathrm{PE}_w - \mathrm{PE}_{w'} + \mathrm{KE}_{w'}) \times (1 - \mathrm{a}),$$

$$\mathrm{PE}_{w'} + \mathrm{KE}_{w'} \geq \mathrm{PE}_{w'} \quad \ldots(3)$$

It is always possible to undergo an on-wall ineffective collision when $\mathrm{PE}_{w'} \leq \mathrm{PE}_w$ When a molecule experiences more of this elementary reaction, it will have more KE transferred to butter. Hence ,the chance of having a worse solution is lower in a subsequent change.

2.3.2.2. Decomposition

Decomposition refers to the situation when a molecule its a wall and then breaks into several parts (for simplicity, we consider two parts in our discussion). Assume that w produces w_1' and w_2'. i.e.,

$$w \rightarrow w_1' + w_2'$$

Any mechanism, which can produce w_1' and w_2' from w is allowed. Theoretically , even generating solutions independent of the existing one (random generation of new solution) is feasible. The idea of decomposition is to allow the system to explore other

regions of the solution space after enough local search by the ineffective collisions. The effectiveness of the solution generation mechanism is problem-dependent .Since more solutions are created, the total sum of PE and KE of the original molecule may not be sufficient. In other words ,we may have

$$\mathrm{PE}_w + \mathrm{KE}_w < \mathrm{PE}_{w_1'} + \mathrm{PE}_{w_2'}.$$

As energy conservation is not satisfied, this decomposition has to be aborted. To increase the chance of having a decomposition completed, we randomly draw a small portion of energy from buffer to support the change. We modify the energy conservation condition for decomposition as follows:

$$\mathrm{PE}_w + KE_w + \delta_1 \times \delta_2 \times buffer \geq \mathrm{PE}_{w_1'} + \mathrm{PE}_{w_2'} \quad (4)$$

This models the situation that some energy from buffer is transferred to the molecule when it hits the wall. If (4) holds, the existing molecule with o is replaced by the two newly generated ones, whose KEs randomly share the remaining energy $\mathrm{E}_{dec} = (\mathrm{PE}_w + KE_w + \delta_1 \times \delta_2 \times buffer) - (\mathrm{PE}_{w_1'} + \mathrm{PE}_{w_2'})$,i.e.,

$$KE_{w_1'} = \mathrm{E}_{dec} \times \delta_3 \quad \ldots(5)$$

$$KE_{w_2'} = \mathrm{E}_{dec} \times (1 - \delta_3) \quad \ldots(6)$$

KE where s3 is a random number generated in [0, 1]. The energy in the buffer is also updated by :

$$buffer' = (1 - \delta_1\delta_2)buffer \quad \ldots(7)$$

2.3.2.3. Inter-molecular ineffective collision

Inter-molecular ineffective collision takes place when multiple molecules collide with each other and then bounce away. The molecularity (assume two) remains unchanged before and after the process, i.e

$$w_1 + w_2 \rightarrow w_1' + w_2'$$

This elementary reaction is very similar to the unmolecu- lar ineffective counterpart; we generate w and w' by $w_1' = N(w_1)$ *and* $w_2' = N(w_2)$. The energy management is similar but no $buffer$ is involved. The energy conservation condition can be stated as

$$\mathrm{PE}_{w_1} + \mathrm{PE}_{w_2} + \mathrm{KE}_{w_1} + \mathrm{KE}_{w_2} \geq \mathrm{PE}_{w_1'} + \mathrm{PE}_{w_2'} \quad (8)$$

As more molecules are involved, the total sum of energy of the molecular sub-system is larger than that of the on- wall ineffective collision. The probability of the molecules to explore their immediate surroundings is higher. In other words, the molecules have higher flexibility to be trans formed to more diverse molecular structures. We can use the same operator for on-wall ineffective collision to produce new solutions. We apply the operator to each molecule to get a new one. If (8) is satisfied, KEs of the transformed molecules share the remaining energy $Einter =$ in the sub-system, i.e.,

$$(\mathrm{PE}_{w_1} + \mathrm{PE}_{w_2} + \mathrm{KE}_{w_1} + \mathrm{KE}_{w_2}) - (\mathrm{PE}_{w_1'} + \mathrm{PE}_{w_2'})$$

$$K\,\mathrm{E}_{w_1'} = E_{\mathrm{inter}} \times \delta_4 \quad \ldots(9)$$

$$K\,\mathrm{E}_{w_2'} = E_{\mathrm{inter}} \times (1 - \delta_4) \quad \ldots(10)$$

where is a random number generated in [0, 1].

2.3.2.4. Synthesis

Synthesis does the opposite of decomposition. A synthesis happens when multiple (assume two) molecules hit against each other and fuse together, i.e.,

$$w_1 + w_2 \rightarrow w'$$

As only one molecule is produced, it is likely to satisfy the energy conservation condition:

$$\mathrm{PE}_{w_1} + \mathrm{PE}_{w_2} + \mathrm{KE}_{w_1} + \mathrm{KE}_{w_2} \geq \mathrm{PE}_{w'} \quad (11)$$

If (11) holds, the resulting KE ing energy, i.e., just takes up all the remain-

$$\mathrm{KE}_{w'} = (\mathrm{PE}_{w_1} + \mathrm{PE}_{w_2} + \mathrm{KE}_{w_1} + \mathrm{KE}_{w_2}) - (\mathrm{PE}_{w'}) \quad (12)$$

We can see that we allow greater change to w' with respect to w_1 and w_2 and $KE_{w'}$ is usually higher thanKE_w. The resulting molecule has a higher "ability" to explore a new solution region. Any mechanism allowing the combination of solutions is allowed, where the resultant molecule is in a region farther away from the existing ones in the solution space .The idea behind synthesis is diversification of solutions. The implementation detail is again problem-dependent.

2.3. Conservation of energy

One of the fundamental assumptions of CROA is conservation of energy, which means that energy cannot be created or destroyed. The whole system refers to all the defined molecules and the container, which is connected to buffer. The total amount of energy of the whole system is determined by the objective function values (i.e. PE) of the initial population of molecules whose size is $PopSize$, the initial KE ($InitialKE$) assigned, and the initial value of $buffer$. Let $\mathrm{PE}w_i(t)$, $K\,Ew_i(\mathrm{t})$, $PopSize(t)$, *and* $buffer\ (t)$ be the PE of molecule i, the KE of molecule i, the number of molecules, and the energy in the central buffer at time t. When the algorithm evolves, the total amount of energy in the system always remains constant, i.e.,

$$\sum_{i=1}^{Popsize} (\mathrm{PE}_{w_i}(\mathrm{t}) + \mathrm{KE}_{w_i}(\mathrm{t}) + buffer(\mathrm{t}) = C \quad (13)$$

where C is a constant. Each elementary reaction manages a sub-system (i.e. a subset of entities of the system); a uni-molecular collision involves a molecule and the container while an inter-molecular collision concerns multiple molecules. After an elementary

reaction, the total energy of the constructed sub-system remains the same. Let k and I be the number of molecules involved before and after a particular elementary reaction, and let aw and w be the molecular structures of an existing molecule and the one to be generated from the elementary reaction, respectively. In general, the elementary reaction can only take place when it satisfies the following energy conservation condition:

$$\sum_{i=1}^{k}(PE_{w_i} + KE_{w_i}) \geq \sum_{i=1}^{l} PE_{w_i'} \quad \ldots(14)$$

We modify this condition for decomposition as it involves $buffer$ on the left-hand side of (14). Note that PE is determined by Eq.(1) according to the molecular structure. If the resultant molecules have very high potential energy, i.e. they give very bad solutions, the reaction will not occur.

Theoretically, energy cannot attain a negative value and any operation resulting in negative energy should be forbidden. However, some problems may attain negative objective function values (i.e. negative PE), but we can convert the problem to an equivalent one by adding an offset to the objective function to make each PE non-negative. The law of conservation of energy is still obeyed and the system works perfectly.

4. Prediction evaluation criteria

To provide a satisfactory evaluation of the compared modeling methods in constructing an efficient QSRR model, the following criteria were performed. The used criteria for the training dataset were mean-squared error of the training dataset (MSE_{train}) and leave-one-out internal validation (Q^2_{int}), which are defined by

$$MSE_{train} = \frac{\sum_{i=1}^{n_{train}}(y_{i,train} - \hat{y}_{i,train})^2}{n_{train}} \quad (15)$$

and

$$Q^2_{\text{int}} = 1 - \left[\frac{\sum_{i=1}^{n_{train}}(y_{i,train} - \hat{y}_{i,train})^2}{\sum_{i=1}^{n_{train}}(y_{i,train} - \bar{y})^2}\right] \quad (16)$$

respectively.

Furthermore, the test dataset was used to validate the model by computing the following criteria, i.e., mean-squared error of the test dataset (MSE_{test}) and external validation (Q^2_{ext}). These criteria are defined by

$$MSE_{test} = \frac{\sum_{i=1}^{n_{test}}(y_{i,test} - \hat{y}_{i,test})^2}{n_{test}} \quad (17)$$

and

$$Q^2_{ext} = 1 - \left[\frac{\sum_{i=1}^{n_{test}}(y_{i,test} - \hat{y}_{i,test})^2}{\sum_{i=1}^{n_{test}}(y_{i,test} - \bar{y}_{train})^2}\right] \quad (18)$$

respectively, where n_{train} and n_{test} represent the training and test sample sizes, the $y_{i,train}$, $y_{i,test}$, $\hat{y}_{i,train}$, and $\hat{y}_{i,\text{test}}$ stand for the RI values of the training dataset, test dataset, and their corresponding predicted RI values. While $\bar{y}$ and $\bar{y}_{train}$ represent the mean of all the RI values and the mean of the training RI values, respectively.

5. RESULTS AND DISCUSSION

To demonstrate the usefulness of the proposed method, CROA, comprehensive comparative experiments with the lasso and the non-adaptive bridge penalty method was fixed as $0 \leq \lambda \leq 100$.

From Table 1, it is obvious that the performance of the CROA is much better than the lasso, in terms of selected descriptors. On the other hand, the result of the proposed method, CROA, is the best one among them. It selected 4 descriptors out of 4071 descriptors comparing with 6 and 9 selected descriptors of Bridge and lasso, respectively. The names of the selected descriptors and their descriptions for each used method are presented in Table 1. The QSRP model by CROA is

$$\hat{y}_{IR} = 5.214 + 0.608Eig09_{AEA(bo)} + 1.674MW - 8.551Mor07m + 3.09SpMaxA\_EA$$

The results of the prediction evaluation of the constructed QSRR models using CROA, Bridge, and lasso are listed in Table 2. We can see that the CROA was superior to Bridge and lasso in terms of prediction performance for the training data. CROA yields the highest Q^2_{int} , Q^2_{LGO} , and Q^2_{Boot} , and the lowest $\text{MSE}_{\text{train}}$.

Furthermore, depending on testing data, it is noted that CROA reveals greater value of Q^2_{ext} and less value of MSE_{test} compared to the other two used methods. This enhancement of CROA, in terms of MSE_{test} , over the Bridge and lasso is 30.92% and 61.95%, respectively. Additionally, the predictive ability in the testing data using of the CROA was

0.944, which was much better than the 0.901 and 0.811 obtained, respectively, by the Bridge and lasso.

Overall speaking, the results demonstrated that the CROA is effective in modeling high-dimensional QSRR. The CROA not only improved the prediction performance but also identified a small subset of descriptors compared to the Bridge and the lasso.

Table 1: The selected descriptor names and their descriptions by the three used methods

Method	Descriptor name	Group type	Description
CROA	Eig09_AEA(bo)	Edge adjacency indices	eigenvalue n. 9 from augmented edge adjacency mat. weighted by bond order
	MW	Constitutional indices	molecular weight
	Mor07m	3D-MoRSE descriptors	signal 07 / weighted by mass
	SpMaxA_EA	Edge adjacency indices	normalized leading eigenvalue from edge adjacency mat.
Bridge	Qindex	Topological indices	quadratic index
	SpMaxA_EA	Edge adjacency indices	normalized leading eigenvalue from edge adjacency mat.
	MW	Constitutional indices	molecular weight
	GATS5m	2D autocorrelations	Geary autocorrelation of lag 5 weighted by mass
	SdO	Atom-type E-state indices	Sum of dO E-states
	SM14_EA(ri)	Edge adjacency indices	spectral moment of order 14 from edge adjacency mat. weighted by resonance integral
lasso	P_VSA_MR_2	P_VSA-like descriptors	P_VSA-like on Molar Refractivity, bin 2
	GATS5m	2D autocorrelations	Geary autocorrelation of lag 5 weighted by mass
	SdO	Atom-type E-state indices	Sum of dO E-states
	SM14_EA(ri)	Edge adjacency indices	spectral moment of order 14 from edge adjacency mat. weighted by resonance integral
	Eig03_AEA(dm)	Edge adjacency indices	eigenvalue n. 3 from augmented edge adjacency mat. weighted by dipole moment
	Mor07m	3D-MoRSE descriptors	signal 07 / weighted by mass
	Mor24p	3D-MoRSE descriptors	signal 24 / weighted by polarizability
	Eig09_AEA(bo)	Edge adjacency indices	eigenvalue n. 9 from augmented edge adjacency mat. weighted by bond order
	MW	Constitutional indices	molecular weight

Table 2: Prediction evaluation criteria values for the training and testing data

Methods	No. of descriptors	Training set				Testing set	
		MSE_{train}	Q^2_{int}	Q^2_{LGO}	Q^2_{Boot}	MSE_{test}	Q^2_{ext}
CROA	4	0.101	0.961	0.957	0.955	0.277	0.944
Bridge	6	0.234	0.914	0.911	0.909	0.401	0.901
lasso	9	0.571	0.827	0.824	0.822	0.738	0.811

5.1 Y-randomization test

The CROA model was further validated by applying the Y-randomization test [32]. This was in order to ensure that the predictive power of the CROA model was not based on chance. This test randomly shuffled the retention indices values several times and applied CROA each time. In each time, the Q^2_{int} was calculated. If all the obtained values were less than the Q^2_{int} of the constructed QSRR by CROA, then the constructed QSRR was not due to chance correlation, indicating that the CROA method could lead to an acceptable method using the training data. Figure 1 shows the results for the Y-randomization test for 500 times of Q^2_{int} values. It can be clearly seen from Figure 1 that the Q^2_{int} values were in the range of 0.0530 to 0.3951. In comparison to true Q^2_{int} values of CROA ($Q^2_{int} = 0.961$), these values indicate that the QSRR model of retention indices of essential oils by CROA was not due to chance correlation or structural dependence of the training data.

5.2 Robustness performance

To further evaluate the ability of CROA to construct a robust QSRR model, the leverage approach was used as an applicability domain (AD) assessment.

AD is defined as "a theoretical region in chemical space, defined by the model descriptors and modeled response, and thus by the nature of the chemicals in the training set, as represented in each model by specific molecular descriptors" [18]. Figure 2 displays the Williams plot of the leverage values against the standardized residuals for each compound for the CROA model (the dotted line indicates the leverage threshold, while the dashed line represents the standardized residual limits). The influential compound can be detected when its leverage value is greater than the leverage threshold ($h^* = 3(p+1)/n$) where p is the number of the selected descriptors in the final QSAR model, and n represents the number of compounds.

It is obvious from Figure 2 that no compounds have a standardized residual higher than the limit ±3, which can be considered as retention indices outliers, or with a high leverage value. Thus, it is clearly demonstrated from Figure 2 that all the results confirm that the constructed QSRR model using the CROA is reliable and robust.

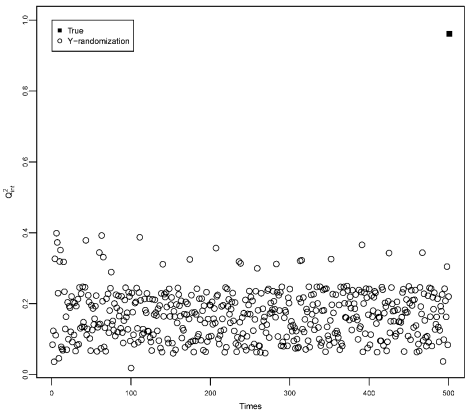

Figure 1. Y-randomization test for CROA over 500 times.

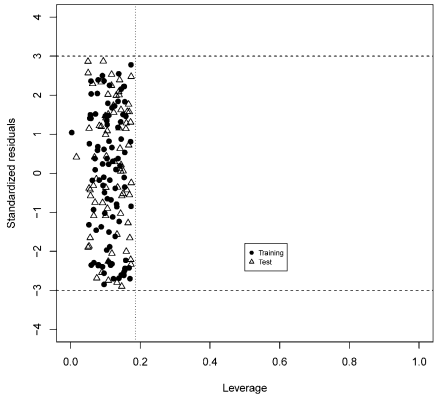

Figure 2. Williams plot for the training and testing data of CROA.

6. CONCLUSION

In the present work, a new QSRR model approach for the prediction of retention indices (RI) of essential oils constituents was developed by proposing CROA. The results gained by the internal validation criteria ($\mathrm{MSE}_{\mathrm{train}}$, Q^2_{int}, Q^2_{LGO}, and Q^2_{Boot}) for training dataset and the external validation parameters ($\mathrm{MSE}_{\mathrm{test}}$ and Q^2_{ext}) for the test dataset prove better predictive power of the CROA model compared with other two developed models. The reduction in $\mathrm{MSE}_{\mathrm{train}}$ was 82.31% and 56.83% of lasso and Bridge methods. Further, using Q^2_{int} criteria, the CROA has the highest value with 0.961. The lasso method is the worst method in constructing the QSRP model. In addition, the obtained results by the applicability domain and Y-randomization test confirm that the CROA model is reliable, robust and not due chance correlation. In conclusion, the current study proposes CROA as a useful modeling approach to be used for predicting RI of new essential oils constituents.

7. REFERENCES

1. Qin L.-T., Liu S.-S., Chen F., Xiao Q.-F. and Wu Q.-S., Chemometric model for predicting retention indices of constituents of essential oils. *Chemosphere*, **90** (2), 300-305 (2013).
2. Okoh O.O., Sadimenko A.P. and Afolayan A.J., Comparative evaluation of the antibacterial activities of the essential oils of Rosmarinus officinalis L. obtained by hydrodistillation and solvent free microwave extraction methods. *Food Chemistry*, **120** (1), 308-312 (2010).
3. Riahi S., Ganjali M.R., Pourbasheer E. and Norouzi P., QSRR Study of GC Retention Indices of Essential-Oil Compounds by Multiple Linear Regression with a Genetic Algorithm. *Chromatographia*, **67** (11), 917-922 (2008).
4. Ibrahim T.A., El-Hela A.A., El-Hefnawy H.M., Al-Taweel A.M. and Perveen S., Chemical Composition and Antimicrobial Activities of Essential Oils of Some Coniferous Plants Cultivated in Egypt. *Iran J Pharm Res*, **16** (1), 328-337 (2017).
5. Riahi S., Pourbasheer E., Ganjali M.R. and Norouzi P., Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine. *Journal of Hazardous Materials*, **166** (2), 853-859 (2009).
6. Resende D.B., Martins H.H.d.A., Souza T.B.d., Carvalho D.T., Piccoli R.H., Schwan R.F. *et al.*, Synthesis and in vitro evaluation of peracetyl and deacetyl glycosides of eugenol, isoeugenol and dihydroeugenol acting against food-contaminating bacteria. *Food Chemistry*, **237**, 1025-1029 (2017).
7. Tohidi B., Rahimmalek M. and Arzani A., Essential oil composition, total phenolic,

flavonoid contents, and antioxidant activity of Thymus species collected from different regions of Iran. *Food Chemistry*, **220**, 153-161 (2017).
8. Deng J., He B., He D. and Chen Z., A potential biopreservative: Chemical composition, antibacterial and hemolytic activities of leaves essential oil from Alpinia guinanensis. *Industrial Crops and Products*, **94**, 281-287 (2016).
9. Fisher K. and Phillips C., Potential antimicrobial uses of essential oils in food: is citrus the answer? *Trends in Food Science & Technology*, **19** (3), 156-164 (2008).
10. Al-Fakih A.M., Algamal Z.Y., Lee M.H., Abdallah H.H., Maarof H. and Aziz M., Quantitative structure–activity relationship model for prediction study of corrosion inhibition efficiency using two-stage sparse multiple linear regression. *Journal of Chemometrics*, **30** (7), 361-368 (2016).
11. Fragkaki A.G., Koupparis M.A. and Georgakopoulos C.G., Quantitative structure–retention relationship study of α-, β1-, and β2-agonists using multiple linear regression and partial least-squares procedures. *Analytica Chimica Acta*, **512** (1), 165-171 (2004).
12. Lu C., Guo W. and Yin C., Quantitative structure-retention relationship study of the gas chromatographic retention indices of saturated esters on different stationary phases using novel topological indices. *Analytica Chimica Acta*, **561** (1), 96-102 (2006).
13. Goodarzi M., Jensen R. and Vander Heyden Y., QSRR modeling for diverse drugs using different feature selection methods coupled with linear and nonlinear regressions. *Journal of Chromatography B*, **910**, 84-94 (2012).
14. Durcekova T., Boronova K., Mocak J., Lehotay J. and Cizmarik J., QSRR models for potential local anaesthetic drugs using high performance liquid chromatography. *Journal of Pharmaceutical and Biomedical Analysis*, **59**, 209-216 (2012).
15. Kritikos N., Tsantili-Kakoulidou A., Loukas Y.L. and Dotsikas Y., Liquid chromatography coupled to quadrupole-time of flight tandem mass spectrometry based quantitative structure–retention relationships of amino acid analogues derivatized via n-propyl chloroformate mediated reaction. *Journal of Chromatography A*, **1403**, 70-80 (2015).
16. Gieleciak R., Hager D. and Heshka N.E., Application of a quantitative structure retention relationship approach for the prediction of the two-dimensional gas chromatography retention times of polycyclic aromatic sulfur heterocycle compounds. *Journal of Chromatography A*, **1437**, 191-202 (2016).
17. Filipic S., Ruzic D., Vucicevic J., Nikolic K. and Agbaba D., Quantitative structure-retention relationship of selected imidazoline derivatives on α1-acid glycoprotein column. *Journal of Pharmaceutical and Biomedical Analysis*, **127**, 101-111 (2016).
18. Algamal Z.Y., Lee M.H. and Al-Fakih A.M., High-dimensional quantitative structure–activity relationship modeling of influenza neuraminidase a/PR/8/34 (H1N1) inhibitors based on a two-stage adaptive penalized rank regression. *Journal of Chemometrics*, **30** (2), 50-57 (2016).
19. Liu J., Zhong W. and Li R., A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, **58** (10), 1-22 (2015).
20. El-Kashef E., El-Shamy A.M., Abdo A., Gad E.A. and Gado A.A., Effect of Magnetic Treatment of Potable Water in Looped and Dead End Water Networks. *Egyptian Journal of Chemistry*, **62** (8), 1467-1481 (2019).
21. Abbas M.A., Zakaria K., El-Shamy A.M. and El Abedin S.Z., Utilization of 1-butylpyrrolidinium Chloride Ionic Liquid as an Eco-friendly Corrosion Inhibitor and Biocide for Oilfield Equipment: Combined Weight Loss, Electrochemical and SEM Studies. *Zeitschrift für Physikalische Chemie*, **1** (ahead-of-print) (2019).
22. El-Shamy A., El-Hadek M., Nassef A. and El-Bindary R., Optimization of the Influencing Variables on the Corrosion Property of Steel Alloy 4130 in 3.5 wt.% NaCl Solution. *Journal of Chemistry*, **2020** (2020).
23. Zohdy K., El-Shamy A., Kalmouch A. and Gad E.A., The corrosion inhibition of (2Z, 2′ Z)-4, 4′-(1, 2-phenylene bis (azanediyl)) bis (4-oxobut-2-enoic acid) for carbon steel in acidic media using DFT. *Egyptian Journal of Petroleum*, **28** (4), 355-359 (2019).
24. Conforti F., Menichini F., Formisano C., Rigano D., Senatore F., Arnold N.A. *et al.*, Comparative chemical composition, free radical-scavenging and cytotoxic properties of essential oils of six Stachys species from different regions of the Mediterranean Area. *Food Chemistry*, **116** (4), 898-905 (2009).
25. https://www.perkinelmer.com/category/chemdraw
26. Todeschini R., Consonni V., Mauri A. and Pavan M. DRAGON, Software version 6.0, Talete srl. (2010). http://www.talete.mi.it/.
27. Qasim M.K., Algamal Z.Y. and Ali H.T.M., A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine. *SAR QSAR Environ Res*, **29** (7), 517-527 (2018).

28. Algamal Z.Y., Qasim M.K. and Ali H.T.M., A QSAR classification model for neuraminidase inhibitors of influenza A viruses (H1N1) based on weighted penalized support vector machine. *SAR QSAR Environ Res*, **28** (5), 415-426 (2017).
29. Siddique N. and Adeli H., Nature-Inspired Chemical Reaction Optimisation Algorithms. *Cognit Comput*, **9** (4), 411-422 (2017).
30. Yan C., Gao S., Luo H. and Hu Z., A Hybrid Algorithm Based on Tabu Search and Chemical Reaction Optimization for 0-1 Knapsack Problem. **9141**, 229-237 (2015).
31. Lam A.Y.S. and Li V.O.K., Chemical Reaction Optimization: a tutorial. *Memetic Computing*, **4** (1), 3-17 (2012).
32. Rücker C., Rücker G. and Meringer M., y-Randomization and its variants in QSPR/QSAR. *Journal of chemical information and modeling*, **47** (6), 2345-2357 (2007).