Orbital: The Electronic Journal of Chemistry
journal homepage: www.orbital.ufms.br
ISSN 1984-6428

| Vol 9 | | No. 4 | | July-September 2017 |

Full Paper

QSPR Study of the Retention/Release Property of Odorant Molecules in Water Using Statistical Methods

Assia Belhassan[a,b], Samir Chtita[a], Tahar Lakhlifi[a], and Mohammed Bouachrine[b*]

[a]*MCNS Laboratory, Faculty of Science, University Moulay Ismail, Meknes, Morocco*
[b]*MEM, High school of technology, University Moulay Ismail, Meknes, Morocco*

Article history: Received: 19 March 2017; revised: 10 July 2017; accepted: 10 August 2017. Available online: 24 September 2017. DOI: http://dx.doi.org/10.17807/orbital.v9i4.978

Abstract: An integrated approach physicochemistry and structures property relationships has been carried out to study the odorant molecules retention/release phenomenon in the water. This study aimed to identify the molecular properties (molecular descriptors) that govern this phenomenon assuming that modifying the structure leads automatically to a change in the retention/release property of odorant molecules. ACD/ChemSketch, MarvinSketch, and ChemOffice programs were used to calculate several molecular descriptors of 51 odorant molecules (15 alcohols, 11 aldehydes, 9 ketones and 16 esters). A total of 37 molecules (2/3 of the data set) were placed in the training set to build the QSPR models, whereas the remaining, 14 molecules (1/3 of the data set) constitute the test set. The best descriptors were selected to establish the quantitative structure property relationship (QSPR) of the retention/release property of odorant molecules in water using multiple linear regression (MLR), multiple non-linear regression (MNLR) and an artificial neural network (ANN) methods. We propose a quantitative model according to these analyses. The models were used to predict the retention/release property of the test set compounds, and agreement between the experimental and predicted values was verified. The descriptors showed by QSPR study are used for study and designing of new compounds. The statistical results indicate that the predicted values are in good agreement with the experimental results. To validate the predictive power of the resulting models, external validation multiple correlation coefficient was calculated and has both in addition to a performant prediction power, a favorable estimation of stability.

Keywords: odorant molecules; retention/release; quantitative structure property relationship; multiple linear regression; artificial neural network

1. INTRODUCTION

The concept of quality of food commonly includes four criteria: Safety, Health, Flavor and Services. Each of these key words refers to the notions of product safety, their nutritional value and health, the organoleptic criteria of taste, odor and all the services associated with the food product. It can be said that if the notions of safety and health are present in the mind of the consumer during a purchase, the organoleptic dimension of a product remains essential [1]. The flavor compounds present in a product must be sensorially perceived to be released from the food phase. The release of odorant molecules from the solid or liquid food matrix and their passage through the vapor phase is therefore the first step before a possible perception due to the activation of the olfactory receptors present in the nasal cavity and to the activation of complex neurophysiological events [2].

Retention/release property of odorant molecules is a phenomenon primarily dependent on the interactions between the solute and the stationary phase of molecules, which included directional force, induction force, dispersion force and hydrogen bond [3]. These forces can be related to the topological structures; therefore, it was possible to predict the solute retention from molecular descriptors.

Molecular descriptors theoretically calculated can be used to construct mathematical models, being related to molecular properties. In this insight, the Quantitative Structure Property Relationships (QSPR) [4] refers to obtain a robust and predictive mathematical model involving response variable with molecular descriptors, calculated through molecular modeling methods.

*Corresponding author. E-mail: m.bouachrine@est-umi.ac.ma

The purpose of this work is to study the retention/release property of odorant molecules in the water by varying their chemical class and molecular structure (linear, branched and/or unsaturated) using QSPR chemical modeling methods

2. MATERIAL AND METHODS

This QSPR study was investigated for predicting, interpreting studied property and for designing new compounds by using linear and nonlinear methods. It consists of four stages: selection of data set and generation of molecular descriptors, descriptive analysis, statistical analysis and suggestion of novel compounds.

The methodology used in this QSPR study is as follows (Fig.1):

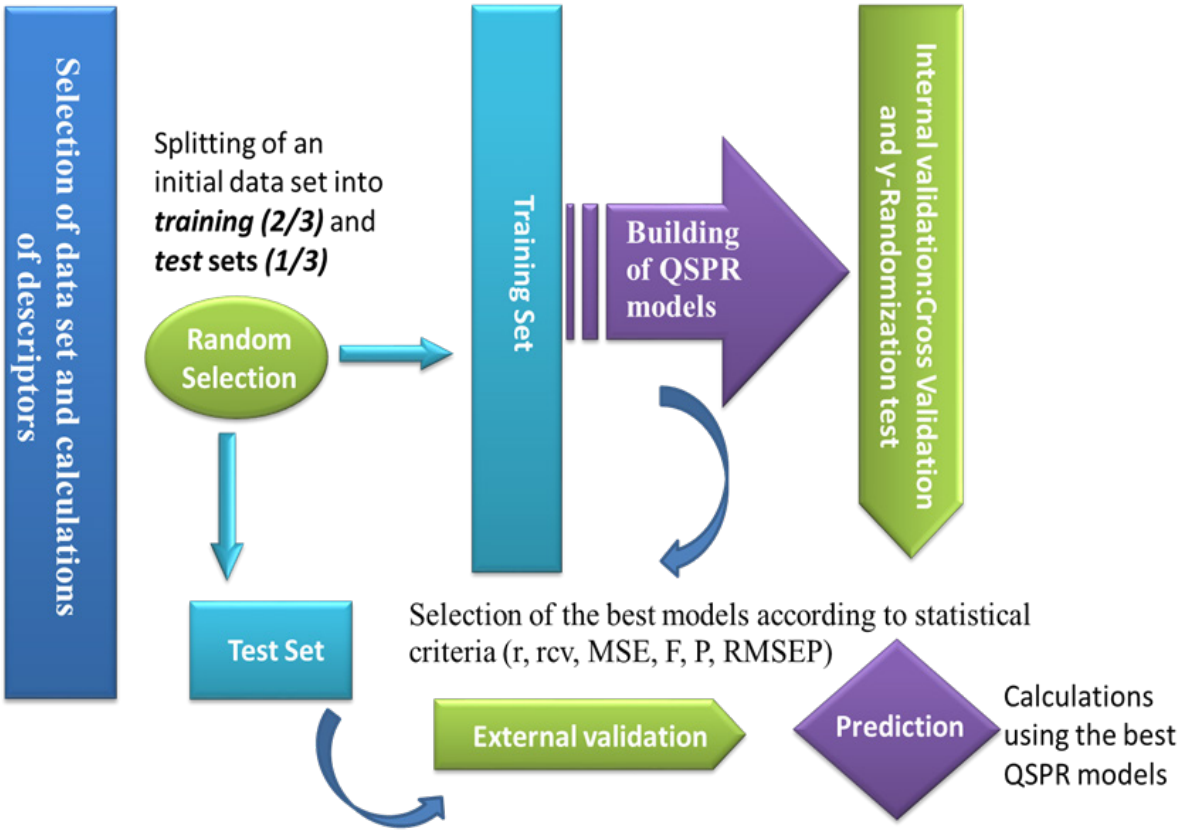


Figure 1. Flow chart of the methodology used in this work.

Experimental data set

In this study, we selected 51 odorant molecules with properties reported in the literature [5] to provide a diversified set of chemical families (alcohols, aldehydes, ketones and esters) and chemical structures (linear, unsaturated and unsaturated-branched). The fragrant molecules were selected by their structures without taking into account their organoleptic qualities. The list of molecules and the Log(1/K) values are displayed in Table 1.

The retention/release property of the selected odorant molecules was examined using pure water, this property was quantified by the vapor-liquid partition coefficient K, and more precisely by the Log(1/K) values [5].

A total of 37 molecules were placed in the training set to build the QSPR models, whereas the remaining, 14 molecules constitute the test set. The division was carried out by random selection using the SPSS 19.0 statistical package [6].

Molecular descriptor generation

A wide variety of molecular descriptors were calculated using ACD/ChemSketch, MarvinSketch and ChemOffice software [7-9] to predict the correlation between these descriptors and the retention/release property of the molecules studied (Table 2). The Table S1 and Table S2 show values of these descriptors for each molecule studied.

Table 1. List of aroma compounds.

No.	Molecule name	Log(1/K)	No.	Molecule name	Log(1/K)
1	Hexan-1-ol	3.153	**27**	Heptan-2-one	2.338
2	Octan-1-ol	2.629	**28**	3-Methylpentan-2-one	2.447
3	Nonan-1-ol	2.330	**29**	4-Methylpentan-2-one	2.361
4	2-Methylpentan-1-ol	3.097	**30**[a]	5-Methylhexan-2-one	2.338
5	3-Methylpentan-3-ol	2.523	**31**	2-Methylheptan-3-one	2.081
6[a]	Heptan-2-ol	2.857	**32**	5-Methylheptan-3-one	2.178
7	2-Ethylhexan-1-ol	2.721	**33**	2,6-Dimethylheptan-4-one	1.968
8	Octan-3-ol	2.582	**34**	Hex-5-en-2-one	2.586
9	3,7-Dimethyloctan-1-ol	2.086	**35**[a]	4-Methylpent-3-en-2-one	2.488
10	(4Z)-Hex-4-en-1-ol	3.480	**36**[a]	Ethyl propanoate	2.244
11	(4Z)-Hept-4-en-1-ol	3.283	**37**[a]	Ethyl butanoate	2.135
12	Hex-1-en-3-ol	3.090	**38**	Ethyl pentanoate	2.078
13	6-Methylhept-5-en-2-ol	3.004	**39**	Ethyl hexanoate	1.983
14	Linalool	2.690	**40**[a]	Butyl pentanoate	1.894
15[a]	Nerol	2.886	**41**	Ethyl heptanoate	1.898
16[a]	Hexanal	2.204	**42**[a]	Ethyl 2-methylpropanoate	2.019
17	Heptanal	2.116	**43**	Isoamylacetate	2.078
18	Octanal	2.030	**44**	Isobutylisovalerate	1.868
19[a]	2-Ethylbutanal	2.099	**45**	2-Methylbutyl 2-methylbutanoate	1.863
20	2-Ethylhexanal	1.935	**46**[a]	Ethylcrotonate	2.348
21	3,5,5-Trimethylhexanal	1.897	**47**	Isopropyltiglate	2.046
22[a]	(2E)-Hex-2-enal	2.695	**48**	Propyl (2E)-2-methylbut-2-enoate	2.160
23	(4Z)-Hept-4-enal	2.477	**49**[a]	Isobutylangelate	1.937
24	(2E)-2-Methylbut-2-enal	2.699	**50**	Isoamyltiglate	1.960
25	(2E)-2-Methylpent-2-enal	2.480	**51**	Hexyltiglate	1.909
26[a]	2-Isopropyl-5-methyl-2-hexenal	1.817			

[a]Test Set.

Table 2. Calculated topological molecular descriptors.

Software	Descriptors	Abbreviation	Software	Descriptors	Abbreviation
ChemOffice	Heat of formation (KJ mol^{-1})	$H°$	**ChemSketch**	Percent ratios of hydrogen	**H%**
	Gibbs free energy (KJmol^{-1})	G		Percent ratios of oxygen	**O%**
	Ideal gas thermal capacity (J mol^{-1} K^{-1})	**IGTC**		Percent ratios of carbon	**C%**
	Melting point (Kelvin)	T		Surface tension	γ
	Critical temperature(Kelvin)	**CT**		Index of refraction	n
	Boiling point(Kelvin)	**TB**		Density	d
	Critical pressure (Bar)	**CP**		Log *P*	**Log *P***
	Henry's law constant	KH		Winner index	W
	Total valence connectivity	**TVC**	**MarvinSketch**	Number of H-Bond acceptors	**NHA**
	Partition coefficient	**PC**		Number of H-Bond donors	**NHD**
	Number of rotatable bonds	**NRB**		Balaban index	J
	Shape coefficient	I		Polar surface area ($A°$)	**PSA**
	Sum of valence degrees	**SVD**			
	Total connectivity	**TC**			

Statistical analysis

In this step, Matrix of correlation was used to determine the non-linearity of variables (descriptors) and to select the descriptors correlated with the property [10].

Consequently, Multiple Linear Regression (MLR) is used to study the relationship between a dependent variable and several independent variables; it minimizes the differences between actual values and predicted values and has been used to select the

descriptors to be used as inputs in Multiple Non-Linear regression (MNLR) and Artificial Neural Network ANN (Multi-Layer Perceptron (MLP) and Radial Basis Function Networks (RBF) types). Multiple linear and nonlinear regressions were used to predict the effects on the property, the equations were justified by the correlation coefficient (r), the mean square error (MSE), the Fisher value (F) and the significance level (p) [11].

MLR, MNLR, and ANN are generated using the SPSS 19.0 statistical package [2].

Cross-Validations, the most commonly used techniques for internal validation, are statistical techniques in which different proportions of chemicals are iteratively held-out from the training set used for model development (an optimal parameters K selection step) and "predicted" as new by the developed model in order to verify internal "predictivity". In this work, the Leave-One-Out is used, this procedure successively removes one molecule from the training set containing 37 molecules. A QSPR model is constructed on an "36" set of compounds and the molecule removed is predicted by the model. This procedure is repeated "37" times in order to predict the property of all molecules [12].

Y-randomization, randomly scrambling the responses, is another internal validation approach that must be used in parallel with Cross-Validations, and must always be applied to test the significance of the derived QSPR model, highlighting the presence of apparent models, obtained only by chance correlation [12]. We performed in this work 100-y-randomization tests for the MLR and MNLR models. In this test, random QSPR models are generated by randomly shuffling the dependent variable while keeping the independent variables as it is. The new QSPR models are expected to have significantly low r^2 and r^2cv values for several trials, which confirm that the developed QSPR models are robust.

The permutation test proved to be a good tool for detecting the presence of the trends in residuals of multivariate regression models. The quality of the permutation test depends on the number of permutations used. A total of 500,000 permutations are enough for reproducibility of the test results [13]. In this work we used the Matlab code for the permutation test algorithm presented in the literature [13]. When the p-value for the test is smaller than the level of significance adopted ($\alpha = 0.05$), the residuals are not random. Otherwise, there are no trends in the residuals [13].

Other useful parameters to be considered are the RMSEP (Root Mean Squared Errors of prediction) calculated on test set. The r and rcv values are good tests for evenly distributed data, but they are not always reliable for unevenly distributed data sets; instead RMSEP provide a more reliable indication of the fitness of the model, independently of the applied splitting. The randomization t-test for the comparison of the predictive accuracy (RMSEP) of methods is useful in this case. In this work we used the Matlab code for the randomization t-test algorithm presented in the literature [14]. When the p-value for the test is smaller than the level of significance adopted ($\alpha = 0.005$ for 199 randomization trials) [14], the difference between methods is significant.

3. RESULTS AND DISCUSSION

Data set for analysis

A QSPR study was carried out for a series of 51 odorant molecules, as indicated above, to determine a quantitative relationship between the structure and the property studied. The values of the 26 descriptors are shown in Table S1 and Table S2, and the correlations between this descriptors and the Log(1/K) value are shown in Table 3 as a matrix of correlation.

Multiple Linear Regressions (MLR)

The results of the PCA analysis are used to select the input data of the MLR. So, at the beginning we have eliminated all variables (descriptors) whose correlations are small (not significant, $r \leq 0.3$) with respect to the dependent variable (Log(1/K)). In order to reduce the redundancy existing in our data matrix, the highly correlated descriptors ($r \geq 0.9$) and which have the low correlation coefficient value in relation to the dependent variable have been excluded (Table 3).

The VIF (Variance Inflation Factor) was defined as $1/(1-r^2)$, where r was the multiple correlation coefficient for an independent variable against all other descriptors in the model. The models with a VIF greater than 5 were unstable and were eliminated; the models with VIF values between 1 and 4 may be accepted.

At this stage VIF values greater than 5 were found, then to improve the results (Table 4), the highly-correlated descriptors ($r \geq 0.8$) and which have the low value of correlation coefficient with the dependent variable were eliminated (Table 4).

Table 3. Matrix of correlation.

	Log(1/k)	H°	G	IGTC	T	CT	TB	CP	KH	TVC	PC	NRB	I	SVD	TC	H%	O%	C%	γ	n	D	log P	W	NHD	J	PSA
Log(1/K)	**1**																									
H°	**0.596**	**1**																								
G	**0.514**	**0.918**	**1**																							
IGTC	**-0.489**	**-0.572**	-0.214	**1**																						
T	0.033	**-0.240**	0.065	**0.667**	**1**																					
CT	-0.138	-0.126	0.180	**0.773**	**0.627**	**1**																				
TB	-0.053	-0.214	0.131	**0.835**	**0.766**	**0.958**	**1**																			
CP	**0.628**	**0.633**	**0.314**	**-0.969**	**-0.587**	**-0.720**	**-0.745**	**1**																		
KH	**0.913**	**0.661**	**0.660**	**-0.334**	0.147	0.018	0.104	**0.491**	**1**																	
TVC	**0.506**	**0.453**	0.294	**-0.672**	-0.289	**-0.709**	**-0.612**	**0.742**	**0.406**	**1**																
PC	**-0.516**	**-0.549**	-0.206	**0.966**	**0.659**	**0.817**	**0.851**	**-0.942**	**-0.381**	**-0.680**	**1**															
NRB	**-0.411**	**-0.592**	**-0.325**	**0.797**	**0.762**	**0.675**	**0.741**	**-0.790**	**-0.380**	**-0.611**	**0.858**	**1**														
I	-0.043	0.080	0.123	0.029	-0.066	0.010	-0.003	-0.015	-0.050	0.162	0.053	-0.108	**1**													
SVD	**-0.617**	**-0.611**	**-0.451**	**0.737**	0.219	**0.629**	**0.546**	**-0.801**	**-0.513**	**-0.882**	**0.716**	**0.526**	-0.078	**1**												
TC	**0.459**	**0.542**	0.236	**-0.921**	**-0.640**	**-0.840**	**-0.851**	**0.938**	**0.344**	**0.841**	**-0.931**	**-0.845**	0.058	**-0.779**	**1**											
H%	0.272	0.050	0.286	0.273	**0.585**	0.106	**0.324**	-0.144	**0.302**	**0.406**	0.256	**0.301**	0.142	**-0.434**	-0.109	**1**										
O%	-0.266	**-0.534**	**-0.794**	-0.246	**-0.451**	**-0.374**	**-0.426**	0.137	**-0.417**	-0.208	-0.234	-0.114	-0.192	**0.303**	0.129	**-0.706**	**1**									
C%	0.232	**0.624**	**0.861**	0.208	**0.355**	**0.415**	**0.408**	-0.118	**0.403**	0.121	0.199	0.042	0.185	-0.226	-0.120	**0.529**	**-0.975**	**1**								
γ	0.275	-0.225	0.007	**0.551**	**0.764**	**0.722**	**0.833**	**-0.431**	**0.328**	**-0.413**	**0.587**	**0.660**	-0.133	0.291	**-0.640**	**0.357**	-0.196	0.122	**1**							
n	**0.344**	0.194	**0.425**	**0.464**	**0.476**	**0.785**	**0.786**	**-0.340**	**0.531**	**-0.435**	**0.434**	0.226	-0.034	**0.377**	**-0.501**	0.084	**-0.394**	**0.446**	**0.719**	**1**						
D	-0.284	**-0.595**	**-0.633**	**0.303**	-0.030	0.287	0.229	**-0.360**	-0.277	**-0.692**	0.287	0.225	-0.201	**0.772**	**-0.434**	**-0.649**	**0.734**	**-0.673**	**0.302**	0.273	**1**					
log P	**-0.506**	**-0.416**	-0.071	**0.928**	**0.603**	**0.771**	**0.789**	**-0.913**	**-0.331**	**-0.601**	**0.932**	**0.728**	0.097	**0.673**	**-0.850**	0.260	**-0.368**	**0.358**	**0.429**	**0.435**	0.159	**1**				
W	**-0.494**	**-0.541**	-0.240	**0.936**	**0.566**	**0.823**	**0.818**	**-0.918**	**-0.349**	**-0.769**	**0.919**	**0.747**	0.000	**0.872**	**-0.887**	-0.001	-0.090	0.109	**0.522**	**0.521**	**0.480**	**0.884**	**1**			
NHD	**0.734**	0.218	**0.368**	0.085	**0.501**	0.198	**0.404**	0.115	**0.809**	0.282	0.038	0.062	-0.037	**-0.358**	-0.009	**0.658**	**-0.445**	**0.325**	**0.654**	**0.547**	-0.209	-0.014	-0.051	**1**		
J	**-0.542**	**-0.390**	-0.229	**0.542**	-0.023	0.182	0.181	**-0.564**	**-0.357**	**-0.334**	**0.432**	0.030	0.126	**0.601**	**-0.363**	-0.124	0.049	-0.019	-0.141	0.162	**0.338**	**0.532**	**0.505**	-0.224	**1**	
PSA	**-0.316**	**-0.819**	**-0.829**	**0.382**	0.089	0.185	0.217	**-0.418**	**-0.363**	**-0.562**	**0.360**	**0.377**	-0.210	**0.698**	**-0.445**	**-0.407**	**0.748**	**-0.767**	**0.357**	0.108	**0.918**	0.196	**0.489**	-0.111	**0.313**	**1**

Table 4. Multicollinearity statistics.

	H°	*KH*	n	J
Tolerance	0.489	0.348	0.620	0.724
VIF	2.043	2.876	1.614	1.381

The relationship obtained using this method corresponds to the linear combination of these descriptors: Heat of formation (H°), Henry's law constant (KH), Index of refraction (n) and Balaban index (J).

The resulting equation is as follows:

$$Log\left(\frac{1}{K}\right) = 5.718 - 1.934 \times 10^{-4} \times H° + 0.588 \times KH - 2.836 \times n - 0.268 \times J \quad \textbf{(Equation 1)}$$

N = 37; r = 0.942; r^2= 0.886; MSE = 0.026; F = 62.460; P < 0.0001.

In this equation, N is the number of compounds, r is the correlation coefficient, r^2 is the coefficient of determination, MSE is the mean squared error, F is the Fisher's criterion and P is the significance level.

It is observed that the correlation coefficient r is very high, and the mean squared error value (MSE) is low, which makes it possible to indicate that the model is more reliable. A P value much smaller than 0.05 indicates that the regression equation is statistically significant, we can conclude, with confidence, that the model provides a significant amount of information [15].

The predicted Log(1/K) values calculated from equation (1) are given in Table 5 in comparison to the observed values.

The correlation between the predicted and observed Log(1/K) and the residue values are shown in Fig. 2.

The residuals should not show any trend. A trend would indicate that the residuals were not independent. In the permutation test, the MLR model showed p-value more than the significance level of 0.05, with result of 0.4999. In this case, the residuals of the MLR model were random.

The descriptors proposed in equation (1) by MLR are therefore used as input parameters in the multiple non-linear regressions (MNLR) and the artificial neural network (ANN) [16].

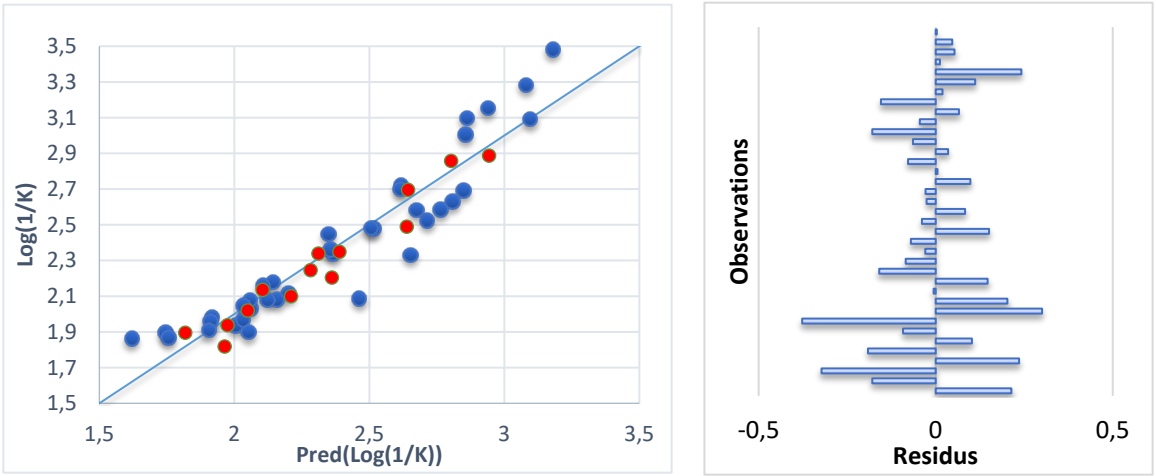


Figure 2. Graphical representation of calculated and observed property and the residues values calculated by MLR (training set in blue; test set in red).

Multiple Non-Linear Regression (MNLR)

We also used multiple non-linear regression model technique to quantitatively improve the structure–property relationships by accounting for several parameters. MNLR is the most commonly used tool for the study of multidimensional data. We applied it to the data matrix constituted from the descriptors proposed by the MLR corresponding to the set of 37 molecules [17].

The resulting equation is as follows:

$$Log\left(\frac{1}{K}\right) = 11.497 + 1.008 \times 10^{-4} \times H° + 0.138 \times \log(-H°) + 1.377 \times KH - 4.279 \times \log(KH) - 5.430 \times n - 2.101 \times J + 0.309 \times J^2 \quad \textbf{(Equation2)}$$

N = 37; r = 0.957; r^2= 0.917; MSE = 0.021; F = 45.568; P < 0.0001.

The predicted Log(1/K) values calculated from equation (2) are given in Table 5 in comparison to the observed values. The correlation between the predicted and observed Log(1/K) and the residue values are shown in Fig. 3. In the permutation test, the MNLR model showed p-value more than the significance level of 0.05, with result of 0.3872. In this case, the residuals of the MNLR model were random.

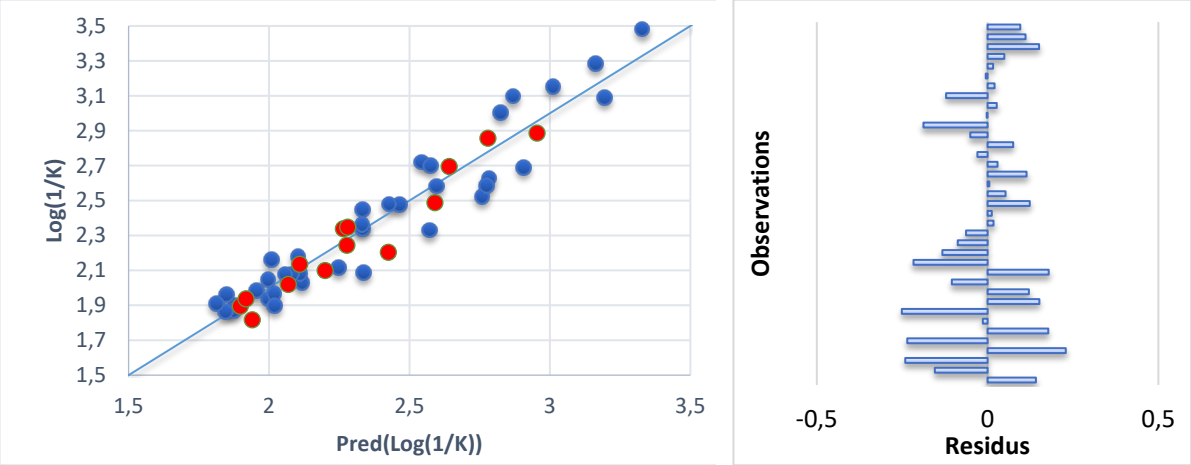


Figure 3. Graphical representation of calculated and observed property and the residues values calculated by MNLR (training set in blue; test set in red).

Artificial Neural Networks (ANN)

In order to increase the probability of a good characterization of the molecules studied, the Artificial Neural Networks (ANN) can generate a predictive model of the QSPR relationship between the descriptors obtained from the MLR and the observed property.

In this study, we used two types of artificial neural networks: Multi-Layer Perceptron (MLP) and Radial Basis Function Networks (RBFs).

Multi-Layer Perceptron (MLP): The ANN model has aroused great interests as its universal function approximators are capable of mapping any linear or nonlinear functions. The multi-layer perceptron (MLP) neuronal network model is a supervised neural network based on the original simple perceptron model with back propagation for training the network. It commonly consists of an input layer of source nodes, an output layer and one or more hidden layers of computation nodes (neurons) that increasing the learning power of the MLP model. The number of hidden neurons determines the learning capacity of MLP network. It is most recommended to select the network which performs best with the least possible number of hidden neurons.

The property model computed by the MLP method was developed using the properties of several molecules studied (Fig. 4). The correlation between the predicted and observed Log(1/K) and the residue values are shown Fig.5. In the permutation test, the MLP model showed p-value more than the significance level of 0.05, with result of 0.2856. In this case, the residuals of the MLP model were random.

The predicted Log(1/K) values calculated by MLP method are given in Table 5 to comparison to the observed values.

Radial Basis Function Networks (RBFs): RBF neural networks are neural networks based on localized basis functions and iterative function approximation. In terms of structure, a RBF is composed of three layers, namely an input layer, an output layer, and a hidden layer (see Fig. 6). These types of networks are of fixed architecture with a single hidden layer; this is while MLP may be of more than one hidden layers. Indeed, a RBF represents a special case of a MLP [18]. Owing to their simple design, extremely strong tolerance to input noises, and fast yet pervasive training capabilities, these networks have attracted a large deal of attention. In RBF, there is a single input layer wherein no processing is undertaken. The hidden layer, however, contains radial basis functions, with the output layer solely containing collectors. In fact, the output layer linearly combines all outputs from neurons in the

hidden layer to generate the network output. Compared to MLP networks, this type of network requires larger number of neurons, even though they enjoy shorter designs, with the principal distinction being the application of activation functions to be used by neurons [19].

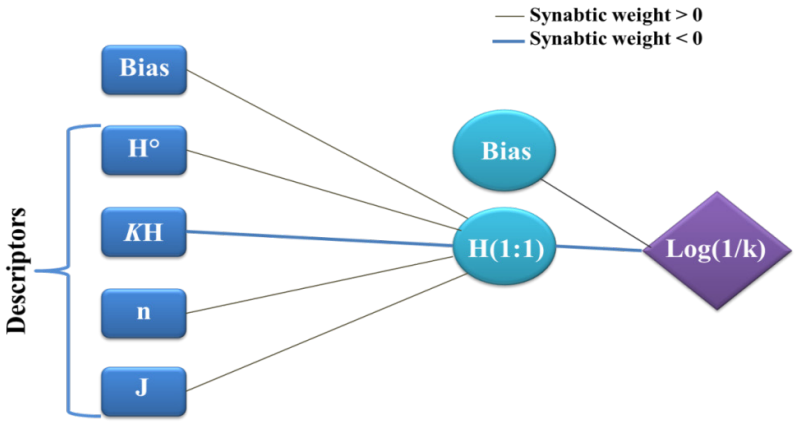


Figure 4. The architecture of the MLP method used (four input variables, one neuron in the hidden layer and one neuron to the output layer).

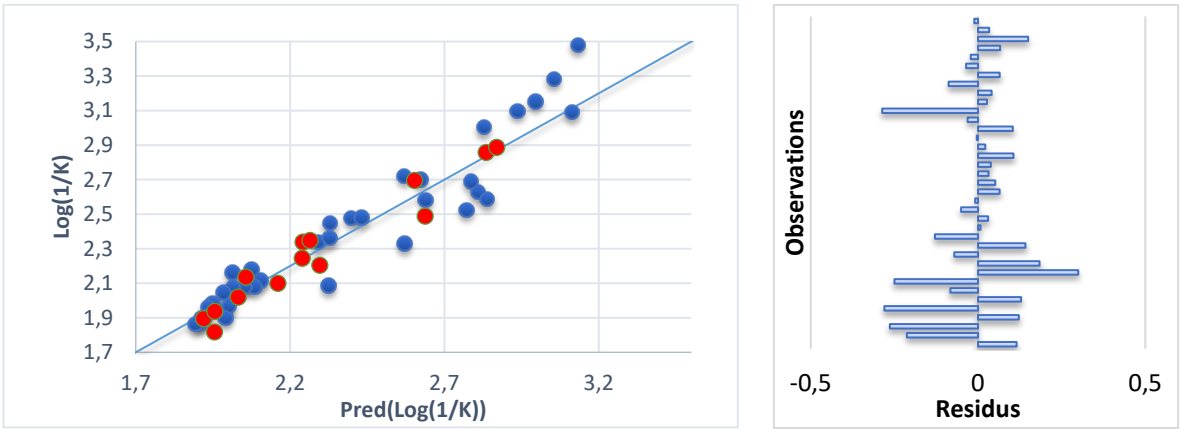


Figure 5. Graphical representation of calculated and observed property and the residues values calculated by MLP method (training set in blue; test set in red).

The property model computed by the RBF method was developed using the properties of several molecules studied (Fig. 6). The correlation of the predicted and observed property and the residue values are illustrated in Fig. 7. In the permutation test, the RBF model showed p-value more than the significance level of 0.05, with result of 0.4512. In this case, the residuals of the RBF model were random.

The predicted Log(1/K) values calculated by RBF method are given in Table 5 in comparison to the observed values.

Internal Validation

Cross-Validation: The Cross-Validation statistical procedure can be used to evaluate the predictive power of QSPR models. The Leave-One-Out procedure successively removes one molecule from the training set containing n molecules. A QSPR model is constructed on an "n-1" set of compounds and the molecule removed is predicted by the model. This procedure is repeated "n" times in order to predict the property of all molecules.

The QSPR model expressed by the equations of MLR and MNLR methods is validated by its appreciable values of r^2cv (Table 6) obtained using the Leave-One-Out (LOO) procedure. The value of r^2cv greater than 0.5 is the basic condition for qualifying a QSPR model as valid.

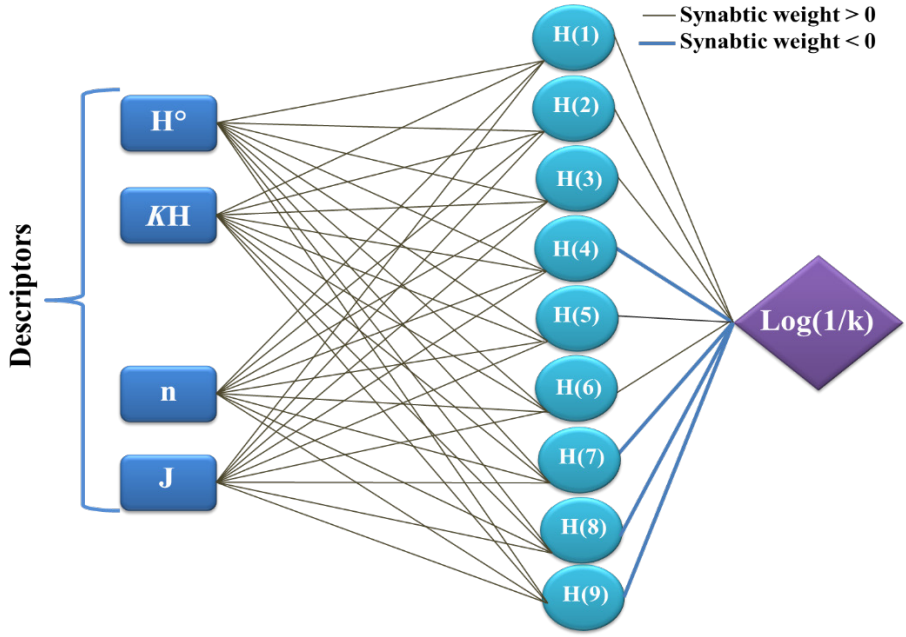


Figure 6. The architecture of the RBF method used (four input variables, nine neurons in the hidden layer and one neuron to the output layer).

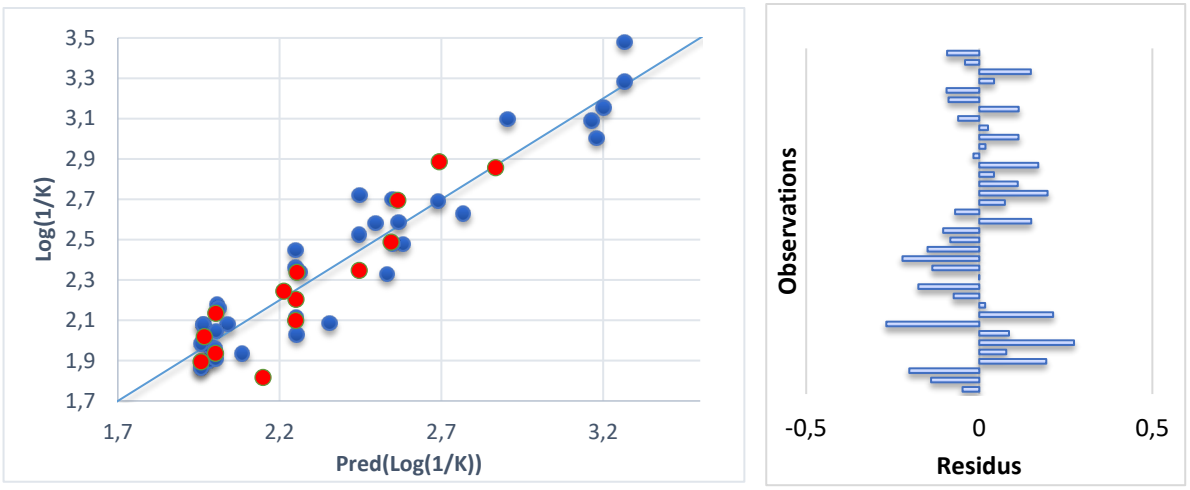


Figure 7. Graphical representation of calculated and observed property and the residues values calculated by RBF method (training set in blue; test set in red).

We use Cross-Validation as an internal test of the quality of MLR and MNLR models. The performance of models was good and was characterized by r^2cv values; 0.846 for the MLR and 0.860 for MNLR method (Table 6).

y-Randomization test: To ensure the developed QSPR model is robust and not derive due to chance, the y-randomization test was performed on the training set data as recommended [20]. In this test, MLR and MNLR models are generated by randomly scrambling the dependent variable (property data) while keeping the independent variable (descriptors) unchanged. The resulting models are expected to have significantly low r^2 and cross validated r^2cv values for several trials, which confirm that the developed models are robust. We performed 100-y-randomization tests and observed that for all the models, the values of r^2 and r^2cv were <0.5 (Fig. 8). This test confirms that the developed models are robust and not derived merely due to chance.

Table 5. Comparison of the observed values with those calculated by MLR, MNLR and ANN (MLP and RBF types) methods.

N°	Log(1/K) (obs.)	Log(1/K) (calc.)				N°	Log(1/K) (obs.)	Log(1/K) (calc.)			
		MLR	MNLR	MLP	RBF			MLR	MNLR	MLP	RBF
1	3.153	2.939	3.012	2.996	3.202	**27**	2.338	2.367	2.333	2.290	2.264
2	2.629	2.808	2.783	2.808	2.768	**28**	2.447	2.349	2.333	2.330	2.250
3	2.330	2.653	2.571	2.571	2.532	**29**	2.361	2.356	2.332	2.328	2.250
4	3.097	2.862	2.868	2.936	2.904	**30**[a]	2.338	2.312	2.265	2.242	2.254
5	2.523	2.715	2.759	2.772	2.446	**31**	2.081	2.159	2.111	2.085	2.039
6[a]	2.857	2.803	2.779	2.835	2.868	**32**	2.178	2.143	2.103	2.075	2.008
7	2.721	2.619	2.543	2.570	2.447	**33**	1.968	2.032	2.019	2.003	1.985
8	2.582	2.674	2.596	2.639	2.496	**34**	2.586	2.765	2.774	2.838	2.568
9	2.086	2.463	2.337	2.325	2.354	**35**[a]	2.488	2.639	2.591	2.638	2.545
10	3.480	3.180	3.329	3.133	3.267	**36**[a]	2.244	2.282	2.277	2.239	2.213
11	3.283	3.081	3.162	3.055	3.266	**37**[a]	2.135	2.105	2.109	2.057	2.003
12	3.090	3.096	3.195	3.114	3.164	**38**	2.078	2.123	2.081	2.053	1.965
13	3.004	2.857	2.825	2.828	3.180	**39**	1.983	1.917	1.956	1.949	1.958
14	2.690	2.849	2.907	2.787	2.690	**40**[a]	1.894	1.819	1.898	1.920	1.957
15[a]	2.886	2.944	2.954	2.869	2.693	**41**	1.898	2.053	2.020	1.991	1.960
16[a]	2.204	2.361	2.425	2.297	2.251	**42**[a]	2.019	2.050	2.069	2.032	1.967
17	2.116	2.200	2.248	2.107	2.251	**43**	2.078	2.058	2.057	2.016	1.964
18	2.030	2.059	2.117	2.004	2.252	**44**	1.868	1.757	1.874	1.912	1.957
19[a]	2.099	2.210	2.200	2.161	2.250	**45**	1.863	1.621	1.847	1.893	1.957
20	1.935	2.005	1.998	1.990	2.085	**46**[a]	2.348	2.390	2.280	2.265	2.446
21	1.897	1.746	1.879	1.913	1.981	**47**	2.046	2.034	1.997	1.985	2.003
22[a]	2.695	2.644	2.642	2.603	2.566	**48**	2.160	2.107	2.009	2.014	2.011
23	2.477	2.516	2.465	2.398	2.582	**49**[a]	1.937	1.975	1.918	1.956	2.002
24	2.699	2.616	2.576	2.623	2.549	**50**	1.960	1.913	1.849	1.934	2.001
25	2.480	2.506	2.427	2.433	2.550	**51**	1.909	1.906	1.813	1.927	2.002
26[a]	1.817	1.964	1.941	1.955	2.149						

[a]Test Set.

Table 6. r^2cv values obtained by the leave-one-out (LOO) method.

	MLR	MNLR
rcv	0.920	0.927
r^2cv	0.846	0.860
MSE	0.032	0.029

External Validation

To estimate the predictive power of the MLR, MNLR and ANN (MLP and RBF types) models, we must use a set of compounds that have not been used in the training set to establish the QSPR model. The models established in the calculation procedure using the odorant molecules are used to predict the property of the remaining 14 molecules. The main performance parameters for the four models are shown in Table 7.

The results obtained by MLR, MNLR and ANN (MLP and RBF types) models, are very sufficient to conclude the performance of models; it's confirmed by the test done with the 14 compounds.

A comparison of the quality of MLR, MNLR and ANN (MLP and RBF types) models shows that the four approaches have better predictive capability gives better results. MLR, MNLR and ANN were able to establish a satisfactory relationship between the molecular descriptors and the retention/release property of the studied compounds, it can be also seen that MLR method yielded the smallest RMSEP but the comparison of the prediction accuracy of four methods by randomization t-test show that the difference

between MLR and the other methods (MNLR and ANN (MLP and RBF types)) is only indicative (p = 0.01 for 199 randomization trials, so p>(α=0.005)), in this case the four methods cannot account for a significant difference in prediction accuracy.

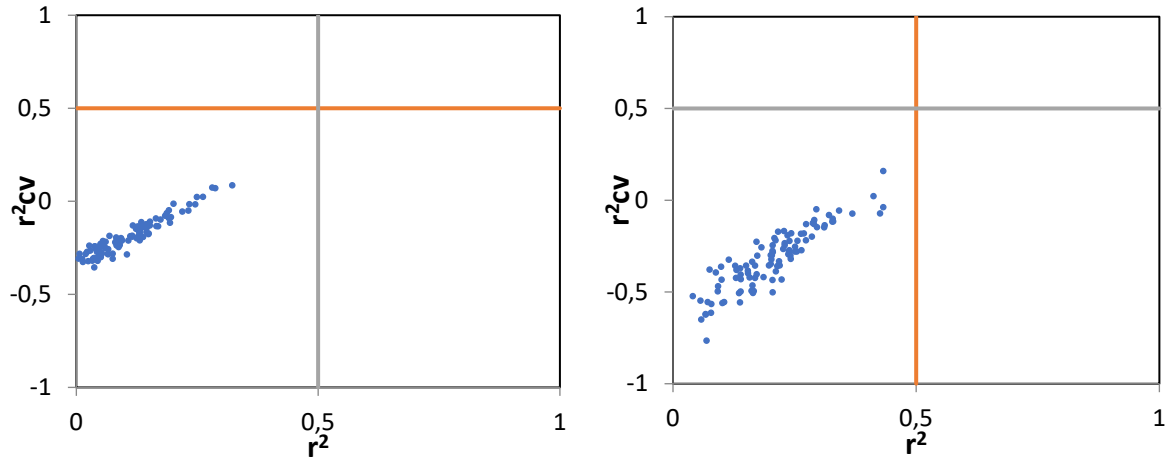


Figure 8. y-Randomization plot of MLR and RNLM model.

Table 7. Comparison of MLR, MNLR and ANN (MLP and RBF types) models.

				ANN	
		MLR	**MNLR**	**MLP**	**RBF**
Training Set	r	0.942	0.957	0.955	0.954
	r^2	0.886	0.917	0.912	0.910
	MSE	0.026	0.021	0.018	0.019
Test Set	r	0.973	0.966	0.972	0.924
	r^2	0.946	0.933	0.945	0.854
	MSE	0.007	0.009	0.007	0.019
	RMSEP	0.083	0.092	0.084	0.137

Domain of applicability

Evaluation of the applicability domain of the QSPR model is considered as an important step to establish that the model is reliable to make predictions within the chemical space for which it was developed [21]. There are several methods for defining the applicability domain of a QSPR model, but we used the most commonly used leverage approach in this study [22]. Leverage of a given chemical compound hi is defined as:

$$\mathbf{hi = xi^T(X^T X)^{-1}xi \quad (i = 1. . . .n)}$$

where xi is the descriptor row of the query compound and X is the descriptor matrix of the training set compounds used to develop the model. As a prediction tool, the warning leverage h* is defined as:

$$\mathbf{h^* = 3(p + 1)/n}$$

where n is the number of training compounds, and p is the number of descriptors in the model.

The test compounds with leverages hi<h* are considered to be reliably predicted by the model. The Williams plot is used to interpret the applicability domain of the model. The domain of reliable prediction for external test set compounds is defined as compounds which have leverage values within the threshold (hi<h*) and standardized residuals no greater than 3 units (±δ). Test set compounds where (hi>h*) are considered to be unreliably predicted by the model due to substantial extrapolation. For the training set, the Williams plot is used to identify compounds with the greatest structural influence (hi>h*) in developing the model.

From the Williams plot (Fig. 9), it is obvious that all compounds in the test set fall inside the domain of the MLR model (the warning leverage limit is 0.405). For all the compounds in the training and test sets, their standardized residuals are smaller than three standard deviation units (3±δ). Therefore, the predicted retention/release property by the developed MLR model is reliable.

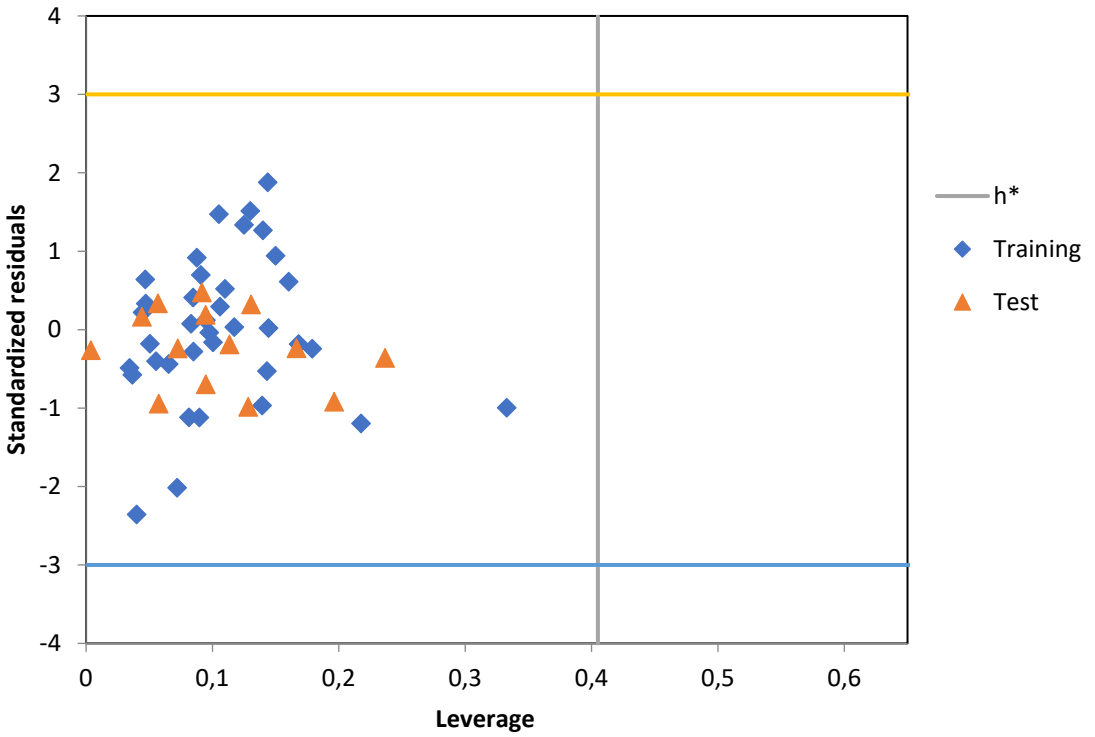


Figure 9. Williams plot to evaluate the applicability domain of MLR model.

Proposed novel compounds

QSPR correlates property data with the physicochemical properties of a group of compounds. It has been frequently used to predict proprieties of new compounds and to design compounds with desired properties.

The developed equation (1) can be used for the designing of new odorant molecules derivatives with improved retention/release property (Log (1/k)).

Comparing *t*-test and standardized coefficient values of descriptors (Table 8) indicates that the influences of the Henry's law constant *K*H on Log (1/k) are stronger than those of the others.

Table 8. t-test and standardized coefficient values of descriptors for equation (1).

	Standardized coefficient	*t*-test	Sign.
H°	-0.053	-0.627	0.535
*K*H	0.923	9.133	< 0.0001
n	-0.098	-1.290	0.206
J	-0.191	-2.732	0.010

The equation (1) of the MLR method indicated the positive correlation of the Henry's law constant KH.

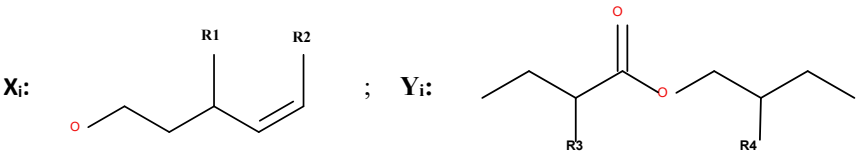
The obtained results show that, to increase retention property of odorant molecules, we will increase Henry's law constant KH. Moreover, to increase release property, we will decrease Henry's law constant KH of this molecule, by adding suitable substituents and calculated their property using the equation (1).

The structures of the designed compounds and their parameter values calculated by the same methods as well as Log(1/k) values theoretically predicted by the MLR model (Equation (1)) are listed in Table 9.

From the predicted properties, it has been observed that the designed compounds X_1, X_2, X_3, and X_5 have higher Log (1/k) values than the existing compounds. Also, the designed compounds Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7, $Y_8$and Y_9 have lower Log (1/k) values than the existing compounds in the case of the 51 studied compounds (Table 1).

The leverage values (h) calculated by equation (1) of the MLR for the new designed compounds are displayed in Table 9, only three compounds X_2, X_3 and Y_6 are defined as outliers and consequently they are not being considered, because they have higher leverage which is greater than h* (h*=0.405) [23].

Table 9. Values of descriptors, retention/ release property (Log(1/K)), and leverages (h) for the new designed compounds.



	Designed compounds	*H*°	*K*H	n	J	Log (1/K)	Leverage (h)
X1	R1=OH; R2=CH_3	-359.69	4.8356	1.476	2.92	3.661	0.380
X2	R1=CH_3; R2=OH	-359.69	7.01	1.476	2.92	4.939	0.677
X3	R1=H; R2=OH	-333.77	7.13	1.48	2.45	5.119	0.667
X4	R1=H; R2=CH_2OH	-354.41	4.84	1.479	2.53	3.756	0.396
X5	R1=CH_2OH; R2=CH_3	-380.33	4.71	1.476	3.18	3.523	0.387
Y1	R3=C_2H_5; R4=CH_3	-598.21	1.16	1.424	3.65	1.498	0.149
Y2	R3=CH_3; R4=C_2H_5	-598.21	1.16	1.424	3.61	1.509	0.143
Y3	R3=C_2H_5; R4=C_2H_5	-618.85	1.04	1.427	3.75	1.396	0.183
Y4	R3=CH_3; R4=F	-753.04	1.11	1.403	3.49	1.601	0.259
Y5	R3=F; R4=CH_3	-753.04	1.11	1.403	3.49	1.601	0.259
Y6	R3=F; R4=F	-928.51	0.93	1.384	3.49	1.583	0.476
Y7	R3=CH_3; R4=$CH(CH_3)_2$	-624.13	1.04	1.426	3.79	1.389	0.189
Y8	R3=$CH(CH3)_2$; R4=CH_3	-624.13	1.04	1.426	3.86	1.371	0.203
Y9	R3=CH $(CH_3)_2$; R4=$CH(CH_3)_2$	-670.69	0.79	1.431	4.08	1.159	0.280

4. CONCLUSION

Multiple linear and non-linear Regression and artificial neural networks (MLP and RBF types) were used to construct quantitative structure property relation models of odorant molecules for their retention/release property. The results show that the models proposed in this paper can predict retention/release property accurately and that the selected parameters are pertinent. The accuracy and predictability of the proposed models were illustrated by comparison of the key statistical terms r or r^2 and the predictive powers of the equations were validated by an internal test (Cross validation and 100-y-randomization) and external test set.

All used models results have substantially good predictive capability, but MLR gives the most important interpretable results. The applicability domain of the MLR model was defined.

We conclude that the most important finding about this research is that we have been able to design and proposed some new compounds with high or lower values property than the existing ones by adding suitable substituents and calculated their property using regression equation. Consequently, the proposed models will reduce the time and cost of synthesis and determination of the retention/release property for the odorant molecules.

5. ACKNOWLEDMENTS

We are grateful to the “Association Marocaine des Chimistes Théoriciens” (AMCT) for its pertinent help concerning the programs.

6. REFERENCES AND NOTES

[1] McEwan, J. A.; Thomson, D. M. H. *Food Quality and Preference* **1988**, *1*, 3. [CrossRef]

[2] Pernollet, J. C.; Sanz, G.; Loïc Briand, L. *C. R. Biol.* **2006**, *329*, 679. [CrossRef]

[3] Ghavami, R.; Faham, S. *Chromatographia* **2010**, *72*, 893. [CrossRef]

[4] Zhang, Y.; Pan,Y.; Jiang, J.; Ding, L. *J. Environ. Chem. Eng.* **2014**, *2*, 868. [CrossRef]

[5] Ayed, C.; Lubbers, S.; Andriot, I.; Merabtine, Y.; Guichard, E.; Tromelin, A. *Journal of Food Research International* **2014**, *62,* 846. [CrossRef]

[6] SPSS 19.0. Available: [Link]

[7] ACDLABS 10., Advanced Chemistry Development. Inc. Toronto. ON. Canada 2015. Available: [Link]

[8] MarvinSketch 5.11.4., Chem Axon 2012. Available: [Link]

[9] ChemBioOffice, PerkinElmer Informatics 2010. [Link]

[10] Larif, M.; Adad, A.; Hmammouchi, R.; Taghki, A.I.; Soulaymani, A.; Elmidaoui, A.; Bouachrine, M.; Lakhlifi, T. *Arabian J. Chem.* **2017**, *10*, S946. [CrossRef]

[11] Chtita, S.; Larif, M.; Ghamali, M.; Bouachrine, M; Lakhlifi, T. *Journal of Taibah University for Science* **2015**, *9*, 143. [CrossRef]

[12] Gramatica, P. *QSAR Comb. Sci.* **2007**, *26*, 694. [CrossRef]

[13] Filgueiras, P. R.; Alves, J. C. L.; Sad, C. M. S .; Castro, E. V. R.; Dias, J. C. M.; Poppi, R. J. *Chemom. Intell. Lab. Syst.* **2014**, *133*, 33. [CrossRef]

[14] Van der Voel, H. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 313. [CrossRef]

[15] Hmamouchi, R.; Larif, M.; Chtita, S.; Adad, A.; Bouachrine, M.; Lakhlifi,T. *Journal of Taibah University for Science* **2016**, *10*, 451. [CrossRef]

[16] Chtita, S.; Hmamouchi, R.; Larif, M.; Ghamali, M.; Bouachrine, M. Lakhlifi, T. *Journal of Taibah University for Science* **2016**, 10, 868. [CrossRef]

[17] Chtita, S.; Hmamouchi, R.; Larif, M.; Ghamali, M.; Bouachrine, M.; Lakhlifi. *Orbital: Electron. J. Chem*. **2015**, *7*, 176. [CrossRef]

[18] Schalkoff, R. J., Artificial Neural Networks. McGraw-Hill, New York, 1997. [Link]

[19] Marini, F. *Anal. Chim. Acta* **2009**, *635*, 121. [CrossRef]

[20] Roy, K.; Mitra, I. *Comb. Chem. High Throughput Screening* **2011**, *14*, 450. [CrossRef]

[21] Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.D.; McDowell, R.M.; Gramatica, P. Environ. Health Perspect. **2003**, *111*, 1361. [CrossRef]

[22] Gramatica, P. *QSAR Comb. Sci.* **2007**, *26*, 694. [CrossRef]

[23] Chtita, S.; Ghamali, M.; Hmamouchi, R.; Elidrissi, B.; Bourass, M.; Larif, M.; Bouachrine, M.; Lakhlifi, T. *Adv. Phys. Chem.* **2016**, 1. [CrossRef]