

QSAR modeling for predicting the larvicidal activity of essential oils targeting *Culex pipiens pallens* (Diptera: Culicidae)

Samia Alem (✉ alem.samia23@gmail.com)

University of Badji Mokhtar Annaba: Universite Badji Mokhtar Annaba

Youssef Driouche

Environmental research center <https://orcid.org/0000-0001-8024-407X>

Hamza Haddag

University of Badji Mokhtar Annaba: Universite Badji Mokhtar Annaba

Zihad Bouslama

Environmental Research Center

Research Article

Keywords: QSAR modeling, larvicidal activity, *Culex pipiens pallens*, essential oils, MLR, SVM

Posted Date: November 30th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-2272165/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

The search for new larvicides suited for vector control of mosquitoes requires considerable time, an enormous budget, and several analytical setups.

Fortunately, the use of quantitative structure–activity relationship (QSAR) modeling allows the prediction of the larvicidal activity of structurally diverse chemicals against mosquitoes in a way quick and costless. This approach can be helpful to study for making biolarvicide with highest ability to destroy mosquito larvae.

We propose a quantitative structure-activity relationship model using two different statistical methods, multiple linear regression (MLR) and Support vector machine (SVM) for predicting the larvicidal activity of 30 compounds of essential oils (EOs) isolated from the root of *Asarum heterotropoides* against *Culex pipiens pallens*. A model with four theoretical descriptors derived from Dragon software was developed applying the genetic algorithm (GA)-variable subset selection (VSS) procedure. The statistical parameters, $R^2 = 0.9716$, $Q^2_{\text{LOO}} = 0.9595$, $s = 0.1690$ of the model developed by MLR showed a good predictive capability for log LC₅₀ values. The comparison between the results of MLR and SVM models showed that the SVM model present a good alternative to construct a QSAR model for the prediction of the larvicidal activity.

Introduction

Human has been suffering from mosquitoes bites and their harmful effects since time immemorial and it is classified as the most important insect pest of man [1].

About 3492 species of mosquitoes are reported, more than a hundred species can transmit several disease to humans and other vertebrates, [2] such as malaria, encephalitis, yellow fever and chikungunya [3], dengue fever, japanese.

Mosquitoes belonging to the large family of Culicidae are widespread across the world and most mosquitoes vectors belong to *Anopheles*, *Aedes* and *Culex* genera [4].

The control of mosquito vectors is becoming difficult because of the resistance of mosquitoes to chemical insecticides and also their systematic and repeated applications. The most significant negative impacts are biological imbalances, toxic risks to humans, animals and non-target organisms more environmental pollution [5–6, 7]. Therefore there is an urgent need to develop safer, environmentally and efficient alternatives to synthetic insecticides. In this context, researchers have encouraged the use of natural products in control of mosquito larvae. Essential oils (EOs) or plants extracts are rich source of bioactive compounds which are less toxic, selective, eco-friendly, low-cost and easily biodegradable insecticides [8].

For example, the exploitation of essential oils as larvicides in new formulations of botanical insecticides by the use of encapsulation procedures has a low impact for the environment and non-target organisms [9].

In recent decades, several studies have been interested in the interactions between plants and insects. This co-evolution has revealed the potential use of plant metabolites for this purpose. Many components of essential oils (EOs) are known for their insecticidal properties [10].

The most promising groups of plants that have insecticidal activity are Meliaceae, Rutaceae, Asteraceae, Annonaceae, Labiatae, Aristolochiaceae and Malvaceae [11].

Some studies have indicated the effectiveness of aromatic plants against mosquitoes such as *Eucalyptus Grandis* [12], *Mentha arvensis L.* [13] and *Mentha pipeirta* [14].

The constituents of essential oils have been valued as insecticides due to their broad action as toxicity, oviposition, feeding deterrence, repellence, and attraction which act on the nervous system of the insect [15].

Furthermore, larvicides have been commonly used to eradicate the populations of *Culex pipiens* for many years [16], so the ideal method for controlling mosquito populations infestation is by preventing mosquitoes at the breeding sites in the aquatic larval stage compared to the adult one [17, 18].

However, the development of plant-based larvicides is limited by the considerable time and huge budget to perform biological and clinical trials [19]. For this reason, several approaches have been employed to find new, quick, and costless methods of research of new bioactive compounds against mosquitoes, which consist in predicting the properties and activities of molecules [20].

Quantitative structure–activity relationship models are quantitative regression methods employed for determining the chemical features contributing to a specific activity [21]. Various studies that reported QSAR models for the prediction of the larvicidal activity against mosquitoes. M. Javidfar and S. Ahmad have improved QSAR models for 62 larvicidal phytocompounds against *Aedes Aegypti* [22]. Laura M. Saavedra *et al.*, have analysed by QSAR the larvicidal activity against *Aedes aegypti* (Diptera: Culicidae) from 60 plant-derived compounds using the RM variable subset selection technique based on MLR exploring 18 326 molecular descriptors [19].

This method attempts to correlate chemical or biological activities with a great number of molecular descriptors, which are real numbers calculated from molecular structures using various approaches and appropriate software [23, 25]. These descriptors represent quantitatively the encoded chemical information [26]. They may also be classified into four categories vis. hydrophobic, electronic, steric, and topological indices [27].

The main aim of the present study was to develop a new mathematical regression model for prediction of larvicidal activity of different compounds of (EOs) extracted from *Asarum heterotropoides* against third-instar larvae of *Culex pipiens pallens* using QSAR model.

Multiple Linear Regression (MLR) and Support Vector Machine (SVM) were applied for modeling the quantitative relationship between the larvicidal activity and the structural descriptors of EOs components. To achieve this goal, genetic algorithm (GA) was used to choose descriptors before to the construction of

the MLR and SVM models. Finally, Internal (cross-validation and Y-scrambling) and external statistical validations were used to verify the suggested models strength and prediction ability.

Results And Discussion

Multiple linear regression

The MLR technique gives an equation that connects structural features to the (log LC₅₀) of the compounds. Therefore, the optimal model was determined to be a four-variable model. The following is the created model's regression equation:

$$\log \text{LC}_{50} = 15.0 + 0.841 \text{ EEig03x} - 0.500 \text{ EEig11x} - 4.17 \text{ BEHm1} - 1.01 \text{ Mor24v} (1)$$

The statistical parameters listed in Table 1 are used to evaluate the model. They show that the model (Eq. (1)) established a strong correlation between the four selected variables and the studied property, with excellent parameters and a very large value of the Fisher $F = 154.1681$, indicating the model's excellence in predicting the log LC₅₀ values, and a low standard error $s = 0.1908$. We can infer that the statistical parameters of the constructed model have excellent predictive performance and that the descriptors used to define the 50% lethal concentration are well chosen. The developed model satisfies the above accept conditions (Eq. (6–9)).

Table 1
Statistical results of the developed model.

Training set					Prediction set				
R^2	Q^2_{LOO}	$Q^2_{\text{L(5)O}}$	F	s	Q^2_{EXT}	$(R^2 - R^2_0)/R^2$	$(R^2 - R^2_0)/R^2$	k	k'
0.9716	0.9595	0.9596	154.1681	0.1908	0.9263	-0.8809	-0.9427	0.9663	1.0012

Table 2 displays the statistics and definitions for the descriptors chosen. The regression coefficients of the descriptors in the models are much bigger than the standard deviation indicated by the high absolute t-values. The values of probability (P) are less than 0.05 for each descriptor that means the presence of every descriptor is statistically significant and indicates that the models are not a result of mere chance. As can be seen from Table 2, the variance inflation factor (VIF) values of all descriptors are less than 5. Thus, there is no multicollinearity between the selected descriptors, and the obtained models are stable.

Table 2
Characteristics of the selected descriptors in the MLR model.

Descriptor	Descriptor description	Coeff	SE Coeff	t- values	P	VIF
Constant		14.9850	0.8004	18.72	0.000	
<i>EEig03x</i>	Eigenvalue 03 from edge adj. matrix weighted by edge degrees	0.8413	0.1219	6.90	0.000	1.951
<i>EEig11x</i>	Eigenvalue 11 from edge adj. matrix weighted by edge degrees	-0.49977	0.08760	-5.71	0.000	1.596
<i>BEHm1</i>	highest eigenvalue n. 1 of Burden matrix / weighted by atomic masses	-4.1688	0.2392	-17.43	0.000	2.181
<i>Mor24v</i>	3D-MoRSE - signal 24 / weighted by atomic van der Waals volumes	-1.0059	0.2619	-3.84	0.001	1.164

The strong correlation between observed and predicted log LC₅₀ values for both the training and prediction sets is another indication of the model's quality.

Figure 1 shows the predicted values of the log LC₅₀ versus the experimental values, the correlation coefficient ($R^2 = 0.9716$) of this plot indicates the good agreement between these values, proving the model's trustworthiness.

In the Williams plot used to examine the model applicability domain, the majority of the 30 compounds fall inside the box and were accurately predicted (Fig. 2). The leverage values (h_{ij}) of all compounds in the training and test sets are less than the critical value ($h^*=0.652$), except one compound Temephos from the training set it is an influential observation.

With the goal of verifying the robustness of the developed model, Y-randomization test was applied. The dependent variable vector (log LC₅₀) was shuffled randomly within the training set by using 100 iterations, knowing that on every iteration a novel model will be generated.

Figure 3 shows a graphic of the statistical coefficients Q^2 and R^2 that allows comparing the outcomes of randomized models (squares) with the developed model (ring). R^2 and Q^2 values are clearly quite high for the real model. These results indicate that the model is not due to chance correlation.

Descriptor Contributions Analysis

Based on a previously described procedure, the relative contributions of the four descriptors to the model were determined as follows: *BEHm1* (39.29%) > *EEig03x* (22.99%) > *EEig11x* (21.62%) > *Mor24v* (16.09%).

As can be seen, the *BEHm1* contribution is greater than the contributions of *EEig03x*, *EEig11x*, and *Mor24v*, while the difference in descriptor contribution is not significant, implying that the *BEHm1* descriptor is more important in generating the predictive model than the other three descriptors, as shown in Fig. 4.

Support vector machine

The quality of SVM for regression depends on several parameters namely, kernel type k , which determines the sample distribution in the mapping space, and its corresponding parameter σ , capacity parameter C , and ϵ -insensitive loss function. SVM parameters were optimized by systematically adjusting their values in the training set and using 5-fold cross-validation to calculate the model's RMSE.

First, the value of C fixed and the epsilon value and σ value varied, a minimal RMSE corresponds to optimal values of ϵ and σ . Once ϵ and σ are optimized, the regularization parameter C that controls the trade-off between maximizing the margin and minimizing the training error be optimized. To find an optimal value of C , the RMSE of SVM models with different C values was calculated. The variation of RMSE versus C values is plotted in Fig. 5.

As shown in this figure, the optimal value of C was 44.45. The optimal values of the three parameters and the final optimal model were determined in Table 3.

Figure 6 shows predicted log LC₅₀ against experimental values (Table 4). The predicted values are, in general, in good agreement with the corresponding experimental values.

Table 3
Results and statistical parameters of GA- SVM.

Parameters	C	ϵ	σ	R^2	Q^2_{LOO}	RMSE	Q^2_{EXT}	RMSE _{EXT}
GA- SVM	44.45	0.009	0.09	0.9778	0.9450	0.1516	0.9928	0.0377

The model established a strong correlation between the four selected variables and the studied property, characterized by excellent parameters $Q^2_{\text{LOO}} > 0.5$. [28] In addition to a good standard error (RMSE) All statistical parameters of the model are satisfying and prove that the model is stable, robust and predictive.

Comparative Results Of Mlr And Svm

The comparative results of the MLR and SVM models are showed in Fig. 7. These results demonstrate that the SVM model was more powerful than the MLR model because it had a higher statistical quality and a lower prediction error.

Conclusion

In this study, linear and nonlinear QSAR models were developed to predict the larvicidal activity (log LC₅₀) for a dataset of 30 compounds against *Culex pipiens pallens* by using MLR and SVM methods. The built models clearly demonstrate good correlations between the structure and larvicidal activity of the studied compounds. A four descriptors linear model was developed by MLR, with R^2 of 0.9716 and RMSE of the 0.1690, SVM with R^2 of 0.9778, RMSE of 0.1516 for training set.

The suggested model's reliability was demonstrated using validation techniques, such as leave-one-out cross validation, randomization testing, and validation through the test set. The performance of MLR and SVM were compared in this QSAR study. The obtained results show that SVM seems to be the best way to relate molecular descriptors to larvicidal activity and to derive statistical models with better quality and better generalization capabilities than the linear regression method. The optimization process of SVM is relatively easy to be implemented. They can be used as alternative non-linear modeling tools in QSAR. Our QSAR model with simply calculated molecular descriptors could be employed to estimate the larvicidal activity for new compounds with unknown activities to avoid testing.

Experimental

Data Set

The larvicidal activity values were previously obtained from the literature (LC_{50}), (Perumalsamy *et al.*, 2009) [29], which represents the concentration at which 50% of third-instar larvae of *Culex pipiens pallens* show lethal effects after 24h of treatment. For QSAR procedure, The 30 LC_{50} values were expressed in ppm and were converted to a logarithmic scale ($LC_{50} = \log LC_{50}$) to reduce standard deviation.

The data set listed in Table 4 was divided into two molecular subsets using the Kennard and Stone algorithm [30], the training set of 23 compounds and 7 compounds for the prediction set.

Softwares

The chemical structures were geometry-optimized in HyperChem 6.03 [31]. Molecular descriptors were calculated using Dragon software [32], MLR regression and variable subset selection were performed by the software MobyDigs [33]. Then SVM regression was realized in the Molegro Data Modeller [34].

Molecular Optimization And Descriptors Generation

The chemical structure of each compound was sketched on a PC using the HYPERCHEM program [31] and pre optimized using MM⁺ molecular mechanics method (Polack-Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by the semi empirical PM3 method at a restricted Hartree-Fock level with no configuration interaction applying a gradient norm limit of $0.001 \text{ kcal} \cdot \text{\AA}^{-1} \cdot \text{mol}^{-1}$ as a stopping criterion.

The Hyperchem output files were imported into the Dragon software [32], aiming the calculation of 3224 molecular descriptors based on the geometrical and electronic structure of the molecules. In a pre-reduction step, constant values and descriptors that were found to be correlated pairwise were eliminated (when there was more than 96 percent pairwise correlation, one variable was deleted) [35]. Only 393 descriptors remain from this procedure.

Table 4

The data set and the corresponding observed and predicted values of log LC₅₀ by MLR and SVM, Hat values and standardized residual of prediction for the training and test sets.

ID	Object	log LC ₅₀ (Exp)	log LC ₅₀ (Pred)		Hat (h_{ij})	e_{istd}
			MLR	SVM		
1	Beta-Asarone	1.3499	1.3601	1.3665	0.090	0.0563
2	Borneol	1.9617	1.8214	1.9444	0.266	-0.8582
3	Camphene	1.8479	2.0109	1.8662	0.139	0.9208
4	Delta-3-Carene	1.1408	0.8600	1.1226	0.271	-1.7241
5	Beta-Caryophyllene	1.9715	1.9457	1.9899	0.315	-0.1631
6	1,8-Cineole	1.8975	1.9002	1.8311	0.087	0.0145
7	3,5-Dimethoxytoluene	1.7413	1.8710	1.9334	0.136	0.7318
8	Estragole	1.7321	1.5732	1.6609	0.064	-0.8612
9	Eucarvone	2.0716	2.1529	2.0881	0.107	0.4511
10	Fenchene	1.8584	1.8586	1.8754	0.236	0.0014
11	(+)-Limonene	1.1225	1.5441	1.4131	0.178	2.4371
12	Linalool	1.9768	1.6449	1.7811	0.288	-2.0625
13	Methyleugenol	1.7267	1.6233	1.6700	0.049	-0.5559
14	Myrcene	1.8214	1.8040	1.8379	0.297	-0.1091
15	Myristicin	1.8864	1.4993	1.683	0.124	-2.1684
16	Pentadecane	1.9894	2.0775	1.973	0.387	0.5900
17	(+)-beta-Pinene	1.3247	1.6359	1.6075	0.125	1.7435
18	Safrole	0.9149	1.3139	1.3269	0.085	2.1856
19	Terpinen-4-ol	1.7657	1.6476	1.7828	0.148	-0.6707
20	3,4,5-Trimethoxytoluene	1.8737	1.7065	1.8571	0.147	-0.9485
21	Verbenone	1.9824	2.0001	1.1451	0.188	0.1028
22	Fenthion	-1.5376	-1.5599	1.7366	0.571	-0.1784
23	Temephos	-1.7959	-1.719	1.7483	0.702	0.7388
24 ^a	Alpha-Phellandrene	1.1411	1.3912	1.1152	0.14	1.4137

^a Test compounds

ID	Object	log LC ₅₀ (Exp)	log LC ₅₀ (Pred)		Hat (h_{ij})	e_{istd}
			MLR	SVM		
25 ^a	(+)-alpha-Pinene	1.7321	1.4889	1.106	0.119	-1.3584
26 ^a	(-)-alpha-Pinene	1.8473	1.4878	2.013	0.119	-2.0073
27 ^a	(-)-beta-Pinene	1.1096	1.5972	1.0695	0.121	2.7259
28 ^a	Gamma-Terpinene	1.1017	1.2682	1.1451	0.171	0.9588
29 ^a	Alpha-Terpineol	2.0173	2.0628	1.7366	0.083	0.2489
30 ^a	Terpinolene	1.0737	1.0761	1.7483	0.255	0.0147
^a Test compounds						

Model Development And Validation

A quantitative structure-activity relationship (QSAR) model should incorporate acceptable metrics of quality of fit, robustness, and predictability, according to the standards of the Organization for Cooperation and Economic Development (OECD). A learning set is used to determine a model's internal performance, and an appropriate test set is used to determine its predictivity [36].

To build an effective QSAR model, reliable and consistent data were needed. The genetic algorithm (GA) of the MobyDigs software [33] is used to choose significant descriptors. GA is a stochastic optimization method that mimics the evolution process by manipulating a collection of data structures [37]. It has been used for the selection of characteristics in QSAR studies [38].

In this study, the cross-validation value leave-one-out (LOO) is the optimized parameter. The GA-MLR model for the training set is obtained using the MobyDigs software. It is a model with various numbers of descriptors were examined. Because of the huge number of descriptors and varying multicollinearity, the created models were checked for over fitting. The possible multicollinearity among the selected descriptors is avoided by applying the *Q* Under the Influence of Krule (QUIK) [39].

The main objective of any QSAR study is to get a model with the best predictive and generalization abilities. In order to attest the prediction capability of the constructed QSAR models, two main methods of validation are used (internal validation and external validation).

Thus, several regularly used statistical terms: correlation coefficient (R^2) (Eq. (2)), leave-one-out (LOO) cross-validated Q^2_{LOO} (Eq. (3)), and root mean squared error (RMSE) (Eq. (4)), are used to assess the proposed model's dependability, robustness and stability

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

2

$$Q_{\text{LOO}}^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_{i/i} - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

3

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 1}}$$

4

Also, leave-many-out (LMO) and Y scrambling techniques are employed. Leave-many-out is a more powerful technique than LOO to avoid over estimating and to verify the predictive ability and stability of a model. Here, LMO is repeated 5000 times with 5 of the objects left out at each time. Then the mean value of Q_{LMO}^2 is reported.

The predictive power of QSAR model can be judged by the external statistical validation by mean of the Q_{EXT}^2 defined as follows:

$$Q_{\text{Ext}}^2 = 1 - \frac{\sum_{i=1}^{n_{\text{ext}}} (\hat{y}_{i/i} - y_i)^2 / n_{\text{ext}}}{\sum_{i=1}^{n_{\text{tr}}} (y_i - \bar{y}_{\text{tr}})^2 / n_{\text{tr}}}$$

5

An additional external validation according to Golbraikh and Tropsha [40] is applied solely to the test set. According to the recommended criteria of Tropsha *et al.*, a predictive QSAR model, must attend the following conditions: [41]

$$Q_{\text{EXT}}^2 > 0.5 \quad (6)$$

$$R^2 > 0.6 \quad (7)$$

$$(R^2 - R_0^2) / R^2 < 0.6 \text{ and } 0.85 < k < 1.15 \quad (8)$$

or

$$(R^2 - R^2_0) / R^2 < 0.6 \text{ and } 0.85 < k' < 1.15 \text{ (9)}$$

where

$$R = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad R^2_0 = 1 - \frac{\sum (y_i - y_i^{r_0})^2}{\sum (y_i - \bar{y})^2} \quad R^2_0 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum (\tilde{y}_i - \bar{\tilde{y}})^2} \quad k = \frac{\sum (y_i \tilde{y}_i)}{\sum (\tilde{y}_i)^2}$$

$$k' = \frac{\sum (y_i \tilde{y}_i)}{\sum (y_i)^2}$$

where R is the correlation coefficient between the calculated and experimental values in the test set; R^2_0 (calculated versus observed values) and R^2_0 (observed versus calculated values) are the coefficients of determination; k and k' are slopes of regression lines through the origin of calculated versus observed and observed versus calculated, respectively; $y_i^{r_0}$ and $\tilde{y}_i^{r_0}$ are defined as:

$$y_i^{r_0} = k \tilde{y}_i$$

10

and

$$\tilde{y}_i^{r_0} = k' y_i$$

11

respectively; and the summations are over all samples in the test set.

The reason to use R^2_0 and require k values that are close to 1 is that when actual *versus* predicted properties are compared, an exact fit is required, not just a correlation.

A randomization test is used to rule out the possibility of increasing the correlation by chance and permutation testing is used to check for reliability and robustness. The novel models are recalculated for randomly reordered responses (Y scrambling).

Support Vector Machine

SVM is a new and very promising classification and regression method developed by Vapnik *et al.*, [42, 43] SVMs were originally developed for classification problems; they can also be extended to solve nonlinear regression problems by the introduction of ε -insensitive loss function. In support vector regression, the input x is first mapped into a higher dimensional feature space by the use of a kernel function, and then a linear model is constructed in this feature space. The kernel functions often used in SVM include linear, polynomial, radial basis function, and sigmoid function. The general form of the SVR-based regression function can be written as below: [44, 45]

$$f(x, w) = f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^N (\alpha_i - \alpha_i^*) K(x_i, x_j) + b$$

12

where both α_i and α_i^* are Lagrange multipliers. According to the Karush- Kuhn- Tucker conditions only the minority sample coefficients are non- zero values, the data points corresponding with them are called support vectors. These support vectors are the samples which can determine the hyper plane. [46, 47] $K(x, x_j)$ is the Kernel function [48]. Any function satisfying Mercer's condition can be used as the Kernel function. [49] In this work, the Gaussian radial basis function (RBF) Kernel was used in the SVM as below:

$$K(x, x_i) = \exp \left(\frac{-\|x - x_i\|^2}{\sigma^2} \right)$$

13

Where σ^2 is the width of the Gaussian function, so the C and σ that are the relative weights of the regression error and the kernel parameter of the RBF kernel should be optimized by the user, to obtain the support vector. The parameters of SVM were optimized by systemically changing their values in the training step and calculating the RMSE and accuracy of the model using 5-fold cross-validation. The optimized values of C , σ , and ε were 44.45, 0.09, and 0.009 respectively obtained based on minimum RMSE and maximum accuracy of model.

A detailed description of the theory of SVM can be referred in several excellent books and tutorials [50, 51].

Applicability Domain Analysis

The applicability domain (AD) [52, 53] is a theoretical region in the space defined by the descriptors of the model and the modeled response, for which a given QSAR should make reliable predictions. In this work, the structural AD was verified by the leverage approach. The leverage h_i [54] is defined as follows:

$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \quad (i = 1; \dots; n) \quad (13)$$

Where $\mathbf{x}_i$ is the descriptor row vector of the i^{th} compound, $\mathbf{x}_i^T$ is the transpose of $\mathbf{x}_i$, $\mathbf{X}$ is the model matrix derived from the calibration set descriptor values.

The warning leverage h^* is, generally, fixed at $3(m+1)/n$, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation. [55]

Declarations

Acknowledgements. The authors would like to thank The General Directorate for Scientific Research and Technological Development (DG-RSDT), the Algerian Ministry of Scientific Research.

Conflict of interest. The authors declare no competing interests.

References

1. Manimaran, A. , Cruz, M. , Muthu, C. , Vincent, S., & Ignacimuthu, S. (2012) Larvicidal and knockdown effects of some essential oils against *Culex quinquefasciatus* Say, *Aedes aegypti* (L.) and *Anopheles stephensi* (Liston). *Advances in Bioscience and Biotechnology*, 3, 855-862. (<https://doi.org/10.4236/abb.2012.37106>)
2. Rueda, L. M. (2007). Global diversity of mosquitoes (Insecta: Diptera: Culicidae) in freshwater. In *Freshwater Animal Diversity Assessment*. Developments in Hydrobiology, vol 198. pp. 477-487. Springer, Dordrecht. (https://doi.org/10.1007/978-1-4020-8259-7_48)
3. Das, K., Vasudeva, C., & Dang, R. (2016). Economical novel formulation and evaluation of herbal oils for mosquito and house fly repellent activities. *Annals of Phytomedicine-An International Journal*, 5, 91-6. (<https://doi.org/10.21276/AP.2016.5.2.11>)
4. Tandina, F., Doumbo, O., Traoré, S. F., Parola, P., & Robert, V. (2018). Mosquitoes (Diptera: Culicidae) and mosquito-borne diseases in Mali, West Africa. *Parasites & vectors*, 11(1), 1-12. (<https://doi.org/10.1186/s13071-018-3045-8>)
5. Reegan, A. D., Kumar, P. S., Asharaja, A. C., Devi, C., Jameela, S., Balakrishna, K., & Ignacimuthu, S. (2021). Larvicidal and ovicidal activities of phenyl acetic acid isolated from *Streptomyces collinus* against *Culex quinquefasciatus* Say and *Aedes aegypti* L.(Diptera: Culicidae). *Experimental Parasitology*, 226. (<https://doi.org/10.1016/j.exppara.2021.108120>)
6. Pavela, R. (2008). Larvicidal effects of various Euro-Asiatic plants against *Culex quinquefasciatus* Say larvae (Diptera: Culicidae). *Parasitology research*, 102(3), 555-559. (<https://doi.org/10.1007/s00436-007-0821-3>)
7. Zoubiri, S., & Baaliouamer, A. (2014). Potentiality of plants as source of insecticide principles. *Journal of Saudi Chemical Society*, 18(6), 925-938. (<https://doi.org/10.1016/j.jscs.2011.11.015>)
8. Manjari, M. S., Karthi, S., Ramkumar, G., Muthusamy, R., Natarajan, D., & Shivakumar, M. S. (2014). Chemical composition and larvicidal activity of plant extracts from *Clausena dentata* (Willd) (Rutaceae) against dengue, malaria, and filariasis vectors. *Parasitology research*, 113(7), 2475-2481. (<https://doi.org/10.1007/s00436-014-3896-7>)
9. Lucia, A., Girard, C., Fanucce, M., Coviella, C., Rubio, R. G., Ortega, F., & Guzmán, E. (2020). Development of an environmentally friendly larvicidal formulation based on essential oil compound blend to control *Aedes aegypti* larvae: Correlations between physicochemical properties and insecticidal activity. *ACS Sustainable Chemistry & Engineering*, 8(29), 10995-11006. (<https://doi.org/10.1021/acssuschemeng.0c03778>)
10. Dias, C. N., & Moraes, D. F. C. (2014). Essential oils and their compounds as *Aedes aegypti* L.(Diptera: Culicidae) larvicides. *Parasitology research*, 113(2), 565-592. (<https://doi.org/10.1007/s00436-013-3687-6>)

11. Amer, A., & Mehlhorn, H. (2006). Larvicidal effects of various essential oils against *Aedes*, *Anopheles*, and *Culex* larvae (Diptera, Culicidae). *Parasitology Research*, 99(4), 466-472. (<https://doi.org/10.1007/s00436-006-0182-3>)
12. Lucia, A., Audino, P. G., Seccacini, E., Licastro, S., Zerba, E., & Masuh, H. (2007). Larvicidal effect of *Eucalyptus grandis* essential oil and turpentine and their major components on *Aedes aegypti* larvae. *Journal of the American Mosquito Control Association*, 23(3), 299-303. ([https://doi.org/10.2987/8756-971x\(2007\)23\[299:leoege\]2.0.co;2](https://doi.org/10.2987/8756-971x(2007)23[299:leoege]2.0.co;2))
13. Manh, H. D., & Tuyet, O. T. (2020). Larvicidal and repellent activity of *Mentha arvensis* L. essential oil against *Aedes aegypti*. *Insects*, 11(3), 198. (<https://doi.org/10.3390%2Finsects11030198>)
14. Kumar, S., Wahab, N., & Warikoo, R. (2011). Bioefficacy of *Mentha piperita* essential oil against dengue fever mosquito *Aedes aegypti* L. *Asian Pacific journal of tropical biomedicine*, 1(2), 85-88. ([https://doi.org/10.1016%2FS2221-1691\(11\)60001-4](https://doi.org/10.1016%2FS2221-1691(11)60001-4))
15. Pavela, R., Kaffkova, K., & KUMŠTA, M. (2014). Chemical composition and larvicidal activity of essential oils from different *Mentha* L. and *Pulegium* species against *Culex quinquefasciatus* say (Diptera: Culicidae). *Plant Protection Science*, 50(1), 36-42. (<https://doi.org/10.17221/48%2F2013-PPS>)
16. Cetin, H., & Yanikoglu, A. (2006). A study of the larvicidal activity of *Origanum* (Labiatae) species from southwest Turkey. *Journal of Vector ecology*, 31(1), 118-122. ([https://doi.org/10.3376/1081-1710\(2006\)31\[118:asotla\]2.0.co;2](https://doi.org/10.3376/1081-1710(2006)31[118:asotla]2.0.co;2))
17. Quiroz-Martínez, H., & Rodríguez-Castro, A. (2007). Aquatic insects as predators of mosquito larvae. *Journal of the American mosquito control association*, 23(sp2), 110-117. ([https://doi.org/10.2987/8756-971x\(2007\)23\[110:aiapom\]2.0.co;2](https://doi.org/10.2987/8756-971x(2007)23[110:aiapom]2.0.co;2))
18. Ciccia, G., Coussio, J., & Mongelli, E. (2000). Insecticidal activity against *Aedes aegypti* larvae of some medicinal South American plants. *Journal of Ethnopharmacology*, 72(1-2), 185-189. ([https://doi.org/10.1016/S0378-8741\(00\)00241-5](https://doi.org/10.1016/S0378-8741(00)00241-5))
19. Saavedra, L. M., Romanelli, G. P., & Duchowicz, P. R. (2018). Quantitative structure–activity relationship (QSAR) analysis of plant-derived compounds with larvicidal activity against Zika *Aedes aegypti* (Diptera: Culicidae) vector using freely available descriptors. *Pest management science*, 74(7), 1608-1615. (<https://doi.org/10.1002/ps.4850>)
20. Devillers, J., Doucet-Panaye, A., & Doucet, J. P. (2015). Structure–activity relationship (SAR) modelling of mosquito larvicides. *SAR and QSAR in Environmental Research*, 26(4), 263-278. (<https://doi.org/10.1080/1062936x.2015.1026571>)
21. De, P., Aher, R. B., & Roy, K. (2018). Chemometric modeling of larvicidal activity of plant derived compounds against zika virus vector *Aedes aegypti*: application of ETA indices. *RSC advances*, 8(9), 4662-4670. (<https://doi.org/10.1039/C7RA13159C>)
22. Javidfar, M., & Ahmadi, S. (2020). QSAR modelling of larvicidal phytocompounds against *Aedes aegypti* using index of ideality of correlation. *SAR and QSAR in Environmental Research*, 31(10), 717-739. (<https://doi.org/10.1080/1062936X.2020.1806922>)
23. Winkler, D. A. (2002). The role of quantitative structure-activity relationships (QSAR) in biomolecular discovery. *Briefings in bioinformatics*, 3(1), 73-86. (<https://doi.org/10.1093/bib/3.1.73>)

24. Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., & Tropsha, A. (2014). QSAR modeling: where have you been? Where are you going to?. *Journal of medicinal chemistry*, 57(12), 4977-5010. (<https://doi.org/10.1021/jm4004285>)
25. Alencar Filho, E. B., Castro Silva, J. W., & Cavalcanti, S. C. (2016). Quantitative structure-toxicity relationships and molecular highlights about *Aedes aegypti* larvicidal activity of monoterpenes and related compounds. *Medicinal Chemistry Research*, 25(10), 2171-2178. (<http://dx.doi.org/10.1007%2Fs00044-016-1650-7>)
26. Roy, K., Kar, S., & Das, R. N. (2015). QSAR/QSPR modeling: introduction. In *A primer on QSAR/QSPR modeling* (pp. 1-36). Springer, Cham. (https://doi.org/10.1007/978-3-319-17281-1_1)
27. Hansch, C., & Verma, R. P. (2009). Larvicidal activities of some organotin compounds on mosquito larvae: A QSAR study. *European journal of medicinal chemistry*, 44(1), 260-273. (<https://doi.org/10.1016/j.ejmech.2008.02.040>)
28. Bouakkadia, A., Driouche, Y., Kertiou, N., & Messadi, D. (2020). Modeling of the Henry constant of a series of pesticides: Quantitative structure-property relationship approach. *International Journal of Safety and Security Engineering*, 10(3), pp. 389- 396. (<https://doi.org/10.18280/IJSSE.100311>).
29. Perumalsamy, H., Kim, N. J., & Ahn, Y. J. (2009). Larvicidal activity of compounds isolated from *Asarum heterotropoides* against *Culex pipiens pallens*, *Aedes aegypti*, and *Ochlerotatus togoi* (Diptera: Culicidae). *Journal of Medical Entomology*, 46(6), 1420-1423.
30. Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, 11(1), 137-148. (<https://doi.org/10.1080/00401706.1969.10490666>).
31. HyperchemTM. Release 6.03 for windows. Molecular Modeling system, (2000) (<http://www.hyper.com/>).
32. Talete srl, DRAGON (Software for Molecular Descriptors calculation) version 5.5, (2007) (<http://www.talete.mi.it/>).
33. Todeschini, R. Ballabio, D. Consonni, V. Mauri, A., & Paven, M. (2009). MobyDigs - version 1.1 - Copyright TALETE srl, 2004 (<http://www.talete.mi.it/>).
34. Molegro Data Modeller (MDM), v.2.0. Copyright Molegro (2009) (<https://www.scientific-computing.com/press-releases/molegro-data-modeller-v20>).
35. Didi, M., Haddag, H., Driouche, Y., & Messadi, D. (2017). Modeling and prediction of flash point of unsaturated hydrocarbons using hybrid genetic algorithm/multiple linear regression approach. *Research journal of pharmaceutical biological and chemical sciences*, 8(4), 379-390. (https://www.rjpbcs.com/2017_8.4.html).
36. Organisation for Economic Co-operation and Development. (2014). *Guidance document on the validation of (quantitative) structure-activity relationship [(Q) SAR] models*. Organisation for Economic Co-operation and Development.
37. Leardi, R., Boggia, R., & Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of chemometrics*, 6(5), 267-281.

38. Liu, P., & Long, W. (2009). Current mathematical methods used in QSAR/QSPR studies. *International Journal of Molecular Sciences*, 10(5), 1978-1998.
39. Todeschini, R., Consonni, V., & Maiocchi, A. (1999). The K correlation index: theory development and its application in chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 46(1), 13-29.
40. Golbraikh, A., & Tropsha, A. (2002). Beware of q²!. *Journal of molecular graphics and modelling*, 20(4), 269-276. ([https://doi.org/10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1)).
41. Kertiou, N. E., Bouakkadia, A., & Messadi, D. (2017). QSPR study of the boiling point of diverse hydrocarbons: hybrid (GA/MLR) approach. *Research journal of pharmaceutical biological and chemical sciences*, 8(6), 251-265. ([https://www.rjpbcs.com/pdf/2017_8\(6\)/\[29\].pdf](https://www.rjpbcs.com/pdf/2017_8(6)/[29].pdf)).
42. Vapnik, V. N. (1998). Statistical Learning Theory, John Wiley & Sons, New York, pp. 375–473 (ISBN: 978-0-471-03003-4).
43. Bouakkadia, A., Kertiou, N., Amiri, R., Driouche, Y., & Messadi, D. (2021). Use of GA-ANN and GA-SVM for a QSPR study on the aqueous solubility of pesticides. *Journal of the Serbian Chemical Society*, 86(7-8), 673-684. <https://doi.org/10.2298/JSC200618066B>.
44. Vapnik, V. N. (2000). The Nature of Statistical Learning Theory. Springer, New York, pp. 267–287 (ISBN: 978-1-4757-3264-1).
45. Lu, C. J., Lee, T. S., & Chiu, C. C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision support systems*, 47(2), 115-125. (<https://doi.org/10.1016/j.dss.2009.02.001>).
46. Jothilakshmi, S., & Gudivada, V. N. (2016). Handbook of Statistics, Vol. 35, Ch. 10, Elsevier, Amsterdam, pp. 301–340 (<http://dx.doi.org/10.1016/bs.host.2016.07.005>).
47. Cristianini, N., & Shawe-Taylor, J. (2004). *An introduction to support vector machines and other kernel-based learning methods*, Publishing House of Electronics Industry, Beijing, pp. 93–122.
48. Amiri, R., Messadi, D., Bouakkadia, A., & Lourici, L. (2019). QSPR Models for the prediction of octanol/water partition coefficient of organophosphorous insecticides. *Egyptian Journal of Chemistry*, 62(9), 1563-1574. (<https://doi.org/10.21608/ejchem.2019.4976.1446>).
49. Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc., pp.70–71 (ISBN: 0-387-94559-8).
50. Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*, Cambridge, UK: Cambridge Univ. Press, pp. 32–42.
51. Schölkopf, B., & Smola, A. (2001). *Learning with Kernels*, MIT Press, Cambridge, MA, pp. 13–17.
52. Bouakkadia, A., Haddag, H., Bouarra, N., & Messadi, D. (2016). QSPR study of the water solubility of a diverse set of agrochemicals: hybrid (GA/MLR) approach. *Synthèse: Revue des Sciences et de la Technologie*, 32, 12-21. (<https://www.ajol.info/index.php/srst/issue/view/13942>).
53. Deeb, O., & Goodarzi, M. (2010). Predicting the solubility of pesticide compounds in water using QSPR methods. *Molecular Physics*, 108(2), 181-192. (<https://doi.org/10.1080/00268971003604575>).
54. Wang, W., Xu, Z., Lu, W., & Zhang, X. (2003). Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing*, 55(3-4), 643-663.

([https://doi.org/10.1016/S0925-2312\(02\)00632-X](https://doi.org/10.1016/S0925-2312(02)00632-X)).

55. Driouche, Y., & Messadi, D. (2019). Quantitative structure–retention relationship model for predicting retention indices of constituents of essential oils of *Thymus vulgaris* (Lamiaceae). *Journal of the Serbian Chemical Society*, 84(4), 405-416. (<https://doi.org/10.2298/JSC180817010D>)

Figures

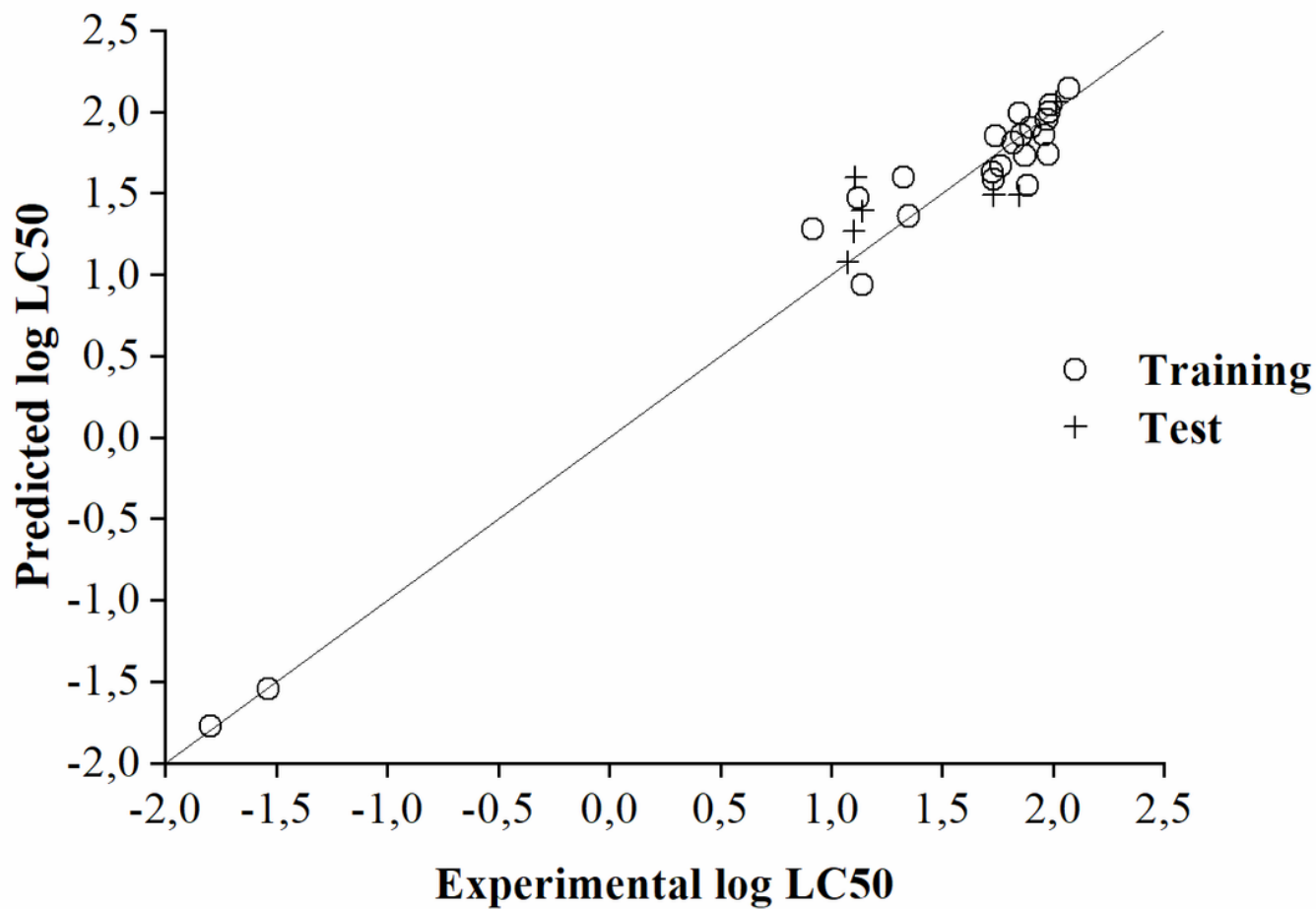


Figure 1

Predicted versus experimental values of log LC₅₀

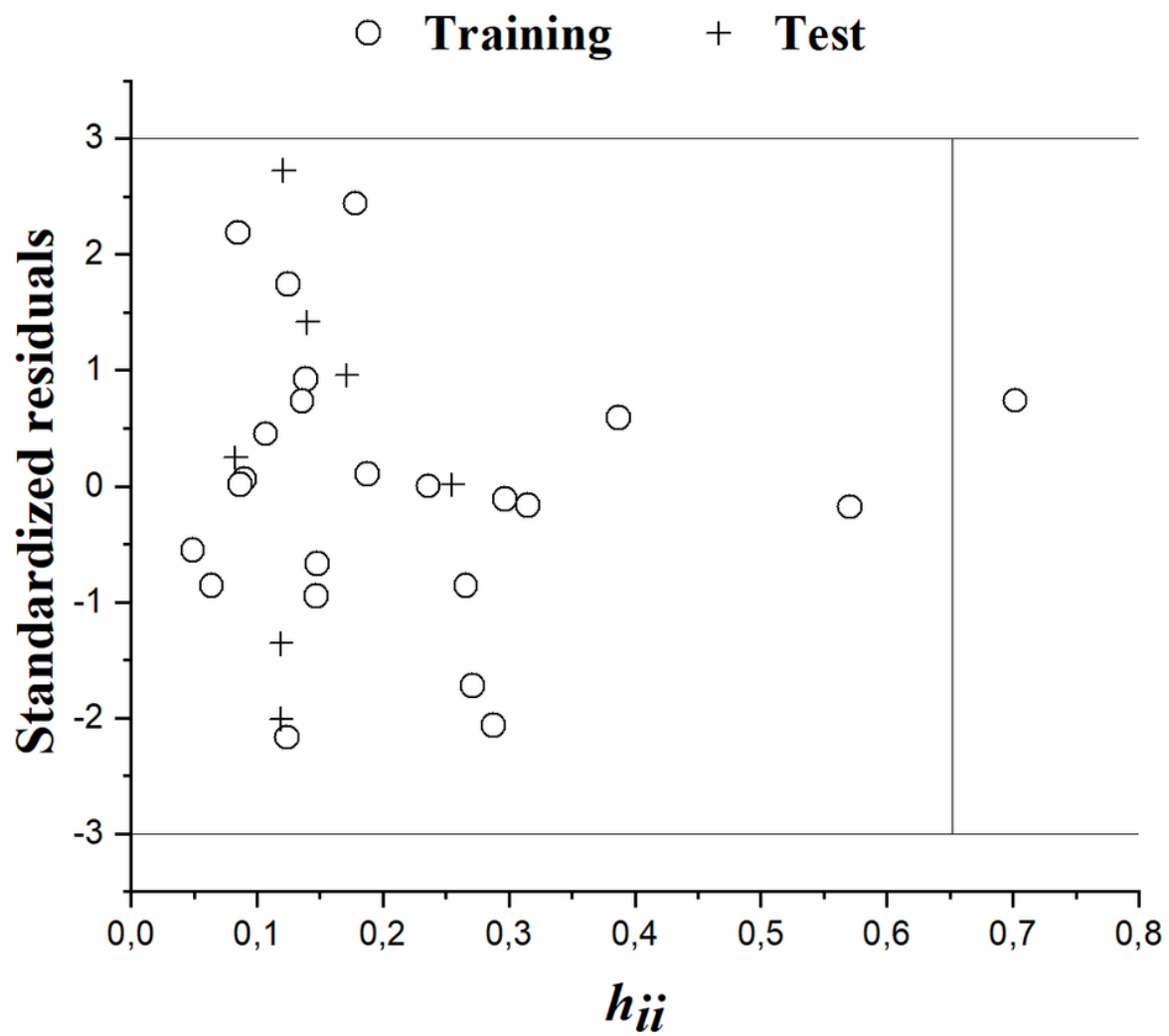


Figure 2

The Williams plot

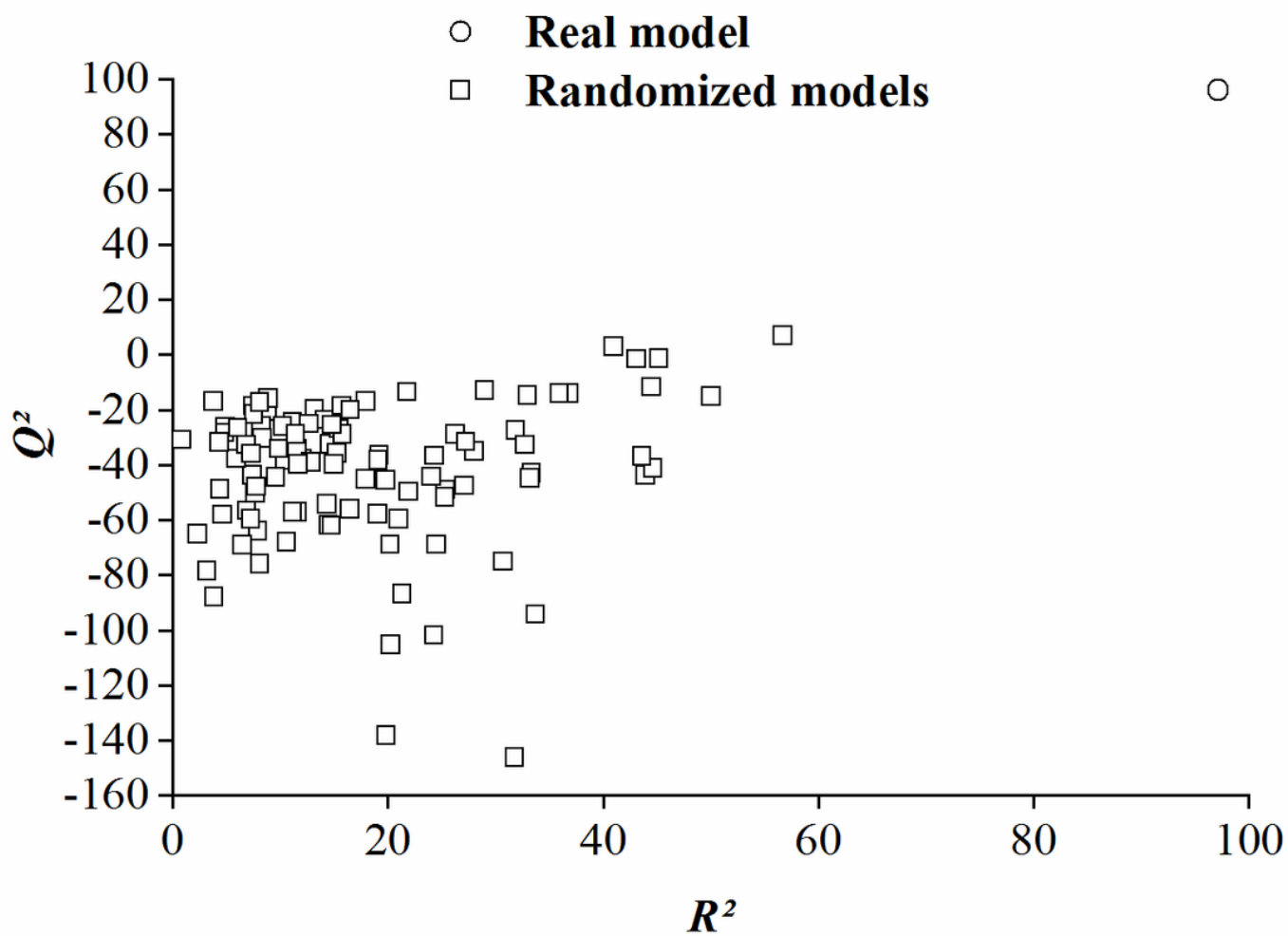


Figure 3

Randomization test associated to previous QSAR model

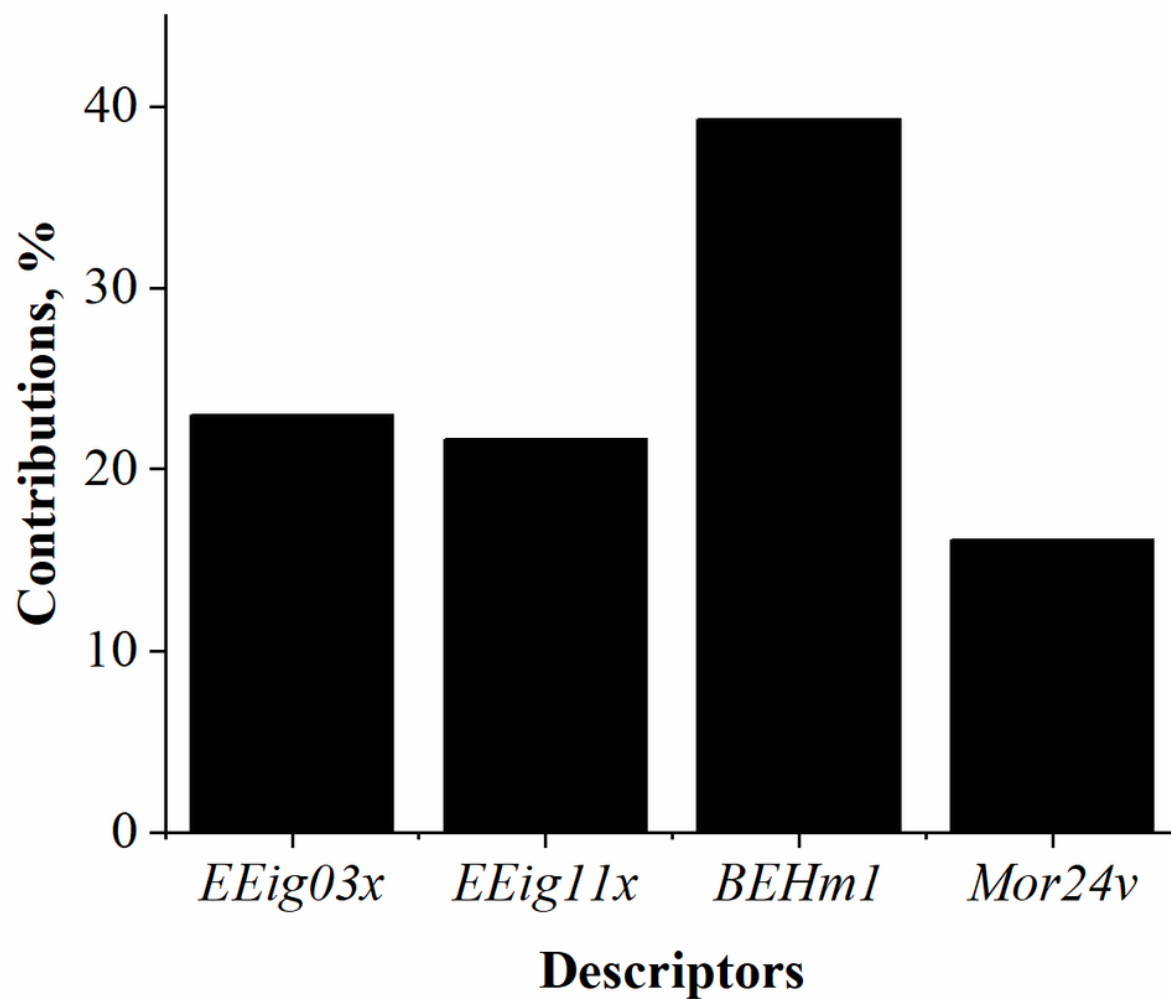


Figure 4

Relative contribution of the selected descriptors in the MLR model

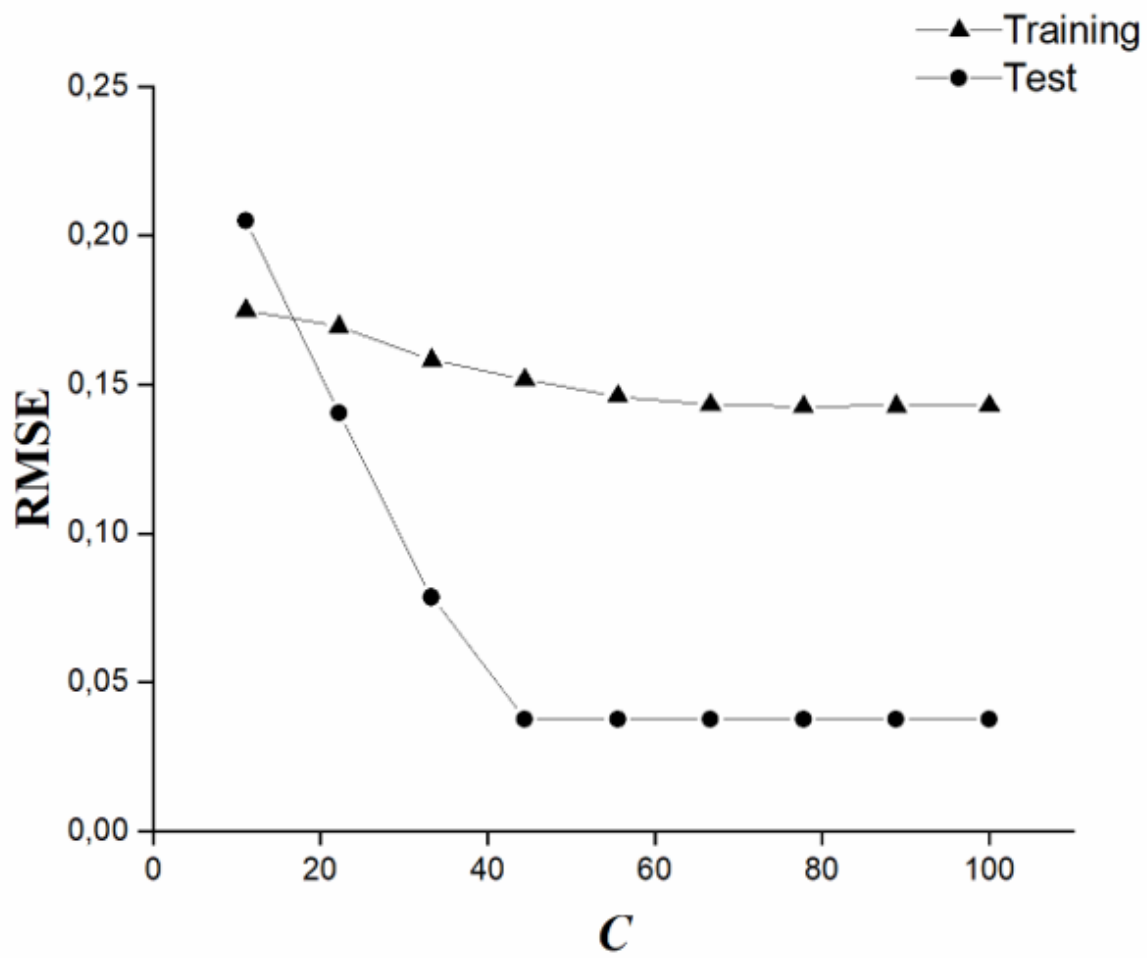


Figure 5

Variation of RMSE vs. C values

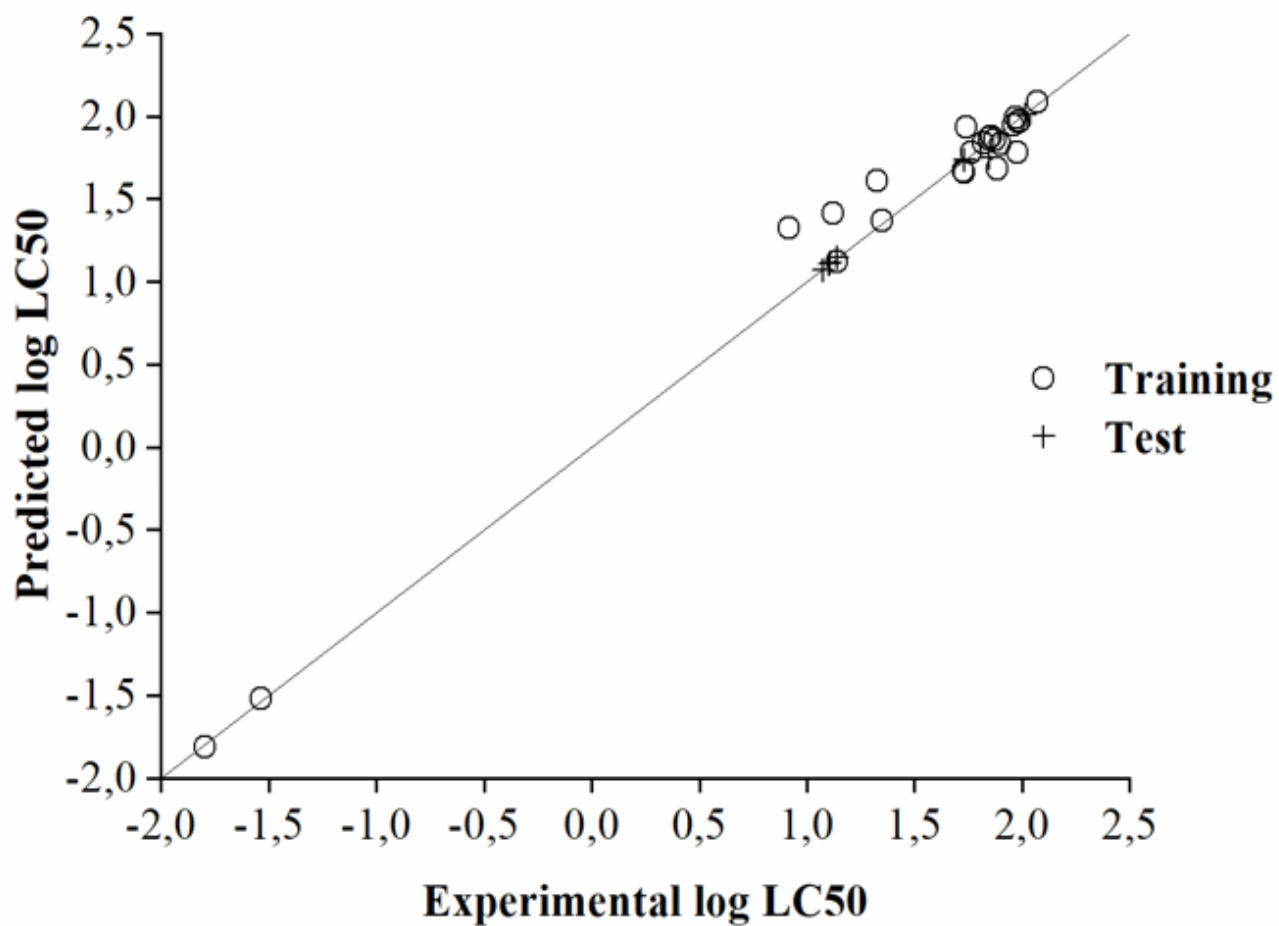


Figure 6

Predicted values vs. Experimental values for the training test sets

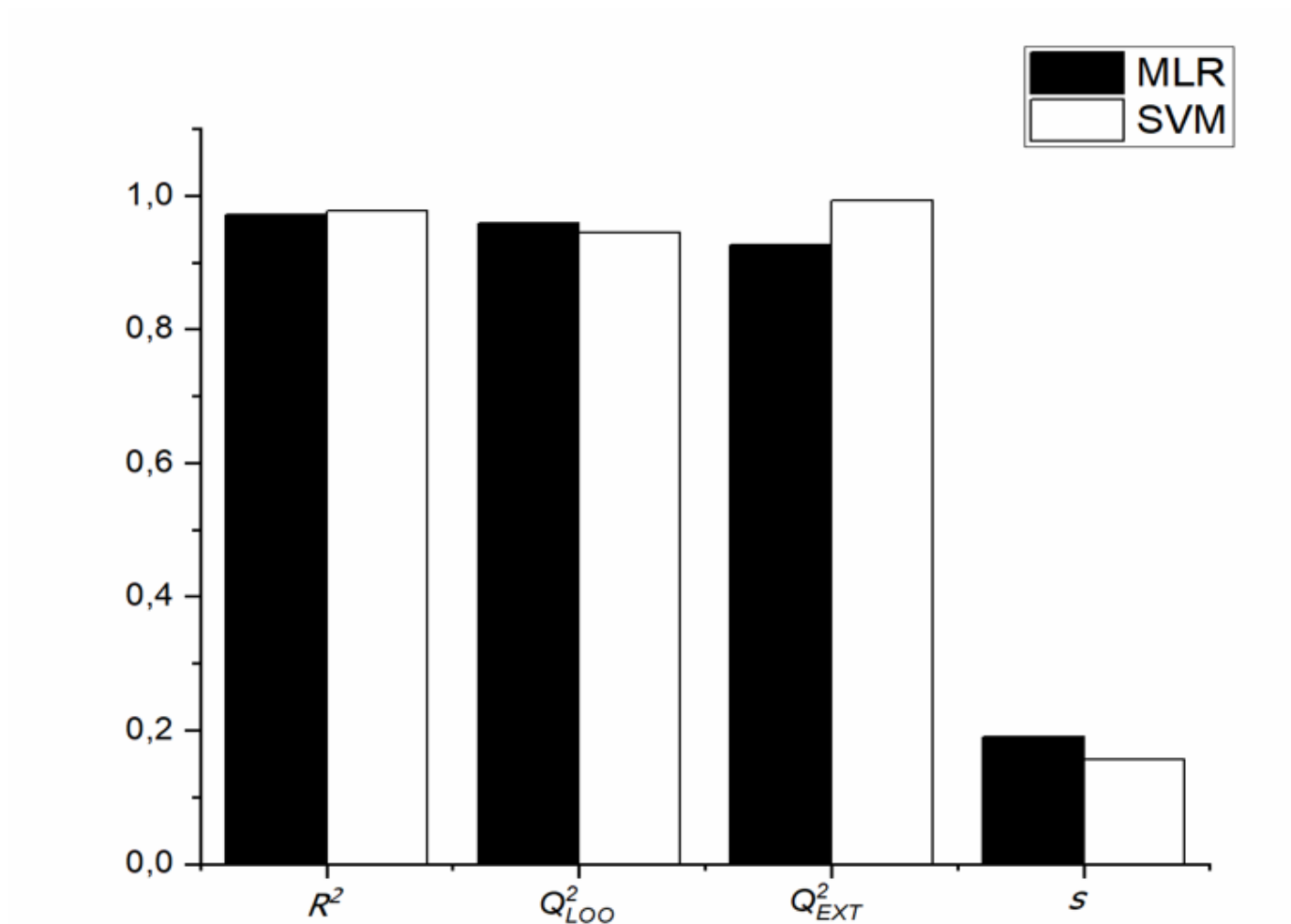


Figure 7

Graphical comparison of MLR and SVM statistics

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [GraphicalabstractSamiaAlem.docx](#)