Egypt.J.Chem. **Vol. 62**, No. 9. pp. 1563 - 1574 (2019)

Egyptian Journal of Chemistry
http://ejchem.journals.ekb.eg/

Qspr Models for The Prediction of Octanol/water Partition Coefficient of Organophosphorous Insecticides

Rana Amiri[1], Djelloul Messadi[1,*], Amel bouakkadia[2], Leila lourici[3]
[1]*Environmental and Food Safety Laboratory, Badji Mokhtar University, Annaba 23000, Algeria.*
[2]*Departement of chemistry Abbes Laghrour University, Khenchela 1252, Algeria.*
[3]*Departement of chemistry Chadli Bendjedid University, Eltaref 36000, Algeria.*

THIS STUDY aims to predict the octanol/water partition coefficient (K_{ow}) of 43 organophosphorous insecticides. Quantitative structure- property relationship analysis was performed on a series of 43 insecticides using Multiple Linear Regression (MLR) and Support Vector Machines (SVM) methods, which correlate octanol- water partition coefficient (K_{ow}) values of these chemicals to their structural descriptors. At first, the data set was separated with duplex algorithm into a training set (22 chemicals) and a test set (21 chemicals) for statistical external validation. The IX'XI ratio for the two data sets was 0.9839 indicating that the volumes of the regions covered by the two data sets were approximately the same. Model with four descriptors was developed using as independent variables theoretical descriptors derived from DRAGON software when applying GA (Genetic Algorithm)- VSS (Variable Subset Selection) procedure . The values of statistical parameters R^2, Q^2_{ext}, $SDEP_{ext}$ and SDEC for MLR and SVM model were: (93.57%; 92.73%; 0.493; 0.463), (98.60%; 96.30%; 0.504; 0.316); obtained for the two approaches are very similar, which confirm that our four parameters model is stable, robust and significant.

Keywords: Octanol/water partition coefficient, Quantitative structure-property/activity relationship, Molecular descriptor, Modeling, Support vector machine, Multiple linear regression.

Introduction

Pesticides are used globally for the protection of food, fiber, human health and comfort. Food safety is an area of growing worldwide concern on account of its direct bearing on human health. The presence of harmful pesticide residues in food and seed oils has caused a great concern among the consumers [1]. Repeated application leads to loss of biodiversity. Many pesticides are not easily degradable, they persist in soil, leach to groundwater and surface water and contaminate wide environment. Depending on their chemical properties they can enter the organism, bioaccumulate in food chains and consequently influence also human health. Overall, intensive pesticide application results in several negative effects in the environment that cannot be ignored. Organophosphous compounds are widely used as insecticides in agricultural and domestic practice, which causes their widespread appearance in the environment [2, 3]. The organophosphorus pesticides most commonly used in the cultivation of essential and fixed oil plants travel through the vascular system of the plant and are absorbed at the cellular level[4].

Fortunately, with the development of quantitative structure–activity relationships (QSARs), the toxicity of chemicals can be predicted based on the knowledge of their structures (even before the chemicals are produced), and QSARs

*Corresponding author e-mail: d_messadi@yahoo.fr

Received 6/9/2018; Accepted 1/4/2019
DOI: 10.21608/ejchem.2019.4976.1446

have proven to be reliable tools for the toxicity assessment of organic chemicals when little or no empirical data are available [2, 3].

The quantitative structure-property/activity relationship (QSPR/QSAR) method is based on the assumption that the variation of the behavior of the compounds, as expressed by many measured physicochemical properties, can be correlated with changes in molecular features of the compounds termed descriptors [5]. This method can be used for the prediction of the properties of new compounds. It can also be applied to identify and describe important structural features of the molecules that are relevant to variations in molecular properties. Computational models are useful because they rationalize a large number of experimental observations and therefore save time and money in the process of drug design [6].

Physicochemical properties of an organic chemical compound play an important role in determining its distribution and fate in the environment. Vapor pressures (Pv), aqueous solubility ($S_{w,L}$) and n-octanol/water partition coefficients (K_{ow}) are key physicochemical properties that can be used for assessing environmental partition and transport of organic substances [7].

K_{ow} is considered to be a good indicator of bioaccumulation of pesticides in organisms and food chains. Pesticides with a positive correlation to Log K_{ow} are more likely to have bioaccumulation effects to organisms and food chains. The paramter is also a good indicator of systemic mode of action of a pesticide.

Properties such as the n-octanol–water partition coefficient (K_{OW}) are important in predicting the environmental fate of organic compounds [8]. Further, the K_{OW} is used as one of the molecular descriptors of the toxic effects of chemicals in quantitative Structure Activity Relationships (QSAR) [9-11]. The partition coefficient (P) is defined as the ratio of the equilibrium concentrations of a dissolved substance in a two-phase system consisting of two largely immiscible solvents, in this case n-octanol and water. Octanol represents a substitute for biotic lipid and hence gives an approximation to a biotic lipid-water partition coefficient [12].

The objective of this study will be to divide the data into two sets which cover approximately the same region and have similar statistical properties. The DUPLEX algorithm assists in accomplishing this objective.

This study was to develop QSAR models to describe the acute toxicity of pesticides and to found a statistical model for the prediction of the n-octanol/water partition coefficient (K_{ow}) of organophosphorous compounds. For this purpose the relationship between molecular descriptors [13] connected to the factors found, experimentally, as affecting the K_{ow} of the compounds was searched. The QSPR model was constructed using Multiple Linear Regression (MLR) and Support Vector machine (SVM). The model obtained shows which descriptors play a significant role in the K_{ow} variation of these pesticides.

Materials and Methods

The 43 organophosphorous insecticides are taken from Hasen [14] and are listed in Table 1. The data were presented as the logarithm of K_{ow} to reduce the range of variation. The data set is separated into a training set of 22 compounds and a test set of 21 compounds.

The chemical structure of each compound was sketched on a PC using the HYPERCHEM program [15] and preoptimized 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 0.01 kcal.Ǻ$^{-1}$. mol-^{1} as a stopping criterion. Then the geometries were used as input for the generation of 1664 descriptors using the Dragon software (version 5.4) [16]. Quantum-chemical descriptors such as HOMO (highest occupied molecular orbital), LUMO (lowest unoccupied molecular orbital), HOMO– LUMO gap (DHL), and ionization potential (Pion), calculated by the semi empirical PM3 method using [15], were added and used for descriptor selection during model development. Constant values and descriptors found to be correlated pairwise were excluded in a prereduction step (when there was more than 98% pairwise correlation, one variable was deleted), and the genetic algorithm was applied for variables selection to a final set of 1230 descriptors.

It is important to rationally define a training set from which the model is built and external test set on which to evaluate its prediction power. The object of this selection should be to generate two sets with similar molecular diversity, in order to be reciprocally representative and to cover all the main structural and physicochemical characteristics of the global data set. Several procedures can be adopted for the selection of the training and test sets, the later should contain between 15 and 40% of the compounds in the full data set.

In this work we used Duplex algorithm [17] to split, arbitrarily, the whole data into a 22 samples training set and a 21samples testing set.

The algorithm begin with a list of the n (=43) observations where the k regressors are standardized to unit length; that is,

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{S_{jj}^{1/2}}, i = 1, 2, ..., n; j = 1, 2, ..., k \quad (1)$$

Where $S_{jj}^{1/2} = \sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2$ is the corrected sum of squares of the jth regressor. The standardized regressors are then orthonormalized. This can be done by factoring the **Z'Z** matrix as:

$$\mathbf{Z'Z=T'T} \quad (2)$$

Where $\underset{\sim}{\mathbf{T}}$**'** is unique **k×k** upper triangular matrix. The elements of $\underset{\sim}{\mathbf{T}}$ can be found using the square root or cholesky method [18]. Then make the transformation

$$\mathbf{W=ZT^{-1}} \quad (3)$$

Resulting in a new set of variables (the **w's**) that are orthogonal and have unit variance. Then the Euclidian distance between all possible pairs of points is calculated. The two points which are farthest apart are assigned to the estimation set. The two points in the remaining list which are farthest are assigned to the prediction set. At the third step the point which is farthest from the two points in the estimation set is added to the estimation set. At the fourth step the point which is farthest from the two points in the prediction set is included in the prediction set. The alternation between the estimation and the prediction set continues until all points in the list have been assigned to one of the two sets. Of course, once a point is assigned to a set, it is deleted from further consideration. This algorithm was applied in the present study to separate data into two independent subsets: a training set of 22 compounds to build the model and a test set of the remained 21 compounds to evaluate its prediction ability. [18] suggests measuring the statistical properties of the estimation and prediction data sets by comparing the *p*th root of the determinants of the **X′X** matrices for these two data sets, where *p* is the number of parameters in the model. The determinant of **X′X** is related to the volume of the region covered by the points. Thus, if $\mathbf{X_E}$ and $\mathbf{X_P}$ denote the **X** matrices for points in the estimation and prediction data sets, respectively, then

$$\left(\frac{\left|X'_E X_E\right|}{\left|X'_P X_P\right|}\right)^{1/P} \quad (4)$$

is a measure of the relative volumes of the regions spanned by the two data sets. Ideally this ratio should be close to unity. It may also be useful to examine the variance inflation factors for the two data sets and the eigenvalue spectra of $\mathbf{X_E'X_E}$ and $\mathbf{X'_PX_P}$ to measure the relative correlation between the regressors. In using any data-splitting procedure (including the DUPLEX algorithm), several points should be kept in mind:

is a measure of the relative volumes of the regions spanned by the two data sets. Ideally this ratio should be close to unity. It can also be used to examine the variance of inflation factors for the two data sets and the spectral eigenvalue of $\mathbf{X_E'X_E}$ and $\mathbf{X'_PX_P}$ to measure the relative correlation between the regressors.

Multiple linear regression analysis and variable selection were performed by the software MobyDigs [19] using the Ordinary Least Square regression (OLS) method and Genetic Algorithm-Variable Subset Selection (GA-VSS) [20]. The outcome of the application of the genetic algorithms is a population of 100 regression models, ordered according to their decreasing internal predictive performance, verified by Q^2. The models with lower Q^2 are those with fewer descriptors. First of all, models with 1-2 variables were developed by the all – subset – method procedure in order to explore all the low dimension combinations. The number of descriptors was subsequently increased one by one, and new models were formed. The best models are selected at each rank, and the final model must be chosen from among them. This has to be sufficiently correlated and, at the same time, protect against any over parameterization, which would lead to a loss of predictive power for molecules outside training set. From a statistical view point the ratio of the number of samples (n) to the number of descriptors (m) should not be too low. Usually, it is recommended that n/ m ≥ 5 [21]. The GA was stopped when increasing the model size did not increase the Q^2 value to any significant degree.

Particular attention was paid to the collinearity of the selected molecular descriptors: by applying the QUIK rule (Q under Influence of K) [22] a necessary condition for the model validity. Acceptable model is only that with a global correlation of [x + y] block (Kxy) greater than the global correlation of the x block (Kxx) variable, x being the molecular descriptors and y the response variable. The collinearity in the original set of molecular descriptors results in many similar models that more or less yield the same predictive power (in MOBYDIGS software 100 models of different dimensionality). Therefore, when there were models of similar performance, those with higher ΔK (Kxy- Kxx) were selected and further verified. The models were justified by the R^2, the adjusted R^2, the cross-validated values of Q^2 by leave-one-out (LOO), the F ratio values and the standard error s. The robustness of the models and their predictivity were evaluated by both Q^2_{LOO} and bootstrap. In this last procedure K n- dimensional groups are generated by a randomly repeated selection of n- objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2 is calculated for each model. The bootstrapping was repeated 8000 times.

The proposed model was also checked for reliability and robustness by permutation testing: new models are recalculated for randomly recorded response (Y- scrambling) by using the same original independent variable matrix. After repeating this test several times (100 times in this work) it is expected to obtain new models that have significantly lower R^2 and Q^2 than the original model. If this condition is not verified the original model is not acceptable, as it was due to a chance correlation or a structural redundancy in the training set. Obtaining a robust model does not give real information about its prediction power. This is evaluated by predicting the compounds included in the test set. The Q^2 external for the test set is determined with equation (5):

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_{i/i} - y_i)^2 / n_{ext}}{\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2 / n_{tr}} \quad (5)$$

Here n_{ext} and n_{tr} are the number of objects in the external set (or left out by bootstrap) and the number of training set objects, respectively.

The applicability domain (AD) [23, 24] is a theoretical region in the space defined by the descriptors of the model and the modeled response, for which a given QSPR should make reliable predictions. In this work, the structural AD was verified by the leverage (hii) approach [25]. 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. The presence of both the response outliers (Y outliers) and the structurally influential compounds (X outliers) was verified by the Williams plot [26], the plot of standardized residuals versus leverage values.

MLR was utilized as linear technique, Whereas support vector machines (SVMs) were employed as nonlinear feature mapping technique for the construction of the QSPR models in this work. SVM is a novel classification and regression method proposed by Vapnik [27]. The SVM theory was described in detail elsewhere [28-30].

In our study we were using support vector machines (SVM), employing the software Molegro Data Modeller [31].

Briefly, assuming that sampled data set is (x_1, y_1), (x_2, y_2),….., (x_m, y_m), where x_k is the independent variables assembly of k- th sample which is a measured value, k=1, 2, …..,m. m is the total number of sample. The main idea of SVM is to make a regression hyper plane y=<w. x> + b, which can best fit samples in space. Based on the ε-insensitive loss function and Lagrange function, the original fitting problem can be transformed as the corresponding dual Lagrangian form. In consideration of kernel function K(.), the space transformation of inner product operation can be realized, and then the decision function can be obtained as below:

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) K(x, x_i) + b^* \quad (6)$$

Where both α_i and $\alpha_i^*$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 [28] . $K(x, x_i)$ is the Kernel function. Any function satisfying Mercer's condition [32] can be used as the Kernel function [27]. In this investigation, the Gaussian (RBF) Kernel function was employed in the SVM as below:

TABLE 1. Data for the studied Organophosphours.

Object	Status	Y Exp.	Y-Pred	Hat	Err.Pred.	Std.Err.Pred.
Azinphos-ethyl	Training	3.18	4.0163	0.178	0.8363	1.7494
Azinphos-methyl	Training	2.56	2.9149	0.105	0.3549	0.7116
Chlorpyrifos methyl	Training	4.24	3.5932	0.123	-0.6468	-1.3102
Cyanophos	Training	2.65	2.9614	0.162	0.3114	0.6454
Dichlorvos	Training	1.16	0.7387	0.373	-0.4213	-1.0095
Dicrotophos	Training	-0.49	-0.1313	0.232	0.3587	0.7762
Ethion	Training	5.07	4.4729	0.454	-0.5971	-1.5335
Fenthion	Training	4.09	4.2346	0.123	0.1446	0.2929
Fonofos	Training	3.94	4.6529	0.352	0.7129	1.6799
Isazofos	Training	3.82	3.7394	0.147	-0.0806	-0.1654
Methamidophos	Training	-0.8	-0.0903	0.298	0.7097	1.6071
Methidathion	Training	2.2	1.8743	0.147	-0.3257	-0.6687
Naled	Training	1.38	1.692	0.092	0.312	0.6209
oxydemeton-methyl	Training	-0.74	-0.2412	0.368	0.4988	1.1905
Phosmet	Training	2.78	1.9899	0.273	-0.7901	-1.7579
Phosphamidon	Training	0.79	1.0244	0.204	0.2344	0.4984
Pirimiphos-methyl	Training	4.2	3.7295	0.241	-0.4705	-1.0248
Profenofos	Training	4.44	3.9063	0.317	-0.5337	-1.2253
Propetamphos	Training	3.82	2.5274	0.055	-1.2926	-2.5226
Temephos	Training	5.96	7.2052	0.36	1.2452	2.9519
Tetrachlorvinphos	Training	3.53	3.4542	0.141	-0.0758	-0.155
Thiometon	Training	3.46	3.2919	0.254	-0.1681	-0.3691
Acephate	Test	-0.89	-0.5037	0.258	0.3863	0.8505
Azamethiphos	Test	1.05	0.861	0.248	-0.189	-0.4134
Chlorfenvinphos	Test	3.95	4.1533	0.35	0.2033	0.4783
Chlorpyrifos	Test	4.7	4.7917	0.25	0.0917	0.2008
Diazinon	Test	3.74	3.8238	0.24	0.0838	0.1823
Dimethoate	Test	0.7	1.4963	0.08	0.7963	1.5747
Disulfoton	Test	3.95	4.4455	0.204	0.4955	1.0534
Ethoprophos	Test	3.59	2.859	0.276	-0.731	-1.6298
Object	Status	Y Exp.	Y-Pred	Hat	Err.Pred.	Std.Err.Pred.
Etrimfos	Test	3.3	3.6712	0.349	0.3712	0.8722
Fenitrothion	Test	3.43	2.951	0.202	-0.479	-1.0171
Formothion	Test	-0.56	0.1923	0.306	0.7523	1.713
Isofenphos	Test	4.12	4.8863	0.432	0.7663	1.9283
Malathion	Test	2.75	2.2869	0.324	-0.4631	-1.0682
Mevinphos	Test	0.13	-0.5225	0.253	-0.6525	-1.4319
Phorate	Test	3.56	4.2034	0.21	0.6434	1.3734
Phosalone	Test	4.3	4.2462	0.182	-0.0538	-0.1128
Phoxim	Test	3.38	3.837	0.144	0.457	0.9368
Pirimiphos-ethyl	Test	4.85	4.4831	0.223	-0.3669	-0.7897
Sulprofos	Test	5.48	6.1676	0.501	0.6876	1.8464
Terbufos	Test	4.48	4.3531	0.356	-0.1269	-0.3
Trichlorfon	Test	0.51	0.6959	0.139	0.1859	0.3799

$$K(x\,,x_i) = \exp\left(\frac{-\|x - x_i\|^2}{\acute{o}^2}\right) \qquad (7)$$

In SVM regression three parameters strongly influence the number of the support vectors, having a close relation with the SVM performance and the training time. The γ parameter controls the amplitude of the RBF function and accordingly, it controls the SVM generalization ability. ε-insensitive parameter, it can prevent the entire training set meeting boundary conditions. In this way, the sparsity possibility in the dual formulation solution is provided. The optimum ε value is significantly affected by the noise type present in the data, which is usually unknown. The capacity parameter C controls the trade-off between the margin maximization and the training error minimization. When the C value is too low, then insufficient stress will be placed on fitting the training data. When the C value is too high, then the algorithm will over-fit the training data [33]. The prediction error was not frequently affected by the C parameter [34].

An additional external validation according to [35] is applied solely to the test set. According to the recommended criteria of Tropsha *et al*; a predictive QSPR model, must attend the following conditions:

1) $Q^2_{EXT} > 0.5$ (8-a)

2) $r^2 > 0.6$ (8-b)

3) $(r^2 - r_0^2)/r^2 < 0.1$ and $0.85 < k < 1.15$ (8-c)

or

$(r^2 - r'^2_0)/r^2 < 0.1$ and $0.85 < k' < 1.15$ (8-d)

where

$$r = \frac{\sum(y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\tilde{y}_i - \bar{\tilde{y}})^2}} \qquad (8\text{-a})$$

$$r_0^2 = 1 - \frac{\sum(y_i - y_i^{r_0})^2}{\sum(y_i - \bar{y})^2} \qquad (8\text{-b})$$

$$r'^2_0 = 1 - \frac{\sum(\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum(\tilde{y}_i - \bar{\tilde{y}})^2} \qquad (8\text{-c})$$

$$k = \frac{\sum(y_i \tilde{y}_i)}{\sum(\tilde{y}_i)^2} \qquad (8\text{-d})$$

$$k' = \frac{\sum(y_i \tilde{y}_i)}{\sum(y_i)^2} \qquad (8\text{-e})$$

where r is the correlation coefficient between the calculated and experimental values in the test set; r_0^2 (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 $y_i^{r_0} = k\,\tilde{y}$ and $\tilde{y}_i^{r_0} = k'\,y$, respectively; and the summations are over all samples in the test set. The reason to use r_0^2 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.

Results and Discussion

Multiple linear regression

A total of 43 compounds represent organophosphorous chemicals which are highly pollutant. The K_{ow} of these insecticides were converted to logarithmic form and listed in Table 1.

MLR is one of the most modeling methods in QSAR/QSPR, MLR method provides equation linking the structural feature to the Log K_{ow} of the compounds. The following equation obtained by this method is:

Log K_{OW} = 1.21 + 0.223 **Polarizability** - 0.899 **O-058** - 0.270 **nHAcc** - 3.89 **E1u** (9)

Volumes of the regions are used in this paper to illustrate the DUPLEX algorithm, and to confirm that the estimation and prediction points appear to be evenly distributed throughout the region and indicates that the prediction data set contains both interpolation and extrapolation points. For these two data sets we find that $|X_P{}'X_E| = |X'_P X_P|$.

Thus $\left(\frac{|X'_E X_E|}{|X'_P X_P|}\right)^{1/P} = 0.9839$ Indicating that the volumes of the two regions are very similar.

The model performances are described by means of the parameters related to the model predictive capability (Q^2_{LOO}, Q^2_{LMO}) and the fitting power (R^2). Standard deviation error in prediction (SDEP) and standard deviation error in calculation (SDEC) with chemicals domain are also reported.

The reported fitting and validation parameters have, as expected and shown in the Table 2, high values indicating that the model has very good predictive performance and the descriptors involved in it well describe the partition coefficient.

A detailed description of the linear model based on compounds in the training set summarized in Table 3. The high absolute t-values shown in Table 3 express that the regression coefficients of the descriptors involved in the MLR model are significantly larger than the standard deviation. The t- probability of a descriptor can describe the statistical significance when combined together within an overall collective QSPR model (i.e. descriptors interactions). Descriptors with t- probability values below 0.05 (95 percent confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance [27] The smaller t- probability suggests the more significant descriptor. The t- probability values of three descriptors are very small, indicating that all of them are highly significant descriptors. Models would not be accepted if they contain descriptors with VIFs above a value of five [27]. Correlation matrix as shown in Table 4 suggests that these descriptors are weakly correlated with each other. Thus, the model can be regarded as an optimal regression equation.

According to the (Fig. 1) it was clear that the calculated log Kow values were very similar to the experimental values.

On analyzing the model applicability domain from Williams plot, all the objects present a leverage smaller than the control value (h*=0.64) represented by the vertical straight line in the plot, and there is no aberrant compound both for training or prediction set (Fig. 2), which means that the model has a good external predictivity.

TABLE 2. The statistical parameters with n_{tr}=22, n_{test}=21.

statistical parameters	R^2	Q^2	Q^2boot	Q^2ext	R^2adj	s	F	$SDEP_{ext}$	Kx	Kxy	SDEP	SDEC
	93.57	89.21	86.13	92.73	92.06	0.527	61.87	0.493	28.43	43.69	0.601	0.463

TABLE 3**.** Characteristics of the selected descriptors in MLR model.

Predictor	Coef	SE Coef	T	P	VIF
Constant	1.208	1.002	1.21	0.244	
Polarizability	0.22317	0.02214	10.08	0	1.708
O-058	-0.8987	0.1948	-4.61	0	1.62
nHAcc	-0.2696	0.101	-2.67	0.016	1.589
E1u	-3.892	1.728	-2.25	0.038	1.1

TABLE 4. Correlation matrix.

	Log K_{OW}	Polarizability	O-058	nHAcc
Polarizability	0.825			
	0			
O-058	-0.717	-0.346		
	0	0.114		
nHAcc	-0.076	0.34	0.355	
	0.736	0.122	0.105	
E1u	0.049	0.274	0.013	0.112
	0.828	0.217	0.955	0.621

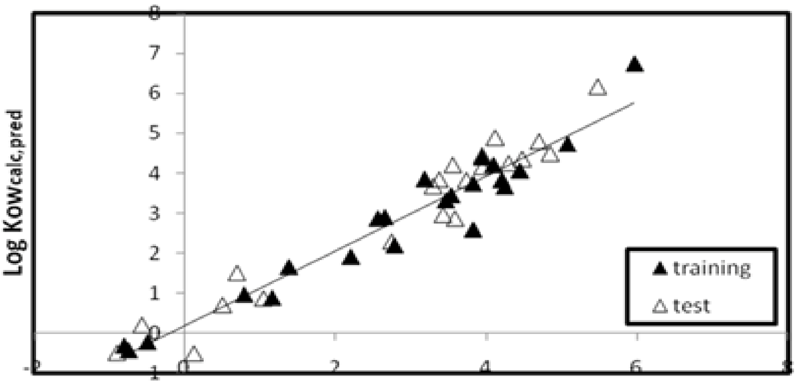

Fig.1. Plot of observed vs calculated of log Kow for the training and test sets.

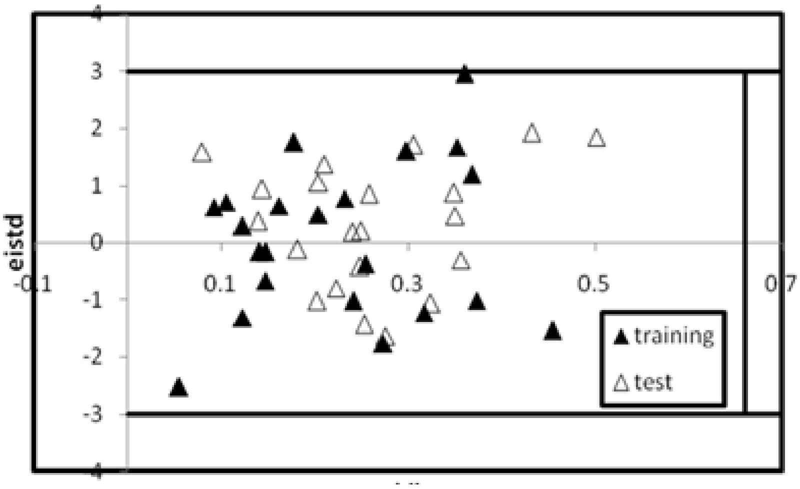

Fig. 2. Williams plot.

All the errors are distributed on both sides of the zero line; one may conclude that there is no systematic error in the model development. (Fig. 3) which represent the diagram of the statistical coefficients Q ^{2} and R ^{2} makes to compare the results obtained for the randomized models (triangles) with the starting model (ring). It is clear that the statistics obtained for the modified vectors of the log K_{ow} are smaller than those of the real models; Q^2 are lower than 0.3, and for the major part one obtains even Q^2<0. This allows ensuring that the established model has a real base, and is not due randomly.

Results of the Support vector regression model

After the establishment of the MLR model, SVM was used to develop a model by the training set compounds, based on the same subset of descriptors.

In our work the SVM model used the radial basis function (RBF). With a fine-tuning procedure, we tried to obtain the lowest Root Mean Square Error (RMSE) related to the best regression parameter using the leave- one- out (LOO) taking into account the RMSE of the test set i.e: similar fitting and predictive powers. The optimal values obtained for the SVM parameters and the results of the regression are presented in Table 5.

Correlation matrix as shown in Table 6 suggests that these descriptors are weakly correlated with each other. Thus, the model can be regarded as an optimal regression equation.

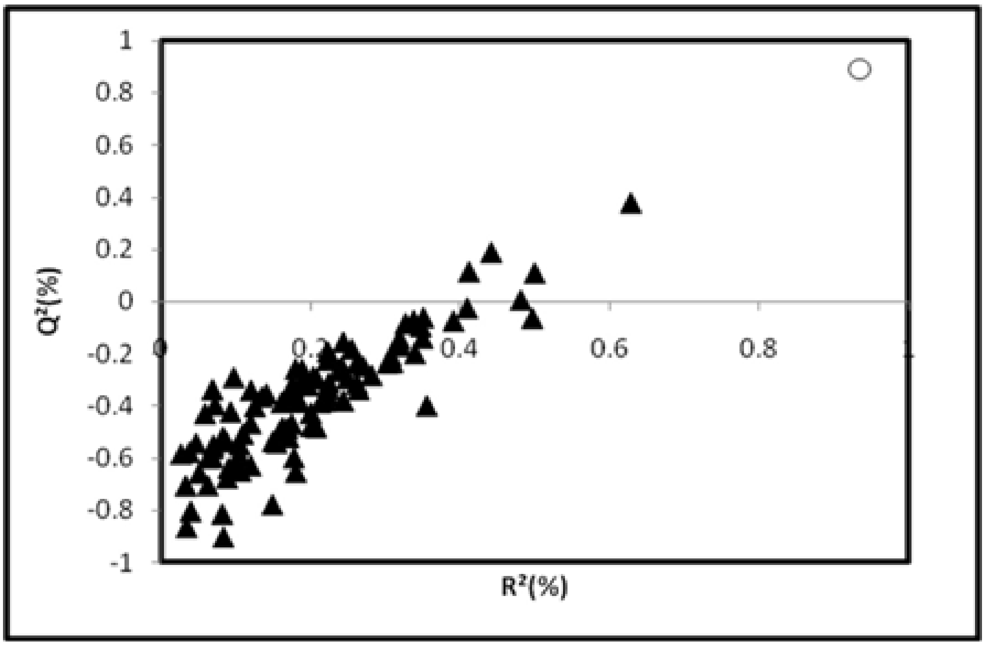

Fig. 3. Random test plot.

TABLE 5. Parameters and results of SVM model.

C	γ	E	R^2	Q^2_{loo}	Q^2_{ext}	RMSE	$RMSE_{ext}$
20	0.22	00	98.60%	90.70%	96.30%	0.316	0.504

TABLE 6. Correlation matrix.

	Log K_{OW}	Polarizability	O-058	nHAcc	E1u
Log K_{OW}		0.681	0.514	0.006	0.002
Polarizability	0.681		0.12	0.115	0.075
O-058	0.514	0.12		0.126	0
nHAcc	0.006	0.115	0.126		0.012
E1u	0.002	0.075	0	0.012	

TABLE 7. Comparative results of MLR and SVM.

Methods	Training set n= 28				Validation set n= 15					
	R^2	Q^2_{ext}	SDEC	$SDEP_{ext}$	R^2_{test}	Q^2_{ext}	$(r^2-r^2_0)/r^2<0.1$	$(r^2-r'^2_0)/r^2<0.1$	0.85 < k< 1.15	0.85< k'< 1.15
MLR	93.57%	92.73%	0.463	0.493	93.69%	92.95%	-0.0595	-0.0644	0.9545	1.0283
SVM	98.60%	96.30%	0.316	0.504	92.66%	92.60%	-0.0786	-0.0789	0.9872	0.9911

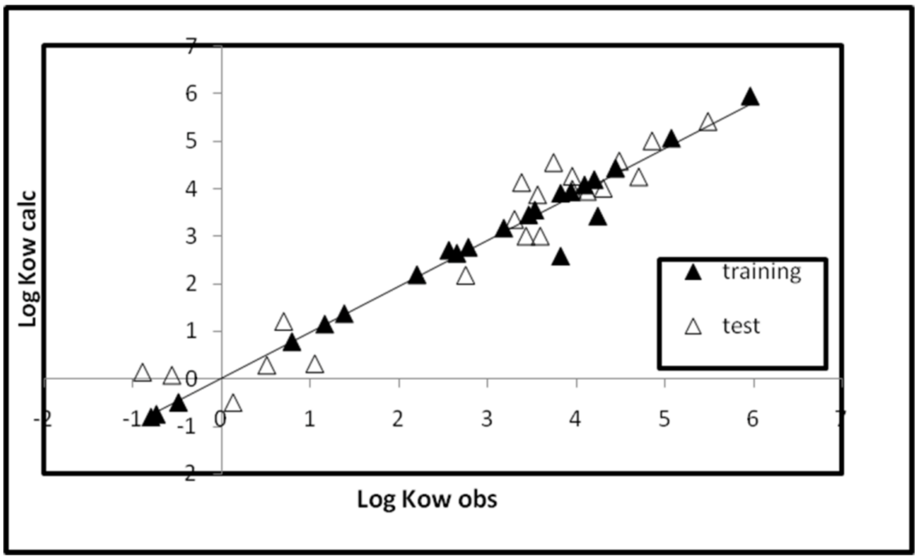

Fig. 4. Plot of observed *vs* calculated log K_{ow} for the training and test sets.

The comparative results of MLR and SVM model are showed in Table 7. The results demonstrated that SVM was more powerful than MLR model, because the SVM model presented a high statistical quality and low prediction error.

Conclusion

Recently, several works reported QSRR studies on the pesticides have been published. QSPR models were developed for the prediction of octanol/organic carbone partition coefficient (Koc) of an heterogeneous set of pesticides. The approaches based on multilinear regression (MLR), artificial neural networks (ANN), (*Bouakkadia, A et al*) [22] and The Duplex algorithm also used to separate data into two independent subsets: a training set of compounds to build the model and a test set of the remained- compounds to evaluate its prediction ability. (*Kertiou, N et al*) [18],this study aims to predict the octanol/water partition coefficient (Kow) of 43 organophosphorous insecticides. Quantitative structure- property relationship analysis was performed on a series of 43 insecticides using Multiple Linear Regression (MLR) and Support Vector Machines (SVM) methods, which correlate octanol- water partition coefficient (Kow) where we use the volume ratio for separate data into two independent subsets were approximately the same. A six-parameter linear model was developed by MLR, with R2 of 93.57%and RMSE of the 0.463, SVM with R2 of 98.60%, RMSE of 0.316 for training set. Several validation techniques, including leave-one-out cross validation, randomization tests, and validation through the test set, illustrated the reliability of the proposed model. All of the descriptors involved can be directly calculated from the molecular structure of the compounds, thus the proposed model is predictive and could be used to estimate the octanol/water partition coefficient of organophosphorous insecticides.

In the present work, the comparative of the performance of MLR and SVM in QSAR study results show that SVM can be used to derive statistical models with better qualities and better generalization capabilities than linear regression methods. The optimization process of SVM is relatively easy to be implemented. They can be used as alternative nonlinear modeling tools in QSAR.

References

1. Farghaly M., Tahaa H., Soliman S.M., Fathy U and Bedair A.H., Effect of Refining Processes on Magnitude and Nature of Fenitrothion and Pirimiphos-Methyl Residues in Maize Oil and Bioavailability of their Cake Residues on Rats. *Egyptian Journal of Chemistry*, **53**(6)**,** 923 – 938 (2010)

2. Myint K.Z and Xiang Q. X., Recent advances in fragment-based QSAR and multi-dimensional QSAR methods. *International Journal of Molecular Sciences,* **11**(10), 3846-3866 (2010).

3. Moyo F., Tandlich R., Wilhelmi B and Balaz S., Sorption of hydrophobic organic compounds on natural sorbents and organoclays from

aqueous and non-aqueous solutions: a mini-review. *International Journal of Environmental Research and Public Health*, **11**(5), 5020-5048 (2014).

4. Abdel-Gawad H., Kamel H. A and Hegazi B., Foliar Application of Calcium Chloride to Decrease Ethion Residues in the Oil of Chamomile Flowers and Soybean Seeds. *Egyptian Journal of Chemistry*, **54**(2), 175 – 187 (2011).

5. Yao X. J., Liu M. C., Zhang X. Y., Hu Z. D and Fan B. T., Radial basis function network-based quantitative structure–property relationship for the prediction of Henry's law constant. *Analytica Chimica Acta*, **462**(1), 101– 117 (2002).

6. Si H., Yuan S., Zhang K., Fu A., Duan Y. B and Hu Z., Quantitative structure activity relationship study on EC50 of anti-HIV drug. *Chemometrics and Intelligent Laboratory Systems*, **90**(1), 15-24 (2008).

7. Xu H. Y., Zhang J. Y., Zou J. W and Chen X. S., QSPR models for the physicochemical properties of halogenated methyl-phenyl ethers. *Journal of Molecular Graphics and Modelling*, **26**(7), 1076–1081 (2008).

8. Barber M. C., Dietary uptake models used for modeling the bioaccumulation of organic contaminants in fish. *Environmental Toxicology and Chemistry*, **27**(4), 755-777 (2008).

9. Devillers J., Prediction of toxicity of organophosphorus insecticides against the midge, Chironomus riparius, via a QSAR neural network model integrating environmental variables. *Toxicology Methods*, **10**(1), 69-79 (2000).

10. Vaal M. A., Van Leeuwen, C. J., Hoekstra J. A., and Hermens J. L., Variation in sensitivity of aquatic species to toxicants: practical consequences for effect assessment of chemical substances. *Environmental Management*, **25**(4), 415-423 (2000).

11. Niculescu S. P., Kaiser K. E., and Schultz T. W., Modeling the toxicity of chemicals to Tetrahymena pyriformis using molecular fragment descriptors and probabilistic neural networks. *Archives of Environmental Contamination and Toxicology*, **39**(3), 289-298 (2000).

12. Åberg A., MacLeod M., and Wiberg K., Physical-chemical property data for dibenzo-p-dioxin (DD), dibenzofuran (DF), and chlorinated DD/Fs: a critical review and recommended values. *Journal of Physical and Chemical Reference Data*, **37**(4), 1997-2008 (2008).

13. Todeschini R., Consonni V and Pavan M., Dragon, Software for the Calculation of Molecular Descriptors. Release 5.3 for windows, Milano, Italy, (2006).

14. Hansen O. C., *Quantitative Structure-Activity Relationships (QSAR) and Pesticides,* Teknologisk Institute. Pesticides Research; Danish, (2004).

15. Hyperchem^{TM}. Release 6.02 for windows, Molecular Modeling system, (2000).

16. Todeschini R., Consonni V., Mauri A and Pavan M., DRAGON Software – version 5.4-TALETE srl, (2005).

17. Bouakkadia A., Haddag H., Bouarra N., and Messadi D., QSPR study of the water solubility of a diverse set of agrochemicals: hybrid (GA/MLR) approach. *Synthese*, **32**(1), 12-21 (2016).

18. Kertiou, N. E., Bouakkadia, A., and Messadi, D., 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. (2017).

19. Todeschni R., Ballabio D., Consonni V., Mauri A and Pavan M., MOBYDIGS – version 1.1 – Copyright TALETE srl (2004).

20. Gramatica P., Principles of QSAR models validation: internal and external. *QSAR & Combinatorial Science*, **26**(5), 694-701 (2007).

21. Xu J., Zhang H., Wang L., Liang G., Wang L., Shen X and Xu W., QSPR study of absorption maxima of organic dye- sensitized solar cells based on 3D descriptors. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **76**(2), 239 –247 (2010).

22. Bouakkadia A., Kertiou N., Bouakkadia H., and Messadi D., Soil contamination by pesticides: molecular modeling of octanol/organic carbone partition coefficient. *Energy Procedia*, **157**, 551-560 (2019).

23. Tropsha A., Gramatica P and Gombar V. K., The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Combinatorial Science*, **22**(1), 69–77 (2003).

24. Shen M., Béguin C., Golbraikh A., Stables J. P; Kohn H and Tropsha A., Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *Journal of Medicinal Chemistry*, **47**(9), 2356 –2364 (2004).

25. Weisberg S., *Applied Linear Regression,* third ed; John Wiley and sons, New Jersey, (2005).

26. SCAN- Software for Chemometric Analysis-version 1.1- for Windows, Minitab USA (1995).

27. Bouakkadia A., Lourici L., and Messadi D., Modeling and prediction of octanol/water partition coefficient of pesticides using QSPR methods. *Management of Environmental Quality: An International Journal*, **28**(4), 579-592 (2017).

28. Nello C and Joh S. T., *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods,* Publishing House of Electronics Industry, fourth ed; Beijing, (2005).

29. Chen N., Lu W., Yang J., Li G., *Support Vector Machine in Chemistry,* World Scientific, Singapore, Publishing, Co. Pte. Ltd, (2004).

30. Molegro, Data Modeller User Manual, Copyright Molegro, (2008).

31. Molegro, Data Modeller (MDM), v.2.1.0. Copyright Molegro, (2009).

32. Darnag R., Minaoui B and Fakir M., QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression. *Arabian Journal of Chemistry*, **10**, S600-S608 (2017).

33. Riahi S., Pourbasheer E., Ganjali M. R and Norouzi P., Investigation of different linear and non linear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine. *Journal of Hazardous Materials*, **166**(2-3), 853–859 (2009).

34. Wang W. J., Xu Z. B., Lu W. Z and Zhang X. Y., Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing*, **55**(3-4), 643–663 (2003).

35. Golbraikh A. and Tropsha A., Beware of q^2!. *Journal of Molecular Graphics and Modelling*, **20,** 269 –276 (2002).