Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS

J. Serb. Chem. Soc. 85 (4) 467–480 (2020)
JSCS–5315

Original scientific paper

QSAR Study of the octanol/water partition coefficient of organophosphorus compounds: The hybrid GA/MLR and GA/ANN approaches

RANA AMIRI[1], DJELLOUL MESSADI[1]* and AMEL BOUAKKADIA[2]

[1]*Environmental and Food Safety Laboratory, Badji Mokhtar-Annaba University, BP.12, 23000 Annaba, Algeria and* [2]*University Abbes Laghrour Khenchela, BP 1252 Route de Batna, Khenchela 40004, Algeria*

(Received 10 June, revised 22 July, accepted 22 August 2019)

Abstract: This study aimed at predicting the *n*-octanol/water partition coefficient (K_{ow}) of 43 organophosphorous insecticides. Quantitative structure–property relationship analysis was performed on the series of 43 insecticides using two different methods, linear (multiple linear regression, MLR) and non-linear (artificial neural network, ANN), which K_{ow} values of these chemicals to their structural descriptors. First, the data set was separated with a duplex algorithm into a training set (28 chemicals) and a test set (15 chemicals) for statistical external validation. A model with four descriptors was developed using as independent variables theoretical descriptors derived from Dragon software when applying genetic algorithm (GA)–variable subset selection (VSS) procedure. The values of statistical parameters, R^2, Q^2_{ext}, $SDEP_{ext}$ and *SDEC* for the MLR (94.09 %, 92.43 %, 0.533 and 0.471, respectively) and ANN model (97.24 %, 92.17 %, 0.466 and 0.332, respectively) obtained for the three approaches are very similar, which confirmed that the employed four parameters model is stable, robust and significant.

Keywords: octanol/water partition coefficient; molecular descriptors; QSPR methods.

INTRODUCTION

Pesticides present the only group of chemicals that are purposely applied to the environment with aim of suppressing plant and animal pests and to protect agricultural and industrial products. However, the majority of pesticides are not specifically targeting the pest only and during their application, they also affect non-target plants and animals. Repeated application leads to loss of biodiversity.[1] Many pesticides are not easily degradable, they persist in the soil, leach to groundwater and surface water and contaminate the wide environment. Depend-

* Corresponding author. E-mail: d_messadi@yahoo.fr
https://doi.org/10.2298/JSC190610090A

ing on their chemical properties, they can enter the organism, bioaccumulate in food chains and consequently influence human health. Overall, intensive pesticide application results in several negative effects in the environment that cannot be ignored. Organophosphorus compounds are widely used as insecticides in agricultural and domestic practice, which causes their widespread appearance in the environment.[2]

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 have proven to be reliable tools for toxicity assessment of organic chemicals when little or no empirical data are available.[1,2]

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.[3] 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.[4]

The physicochemical properties of an organic chemical compound play an important role in determining its distribution and fate in the environment. Vapor pressures (P_V), aqueous solubility ($S_{w,L}$) and *n*-octanol/water partition coefficients (K_{ow}) are key physicochemical properties that could be used for assessing environmental partition and transport of organic substances.[5]

The K_{ow} value 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 parameter is also a good indicator of the systemic mode of action of a pesticide.

Properties such as the *n*-octanol–water partition coefficient are important in predicting the environmental fate of organic compounds.[6] Furthermore, the K_{ow} is used as one of the molecular descriptors of the toxic effects of chemicals in QSAR.[7–9] The partition coefficient value (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,[10] in the present case *n*-octanol and water. Octanol represents a substitute for biotic lipids and hence gives an approximation to a biotic lipid–water partition coefficient.[11]

The objective of this study was to develop QSAR models to describe the acute toxicity of pesticides and to find a statistical model for the prediction of

K_{ow} of organophosphorous compounds. For this purpose, the relationship between molecular descriptors[12] connected to the factors found, experimentally, as affecting the K_{ow} of the compounds. The QSPR model was constructed using multiple linear regression (MLR) and artificial neural networks (ANN). The model obtained shows which descriptors play a significant role in the K_{ow} variation of these pesticides.

Recently, several works reported QSAR used for K_{ow} prediction have been published, such as QSARpy: A new flexible algorithm to generate QSAR models based on dissimilarities. In addition to the log K_{ow} case study by Thomas *et al.*[13] on the modeling and prediction of octanol/water partition coefficient of pesticides using QSPR methods, two methods, MLR and SVM, were used to predict K_{ow} of pesticides by Bouakkadia *et al.*[14] In this article, they constructed a new connectivity index (*mD*) based on the randic branching degree index (*mχ*). Moreover, the first-order index (*1D*) and hydrophobicity parameters (log K_{ow}) were used to estimate the acute toxicity of substituted arenes to aquatic organisms by Feng *et al*.[15]

EXPERIMENTAL

Experimental data

The 43 organophosphorous insecticides are taken from Hasen[16] and are listed in Table I. The data is presented as log K_{ow} to reduce the range of variation. The data set was separated into a training set of 28 compounds and a test set of 15 compounds.

Descriptors generation

The chemical structure of each compound was sketched on a PC using the Hyperchem program[17] and pre-optimized using the 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* Å^{-1} mol^{-1} as the stopping criterion. Then the geometries were used as input for the generation of 1664 descriptors using the Dragon software (version 5.4).[18] Quantum-chemical descriptors such as highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), HOMO–LUMO gap (DHL), and ionization potential, calculated by the semi-empirical PM3 method using Hyperchem,[17] were added and used for descriptor selection during model development. Constant values and descriptors found to be correlated pair wise were excluded in a pre-reduction step (when there was more than 98 % pair wise correlation, one variable was deleted), and the genetic algorithm was applied for variables selection to a final set of 1230 descriptors.

Training set and test set

It is important to rationally define a training set from which the model is built and an 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.

* 1 kcal = 4184 J

TABLE I. Names, CAS, values of log K_{ow} and structures of the insecticides examined in this study

No.	Object	log K_{ow} experiment	log K_{ow} predicted	h_{ii}	Error prediction	Standardized error prediction
1	Acephate[a]	–0.89	–0.5037	0.258	0.3863	0.8505
2	Azamethiphos[a]	1.05	0.861	0.248	–0.189	–0.4134
3	Azinphos-ethyl	3.18	4.0163	0.178	0.8363	1.7494
4	Azinphos-methyl[a]	2.56	2.9149	0.105	0.3549	0.7116
5	Chlorfenvinphos[a]	3.95	4.1533	0.35	0.2033	0.4783
6	Chlorpyrifos[a]	4.7	4.7917	0.25	0.0917	0.2008
7	Chlorpyrifosmethyl[a]	4.24	3.5932	0.123	–0.6468	–1.3102
8	Cyanophos	2.65	2.9614	0.162	0.3114	0.6454
9	Diazinon	3.74	3.8238	0.24	0.0838	0.1823
10	Dichlorvos	1.16	0.7387	0.373	–0.4213	–1.0095
11	Dicrotophos	–0.49	–0.1313	0.232	0.3587	0.7762
12	Dimethoate[a]	0.7	1.4963	0.08	0.7963	1.5747
13	Disulfoton	3.95	4.4455	0.204	0.4955	1.0534
14	Ethion[a]	5.07	4.4729	0.454	–0.5971	–1.5335
15	Ethoprophos	3.59	2.859	0.276	–0.731	–1.6298
16	Etrimfos[a]	3.3	3.6712	0.349	0.3712	0.8722
17	Fenitrothion[a]	3.43	2.951	0.202	–0.479	–1.0171
18	Fenthion	4.09	4.2346	0.123	0.1446	0.2929
19	Fonofos	3.94	4.6529	0.352	0.7129	1.6799
20	Formothion	–0.56	0.1923	0.306	0.7523	1.713
21	Isazofos[a]	3.82	3.7394	0.147	–0.0806	–0.1654
22	Isofenphos[a]	4.12	4.8863	0.432	0.7663	1.9283
23	Malathion[a]	2.75	2.2869	0.324	–0.4631	–1.0682
24	Methamidophos	–0.8	–0.0903	0.298	0.7097	1.6071
25	Methidathion	2.2	1.8743	0.147	–0.3257	–0.6687
26	Mevinphos	0.13	–0.5225	0.253	–0.6525	–1.4319
27	Naled	1.38	1.692	0.092	0.312	0.6209
28	oxydemeton-methyl	–0.74	–0.2412	0.368	0.4988	1.1905
29	Phorate[a]	3.56	4.2034	0.21	0.6434	1.3734
30	Phosalone	4.3	4.2462	0.182	–0.0538	–0.1128
31	Phosmet	2.78	1.9899	0.273	–0.7901	–1.7579
32	Phosphamidon	0.79	1.0244	0.204	0.2344	0.4984
33	Phoxim	3.38	3.837	0.144	0.457	0.9368
34	Pirimiphos-ethyl	4.85	4.4831	0.223	–0.3669	–0.7897
35	Pirimiphos-methyl	4.2	3.7295	0.241	–0.4705	–1.0248
36	Profenofos	4.44	3.9063	0.317	–0.5337	–1.2253
37	Propetamphos	3.82	2.5274	0.055	–1.2926	–2.5226
38	Sulprofos	5.48	6.1676	0.501	0.6876	1.8464
39	Temephos	5.96	7.2052	0.36	1.2452	2.9519
40	Terbufos[a]	4.48	4.3531	0.356	–0.1269	–0.3
41	Tetrachlorvinphos	3.53	3.4542	0.141	–0.0758	–0.155
42	Thiometon	3.46	3.2919	0.254	–0.1681	–0.3691
43	Trichlorfon	0.51	0.6959	0.139	0.1859	0.3799

[a]Validation compounds

Chemometric methods

The duplex algorithm adopted in this study proceeds as follows. In the first step, the two points which are furthest away from each other are selected for the training set. From the remaining points, the two-objects that are furthest away are included in the test set. In the third step, the remaining point which is furthest away from the two previously selected for the training set is included in that set. The procedure is repeated selecting a single point for the test set that is furthest from the existing points in that set. Following this procedure, points are added alternately to each set.[19] This algorithm was applied in the present study to separate data into two independent subsets: a training set of 28 compounds to build the model and a test set of the remained 15 compounds to evaluate its prediction ability.

Multiple linear regression analysis and variable selection were performed by the software MobyDigs[20] using the ordinary least square regression (OLS) method and genetic algorithm––variable subset selection (GA–VSS).[21] 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 were 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 the 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 $n/m \geq 5$.[22] 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 co-linearity of the selected molecular descriptors: by applying the Q under influence of K (QUIK) rule,[23] a necessary condition for the model validity. An acceptable model is only that with a global correlation of [$x + y$] block (K_{xy}) greater than the global correlation of the x block (K_{xx}) variable, x being the molecular descriptors and y the response variable. The co-linearity 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 (K_{xy}–K_{xx}) 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 was 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 were 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 values 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 Eq. (1):

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

where 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.

Applicability domain analysis

The applicability domain (AD)[24,25] 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 (h_{ii}) approach.[26] 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 a Williams plot,[27] a plot of standardized residuals *versus* leverage values.

MLR was utilized as linear technique, whereas ANNs are artificial systems simulating the function of the human brain. Three components constitute an ANN: the processing elements or nodes, the topology of the connections between the nodes, and the learning rule by which new information is encoded in the network. While there are a number of different ANN models, the most frequently used type of ANN in QSAR is the three-layered feed-forward network.[28] In this type of networks, the neurons are arranged in layers (an input layer, one hidden layer and an output layer). Each neuron in any layer is fully connected with the neurons of a succeeding layer and there are no connections are between neurons belonging to the same layer.

An additional external validation according to the lierature[29] 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:

$$Q^2_{\text{ext}} > 0.5 \quad (2a)$$

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

$$(R^2 - R_0^2)/R^2 < 0.1 \text{ and } 0.85 < k < 1.15 \quad (2c)$$

or

$$(R^2 - R_0'^2)/R^2 < 0.1 \text{ and } 0.85 < k' < 1.15 \quad (2d)$$

where

$$R = \frac{\sum (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (3a)$$

$$R_0'^2 = 1 - \frac{\sum (y_i - y_i^{r_0})^2}{\sum (y_i - \bar{y})^2} \quad (3b)$$

$$R_0'^2 = 1 - \frac{\sum\left(\tilde{y}_i - \tilde{y}_i^{r0}\right)^2}{\sum\left(\tilde{y}_i - \bar{\tilde{y}}\right)^2} \tag{3c}$$

$$k = \frac{\sum(y_i \tilde{y}_i)}{\sum(\tilde{y}_i)^2} \tag{3d}$$

$$k' = \frac{\sum(y_i \tilde{y}_i)}{\sum(y_i)^2} \tag{3e}$$

where R is the correlation coefficient between the calculated and experimental values in the test set; R_0^2 (calculated *versus* observed values) and $R_0'^2$ (observed *versus* calculated values) are the coefficients of determination; k and k' are slopes of regression lines through the origin of the calculated *versus* observed and the observed *versus* calculated, respectively, are defined as:

$$y_i^{r0} = k\tilde{y} \tag{4}$$

and

$$\tilde{y}_i^{r0} = k'y \tag{5}$$

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 that are highly pollutant. The K_{ow} of these insecticides were converted to logarithmic form and are listed in Table I.

The application of the GA–VSS led to several good models for the prediction of the K_{ow} of organphosphorus chemicals based on different sets of molecular descriptors. The best model obtained using 43 organophosphorous is a model with a high predictive power, which was established by using the regression equation according to:

$$\log K_{ow} = 1.87 + 0.219\text{Polarizability} - 0.817\text{O-058} - 0.218\text{nHAcc} - 5.44\text{E1u} \tag{8}$$

In Eq. (8), four different kinds of molecular descriptors appear that are described in Table II. 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). The standard deviation error in the prediction (*SDEP*) and the standard deviation error in the calculation (*SDEC*) with the chemicals domain are also reported.

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

TABLE II. Molecular descriptors

Descriptor	Class	Meaning
Polarizability	Hyperchem descriptor	Polarizability defined as the dipole moment of a molecule induced by an electric field of unit intensity
O-058	Atom-centered fragments	Defined hydrophobicity
nHAcc	Functional group counts	Total number of Ns, Os and Fs in the molecule, excluding N with a formal positive charge, higher oxidation states and the pyrrolyl form of N
E1u	WHIM descriptors	1st component accessibility directional WHIM index / unweighted

TABLE III. The statistical parameters with $n_{tr} = 28$, $n_{test} = 15$

Statistical parameter	$n_{tr} = 28$	$n_{test} = 15$
Q^2_{ext} / %		92.43
$SDEP_{ext}$		0.533
R^2 / %	94.09	
Q^2 / %	91.39	
Q^2_{boot} / %	89.36	
Q^2_{adi} / %	93.06	
SDEP	0.569	
SDEC	0.471	
S	0.520	
F	91.53	
K_{xx}	30.91	
K_{xy}	46.59	

The high absolute *t*-values shown in Table IV 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).

TABLE IV. Characteristics of the selected descriptors in the MLR model

Predictor	Coef	SE Coef	*t*	*t*-probability	*VIF*
Constant	1.8705	0.9856	1.9	0.07	–
Polarizability	0.21855	0.01843	11.86	0	1.558
O-058	–0.817	0.1693	–4.83	0	1.707
nHAcc	–0.2175	0.08053	–2.7	0.013	1.528
E1u	–5.436	1.645	–3.3	0.003	1.109

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.[30] A smaller *t*-probability suggests a 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 *VIF* values above five.[31] The correlation matrix as shown in Table V suggests that these descriptors are weakly correlated with each other. Thus, the model can be regarded as an optimal regression equation.

TABLE V. Correlation matrix

	log K_{ow}	Polarizability	O-058	nHAcc
Polarizability	0.847			
	0			
O-058	–0.739	–0.449		
	0	0.017		
nHAcc	–0.181	0.21	0.357	
	0.357	0.284	0.062	
E1u	–0.219	0.066	0.124	0.313
	0.263	0.737	0.53	0.105

On analyzing the model applicability domain from the Williams plot, all the objects present a leverage smaller than the control value ($h^* = 0.54$) represented by the vertical straight line in the plot, and there is no aberrant compound for both the training or prediction set, Fig. 1, which means that the model has a good external predictivity.

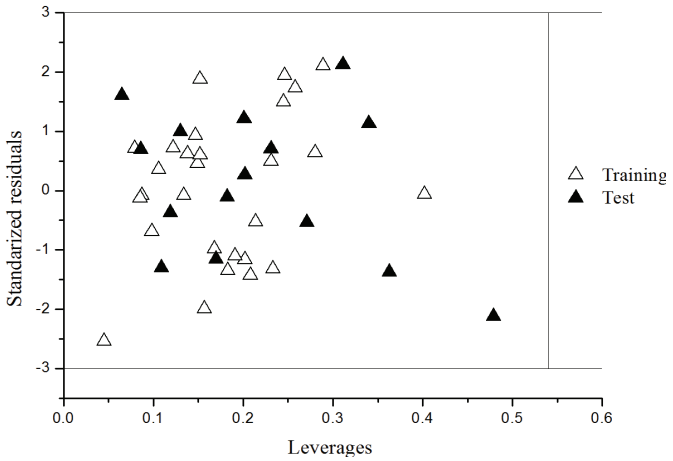

Fig. 1. Williams plot.

As all the errors are distributed on both sides of the zero line; it may be conclude that there is no systematic error in the model development. Fig. 2, which represents the diagram of the statistical coefficients Q^2 and R^2, enables the results obtained for the randomized models (triangles) to be compared with the starting model (ring). It is clear that the statistics obtained for the modified vectors of the

log K_{ow} values are smaller than those of the real models; Q^2 are lower than 0.3, and for the major part even $Q^2 < 0$ is obtained. This ensures that the established model has a real base and is not due randomness.

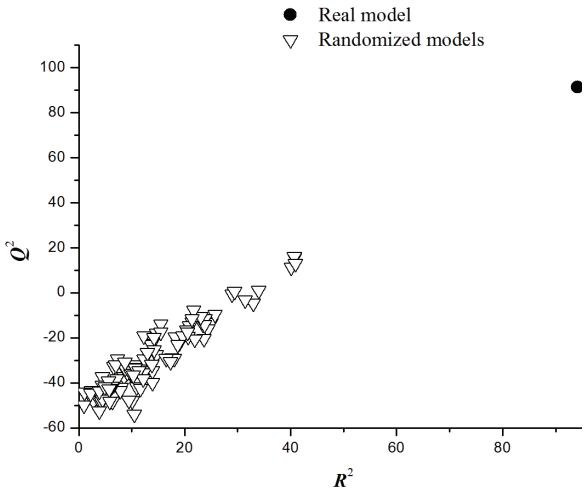

Fig. 2. Random test.

Results of the ANN model

The choice of the number of neurons of the hidden layer was fixed at 2 and it iteration count to 50. Table VI clarify this choice.

TABLE VI. Optimal structure adopted for the neural network

Input	04 (descriptors)
Output	01 (log K_{ow})
Hidden layer	One Hidden layer
Number of neurons	02
Learning algorithm	Retro propagation of the error gradient
Learning function	Hyperbolic tangent (hidden layer)
	Linear (output layer)

$$R^2 = 97.24\ \%;\ Q^2_{ext} = 92.17\ \%;\ SDEC = 0.332;\ SDEP_{ext} = 0.466$$

The inputs of the ANN were the subset of the descriptors selected by a genetic algorithm from a large set of the descriptors. The input layer comprises four descriptors. Usually one hidden layer is enough. After several trials, a hidden layer with two neurons was selected; Fig. 3 explains this choice.

The plot of predicted *versus* experimental values for the data set is shown in Fig. 4 which shows that the K_{ow} values calculated by the ANN are very similar to the experimental ones.

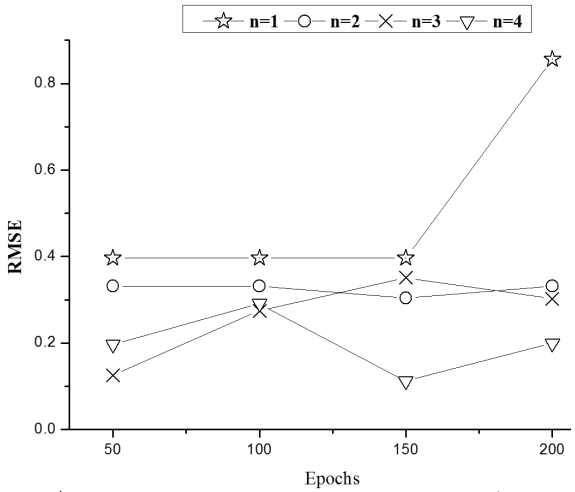

Fig. 3. Choice of the number of neurons of the hidden layer and the number of cycles.

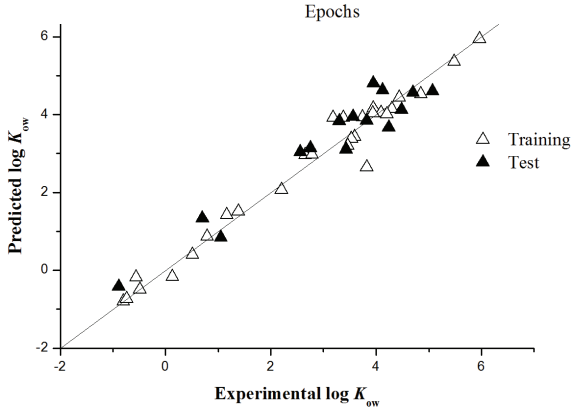

Fig. 4. Plot of the observed *vs.* calculated log *K*ow for the training and test sets (ANN).

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

TABLE VII. Comparative results of the MLR and ANN methods

Set	Parameter	Method	
		MLR	SVM
Training = 28	R^2 / %	94.09	97.24
	Q^2_{ext} / %	92.43	92.17
	SDEC	0.471	0.332
	$SDEP_{ext}$	0.533	0.466
Validation =15	R^2_{test} / %	89.35	92.53
	Q^2_{ext} / %	89.73	92.17
	$(R^2–R_0^2) / R^2<0.1$	–0.1188	–0.075
	$(R^2–R_0'^2) / R^2<0.1$	–0.118	–0.0797
	$0.85 < k < 1.15$	0.9917	0.9701
	$0.85< k' < 1.15$	0.9851	1.0138

CONCLUSIONS

The QSAR method was applied to the prediction of the octanol/water partition coefficient of organophosphorous insecticides. The validation techniques, including leave-one-out cross validation, randomization tests, and validation through the test set, illustrated the reliability of the proposed model. All the descriptors involved could 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 performance of MLR and ANN were compared in the QSAR study. The obtained results show that ANN seems to be the best way to select representative calibration and test data sets in a validation context and could be used to derive statistical models with better qualities and better generalization capabilities than linear regression method. The optimization process of ANN could be relatively easy implemented. They can be used as alternative non -linear modeling tools in QSAR. The ANN approach would seem to have great potential for determining quantitative structure activity relationships and as such is a valuable tool for chemists.

ИЗВОД

QSAR ПРОУЧАВАЊЕ ПОДЕОНИХ КОЕФИЦИЈЕНТА ОКТАНОЛ/ВОДА ОРГАНОФОСФОРНИХ ЈЕДИЊЕЊА: ХИБРИДНИ GA/MLR И GA/ANN ПРИСТУПИ

RANA AMIRI[1], DJELLOUL MESSADI[1] и AMEL BOUAKKADIA[2]

[1]*Environmental and Food Safety Laboratory, Badji Mokhtar-Annaba University, BP.12, 23000 Annaba, Algeria и* [2]*University Abbes Laghrour Khenchela, BP 1252 Route de Batna Khenchela 40004, Algeria*

Ова студија настоји да предвиди подеони коефицијент *n*-октанол/вода (K_{ow}) за 43 органофосфорна инсектицида. Анализа квантитативне релације структуре и особина урађена је на серији од 43 инсектицида коришћењем две различите методе линеарне (*Multiple Linear Regression*, MLR) и нелинеарне (*Artificial Neural Network*, ANN), које корелишу вредности K_{ow} ових једињења са њиховим структурним дескрипторима. Прво су подаци раздвојени помоћу дуплекс алгоритма на скуп за учење (training set, 28 једињења) и скуп за проверу (test set, 15 једињења) за статистичку спољашњу валидацију. Рзвијен је модел са четири дескриптора користећи као независне варијабле теоријске дескрипторе изведене из Dragon софтвера уз примену генетски алгоритам (GA)–променљив избор подскупа (VSS). Вредности статистичких параметара R^2, Q^2_{ext}, $SDEP_{ext}$ и *SDEC* за MLR 94,09 %; 92,43 %; 0,533 и 0,471, редом) и ANN (97,24 %; 92,17 %; 0,46 и 0,332, редом) моделе, добијени за три приступа врло су слични, што потврђује да је четворопараметарски модел стабилан, робустан и значајан.

(Примљено 10 јуна, ревидирано 22. јула, прихваћено 22. августа 2019)

REFERENCES

1. S. P. Bradbury, *Tox. Lett*. **79** (1995) 37(https://doi.org/10.1002/etc.5620160629)
2. A. P. Bearden, T. W. Schultz, *Environ. Toxicol. Chem*.**16** (1997) 7 (https://doi.org/10.1016/0378-4274(95)03374-T)
3. X. J. Yao, M. C. Liu, X. Y. Zhang, Z. D. Hu, B. T. Fan, *Anal. Chim. Acta* **462** (2002) 101 (https://doi.org/10.1016/S0003-2670(02)00273-8)

4. H. Si, S. Yuan, K. Zhang, A. Fu, Y. B. Duan, Z. Hu, *Chemometr. Intell. Lab. Syst.* **90** (2008) 15 (https://doi.org/10.1016/j.chemolab.2007.06.011)
5. H. Y. Xu, J. Y. Zhang, J. W. Zou, X. S. Chen, *J. Mol. Graph. Model.* **26** (2008) 1076 (https://doi.org/10.1016/j.jmgm.2007.09.004)
6. F. Wania, D. Mackay, *Environ. Pollut.* **100** (1999) 223 (https://doi.org/10.1016/S0269-7491(99)00093-7)
7. S. W. Fisher, M. J. Lydy, J. Barger, P. F. Landrum, *Environ. Toxicol. Chem*. **12** (1993) 1307 (https://doi.org/10.1002/etc.5620120721)
8. C. J. Van Leeuwen, P. T. Vanderzandt, T. Aldenberg, H. J. M. Verhaar, J. L. M. Hermens, *Environ. Toxicol. Chem*. **11** (1992) 267 (https://doi.org/10.1002/etc.5620110216)
9. S. P. Niculescu, K. L. E. Kaiser, G. Schüürmann, *Water Qual. Res. J. Can*. **33** (1998) 153 (https://doi.org/10.2166/wqrj.1998.009)
10. S. L. Wong, *Aquat. Toxicol.* **6** (1985) 115 (https://doi.org/10.1016/0166-445X(85)90011-6)
11. M. Chessels, D. W. Hawker, D. W. Connell, *Chemosphere* **22** (1991) 1175 (https://doi.org/10.1016/0045-6535(91)90213-W)
12. R. Todeshini, V. Consonni, M. Pavan, *Dragon, Software for the Calculation of Molecular Descriptors,* Release 5.3 for Windows, Milan, 2006
13. T. Ferrari, A. Lombardo, E. Benfenati, *Sci. Total Environ.* **637** (2018) 1158 (https://doi.org/10.1016/j.scitotenv.2018.05.072)
14. A. Bouakkadia, L. Lourici, D. Messadi, *Manage. Environ. Qual.* **28** (2017) 579 (http://dx.doi.org/10.1108/MEQ-08-2015-0162)
15. H. Feng, Y. Chen, W. Yue, C Feng, *Earth Environ. Sci.* **153** (2018) 022035 (https://doi:10.1088/1755-1315/153/2/022035)
16. O. C. Hansen, *Quantitative Structure-Activity Relationships (QSAR) and Pesticides*, Teknologisk Institute, Pesticides Research No. 94 (2004)
17. *Hyperchem*TM, Release 6.02 for Windows. Molecular Modeling system, 2000 (http://www.hyper.com/)
18. R. Todeschini, V. Consonni, A. Mauri, M. Pavan, *Dragon Software*, version 5.4, Copyright TALETE srl, 2005
19. R. D. Snee, *Technometrics* **19** (1977) 415 (https://doi.org/10.1080/00401706.1977.10489581)
20. R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Paven, *MobyDigs*, version 1.1, Copyright TALETE srl, 2009 (http://www.talete.mi.it/)
21. R. Leardi, R. Boggia, M. Tarrile, *J. Chemom*. **6** (1992) 267(https://doi.org/10.1016/B978-012213810-2/50004-9)
22. J. Xu, H. Zhang, L. Wang, G. Liang, L. Wang, X. Shen, W. Xu, *Spectrochim. Acta, A* **76** (2010) 239 (https://doi.org/10.1016/j.saa.2010.03.027)
23. R. Todeschini, A. Maiocchi, V. Consonni, *Chemom. Int. Lab. Syst*. **46** (1999) 13 (https://doi.org/10.1016/S0169-7439(98)00124-5)
24. A. Tropsha, P. Gramatica, V. K. Grombar, *QSAR Comb. Sci*. **22** (2003) 69 (https://doi.org/10.1002/qsar.200390007)
25. M. Shen, C. Béguin, A. Golbraikh, J. P. Stables, H. Kohn, A. Tropsha, *J. Med. Chem*. **47** (2004) 2356 (https://pubs.acs.org/doi/abs/10.1021/jm030584q)
26. S. Weisberg, *Applied Linear Regression*, 3rded., John Wiley and Sons, Inc., Hoboken, NJ, 2005
27. *SCAN Software for Chemometric Analysis*, version 1.1 for Windows, Minitab, USA 1995

28. J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, Wiley-VCH, Weinheim, 1999
29. A. Golbraikh, A. Tropsha, *J. Mol. Graphics Modell*. **20** (2002) 269 (https://doi.org/10.1016/S1093-3263(01)00123-1)
30. S. R. Gunn, *Support Vector Machines for Classification and Regression*, Technical Report, University of Southampton, 1998 (http://users.ecs.soton.ac.uk/srg/publications/pdf/SVM.pdf)
31. N. Chen, W. Lu, J. Yang, G. Li, *Support Vector Machine in Chemistry*, World Scientific Publishing Co. Pte. Ltd, Singapore, 2004.