Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS
J. Serb. Chem. Soc. *86* (7–8) 673–684 (2021)
JSCS–5453
Original scientific paper

Use of GA-ANN and GA-SVM for a QSPR study on the aqueous solubility of pesticides

AMEL BOUAKKADIA[1,2]*, NOUREDDINE KERTIOU[1,2], RANA AMIRI[1], YOUSSOUF DRIOUCHE[1] and DJELLOUL MESSADI[1]

[1]*Environmental and Food Safety Laboratory, Department of Chemistry, Badji Mokhtar University – Annaba, BP. 12, 23000 Annaba, Algeria and* [2]*Abbes Laghrour University, Faculty of Sciences and Technology – Khenchela, BP 1252 Route de Batna, 40004 Khenchela, Algeria*

(Received 18 June, revised 19 August, accepted 6 October 2020)

Abstract: The partitioning tendency of pesticides, in this study herbicides in particular, into different environmental compartments depends mainly of the physicochemical properties of the pesticides itself. Aqueous solubility (S) indicates the tendency of a pesticide to be removed from soil by runoff or irrigation and to reach surface water. The experimental procedure for determining the aqueous solubility of pesticides is very expensive and difficult. QSPR methods are often used to estimate the aqueous solubility of herbicides. The artificial neural network (ANN) and support vector machine (SVM) methods, always associated with selection of a genetic algorithm (GA) of the most important variable, were used to develop QSPR models to predict the aqueous solubility of a series of 80 herbicides. The values of log S of the studied compounds were well correlated with the descriptors. Considering the pertinent descriptors, a Pearson correlation squared coefficient (R^2) of 0.8 was obtained for the ANN model with a structure of 5-3-1 and 0.8 was obtained for the SVM model using the RBF function for the optimal parameters values: C = 11.12; σ = 0.1111 and ε = 0.222.

Keywords: genetic algorithm; agrochemicals; descriptors; statistical methods.

INTRODUCTION

Pesticide distribution and fate in various environmental media and compartments is strongly influenced by the inherent properties of the compounds themselves, particularly by the basic physicochemical properties, such as solubility in water, vapor pressure (p_v) and partitioning coefficients between the organic matter (in soil or sediment) and water.[1]

* Corresponding author. E-mail: amelbouakkadia@yahoo.fr
https://doi.org/10.2298/JSC200618066B

The water solubility reflects the maximum amount of a chemical that will dissolve in pure water at a given temperature. The water solubility is one of the most important physicochemical properties in ecological hazard and exposure assessments, including environmental fate. The spatial and temporal movement (mobility) of a substance within and between the environmental compartments of soil, water and air depends largely on its solubility in water.

The knowledge of the solubility in water is essential when estimating biological degradation, bioaccumulation, hydrolysis, adsorption and the octanol/ /water partition coefficient. Highly water-soluble chemicals are potentially easier distributed by the hydrologic cycle, as they tend to have relatively low adsorption coefficients (*i.e.*, low adsorption to soil and sediment).[2] However, due to the complexity of analytical methods and the high cost of experiments, the application of theoretical predictive methods, which are fast, convenient and cost-effective, for preliminary assessment and estimation of aqueous solubility of pesticides have gained much attention.

Quantitative structure–property relationships (QSPR) models are found on the basis of the correlation between the experimental values of the physicochemical properties and descriptors reflecting the molecular structure of the compounds. In QSPR studies, a regression model of the form ($y = Xb + e$) may be used to describe a set of predictor variables (X) with a predicted variable (y) by means of a regression vector (b).[3]

The objective of the present study was to develop valid QSPR models for the aqueous solubility, S, of pesticides. Artificial neural network (ANN) and support vector machine (SVM) were applied for modeling the quantitative relationship between the aqueous solubility and the structural descriptors of pesticides. Previous to the generation of the ANN and SVM models, a genetic algorithm (GA) was used for descriptor selection. The strength and the predictive performance of the proposed models were verified using both internal (cross-validation and Y-scrambling) and external statistical validations.

EXPERIMENTAL

Data set

The data set in this study comprises diverse chemical classes of 80 pesticides. The data were collected from the literature,[2] and were compiled in the units of mg L^{-1}, and presented as the logarithm of S, the values of which ranged from -1.04576 to 5.90091.

The data set was randomly divided into a training set of 58 compounds and a test set of 22 compounds. The names of the compound and their aqueous solubility are given in Table S-I of the Supplementary material to this paper.

Software

Geometry optimization was performed using HyperChem 6.03.[4] Dragon software was utilized to calculate the molecular descriptors. ANN and SVM regression were performed in the Molegro software.[5]

Molecular optimization and descriptor generation

The molecular modeling software HyperChem 6.03[4] was used to represent the molecules and then, the semi-empirical AM1 method was used to obtain the final geometry. All calculations were performed under RHF formalism[6] without configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm for criterion with a root mean square gradient of 0.001 kcal* mol^{-1}. The optimized geometries were then transferred to the Dragon computer software Version 5.3[7] to calculate 1201 descriptors belonging to different classes. Using the corresponding DRAGON software options, constant values of the descriptors (standard deviations less than 0.0001), which provide no information, were eliminated first and then those that were highly correlated ($R \geq 0.95$) as they convey redundant information. For each pair of correlated descriptors, the one with the highest cross-correlation with other descriptors was automatically eliminated. In this work, the genetic algorithm (GA) variable subset selection method in the MOBYDIGS release of Todischini[8] was used for the selection of the most relevant descriptors from the pool of remaining descriptors. These descriptors were used as inputs of ANN and SVM for the construction of QSPR models, as in similar work.[9]

Genetic algorithm

Modeling of the genetic process initiated the development of genetic algorithms, which could be exploited in a variety of optimization problems.[10,11] In this case, a potential solution is considered as an individual in a population. The value of the cost function associated with a measurement solution "adaptation" of the individual related to its environment. A genetic algorithm simulates the evolution, over several generations, of an initial population whose individuals are poorly adapted using genetic operators of reproductions and mutations. After a number of generations, the population consists of well adapted individuals, *i.e.*, the supposed "good" solutions to the optimization problem. 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 $n/m \geq 5$.[12] The GA was stopped when increasing the model size did not significantly increase the Q^2 value.

Artificial neural networks

An artificial neural network is modeled on the biological neural networks found in the central nervous system of animals.[13] An artificial neural network is a mathematical and numerical method based on biological neural networks. An ANN consists of some connected neurons and process information. A network is made of one input layer, one output layer and may also consist of some hidden layers. Each layer is made of some neurons connected to other neurons in previous and subsequent layers. A neuron has an input, an output and a transfer function. The sigmoidal transfer function is one of the performed functions, expressed by the following equation:

$$a_j = \frac{1}{1+e^{-S_j}} \tag{1}$$

where a_j is the output of the j^{th} neuron and S_j is the input of the j^{th} neuron, produced by outputs of the previous layer. S_j is given as:

$$S_j = \sum_{i=1}^{n}\left(w_{ij}a_i\right) + b_j \tag{2}$$

* 1 kcal = 4184 J

where a_i are the outputs of the i^{th} neuron from the previous layer, w_{ij} presents the weights applied to the connection of the i^{th} and j^{th} neuron, and b_j is a bias number.[14]

An ANN is an adaptive network that changes its structure based on external or internal information that flows through the network during the learning (training) phase. Estimation of optimum weights and biases of network needs an algorithm, called the propagation method. Several kinds of propagation methods are available and back propagation (BP) is the easiest and simple one with enough reliability. BP and other usual propagation methods are explained completely in mathematical literatures.[15,16]

Artificial neural networks are capable of recognizing highly non-linear relationships contrary to MLR. When compared with the traditional statistical methods, the flexibility of an ANN enables more complex relationships in the experimental data to be discoverd.[17]

Support vector machine

SVM is a new and very promising classification and regression method developed by Vapnik.[18] SVMs were originally developed for classification problems; they can also be extended to solve nonlinear regression problems by the introduction of an ε-insensitive loss function. In the support vector regression, input x is first mapped into a higher dimensional feature space by using 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:[19,20]

$$f(x,w) = f\left(x,\alpha_i,\alpha_i^*\right) = \sum_{i=1}^{N}\left(\alpha_i - \alpha_i^*\right)K\left(x_i,\alpha_j\right) + b \tag{3}$$

where both α_i and α_i^* are Lagrange multipliers. According to the Karush–Kuhn–Tucker conditions, only a minority sample coefficients are non-zero values, the data points corresponding to them are called support vectors. These support vectors are the samples that can determine the hyper plane.[13,21] $K(x,x_i)$ is the Kernel function.[22] Any function satisfying the Mercer condition can be used as the Kernel function.[23] In this work, the Gaussian radial basis function (RBF) Kernel was used in the SVM as below:

$$K(x,x_i) = e^{-\|x-x_i\|^2/\sigma^2} \tag{4}$$

where σ^2 is the width of the Gaussian function, so 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 SVMR 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 11.12, 0.1111 and 0.222, respectively, obtained based on the minimum *RMSE* and maximum accuracy of the model.

A detailed description of the theory of SVM can be found in several excellent books and tutorials.[24,25]

An additional external validation according to Golbraikh and Tropsha[26] was applied solely to the test set. According to the recommended criteria of Tropsha *et al*., a predictive QSPR model must meet the following conditions:[27]

$$Q^2_{EXT} > 0.5 \tag{5}$$

$$R^2 > 0.6 \tag{6}$$

$$(R^2 - R^2_0)/R^2 < 0.6 \text{ and } 0.85 < k < 1.15 \tag{7}$$

or

$$(R^2–R'^2_0)/R^2 < 0.6 \text{ and } 0.85 < k' < 1.15 \quad (8)$$

where

$$R = \frac{\sum(y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\tilde{y}_i - \bar{\tilde{y}})^2}}, R_0^2 = 1 - \frac{\sum(y_i - y_i^{r0})^2}{\sum(y_i - \bar{y})^2}, R_0'^2 = 1 - \frac{\sum(\tilde{y}_i - \tilde{y}_i^{r0})^2}{\sum(\tilde{y}_i - \bar{\tilde{y}})^2},$$

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

where R is the correlation coefficient between the calculated and experimental values in the test set; R_0^2 (calculated *vs*. observed values) and $R_0'^2$ (observed *vs*. calculated values) are the coefficients of determination; k and k' are slopes of regression lines through the origin of calculated *vs*. observed and observed *vs.* calculated, respectively; y_i^{r0} and $\tilde{y}_i^{r0}$ are defined as:

$$y_i^{r0} = k\tilde{y} \quad (5)$$

and

$$\tilde{y}_i^{r0} = k'y \quad (6)$$

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 *vs.* predicted properties are compared, an exact fit is required, not just a correlation.

RESULTS AND DISCUSSION

The GA parameters that were used in this study are: population size 100 and maximum generations 100. Five descriptors were selected by GA (Table S-II of the Supplementary material), which are: MATS8m, RNCG, AlogP2, MAXDN and Mor26u. Table I gives e short descriptions of these descriptors.

TABLE I. Description of the selected descriptors by GA

No	Symbol	Class	Meaning
1	MATS8m	2D autocorrelation indices	Moran autocorrelation – lag 8 / weighted by atomic masses
2	RNCG	Charge descriptors	Relative negative charge
3	AlogP2	Molecular properties	Squared Ghose–Crippen octanol–water partition coeff. log P^2
4	MAXDN	Topological descriptors	Maximal electrotopological negative variation
5	Mor26u	3D-MoRSE descriptors	3D-MoRSE – signal 26 / unweighted

Artificial neural network

The inputs of the ANN were a subset of the descriptors selection by genetic algorithm from a large set of descriptors. The input layer comprised five descriptors. Usually one hidden layer is enough. After several trials, a hidden layer with three neurons was selected; Fig. 1 explains this choice. The calculated sol-

ubility values (log *S*) constitute the output layer (*e.g.*, the NN architecture is 5-3-1). A sigmoid transfer function $1/(1+e^x)$ was used. After 212 epochs, a correlation coefficient of 0.86 (n = 58) between calculated and observed log (S / mg L^{-1}) was obtained.

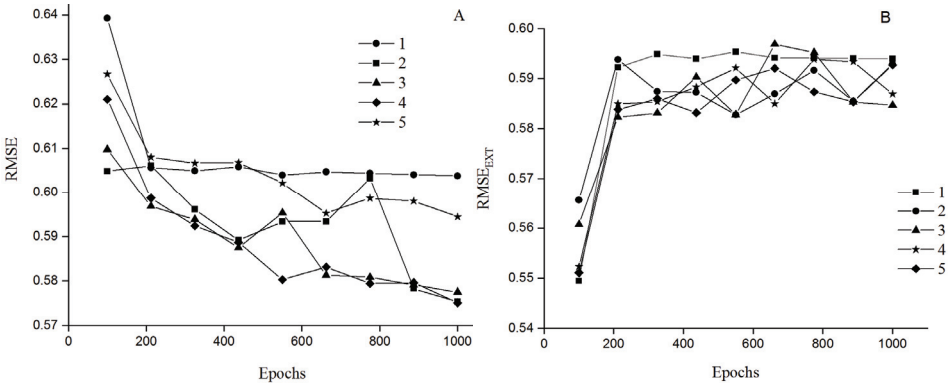

Fig. 1. A – RMSE and B – $RMSE_{EXT}$ values *vs.* number of neurons of the hidden layer and the number of epochs.

The general statistical parameters selected to evaluate the prediction ability of the constructed model are: root mean squared error for the training set (*RMSE*), root mean squared error for the external data set ($RMSE_{EXT}$), cross validated squared CC (Q^2) and the Pearson correlation squared (R^2).

Thus, a 5-3-1 architecture was obtained, the statistic parameters of the final optimal model are given in Table II.

TABLE II. Results and statistical parameters of GA-ANN and GA-SVM

Parameter	Set	GA-ANN	GA-SVM
Number of neurons	–	3	–
Epochs	–	212	–
C	–	–	11.12
E	–	–	0.222
Σ	–	–	0.1111
Q^2_{LOO}	–	0.7748	0.7125
R^2	Training set	0.8097	0.8403
	Validation set	0.7412	0.7068
RMSE	Training set	0.5968	0.5933
	Validation set	0.5823	0.5782

The plot shown in Fig. 2 indicates that there was a significative correlation between calculated and observed log *S* values (Table S-II of the Supplementary material), which shows a weak dispersion of the points around the first bisectrix, which confirm the acceptable performance especially for rough estimations.

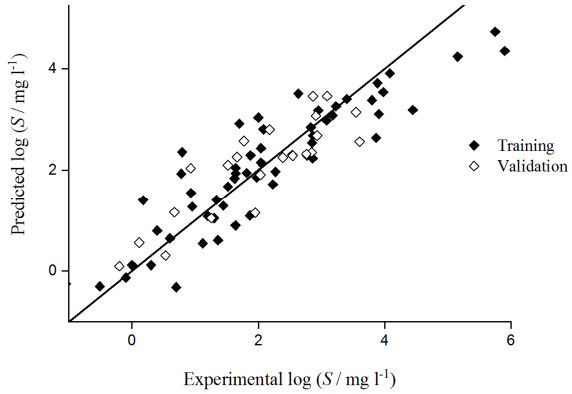

Fig. 2. Predicted values *vs.* experimental values for the training and validation sets.

Support vector machine

The quality of SVM for the 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. Optimization of the SVM parameters was performed by systematically changing their values in the training set and calculating the *RMSE* of the model using 5-fold cross validation.

First, with 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 is optimized. To find an optimal value of C, the *RMSE* of SVM models with different C values was calculated. The variation of *RMSE vs*. C values was plotted in Fig. 3. As shown in this figure, the optimal value of C was 11.12. The optimal values of the three parameters and the final optimal model were determined and are given in Table II.

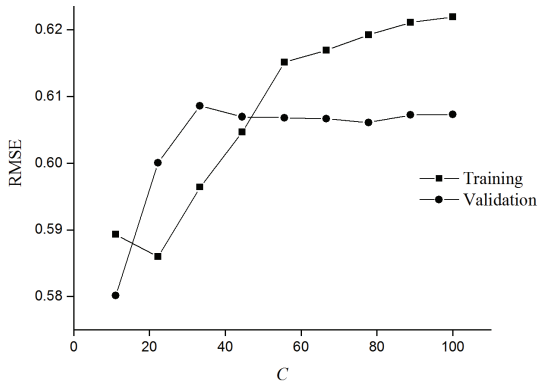

Fig. 3. Variation of *RMSE vs. C* values.

The predicted log *S* values are plotted against the experimental values (Table S-II of the Supplementary materials) in Fig. 4. The predicted values are, in general, in good agreement with the corresponding experimental values.

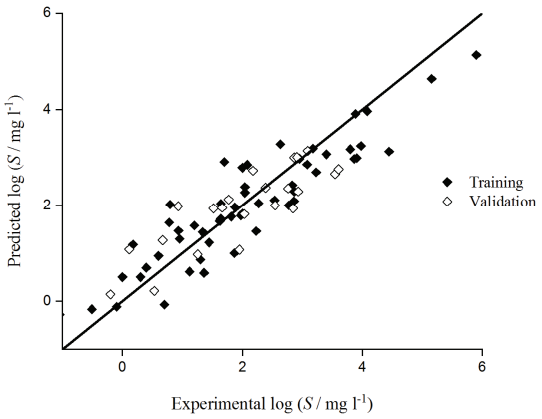

Fig. 4. Predicted values *vs*. experimental values for the training, validation sets.

The statistical parameters show that the models established a strong correlation between the five selected variables and the studied property, characterized by excellent parameters $Q^2_{LOO} > 0.5$.[28] in addition to a good standard error ($RMSE_{ANN}$ = 0.5968, $RMSE_{SVM}$ = 0.5933). All statistical parameters of the model are satisfactory and prove that the models are stable, robust and predictive. Then, the built models were used to predict the test set data.

Analysis of descriptors contribution in the ANN model and interpretation

To evaluate the influence of each descriptor on the calculated solubility, the relevance score was used. The relevance score is calculated by following all paths from the input neuron to the output neuron (including hidden layers). For each path, the product of all the connection weights (in absolute values) were added to the score. Afterwards, all relevance scores are normalized to be in the range between 0 and 100.[5]

The relative value of contributions of the five descriptors of the model was determined (Table III). These values of contributions allow the following classification: RNCG > AlogP2 > MAXDN > MATS8m > Mor26u. It should be noted that the difference in the descriptor contribution between any two descriptors used in the model is not significant, indicating that all of the descriptors are indispensable in generating the predictive model.

These values confirm the great effect of the RNCG and AlogP2 on the solubility. Another advantage of this method is the determination of the impact of each descriptor on the aqueous solubility.

The relative negative charge (*RNCG*), the first important descriptor, is the charge of the most negative atom divided by the total negative charge (Q_{neg}). The

charge density of the ions plays an important role in the interactions of these ions and the water molecule, which reflects the influence of the negative charge on the aqueous solubility. The second important descriptor, the squared Ghose–Crippen–Viswanadhan octanol–water partition coefficient (AlogP2) is calculated from a regression equation based on the hydrophobic character of the molecule. It reflects both the interactions of the solute with the bulk of the surrounding solvent (macroscopic or non-specific solvent effects) and the specific bonding between the solute and individual solvent molecules (microscopic or specific solvent effects).[29]

TABLE III. Relevance score

Index	Name	Relevance score
0	RNCG	100
1	ALOGP2	76
2	MAXDN	39
3	MATS8m	32
4	Mor26u	28

MAXDN, a topological descriptor, is the maximal electrotopological negative variation (MAXDN) is calculated as the maximum negative value of DI_i (topologic distance) in the molecule.

Mats8m is a 2D autocorrelation indices that is calculated by applying the Moran coefficient to the molecular graph.[7]

Mor26u is a descriptor of the 3D-MoRSE descriptor class, the 3D-MoRSE descriptors are 3D molecular representations of structure based on electron diffraction descriptor,[30,31] that are calculated by summing atomic weights viewed by a different angular scattering function. The values of these descriptor functions are calculated at 32 evenly distributed values of scattering angle (s) in the range of 0–31 Å from the three-dimensional atomic coordinates of a molecule. The 3D-MoRSE descriptor is calculated using following expression:

$$\text{Morsw} = \sum_{i=1}^{n\text{AT}-1} \sum_{j=i+1}^{n\text{AT}} w_i w_j (\sin(sr_{ij}) / sr_{ij}) \quad (16)$$

where s is the scattering angle, nAT is the number of atoms, r_{ij} is the interatomic distance between the i^{th} and the j^{th} atoms, w is an atomic property, including atomic number, masses, van der Waals volumes, Sanderson electronegativities and polarizabilities.

The statistical parameters obtained for the test set,[32] demonstrate the power of the predictivity of the models.

Comparison of the results with other modeling methods

The present results were compared with those obtained in previous publications using other modeling methods. The comparisons are summarized in Table

IV, which shows that the presented SVM model gives better predictions than most of the other methods, because fewer descriptors were used than in the the other models, and an almost similar result was obtained. In addition, the herein presented models were evaluated using different statistical parameters compared with the other model in the literature.[33]

TABLE IV. Comparison of the presented results with those obtained using other modeling methods

Reference	Method	Test set	Training set	Number of descriptors	R2tr	R2 test	*RMSE*
Our results	ANN	22	58	5	80.97	74.12	0.5968
	SVM	22	58	5	84.03	70.68	0.5933
Deeb and Goodarzi[3]	PLS	–	219	22	79.98	79.44	–
	PC-ANN	–	219	22	84.35	81.93	–
Bouakkadia *et al*.[29]	MLR	19	58	6	88.95	85.11	0.5200

The difference between this work and the previously published work on this data set is that the number of compounds in the validation set was not the same and also the two training and validation sets did not contain the same compounds, because the method of separation was not the same. Additionally, the descriptors selected by the genetic algorithm are different except for AlogP2.

CONCLUSIONS

A quantitative structure–property relationship analysis was performed on the logarithm of the solubility in water for 80 pesticide compounds using ANN and SVM. The built models clearly demonstrate good correlations between the structure and aqueous solubility of the studied compounds. Five descriptors were selected with genetic algorithm. The selected descriptors, *i.e*., MATS8m, RNCG, AlogP2, MAXDN and Mor26u, were found to be important factors controlling the aqueous solubility. Comparison between the ANN and SVM methods demonstrates that the performance of SVM model is better than that of ANN, but the ANN model is more general than SVM because of the great value of R^2_{test}.The proposed models will help identifying new pesticides and provide insight to guide their development and may be useful for predicting their solubility.

SUPPLEMENTARY MATERIAL

Additional data are available electronically at the pages of journal website: https://www.shd-pub.org.rs/index.php/JSCS/index, or from the corresponding author on request.

ИЗВОД

КОРИШЋЕЊЕ GA-ANN И GA-SVM ЗА QSPR СТУДИЈУ РАСТВОРЉИВОСТИ ПЕСТИЦИДА У ВОДИ

AMEL BOUAKKADIA[1,2], NOUREDDINE KERTIOU[1,2], RANA AMIRI[1], YOUSSOUF DRIOUCHE[1] и DJELLOUL MESSADI[1]

[1]*Environmental and Food Safety Laboratory, Department of Chemistry, Badji Mokhtar University – Annaba, BP. 12, 23000 Annaba, Algeria и* [2]*Abbes Laghrour University, Faculty of Sciences and Technology – Khenchela, BP 1252 Route de Batna, 40004 Khenchela, Algeria*

Тежња пестицида, у овој студији нарочито хербицида, за расподелу по различитим одељцима животне средине, зависи углавном од физичкохемијских особина самих пестицида. Растворљивост у води (S) указује на тенденцију пестицида да се уклоне испирањем или иригацијом да би завршили у површинским водама. Експериментални поступак за одређивање растворљивости пестицида у води је веома скуп и тежак. QSPR методе се често користе за процену растворљивости хербицида у води. Методе вештачке неуронске мреже (ANN) и векторске машине за подршку (SVM), сваки пут повезане са селекцијом помоћу генетичког алгоритма (GA) за избор најзначајније варијабле, биле су коришћене за развој QSPR модела за предвиђање растворљивости у води серије од 80 хербицида. Вредности $\log S$ проучаваних једињења добро су корелисане са дескрипторима. Разматрајући погодне дескрипторе, квадратни Пирсонов коефицијент (R^2) 0,8 добијен је за ANN модел за структуру 5-3-1, а 0,8 је добијен за SVM модел користећи RBF функцију за оптималне вредности параметара: $C = 11{,}12$; $\sigma = 0{,}1111$ и $\varepsilon = 0{,}222$.

(Примљено 18. јуна, ревидирано 19. августа, прихваћено 6. октобра 2020)

REFERENCES

1. P. Gramatica, A. Di Guardo, *Chemosphere* **47** (2002) 947 (https://doi.org/10.1016/S0045-6535(02)00007-3)
2. O. C. Hansen, *Quantitative Structure–Activity Relationships (QSAR) and Pesticides*, Pesticides Research No. 94, Teknologisk Institute, Taastrup, 2004 (https://www2.mst.dk/udgiv/publications/2004/87-7614-434-8/html/helepubl_eng.htm)
3. M. Zine, A. Bouakkadia, L. Lourici, D. Messadi, *J. Serb. Chem. Soc*. **84** (2019) 1405 (https://doi.org/10.2298/JSC190306059Z)
4. *Hyperchem*TM, Release 6.03 for Windows, Molecular Modeling system, 2000 (http://www.hyper.com)
5. *Molegro Data Modeller (*MDM), v.2.0. Copyright Molegro (2009) (https://www.scientific-computing.com/press-releases/molegro-data-modeller-v20)
6. I. N. Levine, *Quantum Chemistry*. 5thed., Prentice Hall, Hoboken, NJ, 2000, pp. 626–739 (ISBN-10:0136855121)
7. R. Todeschini, V. Consonni, M. Pavan, Dragon, *Software for the Calculation of Molecular Descriptors,* Release 5.3 for windows, Milano, 2006 (http://www.talete.mi.it)
8. R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan, *MOBYDIGS*, *Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm*, Release 1.1 for Windows, Milano, 2009 (http://www.talete.mi.it/)
9. R. Amiri, D. Messadi, A. Bouakkadia, *J. Serb. Chem. Soc*. **85** (2020) 467 (https://doi.org/10.2298/JSC190610090A)
10. D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley, London, 1989, pp. 89–94 (https://dl.acm.org/citation.cfm?id=534133)
11. A. Beheshti, E. Pourbasheer, M. Nekoei, S. Vahdani, *J. Saudi Chem. Soc.* **28** (2016) 282 (http://dx.doi.org/10.1016/j.jscs.2012.07.019)

12. 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)
13. S. Jothilakshmi, V. N. Gudivada, *Handbook of Statistics,* Vol. 35, Ch. 10, Elsevier, Amsterdam, 2016, pp. 301–340 (http://dx.doi.org/10.1016/bs.host.2016.07.005)
14. M. Safamirzaei, H. Modarress, M. Mohsen-Nia, *Fluid Phase Equilib.* **266** (2008) 187 (https://doi.org/10.1016/j.fluid.2008.01.022)
15. S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice-Hall, New Delhi, 2006 (ISBN-13: 978-0023527616)
16. R. Rojas, *Neural Network*, Springer-Verlag, Berlin, 1996, pp. 229–261 (http://page.mi.fu-berlin.de/rojas/neural/)
17. O. Deeb, M. Drabh, *Chem. Biol. Drug. Des*. **76** (2010) 255 (https://dx.doi.org/10.4236/aces.2012.21010)
18. V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York,1998, pp. 375–473 (ISBN: 978-0-471-03003-4)
19. V. N. Vapnik, *The Nature of Statistical Learning Theory.* Springer, New York, 2000, pp. 267–287 (ISBN: 978-1-4757-3264-1)
20. C. J. Lu, T. S. Lee, C. C. Chiu, *Decision Support. Syst*. **47** (2009) 115 (https://doi.org/10.1016/j.dss.2009.02.001)
21. N. Cristianini, J. Shawe-Taylor, *An introduction to support vector machines and other kernel- based learning methods*, Publishing House of Electronics Industry, Beijing, 2005, pp. 93–122 (ISBN: 0521 780195).
22. R. Amiri**,** D. Messadi, A. Bouakkadia**,** L. Lourici, *Egypt. J. Chem.* **62** (2019) 1563 (https://doi.org/ 10.21608/ejchem.2019.4976.1446)
23. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer*,* New York, 1995, pp. 70–71 (ISBN: 0-387-94559-8).
24. N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2000, pp. 32–42 (ISBN: 0-521-78019-5).
25. B. Scholkopf, A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002, pp. 13–17 (ISBN: 0262194759)
26. A. Golbraikh, A. Tropsha, *J. Mol. Graph. Model*. **20** (2002) 269 (https://doi.org/10.1016/S1093-3263(01)00123-1)
27. N. Kertiou, A. Bouakkadia**,** D. Messadi, *Res. J. Pharm. Biol. Chem. Sci.* **8** (2017) 251 (https://www.rjpbcs.com/pdf/2017_8(6)/[29].pdf)
28. A. Bouakkadia**,**Y. Driouche, N. Kertiou, D. Messadi, *Int. J. Saf. Secur. Eng.* **10** (2020) 389 (https://doi.org/10.18280/IJSSE.100311)
29. A. Bouakkadia, H. Haddag, N. Bouarra, D. Messadi, *Synthèse: Revue Sci. Technol.* **32** (2016) 12 (https://www.ajol.info/index.php/srst/issue/view/13942)
30. J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, *J. Chem. Inf. Comput. Sci*. **36** (1996) 1030 (https://doi.org/10.1021/CI960343+)
31. J. Schuur, P. Selzer, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **36** (1996) 334 (https://doi.org/10.1021/CI950164C)
32. W. J. Wang, Z. B. Xu, W. Z. Lu, X. Y. Zhang, *Neurocomputing* **55** (2003) 643 (https://doi.org/10.1016/S0925-2312(02)00632-X)
33. O. Deeb, M. Goodarzi, *Mol. Phys.* **108** (2010) 181 (https://doi.org/10.1080/00268971003604575).