Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS
J. Serb. Chem. Soc. *84* (6) 575–590 (2019)
JSCS–5210
UDC 539.16+544.022:547.562–145.2:510.644
Original scientific paper

Modeling of linear and nonlinear quantitative structure property relationships of the aqueous solubility of phenol derivatives

SOUMAYA KHEROUF[1], NABIL BOUARRA[1,2], AMEL BOUAKKADIA[1,3] and DJELLOUL MESSADI[1]*

[1]*Laboratory of Environmental and Food Safety, Department of Chemistry, Badji Mokhtar–Annaba University, PB12, 23000, Annaba. Algeria,* [2]*Center of Scientific and Technical Research in Physicochemical Analyzes (CRAPC), Bp 384, Siège ex-Pasna Zone Industrielle, Bou-Ismail, 42004, Tipaza, Algeria and* [3]*University Abbes Laghrour Khenchela – Algeria – BP 1252 Route de Batna Khenchela 40004, Algeria*

(Received 20 August, revised 16 December 2018, accepted 11 February 2019)

Abstract: Quantitative structure–solubility relationships (QSSR) are considered as a type of Quantitative structure–property relationship (QSPR) study in which aqueous solubility of chemicals are related to chemical structure. In the present work, multiple linear regression (MLR) and artificial neural network (ANN) techniques were used for QSSR studies of the water solubility of 68 phenols (phenol and its derivatives) based on molecular descriptors calculated from the optimized 3D structures. By applying missing value, zero and multicollinearity tests with a cutoff value of 0.95, and a genetic algorithm (GA), the descriptors that resulted in the best fitted models were selected. After descriptor selection, multiple linear regression (MLR) was used to construct a linear QSSR model. The R^2 = 91.0 %, Q^2_{LOO} = 89.33 %, s = 0.340 values of the model developed by MLR showed a good predictive capability for log S values of phenol and its derivatives. The results of MLR model were compared with those of the ANN model. the comparison showed that the R^2 = 94.99 %, s = 0.245 of ANN were higher and lower, respectively, which illustrated an ANN presents an excellent alternative to develop a QSSR model for the log S values of phenols to MLR.

Keywords: QSPR; aqueous solubility; phenols; multiple linear regression; artificial neural network.

INTRODUCTION

Phenols compounds are nearly ubiquitous pollutant in all ecosystems[1] because they are basic materials for industry production and agriculture, which are commonly used in chemical synthesis. They can spread through air and water, with strong carcinogenicity, teratogenicity and mutagenicity,[2,3] which cause great dam-

* Corresponding author. E-mail: d_messadi@yahoo.fr
https://doi.org/10.2298/JSC180820016K

age to the environment, plants, animals and human health. The compounds penetrate ecosystems as the result of drainage of municipal or industrial sewage to surface water.[4] Therefore, it is vital to protect the environment and prevent their behavior by studying their physicochemical properties.

Aqueous solubility is the concentration of a chemical in the aqueous phase, when the solution is in equilibrium with the pure compound in its usual phase (gas, liquid or solid) under standard conditions of temperature and pressure.[5] Aqueous solubility is one of the major physiochemical properties to be optimized in pharmaceutical and environmental studies; it is related to absorption and distribution in ADME-Tox (absorption, distribution, metabolism, excretion-toxicity).

Experimental determination of compound solubility is not easily managed, or even possible, when working with large chemical libraries.[6] Alternately, the quantitative structure–property relationship (QSPR) provides a promising method for the prediction of solubility using descriptors derived solely from the molecular structure to fit experimental data. The QSPR approach attempts to establish simple mathematical relationships to describe the correlation of a given property to molecular structures for a set of compounds.[7]

The advantage of this method lies in the fact that it requires only knowledge of the chemical structure and is not dependent on any experimental property. Moreover, it could be used for the prediction of the properties of new compounds.

Recently, several works reported QSPR studies on the solubility and other properties. Warne *et al.*[8] reported the utility of thirty-nine molecular descriptors and physicochemical properties to model the solubility (S) and octanol–water partition coefficient (K_{ow}) of thirty-one lipophilic organic compounds, which was assessed using least squares regression analysis. Several novel molecular descriptors that yielded high correlation in linear regression equations were the approximate sigma electron density, radius of gyration and the first and second principle moments of inertia. Other useful properties were the sum of all absolute valency charges, density at 20 °C, liquid density and pK_a. Solubility data for 930 diverse organic compounds have been analyzed using linear partial least square (PLS) and nonlinear PLS methods, continuum regression (CR) and neural networks (NN). 1D and 2D descriptors from the Molecular Operating Package (MOE) package in combination with E-state or ISIS keys have been used. The combination between 22 descriptors for MOE with a subset of 65 ISIS keys, used in linear PLS methods, provided the best statistics for the given set of compounds,[9] while ZAHRA Garkani Nejad[10] modeled retention times of phenol derivatives by using linear methods such as PLS and MLR and three different ANN methods, such as learning rate (GDX-ANN), resilent back propagation (RP-ANN) and Levenberg–Marquardt (ML-ANN).

The main objective of this work was to establish new QSSR models for predicting the solubility of phenols from the theoretical derived molecular descript-

tors. Two modeling techniques, *i.e*., MLR and ANN, were used to construct linear and nonlinear models. The performances of the obtained models were compared with each other. Furthermore, the applicability domain (AD) was analyzed based on the Williams plot for the MLR model.

EXPERIMENTAL

Data set

In the present study, the values of log *S* were mainly harvested from the handbook of physicochemical properties and environmental fate for organic chemicals.[11] The experimental data were available for phenol and 68 of its derivatives. The values of log *S* ranged from 0.73 to 5.04. A complete list of the compounds names, and corresponding experimental and predicted solubility by MLR and ANN methods are given in Table I.

Descriptor generation

The chemical structure of each compound was sketched on a PC using the HYPERCHEM program 6.03,[12] 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 0.001 kcal* mol^{-1} as the stopping criterion. The molecular structures were used as input for the generation of 1664 descriptors using Dragon software 5.3[13] for the computation of different types (geometric, topological, 2D autocorrelation, *etc*.). To reduce redundant and non-useful information, constant or near constant values and descriptors found to be highly correlated pairwise (one of any two descriptors with a correlation greater than 0.97 were excluded in a pre-reduction step. Thus, 527 descriptors remained to undergo subsequent descriptor selection.

Selection of the training and test sets (splitting)

It is important to define rationally 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 could 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.[14] In this work, four different splitting techniques were applied: a) random splitting, b) sorted response splitting, c) structural similarity ordered by the first axis of principal component analysis (PCA, PC1 score)[15] and d) by using the Kennard and Stone algorithm (CADEX).[16]

Descriptor selection and model development

Multiple linear regression analysis and variable selection were performed with the help of QSARINS software (version 2.2)[17] using the ordinary least squares (OLS) method and genetic algorithm/variable subset selection (GA/VSS). This "variable selection" procedure generates a "population" of models, ranked according to decreasing R^2 values. The best models were chosen by using Q^2_{LOO} (leave-one-out) as the optimization value, and taking into account the parsimony principle regarding the complexity of the models, which should be as small as possible. For this reason, only up to three descriptors were included in the QSAR generated in this study. Furthermore, the correlation between the modeling descriptors and the

* 1 kcal = 4184 J

modeled response was checked by the Q under influence of K (QUIK) rule, to exclude models with high predictor collinearity and exclude chance correlation.[18]

TABLE I. log (S / g m^{-3}) values, experimental and calculated by MLR and ANN models for 68 phenols

Name	Model		
	Experimental	MLR	ANN
Naphthalen-1-ol	2.6415	2.859	2.717
2,3,4,5-Tetrachlorophenol	2.2201	2.007	2.240
2,3,5,6-Tetrachlorophenol	2.0000	2.337	2.208
2,3,5-Trimethylphenol	2.9031	3.023	2.894
2,3,6-Trichlorophenol	2.6532	2.665	2.708
2,4,6-Trimethylphenol	3.0792	3.094	2.990
Naphthalene; 2,4,6-trinitrophenol	4.1383	3.323	3.494
2,4-Dinitrophenol	2.5250	3.335	3.056
2,6-Dichlorophenol	3.4191	3.157	3.116
2,6-Dimethylphenol	3.7945	3.904	4.044
2-Methoxyphenol	4.3945	4.185	4.233
5-*tert*-Butyl-2-methylphenol	2.6128	2.523	2.646
3-Nitrophenol	4.0626	3.801	3.805
3-*tert*-Butylphenol	3.3160	3.057	2.951
4,5-Dichloro-2-methoxyphenol	2.8500	2.746	2.576
2-Methyl-4,6-dinitrophenol	2.3464	2.188	2.375
4-Butylphenol	2.7903	2.927	2.967
4-Chloro-2-methoxyphenol	3.7300	3.894	4.092
4-Chlorophenol	4.4314	4.771	4.457
4-Hexylphenol	2.5922	2.044	2.130
4-Propan-2-ylphenol	3.5136	3.351	3.376
4-Methoxyphenol	4.2900	4.464	4.499
4-Nitrophenol	4.1303	3.948	4.191
4-Nonylphenol	0.7348	0.715	0.781
4-Octylphenol	1.1004	1.163	1.243
4-Phenylphenol	0.9912	2.151	1.509
4-Propylphenol	3.2375	3.357	3.371
4-Butan-2-ylphenol	2.9823	3.009	2.951
4-*tert*-Butylphenol	2.7634	2.974	2.956
Benzene-1,2-diol	4.6532	4.572	4.889
Benzene-1,4-diol	4.8451	5.090	4.504
2-Methylphenol	4.4150	4.492	4.648
2-Ethylphenol	4.1474	4.069	4.260
4-Methylphenol	4.3010	4.267	4.341
2,3,4,5,6-Pentachlorophenol	1.1461	1.595	1.336
4-Ethylphenol	3.9020	3.831	3.764
Phenol	4.9463	5.071	4.749
2,3,4,5-Tetrachloro-6-methoxyphenol	1.4150	1.390	1.372
Naphthalen-2-ol	2.8692	2.863	2.743
2-Nitrophenol	3.0334	3.450	3.188
3-Ethyl-5-methylphenol	3.3644	3.316	3.565

TABLE I. Continued

Name	Model		
	Experimental	MLR	ANN
2-Phenylphenol	2.8451	2.474	2.915
3,4,5-Trichloro-2-methoxyphenol	2.4914	1.982	2.080
3,4-Dichlorophenol	3.9664	3.564	3.713
3,5-Dichlorophenol	3.8689	3.976	4.098
3,5-Dimethylphenol	3.7404	3.649	3.875
3,5-Di-*tert*-butylphenol	1.1461	1.170	1.008
3-Methoxyphenol	4.8312	4.348	4.534
2,3,5,6-Tetrachlorophenol	2.2625	2.073	2.249
2,3,5-Trichlorophenol	2.6990	2.800	2.786
2,3-Dichlorophenol	3.9146	2.969	3.199
2,3-Dimethylphenol	3.7782	3.719	3.954
2,4,5-Trichlorophenol	2.9768	3.179	3.268
2,4,6-Trichlorophenol	2.6375	3.031	2.937
2,4-Dichlorophenol	3.6532	3.533	3.604
2,4-Dimethylphenol	3.9442	3.684	3.924
2,5-Dimethylphenol	3.5019	3.776	3.985
2-Chlorophenol	4.3918	4.110	4.272
2-Propan-2-ylphenol	3.6457	3.628	3.606
2,3,4-Trichlorophenol	2.6990	2.458	2.580
2,3,4-Trichloro-6-methoxyphenol	1.7324	1.436	1.815
4,5-Dichloro-2-methoxyphenol	2.7597	2.612	2.454
5-Chloro-2-methoxyphenol	3.5977	3.553	3.796
3-Methylphenol	4.3424	4.320	4.458
Benzene-1,3-diol	5.0414	4.647	4.961
3,4,5-Trimethylphenol	3.1875	2.909	2.882
3,4-Dimethylphenol	3.7076	3.563	3.632
3-Chlorophenol	4.3424	4.267	4.345

The best modeling descriptors were selected using the all subset procedure of QSARINS software.[17,18] The search of the best solution is realized by maximizing a selected fitness function, in the present case Q^2_{LOO}.

Artificial neural network (ANN) method

An ANN is suitable for modeling a nonlinear relationship; the major advantage is that models can be developed without knowing the exact form of the analytical function on which the model should be built. The theory of ANN and application in QSPR studies has been extensively discussed in many reviews.[19,20]

The descriptors selected from MLR were submitted to a three-layer feed-forward ANN with a back propagation learning algorithm[21] for the prediction of the solubility of 68 phenols. The number of hidden neurons was optimized by a trial and error procedure on the training process. One output neuron was used to represent the experimental solubility.

Model evaluation and validation

Model performance was evaluated by the following statistical parameters (R^2 and Q^2_{LOO}). R^2 is the coefficient of determination in the training set that measures the adequacy

between the model and the observed data. In other words, R^2 verifies the goodness-of-fit of the developed model. For the quantitative assessment of model's robustness, the cross-validation (CV) coefficient (Q^2_{LOO}), which is one of the most commonly applied internal validation techniques, was calculated. In this method, the process of removing a molecule, and creating and validating the model against the individual molecules was performed for the entire training set.[22-24]

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

$$Q^2_{\mathrm{LOO}} = 1 - \frac{\sum_{i=1}^{n} \left(\hat{y}_{i/i} - y_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} \tag{2}$$

where y_i is the observed dependent variable (the experimental response), $\hat{y}_i$ is the value calculated by the model, $\bar{y}$ is the mean value of the dependent variable, n is the number of compounds in the training set, and $\hat{y}_{i/i}$ is the value of log S predicted by the model built without compound i according to the LOO method.

The leave-many-out (LMO) procedure is a strong internal validation. By design, model validation by LMO employs smaller training sets than the LOO procedure and can be repeated many more times due to the possibility of larger combinations in leaving many compounds out from the training set. The premise being that if a QSPR model has a high average in the Q^2_{LOO} validation, it could reasonably be concluded that the obtained model is robust.[18] In the present case, 20 of the chemicals were set aside from the training set with 2000 iterations.

The predictive power of the developed model was evaluated by the external validation set on a series of coefficients: R^2_{Val} (characterizing the correlation between the predicted and experimental values in the validation set) the coefficients Q^2_{F1},[25] Q^2_{F2},[26] Q^2_{F3} [27,28] and the concordance correlation coefficient CCC_{EXT},[24,29,30] the last verifies how small the differences are between experimental data and external data set in predictions.[24,30]

In addition, the root mean squared error (*RMSE*), which summarizes the overall error of MLR and ANN models, was used to measure and compare the prediction accuracy in the training ($RMSE_{\mathrm{tr}}$) and in the prediction ($RMSE_{\mathrm{val}}$) set, defined as follows:

$$RMSE\ \mathrm{tr(val)} = \sqrt{\frac{1}{n_{\mathrm{tr(val)}}} \sum_{i=1}^{n_{\mathrm{tr(val)}}} \left(y_i - \hat{y}_i\right)^2} \tag{3}$$

The standard error of the estimate s is a measure of the accuracy of the predictions made with a regression line defined by the following equation:

$$s = \sqrt{\frac{\left(y_{\mathrm{obs}} - y_{\mathrm{calc}}\right)^2}{N - P - 1}} \tag{4}$$

where $N–P–1$ is the degree of freedom

A variance inflation factor (VIF) was calculated to test if multicollinearities existed among the descriptors, which is defined as below:

$$VIF = \frac{1}{1 - R^2} \tag{5}$$

where R^2 is the correlation coefficient of the multiple regression equation between the descriptors of the model. If *VIF* equals one, no intercorrelation exists for each descriptor; if *VIF* maintains within the range 1.0–5.0, the corresponding model is acceptable; if *VIF* is larger than 10.0, the corresponding model is unstable.[31]

Applicability domain (AD) analysis

The AD of a QSPR model[23] must be defined if the model is to be used for screening new compounds. The AD is a theoretical region in space defined by the descriptors of the model and the modeled response, for which a given QSPR should make reliable predictions. This region is defined by the nature of the compounds in the training set and can be characterized in various ways. In this work, the structural AD was verified by the leverage approach. The leverage h_i and warning leverage h^*[32] are defined by the following expressions:

$$h_i = x_i^T \left(X^T X \right)^{-1} x_i \ (i = 1,\ldots,n) \tag{6}$$

$$h^* = \frac{3\left(p+1\right)}{n_{\text{tr}}} \tag{7}$$

where x_i is the descriptor vector of the considered compound, X is the model matrix derived from the training set descriptor values, n_{tr} is the number of training compounds and p is the number of model descriptors. A Williams plot, ta plot of the leverage values *vs*. the standardized residuals, was used to detect the responses of outliers (Y outliers, *i.e*., chemicals with absolute standardized residuals greater than 3 standard deviation units) and the structurally influential chemicals (X outliers, *i.e.*, chemicals with leverage values greater than the threshold value $h_i > h^*$).

RESULTS AND DISCUSSION

In the following sections, the best developed models predicting the solubility for the training set of 48 phenols are reported, which are then validated on the test set of 20 compounds; using linear and nonlinear methods.

MLR model

AG/MLR by using QSARINS software was applied to the training set to select the optimum subset of descriptors and to develop a linear model.

To achieve the best model, four different splitting techniques were applied (Table II)

TABLE II. Statistical parameters of the obtained models based on different splitting techniques

Splitting	Q^2_{LOO} / %	R^2 / %	Q^2_{LMO} / %	Q^2_{F3} / %
Random splitting	88.66	90.82	88.04	88.81
Sorted by response	90.32	91.97	89.63	87.16
Structural similarity	90.10	90.01	89.01	79.08
CADEX	89.33	91.00	89.00	92.36

It is evident from the statistical values in Table II that the models show similar good performances and have excellent predictive capabilities as verified by various criteria in different splittings. Based on the above results, the model constructed by the Kennard and Stone algorithm splitting was chosen.

The equation of the selected model is defined as:

$$\log S = 8.59 - 0.240\times\text{Polarizability} - 2.20DISPe - 0.290nCb- \quad (8)$$

n_{tr} = 48, R^2 = 91.00 %, Q^2_{LOO} = 89.33 %, Q^2_{LMO} = 89.00 %, Q^2_{F1} = 87.24 %, Q^2_{F2} = 85.76 %, Q^2_{F3} = 92.36 %, CCC_{ext} = 92.80 %, $RMSE_{tr}$ = 0.325, $RMSE_{val}$ = = 0.300, s = 0.3404, R^2ys = 0.0668.

The statistical parameters show that the model (Eq. (8)) established a strong correlation between the 3 selected variables and the studied property, characterized by excellent parameters, in addition to a very large value of the Fisher F = 148.2342, which indicates the excellence of the model in the prediction of solubility values, and a good standard error s = 0.3404. Eq. (8) presented a value of R^2 = 91.00 %, indicating excellent agreement between correlation and variation of the data, also the low value of R^2. indicates that the obtained model has no chance correlation. All statistical parameters of the selected model are satisfying and prove, at the same time, that the obtained model is stable, robust and predictive. Then, the selected model was used to predict the test set data.

In Eq. (8), three different kinds of molecular descriptors appear: 1) polarizability defined as the dipole moment of a molecule induced by an electric field of unit intensity;[33] 2) *DISPe*, d COMMA2 value/weighted by atomic Sanderson electronegativities, which is the displacement (DISPe) between the geometric centre and the centre of the considered property field, calculated with respect to molecular principal axes;[13] (3) nCb - the number of substituted benzene C(sp2).[13]

Some important statistical parameters (as given in Table III) were used to evaluate the involved descriptors.

TABLE III. Characteristics of the descriptors in the optimal MLR model

Descriptor	Descriptor type	Coefficient	Standard error coefficient	*t*	*t*-probability	*VIF*
Constant	Electronic	8.58640	0.2647	32.43	0.000	
Polarizability		–0.24015	0.0138	–17.35	0.000	1.079
DISPe	Geometrical	–2.2032	0.3572	–6.17	0.000	1.056
nCb-	Functional group counts	–0.29028	0.0503	–5.77	0.000	1.135

The *t* of a descriptor measures the statistical significance of the regression coefficients. The high absolute values of *t* shown in Table III 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.*, interactions of the descriptors). Descriptors with *t*-probability values below 0.05 (95 % confidence) are usually considered statis-

tically significant in a particular model, which means that their influence on the response variable is not merely by chance.[34]

A smaller *t*-probability suggests a more significant descriptor. The *t*-probability values of the three descriptors are very small, indicating that all of them are highly significant descriptors. The *VIF* values of these descriptors (less than five) suggest that these descriptors are not correlated with each other. Thus, the model could be regarded as an optimal regression equation.

AD of the developed model

Analyzing the applicability domain of a model is another stage of validation. The so-called Williams plot (Fig. 1) presents the relationship between the leverage values (expressing the similarity of a given compound to the training set and standardized residuals). Analyzing the plot, all residuals were located within the range of ±3*s* (horizontal lines), excepting one compound (4-phenylphenol) in the training set was found to be a Y outlier (response outlier). In a Williams plot, Y outliers can be explained as compounds with errors in the experimental values. However, the leverage values (h_i) of all compound in the training and test sets are less than the critical value ($h^* = 0.25$). Therefore, the solubility predicted by the developed MLR model is reliable. The proposed model could be used to screen existing databases or virtual chemical structures to identify the solubility of phenols. In this case, the applicability domain will serve as a valuable tool to filter out "dissimilar" chemical structures.

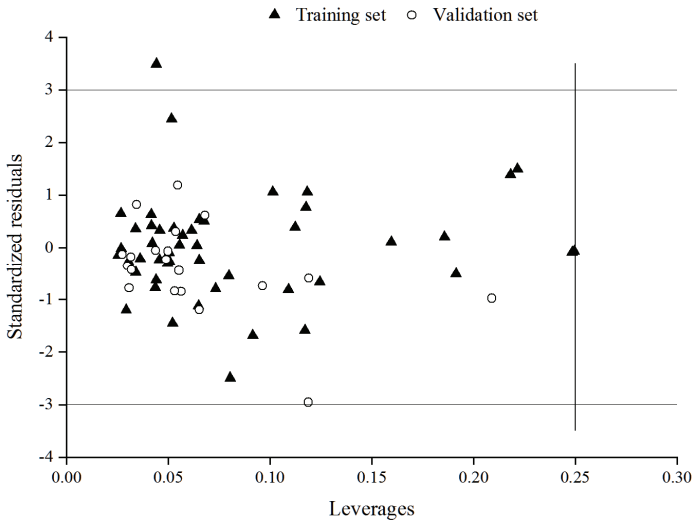

Fig. 1. Williams plot of the developed model.

The predicted values of aqueous solubility are plotted *vs*. their experimental values in Fig. 2. The agreement between these values is good ($R^2 = 91.0$ %, R^2_{val} =

= 88.4 %) and proof of the model quality is the strong correlation between observed and predicted log *S* for both the training and validation sets.

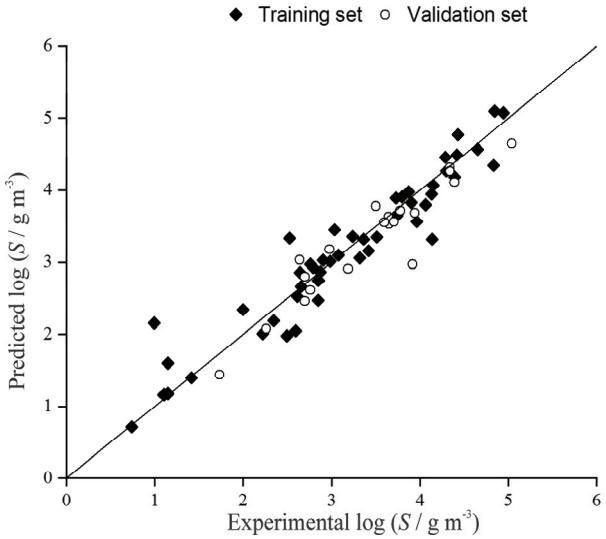

Fig. 2. Plot of predicted *vs*. experimental aqueous solubility.

ANN model

ANN models have become important and widely used nonlinear modeling techniques for QSPR studies. Thus, in order to compare the predictive ability of the MLR model with an ANN model, the dataset was modeled by ANN using the descriptors selected by the MLR model as input variables for a three-layered feed forward ANN model with a back propagation learning algorithm.[35] After optimization of the model several times. The number of neurons in the hidden layers was 4 and the number of iterations was 20.

TABLE IV. Optimal structure adopted for the neuron network

Entries	3 descriptors
Exit	log *S*
Hidden layer	1 hidden layer
Number of neurons in the hidden layer	4 neurons
Algorithm for learning	Retro propagation three-layered feed forward ANN

Multiple calculations were performed in order to obtain the global best results. Finally, the developed ANN model that achieved the goal of the training set and had good performance for the validation set was selected as the prediction model for the solubility of phenols. The test set, which did not contribute to the development of the model, was calculated to test the prediction capability of the ANN model.

Graphs of the *RMSE* values as a function of the number of neurons in the hidden layer for the model are shown in Fig. 3. This figure shows that the smallest *RMSE* values of the learning, validation and test sets are close and the smallest when using four neurons in the hidden layer.

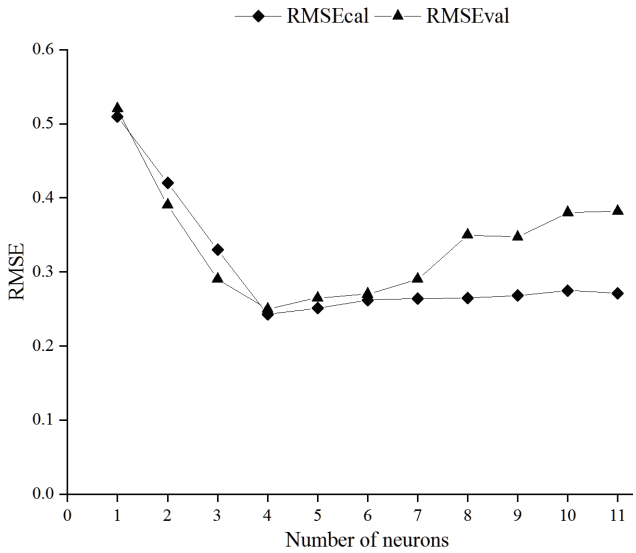

Fig. 3. RMSE values in dependence on the number of neurons in the hidden layer.

The number of neurons in the hidden layer is an important parameter that influences the performance of an ANN model. The usual rule is that the weights and the bias should be smaller than the samples so that the model obtained by the network is stable.

In the situation of this work, the number of hidden neurons should not be greater than 10 with 48 samples in the training set. Better results could be obtained by using four hidden neurons after optimization of the network architecture with respect to the number of hidden neurons. Thus, an architecture (3–4––1) was obtained, with R^2 = 94.992 %, $RMSE_{val}$ = 0.250, $RMSE_{test}$ = 0.224, $RMSE_{tr}$ = 0.243, and s = 0.245 for the training set.

The quality of fit was verified by representation of the predicted values of the solubility as a function of the experimental ones in Fig. 4, which shows a weak dispersion of the points around the first bisectrix, which indicates a good agreement between these values.

Comparison between the MLR and ANN model

A comparison between the statistical parameters obtained by the MLR method and the ANN method was made (Table V and Fig. 5).

Comparison of the MLR and ANN models shows that the *RMSE* for the training set and the entire dataset are lower for the nonlinear ANN model, which confirms the non-linear relationship between structural information and the solubility values of the compounds.

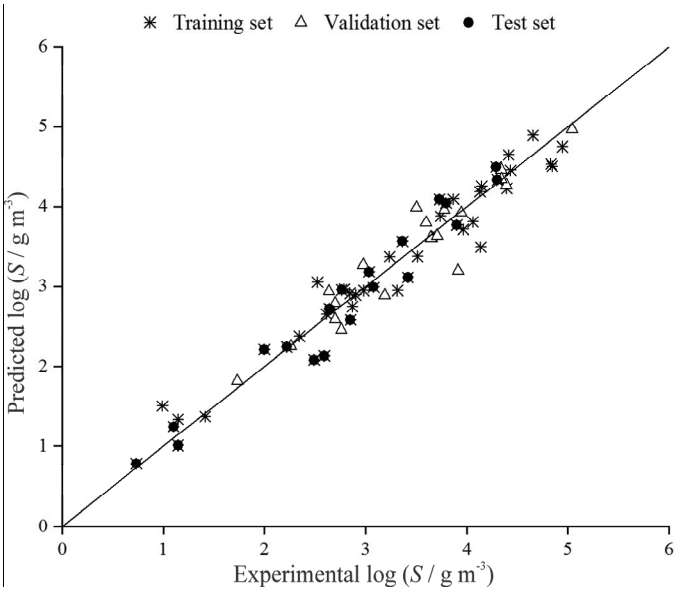

Fig. 4. Predicted values *vs*. experimental values for the training, validation and test sets.

TABLE V. Statistical parameters for the MLR and ANN models

Parameter	Number of descriptors	$RMSE_{tr}$	$RMSE_{val}$	*s*	R^2 /%	Q^2_{ext} / %
MLR	3	0.325	0.3003	0.34	91	92.3
ANN	3	0.243	0.25	0.245	94.99	94.7

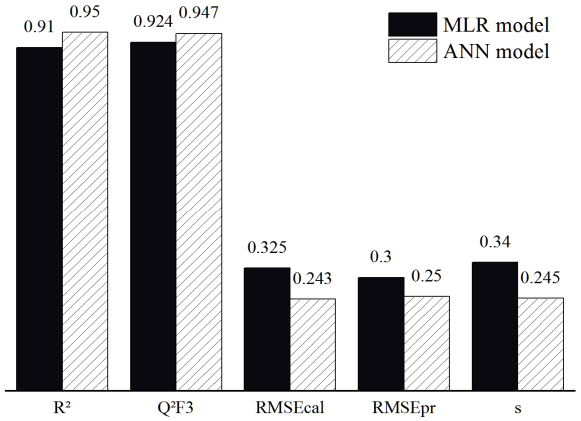

Fig. 5. Values of the statistical parameter for the MLR and RNA models.

Descriptor contribution analysis and interpretation

It is not a trivial task to interpret nonlinear models due to their complex modeling procedure and vague output, although the nonlinear models could give better predictive results. Thus, the descriptor interpretation was realized based on the MLR model. According to a previously described procedure,[36,37] the relative contributions of the three descriptors to the MLR model were determined and are plotted in Fig. 6.

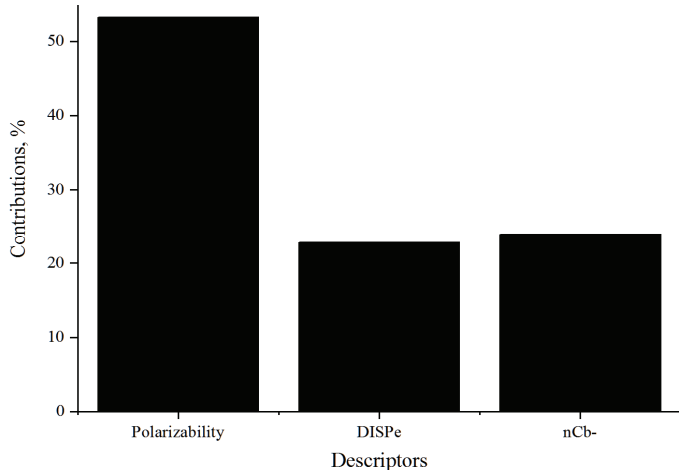

Fig. 6. Relative contributions of the descriptors to the MLR model.

The significance of the descriptors involved in the MLR model decreases in the following order: polarizability (53.23 %)> *nCb-* (23.95 %) > *DISPe* (22.82 %).

CONCLUSIONS

In this study, linear and nonlinear QSPR models were developed to predict the solubility (log *S*) for a dataset of 68 phenols using MLR and ANN. The detailed validation procedure demonstrated the accuracy and robustness of the proposed models not only by calculating their fitness on the training set, but also by testing their predicting ability. The MLR provided a transparent result for modeling the solubility, which indicates that the polarizability is closely related to the property. Nonlinear model was proved to give better results and predict the water solubility of phenols more accurately than the MLR approach. Thus, it could be concluded that this investigation extended the research method to predict solubility. It could also identify and provide some insight into what structural features are responsible for the solubility.

SUPPLEMENTARY MATERIAL

Additional data and considerations are available electronically from http://www.shd.org.rs/JSCS/, or from the corresponding author upon request.

ИЗВОД

ЛИНЕАРНЕ И НЕЛИНЕАРНЕ КВАНТИТАТИВНЕ РЕЛАЦИЈЕ СТРУКТУРЕ И ОСОБИНА ЗА МОДЕЛОВАЊЕ РАСТВОРЉИВОСТИ ДЕРИВАТА ФЕНОЛА

SOUMAYA KHEROUF[1], NABIL BOUARRA[1,2], AMEL BOUAKKADIA[1,3] и DJELLOUL MESSADI[1]

[1]Laboratory of Environmental and food safety, department of chemistry, Badji Mokhtar-Annaba university, PB12, 23000, Annaba. Algeria, [2]Center of scientific and technical research in Physico–Chemical analyzes (CRAPC), Bp 384, Siège ex-Pasna Zone Industrielle, Bou-Ismail, 42004, Tipaza, Algeria и [3]University Abbes Laghrour Khenchela – Algeria – BP 1252 Route de Batna Khenchela 40004, Algeria

Квантитативне релације структуре и растворљивости (QSSR) разматране су као студије квантитативних релација структуре и особина (QSPR), где се растворљивост у води доводи у везу са хемијском структуром. У овом раду се коришћене технике вишеструке линеарне регресије (MLR) и вештачких неуралних мрежа (ANN) за студије квантитативне релације структре и растворљивости (QSSR) растворљивости у води 68 фенола (фенол и његови деривати) на бази молекулских дескриптора израчунатих из оптимизованих 3D структура. Примењујући испуштање вредности, тестове нуле и мултиколинеарности са граничном вредношћу 0,95, и генетички алгоритам (GA) за избор дескриптора добијени су модели са најбољим фитовањем. Након избора дескриптора, коришћена је вишеструка линеарна регресија (MLR) за конструисање линеарног QSSR модела. Вредности R^2 = 91,0 %, Q^2_{LOO} = 89,33 %, s = 0,340 за модел развијен са MLR показују добру способност предвиђања за log S вредности фенола и његових деривата. Резултати MLR модела су упоређени са онима из ANN модела. Поређење показује да су R^2 = 94,99 %, s = 0,245 за ANN више, односно ниже, илуструјући да ANN представља изврснију алтернативу за развијање QSSR модела за log S вредности фенола, него MLR.

(Примљено 20. августа, ревидирано 16. децембра 2018, прихваћено 11. фебруара 2019)

REFERENCES

1. J. Devillers, *SAR QSAR Environ. Res*. **15** (2004) 237. (https://doi.org/10.1080/10629360410001724905)
2. Y. B. Zang, Chin, *Agric. Sci. Bull*. **28 (**2012) 282 (http://caod.oriprobe.com/articles/28599226/Research_Advance_of_Phenol_Adsorption_of_Modified_Bentonite.htm)
3. P. R. Zhan, H. T. Wang, Z. X. Chen, *J. Agro-Environ. Sci*. **27** (2008) 801 (http://en.cnki.com.cn/Article_en/CJFDTotal-NHBH200802077.htm)
4. J. Micha Lowicz , R. Ożadowicz Wirgiliusz Duda, *Water Air Soil Poll*. **16** (2005) 205 (https://doi.org/10.1007/s11270-005-3022-7)
5. S. D. Palmer, N. M. O'Boyle, R. C. Glen, J. B. O. Mitchell, *J. Chem. Inf. Model.* **47** (2007) 150 (https://pubs.acs.org/doi/abs/10.1021/ci060164k)
6. R. Gozalbes, A. Pineda-Lucena, *Bioorg. Med. Chem*. **18** (2010) 7078 (https://doi.org/10.1016/j.bmc.2010.08.003)
7. 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)
8. M. St. J. Warne, D. W. Connel, D. W. Hawker, G. Schüürmann, *Chemosphere* **21** (1990) 877 (https://doi.org/10.1016/0045-6535(90)90168-S)
9. C. Catana, H. Gao, C. Orrenius, P. F. W. Stouten, *J. Chem. Inf. Model*. **45** (2005) 170 (https://doi.org/10.1021/ci049797u)
10. Z. Garkani-Nejad, M. Ahmadvand. *Sep. Sci. Technol*. **46** (2011) 1034 (http://doi.org/10.1080/01496395.2010.539587)

11. D. Mackay, W. Y. Shiu, K. C. Ma, S. C. Lee, *Handbook of physical-chemical properties and environmental fate for organic chemicals*, 2nd ed., CRC Press, Boca Raton, FL, 2006 (https://www.taylorfrancis.com/books/9781420044393)
12. E. Benfenati, J. R. Chrétien, G. Gini, N. Piclin, M. Pintore, A. Roncaglioni, in *Quantitative Structure–Activity Relationships (QSAR) for Pesticide Regulatory Purposes*, Elsevier, Amsterdam, 2007, p. 185–199. (https://doi.org/10.1016/B978-044452710-3/50008-2)
13. HyperChem 6.03 Package. Hypercube, Inc., Gainesville, FL, 1999; software available at: http://www.hyper.com
14. Talete Srl. Dragon for windows (Software for Molecular Descriptor Calculation), version 5.5, Milano, 2007; software available at: http://www.talete.mi.it
15. J. E. Jackson, *A User's Guide to Principal Component*, Wiley, New York, 1991 (https://doi.org/10.1002/0471725331)
16. R. Kennard, L. A. Stone, *Technometrics* **11** (1969) 137 (https://doi.org/10.1080/00401706.1969.10490666)
17. P. Gramatica, N. Chirico, E. Papa, *S.* Cassani, *S.* Kovarich, *QSARINS, Software for the Development and validation of QSAR MLR Models*, available on request at http://www.qsar.it
18. P. Gramatica, N. Chirico, E. Papa, *S.* Kovarich, *S.* Cassani, *J. Comput. Chem.* **34** (2013) 2121 (https://doi.org/10.1002/jcc.23361)
19. J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*; Wiley–VCH, Weinheim, 1999 (https://www.wiley.com/en-us/Neural Networks in Chemistry and Drug Design%3A An Introduction%2C 2nd Edition-p-9783527297795)
20. *S.* Haykin, *Neural Networks. A Comprehensive Foundation*, Perarson Prentice Hall, New Delhi, 2006 (ISBN-13: 978-0023527616)
21. D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Nature* **323** (1986) 33 (https://doi.org/10.1038/323533a0)
22. OECD. *Guidance Document on the Validation of (Quantitative) Structure–Activity Relationships [(Q)SAR] Models*, Organisation for Economic Co-Operation and Development, Paris, 2007 (https://oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm)
23. A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci*. **22** (2003) 70 (https://doi.org/10.1002/qsar.200390007)
24. N. Chirico, P. Gramatica, *J. Chem. Inf. Model*. **51** (2011) 2320 (https://doi.org/10.1021/ci200211n)
25. P. Gramatica, *Mol. Inf*, **33** (2014) 311 (https://doi.org/10.1002/minf.201400030)
26. G. Schüürmann, R. Ebert, J. Chen, B. Wang, R. Kühne, *J. Chem. Inf. Model*. **48** (2008) 2140 (https://doi.org/10.1021/ci800253u)
27. V. Consonni, D. Ballabio, R. Todeschini, *J. Chem. Inf. Model*. **49** (2009) 1669 (https://doi.org/10.1021/ci900115y)
28. V. Consonni, D. Ballabio, R. Todeschini, *J. Chemom.* **24** (2010) 194 (https://doi.org/10.1002/cem.1290)
29. L. I. Lin, *Biometrics*. **45**(1989) 255 (https://doi.org/10.2307/2532051)
30. N. Chirico, P. Gramatica, *J. Chem. Inf. Model*. **51** (2011) 2320 (https://doi.org/10.1021/ci200211n)
31. T. I. Netzeva, A. P. Worth, T. Aldenberg, R. Benigni, M. T. D. Cronin, P. Gramatica, J. *S.* Jaworska, S. Kahn, G. Klopman, C. A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G. Y. Patlewicz, R. Perkins, D. W. Roberts, T. W. Schultz, D. T. Stanton, J. J. M. Van de

Sandt, W. Tong, G. Veith, C. Yang, *ATLA. Altern. Lab. Anim*. **33** (2005) 155 (https://www.ncbi.nlm.nih.gov/pubmed/16180989)
32. N. Chirico, P. Gramatica. *J. Chem. Inf. Model*. **52** (2012) 2044 (https://doi.org/10.1021/ci300084j)
33. G. R. Famini, C. A. Penski, L. Y. Wilson, *J. Phys. Org. Chem*. **5** (1992) 395 (https://doi.org/10.1002/poc.610050704)
34. J. Xu, H. Liu, W. Li, H. Zou, W. Xu, *Macromol. Theory Simul*. **17** (2008) 470 (https://doi.org/10.1002/mats.200800063)
35. D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Nature* **323** (1986) 33 (https://doi.org/10.1038/323533a0)
36. F. Zheng, E. Bayram, *S*. P. Sumithran, J. T. Ayers, C. Zhan, J. D. Schmitt, L. P. Dwoskin, P. A. Crooks, *Bioorg. Med. Chem.* **14** (2006) 3017 (https://doi.org/10.1016/j.bmc.2005.12.036)
37. R. Guha, P. C. Jurs, *J. Chem. Inf. Model*. **45** (2005) 800 (https://doi.org/10.1021/ci050022a).