Algerian Journal of Environmental Science and Technology
September edition. Vol.7. N°3. (2021)
ISSN : 2437-1114
www.aljest.org

Computer-assisted methods useful for the modeling of phenolic dyes wavelengths (λ_{max}) using MLR and ANN methods

N. Bouarra[1,2*], N. Nadji[1], L. Nouri[1], A. Boudjemaa[1], K. Bachari[1], D. Messadi[2]

[1]Centre de Recherche Scientifique et Technique en Analyses Physico-Chimiques, BP 384, Siège ex-Pasna Zone Industrielle, Bou-Ismail CP 42004, Tipaza, Algeria.
[2]Laboratory of environmental and food safety, department of chemistry, Badji Mokhtar-Annaba university, PB12, 23000, Annaba. Algeria.

*Corresponding author: bouarranabil@yahoo.com; Tel.: +213 670 18 52 79; Fax: +21300 00 00

ARTICLE INFO

Article History :

Received : 29/09/2019
Accepted : 21/06/2020

***Key Words**:*

Dyes;
QSPR;
Genetic algorithm;
MLR;
Prediction.

ABSTRACT/RESUME

Abstract: *In this work, a quantitative structure-property relationship (QSPR) was built by using multiple linear regression (MLR) and artificial neural networks (ANN) to predict the wavelengths (λ_{max})of phenolic dyes. After many procedures to reduce the number of descriptors, a hybrid genetic algorithm and multiple linear regression (GA/MLR) method was used to select the descriptors that resulted in the best fitted models. The statistical parameters of the MLR model (R^2 = 89.01 %, Q^2_{LOO} = 85.39 %, s = 24.763) showed a good predictive capacity of λ_{max}. The comparison between statistical parameters obtained by MLR and ANN models indicates the superiority of the ANN over that the MLR model, which illustrates that the ANN method is an excellent alternative for developing QSPR models for λ_{max} than MLR method.*

I. Introduction

Nowadays, dyes dominate the market especially since their properties can be precisely adapted in many sectors such as; textile, plastic, food, paper, printing, pharmaceutical, and cosmetic industries [1-4]. All these dyes are synthesized mainly from petroleum products, especially benzene and its derivatives such as toluene, naphthalene, xylene and anthracene [5].

Since the discovery of the first synthetic dye in 1856 by Perkin, the synthetic dyes gained huge popularity and began to be synthesized on a large scale [6-8]. In fact, it has reached to a level of annually over 7.0 x 10^5 and nearly 1000 different types of dyes are produced worldwide [9-12].

The dye molecules absorb electromagnetic radiation, but differ in the specific wavelengths absorbed. Some dye absorbs light in the visible spectrum (400 – 800 nm) [13]. The dye molecules have delocalized electron systems with conjugated double bonds consists of two groups; the chromophore and the auxochrome groups [14, 15]. Dyes exhibit considerable structural diversity and are classified in several ways [16, 17]. So, they can be classified by both their chemical structure (azo, anthraquinone, sulfur, indigoid, triphenylmethyl (trityl), and phthalocyanine) [16, 18] and their applications (acid, basic, direct, disperse, mordant, reactive, and vat dyes) [19-22]. On the other hand, the spectroscopic property is related to the color property and determined by the dyes structure. It is noted that the maximum absorption wavelengths (λ_{max}) [23]. There have been some reports on the applications of quantitative structure-property relationship (QSPR) methods to investigate the relationship between λ_{max} and the dyes structure [24]. Briefly, QSPR is a promising method applied to quantify the relationship between the molecular structural information and their physicochemical properties [25]. The applications of QSPR are found in all major chemical disciplines including physical organic, physical, medicinal, agricultural, biological, environmental, and polymer chemistry [26-31]. QSPR model consists of a mathematical relationship between the property of interest and variety of molecular features (named descriptors) derived from the structure of the molecule, which ranged from structural and topological indices to electronic and quantum chemical properties [32]. The main steps involved to include the data collection, the molecular descriptor selection and obtaining, the correlation model development, and finally the model evaluation [33]. Many different chemometrics methods, such as multiple linear regressions (MLR) [34], different types of artificial neural networks (ANN) [35, 36] and genetic algorithms (GAs) [37] can be employed to derive correlation models between the molecular structures and their properties. Multiple linear regression (MLR) are the simplest and most commonly used approach in QSPR since it assumes a simple linear relation between the property and each molecular descriptor. However, nonlinear approaches, such as artificial neural networks (ANN), can also be investigated. These approaches can “catch” the hidden nonlinearity between the property and the descriptors, which make them better predictors than MLR models in the most cases [38, 39].

The aim of the present study is to develop a QSPR model between the dye structure and λ_{max} of a group of typical industry organic dyes. Furthermore, QSPR model will be used to predict λ_{max} of 69 phenolic dyes and help to understand the physical mechanisms determining the maximum absorption λ_{max} of phenolic dyes.

II. Materials and methods

II.1. Dataset

A total set of 69 phenolic dyes with a wide structural diversity (see Table S1) are selected as a dataset. The diversity of dataset assures the quality and the robustness of the predictive power of the QSPR model. To select significant descriptors for the QSPR model that captures all the underlying interaction mechanisms, it is advisable to have as many structural characteristics as possible in the dataset. The dataset of this work included phenolic dyes, all values of λ_{max} are taken from the literature [40]. The reported λ_{max} values of different dyes are between 347 and 618 nm (Table 1).

II.2. Structure and descriptors generation

The chemical structures of all compounds are designed using ChemDraw 7.0 program [41] and their three dimensional geometries are pre-optimized with the semi-empirical PM3 method using Hyperchem program [42]. The final geometries are then used as input for the generation of more than 1600 descriptors using Dragon 5.3 software [43]. The generated molecular descriptors include topological descriptors, molecular counts, connectivity indices, information indices, 2D autocorrelations, edge connectivity indices…etc. A preselection of descriptors is performed with the aim to reduce the pool of descriptors by eliminating those that satisfy the following conditions: (a) the descriptor has a constant or near-constant value for all molecules investigated; (b) in the mono parametric correlations with λ_{max} the descriptors has a squared correlation coefficient lower than 0.1 and (c) the descriptors has an inter-correlation coefficient higher than 0.95 with another descriptor [44]. The pre-selection is performed in DRAGON software [43].

II.3. Model development and validation

According to the principles of the Organization for cooperation and Economic Development (OECD), a quantitative model of structure-activity (property) relationship (QSAR/QSPR) should include appropriate measures of quality of fit, robustness and predictability. While the internal performance of a model is determined using a learning set and the predictivity is determined using an appropriate test set [45].

To develop a powerful QSPR model, robust and consistent data is required. Significant descriptors are selected via the genetic algorithm (GA) in the Mobydigs software [46]. GA is a stochastic optimization method that mimics the evolution process by manipulating a collection of data structures [47]. It has been used for the selection of characteristics in QSAR studies [48]. The cross-validation value leave-one-out (LOO) is the optimized parameter in this study. The GA-MLR model for the training set is obtained using the Mobydigs software. Models with varying numbers of descriptors are examined. The developed models have been verified for over-adjustment due to the large number of descriptors and the variable multicollinearity. The possible multicollinearity among the selected descriptors is avoided by applying the rule Q Under the Influence of K (QUIK) [49]. The parameter of QUIK rule has been set to 0.05 to avoid multicollinearity.

The main objective of any QSPR study is to get a model with the highest predictive and generalization abilities. Therefore, two principal parameters (internal validation and external validation) are carried out in order to judge the predictive power of the developed QSAR models. Several commonly used statistic terms are adopted to check the reliability, robustness and stability of the proposed model such as correlation coefficient (R^2) (see eq.1), leave-one-out (LOO) cross-validated $Q^2{}_{LOO}$ (eq.2), and root mean squared error (RMSE) (eq.3):

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

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N-1}} \qquad \textbf{(3)}$$

Also, leave-many-out (LMO) and Y scrambling techniques are also employed. Leave-many-out (LMO) is a more powerful technique than LOO to avoid over estimation and to verify the predictive ability and stability of a model. Here, LMO is repeated for 2000 times with 30 % of the objects left out randomly from the training set at each time. Then a mean value of $Q^2{}_{LMO}$ is reported. Randomization test is applied to exclude the possibility of increasing the correlation by chance and to check for reliability and robustness by permutation testing. New models are recalculated for randomly reordered responses (Y scrambling). The predictive power of QSPR model can be estimated by the external Q^2_{ext} defined as follows:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_{i/i} - y_i)^2 / n_{ext}}{\sum_{i=1}^{n_{tr}}(y_i - \bar{y})^2 / n_{tr}} \quad \textbf{(4)}$$

Tropsha et al. are suggested some criteria which are satisfied by the proposed model **[50]**. These criteria include:

$$Q^2 > 0.5 \quad \textbf{(5)}$$

$$\frac{r^2 - r_0^2}{r^2} < 0.1, \ \frac{r^2 - r_0'^2}{r^2} < 0.1 \quad \textbf{(6)}$$

$$0.85 < k < 1.15 \ or \ 0.85 < k' < 1.15 \quad \textbf{(7)}$$

Here Q^2 is the correlation coefficient between the calculated and the experimental values in the validation set. r_0^2 and $r_0'^2$ are the coefficients of determination. k and k' are slopes of regression lines through the origin of predicted vs. observed, and observed vs. predicted, respectively.

The applicability domain (AD) of QSPR model must be defined if the model is to be used for screening new compounds [51]. The AD is the 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. 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 is verified by the leverage approach. The Williams plot, the plot of leverage values *vs.* standardized residuals, is used to give a graphical detection of both the response outliers (Y outliers) and the structurally influential compounds (X outliers) [51]. In this plot, the two horizontal lines indicate the limit of normal values for Y outliers (i.e. compound with standardized residuals greater than 3 standard deviation units, ± 3s); the vertical straight lines indicate the limits of normal values for X outliers (i.e. compound with leverage values greater than the threshold value, *h**). In general, *h** is set to *3(p+1) /n*, where *p* is the number of descriptors in the developed model and *n* is the number of compounds in the training set [52]. A composite predicted value greater than three normalized residuals is considered an outlier.

II.4. Artificial neural networks (ANN)

Neural networks have been studied since the 1940s [53]. The basic ideas of this technique come from cognitive research, from which comes the name "neural networks". The technique inspired many researchers, but much of the interest disappears after an article by Minsky and Papert [54]. Finally, it is published in the early 80s after a virtual forgetfulness of twenty years. The cause of the sudden interest is the appearance of new architectures of neural networks.

Artificial neural network (ANN) is an information-processing pattern that is inspired by the way biological nervous systems, such as the brain, process information [55]. The majority of the networks contain at least three layers: input, hidden and output. Based on the function, there are different types of neural networks like feedforward backpropagation, counter propagation, probabilistic neural network, self-organizing map, etc. In the present study, for the development of our nonlinear model, feed-forward backpropagation method was used. Multilayer feed-forward network is a type of ANN widely used, which is trained by the back-propagation (BP) learning algorithm [56]. ANN consists of several “neurons” that receive data from the outside, process the data, and output a signal. A “neuron” is essentially a regression equation with a nonlinear output. When more than one of these neurons is used, nonlinear models can be fitted. BP-ANN receives a set of inputs, which are multiplied by each neuron’s weight. These products are summed for each neuron, and a nonlinear transfer function is applied. In this study, the log sigmoid function is defined as follow:

$$f(x) = \frac{1}{1+e^{-x}} \quad \textbf{(8)}$$

This equation is used as a transfer function. The transformed sums are then multiplied by the output weights where they are summed a final time, transformed, and interpreted. Each input and output value is scaled between 0 and 1 by using the following equation:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \textbf{(9)}$$

Here X′ is the normalized value; x is any one of the descriptor vectors. X_{max} and X_{min} are the maximum and minimum values of the descriptor vector in the dataset.

Since a back-propagation network is a supervised method, the desired output must be known for each input vector so an error can be calculated. This error is propagated backward through the network, adjusting the weights so that the next time the network sees the same input patterns, it will come closer to the desired output. The patterns are repeated many times until the network learns the relationship.

III. Results and discussion

III.1. Dataset for analysis

The experimental data of λ_{max} values are divided randomly into training and validation sets. As a first step, GA-MLR is applied to the training set to select the best subset of descriptors and build a linear model. The six selected descriptors for the selected model are regrouped in the Table 1.

III.2. Multiple linear regressions (MLR) model

In order to predict λ_{max}, a mathematical linear model is proposed using multiple linear regression. The whole data set was split into a training set with 49 compounds and into a prediction set with 20 compounds. The optimal model obtained contains six molecular descriptors: *EHOMO, X2A, X5A, RDF080e, RDF135e and R8u+*, is defined by the following equation:

$$\boldsymbol{\lambda_{max}} = 1913 + 42.1\,\boldsymbol{E_{HOMO}} + 2063\,\boldsymbol{X2A} + 6423\,\boldsymbol{X5A} + 1.96\,\boldsymbol{RDF080e} + 2.60\,\boldsymbol{RD135e} + 1674\,\boldsymbol{R8u^{+}} \qquad \textbf{(10)}$$

$n_{tr} = 49$, $R^2 = 89.01\%$, $Q^2_{LOO} = 85.39\%$, $Q^2_{LMO} = 84.28\%$, $Q^2_{ext} = 85.73\%$, $RMSE_{tr} = 22.926$,

$RMSE_{val} = 26.121$, $s = 24.762$, $F = 56.676$, $Q^2_{Yscrambling} = -0.1972$, $R^2_{Yscrambling} = 0.1262$.

The statistical parameters of the model prove that the established model is stable, robust and predictive. Thus, the model was approved, R^2 and Q^2_{LOO} value is greater than 0.7. Additionally, this model has a smaller $RMSE_{val}$ values and the greatest Q^2_{ext} values, which indicate that this model presented the least error and the smallest differences between the experimental and predicted data. Also, Q^2_{LMO} values for the model is greater than 0.6 and close to R^2. Additionally, the value of the Fisher statistic (F = 56.676), which indicates that the model is good in the prediction of the values of λ_{max}. The low value of $Q^2_{Yscrambling}$ and $R^2_{Yscrambling}$ indicates that the obtained model has no chance correlation. The proposed MLR model satisfies the Golbraikh and Tropsha requirements of the test set;

$Q^2 = 0.8543 > 0.5$ $r_0^2 = 0.8539$ $r_0'^2 = 0.8368$

$r^2 - r_0^2 / r^2 = 0.0005 < 0.1$ $r^2 - r_0'^2 / r^2 = 0.0205 < 0.1$

$0.85 \leq k = 1.0001 \leq 1.15$ $0.85 \leq k' = 0.9971 \leq 1.15$

The results of prediction by the developed model are regrouped in Table 1.

Table 1. *Names, experimental and predicted* λ_{max} *and the calculated descriptors of the model*

N°	Name	λmax exp	λmax Pred	EHOMO	X2A	X5A	RDF080e	RDF135e	R8u+
1	Acid blue 45	595	582.52	-8.946	0.282	0.06	13.276	0	0.023
2	Acid red 97	498	523.06	-8.67	0.285	0.071	29.45	15.738	0.021
3	Acid red 114	514	537.86	-8.709	0.291	0.07	30.303	26.941	0.018
4	Acid red 151	512	480.47	-8.716	0.293	0.078	8.472	6.604	0.023
5	Acid red 183	494	477.91	-9.152	0.297	0.073	11.438	7.023	0.021
6	Acid violet 7	520	540.94	-8.831	0.282	0.072	6.783	3.029	0.029
7	Allura Red AC	504	494.72	-9.398	0.293	0.069	12.689	3.419	0.024
8	Biebrich Scarlet	505	481.88	-8.969	0.296	0.075	14.767	9.032	0.026
9	Brilliant Black BN	570	544.41	-8.921	0.296	0.065	35.365	25.108	0.023
10	Brilliant Crocein	510	513.14	-9.072	0.297	0.071	15.769	8.598	0.035
11	Brilliant yellow	397	443.98	-8.94	0.303	0.079	28.039	24.614	0.018
12	Chicago sky blue 6b	618	612.42	-8.974	0.285	0.063	23.609	36.674	0.012
13	Chromotrope 2B	514	511.91	-9.309	0.298	0.068	13.142	7.218	0.029
14	Crystal scarlet G	510	525.26	-8.607	0.289	0.068	15.248	2.518	0.018
15	Direct blue 71	594	574.28	-8.535	0.285	0.067	45.267	37.918	0.017
16	Direct red 23	507	544.83	-8.778	0.292	0.07	25.615	27.995	0.018
17	Direct red 75	522	539.21	-8.557	0.294	0.071	29.355	28.494	0.019
18	Direct red 80	528	506	-9.053	0.296	0.07	58.237	50.493	0.01
19	Direct violet 51	549	528.62	-8.529	0.29	0.072	30.018	25.048	0.017
20	Disperse yellow 7	385	395.67	-8.906	0.299	0.087	6.846	6.69	0.017
21	Eriochrome black T	503	554.25	-8.931	0.284	0.068	7.57	4.644	0.025
22	Eriochrome blue black B	528	527.4	-8.713	0.282	0.071	11.693	1.01	0.023

23	Evans blue	611	624.91	-8.882	0.287	0.063	27.062	42.904	0.014
24	Mordant brown 24	373	462.96	-9.019	0.305	0.074	7.757	0	0.029
25	Mordant brown 33	442	369.07	-9.718	0.306	0.082	4.025	0.017	0.018
26	Mordant orange 1	385	392.23	-8.621	0.306	0.086	4.332	0	0.02
27	Mordant yellow 12	380	566.39	-8.874	0.293	0.068	17.173	20.229	0.029
28	Naphthol blue black	618	521.86	-8.576	0.284	0.078	16.093	8.573	0.039
29	Oil red EGN	521	501.54	-8.557	0.285	0.078	23.033	12.925	0.029
30	Oil red O	518	323.21	-8.974	0.31	0.094	3.256	0	0.022
31	4-phenylazophenol	347	502.34	-8.485	0.286	0.077	9.293	0	0.029
32	Sudan II	493	479.21	-8.654	0.287	0.081	6.24	8.657	0.019
33	Sudan III	507	471.01	-9.218	0.3	0.072	8.831	2.729	0.022
34	Sunset Yellow FCF	482	516.63	-8.754	0.288	0.078	6.22	0	0.047
35	Toluidine Red	507	554.45	-8.738	0.289	0.066	37.258	30.797	0.013
36	Trypan Blue	520	491.89	-8.895	0.297	0.067	16.156	0.836	0.015
37	Xylidine Ponceau 2R	503	493.26	-8.712	0.297	0.069	45.766	28.309	0.011
38	Cibacron Brilliant Red 3B-A	517	572.87	-8.079	0.284	0.07	1.95	0	0.023
39	Methylene violet	580	480.82	-8.934	0.293	0.076	4.907	0.047	0.027
40	Acid orange 8	490	352.99	-9.575	0.306	0.084	4.455	0	0.013
41	Alizarin yellow GG	362	499.33	-8.889	0.289	0.068	23.45	1.369	0.021
42	Bordeaux R	518	461.34	-8.598	0.287	0.079	12.12	4.236	0.013
43	Disperse orange 13	427	477.29	-8.718	0.298	0.075	7.583	0	0.025
44	Mordant brown 48	492	377.95	-9.588	0.31	0.081	7.043	2.382	0.021
45	Mordant yellow 10	354	504.21	-8.527	0.284	0.078	9.68	0	0.033
46	Orange OT	505	549.64	-8.811	0.282	0.071	5.98	1.041	0.032
47	Platine chrome black 6BN	569	472.47	-8.559	0.286	0.08	6.286	0	0.021
48	Sudan I	476	376.8	-8.842	0.306	0.088	4.323	0	0.024
49	Sudan Orange G	388	543.36	-8.878	0.294	0.073	13.049	21.466	0.029
50	Direct red 81*	508	420.99	-9.259	0.302	0.08	7.684	9.151	0.015
51	Mordant orange 10*	386	531.24	-9.137	0.293	0.065	15.572	8.764	0.019
52	New coccine*	506	511.88	-9.272	0.299	0.068	8.322	1.177	0.033
53	Orange G*	475	491.41	-9.018	0.308	0.07	13.787	16.368	0.016
54	Reactive Orange 16*	494	413.84	-9.216	0.311	0.08	7.235	2.197	0.031
55	Tropaeolin O*	490	430.33	-9.228	0.292	0.076	18.493	0	0.019
56	Mordant red 19*	413	442.01	-9.198	0.309	0.075	15.846	4.019	0.033
57	Acid yellow 99*	445	519.1	-8.848	0.282	0.072	8.928	1.026	0.022
58	Acid red 88*	505	493.27	-9.291	0.292	0.066	22.428	2.855	0.02
59	Amaranth*	521	522	-9.236	0.287	0.068	12.058	6.165	0.02
60	Chromotrope FB*	515	521.67	-9.144	0.297	0.068	12.05	2.816	0.035
61	Chromotrope 2r*	510	485.59	-8.982	0.294	0.074	8.581	0	0.029

62	Crocein orange G*	482	364.95	-8.876	0.312	0.088	5.163	0.1	0.026
63	Disperse yellow 3*	357	364.950 3	-8.876	0.312	0.088	5.163	0.1	0.026
64	Mordant brown 4*	500	469.19	-8.552	0.298	0.075	9.298	0	0.018
65	Orange II*	483	472.99	-8.798	0.294	0.076	7.463	0.135	0.023
66	Plasmocorinth B*	527	514.7	-9.089	0.296	0.068	15.964	2.696	0.033
67	Ponceau SS*	514	494.09	-9.255	0.296	0.071	18.633	9.464	0.029
68	Sudan IV*	520	507.75	-8.558	0.285	0.078	12.718	9.47	0.026
69	Sudan Red B*	521	511.89	-8.586	0.287	0.078	9.101	7.794	0.03

*Compounds with the validation set.

III.3. Variable analysis

The multi-collinearity between the above six descriptors for the model is detected by calculating their variation inflation factors VIF as shown in Table 2. Consequently, it has been found that the descriptors employed in the proposed models have low-inter-correlation. The VIF is defined as follow [57]:

$$VIF = \frac{1}{1-R^2} \tag{11}$$

Where R^2 is the squared correlation coefficient between the coefficient regressed against all the other descriptors in the developed model. If VIF value is bigger than 5.0 indicates a more serious multi-colinearity problem. Where, a value less than 5 indicates that they are all highly significant descriptors. As shown in Table 2, VIF value for each descriptor is less than 5, which indicates absence of any inter-correlation among the descriptors and the model had evident statistical significance.

***Table 2.** The coefficients Analysis of MLR Model*

Predictor	Coef	SE Coef	T	P	VIF
Constant	1913.5	144.4	13.25	0	-
EHOMO	42.06	16.19	2.60	0.013	1.988
X2A	-2062.9	739.3	-2.79	0.008	2.855
X5A	-6422.6	771.8	-8.32	0	2.319
RDF080e	-1.963	0.6172	-3.18	0.003	4.834
RDF135e	2.5966	0.5652	4.59	0	4.641
R8u+	1673.8	576.8	2.90	0.006	1.431

P-value is smaller than 0.05 means the obtained equation is statistically important at 95 % level. Predicted versus experimental values for λ_{max} values of training and validation set obtained by MLR modeling is shown in Figure 1. An agreement between the experimental and predicted λ_{max} for each set is observed. Furthermore, the data show a low scattering around the first bisector.

The applicability domain (AD) of the model is analyzed by Williams plots (shown in Figure 2). As can be seen, there is an outlier compound in the training set and all validation compounds located between two vertical lines.

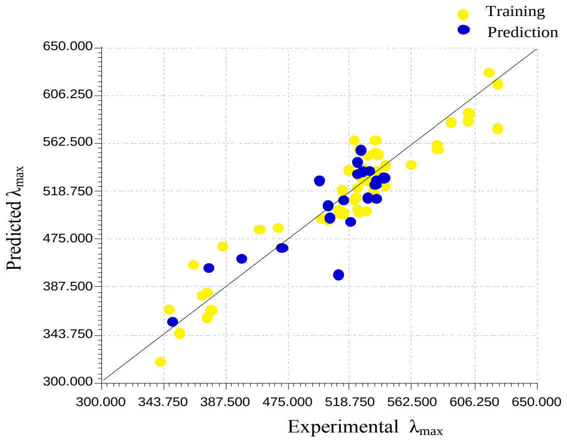

***Figure 1.** Predicted λ_{max} versus experimental plot*

The standard residual value of (***Trophaeolin O***) is greater than 3s. This compound can be considered as response outlier (Y outlier), which could be associated with errors in the experimental values. However, the majority of the compounds within the model applicability domain are calculated accurately, due to the further the reliability of the prediction.

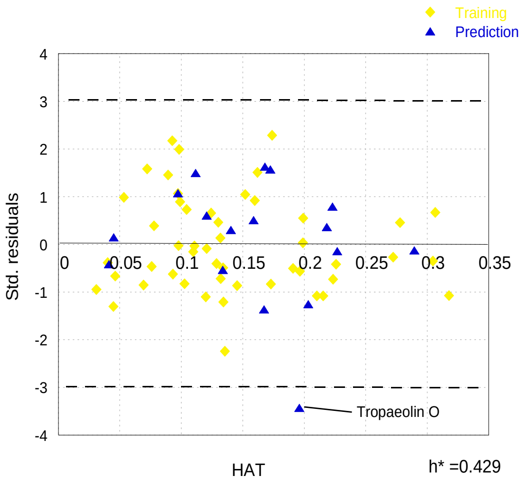

Figure 2. *Williams plots of the developed MLR model*

The reliability and robustness of the proposed model have also verified the response permutation test, also known as Y-scrambling. The Y-scrambling test is a helpful tool used to verify the possibility that the obtained model could suffer from a chance correlation [52]. This procedure involves fitting several models. The correlation is obtained for the permuted models having R² and Q² significantly lower than the original model.

If the original QSPR model is statistically significant, its results should to be significantly better than those from permuted data. R² and Q² values of the original model were a lot of higher than any of the trials using permuted data. It is showed in Figure 3 that the results obtained for all randomised models are of bad quality compared to original model. So, the proposed model is statistically significant and strong.

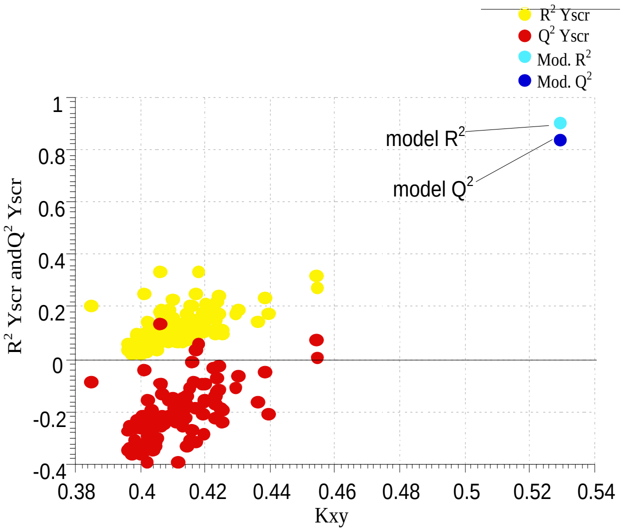

Figure 3. *Y-Scramble plot of R² and Q² vs. Kxy for random models (Kxy: correlations among the block of the descriptors and the experimental data).*

III.4. Artificial neural networks (ANN) results

Another way to search out a relationship between the λ_{max} and descriptors, nonlinear modeling using descriptors as input and ANN as a regression tool was employed. During this nonlinear modeling, a network including a totally connected three-layer, feed-forward ANN model trained with a back-propagation learning algorithm was used. The descriptors selected by the MLR model were used as input variables for the BP ANN model, and λ_{max} is the output variable.

Recall that the two sets (training and validation) and the descriptors are those used for RLM model. The descriptors are used for the configuration of the neural network, which is perfected during the learning phase; the operating parameters are determined so as to obtain a good match between the simulated values and the training data, combined with a correct generalization of these simulations.

Before training the network, the number of neurons within the hidden layer should be optimized. For this purpose, a lot of training of network is performed with totally different numbers of hidden neurons from one to eight. The root means square error for training and validation sets are obtained for various numbers of neurons at the hidden layer, and therefore the minimum value of RMSE is verified as the optimum value. The plot of RMSE for the training set and the test set versus the number of neurons within the hidden layer has been shown in Figure 4. It is clear that six nodes within the hidden layer are the optimum value.

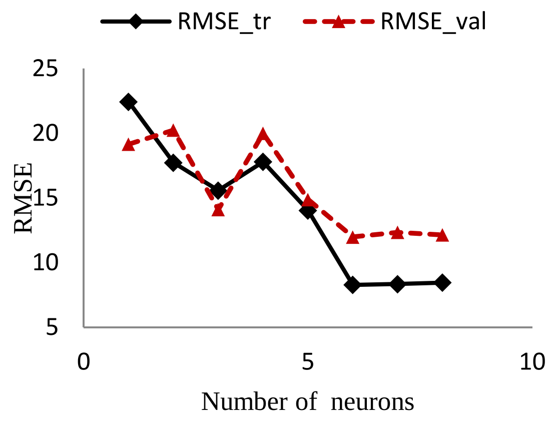

Figure 4. *Plot of RMSE for training (RMSET) and validation (RMSEV) sets versus the number of nodes in hidden layer*

Figure 5 shows a plot of RMSE for the training set and validation set versus the number of interations that represents the estimation of the extent of the training amount. It can be seen from this figure that while training the network is performed for the training set, with increasing iteration the RMSEtr is reduced, at first quickly and later slowly. However, the RMSEval for the validation (prediction) set initially decreases and then starts to extend after approximately 2000 iterations. This position is the commencing point of overtraining of network and then 2000 are chosen as the number of iteration.

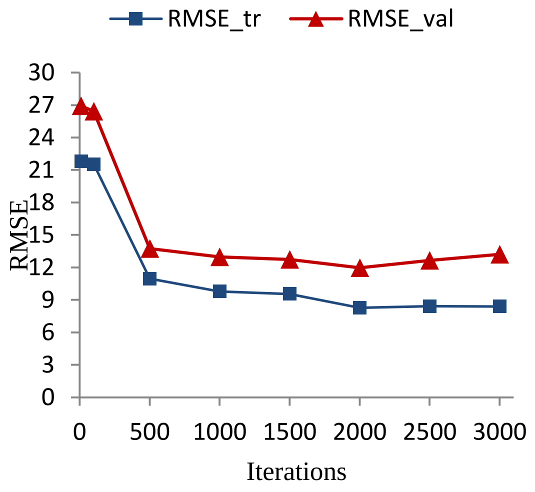

Figure 5*. Plot of RMSE for training (RMSE$_{tr}$) and validation (RMSE$_{val}$) sets versus the number of iterations.*

Learning the neural network represents a fragile balance between all these parameters, hence the difficult is to attain it. Table 3 presents the optimal structure of the neural network used in this work.

Table 3. *Optimal structure of the neural network.*

Number of entries	**06 descriptors**
Number of exits	01 (λ_{max})
Number of hidden layers	One hidden layer
Number of neurons in the hidden layer	06 neurons
Number of iterations	2000 iterations
Learning Algorithm	Retro propagation of the error gradient
Learning function	Hyperbolic tangent

The statistical parameters obtained are regrouped in the Table 4. The value of the coefficient R² (= 98.567 %) indicating excellent agreement between correlation and variation of the data. The RMSE of the training, the test and the validation sets values are 8.275, 11. 979 and 21.561, respectively. According to the Figure 6, λ_{max} values calculated using ANN model are similar to the experiment data. The training set is used to construct ANN model. So, the high values of R² and the small RMSE suggested that ANN model is able to fit λ_{max} of phenolic dyes. The validation set is used to confirm the parameters of ANN model. These results suggest that ANN model could be used to predict λ_{max} of phenolic dyes.

Table 4. *Statistical Parameter for ANN model.*

Training Set			**Test Set**	**Validation Set**	
R² (%)	**s**	**RMSE$_{tr}$**	**RMSE$_{test}$**	**Q²$_{ext}$ (%)**	**RMSE$_{val}$**
98.567	8.361	8.275	11.979	90.277	21.561

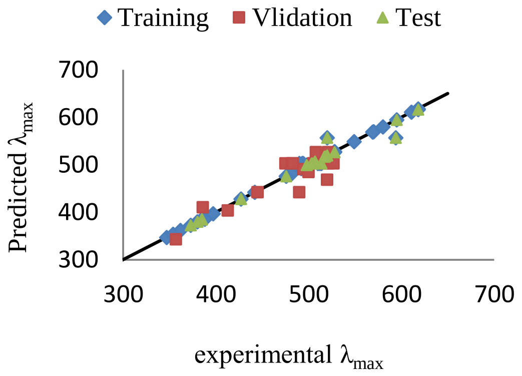

Figure 6. *The plots of λ_{max} predicted by ANN model versus the experimental values.*

III.5.Comparisons between MLR and ANN methods

In order to compare the performance of MLR and ANN in predicting of λ_{max} of phenolic dyes, the descriptors that used in MLR model should be the same as the input variables for the generation of the network (Figure 7). It can be seen from this comparison the dissimilarity between MLR and ANN statistical results. So, ANN has more efficiency as well as MLR to map the relationship between input objects and response values.

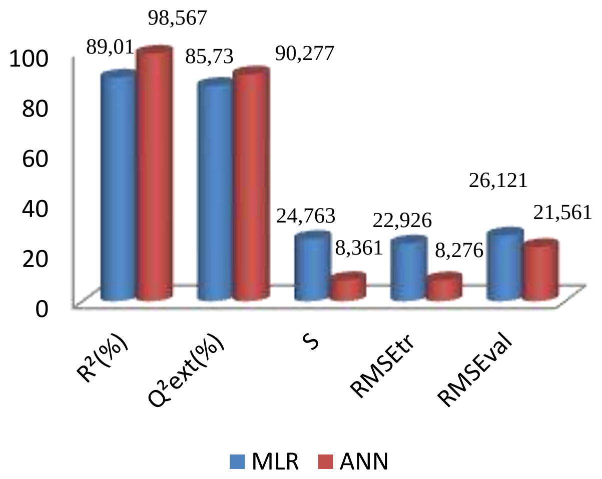

Figure 7. *Comparison between MLR and ANN performances.*

III.6. Description of model descriptors

By analyzing the selected descriptors contained physical and chemical properties, it is possible to obtain some main structural factors relating to λ_{max}. Hence, six descriptors indicated different aspects information of the molecular structure. The descriptor ***E_{HOMO}*** (highest occupied molecular orbital energy) is an important parameter for modeling λ_{max} [58]. ***RDF080e*** and ***RDF150e*** (Radial Distribution Function-8.0 and 15.0, respectively), weighted by atomic Sanderson electronegativities. ***RDF080e*** and ***RDF150e*** descriptors are among RDF descriptors, based on the distance distribution in the geometrical representation of a molecular and constitute a radial distribution function code (RDF code) that shows certain characteristics in common with the 3D-Morse code [59]. The presence of RDF080e and RDF150e in the model equation illustrates the influence of atomic electronegativities on λ_{max}. The general form of the radial distribution function is represented by:

$$RDFR_w = f.\sum_{i=1}^{nAT-1}\sum_{j=i+1}^{nAT} w_i w_j e^{-\beta(R-r_{ij})^2} \qquad \textbf{(12)}$$

Where *f* is a scaling factor (assumed equal to one in the calculations), w is a characteristic properties of the atoms i and j, r_{ij} is the interatomic distance and nAT is the number of atoms in the molecule. The exponential term contains the interatomic distance r_{ij} and the smoothing parameter β, which defines the probability distribution of the individual interatomic distance; β can be interpreted as a temperature factor that defines the movement of atoms. *w* can be an atomic mass (m), the van der Waals volume (v), the Sanderson atomic electronegativity (e) and, the atomic polarizability (p).

The ***X2A*** and ***X5A*** (average connectivity index chi-2 and chi-5, respectively) [60] are calculated based on the graph representation of the molecule (hydrogen-depleted molecular graph) [60]. A path of length represents the connection of two atoms with the bond. It relates to the valence electrons and the number of atoms in the molecule. The general formula for connectivity indices is:

$$\chi_q^m = \sum_{k=1}^{k}\left(\prod_{a=1}^{n}\delta_a\right)^{-1/2} \qquad \textbf{(13)}$$

Where *k* runs over all of the *m*th order subgraphs constituted by *n* atoms (*n* = *m* + 1 for acyclic subgraphs); *K* is the total number of *m*th order subgraphs present in the molecular graph. The product is over the vertex degrees δ of all the vertices involved in each subgraph. The subscript "*q*" refers to the type of molecular subgraph and is *ch* for chain or ring, *pc* for path-cluster, *c* for cluster, and *p* for path (that can also be omitted).

The average valence connectivity indices X*k*Av are obtained by dividing each valence connectivity index by the number of paths involved in its calculation.

The descriptor ***R8u⁺*** (GETAWAY descriptor) which may be defined as > R maximal autocorrelation of lag 8 / unweighted. *Rkw* is defined as the following equation:

$$Rkw = \sum_{i=1}^{nAT-1}\sum_{j>1}\frac{\sqrt{(h_{ii}-h_{jj})}}{r_{ij}}w_i w_j \delta(k,d_{ij})$$

$$k = 1{,}2\ldots{,}8 \qquad \textbf{(14)}$$

nAT is the number of molecule atoms; d_{ij} is the topological distance between atoms i and j; w_i is a physico-chemical atomic weight; d is the topological diameter; δ(k; d_{ij}) is a Dirac-delta function (δ=1 if d_{ij} = k, zero otherwise); δ(k; d_{ij}; h_{ij}) is another Dirac-delta function (δ = 1 if d_{ij} = k and h_{ij}>0, zero otherwise). The atomic properties *w* used for GETAWAY descriptor calculation are atomic mass (m), atomic polarizability (p), atomic electronegativity (e), van der Waals atomic volume (v), and the unit weight (u). In the Eq.10, ***R8u⁺*** shows the positive contribution. ***R8u⁺*** is one of GETAWAY descriptors, which are defined as influence/distance matrix R. However, R-GETAWAY belongs to GEAWAY descriptors that have been proposed with the aim of matching 3D molecular geometry, atom relatedness, and chemical information [59]. In summary, all selected molecular descriptors have a clear chemical and physical meaning. Hence, it could be concluded that λ_{max} of phenolic dyes is closely related to their connectivity index, the atomic Sanderson electronegativities and 3D information of molecular structure.

IV. Conclusion

MLR was performed to study the relationship between λ_{max} and theoretical descriptors. The model validation was achieved by using rigorous internal and external validation methods. To prove the absence of the chance correlation between independent and dependents variables MLR modeling was performed on the randomized data. The low values of $Q^2_{Yscrambling}$ and $R^2_{Yscrambling}$ confirm that the obtained model not due by chance. The ANN modeling was used to handle the probable nonlinear relationship between descriptors and λ_{max}. The six descriptors that are appeared in the MLR model were used as input parameters of the network. Comparison of the MLR and 6-6-1 ANN models showed that the ANN method seems to be the best way to select representative calibration and test data sets in a validation context. Therefore, the ANN method could be used to derive statistical models with better qualities and better generalization ability than the linear regression method. The proposed model in this study provided a simple and straightforward way to predict the λ_{max} just from the molecular structures and give some insight into the structural features related to λ_{max} of the phenolic dyes.

Acknowledgements

The authors are grateful to the Directorate General for Scientific Research and Technological Development DGRSDT of Algeria for the financial support to this work.

V. References

1. Oze, A.; Akkaya, G.; Turabik M. The biosorption of acid red 337 and acid blue 324 on enteromorpha prolifera: The application of non linear regression analysis to dye biosorption. *Journal Chemical Engineering* 112 (2005) 181-190.
2. Jaikumar, V.; Kumar, K. S.; Prakash D. G. Biosorption of acid dyes using spent brewery grains : Characterization and modeling. *Journal of Applied Science and Engineering* 7 (2009) 115-125.
3. Gupta, V. K.; Kumar, R.; Nayak, A.; Saleh, T.A.; Barakat, M.A. Adsorptiv removal of dyes from aqueous solution onto carbon nanotubes: a review. *Journal Colloid and Interface Science* 193-194 (2013) 24-33.
4. Sapra, N. Treatment reuse of textile wastewater by overland flow. *Journal Desalination* 106 (1996) 179-182.
5. Sharma,B.k. Fuel and petroleum processing, 1st ed.; Krishna prakashan media(P) LTD, Meerut, India, 1998; p131.
6. Welham, A. Theory of dyeing (and the secret of life). *Journal of the Society of Dyers and Colourists* 116 (2000) 140-143.
7. Atabati, M.; Khandani, F. Ant colony optimization as a descriptor selection in QSPR modeling for prediction of λmax of azo dyes. *Journal Chinese Chemical Letters* 23 (2012) 1209-1212.
8. Ezeokonokwo, M.A.; Okoro, U.C. 3D-QSPR method of computational technique applied on red reactive dyesby using CoMFAstrategy, *International Journal of Molecular Sciences* 12 (2011) 8862-8877.
9. Robinson, T.; McMullan, G.; Marchant, R.; Nigam, P. Remediation of dyes in textile effluent: a critical review on current treatment technologies with a proposed alternative. *Journal Bioresource Technology* 77 (2001) 247-255.
10. Tunç, O.; Tanaci, H.; Aksu, Z. Potential use of cotton plant wastes for the removal of Remazol Black B reactive dye. *Journal of Hazardous Materials* 163 (2009) 187-198.
11. Auta, M.; Hameed, B.H. Preparation of waste tea activated carbon using potassium acetate as an activating agent for adsorption of acid blue 25 dye. *Chemical Engineering Journal* 171, (2011) 502-509.
12. Sen, T.K.; Afroze, S.; Ang H. Equilibrium, kinetics and mechanism of removal of methylene blue from aqueous solution by adsorption onto pine cone biomass of Pinus radiata. *Journal Water, Air & Soil Pollution* 218 (2011) 499-515.
13. Rangabhashiyam, S.; Anu, N.; Selvaraju, N. Sequestration of dye from textile industry wastewater using agricultural waste products as adsorbents. *Journal of Environmental Chemical Engineering* 1 (2013) 629-641.
14. Gupta, V. K.; Mittal, A.; Malviya, A.; Mittal, J. Adsorption of carmoisine A from wastewater using waste materials-Bottom ash and deoiled soya. *Journal of Colloid and Interface Science* 335 (2009) 24-33.
15. Suyambo, K.B.; Perumal, R.S. Equilibrium, Thermodynamic and kinetic studies on Adsorption of a basic dye by Citrullus Lanatus Rind. *Journal Energy and Environment* 3 (2012) 23-34.
16. Forgacs, E.; Cserhati, T.; Oros Gyula. Removal of synthetic dyes from wastewaters: a review. *Journal Environment International* 30 (2004) 953-971.
17. Gupta, V.K.; Suhas. Application of low-cost adsorbents for dye removal: a review. *Journal of environmental management* 90 (2009) 2313-2342.
18. Ghaly AE.; Ananthashankar R.; Alhattab M.; Ramakrishnan. Production, characterization and treatment of textile effluents: a critical review. *Journal Chemical Engineering and Process Technology* 5 (2014) 2-18.
19. Demirbas, A. Agricultural based activated carbons for the removal of dyes from aqueous solution:A review. *Journal of Hazardous Materials* 167 (2009) 1-9.
20. Hameed, B.H.; El-Khaiary, M.I. Removal of basic dye from aqueous medium using a novel agricultural waste material: Pumpkin seed hull. *Journal of Hazardous Materials* 155 (2008) 601-609.
21. Hao, J.; Kim, H.; Chiang, P. C. Decolorization of wastewater. *Critical Reviews in Environmental Science and Technology* 30 (2000) 449-505.
22. Yagub, M. T.; Sen, T. K.; Afroze, S.; Ang, H.M. Dye and its removal from aqueous solution by adsorption: A review. *Journal of Advances in Colloid Interface science*. 209 (2014) 172-184.
23. Xu, J.; Zheng, Z.; Chen, B.; Zhang, Q. A Linear QSPR model for prediction of maximum absorption wavelength of second-order NLO chromophores. *QSAR & Combinator Science* 25 (2006) 372-379.
24. Luan, F.; Xu, X.; Liu, H.; Cordeiro, M.N.D. Review of quantitative structure-activity/proprety relationship studies of dyes: recent advances and perspectives. *Society of Dyes and Colourists, Coloration Technology* 129 (2013) 173-186.
25. Xu, Y.; Chen, X.Y.; Li, Y.; Ge, F.; Zhu, R.L. Quantitative structure-property relationship (QSPR) study for the degradation of dye wastewater by Mo-Zn-Al-O catalyst. *Journal of Molecular liquids* 215 (2016) 461-466.
26. Pinheiro, L.M.V.; Ventura, M.C.M.M.; Li, Y.; Moita, M.L.C.J. Application of QSPR/MLR methodology to solvatochromic behavior of quinoline in binary solvent HBD/DMF mixtures. *Journal of Molecular liquids* 154 (2010) 102-110,.
27. Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity: a Review. *QSAR & Combinatorial Science* 22 (2003) 1006-1026.

28. Venkatraman, V.; Alsberg, B.K.; Pinheiro, L.M.V. A quantitative structure-property relationship study of the photovoltaic performance of phenothiazine dyes. *Journal of Dyes and Pigments* 114 (2015) 69-77.
29. Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C.D. Chem-Bioinformatics: Comparative QSPR at the interface between chemistry and biology. *Chemical Reviews* 102 (2002) 783-812.
30. Pasha, F.A.; Muddassar, M.; Chung, H.W.; Cho, S.J.; Cho H. Hologram and 3D-quantitative structure toxicity relationship studies of azo dyes. *Journal of Molecular Modeling* 14 (2008) 293-302.
31. Ding, G.H.; Li, X.; Zhang, F.; Chen, J.W.; Huang, L.P.; Qiao, X.L. Mechanism-based quantitative structure-activity relationships on toxicity of selected herbicides to Chlorella vulgaris and Raphidocelis subcapitata. *Bulletin of Environmental Contamination and Toxicology* 83 (2009) 520-524.
32. Gadaleta, D.; Mangiatordi, G. F.; Catto, M.; Carotti, A.; Nicolotti, O. Applicability Domain for QSAR Models: Where Theory Meets Reality. *International Journal of Quantitative Structure-Property Relationships* 1 (2016) 45-63.
33. Yao, X.J.; Panaye, A.; Doucet, J.P.; Zhang, R.S.; Chen, H.F.; Liu, M.C.; Hu, Z.D.; Fan, B.T. Comparative study of QSAR/QSPR correlations using support Vector Machines, Radial Basis function Neural Networks, and Multiple Linear Regression. *Journal of Chemical Information and Computer Sciences. Comput. Sci* 44 (2004) 1257-1266.
34. Katritzky, A.R.; Lobanov, V.S.; Karelson, M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. *Chemical Society Reviews* 24 (1995) 279-287.
35. Taskinen, J.; Yliruusi, J. Prediction of physicochemical properties based on neural network modelling. *Advanced drug delivery reviews* 55 (2003) 1163-1183.
36. Livingstone, D.J.; Manallack, D.T. Neural Networks in 3D QSAR. *QSAR & Combinatorial Science* 22 (2003) 510-518.
37. Leardi, R. Genetic algorithms in chemometrics and chemistry: a review. *Journal of Chemometrics* 15 (2001) 559-569.
38. Katritzky, A.R.; Kuanar, M.; Slavov, S.; Hall, D. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chemical Society Reviews* 110 (2010) 5714-5789.
39. Banchero, M.; Manna, L. Comparative between multilinear and radical basis function Neural Network based QSPR models for the prediction of the critical temperature, critical pressure and acentric factor of organic compounds. *Molecules* 23 (2018) 1379-1391.
40. Rappoport, Z. The Chemistry of phenols.; John Wiley & Sons, Chichester, England ,2004; pp921-924.
41. ChemDraw Utra "Ultra-chemical structure drawing standard". Version 7. 2002. Copyright CambridgeSoft Coperation.
42. HyperChem Pro. Molecular Modeling system.Version 8. 2008. Copyright Hypercube, Inc.
43. Todeschini, R.; Consonni, V.; Pavan, M.; DRAGON, Version 4.5, 2005, Copyright TALETE srl.
44. Liu, H.; Gramatica, P. QSPR study of selective ligands for the thyroid hormone receptor beta. *Bioorganic & medicinal chemistry* 15 (2007) 5251-5261.
45. Organization for Economic Cooperation and Development, Guidance Document on the Validation of Quantitative Structure Activity Relationships (QSPR) Models, ENV/JM/MONO, *OECD Publishing*, Paris, 2 (2007).
46. Todeschini, R.; Ballabio, D.; Consonni, V.; Mauri, A.; Pavan, M. MOBYDIGS Software for Multilinear Regression Analysis and variable Subset Selection by Genetic Algorithm. Release I.1 for Windows, Milano, 2009.
47. Leardi, R.; Boggia, R.; Terrible, M. Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics* 6 (1992) 267-281.
48. Liu, P.; Long, W. Current mathematical methods used in QSPR/QSPR studies. *Intnational Journal of Molecular Sciences* 10 (2009) 1978-1998.
49. Todeschini, R.; Consonni, V.; Maiocchi, A. The K correlation index : theory development and its application in chemomerics. *Chemometrics and Inteligent Laboratory Systems* 46 (1999) 13-29.
50. Golbraikh, A.; Tropsha, A. Beware of q^2!. *Journal of molecular graphics and modeling* 20 (2002) 269-276.
51. Li, J.; Gramatica, P. The importance of molecular structures, endpoints' values, and predictivity parameters in QSPR research: QSPR analysis of a series of estrogen receptor binders. *Molecular diversity* 14 (2010) 687-696.
52. Tropsha, A.; Gramatica, P.; Gombar, V. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science* 22 (2003) 69–77.
53. Warren, S.M.; Walter, H. P. A logical calculus at the ideas immanent in Nervous Activity. *Bulletin of mathematical Biophysics* 5 (1943) 115-13.
54. Minsky, M.; Papert S. Perceptrons:An Introduction to Computational Geometry.; MIT Press: Cambridge, MAS, USA, 1969; pp.16-19.
55. J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, WileyVCH, Weinheim, 1999.
56. Svozil, D; Kvasnicka, V; Pospichal. Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems* 39 (1997) 43–62.
57. Veerasamy, R.; Rajak; H.; Jain A.; Sivadasan, S.; P, Christapher.; Ram, V.; A, Kishore. Validation of QSAR models-strategies and importance. *International Journal of Drug Design and Discovery* 2 (2011) 511–519.
58. Atabati, M.; Zarei, K.; Mohsennia, M. Prediction of λmax of 1,4-naphthoquinone derivatives using ant colony optimization. *Analytica Chimica Acta* 663 (2010) 7-10.
59. Todeschini, R.; Cosonni, V. Molecular Descriptors for Chemoinformatics. Wiley-VCH, Weinheim, Germany, 2009; p. 1257.
60. Bordás, B.; Bélai, I.N.; Kőmíves, T.S.; Theoretical molecular descriptors relevant to the uptake of persistent organic pollutants from soil by Zucchini. A QSAR study. *Journal of Agricultural and Food Chemistry* 59 (2011) 2863–2869.

Please cite this Article as:

Bouarra N., Nadji N., Nouri L., Boudjemaa A., Bachari K., Messadi D., Computer-assisted methods useful for the modeling of phenolic dyes wavelengths (λmax) using MLR and ANN methods, ***Algerian J. Env. Sc. Technology***, *7*:**3 (2021) 2013-2023**