Available online at www.derpharmachemica.com

ISSN 0975-413X
CODEN (USA): PCHHAX

Der Pharma Chemica, 2016, 8(19):8-21
(http://derpharmachemica.com/archive.html)

DFT-based QSAR Studies of cytotoxicity of phenothiazine derivatives as *in vitro* anti-cancer agents using the statistical analysis methods

Youness Boukarai[1*], Fouad Khalil[1] and Mohamed Bouachrine[2]

[1]LAC, Laboratory of Applied Chemistry, Faculty of Science and Technology, University Sidi Mohammed Ben Abdellah, Fez, Morocco
[2]ESTM, University Moulay Ismail, Meknes, Morocco

ABSTRACT

Phenothiazine and its derivatives are potent anticancer agents, these compounds inhibit cancer cells proliferation and tumor growth. A study of quantitative structure-activity relationship (QSAR) is applied to a set of 18 molecules derived from phenothiazine, in order to predict the anticancer biological activity of the test compounds and find a correlation between the different physic-chemical parameters (descriptors) of these compounds and its biological activity, using principal components analysis(PCA), multiple linear regression (MLR), multiple non-linear regression (MNLR) and the artificial neural network (ANN). We accordingly propose a quantitative model (non-linear and linear QSAR models), and we interpret the activity of the compounds relying on the multivariate statistical analysis. Density functional theory (DFT) with Becke's three parameter hybrid functional using the LYP correlation functional (B3LYP/6–31G (d)) calculations have been carried out in order to get insights into the structure, chemical reactivity and property information for the study compounds. The topological descriptors and the electronic descriptors were computed, respectively, with (ACD/ChemSketch; ChemBioOffice 14.0) and Gaussian 03W programs. A good correlation was found between the experimental activity and those obtained by MLR and MNLR respectively such as (R = 0,94 and R^2 = 0,885) and (R = 0,986 and R^2 = 0,973), this result could be improved with ANN such as (R = 0,988 and R^2 = 0,976) with an architecture ANN (6-1-1). To test the performance of the neural network and the validity of our choice of descriptors selected by MLR and trained by MNLR and ANN, we used cross-validation method (CV) such as (R = 0,975 and R^2 = 0,95) with the procedure leave-one-out (LOO). This study show that the MLR and MNLR have served to predict activities, but when compared with the results given by an 6-1-1 ANN model we realized that the predictions fulfilled by this latter was more effective and much better than other models. The statistical results indicate that this model is statistically significant and shows very good stability towards data variation in leave-one-out (LOO) cross validation.

Keywords: Anti-cancer,phenothiazine derivatives, DFT, QSAR, PCA, MLR, MNLR, ANN, CV.

Abbreviations: *QSAR,Quantitative Structure-Activity Relationship; PCA, Principal Component Analysis; MLR, Multiple Linear Regression; MNLR, Multiple Non-Linear Regression;ANN, Artificial Neural Networks; CV,Cross Validation; LOO-CV, Leave One Out Cross-Validation; R, Correlation Coefficient; R^2, Coefficient of Determination;R^2_{aj}, Adjusted Coefficient of Determination; q^2, Coefficient of Prediction;SD, Standard Deviation;DFT, Density Functional Theory; HOMO, Highest Occupied Molecular Orbital; LUMO, Lowest Unoccupied Molecular Orbital;µ, Total Dipole Moment;MW, Molecular Weight; MR, Molar Refractivity;MV, Molar Volume;Pc, Parachor; n, Refractive Index; γ, Surface Tension;D, Density;αe, Polarizability;LogP,Lipophilic; HBA, Hydrogen Bond Acceptor; HBD, Hydrogen Bond Donor;SSE, Sum of residual (Error) Squares; SSF, Sum of regression (Factor) Squares; MSE (V_E:Error Variance), Mean Squared Error;MSF (V_F: Factor Variance), Mean Squared Factor; F, Fishers F-statistic; F value, Significance level;p-value, Critical Probability*

INTRODUCTION

Drugs from phenothiazine family exhibit a wide range of biological activities which depend on their real structure:**[1]** neuroleptic action,**[2,3]** antidepressant,**[4]** and anticancer, antibacterial, antiviral activities,**[5,6]** anti-CaM activity, inhibition of the PKC activity, decrease of cell proliferation, and inhibition of the Pgp transport function **[7]**. Apart of their well known activity in nerve cells some phenothiazine derivatives were discovered to be anti-MDR effective chemo sensitizers in multidrug resistant (MDR) tumor cells. Their MDR reversing activity has been assessed in different resistant tumor cell lines**[8]**.

Quantitative structure-activity relationship (QSAR) tries to investigate the relationship between molecular descriptors that describe the unique physicochemical properties of the set of compounds of interest with their respective biological activity or chemical property **[9,10]**.

In this work we attempt to establish a quantitative structure-activity relationship between anticancer activity of a series of 18 bioactive molecules derived from phenothiazine and structural descriptors. Thus we can predict the anticancer activity of this group of organic compounds. Therefore we propose a quantitative model, and we try to interpret the activity of these compounds based on the different multivariate statistical analysis methods include:

* The Principal Components Analysis (PCA) has served to classify the compounds according to their activities and to give an estimation of the values of the pertinent descriptors that govern this classification. * The Multiple Linear Regression (MLR) has served to select the descriptors used as the input parameters for the Multiples Non-Linear Regression (MNLR) and Artificial Neural Network (ANN). * The artificial neural network (ANN) which is a nonlinear method, which allows the prediction of the activities.* Cross-validation (CV) to validate models used with the process leave-one-out (LOO).

MATERIAL AND METHODS

Experimental data

The Biological data used in this study were anti-cancer activity against MDR in P388 sensitive cell line(inhibition of multidrug resistant (MDR) tumor cells.(ED_{50})), a set of eighteen derivatives of phenothiazine. We have studied and analyzed the series of phenothiazine molecule consists of 18 selected derivatives that have been synthesized and evaluated for their anticancer activity in vitro against P388(in terms of -log (ED_{50})) **[11,12]**.This in order to determine a quantitative structure-activity relationship between anticancer activity and the structure of these molecules that are described by their substituents R and X.

The chemical structure of phenothiazine is represented in **Figure1**.

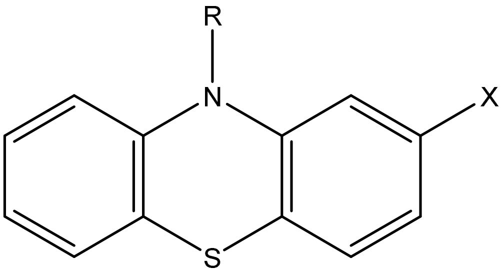
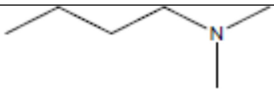
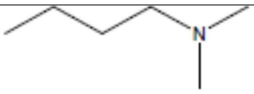
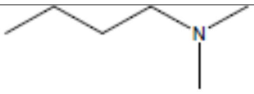
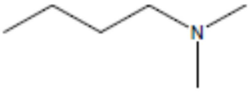
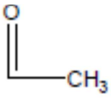
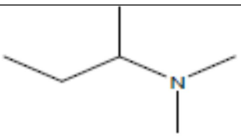
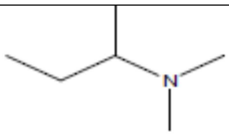
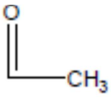
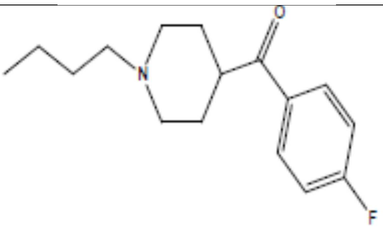
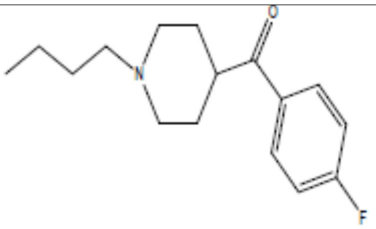
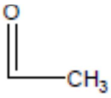
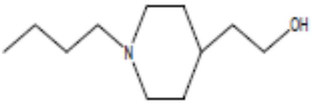
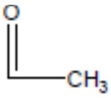
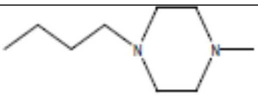
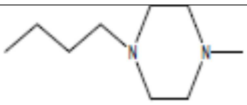
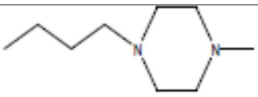
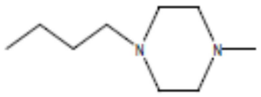
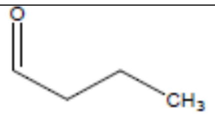
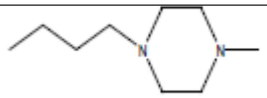
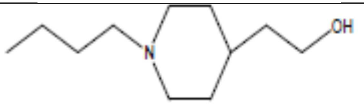
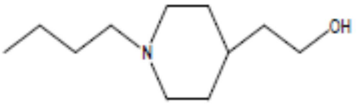
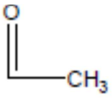
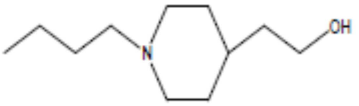
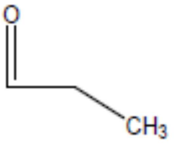
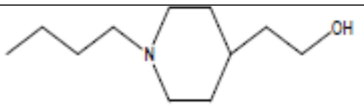


Figure1: The general structure of phenothiazine

The chemical structures of 18 compounds of phenothiazine used in this study and their experimental anti-cancer biological activity observedED_{50}(Cytotoxic concentration of drug effective required to inhibit the growth of P388than 50%) are collected from recent publications**[11,12]**.The observations are converted into logarithmic scale-log (ED_{50})in molar units (M) and are included in **Table1**.

Table1: Chemical structure and activity observed of phenothiazine derivatives against P388

N°	**Compound**	**R**	**X**	**$pED_{50}^{a}{}_{Obs}$**
1	Promazine		H	-1,602
2	Chlorpromazine		Cl	-1,079
3	Triflupromazine		CF_3	-1,079
4	Acepromazine			-1,477
5	Promethazine		H	-1,778
6	Acepromethazine			-1,602
7	Duoperone		CF_3	-1
8	AHR 06601			-1
9	Piperacetazine			-1,301
10	Perazine		H	-1,079
11	Prochlorperazine		Cl	-0,653
12	Trifluoperazine		CF_3	-0,653

13	Butaperazine			-0,653
14	Thioproperazine		$SO_2N(CH_3)_2$	-0,903
15	Perphenazine		Cl	-0,903
16	Acetophenazine			-1,176
17	Carphenazine			-1,079
18	Fluphenazine		CF_3	-0,903

[a] *pED$_{50}$ = -log (ED$_{50}$).*

Computational methods

An attempt has been made to correlate the activity of these compounds with various physicochemical parameters. DFT (density functional theory) methods were used in this study. The 3D structures of the molecules were generated using the Gauss View 3.0, and then, all calculations were performed using Gaussian 03W program series, Geometry optimization of 18 compounds was carried out by B3LYP functional employing 6–31G (d) basis set **[13,14]**. The geometry of all species under investigation was determined by optimizing all geometrical variables without any symmetry constraints. ChemSketch and ChemBioOffice programs**[15-17]**are employed to calculate the others molecular descriptors.

Calculation of molecular descriptors

Calculation of descriptors using Gaussian 03W

From the results of the DFT calculations, the quantum chemistry descriptors were obtained for the model building as follows: the highest occupied molecular orbital energy (E_{HOMO} (eV)), the lowest unoccupied molecular orbital energy (E_{LUMO} (eV)), the total dipole moment of the molecule (μ (Debye)).

Calculation of descriptors using ACD/ChemSketch and ChemBioOffice 14.0

Advanced chemistry development's ACD/ChemSketch program was used to calculate Molecular Weight (MW), Molar Refractivity (MR (cm^3)),Molar Volume (MV (cm^3)), Parachor(Pc (cm^3)), Density (D (g/cm^3)), Refractive Index (n), Surface Tension(γ (dyne/cm)) and Polarizability (α_e (cm^3)) **[15,16]**.

Steric, thermodynamic descriptors are calculated using ACD/ChemSketch and ChemBioOffice 14.0**[17]**after optimization of the energy for each compound using the MM2 method (force field method with Gradient Setting Root Mean Square (RMS) 0.1 kcal mol^{-1}) **[18]**.

In this work 14 descriptors were chosen to describe the structure of the molecules constituting the series to study: the highest occupied molecular orbital energy (E_{HOMO} (eV)), the lowest unoccupied molecular orbital energy (E_{LUMO} (eV)),the total dipole moment (μ (Debye)), the molecular weight (MW), the molar refractivity (MR (cm^3)),the molar volume (MV (cm^3)), the parachor (Pc(cm^3)), the refractive index (n), the surface tension(γ (dyne/cm)), the density (D (g/cm^3)), the polarizability (α_e (cm^3)), the lipophilic (LogP), the hydrogen bond acceptor (HBA) and the hydrogen bond donor (HBD).

Statistical analysis

To explain the structure-activity relationship, these 14 descriptors are calculated for 18 molecules (**Table2**) using the Gaussian03W, Gauss View, ChemSketch and ChemBioOffice 14.0.

Table2: The values of the 14 chemical descriptors

	MW	MR	MV	Pc	n	γ	D	α_e	LogP	HBA	HBD	E_{HOMO}	E_{LUMO}	µ
1	298,445	92,490	267,4	689,80	1,608	44,2	1,115	36,66	4,138	2	0	-0,181	-0,012	2,973
2	332,890	97,380	279,4	726,90	1,614	45,8	1,191	38,60	4,696	2	0	-0,191	-0,023	2,564
3	366,443	97,470	300,9	751,80	1,561	38,9	1,217	38,64	5,059	5	0	-0,196	-0,041	3,337
4	340,482	102,51	298,9	773,30	1,601	44,7	1,138	40,64	3,450	3	0	-0,191	-0,062	4,904
5	298,445	92,450	267,8	687,80	1,606	43,4	1,114	36,65	4,003	2	0	-0,185	-0,009	2,081
6	340,482	102,47	299,3	771,30	1,600	44,0	1,137	40,62	3,315	3	0	-0,191	-0,061	1,833
7	528,604	138,76	415,4	1066,9	1,582	43,4	1,272	55,00	6,766	7	0	-0,197	-0,071	2,736
8	502,642	143,80	413,4	1088,5	1,612	48,0	1,215	57,01	5,157	5	0	-0,194	-0,071	3,760
9	424,598	125,03	367,3	959,20	1,596	46,4	1,155	49,56	3,731	4	1	-0,196	-0,062	4,330
10	353,524	107,80	311,7	806,90	1,607	44,8	1,133	42,73	3,894	3	0	-0,167	-0,008	2,320
11	387,969	112,69	323,7	844,00	1,613	46,2	1,198	44,67	4,452	3	0	-0,171	-0,019	2,834
12	421,522	112,78	345,3	868,90	1,566	40,0	1,220	44,71	4,815	6	0	-0,172	-0,039	3,874
13	423,614	127,09	376,3	970,50	1,590	44,2	1,125	50,38	4,277	4	0	-0,171	-0,060	4,709
14	460,655	130,65	377,3	991,20	1,609	47,6	1,220	51,79	3,173	5	0	-0,165	-0,084	4,953
15	417,007	119,89	347,7	912,80	1,605	47,4	1,199	47,53	4,977	3	1	-0,172	-0,019	2,168
16	424,598	125,03	367,3	959,20	1,596	46,4	1,155	49,56	3,731	4	1	-0,173	-0,061	3,842
17	438,625	129,66	383,8	999,30	1,590	45,9	1,142	51,40	4,385	4	1	-0,173	-0,060	3,708
18	450,560	119,98	369,3	937,70	1,563	41,5	1,219	47,56	5,340	6	1	-0,174	-0,040	3,493

The study we conducted consists of:

-The principal component analysis (PCA), the multiple linear regressions (MLR), and the non-linear regression (MNLR) available in the XLSTAT 15software**[19]**.

-The Artificial Neural Network (ANN) and the leave-one-out cross validation (CV-LOO)are done on Matlab 7 using a program written in C language.

The structures of the molecules based on phenothiazine derivatives were studied by statistical methods based on the principal component analysis (PCA). PCA is a statistical technique useful for summarizing all the information encoded in the structures of the compounds. It is also very helpful for understanding the distribution of the compounds. This is an essentially descriptive statistical method which aims to present, in graphic form, the maximum of information contained in the data **Table2** and**Table3**.

Table3: The correlation matrix (Pearson (n)) between different obtained descriptors

Variables	MW	MR	MV	Pc	n	γ	D	α_e	LogP	HBA	HBD	E_{HOMO}	E_{LUMO}	µ	pED50
MW	**1**														
MR	0,953	**1**													
MV	**0,977**	**0,986**	**1**												
Pc	**0,965**	**0,997**	**0,995**	**1**											
n	-0,284	-0,067	-0,232	-0,142	**1**										
γ	0,204	0,431	0,282	0,373	0,821	**1**									
D	0,687	0,466	0,518	0,485	-0,380	-0,148	**1**								
α_e	0,953	**1**	**0,986**	**0,997**	-0,067	0,431	0,466	**1**							
LogP	0,519	0,328	0,398	0,352	-0,442	-0,315	0,700	0,328	**1**						
HBA	0,829	0,654	0,754	0,695	-0,699	-0,343	0,756	0,654	0,576	**1**					
HBD	0,297	0,342	0,370	0,380	-0,205	0,234	-0,025	0,342	0,017	0,111	**1**				
E_{HOMO}	0,052	0,124	0,112	0,122	0,071	0,167	-0,135	0,124	-0,272	-0,046	0,198	**1**			
E_{LUMO}	-0,663	-0,687	-0,695	-0,695	0,173	-0,205	-0,297	-0,687	0,042	-0,581	-0,101	0,212	**1**		
µ	0,378	0,423	0,441	0,438	-0,191	0,100	0,051	0,424	-0,244	0,368	0,098	0,127	-0,648	**1**	
pED50	0,610	0,535	0,569	0,546	-0,283	-0,004	0,587	0,535	0,407	0,530	0,066	0,470	-0,145	0,299	**1**

The multiple linear regression statistic technique is used to study the relation between one dependent variable and several independent variables. It is a mathematic technique that minimizes differences between actual and predicted values. It has served also to select the descriptors used as the input parameters in the multiple non-linear regression (MNLR)and artificial neural network (ANN).

The (MLR) and the (MNLR) were generated to predict cytotoxic effects ED_{50} activities of phenothiazine derivatives. Equations were justified by the correlation coefficient (R), the coefficient of determination (R^2), the mean squared error (MSE), the Fishers F-statistic (F) and the significance level(F value) **[20-21]**.

ANN is artificial systems simulating the function of the human brain. Three components constitute a neural network: the processing elements or nodes, the topology of the connections between the nodes, and the learning rule by which new information is encoded in the network. While there are a number of different ANN models, the most frequently used type of ANN in QSAR is the three-layered feed-forward network **[22]**. In this type of networks, the neurons are arranged in layers (an input layer, one hidden layer and an output layer). Each neuron in any layer is fully connected with the neurons of a succeeding layer and no connections are between neurons belonging to the same layer.

Cross-validation is a popular technique used to explore the reliability of statistical models. Based on this technique, a number of modified data sets are created by deleting in each case one or a small group of molecules, these procedures are named respectively “leave-one-out” and “leave-some-out” **[23-25]**. For each data set, an input-output model is developed. In this study we used, the leave-one-out (LOO) procedure.

RESULTS AND DISCUSSION

Data set for analysis

The QSAR analysis was performed using the -log (ED_{50}) of the 18 selected molecules that have been synthesized and evaluated for their anticancer activity in vitro against MDR in P388 sensitive cell line(experimental values) **[11,12]**. The exploitation of experimental data observed by the use of mathematical and statistical tools is an effective method to find new chemical compounds with high anticancer activity. The values of the 14 chemical descriptors as shown in Table2.

The principle is to perform in the first time, a main component analysis (PCA), which allows us to eliminate descriptors that are highly correlated(dependent), then perform a decreasing study of MLR based on the elimination of descriptors aberrant until a valid model (including the critical probability: **p-value < 0.05** for all descriptors and the model complete).

Principal Components Analysis (PCA)

The totality of the 14 descriptors (variables) coding the 18 molecules was submitted to a principal components analysis (PCA). 15 principal components were obtained (**Figure2**). The first three axes F1, F2 and F3 contributing respectively 49.34 %, 18.54 % and 9.88 % to the total variance, the total information is estimated to a percentage of 77.76%.

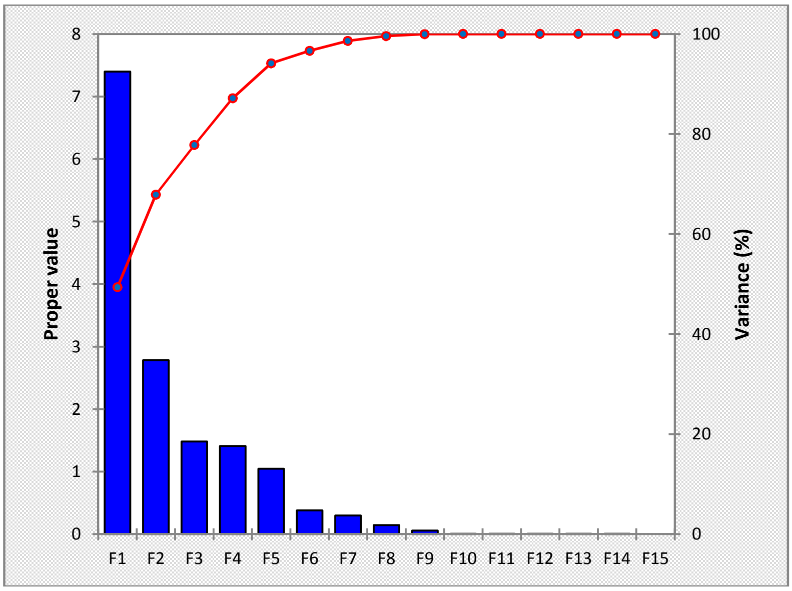


Figure2: The principal components and there variances

The Pearson correlation coefficients are summarized in the above Table3. The obtained matrix provides information on the negative or positive correlation between variables. The principal component analysis (PCA) was conducted to identify the link between the different variables. Correlations between the 14 descriptors are shown in Table3 as a correlation matrix and in **Figure3** these descriptors are represented in a correlation circle.

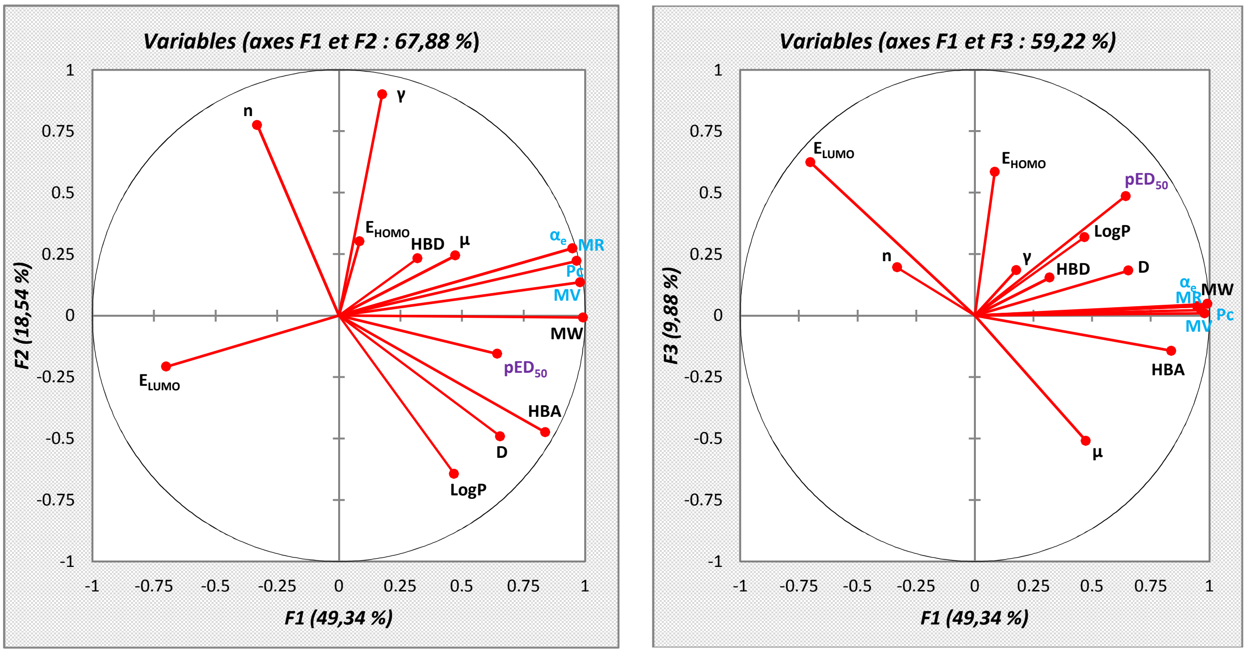


Figure3: Correlation circles

(MR, α_e) are perfectly correlated (r=1), both variables are redundant.
Pc, MR andα_e are highly correlated(r (Pc, MR) = 0,997; r (Pc, α_e) = 0,997).
Pc, MV and MW are highly correlated(r (Pc, MV) = 0,995; r (Pc, MW) = 0,965).
MV, MR and α_e are highly correlated(r (MV, MR) = 0,986; r (MV,α_e) = 0,986).
MV and MW are highly correlated (r (MV, MW) = 0,977).
The following variables then removed are: (α_e), (Pc) and (MV).

Multiple Linear Regression (MLR)

In order to propose a mathematical model linking the descriptors and activity, and to evaluate quantitatively the substituent's physicochemical effects on the activity of the totality of the set of these 18 molecules, we presented the data matrix which is the corresponding physicochemical variables different substituent’s from 18 molecules to a multiple linear regression analysis. This method used the coefficients R, R^2, R^2_{aj}, q^2, SD, MSE, MSF and the F-values to select the best regression performance. Where R is the correlation coefficient; R^2 is the coefficient of determination; R^2_{aj} is the adjusted coefficient of determination; q^2 is the coefficient of prediction; SD is the standard deviation; MSE is the mean squared error; MSF is the mean squared factor; F is the Fisher F-statistic.

Treatment with multiple linear regression is more accurate because it allows you to connect the structural descriptors for each activity of 18 molecules to quantitatively evaluate the effect of substituent. The selected descriptors are:

MR, **n, D, HBA, HBD** and **E_{HOMO}**.
The QSAR model built using multiple linear regression (MLR) method is represented by the following equation:

$$\boxed{\mathbf{N = 18R = 0.94R^2 = 0.885 \quad F = 14.073 \quad MSE = 0.02}}$$

$$\mathbf{pED_{50MLR}} = 36{,}335 + 3{,}139E\text{-}02\mathbf{MR} - 28{,}023\mathbf{n} + 7{,}259\mathbf{D} - 0{,}480\mathbf{HBA} - 0{,}423\mathbf{HBD} + 15{,}901\mathbf{E_{HOMO}}$$

(**Equation 1**)

Higher correlation coefficient and lower mean squared error (MSE) indicate that the model is more reliable. And the Fisher's F test is used. Given the fact that the probability corresponding to the F value is much smaller than **0.05**, it mean that we would be taking a lower than 0.01 % risk in assuming that the null hypothesis is wrong. Therefore, we can conclude with confidence that the model do bring a significant amount of information.

The elaborated QSAR model reveals that the anticancer activity could be explained by a number of topologic factors. The negative correlation of the Refractive Index(n), the Hydrogen Bond Acceptor (HBA)and the Hydrogen Bond Donor (HBD)with the ability to displace the phenothiazine activity reveals that a decrease in the value of pED_{50}, While the positive correlation of the descriptors (Molar Refractivity (MR), the Density (D)and the Highest Occupied Molecular Orbital Energy (E_{HOMO})) with the ability to displace the phenothiazine activity reveals that an increase in the value of pED_{50}.

With the optimal MLR model, the values of predicted activities **$pED_{50\ MLR}$** calculated from equation1 and the observed values are given in **Table4**. The correlations of predicted and observed activities are illustrated in **Figure4**. The descriptors proposed in equation1 by MLR were, therefore, used as the input parameters in the multiples non-linear regression (MNLR) and artificial neural network (ANN).

The correlation between MLR calculated and experimental activities are very significant as illustrated in Figure4 and as indicated by R and R^2 values.

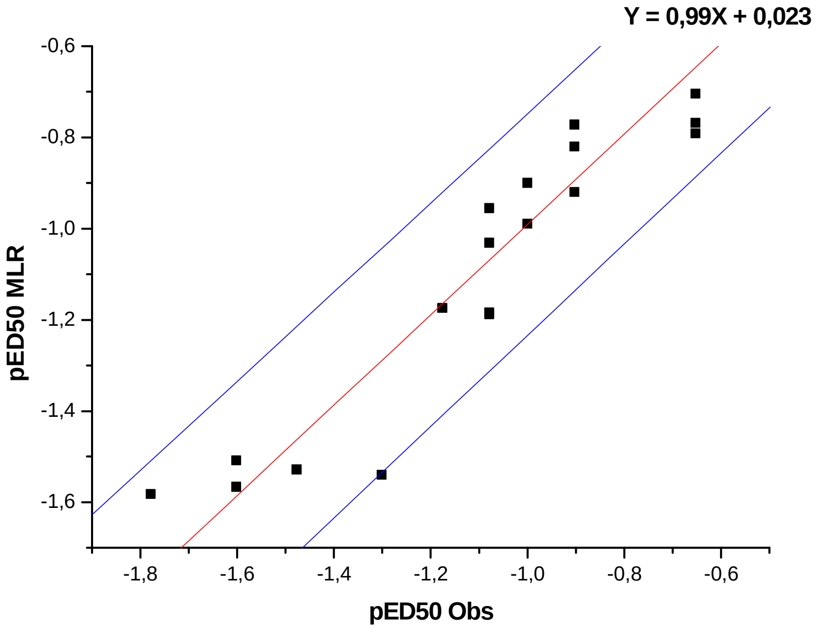


Figure4: Correlations of observed and predicted activities calculated using MLR

Table4: The observed, the predicted activities (pED_{50}), according to different methods MLR, MNLR, ANN and CV for the 18 derivatives of phenothiazine

N°	**pED_{50Obs}**	**pED_{50MLR}**	**pED_{50MNLR}**	**$pED_{50\ ANN}$**	**pED_{50CV}**
1	-1,602	-1,566	-1,703	-1.597	-1,522
2	-1,079	-1,188	-1,069	-1.118	-1,159
3	-1,079	-1,031	-1,100	-1.095	-1,152
4	-1,477	-1,528	-1,491	-1.621	-1,433
5	-1,778	-1,582	-1,715	-1.673	-1,628
6	-1,602	-1,508	-1,482	-1.609	-1,703
7	-1,000	-0,900	-1,011	-0.932	-0,986
8	-1,000	-0,989	-0,991	-1.035	-0,889
9	-1,301	-1,540	-1,357	-1.286	-1,222
10	-1,079	-1,184	-1,062	-1.078	-1,120
11	-0,653	-0,791	-0,707	-0.663	-0,754
12	-0,653	-0,768	-0,701	-0.653	-0,638
13	-0,653	-0,704	-0,660	-0.667	-0,634
14	-0,903	-0,820	-0,865	-0.942	-0,807
15	-0,903	-0,772	-0,877	-0.897	-0,901
16	-1,176	-1,174	-1,249	-1.170	-1,137
17	-1,079	-0,955	-1,047	-1.073	-1,020
18	-0,903	-0,920	-0,832	-0.937	-0,910

Validation criteria of the MLR model (Anova: Analysis Of Variance)

To validate the correlation equation provided by the statistical method of multiple linear regression (MLR), different criteria may be used**[26,27]**.

Overall assessment of the regression

Table 5 summarizes the variances, the degrees of freedom (df), the sums of squares (SS), Fisher's F value (F_{exp}) and overall p-value value of the model.

Table 5: Variance analysis

Source	SS	df	Variance	F-exp	p-Value
Regression (Factor)	**1.671**	**6**	**0.279**	**14.073**	**0.000**
Residual (Error)	**0.218**	**11**	**0.020**	**-**	**-**
Total	**1.889**	**17**	**0.299**	**-**	**-**

-The variability not explained by the model is the sum of residual squares **SSE** = **0.218** with a degree of freedom equal to **11** (**N-p-1= 18-6-1**).
-The variability explained by the model is the sum of regression squares **SSF** = **1.671** with a degree of freedom equal to **6** (**N-(N-p-1)-1= p =18-11-1**).
- The results seem excellent and the model is significant because we achieved good results for **F-exp** Fisher (**14,073**) and lower overall p-value at **α (F value) = 0.05** level (**p-value <0.05**).

Test for significance
-The first test that comes to mind is the significance of the correlation i.e. the correlation coefficient **R** is it significantly different from (**0**)?

-The test is:$\mathbf{H_0}$: R = 0
$\mathbf{H_1}$:R ≠ 0
-If the correlation coefficient is zero, we reject the hypothesis $\mathbf{H_0}$ (null hypothesis) and accept $\mathbf{H_1}$ (not null hypothesis). So the model is significant.

Confidence Interval (CI)
-The confidence interval (**CI**) **1-α** is a range of values that has a chance of **1-α** to contain the true value of the estimated parameter.

-If the p-value value exceeds (**0.05**), we reject H_1 and $H_0$is accepted. So the model is not significant.
-If α> p-value, reject H_0 (H_1 acceptance).
- If α <p-value, H_0 acceptance (reject H_1).

Student test
-The Student law with (**N-p-1**) degree of freedom $\boldsymbol{t_{calc}}$ is written:

$$\boldsymbol{t_{calc}=}\left(\frac{\boldsymbol{R}}{\sqrt{\frac{\boldsymbol{1-R^2}}{\boldsymbol{N-p-1}}}}\right)$$

-H_0 is rejected (null hypothesis) where: $\boldsymbol{t_{calc}} > \boldsymbol{t}_{\left(\mathbf{1-\frac{\alpha}{2}}\right),(\mathbf{N-p-1})}$

Where $\boldsymbol{t}_{\left(\mathbf{1-\frac{\alpha}{2}}\right),(\mathbf{N-p-1})}$ is the value of the Student law for $(\mathbf{N-p-1})$ degree of freedom, a probability$\left(\mathbf{1-\frac{\alpha}{2}}\right)$.

-In our case we have **N = 18** and **R = 0.94**. This corresponds to $\boldsymbol{t_{calc}}$ **= 9.193**, one rejects H_0 (null hypothesis) where: $\boldsymbol{t_{calc}} > \boldsymbol{t}_{\left(\mathbf{1-\frac{\alpha}{2}}\right),(\mathbf{N-p-1})}$.

-According to the Student table $\left(\mathbf{1-\frac{\alpha}{2}}\right)$**= 0.975** and **N = 18**isobtained $\boldsymbol{t}_{(\mathbf{0.975,11})}$ **= 2.201**.
$\boldsymbol{t_{calc}} > \boldsymbol{t}_{(\mathbf{0.975,11})}$ then we reject the null hypothesis H_0.

Fisher test
Analysis of variance (**V**) was used to test the equality of means, is called the **F statistic of Fisher**.

-Hypothesis $\mathbf{H_0}$: SSF=SSE ($V_F = V_E$) Where (Error Variance) $\mathbf{V_E}$ **= MSE**
-against hypothesis $\mathbf{H_1}$:SSF> SSE ($V_F > V_E$) Where (Factor Variance) $\mathbf{V_F}$**= MSF**

-The Fisher F is calculated according to the following equation:

$$\left\{ F_{exp}=\frac{\boldsymbol{VF}}{\boldsymbol{VE}}=\frac{\boldsymbol{MSF}}{\boldsymbol{MSE}}=\frac{\boldsymbol{SSF}/\boldsymbol{p}}{\boldsymbol{SSE}/_{\boldsymbol{N-p-1}}}\right.$$

-To a threshold of (**0.05**) comparing$\boldsymbol{F_{exp}}$ obtained by the theoreticalcalculation and thatobtainedfromFisher's table$\boldsymbol{F_{(p,N-p-1)}}$ for one degree of freedom (**p, N-p-1**) with**p = 6** and **N = 18**, such as **(N-p-1) = 11**.
-We Accept H_1 if $\boldsymbol{F_{exp}} > \boldsymbol{F_{(6,11)}}$.
-We then find$\boldsymbol{F_{(6,11)}}$**= 3.09**and$\boldsymbol{F_{exp}}$ **= 14.073**, so we accept H_1 and H_0isrejected.

Correlation Coefficient: R
This coefficient determines the variance of the target activity is explained by the model of QSAR i.e. by the regression of target activity based on the initial activity.

$$\mathbf{R} = \sqrt{1 - \frac{SSE}{SST}}$$

-A good correlation between the target activity and initial activity if **R** is closer to **1**.
-A non-linear correlation between the target activity and initial activity if **R** is closer to **0**.
-In our case we have **R = 0.94**, so a good correlation was shown between the observed activity and that obtained by **MLR**.

Coefficient of Determination: R²
The coefficient of determination $\mathbf{R^2}$, gives the rate of explanation or percentage of the variation of **Y** (endogenous variables) explained by the variation in **X** (exogenous variable).

$$\mathbf{R^2} = \frac{SSF}{SST}$$

-In our case we have$\mathbf{R^2}$ **= 0.885**, this figure means that **88.5%** of the variable Y (activity) is attributable to the variation in the variable X (descriptors), which indicates that this model is statistically explanatory.

Adjusted Coefficient of Determination: R²aj
The overall quality of the linear regression is measured by the coefficient of determination ($\mathbf{R^2_{aj}}$) "adjusted" taking into account the degree of freedom.

$$\mathbf{R^2_{aj} = 1 - \frac{N-1}{N-p-1}(1-R^2)}$$

-With: N = 18, p = 6 and R^2 = 0.885.
-In our case we have$\mathbf{R^2_{aj}}$ **= 0.822**, so the overall quality of the MLR is best. This indicates that this model is statistically significant.

Coefficient of Prediction: q²
The $\mathbf{q^2}$ value is used as the determining factor in selection of optimal models. The coefficient of prediction ($\mathbf{q^2}$) was calculated using:

$$\mathbf{q^2 = 1 - \frac{VE}{SST} = 1 - \frac{MSE}{SST}}$$

-SST: sum of total squares.
-In our case we have$\mathbf{q^2}$ **= 0.989> 0.6**, So the predictive power of this model is very significant, which shows that the model proposed in this paper is able to predict activity with a great performance, and that the selected descriptors are pertinent. This means that the prediction of the new compounds is feasible.
-we can enjoy the performance of the predictive power of this model to explore and propose new molecules could be active.

Standard Deviation: SD
The standard deviation (**SD**) measures the variation in the target activity is not explained by the QSAR model. In particular, over the standard deviation is small, the correlation is best.

$$SD = \sqrt{\frac{SSE}{N-p-1}} = \sqrt{VE} = \sqrt{MSE}$$

-N: (**N = 18**) number of data points considered.
-p: (**p = 6**) number of restrictions on the degrees of freedom (equal to the number of parameters).
-In our case we have **SD = 0.141**, so the correlation between the observed activity and that obtained by MLR is best.

Multiples Non-Linear Regression (MNLR)
We have used also the technique of nonlinear regression model to improve the structure-activity relationship to quantitatively evaluate the effect of substituent. We have applied to the data matrix constituted obviously from the

descriptors proposed by MLR corresponding to the 18 molecules. The coefficients R, R^2, and the F-values are used to select the best regression performance. We used a pre-programmed function of XLSTAT following:
$\mathbf{Y = a + (bX_1 + cX_2 + dX_3 + eX_4 \ldots) + (fX_1^2 + gX_2^2 + hX_3^2 + iX_4^2 \ldots)}$

Where a, b, c, d…represent the parameters and X_1, X_2, X_3, X_4…: represent the variables. The resulting equations:
$\mathbf{pED_{50MNLR}}$= 1053,755+ 0,178**MR**-1361,509**n**+98,785**D**-0,956**HBA** -0,531**HBD**+ 220,800 $\mathbf{E_{HOMO}}$-5,931E-04(**MR**)^2+419,680 (**n**)^2-40,574 (**D**)^2+5,896E-02 (**HBA**)^2+ 585,623 ($\mathbf{E_{HOMO}}$)^2

(**Equation 2**)

N = 18R = 0.986R^2 = 0.973MSE = 0.010

With the optimal MNLR model, the values of predicted activities $\mathbf{pED_{50\ MNLR}}$ calculated from equation2 and the observed values are given in Table4. The correlations of predicted and observed activities are illustrated in **Figure5.** The correlation between MNLR calculated and experimental activities are very significant as illustrated in Figure5 and as indicated by R and R^2 values.

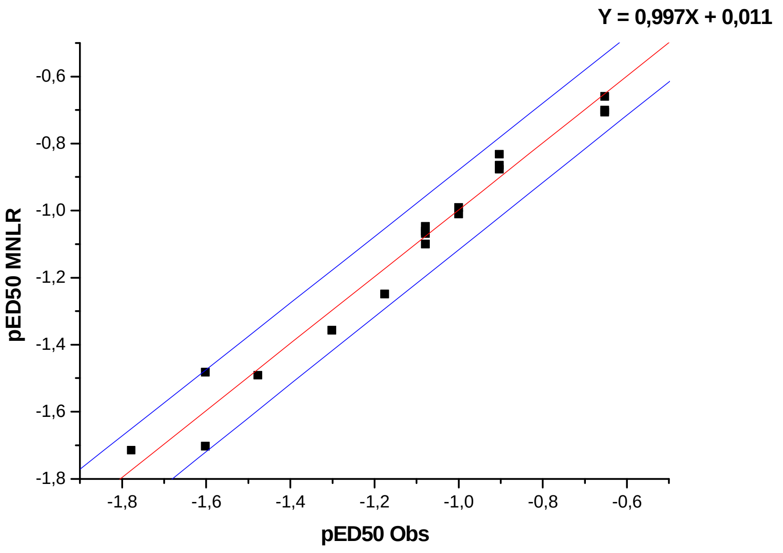


Figure5: Correlations of observed and predicted activities calculated using MNLR

Artificial Neural Networks (ANN)

In order to increase the probability of good characterization of studied compounds, artificial neural networks (ANN) can be used to generate predictive models of quantitative structure-activity relationships (QSAR) between a set of molecular descriptors obtained from the MLR, and observed activity. The ANN calculated activities model were developed using the properties of several studied compounds. Some authors **[28,29]** have proposed a parameter **ρ**, leading to determine the number of hidden neurons, which plays a major role in determining the best ANN architecture defined as follows:

ρ = (Number of data points in the training set /Sum of the number of connections in the ANN)
In order to avoid over fitting or under fitting, it is recommended that **1.8 < ρ < 2.3[30]**.The output layer represents the calculated activity values pIC_{50}. The architecture of the ANN used in this work (**6-1-1**), **ρ =2**.

The values of predicted activities $\mathbf{pED_{50\ ANN}}$ calculated using ANN and the observed values are given in Table4. The correlations of predicted and observed activities are illustrated in **Figure6**.

The correlation between ANN calculated and experimental activities are very significant as illustrated in Figure6 and as indicated by R and R^2 values.

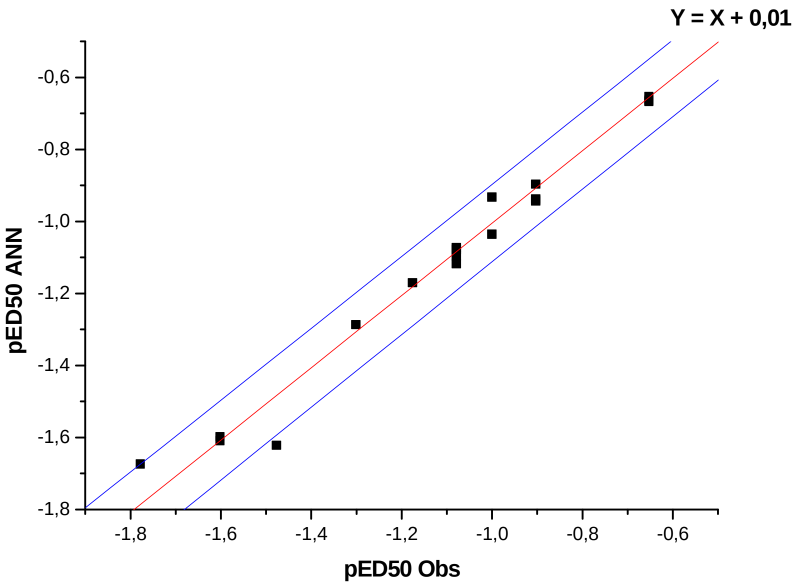


Figure6: Correlations of observed and predicted activities calculated using ANN

N = 18R = 0.988R^2 = 0.976

The obtained squared correlation coefficient (R^2) value confirms that the artificial neural network result were the best to build the quantitative structure activity relationship models.

It is important to be able to use ANN to predict the activity of new compounds. To evaluate the predictive ability of the ANN models, '**Leave-one-out**' is an approach particularly well adapted to the estimation of that ability.

Cross Validation (CV)

To test the performance of the neural network and the validity of our choice of descriptors selected by MLR and trained by MNLR and ANN, we used cross-validation method (CV) with the procedure leave-one-out (LOO). In this procedure, one compound is removed from the data set, the network is trained with the remaining compounds and used to predict the discarded compound. The process is repeated in turn for each compound in the data set.

In this paper the 'leave-one-out' procedure was used to evaluate the predictive ability of the ANN.

The values of predicted activities $\mathbf{pED_{50\ CV}}$ calculated using CV and the observed values are given in Table4. The correlations of predicted and observed activities are illustrated in **Figure7**.

The correlation between CV calculated and experimental activities are very significant as illustrated in Figure7 and as indicated by R and R^2 values.

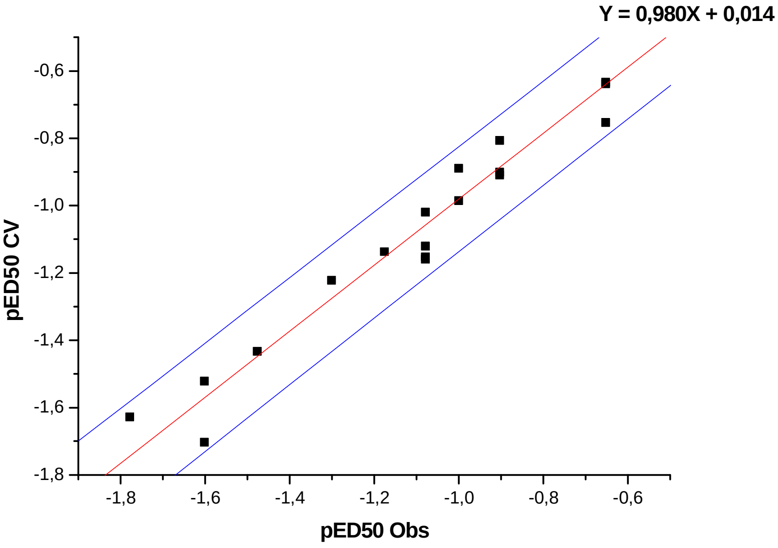


Figure7: Correlations of observed and predicted activities calculated using CV
N = 18R = 0.975R^2 = 0.95

The good results obtained with the cross validation, shows that the model proposed in this paper is able to predict activity with a great performance, and that the selected descriptors are pertinent.

The results obtained by MLR and MNLR are very sufficient to conclude the performance of the model. Even if it is possible that this good prediction is found by chance we can claim that it is a positive result. So, this model could be applied to all derivatives of phenothiazine accordingly to Table1 and could add further knowledge in the improvement of the search in the domain of inhibitors of anti-cancer agents.

A comparison of the quality of MLR, MNLR and ANN models shows that the ANN models have substantially better predictive capability because the ANN approach gives better results than MLR and MNLR. ANN was able to establish a satisfactory relationship between the molecular descriptors and the activity of the studied compounds. A good correlation was obtained with cross validation **R_{CV}= 0.975**. So the predictive power of this model is very significant.The results obtained in this study, showed that models MLR, MNLR and ANN are validated, which means that the prediction of the new compounds is feasible.

CONCLUSION

In this study, three different modelling methods, MLR, MNLR and ANN were used in the construction of a QSAR model for the anti-cancer agents and the resulting models were compared. It was shown the artificial neural network ANN results have substantially better predictive capability than the MLR and MNLR, yields a regression model with improved predictive power, we have established a relationship between several descriptors and the anticancer activity in satisfactory manners. The good results obtained with the cross validation CV, shows that the model proposed in this paper is able to predict activity with a great performance, and that the selected descriptors are pertinent.

The accuracy and predictability of the proposed models were illustrated by the comparison of key statistical terms like R or R^2 of different models obtained by using different statistical tools and different descriptors has been shown in Table4. It was also shown that the proposed methods are a useful aid for reduction of the time and cost of synthesis and activity determination of anti-cancer agents(compounds based on phenothiazine).

Furthermore, we can conclude that studied descriptors, which are sufficiently rich in chemical, electronicand topological information to encode the structural feature and have a great influence on the activity may be used with other descriptors for the development of predictive QSAR models.

Previous studies QSAR already performed on the same set of phenothiazine using cross validation, obtained a correlation coefficient (**R = 0.897**) **[31].** In this study the correlation coefficient obtained from the MLR (**R_{MLR} = 0.94**), by using a variety of descriptors, is very important and this coefficient improved by using MNLR and ANN respectively($\mathbf{R_{MNLR} = 0.986}$) and ($\mathbf{R_{ANN} = 0.988}$) so the proposed model is very significant and its performance is tested by cross-validation method CV ($\mathbf{R_{CV} = 0.975}$).

Thus, grace to QSAR studies, especially with the ANN that has allowed us to improve the correlation between the observed biological activity and that predicted, we can enjoy the performance of the predictive power of this model to explore and propose new molecules could be active.

Acknowledgment
We gratefully acknowledge to the group of Applied Chemistry Laboratory.

REFERENCES

[1] B. M. Mlodawska, K. Suwinska, K. Pluta, and M. Jelen, *J. Mol. Struct*. 1015, 94, (**2012)**.
[2] H. Laborit, P. Huguenard, and R. Allaume, *Presse. Med*. 60, 206, (**1952**).
[3] Y. Mizuno, K. Sato, T. Sano, R. Kurihana, T. Kojima, Y. Yamakawa, A. Ishii, and Y. Katsumata, *Legal Medecine*4, 207, **2002**.
[4] I. M. Tsakovska, Bioorg. *Med. Chem*. 11, 2889, (**2003**).
[5] N. Motohashi, M. Kawase, S. Saito, and H. Sakagami, Curr. *Drug Targets*7, 237, (**2000**).
[6] S. G. Dasgupta, Y. Dastridara, and N. Shirataki, *Top Heterocycl*. 15, 67, (**2008**).
[7] A. Jaszczyszyn, K. Gasiorowski, P. Swiatek, W. Malinka, K. Cieslik-Boczula, J. Petrus, and B. Czarnik-Matusewicz, *Pharmacological Reports* 64, 16, (**2012**).
[8] A. Poła, K. Michalak, A. Burliga, N. Motohashi, and M. Kawase, *European Journal of Pharmaceutical Sciences*. 21, 421, (**2004**).
[9] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna& V. Prachayasittikul, *J. Excli*. 8 (**2009**) 74-88.
[10] C. Nantasenamat, C. Isarankura-Na-Ayudhya& V. Prachayasittikul, *J. Expert Opin. Drug Discov*. 5(7) (**2010**) 633-654.
[11] I. M. Tsakovska, Bioorg. *Med. Chem*. 11, 2889, (**2003**).
[12] A. Ramu and N. Ramu, N. Cancer Chemother. *Pharmacol*.30, 165, (**1992**).
[13] C. Adamo& V. Barone, J. Chem. Phys. Lett., 330 (**2000**) 152. M. Parac& S. Grimme, *J. Phys. Chem*., 106 (**2003**) 6844. Y. Yamaguchi, S. Yokoyama & S. Mashiko, *J. Chem. Phys*., 116 (**2002**) 6541.
[14] L. Becker, K. Hinrichs& U. Finke, *A New Algorithm for Computing Joins with Grid Files, In Proc. of the 9th International Conference on Data Engineering, Vienna, Austria*. (**1993**) 190-197. S.J. Lee, J. Fink, A.B. Balantekin, M.R. Strayer, A.S Umar, P.G. Reinhard, J.A. Maruhn, & W. Greiner, *Reply, Phys. Rev. Lett*. 60 (**1988**) 163.
[15] Advanced Chemistry Development Inc., Toronto, Canada. (**2009**). http://www.acdlabs.com/resources/freeware/chemsketch.
[16] ACD/ChemSketch Version 4.5 for Microsoft Windows User's Guide.
[17] ACD/Labs Extension for ChemBioOffice Version 14.0 for Microsoft Windows User's Guide.
[18] A, N. L. Conformational Analysis 130. MM2. A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms, *J. Am. Chem. Soc*. Vol. 99, pp.8127-8134, (**1977**).
[19] XLSTAT 2015 Add-in software (XLSTAT Company). www.xlstat.com.
[20] Y. Boukarai, F. Khalil, M. Bouachrine, *International Journal of Scientific & Engineering Research, Volume 6, Issue 11*, (**2015**) 159-166.
[21] Y. Boukarai, F. Khalil, M. Bouachrine, *Journal of Chemical and Pharmaceutical Research*, 8(3) 1000-1013, (**2016**).
[22] V.J. Zupan& J. Gasteiger, *Neural Networks for Chemists - An Introduction*, VCH Verlagsgesellschaft, Weinheim/VCH Publishers, New York. 106(12) (**1993**) 1367-1368.
[23] B. Efron, *J. Am.Stat. Assoc*. 78 (**1983**) 316-331.
[24] M.A. Efroymson, Multiple regression analysis, In Mathematical Methods for Digital Computers, Ralston, A., *Wilf, H.S., Eds,WileyNewYork*, (**1960**).
[25] D.W. Osten, *J. Chemom*. 2(**1998**) 39-48.
[26] Y. Boukarai, F. Khalil, M. Bouachrine, *QSAR study of flavonoid derivatives as in vitro inhibitors agents of aldose reductase (ALR2) enzyme for diabetic complications* (**2016**).
[27] Y. Boukarai, F. Khalil, M. Bouachrine, *QSAR study of diarylaniline analogues as in vitro anti-HIV-1 agents in pharmaceutical interest***(2016)**.
[28] S-S. So & W.G. Richards, *J. Med. Chem*. 35 (**1992**) 3201-3207.
[29] T.A. Andrea & H. Kalayeh, *J. Med. Chem*. 34 (**1991**) 2824–2836.
[30] M. Elhallaoui, *Modélisatricemoléculaire et étude QSAR d'antagonistes non compétitifs durécepteur NMDA par les méthodesstatistiques et le réseau de neurones*, (**2002**) 106.
[31] I. Pajeva, M. Wiese, *J. Med. Chem*. (**1998**), 41, 1815-1826.