Quantitative structure activity relationship studies of anti-proliferative activity of some indole derivatives combining DFT calculations and statistical results

Halima Hazhazi[1], Hanane Zaki[2,3*], Mohamed Benlyas[2], Youness Filali -Zegzouti[2], Youcef Boumedjane[1], Mohammed Bouachrine[3*]

[1]Research team computatinal & pharmaceutical chemistry,LMCE laboratory, department of chemistry, Faculty of sciences, Biskra University, Biskra , Algeria.
[2]Research team Biologiy, Environnement & health. Department of biology Faculty of Science and Technics. My Ismail University. Meknes. Morocco
[3]Research team Matériaux. Environnement & Modélisation. ESTM. My Ismail University. Meknes. Morocco

Abstract*— *Many studies have focused on indole derivatives mainly their antiproliferative effect. The therapeutic effect of this group of molecule is very important. Quantitative structure–activity relationships (QSAR) have been applied for development relationships between physicochemical properties and their biological activities.

A series of 30 molecules derived from indole is based on the quantitative structure-activity relationship (QSAR). This study was carried out using the principal component analysis (PCA) method, the multiple linear regression method (MLR), non-linear regression (RNLM), the artificial neural network (ANN) and it was validated using cross validation analysis (CV). We accordingly propose a quantitative model and we try to interpret the activity of the compounds relying on the multivariate statistical analyses. A theoretical study of series was studied using density functional theory (DFT) calculations at B3LYP/6-31G(d) level of theory for employing to calculate electronic descriptors when, the topological descriptors were computed with ACD/ChemSketch and ChemDraw 8.0 programs. The best QSAR model was found in agreement with the experimental by ANN (R = 0,99).

Keywords*— *Breast cancer, anti-proliferative, indole derivatives, QSAR, MLR, MNLR, ANN, CV.

I. INTRODUCTION

Breast cancer is considered as one of the major and widespread reasons that cause death among women all over the world, in case of the late diagnosis [1]. Despite the improvements and the efficiency in early detection, chemotherapy and radiotherapy breast cancer is still at high risks [2-4]. Therefore, it is necessary to find a cure for this disease, a lot of scientific researches were carried out to determine a particular molecule to this treatment [5]. Among the great number of compounds that occur in nature, Indole is the main component. Moreover, indole derivatives have many applications, in the pharmaceutical, industry in the treatment of various diseases [6]. Indole derivatives are one of the most promising heterocyclic, which have active sites in treating various diseases [7]. In addition, these compounds have broad spectrum of biological activities involving anticancer, antioxidant, antimicrobial, anticonvulsant, anti-leishmanial, antidepressants, anti-inflammatory activities and they were found to have capabilities of anti-proliferative activity on cancer cells lines [8-14].

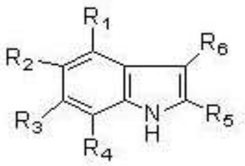


Fig.1: Studied compounds (indole)

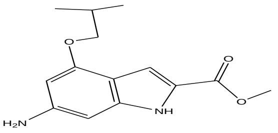
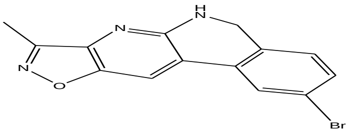
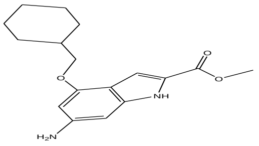
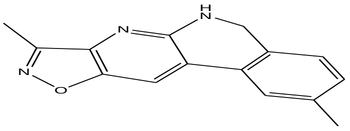
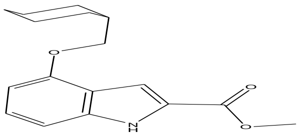
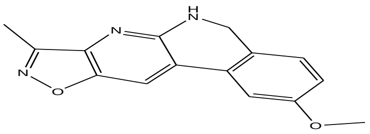
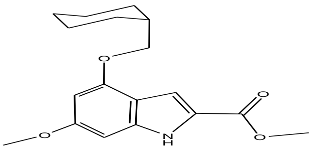
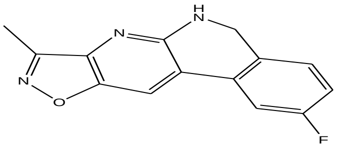
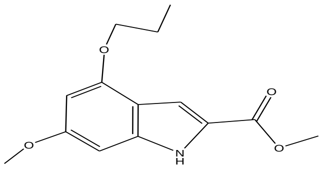
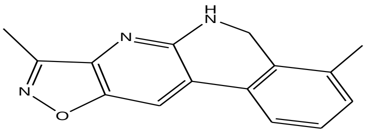
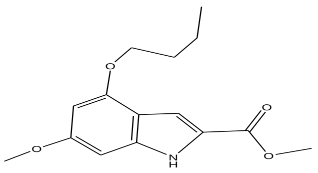
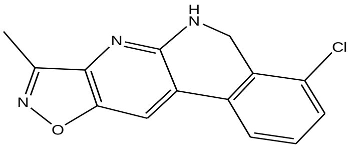
On the other hand, Quantitative structure-activity relationship (QSAR) seeks to inquire into the relationship between molecular descriptors which describe the physicochemical properties correlated with biological activity of the set of compounds [15, 16]. The QSAR study is an important step in the development of new drugs. In this paper we have studied a quantitative structure- activity relationship (QSAR) of indole against human breast cancer cells (MCF-7) based on 30 indole derivatives taken from the literature [17-20]. Therefore, we propose to develop a quantitative model, and we try to predict the activity of these compounds based on the several statistical methods: Principal Component Analysis (PCA), Multiple Non-Linear Regression (MNLR) and Multiple Linear Regression (MLR), Artificial Neural Network (ANN) and Cross Validation analyses (CV). The development of a performant model will help to explain the role of indole derivatives in chemotherapy against breast cancer and also propose other molecules, then predict their anti-cancer activity.

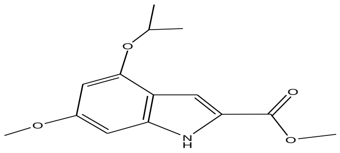
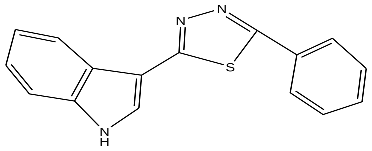
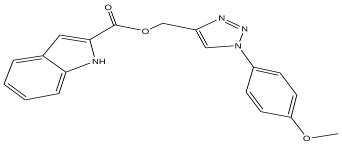
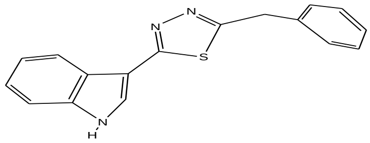
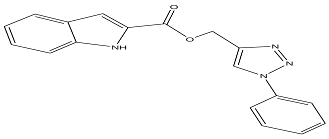
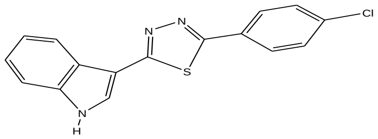
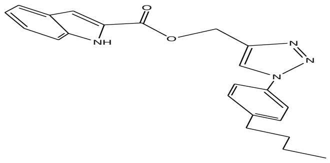
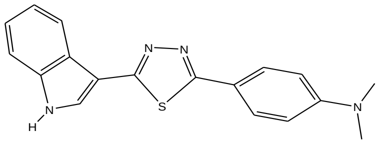
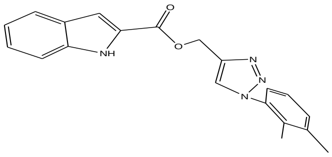
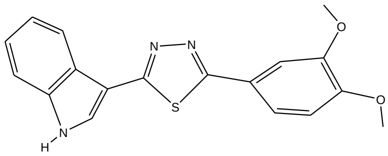
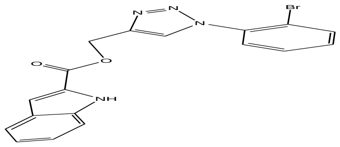
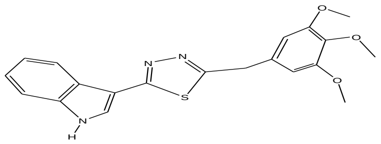
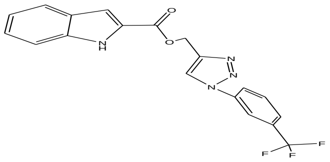
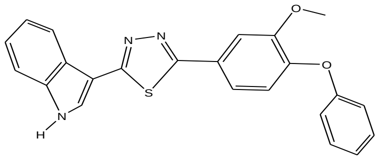
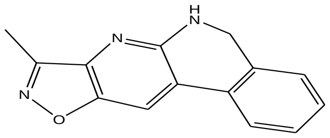
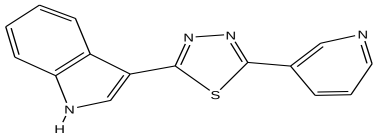
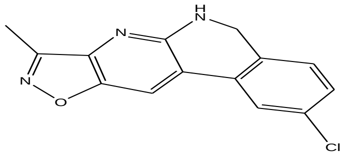
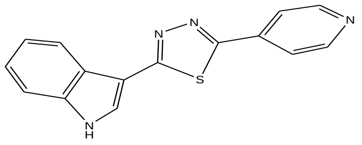
II. MATERIALS AND METHODS

2.1 Chemical data

A dataset of the series of indole Compounds collected from literature [17-20], are listed in table 1. A total of 30 derivatives of indole were studied and analyzed in order to find quantitative structure activity relationship between the anti-proliferative activity and the structure of these molecules. The IC50 values in µM units exhibiting 50% inhibition of cell growth for human breast cancer (MCF 7) were converted in pIC50 by taking logarithm (pIC50 = log10 IC50) for QSAR stady.

Table.1: Observed IC50 of the indole derivatives anti-proliferative agents

S.N	Structure	pIC50	S.N	Structure	pIC50
M1		1.988	M16		1.690
M2		0.577	M17		1.781
M3		1.953	M18		1.274
M4		1.273	M19		1.684
M5		1.908	M20		1.703
M6		1.273	M21		1.674

M7		1.621	M22		1.745
M8		1.105	M23		2.447
M9		1.209	M24		1.130
M10		1.484	M25		2.174
M11		1.459	M26		1.089
M12		1.233	M27		0.832
M13		1.588	M28		2.161
M14		1.218	M29		0.812
M15		1.727	M30		1.961

2.2. Molecular descriptor

The present work is necessary for us to determine several different descriptors to estimate in the QSAR model. The quantum chemical calculations are performed at the B3LYP/6-31G(d) level of theory using GAUSSIAN 03 of programs [21] to calculate some electronic descriptors such as: Frontier molecular orbital's highest occupied molecular orbital: EHOMO (eV); lowest unoccupied molecular orbital energy :ELUMO (eV) ; The Gap energetic (Gap) (eV), (the difference between EHOMO and ELUMO); Total Energy TE (ua); The absolute electronegativity χ (eV), χ = (EHOMO + ELUMO)/2; the absolute hardness η (eV): η = (EHOMO - ELUMO)/2; The Softness S (eV), it is the reactivity index and defined reciprocal of hardness S= 1/η; The electrophilicity index ω (eV), $\omega = 2\chi/2\ \eta$ [22] and The dipole moment μ (Debye). On the other side, we have chosen some physico-chemical descriptors, which were computed with Advanced chemistry development's ACD/ Chem Sketch [23] and ChemDraw Ultra8.0 [24] programs was employed to calculate: Molecular Weight (MW), Torsion energy (TE), Repulsion energy (RE), electronic energy (EE), the octanol/water partition coefficient (log P), Parachor (Pc) and Density (D) Thus 12 descriptors. Data was presented in **Table 2**.

2. 3- Statistical methods

To explain the structure-activity relationship, The 12 quantitative descriptors of the compounds of indole (1 to 30) are studied using different statistical methods:

The principal component analysis (PCA) [25] using the software XLSTAT version 2013 [26]. This is an essentially descriptive statistical method which aims to present, in graphic form. The large information contained in a data, as shown in **table 1**. PCA is a helpful statistical technique for summarizing the maximum of information encoded in the structures of compounds. This method is very useful for understanding the distribution of the compounds. The Multiple Linear Regression (MLR) statistical technique is used to study the relation between one dependent variable and several independent variables. It is a mathematic technique that minimizes the differences between actual and predicted values. The multiple linear regression model (MLR) was performed to predict pIC50. and it served to select the used descriptors as the input parameters for (NLMR). MLR and MNLR were generated using the software XLSTAT version 2014. The obtained equations were justified by the determination coefficient (R^2) correlation coefficient (R) , mean squared error (MSE), Mean Absolute Error (MAE) and Fisher's criterion (F). [27,28].

The ANN analysis was performed with the use of Matlab software version 2009. A Neural Fitting tool (nftool) toolbox on a data set of the indole compounds [29]. Three components constitute a neural network: the topology of the connections between the nodes, the processing elements or nodes and the learning rule by which new information is encoded in the network. However, there are a number of different ANN models; the most frequent type of ANN in QSAR is the three-layered feed-forward network [30]. In this kind of networks, the neurons are arranged in layers (an input layer, one hidden layer and an output layer). the neurons 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 (CV) 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" [31-33]. For each data set, an input-output model is developed. In this study we used, the Leave-One-Out (LOO) procedure.

III. RESULTS AND DISCUSSION

3.1. Data set for analysis

A QSAR study was performed on 30 indole derivatives as reported previously, in order to identify a quantitative relationship between the structure and anti-proliferative activity against breast cancer cells lines (MCF7). The values of the 12 descriptors (2D and 3D descriptors) are shown in **Table 2**.

Table.2: Dataset used for QSAR analysis of series of indole derivatives

molecules	EHOMO	ELUMO	ΔE	µ	χ	TotE	*Log P*	*RE*	*TE*	*Kow*	MW	*D*
M1	-4.907	-0.846	4.061	2.855	2.030	-879.519	1.744	16622.8	7.5280	2.860	262.31	1.214
M2	-4.905	-0.064	4.841	2.796	2.484	-996.254	2.498	20936.1	14.036	4.053	302.37	1.216
M3	-5.395	-1.403	3.992	2.423	3.399	-940.636	3.300	17066.4	16.621	5.171	287.36	1.166
M4	-5.229	-1.297	3.932	1.239	3.263	-1055.11	3.174	20431.2	0.9840	5.137	317.38	1.174
M5	-5.171	-1.158	4.013	3.782	3.164	-899.114	2.022	15064.2	24.456	3.545	263.29	1.192

M6	-5.013	-0.882	4.131	3.402	2.947	-938.687	2.439	16703.9	24.443	4.074	277.32	1.168
M7	-5.032	-0.991	4.041	1.192	3.011	-899.376	1.854	15479.6	23.082	3.325	263.29	1.190
M8	-0.218	-0.160	0.058	1.964	0.189	-1175.25	3.086	21693.7	17.771	4.112	348.36	1.340
M9	-0.220	-0.141	0.079	1.242	0.180	-1138.07	3.213	18040.2	5.7020	3.884	318.33	1.340
M10	-0.210	-0.044	0.166	5.920	0.127	-1221.04	4.952	24448.4	19.596	5.970	374.44	1.240
M11	-0.213	-0.048	0.165	1.881	0.130	-1142.41	4.187	22242.3	19.484	4.832	346.39	1.290
M12	-5.820	-1.331	4.489	2.474	3.575	-3634.88	4.042	20755.0	22.713	4.918	397.23	1.590
M13	-0.217	-0.054	0.163	3.348	0.135	-1400.82	4.134	25558.2	8.8150	5.067	386.33	1.440
M14	-5.564	-1.687	3.877	4.393	3.625	-779.571	2.636	13236.6	11.444	3.202	237.26	1.305
M15	-5.748	-1.905	3.843	4.656	3.826	-1239.16	3.195	14469.8	11.896	3.918	271.70	1.402
M16	-5.741	-1.900	3.840	4.592	3.820	-3350.67	3.465	14425.7	11.537	4.068	316.15	1.597
M17	-5.509	-1.646	3.863	4.521	3.577	-818.890	3.123	14613.0	11.395	3.701	521.28	1.269
M18	-5.485	-1.633	3.852	5.655	3.559	-894.094	2.510	16064.6	10.890	3.213	267.28	1.299
M19	-5.709	-2.029	3.680	5.824	3.869	-878.558	2.794	14600.0	11.608	3.348	255.25	1.372
M20	-5.508	-1.628	3.880	4.839	3.568	-818.888	3.123	14634.9	14.399	3.701	251.28	1.269
M21	-5.696	-2.044	3.652	3.172	3.870	-1238.92	3.195	14502.5	11.543	3.918	271.70	1.402
M22	-5.489	-1.612	3.877	4.567	3.550	-1178.76	4.175	14005.1	28.720	3.652	277.35	1.331
M23	-5.509	-1.181	4.328	4.793	3.345	-1218.06	4.554	14998.9	11.231	3.621	291.37	1.309
M24	-5.593	-1.808	3.785	5.518	3.7005	-1638.35	4.734	15919.8	13.132	4.368	311.79	1.415
M25	-5.463	-1.579	3.884	3.715	3.521	-1312.71	4.46	18832.6	12.738	3.984	320.42	1.300
M26	-5.506	-1.659	3.847	3.898	3.5825	-1407.78	3.923	19998.1	10.607	3.433	337.4	1.316
M27	-5.429	-1.57	3.859	4.217	3.4995	-1522.31	4.175	25455.7	53.488	2.921	381.45	1.295
M28	-5.468	-1.633	3.835	3.813	3.5505	-1599.54	5.586	26277.4	11.423	5.458	399.47	1.318
M29	-5.627	-1.849	3.778	6.858	3.738	-1194.79	2.838	14662.8	26.266	2.255	278.33	1.380
M30	-5.718	-2.101	3.617	6.78	3.9095	-1194.72	2.838	14721.1	28.986	2.255	278.33	1.380

3.2. Data Modeling.

3.2.1 Principal component analysis

The 12 descriptors (variables) describing the 30 molecules were submitted to Principal Components Analysis (PCA). The first two principal axes are sufficient to describe the information provided by the data matrix. Figure.2 presents the percentages of variance: F1= 40, 50 %. F2= 21,23% and the total information is estimated on 61,74 %.

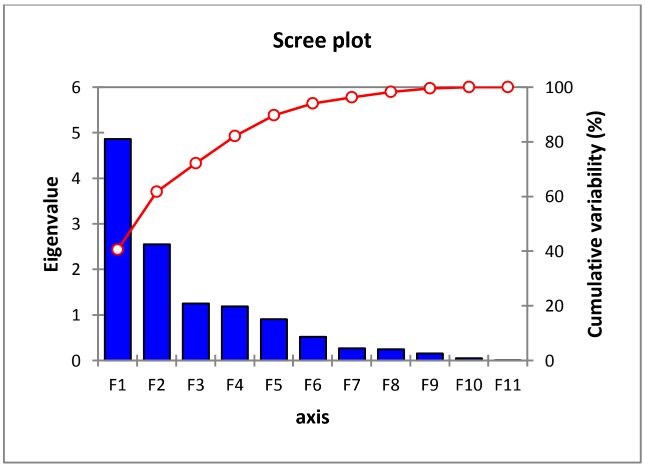


Fig. 2. The principal components and their variances

The principal component analysis (PCA) was carried out to have an idea about the link between the different variables.

The obtained matrix (Table3) summarizes the correlations between the 12 descriptors and provides information on the negative or positive correlation between variables. Figure 3 shows these descriptors in a correlation circle. In general the correlation matrix shows a low interrelationship between most of the descriptors, Good co-linearity (r>0.5) was observed between some of the variables. Hight interrelationship was observed between E_{HOMO} and χ (r = -0.979), E_{HOMO} and ΔE (R= -0 .96), E_{LUMO} and χ (R=-0.91) and ΔE and χ (R= 0.90), the variables ΔE and χ are removed to decrease the correlations.

Table.3: The correlation matrix (pearson (n)) between different obtained descriptors

Variables	**EHOMO**	**ELUMO**	**ΔE**	**µ**	**χ**	**TotE**	**Log P**	**RE**	**TE**	**Kow**	**MW**	**D**
EHOMO	**1**											
ELUMO	**0.845**	**1**										
ΔE	**-0.969**	**-0.688**	**1**									
µ	-0.347	**-0.536**	0.224	**1**								
χ	**-0.979**	**-0.918**	**0.907**	**0.419**	**1**							
TotE	0.106	0.142	-0.078	0.020	-0.132	**1**						
Log P	0.194	0.018	-0.255	0.149	-0.114	**-0.378**	**1**					
RE	**0.538**	**0.559**	**-0.474**	-0.317	**-0.549**	-0.186	**0.541**	**1**				
TE	-0.106	-0.093	0.101	0.219	0.125	-0.090	0.042	0.142	**1**			
Kow	**0.428**	**0.436**	**-0.381**	**-0.380**	**-0.413**	-0.257	**0.564**	**0.583**	-0.312	**1**		
MW	0.288	0.225	-0.288	-0.119	-0.259	-0.312	**0.509**	**0.576**	0.049	**0.428**	**1**	
D	-0.040	-0.293	-0.080	0.267	0.128	**-0.790**	0.326	-0.042	-0.028	-0.009	0.178	**1**

The correlation circle (Figure 3**)** which shows that the F1 axis (40.50 % of the variance) appears to represent the Density (D) and the Total energy (TE). The F2 axis (21.24% of the variance) seems to represent the EHOMO and gap Energy (ΔE).

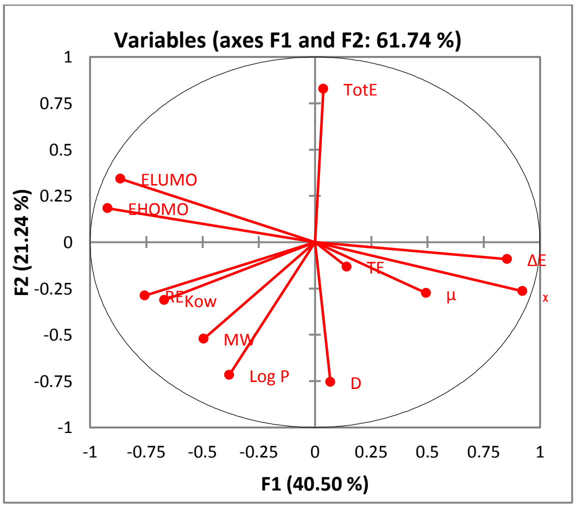


Fig. 3: Correlation circle between descriptors

From other side the analysis of diagrams according to the planes F1 and F2 (of the total variance) of the studied series are presented in Figure 4 we can discern three groups of molecules:

- Group 1: contains the molecules: M8, M9, M11, M10, and M13. (Green color)
- Group 2: contains the molecules M27, M24, M16, M28 and M12. (Red color)
- Group 3: contains the rest of the molecules. (Blue color)

When we return to the structures of molecules M8, M9, M10 and M11 (group 1) , we note that all these molecules are alike in their structures, and have as basic structure compound 1- Aryl- 1H- 1,2,3- Triazol-4-yl methyl 1H indole-2-carboxylate. The molecules M24, M27 and M28 (group 2) have the same basic derivative which is 5- (3-indolyl) -2-Substituted-1,3,4-thiadiazoles. Group 3 is the most important of the groups because it contains a large number of molecules (20 molecules) which have the same behavior.

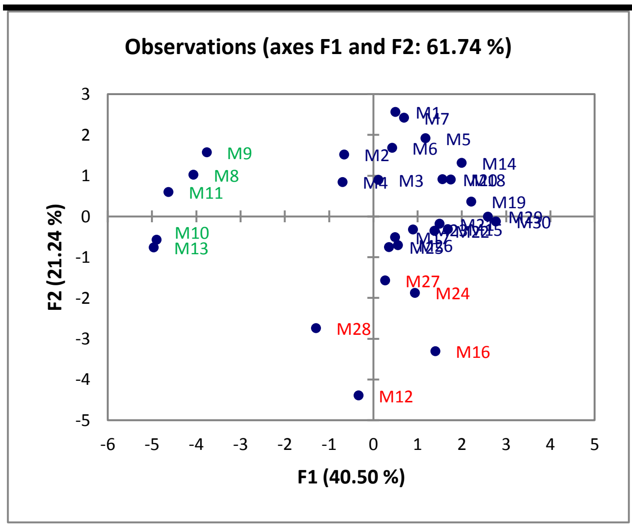


Fig.4: Correlation plot between the different molecules

3.2.2 Multiple linear regressions MLR

Our work is based on the development of the best QSAR model to clarify the correlation between the different descriptors and the biological activities pI50 values of the indole derivatives. This method utilised several coefficients: R is the correlation coefficient, R^2 is the coefficient of determination, MSE mean squared error , MAE Mean Absolute Error and F is the Fisher F-statistic those coefficients adopt the best regression performance.

The obtained relationship in this model by the linear combination of the essential descriptors: E_{HOMO}, E_{LUMO}, µ, TE, Log P, RE, TE, Kow, MW, D.

The QSAR models using multiple linear regressions method is represented by the following equations:

$$\mathbf{pIC50} = -2.91 - 0.45 * \mathbf{E_{HOMO}} + 1.15 * \mathbf{E_{LUMO}} - 6.44\,E\text{-}02 * \boldsymbol{\mu} + 8.74\,E\text{-}04 * \mathbf{TE} + 0.20 * \mathbf{LogP} - 4.12\,E\text{-}06 * \mathbf{RE} - 9.31\,E\text{-}03 * \mathbf{TE} + 0.22 * \mathbf{Kow} + 3.03\,E\text{-}04 * \mathbf{MW} + 2.84 * \mathbf{D}.$$

N= 24 N_{test}= 6 R= 0.80 R^2= 0.641
MSE = 0.48 MAE = 0.44 F = 2.326

The model shows a good correlation coefficient (R =0.800) between ten descriptors and the anti-proliferative activity. This equation shows that the anti-proliferative activity of the indole derivatives depends on the electronic and the topological side of the molecule. Anti-proliferative activity increases by increasing the topological properties, Log P, RE, Kow, MW, D and by diminishing the electronic properties E_{HOMO}, µ, TE, RE, TE. Figure 5 presents the graphical representations of graphical calculated and observed pIC50 by MLR.

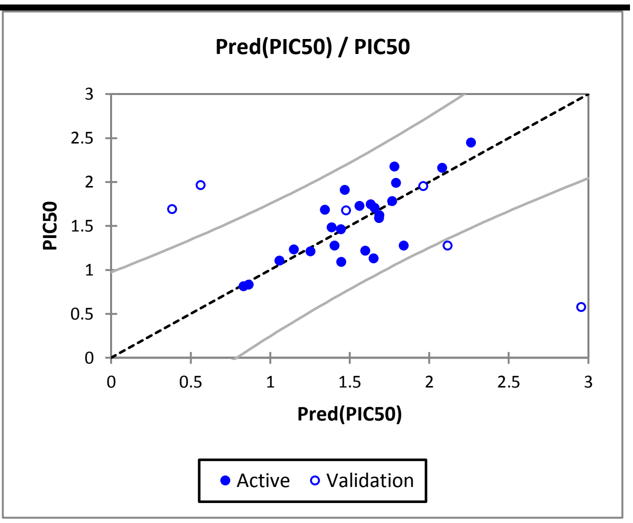


Fig. 5: Graphical representation of calculated and observed pIC50 by MLR

As illustrated in Figure 5, the correlation between calculated and experimental activities is very remarkable.

3.2.3 Multiple nonlinear regressions MNLR

We have utilized the technique of nonlinear regression model to improve the structure activity relationship in a quantitative way, the selected descriptors from the MLR model are used like data base matrix for the MNLR. The resulting equation is:

$$\mathbf{PIC50} = 120.18 + 2.44 * \mathbf{E_{HOMO}} + 1.01 * \mathbf{E_{LUMO}} + 0.75 * \boldsymbol{\mu} + 6.63\,E\text{-}03 * \mathbf{TE} - 4.56 * \mathbf{LogP} - 5.06\,E\text{-}05 * \mathbf{RE} - 4.73\,E\text{-}02 * \mathbf{TE} - 1.44 * \mathbf{Kow} + 0.16 * \mathbf{MW} - 203.25 * \mathbf{D} + 0.44 * \mathbf{E_{HOMO}}^2 + 0.43 * \mathbf{E_{LUMO}}^2 - 0.11 * \boldsymbol{\mu}^2 + 6.27\,E07 * \mathbf{TE}^2 + 0.74 * \mathbf{LogP}^2 - 5.9\,E\text{-}09 * \mathbf{RE}^2 + 1.13\,E\text{-}03 * \mathbf{TE}^2 + 0.12 * \mathbf{Kow}^2 - 2.15\,E\text{-}04 * \mathbf{MW}^2 + 79.02 * \mathbf{D}^2$$

N= 24 N_{test} = 6 R= 0.95 R^2= 0.90 MSE = 0.13
MAE= 0.24

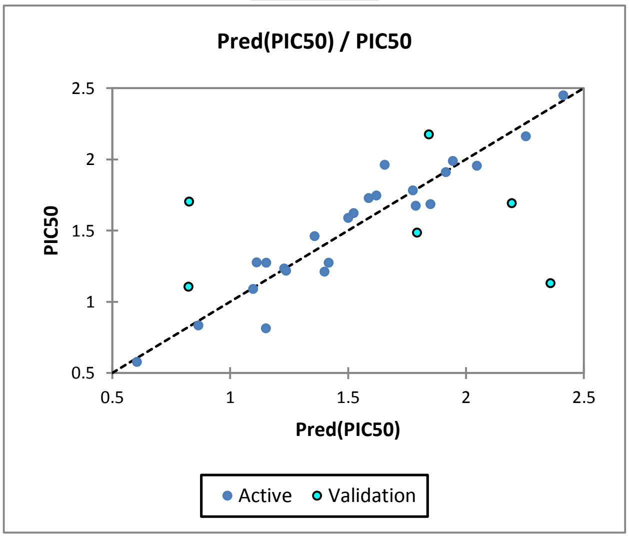


Fig. 6: Graphical representation of calculated and observed pIC50 by MNLR

The obtained correlation coefficient was significant R = 0,95. **Figure 6** shows a regular distribution of the PIC50 observed values depend on the experimental values.

3.2.3 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 the observed activity. The calculated activities model was developed using the properties of several studied compounds. Some authors **[34, 35]** have proposed a parameter ρ, leading to determine the number of hidden neurons, which play a major role in determining the best ANN architecture. These are defined as follows:

ρ = (Number of data points in the training set /Sum of the number of connections in the ANN)

The values of predicted activities (pIC50) using ANN and the observed values are given in **Table 4**. The correlation between calculated ANN and experimental anti-proliferative values is very significant as indicated by R and R^2 values illustrated in **figure 8.**

Table.4: The observed and ANN predicted activities

Molecules	pIC50	Pred (pIC50)
M1	1,988	2.0078
M2	0,577	0.5947
M3	1,953	2.0063
M4	1,273	1.2653
M5	1,908	1.7715
M6	1,273	1.4139
M7	1,621	1.5753
M8	1,105	1.0249
M9	1,209	1.2162
M10	1,484	1.4810
M11	1,459	1.5133
M12	1,233	1.2304
M13	1,588	1.5925
M14	1,218	1.3461
M15	1,727	1.7242
M16	1,690	1.7099
M17	1,781	1.7874
M18	1,274	1.2230
M19	1,684	1.7271
M20	1,703	1.5881
M21	1,674	1.6810
M22	1,745	1.7352
M23	2,447	2.4466
M24	1,130	1.0820
M25	2,174	2.1477
M26	1,089	1.1153
M27	0,832	0.8418
M28	2,161	2.1485
M29	0,812	0.8772
M30	1,961	1.9345

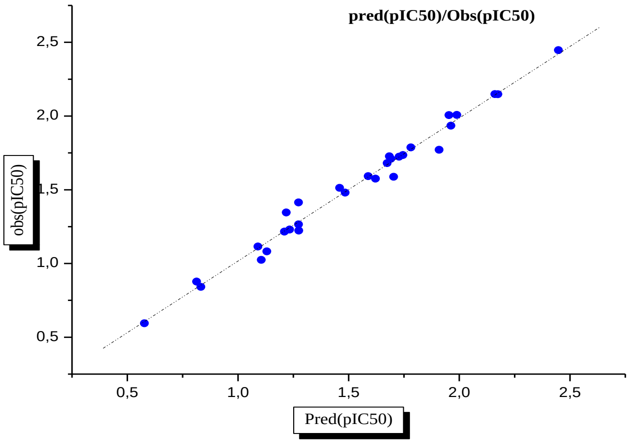


Fig. 8: Correlations of observed and predicted activities $TC_{50\ (2)}$ calculated using ANN

N=30 R= 0.99 R^2 = 0.98 MSE=0.003

MAE= 0.03

The obtained correlation coefficient R value confirms that the artificial neural network result was the best to build the quantitative structure activity relationship models.

A comparison of the quality of MLR, MNLR and ANN models **table 5** 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.

Table.5: observed, predicted activities according to different used methods

Molecules	Obs (pIC50)	(pIC50) RLM	(pIC50) RNLM	(pIC50) ANN	(pIC50) CV
M1	1,989	1,791	1,944	2,0078	1,78
M2	1,909	1,470	0,605	0,5947	0,61

M3	1,273	1,840	2,046	2,0063	1,39	
M4	1,621	1,688	1,154	1,2653	1,2	
M5	1,106	1,058	1,915	1,7715	1,79	
M6	1,210	1,254	1,418	1,4139	1,34	
M7	1,484	1,386	1,524	1,5753	1,63	
M8	1,459	1,446	1,400	1,0249	1,21	
M9	1,233	1,149	1,358	1,2162	1,37	
M10	1,588	1,686	1,230	1,481	1,43	
M11	1,218	1,598	1,500	1,5133	1,46	
M12	1,728	1,563	1,237	1,2304	1,22	
M13	1,781	1,768	1,588	1,5925	1,57	
M14	1,275	1,405	1,774	1,3461	1,32	
M15	1,684	1,344	1,113	1,7242	1,68	
M16	1,703	1,658	1,850	1,7099	1,72	
M17	1,745	1,632	1,787	1,7874	1,61	
M18	2,448	2,263	1,620	1,223	1,39	
M19	1,130	1,651	2,413	1,7271	1,67	
M20	2,175	1,782	1,097	1,5881	1,63	
M21	1,090	1,447	0,865	1,681	1,58	
M22	0,833	0,865	2,254	1,7352	1,64	
M23	2,161	2,081	1,151	2,4466	1,99	
M24	0,813	0,833	1,656	1,082	1,23	
M25	0,577	2,957	0,823	2,1477	1,95	
M26	1,953	1,964	1,793	1,1153	1,42	
M27	1,273	2,116	2,195	0,8418	1,71	
M28	1,691	0,383	0,827	2,1485		1,93
M29	1,674	1,477	2,358	0,8772	1,71	
M30	1,962	0,562	1,843	1,9345	1,82	

3.2.4 Cross Validation

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 which is well adapted to the estimation of that ability. A good correlation was obtained with cross validation RCV = 0,74. So, the predictive power of this model is very significant. The results obtained showed that models MLR, MNLR and ANN are validated, which means that the prediction of the new compounds is feasible

In this study, three different modelling methods, MLR, MNLR and ANN were used in the construction of a QSAR model for 30 derivatives of indole and the resulting models were compared (**table 5 - table 6)**. It was shown that the artificial neural network ANN results have better predictive capability than the MLR and MNLR. we established a relationship between several descriptors and the anti-proliferative activity p**IC$_{50}$** in satisfactory manners. The good results obtained with the cross validation (CV) shows that the model proposed in this paper are able to predict activity with a good performance, and that the selected descriptors are pertinent.

Table.6: Statistical values obtained by different methods

	RLM	RNLM	ANN	CV
R	0.80	0.95	0.99	0.74
MSE	0.48	0.13	0.003	0.08
MAE	0.44	0.24	0.03	0.18

Correlation coefficient (R), Mean squared error (MSE), Mean Absolute Error (MAE)

IV. CONCLUSION

In this paper we have used different statistical methods: MLR, MNLR, ANN, cross validation CV and various electronic and topologic descriptors for construction of QSAR model for the anti-proliferative activity of indole derivatives, also, were compared the statistical terms R, R2 , MAE, MSE Resulting models. Moreover, the neural network ANN results (R= 99, MAE= 0.03 MSE= 0.003) have better predictive capability than the MLR and MNLR. A good correlation was obtained with cross validation RCV = 0,74 that confirms the great ability of our model to predict the activity. we established a relationship between several descriptors and inhibition values pIC50 of several organic compounds based on substituted indole in satisfactory manners. That studied model which is sufficiently rich in chemical, electronic and topological information may be utilized for predicting and developing new molecules with better effect. Thus, thanks to QSAR studies, especially with the ANN that 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 that could be active in experiment.

REFERENCES

[1] Siegel R, Naishadham D and Jemal A: Cancer statistics, 2013. CA cancer J clin 63: 11-30, 2013.

[2] Aliabadi HM, Maranchuk R, Kucharaki C, Mahdipoor P, Hugh J, Uludag H. Effective response of doxorubicin- sensitive and resistant breast cancer cells to combinational siRNA therapy. J Contr Release Official: J Contr Release Soc 2013; 172(1): 219-228.

[3] Hong SE, Kim EK, Jin HO, Kim HA, Lee JK, Koh JS, et al S6K1 inhibition enhances tamoxifen-induced cell death in MCF-7 cells through translational inhibition o Mcl-1 and survinn. Cell Biol Toxicol 2013: 29(4): 273-282.

[4] Choi HN, Jin HO, Kim JH, Hong SE, Kim HA, Kim EK, et al. Inhibition of S6K1 enhances glucose deprivation-induced cell death via downregulation of anti-apoptotic proteins in MCF-7 breast cancer

[5] celles Biochem Biophys Res Comm 2013; 432(1): 123-128

[6] Hao Wang, Yi-Feng Ye, Effect of survivin siRNA on biological behaviour of breast cancer MCF7 cells. J Tropical Medicine (2015)225-228.

[7] T.M.T. Carvalho, L.M.P. Amaral, V.M.F. Morais, M.D.M. Ribeiro da Silva, Thermodynamic properties of alkyl 1*H*-indole carboxylate derivatives: a combined experimental and computational study, *J. Chem. Thermodynamics* (2016).

[8] Sirassu Narsimha, Nukala Satheesh Kumar, Kumara Swamy Battula, Vasudeva Reddy Nagavelli ,S. K. Althaf Hussain , M. Srinivasa Rao. Indole-2-carboxylic acid derived mono and bis 1,4 disubstituted 1,2,3-triazoles: Synthesis, characterization and evaluation of anticancer, antibacterial, and DNA-cleavage activities. Bioorg. Med.Chem. Lett 26 (2016) 1639–1644

[9] Daly, S; Hayden, K.; Malik, I.; Porch, N.; Tang, H.; Rogelj, S.; Frolova, L. V.; Lepthien, K.; Kornienko, A.; Magedov, I.V. Bioorg Med. Chem. Lett. 2011, 21, 4720

[10] Holla, B. S.; Mahalinga, M.; Karthikeyan, M. S.; Poojary, B.; Akberali, P. M.; Kumari, N. S. Eur. J. Med. Chem. 2005, 40, 1173.

[11] . Mady, M. F.; Awad, G. E. A.; Jørgensen, K. B. Eur. J. Med. Chem. 2014, 84, 433.

[12] Rao, P. S.; Kurumurthy, C.; Veeraswamy, B.; Kumar, G. S.; Poornachandra, Y.; Kumar, C. G.; Babu, V. S.; Kotamraju, S.; Narsaiah, B. Eur. J. Med. Chem. 2014, 80, 184.

[13] H.N. Dogan, A. Duran, S. Rollas, G. Sener, M.K. Uysal, D. Gülen, Bioorg. Med. Chem. 10 (2002) 2893e2898.

[14] F. Poorrajab, S.K. Ardestani, S. Emami, M. Behrouzi-Fardmoghadam, A. Shafiee, A. Foroumadi, Eur. J. Med. Chem. 44 (2009) 1758e1762.

[15] M. Yusuf, R.A. Khan, B. Ahmed, Bioorg. Med. Chem. 17 (2008) 8029e8034.

[16] Hansch C., Muir R.M., Fujita T., Maloney P.P, Geiger F., Streich M., J. Am. Chem. Soc. 85 (1963) 2817–2825.

[17] Bodor N., Current Medicinal Chemistry 5 (1988) 353-380. From book: Biochemistry of Redox Reactions, by Bernard Testa, editor: London [u, a], Acad. Press (1995).

[18] Xing-yue Ji , Si-tu Xue, Yue-chen Zhan, Jia-jia Shen, Lin-tao Wu, Jie Jin, Zhen Wang, Zhuo-rong Li , J. Med. Chem. 83 (2014) 409-418.

[19] Sirassu Narsimha, Nukala Satheesh Kumar, Kumara Swamy Battula , Vasudeva Reddy Nagavelli, S. K. Althaf Hussain, M. Srinivasa Rao. Bioorg. Med. Chem , 26 (2016) 1639–1644.

[20] Dalip Kumar , N. Maruthi Kumar , Kuei-Hua Chang , Kavita Shah, Eur. J. Med. Chem. 45 (2010) 4664-4668.

[21] E. Rajanarendar, K. Govardhan Reddy, S. Ramakrishna, M. Nagi Reddy, B. Shireesha, G. Durgaiah, Y. N. Reddy. Bioorg . Med. Chem . 22 (2012) 6677–6680.

[22] M.J. Frisch & al, Gaussian 03, Revision B.01, Gaussian, Inc., Pittsburgh, PA, (2003).

[23] U. Sakar, R. Parthasarathi, V. Subramanian & P.K. Chattaraji, Toxicity analysis of polychlorinated dibenzofurans through global, J. Mol. Des. IECMD, (2004) 1-24.

[24] Advanced Chemistry Development Inc., Toronto, Canada, (2009), (www.acdlabs.com/resources/freeware/chemsketch/).

[25] ACD/Labs Extension for ChemDraw Version 8.0 for Microsoft Windows User's Guide.

[26] M Larif, A Adad , R Hmamouchi, A I Taghki, A Soulaymani , A Elmidaoui, M Bouachrine, T Lakhlifi article in press in *Arabian Journal of Chemistry* **2013**.

[27] SYSTAT 13 Add-in software (SYSTAT Company). www.systat.com.

[28] Y. Boukarai, F. Khalil, M. Bouachrine, QSAR study of isatin (1H-indole-2-3-dione) analogues as in vitro anticancer agents using the statistical analysis methods, *International Journal of Scientific & Engineering Research* (**2015**)

[29] Youness Boukarai, Fouad Khalil and Mohamed Bouachrine, QSAR study of 5,6-bicyclic heterocycles analogues as anti-Alzheimer's agents using the statistical analysis methods. J. Chem. Pharm. Res., 2016, 8(3):1000-1013

[30] Adad A., Hmamouchi R., Taghki A. I., Abdellaoui A., Bouachrine M., Lakhlifi T., Atmospheric half-lives of persistent organic pollutants (POPs) study combining DFT and QSPR results, J. of Chemical and Pharmaceutical Research, Vol. 5, Issue 7 (2013) 28-41 .

[31] V.J. Zupan, J. Gasteiger, Neural Networks for Chemists - An Introduction, VCH Verlagsgesellschaft, Weinheim/VCH Publishers, New York. 106(12) (**1993**) 1367-1368.

[32] B. Efron, Estimating the error rates of a predictive rule: improvement on crossvalidation, J. Am.Stat. Assoc. 78 (1983) 316-331

[33] M.A. Efroymson, Multiple regression analysis, In Mathematical Methods for DigitalComputers, Ralston, A., Wilf, H.S., Eds,WileyNewYork, (1960).

[34] D.W. Osten, Selection of optimal regression models via cross-validation, J. Chemom. 2(1998) 39-48.

[35] S-S. So & W.G. Richards, Application of neural networks: quantitative structure-activity relationships of the derivatives of 2, 4 diamino (substituted-benzyl) pyrimidines as DHFR inhibitors, J. Med. Chem. 35 (1992) 3201-3207.

[36] T.A. Andrea & H. Kalayeh, Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors, J. Med. Chem. 34 (1991) 2824–2836.