Chemical Methodologies 7(2023) 719-735

Chemical Methodologies

Journal homepage: http://chemmethod.com

Original Research article

Study of Quantitative Structure-Activity Relationship (QSAR) of Diarylaniline Analogues as *in Vitro* Anti-HIV-1 Agents in Pharmaceutical Interest

Mohammed Bouachrine[1,*], Youness Boukarai[1], Fouad Khalil[2]

[1]*Equipe Matériaux, Environnement & Modélisation, ESTM, University Moulay Ismail, Meknes, Morocco*
[2]*LAC, Laboratory of Applied Chemistry, Faculty ofScience and Technology, University Sidi Mohammed Ben Abdellah, Fez, Morocco*

ARTICLE INFO

Article history

Submitted: 2023-06-18
Revised: 2023-08-10
Accepted: 2023-08-30
Manuscript ID: CHEMM-1710-1016
Checked for Plagiarism: Yes
Language Editor:
Ermia Aghaie
Editor who approved publication:
Dr. Sami Sajjadifar

DOI: 10.48309/chemm.2023.53807

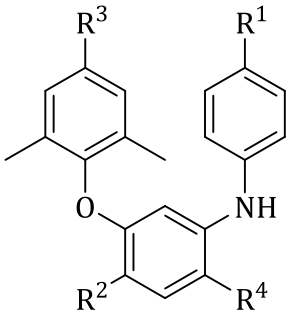
KEYWORDS

HIV-1 virus
Reverse transcriptase (RT)
Diarylaniline derivatives
QSAR
PCA

ABSTRACT

A study of quantitative structure-activity relationship (QSAR) is applied to a set of 24 molecules derived from diarylaniline to predict the anti-HIV-1 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 proposed a quantitative model (non-linear and linear QSAR models), and we interpreted the activity of the compounds relying on the multivariate statistical analysis. The topological descriptors were computed with ACD/ChemSketch and ChemBioOffice14.0 programs. A correlation was found between the experimental activity and those obtained by MLR and MNLR such as (R_{train} = 0.886; R^2_{train} = 0.786) and (R_{train} = 0.925; R^2_{train} = 0.857) for the training set compounds, and ($R_{MLR\text{-}test}$ = 0.6) and ($R_{MNLR\text{-}test}$ = 0.7) for a randomly chosen test set of compounds, this result could be improved with ANN such as (R = 0.916 and R^2 = 0.84) with an architecture ANN (6-1-1). To evaluate 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) including (R = 0.903 and R^2 = 0.815) with the procedure leave-one-out (LOO). The results showed that the MLR and MNLR have served to predict activities, but when compared with the results given by a 6-1-1 ANN model. We realized that the predictions fulfilled by the latter model were more effective than the other models. The statistical results indicated that this model is statistically significant and showing a very good stability towards the data variation in leave-one-out (LOO) cross validation.

GRAPHICAL ABSTRACT



* Corresponding author: Mohammed Bouachrine
E-mail: bouachrine@gmail.com

Introduction

Acquired immunodeficiency syndrome (AIDS) is the major cause of death worldwide. Human immunodeficiency virus (HIV) is a retrovirus that can lead to AIDS, which is a condition in humans where the immune system begins to fail, leading to life-threatening opportunistic infections resulting in progressive destruction of $CD4^{+}$T lymphocytes and inexorable collapse of immune function. The treatment with antiretrovirals may increase the life expectancy of the infected individuals [1, 2]. HIV-1 is a retrovirus that encodes a reverse transcriptase (RT) required for viral replication [3]. (RT) is a key enzyme that plays an essential and multifunctional role in the replication of HIV-1 and considered to be an attractive target for inhibition of HIV replication [4]. HIV-1 RT catalyses the conversion of single-stranded RNA into double-stranded DNA that is integrated into the host cell's genome [5]. Blocking each of the (RT) activities can protect target cells from infection by HIV-1 [6].Non-nucleoside reverse transcriptase inhibitors (NNRTIs) are key components of most current combination therapies used to fight HIV-1 infections [7, 8]. NNRTIs consist of a variety of hydrophobic compounds that are potent non-competitive inhibitors of the DNA polymerase activity of HIV-1 RT. Most NNRTIs are highly specific against HIV-1 RT, and, therefore, they are typically non-toxic to human cells [6]. NNRTIs are highly lipophilic compounds that are able to enter intact HIV-1 virions [9].

NNRTIs are a family of selective chemically diverse RT targeting agents [10]. In anti-AIDS agent research, NNRTIs have gained a definitive and important position due to their unique mechanism of action, antiviral potency, high specificity, low toxicity and favorable pharmacokinetic properties [11, 12]. To date, more than 30 different classes of NNRTIs have been reported. Among these classes, three compounds including nevirapine, delavirdine, and efavirenz have been approved by the food and drug administration (FDA) to treat HIV-1 infected adults in combination with nucleoside and protease analogues [4, 11]. The development of new and more potent mutation-resistant NNRTIs remains a challenging task for the treatment of HIV-1 infected patients due to the drug compliance, adverse effects, and cross-resistance [13]. Diarylaniline analogues have attracted considerable attention due to their excellent activity and led to the identification of highly potent compounds against both wild type and HIV-1 RT resistant viral strains [14, 15].

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 [16, 17].

In this work, we attempt to establish a quantitative structure-activity relationship between anti-HIV-1 activities of a series of 24 bioactive molecules derived from diaryl aniline and structural descriptors.

Thus, we can predict the anti-HIV-1 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 component analysis (PCA) has served to classify the compounds according to their activities and the variability of the descriptors. It allows reducing the number of variables not significant and making the information less redundant. The multiple linear regression (MLR) was used 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. 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).

Materials and Methods

The biological data used in this study were inhibitory activity against HIV-1 virus (inhibition

of reverse transcriptase (RT) enzyme. (EC_{50})), a set of twenty-four derivatives of diarylaniline. We have studied and analyzed the series of diarylaniline molecule consists of 24 selected derivatives that have been synthesized and evaluated for the irinhibitory activity in vitro against HIV-1 (in terms of -log (EC_{50})) [14, 15].

18 molecules were selected to propose the quantitative model (training set), and 6 compounds were selected randomly to test the performance of the proposed model (test set). The quantitative structure-activity relationship between the inhibitory activity (RT) and the molecules structure was determined. The chemical structure of diarylaniline is represented in Figure 1.

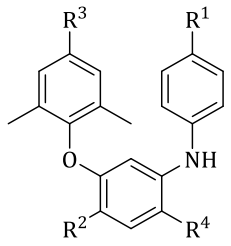


Figure 1: The general structure of diarylaniline

The chemical structures of 24 compounds of diarylaniline used in this study and their experimental HIV-1 inhibitory biological activity observed EC_{50} (Cytotoxic concentration required to inhibit the reverse transcriptase enzyme (RT) than 50%, this is a concentration of diarylanilines where 50% of the maximum effect is observed) were collected from recent publications [14, 15]. The observations converted into logarithmic scale -log (EC_{50}) in molar units (M) are presented in Table 1.

Table 1: Chemical structures and observed activities of diarylaniline derivatives (Training set and test set) against HIV-1 virus

N°	Position of substituents				pEC_{50}[a]Obs
	R_1	R_2	R_3	R_4	
1	OCH_3	NO_2	H	NO_2	5,42
2[b]	OCH_3	NO_2	CH_3	NO_2	5,52
3	OCH_3	NO_2	Br	NO_2	5,44
4	CH_3	NO_2	Br	NO_2	5,37
5[c]	CN	NO_2	Br	NO_2	6,76
6	CN	NO_2	H	NO_2	6,26
7[b]	CN	NO_2	CN	NO_2	5,38
8	CN	NO_2	CH_3	NO_2	6,55
9[c]	CN	NO_2	CHO	NO_2	5,82
10[c]	CN	H	Br	NO_2	6,50
11[b]	CN	H	H	NO_2	5,50
12	CN	H	CN	NO_2	6,68
13	CN	H	CH_3	NO_2	7,17
14	CN	H	CHO	NO_2	5,66
15[c]	CN	H	Br	NH_2	7,33
16	CN	H	CN	NH_2	7,16
17[b]	CN	H	CH_3	NH_2	7,14
18	CN	NH_2	Br	NH_2	6,79
19[b,c]	CN	NH_2	H	NH_2	5,49
20[b]	CN	NH_2	CN	NH_2	7,52
21	CN	NH_2	CH_3	NH_2	7,16
22	CN	NO_2	Br	NH_2	7,80
23[c]	CN	NO_2	CN	NH_2	8,52
24	CN	NO_2	CH_3	NH_2	7,21

[a] Pec_{50}=log(EC_{50}). [b]Test set molecules(MLR). [c] Test set molecules(MNLR)

Calculation of molecular descriptors

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)) [18, 19].

Steric, thermodynamic descriptors are calculated using ACD/ChemSketchand ChemBioOffice14.0 [20] 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}) [21].

In this study, 11 descriptors were chosen to describe the structure of the molecules constituting the series to evaluate: 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, the 11 descriptors were calculated for 24 molecules (Table 2) through software ChemSketch and ChemBioOffice14.0.

Table 2: Values of the 11 chemical descriptors

N°	MW	MR	MV	Pc	n	γ	D	α_e	LogP	HBA	HBD
1	409,392	111,49	306,2	834,6	1,648	55,1	1,33	44,19	6,417	5	1
2	423,418	116,31	322,5	872,3	1,640	53,4	1,31	46,11	6,832	5	1
3	488,288	119,18	322,4	885,1	1,660	56,7	1,51	47,24	7,305	5	1
4	472,288	117,32	314,7	866,1	1,668	57,3	1,50	46,51	7,589	4	1
5	483,271	115,15	299,6	887,4	1,694	76,9	1,61	45,65	6,680	5	1
6	404,375	107,43	286,9	836,4	1,671	72,1	1,40	42,59	5,792	5	1
7	429,385	112,00	296,1	884,0	1,680	79,3	1,44	44,40	5,298	6	1
8	418,402	112,06	302,6	874,6	1,662	69,7	1,38	44,42	6,207	5	1
9	432,385	112,40	299,5	883,5	1,673	75,6	1,44	44,56	5,269	6	1
10	438,274	109,12	288,6	830,4	1,680	68,5	1,51	43,26	6,720	4	1
11	359,377	101,40	275,9	779,3	1,656	63,6	1,30	40,20	5,832	4	1
12	384,387	105,97	285,1	826,9	1,665	70,7	1,34	42,01	5,338	5	1
13	373,404	106,02	291,6	817,6	1,647	61,7	1,28	42,03	6,247	4	1
14	387,388	106,37	288,6	826,5	1,658	67,2	1,34	42,17	5,309	5	1
15	408,291	106,71	280,6	801,2	1,685	66,4	1,45	42,30	6,070	4	2
16	354,404	103,55	277,1	797,8	1,670	68,6	1,27	41,05	5,274	5	2
17	343,421	103,61	283,5	788,5	1,651	59,7	1,21	41,07	5,728	4	2
18	423,305	110,32	283,6	829,2	1,705	73,0	1,49	43,73	5,267	5	3
19	344,409	102,60	270,9	778,1	1,681	68,0	1,27	40,67	4,438	5	3
20	369,419	107,17	280,2	825,8	1,690	75,4	1,31	42,48	4,471	6	3
21	358,436	107,23	286,6	816,4	1,671	65,8	1,25	42,50	4,925	5	3
22	453,288	112,74	291,6	858,3	1,700	74,9	1,55	44,69	5,907	5	2
23	399,402	109,59	288,2	854,9	1,685	77,4	1,38	43,44	4,525	6	2
24	388,419	109,64	294,6	845,5	1,666	67,8	1,31	43,46	5,434	5	2

The statistical 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

15 software [22]. 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 diarylaniline derivatives were studied by statistical methods based on the principal component analysis. The PCA is a useful multivariate statistical technique to describe 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 Table 2 and Table 3.

Table 3: The correlation matrix (Pearson (n)) between different studied descriptors

Variables	MW	MR	MV	Pc	n	γ	D	α_e	LogP	HBA	HBD	pEC_{50}
MW	1											
MR	0,902	1										
MV	0,721	0,904	1									
Pc	0,826	0,902	0,788	1								
n	0,262	0,057	-0,375	0,106	1							
γ	0,040	-0,134	-0,457	0,187	0,762	1						
D	0,913	0,675	0,378	0,639	0,571	0,317	1					
α_e	0,902	1	0,904	0,903	0,057	-0,134	0,676	1				
LogP	0,678	0,629	0,731	0,407	-0,345	-0,591	0,480	0,630	1			
HBA	0,075	0,223	0,067	0,468	0,325	0,591	0,054	0,222	-0,551	1		
HBD	-0,462	-0,369	-0,554	-0,460	0,513	0,237	-0,290	-0,370	-0,650	0,186	1	
pEC_{50}	-0,236	-0,254	-0,419	-0,166	0,430	0,429	-0,065	-0,255	-0,399	0,091	0,494	1

The multiple linear regression method is used to study the relation between one dependent variable and several independent variables. It is a mathematic technique that minimizes differences between the actual and predicted values. It was also used to select the significant 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 the cytotoxic effects EC_{50} of the diarylaniline derivatives. The equations were justified by the coefficient of determination (R2), the mean squared error (MSE), the test- F (Fisher), and the significance level (p-value) [23, 24].

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 many different ANN models. The most frequently used ANN in QSAR is the three-layered feed-forward network [25]. 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 the neurons belonging to the same layer.

According to the supervised learning adopted, the networks are taught by giving them examples of input patterns and the corresponding target outputs. Through an iterative process, the connection weights are modified until the network gives the desired results for the training set of data. A back-propagation algorithm was used to minimize the error function. This algorithm has been described previously with a simple example of application [26] and a detail of this algorithm is given elsewhere [27].

Cross-validation is a popular technique utilized to explore the reliability of statistical models. Based

on this technique, a great number of modified data sets were created by deleting in each case one or a small group of molecules. These procedures are named "leave-one-out" and "leave-some-out" [28-30]. For each data set, an input-output model was developed. The model is evaluated by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been used in the development of the model). 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 (EC_{50}) of the 24 selected molecules that have been synthesized and evaluated for their inhibitory activity in vitro against the HIV-1 (experimental values) [14, 15]. The exploitation of experimental data observed using mathematical and statistical tools is an effective method to find new chemical compounds with high HIV-1 inhibitory activity. The values of the 11 chemical descriptors as shown in Table 2. The principle is to perform in the first time, a PCA, which allows us to choose the relevant descriptors from several correlated descriptors (dependent), then perform a decreasing study of MLR based on the elimination of no significant descriptors until approaching a valid model (including the critical probability: p-value < 0.05 for the all descriptors and the model complete).

Principal Components Analysis (PCA)

The totality of the 11 descriptors (variables) coding the 24 molecules was submitted to a principal components analysis. The 12 principal components were obtained (Figure 2). The first three components F1, F2, and F3 contributing 49.73 %, 26.66 %, and 10.86 % respectively to the total variance. In addition, the total estimated information was 87.25%.

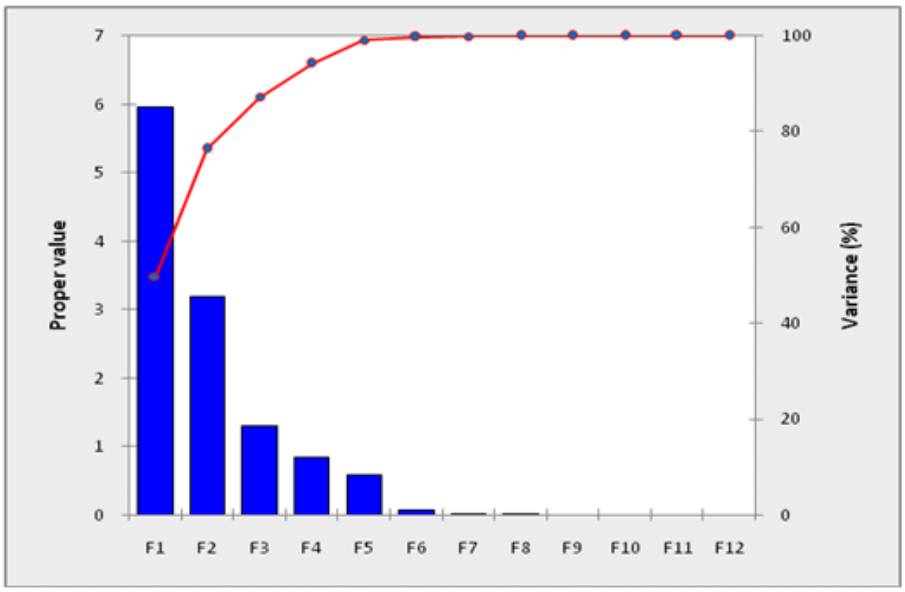


Figure 2: The principal components and their variances

The Pearson correlation coefficients are summarized in Table 3. The obtained matrix provides information on the negative or positive correlation between the variables. The principal component analysis (PCA) was conducted to identify the link between the different variables. Correlations between the 11 descriptors are demonstrated in Table3 as a correlation matrix, and in Figure 3 these descriptors are represented in a correlation circle.

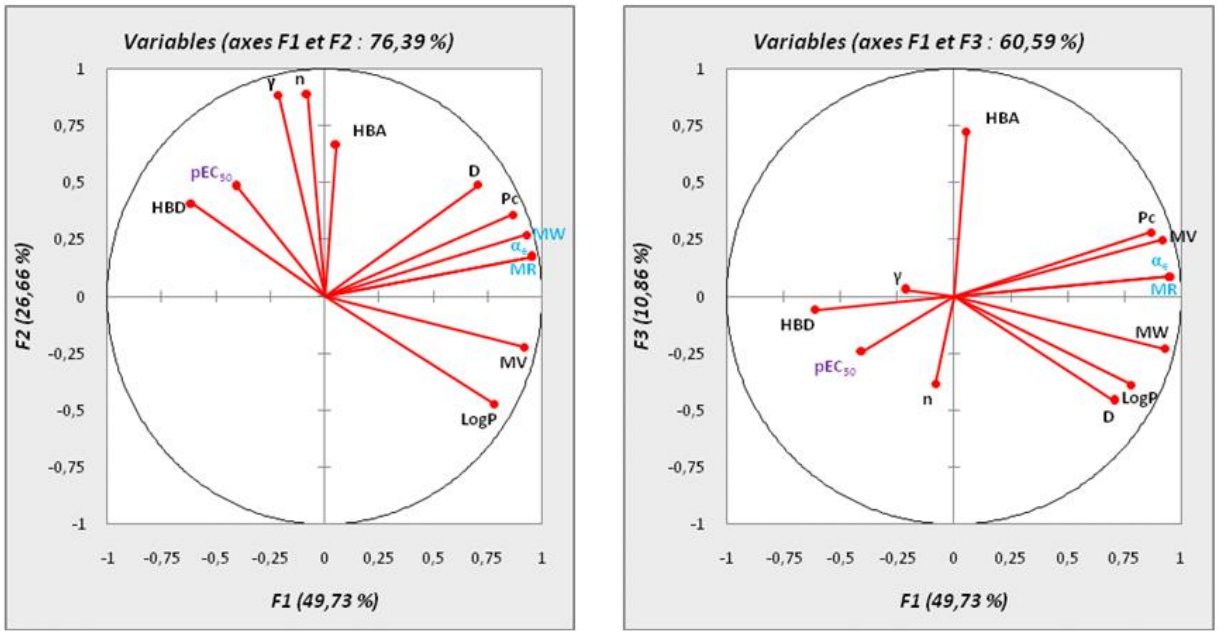


Figure 3: Correlation circles

To reduce the number of descriptors not significant (aberrant) and according to the matrix and the circle of correlations we observed the following correlations: (MR, α_e) are perfectly correlated (r=1), both variables are redundant. MW, MR and α_e are highly correlated (r (MW, MR) = 0,902); r (MW, α_e) = 0,902). The following variables then removed are: (α_e) and (MW).

Multiple Linear Regression (MLR)

To propose a mathematical model linking the descriptors and activity, and to quantitatively evaluate the substituent's physicochemical effects on the activity of the totality of the 24 molecules, we submitted the data matrix constituted obviously from the 9 variables corresponding to the 18 molecules (training set), to a descendent multiple regression analysis. This method used the coefficients R^2, R^2_{aj}, q^2, and the p-values to select the best regression performance.

Treatment with multiple linear regressions was more accurate because it allows us to connect the structural descriptors for each activity of 18 molecules to quantitatively evaluate the effect of substituent. The selected descriptors are: MR, MV, Pc, γ, HBA and HBD.

The QSAR model built using multiple linear regression (MLR) method is represented by the following equation:

pEC_{50MLR}= -127,329-0,678MR + 1,834MV - 0,550Pc + 2,073γ - 1,871HBA + 1,498 HBD (Equation 1)

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 [31, 32].

Overall assessment of the regression

Table 4 summarizes the results of the regression analysis of variance such as Fisher's value (Fexp) and overall p-value of the model.

The variability not explained by the model is the sum of residual squares SSE = 2.760 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 = 10.127 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 (6,728) and lower overall p-value at α (F value) = 0.05 level (p-value <0.05).

Table 4: Variance analysis

Source	SS	df	MS	F-exp	p-Value
Regression	10.127	6	1.688	6.728	0.003
Residual	2.760	11	0.251	-	-
Total	12.887	17	1.939	-	-

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: $\begin{cases} \mathbf{H_0} : R = 0 \\ \mathbf{H_1} : R \neq 0 \end{cases}$

If the correlation coefficient is zero, we reject the hypothesis H_0 (null hypothesis) and accept 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 t_{calc} is written:

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

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

Where $t_{\left(1-\frac{\alpha}{2}\right),(N-p-1)}$ is the value of the Student law for $(N-p-1)$ degree of freedom, a probability$\left(1-\frac{\alpha}{2}\right)$.
In our case we have N = 18 and R = 0.886. This corresponds to t_{calc} = 6.352, one rejects H_0 (null hypothesis) where: $t_{calc} > t_{\left(1-\frac{\alpha}{2}\right),(N-p-1)}$.
According to the Student table $\left(1-\frac{\alpha}{2}\right)$ = 0.975 and N = 18 is obtained $t_{(0.975,11)}$ = 2.201.
$t_{calc} > t_{(0.975,11)}$ 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 H_0 : SSF = SSE ($V_F = V_E$) Where (Error Variance) V_E = MSE

-against hypothesis H_1 : SSF > SSE ($V_F > V_E$) Where (Factor Variance) V_F = MSF

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

$$F_{exp} = \frac{VF}{VE} = \frac{MSF}{MSE} = \frac{SSF/p}{SSE/N-p-1}$$

-To a threshold of (0.05) comparing F_{exp} obtained by the theoretical calculation and that obtained from Fisher's table $F_{(p,N-p-1)}$ 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 $F_{exp} > F_{(6,11)}$.
-We then find $F_{(6,11)}$ = 3.09 and F_{exp} = 6.728, so we accept H_1 and H_0 is rejected.

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.

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

-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 = 0.886, so a good correlation was shown between the observed activity and that obtained by MLR.

Coefficient of Determination: R^2

The coefficient of determination R^2, gives the rate of explanation or percentage of the variation of Y (endogenous variables) explained by the variation in X (exogenous variable).

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

-In our case we have $R^2 = 0.786$, this figure means that 78.6% 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: R^2_{aj}

The overall quality of the linear regression is measured by the coefficient of determination (R^2_{aj}) "adjusted" taking into account the degree of freedom.

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

With: N = 18, p = 6 and $R^2 = 0.786$. In our case we have $R^2_{aj} = 0.669$, so the overall quality of the MLR is best. This indicates that this model is statistically significant.

Coefficient of Prediction: q^2

The q^2 value is used as the determining factor in selection of optimal models. The coefficient of prediction (q^2) was calculated using:

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

-SST: sum of total squares. In our case we have $q^2 = 0.98 > 0.6$, so the predictive power of this model is very significant, showing that the proposed model in this study paper is able to predict the activity with a great performance. 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 = 0.5, so the correlation between the observed activity and that obtained by MLR is the best.

N = 18 R = 0.886 $R^2 = 0.786$ F = 6.728 MSE = 0.251

Due to the fact that corresponding to the F value is much smaller than 0.05 it means that we would be taking a lower than 0.01 % risk assuming that the null hypothesis is wrong. Therefore, we can conclude that this model can provide a significant amount of information.

The elaborated QSAR model revealed that the HIV-1 inhibitory activity could be explained by many topologic factors. The negative correlation of the molar refractivity (MR), the parachor (Pc) and the hydrogen bond acceptor (HBA) with the ability to displace the diarylaniline activity reveals that a decrease in the value of pEC_{50}, while the positive correlation of the descriptors (molar volume (MV), the surface tension (γ) and the hydrogen bond donor (HBD)) with the ability to displace the diarylaniline activity reveals an increase in the value of pEC_{50}.

With the optimal MLR model, the values of predicted activities $pEC_{50\ MLR}$ calculated from Equation1 and the observed values are presented in Table 5. The correlations of predicted and observed activities are illustrated in Figure 4. 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.

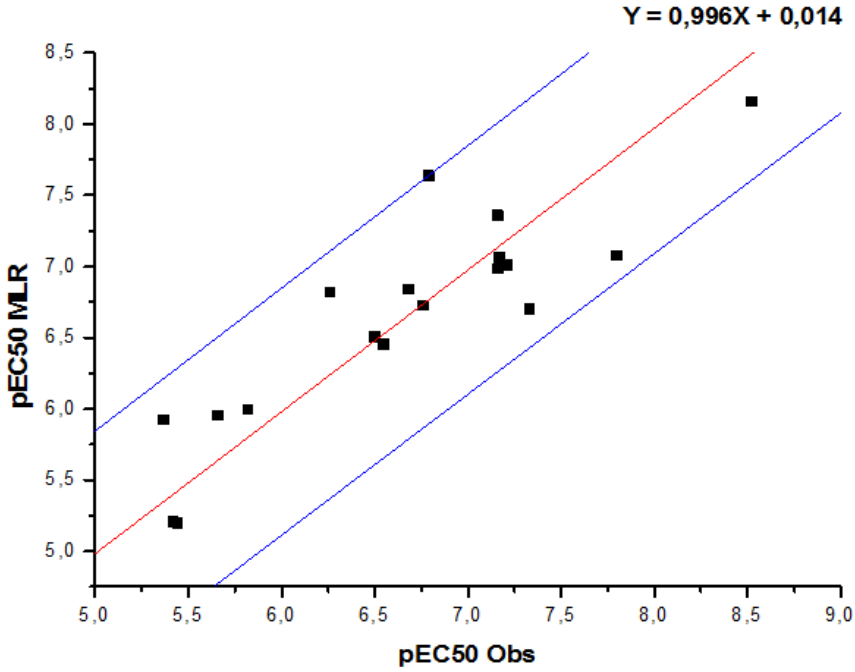


Figure 4: Correlations of observed and predicted activities calculated using MLR

Table 5: The observed, the predicted activities (**pEC_{50}**), according to different methods MLR, MNLR, ANN and CV-MLR for the 18 derivatives of diarylaniline (training set)

N°	pEC_{50Obs}	pEC_{50MLR}	pEC_{50ANN}	$pEC_{50\ CV\text{-}MLR}$	N°	pEC_{50Obs}	pEC_{50MNLR}
1	5,420	5,207	5.439	5,40	1	5,420	5,155
3	5,440	5,198	5.440	5,48	2	5,520	5,667
4	5,370	5,921	5.505	5,26	3	5,440	5,233
5	6,760	6,729	6.792	6,66	4	5,370	5,708
6	6,260	6,819	6.550	7,15	6	6,260	6,507
8	6,550	6,451	6.574	7,08	7	5,380	5,728
9	5,820	5,993	5.663	5,87	8	6,550	6,116
10	6,500	6,504	6.455	6,40	11	5,500	5,475
12	6,680	6,839	6.499	6,02	12	6,680	6,363
13	7,170	7,062	7.094	6,99	13	7,170	6,761
14	5,660	5,951	5.615	5,35	14	5,660	6,236
15	7,330	6,698	7.288	6,45	16	7,160	7,174
16	7,160	6,985	7.286	7,40	17	7,140	7,236
18	6,790	7,636	7.515	6,36	18	6,790	7,061
21	7,160	7,358	7.512	7,48	20	7,520	7,172
22	7,800	7,076	7.429	7,99	21	7,160	7,237
23	8,520	8,161	7.503	8,59	22	7,800	7,319
24	7,210	7,012	7.432	7,18	24	7,210	7,581

Multiples Non-LinearRegression (MNLR)

The nonlinear regression model was used to improve the structure-activity relationship, and 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 (training set) molecules. The coefficients R, R^2, and the F-values were used to select the best regression performance. We used a pre-programmed function of XLSTAT following:

$$Y = a + (bX_1 + cX_2 + dX_3 + eX_4 \ldots) + (fX_1^2 + gX_2^2 + hX_3^2 + iX_4^2 \ldots)$$

Where a, b, c, d: represent the parameters and X_1, X_2, X_3, X_4: represent the variables.

The resulting equations are as following:

$$pEC_{50MNLR} = 97{,}173 - 0{,}849MR - 5{,}760MV + 2{,}593Pc - 5{,}716\gamma - 1{,}162HBA + 3{,}555\ HBD + 2{,}518E\text{-}03(MR)^2 + 5{,}363E\text{-}03(MV)^2 - 9{,}763E\text{-}04\ (Pc)^2 + 2{,}118E\text{-}02\ (\gamma)^2 + 4{,}729E\text{-}02\ (HBA)^2 - 0{,}656\ (HBD)^2$$

N=18 R=0.925 R^2=0.857 MSE=0.360

With the optimal MNLR model, the values of predicted activities $\mathbf{pEC_{50\ MNLR}}$ calculated from the Equation2 and the observed values are given in Table5. The correlations of predicted and observed activities are illustrated in Figure 5.

The correlation between MNLR calculated and experimental activities are very significant as illustrated in Figure5 and as indicated by R and R^2 values.

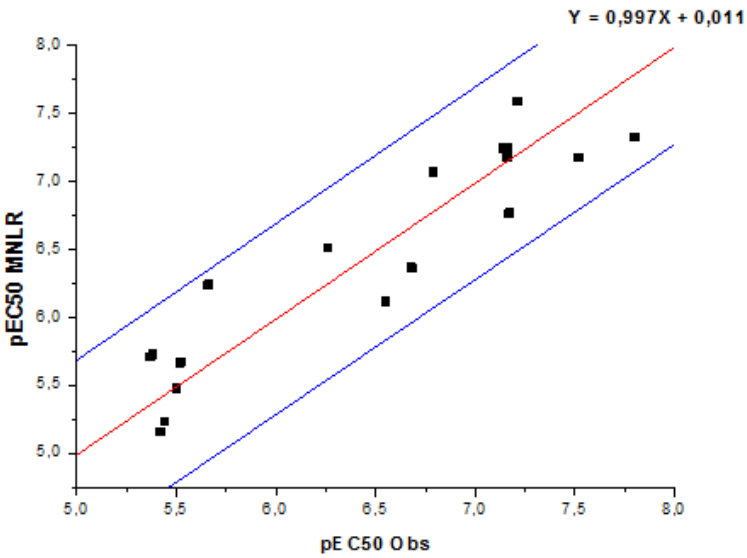


Figure 5: Correlations of observed and predicted activities calculated using MNLR

True predictive power of a QSAR model is to test their ability to predict accurately the activities of compounds from an external test set (compounds which were not used for the model development), the activities of the remained set of 6 compounds (1-6) are deduced from the quantitative model proposed with the 18 molecules (training set) by MLR and MNLR. Their structures and the observed and calculated pEC_{50}values are given in Tables 1 and 6.

Table 6: The observed, the predicted activities (pEC_{50}), according to methods MLR and MNLR for the 6 tested compounds (test set)

N°	pEC_{50Obs}	$pEC_{50MLR\text{-}test}$	N°	pEC_{50Obs}	$pEC_{50MNLR\text{-}test}$
2	5,520	7,538	5	6,760	5,235
7	5,380	7,424	9	5,820	5,465
11	5,500	6,441	10	6,500	6,244
17	7,140	7,225	15	7,330	7,005
19	5,490	7,366	19	5,490	6,156
20	7,520	8,515	23	8,520	7,243

The comparison of the values of pEC_{50test} to $pEC_{50\ Obs}$ shows that a good prediction has been obtained for the 6 compounds:

MLR:

$N = 6R_{test} = 0.6$

MNLR:

$N = 6R_{test} = 0.7$

Artificial Neural Networks (ANN)

To increase the probability of good characterization of the studied compounds, artificial neural networks (ANN) was 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 was developed using the properties of several studied compounds. Some authors [33, 34] have proposed a parameter ρ, leading to determine the number of hidden neurons, playing 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)

To avoid over fitting or under fitting, it is recommended that $1.8 < \rho < 2.3$. The output layer represents the calculated activity values pEC_{50}. The architecture of the ANN used in this work (6-1-1), ρ =2.

The values of predicted activities $pEC_{50\ ANN}$ calculated using ANN and the observed values are given in Table4. The correlations of predicted and observed activities are illustrated in Figure 6.

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

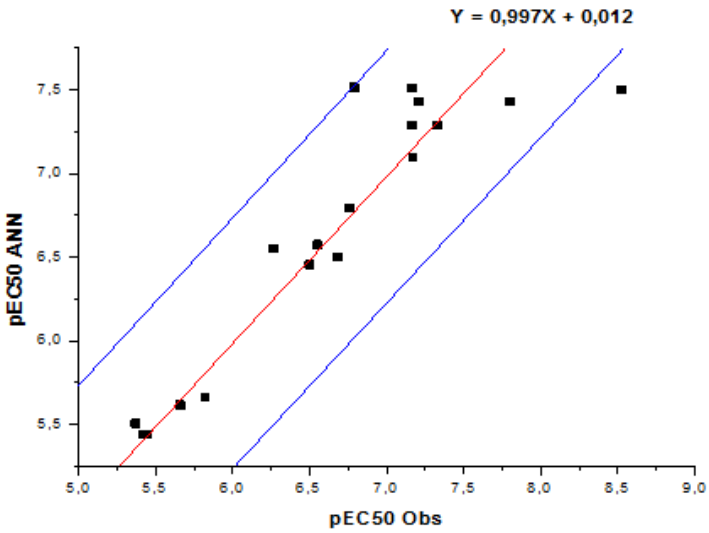


Figure 6: Correlations of observed and predicted activities calculated using ANN
$N = 18R = 0.916R^2 = 0.84$

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

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 evaluate 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). So, one compound was removed from the data set, the network was trained with the remaining compounds and used to predict the discarded compound. The process was repeated in turn for each compound in the data set.

In this study, the 'leave-one-out' procedure was used to assess the predictive ability of the ANN.

The values of predicted activities $pEC_{50\ CV}$ calculated using CV and the observed values are presented in Table5. The correlations of predicted and observed activities are illustrated in Figure 7.

The correlation between the calculated CV and the experimental activities are illustrated in Figure 7.

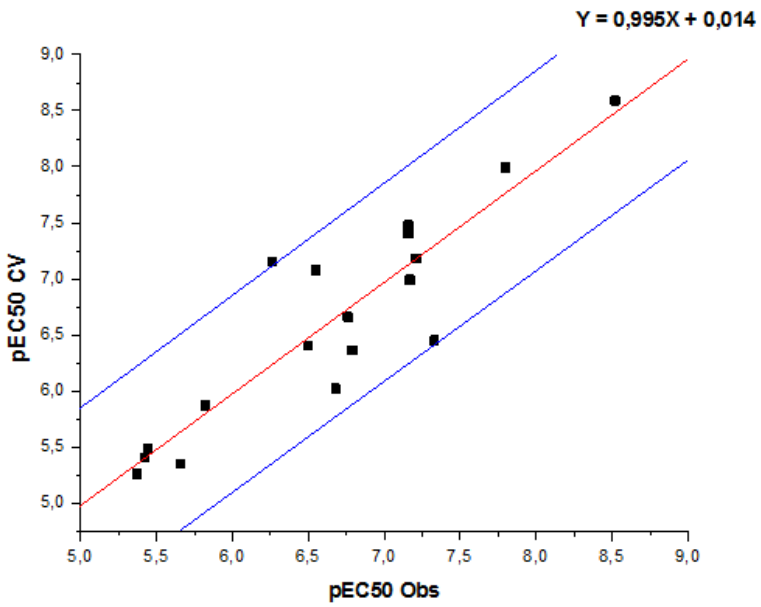


Figure 7: Correlations of observed and predicted activities calculated using CV

N = 18 R = 0.903 R^2 = 0.815

The results obtained with the cross validation and with the prediction of activities of the 6 compounds (test set), shows that the proposed model in this study is able to predict the activity with a great performance. The test conducted with the 6 compounds confirmed the performance of the model. So, this model could be applied to all derivatives of diarylaniline accordingly to Table 1 and could add further knowledge in the improvement of the search in the domain of inhibitors of HIV-1 virus. A comparison of the quality of MLR, MNLR and ANN models revealed that the ANN models had a better predictive capability. 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}= 0.903. So the predictive power of this model was very accurate. The results obtained in this study, revealed that the MLR, MNLR, and ANN are models validated,

representing the fact that the prediction of the new compounds was feasible.

Conclusion

In this study, three different modelling methods, including MLR, MNLR, and ANN were used to construct a QSAR model for the inhibitors of reverse transcriptase (RT) enzyme. The artificial neural network ANN revealed a better predictive capability in comparison with the MLR and MNLR model. We established a relationship between several descriptors and the HIV-1 inhibitory activity in satisfactory manners. The good results obtained with the cross- validation CV and with the prediction of activities of the test compounds (test set), shows that the proposed model in this study is able to predict activity with a great performance.

The accuracy and predictability of the proposed models were illustrated by the comparison of the key statistical terms including R and R^2 of different models obtained by using different statistical tools. It was also shown that the proposed methods were a useful to reduce the time and cost of synthesis and activity determination of inhibitors of (RT) (compounds based on diarylaniline).

He studied descriptors, which were sufficiently rich in chemical and topological information to encode the structural feature might be used with other descriptors for the development of predictive QSAR models.

Previous studies QSAR already performed on the same set of diarylaniline using multiple linear regression, obtained a correlation coefficient (R = 0.866). In this study the correlation coefficient obtained from the MLR (R_{MLR} = 0.886), by using a variety of descriptors, the coefficient improved by using MNLR (R_{MNLR} = 0.925) and ANN (R_{ANN} = 0.916) so the proposed model was very significant and was performance is tested by cross-validation method CV (R_{CV} = 0.903).

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.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

All authors contributed toward data analysis, drafting, and revising the paper and agreed to responsible for all the aspects of this work.

Conflict of interest

The authors declare that they have no conflicts of interest in this article.

References

[1]. Mandal A.S., Roy K., Predictive QSAR modeling of HIV reverse transcriptase inhibitor TIBO derivatives, *European journal of medicinal chemistry*, 2009, **44**:1509 [Crossref], [Google Scholar], [Publisher]

[2]. Vadivelan S., Deeksha T.N., Arun S., Machiraju P.K., Gundla R., Sinha B.N., Jagarlapudi S.A., Virtual screening studies on HIV-1 reverse transcriptase inhibitors to design potent leads, *European journal of medicinal chemistry*, 2011, **46**:851 [Crossref], [Google Scholar], [Publisher]

[3]. Corbett J.W., Gearhart L.A., Ko S.S., Rodgers J.D., Cordova B.C., Klabe R.M., Erickson-Viitanen S.K., Novel 2, 2-dioxide-4, 4-disubstituted-1, 3-H-2, 1, 3-benzothiadiazines as non-nucleoside reverse transcriptase inhibitors, *Bioorganic & medicinal chemistry letters*, 2000, **10**:193 [Crossref], [Google Scholar], [Publisher]

[4]. Masuda N., Yamamoto O., Fujii M., Ohgami T., Fujiyasu J., Kontani T., Moritomo A., Orita M., Kurihara H., Koga H., Kageyama S., Ohta M., Inoue H., Hatta T., Shintani M., Suzuki H., Sudo K.,

Shimizu Y., Kodama E., Matsuoka M., Fuji-wara M., Yokota T., Shigeta S., Baba M., Studies of non-nucleoside HIV-1 reverse transcriptase inhibitors. Part 2: Synthesis and structure–activity relationships of 2-cyano and 2-hydroxy thiazolidenebenzenesulfonamide derivatives, *Bioorganic & medicinal chemistry*, 2005, **13**:949 [Crossref], [Google Scholar], [Publisher]

[5]. Spallarossa A., Cesarini S., Ranise A., Schenone S., Bruno O., Borassi A., Colla P.L., Pezzullo M., Sanna G., Collu G., Secci B., Loddo R., Parallel synthesis, molecular modelling and further structure–activity relationship studies of new acylthiocarbamates as potent non-nucleoside HIV-1 reverse transcriptase inhibitors, *European journal of medicinal chemistry*, 2009, **44**:2190 [Crossref], [Google Scholar], [Publisher]

[6]**.** Weitman M., Lerman K., Nudelman A., Major D.T., Hizi A., Herschhorn A., Structure–activity relationship studies of 1-(4-chloro-2, 5-dimethoxyphenyl)-3-(3-propoxypropyl) thiourea, a non-nucleoside reverse transcriptase inhibitor of human immunodeficiency virus type-1, *European journal of medicinal chemistry*, 2011, **46**:447 [Crossref], [Google Scholar], [Publisher]

[7]. De La Rosa M., Kim H.W., Gunic E., Jenket C., Boyle U., Koh Y.H., Korboukh I., Allan M., Zhang W., Chen H., Xu W., Tri-substituted triazoles as potent non-nucleoside inhibitors of the HIV-1 reverse transcriptase, *Bioorganic & medicinal chemistry letters*, 2006, **16**:4444 [Crossref], [Google Scholar], [Publisher]

[8]. Ellis D., Kuhen K.L., Anaclerio B., Wu B., Wolff K., Yin H., Bursulaya B., Caldwell J., Karanewsky D., He Y., Design, synthesis, and biological evaluations of novel quinolones as HIV-1 non-nucleoside reverse transcriptase inhibitors, *Bioorganic & medicinal chemistry letters*, 2006, **16**:4246 [Crossref], [Google Scholar], [Publisher]

[9]. Rios A., Quesada J., Anderson D., Goldstein A., Fossum T., Colby-Germinario S., Wainberg M.A., Complete inactivation of HIV-1 using photo-labeled non-nucleoside reverse transcriptase inhibitors, *Virus research*, 2011, **155**:189 [Crossref], [Google Scholar], [Publisher]

[10]. Spallarossa A., Cesarini S., Ranise A., Bruno O., Schenone S., La Colla P., Collu G., Sanna G., Secci B., Loddo R., Novel modifications in the series of O-(2-phthalimidoethyl)-N-substituted thiocarbamates and their ring-opened congeners as non-nucleoside HIV-1 reverse transcriptase inhibitors. *European journal of medicinal chemistry*, 2009, **44**:1650 [Crossref], [Google Scholar], [Publisher]

[11] He Y., Chen F., Sun G., Wang Y., De Clercq E., Balzarini J., Pannecouque C., 5-Alkyl-2-[(aryl and alkyloxylcarbonylmethyl) thio]-6-(1-naphthylmethyl) pyrimidin-4 (3*H*)-ones as an unique HIV reverse transcriptase inhibitors of S-DABO series. *Bioorganic & Medicinal Chemistry Letters*, 2004, **14**:3173 [Crossref], [Google Scholar], [Publisher]

[12]**.** Monforte A.M., Logoteta P., De Luca L., Iraci N., Ferro S., Maga G., De Clercq E., Pannecouque C., Chimirri A., Novel 1, 3-dihydro-benzimidazol-2-ones and their analogues as potent non-nucleoside HIV-1 reverse transcriptase inhibitors, *Bioorganic & medicinal chemistry*, 2010, **18**:1702 [Crossref], [Google Scholar], [Publisher]

[13]**.** Liang Y.H., He Q.Q., Zeng Z.S., Liu Z.Q., Feng X.Q., Chen F.E., Balzarini J., Pannecouque C., De Clercq E., Synthesis and anti-HIV activity of 2-naphthyl substituted DAPY analogues as non-nucleoside reverse transcriptase inhibitors, *Bioorganic & medicinal chemistry*, 2010, **18**:4601 [Crossref], [Google Scholar], [Publisher]

[14]**.** Zeng Z.S., He Q.Q., Liang Y.H., Feng X.Q., Chen F.Er., DeClercq E., Balzarini J., Pannecouque C., Hybrid diarylbenzopyrimidine non-nucleoside reverse transcriptase inhibitors as promising new leads for improved anti-HIV-1 chemotherapy, *Bioorganic & medicinal chemistry*, 2010, **18**:5039 [Crossref], [Google Scholar], [Publisher]

[15]. Qin B., Jiang X., Lu H., Tian X., Barbault F., Huang L., Qian K., Chen C.H., Huang R., Jiang S., Lee K.H., Diarylaniline derivatives as a distinct class of HIV-1 non-nucleoside reverse transcriptase inhibitors, *Journal of medicinal*

chemistry, 2010, **53**:4906 [Crossref], [Google Scholar], [Publisher]
[16]. Nantasenamat C., Isarankura Na Ayudhya C, Naenna T, Prachayasittikul V. *A practical overview of quantitative structure-activity relationship. Excli J*, 2009, **8**:74 [Google Scholar]
[17]. Nantasenamat C., Isarankura-Na-Ayudhya C., Prachayasittikul V., Advances in computational methods to predict the biological activity of compounds, *Expert opinion on drug discovery,* 2010, **5**:633 [Crossref], [Google Scholar], [Publisher]
[18]. Mustafa Y.F., Chehardoli G., Habibzadeh S., Arzehgar Z., Electrochemical detection of sulfite in food samples. *Journal of Electrochemical Science and Engineering*, 2022, **12**:1061 [Crossref], [Google Scholar], [Publisher]
[19]. ACD/ChemSketch Version 4.5 for Microsoft Windows User's Guide [PDF]
[20]. Borowiecki P., Chemoenzymatic Synthesis of Optically Active Ethereal Analog of iso-Moramide—A Novel Potentially Powerful Analgesic, *International Journal of Molecular Sciences*, 2022, **23**:11803 [Crossref], [Google Scholar], [Publisher]
[21]. Allinger N.L., Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms, *Journal of the American Chemical Society*, 1977, **99**:8127 [Crossref], [Google Scholar], [Publisher]
[22]. XLSTAT 2015 Add-in software (XLSTAT Company)
[23]. Boukarai Y., Khalil F., Bouachrine M., QSAR study of isatin (1*H*-indole-2, 3-dione) analogues as in vitro anti-cancer agents using the statistical analysis methods and the artificial neural network. *International Journal of Scientific & Engineering Research*, 2015, **6**:159 [Google Scholar]
[24]. Boukarai Y., Khalil F., Bouachrine M., QSAR study of 5, 6-bicyclic heterocycles analogues as anti-Alzheimer's agents using the statistical analysis methods. *Journal of Chemical and Pharmaceutical Research*, 2016, **8**:1000 [Google Scholar]
[25]. Zupan J., Gasteiger J., Neural networks for chemists: an introduction. John Wiley & Sons, 1993 [Google Scholar], [Publisher]
[26]. Cherqaoui D., Villemin D., Use of a neural network to determine the boiling point of alkanes, *Journal of the Chemical Society, Faraday Transactions*, 1994, **90**:97 [Crossref], [Google Scholar], [Publisher]
[27]. Freeman J.A., Skapura D.M., *Neural networks: algorithms, applications, and programming techniques*. Addison Wesley Longman Publishing Co., Inc. 1991 [Google Scholar], [Publisher]
[28]. Efron B., Estimating the error rate of a prediction rule: improvement on cross-validation, *Journal of the American statistical association*, 1983, **78**:316 [Crossref], [Google Scholar], [Publisher]
[29]. Efroymson M.A., Multiple regression analysis. *Mathematical methods for digital computers*, 1960, 191 [Google Scholar], [Publisher]
[30]. Osten D.W., Selection of optimal regression models via cross-validation, *Journal of Chemometrics*, 1998**, 2:**39 [Crossref], [Google Scholar], [Publisher]
[31]. Boukarai Y., Khalil F., Bouachrine M., Application of QSAR Methods on the Study of Bioactive Molecules Derived from Isatin. *Journal of Chemical and Pharmaceutical Research*, 2016, **8**:136 [Google Scholar]
[32]. Boukarai Y., Khalil F., Bouachrine M., QSAR study of flavonoid derivatives as in vitro inhibitors agents of Aldose Reductase (ALR2) enzyme for diabetic complications. *Journal of Materials and Environmental Science*, 2017, **8**:1532 [PDF], [Google Scholar]
[33]. So S.S., Richards W.G., Application of neural networks: quantitative structure-activity relationships of the derivatives of 2, 4-diamino-5-(substituted-benzyl) pyrimidines as DHFR inhibitors, *Journal of Medicinal Chemistry*, 1992, **35**:3201 [Crossref], [Google Scholar], [Publisher]
[34]. Andrea T.A., Kalayeh H., Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase

inhibitors, *Journal of medicinal chemistry*, 1991, **34**:2824 [Crossref], [Google Scholar], [Publisher]

HOW TO CITE THIS ARTICLE
Mohammed Bouachrine*, Youness Boukarai, Fouad Khalil. Study of Quantitative Structure-Activity Relationship (QSAR) of Diarylaniline Analogues as in Vitro Anti-HIV-1 Agents in Pharmaceutical Interest. *Chem. Methodol.*, 2023, 7(9) 719-735
DOI: https://doi.org/10.48309/chemm.2023.53807
URL: https://www.chemmethod.com/article_53807.html