British Journal of Multidisciplinary and Advanced Studies,7(3)1-25, 2025
Health and Medical Sciences
Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

Exploring QSARs for Inhibitory Activity of some Antimalarial Compounds by MLR and PC-ANN

Omar Deeb[1,*] and A. Ashour[2]
[1]Faculty of Pharmacy, Al-Quds University, PO Box 20002, Jerusalem, Palestine
[2]Faculty of Science and Technology, Industrial Applied Technology program, Al-Quds University, PO Box 20002, Jerusalem, Palestine
*Corresponding author: Omar Deeb, omardeeb@staff.alquds.edu, deeb.omar@gmail.com

doi: https://doi.org/10.37745/bjmas.2022.0548 Published May 08, 2026

Citation: Omar Deeb and A. Ashour (2026) Exploring QSARs for Inhibitory Activity of some Antimalarial Compounds by MLR and PC-ANN, *British Journal of Multidisciplinary and Advanced Studies*,7(2)1-25

Abstract: *As malaria disease is continuous to be one of the major health problems, and until now no effective vaccines or drugs are available due to the mutation of the plasmodium. So, in order to help in designing a new antimalarial agent, a quantitative structure activity relationship was performed to study the Activity of 79 compound as antimalarial agents. The QSAR models were developed using the multiple linear regression (MLR) as a linear method. The principal component – artificial neural network (PC-ANN) was used as nonlinear method for modeling. The models resulted have a good prediction power. The MLR models (13-17) which have R^2 >0.6, the best model was model number 17 with correlation coefficient R= 0.889, R^2= 0.791, and R^2adj.= 0.733. Cross validation LOO and LMO were performed on the resulted MLR models 13 - 17 in which they showed a good predictive power. The PCA was performed to divide the data into three data sets; training, validation and test set. Then the ANN performed on the models 13-17. The resulted ANN models were validated by randomization test, then the conditions that proposed by Golbraikh and Tropsha were applied to confirm that the QSAR models have acceptable prediction power or not. However, the best ANN model with the best predictive power was model number 17, with R value 0.8138.*

Keywords: QSAR, MLR, PC- ANN, inhibitory activity, antimalarial compounds.

INTRODUCTION

Malaria continues to be one of the major public health problems in many tropical countries causing extensive morbidity and loss of life [1]. Annual malaria mortality due to Plasmodium falciparum costs 1.5 to 2.7 million lives in Africa alone, comprising of mainly young children [2]. A four main species of parasites belonging to the genus Plasmodium are found to be the main causes of malaria which are; P.falciparum, P.vivax, P.malariae and P.ovale. These are human malaria species that have the ability to spread from one person to another *via* the bite of female mosquitoes of the genus Anopheles [3]. Plasmodium falciparum is the most lethal protozoan parasite of the genus, which is responsible for malaria complications such as cerebral malaria or severe anemia [4, 5]. At present, because of the high mutability of the genome of P. falciparum, there is no effective vaccines available [6], meanwhile, resistance of malaria parasites has also quickly developed to a variety of quinoline analogs (e.g., chloroquine), antifolates (e.g., sulfadoxine-pyrimethamine) and inhibitors of electron transport (e.g., atovaquone). Chloroquine which is the only synthetic antimalarial drug that

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

cured malaria for decades, rather than centuries, although has fallen to resistance. [7]. Chalcone, a biosynthetic product of shikimate pathway, is a class of privileged structure that has a wide range of biological properties. Chalcones are precursors of various flavones and key intermediates for combinatorial assembly of different heterocyclic scaffolds.

Despite of too much researches during the last 40 years, the exact mechanism by which chloroquine kills the malaria parasite remains controversial [8-10]. This drug inhibits DNA and RNA biosynthesis and induces the rapid degradation of ribosome'sand the dissimilation of ribosomal RNA. The inhibition of protein synthesis is also observed evidently as a secondary effect. It has been proposed that the inhibition of DNA replication is the general antimicrobial mechanism of action of chloroquine. Chloroquine accumulates in very high concentrations in the parasite food vacuole [11]. Once in the food vacuole, chloroquine is thought to inhibit the detoxification of heme. Chloroquine becomes protonated (to CQ2+) because the digestive vacuole is acidic (pH 4.7) and subsequently cannot leave the vacuole by diffusion. Chloroquine caps hemozoin molecules and prevents the further biocrystallization of heme, thus leading to heme buildup. Chloroquine binds to heme (or FP) to form what is known as the FP-chloroquine complex; this complex is highly toxic to the cell and disrupts membrane function. The actions of the toxic FP-chloroquine complex and FP result in cell lysis and ultimately the auto-digestion of the parasite cell. In essence, the parasite cell drowns in its own metabolic products.

MATERIALS AND METHODS

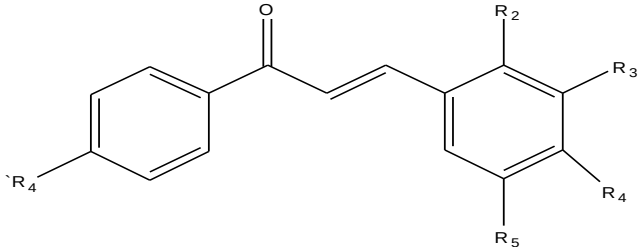
Software

Geometry optimizations were performed using HyperChem (Version 7.5; Hypercube, Inc, USA, http://www.hyper.com) at the AM1 level of theory. An AM1 optimization was chosen because it was developed and parameterized for common organic structures. Descriptors were calculated using HyperChem and DRAGON (Milano Chemometrics and QSAR Group, USA, evaluation version 5.0, http://www.disat.unimib.it/vhml) software. SPSS software (version 13.0, SPSS, Inc.) was used for the simple MLR analysis while ANN analysis was performed using MATLAB (Version 7.0.1 (R14), http://www.mathworks.com).

Chemical data and descriptors

A data set containing 79-compounds with their IC_{50} (µg/ml) against malaria disease taken from the literature [12-17], the activity of them were determined in the same way (candle jar method) and have the same activity unit. The compounds have seven different chemical structures, these all summarized with the biological activities in the Table 1.

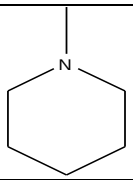
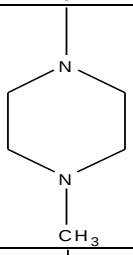
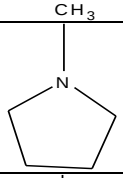
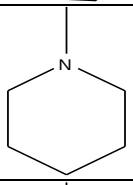
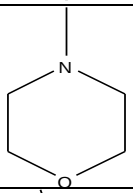
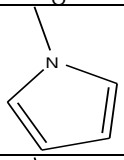
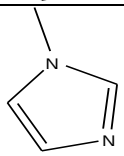
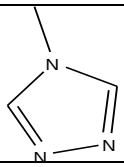
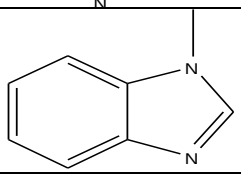
Table: 1**: Dataset, the compounds and antimalarial activity**



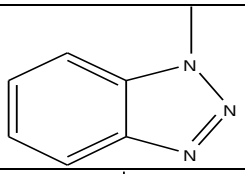
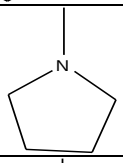
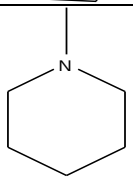
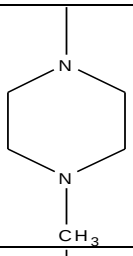
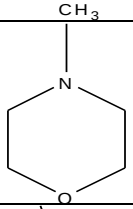
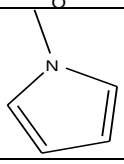
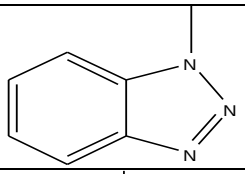
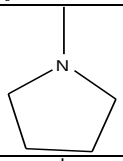
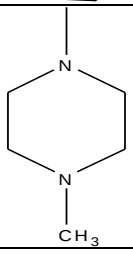
(Compounds 1-27)

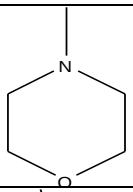
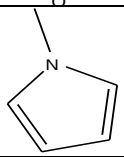
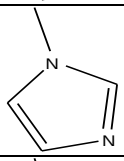
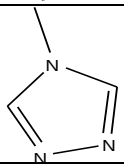
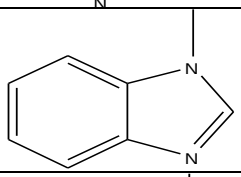
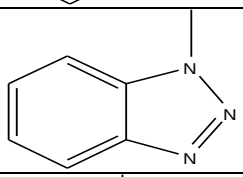
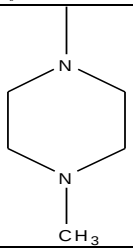
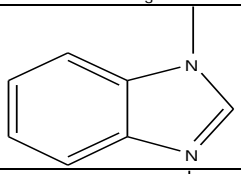
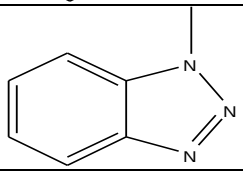
Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

Chemical structure of Chalcone

Compound No.	R_4`	R_2	R_3	R_4	R_5	IC_{50} (µg/ml)
001		H	H	OCH_3	H	6.92
002		H	H	OCH_3	H	6.9
003		OCH_3	H	OCH_3	H	2.37
004		OCH_3	H	OCH_3	H	7.68
005		OCH_3	H	OCH_3	H	2.95
006		OCH_3	H	OCH_3	H	5.98
007		OCH_3	H	OCH_3	H	6.7
008		OCH_3	H	OCH_3	H	3.38
009		OCH_3	H	OCH_3	H	1.1

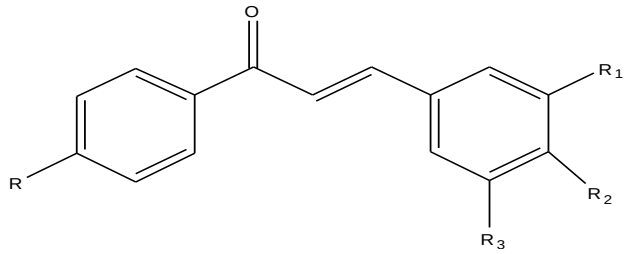
Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

010		OCH_3	H	OCH_3	H	7.22
011		OCH_3	H	H	OCH_3	5.53
012		OCH_3	H	H	OCH_3	6.13
013		OCH_3	H	H	OCH_3	10.1
014		OCH_3	H	H	OCH_3	3.26
015		OCH_3	H	H	OCH_3	6.36
016		OCH_3	H	H	OCH_3	12.73
017		H	OCH_3	OCH_3	H	2.91
018		H	OCH_3	OCH_3	H	4.96

019		H	OCH_3	OCH_3	H	5.16
020		H	OCH_3	OCH_3	H	6.28
021		H	OCH_3	OCH_3	H	7.34
022		H	OCH_3	OCH_3	H	5.85
023		H	OCH_3	OCH_3	H	5.04
024		H	OCH_3	OCH_3	H	3.5
025		H	OCH_3	OCH_3	OCH_3	4.7
026		H	OCH_3	OCH_3	OCH_3	7.6
027		H	OCH_3	OCH_3	OCH_3	8.15

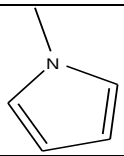
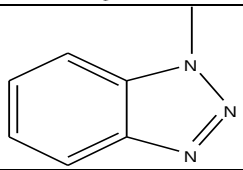
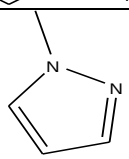
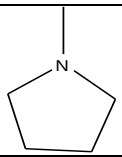
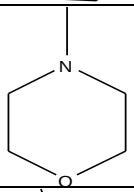
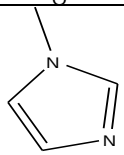
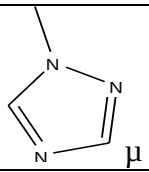
Ref. [12]

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

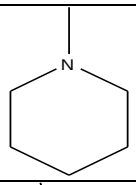
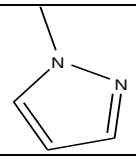
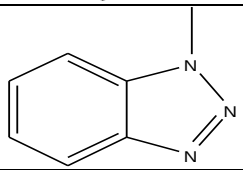
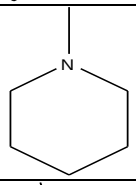
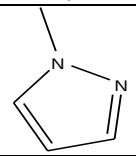
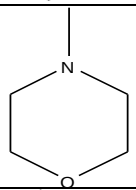
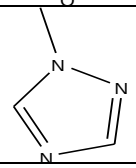


Compounds 28-41

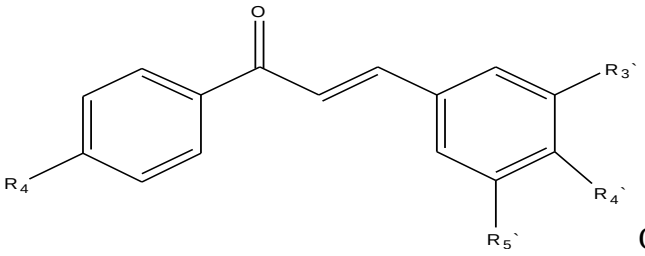
Chemical structure of Chalcone

Compound No.	R	R_1	R_2	R_3	IC_{50} (µg/ml)
028		H	Cl	H	2.93
029		H	Cl	H	2.5
030		H	Cl	H	7.76
031		H	Cl	H	6.01
032		H	Cl	H	9.1
033		H	Cl	H	8.26
034	 µ	H	Cl	H	1.52

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

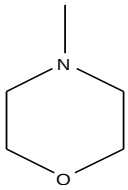
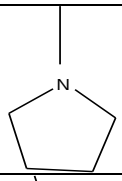
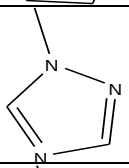
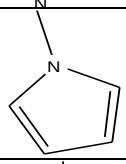
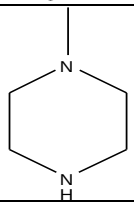
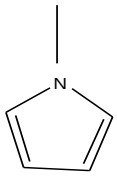
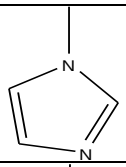
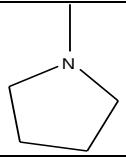
035		H	Cl	H	5.15
036		H	OCH_3	H	12.33
037		H	OCH_3	H	6.8
038		OCH_3	OCH_3	OCH_3	7.10
039		OCH_3	OCH_3	OCH_3	6.0
040		OCH_3	OCH_3	OCH_3	4.6
041		OCH_3	OCH_3	OCH_3	8.03

Ref. [13]

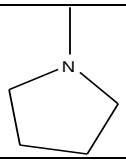
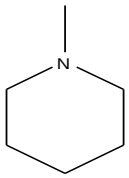
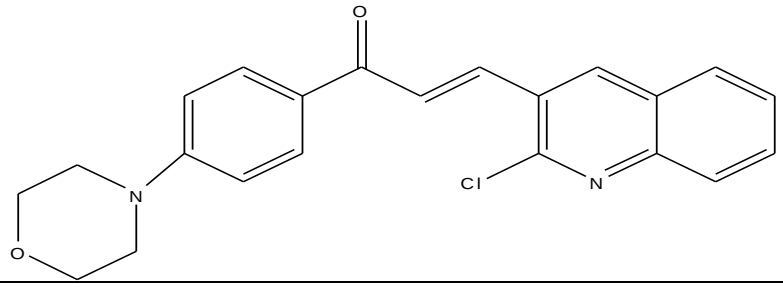


Compounds 42-52

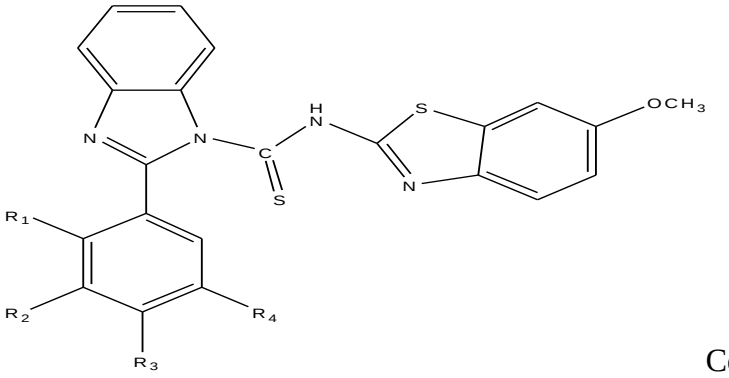
structure of Chalcone

Compound No.	R_4	R_3`	R_4`	R_5`	IC_{50} (µg/ml)
042		H	OCH_3	H	7.9
043		H	OCH_3	H	6.3
044		H	OCH_3	H	4.56
045		H	OCH_3	H	1.61
046		OCH_3	OCH_3	OCH_3	9.0
047		OCH_3	OCH_3	OCH_3	2.03
048		OCH_3	OCH_3	OCH_3	2.48
049		OCH_3	OCH_3	OCH_3	13.0

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

050		OCH_3	H	OCH_3	8.43
051		OCH_3	OCH_3	H	3.13
052					17.03

Ref. [14]



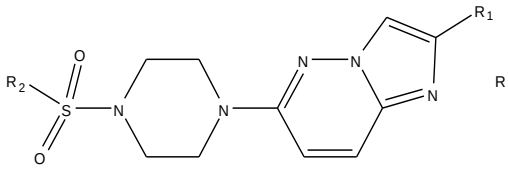
Compounds 53-60

Chemical structure of N-(6-methoxybenzo[*d*]thiazol-2-yl)-2-substituted phenyl-1*H*-benz[*d*]imidazole-1-carbothioamide derivatives

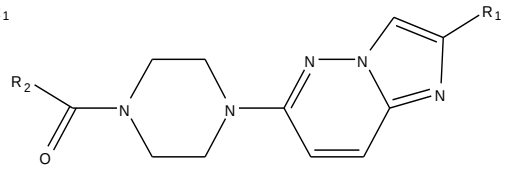
Compound No.	R_1	R_2	R3	R_4	IC_{50} (µg/ml)
053	H	NH_2	H	H	1.95
054	Cl	H	H	H	1.40
055	H	OCH_3	OCH_3	OCH_3	0.18
056	H	H	OCH_3	OCH_3	0.72
057	H	H	Cl	H	0.56
058	F	H	H	H	1.42
059	H	H	NH_2	H	1.10
060	H	H	NO_2	NO_2	0.11

Ref. [15]

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index



formula B

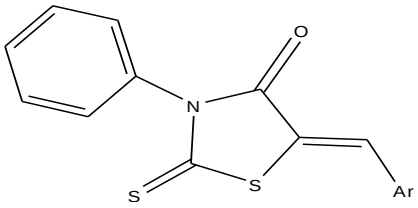


formula A

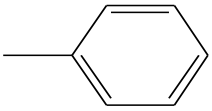
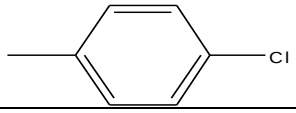
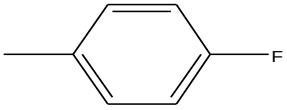
Chemical structure of amide and sulfonamide derivatives (formula A: 61-66, formula B: 67-69)

Compound No.	Formula	R_1	R_2	IC_{50} (µg/ml)
061	A	CF_3	C_2H_5	0.77
062	A	CF_3	$4CH_3$-phenyl	0.97
063	A	CF_3	$4CF_3$-phenyl	0.79
064	A	$4CH_3$-phenyl	C_2H_5	0.80
065	A	$4CF_3$-phenyl	C_2H_5	0.89
066	A	2,5-dichlorophenyl	C_2H_5	0.98
067	B	CF_3	C_2H_5	1.01
068	B	CF_3	$4CH_3$-phenyl	1.12
069	B	2,5-dichlorophenyl	C_2H_5	0.96

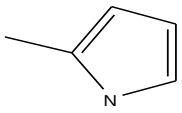
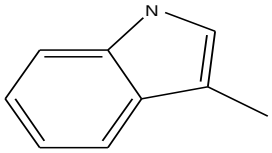
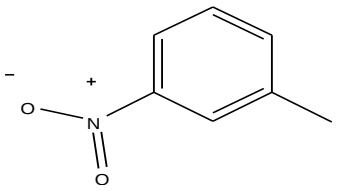
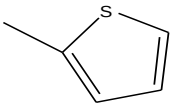
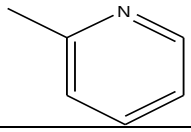
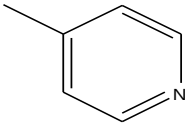
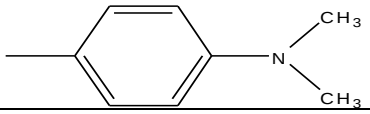
Ref. [16]



Chemical structure of 3-phenyl-2-thioxothiazolidin-4-one Compounds 69-79

Compound No.	Aryl group	IC_{50} (µg/ml)
070		1.16
071		0.90
072		1.28

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

073		1.14
074		1.22
075		0.98
076		1.06
077		1.15
078		0.85
079		0.94

Ref. [17]

The structures of the compounds are drawn by hyperchem software. The resultant structures are 2D then we convert them to 3D. HyperChem software was used to optimize the different compound structures using AM1 semi-empirical level. The optimization was preceded by the Polak-Rebiere algorithm. To be sure that we reached global minima, geometry optimization was run multiple times with different starting points for each molecule.

In this study, a pool of 1235 descriptors classified into 18 different groups was calculated using Dragon software. The constant or nearly constant descriptors for all the 79 compounds were discarded from further analysis. Furthermore, chemical descriptors such as HOMO, LUMO and polarizability were calculated using HyperChem software. Depending on the HOMO and LUMO values, electrophilicity, electronegativity, hardness, and softness descriptors were calculated. Other descriptors such as surface area approximate, surface area grid, volume, mass, polarizability, hydration energy, octanol-water partition coefficient (logP), and refractivity were calculated. Discarding highly inter-correlated ($r>0.95$) descriptors and following the procedure described in the next section, this number of descriptors was declined to 17 descriptors in the "final" MLR regression model (model 17 in Table 2).

Multiple linear regression (MLR) analysis

Multiple linear regression analysis with stepwise selection and elimination of variables was employed to model the inhibitory activity (IC_{50}) relationships with each group of descriptors separately. IC_{50} is the dependent variable and the set of descriptors as independent variables. Then, the “optimal” descriptors for each group were selected and gathered in one group to perform final MLR analysis.

Principal components analysis (PCA)

Collinear descriptors add redundancy to the input data matrix and consequently the performances of the models obtained by using these descriptors would be degraded. PCA and more specifically factor analysis, groups together variables that are collinear to form a composite indicator capable of capturing as much of common information of those indicators as possible. Each factor reveals the set of variables with the highest relationship. The idea under this approach is to explain the highest possible variation in the indicators set using the smallest possible number of factors. Consequently, the index no longer depends upon the dimensionality of the data set but it is rather based on the 'statistical' dimensions of the data. Application of PCA on a descriptor data matrix result in a loading matrix containing factors or PCs, which are orthogonal and therefore have no correlation with each other.

The PC’s were calculated by singular value decomposition (SVD) method in MATLAB environment (MathWork Inc. Version 7.0.1 (R14)). Due to the quality of data, a previous treatment of the data is essential before applying the multivariate analysis methods. Scaling and centering is one of the pre-processing methods needed before performing the regression methods joint with feature extraction. Projection methods results depend on the normalization of the data. Descriptors with small absolute values have a small contribution to overall variances leading to biased PC’s caused by the presence of other descriptors with higher values. In order to have the focus on the important variables in the model, equal weights are assigned to each descriptor, with appropriate scaling. Furthermore, descriptors were standardized to unit variance and zero mean (autoscaling) to give all variables the same importance. Then, the data matrix containing the entire set of descriptors and activity were simultaneously subjected to PCA.

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

Principal component-artificial neural network (PC-ANN) analysis

ANNs are computer-based models in which a number of nodes, also called neurons are interconnected by links forming netlike structure ‘‘layers.’’ A variable value is assigned to every neuron.

There are three kinds of neurons: (a) the input neurons which receive their values from independent variables and constitute the input layer, (b) the hidden neurons which collect values from other neurons, giving a result that is passed to a successor neuron, (c) the output neurons which take values from other units and correspond to different dependent variables, forming the output layer. In this sense, network architecture is commonly represented as I–H–O, where I, H, and O are the number of neurons in the input, hidden, and output layers, respectively.

The weights are links between units that condition the values assigned to the neurons. The weights are adjusted through a training process in order to minimize network error. For this, a non-linear transfer function relates the input parameters with the outputs. Commonly neural networks are adjusted, or trained, so that a particular input leads to a specific target output.

In PC-ANN analysis, as a preliminary treatment, the input data (i.e., molecular descriptors) were normalized to have zero mean and unity variance, and then were subjected to PCA before being introduced into the neural network. It should be illustrated that for each MLR resulted model, separate ANN models were developed so that the input's descriptors were the subsets selected by the stepwise MLR methods. In the case of each MLR model, a feed-forward neural network with back-propagation of error algorithm was constructed to model the activity–structure relationships between the descriptors on one hand and inhibitory activity on the other hand. The model development in ANN and the network architecture is fully described by us [18] and others [19]. The data set was divided into three subsets: training, validation and external test sets. The training and the validation sets are the norm in all model training processes. The test set is used to test the trend of the prediction precision of the model trained at some point of the training evolution. The extracted PC’s for each MLR model were classified homogenously, based on the factors space of the descriptors, into training set (60%), validation set (20%) and external test set (20%) according to the PCA and the two PC’s were plotted against each other (see Figure 1). Afterward, the training set was used to optimize the network performance. The regression between the network output and the observed activity was calculated for each set individually. The training function 'trainscg' was used to train the network. To find models with lower errors, the ANN algorithm was run many times, with different geometry and initial weights each time.

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

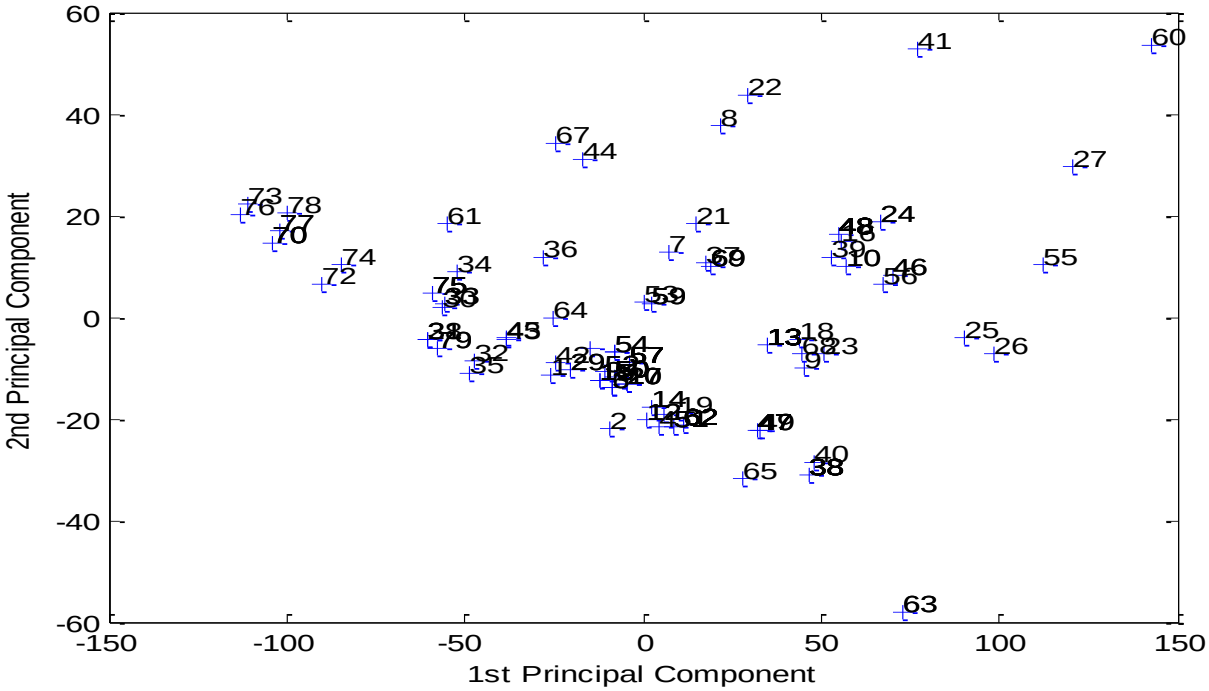


Figure 1. First and second principal components plot for the factor spaces of the descriptors and antimalarial ligands inhibitory activity.

RESULTS AND DISCUSSION

MLR analysis

In continuation to recent QSAR studies [20-24] done using similar methods, we developed an ANN-QSAR model that describes the inhibitory activity of a series of compounds using large number of different descriptors. MLR were performed on each one of the groups of descriptors individually (individual approach described in Ref. [25] by Deeb) where IC_{50} is the dependent variable. Stepwise method is used to develop multilinear equation by correlating dependent variable (activity) and the best independent variables.

Next, a new or "final" MLR analysis was performed by correlating the dependent variable (activity) and the optimal descriptors selected from the individual MLR models 6-17. Table 2 shows the regression models suggested from the "final" MLR analysis. The number of descriptors in these models is varied between 6 and 17. The highest coefficient of determination (R^2) obtained, is 0.791 for a regression model with 17 descriptors (model **17**). Table 3 shows a key for the different descriptors used in the final MLR model.

Table 2: MLR models resulted from the descriptors group

Model No.	No. of descriptors	R	R^2	R^2adj.	Selected descriptors
6	6	0.785	0.616	0.584	n=CHR, nC=N, RDF125e, nC=NPh, Log P, G(N..O)
7	7	0.800	0.640	0.604	n=CHR, nC=N, RDF125e, nC=NPh, Log P, G(N..O), SP10
8	8	0.818	0.668	0.631	n=CHR, RDF125e, nC=NPh, Log P, G(N..O), SP10, GATS3P, Mor32v
9	9	0.830	0.688	0.648	n=CHR, RDF125e, nC=NPh, LogP, G(N..O), SP10, GATS3P, Mor32v, Mor13u
10	10	0.837	0.700	0.656	n=CHR , RDF125e, nC=NPh, LogP, G(N..O), SP10, GATS3P, Mor32v, Mor13u, Mor09u
11	11	0.843	0.711	0.663	n=CHR , RDF125e, nC=NPh, LogP, G(N..O), SP10, GATS3P, Mor32v, Mor13u, Mor09u, BIC1
12	12	0.848	0.718	0.667	n=CHR, RDF125e, nC=NPh, LogP, G(N..O), SP10, GATS3P, Mor32v, Mor13u, Mor09u, BIC1, RDF150v
13	13	0.862	0.744	0.693	n=CHR ,RDF125e,nC=NPh, LogP, G(N..O), SP10, GATS3P, Mor32v, Mor13u, Mor09u, RDF150v, GATS4e, D/D
14	14	0.869	0.755	0.702	n=CHR , RDF125e, nC=NPh, LogP, G(N..O), SP10, GATS3P, Mor32v, Mor13u, Mor09u, RDF150v, GATS4e, D/D, G1e
15	15	0.875	0.766	0.710	n=CHR, RDF125e, nC=NPh, LogP, G(N..O), SP10, GATS3P, Mor32v, Mor13u, Mor09u, RDF150v, GATS4e, D/D, G1e,G1m
16	16	0.883	0.780	0.723	n=CHR , RDF125e, nC=NPh, LogP, G(N..O), SP10, GATS3P, Mor32v, Mor13u, Mor09u, RDF150v, GATS4e, D/D, G1e, G1m, BELm4
17	17	0.889	0.791	0.733	n=CHR, RDF125e, nC=NPh, LogP, G(N..O), SP10, GATS3P,Mor32v, Mor13u, Mor09u, RDF150v, GATS4e, D/D, G1e, G1m, BELm4, C025

The equation below represents the equation of the best MLR model number 17
The equation:
IC_{50} = -3.044(±21.445) + 4.738 (±1.132) n=CHR + 0.455(±0.120) RDF125e + 3.474(±1.197) nC=NPh +1.150 (±0.320) Log P +0.058 (±0.014) G(N..O) - 1.151 (±0.542) SP10 – 3.205(±1.995) GATS3P +4.596 (±2.702) Mor32v – 2.249(±0.706) Mor13u +0.926(±0.592) Mor09u +2.435(±0.777) RDF150v- 4.464(±1.066) GATS4e – 0.029 (±0.016) D/D -174.925 (±75.241) G1e +306.561(±110.090) G1m +11.006(±4.564) BELm4 + 1.502(±0.847) C025
Where R= 0.889, R^2= 0.791, R^2adj.= 0.733 for the best model 17, and the descriptors are mentioned with a brief description in the Table 3 below;

Table 3: Brief description of the descriptors for the best MLR model 17

Descriptor	Description	Descriptor group
n=CHR	number of secondary C(sp2)	Functional group count
RDF125e	Radial Distribution Function - 125 / weighted by Sanderson electronegativity	RDF descriptors
nC=NPh	number of immines (aromatic)	Functional group count
Log P	describes lipophilicity for neutral compounds	Quantum chemical
G(N..O)	sum of geometrical distances between N..O	3D Atom Pairs
SP10	shape profile no. 10	Randic molecular profiles
GATS3P	Geary autocorrelation of lag 3 weighted by polarizability	2D autocorrelations
Mor32v	signal 32 / weighted by van der Waals volume	3D-MoRSE descriptors
Mor13u	signal 13 / unweighted	3D-MoRSE descriptors
Mor09u	signal 09 / unweighted	3D-MoRSE descriptors
RDF150v	Radial Distribution Function - 150 / weighted by van der Waals volume	RDF descriptors
GATS4e	Geary autocorrelation of lag 4 weighted by Sanderson electronegativity	2D autocorrelations
D/D	Wiener-like index from distance/detour matrix	2D matrix-based descriptors
G1e	1st component symmetry directional WHIM index / weighted by Sanderson electronegativity	WHIM descriptors
G1m	1st component symmetry directional WHIM index / weighted by mass	WHIM descriptors
BELm4	lowest eigenvalue n. 4 of Burden matrix / weighted by atomic masses	BCUT descriptors
C025	R--CR—R	Atom-centred fragments

According to the equation that mentioned previously we notice that a group of descriptors that have a positive effect on the compound activity are;
n=CHR, RDF125e, nC=NPh, Log P, G(N..O), Mor32v, Mor09u, RDF150v, G1m, BELm4, C025.
While the following descriptors have a negative effect on the compound activity;
GATS3P, SP10, Mor13u, GATS4e, D/D , G1e.

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

Then, leave one out (LOO) and leave many out (LMO) cross validation was performed on models **6-17** since these models have coefficients of determination larger than 0.6 [26]. The results of cross validation LOO and LMO are summarized in table 4 and 5 respectively.

Table 4: Leave one out cross validation results

Model	No. desc.	PRESS	SPRESS	SST	R^2cv	PRESS/SST	PSE	RSEP
6	6	376.6555	2.2872	618.8843	0.3914	0.6086	2.1835	38.6926
7	7	357.7575	2.2447	635.9207	0.4374	0.5626	2.1280	37.7094
8	8	329.2860	2.1689	664.3843	0.5044	0.4956	2.0416	36.1778
9	9	309.6216	2.1183	683.8274	0.5472	0.4528	1.9797	35.0809
10	10	297.0986	2.0902	695.4475	0.5728	0.4272	1.9393	34.3642
11	11	287.5422	2.0716	705.9785	0.5927	0.4073	1.9078	33.8070
12	12	279.6097	2.0583	711.3247	0.6069	0.3931	1.8813	33.3374
13	13	255.7102	1.9834	738.9271	0.6539	0.3461	1.7991	31.8808
14	14	243.1008	1.9490	750.6322	0.6761	0.3239	1.7542	31.0848
15	15	232.7692	1.9222	760.9981	0.6941	0.3059	1.7165	30.4171
16	16	218.4760	1.8772	775.2674	0.7182	0.2818	1.6630	29.4685
17	17	207.7593	1.8455	786.0047	0.7357	0.2643	1.6217	28.7366

Table 5: Leave many out cross validation results

Model	No. desc.	PRESS	SPRESS	SST	R^2cv	PRESS/SST	PSE	RSEP
6	6	425.3771	2.4306	588.1381	0.2767	0.7233	2.3205	41.1498
7	7	347.1467	2.2956	579.5446	0.3544	0.6456	2.1762	38.5924
8	8	362.2698	2.2749	622.2931	0.4178	0.5822	2.1414	37.9749
9	9	312.5505	2.1283	667.7622	0.5319	0.4681	1.9891	35.2728
10	10	318.1504	2.1630	687.5257	0.5373	0.4627	2.0068	35.5874
11	11	305.8569	2.1366	722.4868	0.5767	0.4233	1.9676	34.8931
12	12	310.4324	2.1688	733.8627	0.5770	0.4230	1.9823	35.1531
13	13	297.0992	2.1379	835.1887	0.6443	0.3557	1.9393	34.3899
14	14	305.6703	2.1854	885.4563	0.6548	0.3452	1.9670	34.8825
15	15	300.6718	2.1846	872.6149	0.6554	0.3446	1.9509	34.5961
16	16	305.1025	2.2183	904.7566	0.6628	0.3372	1.9652	34.8500
17	17	301.3404	2.2226	985.5938	0.6943	0.3057	1.9531	34.6345

Where: PRESS (Predictive residual sum of squares), it's a standard index to measure the accuracy of the model, SST (Total sum of squares), R2CV (Cross validated correlation coefficient), SPRESS (Uncertainty of prediction), PSE (Predictive square error) and RSEP (Relative standard error of prediction).

According to the values in Tables 4 and 5 we note that the models 13-17 have a good predictive power, because these models having high R2CV and PRESS/SST less than 0.4. So these models were chosen for artificial neural network analysis (ANN).

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

PCA

The Principle Component Analysis (PCA) was performed to divide the data set or the molecules group into training, validation and test set. The PCA was performed on the 79 compounds, 17 descriptors and we plot the first and second principles, first and third principles and second and third principles. So we divide the data into 60% training set, 20% test set and 20% validation set by choosing one molecule from each zone to each set.

The first and second principles plot have the best data distribution in comparison with the first and third principles, and second and third principles which they have a condensed data plot.
So that and relying on the first and second principles plot (Figure. 1), we exclude compounds 60, 63 and 41 as outliers from the data analysis, because they seem to behave in a different way in comparison to the other compounds. So, the division of the data become as following; 60% (46 compounds) training set, 20% (15 compounds) for each validation and test set

ANN

In this study, a three-layered feed-forward ANN model with back propagation learning algorithm [27] was employed. The first ANN was performed on the models chose (13-17) from the cross validation (LOO and LMO). The ANN was done for each model with hidden node 7. The results shows that the model number 16 have the highest correlation coefficient R for the test set which equal 0.854211 so this indicates that the model 16 have a high predictive power, also models number 14 and 17 have a good predictive power.

In (Figure. 2) which shows the relation of the PRESS values for training, validation and test sets versus the model number. The figure shows that the minimum PRESS value of the training set obtained for model 13 and the model after is 17. While the minimum PRESS value of the test set was obtained for models 16, 14 and 17 respectively.

In (Figure. 3) which shows the relation of correlation coefficient (R) values for training, validation and test sets versus the model number. the figure shows that the highest R value of the training set was obtained for models 13 and 17. While the highest R value of the test set was obtained for models 16, 17 and 14 respectively.

And finally, in (Figure. 4) which shows the relation of cross validated correlation coefficient (R2CV) values for training, validation and test sets versus the model number. the figure shows that the highest (R2CV) value of the training set was obtained for models 17 and 13. While the highest (R2CV) value of the test set was obtained for models 17 and 14.

According to the previous notes, models 14,16 and 17 are subjected for further analysis by optimizing the number of hidden nodes, because these models have the highest R, R2CV values, and lowest PRESS values for test set.

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

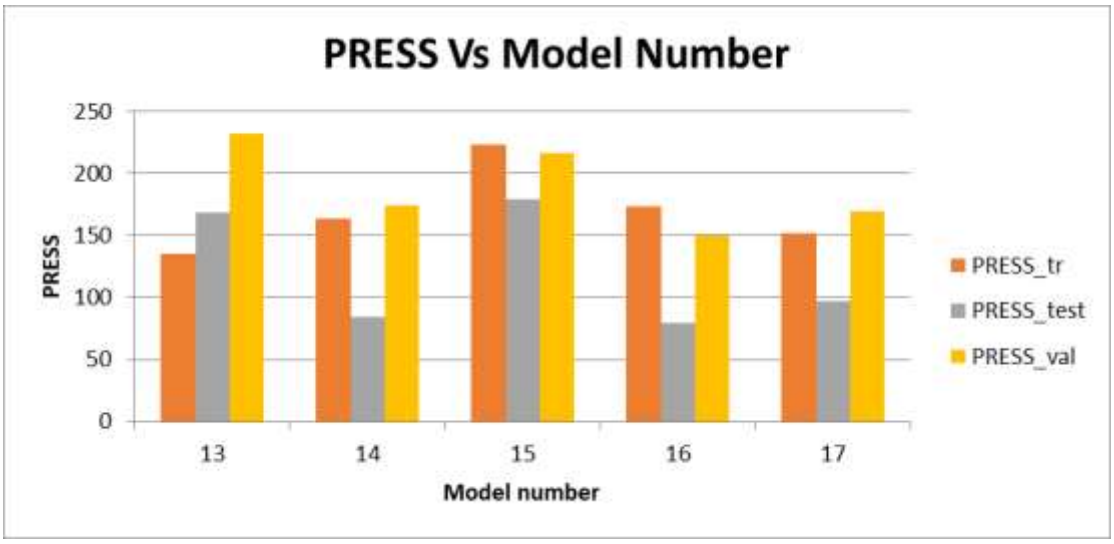


Figure 2: Plots of ANN Predictive Residual Sum of Squares(PRESS) values for the training, test and validation sets versus model number.

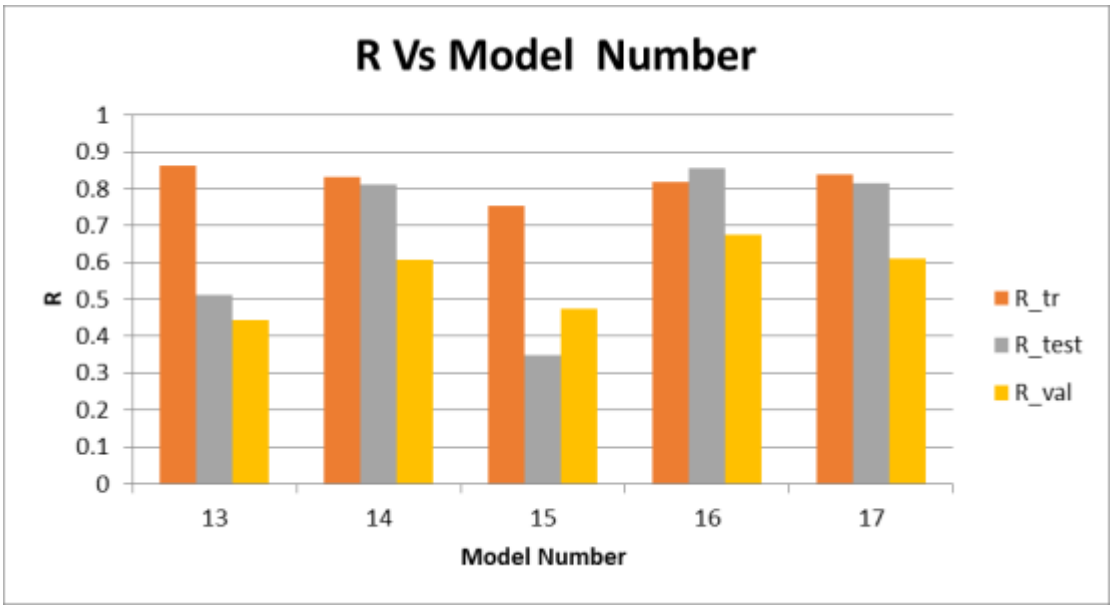


Figure 3: Plots of ANN correlation coefficient (R) values for the training, test and validation sets versus model number.

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index

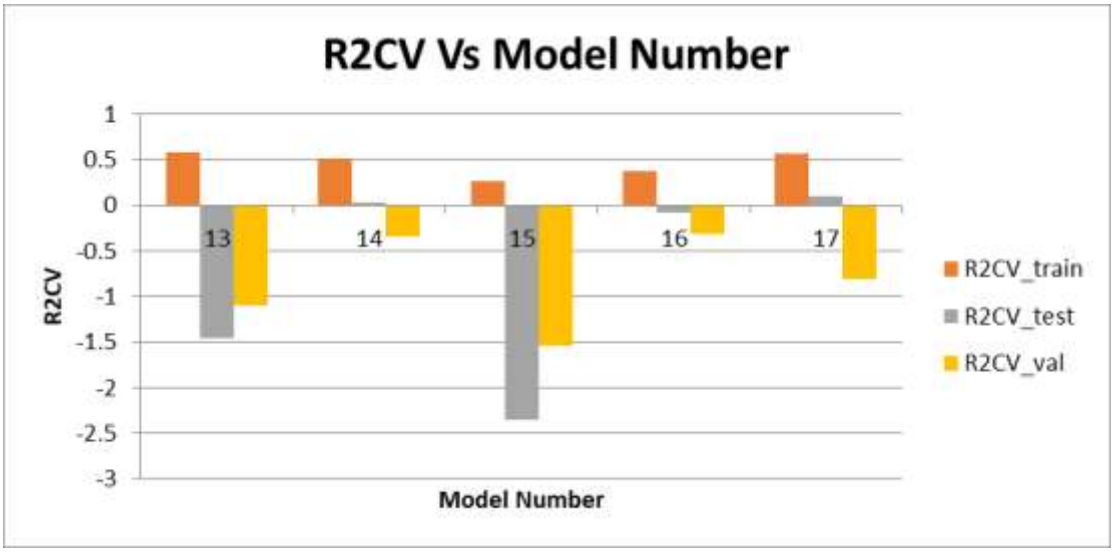


Figure 4: Plots of ANN cross validated correlation coefficient (R2CV) values for the training, test and validation sets versus model number.

The second ANN was performed on the chose models 14, 16, 17 which have the highest correlation coefficient for test set (R-test). The ANN performed with different hidden nodes from 5 to 20. From the results, the best model with the optimal hidden nodes were as follows; model 14 with hidden node 7, model 16 with hidden nodes 7 and 10, and model 17 with hidden nodes 5 and 9, these are chose because they a high prediction power (R), minimum PRESS value of the test set and minimum number of hidden nodes.

The best of models that mentioned above are summarized in Table 6 with their parameters and correlation coefficients. From the table we choose models 14 hn 7, 16 hn 7, and 17 hn 5 to continue with the randomization test (chance correlation test)

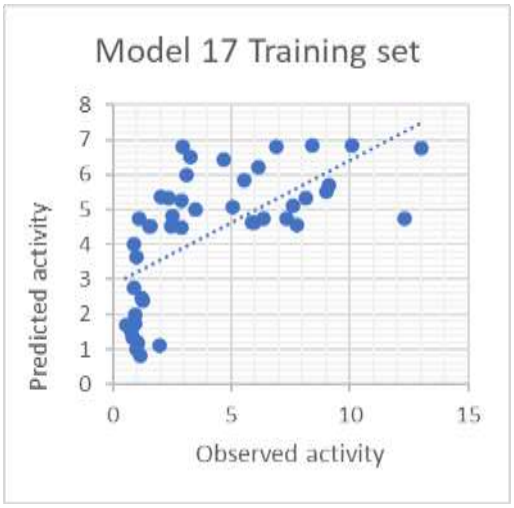
Table 6: Summary of correlation coefficient and cross validation parameters of the optimal number of hidden nodes for each model

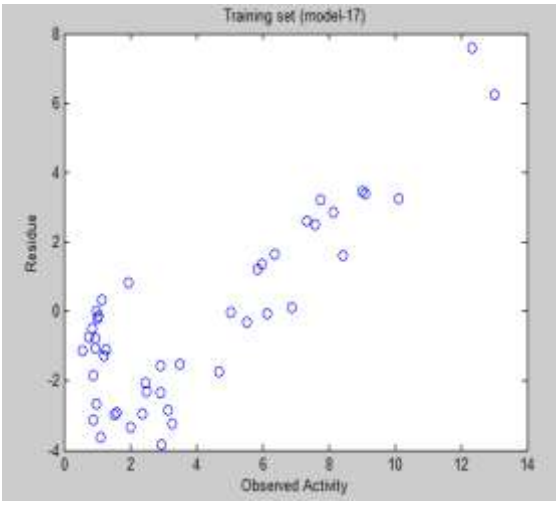
Mo.#	14	16	16	17	17
hn	7	7	10	5	9
nPCs	6	6	6	5	5
R_tr	0.83288	0.81804	0.84088	0.64577	0.65330
PRESS_tr	163.8919	173.3804	151.5737	305.6745	316.0515
R2CV_tr	0.5192	0.37357	0.55335	-0.94013	0.16659
RSEP_tr	35.8703	36.8941	34.4960	48.9876	49.8122

R_test	0.81114	0.85421	0.84024	0.80426	0.72857
PRESS_test	84.0964	79.2825	69.2688	83.2702	85.5780
R2CV_test	0.0316	-0.07786	0.19661	-0.73831	0.13862
RSEP_test	39.5552	38.4064	35.8992	39.3605	39.9022
R_val	0.60846	0.67386	0.70119	0.68722	0.59802
PRESS_val	173.9223	150.0469	147.6169	156.4339	172.8524
R2CV_val	-0.3412	-0.31686	-0.02047	-2.4556	-0.74573
RSEP_val	52.1545	48.4426	48.0488	49.4629	51.9938

Randomization test for the resulted ANN models, its performed for the models as validation test to ensure that the ANN analysis for the resulted models is not by chance. From the results we noted that the correlation coefficients that resulted from chance correlation have low values in general, but the PRESS values are high. This indicates that the models 14, 16 and 17 which resulted from PC_ANN are better than the resulted by randomization, so the results of ANN are not by chance.

Figure 5 shows regressions between observed and predicted activity as well as their residuals for the training, validation, and test sets for model 17.





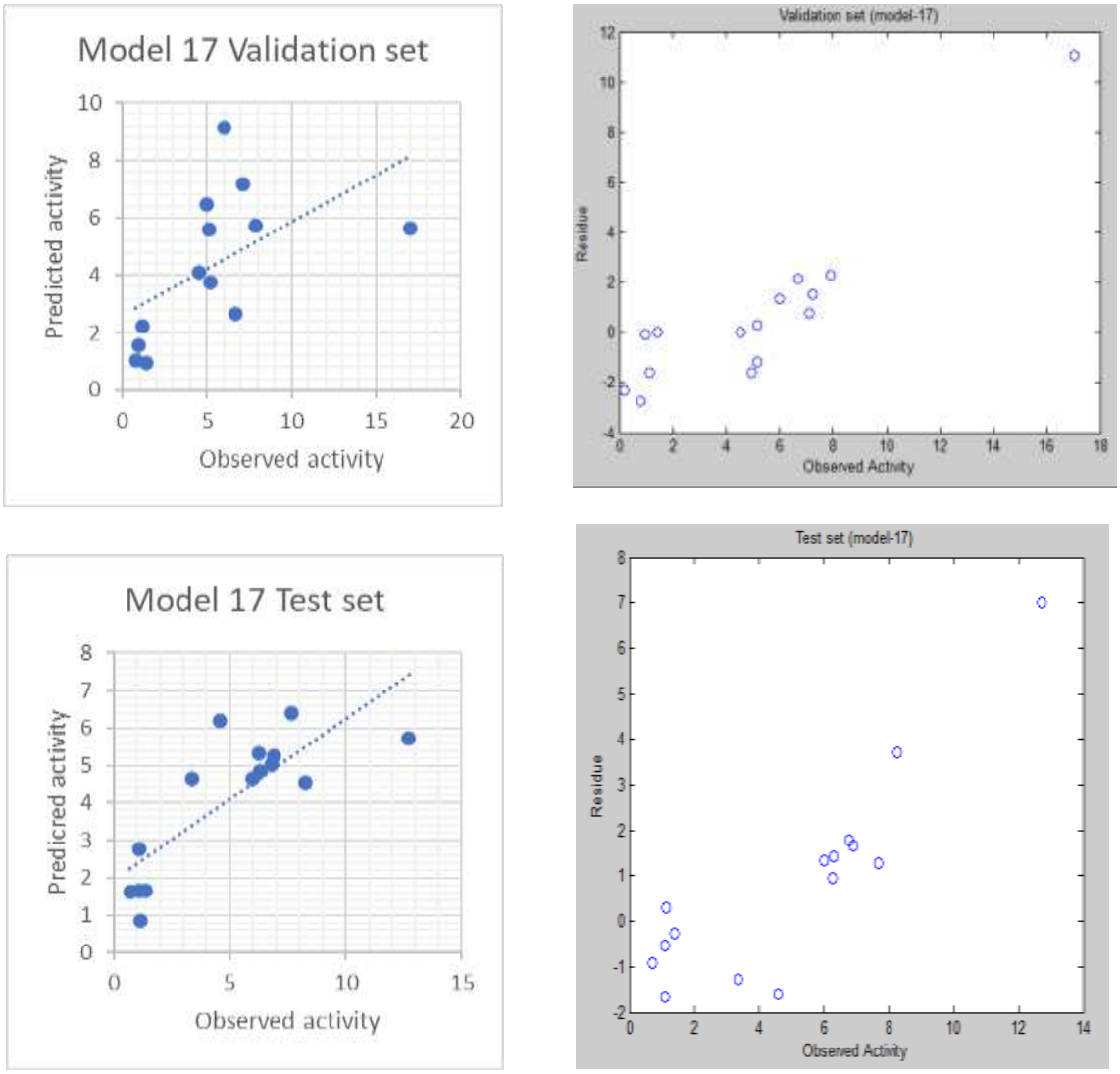


Figure 5: Plot of predicted activity against observed activity as well as their residual for model 17 using 5 hidden nodes. Training, validation, and test set.

Comparison with other QSAR studies

There are many Quantitative structure activity relationship (QSAR) studies about antimalarial compounds that are done by researchers, they take a small group of compounds to build the model, and they use different techniques to get the results. But in this study, we collect 79 compounds which is a large group from different references in comparison with the other studies to make QSAR model for them by MLR and PC-ANN. So, our study will give a model with a high predictive power in comparison with the previous studies.

A study done for compounds 28-41 in Table 1, this a 3D-QSAR study by Sharma and Patil, they found a model showed that steric (S_584), and electrostatic (E-295) interactions play important role in determining DPP IV inhibitory activity [28]

Also a 3D QSAR analyses of antimalarial alkoxylated and hydroxylated chalcones were first conducted by Comparative molecular field analysis (CoMFA) and Comparative similarity indices analysis (CoMSIA). Satisfactory results were obtained after performing a leave-one-out (LOO) cross-validation study with cross-validation *q*2 and conventional *r*2 values of 0.740 and 0.972 by the CoMFA model, 0.714 and 0.976 by the CoMSIA model, respectively [29].

The disadvantages of the previous studies that are study the QSAR for a small group of compounds, but in this study we collect a large group of compounds from different papers to get a better model for designing a new antimalarial agent. Also in the current study we calculated all descriptors for all compounds to build a MLR model. And in this study we use also the PC-ANN as a nonlinear to get more powerful model and good prediction power.

CONCLUSIONS

A quantitative structure activity relationship analysis of 79 antimalarial compounds that are collected from literature and their inhibition activities were calculated experimentally, was performed using the multiple linear regression (MLR) and principle component-artificial neural network (PC-ANN) methods. The cross validation and y-randomization methods were used to verifying the resulted best models.

The results obtained from the MLR were a group of models which have a good predictive power (R^2 > 0.6), the best model was model number 17. Model 17 with 17 descriptors, and the results: R= 0.889, R^2= 0.791, and R^2adj.= 0.733.

The cross validation methods (LOO and LMO) were performed on the resulted MLR models, models (13-17) showed a good predictive power because of having high values of R^2cv and PRESS/SST less than 0.4. so that these models are chose to complete with the ANN analysis.

The Principle component analysis (PCA) was performed to divide the data (79 compound) into three sets (validation, training and test set), then ANN was performed on the best resulted models (13-17) from LOO and LMO validation methods.

The ANN results shows that model 16 have the highest correlation coefficient for test set (0.8542119) which indicates that it has a high predictive power. Also models 14 and 17 have good predictive power. So that models (14,16,17) chose to continue with ANN to find the optimal number of hidden nodes for each one of these models.

From the final result of ANN, model 14 with hidden node 7, model 16 with hidden nodes 7 and 10, and model 17 with hidden nodes 5 and 9 were chose as the best models with the optimal hidden nodes due to the high predictive power (R), minimum number of hidden nodes and minimum PRESS value for the test set.

Then the ANN results were validated by randomization test (chance correlation). Golbraikh and Tropsha proposed conditions were applied to conclude that the QSAR models have acceptable prediction power or not. The best ANN model with the best predictive power was model number 17.

REFERENCES

[1] R.W. Snow, C.A. Guerra, A.M. Noor, H.Y. Myint, S.I. Hay, Nature 434 214(2005).
[2] R.S. Phillips, Clin. Microbiol. Rev. 14 208;(2001).
[3] World Health Organization. World malaria report. Geneva, Switzerland: World Health Organization. Available from http://www.who.int/ith/diseases/malaria/en/:report-2018/en/. [1 April 2018].
[4] B.M. Greenwood, D.A. Fidock, D.E. Kyle, S.H. Kappe, P.L. Alonso, F.H. Collins, et al. Malaria: progress, perils, and prospects for eradication. J Clin Invest 118:1266-76(2008).
[5] A. Trampuz, M. Jereb, I. Muzlovic, R.M. Prabhu. Clinical review: severe malaria. Crit Care7:315-23; (2003).
[6] D.A. Fidock, P.J. Rosenthal, S.L. Croft, R. Brun, S. Nwaka. Antimalarial drug discovery: efficacy models for compound screening. Nat Rev Drug Discovery3:509-20;(2004).
[7] J. May, C.G. Meyer, Trends Parasitol. 19 432(2003).
[8] M. Foley, L. Tilley, Quinoline antimalarials: Mechanisms of action and resistance. Int. J. Parasitol. 27:231-240(1997).
[9] S. Foote, A. Cowman, The mode of action and the mechanism of resistance to antimalarial drugs. Acta Trop. 56:157-171(1994).
[10] W. Peters, Drug resistance in malaria parasites of animals and man. Adv. Parasitol. 41: 1-62(1997).
[11] T. Geary, L.Bonanni , J. Jensen, H. Ginsburg, Effects of combinations of quinoline-containing antimalarials on Plasmodium falciparum in culture. Ann. Trop. Med. Parasitol.
[12] N.Yadav, S. K. Dixit, A. Bhattacharya, L. C. Mishra, M. Sharma, S. K. Awasthi and V. K. Bhasin, Antimalarial Activity of Newly Synthesized Chalcone Derivatives In Vitro, Chem Biol Drug Des 80: 340–347(2012).
[13] N. Mishra, P. Arora, B. Kumar, L. C. Mishra, A. Bhattacharya, S. K. Awasthi V. K. Bhasin ; Synthesis of novel substituted 1,3-diaryl propenone derivatives and their antimalarial activity in vitro;European Journal of Medicinal Chemistry 43 1530e1535(2008).
[14] S. K. Awasthi, N. Mishra, B. Kumar, M. Sharma, A. Bhattacharya, L. C. Mishra, V. K. Bhasin; Potent antimalarial activity of newly synthesized substituted chalcone analogs in vitro; Med Chem Res 18:407–420(2009).
[15] P. C. Sharma, S. Padwal, K. K.Bansal, A. Saini; Synthesis, characterization 1 of benzimidazole clubbed benzothiazole derivatives Chem. Biol. Lett. 4(2), 63-68(2017).
[16] A. Bhatt, R. Kant, R.K. Singh; Synthesis of Some Bioactive Sulfonamide and Amide Derivatives of Piperazine Incorporating Imidazo[1,2-B] Pyridazine Moiety. Med chem (Los Angeles) 6: 257-263 (2016).
[17] M. Zavri, N. Kawthekar; International Journal of Current Pharmaceutical Research ISSN- 0975-7066 Vol 9, Issue 3 (2017).
[18] O. Deeb, B. Hemmateenejad. "ANN-QSAR model of drug-binding to human serum albumin", Chem. Biol. Drug Des. (2007), **70**: 19–29.
[19] B. Hemmateenejad, M. A.. Safarpour, R. Miri, N. Nesari, " Toward an optimal procedure for PC-ANN model building: prediction of the carcinogenic activity of a large set of drugs", J. Chem. Inf. Model. (2005), **45**: 190–199.

[20] Omar Deeb, Manal Muhtaseb, Basheerulla Shaik (2024), "Exploring QSARs for inhibiting activity of a set of EGFR tyrosine kinase inhibitors by GA-MLR and molecular Docking simulations, BJMAS- British Journal of Multidisciplinary and Advanced Studies: Health and Medical Sciences, 2024, **5 (2**), 12-40.

[21] O. Deeb and M. Drabh, "Exploring QSARs of Some Analgesic Compounds by PC-ANN", Chem Biol Drug Des; (2010), **76**: 255–262.

[22] P. V. Khadikar, O. Deeb, A. Jaber, J. Singh, V. K. Agrawal, S. Singh and M. Lakhwani. "Development of Quantitative Structure-Activity Relationship for a set of Carbonic Anhydrase Inhibitors: Use of Quantum and Chemical Descriptors". Letters in Drug Design & Discovery; 2006, **3**(9): 622-635

[23] O. Deeb, B. Hemmateenejad , A. Jaber, R. Garduno-Juarez and R. Miri. "Effect of the electronic and physicochemical parameters on the carcinogenesis activity of some sulfa drugs using QSAR analysis based on genetic-MLR and genetic PLS". Chemosphere (2007), **67**(11): 2122-2130

[24] O. Deeb, K. M. Youssef and B. Hemmateenejad, "QSAR of Novel Hydroxyphenylureas as Antioxidant Agents". QSAR and Combinatorial Sciences; 2008, **27**(4): 417-424.

[25] O. Deeb, "Correlation ranking and stepwise regression procedures in PC-ANN modeling and application to predict the toxic activity and HSA binding affinity". Chemometrics and Intelegent Laboratory Systems.; 2011, **104**: 181-194.

[26] A. Golbraikh, A. Tropsha. "Beware of q2!". J Mol Graph Model; 2002, **20**: 269–276.

[27] D. E. Rumelhart, G. E. Hinton, R. J. Williams. "Learning representations by back-propagating errors". Nature; 1986, **323**: 33–536.

[28] R. Sharma and S. Patil, three dimensional quantitative structure analysis substituted 1,3-diaryl propenone derivatives as antimalarial activity, Der Pharma Chemica, 5(4):80-86 (2013).

[29] C.X. Xue a, S.Y. Cui a, M.C. Liu a, Z.D. Hu a, B.T. Fan, QSAR studies on antimalarial alkoxylated and hydroxylated chalcones by CoMFA and CoMSIA; European Journal of Medicinal Chemistry 39 745–753 3D (2004).