Deanship of Graduate Studies

Al-Quds University

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

Alaa Ishaq Hasan Ashour

M.Sc. Thesis

Jerusalem-Palestine

1440 / 2019

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

Prepared by:

Alaa Ishaq Hasan Ashour

B.Sc. Applied Chemistry Palestine Polytechnic Universtiy

Supervisor: Professor Omar Deeb

Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Industrial Applied Technology at Al-Quds University

1440 / 2019

Al-Quds University

Deanship of Graduate Studies

Industrial Applied Technology Program

Faculty of Science and Technology

Thesis Approval

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

Prepared by : Alaa Ishaq Hasan Ashour

Registration No.: 21412335

Supervisor: Prof. Omar Deeb

Master thesis Submitted and accepted Date: 2019/5/11

The names and signatures of the examining committee members are as follows:

Head of committee: Prof. Omar Deeb Signature:.....

Internal Examiner: Dr. Sawsan Salamah Signature:.....

External Examiner: Dr. Hatem Hejaz Signature:.....

Jerusalem- Palestine

1440/2019

Dedication

I would like to dedicate this work to my family and my parents, for anyone who support me during the thesis work and challenges, to my teachers, and to my homeland Palestine.

Declaration

I certify that this thesis submitted for the degree of Master, is the result of my own research, except where otherwise acknowledged, and that this study (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Signed: .....

Alaa Ishaq Hasan Ashour

Date: 11 /5 /2019

Acknowledgment

At first, I would like to thanks the God who brought me to this stage after a period of working and facing challenges.

And I was grateful for my supervisor Professor Omar Deeb for his continuous support and help during each stage of the research stages. He didn't hesitate any time to apply the advices and guidance for me. I would like to thank him greatly. And a special thanks for the dean Dr. Mohammad Abu Alhaj for his support also and any one supported me to complete my research.

And finally, all thanks for my second house Al-Quds University.

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 agents, 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. Also the principle component –artificial neural network (PC-ANN) was used as nonlinear method for modeling. The models resulted have a good prediction power. The MLR resulted with 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.

The cross validation LOO and LMO were performed on the resulted MLR models, the models 13 - 17 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 choosed 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 test value 0.8138. A new suggested compounds with

IC_{50} 7.057 and 3.336 µg/ml. From the above result, now it's possible to design a new potent antimalarial drug by the application of the best model equation of MLR.

Table of content

List of Tables

List of Figures

List of abbreviations

QSAR: Quantitative Structure Activity Relationship

MLR: Multiple Linear Regression

PCA: Principle Component Analysis

ANN: Artificial Neural Network

LOO: Leave One Out

LMO: Leave Many Out

IC_{50}: Minimum Inhibition Concentration

AM1: Austin Model 1

Hn: Hidden Nodes

HOMO: Highest Occupied Molecular Orbital

LUMO: Lowest Unoccupied Molecular Orbital

R: Correlation Coefficient

R^2: Coefficient Determination

R^2 adj.: Adjusted R^2

RSEP: Relative Standard Error of Prediction

CADD: Computer Aided-Drug Design

CAMD: Computer Aided Molecular Design

F: Fisher Ratio Value

RMSE: Root Mean Square Error

CV: Cross Validation

CoMFA: Comparative Molecular Field Analysis

CoMSIA: Comparative Similarity Indices Analysis

WHIM: Weighted Holistic Invariant Molecular descriptors)

PSE: Predictive square Error

Es: Steric Substituent constant

RDF: Radial Distribution Function descriptors

QSPR: Quantitative Structure-Property Relationship

SPSS: Statistical Package for the Social Sciences

MTD: Minimum Topological Difference

HQSAR: Holographic Quantitative Structure Activity Relationship

PLS: Partial Least Square

PC1: primary Principal Component

SST: Total Sum of Squares

FP: Fe(II)-Protoporphyrin IX

R2CV: Cross Validated Correlation Coefficient

R-val: Correlation Coefficient of validation set

R-tr: Correlation Coefficient of training set

Chapter One

Introduction

Chapter One

Introduction:

1.1 Antimalrial agent preview

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 1e2.7 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 anaemia [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 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's and 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.

1.2 Overview of Computational Chemistry

One of the major sciences that taking a great place recently, in the research development and drug design improvement by using a set of software programs is the computational chemistry.

Two ways have been approached the chemistry problems which are; computational quantum chemistry and non-computational quantum chemistry [12].

Computational theoretical chemistry is mainly concerned with the numerical computation of the molecular electronic structure and molecular interaction and Non-computational

quantum chemistry that deals with formulation of analytical expressions for the properties of the molecules and their reactions [12].

Computational chemistry (also called molecular modeling; two terms with the same meaning) is a set of techniques for investigating chemical problems on a computer [13].

The term **computational chemistry** is usually used when a mathematical method is sufficiently well developed that it can be automated for implementation on a computer. So, the computational chemistry is the application of chemical, mathematical and computing skills to find and choose the solution of interesting chemical problems [12].

Computational chemistry has become a useful way to investigate materials that are too difficult to find or too expensive to purchase. It also helps chemists to make predictions before running or implantation the actual experiments so that they can be better prepared for making observations and predictions [13].

This branch of chemistry can generates data which complements experimental data on the structures, properties and reactions of substances. The calculations are based primarily on Schroedinger's equation [14] and include:

1. Calculation of electron and charge distributions
2. Molecular geometry in ground and excited states
3. Potential energy surfaces
4. Rate constants for elementary reactions
5. Details of the dynamics of molecular collisions.

Computational chemistry methods encompass a variety of mathematical methods which summarized into:

Molecular mechanics: which applies the laws of classical physics to molecular nuclei without explicit consideration of electrons.

Quantum mechanics: which relies on the Schrödinger equation to describe a molecule with explicit treatment of electronic structure [14].

The quantum mechanical methods can be subdivided into two classes:

- Ab-initio, (Latin for "from scratch") a group of methods in which molecular structures can be calculated using nothing but the Schrödinger equation, the values of the fundamental constants and the atomic numbers of the atoms present [14].

- Semi-empirical techniques use approximations from empirical (experimental) data to provide the input into the mathematical models, which is followed in this study [14].

In theoretical chemistry, chemists, physicists, and mathematicians develop algorithms and computer programs to predict atomic and molecular properties and reaction paths for chemical reactions. In contrast, computational chemists may apply existing computer programs and methodologies in a simple way to answer and explain specific chemical questions.

Several major areas may be distinguished within computational chemistry:

- The prediction of molecular structure of molecules by the use of the simulation of forces, or more accurate quantum chemical methods, to find stationary points on the energy surface as the position of the nuclei is varied.
- Storing and searching for data on chemical entities.

- Identifying correlations between chemical structures and properties (quantitative structure–property relationship (QSPR) and quantitative structure–activity relationship (QSAR)).
- Computational approaches to help in the efficient synthesis of compounds.
- Computational approaches to design molecules that interact in specific ways with other molecules (e.g. drug design and catalysis) [15].

The design of new drugs with improved properties and diminished side-effects for treating human diseases such as malaria and others became one of the most important challenges and problems that face the medicinal chemists in these days. Medicinal chemists begin the process by taking a lead compound and then finding analogs which have the preferred biological activities. Next, they used their experiences and chemical insight to eventually choose a nominee analog for further development. This process is difficult, expensive and took a long time. The conventional ways of drug discovery are now being supplemented by shortest approaches made possible by the accepting of molecular processes involved in the original disease. In this view, the preliminary point in drug design is the molecular target which is receptor or enzyme in the body as an option of the existence of known lead structure [16].

The effective design of chemical structures with the required therapeutic properties is directed towards computer aided-drug design (CADD) a well established area of computer aided molecular design (CAMD). These techniques cover a new methodologies, such as molecular modeling and quantitative structure-activity relationships (QSAR). Molecular modeling can be simply used to predict molecular and biological properties [16]. QSAR mainly suppose that the chemical structure of a compound is mainly affected its biological activity. Within the QSAR approach, the descriptor variable are not physically

measured but computed, therefore, they are easy and cheap to generate even for large molecular sets.

1.3 Quantitative structure activity relationships (QSAR)

QSAR is a way to find a simple equation that can be used to calculate some property from the molecular structure of a compound by using a set of software programs. QSAR attempt to correlate structural molecular features (descriptors) with physicochemical properties such as biological activities for a set of compounds, by means of statistical methods. As a result, a simple mathematical relationship is established [16].

1.3.1 QSAR History

More than a century ago, Crum-Brown and Fraser expressed the idea that the physiological action of a substance was a function of its chemical composition and constitution [17]. A few decades later, in 1893, Richet showed that the cytotoxicities of a diverse set of simple organic molecules were inversely related to their corresponding water solubilities [18]. At the turn of the 20th century, Meyer and Overton independently supposed that the narcotic (depressant) action of a group of organic compounds paralleled their olive oil/water partition coefficients [19,20]. In 1939 Ferguson introduced a thermodynamic generalization to the correlation of depressant action with the relative saturation of volatile compounds in the vehicle in which they were administered [21]. The extensive work of Albert, Bell and Roblin established the importance of ionization of bases and weak acids in bacteriostatic activity [22–24]. Taft devised a way for separating polar, steric, and resonance effects and introducing the first steric parameter, ES [25]. The contributions of Hammett and Taft together laid the mechanistic basis for the development of the QSAR paradigm by Hansch and Fujita. In 1962 Hansch and Muir published their brilliant study on the structure-activity relationships of plant growth regulators and their dependency on

Hammett constants and hydrophobicity [26]. The Kubinyi bilinear model is a refinement of the parabolic model and, in many cases, it has proved to be superior [27] Variations on this activity-based approach have been extended by Klopman et al. [28] and Enslein et al. [29]. Topological methods have also been used to address the relationships between molecular structure and physical/biological activity. The minimum topological difference (MTD) method of Simon and the extensive studies on molecular connectivity by Kier and Hall have contributed to the development of quantitative structure property/activity relationships [30,31]. Other recent developments in QSAR include approaches such as HQSAR, Inverse QSAR, and Binary QSAR [32–35]. Improved statistical tools such as partial least square (PLS) can handle situations where the number of variables overwhelms the number of molecules in a data set, which may have collinear X-variables [36].

1.3.2 QSAR advantages and disadvantages

It's not necessary to have a complete understanding of the phenomenon being studied to use QSAR models; its sufficient to feed the model with representative training data of the profile being searched. They are used as a primary filter to shrink the number of the evaluated candidates; there design is not complex and there efficiency depend on the searched profile in the training data set.

On the other hand, as the QSAR have a benefits it have a limitation. Sometimes the QSAR don’t have a small number of candidates, therefore the efficiency is not frequently documented in details, to the detriment of their attributes, as the quality is not depend on the model, but on the profile being searched. Their main disadvantages is that always produce candidates because of these are models for approximation, therefore it advisable to index the candidates with a percentage of relative effectiveness [37].

1.4 QSAR model development steps

The QSAR model development process can be generally divided into three stages: data preparation, data analysis and model validation . These steps represent the main standard practice of any QSAR modeling and their implementations are often determined by the researcher's interests, experience and software availability.

1.4.1 Data preparation :

This step is mainly depend on the collecting and selecting of data that's described in literature. These data which composed of a chemical compounds and their activities were experimentally calculated in labs and the same way of determination was followed for the selected data set for QSAR study. Also it must be have the same activity unit. After that you need to find the geometry optimization of the compound to be in the stable form using a special software such as Hyperchem software which used to optimize the compounds in the 3D structure in our study.

1.4.2 Data analysis

To analyze the data you need a mathematical methods, the applications of these methods are combined with the main aim of explanation and prediction of non-synthesized test compounds. Many different statistical methods are available in the literature and the selection of the appropriate method is critical [38]. In our study we apply the Multiple Linear Regression (MLR) for linear correlation and the Principle component-Artificial Neural Network (PC-ANN) as nonlinear correlation.

1.4.2.1 Linear models

Multiple linear regression (MLR)

It is a mathematical technique to study the relation between one dependent variable and several independent variables. the regression method is based on three criteria: correlation of determination (R^2), the Fisher ratio value (F), and the root mean square error (RMSE).

The regression model assumes a linear relationship between *m* molecular descriptors and the response (biological activity) variable. This relationship can be expressed with the single multiple-term linear equation:

$$\mathbf{Y = b0 + b1X1 + b2X2 + \ldots. + bmXm + e}$$

The MLR analysis calculates the regression coefficients, b1, by minimizing the residuals, **e**, which quantify the deviations between the data (Y) and the model (Y'), as in the case of simple linear regression [39]. Its assumed that a linear relationship between the data so it's the first step of analysis applied.

1.4.2.2 Non linear models

- **Principle component analysis (PCA)**

This technique seeks to get a new set of variables named Principal Components (PC) showing the data in order to decrease the variance with the aim to state the main information in the variables by the principal components of X**.** The primary Principal Component (PC1) describes the maximum deviation in the whole data set. The subsequent principal component (PC2) describes the maximum remaining variance, and so forth, with each axis linearly independent, to the preceding axis. Some of the last components may be discarded to shrink the size of the model and stay far away from over-fitting [40].

- **Artificial neural network.**

An Artificial Neural Network (ANN) is a mathematical model that tries to simulate the structure and functionalities of biological neural networks. Artificial neuron is sorted as the basic building block of every artificial neural network, that is a simple mathematical model (function). Its consists of a number of "neurons" or "hidden units" that receive data from outside, process the data, and output a signal. A "neuron" is essentially a regression equation with a non-linear output. A non-linear models produced when more than one of the neurons is used. These networks have been shown to work well for modeling a number of different problems, including QSAR [41].

In order to get an artificial neural network, you need to combine two or more artificial neurons. Its proved that the artificial neural networks have the ability to solve the real-life problems, if the single artificial neuron isn't useful. In fact artificial neural networks are capable of solving complex real-life problems by processing information in their basic building blocks (artificial neurons) in a non-linear, distributed, parallel and local way.

The topology, architecture or graph of an artificial neural network is defined as the way in which the individual artificial neurons are interconnected.

The learning process consists of adjusting the coefficient so that the network provided as an output the suitable results. In neural networks, a training set is used to train the network, and then the network is used to predict the property (biological activity) that it was trained to predict. This technique can be associated with principal component analysis in which it is referred as **PC-ANN.**

1.4.3 Model validation

Validation is a one of the crucial aspect of any QSAR modeling. It is the process by which the reliability and relevance of a procedure are produced for a specific goal [42]. In the QSAR community, the validation of a model is little more than an assessment of statistical fit and, occasionally, predictivity using cross validation techniques. However, it is now being accepted that validation is a more important process that includes assessment of issues such as data quality, applicability of the model and mechanistic interpretability in addition to statistical assessment [43]. In principle, two reasonable approaches of validation can be envisaged, one based on prediction and the other based on the fit of the predictor variables to rearranged response variables.

1.4.3.1 Cross-validation

The cross-validation is a common method for internally validating a QSAR model (CV, Q^2, q^2, or jack-knifing). The process of CV repeats the regression many times on subsets of data. Usually each molecule is left out once (only), in turn, and *R* is computed using the predicted values of the missing molecule. Sometimes more than one molecule (leave many out, LMO) is left out at a time. CV is often used to determine how large a model can be used for a given data set. A cross-validated R^2 is usually smaller than the overall R^2 for a QSAR equation. It is used for the evaluation of the predictive power of an equation as a diagnostic tool.

CV used to measure a model's predictive ability and give attention to the possibility a model has been over fitted. Over-fitting related to the phenomenon in which a predictive

model may well describe the relationship between predictors and response, but may be after that fail to provide valid predictions for new compounds.

In the leave-one-out (LOO) method, a process is performed for the entire training set to removing a molecule, and creating and validating the model against the individual molecules. Once complete, the mean is taken of all the Q^2 values and reported. The data utilized in obtaining Q^2 is an augmented training set of the compounds (data points) used to determine R^2. The method of removing one molecule from the training set is considered to be an inconsistent method [44,45]**.** A more correct method is leave-many-out (LMO), where a group of compounds is selected for validation of the model. The most of researchers apply the LOO or LMO CV procedures in order to validate a QSAR model. The cross-validated correlation coefficient R^2 (Q^2) is an outcome from this procedure. Frequently, Q^2 is used as a criterion of both robustness and predictive ability of the model. A high values of Q^2 (for instance, $Q^2 > 0.5$) were considered by many authors as an indicator or even as the ultimate proof of the high predictive power of the QSAR model.

Randomization test (Scrambling model)

When the observations are not independent of each other sufficiently, so the predictive power of the equation become poor. One way to test for this is by randomization of the dependent variables. This procedure ensures that the model is not due to a chance. The set of activity values is reassigned randomly to different molecules, and repeating the entire modeling procedure. We repeated the process many times. If the random models activity prediction is comparable to the original equation, the set of observations is not sufficient to support the model.

The unique method of checking the descriptors is the creation of a Scrambled Model [46,47], it is a used in the model because the bioactivities are randomized ensuring the new

model is created from a bogus data set. The basis for this method is to test the validity of the original QSAR model and to ensure that the selected descriptors are appropriate. These new models (Scram-models) are created using the same descriptors as the original model, yet the bioactivities are changed. The methods mentioned previously are used to perform validation after the creation of each Scram-model. The process of changing the bioactivities can be repeated to ensure that the Scram-models are truly random, and as each new Scram-model is created its R^2 and Q^2 values. Each time the R^2 and Q^2 values of the Scram-models are substantially lower further enforces that the true QSAR model is found. The basis of using this method is to validate the original QSAR model because the Scram models are created using the original descriptors and bogus bioactivities. The model would be in question if there was a strong correlation ($R^2 > 0.50$) [48] between the randomized bioactivities and the predicted bioactivities, specifically that the model is not responsive to the bioactivities.

1.5 Software used in QSAR study

In order to carry the QSAR study, the software is the main component. Many software are common for that according to the techniques used. These are special programs to draw compounds and optimize them, and find linear or non linear relationship to get the best model.

In our study we used the following programs ;

HyperChem (version 7.0 Hyperchem, Inc.), Dragon software (version 2.1, Todeschini, R., Milano Chemometrics and QSAR Group, http://www.disat.unimib.it/chm/), and different statistical packages such as SPSS software (version 11.50, SPSS Inc.), MATLAB (version 6.50, Math works Inc.).

1.5.1 Hyperchem

HyperChem is an exceptionally powerful and easy to use desktop molecular modeling program. It is a popular and well-known molecular modeling solution for researchers, educators, and students alike. It is used to draw the chemical compound and appears them in the 3D form and the most stable state (Fig.1-1). Also it is used to calculate some descriptors and it's easy to use. The output file of it is an input for dragon to calculate descriptors, it provides a wide range of computation and visualization methods.

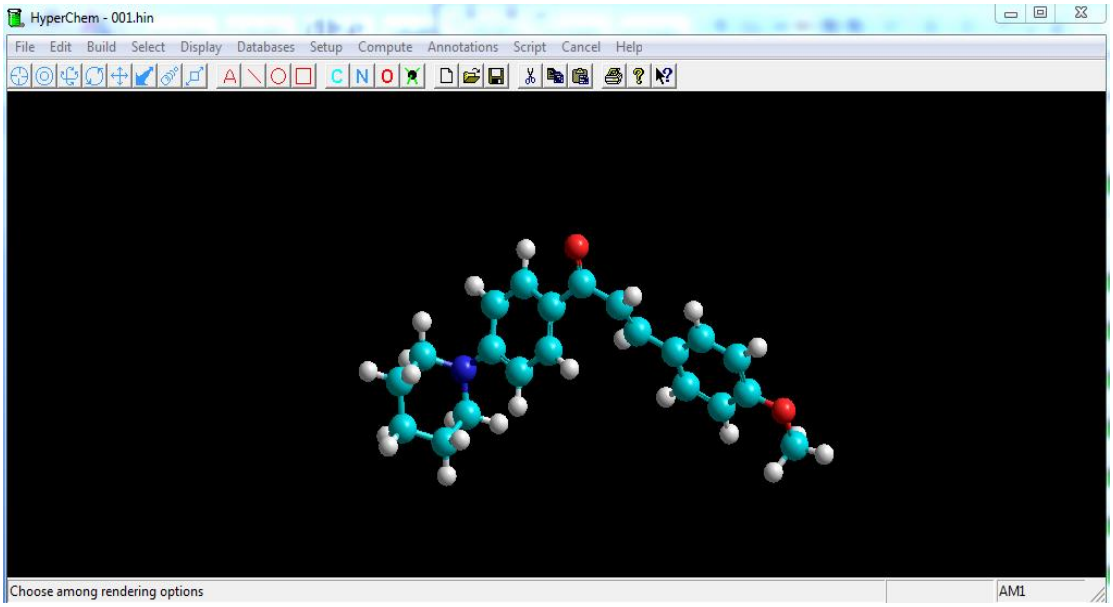


Figure:1-1: The Hyperchem display screen of a compound

1.5.2 Dragon

The first release of DRAGON was developed in 1994 by Milano Chemometrics and QSAR Research Group with the name "WHIM/3D QSAR", being specific for the calculation of the WHIM descriptors [49]. Since 1997 DRAGON has been regularly updated by inclusions of new molecular descriptors in order to advance research in QSAR and new algorithms for optimizing precision and time performances as well as its capability to read

different molecular file formats. Actually, DRAGON [50] allows the calculation of 1,664 molecular descriptors and it is designed to work both for Windows and Linux system. However, by DRAGON it is possible to merge calculated molecular descriptors and user-defined properties for a set of molecules, providing a complete output file which is easily loaded by any correlation analysis application (Fig.1-2).

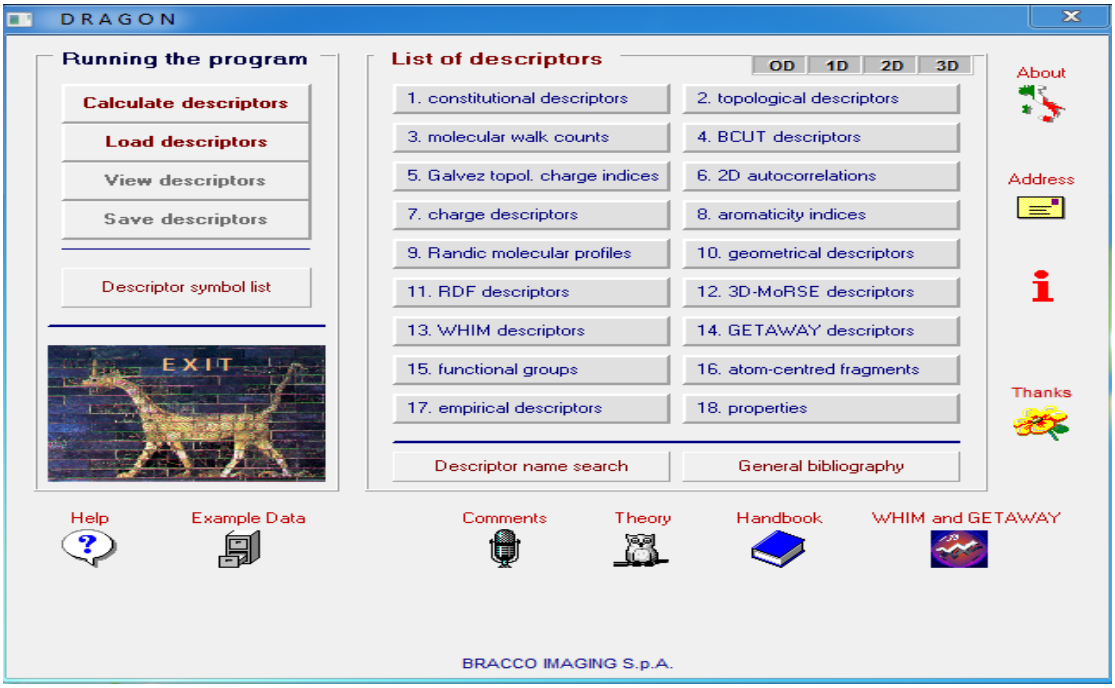
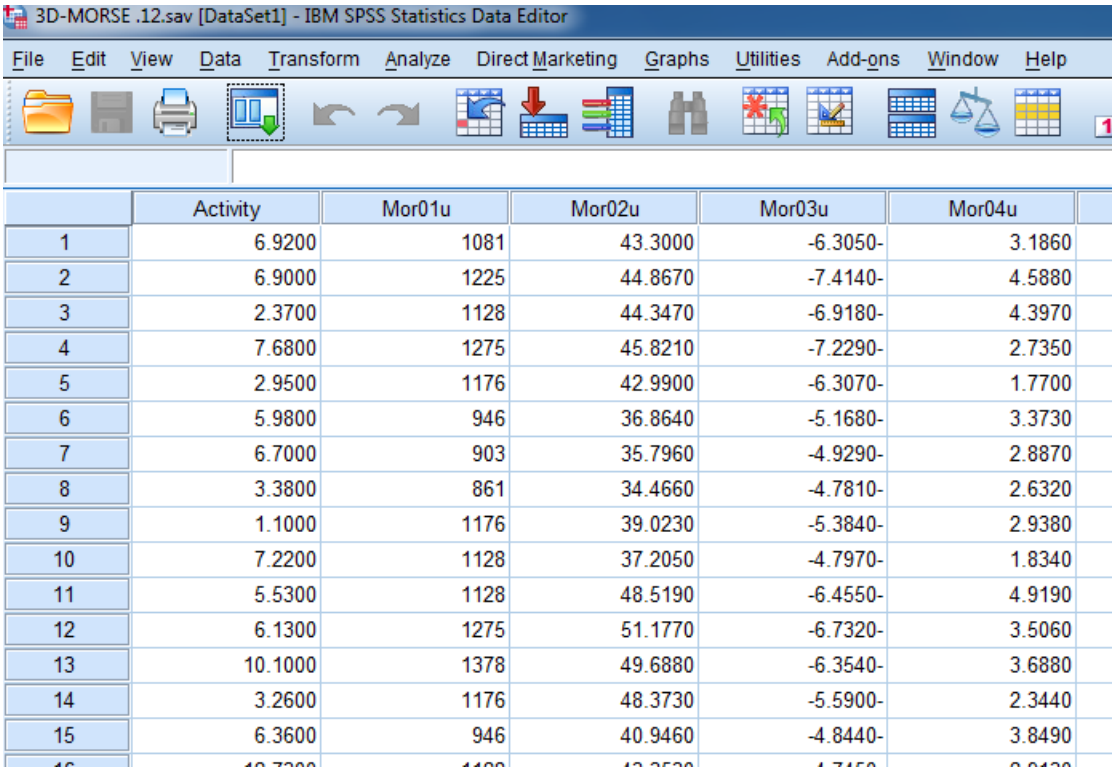


Figure 1-2: Dragon software screen

1.5.3 SPSS

SPSS (currently officially "IBM® SPSS® Statistics") is a commercially distributed software suite for data management and statistical analysis and the name of the company originally developing and distributing the program. **SPSS** means "**S**tatistical **P**ackage for the **S**ocial **S**ciences" introduced in 1968, it helped revolutionize research practices in the social sciences, enabling researchers to conduct complex statistical analyses on their own. SPSS is a relatively easy-to-handle statistics program providing commonly used

procedures. As such, it is widely used in academia including communication studies, although facing increasingly tough competition from more comprehensive and free open-source software [51]. In our study we use SPSS version 20 to perform MLR analysis (Fig.1-3).



	Activity	Mor01u	Mor02u	Mor03u	Mor04u
1	6.9200	1081	43.3000	-6.3050-	3.1860
2	6.9000	1225	44.8670	-7.4140-	4.5880
3	2.3700	1128	44.3470	-6.9180-	4.3970
4	7.6800	1275	45.8210	-7.2290-	2.7350
5	2.9500	1176	42.9900	-6.3070-	1.7700
6	5.9800	946	36.8640	-5.1680-	3.3730
7	6.7000	903	35.7960	-4.9290-	2.8870
8	3.3800	861	34.4660	-4.7810-	2.6320
9	1.1000	1176	39.0230	-5.3840-	2.9380
10	7.2200	1128	37.2050	-4.7970-	1.8340
11	5.5300	1128	48.5190	-6.4550-	4.9190
12	6.1300	1275	51.1770	-6.7320-	3.5060
13	10.1000	1378	49.6880	-6.3540-	3.6880
14	3.2600	1176	48.3730	-5.5900-	2.3440
15	6.3600	946	40.9460	-4.8440-	3.8490

Figure 1-3: SPSS display screen as an example

1.5.4 MATLAB

The name MATLAB stands for MATrix LABoratory. MATLAB was written originally to provide easy access to matrix software developed by the LINPACK (linear system package) and EISPACK (Eigen system package) projects [52]. It is a high-performance language for technical computing, which integrates computation, visualization, and programming environment. It has powerful built-in routines that enable a very wide variety of computations. It also has easy to use graphics commands that make the visualization of results immediately available. The software package has been commercially available

since 1984 and is now considered as a standard tool in most universities and industries worldwide. In this work the MATLAB version 2015 were used to perform the cross validation and PC-ANN (Fig.1-4).

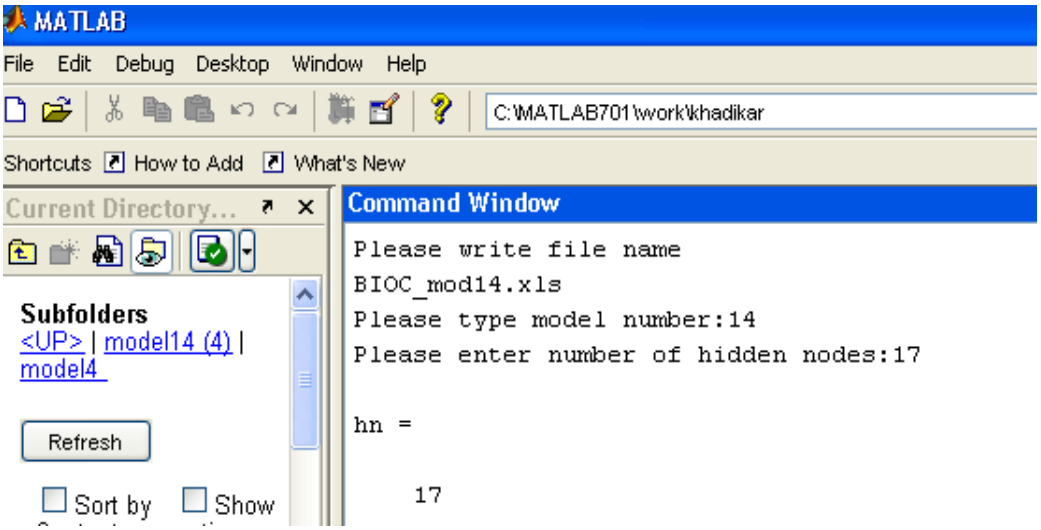


Figure 1-4: MATLAB command window

1.6 Previous studies

V. K. Agrawal, et al.(2001) has studied the antimalarial activity of a series of sulfonamide derivatives (2,4-diamino-6-quinazoline sulfonamides) and they modeled them topologically using Wiener (W)-, and Szeged (Sz)-indices. They obtained regression analysis of the data which shown that better results are in multiparametric regressions upon introduction of indicator parameters. The r2 CV values was used for testing the predicting ability of the models, and the W index based models gave a little bit better results in comparison to those involve Sz [53].

C.X. Xue , et al.(2004) has studied the 3D QSAR analyses of 47 compound (antimalarial alkoxylated and hydroxylated chalcones), were first conducted by Comparative molecular field analysis (CoMFA) and Comparative similarity indices analysis (CoMSIA) to

determine the factors required for the activity of these compounds. And they obtained the satisfactory results 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. These results provided the tools to expect the affinity of related compounds, and for guiding the design and synthesis of novel and more effective antimalarial agents [54]. Also L.F. Motta ,et al. (2006) has performed QSAR analysis on a series of chalcone derivatives as antimalarial agent. A computational package Molecular Modeling Pro 4.0, ChemSite Pro 5.0 and Arguslab 4.0 programs were used to calculate the different physical-chemical parameters such as hydrophobic, electronic, steric, thermodynamic and structural properties. QSAR models with up to four variables were generated employing multiple linear regression method using Build QSAR program. Statistically significant models with *R*–values 0.931 and 0.958 were obtained. And they found that hydrophobic and steric properties seem to play an important role in the explanation of the activity [55].

However, O. Deeb et al. (2012) has modeled the antimalarial activities of two series of farnesyltransferase (FTase) inhibitors by means of multivariate image analysis applied to quantitative structure-activity relationship (MIA-QSAR). A reliable model was achieved, with *r*2 for calibration, external prediction and leave-one-out cross-validation of 0.96, 0.87 and 0.83, respectively. The MIA-QSAR was used to predict the bioactivities using model of the new compounds based on the miscellany of sub-structures of the two classes of FTase inhibitors and the most promising ones were submitted to ADME (absorption, distribution, metabolism and excretion) and docking evaluation. The proposed compounds have substructures contemplated in calibration and, therefore, the modeling of their activities is reliable. Hydrogen bond between a couple of proposed substrates, explain the high affinity of these ligands to the FTase enzyme [56].

Also Nitendra K. SahuI et al. (2012) has performed quantitative structure-activity relationship (QSAR) analysis of some synthesized substituted 4- quinolinyl and 9-acridinyl hydrazone derivatives to find out the structural requirements of their antimalarial activities. Various 2D descriptors were calculated and used in the present analysis. Its performed using three statistical methods: the multiple linear regressions, giving square of correlation coefficient (r2) = 0.8456, cross validated squared correlation coefficient (q2) = 0.7814 and predictable ability (pred_r2) = 0.7443; the partial least squares regression, with r2 = 0.8402, q2 = 0.7879 and pred_r2 = 0.7680; and principle component regression, giving r2 = 0.8392, q2 = 0.7979 and pred_r2 = 0.6440. Best equation was selected on the basis of high r2, q2 and pred_r2. The QSAR model indicated that the T_N_O_3, Id Average, chiV6chain, Most-ve Potential and T_C_N_6 played an important role in determining antimalarial activities. The results of the present study may be useful in the designing of more potent and effective analogues as antimalarial agents [57].

And Apilak Worachartcheewan et al. (2013) has employed a data set of amidino bis-benzimidazoles, in particular 2′-arylsubstituted-1*H*,1′*H*-[2,5′]bisbenzimidazolyl-5-carboximidine derivatives with anti-malarial activity against *Plasmodium falciparum* in investigating the quantitative structure-activity relationship (QSAR). Quantum chemical and molecular descriptors were obtained from calculations and Dragon software. Using multiple linear regression (MLR) and artificial neural network (ANN), a QSAR models were constructed. The results indicated that the predictive models for both the MLR and ANN approaches using leave-one-out cross-validation afforded a good performance in modelling the anti-malarial activity against *P.falciparum* as observed by correlation coefficients of leave-one-out cross-validation (*R* LOO-CV) of 0.9760 and 0.9821,

respectively, and predictivity of leave-one-out cross-validation (Q LOO-CV 2) of 0.9526 and 0.9645, respectively. Model validation was performed using an external testing set and the results suggested that the model provided good predictivity for both MLR and ANN models with correlation coefficient of the external set (RExt) values of 0.9978 and 0.9844, respectively.and predictivity of the external set (Q Ext 2) of 0.9956 and 0.9690, respectively. Furthermore, the robustness of the QSAR models is corroborated by a number of statistical parameters. So the QSAR models that constructed provide pertinent insights for the future design of anti-malarial agents [58].

While Ruslin Hadanu et al. 2015 has performed a quantitative structure and activity relationship (QSAR) analysis of 13 benzothiazoles derivatives as antimalarial compounds, using electronic descriptor of the atomic net charges (q), dipole moment (μ), ELUMO, EHOMO and polarizability (α). The antimalarial activity (IC50) were taken from literature. The best model of QSAR model was determined by multiple linear regression approach and giving equation of QSAR: Log IC50 = 23.527 + 4.024 (qC4) + 273.416 (qC5) + 141.663 (qC6) – 0.567 (ELUMO) – 3.878 (EHOMO) – 2.096 (α). The equation was significant on the 95% level with statistical parameters: n = 13, r = 0.994, r 2 = 0.987, SE = 0.094, Fcalc/Ftable = 11.212, and gave the PRESS = 0.348. This means that only a relatively few deviations between the experimental and theoretical data of antimalarial activity [59].

A. Jarrahpour et al. (2018) has described the synthesis of some new β-lactam derivatives containing the 1,2,3-triazole moiety. An *in vitro* antimicrobial and antimalarial activities were evaluated for all of the compounds. And its show moderate to excellent antimalarial activities. The results showed that compound (**5a)** 1,4-Bis(4-((1-benzyl-1H-1,2,3-triazol -4-yl)methoxy)phenyl)-3-methoxyazetidin-2-one, exhibited the most potent

antimalarial activity with IC50 values of 0.85 μM against chloroquine-resistant *P. falciparum* K1 strain. QSAR study that correlate the observed antimalarial activities with changes in the compounds' structural features. (QSAR) modeling not only showed a systematic relationship between the structures and the observed antimalarial activities, but also pointed to the positive role of electronegative groups and negative role of polarizability and molecular volume on the pIC50 (− log IC50) values. This suggests the use of non-bulky, highly electronegative side chains with a symmetric or near-symmetric position in the structure of these molecules (to reduce the polarizability) as a potential means of improving the bioactivity [60].

Ruslin Hadanu et al. (2018) has studied the Quantitative Structure and Activity Relationship (QSAR) for a series of 13 nitrobenzothiazole derivatives as antimalarial compounds to find out the structural relationship of their antimalarial activities against the W2 Plasmodium falciparum strain. The atomic net charges (q), dipole moment (μ), E LUMO, E HOMO, polarizability (α) and Log P were used to determine the electronic descriptors. In addition, the HyperChem for Windows 8.0 using the PM3 semi - empirical method were used to calculate the descriptors. The antimalarial activities (IC 50) of the compounds were collected from literature. Furthermore, the multiple linear regression (MLR) approach was used to determine the QSAR model, giving equation model . The most statistically significant QSAR model with correlation coefficients n = 13, (r) = 1.00 , (r^2) = 1.00 , SE = 0 , and PRESS = 3.40 were developed by MLR. According to the model of the QSAR equation, 43 new nitrobenzothiazole derivatives were modeled and 24 of these compounds showed high antimalarial activity. It is recommended that these are synthesized for further investigation. Four of the new compounds show equivalent activity to that achieved with chloroquine antimalarial drugs [61].

1.7 Study objectives

As the malaria disease continuous to be one of the major public health problem. It's must be stopped by vaccine and drugs to reduce the mortality. Many previous studies were done on different antimalarial agent but there's no study found or help in the synthesis of the optimum antimalarial drug. So the aim of our study is to design a model used to expect the inhibitory activity of a new suggested compounds by the application of the resulted equation for the best model (good prediction power). In order to do this, a 79 antimalarial compounds were collected from different scientific papers to develop QSAR model by the application of these techniques; MLR as a linear method and PC-ANN as non linear method in order to help in designing a new antimalarial agent with the optimum inhibitory activity which is potent, non toxic and less side effects in comparison with the used one.

Chapter Two

Methodology

Chapter Two

Methodology

2.1 QSAR methodology

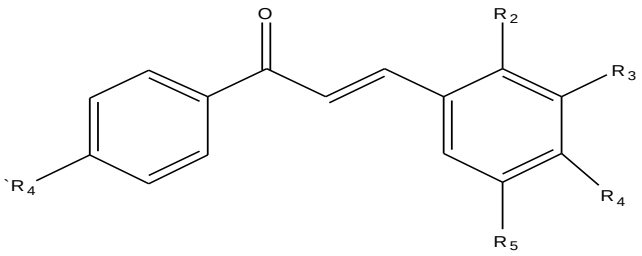
The quantitative structure activity relationship is an analytical way to design and choose the best active agent after calculating the relationship between the chemical structure and the experimental value of activity (biological activity). A lot of steps take place in QSAR done by a specific computer software.

As we mentioned in the previous chapter one; a three main steps are necessary to develop a QSAR model. These are; data preparation, data analysis and model validation.

2.2 Data preparation

2.2.1 Data set

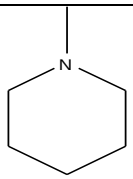
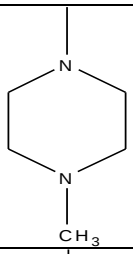
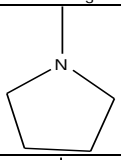
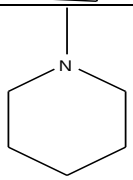
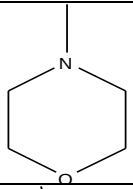
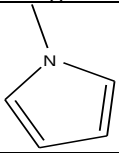
A data set containing 79-compounds with their IC_{50} (µg/ml) against malaria disease taken from the literature [62-67], the activity of them were determined in the same way (candle jar method) and have the same activity unit. The compounds have seven different chemical structures, these all summarized with the biological activities in the following Table 2-1.

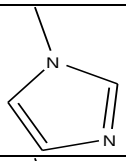
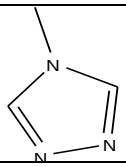
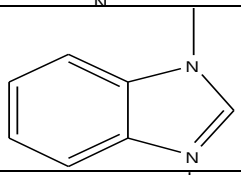
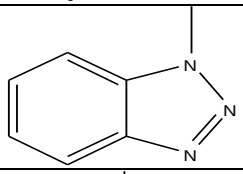
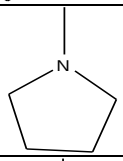
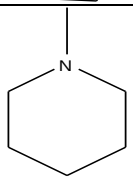
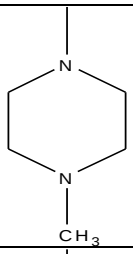
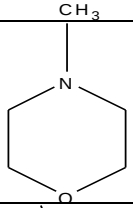
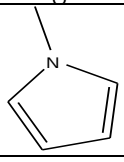
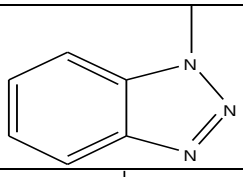
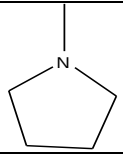


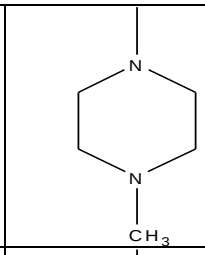
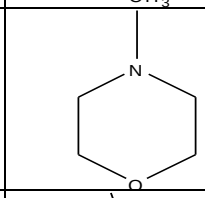
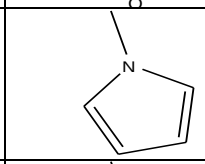
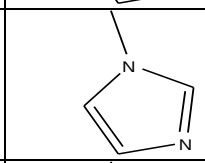
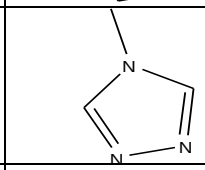
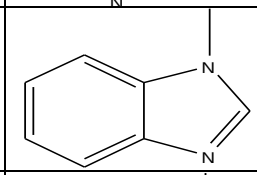
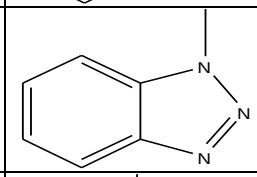
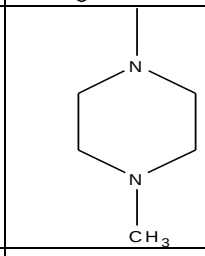
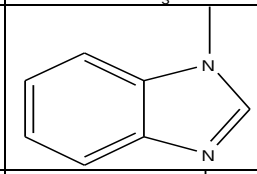
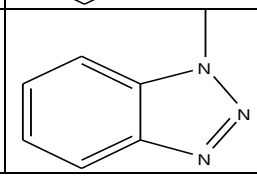
1-27

Chemical structure of Chalcone

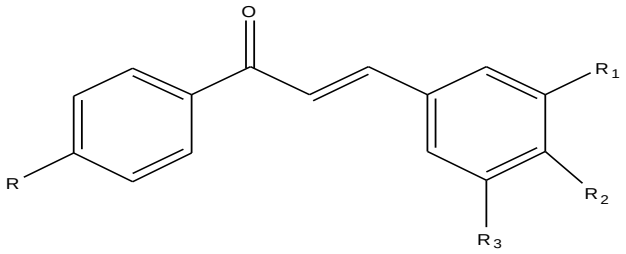
Table:2-1:Dataset,the compounds and antimalarial activity

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
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

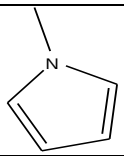
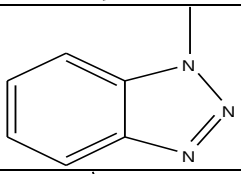
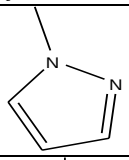
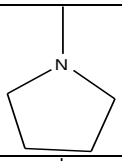
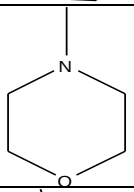
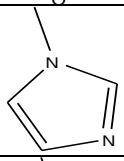
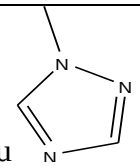
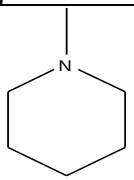
018		H	OCH_3	OCH_3	H	4.96
019		H	OCH_3	OCH_3	H	5.16
020		H	OCH_3	OCH_3	H	6.28
021		H	OCH_3	OCH_3	H	7.34
022		H	OCH_3	OCH_3	H	5.85
023		H	OCH_3	OCH_3	H	5.04
024		H	OCH_3	OCH_3	H	3.5
025		H	OCH_3	OCH_3	OCH_3	4.7
026		H	OCH_3	OCH_3	OCH_3	7.6
027		H	OCH_3	OCH_3	OCH_3	8.15

Ref. [62]

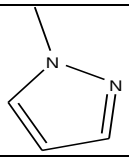
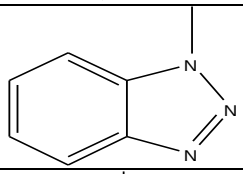
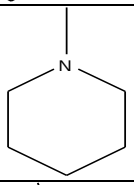
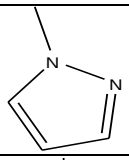
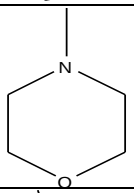
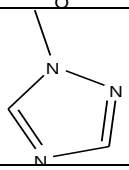


28-41

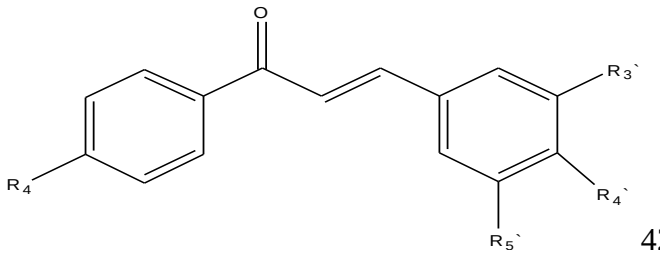
Chemical structure of Chalcone

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		H	Cl	H	1.52
035		H	Cl	H	5.15

µ

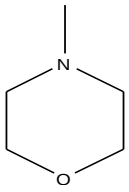
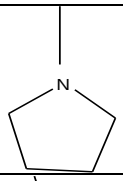
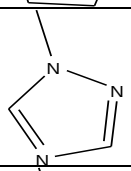
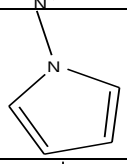
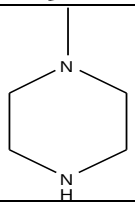
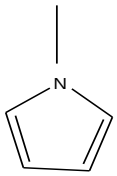
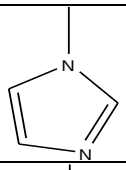
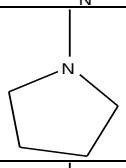
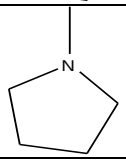
036		H	OCH_3	H	12.33
037		H	OCH_3	H	6.8
038		OCH_3	OCH_3	OCH_3	7.10
039		OCH_3	OCH_3	OCH_3	6.0
040		OCH_3	OCH_3	OCH_3	4.6
041		OCH_3	OCH_3	OCH_3	8.03

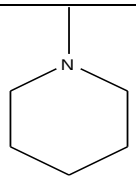
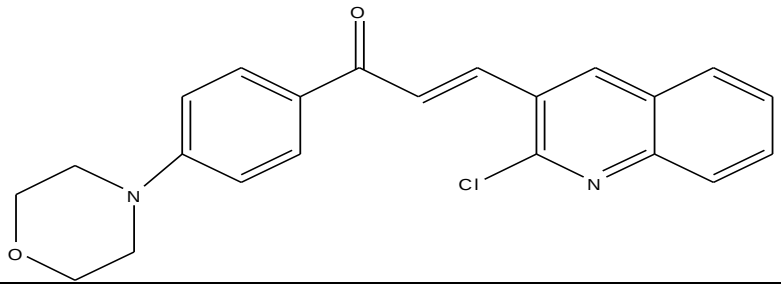
Ref. [63]



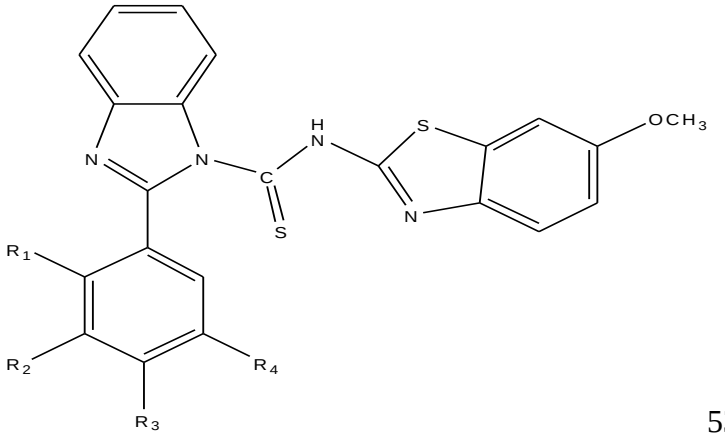
42-52

Chemical 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
050		OCH_3	H	OCH_3	8.43
051		OCH_3	OCH_3	H	3.13

					
052					17.03

Ref. [64]

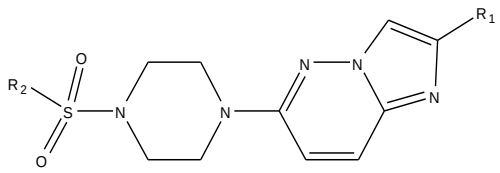


53-60

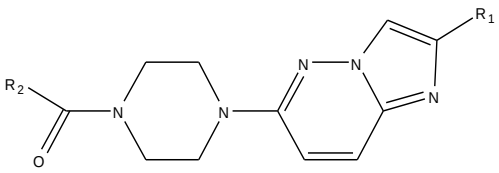
Chemical structure of N-(6-methoxybenzo[*d*]thiazol-2-yl)-2-substituted phenyl-1*H*-benz[*d*]imidazole-1-carbothioamide derivatives

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. [65]



61-66 formula B

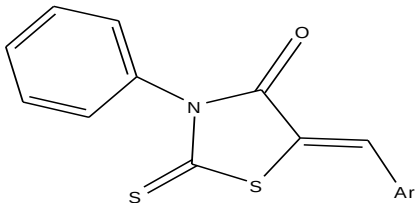


67-69 formula A

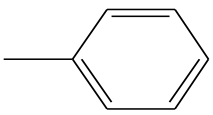
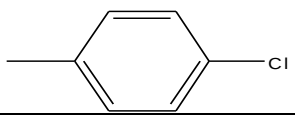
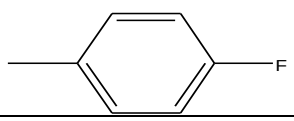
Chemical structure of amide and sulfonamide derivatives

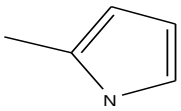
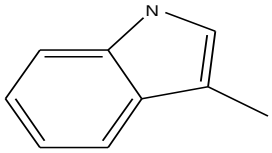
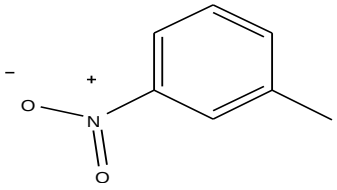
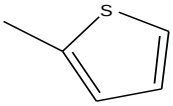
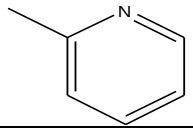
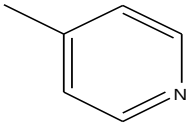
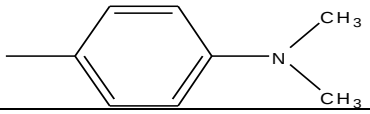
Compound No.	Formula	R_1	R_2	IC_{50} (µg/ml)
061	A	CF_3	C_2H_5	0.77
062	A	CF_3	$4CH_3$-phenyl	0.97
063	A	CF_3	$4CF_3$-phenyl	0.79
064	A	$4CH_3$-phenyl	C_2H_5	0.80
065	A	$4CF_3$-phenyl	C_2H_5	0.89
066	A	2,5-dichlorophenyl	C_2H_5	0.98
067	B	CF_3	C_2H_5	1.01
068	B	CF_3	$4CH_3$-phenyl	1.12
069	B	2,5-dichlorophenyl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

Ref. [66]



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

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

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

Ref. [67]

2.2.2 Compounds optimization

The collected molecules to be used in the study and to calculate some properties , the 3-D structures are required, which represent a minimum potential energy and most stable state. 3-D structures are usually generated using Hyperchem.

Compounds optimizations steps using Hyperchem:

1.Open the Hyperchem program, then start to draw the compound structure using the draw icon and the atoms in the workspace. see (Fig. 2-1).

2. Then convert the drawn compound from 2D (two-dimensional) form into 3D using the Hyperchem model builder, and from the build menu select add H and model build.

3. Save the drawn structure by choosing start log from the file menu. then name the file and choose the appropriate folder to save it.

4. In order to optimize the compound structure; choose semi-empirical from the setup menu. A dialog box of the semi-empirical method will appear, then choose AM1 and click ok.

5. From the compute menu choose geometry optimization, a dialog box appear. then set algorithm on Polak Ribiere, RMS gradient (0.01 kcal/mol) and maximum cycles (10000). Then click ok to start optimization process.

6. After the optimization converged, click on stop log from the file menu in order to save the calculation output as log file. This output file saved as (hin) format.

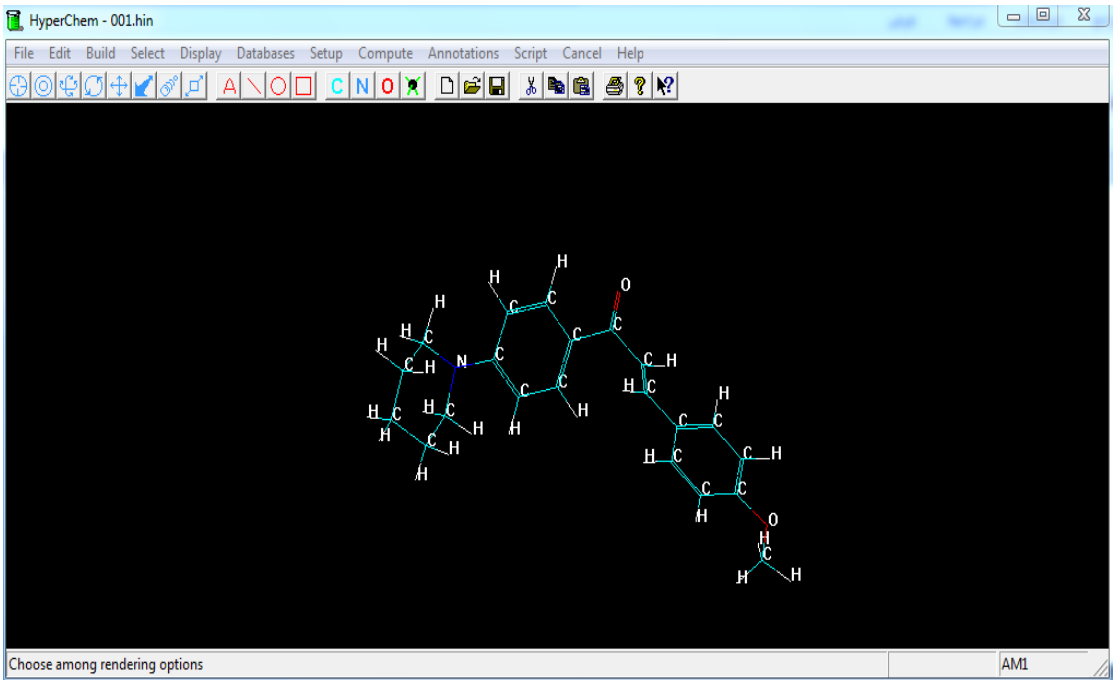


Figure 2-1: Drawing in Hyperchem working space

2.2.3 Descriptors calculation

The fundamental assumption of QSAR modeling is that molecular structure can be correlated to physical or biological properties. Thus the fundamental requirement is some method to encode various structural features in a molecule. Molecular descriptors fulfill this requirement.

Descriptors are (in general) numerical representations of specific molecular features. Such features can range from very simple ones such as the number of carbons or number of halogen atoms to more complex and abstract features such as graph invariants of the molecular graph or the information content of a molecule as characterized by entropy.

In the present study we used two software's to calculate a different descriptors; Hyperchem and Dragon.

2.2.3.1 Descriptors calculated by Hyperchem

a. Descriptors mentioned in the output log file

In the Hyperchem many descriptors are calculated, such as; quantum chemical and more. After the opening of the output log file for each optimized structure, extract the following values to fill them in an excel file:

- HOMO (Highest Occupied Molecular Orbital)
- LUMO (Lowest Unoccupied Molecular Orbital)
- Heat of formation (Kcal/mol)
- Dipole moment (Debyes)

From the values of HOMO and LUMO, calculate the following descriptors;

- Hardness [0.5*(LUMO-HOMO)]
- Softness (1/Hardness)
- Electronegativity [-0.5*(LUMo+HOMO)]
- Electrophilicity (Electronegativity* Electronegativity/(2*Hardness)

b. Descriptors calculated in the Hyperchem for the optimized structure

A selected descriptors calculated in the Hyperchem by the following step

1. Open the file of the optimized structure for each compound that resulted from step 6 in 2.2.2
2. Choose QSAR properties from the compute menu, a dialog box will opened which contain the following properties (Fig.2-2):

- Surface Area (Approx)

- Surface Area (Grid)
- Volume
- Hydration Energy
- Log P
- Refractivity
- Polarizability
- Mass

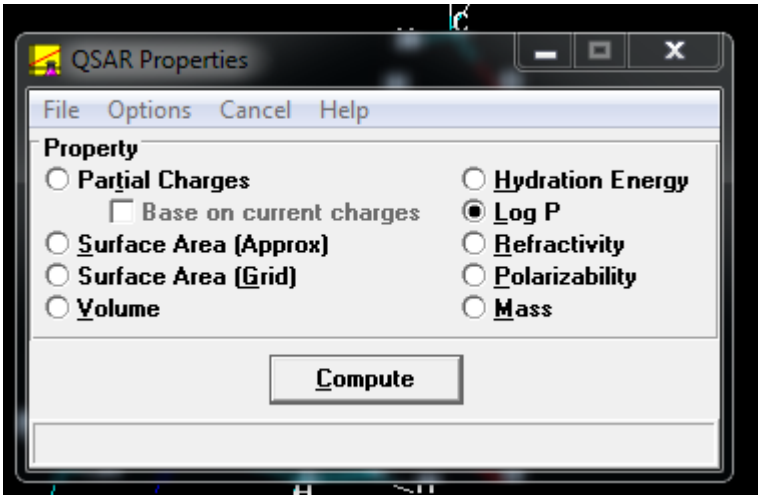


Figure 2-2: Dialog box of QSAR properties in Hyperchem

3. Choose one of them and press on compute bottom, then save the result in a table in excel file.

4. Repeat the above steps to calculate the rest of properties one by one for each optimized structure file.

2.2.3.2 Descriptors calculated by Dragon software

DRAGON virgin 2.1 was used for the calculation of thousands of descriptors that are divided into eighteen group that mentioned in the previous chapter.

DRAGON software has been conceived to provide the user with a variety of molecular descriptors derived from different molecular representations, allowing the user to choose those molecular descriptors which are more suitable for his specific research.

2.2.3.2.1 Brief description about Dragon descriptors

Constitutional descriptors (block 1) are the most simple and commonly used descriptors, reflecting the composition of a molecule without any geometrical information. Examples of these descriptors are the number of atoms, bonds, rings, specific atom types, rotatable bonds, etc. Enumerative descriptors are also counts of functional groups (block 17) and Ghose-Crippen atom-centred fragments (block 18).

The descriptor blocks 2–10 contain topological and topographic descriptors. Topological descriptors are based on a graph representation of the molecule. They are numerical quantifiers of molecular topology obtained by the application of algebraic operators to matrices representing molecular graphs and whose values are independent of vertex numbering or labeling. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. Topographic indices are derived from the graph representation of molecules in the same way as the topological indices, but using the geometric distances between atoms instead of the topological distances.

The blocks 11–16 include descriptors derived from the knowledge of the 3D structure of the molecule.

The molecular descriptors are classified into:

- 0D : Constitutional descriptors
- 1D : Functional group, Atom-centered fragment, Empirical, properties
- 2D : Topological descriptors, Molecular walk count, BCUT descriptors, Galvez topol. charge indices, 2D autocorolations.
- 3D : Charge descriptors, Aromaticity indices, Randic molecular profiles, Geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors.

2.2.3.2.2. Steps to perform descriptors calculation in DRAGON software

1. Start the Dragon software, then click on calculate descriptors bottom in the left of the program window, a dialog box appear for calculation.

2. Select the output files of compounds optimization from Hyperchem and choose the file type as (hin) format. After that choose the descriptors group to be calculated and press run.
3. An output file resulted from the previous step ,this file saved in notepad format.

4. Convert the file format from notepad to excel to use it in SPSS analysis.

5. Repeat the previous steps for all compounds for each group of descriptors.

2.3 Data analysis

2.3.1 Multiple linear regression (MLR)

The SPSS software was used to find a linear relationship between the biological activity (dependent variable) and the molecular descriptors (independent variables). The MLR is the first statistical step performed because the of the assumption of the linear relationship between the variables.

2.3.1.1 MLR performing steps for each descriptor group using SPSS

1. Import the output files of Dragon and Hyperchem which contains the calculated descriptors like topological descriptors for all 79 compounds in excel file format. This output file will contain the calculated descriptors and the biological activity of all compounds as in (Fig. 2-3)

	Activity	ATS1m	ATS2m	ATS3m	ATS4m
1	6.9200	.6010	.4680	.3680	.3030
2	6.9000	.5880	.4610	.3630	.2760
3	2.3700	.6220	.4920	.4140	.3730
4	7.6800	.6090	.4750	.4050	.3430
5	2.9500	.6420	.5070	.4520	.3740
6	5.9800	.6690	.5620	.4930	.4270
7	6.7000	.6890	.5810	.5170	.4390
8	3.3800	.7110	.6000	.5360	.4520
9	1.1000	.6960	.6010	.5400	.4510
10	7.2200	.7150	.6190	.5610	.4720
11	5.5300	.6220	.4920	.4140	.3690
12	6.1300	.6090	.4750	.4050	.3400
13	10.1000	.6150	.4840	.4280	.3220

Figure 2-3: SPSS data screen

2. After that click on analyze , regression then linear to perform MLR. An dialog box will open as shown in (Fig.2-4) to set the following;

- Set IC50 as dependent variable and the descriptors as independent variables
- Choose the method as stepwise for the analysis

- Click on options button , another dialog box opened to set F value (entry and removal F value),change them to get proper wanted results see (Fig.2-5).
- Then click on statistics button and choose model fit and estimates then continue.
- Finally click on save button and choose unstandardized predicted value then continue .

3. After that click on Ok to complete the MLR process.

4. Repeat the MLR steps discussed above for each descriptors group alone.

5. From the output file of MLR for each group choose the best model, which has the higher R value and minimum number of descriptors.

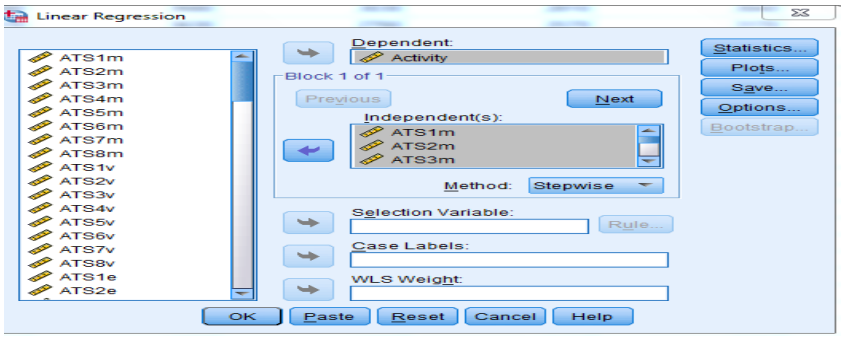


Figure 2-4: Dialog box for MLR analysis

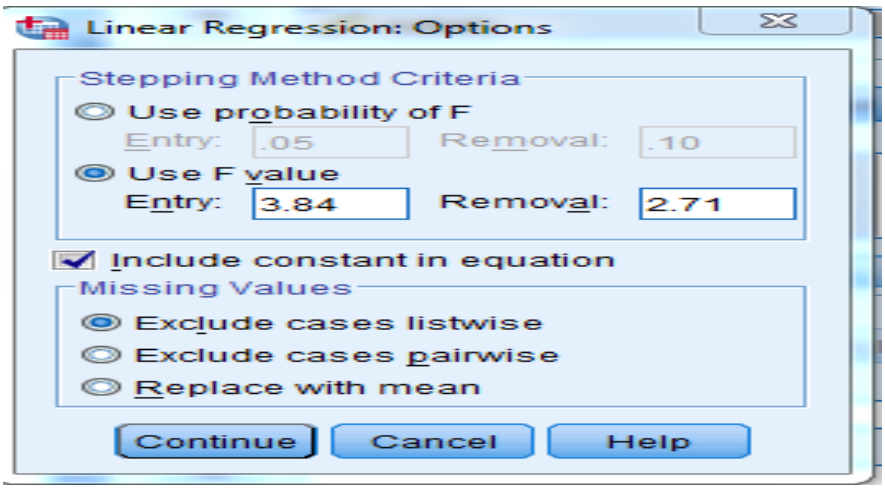


Figure 2-5: Dialog box for options to change F value

2.3.1.2 Steps to perform MLR for all descriptors resulted from the first MLR for best models using SPSS

1. Create an excel file for all descriptors of the best models resulted from the first MLR.

2. Import the prepared file to SPSS and apply the MLR steps mentioned in section 2.3.1.1

3. After getting the results from the MLR choose the models with $R^2 \geq 0.6$ [68].

2.3.2 MLR model validation

Validation methods are needed to establish the predictiveness of a model on unseen data and to help determine the complexity of an equation that the amount of data justifies. Using the data that created the model (an internal method) or using a separate data set (an external method) can help validate the QSAR model. In our study we use two different methods to validate the models which are; cross validation ($Q2$) and scrambling (Y-Randomization). These are internal methods of validating a model.

2.3.2.1 Cross validation

This method was used in the research to validate the best models resulted from the MLR in SPSS. The two types of methods was used are leave one out (LOO) and the leave many out (LMO) by using special scripts in the MATLAB.

2.3.2.1.1 Leave one out (LOO) method performing steps

1. Open an excel file to prepare the input file for the (LOO), copy the observed activities of the compounds in the first column, and then copy the predicted activities column from the MLR results in SPSS for each models that have $R^2 \geq 0.6$ [68].

2. Run the MTLAB script for the LOO validation, then enter the file name prepared in step 1, then the MATLAB will ask for the model number and then the number of descriptors for each models respectively.

3. The results will appear in an output file as in (Fig. 2-6), in the same working directory.

4. Choose the models which have the PRESS/SST value < 0.4, and compare it with (LMO) results [68].

CV_LOO.dat

	A	B	C	D	E	F	G	H
	Model	**PRESS1**	**SPRESS1**	**SST1**	**R2CV1**	**PRESSSST**	**PSE1**	**RSEP1**
	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER	NUMBER
1	Model	PRESS	SPRESS	SST	R2CV	PRESS/SST	PSE	RSEP
2	----------							
3	1	1927.0775	5.1735	17.7722	-107.4324	108.4324	4.9390	87.5216
4	2	2480.6975	5.9110	2.8861	-858.5399	859.5399	5.6037	99.3007
5	3	1193.7548	4.1004	797.4002	-0.4971	1.4971	3.8873	68.8848

Figure 2-6: LOO results

2.3.2.1.2 Leave many out (LMO) method performing steps

1. Prepare a file for each model which containing the observed activity and the predicted activity.

2. Run the MATLAB script for LMO validation, then enter the file name prepared in step 1, then the MATLAB will ask for the model number and then the number of descriptors.

3. The results will appear in an output file in the same working directory.

4. Repeat the previous steps for each model

5. Choose the models which have the PRESS/SST value < 0.4, and compare it with (LOO) results [68].

2.3.3 Principal component analysis (PCA)

The principle component analysis technique is widely used for data analysis and data reduction, in this research its used to divide the data set into three main sets; training set, validation set and test set. Each set have a different percent as the following; 60% training set, 20% validation set and 20% test set. To get the appropriate analysis draw the relation between the 1^{st} with 2^{nd} or 2^{nd} and 3^{rd} principle component until you get the plot to do the division of compounds into sets.

Performing steps of PCA in MATLAB:

1. Open the MATLAB program and run the script of principle component analysis.
2. After the MATLAB script run it will ask for the file name needed to PC, then enter the file name that containing the compounds activities and all descriptors which is resulted from the first MLR.
3. Then a figure of the first two PCs will appear as an output of the script.
4. The figure shows the data distribution that you need to select the molecules for the training , validation and the test set.

2.3.4 Artificial Neural Network (ANN)

The Artificial Neural Network technique is the non linear method, used if the linear method of the analysis doesn't give better model of the study.

2.3.4.1 Performing steps of ANN for each model using MATLAB

1. Open the MATLAB program, and run the special script of ANN.

2. After you divide the data in the PCA, fill the compounds number of each set (validation and test set) in the script window.
3. Prepare an excel file for each model alone, which contain the activity and the descriptors.
4. Then the MATLAB will ask you for the file name, model number and the hidden nodes, we use 7 hidden node.
5. After the ANN process stop, take the last value that appear like a file in the data column (CV_model No._hn.7).
6. From the result that you get, choose the best model which have the higher R-test value and low PRESS and RSEP.

2.3.4.2 Performing steps of ANN for the best models with a range of hidden nodes using MATLAB:

1. Open the MATLAB program and run the ANN script.
2. Then the MATLAB will ask you for the file name, enter the file name that prepared for the previous process.
3. Then enter the model number and the hidden nodes; start with hidden node 5.
4. After the ANN process stop, take the last value that appear like a file in the data column (CV_model No._hn.5).
5. Repeat the above steps for hidden nodes from 5-20 for each model choosed.

2.3.5 Randomization test (chance correlation or scrambling model)

The randomization test was choosed to confirm that the ANN models does not resulted by chance.

Performing steps of Randomization test using MATLAB:

1. Prepare a file for each model resulted from the ANN, which contain the activity and the descriptors.
2. Open the MATLAB and run the script for the chance correlation test (Randomization test), then enter the file name, the model number and the trial number, the result appear in the files list as (CC_model NO._hn.No).
3. Repeat the second step until 10 trials in the same MATLAB window that you open for each model alone.

2.4 Summary of QSAR process:

We performed the following menthined previously;

- Dataset preparation (Compounds chemical structure and their activity)
- Geometry optimization using Hyperchem
- Descriptors calculation using Hyperchem and Dragon software
- MLR of models using SPSS as well as validation of them
- PC-ANN statistical models and validation of them using MATLAB.

Chapter Three

Results and Discussion

Chapter Three

Results and Discussion

Antimalarial drugs are effective for decades especially the chloroquine but due the mutation of the parasite that cause malaria, the antimalarial agent become non potent. Many studies have been designed drugs for malaria but in our study we collect a group of compounds with different structure to build a QSAR model to design a new antimalarial agent. A QSAR models were developed as a result of the study using the 79 compounds and their observed activities as antimalarial agents.

- ✓ **Compounds optimization using HyperChem**

A 79 compounds were optimized using HyperChem through the semi-empirical AM1 method. using this method we get very fast and accurate results. Its more appropriate for large molecules.

- ✓ **Descriptors calculation using HyperChem**

Agroup of descriptors were calculated using HyperChem called G-16 quantum chemical descriptors. These descriptors are mentioned below and in the previous chapter in 2.2.3.1

- HOMO (Highest Occupied Molecular Orbital)
- LUMO (Lowest Unoccupied Molecular Orbital)
- Heat of formation (Kcal/mol)
- Dipole moment (Debyes)
- Hardness [0.5*(LUMO-HOMO)]

- Softness (1/Hardness)
- Electronegativity [-0.5*(LUMo+HOMO)]
- Electrophilicity [Electronegativity* Electronegativity/(2*Hardness)]
- Surface Area(Approx)
- Surface Area(Grid)
- Volume
- Hydration Energy
- Log P
- Refractivity
- Polarizability
- Mass

✓ **Descriptors calculation using Dragon**

In Dragon a large number of descriptors is calculated (1235 descriptor). These descriptors were divided into eighteen groups. the results explained below;

- Two groups (Empirical and Properties descriptors) were constant, so the Dragon discard these two groups due to the correlation with each other's.

- The other groups of descriptors which are; Constitutional descriptors, Functional group, Atom-centered fragment, Topological descriptors, Molecular walk count, BCUT descriptors, Galvez topological charge indices, 2D autocorolations, Charge descriptors, Aromaticity indices, Randic molecular profiles, Geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors.

✓ **Performing the first MLR using SPSS**

A Multiple linear regression were performed for each group of descriptors alone. The results of the MLR is summarized in Table 3-1 bellow. where No. refers to the group number, (R) refers to the correlation coefficient, (R^2) refers to the coefficient of determination, (R^2adj.) refers to the adjusted R^2 , and selected descriptors for the best MLR model for each group.

Table 3-1: First MLR models resulted for the descriptors group in SPSS

No.	Group name	#of calculated descriptors	R	R^2	R^2adj.	Standard error of estimation	Selected descriptors
1	Constitutional	33	0.764	0.584	0.549	2.3972324	nS,nF,nN,nR06,nBM,nR09,nCIR,nCIC,nBnz,AMW
2	Topological	227	0.811	0.657	0.607	2.2390110	BIC1,TI2,D/D,HVcpx,IAC,IDDE,HyDp,D/Dr09,piPC06,T(N..CI),VEA1,AECC
3	Molecular walk	19	0.689	0.457	0.447	2.6544048	SRW09,MWC01,MWC06,MWC09,MWC05
4	BCUT	64	0.758	0.575	0.533	2.4386725	BELe1,BEHm3,BEHp1,BELm4,BELp5,BELp6,BELv6,BEHv3,BEHv2
5	Galvez topological charge indices	21	0.678	0.459	0.406	2.7515454	JGI3,JGI1,GGI2,JGI10,JGI2,GGI8,JGI8,JGI6
6	2D autocorrelations	96	0.946	0.896	0.419	2.7203349	GATS8p,ATS5p,ATS1e,MATS3v,MATS3e,MATS1e,MATS5v,GATS4e,MATS8e,GATS3m,GATS2p,MATS1e,MATS7p,MATS8m,MATS5e,GATS8v,GAT

							S6v,GATS3p,MATS7e,MATS6e,MATS1v,GATS4m,GATS4p,GATS5m,GATS1m,MATS4m,MATS2e,ATS7e,GATS8m,MATS4e,GATS2m,GATS7m,GATS8e,GATS3e,GATS6e,MATS3m,GATS6MATS2m,GATS7v,MATS8p,ATS8m,GATS3v,MATS6p,MATS7m,MATS5p,ATS6e,MATS1p,GATS5v,GATS7p,GATS7e,ATS7m,GATS2v,ATS3v,GATS5e,ATS3e,MATS4p,GATS6p,ATS2e,MATS2p,ATS8e,GATS4v,ATS8v,ATS7v,MATS7v
7	Charge descriptors	14	0.578	0.334	0.298	2.9904141	RNCG,Qneg,Q2,PCWTe
9	Randic molecular profile	41	0.625	0.391	0.375	2.8229059	SP10,DP01,SP01
10	Geometrical	35	0.761	0.578	0.502	2.5192790	G(N..S),FDI,G(N..F),ASP,SPH,G(N..N),G(N..O),G1,J3D,SPAM,PJI3,G(O..Cl),G(N..Cl),G(C..Cl),G(S..Cl)
11	RDF	132	0.803	0.645	0.581	2.3115833	RDF040m,RDF105u,RDF055u,RDF090m,RDF065m,RDF075e,RDF115v,RDF125m,RDF150v,RDF040e,RDF140u,RDF125e,RDF015e
12	3D-MorSE	160	0.811	0.657	0.601	2.2546173	Mor30v,Mor03m,Mor03v,Mor14v,Mor29u,Mor30m,Mor13u,Mor05p,

							Mor32v,Mor09u, Mor10m,Mor14p
13	WHIM	99	0.719	0.517	0.438	2.6766245	L1u,L1s,E1m,L1e ,E2u,G1e,G1m,L 3s,E2p,L3m,Av,V p,E1v
14	GETAWAY	197	0.792	0.627	0.545	2.4064329	R2m,R6p+,HATS 6m,H8m,R6m+,H ATS8m,HATS3m ,R4e+,R6u+,R8u +,R3e+,R3u+,HA TS1e,HATS6p
15	Functional	24	0.753	0.567	0.544	2.4103676	N=CHR,nC=N,n C=NPh,nCaH
16	Atom-centered fragments	43	0.751	0.565	0.528	2.4513895	C-016,C-019,N-075,C-042,C-027,C-025
17	Quantum chemicals	16	0.636	0.405	0.364	2.8462895	LOMO(ev),mass(amu),volume,HO MO(ev),Refractiv ity,LogP, Hydration energy

- ✓ **Performing the second MLR using SPSS**

MLR were applied for a group of descriptors resulted from the first MLR together. The results of second MLR is summarized in Table 3-2, in this table only the models that have $R^2 \geq 0.6$ were taken to continue to the next step. So we take models (6 -17) which have $R^2 \geq 0.6$ to do cross validation (leave one out and leave many out methods) [68].

Table 3-2: Second 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, Log P,G(N..O),SP10,GATS3P,Mor32v,Mor13u
10	10	0.837	0.700	0.656	n=CHR ,RDF125e,nC=NPh, Log P,G(N..O),SP10,GATS3P,Mor32v,Mor13u, Mor09u
11	11	0.843	0.711	0.663	n=CHR ,RDF125e,nC=NPh, Log P,G(N..O),SP10,GATS3P,Mor32v,Mor13u, Mor09u,BIC1
12	12	0.848	0.718	0.667	n=CHR ,RDF125e,nC=NPh, Log P,G(N..O),SP10,GATS3P,Mor32v,Mor13u, Mor09u,BIC1, RDF150v
13	13	0.862	0.744	0.693	n=CHR ,RDF125e,nC=NPh, Log P,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, Log P,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, Log P,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, Log P,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, Log P,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 $\underline{R= 0.889}$, $\underline{R^2= 0.791}$, $\underline{R^2adj.= 0.733}$ for the best model 17, and the descriptors are mentioned with a brief description in the Table 3-3 below ;

Table 3-3: 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.

- ✓ **Cross validation performed on the MLR resulted models (6-17), using MATLAB Software.** The results of cross validation done for the models in both methods (Leave one out and leave many out) are summarized in Table 3-4 and 3-5 respectively. 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 3-4 and 3-5 we note that the models 13-17 have a good predictive power, because this models having high R2CV and PRESS/SST less than 0.4. So these models were chosen for artificial neural network analysis (ANN).

Table 3-4: 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 3-5: Leave many out cross validation results

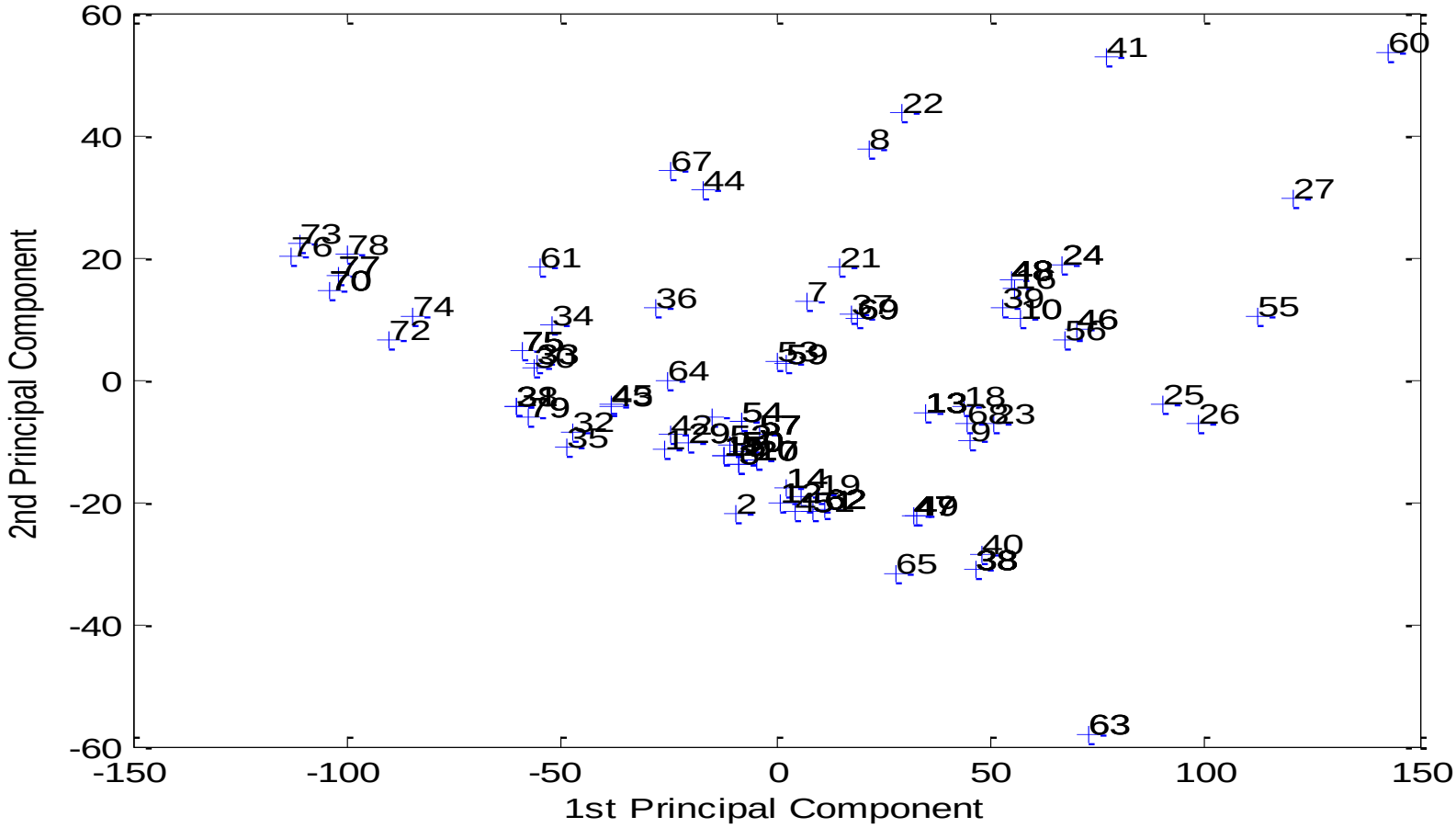
Model	No. desc.	PRESS	SPRESS	SST	R^2cv	PRESS/SST	PSE	RSEP
6	6	425.3771	2.4306	588.1381	0.2767	0.7233	2.3205	41.1498
7	7	347.1467	2.2956	579.5446	0.3544	0.6456	2.1762	38.5924
8	8	362.2698	2.2749	622.2931	0.4178	0.5822	2.1414	37.9749
9	9	312.5505	2.1283	667.7622	0.5319	0.4681	1.9891	35.2728
10	10	318.1504	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

- ✓ 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 plots.

  So that and relying on the first and second principles plot (Fig.3-1), 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.

Figure 3-1: First and second principle component analysis



- ✓ **Artificial Neural Network**

The first ANN was performed on the models choosed (13-17) from the cross validation (LOO and LMO). The ANN was done for each model with hidden node 7. Table 3-6 shows the results of first ANN. 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 (Fig.3-2) 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 (Fig.3-3) 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 (Fig.3-4) 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.

Table 3-6: Correlation coefficient and cross validation parameters for ANN models 13-17 with hidden node 7

Mo.#	Hn	nPCs	R_tr	PRESS_tr	R2CV_tr	R_test	PRESS_test	R2CV_test	R_val	PRESS_val	R2CV_val
13	7	6	0.861082	135.10292	0.588299	0.511026	168.30021	-1.454525	0.443640	231.61768	-1.094348
14	7	6	0.832886	163.89197	0.5192734	0.811143	84.096421	0.0316320	0.608462	173.92235	-0.341211
15	7	6	0.752667	222.93522	0.2594241	0.350165	179.24653	-2.354871	0.4755170	216.39751	-1.541993
16	7	6	0.81804201	173.380427	0.373579	0.8542119	79.282583	-0.077866	0.6738605	150.046951	-0.316826
17	7	5	0.84044	151.533032	0.569134	0.8162310	96.69016	0.099469	0.611081	169.73501	-0.8109936

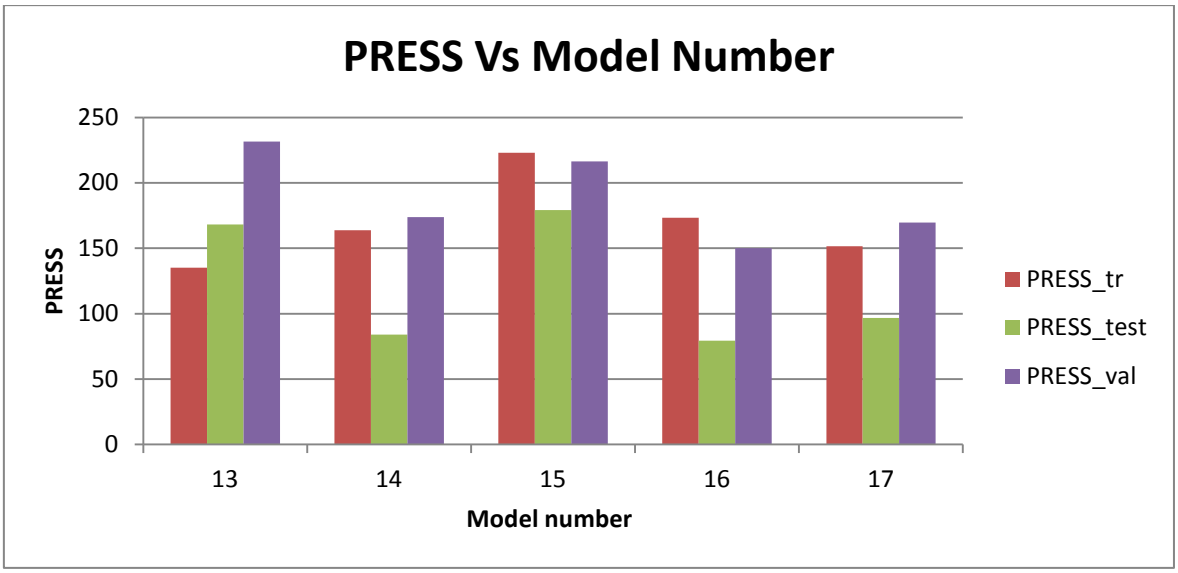


Figure 3-2: Plots of ANN Predictive Residual Sum of Squares(PRESS) values for the training, test and validation sets versus model number.

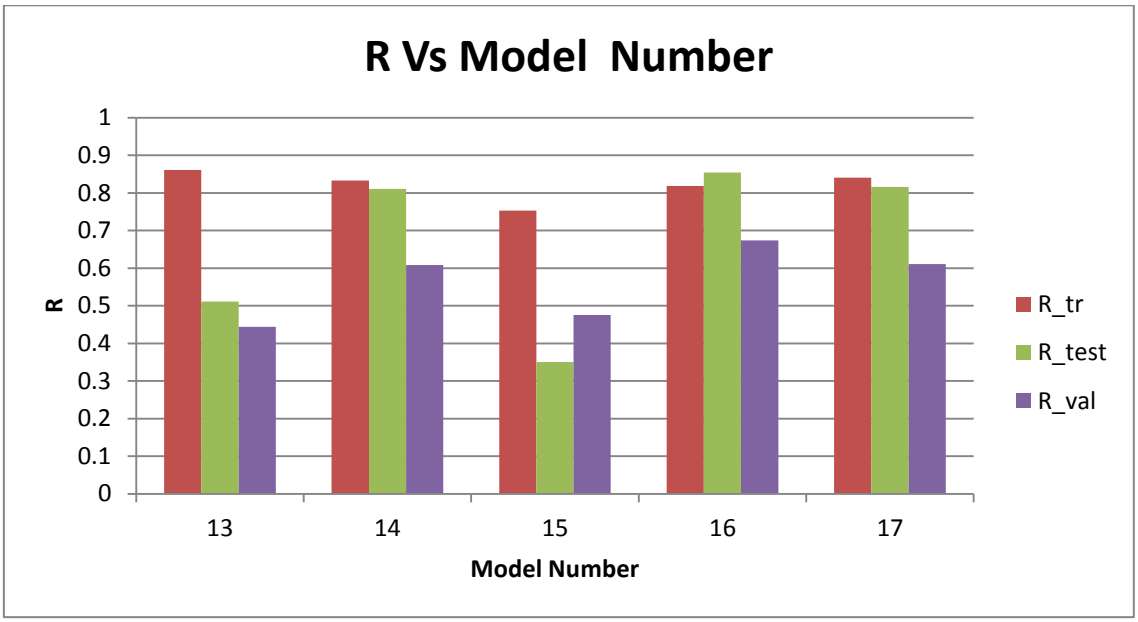


Figure 3-3: Plots of ANN correlation coefficient (R) values for the training, test and validation sets versus model number.

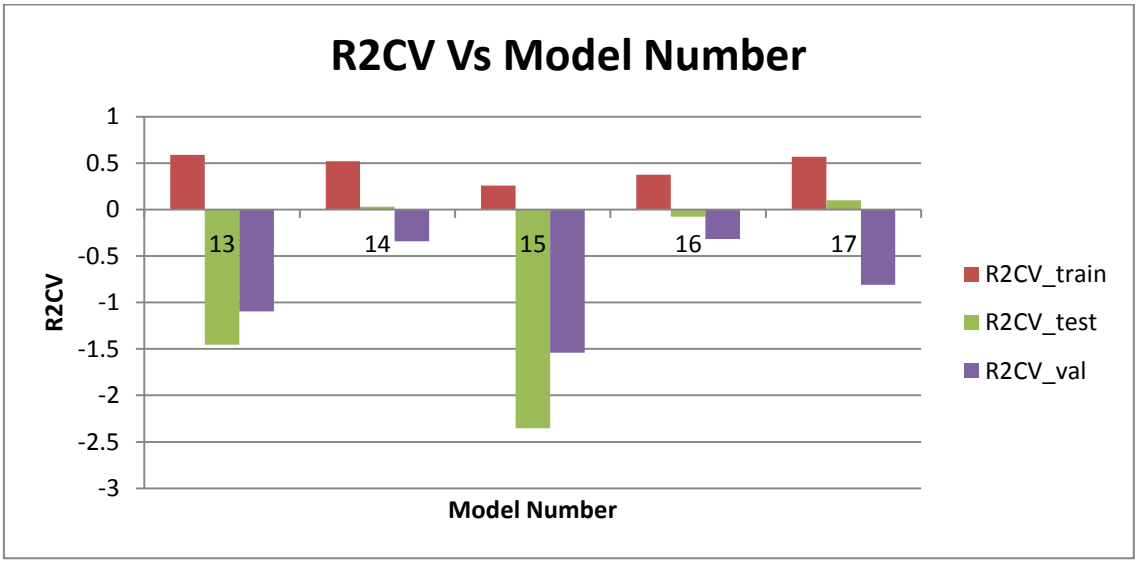


Figure 3-4: Plots of ANN cross validated correlation coefficient (R2CV) values for the training, test and validation sets versus model number.

✓ The second ANN was performed on the choosed 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, the results were noted in Tables 3-7, 3-8, and 3-9 respectively.

From the results shown in the tables, 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 choosed 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 3-10 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 3-7: Correlation coefficient and cross validation parameters of model 14 with hidden nodes (5-20)

Mo.#	hn	nPCs	R_tr	PRESS_tr	R2CV_tr	RSEP_tr	R_test	PRESS_test	R2CV_test	RSEP_test	R_val	PRESS_val	R2CV_val	RSEP_val
14	5	6	0.73345	271.6155	-0.4651	46.1779	0.41107	145.2967	-2.7123	51.9928	0.43922	208.6591	-1.6363	57.12593
14	6	6	0.70428	259.4590	-0.1209	45.1327	0.60052	126.1643	-1.6630	48.4489	0.48697	204.8466	-1.9764	56.60163
14	7	6	0.83288	163.8919	0.5192	35.8703	0.81114	84.0964	0.0316	39.5552	0.60846	173.9223	-0.3412	52.1545
14	8	6	0.76019	226.5430	0.2164	42.1727	0.46158	151.6762	-1.8160	53.1220	0.40981	242.4951	-1.0613	61.5837
14	9	6	0.80215	183.4116	0.4104	37.9463	0.71028	142.3337	-0.5377	51.4600	0.57102	189.4072	-0.5903	54.4268
14	10	6	0.74213	239.1773	0.0030	43.3328	0.65319	108.1027	-1.0725	44.8470	0.62393	162.0400	-1.2466	50.3414
14	11	6	0.74036	232.5363	0.1181	42.7269	0.37753	184.5301	-3.6773	58.5934	0.43406	226.8897	-2.6990	59.5692
14	12	6	0.75403	226.1157	0.3881	42.1330	0.31056	188.8565	-1.4218	59.2764	0.41137	252.1569	-0.6488	62.7986
14	13	6	0.73338	239.8440	0.2720	43.3931	0.62763	112.2192	-0.3881	45.6929	0.39337	251.6858	-0.9331	62.7399
14	14	6	0.80330	183.1268	0.4449	37.9169	0.80564	98.0033	-0.1941	42.7008	0.60950	184.1951	-0.7108	53.6727
14	15	6	0.57213	349.7204	-0.6487	52.3983	0.71248	127.3309	0.2179	48.6723	0.45472	220.3730	-1.5262	58.7075
14	16	6	0.81715	171.3627	0.4551	36.6788	0.73181	124.8039	-1.0524	48.1870	0.33073	286.8721	-1.7754	66.9821
14	17	6	0.82592	164.4892	0.4959	35.9356	0.80347	92.1889	-0.1272	41.4147	0.61170	178.4192	-0.6307	52.8244
14	18	6	0.82066	168.2416	0.5532	36.3432	0.80526	113.4833	0.0816	45.9496	0.66548	167.7144	-0.1202	51.2153
14	19	6	0.71875	254.5516	0.2267	44.7038	0.60374	150.2888	-1.0562	52.8785	0.41255	277.9450	-0.6034	65.9316
14	20	6	0.83310	157.1317	0.5517	35.1227	0.82694	85.6553	0.1913	39.9202	0.64446	169.4435	0.0203	51.4786

Table 3-8: Correlation coefficient and cross validation parameters of model 16 with hidden nodes (5-20)

Mo.#	hn	nPCs	R_tr	PRESS_tr	R2CV_tr	RSEP_tr	R_test	PRESS_test	R2CV_test	RSEP_test	R_val	PRESS_val	R2CV_val	RSEP_val
16	5	6	0.72229	245.6204	0.12176	43.9126	0.36745	168.2919	-2.0874	55.9561	0.42989	231.3034	- 1.4981	60.1458
16	6	6	0.66070	290.6015	-0.11412	47.7645	0.69386	102.0568	-0.30198	43.5749	0.60871	180.1090	-0.77495	53.0740
16	7	6	0.81804	173.3804	0.37357	36.8941	0.85421	79.2825	-0.07786	38.4064	0.67386	150.0469	-0.31686	48.4426
16	8	6	0.81429	174.6599	0.40537	37.0299	0.61971	148.8169	-1.7095	52.6189	0.52863	200.1657	-0.59171	55.9512
16	9	6	0.63487	308.8249	-0.36210	49.2394	0.55770	122.3370	-0.91307	47.7083	0.52017	193.4179	-1.52546	55.0000
16	10	6	0.84088	151.5737	0.55335	34.4960	0.84024	69.2688	0.19661	35.8992	0.70119	147.6169	-0.02047	48.0488
16	11	6	0.82420	166.2533	0.44896	36.1278	0.80270	87.0154	-0.17479	40.2359	0.61955	181.9819	-0.11586	53.3492
16	12	6	0.84724	147.1060	0.59741	33.9838	0.80052	100.3499	-0.02733	43.2090	0.63652	169.3349	-0.48076	51.4621
16	13	6	0.67650	287.1009	-0.04198	47.4760	0.67131	104.6418	0.12072	44.1233	0.73787	138.8849	-0.36294	46.6060
16	14	6	0.84916	143.9715	0.59103	33.6198	0.82729	79.57338	0.12978	38.4768	0.63994	157.2467	-0.22235	49.5912
16	15	6	0.81279	177.1314	0.36497	37.2910	0.75953	103.2818	-1.0183	43.8356	0.45091	231.6704	-0.96511	60.1935
16	16	6	0.73238	241.6967	-0.12075	43.5604	0.17005	202.9744	-3.4819	61.4520	0.24423	279.8247	-2.03481	66.1542
16	17	6	0.84292	149.5061	0.62883	34.2599	0.81788	91.4879	0.21372	41.2570	0.68949	150.8313	-0.06679	48.5691
16	18	6	0.91157	87.6791	0.77417	26.2364	0.80445	108.6759	-0.10903	44.9658	0.67788	155.3797	0.28172	49.2960
16	19	6	0.77511	205.0582	0.32365	40.1232	0.57108	127.8923	-0.61684	48.7795	0.43317	248.1316	-0.69755	62.2953
16	20	6	0.83814	156.4284	0.46412	35.0440	0.85077	79.6822	0.08189	38.5031	0.63033	164.1115	-0.64220	50.6622

Table 3-9: Correlation coefficient and cross validation parameters of model 17 with hidden nodes (5-20)

Mo.#	hn	nPCs	R_tr	PRESS_tr	R2CV_tr	RSEP_tr	R_test	PRESS_test	R2CV_test	RSEP_test	R_val	PRESS_val	R2CV_val	RSEP_val
17	5	5	0.64577	305.6745	-0.94013	48.9876	0.80426	83.2702	-0.73831	39.3605	0.68722	156.4339	-2.4556	49.4629
17	6	5	0.66416	440.7048	-1.25446	58.8207	0.39429	145.9853	-2.41834	52.1159	0.54197	190.5388	-1.3452	54.5891
17	7	5	0.84044	151.5330	0.56913	34.4913	0.81623	96.6901	0.09946	42.4137	0.61108	169.7350	-0.81099	51.5228
17	8	5	0.79503	188.9730	0.42978	38.5173	0.63933	137.0632	-0.87797	50.4982	0.50047	221.3367	-0.27436	58.8357
17	9	5	0.65330	316.0515	0.16659	49.8122	0.72857	85.5780	0.13862	39.9022	0.59802	172.8524	-0.74573	51.9938
17	10	5	0.83617	155.5586	0.60033	34.9465	0.62352	152.0312	-0.50279	53.1841	0.57340	187.6643	-0.26952	54.1758
17	11	5	0.71338	252.1260	-0.00772	44.4903	0.69103	115.1864	-0.71964	46.2931	0.71921	151.5797	-0.99933	48.6894
17	12	5	0.76992	218.9339	0.08662	41.4584	0.61896	114.8009	-1.33220	46.2155	0.63114	157.9157	-0.66550	49.6966
17	13	5	0.90361	94.4966	0.76910	27.2373	0.80492	138.4780	-0.33500	50.7582	0.68404	159.4459	-0.22350	49.9368
17	14	5	0.82151	168.1023	0.45229	36.3281	0.81736	95.4500	-0.04931	42.1409	0.63549	167.4812	-0.72918	51.1796
17	15	5	0.83162	158.4690	0.53129	35.2719	0.82830	83.7453	-0.22528	39.4726	0.65173	167.1908	-0.47310	51.1352
17	16	5	0.81410	173.3517	0.49338	36.8910	0.83237	76.4917	0.29112	37.7244	0.64993	167.5848	-0.10286	51.1955
17	17	5	0.93825	61.8014	0.86407	22.0270	0.80986	92.2901	0.53198	41.4374	0.67539	148.5681	0.11318	48.2033
17	18	5	0.83991	152.1394	0.53195	34.5603	0.80264	102.6347	-0.07971	43.6981	0.65159	168.0263	-0.22191	51.2629
17	19	5	0.81019	176.5983	0.44442	37.2349	0.81310	90.0946	-0.21199	40.9416	0.65290	158.9096	-0.34817	49.8528
17	20	5	0.81755	170.3231	0.49152	36.5673	0.80905	89.4074	-0.09069	40.7852	0.62521	167.4861	-0.42994	51.1804

Table 3-10: Summary of correlation coefficient and cross validation parameters of the optimal number of hidden nodes for each model

Mo.#	hn	nPCs	R_tr	PRESS_tr	R2CV_tr	RSEP_tr	R_test	PRESS_test	R2CV_test	RSEP_test	R_val	PRESS_val	R2CV_val	RSEP_val
14	7	6	0.83288	163.8919	0.5192	35.8703	0.81114	84.0964	0.0316	39.5552	0.60846	173.9223	-0.3412	52.1545
16	7	6	0.81804	173.3804	0.37357	36.8941	0.85421	79.2825	-0.07786	38.4064	0.67386	150.0469	-0.31686	48.4426
16	10	6	0.84088	151.5737	0.55335	34.4960	0.84024	69.2688	0.19661	35.8992	0.70119	147.6169	-0.02047	48.0488
17	5	5	0.64577	305.6745	-0.94013	48.9876	0.80426	83.2702	-0.73831	39.3605	0.68722	156.4339	-2.4556	49.4629
17	9	5	0.65330	316.0515	0.16659	49.8122	0.72857	85.5780	0.13862	39.9022	0.59802	172.8524	-0.74573	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. The results of the test of the models; model 14 with hn 7, model 16 with hn 7, and model 17 with hn 5 are shown in Tables 3-11, 3-12 and 3-13 respectively. From the tables 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.

Table 3-11: Chance correlation result of model 14 with hn 7

Trial No.	hn	nPCs	R_tr	PRESS_tr	R2CV_tr	R_test	PRESS_test	R2CV_test	R_val	PRESS_val	R2CV_val
1	7	6	-0.35340	391.1267	-3.4934	0.14706	905.9826	-1.6198	0.24477	395.7684	-1.6836
2	7	6	0.15675	192.2030	-9.6360	-0.21211	752.7547	-5.2348	0.10414	295.4529	-6.2640
3	7	6	0.06030	615.8436	-5.4162	-0.28439	2055.9418	-4.5855	-0.40205	955.0102	-7.0999
4	7	6	-0.41321	314.9225	-6.6842	-0.44777	915.0712	-5.5108	-0.50719	446.2870	-10.0806
5	7	6	-0.35826	270.9327	-14.0544	-0.41005	750.4385	-12.7556	-0.37624	392.8095	-15.0512
6	7	6	-0.20431	292.8178	-7.0254	-0.31964	768.4146	-7.0003	-0.13115	373.7129	-6.2113
7	7	6	-0.48196	423.4074	-2.5675	-0.29074	1387.2103	-1.4717	-0.15387	519.6992	-1.6124
8	7	6	-0.11408	253.9217	-3.1061	0.28392	924.0196	-3.8563	-0.00495	334.3335	-3.1078
9	7	6	-0.46710	231.8775	-15.1228	0.21449	562.5747	-4.2598	-0.26421	326.2864	-10.0836
10	7	6	-0.11408	253.9217	-3.1061	-0.28392	924.0196	-3.8563	-0.00495	334.3335	-3.1078

Table 3-12: Chance correlation result of model 16 with hn 7

Trial No.	hn	nPCs	R_tr	PRESS_tr	R^2CV_tr	R_test	PRESS_test	R2CV_test	R_val	PRESS_val	R2CV_val
1	7	6	-0.25640	340.2921	-4.5544	0.26749	755.0782	-1.2548	-0.00922	497.5021	-2.9842
2	7	6	0.28843	533.3192	-4.5582	-0.16118	1808.9678	-3.6668	-0.07146	809.2680	-4.1176
3	7	6	-0.07985	214.5892	-22.7850	-0.33008	753.5146	-7.4861	0.00645	311.0815	-6.3129
4	7	6	0.23397	191.9620	-4.1360	0.45067	714.7772	-4.3344	0.64659	184.7672	-5.0898
5	7	6	0.08546	196.7719	-3.7714	0.35154	570.3164	-1.4581	0.43862	208.6667	-4.6916
6	7	6	-0.44789	242.8537	-12.6966	-0.01632	631.6679	-4.7583	-0.27693	365.6639	-7.2425
7	7	6	0.38815	163.2207	-15.3024	-0.24392	619.1597	-15.3775	-0.21100	336.2735	-13.3968
8	7	6	-0.12455	326.5620	-4.1954	-0.32634	971.1688	-4.4578	0.24972	324.1440	-2.6356
9	7	6	-0.28195	225.8472	-8.9736	-0.19596	745.2070	-4.8319	0.01213	278.6982	-8.7813
10	7	6	0.15721	185.8041	-4.1585	-0.25241	731.3327	-6.1927	-0.24132	367.7104	-6.9925

Table 3-13: Chance correlation result of model 17 with hn 5

Trial No.	hn	nPCs	R_tr	PRESS_tr	R2CV_tr	R_test	PRESS_test	R2CV_test	R_val	PRESS_val	R2CV_val
1	5	5	-0.07316	224.9415	-26.6313	-0.09641	642.9757	-7.2671	-0.00680	319.2681	-22.6326
2	5	5	-0.15115	293.9882	-3.1571	-0.15795	1386.0425	-1.0435	-0.14009	525.2266	-1.5540
3	5	5	-0.01062	252.5905	-2.4648	-0.00101	843.5534	-1.7481	-0.34184	516.9887	-3.0778
4	5	5	0.14753	176.4427	-15.7686	0.05227	540.6243	-14.0441	-0.04025	284.7066	-19.5725
5	5	5	0.27435	226.3401	-1.0223	0.22653	564.7042	-1.9235	0.57349	209.7798	-1.3166
6	5	5	0.27517	167.7000	-2.8668	0.25575	636.4812	-7.2466	0.42092	214.5437	-3.9469
7	5	5	-0.31034	354.8747	-2.3936	0.42901	511.2426	-0.8209	0.07881	328.0043	-2.2318
8	5	5	0.08131	214.6970	-5.6140	0.18284	563.6120	-4.1685	-0.19702	382.7281	-8.2662
9	5	5	-0.38747	343.2462	-3.2084	0.32822	710.8494	-0.5928	-0.01512	363.8943	-2.3913
10	5	5	-0.36572	292.9990	-14.1189	0.32651	527.5757	-9.2845	-0.13980	365.9283	-7.7939

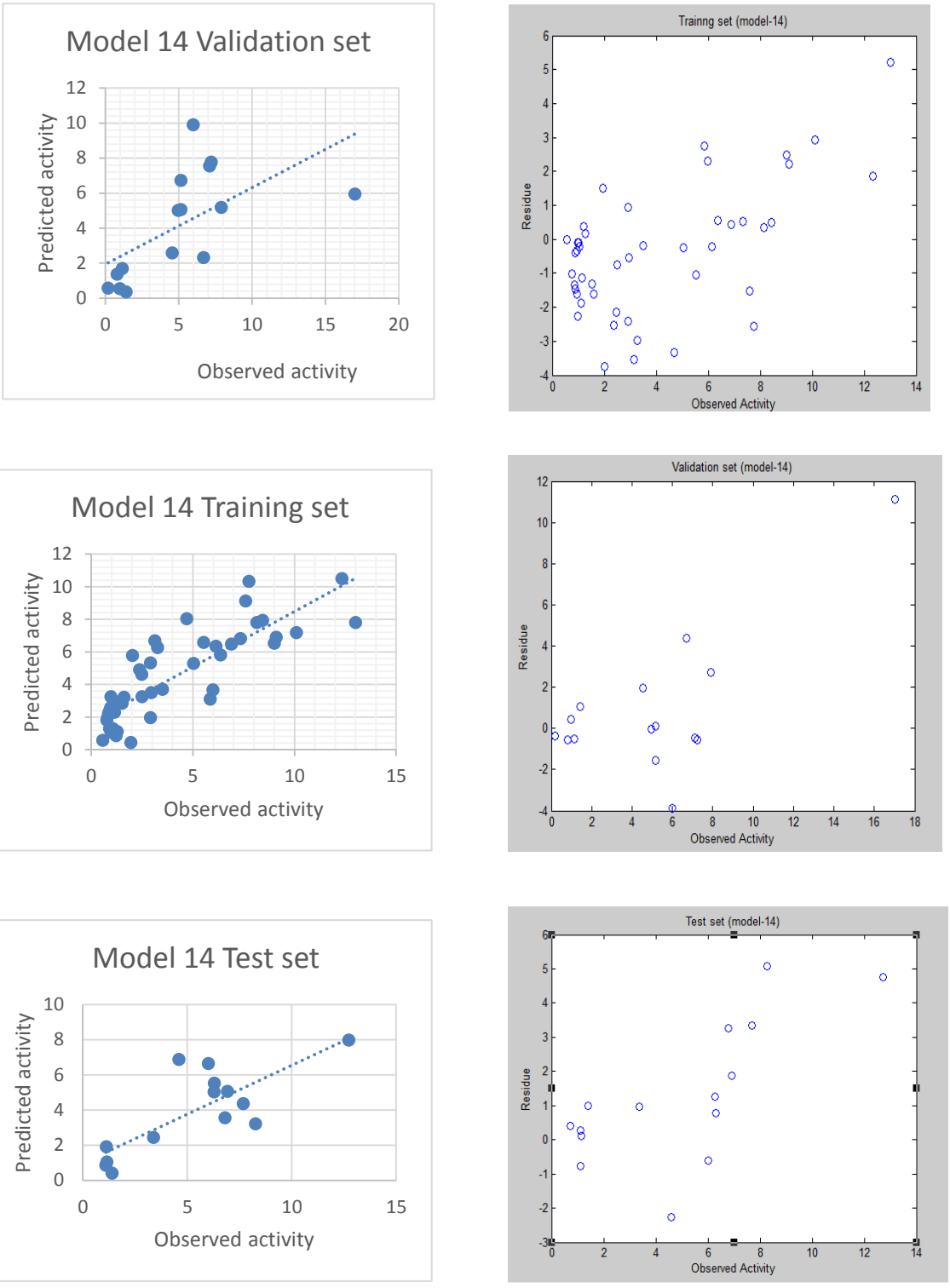


Figure 3-5: Plot of predicted activity against observed activity as well as their residual for model 14 using 7 hidden nodes. Training, validation, and test set.

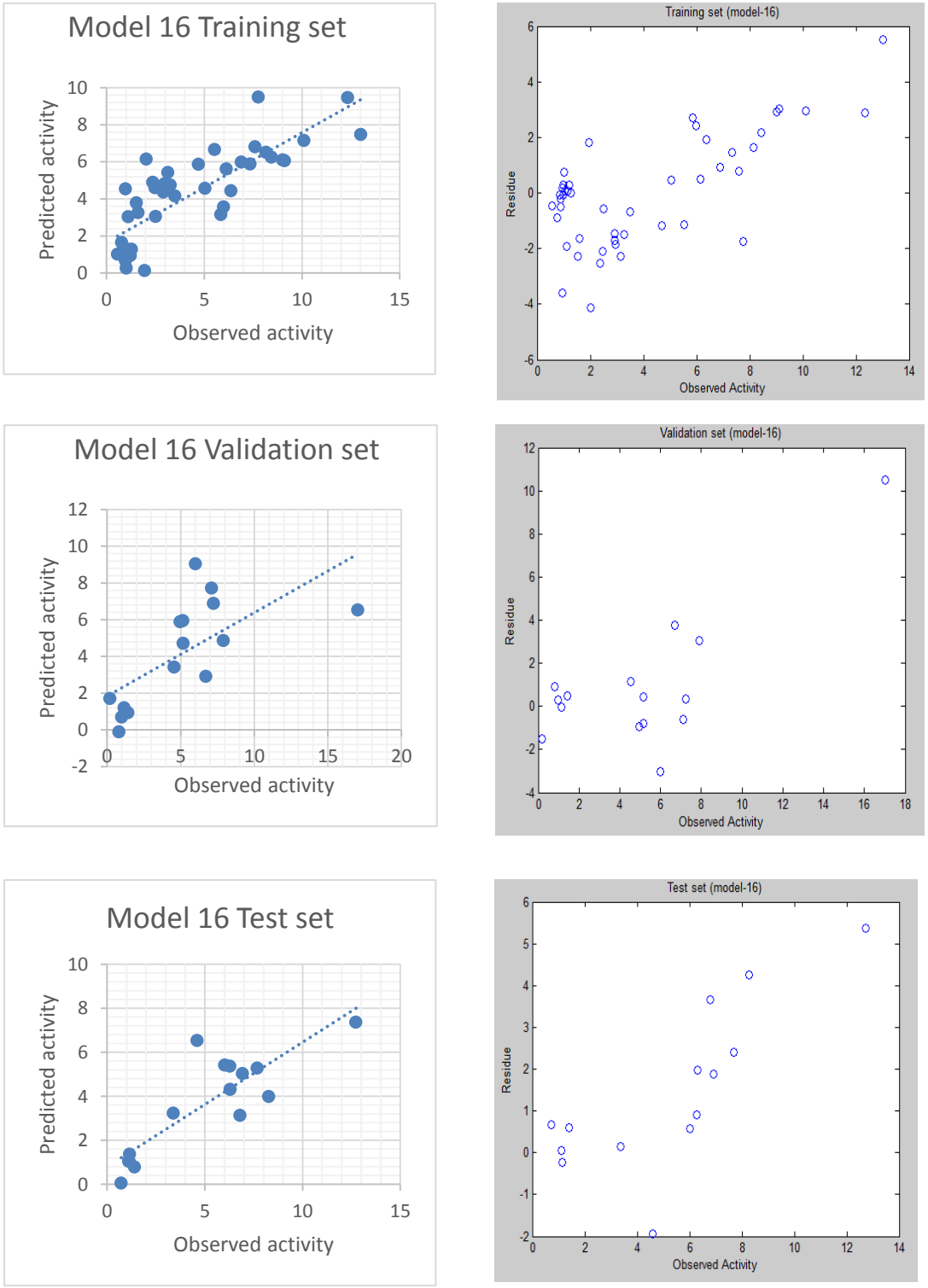
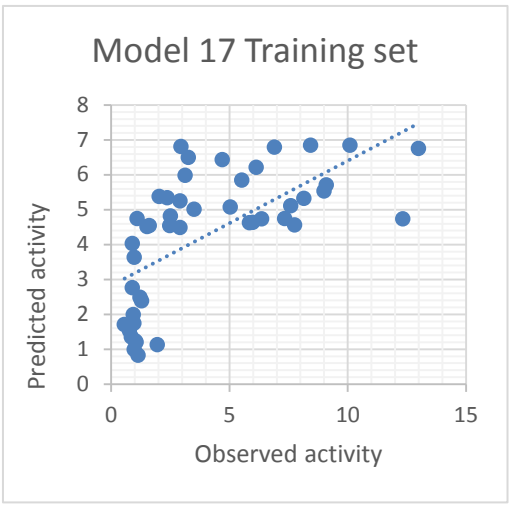
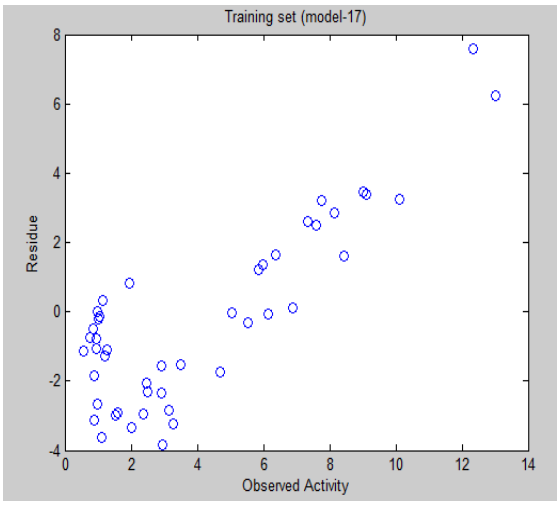
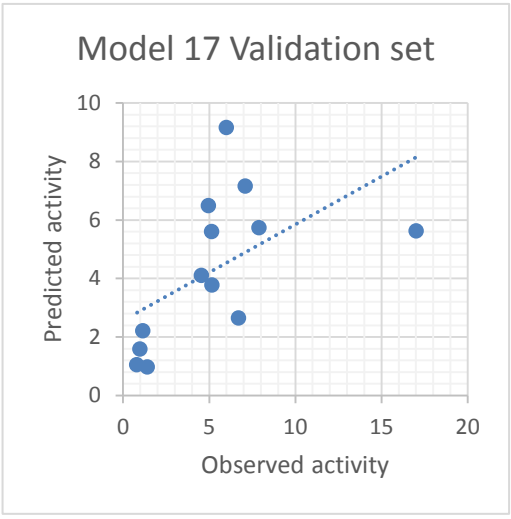
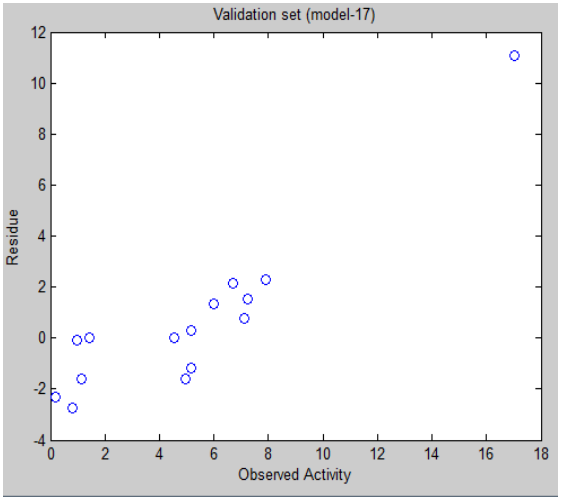


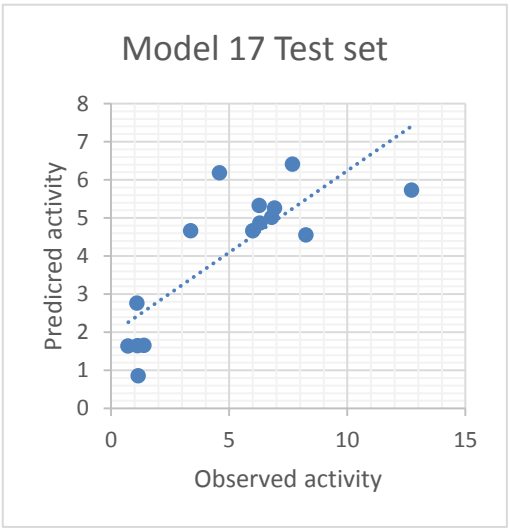
Figure 3-6: Plot of predicted activity against observed activity as well as their residual for model 16 using 7 hidden nodes. Training, validation, and test set.











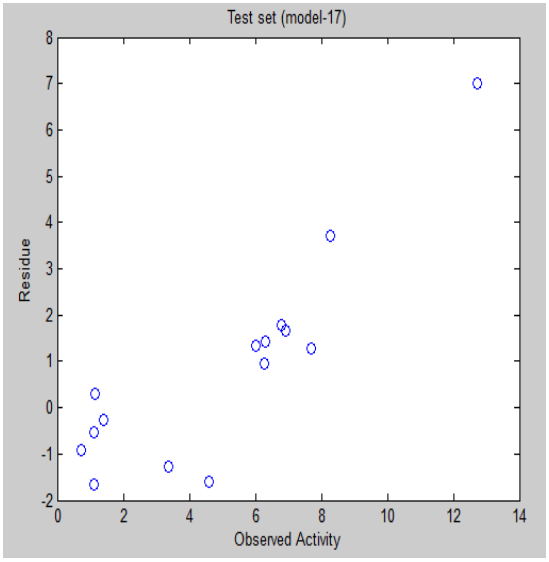


Figure 3-7: Plot of predicted activity against observed activity as well as their residual for model 17 using 5 hidden nodes. Training, validation, and test set.

The following conditions proposed by Golbraikh and Tropsha [68] were applied to conclude that the QSAR model has acceptable prediction power if:

1. $R^2cv > 0.5$
2. $R^2 > 0.6$
3. $(R^2 – R^2_0) / R^2 < 0.1$ and $0.85 < k < 1.15$

   Or

   $(R^2 – R`^2_0) / R^2 < 0.1$ and $0.85 < k` < 1.15$

Where R^2_0 and $R`^2_0$ are the coefficients of determination characterizing linear regression with Y-intercept at zero, the first associated with observed vs. predicted values, the second related to predicted vs. observed values; k and k` are the slopes of the regression lines forced through zero, relating observed vs. predicted and predicted vs. observed values.

4. $\left| R^2_0 – R`^2_0 \right| < 0.3$

Alternatively, the parameters R^2_m (R^{2*} (1- (R^2 - R^2_0)$^{1/2}$)) can be used. This parameters penalizes a model for large differences between observed and predicted values, was also calculated. R^2_m should be larger than 0.5 for a good external prediction.

If a model shows good statistical performance for all these criteria, on both the training and the test set, its reliability and robustness are high.

Model **17** validated according to these criteria, and shows to have acceptable prediction power.

Structure Activity Relationship of the Dataset

- ✓ **Compounds 1-27 in Table 2-1 SAR [62]:**

From the tables we note that the substitution on the main chemical structure affected the half maximal inhibition concentration (IC_{50}). The substituent may increase or decrease IC_{50} , according to substituent type and its position.

In Table 2-1; its noted that the functional groups that are substituted on the main chemical structure of quinoline affected IC_{50}, the pyrrolidine is better than pyrrole as functional group in compounds (03 and 06). Also compounds (12 and 14) the substituent morpholine have better inhibitory activity than piperidine, this due to the increase in the electronegativity of oxygen which decrease IC_{50}. Compounds (11 and 17) have the same functional group pyrrolidine, but they differ in the position of the other substituent (methoxy group -OCH_3), compound 17 have a better results in the inhibitory activity than compound 11 in which the methoxy groups in the ortho position and that increase the vader waals attraction, but in 11 the methoxy groups in the meta position. Compound 9 which have benzimidazole as functional group this decrease the IC_{50}.

- ✓ **Compounds 28-41 in Table 2-1 SAR [63]:**

These compounds have different IC_{50} according to the difference in the substituent which are nitrogen cyclic compounds. Its noted that the chlorine atom in general give better half maximal inhibition concentration than methoxy group. This due to the higher electronegativity of chlorine, which decrease IC_{50}. for example; compounds (34 and 41), they have the same functional 4-nitro pyrazole group but they differ in the other substituent, the chlorine atom give better results than the methoxy group, due to the higher electronegativty also.

- ✓ **Compounds 42-52 in Table 2-1 SAR [64]:**

From the table it's obvious that compounds (45 and 47) have decrease in the IC_{50}, these are have same functional group pyrrole, but compound 45 is better than 47, this because the three methoxy group which make it more bulky than compound 45 which have only one methoxy group.

- ✓ **Compounds 53-60 in Table 2-1 SAR [65]:**

In this table the compounds are differ in the substituent and also the IC_{50}, compounds (54 and 57) have the chlorine atom as substituent but in different position, compound 54 the chlorine in the ortho position, compound 57 in the para position and this is give better result. Compounds (59 and 60) they have different closely functional groups. But compound 60 have lower IC_{50} which have the nitro group, but compound 59 have an increase in IC_{50} which have amino group as functional group in comparison to compound 60, this because the oxygen increase the electronegativty.

- ✓ **Compounds 61-69 in Table 2-1 SAR [66]:**

According to the data in the table, there is a small differences in the IC_{50}. Compounds (61 and 65) they have the same substituent ethylene group, but differ in the other substituent. Compound 61 have triflouromethyl and give smaller IC_{50} than compound 65, which have tert-triflouromethyl phenyl as substituent, this mean that the phenyl group increase IC_{50} in small amount. Compounds (62 and 68) these compounds have the same two substituents triflouromethyl and tert-methyl phenyl group. But they differ in the main formula, compound 62 have better value of IC_{50} than compound 68 in which they replace the carbon atom by sulfur oxide which increase IC_{50}.

- ✓ **Compounds 70-79 in Table 2-1 SAR [67]:**

The rest of the compounds also differ in the IC_{50} due to the difference in the aryl group structure. Compounds (70 and 71) differ in the substituent, compound 70 have a benzene ring and higher IC_{50} than compound 71 which have chlorobenzene as substituent. Also compounds (77 and 78) they differ in the linkage position of pyridine, compound 77 have the higher IC_{50} which linked in the ortho position than compound 78 which linked in the para position.

- ✓ **Suggestion of new chemical structure with better activity than the available one**

According to the previous SAR and the resulted equation of MLR model, the QSAR of the antimalarial should have the following in ;

1. More number of imines group.
2. Less electonegativity
3. More lipophilicity
4. Less polarizability
5. More R-CR-R group
6. More Vander Waal attraction

Below suggested two compounds as antimalarial agents:

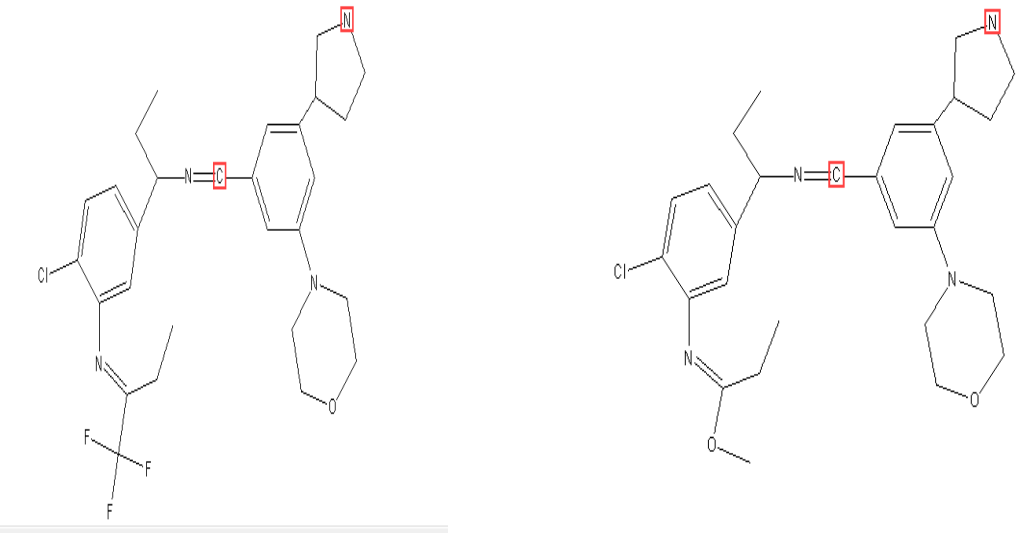


Figure 3-8: The chemical structure of the proposed compounds as antimalarial agents

According to model 17 MLR equation, the calculated IC_{50} of the suggested compounds are 7.057 and 3.336 µg/ml respectively. these values which calculated according to the resulted MLR equation are good and in the range of the IC_{50} values of the studied compounds. but its preferred to get smaller values than this to get a better design.

Comparison with previous 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 2-1, 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 [69]

- ✓ 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 $q2$ and conventional $r2$ values of 0.740 and 0.972 by the CoMFA model, 0.714 and 0.976 by the CoMSIA model, respectively [54].

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 agents. 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.

Chapter Four

Conclusions

Chapter Four

Conclusions:

A quantitative structure activity relationship analysis of 79 antimalarial compounds that are collected from literature and there 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. Which group a 17 descriptors, and the results was 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 choosed 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) choosed 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 choosed 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.

A new suggested compounds with IC_{50} 7.057 and 3.336 µg/ml for both compounds.

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. 80(3): 285-291(1986).

[12] D. Young "Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems", John Wiley & Sons , (2001).

[13] E.G. Lewars, Computational Chemistry, DOI 10.1007/978-90-481-3862-3_1, # Springer ScienceþBusiness Media B.V. (2011).

[14] http://www.shodor.org/chemviz/overview/ccbasics.html 29/9/2018.

[15] https://en.wikipedia.org/wiki/Computational_chemistry 3/10/2018.

[16] O. Deeb, D. Ekinci, Recent Applications of Quantitative Structure-Activity Relationships in Drug Design, Medicinal Chemistry and Drug Design,ISBN: 978-953-51-0513-8, (2012).

[17] A. Crum-Brown and T. R. Fraser, Trans. R. Soc. Edinburgh, 25, 151 (1868).

[18] C. Richet and C. R. Seances, Soc. Biol. Ses. Fil., 9, 775 (1893).

[19] H. Meyer, Arch. Exp. Pathol. Pharmakol., 42, 109 (1899).

[20] E. Overton, Studien Uber die Narkose, Fischer, Jena, Germany, (1901).

[21] J. Ferguson, Proc. R. Soc. London Ser. B, 127, 387 (1939).

[22] A. Albert, S. Rubbo, R. Goldacre, M. Darcy, and J. Stove, Br. J. Exp. Pathol., 26, 60 (1945).

[23] A. Albert, Selective Toxicity: The Physicochemical Bases of Therapy, 7th ed., Chapman and Hall, London, p. 33, (1985).

[24] P. H. Bell and R. O. Roblin, Jr. J. Am. Chem. Soc., 64, 2905 (1942).

[25] R. W. Taft, J. Am. Chem. Soc., 74, 3120 (1952).

[26] C. Hansch, P. P. Maloney, T. Fujita, and R. M. Muir, Nature, 194, 178 (1962).

[27] H. Kubinyi, Arzneim.-Forsch., 26, 1991 (1976).

[28] G. Klopman, J. Am. Chem. Soc., 106, 7315 (1984).

[29] B. W. Blake, K. Enslein, V. K. Gombar, and H. H. Borgstedt, Mutat. Res., 241, 261 (1990).

[30] Z. Simon, Angew. Chem. Int. Ed. Eng., 13, 719 (1974).

[31] L. H. Hall and L. B. Kier, J. Pharm. Sci., 66,642 (1977).

[32] W. Tong, D. R. Lowis, R. Perkins, Y. Chen, W. J. Welsh, D. W. Goddette, T. W. Heritage, and D. M. Sleehan, J. Chem. Inf. Comput Sci., 38, 669 (1998).

[33] S. J. Cho, W. Zheng, and A. Tropsha, Pac. Symp. Biocomput., 305 (1998).

[34] H. Gao and J. Bajorath, J. Mol. Diversity, 4, 115 (1999).

[35] H. Gao, C. Williams, P. Labute, and J. Bajorath, J. Chem. Inf. Comput. Sci., 39, 164 (1999).

[36] W. J. DunnIII, S. Wold, U. Edlund, S. Hellberg, and J. Gasteeger, Quant. Struct.-Act. Relat., 3, 131 (1984).

[37] C. polanco, Polarity index in proteins-Abioinformatics tool, benthan science publisher ,Sharjah, UAE, 978-1-68108-269-1 (2016).

[38] L. Xu, and W.J. Zhang, Comparison of different methods for variable selection. Anal.Chim. Acta, 446, 477–483(2001).

[39] D.C. Montgomery, and E.A.Peck, Introduction to linear regression analysis. Wiley: NewYork(1992).

[40] I.T. Jolliffe, Principal Component Analysis. Springer-Verlag: New York(1986).

[41] A. Krenker, J. Bešter and A. Kos, K. Suzuki, Introduction to the Artificial Neural Networks, Artificial Neural Networks - Methodological Advances and Biomedical Applications, ISBN: 978- 953-307-243-2, (2011).

[42] T. Ghafourian, M.T.D.Cronin, SAR QSAR Environ.Res., 16, 171-190 (2005).

[43] K. Roy, J.T. Leonard,. QSAR Comb. Sci., 25, 235- 251(2006).

[44] J. J. Shao, Am. Stat. Assoc., 88, 486-494 (1993).

[45] E. J. Besal, Math. Chem., 29, 191-195 (2001).

[46] S. Wold, L. Ericksson, Partial least squares projections to latent structures (PLS) in chemistry. In Encyclopedia of computationalchemistry, Ragu & Schleyer, P. (ed.), John Wiley & Sons, Chichester, Vol. 3, 2006–2021, (1998).

[47] A. Yasri, D. J. Hartsough, Chem. Inf. Comput. Sci., 41, 1218-1227 (2001).

[48] R. Guha, P.C. Jurs, J. Chem. Inf. Model., 45, 65-73 (2005).

[49] R. Todeschini, M. Lasagni, and E. Marengo, New Molecular Descriptors for 2D- and 3Dstructures. Theory. J. Chemometrics, 8, 263-273 (1994).

[50] Talete srl, Dragon (ver. 5.4), Milano, Italy. Web site: www.talete.mi.it/products/software.htm:

[51] https://www.researchgate.net/publication/311101660_SPSS_software [Oct 28 2018].

[52] The MathWorks Inc. MATLAB 7.0 (R14SP2). The MathWorks Inc., (2005).

[53] V. K. Agrawal, R. Srivastava and P. V. Khadikarb, QSAR Studies on Some Antimalarial Sulfonamides ; Bioorganic & Medicinal Chemistry 9 3287–3293(2001).

[54] 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).

[55] L. F. Motta, A. C.Gaudio ,Y. Takahata , Quantitative Structure–Activity Relationships of a Series of Chalcone Derivatives (1,3–Diphenyl–2–propen–1–one) as Anti Plasmodium falciparum Agents (Anti Malaria Agents), Internet Electronic Journal of Molecular Design, 5, 555–569;2006.

[56] O. Deeb, S. Alfalah, M. P. Freitas, E. da Cunha, T. C. Ramalho ;Exploring MIA-QSARs for farnesyltransferase inhibitory effect of antimalarial compounds refined by docking simulations, Journal of Biophysical Chemistry 3 58-71(2012).

[57] K. NITENDRA SAHU,QSAR STUDY OF SOME SUBSTITUTED 4-QUINOLINYL AND 9 ACRIDINYL HYDRAZONES AS ANTIMALARIAL AGENTS, Acta Poloniae Pharmaceutica, Drug Research, Vol. 69 No. 6 pp. 1153.1165, (2012).

[58] A. Worachartcheewan, C. Nantasenamat, C. Isarankura, V. Prachayasittikul, QSAR study of amidino bis-benzimidazole derivatives as potent anti-malarial agents against Plasmodium falciparum,chemical papers, Volume 67, Issue 11, pp 1462–1473 November (2013).

[59] R. Hadanu, S. Idris, W. Sutapa, QSAR ANALYSIS OF BENZOTHIAZOLE DERIVATIVES OF ANTIMALARIAL COMPOUNDS BASED ON AM1 SEMI-EMPIRICAL METHOD, Indones. J. Chem., 15 (1), 86 – 92,(2015).

[60] A. Jarrahpour, M. Aye, J. A. Rad , S.Y.inejad, V. Sinou, C. Latour, J. M. Brunel, E. Turos.Design, synthesis, activity evaluation and QSAR studies of novel antimalarial 1,2,3‐triazolo‐β‐lactam derivatives, Journal of the Iranian Chemical Society, pp 1–16(2018).

[61] QSAR Studies of Nitrobenzothiazole Derivatives as Antimalarial Agents. Available from:https://www.researchgate.net/publication/324158196_QSAR_Studies_of_Nitrobenzothiazole_Derivatives_as_Antimalarial_Agents [Nov 04 2018].

[62] 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).

[63] 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).

[64] 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).

[65] 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).

[66] 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).

[67] M. Zavri, N. Kawthekar; International Journal of Current Pharmaceutical Research ISSN- 0975-7066 Vol 9, Issue 3 (2017).

[68] A. Golbraikh,and A. Tropsha, Beware of q2! Journal of Molecular Graphics and Modelling,.20(4): p.269-276 (2002).

[69] 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).

دراسة العلاقة الكمية بين الفاعلية و الصيغة البنائية لبعض المركبات التي لها فعالية ضد الملاريا باستخدام طريقتي MLR وPC-ANN

إعداد: الاء إسحق حسن عاشور

إشراف: أ.د عمر ديب

الملخص:

يعد مرض الملاريا احد المشاكل الصحية الرئيسية التي تزيد الوفيات سنويا.والى الان لا يوجد تطعيم او دواء فعال ضدها بسبب كثرة الطفرات التي يحدثها العائل. لذلك تمت دراسة العلاقة الكمية بين فعالية 79 مركب كعوامل مضادة للملاريا والصيغة البنائية لهم. تم تطوير نماذج QSAR باستخدام الانحدار الخطي المتعدد (MLR) كطريقة خطية. بينما استخدم المكون الأساسي – الشبكة العصبية الاصطناعية (PC-ANN) كطريقة غير خطية. النماذج الناتجة كانت لديها قوة تنبؤ جيدة. حيث نتج من طريقة MLR النماذج (13–17) التي لها المعامل R^2> 0.6 ، وكان أفضل نموذج هو رقم 17 مع معامل الارتباط R = 0.889، R^2 = 0.791 ، و R2adj. = 0.733.

تم التحقق من قدرة النماذج على التنبؤ باستخدام طريقتي LOO و LMO ، وأظهرت النماذج 13–17 قدرة تنبؤية جيدة. وتم إجراء PCA لتقسيم البيانات إلى ثلاث مجموعات. ثم استخدمت ANN على النماذج المختارة 13–17.

تم التحقق من صحة نماذج ANN الناتجة عن طريق اختبار التوزيع العشوائي ، ثم تم تطبيق الشروط التي اقترحها Golbraikh و Tropsha للتأكد من قوة نماذج QSAR في القدرة على التنبؤ. ومع ذلك، كان أفضل نموذج ANN مع أفضل قوة تنبؤية هو النموذج رقم 17، مع قيمة معامل ارتباط R 0.8138 للمجموعة الضابطة test set.وتم اقتراح مركبين لهما فعالية ضد الملاريا تساوي 7.057

و 3.336 µg/ml. ومن خلال النتائج السابقة أصبح من الممكن اقتراح مركبات جديدة لها فعالية من خلا تطبيق معادلة افضل نموذج الناتج من MLR.