Al- Quds University

Deanship of Graduate Studies

Applied and Industrial Technology

Thesis Approval

Exploring QSARs for inhibiting activity of epidermal growth factor receptor (EGFR) tyrosine kinase by MLR and PC-ANN

Prepared by: Manal (M.Wael) AbdelHafez Muhtaseb

Registration No: 21220212

Supervisor: Prof. Omar Deeb

Master Thesis submitted and accepted Date 6 / 5/ 2017, The names and signatures of the examining committee members are as follows:

1- Head of Committee: Prof. Omar Deeb signature…….

2- Internal Examiner: Dr. Hatem Hejaz signature…….

3- External Examiner: Dr. Nasr Shraim signature…….

Jerusalem- Palestine

1438 /2017

Faculty of Graduate Studies

Al- Quds University

Applied & Industrial Technology program

Exploring QSARs for inhibiting activity of Epidermal growth factor receptor (EGFR) tyrosine kinase by MLR and PC-ANN

Manal Muhtaseb

Prof. Omar Deeb

M. Sc. Thesis

Winter 2017

Exploring QSARs for inhibiting activity of Epidermal growth factor receptor (EGFR) tyrosine kinase by MLR and PC-ANN

Prepared By :

Manal Muhtaseb

B. Sc. Chemistry

Al-Quds University (Palestine)

Supervisor: Prof. Omar Deeb

A Thesis Submitted in Partial Fulfillment of Requirements for the Degree of Master of Applied and Industrial Technology Program for Postgraduate studies in Applied and Industrial Technology

Faculty of Science and Technology

Al-Quds University

1438/2017

Al- Quds University

Deanship of Graduate Studies

Applied and Industrial Technology

Department of Science and Technology

Thesis Approval

Exploring QSARs for inhibiting activity of Epidermal growth factor receptor (EGFR) tyrosine kinase by MLR and PC-ANN

Prepared by: Manal Muhtaseb

Registration number: 21220212

Supervisor : Prof. Omar Deeb

Master Thesis submitted and accepted Date / / 2017

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

1- Head of Committee: Prof. Omar Deeb signature…….

2- Internal Examiner: signature…….

3- External Examiner: signature…….

Jerusalem- Palestine

1438 /2017

Dedication

To all of my family, especially my parents for their unconditioned support and love, my husband for the constant support and understanding.

Declaration :

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

Signed ……..

Manal Muhammad Waal Muhtaseb

Date……

Acknowledgments

Above all I would like to express my endless thanks to God for His blessing, and because of him I Made this through against all challenges.

Also I would like to thank my supervisor Prof. Omar Deeb for his guidance, encouragement, patience, advice, during the different stages of this research .

Exploring QSARs for inhibiting activity ofEpidermal growth factor receptor (EGFR) tyrosine kinase by MLR and PC-ANN

Abstract

Quantitative structure- activity relationship (QSAR) study was preformed to understand the activity of a set of 113 compounds of Epidermal Growth Factor Receptor (EGFR) inhibitors.

QSAR models were developed using multiple linear regression (MLR) as linear method. While principle component- artificial neural network (PC-ANN) modeling method was performed as nonlinear method.

The MLR resulted with models (12-23) which have coefficient of determination (R^2)>0.6, the best model (model 23) resulted with correlation coefficient (R) = 0.878, coefficient of determination (R^2) =0.771, and adjusted coefficient of determination (R^2adj) =0.719. Cross validation leave one out (LOO) and leave many out (LMO) were performed on the resulted MLR models, models 19-23 showed a good predictive power. After that principle component analysis (PCA) performed to divide the data into three data sets. Then the ANN performed on the chosen models (19-23) from leave one out (LOO) and leave many out (LMO) cross validation. ANN resulted models were validated through randomization test.

The best ANN model with good predictive power was model 19 with R=0.812 for the test set.

Table of contents

List of Tables:

List of Figures:

List of ABBREVIATIONS :

AM1	Austin Model 1
ANN	Artificial Neural Networks
COMFA	Comparative molecular field analysis
COMSIA	Comparative molecular similarity indices analysis
EGFR	Epidermal growth factor receptor
HER	Human growth factor
Hn	Hidden nodes.
HOMO	Highest occupied molecular orbital
LMO	Leave many out
LOO	Leave one out
LUMO	Lowest unoccupied molecular orbital
MLR	Multiple Linear regression
NSCLC	Non-small cell lung cancer
PCA	Principle component analysis
PIC_{50}	Half maximal inhibitory concentration
PRESS	Predicted residual sum of squares
PSE	Predictive square errors
PE	Processing elements
PDB	Brookhaven protein data bank
QSAR	Quantitative structure activity relationship
QSPR	Quantitative structure property relationships
R	Correlation coefficient

R^2	Coefficient of Determination
R^2adj	Adjusted $R^{2.}$
R^2cv or Q^2	Cross- validation coefficient of determination
RHF	Restricted Hartree-Fock
SPSS	Statistical package for social science
QSAR	Quantitative structure activity relationships
RMSE	Root mean- squared error
RSEP	Relative standard error of prediction
SPRESS	Uncertainty of prediction
SST	Total sum of squares
SSE	Error sum of squares

Chapter one

Introduction

1.Introduction :

1.1 Computational and theoretical chemistry:

Computational quantum chemistry has been in development for almost nine decades [1]. High quality, original reports have been published in computational and theoretical chemistry including those thatdeal with problems of structure, properties, energetic, weak interactions, reaction mechanisms, catalysis, and reaction rates involving atoms, molecules, clusters, surfaces, and bulk matter [2].

Computational chemistry is rapidly emerging as a subfield of theoretical chemistry, where the primary focus is on solving chemically related problems by calculations [3].

For the newcomer to the field, there are three main problems:

1. Deciphering the code. The language of computational chemistry is littered with acronyms, what do these abbreviations stand for in terms of underlying assumptions and approximations?
2. Technical problems. How does one actually run the program and what to look for in the output?
3. Quality assessment. How good is the number that has been calculated? The quantum and classical mechanics as well as statistical physics and thermodynamics are the foundation for most of the computational chemistry theory and computer programs [3].

This branch of calculation is based primarily on Schrödinger's equation (Equation (1-1)) [4] and include :

1-Calculation of electron and charge distribution

2-Molecular geometry in ground and excited states

3-Potential energy surface

4-Rate constants for elementary reactions

5- Details of the dynamics of molecular collisions

Hψ= Eψequation (1.1)

Where ,H: Hamiltonian operator

Ψ: psi,the wave function

E: total energy of the system

But the biggest mistake that a computational chemists can make is to assume that any computed number is exact. However, just as not all spectra are perfectly resolved, often a qualitative or approximate computation can give useful insight into chemistry if you understand what it tells you and what it doesn't [5] .

The computational studies can give better result when many experimental data are available, providing a strong background for the calculation. So if we have an experimental data of molecules which have shown good activity, we can know the groups which are responsible for the activity by doing calculations on these molecules using softwares, this relation between the structure and activity is called QSAR .

1.2 Quantitative structure activity relationships (QSAR) :

1.2.1 QSAR history

The evolution of QSAR is traced from the insightful observations of Crum-Brown and Frazier to Hammett's critical equations and finally Hansch's seminal contributions on hydrophobicity and modeling of biological activity based on extrathermodynamic principles [6].

QSAR has its origins in the field of toxicology whereby Cros in 1863 proposed a relationship which existed between the toxicity of primary aliphatic alcohols with their water solubility, likewise, Crum-Brown and Fraser (1868-1869) postulated the linkage between chemical constitution and physiological action in their pioneering investigation [7].

Meyer (1899), and Overton (1895) independently suggested that the narcotic (depressant) action of a group of organic compounds paralleled their olive oil/water partition coefficients [8,9].

Fieser, an eminent organic chemist of the mid-1900s, showed graphically the relationship between the antimalarial potency of naphthoquinones and their ether-water distribution coefficients. He also observed a constant optimum lipophilicity for different series of molecules [10].

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 [11].

Although Kauzmann’s 1959 prompted biochemists to endorse the central role of hydrophobicity in determining protein structure [12].

In the following years on the physical organic front the seminal work of Hammett gave rise to “sigma-rho”[13].

In 1956 Taft proposed an approach for separating polar, steric, and resonance effects of substituents in aliphatic compounds . The contributions from Hammett and Taft set forth the mechanistic basis for QSAR/QSPR development by Hansch and Fujita (1964) in their seminal development of the linear Hansch equation which integrated hydrophobic parameters with Hammett's electronic constants [14].

An early example of QSAR in drug design involves a series of 1-(X-phenyl)-3,3-dialkyl triazenes, these compounds are the interest of their anti-tumor activity but also were mutagenic [15].

A lot of QSARs studies done in AL-Quds computational chemistry laboratory in the last few years, these studies done to predict compounds properties , including biological activity, physical properties [16-18].

1.2.2 QSAR overview :

Quantitative structure–activity relationship (QSAR) modeling has matured over the past 50 years and has been very useful in discovering and optimizing drug leads [19].

Experimental methods based on receptors and other biological materials of human, rat, mouse and so on at least have been available for screening the biological activity of compounds, they are too costly and time-consuming [20]. Computational methods, especially quantitative structure-activity relationship (QSAR) analysis, provides an effective and powerful tool for achieving the same destination with much lower cost [21].

QSAR are mathematical model which relates the physico-chemical property / biological activity of compounds to their chemical structures [22,23]. This model is based on changes in molecular structure that would reflect changes in observed biological activity so it involves chemistry biology and statistics field for analysis [24].

QSAR has now been extensively applied to predict compounds' properties, including biological activity, physical property and even toxicity [25-28].

The biological activity of the candidate drug molecules can be predicted before the actual chemical synthesis. The prediction is based on the structural descriptors which contribute to the biological activity [29].

1.2.3 QSAR advantages [24]:

• Model gives better understanding about the interaction or reaction between molecules and its activity.

• It can provide useful information about biological effect of the compound which would help in drug research.

• It can also be used to predict the property or activity of the compound before synthesis. It mainly helps in reducing or replacing the molecule taken for testing in wet lab.

• It is becoming more useful and reliable. Computer based mathematical QSAR/QSPR model are based on chemical information extracted based on chemical structure not based on experimental values.

Hansch pioneered this field by demonstrating that the biological activities of drug molecules can be correlated to a few variables (properties) using simple regression equation (Equation 1-2) [30].

Log (1/C)= a (lipophilic descriptor) + b (electronic descriptor) + c (steric descriptor) + d (other descriptor) + etc…….(1.2).

Where,

1/C =Measure of biological activity.

a,b,c,etc.=Regression coefficients.

1.3 QSAR model development steps

QSAR model development has three steps :

1. Data preparation .
2. Data analysis .
3. Model validation [31].

1.3.1 Data preparation

Data preparation is an important step in presenting QSAR. It starts by selection of the data set to be used (compounds and their activities) which taken from literature. After this step we should do geometry optimization for the compounds we have chosen. The geometry optimization which would be done by software as HyperChem could find coordinates that represent the minimum potential energy of the molecular structure in its 3D form .

computational optimization, modeling and simulation form an integrated part of the modern design practice in engineering and industry, to minimize the cost and energy consumption, and to maximize the performance, profits and efficiency can be crucially important in all designs [32].

There are two types of molecular modeling - molecular mechanics and quantum mechanics.

Molecular mechanics: a classical mechanical model that represents a molecule as a group of atoms held together by elastic bonds. Molecular mechanics methods give predictions of molecular geometries.

Quantum mechanics: a quantum mechanical model of the electronic structure of a molecule, which involves solving the Schrödinger equation. Quantum mechanics can be used to predict electronic properties of molecules, such as dipole moments and spectroscopy [33].

Quantum mechanics divided into two methods of calculations :

1- Ab initio ,the term is Latin for (from scratch) it is first used by Robert Parr and coworkers . ab initio is a group of methods in which molecular structure can be calculated using nothing but Schrodinger equation [34]. Ab initio calculations give the absolute energy of the system of fixed nuclei and moving electrons [34].

2- Semi–empirical techniques use approximation from empirical (experimental) data to provide the input into the mathematical models. it is faster than ab initio [35].

In our study we have used semi-empirical method for geometry optimization , HyperChem and dragon software will be used to calculate all descriptors (properties).

1.3.2 Data analysis

Many steps uses to build a model which have a correlation between the endpoint and certain descriptors. If the correlation models to be built are linear then the multi linear regression (MLR) is used, however if it is nonlinear then the artificial neural network (ANN) is used after MLR .

1.3.2.1 Linear models

The most widely used in QSAR analysis is multiple linear regression analysis, which is a powerful means for establishing a correlation between independent variables and dependent variable such as biological activity [13].

MLR models are extremely powerful, and have the power to empirically tease out very complicated relationships between variables. Generally speaking, the technique is useful in helping explain observations of a dependent variable, usually denoted *y,* with observed values of more than one independent variables, usually denoted $x_1, x_2, ...$ A key feature of all regression models is the error term (fitted correlation coefficient r) , which is included to capture sources of error that are not captured by other variables. Linear regression models have been heavily studied, and are very well-understood [36].

The relationship between the dependent variable and independent variables represented by the following equation (equation1.3):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i \quad (\text{equation 1.3})$$

Where:

β_0 is the constant term and β_1 to β_p are the coefficients relating the independent variables to the variable of interest. $e_{i:}$ is an error term.

The term 'linear' is used because in multiple linear regression we assume that y is directly related to a linear combination of the independent variables [36].

1.3.2.2 Nonlinear model

Principle component analysis(PCA)

PCA which is a useful tool for reducing the number of variables in a data set and for obtaining useful two dimensional views of a multi-dimensional data set. Thus, irrelevant and unstable information is discarded from the regression analysis. Principal component-artificial neural network (PC-ANN) joins PCA with artificial neural networks (ANN), the flexibility of ANN for finding out relationships that are more complex allows this method to be widely applied in QSAR studies [37].

Artificial Neural Networks(ANN)

ANN is biologically inspired prediction methods based on the architecture of a network of neurons. This method fall into the category of feed-forward networks, in which, during the prediction, the information flows only in direction from the input descriptors, through a set of layers, to the output of the network [36].

An ANN is formed from hundreds of single units, artificial neurons or processing elements (PE), connected with coefficients (weights), which constitute the neural structure and are

organized in layers. The power of neural computations comes from connecting neurons in a network.

Each PE has weighted inputs, transfer function and one output. The behavior of a neural network is determined by the transfer functions of its neurons, by the learning rule, and by the architecture itself. The weights are the adjustable parameters and, in that sense, a neural network is a parameterized systemThe weighed sum of the inputs constitutes the activation of the neuron (figure 1.1) [38].

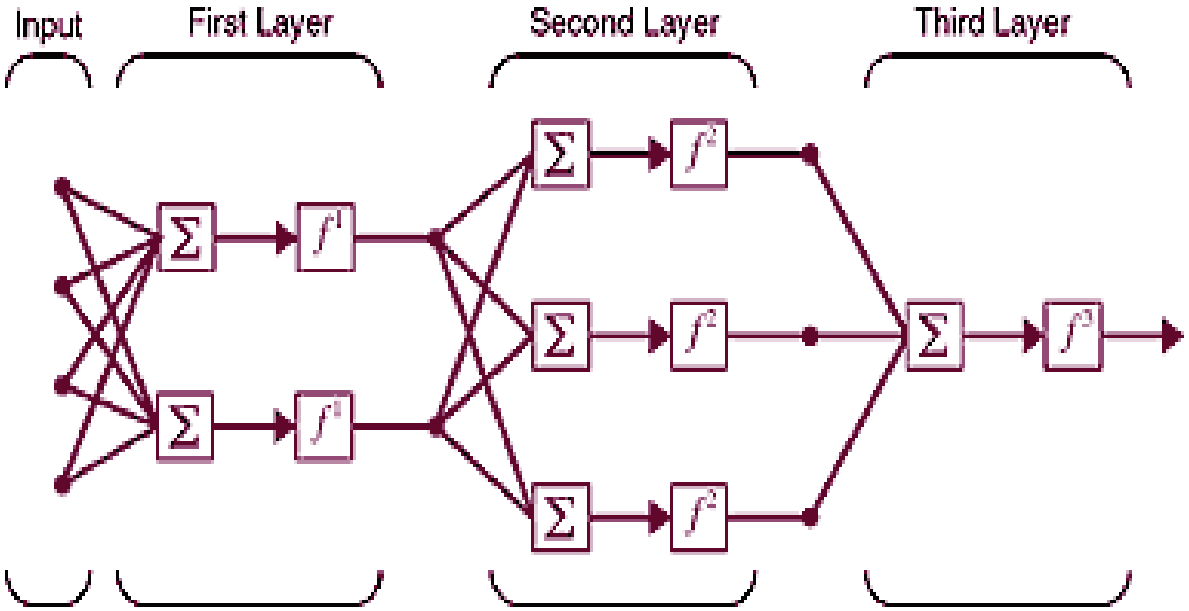


Figure (1.1): Artificial Neural Network(Here is an example of a simple three- layer, feed-forward network. It has four inputs, two neurons in the first layer, three in the second layer, and one in the third (output) layer. There is a connection from each neuron to all the neurons in the previous layer, and each connection has a weight associated with it. The neuron takes signals from previous layers, multiplies each signal by the connection's weight, and adds them together. The neuron then passes the sum through a transfer function. The result becomes the neuron's output [38]).

Neural network simulation appear to be a recent development, however , research in the field of neural networks has been attracting attention before many years . Since 1943, when Warren McCulloch and Walter Pitts presented the first model of artificial neurons, new and more sophisticated proposals have been made from decade to decade.

Neural network strength lies in their ability to make sense out of complex, noisy, or nonlinear data. And can provide robust solutions to problems in a wide range of disciplines, particularly areas involving classification, prediction, filtering, optimization, pattern recognition, and function approximation [39].

The potential applications of ANN methodology in the pharmaceutical sciences range from interpretation of analytical data, drug and dosage form design through biopharmacy to clinical pharmacy [38].

1.3.3 Model validation

In the last years, external validation of QSAR models was the subject of intensive debate in the scientific literature. Different groups have proposed different metrics to find “the best” parameter to characterize the external predictivity of a QSAR model [40].

QSAR is based on the hypothesis that changes in molecular structure reflect changes in the observed response or biological activity. The success of any QSAR model depends on the accuracy of the input data, selection of appropriate descriptors, statistical tools and the validation of the developed model. Validation is a crucial aspect of QSAR modeling. Validation is the process by which the reliability and significance of a procedure are established for a specific purpose [41].

QSAR model validation performed either by using the data that created the model (an internal validation) or by using a separated data set (an external validation). The internal validation are: least squares fit (R^2), cross validation (Q^2),adjusted R^2 [42].Chi-squared test

(x^2), root mean-squared error (RMSE), bootstrapping and scrambling (Y-Randomization) [43].

An external method performed by comparing the predicted and observed activities of an external test set of compounds that were not used in model development .

In current research we will do an external validation and an internal validation method,cross validation and scrambling (Y-Randomization).

Cross validation

Internal validation of QSAR model such as cross validation (CV,Q^2,$q^{2,}$ or jack –knifing) have a simple process which repeats the regression many times on subset of data usually each molecule is left out once (leave one out ,LOO) in turn. Sometimes more than one molecule (leave many out ,LMO) is left out at a time .

Leave-one-out is one of the best known methods. The goal of it to obtain as honest estimation as possible about the classification accuracy of the system.

The most common outcome parameters resulted from cross validation procedure are cross-validation coefficient q^2 (R^2cv) and root mean square error (RMSE) (equation 1.4)

High R^2cv and low RMSE values is a result of good and more predictive model and better description of the observed data .

$$RMSE = PSE = \sqrt{\frac{PRESS}{n}}$$ (equation1.4)

Where PRESS:Prediction error sum of squares,n:is compounds number

Prediction error sum of squares (PRESS) is a standard index to measure the accuracy of a modeling method based on the cross-validation technique. Based on the PRESS and SST (sum of squares of deviations of the experimental values from their mean) statistics,

the R^2, R^2cv (or Q^2) and S_{PRESS} values can be calculated easily. PRESS and Q^2 have good properties, which render them, appropriate for statistical testing with critical distributions. (figure1.2)

The formulas used to calculate all the mentioned statistics

$$R^2 = \frac{SSR}{SST} = 1 - \frac{PRESS}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_{obs} - y_{pred})^2}{\sum_{i=1}^{n}(y_{obs} - \overline{y}_{obs})^2}$$

are: $R_A^2 = \frac{R^2 - (k-1)}{(n-k) * (1 - R^2)}$

$$R_{CV}^2 (Q^2) = 1 - \frac{PRESS}{SSR} = 1 - \frac{\sum_{i=1}^{n}(y_{obs} - y_{pred})^2}{\sum_{i=1}^{n}(y_{pred} - \overline{y}_{pred})^2}$$

Figure(1.2) : Cross validation equation ,where PRESS is the predictive residual sum of the squares ,y_{obs} is the experimental activity for the individual compound in the training set , y_{pred} is the predictive activity for the compound in the training set, k is descriptors number in the regression model ,and ^2_A (adjusted coefficient of determination).

For a good prediction ability of the model , many authors consider high Q^2 (for instance $Q^2 > 0.5$) as an indicator or even as the ultimate proof of a high prediction power of a QSAR model [41].

Y-randomization (chance correlation) test:

Randomization test is the second internal validation test performed in this research, which we have to do it to ensure that the model is not due to a chance .

This test performed by randomization of the dependent variables , in which the set of activity values is reassigned randomly to different molecules and repeating the entire modeling procedure .This process is repeated many times [41].

The new QSAR models (after several repetitions) are expected to have low R^2 and R^2_{CV} values. If the opposite happens, then an acceptable QSAR model cannot be obtained for the specific modeling method and data, or simply it is obtained by chance.

1.4 Software used in QSAR process

A lot of software used in development of QSAR models ,these include software for drawing chemical structures , drawing 3D structures , also software to calculate descriptors and software for data analysis

In our research, four software were used :

1- HyperChem (version 8.3 HyperChem ,Inc),

2-Dragon software (vergion 2.1 ,Todeschini , R , Milano Chemometrics and QSAR Group)

3-Statistical package such as SPSS (version 13 ,SPSS Inc .).

4- MATLAB (version 7.0.1,Mathworks Inc) .

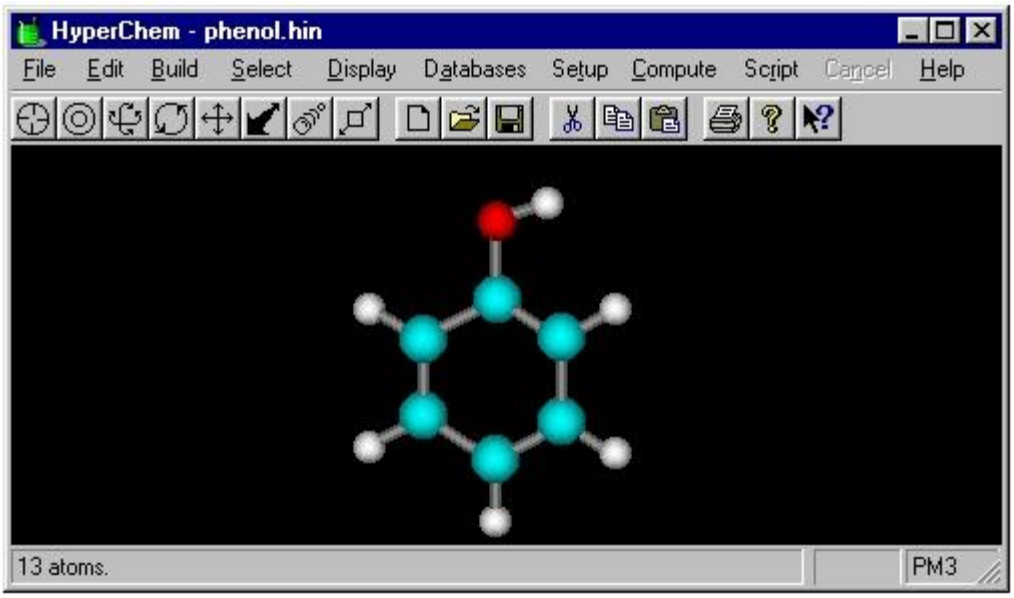
1.4.1 HyperChem

HyperChem is an excellent molecular modeling environment, it can be described as user-friendly , high quality , flexibility and accessibility. (Figure 1.3)

HyperChem includes these functions:

1. Drawing molecules from atoms and converting them to three dimensional (3D) models. Constructing proteins and nucleic acids from standard residues.
2. Using molecules from other sources; for example, Brookhaven Protein Data Bank (PDB) files
3. Rearranging molecules by, for example, rotating and translating them
4. Changing display conditions, including stereo viewing, rendering models, and structural labels.
5. Setting up and directing chemical calculations, including molecular dynamics, by various molecular mechanical or ab initio or semi-empirical quantum mechanics methods

6. Determination of isotope effects in vibration analysis calculations for semi-empirical and ab initio SCF methods
7. Graphing the results of chemical calculations
8. Solvating molecules in a periodic box [45].



Figure(1.3): HyperChem display screen

1.4.2 Dragon

Dragon software is an application for the calculation of molecular descriptors developed by the Milano Chemometrics and QSAR Research Group of Prof. R. Todeschini. These descriptors can be used to evaluate molecular structure-activity or structure-property relationships, as well as for similarity analysis and high throughput screening of molecule databases.

It provides more than 1,600 molecular descriptors that are divided into 18 groups [45].

Table(1.1) Molecular descriptors in dragon software

ID	Block descriptors
1	Constitutional descriptors
2	Topological descriptors
3	Molecular walk counts
4	BCUT descriptors
5	Gavez topological charge indices
6	2D autocorrelation
7	Charge descriptors
8	Aromaticity indices
9	Randic molecular profiles
10	Geometrical descriptors
11	RDF descriptors
12	3D-MORSE descriptors
13	WHIM descriptors
14	GETAWAY descriptors
15	Functional group descriptors

16	Atom –centered fragments
17	Empirical descriptors
18	Properties

1.4.3 SPSS

SPSS is a software for managing data and calculating a wide variety of statistics .SPSS packages provide more capabilities and advanced functions for in-depth statistical analyses.

SPSS stands for Statistical Package for the Social Sciences. First version of SPSS was released in 1968, after being developed by Norman H. Nie, Dale H. Bentand C, Hadlai Hull. SPSS Incorporated is a leading worldwide provider of predictive analytics software and solutions [46].

SPSS is a Window based on full-featured data analysis program that offers a variety of applications such as statistical analysis, graphics, reporting, and data base management. It is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instruction.

SPSS is comprehensive, easy-to-use and predictive analytics tools for agricultural users, business users, analysts and statistical programmers. It consists of a set of software tools for data entry, data management, statistical analysis and presentation [47].

In this research the SPSS will be used to do MLR analysis .

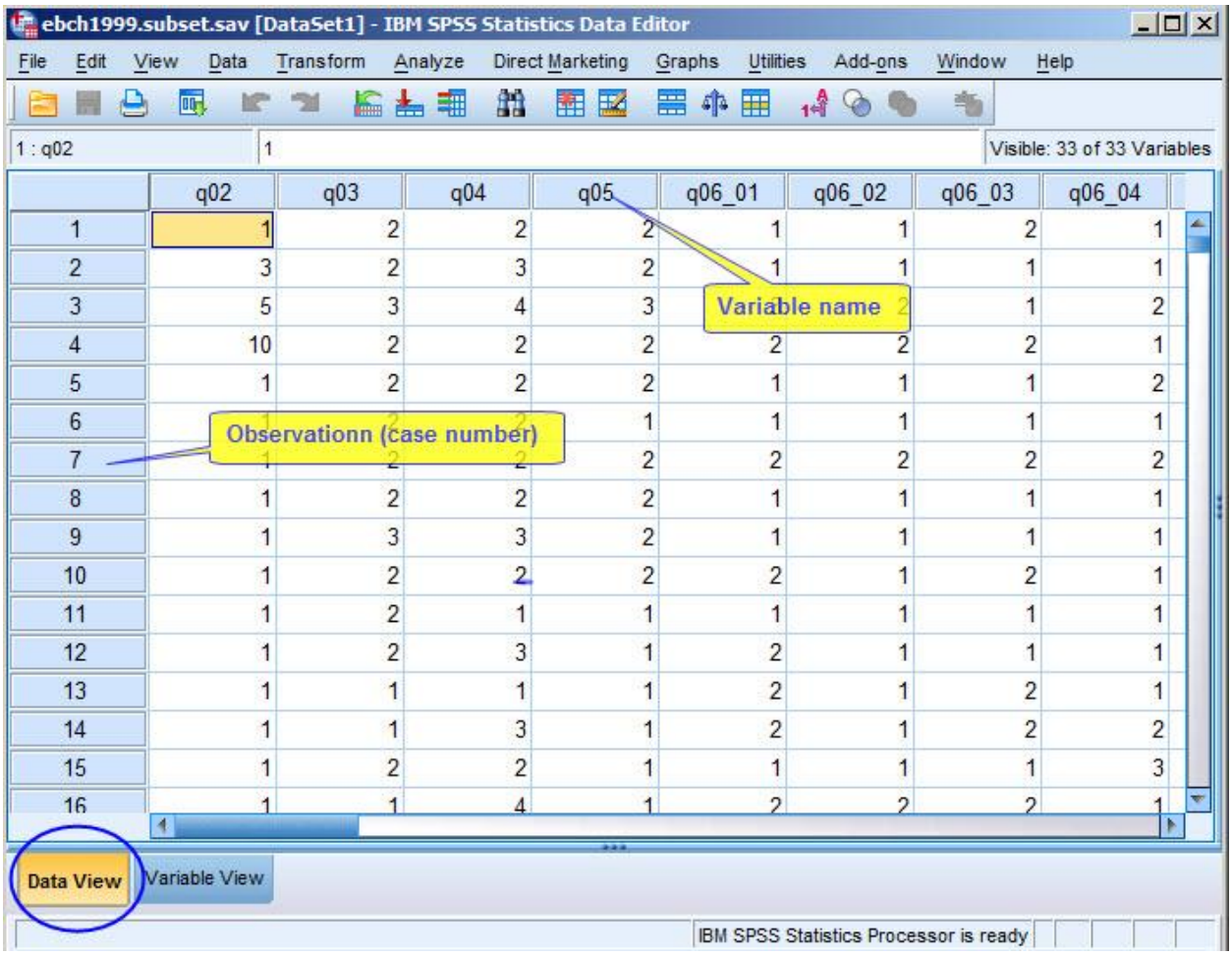


Figure (1.4): SPSS display screen

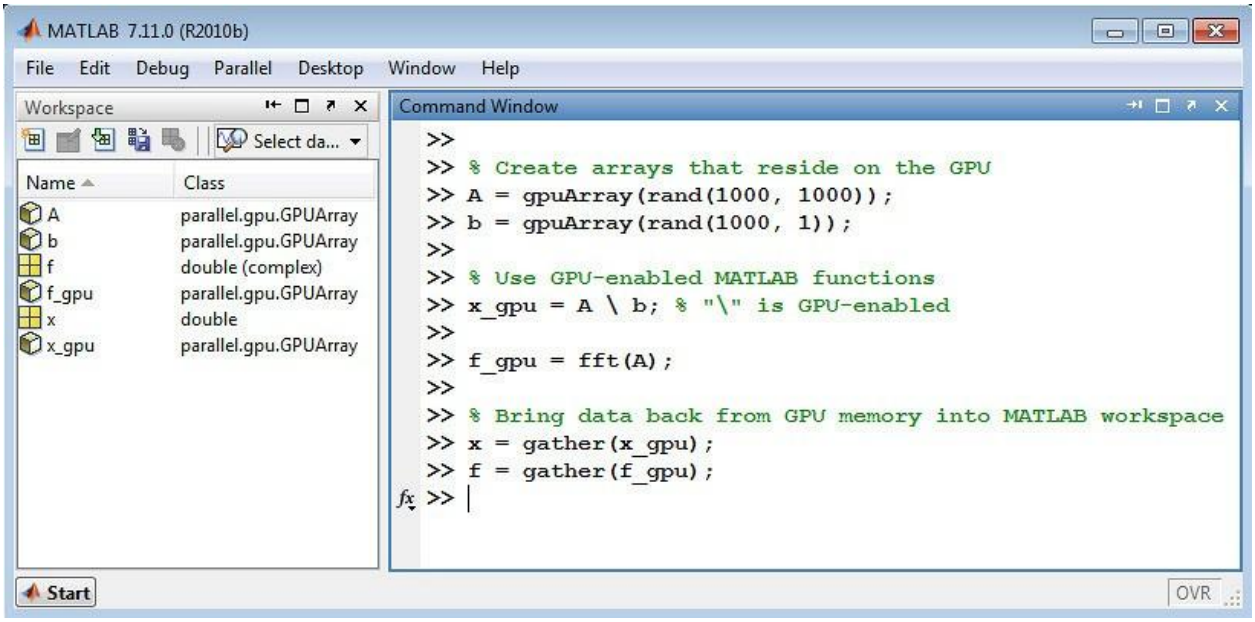
1.4.4 MATLAB

MATLAB is a vector / matrix language, which means that MATLAB thinks in terms of vectors and matrices [48], this make the software particularly useful for linear algebra but MATLAB is also a great tool for solving algebraic and differential equation and for numerical integration. MATLAB has a powerful graphic tools that can produce nice picture in both 2D and 3D [49].

MATLAB includes a variety of different windows for displaying different types of information and performing specific tasks. Each window can generally be opened/closed,

docked in the main window or popped out, and repositioned/resized depending on current needs/preferences.

In our research We have used MATLAB to perform the cross validation ,PCA and ANN .



Figure(1.5): MATLAB display screen

1.5 Epidermal growth factor receptor (EGFR)

Epidermal growth factor receptor (EGFR) belongs to a family of four different receptors, including EGFR (ErbB-1; human epidermal growth factor receptor 1 [HER1]), HER2 (c-ErbB-2), HER3 (c-ErbB-3), and HER4 (c-ErbB-4) [50,51,52]. These proteins are coded by distinct genes that are expressed on chromosomes 7, 17, 12, and 2, respectively [53,54].

There are approximately 500 protein kinases in the human genome, the EGF receptor (EGFR) is one of about 60 transmembrane proteins that have a tyrosine kinase domain within their intercellular region [55].

Receptor tyrosine kinases (RTKs) are essential components of signal transduction pathways that mediate cell-to-cell communication. These single-pass transmembrane receptors, which bind polypeptide ligand - mainly growth factors - play key roles in processes such as cellular growth, differentiation, metabolism and motility [56].

EGFRfamily plays a critical role in vital cellular processes and in various cancers [57], EGFR is over expressed in approximately 50–60% of glioblastoma (GBM) tumors [58].

EGFR and HER2 are over expressed in many solid tumors ,including lung ,head and neck ,breast ,kidney ,colon ,ovary ,prostate, brain and bladder cancers [53]. Activating tumor mutations have been identified in the intracellular and extracellular regions of EGFR [54].

In tumor cells, the TK activity of EGFR may be detected by various oncogenic mechanisms, including EGFR gene mutation, increased gene copy number and EGFR protein overexpression [59]. Improper activation of EGFR TK results in increased malignant cell survival, proliferation, invasion and metastasis. EGFR overexpression is observed in tumors from more than 60% of patients with metastatic non-small-cell lung cancer (NSCLC) and is correlated with poor prognosis [60]. These findings have provided a rationale for the development of novel anticancer agents that target EGFR [61].

1.5.1 EGFR structure

All family members are type I transmembrane glycoprotein that has an extracellular domain which contains two cysteine-rich domains separated by a spacer region that is involved in ligand-binding, and a cytoplasmic domain which has a membrane proximal tyrosine kinase domain and a C-terminal tail with multiple tyrosine autophosphorylation

sites. The human EGFR gene encodes a 1210 amino acid (aa) residue precursor with a 24aa putative signal peptide, a 621aa extracellular domain, a 23aa transmembrane domain, and a 542aa cytoplasmic domain [62,63].

When the EGFR extracellular domain binds to its ligand, such as epidermal growth factor (EGF) and transforming growth factor-α (TGF-α), it forms dimmers with other EGFR or other HER family members and undergoes autophosphorylation at the key tyrosine residues, thus activating several downstream signaling pathways such as protein kinase B (AKT/PKB) and mitogen-activated protein kinases (MAPK), which regulate multiple cellular processes, including proliferation, survival and apoptosis .The constitutive activation of EGFR signaling, caused by gene mutations or by gene amplification or both [64,65].

1.5.2 EGFR inhibitors

The first synthetic tyrosine kinase inhibitors (tyrphostins) was described in 1988 [66]. EGFR inhibitors get their name from a gene called EGFR, Many lung cancer tumors have mutations in this gene. These mutations convert EGFR from a normal gene into a cancer gene that initiates and promotes cancer growth. Approximately 10% to 15% of white and 30% to 35% of Asian patients with NSCLC have EGFR mutations [67].

Drugs targeting mutant EGFR have been developed with the hope that blocking its activation will stop, or at least slow down, the growth of tumors [64,56,57,68].

1.5.3 How EGFR inhibitors work

EGFR inhibitors are targeted specifically against EGFR on the outside of the cell. Protein kinase inhibitors which they orally administered, low molecular weight. Compounds target protein kinases on the inside of the cell [66]. The specific receptors found on the cancer determine which drug is likely to be of benefit.

EGFR are located in the cellmembrane. On binding with an epidermal growth factor ligand outside the cell, they activate a protein kinase inside the cell [68,69]. They prevent the phosphoryation of the specific amino acid, e.g. tyrosineEGFR, and kinase inhibitors may be particularly useful in cancer therapy as they are less toxic than more traditional chemotherapies [68].

These inhibitors all falls in three broad category:

(1) inhibitor of EGFR tyrosine kinase, e.g.Iressa and Tarceva.

(2) inhibitor of split kinase receptor tyrosine kinase e.g. PTK787.

(3) inhibitor of tyrosine kinase from multiple 8, 9, 57 subgroup e.g. Glivec.

1.6 Research objective

The main objective of this study is to develop QSAR models for the inhibition activity of 113 chemical compounds of epidermal growth factor receptor by applying different statistical qualities such as MLR, PC-ANN. These methods will be used to design new inhibitors.

Chapter two

Methodology

2. QSAR working steps

The relation between the biological activities and particular molecular structure can be set by performing QSAR (quantitative –structure activity relationship). The molecular structures and their activities against certain target should be known by experiment.

As mentioned in chapter 1 QSAR has three developmental steps:

- Data preparation
- Data analysis
- Model validation

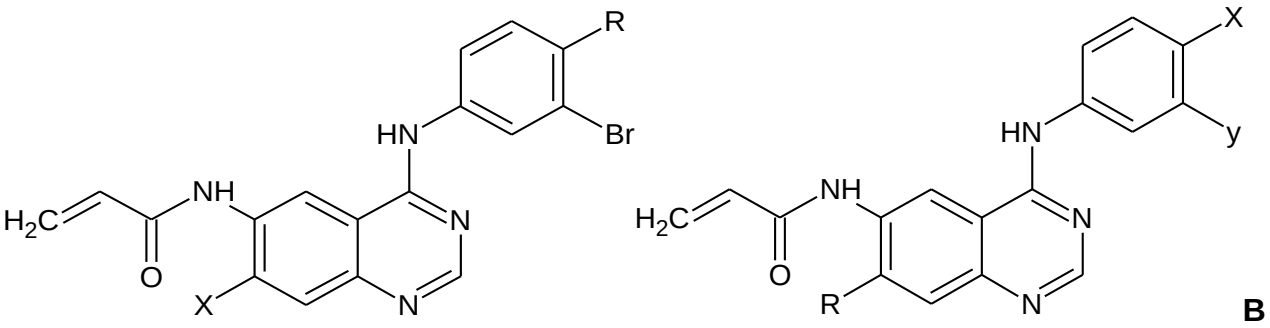
In this chapter, the methodology for each step will be discussed.

2.1 Data preparation

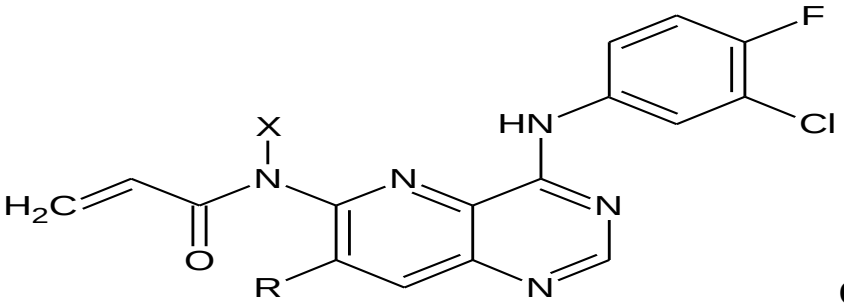
2.1.1 Dataset

A dataset containing 113 compounds along with their EGFR activities that have been taken from the literature [69-72], these shared the same experimental conditions.

The chemical structures and the biological activities of the molecules are summarized in table (2.1).



A **B**



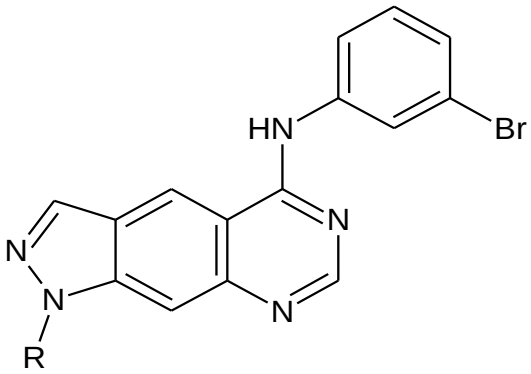
C

Table 2.1 : Dataset compounds and their activity.

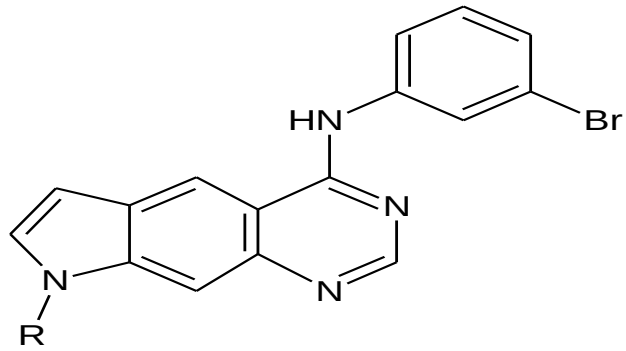
Compound Number	Index	R	X	Y	IC_{50}	PIC_{50}
001	A	H	H		0.70	9.155
002	A	CH_2NMe_2	H		45.00	7.346
003	A	$OCH_2CH_2NMe_2$	H		27.00	7.568
004	B	$O(CH_2)_3$4-Mepip	H	Br	1.70	8.769
005	B	$O(CH_2)_3$morpholide	H	Br	3.60	8.444
006	B	$O(CH_2)_4NMe_2$	H	Br	3.90	8.409
007	B	$O(CH_2)_3$imidazoy1	H	Br	3.00	8.523
008	B	$S(CH_2)_3NEt_2$	H	Br	0.78	9.108
009	B	H	H	CH	0.42	9.376
010	B	$O(CH_2)_3$4-Mepip	H	CH	2.00	8.698
011	B	$O(CH_2)_3$morpholide	H	CH	1.50	8.823
012	B	H	H	Br	0.69	9.161

013	B	$O(CH_2)_3$morpholide	H	Br	1.80	8.745
014	B	H	H	C1	0.75	9.125
015	B	$O(CH_2)_3$morph	F	C1	1.50	8.823
016	B	$[O(CH_2)_2]_2(CH_2)_2OH$	F	C1	1.70	8.769
017	C	H	F		0.75	9.125
018	C	$CH=CH(CH_2)_2$morpholide	F		0.16	9.795
019	C	$(CH_2)_4$morpholide	F		2.70	8.568
020	C	OMe	H		0.95	9.022
021	C	$O(CH_2)_2OMe$	H		0.97	9.013
022	C	$O(CH_2)_3$morpholide	H		1.50	8.823
023	C	$O(CH_2)_3$morpholide	Me		20.00	7.698
024	C	$O(CH_2)_3$4-Mepip	H		6.60	8.180

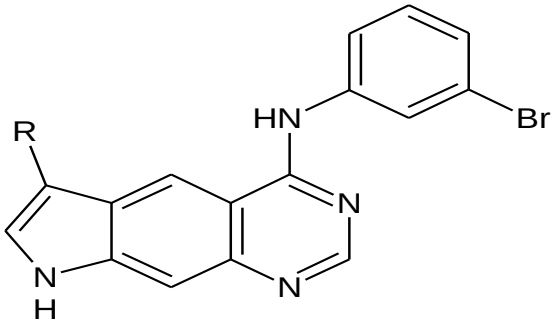
Reference [69]



2a-2b, 2f, 2h



3a-3h

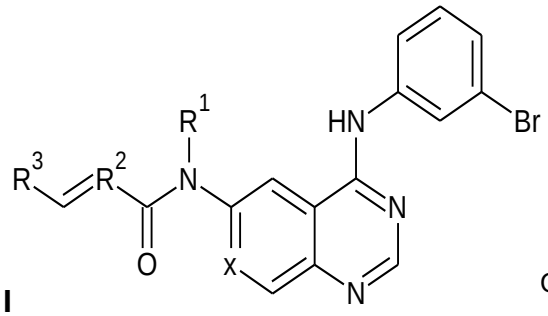


9-4

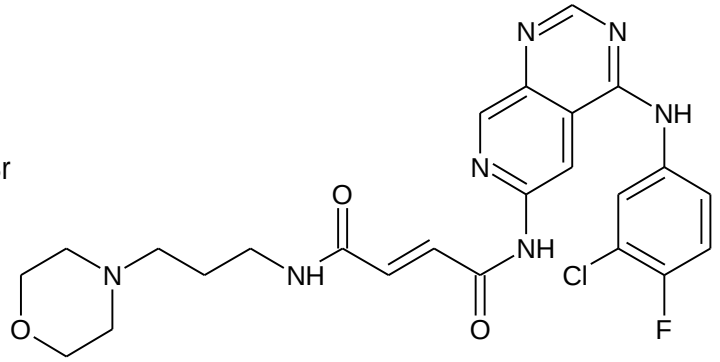
Compound number	Index	R	IC_{50}	PIC_{50}

025	2a	H	0.44	9.356
026	2b	Me	0.37	9.432
027	2c	CH $_2$CH(OH)CH2OH	12.00	7.920
028	2d	$(CH_2)_2$NMe 2b	40.00	7.398
029	2f	$(ch_2)_2$Nmorpholidec	3.70	8.432
030	2h	CH_2COOH	53.00	7.275
031	3a	H	0.44	9.356
032	3b	Me	0.80	9.097
033	3c	$CH_2CH(OH)CH_2OH$	1.60	8.795
034	3d	$(CH_2)_2NMe_2$	41.00	7.387
035	3e	$(CH_2)_3NMe_2$	21.00	7.677
036	3f	$(ch_2)_2$Nmorpholide	3.70	8.432
037	3g	$(ch_2)_3$Nmorpholide	8.80	8.055
038	3h	CH_2COOH	5.10	8.292
039	4	$CH_2N(CH2CH_2OH)_2$	3.50	8.455
040	5	CH_2 NMe_2	2.60	8.585
041	6	Ch_2Nmorpholide	4.80	8.318
042	7	$CH_2N(Me)(CH_2)$ $_2$Nme	7.50	8.124
043	8	$CH_2N(Me)CH_2COOMe$	3.40	8.468
044	9	$CH_2N(Me)CH$ $_2COOH$	0.72	9.143

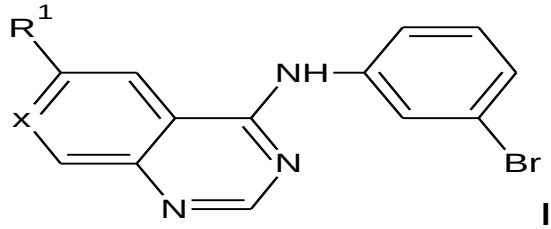
Reference [70]



I



II

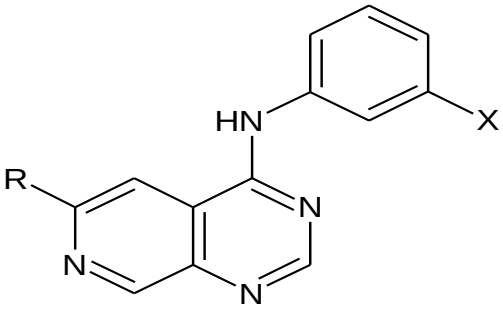


III

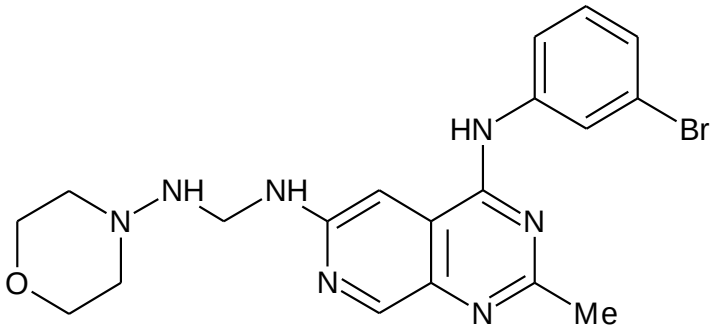
Compound Number	Index	X	R_1	R_2	R_3	IC_{50}	PIC_{50}
045	I	N	H	H	H	0.91	9.040
046	I	C	H	H	H	0.70	9.154
047	II					1.50	8.823
048	I	N	Me	H	H	0.17	9.769
049	I	N	H	Me	H	1.60	8.795
050	I	C	H	Me	H	1.20	8.920
051	I	N	H	H	Me	0.50	9.301
052	I	C	H	H	Me	0.55	9.259
053	I	N	H	H	Cis-Cl	0.69	9.161
054	I	C	H	H	CF_3	1.75	8.756
055	I	N	H	H	$CH=CH_2$	1.10	8.958
056	I	C	H	H	$=CH_2$	1.60	8.795
057	I	N	H	H	Ph	9.10	8.040
058	I	C	H	H	COMe	1.20	8.920
059	I	C	H	H	COOH	0.37	9.431
060	I	C	H	H	COOEt	2.70	8.568
061	I	N	H	H	COOEt	1.50	8.823

062	I	N	$(CH_2)_2NMe_2$	H	H	4.20	8.376
063	I	N	$(CH_2)_3$-N-morpholiny1	H	H	2.70	8.568
064	I	C	$(CH_2)_3$-N-morpholiny1	H	H	3.30	8.481
065	I	C	H	H	$COO(CH_2)_3NMe_2$	2.40	8.619
066	I	C	H	H	$CONH(CH_2)_3NMe_2$	0.44	9.356
067	I	N	H	H	$CONH(CH_2)_3NMe_2$	1.10	8.958
068	I	N	H	H	$CONH(CH_2)_3NEt_2$	0.73	9.136
069	I	N	H	H	$CONH(CH_2)_3$-N-morpholiny1	0.81	9.091
070	I	N	H	H	$CONH(CH_2)_3$-N-imidazoly1	0.56	9.251
071	I	N	Me	H	$CONH(CH_2)_3NMe_2$	1.45	8.838
072	II					0.61	9.214
073	III	N	$NHSO_2CH=CH_2$			0.76	9.119
074	III	C	$NHSO_2CH=CH_2$			1.40	8.853
075	III	N	$SO_2CH_2CH_2OH$			93.5	7.029
076	III	N	$SO_2CH=CH_2$			0.43	9.366
077	III	N	$SOCH=CH_2$			4.60	8.337

Reference [71]



Compound 5-9



10n

Compound number	Index	R	X	IC_{50}	PIC_{50}
078	5a	NH_2	Br	0.130	9.886
079	5b	NHMe	Br	0.008	11.096

080	5c	NMe_2	Br	0.006	11.221
081	5d	$NHCH_2CH_2OH$	Br	0.190	9.721
082	5e	$N(Me)CH_2CH_2OH$	Br	0.220	9.657
083	5f	$NHCH_2CH(OH)CH_2OH$	Br	0.180	9.744
084	5g	$N(Me)CH_2CH(OH)CH_2OH$	Br	0.560	9.252
085	5h	$NH(CH_2)_2NMe_2$	Br	1.100	8.958
086	5i	$NH(CH_2)_3NMe_2$	Br	1.200	8.920
087	5j	$NH(CH_2)_4$ NMe_2	Br	1.800	8.744
088	5k	$NHCH_2CH(OH)CH_2NEt_2$	Br	4.600	8.337
089	5l	$NH(CH_2)_2N(Me)CH_2CH_2OH$	Br	1.700	8.769
090	5m	$N(Me)(CH_2)_2NMe_2$	Br	8.100	8.091
091	5n	$NH(CH_2)_3$morpholinyl	Br	0.650	9.187
092	5o	$NH(CH_2)_2$morpholinyl	Br	1.000	9.000
093	5p	$NH(CH_2)_3$Nmepiperazinyl	Br	3.900	8.408
094	5q	$NH(CH_2)_2N(CH_2CH_2OH)_2$	Br	0.930	9.032
095	5r	$NH(CH_2)_3N(CH_2CH_2OH)_2$	Br	0.350	9.455
096	5s	$NHCH_2$(3-pyridyl)	Br	1.500	8.824
097	5t	$NHCH_2CH_2$(2-pyridyl)	Br	1.200	8.920
098	5u	$NH(CH_2)_2$(4-imidazolyl)	Br	0.780	9.107
099	5v	$NH(CH_2)_3$(1-imidazolyl)	Br	1.700	8.769
100	5w	4-Mepiperazinyl	Br	6.400	8.193
101	5x	$NHCH_2COOH$	Br	0.280	9.553
102	5y	$N(Me)CH_2COOH$	Br	0.440	9.356
103	5z	$NH(CH_2)_2COOH$	Br	0.270	9.568
104	6b	NHMe	H	9.000	8.045
105	7b	NHMe	Cl	0.190	9.721
106	8b	NHMe	CF_3	1.100	8.958
107	9a	NH_2	Me	3.100	8.508
108	9b	NHMe	Me	0.450	9.346

109	9c	NMe_2	Me	2.800	8.553
110	9n	$NH(CH_2)_2$morpholinyl	Me	1.500	8.824
111	9o	$NH(CH_2)_3$morpholinyl	Me	1.800	8.744
112	9u	$NH(CH_2)_2$(4-imidazolyl)	Me	1.300	8.886
113	10n			117.000	6.931

Reference [72]

2.1.2 Compound optimization

Drawing and optimizing each compound from our dataset is an essential step for a well-defined structure. The minimum potential energy surface obtained enables a calculation of the properties of each molecule. This has been perceived by the following steps using Hyperchem.

1- Draw the compound structure on HyperChem Workspace using drawing tools.

2- Model builder (add hydrogen and model build) existing in the Build menu is used to convert structure from two dimensional 2D to three-dimensional 3D.

3- Click start log on the File menu to save the new drawn structure, name the file, and choose a directory to save in it.

4- Then in order to perform the optimization of the compound structure, choose the semi-empirical calculation from the setup tab, a dialog box of types of semi-empirical methods will open. Thus choose from it the AM1 method and after that press on the option button to determine the geometry optimization parameters, total charge =0, spin multiplicity =1,Spain is pairing =RHF (Restricted Hartree-Fock), convergence limit =0. 1.

These parameters mean that the calculation ends when the difference in energy after two consecutive iterations is less than 0.1 kcal/ mol).

5- Click ok to close the option screen, and then click ok to close semi empirical method dialog box .

6- To start optimization process, click compute from the menu,choose geometry optimization the semi empirical optimization screen appears, Polak –Ribiere as algorithm

method is chosen, RMS =0.01 , then maximum cycle should be chosen according to needs.Then click ok so the optimization process initiate.

7-When this process stopped, select stop log from file menu to save the calculation output as log file, the output file will be saved in (.hin) format .

2.1.3 Calculating descriptors

The main concept of QSAR is to find the relationship between the chemical structure and the properties (descriptors), the first step was drawing the chemical structure of the compounds and optimizing them, the next step to calculate the descriptors.

We can define molecular descriptor as the final result of a logic and mathematical procedure. This transforms chemical information encoded within a symbolic representation of a molecule into a useful result of some standardized experiment.

In current research, many types of descriptors have been calculated using HyperChem and Dragon software's.

2.1.3.1 Descriptors by HyperChem

Two types of descriptors will be calculated using HyperChem

1- Descriptors taken from the output log file.

2- Descriptors calculated from the optimized structure.

1- Descriptors taken from the output log file.

Output log file is opened for each compound after doing the optimization and obtaining the following values:

- Heat of formation (kcal/mol).
- Dipole moment (Debyes).
- HOMO (highest occupied molecular orbital).
- LUMO (lowest unoccupied molecular orbital).

Finally, values are collected and saved in excel file.

- Descriptorscalculated from the HOMO, LUMO values:
- Hardness (0.5*(LUMO-HOMO))
- Softness (1/Hardness).
- Electronegativity (-0.5* (LUMO+HOMO))
- Electrophlicity (Electronegativity *Electronegativity /(2* Hardness)) [73]

2- Descriptors calculated from the optimized structures using HyperChem.

We can calculate the descriptors we need from the HyperChem by performing the following steps:

1- After optimizing the 3D structure of each compound we have to open the HyperChem file for each compound.

2- The QSAR properties button should be clicked from the computer tab, which in turn opens a dialog box containing the following properties:

- Surface area (Approx)
- Surface area (Grid)
- Volume
- Hydration Energy
- Log p
- Refractivity
- Polarizability

- Mass

3- calculate the value of each one of these properties for all compounds. As the results appear copy into excel file. This step is repeated for all properties.

2.1.3.2. Descriptors calculated by Dragon

DRAGON software has been chosen to provide the user with a variety of molecular descriptors, derived from their representations. Dragon software calculates thousands of descriptors in a technical way. These descriptors are divided into 18 blocks as mentioned in chapter 1.

Types of molecular descriptors:

- Simple molecular descriptors are derived by counting some atom-types or structural fragments in the molecule.
- Topological or 2D descriptors are derived from algorithms thus applied to a topological representation (molecular graph).
- Geometrical or 3D descriptors are derived from a geometrical representation.

All these types should contain chemical information, requirements and well established procedures enabling them to be calculated [74].

Descriptors are divided into four categories according to their dimensions:

3D: Geometrical, Randic molecular profile, WHIM, GETAWAY, RDF, 3D- MoRSE, Charge descriptor.

2D: Autocorrelation, Topological, Molecular walk counts, Galves topological charge indices, BCUT descriptor.

1D: Empirical, Functional groups, Properties, Atom- centered fragment descriptors.

0D: Constitutional descriptor.

2.1.3.2.1 Steps to calculate descriptors using Dragon software:

1- First we must open dragon software, a screen will appear, choose calculate descriptors.

2- Select the output file resulted from the HyperChem structures optimization process, and choose the type of the file to be in (.hin) format then choose the type of descriptor group to be calculated, then press run.

3- Soon as the calculation is done, save the output file in notepad format.

4- Open the notepad file in excel file after changing its format, so the file could be opened in SPSS and other software.

5- Make sure that the excel file, has all the descriptors for all blocks and all compounds are calculated.

Table 2.2 Brief description of the descriptors that will be used in this study.

Descriptors Type	Molecular Descriptors

Constitutional	Molecular weight (MW), number of atoms (nAT), number of non H-atoms (nSK), number of bonds (nBT), number of multiple bonds (nBM), number of rings (nCIC), number of circuits (nCIR), number of H-bond donor (nHDon), number of H-bond acceptor (nHAcc).
Topological Indices	Information index molecular size (ISIZ), connectivity indices(X), average connectivity index (XA), kier symmetry index (S0K), total walk count (TWC), Zagreb index (Z), Schultz molecular topological index, Balaban j index (J), Wiener w index (W)
Quantum Chemical	Highest occupied molecular orbital energy(E_{HOMO}), Lowest unoccupied molecular orbital energy (E_{LUMO}), Most positive charges(MPC), Least negative charges (LNC), Most negative charges(MNC), Sum of positive charges(SPC), Sum of negative charges (SNC), Sum of squares of positive charges (SSPC), Sum of squares of negative charges(SSNC),Sum of squares of charges (SSC), Sum of absolute of charges (SAC) ,molecular Dipole moment (DM) , Electronegativity ($\chi=-0.5(E_{HOMO}-E_{LUMO})$) .Hardness($\eta=0.5(E_{HOMO}+E_{LUMO})$). Softness ($S$=1\η).Electrophilicity ($\omega=\chi^2/2\eta$). Heat of formation ($H_f$).
Chemical descriptors	Octanol-water partition coefficient (LogP), hydration energy (HE) polarizability (Pol), refractivity (Ref), volume (V), surface area (SA),

2.2 Data analysis

2.2.1 Multiple Linear Regression (MLR)

MLR is the first statistical step done because the hypothesis is that there is a linear correlation between the independent variables (descriptors) and the dependent variable (biological activity), MLR statistic is done using SPSS software.

2.2.1.1 Steps to perform MLR using SPSS:

1- Import to SPSS the chosen output file (the file that contains the biological activity and the descriptors from a specific group which resulted from dragon or HyperChem software).

2- After the file is opened in SPSS, press Analysis from the menu, “Regression”, then “Linear”. See figure 2.1.

3- From the list in the variable box choose variables for analysis and move to the independent or dependent box by choosing the variables and clicking the arrow. See figure 2-2.

4- Press option button; a dialog box will open to set the F value, which must be changed to get the best result.

5- Press the OK button to start the MLR process.

6- This process undergoes repetition until all groups of descriptors are finished.

7- From the output file, choose the highest R value models and minimum number of descriptors.

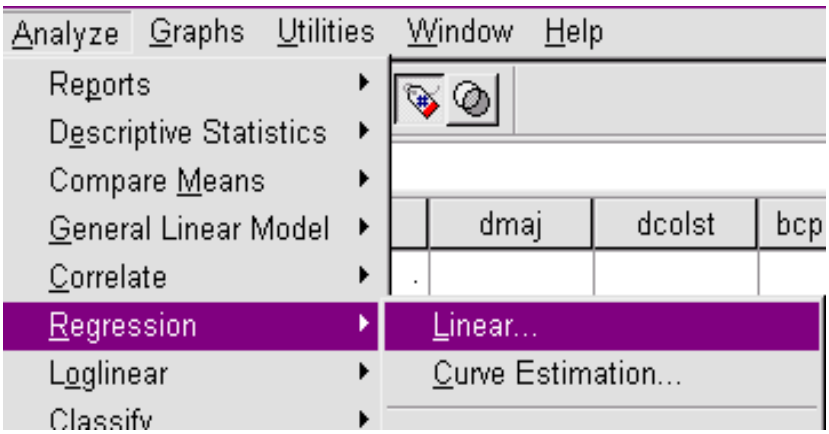
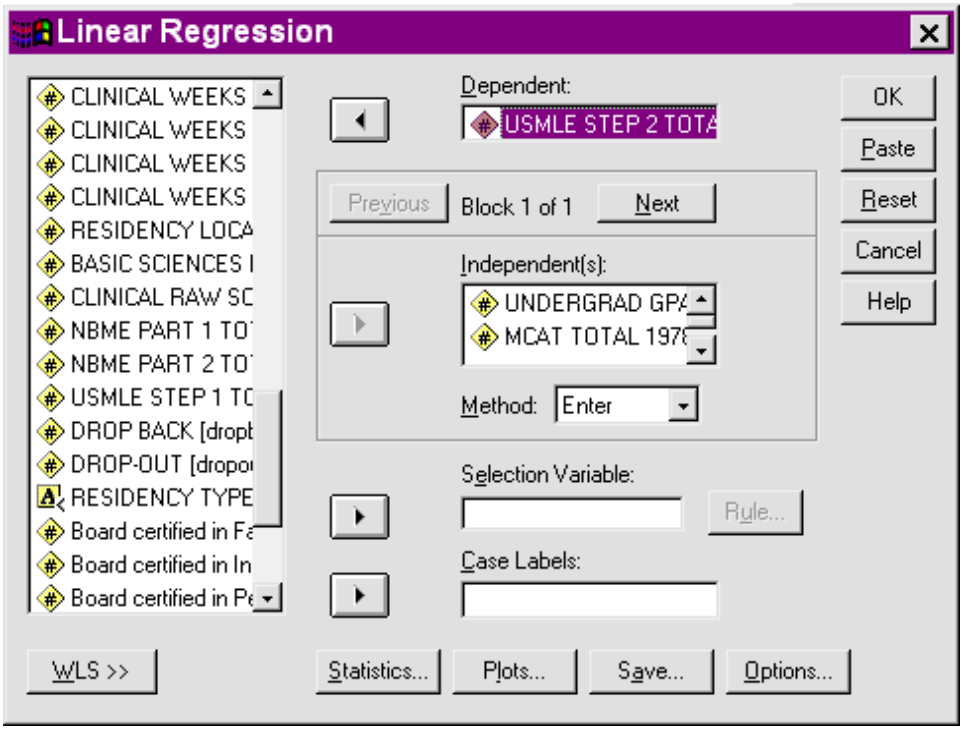


Figure (2.1): Choosing linear regression analysis .



Figure(2.2): Dialog box of dependent and independent variables

2.2.1.2 Performing MLR of all descriptors resulted from the first MLR using SPSS:

1- Follow the first step by performing MLR for the best models of descriptors groups. All should be gathered in one excel file starting with their activity.

2- Import this file to SPSS and perform MLR process as mentioned before.

Choose all models which have $R^2 \geq 0.6$. from the output [73].

2.2.2 Model Validation

Model validation is the most important and overlooked step in the model building sequence of this research; so external and internal validation methods have been performed , two internal validation have been done to validate the MLR and the ANN resulted method, cross-validation and scrambling (Y-Randomization) respectively .

2.2.2.1 Cross validation is divided into two types, leave one out (LOO), and leave many out (LMO) cross validation.

2.2.2.1.A Leave one out (LOO) steps using MATLAB

1- Prepare an excel file which have the observed activity and the predicted activity of all chosen models taken from the final MLR.

2- To perform LOO we have to run a specific MATLAB script, the MATLAB will ask about the file name, model number, number of descriptors, enter the required information

A good output should look like:

```
Model1  PRESS    SPRESS   SST     R2CV    PRESS/SST   PSE     RSEP

-------------------------------------------------------------

 1    23.256  0.7155   15.235   -0.4807   1.2635    0.6338   61.1801
```

3- Choose the models which have PRESS/SST values <0.4.

2.2.2.1.B Leave many out (LMO) steps using MATLAB

1- Prepare an excel file for each model which have the observed activity and descriptors of the model.

2- Run the MATLAB script.

A good output should look like:

```
PRESS       SPRESS      SST         R2CV      PRESS/SST     PSE         RSEP

-------------------------------------------------------------------------------

44.7577      1.0198      48.6208      -0.0736      1.0736      0.3292      44.1042
```

Choose models with PRESS /SST <0.4 [74].

A comparison of LOO results should be performed to choose the models for PCA and ANN.

2.2.3 Principle component analysis (PCA)

Principle component analysis is an important step used to divide the data into three groups (training set, test set, validation set); this division should not be done randomly, instead, their factor spaces and activity data should be used. Descriptors and activities should be gathered in a single matrix (X) then principal component analysis (PCA) done on X and plot the first score against the second. Then select the training set molecules from the scattered distribution data to cover the space of the entire data. The validation, test sets would get 20%, while the data in the training set would get 60%.

Steps of principle component analysis:

Open MATLAB, run the specific script to plot the PCs.

1- The MATLAB will ask you the file name, enter the file name which has all activities and all descriptors for the chosen model from the MLR validation.

2- Figure will be produced where x label ('Xth Principal Component'), y label ('Yth Principal Component').

3-Divide the data into training set, test set, validation set according to the figure produced.

2.2.4 Artificial Neural Networks (ANN)

In our research, we will perform ANN after doing the MLR.

2.2.4.A Steps of ANN for each MLR model using MATLAB:

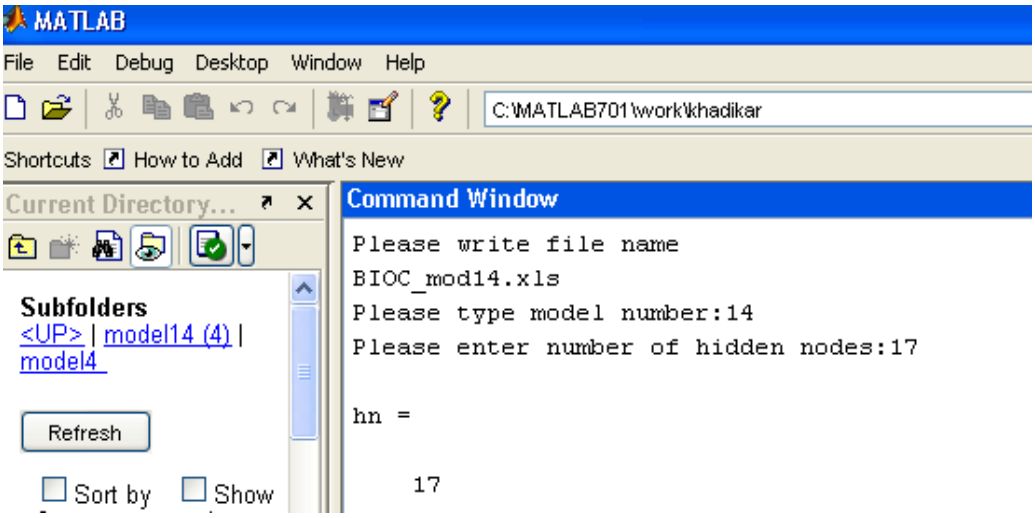
1- Prepare an excel file for each model (the same models used in PCA) the file should have the activities and the descriptors.

2- Open MATLAB and run special script for ANN.

3- The MATLAB will ask for the file name, model number, number of hidden nodes see figure 2.4.

4- Enter the file name, model number and number of hidden nodes.

5- Choose the best models according to the high R for test set and low PRESS and RESP values.



Figure(2.3): MATLAB command window.

2.2.4.B Steps to do ANN for the best models of a particular range of hidden nodes (Hn) using MATLAB

1- Perform ANN for the best models based on the previous ANN with range of hidden nodes 5-20 by repeating the same procedure as in section 2.2.4.A.

2- Choose the best models which have high R, low PRESS, low RESP values, with a small number of hidden nodes.

2.2.5 Randomization test (chance correlation)

In our research the randomization test is done to ensure accuracy.

Steps to do randomization test using MATLAB

1- Open MATLAB, run the special script, the MATLAB will ask about the file name and number of trail.

2- Enter the file name, the files should be prepared for each model resulted from ANN, the file content similar to LMO data files, also enter the trail number.

3- Repeat the test for each model more than 10 times.

Chapter Three

Results and Discussion

3. Results and Discussion :

In this study, we developed QSAR models for the 113 chemical compounds of the inhibition activity of epidermal growth factor receptor (EGFR).

As we have already discussed in chapter 2 there are three steps for the development of QSAR models, data preparation, data analysis, and model validation .

3.1 Data preparation results

113 optimized compounds were obtained using Hyperchem, through semi-empirical AMI (Austin Model 1) method, where AMI method is found to be more reliable, faster than other semi-empirical methods [75].

Descriptors were calculated using HyperChem and Dragon softwares.

All the descriptors calculated using HyperChem is mentioned below :

- Heat of formation (kcal/mol)
- Dipole moment (Debyes)
- HOMO (highest occupied molecular orbital)
- LUMO(lowest unoccupied molecular orbital)
- Hardness (0.5*(LUMO-HOMO))
- Softness (1/Hardness).
- Electronegativity (-0.5*(LUMO+HOMO))
- Electrophlicity (Electronegativity *Electronegativity /(2* Hardness))
- Surface area (Approx)
- Surface area (Grid)

- Volume
- Hydration Energy
- Log p.
- Refractivity .
- Polarizability .
- Mass .

There are 1266 descriptors calculated using dragon, which is divided into 18 groups .

Each group has certain numbers of descriptors ,groups such as constitutional ,topological ,molecular walk ,BCUT ,functional …..etc.

For example: A 35 descriptors were calculated within the constitutional group , 229 descriptors within the topological group and so on.

Performing the first MLR for each group of descriptors, except the groups which have small numbers of descriptors; these groups such as Charge descriptors, Randic molecular profiles, Geometrical descriptors all gather in one group .

Results of first MLR are in table 3.1,where (R) refers to correlation coefficient, (R^2) refers to coefficient of determination, (R^2adj) refers to adjusted R^2, and select (the selected) descriptors refers to the descriptors chosen by MLR (the MLR) model.

Table 3.1: MLR Models resulted from each group of descriptors.

Group name	#of calculated descriptors	R	R^2	R^2adj	Standard error of estimation	Selected descriptors
Constitutional descriptors	35	0.560	0.314	0.164	0.616	nCIR, nH, nS, RBF, RBN, nC, Mv, nBR, AMW, Ms, nCL, nBM, nAB, nBnz, nX, nR06, Mp, Ss, Se, nR09.
GETAWAY descriptors.	197	0.686	0.471	0.430	0.509	R4e, R6p+, R7e+, R7u, H3e, R3v+, ISH, R6u
Charge indices ,2D autocorrelation.	117	0.803	0.645	0.577	0.438	GGI5, ATS7e, ATS6e, ATS5e, MATS4e, GATS6m, MATS8m, GATS3e, JGI4, GATS5e, MATS1e, MATS3e, MATS6e, JGI9, ATS6m, ATS1p, GATS1e, ATS3p
		0.806				MPC10, BIC3, ww, CSI, Eig1Z, X5sol, X0Av, Xt, X3A, TI2, T(N..O), IC2,

Topological descriptors.	229		0.650	0.574	0.440	X5, T(N..S), BIC4, T(S..Br), MPC09, IDE, TPC, X4Av
Molecular walk ,BCUT descriptors	83	0.799	0.638	0.539	0.458	MWC07, BEHp2, MWC08, BELm4, BEHp6, BELv5, BEHv1, BEHe1, BEHm3, BEHm7, BEHm6, BELm7, BEHe6, BEHv6, BEHe8, BEHp8, BELe4, BELe8, BELp2, BELv8, BELv3, BEHv3, BELp5, BELm5
RDF	150	0.713	0.508	0.420	. 0.513	RDF055e, RDF115m, RDF075u, RDF105m, RDF090m, RDF100m, RDF100u, RDF065m, RDF115p, RDF115v, RDF085m, RDF015m, RDF070e, RDF155e, RDF130v, RDF125e, RDF135p

Group name	#of calculated descriptors	R	R^2	R^2adj	Standard error of estimation	Selected descriptors
WHIM descriptors	109	0.738	0.545	0.458	0.496	G2v, As, P2e, G2s, G2p, E3s, E3v, Gu, L2m, G3v, G3p, G2u, G1e, E3m, E2e, L3u,

						L3e, Am
3D-MORSE descriptor	160	0.739	0.546	0.464	0.493	Mor19m, Mor25p, Mor16m, Mor18u, Mor28u, Mor26m, Mor30e, Mor29p, Mor28e, Mor04v, Mor22e, Mor22p, Mor24v, Mor24p, Mor04p, Mor22m, Mor22v
Charge descriptors, Aromaticit indices ,Randic molecular profiles, Geometrical descriptors	95	0.727	0.528	0.449	0.500	G1, G(N..O), J3D, H3D, W3D, AGDD, SHP2, RGyr, RNCG, MAXDN, G(O..O), G(O..Br), FDI, SPAN, G(O..Cl), G(O..S)
Functional groups , Atom centered fragments, Empirical descriptors, properties ,Quantum chemical descriptors.	101	0.730	0.533	0.449	0.500	nNR2, C-024, nSO2, H-053, surface area approx , C-040, nOHt, Br-094, dipole moment, nNR2Ph, N-070, H-051, C-007, C-002, nRORPh, C-017, Nconhr

After doing the first MLR, we gather all resulted descriptors from all groups and perform second MLR .

Results of second MLR are presented in table (3.2) where only models with R^2> 0.6 were chosen for farther analysis [73].

So models (12-23) were chosen, so we perform cross validation on them .

Table 3.2: MLR Models resulted from all groups of descriptor together.

Model NO	R	R^2	R^2adj	Descriptors used in MLR model
12	0.776	0.602	0.563	GATS6m, MATS1e, MATS8m, R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u ,G1e
13	0.790	0.624	0.583	GATS6m, MATS1e, MATS8m, R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u, G1e, G(O..S)

14	0.803	0.645	0.602	GATS6m, MATS1e, MATS8m, R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u, G1e, G(O..S), Mor18u.
15	0.812	0.660	0.615	GATS6m, MATS1e, MATS8m, R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u, G1e, G(O..S), Mor18u, Mor19m
16	0.820	0.672	0.625	GATS6m, MATS1e, MATS8m, R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u, G1e, G(O..S), Mor18u, Mor19m, RDF155e
17	0.830	0.689	0.641	GATS6m, MATS1e, MATS8m R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u, G1e, G(O..S), Mor18u, Mor19m, RDF155e, Gu
18	0.841	0.707	0.658	GATS6m, MATS1e, MATS8m, R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u, G1e, G(O..S), Mor18u, Mor19m, RDF155e, Gu, R3v+
19	0.851	0.725	0.675	GATS6m, MATS1e, MATS8m, R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u, G1e, G(O..S),Mor18u, Mor19m, RDF155e, Gu, R3v+, BELe4
			0.691	

20	0.861	0.741		. GATS6m, MATS1e, MATS8m, R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u, G1e, G(O..S), Mor18u, Mor19m, 1RDF155e, Gu, R3v+, BELe4, G2p
21	0.869	0.755	0.705	GATS6m, MATS1e, MATS8m, R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u, G1e, G(O..S), Mor18u, Mor19m, RDF155e, Gu, R3v+, BELe4, G2p, G(N..O).
22	0.874	0.764	0.712	GATS6m, MATS1e, MATS8m, R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u, G1e, G(O..S), Mor18u, Mor19m, RDF155e, Gu, R3v+, BELe4, G2p, G(N..O), N-070
23	0.878	0.772	0.719	GATS6m, MATS1e, MATS8m, R7e+, RDF090m, R7u, nOHt, Mor22e, RDF075u, G1e, G(O..S), Mor18u, Mor19m, RDF155e, Gu, R3v+, BELe4, G2p, G(N..O), N-070, ATS3p

The equation which represent best MLR model number 23 :

PIC50 = 27.423(±5.249) - 44.938(±10.050)GATS6m -3.754(±0.931)MATS1e - 22.191(±4.330)MATS8m+12.546(±2.175)$R7e^{+}$ +0.57(±0.20)RDF 090m -1.799(±0.617)R7u -0.938(±0.300)noHT -1.200(±0.252)Mor22e +0.62(±0.014) RDF075u

+20.618(±12.578)G1e -0.228(±0.061)G(O.S) -0.662(±0.185)Mor18u -0.449(±0.181)Mor19m +0.068(±0.018)RDF155e +8.738(±1.841)Gu +6.845(±1.884)$R3V^+$ -2.427(±0.810)BELe4 -26.982(±9.783)G2P +0.005(±0.002)G(N.O)+0.358(±0.138)N_070 +4.633(±2.596)ATS3P.

Where R = 0.878, R^2= 0.772 ,R^2adj =0.719, and standard error of estimate STD =0.357.

According to equation of the best MLR there are descriptors with positive correlation and others with negative correlation.

Descriptors with positive correlation such as $R7e^+$,RDF090m ,RDF075u ,G1e ,RDF100e ,Gu ,$R3V^+$,G(N.O) ,N-070 ,ATS3P .

Descriptors with negative correlation such as GATS6m ,MATS1e ,MATS8m ,R7u ,noHT ,Mor22e ,G(O.S) , Mor18u ,Mor19m ,BELe4 ,G2P .

LOO and LMO cross validation performed on the MLR best models (12-23) using MATLAB software. The results of LOO is in table (3-3) where : PRESS (Predictive residual sum of squares) also called SSE (Error sum of squares) ,PRESS is used to measure the accuracy of the models .SST (Total sum of square) R^2CV or Q^2 (cross validation correlation coefficient), SPRESS (uncertainty of prediction) ,PSE (Predictive Square Error) also called RMSE (Root Mean Square Error), RSEP (Relative Standard Error of Prediction).

Table (3.3): LOO cross validation results.

Model	No.Descriptor	PRESS	SPRESS	SST	R^2CV	PRESS/SST	PSE	RSEP
12	10	20.27	0.427	30.688	0.33	0.66	0.423	4.78
13	11	19.14	0.417	31.81	0.39	0.60	0.411	4.65
14	12	18.10	0.407	32.85	0.44	0.55	0.400	4.52
15	13	17.34	0.4	33.61	0.48	0.51	0.391	4.43

16	14	16.73	0.395	34.22	0.51	0.48	0.384	4.35
17	15	15.83	0.386	35.12	0.54	0.45	0.374	4.23
18	16	14.93	0.377	36.02	0.58	0.41	0.363	4.11
19	17	14.03	0.367	36.93	0.62	0.37	0.352	3.98
20	18	13.20	0.358	37.75	0.65	0.34	0.341	3.86
21	19	12.50	0.35	38.45	0.67	0.32	0.332	3.76
22	20	12.04	0.345	38.91	0.69	0.30	0.326	3.69
23	21	11.63	0.341	39.32	0.70	0.29	0.320	3.62

The result of LMO is agree with the result of LOO.

The PCA was performed to divide the molecules into training ,test, and validation set , doing PCA on all data 113 compound,23 descriptors and plotting the first and second principles. The data will be divide into 60% training set,20% test set, 20% validation set ,the division should be in equal manner in which choosing one compound from each zone to each set .

So the division according to PCA 60% (67 compounds) training set, 20% (23 compounds) test set ,20% (23 compounds) validation set .

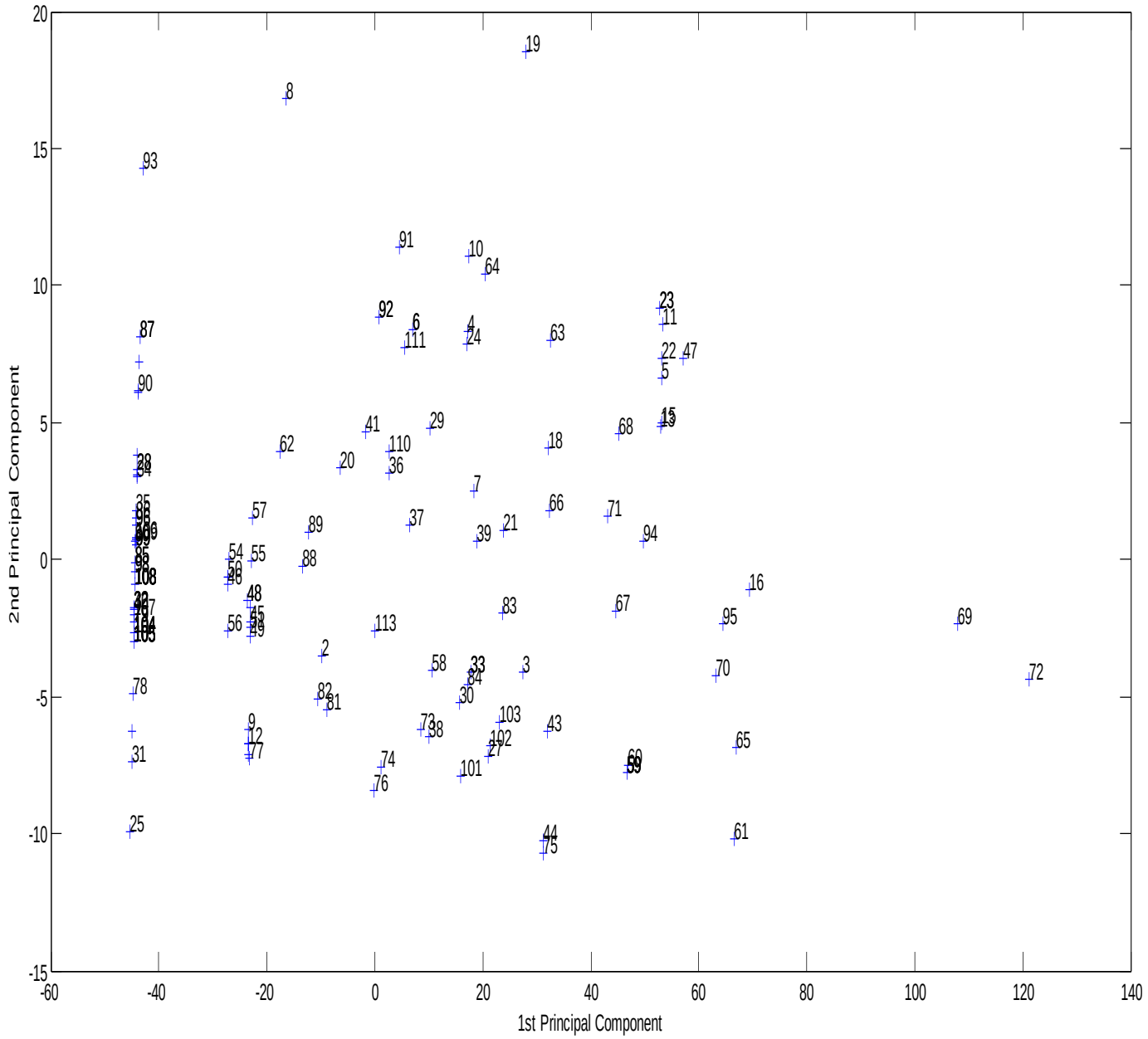


Figure (3.1): First and Second principle component plot .

The models (19-23) were chosen to perform Artificial Neural Networks (ANN)according to LOO and LMO cross validation results. First doing ANN for each model with 7 hidden nodes.

The results for ANN are in table (3.4), the table shows that model 19 has the highest correlation coefficient for the R test (0.805) indicating its high predictive power,and then model 20 with r test (0.763) .

Figure 3.2 shows the relation between the correlation coefficient (R) values for training ,test, and validation sets versus the model number . This figure shows that the higher (R) value for training set is obtained for models 23 then 22 then 19. While the highest (R) values for the test set is for model 19 then 20.

Figure 3.3 shows the relation between the PRESS value for training, test, and validation sets versus model number. This figure shows that the minimum PRESS values for training set is found in model 23 then model 22 then model 19. while the minimum PRESS values for test set is obtained in model 19 then 20 .

Accordingly,models 19,20 were subjected for more analysis by optimizing the number of hidden nodes,because these models have the highest R, R^2cv and low PRESS values for test set .

Table 3.4: Correlation Coefficient and Cross Validation Parameters for ANN models 19-23.

Model no	nPCs	Hn	R_tr	PRESS_tr	R^2CV _tr	R_ test	PRESS_ test	R^2CV_ test	R_val	PRESS-val	R^2CV_ val
19	6	7	0.89	4.749	0.639	0.805	3.419	-0.408	0.717	11.561	-2.006
20	6	7	0.88	5.13	0.604	0.763	3.700	-0.500	0.72	11.66	-2.2
21	6	7	0.87	5.527	0.555	0.760	3.974	-1.216	0.70	12.017	-2.39
22	6	7	0.89	4.73	0.645	0.711	4.189	-0.855	0.67	12.458	-2.473

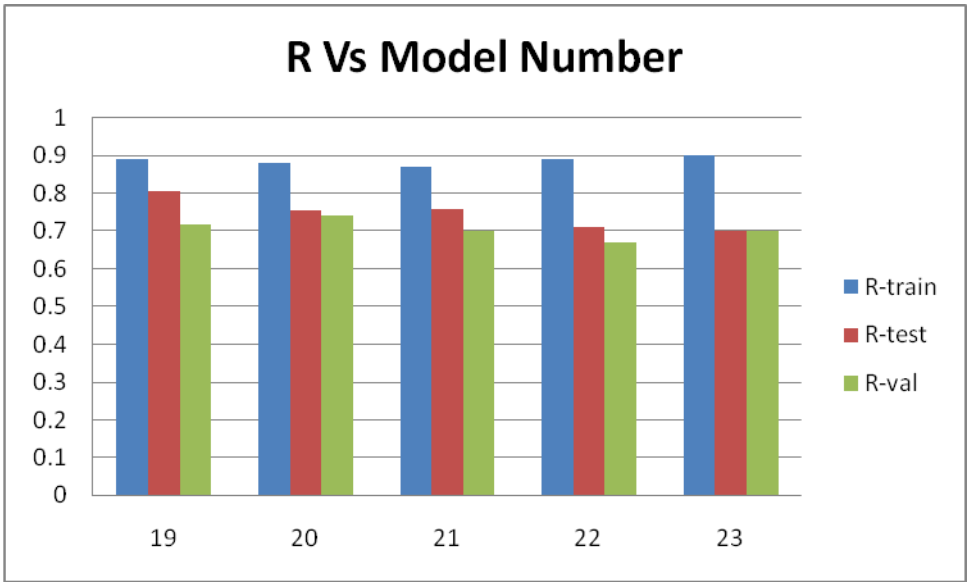
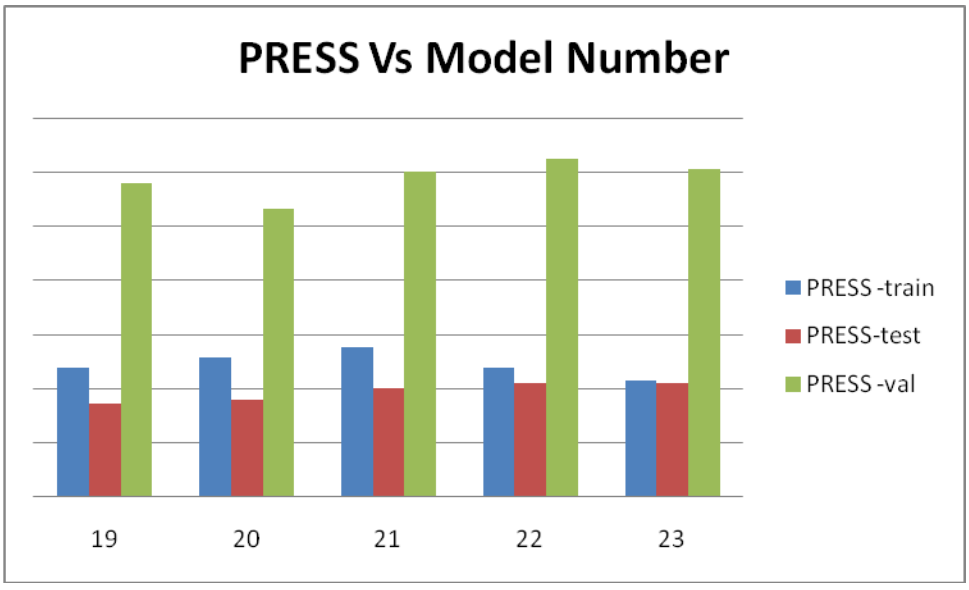


Figure (3.2): Plots of ANN correlation coefficient (R) values for the training,test, and validation sets versus model number .



Figure(3.3): Plots of ANN PRESS (Predictive Residual Sum of Square) values for the training ,test, and validation sets versus model number .

Second ANN performed on models 19, 20 ,21 each model with a range of hidden nodes starting from 5-20. The results shows in tables 3.5, 3.6, and 3.7 respectively.

According to the resulted tables model 21 with 7 nodes,model 20 with 9 nodes and 12 nodes, model 19 with 10 nodes were chosen as the best models with the optimal hidden nodes see table 3.8. This models were chosen because they have high prediction power (R), minimum PRESS values for the test set and minimum number of hidden nodes.

So models 19 and 20 chosen as the best models to continue to randomization test.

Randomization test were done for models 19 with 10 nodes and model 20 with 9 and 12 nodes see tables 3.9 ,3.10, and 3.11 respectively. This test is to ensure that the ANN resulted models are accurate not due to chance .

Referred tables show that the correlation coefficientobtained by chance are low in general while PRESS values are high.

Models 19 and 20 proved to be good models and results are not due to chance .

Table3.5: Correlation Coefficients and cross validation parameter of number of hidden nodes of model #19

Hn	NPCs	R-train	PRESS-tr	R2CV-tr	RSEP-tr	R-test	PRESS-test	R2CV-test	RSEP-test	R-VAL	PRESS-val	R2CV-val	RSEP-val
5	6	0.856	6.097	0.466	3.419	0.762	3.858	-0.897	4.552	0.740	11.920	-3.140	8.030
6	6	0.877	5.193	0.597	3.156	0.764	3.696	-0.550	4.455	0.737	11.410	-2.280	7.864
7	6	0.890	4.749	0.639	3.010	0.805	3.419	-0.408	4.290	0.717	11.561	-2.006	7.900
8	6	0.890	4.687	0.662	2.998	0.802	3.369	-0.266	4.253	0.744	10.818	-1.557	7.656
9	6	0.840	6.770	0.463	3.603	0.832	3.166	-0.350	4.123	0.732	11.607	-2.463	7.930
10	**6**	**0.860**	**5.740**	**0.587**	**3.319**	**0.812**	**2.939**	**0.120**	**3.973**	**0.778**	**10.170**	**-1.422**	**7.423**
11	6	0.880	4.992	0.710	3.094	0.800	3.233	0.012	4.167	0.775	9.726	-0.842	7.259
12	6	0.860	5.543	0.640	3.260	0.818	3.101	-0.199	4.081	0.754	10.483	-1.317	7.536
13	6	0.874	5.080	0.686	3.120	0.814	3.032	0.092	4.030	0.779	9.912	-1.050	7.328
14	6	0.870	5.370	0.630	3.209	0.814	2.900	0.146	3.970	0.765	10.217	-1.188	7.440
15	6	0.860	5.900	0.673	3.367	0.817	3.427	-0.534	4.290	0.765	9.769	-0.832	7.275

16	6	0.870	5.200	0.687	3.160	0.810	3.120	0.114	4.090	0.766	10.250	-1.042	7.452
17	6	0.899	4.153	0.750	2.822	0.820	3.122	0.063	4.094	0.753	10.385	-1.210	7.500
18	6	0.880	4.892	0.699	3.063	0.822	2.922	0.011	3.962	0.768	10.711	-1.840	7.620
19	6	0.910	3.854	0.790	2.720	0.820	2.900	0.128	3.947	0.780	9.796	-0.920	7.285
20	6	0.860	5.780	0.600	3.330	0.814	2.920	0.289	3.962	0.770	10.357	-1.430	7.490

Table3.6: Correlation Coefficients and cross validation parameter of number of hidden nodes of model #20

Hn	nPCs	R-train	PRESS-tr	R2CV-tr	RSEP-tr	R-test	PRESS-test	R2CV-test	RSEP-test	R-VAL	PRESS-val	R2CV-val	RSEP-val
5	6	0.820	7.48	0.320	3.80	0.76	3.60	-0.35	4.398	0.70	12.58	-3.47	8.25
6	6	0.874	5.40	0.560	3.20	0.76	3.89	-0.74	4.569	0.71	12.02	-2.68	8.07
7	6	0.880	5.13	0.604	3.14	0.76	3.70	-0.50	4.461	0.71	11.66	-2.21	7.95
8	6	0.860	5.68	0.570	3.30	0.76	3.67	-0.30	4.440	0.73	11.59	3.40	7.90
9	**6**	**0.880**	**4.73**	**0.670**	**3.01**	**0.77**	**3.06**	**-0.17**	**4.400**	**0.72**	**11.19**	**-1.50**	**7.78**
10	6	0.890	4.61	0.650	2.97	0.78	3.53	-0.02	4.300	0.73	10.77	-1.29	7.63
11	6	0.896	4.43	0.680	2.90	0.77	3.52	-0.26	4.340	0.73	10.99	-1.50	7.72
12	**6**	**0.860**	**5.62**	**0.644**	**3.28**	**0.79**	**3.16**	**0.35**	**4.123**	**0.74**	**10.55**	**-1.20**	**7.56**
13	6	0.903	4.31	0.669	2.90	0.77	3.6	-0.21	4.390	0.73	10.88	-1.40	7.60
14	6	0.850	6.19	0.600	3.44	0.81	2.95	0.11	3.980	0.75	10.12	-0.90	7.40
15	6	0.880	5.20	0.600	3.16	0.82	2.83	0.34	3.900	0.74	10.37	0.90	7.49
16	6	0.903	4.04	0.730	2.08	0.79	3.33	-0.02	4.230	0.74	10.27	-0.92	7.46
17	6	0.880	4.88	0.709	3.06	0.79	3.90	0.22	4.070	0.73	10.11	-0.63	7.40

18	6	0.890	4.66	0.730	2.99	0.78	3.31	-0.08	4.210	0.73	11.20	-1.86	7.79
19	6	0.890	4.67	0.650	2.99	0.79	3.34	-0.27	4.240	0.78	9.98	-1.30	7.35
20	6	0.910	3.92	0.765	2.74	0.78	3.26	0.11	4.180	0.73	10.62	-0.82	7.58

Table3.7: Correlation Coefficients and cross validation parameter of number of hidden nodes of model #21

Hn	NPCs	R-train	PRESS-train	R2CV-train	RSEP-train	R-test	PRESS-test	R2CV-test	RSEP-test	R-val	PRESS-val	R2CV-val	RSEP-val
5	6	0.83	6.89	0.42	3.63	0.76	3.55	-0.10	4.36	0.70	12.34	-3.90	8.18
6	6	0.83	6.80	0.45	3.60	0.75	4.01	-0.90	4.64	0.70	12.36	-3.10	8.18
7	**6**	**0.88**	**4.90**	**0.62**	**3.09**	**0.73**	**4.05**	**-0.82**	**4.66**	**0.72**	**11.34**	**-1.90**	**7.80**
8	6	0.81	7.80	0.20	3.80	0.80	3.20	-0.13	4.20	0.74	11.12	-1.86	7.76

9	6	0.85	6.02	0.60	3.40	7.60	3.60	-0.20	4.40	0.73	11.40	-2.20	7.85
10	6	0.86	5.80	0.60	3.30	0.75	3.83	-0.60	4.50	0.74	11.14	-2.11	7.77
11	6	0.83	6.70	0.50	3.58	0.75	3.50	-0.15	4.30	0.75	10.82	-1.60	7.60
12	6	0.83	6.74	0.54	3.59	0.76	3.80	-0.43	4.50	0.78	9.60	-0.90	7.20
13	6	0.86	5.90	0.54	3.37	0.77	3.23	0.21	4.20	0.79	9.24	-0.80	7.07
14	6	0.85	5.97	0.60	3.38	0.78	3.16	0.23	4.12	0.77	10.46	-1.40	7.50
15	6	0.87	5.20	0.70	3.15	0.78	3.32	-0.07	4.22	0.76	10.50	-1.38	7.53
16	6	0.85	6.10	0.61	3.40	0.76	3.56	-0.20	4.30	0.74	11.36	-2.20	7.84
17	6	0.88	4.90	0.71	3.06	0.77	3.28	0.18	4.20	0.75	9.90	-0.58	7.34
18	6	0.89	4.67	0.73	2.99	0.75	3.49	0.35	4.33	0.73	10.30	-0.70	7.46
19	6	0.91	3.90	0.78	2.70	0.75	3.69	-0.12	4.42	0.72	10.84	-0.80	7.66
20	6	0.91	3.70	0.77	2.70	0.74	4.02	-0.09	4.74	0.70	11.35	-1.02	7.80

Table 3.8: Summary of the correlation coefficients and cross validation parameters of the optimal number of hidden nodes of each model.

Model no	Hn	nPCs	R-train	PRESS-train	R^2CV-train	RSEP-train	R-test	PRESS-test	R2CV-test	RSEP-test	R-val	PRESS-val	R2CV-val	RSEP-val
19	10	6	0.86	5.74	0.587	3.319	0.812	2.94	0.120	3.973	0.77	10.17	-1.42	7.42
20	9	6	0.88	4.73	0.670	3.013	0.77	3.06	-0.166	4.400	0.72	11.19	-1.5	7.78
20	12	6	0.86	5.62	0.644	3.280	0.79	3.16	0.358	4.123	0.74	10.55	-1.2	7.56
21	7	6	0.88	4.90	0.620	3.090	0.73	4.05	-0.82	4.660	0.72	11.34	-1.9	7.80

Table 3.9: chance correlation test for model 19 with 10 nodes.

Trail no	NPCs	R-tr	PRESS-tr	R2CV-tr	R-test	PRESS-test	R2CV-test	R- VAL	PRESS-val	R2CV-val
1	6	0.27	43.80	-38.08	0.12	0.70	-10.51	-0.40	4.69	-86.70
2	6	-0.03	54.35	-6.82	0.13	1.40	-0.75	-0.53	4.70	-27.90
3	6	0.42	38.79	-9.59	-0.02	1.02	-2.02	-0.90	5.60	-26.29
4	6	0.05	48.60	-12.07	-0.36	1.40	-2.68	-0.90	8.06	-18.32
5	6	0.07	48.48	-10.70	-0.70	1.60	-4.50	-0.33	5.20	-20.56
6	6	-0.36	60.86	-13.59	-0.02	1.20	-2.42	0.02	3.99	-30.36
7	6	0.10	46.19	-21.42	0.15	1.34	-1.31	0.40	3.90	-87.04
8	6	0.14	48.50	-5.51	-0.04	1.61	-0.89	0.08	4.00	-52.50
9	6	-0.28	55.50	-19.50	-0.06	1.57	-1.12	0.04	4.72	-44.95
10	6	-0.10	61.32	-5.23	-0.03	5.04	-0.18	-0.70	5.15	-8.97

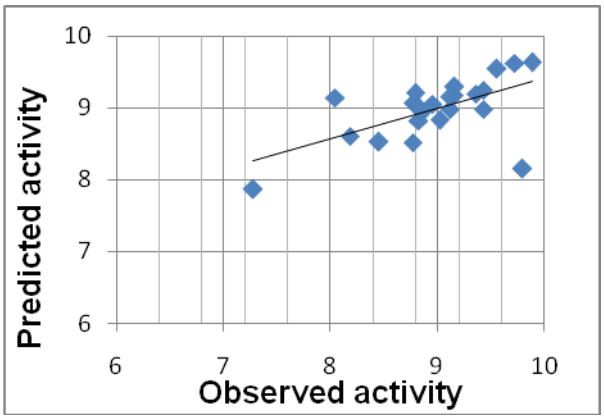
Table 3.10: chance correlation test for model 20 with 12 nodes.

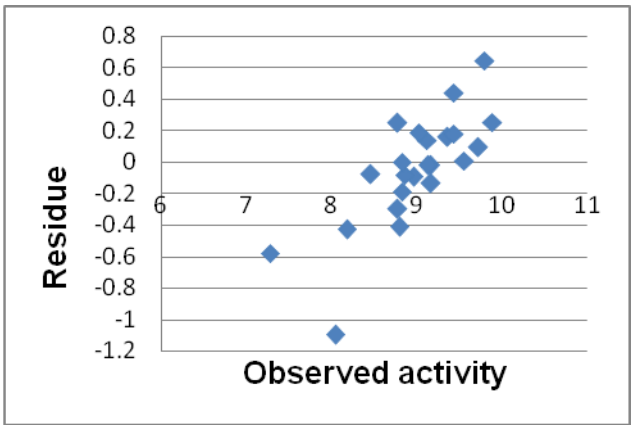
Trail no	nPCs	R-tr	PRESS-tr	R2CV-tr	R-test	PRESS-test	R2CV-test	R- VAL	PRESS-val	R2CV-val
1	6	-0.11	53.45	-121.2	0.05	0.95	-1.80	-0.63	3.80	-41.35
2	6	-0.27	59.30	-11.40	-0.02	0.90	-7.12	0.60	3.50	-16.90
3	6	-0.22	58.70	-9.96	-0.64	1.20	-6.61	0.60	2.40	-9.60
4	6	0.04	53.65	-4.74	-0.01	3.02	-0.33	0.27	3.70	-19.63
5	6	0.11	47.40	-9.78	-0.36	1.80	-1.77	0.90	2.20	-7.08
6	6	-0.10	68.20	-3.42	-0.51	3.38	-1.42	-0.90	6.50	-11.24
7	6	0.01	56.40	-4.20	-0.60	4.70	-1.54	-0.40	4.02	-43.40
8	6	0.12	46.40	-18.40	-0.60	2.15	-3.42	-0.90	8.70	-10.14
9	6	-0.17	56.60	-10.02	0.12	0.70	-3.74	-0.06	4.18	-279.50
10	6	0.21	45.40	-5.99	-0.70	1.70	-4.23	-0.18	5.48	-14.14

Trail no	nPCs	R-tr	PRESS-tr	R2CV-tr	R-test	PRESS-test	R2CV-test	R- VAL	PRESS-val	R2CV-val

Table 3.11: chance correlation test for model 20 with 9 nodes.

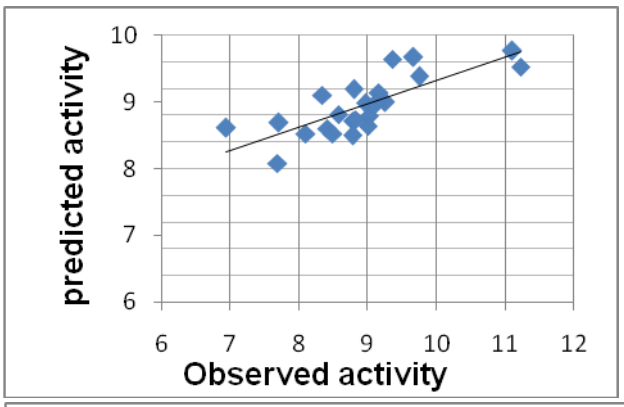
1	6	0.27	43.80	-38.08	0.12	0.7	-10.51	-0.40	4.69	-86.70
2	6	-0.03	54.35	-6.82	0.13	1.4	-0.75	-0.53	4.70	-27.90
3	6	0.42	38.79	-9.59	-0.02	1.0	-2.02	-0.90	5.60	-26.29
4	6	0.05	48.60	-12.07	-0.36	1.4	-2.68	-0.90	8.06	-18.32
5	6	0.07	48.48	-10.70	-0.70	1.6	-4.50	-0.33	5.20	-20.56
6	6	-0.36	60.86	-13.59	-0.02	1.2	-2.42	0.02	3.99	-30.36
7	6	0.10	46.19	-21.42	0.15	1.3	-1.31	0.40	3.90	-87.04
8	6	0.14	48.50	-5.51	-0.04	1.6	-0.89	0.08	4.00	-52.50
9	6	-0.28	55.50	-19.50	-0.06	1.6	-1.12	0.04	4.72	-44.95
10	6	-0.10	61.32	-5.23	-0.03	5.0	-0.18	-0.70	5.15	-8.97

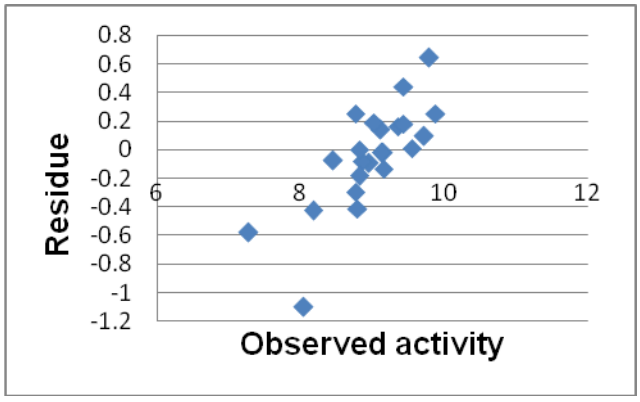




Test set (model 19)

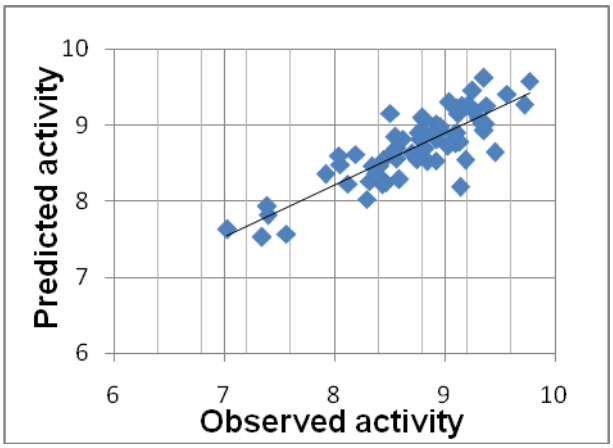
Test set (model 19)

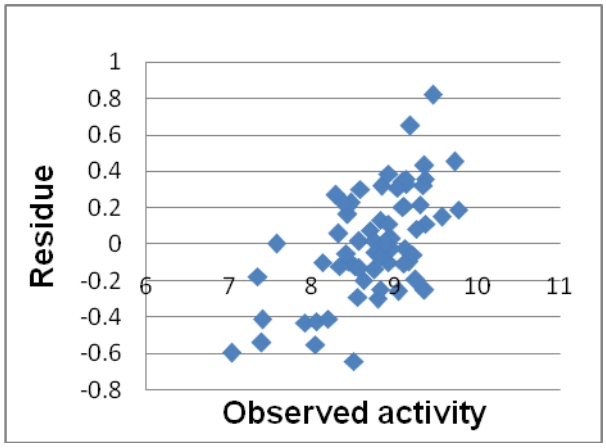




Validation set (model 19)

Validation set (model 19)





Training set (Model19) Training set (Model19)

Figure(3.4): Plot of the predicted against observed one as well as their residue for model 19 using 10 hidden nodes. Training set, Validation set, and external test set.

The following conditions proposed by Golbraikh and Tropsha [76] were applied to conclude that the QSAR model has acceptable prediction power if:

(1) $R^2_{CV} > 0.5$

(2) $R^2 > 0.6$

(3) $(R^2 - R^2_0)/ R^2 < 0.1$ and $0.85 < k < 1.15$

Or

$(R^2 – R'^2_0)/R2 < 0.1$ 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 set 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) $| R^2_0 - R'^2_0 | < 0.3$

Alternatively, the parameter R^2_m ($R^2 * (1 - (R^2 - R^2_0)^{1/2})$) can be used. This parameter 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 sets, its reliability and robustness are high. Model **(19)** validated according to these criteria, and shows to have acceptable prediction power.

.

Structure _ Activity Relationships of the Dataset:

Compounds 1-24 in table 2-1:

Compound 18 has high activity. The presence of the ($CH=CH (CH_2)_2$morph) group in the compound increases the EGFR activity.

The presence of fluorine atom on the first ring has good effect on the activity.

Compounds 25- 44. This group of 4 rings. The structure of the compounds in his group shows good activity in average.

Compounds 78-112

Compound 79 ,80 with high EGFR activity has R group (NMe_2 ,$NHMe_2$). The presence of amino methyl group produces the high activity .

When R group is changed in the compounds 81-112, the activity is decreased.

Compound 113 has low activity; this compound has a different structure than other compounds.

There is a suggestion of a new chemical structure with better activity than the available ones.

According to the previous SAR ,the QSAR for the EGFR inhibitors should have :

1- Four six- membered rings .

2- A fluorine atom and bromine atom attached to the first ring.

3-The (NMe2) group in the compound is useful.

4- The ($CH=CH(CH_2)_2$morph) group is also good for the activity .

5- The existence of N atom in the second and fourth rings is also useful.

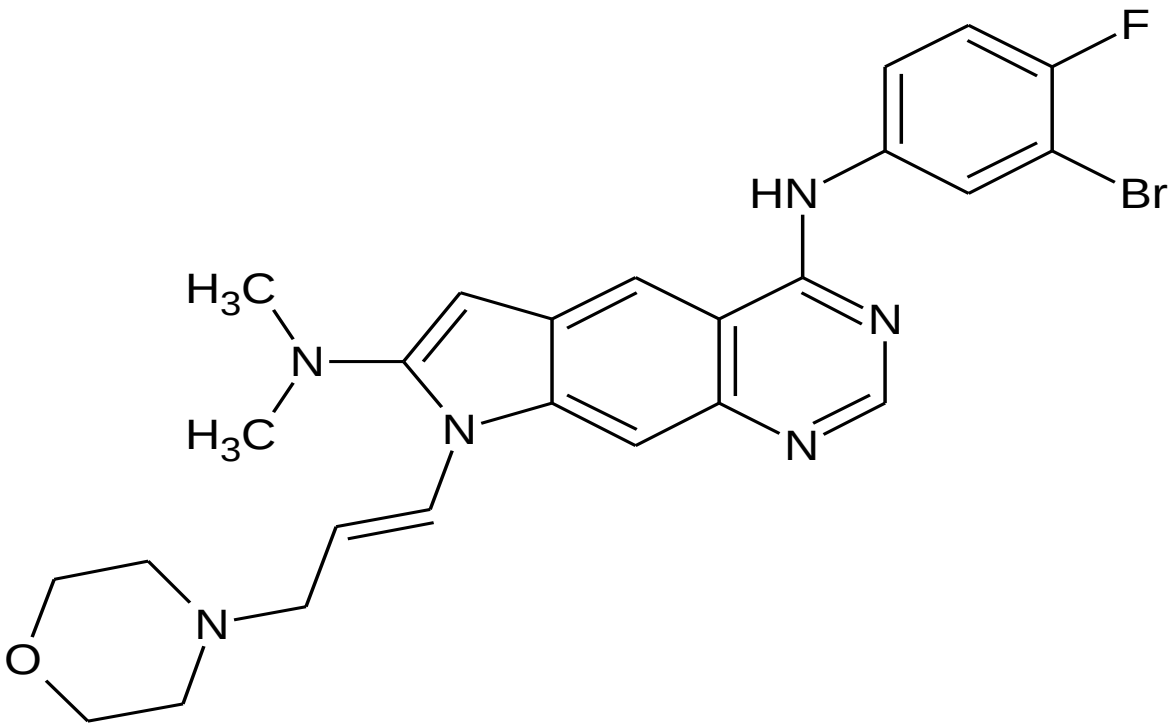


Figure (3-5): suggested compound as EGFR inhibitors

According to model 23 equation , the PIC_{50} of the suggested compound is 30.7. Whereas the PIC_{50} of all the compounds used in the dataset was ranging between 6.93-11.22 .

QSAR studies on EGFR inhibitors:

There are a lot of QSAR studies done on EGFR inhibitors but few of them have similar compounds as the one we used.

In 2003 a three-dimensional QSAR studies for one large set of quinazoline type epidermal growth factor receptor (EGFR) inhibitors were conducted using two types of molecular field analysis techniques: comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA). These compounds belonging to six different structural classes were randomly divided into a training set of 122 compounds and a test set of 13 compounds. The statistical results showed that the 3D-QSAR models derived from CoMFA were superior to those generated from CoMSIA. Where **(R^2_{cv} of 0.60** and conventional **R^2 of 0.92**) [77].

In our work we use similar compounds to the previous study and get good results (R=0.878 ,R^2=0.772 ,R^2adj =0.719).

In 2004 a training set was derived using twenty-eight compounds. Six were used as test sets to evaluate external predictability of the model. The results indicate that thermodynamic and electronic parameters are major contributors to the activity. The same compounds were studied using CoMFA to which the molecules were first aligned to a template molecule. which were most active in the series. Later steric and electrostatic fields were determined across a 3D grid. The CoMFA model was evaluated for both internal and external predictabilities having **(R^2_{cv}=0.456 and R^2_{pred}=0.519)** respectively [78], the results of this study is not good comparing to our results (R=0.878 ,R^2=0.772 ,R^2adj =0.719).

IN 2010 QSAR studies were performed on a set of 61 analogs of 4-anilino quinazoline, A QSAR model was generated by a training set of 42 molecules with correlation coefficient **(R2 of 0.912** , significant cross validated correlation coefficient **(q2) of 0.800 , F test of 60.5149** , r 2 for external test set **(pred_R2) 0.6042** , coefficient of determination (pred_r2 se) 0.7438 and degree of freedom 38 by Multiple Linear Regression (MLR) method [79]. This study have good results but the number of compounds used are small comparing to our results.

In 2017 QSAR studies were performed on a set of 137 analogs of quinazoline, using Multiple Linear Regression (MLR), Principal Component Regression (PCR) and Partial Least Squares (PLS) Regression methods. Among these, MLR method has shown a very promising result as compared to other two methods and a QSAR model was generated by a training set of 100 molecules with correlation coefficient **(R2= 0.884)**, significant cross validated correlation coefficient **(q2= 0.800)**, F test of 39.5149, r2 for external test set (pred_r2) 0.5902, coefficient of correlation of predicted data set (pred_r2se) 0.7438 and degree of freedom 83 [80]. This study have good results as our study (R=0.878, R^2=0.772 ,R^2adj =0.719).

The disadvantage of the previous QSAR studies on EGFR inhibitors was the number of used data set . While in the current study 113 compounds is used. Also the methods used in the current study are MLR and PC-ANN, while in the previous studies MLR alone or other methods which are having less powerful and predication capabilities.

Chapter Four

Conclusions

CONCLUSION:

A quantitative-structure-activity relationship analysis, has been done on a set of 113 compounds of Epidermal Growth Factor Receptor (EGFR) inhibitors. This study is done using MLR and principle component-artificial neural network (PC-ANN) modelling methods. The power and the predictive performance of the models were verified using internal (cross validation and Y-scrambling) and external validation.

The results which we have from MLR were good at a number of models. Models (12-23) have a good (R^2) >0.6, the best was model 23 which included21 descriptors, with R=0.878 ,R^2=0.771 ,R^2adj =0.719 .

Cross validation LOO,LMO were preformed on the chosen models from MLR. Models 19-23 showed a good predictive power because it has a high R^2cv and PRESS/SST less than 0.4 thus models 19 -23 were chosen for ANN analysis.

PCA was performed to divide the data into training, test, and validation sets then ANN performed on the chosen models (19-23) from LOO, LMO validations.

The results show that model 19 have the highest correlation coefficient to the test set (0.805) indicates high predictive power. Models 20 ,21 have good predictive . these models was chosen to continue to ANN to find the optimal number of hidden nodes for each one of these models .

The results were that model 19 of 10 nodes, model 20 of 9,12 nodes ,model 21 of 7 nodes were chosen as best models. These models, have high prediction power, minimum PRESS values of the test set and minimum hidden nodes.

Finally the randomization test was done on chosen models .

A new suggested compound with predicted PIC_{50} of 30.7 was suggested .

References:

1-McDouall, J.J., Computational quantum chemistry molecular structure and properties in silica,Royal Society of Chemistry, 2013. 1st ed.

2- Thakkar, A.J., Yáñez, M.,Wilson, A.K., Computational and Theoretical Chemistry.Computational and Theoretical Chemistry,2016, Volume 1101.

3- Jensen, F., Introduction to Computational Chemistry . Department of Chemistry,University of Southern Denmark,2007. 2nd ed.

4-Schrodinger Equation Available from http://hyperphisics .phy-astr.gsu/hbase/quantum/schr.htmI#c1.

5- Young, D., A Practical Guide for Applying Techniques to Real World Problems.Computational Chemistry,2001. 3rd ed.

6-Selassie, C.D., Mekapati, S.B., Verma, R.P., QSAR :then and now, Current Topics in Medicinal Chemistry, 2002, vol. 2, issue 12,pp. 357-79.

7- Ceyda, O., Xue, Z., Wang, Quantitative Structure Activity Relationship (QSAR) models. Research infrastructure Quality Nano, University of Leeds, 2013.

8-Meyer,H.,Welche Eigenschaft der Anaesthetica bedingt ihre narkotische Wirkung. Arch Expo Pathol Pharmakol (Naunyn –Schmiedebergs),1899,vol. 42, pp.109-118.

9- Overton, E., Studien uber die Narkose Fisher. Jena ,Germany,1901. pp. 209-213.

10- Chekasov, A., Muratov, E.N., Fourches D., Varnek, A., Baskin, I., Cronin, M., Dearden, J., Gramatica, P., Martin, C., Todeschini, R., Consonni, V., Kuz'min, V.E., Cramer,R., Benigni, R., Yang, C., Rathman, C.,.Terfloth, L., and Tropsha, A., QSAR Modeling: Where have you been? Where are you going to?. Journal of Medical Chemistry, 2014. Vol. 57, issue 12, pp. 4977-5010

11- Ferguson, J., P.R.S.B., London Ser, 1939

12-Hansch, C., Quantitative Approach to Biochemical Structure-Activity Relationships. Acc. Chem. Res ,1969;2:232–239.

13- Gramatica, P.,A SHORT STORY OF QSAR EVOLUTION. QSAR Research Unit in Environmental Chemistry and Ecotoxicology ,2011.

14-Hansch, C., Muir, R., Fujita, T., Maloney, P., Geiger, F., Streich, M., The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. J. Am. Chem,1963, Soc. 1963;85:2817–2824.

15- Sood ,A., Computational Chemistry Book and Application, DRUG design by computer ,2010.

16-Deeb,O., Jawabreh, S., and Goodarzi, M., Exploring QSARs of vascular endothelial growth factor receptor-2 (VEGFR-2) tyrosine kinase inhibitors by MLR,PLS and PC-ANN. Current pharmaceutical design ,2013. Vol.19, issue 12, pp.2237-2244.

17-Deeb ,O.,Shaik, B., and Agrawal, V.K., Exploring QSARs of the interaction of flavonoids with GABA(A) receptor using MLR ,ANN and SVM techniques. Journal of enzyme inhabition and medical chemistry ,2014 .vol. 29, issue 5, pp.670-676.

18- Ramerez-Gallicia,G., Deeb. O., Exploring QSAR of antiamoebic agents of isolated natural products by MLR ,ANN, and RTO,Medicinal Chemistry Research ,2012. 12(9): p.2501-2516.

19-T. Fujita ,D. A. Winkler,Understanding the Roles of the “Two QSARs,J. Chem. Inf. Model., 2016, pp 269–274

.20-. Sun M., Chen J., Cai J., Cao M., Yin S. and Ji M. , Simultaneously OptimizedSupport Vector Regression Combined With Genetic Algorithm for QSAR Analysis of KDR／VEGFR-2 Inhibitors. Chem Biol Drug Des, 2010, 75: pp 494–505.

21-. Wegner, J.K. Frohlich, H. Zell, A. Feature selection for descriptor basedclassification models. 1. Theory and GA-SEC algorithm. J. Chem. Inf. Comput. Sci,2004, 44: pp 921–930.

22- Ul-Haq Z., Mahmood U.,Reza S., Uddin R., and Aleem M., Ligand-Based 3D-QSAR Studies of Diaryl Acylsulfonamide Analogues as Human Umbilical Vein Endothelial Cells Inhibitors Stimulated by VEGF. Chem Biol Drug Des,2011, 77: pp 288–294.

23- Dua J., Lei B., Qin J. , H. Liu , Yao X., Molecular modeling studies of vascular endothelial growth factor receptor tyrosine kinase inhibitors using QSAR and docking. Journal of Molecular Graphics and Modeling, 2009, 27: pp 642–654.

24- Dudek A. Z., Arodz T. and Galvez J.,Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review, Combinatorial Chemistry & High Throughput Screening,2006, pp 213-228 213.

25- Deeb, O., and Hemmateenejad B.,"ANN-QSAR model of drug-binding to human serum albumin". Chemical Biology & Drug Design,2007 , 70: pp 19-29.

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

27- Deeb. O., and Goodarzi.M., Exploring QSARs for Inhibitory Activity of Nonpeptide HIV-1 Protease Inhibitors by GA-PLS and GA-SVM", Chemical Biology and Drug Design. ,2010 ,75(5): pp 506-514.

28-Deeb. O., and Drabh. M.,Exploring QSARs of Some Analgesic compounds by PC-ANN". Chemical Biology and Drug Design ,2010, 76(3): pp 255-262.

29- BALEKUNDRI, U., MADAGI, S., A PIC50 Prediction Tool For 5-ALPHA Reductase Enzyme. ASIAN JOURNAL OF PHARMACUTICAL AND CLINICAL RESEARCH ,2016, vol 9.

30-Clark,D.E.,ed.Evolutionary Algorithms in Molecular Design. Methods and Principles in Medicinal Chemistry , Wiley –VCH ,2000,vol 8.

31- Deeb,O.,and Jawabreh, M., Exploring QSARs for inhibitory Activity of Cyclic Urea and Nonpeptide –Cyclic Cyanoguanidine Derivatives HIV-1 protease Inhibitors by Artificial Neural Network , 2012 .

32- Yanga, X,. Kozielb, S.,and Leifssonb, L., Computational Optimization, Modelling and Simulation: Recent Trends and Challenges. Procedia Computer Science,2013, vol. 18, pp. 855-860.

33-Krieger, J.H., Computational Chemistry Impact. C&E News, 1997, p 30.

34-Brouwer, F., Chapter 5. Quantum chemistry in Molecular Modeling of Organic Chemistry, University of Amsterdam,1995.

35-Feynman ,R.P., and Hibbs, A.R., Quantum mechanics and path integrals.McGraw-Hill New York , 1956.Emended Edition.

36-Tranmer M., and Elliot M., Multiple Linear Regression, www.ccsr.ac.uk/publications/teaching/mlr.pdf.

37- Wegner, J.K. Frohlich, H. Zell A., Feature selection for descriptor basedclassification models. 1. Theory and GA-SEC algorithm, Journal of Chemical Information and modeling, 2004, Sci,20044.pp 921–930.

38- Kusrin, A., Beresford,R.,Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis,2000, Vol 22. issue 5, pp. 717-727.

39- Rojas, R.,Neural Networks A Systematic Introduction(Neural Networks Provide Solutions to Real-World Problems: Powerful new algorithms to explore, classify, and identify patterns in data), Neural network ,Springer, Berlin,1996.

40- Matthew, J., Simoneau, A., Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. Technical Articles and Newsletters ,1998.

41-Veerasamy,R.,et al., Validation of QSAR models- strategies and importance .International Journal of Drug & Discovery ,2011. Vol. 3, pp. 511-519.

42-Leach,A.R.,Molecular modeling : principles and applications. person education, 2001. 2nd ed.

43-Wold, S.,M. Sjostrom, and L. Eriksson,Partial Least squares projections to latent structures (PLS) in chemistry .Encyclopedia of computational chemistry ,1998. pp. 2006-2021.

44- HyperChem® Release 5.0 for Windows,1996,test.kirensky.ru/.../GetStart.pdf.

45- Todeshini, R., Consonni, V., Molecular Descriptors for Chemoinformatics ,WILEY-VCH,Weinheim,2009 . 2nd ed.

46-Nie,N.H., Bent, D.H., Hell, C.H., SPSS :Statisticalpackage for the social science . McGraw-Hill NEW York ,1975.

47- Krishan Lal, An Overview to SPSS, Agricultural Statistics Research Institute Library Avenue, New Delhi,klkalra@iasri.res.in.

48-Kalnins, L.M., MATLAB Basics,2010.

49-Martinez ,W.L. and Martinez, A.R., Computational statistical handbook with MATLAB2007 : CRC press .

50- Ferguson, K.M., Structure-based view of epidermal growth factor, receptor regulation, Annual Review of Biophisics, 2008; vol. 37, pp. 353–373.

51- Goyal, S., Jamal, S., Shanker A., and Grover, A.,Structural investigations of T854A mutation in EGFR and identification of novel inhibitors using structure activity relationships, BMC Genomics, 2005, Vol 16, pp. 5-8.

52-Morgillo, F,. C. M. D. Corte, M Fasano, F. Ciardiello,Mechanisms of resistance to EGFR-targeted drugs: lung cancer,British Medical Journal, 2016. Vol.1, issue 3.

53-Cappuzzo, F., The Human Epidermal Growth Factor Receptor (HER) Family: Structure and Function,Guide to Targeted Therapies: EGFR mutations in NSCLC,Springer International Publishing Switzerland ,2014. pp. 978-1007.

54-Ferguson, K.M., Structure-based view of epidermal growth factor receptor regulation, Annual Review of BiophisicSs, 2008. Vol.37, pp. 353-373.

55- Eigenbrot, C., structure –function of EGFR kinase Domain and Its Inhibitors, Chapter in EGFR Signaling Networks in Cancer Therapy, Part of the series Cancer Drug Discovery and Development, 2008. pp. 30-44.

56-Hubbard SR[1], Miller WT, Receptor tyrosine kinases: mechanisms of activation and signaling, Journal of Translational Medicine, 2007. Vol. 19, issue 2, pp. 117-123.

57- S. Sogabe, Y., Kawakita, S,. Igaki, H., Iwata, H. Miki, D.R., Cary, T., Takagi, S., Takagi, Y., Ohta, and Ishikawa,T.,Structure-Based Approach for the Discovery of Pyrrolo[3,2-d]pyrimidine-Based EGFR T790M/L858R Mutant Inhibitors, ACS Medical Chemistry Letters ,2012. Vol. 4, issue 2, pp. 201-205.

58-Heimberger, A.B., Suki, D., Yang, D., The natural history of EGFR and EGFRvIII in glioblastoma patients,Journal of Translational Medicine, 2005. Vol. 3 issue 38.

59- Ciardello, F., Tortora, G., EGFR antagonists in cancer treatment. N Engl J Med , 2008, 358,pp. 1160–1174.

60-Sharma, S.V., Bell, D.W., Settleman, J., Haber, D.A., Epidermal growth factor receptor mutations in lung cancer. Nat Rev Cancer ,2007, vol. 7, pp.169–181.

61- Gazdar, A.F., Review Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors, Oncogene, 2009. Vol. 28, pp. 24–31.

62-Roger, J., Daly, Take Your Partners, Please — Signal Diversification by the erbB Family of Receptor Tyrosine Kinases,PupMed Journal,1999. Vol. 16, issue 14, pp. 255-263

63- Schlessinger, J., Cell Signaling by Receptor Tyrosine Kinases,Cell Journal, 2000, Vol. 103, issue 2, pp. 211-225.

64- Morgillo, F., Carminia, M. D., Corte, M.Fasano, F. Ciardiello ,Mechanisms of resistance to EGFR-targeted drugs: lung cancer,British Medical Journal, 2016. Vol. 1, issue 3.

65-Huang, L.,Liwu Fu ,Mechanisms of resistance to EGFR tyrosine kinase inhibitors, Acta Pharmaceutica Sinica B,2015, vol 5 .issue 5, PP. 390-401.

66- Bar,i S.B., Adhikari, S., Surana, S.J.,Tyrosine Kinase Receptor Inhibitors: A New Target for Anticancer Drug Development, Journal of PharmaSciTech, 2012. Vol. 1, issue 2, pp. 36-45.

67- Shtivelman,E.,New Drugs Aim to Defeat Tumor Resistance to EGFR Inhibitors,2015. Vol. 6, issue 29, pp. 814-825.

68-Robert, C., Soria, J.S., Cutaneous side-effects of kinase inhibitors and blocking antibodies, PubMed PMID, 2005. Vol. 6, issue 7, pp. 491-500.

69- Smaill, J., Gorden, W., Loo, L., Greis, K., Chan, H., Reyner, E., Lipka, E., Showalter, H., Vincent, P., Elliott, W., Tyrosine Kinase Inhibitors. 17. Irreversible Inhibitors of the Epidermal Growth Factor Receptor: 4-(Phenylamino)quinazoline- and (Phenylamino)pyrido[3,2-d]pyrimidine-6-acrylamides Bearing Additional Solubilizing Functions. Journal of Medical Chemistry, 2000. Vol. 43, issue 7, pp. 1380-1390.

70- Palmer, D., B., Trumpp-Kallmeyer, S., Fry, D., Nelson, J., Showalter, H.D. O., and Denny, W., 2-Tyrosine Kinase Inhibitors. 11. Soluble Analogues of Pyrrolo- and Pyrazoloquinazolines as Epidermal Growth Factor Receptor Inhibitors: Synthesis, Biological Evaluation, and Modeling of the Mode of Binding . Journal of Medical Chemistry, 1997. Vol. 40, issue 10, pp.1519-1529.

71- Smaill, J., Hollis,H.D., Zhou, S., Bridges, A,. McNamara, D., Fry, Nelson, Veronika, J., Sherwood, Vincent, P., Roberts, B., Elliott, W., and. Denny, W., Tyrosine Kinase Inhibitors. 18. 6-Substituted 4-Anilinoquinazolines and 4-Anilinopyrido[3,4-d]pyrimidines as Soluble, Irreversible Inhibitors of the Epidermal Growth Factor Receptor. Journal of Medical Chemistry, 2001. Vol.44, issue 3, pp. 429-440.

72- Rewcastle, G., Murray, D., Elliott, W., Fry, D., Howard, C., Nelson, J., Roberts, B., Vincent, P., Showalter, H., Winters, R., and. Denny, W., Tyrosine Kinase Inhibitors. 14. Structure-Activity Relationships for Methylamino- Substituted Derivatives of 4-[(3-Bromophenyl)amino]-6-(methylamino)- pyrido[3,4-d]pyrimidine (PD 158780), a Potent and Specific Inhibitor of the Tyrosine Kinase Activity of Receptors for the EGF Family of Growth Factors .Journal of Medical Chemistry ,1998. Vol. 41, issue 4, pp. 742-751.

73- Rumsey, D., J., Statistics F.or Dummies.The Ohio State University, 2009. 1st. ed.

74- Carbo-Dorca, R., Robert, D., Amat, L., Girona, X., Besalu, E., Molecular Quantum Similarity in QSAR and Drug Design, 2000, ed. 73.

75-Dewar, M., Zoebisch, M., Healy, H., Stewart, J., Development and use of quantum mechanical molecular models. 76. AM1: A new general purpose quantum mechanical molecular model. Journal of the American Chemical Society, 1985, vol. 107, issue 13. P. 3902.

76-.Golbraikh, A. and A. Tropsha, Beware of q 2! Journal of Molecular Graphics and Modelling, 2002. Vol. 20, issue 4 : pp. 269-276.

77-Tingjun, H., Lili, Z., Lirong, C., and Xiaojie X., Mapping the Binding Site of a Large Set of Quinazoline Type EGF-R Inhibitors Using Molecular Field Analyses and Molecular Docking Studies. J. Chem. Inf. Comput. Sci, 2003, 43 (1), pp 273–28.

78- Pednekar, D.V., Kelkar, M.A., Pi.,ple, S.R, Akamanchi, K.J., 3D QSAR STUDIES OF INHIBITORS OF EPIDERMAL GROWTH FACTOR RECEPTOR [EGFR] USING CoMFA AND GFA METHODOLOGIES. Medicinal Chemistry Research, 2004. volume 13,issue 8,pp 605-618.

79- NOOLVI, N., PATEL, M., BHARDWAJ, V., 2D QSAR STUDIES ON A SERIES OF 4-ANILINO QUINAZOLINE DERIVATIVES AS TYROSINE KINASE (EGFR) INHIBITOR: AN APPROACH TO DESIGN ANTICANCER AGENTS. Digest Journal of Nanomaterials and Biostructures, 2010. Vol. 5, issue 2, pp. 387 – 401.

80- Noolvi, N., Patel, M., 2D QSAR Studies on a Series of Quinazoline Derivatives as Tyrosine Kinase (EGFR) Inhibitor: An Approach to Design Anticancer Agents. Letters in Drug Design & Discovery, 2017. Volume 7, Issue 8, pp.556-586.

دراسة العلاقة الكمية بين الفعالية والصيغة البنائي

لبعض المركبات المثبطة للبروتين Epidermal Growth Factor Receptor (EGFR)

باستخدام طريقتي PC- ANN وMLR

إعداد : منال المحتسب

إشراف: أ.د. عمر ديب

الملخص

يتناول موضوع هذا البحث دراسة العلاقة الكمية بين فعالية 113 مركب وصيغتها البنائية QSAR على بروتين Epidermal Growth Factor. وقد تم الحصول على نماذج QSAR باستخدام الانحدار الخطي المتعدد (MLR) كطريقة خطية. بينما تم استخدام الشبكات العصبية (PC-ANN) كطريقة غير خطية . النتائج التي تم الحصول عليها هي نماذج ذات قدرة تنبؤ جيدة . النماذج التي نتجت عن MLR والتي حصلت على معامل ارتباط أعلى من 0.6 هي النماذج 12-23 وكان الأفضل من بينها نموذج رقم 23 مع معامل ارتباط يساوي 0.878 , وتم التحقق من قدرة النماذج على التنبؤ عن طريق استخدام LOO وLMO التي أظهرت أن النماذج 19-23 هي أفضل النماذج . ومن ثم تم تقسيم المركبات إلى ثلاث مجموعات عن طريق استخدام PCA . وقد تم دراسة النماذج 19-23 في الطريقة الغير خطية ANN .

ومن ثم تم التحقق من قدرة وأداء كل النماذج المقترحة من ANN باستخدام (randomization test) وقد وجد أن النموذج رقم 19 هو الأفضل مع معامل ارتباط 0.812 .