Deanship of Graduate Studies
Al-Quds University

Exploring QSARs of Vascular Endothelial Growth Factor Receptor-2 (VEGFR-2) Tyrosine Kinase Inhibitors by MLR, PLS and PC-ANN

Sana' Mahmud Jawabreh

M. Sc. Thesis

Jerusalem-Palestine

1434/2012

Exploring QSARs of Vascular Endothelial Growth Factor Receptor-2 (VEGFR-2) Tyrosine Kinase Inhibitors by MLR, PLS and PC-ANN

Prepared By:

Sana' Mahmud Jawabreh

B.Sc. Chemistry Al-Quds University (Palestine)

Supervisor: Dr. 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

1434/2012

Al-Quds University
Deanship of Graduate Studies
Applied and Industrial Technology
Department of Science and Technology

Thesis Approval

Exploring QSARs of Vascular Endothelial Growth Factor Receptor-2 (VEGFR-2)Tyrosine Kinase Inhibitors by MLR, PLS and PC-ANN

Prepared by:
Student Name: Sana' Mahmud Jawabreh
Registration number: 20912701

Supervisor: Dr. Omar Deeb

Master thesis submitted and accepted Date:
The names and signatures of the examining committee members are as follows:

1- Head of Committee / Dr. Omar Deeb signature.....
2- Internal Examiner / Dr. Mohammad Abu alhaj signature.....
3- External Examiner / Dr.Sameer AL-Najdi signature.....

Jerusalem – Palestine

1434/2012

To my parents for their love and encouragement

and

To my brothers, sisters, and my husband for constant support

and

To my dearest friends for the best

time we shared

Declaration

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

Signed:

Sana' Jawabreh

Date:

Acknowledgments:

I would like to express my endless thanks to ALLAH.

Then I would like to thank my supervisor Dr. Omar Deeb for his endless help, patience, and encouragement during my research.

Finally I would like to thank everyone who helped me at Al-Quds University.

Abstract

Vascular endothelial growth factor receptor-2 (VEGFR-2) is attractive target for the development of novel anticancer agents. In this work quantitative structure-activity relationship (QSAR) study was performed to understand the inhibitory activity of a set of 192 VEGFR-2 compounds. While we chose those compounds we didn't depend on some core or some compound derivatives, we chose varied compounds and that is to get a comprehensive model can be applied on a large number of compounds. QSAR models were developed using multiple linear regression (MLR) and partial least squares (PLS) as linear methods. While principal component - artificial neural networks (PC-ANN) modeling method was used as nonlinear method. The results obtained offer good regression models having good prediction ability. The results obtained by MLR and PLS are close and better than those obtained by principal component- artificial neural network. The best MLR model was obtained with a correlation coefficient of 0.870, and the correlation coefficient value of the best PLS model is 0.860, while the correlation coefficient value of the best PC-ANN model is 0.752. The strength and the predictive performance of the proposed models was verified using both internal (cross-validation and Y-scrambling) and external statistical validations.

دراسة العلاقة بين الصيغة البنائية والفاعلية باستخدام طريقة MLR,PLS, PC-ANN لمثبطات مستقبلات عامل نمو بطانة الأوعية الدموية.

إعداد الطالبة: سناء محمود جوابرة

إشراف: د. عمر ديب

الملخص

لقد تمت دراسة العلاقة الكمية بين الصيغة البنائية بأبعادها الثلاثة والفاعلية لمئة واثنان وتسعون مركبا كمثبطات لمستقبلات عامل نمو بطانة الأوعية الدموية. وقد تم مراعاة التنوع في اختيار المركبات وعدم الاعتماد على تركيب كيميائي معين او التقيد بفصيلة مركبات معينة وذلك للحصول على علاقة عامة نستطيع تطبيقها على اكبر قدر من المركبات المتنوعة . لدراسة هذه العلاقة تم بناء علاقات كمية خطية باستخدام (MLR, PLS), كما تم استخدام (PC-ANN) لبناء علاقات كمية غير خطية. وبمقارنة النتائج التي تم الحصول عليها باستخدام العلاقات المختلفة التي نفذناها لاحظنا ان نتائج العلاقات الخطية كانت متقاربة وأفضل من تلك التي تم الحصول عليها بواسطة العلاقات غير الخطية. حيث أن معامل الارتباط للعلاقة الافضل التي تم الحصول عليها بواسطة طريقة MLR هو 0.870 بينما قيمته لأفضل علاقة تم الحصول عليها بواسطة PLS هو 0.860 أما قيمة معامل الارتباط لأفضل علاقة باستخدام PC-ANN هي 0.752 . وكانت العلاقات التي تم التوصل اليها جيدة وذات قدرة على التنبؤ بفاعلية مركبات اخرى لم تستخدم في بناء هذه العلاقة.

Table of contents:

List of tables:

List of figures:

List of abbreviation:

Abbreviation	Meaning
EC	Endothelial cells
BM	Basement membrane
DR	Diabetic retinopathy
AMD	Age-related macular degeneration
RA	Rheumatoid arthritis
VEGF	Vascular endothelial growth factor
VEGFR-1	Vascular endothelial growth factor receptor- 1
VEGFR-2	Vascular endothelial growth factor receptor - 2
KDR	Kinase insert domain receptor
MVD	Microvascular density,
QSAR	Quantitative structure activity relationships
MLR	Multiple linear regression
PLS	Partial least squares
PC-ANN	Principle component artificial neural networks
R	Correlation coefficient.
R^2	Coefficient of determination
PRESS	Predictive residual sum of squares
SSR	Regression sum of squares
SST	Total sum of squares
SSE	Error sum of squares
LV	Latent variables
PCR	principal component regression
SPSS	Statistical Package for the Social Sciences
LOO	Leave one out
LMO	Leave many out
pIC_{50}	Log (The half maximal inhibitory concentration)

AM1	Austin Model 1
EHOMO	Highest occupied molecular orbital energy
ELUMO	Lowest unoccupied molecular orbital energy
DM	Molecular Dipole moment
PSE	Predictive Square Errors
RMSE	Root mean square error
CoMFA	Comparative molecular field analysis
CoMSIA	Comparative molecular similarity indices analysis
LSSVMs	Least Squares Support Vector Machines

Chapter one
Introduction

1.1 Angiogenesis:

1.1.1 Definition:

Angiogenesis is one of the most pervasive and fundamentally essential biological processes encountered in mammalian organisms. (1) It is the process of blood vessel sprouting and generating new capillaries from existing vasculature, and an important event in a variety of physiological processes including ovulation, embryonic development, wound repair and collateral generation in the myocardium. Angiogenesis is tightly regulated in both time and space. It is driven by a cocktail of growth factors and proangiogenic cytokines and tempered by an equally diverse group of inhibitors of neovascularization. (2)

There are number of key steps in the angiogenic cascade including the reactivation of endothelial cells (EC), rupture of basement membrane (BM), adhesion, migration, proliferation, table formation, and sprouting of new capillary blood vessels of preexisting vessels. The rate of basement membrane (BM) synthesis has been shown to directly correlate with the formation of new blood vessels. Several studies suggested that different BM play a pivotal role in angiogenesis. BM is not only an essential element of all blood vessels, but it also plays the role of a local hormone for activated EC. In that regard, BM biosynthesis might represent an ideal target for developing suppressors of angiogenesis. (1)

In adult organs, the turnover of EC is an extremely slow process, which accelerates only in a few physiological situations such as embryogenesis, ovulation, and wound healing. Under these special circumstances angiogenesis lasts for a relatively short time (days-weeks) then return to a quiescent state in a self limited and well regulated both temporally and spatially through a well coordinated and balanced angiosuppressors and angiopromoters. In contrast, pathological angiogenesis can last for years and somewhat out of control due to unrestricted production of normal or aberrant forms of proangiogenic mediators and /or the result of a relative deficiency in angiogenic inhibitory molecules. This can result in various forms of pathological angiogenesis including: (1)

1.Ocular neovascularization-mediated diseases: diabetic retinopathy (retinal neovascularization) and age-related macular degeneration (choroidal neovascularization), DR & AMD, respectively.

2. Cancer: metastasis of solid tumor.

3. Chronic inflammatory disorders: rheumatoid arthritis (RA), psoriasis.

4. Vascular diseases—ischemic heart diseases, atherosclerosis.

Hence understanding the mechanisms involved in the regulation of angiogenesis could have a major impact in the prevention and treatment of pathological angiogenesis processes.

1.1.2 Regulation of Angiogenesis

The organ microenvironment controls the extent of vascularization under physiological and pathological conditions. It has been demonstrated that in tissues, Vascular endothelial growth factor (VEGF) and corresponding receptors VEGFR-1 (Flt-1) and VEGFR-2 [kinase insert domain receptor (KDR)] are critical regulators of angiogenesis and that specific binding of VEGF to VEGFR-2 would initiate effective downstream cell proliferation signaling pathways and even leads to tumor vascularization, which promotes tumor infiltration and metastasis. (3) Conversely, inhibition of VEGF action by specific monoclonal antibodies or soluble receptors results in suppression of neovascularization associated with tumors and retinopathies. Endothelial cells play a major role in the modeling of blood vessels. The interplay of growth factors, cell adhesion molecules and specific signal transduction pathways either in the maintenance of the quiescent state or in the reactivation of endothelial cells is well coordinated. (1)

Vascular Endothelial Growth Factor (VEGF):

It is a member of angiogenic mediators that act directly on the EC .VEGF is a potent mitogen and angiogenesis-promoting factor in vivo. VEGF has been shown to be secreted in response to hypoxic or ischemic insults with the subsequent initiation and amplification of neovascularization. VEGF appears to be a crucial mediator of blood vessel growth associated with tumors and proliferative retinopathies. Strong experimental evidence indicates that VEGF

is a major mediator of angiogenesis associated with most human tumors and also ischemic retinal disorders. (2)

Anti-VEGF monoclonal antibodies or soluble receptors suppress neovascularization in a variety of animal models. Therefore, a humanized anti-VEGF antibody has therapeutic value for a variety of disorders where angiogenesis plays a significant role.(1)

Tyrosine Kinases

Two tyrosine kinases have been identified as putative VEGF receptors including VEGFR-1 (Flt-1) and VEGFR-2 tyrosine kinases. Both are located in EC. VEGFR-1 can be secreted and competitively inhibit VEGF-induced angiogenesis. VEGFR-1 (Flt-1) and VEGFR-2 form a new subfamily of receptor tyrosine kinases and both are predominantly expressed in vascular endothelial cells during blood vessel formation and remodeling. Knockout mice for VEGFR-1 (Flt-1) and VEGFR-2 genes demonstrated the critical role of VEGFR-1 in vascular formation and VEGFR-2 in sprouting of new capillary blood vessel from preexisting blood vessels. (1)

1.1.3 Angiogenesis and Cancer

Cancer researchers studying the conditions necessary for cancer metastasis have discovered that one of the critical events required is the growth of a new network of blood vessels. Tumor angiogenesis is the proliferation of a network of blood vessels that penetrates into cancerous growths, supplying nutrients and oxygen and removing waste products. Tumor angiogenesis actually starts with cancerous tumor cells releasing molecules that send signals to surrounding normal host tissue. This signaling activates certain genes in the host tissue that, in turn, make proteins to encourage growth of new blood vessels. (2)

Tumor progression and metastasis are classically conditioned where cells escape normal growth and adhesion controls and invade, migrate, attach and grow at inappropriate sites. Angiogenic factors including growth factors, cytokines, and cell adhesion molecules are known to control many of those events. Substantial evidence has been accumulated over the years pointing to the dependency of solid tumors on angiogenesis which was first proposed by Folkman. (3)

As the advancing edge of the tumor approach adjacent microvessels, proangiogenic factors are released from the tumor stimulating EC to grow and migrate toward the tumor and organize into a capillary network. This switch from the prevascular to vascular phase is accompanied by exponential growth of the tumor. An increase in the microvascular density, (MVD) in breast and prostate carcinoma has been shown to correlate with malignant and metastatic potential and hence with a poor prognosis.

Tumor cells recruit new blood vessels via various angiogenic factors and are further amplified by the release of cytokines, which attract and activate macrophages, mast cells and neutrophils. Research involving tumor-associated angiogenesis continues to yield new insights into the pathogenic mechanisms of this process. Based on this found understanding, innovative and novel therapeutic approaches targeting various steps in this process as well as the neovasculature itself may be developed.(1)

In addition, as more is learned about the biology of angiogenesis, biological markers may be developed that can facilitate of clinical trials. Specific agents currently in clinical trials, as well as other approaches under development that act at various points in the complex process involved in neovascularization, may soon have an impact on the treatment of neoplastic diseases. (3)

1.2 Computational chemistry

Chemists have been some of the most active and innovative participants in this rapid expansion of computational science. Computational chemistry is simply the application of chemical, mathematical and computing skills to the solution of interesting chemical problems. It uses computers to generate information such as properties of molecules or simulated experimental results. (4)

Computational chemistry has become a useful way to investigate materials that are too difficult to find or too expensive to purchase. It also helps chemists make predictions before running the actual experiments so that they can be better prepared for making observations. The Schroedinger equation is the basis for most of the computational chemistry scientists' use.

This is because the Schroedinger equation models the atoms and molecules with mathematics. (5)

Currently, there are two ways to approach chemistry problems: computational quantum chemistry and non-computational quantum chemistry Computational quantum chemistry is primarily concerned with the numerical computation of molecular electronic structures by ab initio and semi-empirical techniques and non-computational quantum chemistry deals with the formulation of analytical expressions for the properties of molecules and their reactions. (4)

The following three ways are used to perform numerical computation:

- Ab Initio: The term "Ab Initio" is latin for "from the beginning". This name is given to computations which are derived directly from theoretical principles, with no inclusion of experimental data, calculating molecular structures using nothing but the Schroedinger equation. Most of the time this is referring to an approximate quantum mechanical calculation. The approximations made are usually mathematical approximations, such as using a simpler functional form for a function or getting an approximate solution to a differential equation. (5)

  The good side of ab initio methods is that they eventually converge to the exact solution, once all of the approximations are made sufficiently small in magnitude. However, this convergence is not montonic. Sometimes, the smallest calculation gives the best result for a given property. The bad side of ab initio methods is that they are expensive. These methods often take enormous amounts of computer cpu, time, memory and disk space. In general, ab initio calculations give very good qualitative results and can give increasingly accurate quantitative results as the molecules in question become smaller. (4)

- Semi-empirical techniques: use approximations from empirical (experimental) data to provide the input into the mathematical models. (5)

The good side of semiempirical calculations is that they are much faster than the ab initio calculations. The bad side of semiempirical calculations is that the results can be erratic. If the molecule being computed is similar to molecules in the data base used to parameterize the method, then the results may be very good. If the molecule being computed is significantly different from anything in the parameterization set, the answers may be very poor.

Semiempirical calculations have been very successful in the description of organic chemistry, where there are only a few elements used extensively and the molecules are of moderate size. However, semiempirical methods have been devised specifically for the description of inorganic chemistry as well. (4)

- Molecular mechanics uses classical physics to explain and interpret the behavior of atoms and molecules. (5)

1.3 Quantitative structure activity relationships (QSAR)

In the process of drug design and development, biological activity evaluation forms the premise for compound screening and optimization.(6) Although several experimental methods based on receptors and other biological materials of human, rat, mouse and so on, they are too costly and time-consuming.(7)

QSAR is mathematical model relating chemical structures to their biological activity; give useful information about drug design and medicinal chemistry (8). In1962 Hansch et. al. developed quantitative relationships between biological activity and the octanol water partition coefficient .(9)

Computational methods, especially quantitative structure-activity relationship (QSAR) analysis, provides an effective and powerful tool for achieving the same destination with much lower cost. Actually, QSAR has now been extensively applied to predict compounds' properties, including biological activity, physical property and even toxicity.(10-13)

QSAR studies aimed to:

- Quantitatively correlate the relationships between chemical structure alterations and respective changes in biological endpoint for clarifying which chemical properties are most likely determinants for their biological activities.

- Screening the library for most promising candidates, or optimize the existing leads so as to improve their biological activities.

- Predict the biological activities of untested and sometimes yet unavailable compounds.(14)

- QSAR studies can reduce the costly failures of drug candidates in clinical trials by filtering the combinatorial libraries. Virtual filtering can eliminate compounds with predicted toxic of poor pharmacokinetic properties. It also allows for narrowing the library to drug-like or lead-like compounds and eliminating the frequent-hitters, i.e., compounds that show unspecific activity in several assays and rarely result in leads. Including such considerations at an early stage results in multidimensional optimization.

The importance and difficulty of the above-described tasks facing QSAR models has inspired many chemoinformatics researchers to borrow from recent developments in various fields, including pattern recognition, molecular modeling, machine learning and artificial intelligence. This results in large family of conceptually different methods being used for creating QSARs. (15)

1.3. 1 General Scheme of a QSAR Study

To build QSAR model you need to follow four successive steps; preparing the compounds list, extracting descriptors from molecular structure, choosing the informative descriptors in the context of the analyzed activity, and, finally, using the values of the descriptors as independent variables to correlate them with the activity in question.

Preparing the compounds list and structure optimization:

Compiling a list of compounds for which the experimentally determined activity is known. Then we must obtain the geometry optimization (minimization) of the compounds by using one of two methods: ab initio or semi empirical. The latter one derived from experimental values that simplify theoretical calculations, and usually do not require long computation times, as in the ab initio (10).

Generation of Molecular Descriptors from Structure

Molecular descriptors must then be computed. Any numerical value that describes the molecule could be used. The structure cannot be directly used for creating structure-activity relationship for reasons arising from chemistry and computer science. (8).

First, the chemical structures do not usually contain in an obvious form the information that relates to activity. This information has to be extracted from the structure. Various rationally designed molecular descriptors clarify different chemical properties implicit in the structure of the molecule. Only those properties may correlate more directly with the activity. Such properties range from physicochemical and quantum-chemical to geometrical and topological features. There is a brief description of those features in table (1.1)

Second, most methods employed to predict the activity require as input numerical vectors of features of uniform length for all molecules. Chemical structures of compounds are diverse in size and nature and do not fit into this model directly. To solve this obstacle, molecular descriptors convert the structure to the form of well defined sets of numerical values. (15)

Table (1.1) Brief descriptions of the descriptors 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 index (J), Wiener 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 (χ=-0.5(E_{HOMO}+E_{LUMO})). Hardness (η= 0.5(E_{LUMO} - E_{HOMO})). Softness (S=1\η). Electrophilicity (ω=χ^2/2η). Heat of formation (H_f).
Chemical descriptors	Octanol-water partition coefficient (LogP), hydration energy (HE) polarizability (Pol), refractivity (Ref), volume (V), surface area (SA),

Selection of Relevant Molecular Descriptors

Many applications are capable of generating hundreds or thousands of different molecular descriptors. Typically, only some of them are significantly correlated with the activity. Furthermore, many of the descriptors are intercorrelated. This has negative effects on several aspects of QSAR analysis. Some statistical methods require that the number of compounds is significantly greater than the number of descriptors. Using large descriptor sets would require

large datasets. Other methods, while capable of handling datasets with large descriptors to compounds ratios, nonetheless suffer from loss of accuracy, especially for compounds unseen during the preparation of the model. Large number of descriptors also affects interpretability of the final model. To solve these problems, a wide range of methods for automated narrowing of the set of descriptors to the most informative ones is used in QSAR analysis. (15)

Modeling the Descriptors to Activity

Once the relevant molecular descriptors are computed and selected, the final task of creating a function between their values and the analyzed activity can be carried out. The value quantifying the activity is expressed as a function of the values of the descriptors. The most accurate modeling function from some wide family of functions is usually fitted based on the information available in the training set, i.e., compounds for which the activity is known. A wide range of modeling function families can be used, including linear or non-linear ones, and many methods for carrying out the training to obtain the optimal function can be employed.(15)

1.4 Statistical methods:

Among the widely utilized algorithms applied for model construction in QSAR, in our study we used multiple linear regression (MLR), partial least squares (PLS), Principle Component artificial neural networks (PC-ANN).

1.4. 1 Multiple linear regression (MLR)

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. (10)

Multiple linear regression models are extremely powerful, and have the power to empirically form 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, ... Linear regression models have been heavily studied, and are very well-understood.

The relationship between the dependent variable and independent variables represented by the following equation:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i \quad (1)$$

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. (16)

The explanatory power of the regression is summarized by its "R-squared" value, computed from the sums-of-squares terms as:

$R^2 = SSR/SST = 1- (SSE/SST)$ (2)

Where:

$$SSE = \sum_{i=1}^{n} (y\ obs - y\ pred)^2$$

$$SST = \sum_{i=1}^{n} (y\ obs - \bar{y}\ obs)^2$$

$$SSR = \sum_{i=1}^{n} (y\ pred - \bar{y}\ pred)^2$$

$y\ obs$: Dependent variable found by experiments, $y\ pred$: dependent variable found by calculation. $\bar{y}$: average of y.

R^2, the coefficient of determination, is often described as the proportion of variance described by regression. It is important to keep in mind that a high R^2 does not imply causation. The relative sizes of the sums-of-squares terms indicate how "good" the regression is in terms of fitting the calibration data. If the regression is perfect, all residuals are zero, SSE is zero, and

R^2 is 1. If the regression is a total failure, the sum-of-squares of residuals equals the total sum-of-squares, no variance is accounted for by regression, and R^2 is zero. (17)

When the factors are few in number, are not significantly collinear, and have a well understood relationship to the responses, then multiple linear regression (MLR) can be a good way to turn data into information. However, if any of these three conditions breaks down, MLR can be inefficient or inappropriate. In such so-called soft science applications, the researcher is faced with many variables and ill-understood relationships, and the object is merely to construct a good predictive model. (18)

1.4.2 Partial Least Squares (PLS)

PLS statistical analysis module performs model construction and prediction of activity using the Partial Least Squares (PLS) regression technique. It is based on linear transition from a large number of original descriptors to a small number of orthogonal factors (latent variables) providing the optimal linear model in terms of predictivity. (19)

In principle, MLR can be used with many factors. However, if the number of factors gets too large (for example, greater than the number of observations), you are likely to get a model that fits the sampled data perfectly but that will fail to predict new data well. This phenomenon is called over-fitting. In such cases, although there are many manifest factors, there may be only a few underlying or latent factors that account for most of the variation in the response. The general idea of PLS is to try to extract these latent factors, accounting for as much of the manifest factor variation as possible while modeling the responses well. (18)

Unlike some similar approaches (e.g. principal component regression PCR), latent variables are chosen in such a way as to provide maximum correlation with dependent variable; thus, PLS model contains the smallest necessary number of factors. With increasing number of factors, PLS model converges to ordinary multiple linear regression model. In addition, PLS approach allows one to detect relationship between activity and descriptors even if key descriptors have little contribution to each other.

Because latent variables are the linear combinations of original descriptors (with coefficients represented by loading vector ***p***), factor model indirectly describes the effect of each descriptor on activity. (19)

Partial least squares regression is an extension of the multiple linear regression models. In its simplest form, a linear model specifies the (linear) relationship between a dependent variable Y, and a set of independent variables denoted as $X_1, X_2, \ldots$ so that

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p \qquad (3)$$

Where:

b_0 is the regression coefficient for the intercept and the b_i values are the regression coefficients (for variables 1 through p) computed from the data.(16)

In short, partial least squares regression is probably the least restrictive of the various multivariate extensions of the multiple linear regression models. This flexibility allows it to be used in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than independent variables. Furthermore, partial least squares regression can be used as an exploratory analysis tool to select suitable independent variables and to identify outliers before classical linear regression. (8)

1.4.3 Principal component-Artificial Neural Networks (PC-ANN)

Artificial Neural Networks (ANNs) are a data processing system consisting of a large number of simple, highly interconnected processing elements inspired by the biological system and designed to simulate neurological processing ability of human brain. (8)
Computationally, ANN is an approach for handling multivariate and multi-response data and hence suitable for modeling. Unlike standard modeling techniques where the mathematical function is required to be known in advance, ANN models do not require knowledge of the mathematical function in advance and are called ‘soft models’, i.e. the models are able to represent the experimental behavior of the system when the exact description is missing or too

complex. ANNs adapt to any relation between input and output data on the basis of their supervised training. The characteristics that make ANN systems different from traditional computing are: learning by example, distributed associative memory, fault tolerance and pattern recognition. The flexibility of ANNs and their ability to maintain their performance even in the presence of significant amounts of noise in the input data are highly desirable since perfectly linear and noise free data sets are seldom available in practice, thus making it suitable for multivariate calibration modeling. (20)

The basic processing elements of neural networks are called artificial neurons, or simply neurons or nodes. In a simplified mathematical model of the neuron, the effects of the synapses are represented by connection weights that modulate the effect of the associated input signals, and the nonlinear characteristic exhibited by neurons is represented by a transfer function. The neuron impulse is then computed as the weighted sum of the input signals, transformed by the transfer function. The learning capability of an artificial neuron is achieved by adjusting the weights in accordance to the chosen learning algorithm. (8)

The neurons (hidden units) are non-linear transformation functions. Non linear models can be constructed when more than one of these neurons is used. ANN can model a wide set of functions, as in figure (1.1). The input is multiplied by the connection weight, while products are summed at each neuron where a nonlinear transfer function is applied. The output of each neuron is then multiplied by the connection weight and summed and interpreted until having the minimum error (20).

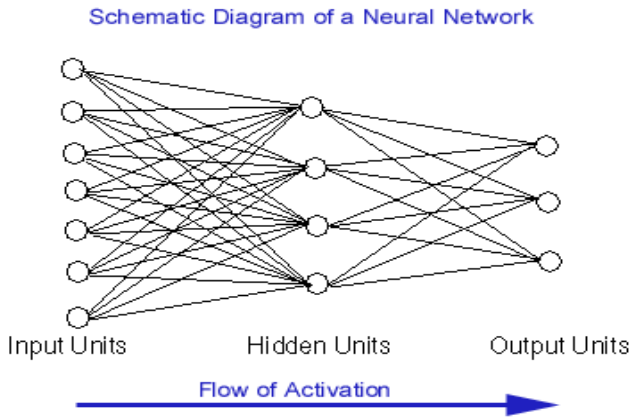


Figure (1.1): Schematic Diagram of a Neural Network.

For better predictive model we used Principle Component Analysis (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 analysis (PCA) groups together variables that are collinear to form a composite indicator capable of capturing as much of common information of those indicators as possible. Each factor reveals the set of variables having the highest association with it. The idea under this approach is to account for the highest possible variation in the indicators set using the smallest possible number of factors. Therefore, the index no longer depends upon the dimensionality of the dataset but it is rather based on the "statistical" dimensions of the data. Application of PCA on a descriptor data matrix results in a loading matrix containing factors or principal components, which are orthogonal and therefore do not correlate with each other. We used these factors as the inputs of ANN instead of the original descriptors.

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. (21)

1.5 Software in QSAR

A huge number of computer programs to serve QSAR produced in the last 50 years, these programs help in the progression of QSAR and make it easy to achieve all QSAR goals.

We used in our study the following four softwares:

HyperChem (version 8.0 HyperCub, Inc.), Dragon (version 3, Milano Chemometrics and QSAR research group, http://www.disat.unimib.it/chm/Dragon.htm), SPSS (version 11.50, SPSS Inc.), Matlab (version 7.0, Math works Inc, http://www.mathworks.com)

1.5.1 Hyperchem

Hyperchem software is a sophisticated molecular modeling environment that is flexible, easy to use, and high quality program. Hyperchem integrates 3D visualization and animation with quantum chemical calculations, molecular mechanics, and dynamics.

Hyperchem can be used to: building molecular structures, structure optimization, calculating some QSAR descriptors, computing some structural properties, studying dynamic behavior, etc.

In this research we used the hyperchem software to build the structure, optimizing the structure, and to calculate some structural properties and molecular descriptors. Before calculation of any property of the molecule, the structure must be optimized (minimized).

The tool menu is shown in figure (1.2) where we can use any tool to draw, display, optimize, calculate, and then we can use the model builder to transform (2D) structures to (3D) structures.

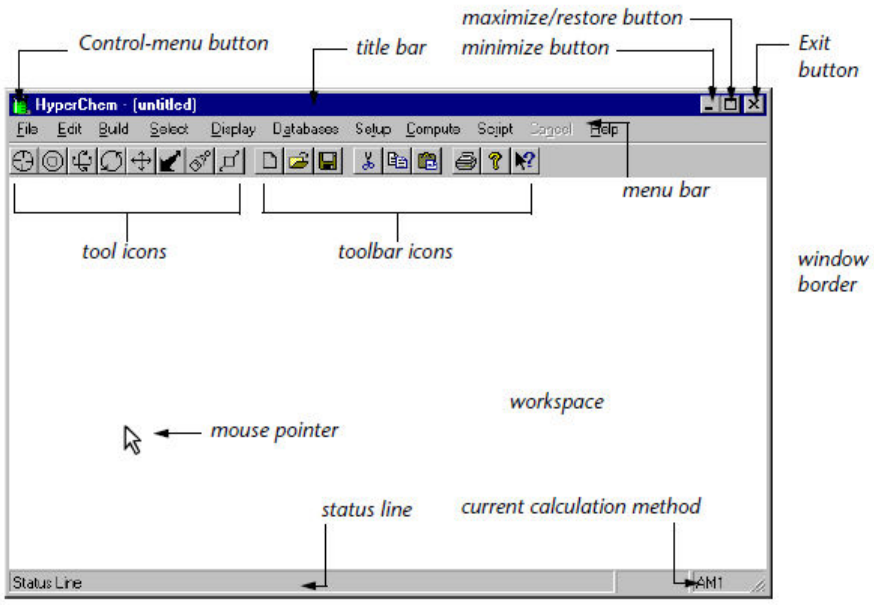


Figure (1.2): Hyperchem main menu.

1.5.2 Dragon software

Dragon software is designed for the calculation of theoretical molecular descriptors, it is developed by Milano chemometrics and QSAR research group to calculate molecular descriptors for molecules containing the following atoms: C, H, O, N, S, P, F, Cl, Br, I, B, Si, Ni, Fe, Co, Al, Cu, Zn, Sn, Gd.

The number of different descriptors that can be calculated by Dragon software is 1481 descriptor, and this huge number of descriptors is divided into 18 groups such as: topological and geometrical descriptors …etc.

The Descriptors calculated by Dragon represent an input to the QSAR analysis programs, these descriptors are the independent variables in the models that we aim to build.

1.5.3 SPSS software

SPSS (Statistical Package for the Social Sciences) was released in 1968 after being developed by Norman H. Nie and C. Hadlai Hull. SPSS is the most widely used program for statistical analysis, it uses two main windows: data editor and output viewer.

1.5.3.1 Data Editor

This is a spreadsheet-like window which contains the data to be analyzed. The data editor has two views:

Data View which contains the data and it is the view we see when we open the data editor, figure (1.3). When we click the tab at the bottom of the window brings up the **Variable View** which does not contain data, but displays information about the dataset that is stored, figure (1.4). We can control how SPSS displays data from this window. Each data editor contains one dataset. Multiple data editors can be opened at one time, in which each one contains a separate dataset. Datasets that are currently open are called working datasets. All data manipulations, statistical functions, and other SPSS procedures operate on these datasets.

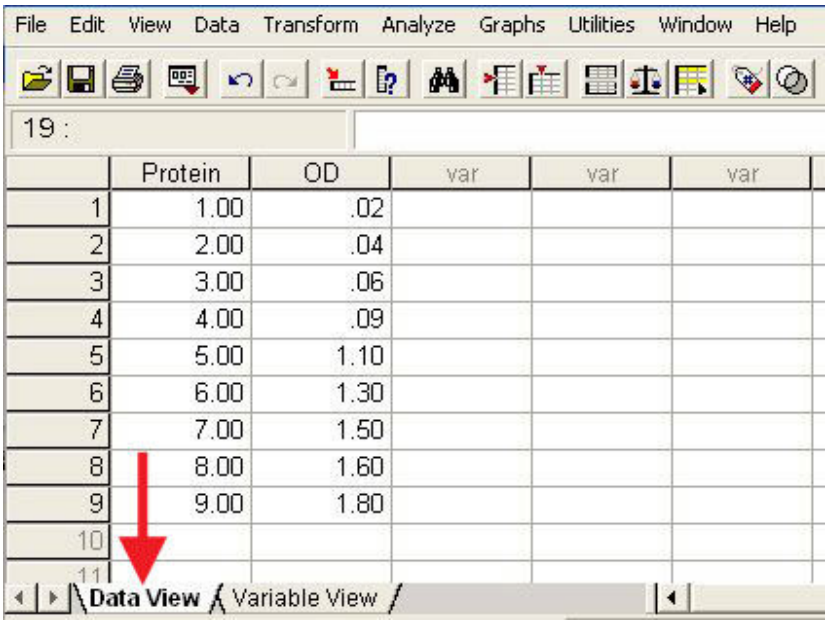


Figure (1.3): Data view window

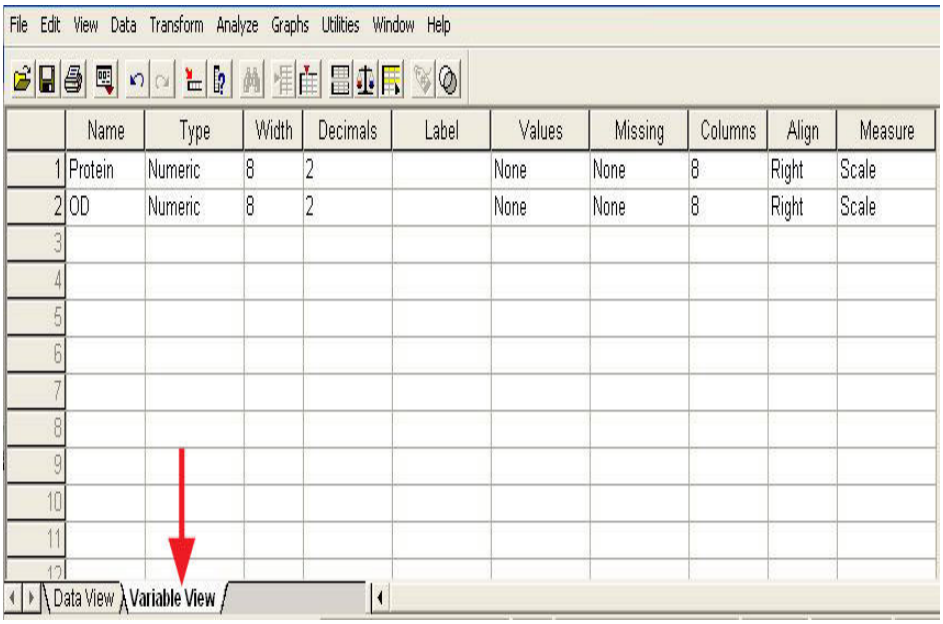


Figure (1.4): Variable view window

1.5.3.2 Output Viewer

This is where the results of any analysis appear. From the viewer, we can format the output in a wide range of ways, and export results in a variety of formats, e.g. text, SPSS, MSWord, MS Excel, etc.

In this study we will use SPSS software to perform MLR analysis.

1.5.4 MATLAB software

A high-level language and interactive environment program that enables us to perform computationally intensive tasks faster than with traditional programming languages such as C, C++, and Fortran. Wide range of applications can be performed by matlab, including image processing, test and measurements, control design, and computational biology.

MATLAB provides a range of numerical computation methods for analyzing data, developing algorithms, and creating models. The MATLAB language includes mathematical functions that support common engineering and science operations. Core math functions use processor-optimized libraries to provide fast execution of vector and matrix calculations.

Available methods include:

- Interpolation and regression
- Differentiation and integration
- Linear systems of equations
- Fourier analysis
- Eigenvalues and singular values
- Ordinary differential equations (ODEs)
- Sparse matrices

MATLAB add-on products provide functions in specialized areas such as statistics, optimization, signal analysis, and machine learning.

The use of matlab depends upon the script used and the input files. We used matlab in our study to do MLR cross validation by applying (leave one out (LOO) and leave many out (LMO) methods), to perform principal component analysis in order to divide the data before starting ANN, to perform the PC- ANN model and to perform PLS model.

1.6 Objective:

The main objective of this study is to develop QSAR models for the inhibition activity of 192 chemical compounds of vascular endothelial growth factor receptor-2 (VEGFR-2) by applying different statistical qualities such as MLR, PLS and PC-ANN. These models will be used to design new inhibitors.

Chapter two
Methodology

In this study we aimed to build QSAR model that can be used to calculate the activity of a chemical compound as anti VEGFR-2, using several statistical methods performed by several softwares (mentioned before). This is achieved by the steps shown in this chapter.

To build QSAR model you need to follow four successive steps as mentioned bellow:

- ✓ preparing the compounds list
- ✓ extracting descriptors from molecular structure
- ✓ choosing the informative descriptors
- ✓ modeling the Descriptors to Activity

2.1 Data preparation

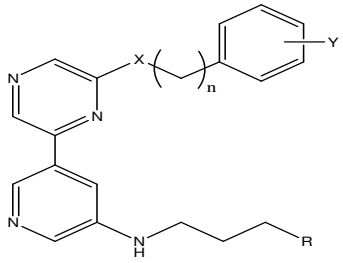
2.1.1 Compiling the compounds list

A data set of 192 vascular endothelial growth factor receptor-2 (VEGFR-2) tyrosine kinase inhibitors and their activity (pIC_{50}) were obtained from the literature (22 -31) and used in this study. The chemical structure and the biological activity of these compounds are summarized in table (2.1)

We aimed to select the compounds from a group of studies so we end up with a list of compounds with variety of cores. This list was used to have a comprehensive model that can be applied on many compounds with different cores.

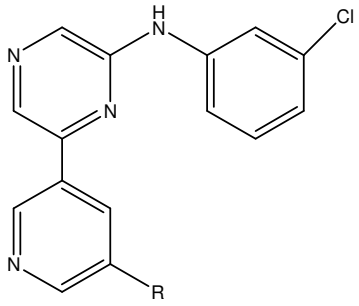
The activity in the list is expressed as pIC_{50} (- log (The half maximal inhibitory concentration)) which is indicator of how active the compound is.

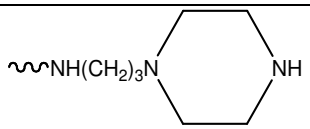
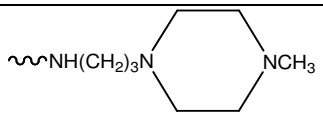
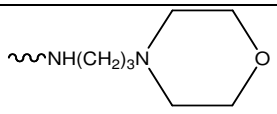
Table (2.1): The chemical structure and the biological activity of the compounds used in this study.

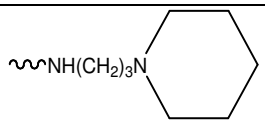
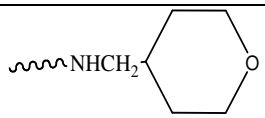
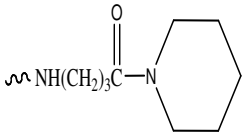


Compound number	Index*	n	X	Y	Substituent (R)	pIC_{50} (M) VEGFR-2
1	7	0	NH	3-Cl	OH	7.08
2	11	0	NH	3-F	OH	6.77
3	15	0	NH	3-OCH_3	OH	6.13
4	19	0	NH	2-Cl	OH	6.00
5	20	0	NH	3-Cl	4-pyridine	7.12
6	24	0	NH	4-Cl	4-pyridine	6.00
7	28	0	NH	3,4-Cl_2	4-pyridine	6.72
8	32	1	NH	H	OH	5.98

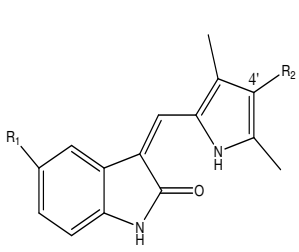
*ref 22



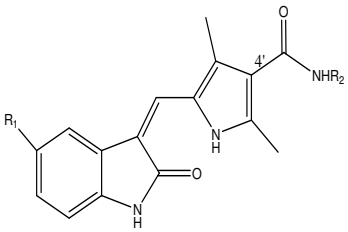
Compound number	Index*	Substituent (R)	pIC_{50} (M) VEGFR-2
9	37	NH_2	5.94
10	38	$NH(CH_2)_2OH$	6.56
11	39	$NH(CH_2)_4OH$	7.20
12	40	$NH(CH_2)_3N(CH_3)_2$	6.87
13	41	$NH(CH_2)_2N(CH_3)_2$	6.98
14	42	$CONH(CH_2)_2N(CH_3)_2$	5.37
15	43	$CONH(CH_2)_3OH$	5.62
16	44	$NHCOCH_2OH$	6.46
17	45	$NHCOCH_2OCH_3$	6.52
18	46	$NHCO(CH_2)_2OCH_3$	6.23
19	50		6.63
20	51		6.91
21	52		6.74

22	53		6.40
23	54		6.41
24	55	$NH(CH_2)_4$(4-pyridine)	6.56
25	56	$NH(CH_2)_3$(3-pyridine)	7.12
26	57	$NH(CH_2)_3$(1-pyrazole)	7.03
27	58	$NH(CH_2)_3$(1,2,4-triazole)	7.17
28	59	$NH(CH_2)_3Ph$	5.64
29	61	$NH(CH_2)_3CO_2H$	6.25
30	62		6.21

*ref 22



5

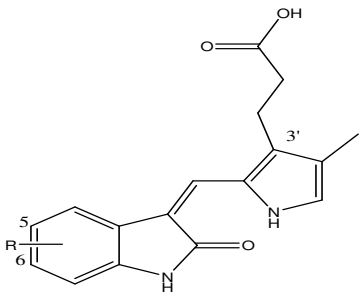


12

Compound number	Index*	Substituent (R_1)	Substituent (R_2)	pIC_{50} (M) VEGFR-2
31	5a	H	H	5.91
32	5b	H	$(CH_2)_2COOH$	5.62

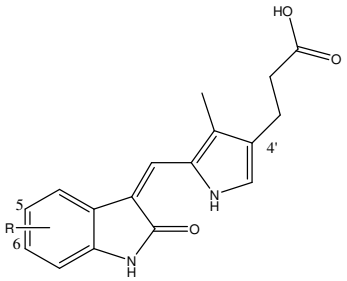
33	5c	H	$(CH_2)_3N(CH_2CH_2)_2NCH_3$	6.52
34	12a	H	$(CH_2)_2N(C_2H_5)_2$	7.30
35	12b	F	$(CH_2)_2N(C_2H_5)_2$	7.10
36	12c	Cl	$(CH_2)_2N(C_2H_5)_2$	7.57
37	12d	Br	$(CH_2)_2N(C_2H_5)_2$	7.49
38	12e	F	$(CH_2)_2N(CH_3)_2$	7.10
39	12f	F	$(CH_2)_2$-pyrrolidin-1-yl	7.22
40	12g	F	$CH_2CH(CH_2CH_2)_2N$-CH_3	7.60
41	12h	F	$(CH_2)_2$-morpholin-4-yl	7.05
42	12j	F	$(CH_2)_2$-triazol-1-yl	7.07

*ref 23



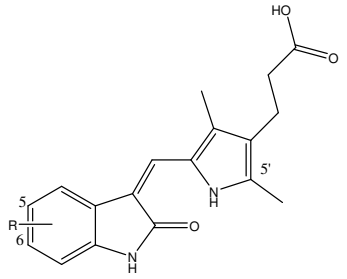
Compound number	Index*	Substituent (R)	pIC_{50} (M) VEGFR-2
43	1	H	7.70
44	9a	4-CH_3	6.70
45	9b	5-Br	6.46
46	9c	6-(3-OCH_3phenyl)	4.55
47	9d	6-(3-OC_2H_5phenyl)	6.00

*ref 24



Compound number	Index*	Substituent (R)	pIC_{50} (M) VEGFR-2
48	11a	H	5.67
49	11b	5-COOH	6.62
50	11c	5-SO_2NH_2	6.04
51	11d	6-OCH_3	5.87
52	11e	6-phenyl	6.52
53	11f	6-(3-OCH_3phenyl)	7.05
54	11g	6-(2-OCH_3phenyl)	5.83
55	11h	6-(4-OCH_3phenyl)	6.35

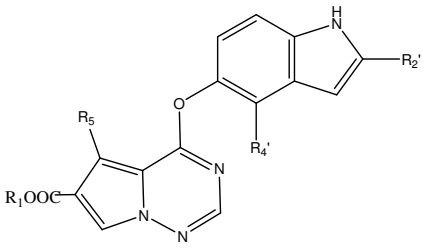
*ref 24



Compound number	Index*	Substituent (R)	pIC_{50} (M) VEGFR-2
56	16b	5-Br	5.76
57	16c	5-COOH	7.15

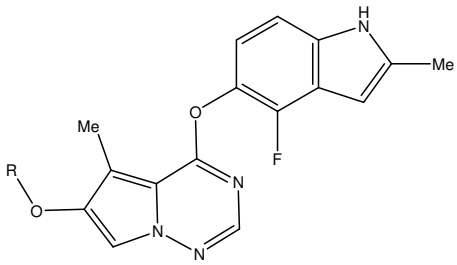
58	16d	5-SO_2NH_2	5.90
59	16e	6-OCH_3	5.08
60	16f	6-phenyl	6.85
61	16g	6-(3-OCH_3phenyl)	6.52
62	16h	6-(2-OCH_3phenyl)	5.36
63	16i	6-(4-OCH_3phenyl)	6.28

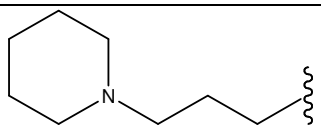
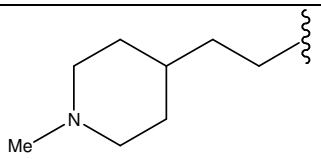
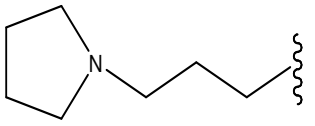
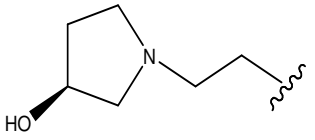
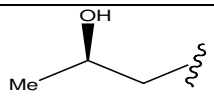
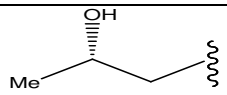
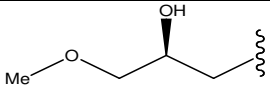
*ref 24

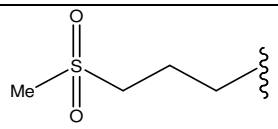
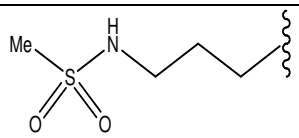


Compound number	Index*	R_1	$R_{2'}$	$R_{4'}$	R_5	pIC_{50} (M) VEGFR-2
64	2	Et	H	H	Me	7.06
65	3	Et	H	H	Et	6.51
66	4	Et	H	H	i-Pr	6.33
67	5	Me	Me	H	Me	7.11
68	6	Me	Me	F	Me	7.77

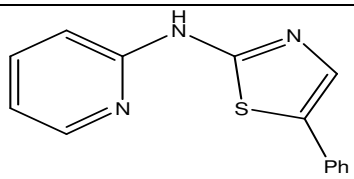
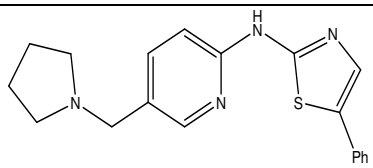
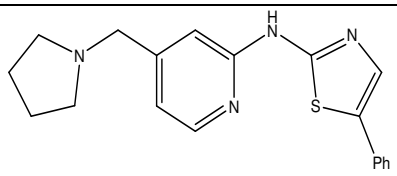
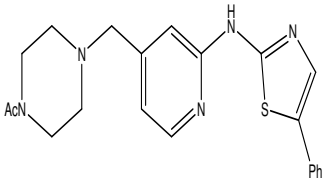
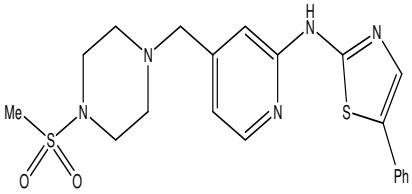
*ref 25



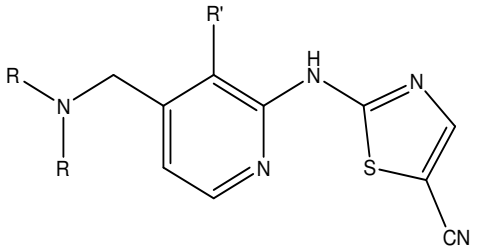
Compound number	Index*	Substituent (R)	pIC_{50} (M) VEGFR-2
69	7	H	7.62
70	8		7.77
71	9		7.70
72	10		7.70
73	11		7.62
74	12		7.60
75	13		7.40
76	14		7.38

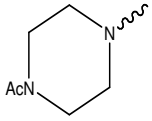
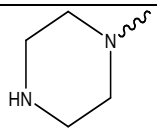
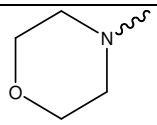
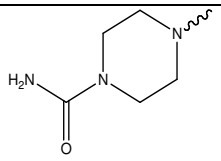
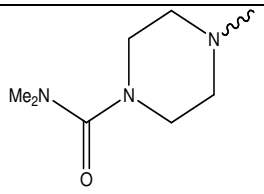
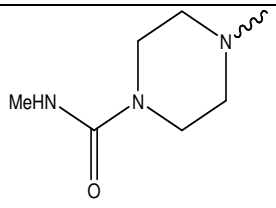
77	15		7.15
78	16		7.18

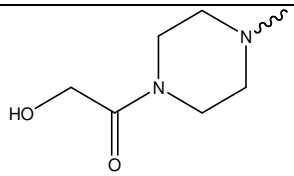
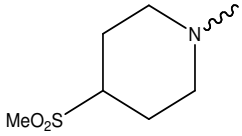
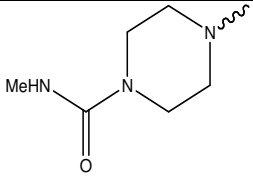
*ref 25

Compound number	Index*	Substituent (R)	pIC_{50} (M) VEGFR-2
79	1		8.15
80	2		8.22
81	12[c]		8.52
82	13[c]		8.10
83	14		8.70

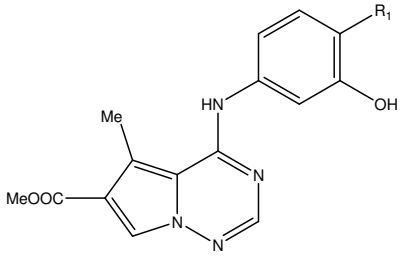
*ref 26



Compound number	Index*	R_2N-	R’	pIC_{50} (M) VEGFR-2
84	15		H	7.89
85	16		H	7.82
86	17		H	8.10
87	18		H	8.10
88	19		H	8.00
89	20		H	7.92

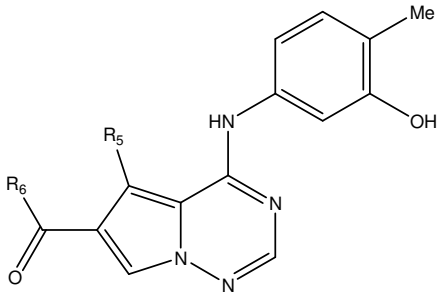
90	21		H	8.05
91	22		H	8.15
92	23		Me	7.92

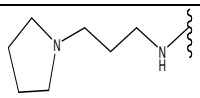
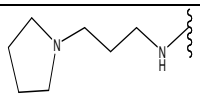
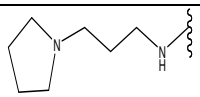
*ref 26



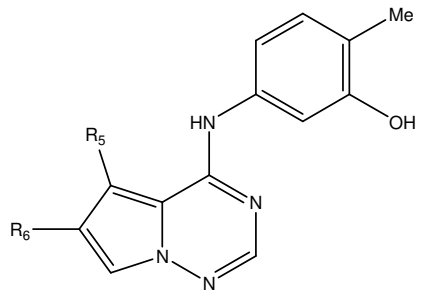
Compound number	Index*	R_1	pIC_{50} (M) VEGFR-2
93	14	H	7.21
94	15	Me	8.00
95	16	Et	7.26
96	17	n-Pr	6.72
97	18	i-Pr	6.42
98	19	CF_3	6.43
99	20	Br	7.48

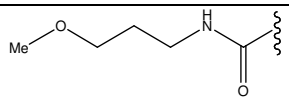
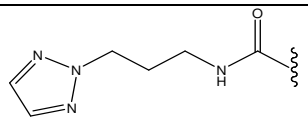
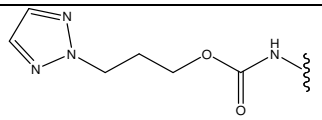
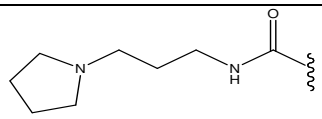
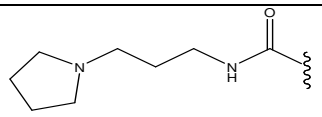
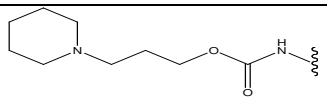
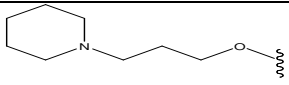
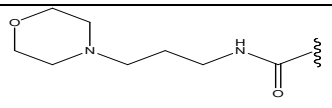
* ref 27

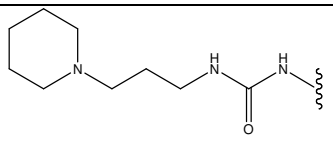


Compound number	Index*	R_5	R_6	pIC_{50} (M) VEGFR-2
100	21	n-Pr	-OEt	8.52
101	22	OEt	-OEt	7.19
102	23	i-Pr	-OEt	7.85
103	24	t-Bu	-OMe	6.51
104	25	Me		7.40
105	26	n-Pr		7.05
106	27	i-Pr		7.74

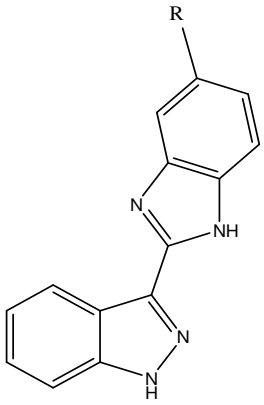
* ref 27

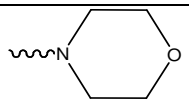
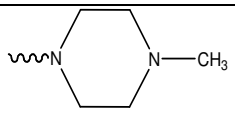
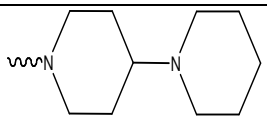
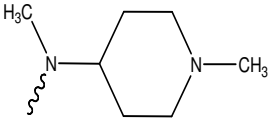
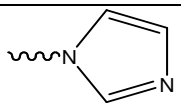


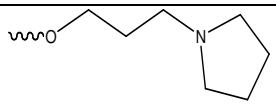
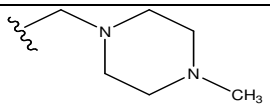
Compound number	Index*	R_5	R_6	pIC_{50} (M) VEGFR-2
107	28	Me		6.82
108	29	Me		7.52
109	30	i-Pr		8.15
110	25	Me		7.40
111	27	i-Pr		7.74
112	31	i-Pr		8.30
113	32	i-Pr		8.40
114	33	Me		7.28

115	34	i-Pr		8.30

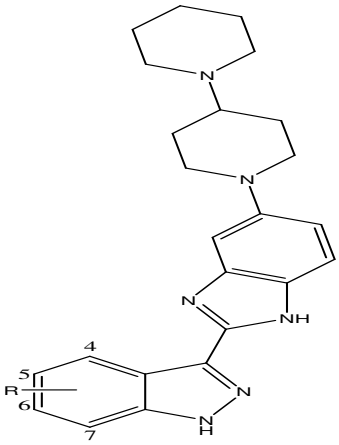
* ref 27

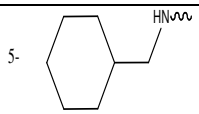


Compound number	Index*	R’	pIC_{50} (M) VEGFR-2
116	2	H	5.48
117	4		6.80
118	5		6.89
119	6		7.11
120	7		6.82
121	8		7.08

122	9		6.77
123	10		6.59

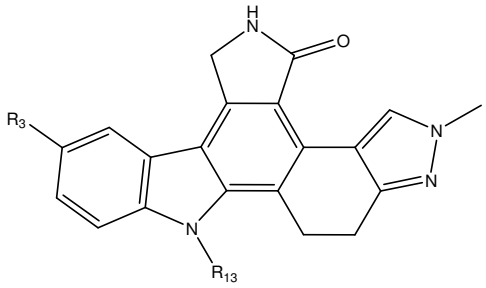
* ref 28



Compound number	Index*	Substituent (R)	pIC_{50} (M) VEGFR-2
124	11	4-OBn	6.18
125	12	4-NH(CO)NH*t*-Bu	7.15
126	13	5-OBn	7.52
127	14	5-OPh	7.36
128	15		7.36
129	16	5-NH(CO)NH*t*-Bu	8.30
130	17	5-CO_2Me	7.51

131	18	6-F	7.55
132	19	6-OBn	7.68
133	20	6-CF_3	7.28
134	21	7-F	6.48

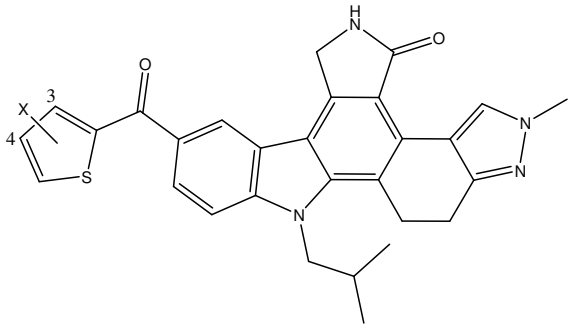
*ref 28



Compound number	Index*	R^3	R^{13}	pIC_{50} (M) VEGFR-2
135	5a	H	H	7.80
136	5b	H	Et	7.82
137	5c	H	nPr	7.80
138	5d	H	i-Bu	7.77
139	6a	Ac	H	7.46
140	6b	Ac	Me	8.00
141	6c	Ac	Et	7.89
142	6d	Ac	Pr	8.52
143	6e	Ac	Bu	7.59
144	6f	Ac	i-Pr	7.12
145	6g	Ac	i-Bu	7.68
146	7a	2-Thiophene-CO	H	8.70

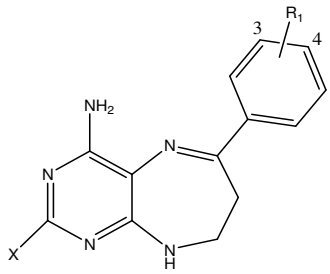
147	7b	2-Thiophene-CO	Et	9.00
148	7c	2-Thiophene-CO	Pr	8.30
149	7d	2-Thiophene-CO	i-Pr	7.82
150	7e	2-Thiophene-CO	i-Bu	8.22
151	8	3- Thiophene-CO	i-Bu	7.89
152	9	2-Furan-CO	i-Bu	8.40
153	10	3-Furan-CO	i-Bu	8.10

* ref 29



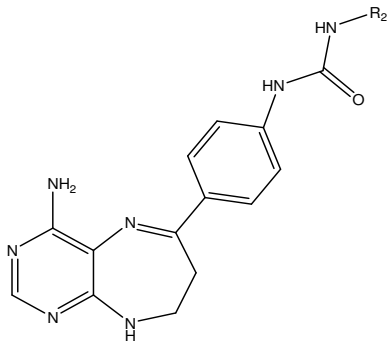
Compound number	Index*	X	pIC_{50} (M) VEGFR-2
154	7f	3-Cl	8.15
155	7g	3-Br	7.6
156	7h	3-Me	8.00
157	7i	4-Me	7.96

* ref 29



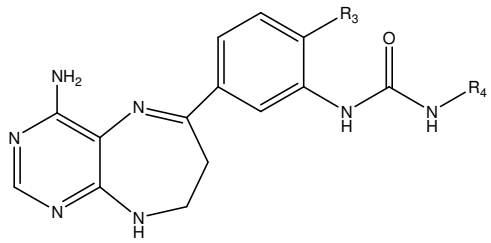
Compound number	Index*	X	R^1	pIC_{50} (M) VEGFR-2
158	1a	H	H	5.52
159	1b	H	3-Me	5.52
160	1c	H	4-Me	5.40
161	1d	H	3-Cl	6.22
162	1e	H	4-Cl	5.52
163	2a	NH_2	H	4.46
164	2b	NH_2	3-Cl	4.68
165	2c	NH_2	4-Cl	5.05

* ref 30



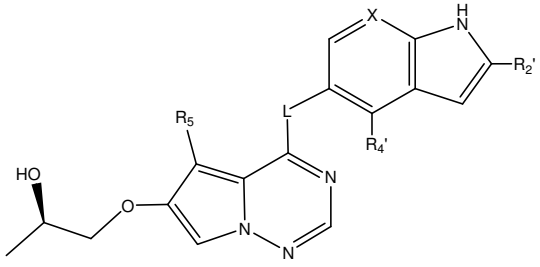
Compound number	Index*	R^2	pIC_{50} (M) VEGFR-2
166	1f	Ph	6.02
167	1g	3-CF_3-Ph	8.52
168	1h	4-CF_3-Ph	6.96
169	1i	3-Cl-Ph	7.72
170	1j	2-F-3-CF_3-Ph	7.31
171	1k	4-F-3-CF_3-Ph	8.30
172	1l	4-Cl-3-CF_3-Ph	8.22
173	1m	2-F-5-Me-Ph	7.19
174	1n	2-F-5-CF_3-Ph	8.40

* ref 30



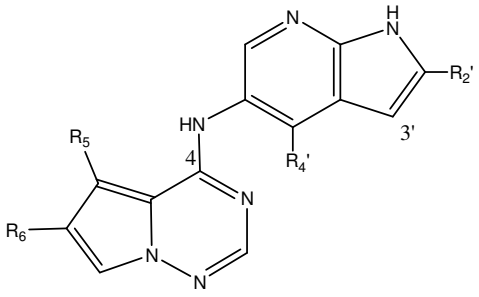
Compound number	Index*	R^3	R^4	pIC_{50} (M) VEGFR-2
175	1o	H	Ph	6.05
176	1p	H	3-CF_3-Ph	6.90
177	14	Me	3-CF_3-Ph	8.22

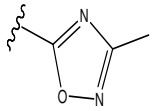
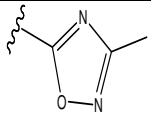
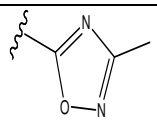
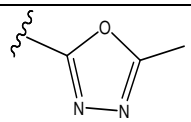
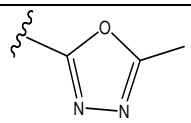
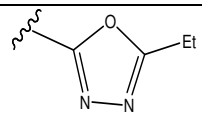
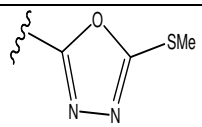
* ref 30

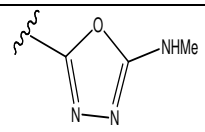


Compound number	Index*	X	L	$R^{2'}$	$R^{4'}$	R^5	pIC_{50} (M) VEGFR-2
178	1	CH	O	Me	F	Me	7.60
179	2	N	O	Me	F	Me	7.72
180	4	N	O	Me	H	Me	6.91
181	5	N	NH	Me	H	Me	6.65
182	6	N	NH	H	H	i-Pr	6.78

*ref 31



Compound number	Index*					pIC_{50} (M) VEGFR-2
183	7	COOMe	Me	H	H	6.92
184	8	COOMe	*i*-Pr	H	H	7.77
185	9		*i*-Pr	H	F	7.52
186	10		*i*-Pr	H	H	7.02
187	11		*i*-Pr	Me	H	6.95
188	12		*i*-Pr	H	H	7.51
189	13		*i*-Pr	Me	H	7.60
190	15		*i*-Pr	Me	H	7.44
191	16		*i*-Pr	Me	H	6.82

192	17		*i*-Pr	Me	H	7.60

*ref 31

2.1.2 Structure drawing and optimization:

The Chemical structures of the whole 192 compounds were taken from the references (22-31). And then they were drawn using hyperchem software.

To perform geometry optimization for the chemical structure you have to follow the steps below:

1. First draw the structure using drawing tools. After drawing we choose "add H &model building" from build menu to have the 3D structure.
2. Then Click on "start log" in the file menu to give it a name, and choose a directory to save it.
3. From the setup menu choose "semi-empirical" method of calculation and then select "AM1" from the semi-empirical window as shown in figure (2.1).

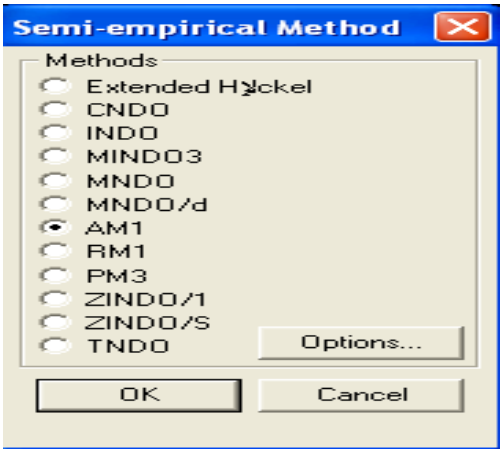


Figure (2.1): semi-empirical method window.

4. Click on the options button of the semi-empirical window and select geometry optimization parameters, choose total charge= 0, spin multiplicity= 1, spin pairing = RHF, convergence limit=0.1, and accelerate convergence= yes as shown in figure (2.2).

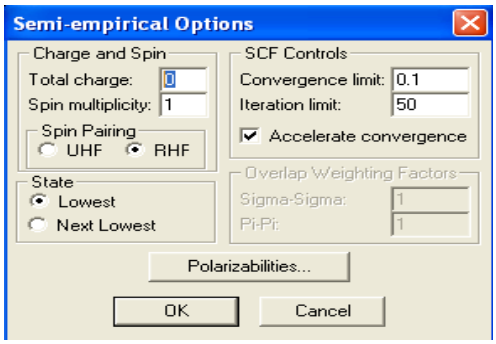


Figure (2.2): Semi-empirical options window.

5. Click OK to close the semi-empirical options dialog box, and then click OK to close the semi-empirical method dialog box.

6. Choose "geometry optimization" from compute menu, this opens semi-empirical optimization dialog box as shown in figure (2.3).

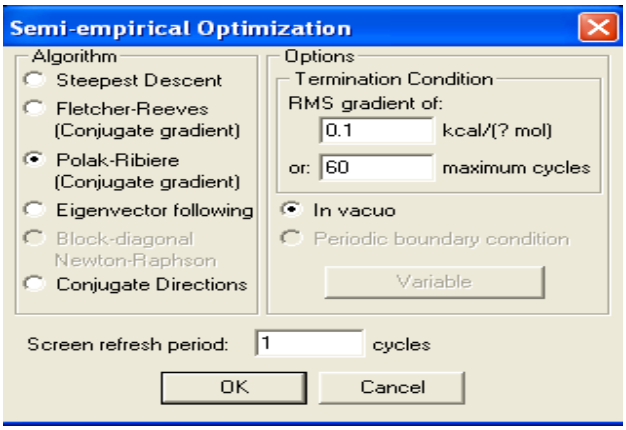


Figure (2.3): Semi-empirical optimization window

7. Select Polak-Ribiere as algorithm method, choose 0.01 for RMS gradient condition, and the default values for the other variables, then click OK to initiate the optimization and close the dialog box. The software will choose the maximum cycles, but we can increase, if needed.

8. Finally, when the program finish the optimization, select "stop log" from the file menu to save the calculation output as log file. And then save the structure as HIN file.

By finishing these steps we had 192 HIN files that represent the optimized chemical structures of the compounds, which is the input of Dragon software. And 192 log files that represent the calculations output.

2.2 Extracting descriptors from molecular structure

The chemical structure can't be linked directly with the activity, so theoretical descriptors are the way of how the activity is linked with the chemical structure of the compounds. Nineteen groups of descriptors were calculated directly and indirectly using Hyperchem Dragon softwares.

2.2.1 Descriptors calculated by Hyperchem

First, from the output file of hypechem Calculation, we can get some descriptors.

We got highest occupied molecular orbital energy (E_{HOMO}), lowest unoccupied molecular orbital energy (E_{LUMO}), molecular Dipole moment (DM), and Heat of formation.

Second, using hyperchem we can calculate some other descriptors.

We computed Surface Area (Approx), Surface Area (Grid), Volume, Mass, Hydration Energy, Octanol- water partition coefficient (Log P), Refractivity, Polarizability.

We calculated them according to the steps below:

1. Open the Hyperchem file of the chemical structure of the compound (HIN file)
2. Open compute menu, select "QSAR properties" , this will open widow as in figure (2.4)

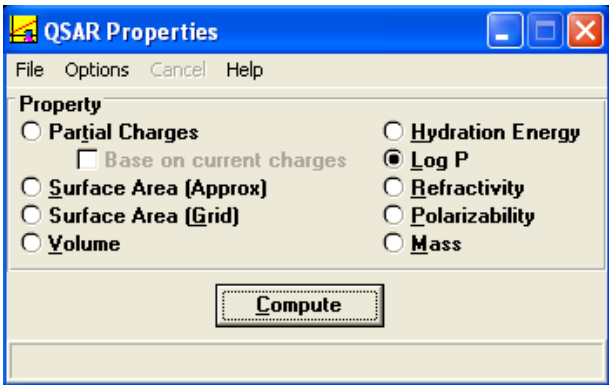


Figure (2.4): QSAR properties window.

3. From options menu, select output to "result window".
4. Then click on the property, then compute button for each property.
5. Finally save the result window as "log file"

2.2.2 Descriptors calculated manually

A group of four quantum descriptors was calculated manually by using excel software according to the equations below:

Electronegativity (χ) = - 0.5 (E_{HOMO} + E_{LUMO})

Hardness (η) = 0.5 (E_{LUMO} - E_{HOMO})

Softness ((S=1\η).

Electrophilicity (ω) = χ^2/2 η

We gathered the whole quantum descriptors in one excel file ready to be used in the further steps.

2.2.3 Descriptors calculated by Dragon software:

By Hyperchem we calculated one group of descriptors which is quantum descriptors. The other eighteen groups (shown in figure 2.5) was calculated using Dragon software. To calculate the descriptors using dragon:

1. Prepare folder that has the chemical structure (HIN files from hyperchem) of all the compounds you want to use in the study (192 in our case).
2. Open dragon software, the program will open a window as in figure (2.5).

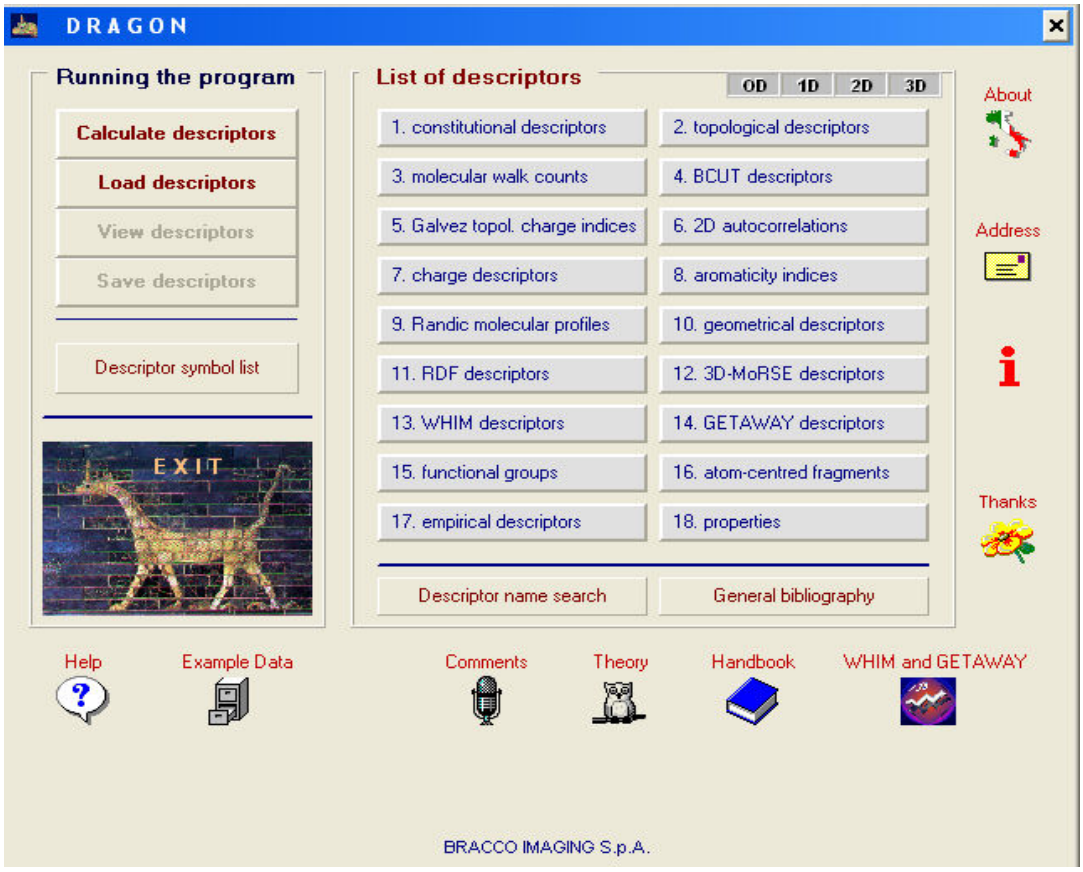


Figure (2.5): Dragon software window.

3. Click on calculate descriptors.
4. Then open the compounds folder, open the files type (HIN files), and select all the compounds to inter the calculation.
5. Click on descriptor selection button
6. Now open the descriptors group/ groups “we calculated each descriptors groups for all the compounds alone”.
7. Select the calculation terms; click on “stop calculation in error”
8. Press Run button to start calculation.

9. Finally, after the software calculation finished name and save the output.

At the end of this work, we had separate file for each group of descriptors ready to be used in the next step.

2.3 Choosing the informative descriptors:

The descriptors which calculated by dragon and hyperchem software were 1497 descriptors. In this step, the subgroup of the descriptors that can provide the best model that can predict the activity is chosen.

SPSS software was used to perform this step. By using each group of descriptors (dragon and Hyperchem descriptors) to built one MLR model. And then collect the descriptors chosen by each model of the MLR models in one excel file (final MLR file). And perform MLR another time for the final file to choose the best subgroup of the descriptors.

To perform MLR analysis using SPSS software you need to follow the following steps:

1. Prepare the files to be used as SPSS input; excel files that has the experimental activity (dependent variable) as the first column, and the descriptors (the independent variables) as the rest of the columns.
2. Open the file containing the dependent variable and independent variables using SPSS, then go to analyze menu and choose regression and select linear as in figure (2.6).

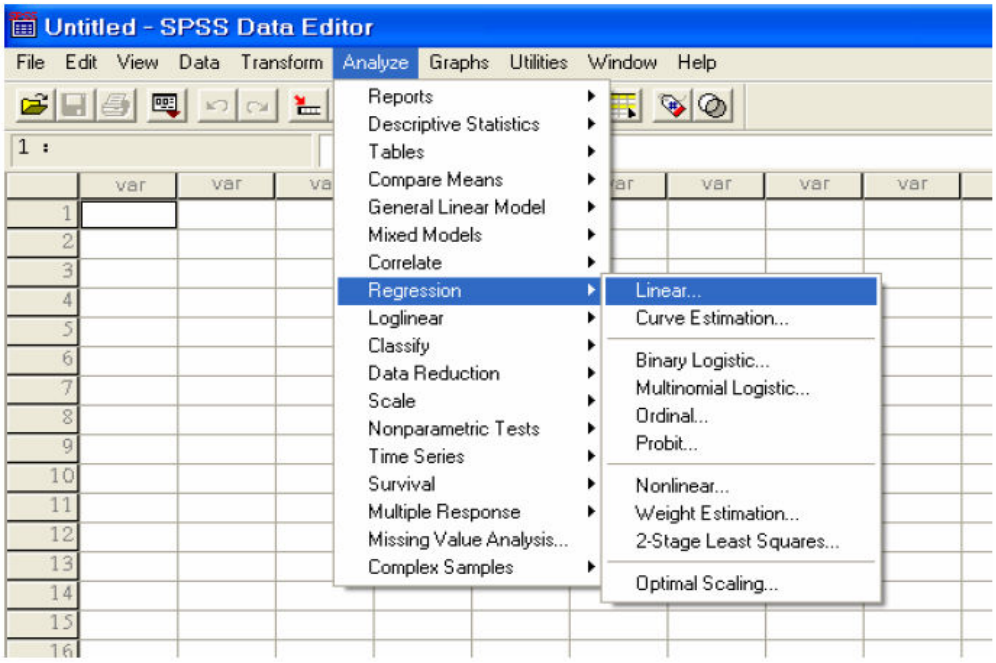


Figure (2.6): SPSS Data Editor Menu

3. Set activity as the dependent variable and set the descriptors in the input file as the independent variables in the linear regression dialog box, and then press on the options button of the same dialog box as shown in figure (2.7).

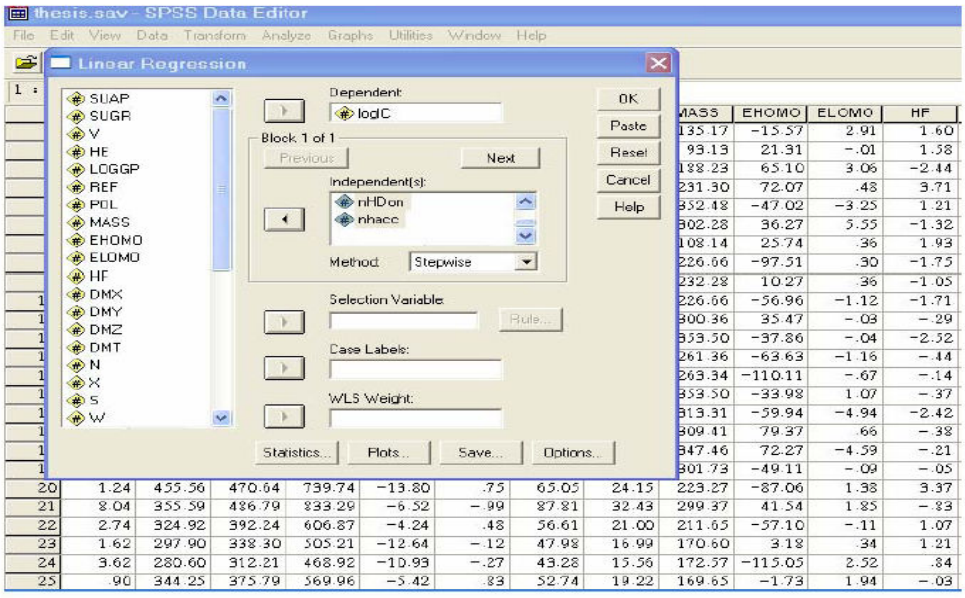


Figure (2.7): Linear regression box.

4. Select use F value and set F Entry and F Removal values and leave other parameters without any change in the linear regression option dialog box.
5. Choose the method to be stepwise, click save to store results back to the input sheet, choose the predicted values to be unstandarized and then click continue.
6.Click statistics to generate additional statistics for variables then click continue, and finally click OK in the linear regression dialog box.
By finishing these steps, we will have linear model for each group of descriptors, which contain the best informative descriptors from each group. So these steps must be performed again on a file that has the whole descriptors chosen by the MLR models of each group of descriptors (final MLR) to result with final model that has the best informative descriptors from the whole 1497 descriptors.
We aimed in this part of the work to extract the informative descriptors, at the same time we built the linear model that link the activity with the descriptors.

2.4 Modeling the Descriptors to Activity

2.4.1 MLR validation:

At the end of the previous step, part of the statistical analysis work will be finished. The work end with several final MLR best models that relate the activity with the descriptors linearly, so those models need to be validated.
Cross validation is a technique taken to evaluate the ability of the QSAR model to predict the activity of chemical compound that not used in building it. (32)
R^2 (coefficient of determination) which is a statistical parameter that can indicate the goodness of the model. R^2 is a measurement of how the regression line perfectly fits the real data points. If the regression is perfect R^2 is 1, but if the regression is a total failure then R^2 is zero.
R^2 does not indicate the quality of the model, so focusing on maximizing the value of R^2 alone is not a good idea. We need other parameters to evaluate the ability of the model.
R^2 is calculated by SPSS program while building the model.

2.4.1.1Cross validation by using MATLAB

Leave One Out Cross Validation

Leave-one-out cross validation done by using a single observation from the original data set as the validation data, and the remaining observations as the training data. This is repeated as the number of data set, such that each observation in the sample is used once as the validation data. This type of cross-validation is usually expensive from a computational point of view because of the large number of times the training process is repeated (33).

This type of validation is performed by the following steps:

1- Prepare the input file by copying the observed and predicted activity columns from the SPSS data editor and paste them in an excel file and save it. The observed activity should be the first column and then comes the predicted activities.

2- Copy excel file to Matlab working directory (C:\MATLAB701\work) or any directory you are working in. In the same directory, there should be a file (script) with the name (cross_val_param_loop.m) which will perform the validation.

3- Open the script file, then you will have a message on Matlab window says "what is the file name", enter the excel file name with or without the (.xls) extension. Then you will have a message says "model number", where the next line contains number that is the model number.

4- Matlab script will ask you for the number of descriptors for each individual model. Towards the end, cross validation results for all models will be saved in a file called "CV_LOO.dat" on the directory (C:\MATLAB701\work) or the directory you are working in.

Leave Many Out Cross Validation

In leave many out cross-validation, the original sample is randomly partitioned into X subsamples. Of the X subsamples, a single subsample is considered as the validation data for testing the model, and the remaining (X – 1) subsamples are used as training data. The cross-validation process is then repeated X times, using each of the X subsamples once as the validation data (33).

This type of cross validation is done by the following steps:

1- Prepare an excel file that contains the activity (first column) and the descriptors entered in the regression model of interest.
2- Then run Matlab script "lgocv.m". This script performs leave-group-out cross validation where 20% of the data are classified as test set so that each compound is entered only once in the test set.
3- Enter excel filename and number of compounds to be used in the training set when you are asked for these information and press enter.
4- The output file: "CV_LGO.dat", appears in the same directory in addition to printing cross validation parameters on the screen.

2.4.2 Principal component-Artificial Neural Networks (PC-ANN)

We consider the relation between the activity and the chemical descriptors more complicated than a linear relationship so we perform PC-ANN analysis, hoping to end up with better models than MLR models.

After using SPSS software to perform MLR analysis, we finished up with several good MLR models; those models will be the input to PC- ANN step. The choice of the best models depends on their validation parameters.

After choosing which MLR models are the best, you need to perform the following steps:

First you must perform Principal component analysis (PCA):

Before starting ANN analysis, you should divide the data into training, validation and external test set. We use PCA to perform this step, because the division should not be done randomly. Data division should be done as to have 60% of the data in the training set and 20% for each of the validation and test sets.

1- Prepare an excel file containing the experimental activity as the first column, and the descriptors of the whole chosen models as the rest of the columns.
2- Use Matlab script "calcpcaplot.m" to perform the analysis (the script file and the file you work on should be in the same directory). Open the script file, by going to Matlab menu and click on the open file icon, and then run the script.

3- After running the script you will be asked about the file name, Enter the name of the file that you prepared.

4- Then you will obtain a figure with a scatter distribution of the data (compounds), each compound indicated by a point when you press it, the compound number appears. Select the training, validation and test sets molecules from these data points so they span the same space of the entire data.

Second performing ANN model:

1- Prepare an excel file for each model of the chosen MLR models. Each file should contain the experimental activity as the first column and the descriptors used in the model as the other columns.

2- To implement the data division on ANN analysis, you should edit the matlab script "ann_ext_test_4loop.m" to indicate the training, validation, and test sets in the calculation.

3- Then, you must open and run the matlab script "nnloop.m" which read the previous script and perform the analysis to end up with the model.

You may also need to modify the R (regression coefficient) value in the script "nnloop.m" to make it stop, because it will still working until it reach the value in the script (we chose it to be more than 0.75 , 0.8 according to test set and training set respectively). When you run the "nnloop.m" script, you will be asked for the excel file name for the model of interest, model number (to be inserted) and number of hidden nodes. To run the "ann_ext_test_4loop.m" file, a Matlab script named "subplotspace.m" should be in the same directory.

When the work is done, the cross validation results will be printed out on Matlab screen at the end of optimization and saved in a file with the name "CV_model_"N"_hn"H".dat", where "N" is the model number and "H" is the number of hidden nodes.

The steps are done for each one of the chosen models and cross validation results for each model will be compared with each other in order to choose the best one.

The script "ann_ext_test_4loop.m" produces three regression figures, one for each data set. These files are named as "'mlr_"test"_model_"N"_hn"H".fig", where N and H are the same as in the cross validation file and "test" is replaced with "train" and "validation" for the training and validation sets, respectively.

The residue, the difference between predicted and observed activities for each data is saved to a file named as “pred_obs_”test”_model_”N”_hn”H”.dat” while for the complete data without division, the file is named “ “pred_obs_all_model_”N”_hn”H”.dat”. Residue figures for each data set are named “residue_”test”_model_”N”_hnee”H”.fig”

4- After choosing the optimal models, you have to optimize the number of hidden nodes for these models. To do so, you have to choose a range for hidden nodes numbers (in our case 3-20) and optimize the network for each number of hidden nodes. The optimal model choice is based on cross-validation results.

After these steps you will end up with several PC_ANN models, you choose the best of them according to the cross validation parameters.

Third: Randomization

Chance correlation

For further validation, run chance correlation test for the optimal models. In this test, the activity column is being randomized and the network performance is checked. To perform this test, run Matlab script “nn_chance_corr_new.m”. When you are asked for; enter the excel file name for the model of interest, model number, number of hidden nodes and trial number for chance correlation test. The output is similar to that original model.

Then the cross validation parameters values of the original models compared with the ones of the chance correlation models. They shouldn’t be the same to prove that our work doesn’t produced by chance.

Y- randomization:

And to ensure the robustness of the optimal models we applied Y- randomization. The Y-randomisation technique proceeds with scrambling of the Y-column data, keeping the descriptor matrix unchanged. The models were built using the scrambled data (by the same steps as with ANN) and the values of correlation coefficients were calculated. If the correlation coefficients are lower than the results of the original models. This indicates that the developed models considered being robust enough. And the value of cR^2p is calculated. These

values should be above 0.5 to consider the model acceptable and it is calculated by the equation below. (34)

$$cR^2p = R\sqrt{(R^2 - R^2r)}$$

Where:

R: correlation coefficient.

R^2: coefficient of determination.

r: for randomization.

2.4.3 Partial Least Squares (PLS)

In this part of the work we aimed to build a linear relation between the activity and the descriptors using another linear method other than MLR.

In this method the descriptors which have correlation will be gathered in one latent variable which indicates them and that's to avoid the intercorrelation that may be happen in MLR. PLS is performed using Matlab software.

PLS is done by the following steps:

1.In the previous work (PC-ANN) the data set (188 compounds) was divided into three sets, training, validation and test sets. In this part of work we need training and test sets only, so we combine the training and the validation sets together to form the new training set (80% of the data set) for the PLS work. While test set used as test set (20% of the data set) for the PLS work.

2.After dividing the data set, prepare four notepad files for each model of the chosen MLR models, as following:

Xcal: that has the columns of descriptors (independent variables) of training (cal) set.

Ycal: that has the column of activity (dependent variable) of training set.

Xtest: that has the columns of descriptors (independent variables) of test set.

Ytest: that has the column of activity (dependent variable) of test set.

3.Open the Matlab software in the pathway in which the script needed is present, which is "PLS.M".

4.Type the following commands in the Matlab sheet:

>> load xcal.txt

>> load ycal.txt

>> load ytest.txt

>> load xtest.txt

This commands to load the files that have our data.

Then type the following commands in Matlab sheet then click inter.

```
>> [p, q, w, b, t, u, x, y, l] = pls(xcal', ycal', 10);
>> plsprs = plspress(xcal', ycal', p, q, w, b, 10);
```

Then type the command : >> plot(plsprs,'*') then click inter.
Those commands typed to run the needed script.

5.At this time a plot is appeared. You should choose the point in which the curve remains constant after it, which define the number of Latent variable.

6.Type

>> [c, x] = plspred(xtest', p, q, w, b, 2);

Before clicking enter, You should change the number 2 in the command with the point obtained from the step 5 and then click enter in the Matlab program.

Choosing this point is dependent on the respective curve and can vary in diverse datasets.

This command is used to bring out the predicted activity values of the test set.

7.Type the command : >> predtest=c'; then click enter

From the workspace menu you can find the predicted results in the predtest.

8.Type

>> [c, x] = plspred(xcal', p, q, w, b, 2);

Before clicking enter, You should change the number 2 in the command with the point obtained from the step 5 and then click enter in the Matlab program.

Choosing this point is dependent on the respective curve and can vary in diverse datasets.

This command is used to bring out the predicted activity values of the training set.

9.Type the command : >> predcal=c'; then click enter

From the workspace menu you can find the predicted results in the training set.

10. The cross validation of this part is done using Leave one out way as clarified before.

Chapter Three

Results and Discussion

To improve the studies that concern about tumor growth, we developed MLR- QSAR model that relates the activity of 192 anti vascular endothelial growth factor receptor-2 (VEGFR-2) to their structures using their theoretical descriptors as structure indicators. It is evident that direct inhibition of the kinase activity of VEGFR-2 will result in the reduction of angiogenesis and the suppression of tumor growth. The work was done by following successive steps using several softwares to build the linear and non linear models and perform their validations.

The methodology was explained in the previous chapter and the results are discussed in this chapter.

Descriptors calculation

We take the structures and experimental activity values of our 192 compounds from the literature (22-31). The compounds were gathered together to deal with them as one set. The compounds structures and their activities as pIC_{50} are summarized in table (2.1). Hyperchem software was used to build the structure of each one of the compounds and AM1 semi-empirical method was used to optimize the chemical structures. Then the same software was used to calculate quantum chemical descriptors group directly and indirectly and they all gathered in one excel file, and this is the first group of 19 descriptors groups.

Dragon software was used to calculate the other eighteen groups of descriptors. We discarded the constant or near constant descriptors because they cannot differentiate between the different compounds. Dragon discarded the whole descriptors of three groups which are empirical, properties and aromaticity indices descriptors because the previous reason. Our output were fifteen files each one has the results of one group of the descriptors which are constitutional, topological, molecular walk counts, BUCT, galvez topological & charge indices, charge descriptors, 2D autocorrelations, randic molecular profiles, geometrical, RDF, 3D-MoRSE, WHIM, getaway, functional groups and atom-centred fragments descriptors.

At the end we had sixteen groups of descriptors; we gathered the small groups of descriptors together. We gathered five groups of descriptors; charge descriptors, galvez topological &

charge indices, molecular walk counts, randic molecular profiles and quantum descriptors in one excel file to end up with twelve groups.

This part of work end up with twelve excel files, ready to be used in the next step.

MLR

The twelve files of calculated theoretical descriptors were used to perform MLR analysis using SPSS software by stepwise regression method. MLR analysis were performed on the twelve descriptors groups individually rather than dealing with them as one group according to the work done by Deeb (35). This method constructs the model in stages. It starts by trying out one independent variable at a time and including it in the regression model if it is statistically significant, and eliminating those that are not statistically significant.

The results of performing MLR regression on the twelve groups of descriptors are summarized in table (3.1), with each group we tried to reach a model with the highest correlation coefficient (R) and in the same time the lowest number of descriptors. That is because we wanted to have only the descriptors that have a direct relation with the activity and we wanted to reduce the inter correlation between descriptors in the model as much as we can.

Table (3.1): The descriptors and correlation coefficient values of MLR models for each group of descriptors.

No *	Group name	R*	R^2 *	R^2 adj.*	Selected descriptors.*
1	2D descriptors	0.86	0.74	0.70	MATS1e,MATS7e,GATS8p,ATS8e, GATS8v,MATS1v,MATS1p,MATS6m, ATS6v,GATS6m,MATS7p,ATS2v, GATS1m,ATS3p,GATS6v,GATS5v, ATS2m,MATS4v,ATS7v,ATS6e, MATS4e,GATS4e,ATS4m,MATS6p, ATS6p,ATS2e,MATS7v,MATS4m

2	3D morse	0.83	0.70	0.65	Mor27u,Mor05u,Mor05m,Mor20v, Mor04v,Mor10u,Mor31v,Mor31m, Mor02u,Mor03m,Mor06v,Mor08u, More28u,Mor21m,Mor25v,Mor29u, Mor29m,Mor23u,Mor25u,Mor19v, Mor15u,Mor25m,Mor31u,Mor17u, Mor04u
3	Atom-centred	0.83	0.68	0.65	C-030, N-069, C-041, C-016, C-025, C-006, C-013, F-084, O-059, C-032, C-005, N-068, N-075, C-026, C-028, H-050, O-060, O-057, H-049
4	BUCT desc.	0.77	0.60	0.52	BELp8, BELm3, BEHp7, BEHv2, BEHe2, BELm6, BELe3, BELe1, BELm2, BELe8, BELm7, BELe7, BELm4, BEHm8, BELm1, BELv6, BELm8, BELm5, BEHv5, BEHm5, BEHm4, BEHp3, BEHe3, BEHm2, BEHm6, BEHv6, BEHm3, BEHm1, BELe2, BELp4, BEHe7, BEHm7
5	Constitutional	0.79	0.62	0.57	nS, nCIR, RBF, nF, nR07, Me, nBM, Mp, nO, nR09, nCIC, nH, nBnz, Mv, Ms, Ss, Sp, nR05, nTB, AMW
6	Function desc.	0.81	0.65	0.62	nCOOH, nCOPh, nHAcc, nNHRPh, nNH2, nNHR, nROR, nCN, nCONN, nCO, nCONHR, nSO2, nRCX3, nCq, nCrH2, nCaH, nCs, nCp
7	Geometrical	0.75	0.55	0.48	G1, G(O..O), G(N..O), G(N..S), G(O..Cl), SPH, MAXDP, G(S..F), L/Bw, RGyr, FDI,

					ADDD, SPAM, PJI3, G(N..F), SPAN, G(O..Br), MAXDN, G(S..S), J3D, W3D, MEcc, G(N..N), TIE, G(N..Br), ASP
8	Getaway	0.80	0.64	0.61	R6v+, HATS1u, R4e+, H6m, R1p+, R3p+, HGM, R5m, HATS5v, HATS6m, H6u, R8v, R2p, R8v+, R8p+, R1v+
9	RDF desc.	0.78	0.60	6	RDF025m, RDF010m, RDF050v, RDF075m, RDF050u, RDF140v, RDF145u, RDF030m, RDF050m, RDF020p, RDF080u, RDF040m, RDF090m, RDF070v, RDF080v, RDF110m, RDF135m, RDF020u, RDF075v,
10	Topological	0.83	0.68	0.65	AAC, SEigZ, IC1, T(O..Cl), Xt, PJI2, D/Dr10, Jhetv, SPI, T(O..F), X4Av, Yindex, ZM2V, VEA1, ww, BIC4, T(N..N), IVDM
11	Whim desc.	0.72	0.52	0.43	G2s, Dm, E1u, Ds, G1s, G2m, E1m, G2u, As, L2p, Ap, G1e, Du, G2e, G1p, G1u, G2p, G1m, E3v, E3u, E2m, P1p, E2v, P2s, E2s, L3s, Ts, L2s, Tv, E3s, L3v,
12	Charge extra**	0.80	0.64	0.60	RPCG, JGI2, GGI4, qneg,GGI2, Hydration, Refractivity, Qneg, DP01, GGI5, PCWTe, GGI10, SHP2, dipole, RNCG, SRW05, SP04, LUMO, JGI10, GGI8, Heat of formation, JGI5, GGI6

* No. refers to group number, R refers to correlation coefficient, R^2 refers to coefficient of determination, $R^2_{adj:}$ refers to adjusted R^2, Selected descriptors refer to descriptors chosen by the last MLR model.

**Charge extra refers to: galvez topological & charge indices, molecular walk counts, randic molecular profiles and quantum chemical descriptors.

After having those MLR models with the best descriptors in each group we gathered those descriptors (as the independent variables) in one SPSS data file with the activity (p IC_{50}) as the dependent variable. And then final MLR model was performed using this file. The results of this step are summarized in the table (3.2). The table shows the models that had R^2 over than 0.6 because with lower R^2 the model not accepted. (35)

Table (3.2): The best models of the final MLR: "models have R^2 over than 0.6"

Model No.	No. of descriptors	R	R^2	R^2 adj.	Selected desc.
6	6	0.78	0.61	0.60	BELp8, nNHRPh,ATS4m,nCOOH, C-032, Mor27u
7	7	0.79	0.63	0.62	BELp8, nNHRPh,ATS4m,nCOOH, C-032, Mor27u, C-005
8	8	0.80	0.64	0.63	BELp8, nNHRPh,ATS4m,nCOOH, C-032, Mor27u, C-005, MATS7e
9	7	0.80	0.64	0.63	BELp8, nNHRPh,ATS4m, C-032, Mor27u, C-005, MATS7e
10	8	0.81	0.65	0.64	BELp8, nNHRPh,ATS4m, C-032, Mor27u, C-005, MATS7e, BELe3
11	9	0.82	0.67	0.65	BELp8, nNHRPh,ATS4m, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u
12	10	0.82	0.68	0.66	BELp8, nNHRPh,ATS4m, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u, G2u
13	11	0.83	0.69	0.67	BELp8, nNHRPh,ATS4m, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u, G2u,PJI2

14	12	0.84	0.70	0.68	BELp8, nNHRPh,ATS4m, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u, G2u,PJI2, G2s
15	13	0.85	0.72	0.69	BELp8, nNHRPh,ATS4m, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u, G2u,PJI2, G2s, nROR
16	14	0.85	0.72	0.70	BELp8, nNHRPh,ATS4m, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u, G2u,PJI2, G2s, nROR, BELp4
17	15	0.86	0.73	0.71	BELp8, nNHRPh,ATS4m, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u, G2u,PJI2, G2s, nROR, BELp4, RDF090m
18	14	0.86	0.73	0.71	nNHRPh,ATS4m, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u, G2u,PJI2, G2s, nROR, BELp4, RDF090m
19	15	0.86	0.74	0.72	nNHRPh,ATS4m, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u, G2u,PJI2, G2s, nROR, BELp4, RDF090m, Mor08u
20	14	0.86	0.74	0.72	nNHRPh, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u, G2u, PJI2, G2s, nROR, BELp4, RDF090m, Mor08u
21	15	0.87	0.75	0.73	nNHRPh, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u, G2u,PJI2, G2s, nROR, BELp4, RDF090m, Mor08u,GGI4
22	16	0.87	0.76	0.74	nNHRPh, C-032, Mor27u, C-005, MATS7e, BELe3, Mor02u, G2u,PJI2, G2s, nROR, BELp4, RDF090m, Mor08u,GGI4, nCONHR

As it's obvious in table (3.2) the best model with highest R^2, R^2_{adj} is model 22. Model 22 represented by the following equation:

pIC_{50} = 9.478 (±2.848) + 2.272 (± 0.284) nNHRPh – 1.053 (± 0.249) C-032 - 0.964 (±0.155) Mor27u – 0.140 (± 0.069) C-005 + 0.908 (± 0.315) MATS7e – 3.236 (± 0.644) BELe3 + 0.015 (± 0.007) Mor02u + 21.013 (± 5.348) G2u – 2.186 (± 0.551) PJI2 – 21.897 (± 9.297) G2s – 0.333 (± 0.082) nROR + 2.328 (± 0.621) BELp4 +0.046 (± 0.014) RDF 090m – 0.233 (± 0.060) Mor08u + 0.278 (± 0.104) GGI4 – 0.219 (± 0.095) nCONHR.

According to the above equation, the most important descriptors are G2s and G2u which reflect the molecular geometrical coordinates of the compounds; G2s is inversely proportional to the inhibitory activity of the compounds while G2u is directly proportional to the inhibitory activity of the compounds.

R^2 is statistical parameters that give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1.0 indicates that the regression line perfectly fits the data, while R^2 of zero indicates no 'linear' relationship between the dependent variable and independent variables. R^2 is only a descriptive measure and it does not measure the quality of the regression model. Accordingly, focusing solely on maximizing R^2 is not a good idea. So we need extra validation for MLR models to judge the productivity of the models.

We applied LOO and LMO cross validation on the MLR models in table (3.2) using matlab software and the results are summarized in tables (3.3), (3.4) below:

Table (3.3): LOO cross validation results.

model	No. descriptors	PRESS	SPRESS	SST	R^2_{CV}	PRESS/SST	PSE	RSEP
6	6	58.58	0.56	92.64	0.37	0.63	0.55	7.73
7	7	55.91	0.55	95.31	0.41	0.59	0.54	7.55
8	8	53.92	0.54	97.30	0.45	0.55	0.53	7.41
9	7	54.52	0.54	96.70	0.44	0.56	0.53	7.45
10	8	52.34	0.53	98.88	0.47	0.53	0.52	7.30
11	9	50.53	0.53	100.69	0.50	0.50	0.51	7.18
12	10	48.73	0.52	102.49	0.52	0.48	0.50	7.05
13	11	47.08	0.51	104.14	0.55	0.45	0.50	6.93
14	12	44.82	0.50	106.40	0.58	0.42	0.48	6.76
15	13	43.11	0.49	108.11	0.60	0.40	0.47	6.63
16	14	41.80	0.49	109.43	0.62	0.38	0.47	6.53
17	15	40.32	0.48	110.90	0.64	0.36	0.46	6.41
18	14	40.50	0.48	110.72	0.63	0.37	0.46	6.42
19	15	38.89	0.47	112.33	0.65	0.35	0.45	6.29
20	14	39.09	0.47	112.14	0.65	0.35	0.45	6.31
21	15	37.75	0.46	113.48	0.67	0.33	0.44	6.20
22	16	36.64	0.46	114.59	0.68	0.32	0.43	6.11

Table (3.4): LMO cross validation results

model	No. descriptors	PRESS	SPRESS	SST	R^2_{CV}	PRESS/SST	PSE	RSEP
6	6	55.95	0.55	93.92	0.40	0.60	0.54	7.54
7	7	55.56	0.55	97.46	0.43	0.57	0.54	7.51
8	8	54.99	0.55	101.13	0.46	0.54	0.54	7.47
9	7	54.99	0.55	101.23	0.46	0.54	0.54	7.47
10	8	54.60	0.55	103.70	0.47	0.52	0.53	7.45
11	9	53.09	0.54	105.61	0.50	0.50	0.53	7.34
12	10	52.52	0.54	108.17	0.51	0.49	0.52	7.30
13	11	50.53	0.53	110.52	0.54	0.46	0.51	7.17
14	12	47.86	0.52	108.57	0.56	0.44	0.50	6.97
15	13	46.23	0.51	110.02	0.58	0.42	0.49	6.85
16	14	45.60	0.51	111.73	0.59	0.41	0.49	6.81
17	15	46.07	0.51	111.92	0.59	0.41	0.49	6.84
18	14	45.70	0.51	110.51	0.59	0.41	0.49	6.81
19	15	43.06	0.49	111.65	0.61	0.39	0.47	6.61
20	14	42.44	0.49	110.69	0.62	0.38	0.47	6.57
21	15	41.57	0.49	113.49	0.63	0.35	0.47	6.50
22	16	41.57	0.48	113.94	0.64	0.36	0.47	6.50

PRESS (predictive residual sum of squares) which is a standard index to measure the accuracy of the model. It is also called SSE (error sum of squares), SST (total sum of squares), R^2_{cv} (cross-validated correlation coefficient), SPRESS (uncertainty of prediction), PSE (Predictive Square Errors) and also called RMSE (root mean square error), and RSEP is relative standard error of prediction.

From the results of LOO and LMO cross validation in tables (3.3) and (3.4) we can see that PRESS values always less than SST values and this means that the models predicting ability better than chance.

We use the values of PSE and R^2_{CV} in the tables to draw the following two graphs below. Figure (3.1) which is R^2_{CV} Vs model no. of both LOO & LMO, and figure (3.2) that shows PSE Vs model no. of both LOO & LMO.

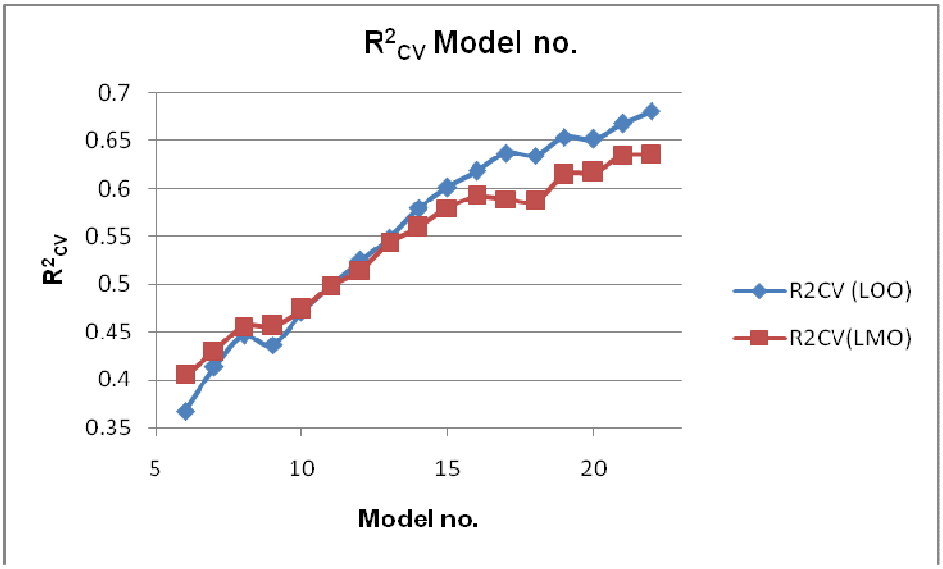


Figure (3.1): R^2_{CV} Vs model no. of both LOO & LMO

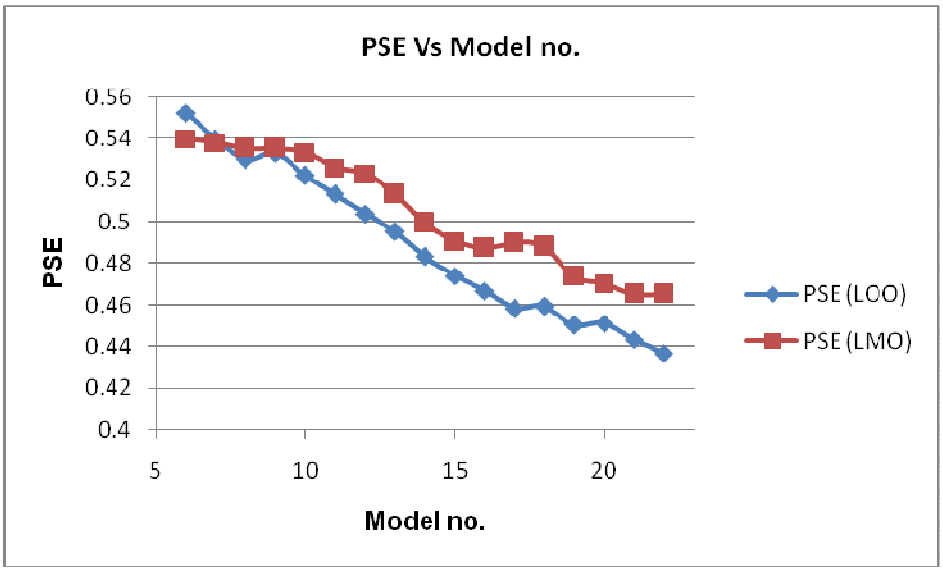


Figure (3.2): PSE Vs model no. of both LOO & LMO

We can see from the values and the graphs that models (15-22) are the best models having the best validation parameters with the lowest values of PSE and highest values of R^2_{CV}. In both graphs we can see the slightly change in the curves from model 15 till model 22. So we picked those models as the best models to be the candidates to next steps, as inputs to PC-ANN and PLS.

PC-ANN

Before running ANN we applied PCA on the data to get rid of the inter correlation between descriptors, divide the descriptors into training, validation and test sets and get rid of outliers compounds that disturb the model. So we used the proper matlab script and apply it on a file that had the activity as the first column and all the descriptors of the picked models (15-22) as the other columns. The result of PCA application is shown in the figure (3.3) below, the graph shows the compounds disruption in the space of the first and the second principle components.

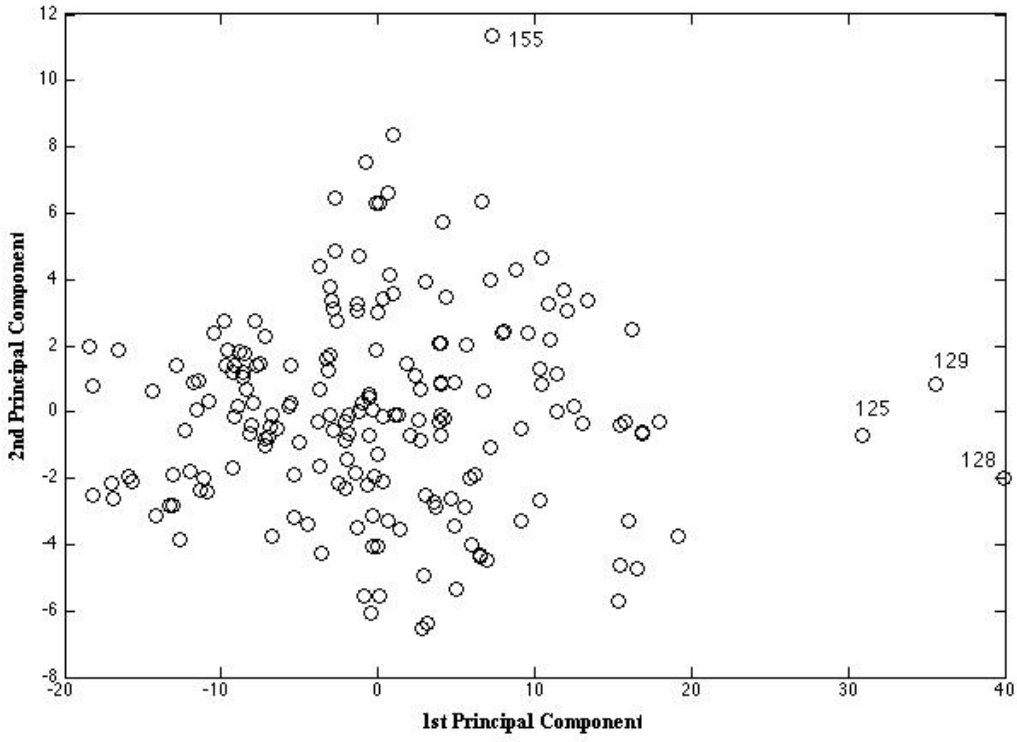


Figure (3.3): correlation between 1st and 2nd principle components.

According to the above figure, compounds 125,128,129,155 are considered as outliers; outlier compounds are the compounds that lie far from the compounds cluster. This means that those four compounds act in different way from other compounds with respect to activity and descriptors. The other 188 compounds were partitioned into validation set 20%, test set 20% and the other 60% for training set. The compounds of each set were picked from the whole area of the compounds cluster. Those 188 compounds used as data points to ANN.

PC-ANN models were built using the proper matlab scripts. We applied the script on each one of the picked models (15-22) using constant hidden nodes for all the models. The table (3.5) below shows a summary of the cross validation parameters of the models.

Table (3.5): Correlation coefficient and cross validation parameters for ANN models (15-22).

Mo no.	nPCs	R_tr	PRESS _tr	R^2_{CV}_tr	R_test	PRESS _test	RSEP _test	R_val	PRESS _val	R2CV _val	RSEP _val
15	6	0.810	32.86	0.29	0.76	13.85	8.52	0.67	13.67	-0.10	8.48
16	6	0.817	31.00	0.42	0.75	14.00	8.57	0.72	11.77	0.14	7.87
17	6	0.814	31.34	0.43	0.77	12.91	8.23	0.78	12.75	0.14	8.18
18	7	0.841	27.50	0.50	0.78	12.89	8.22	0.71	12.12	0.10	7.98
19	6	0.813	31.92	0.35	0.75	14.32	8.67	0.60	13.44	0.01	8.40
20	7	0.828	29.46	0.44	0.75	14.03	8.58	0.76	10.67	0.31	7.49
21	6	0.802	33.42	0.31	0.75	14.26	8.65	0.70	12.72	0.03	8.18
22	7	0.825	29.84	0.44	0.75	14.10	8.60	0.73	11.44	0.25	7.75

Mo: model.

We used those values in the table above to draw several graphs to judge the results, and to choose the best models for the next step in ANN work.

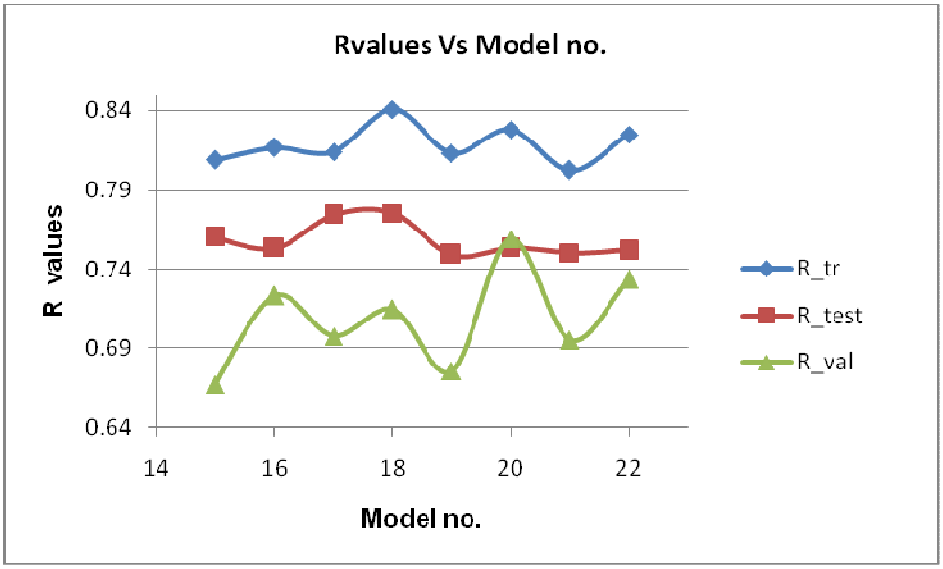


Figure (3.4): correlation coefficient values against ANN model numbers.

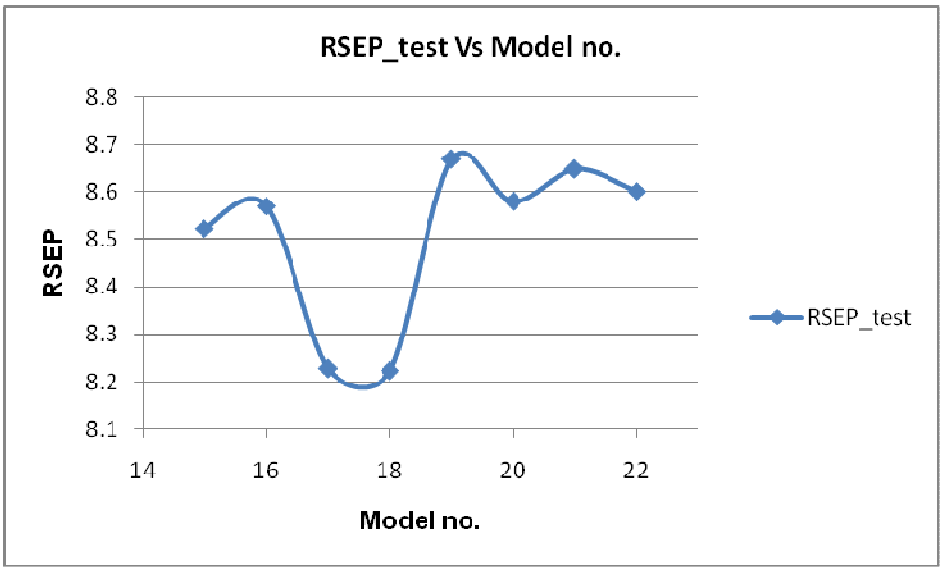


Figure (3.5): RSEP values of the test set against the model numbers.

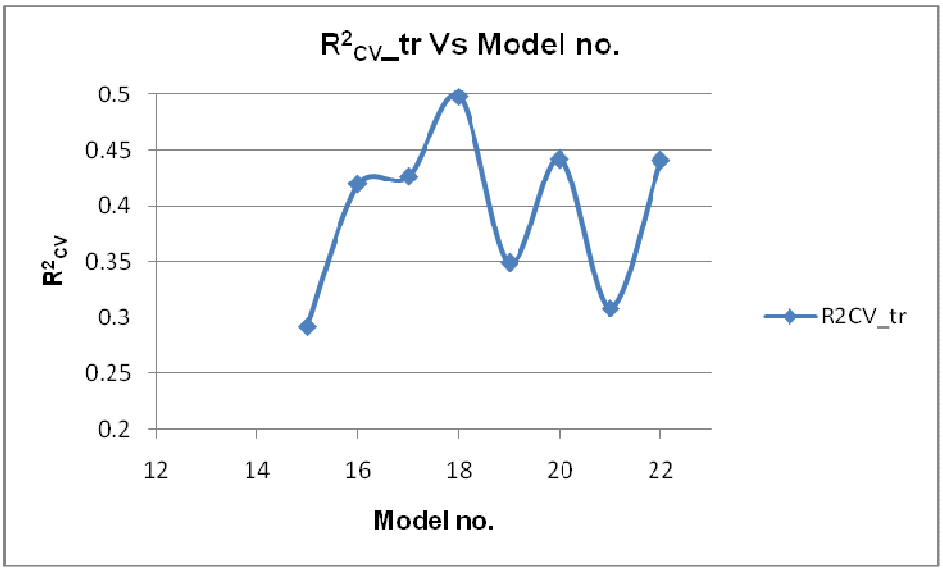


Figure (3.6): R^2_{CV} values of the training set against the model numbers.

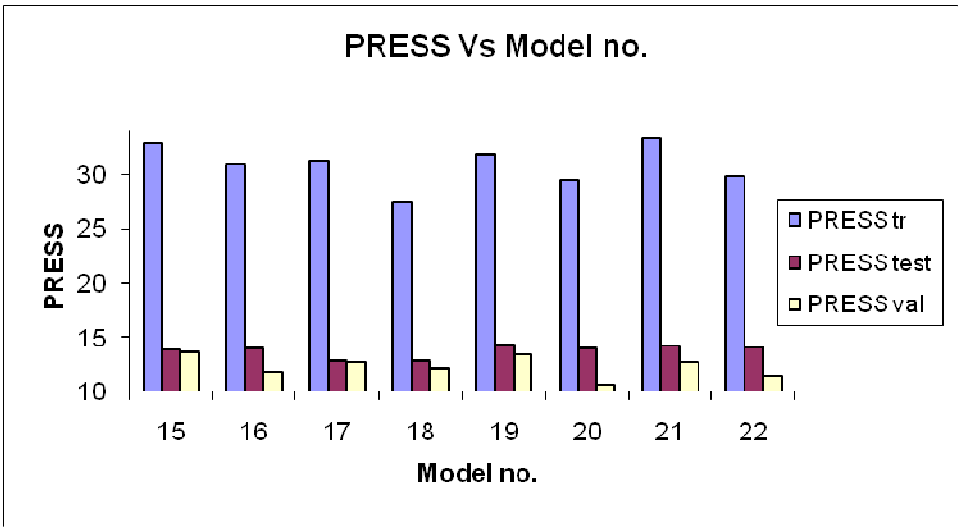


Figure (3.7): PRESS values against the model numbers.

The results are somehow regular. From figure (3.4), we can see that the model with the highest R of the test set is model 18, and the one after it is model 17, but when we compared the results according to R for the training set, model 18 is the highest but 20 is the one after, and the results aren't the same according to R of the validation set model 20 is the best then model 22.

While when we look for figure (3.5) the models with the lowest RSEP values of the test set are 18, 17. From figure (3.6): the models with the highest R^2_{CV} for the training set are 18, 22, and

20 respectively. At the same time, figure (3.7) shows that the models with the lowest PRESS values of the training set are 18, 20, and 22 respectively.

Depending on the previous discussion we picked the models 18, 20, 17, and 22, to be the input for the next step to choose the best PC- ANN model.

Each one of those models was used to train the ANN model using different numbers of hidden nodes from (3-20). And we determined that we want R for the test set to be more than 0.75. The results of the four models are summarized in the tables (3.6), (3.7), (3.8), and (3.9). We take R-of the test set as our prior parameter to pick the best trial in each model because it reflects the predictivity of the models.

Table (3.6): correlation coefficients and cross validation parameters of model 17.

Hn no.	nPCs	R_tr	PRESS_tr	R^2_{CV}_tr	R_test	PRESS_test	RSEP_test	R_val	PRESS_val	RSEP_ val
5	6	0.80	33.07	0.36	0.77	13.13	8.30	0.70	12.63	8.15
6	6	0.81	31.34	0.43	0.77	12.91	8.23	0.70	12.75	8.18
7	6	0.81	32.35	0.33	0.76	14.10	8.60	0.66	14.18	8.63
8	**6**	**0.83**	**29.81**	**0.44**	**0.78**	**12.71**	**8.16**	**0.71**	**12.00**	**7.94**
9	6	0.85	26.10	0.54	0.76	13.9	8.54	0.68	13.62	8.46
10	6	0.80	33.24	0.34	0.76	13.81	8.51	0.65	14.83	8.83
11	6	0.81	32.06	0.44	0.76	13.75	8.49	0.73	11.77	7.86
12	6	0.81	31.30	0.46	0.76	13.76	8.50	0.70	12.96	8.25
13	6	0.81	32.71	0.48	0.76	14.50	8.72	0.73	12.12	7.98
14	6	0.82	30.74	0.45	0.76	13.46	8.40	0.72	11.83	7.88
15	6	0.87	23.51	0.59	0.78	12.82	8.20	0.61	15.99	9.16
16	6	0.80	32.82	0.42	0.76	13.79	8.50	0.61	16.33	9.26
17	6	0.81	31.58	0.49	0.77	13.51	8.42	0.72	12.08	7.97
18	6	0.85	25.77	0.59	0.76	14.00	8.57	0.65	15.12	8.91
19	6	0.80	33.14	0.52	0.76	14.02	8.57	0.62	16.62	9.35
20	6	0.80	33.11	0.53	0.77	14.46	8.71	0.66	14.61	8.76

Hn no.: number of hidden nodes.

With model 17 at 3 and 4 hidden nodes we did not get R of the test set more than 0.75 so I neglected the trials. The trail with the highest R of the test set is with 8 hidden nodes and by looking at all the parameters we notice that they are acceptable. So from this model we picked the trial with 8 hidden nodes to be compared with other models.

Table (3.7): correlation coefficients and cross validation parameters of model 18.

Hn no.	nPCs	R_tr 2	PRESS_tr	R2CV_tr	R_test	PRESS _test	RSEP_ test	R_val	PRESS _val	RSEP _ val
4	7	0.81	32.77	0.32	0.76	13.87	8.53	0.70	12.69	8.17
5	7	0.82	30.83	0.33	0.75	14.04	8.58	0.70	12.53	8.11
6	**7**	**0.84**	**27.50**	**0.50**	**0.78**	**12.89**	**8.22**	**0.71**	**12.12**	**7.98**
7	7	0.82	31.40	0.37	0.75	14.19	8.63	0.68	13.25	8.34
8	7	0.83	29.83	0.41	0.76	13.62	8.45	0.68	13.39	8.39
9	7	0.83	28.60	0.50	0.75	14.28	8.65	0.69	13.65	8.47
10	7	0.83	29.99	0.41	0.76	13.59	8.44	0.68	13.39	8.39
11	7	0.83	29.03	0.45	0.77	13.07	8.28	0.68	13.30	8.36
12	7	0.82	31.31	0.40	0.77	13.34	8.36	0.68	13.39	8.39
13	7	0.80	33.85	0.31	0.75	14.01	8.57	0.69	12.92	8.24
14	7	0.81	31.31	0.43	0.75	14.16	8.62	0.68	13.75	8.50
15	7	0.81	32.48	0.28	0.77	13.34	8.36	0.65	14.56	8.75
16	**7**	**0.83**	**29.45**	**0.42**	**0.78**	**12.83**	**8.20**	**0.68**	**14.15**	**8.62**
17	7	0.81	31.49	0.45	0.75	14.20	8.63	0.67	15.44	9.01
18	7	0.80	34.61	0.19	0.75	14.23	8.64	0.65	14.31	8.67
19	7	0.85	25.69	0.57	0.76	14.03	8.58	0.66	14.82	8.82
20	7	0.85	26.25	0.56	0.77	13.10	8.29	0.65	15.66	9.07

Hn no.: number of hidden nodes.

In model 18 the trail with 3 hidden nodes failed to reach 0.75 as R-test so the trial was neglected. In this model R-test of the trials with 6 and 16 hidden nodes are the best. We picked the trial with the 6 hidden nodes because of the risk of over fitting which increases with increasing the number of hidden nodes. And by looking at the other parameters in general, we notice that the trial with 6 hidden nodes better than that with 16 hidden nodes.

Table (3.8): correlation coefficients and cross validation parameters of model 20.

Hn no.	nPCs	R_tr	PRESS_tr	R2CV_tr	R_test	PRESS_test	RSEP_ test	R_val	PRESS_val	RSEP_ val
5	7	0.80	33.50	0.32	0.76	13.56	8.43	0.77	10.01	7.25
6	7	0.83	29.46	0.44	0.75	14.03	8.58	0.76	10.67	7.49
7	7	0.84	28.10	0.49	0.76	13.97	8.56	0.75	10.86	7.55
8	7	0.83	29.17	0.49	0.75	14.05	8.58	0.75	10.72	7.50
9	7	0.80	32.90	0.39	0.75	14.00	8.57	0.74	11.33	7.71
10	7	0.82	31.16	0.37	0.76	13.90	8.54	0.65	14.38	8.69
11	7	0.80	33.26	0.39	0.76	13.49	8.41	0.71	12.11	7.98
12	7	0.80	32.74	0.40	0.76	13.56	8.43	0.71	12.15	7.99
13	**7**	**0.86**	**24.34**	**0.59**	**0.77**	**13.02**	**8.26**	**0.74**	**11.40**	**7.74**
14	**7**	**0.85**	**25.70**	**0.54**	**0.78**	**13.23**	**8.33**	**0.74**	**11.39**	**7.74**
15	7	0.86	24.27	0.60	0.75	14.31	8.66	0.74	11.15	7.65
16	7	0.86	23.74	0.59	0.77	14.09	8.60	0.71	12.35	8.06
17	7	0.81	33.74	0.24	0.77	13.16	8.31	0.76	10.87	7.56
18	7	0.86	24.11	0.64	0.77	13.36	8.37	0.70	13.27	8.35
19	7	0.87	21.86	0.71	0.77	13.64	8.46	0.74	12.18	8.00
20	7	0.83	28.27	0.55	0.77	13.67	8.47	0.64	15.51	9.03

Hn no.: number of hidden nodes.

In model 20 the trails with 3 and 4 hidden nodes are neglected for the same reason as before. The trial with the best R-test is with 14 hidden nodes and the one after it is with 13 hidden nodes. We preferred to pick the trial with 13 hidden nodes because the other parameters are better for this trial.

Table (3.9): correlation coefficients and cross validation parameters of model 22.

Hn no.	nPCs	R_tr	PRESS _tr	R^2_{CV} _tr	R_test	PRESS _test	RSEP _ test	R_val	PRESS _val	RSEP _ val
3	7	0.80	33.36	0.32	0.75	14.00	8.57	0.72	12.12	7.98
4	7	0.84	27.91	0.47	0.76	13.69	8.47	0.68	13.38	8.39
5	7	0.86	25.34	0.52	0.77	13.04	8.27	0.76	10.64	7.48
6	7	0.83	29.84	0.44	0.75	14.10	8.60	0.73	11.44	7.75
7	**7**	**0.84**	**27.18**	**0.50**	**0.81**	**11.43**	**7.74**	**0.77**	**10.03**	**7.26**
8	7	0.87	23.05	0.60	0.76	13.73	8.49	0.73	11.37	7.73
9	**7**	**0.87**	**22.40**	**0.63**	**0.78**	**12.69**	**8.16**	**0.71**	**12.86**	**8.22**
10	7	0.86	24.56	0.60	0.77	13.43	8.39	0.71	12.59	8.13
11	7	0.90	18.48	0.72	0.77	13.10	8.29	0.73	11.82	7.88
12	7	0.89	19.78	0.69	0.76	13.43	8.39	0.75	10.92	7.58
13	**7**	**0.89**	**19.79**	**0.71**	**0.79**	**12.58**	**8.12**	**0.72**	**12.64**	**8.15**
14	7	0.88	21.50	0.68	0.77	13.16	8.31	0.77	10.22	7.33
15	7	0.90	18.42	0.73	0.75	14.00	8.57	0.76	10.45	7.41
16	7	0.89	20.16	0.69	0.79	12.15	7.98	0.75	11.52	7.78
17	7	0.91	16.00	0.78	0.80	11.97	7.92	0.66	14.81	8.82
18	7	0.89	18.74	0.74	0.80	11.91	7.90	0.71	13.06	8.28
19	7	0.85	25.45	0.57	0.77	13.56	8.43	0.74	11.43	7.75
20	7	0.89	20.11	0.67	0.78	12.79	8.19	0.70	12.93	8.24

Hn no.: number of hidden nodes.

With model 22 the one with best results as R test set is with 7 hidden nodes but it has R^2_{CV} of the training set equal to 0.5 and it must be greater than 0.5. So there are the models with 13 and 9 hidden nodes with good results of R-test and they have R^2_{CV} over than 0.5, we chose the trial with 9 hidden nodes because the over fitting risk with 13 hidden nodes.

We summarized the results of the hidden nodes optimization step in table (3.10) below.

Table (3.10): Correlation coefficients and cross validation parameters of the optimal number of hidden nodes of each model.

Mo no.	nHn	nPCs	R_tr	PRESS_tr	R^2_{CV}_tr	R_test	PRESS_test	RSEP_test	R_val	PRESS_val	RSEP_val
17	8	6	0.83	29.81	0.44	0.78	12.71	8.16	0.71	12.00	7.94
18	6	7	0.84	27.50	0.50	0.78	12.89	8.22	0.71	12.12	7.98
20	13	7	0.86	24.34	0.59	0.77	13.02	8.26	0.74	11.40	7.74
22	9	7	0.87	22.40	0.63	0.78	12.69	8.16	0.71	12.86	8.22

Mo refers to: model.

Hn no.: number of hidden nodes.

By gathering the best results, we found that Model 22 with 9 hidden nodes is the best with the highest R training, R test, R^2_{CV} test, and lowest PRESS test set, RSEP test set, PRESS training set. But model 20 with 13 hidden nodes has higher R val, and lower RSEP val, PRESS val set, and accepted value of R^2_{CV} of training set over than 0.5. We can see that model 22 with 9 hidden nodes and model 20 with 13 hidden nodes are the best with respect to correlation coefficients and cross validation parameters.

Models **20** and **22** were examined to inspect the presence of outliers that may affect models validity. By inspecting the residuals of these models there were no outliers as it is shown in figures (3.8) and (3.9). In both figures there is no compounds lie far from the compounds cluster in training, validation, and test set graphs.

Figure (3.8) and figure (3.9) show regressions between observed and predicted activity as well as their residuals for the training, validation, and test sets for these two models.

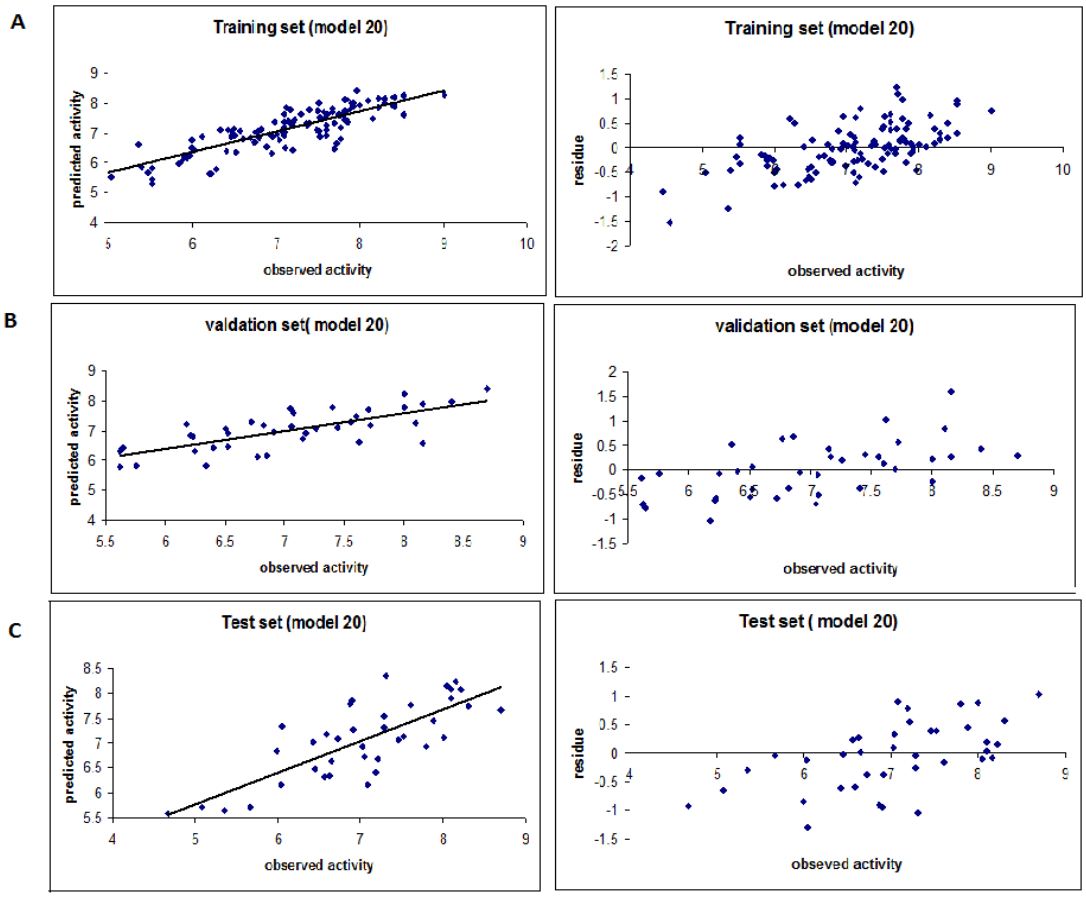


Figure (3.8): Plot of the predicted activity against observed one as well as their residues for model 20 using 13 hidden nodes. (a) training set, (b) validation set, and (c) external test set.

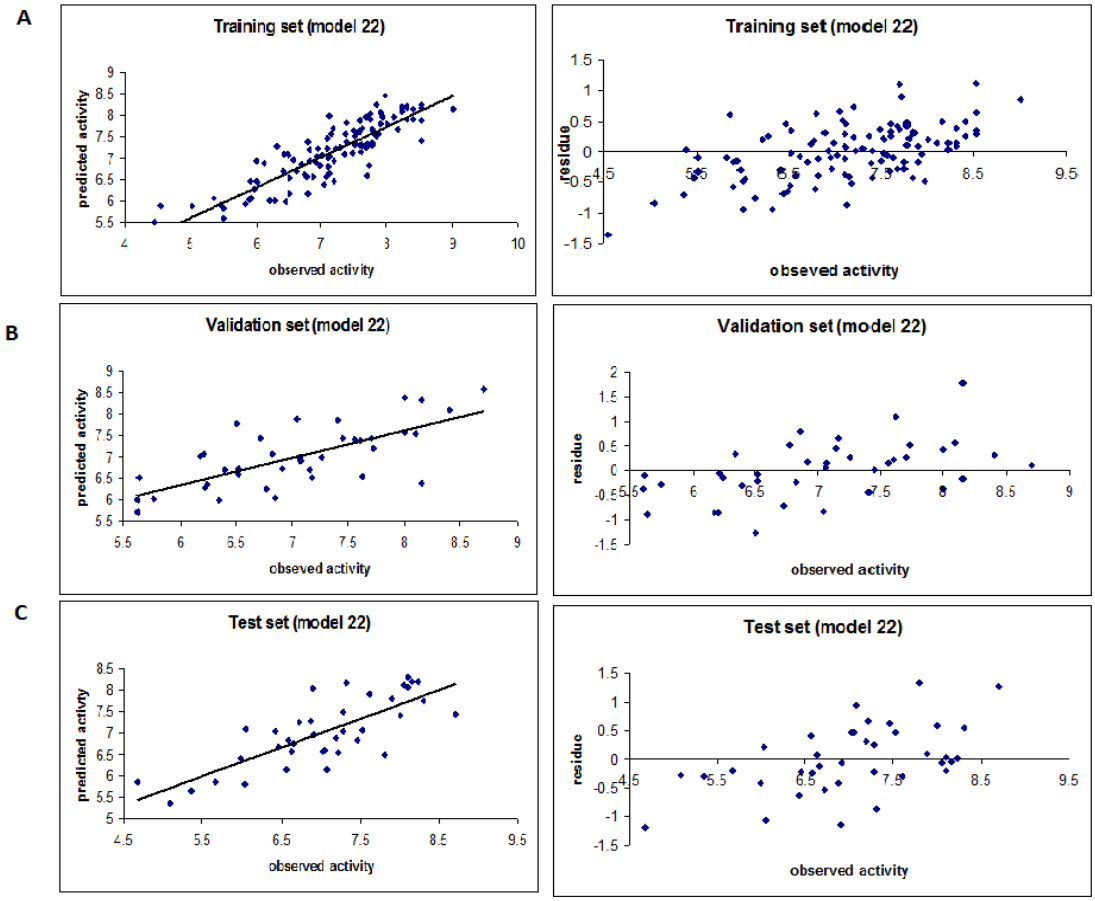


Figure (3.9): Plot of the predicted activity against observed one as well as their residues for model 22 using 9 hidden nodes. (a) training set, (b) validation set, and (c) external test set.

Randomization

The results for randomization test were performed to investigate the probability of chance correlation for the optimal models (**20** with 13 hidden nodes and **22** with 9 hidden nodes). Tables (3.11) and (3.12) shows that the correlation coefficients obtained by chance are low in general while PRESS values are high. This indicates that the model obtained from PC-ANN is better than those obtained by chance.

Table (3.11): The results of chance correlation of model 20 with 13 hidden.

Trial no.	nPCs	R_ train	PRESS_ train	R2CV_ Train	R_ test	PRESS_ test	R_ val	PRESS_ val	R2CV_ Val
1	7	-0.36	119.4	-14.18	-0.02	35.07	-0.07	28.21	-10.32
2	7	-0.14	110.39	-10.62	-0.21	41.09	-0.25	32.61	-9.06
3	7	0.17	90.53	-12.04	0.02	34.11	-0.03	27.34	-14.63
4	7	-0.36	119.4	-14.18	-0.02	35.07	-0.07	28.21	-10.32
5	7	-0.16	105.92	-16.68	-0.13	39.69	-0.3	30.73	-14.97
6	7	-0.07	98.86	-24.69	0.07	33.8	0.06	25.91	-12.91
7	7	0.09	97.64	-8.47	0.04	34.71	0.06	27.49	-5.8
8	7	0.17	90.53	-12.04	0.02	34.11	-0.03	27.34	-14.63
9	7	-0.14	110.39	-10.62	-0.21	41.09	-0.25	32.61	-9.06
10	7	-0.29	121.71	-9.926	0.04	36.61	0.01	29.79	-5.43

Table (3.12): The results of chance correlation of model **22** with 9 hidden nodes.

Trial no.	nPCs	R_ train	PRESS_ train	R2CV_ train	R_ test	PRESS_ test	R_ val	PRESS_ val	R2CV_ Val
1	**7**	**-0.26**	**117.14**	**-11.31**	**-0.15**	**38.23**	**0.16**	**24.19**	**-17.28**
2	**7**	**0.01**	**98.77**	**-21.18**	**-0.03**	**33.9**	**-0.21**	**28.42**	**-19.14**
3	**7**	**0.06**	**99.38**	**-8.21**	**-0.1**	**39.3**	**-0.19**	**35.26**	**-5.03**
4	**7**	**0.07**	**94.62**	**-19.3**	**-0.19**	**36.96**	**-0.13**	**26.88**	**-25.57**
5	**7**	**-0.33**	**127.41**	**-9.96**	**-0.28**	**42.8**	**0.11**	**26.65**	**-5.84**
6	**7**	**-0.25**	**104.58**	**-29.81**	**-0.3**	**37.11**	**-0.29**	**29.01**	**-22.78**
7	**7**	**-0.26**	**123.54**	**-9.1**	**0.12**	**34.4**	**0.02**	**26.91**	**-9.11**
8	**7**	**-0.32**	**131.84**	**-9.79**	**-0.16**	**43.47**	**-0.1**	**31.45**	**-5.9**
9	**7**	**-0.26**	**117.14**	**-11.31**	**-0.15**	**38.22**	**0.16**	**24.19**	**-17.28**
10	**7**	**0.01**	**98.77**	**-21.18**	**-0.03**	**33.9**	**-0.21**	**28.42**	**-19.14**

And to ensure the robustness of the optimal models (20 with 13 hidden nodes and 22 with 9 hidden nodes) we applied Y- randomization. The results are shown in table (3.13) below. And it's obvious in the table that the correlation coefficients are lower than the results of the original models. This indicates that the developed models considered being robust enough. The values of cR^2_p are above 0.5 so both models are acceptable. (34)

Table (3.13): Values of correlation coefficients of randomized models and cR^2_p of the best PC-ANN models (20, 22):

Mo no.*	**R_train r***	**R_train**	**R^2 r ***	**R^2**	**cR^2_p ****
Model 20	0.58	0.86	0.33	0.74	0.55
Model 22	0.6	0.87	0.39	0.76	0.53
	R_test r *	**R_test**			
Model 20	0.07	0.78	0.01	0.60	0.60
Model 22	0.12	0.78	0.02	0.61	0.60

*Mo refers to: model, r refers to: randomized, R^2: coefficient of determination, R^2r: coefficient of determination for randomized trial.

** $cR^2p = R\sqrt{(R^2 - R^2r)}$

PLS

PLS analysis with cross validation was carried out for advance investigation of the linear relationships of the obtained regression models. (Table 3.14) summarizes correlation coefficients and cross validation parameters for models 15-22 and also gathered with MLR and PC-ANN results. From this table, we can see that the PLS results are close to MLR results and better than PC-ANN results. Also, model 20 and 22 are the best according to PLS in which the correlation coefficient for the prediction set is 0.866 and 0.860 for these models

respectively. Also the root mean square error of prediction (RMSE ^{p}) is 0.467 and 0.476 for model 20 and 22 respectively.

Table (3.14): Correlation coefficient for MLR, PLS and PC- ANN models (15-22) and cross validation parameters obtained from PLS and PC- ANN analysis.

Mod $_{no.}$	MLR		PC-ANN					PLS				
	R	SE	PC $_{No.}$	R^{c}	$R^{2}{}_{CV}{}^{c}$	R^{p}	$RMSE^{p}$	LV	R^{c}	$R^{2}{}_{CV}{}^{c}$	R p	$RMSE^{p}$
15	0.846	0.492	6	0.809	0.292	0.760	8.522	3	0.816	0.604	0.794	0.563
16	0.851	0.486	6	0.817	0.420	0.753	8.570	3	0.823	0.616	0.793	0.564
17	0.856	0.479	6	0.814	0.426	0.774	8.228	5	0.854	0.662	0.852	0.498
18	0.856	0.478	7	0.841	0.498	0.776	8.222	4	0.847	0.645	0.846	0.497
19	0.862	0.470	6	0.813	0.350	0.749	8.668	4	0.852	0.655	0.866	0.469
20	0.861	0.470	7	0.828	0.442	0.753	8.579	4	0.851	0.656	0.866	0.467
21	0.866	0.463	6	0.802	0.309	0.750	8.647	4	0.857	0.667	0.860	0.476
22	0.870	0.458	7	0.825	0.440	0.752	8.601	4	0.861	0.670	0.860	0.476

C: calibration set: training set.

P: Prediction set: test set.

SE: Standard error.

RMSE: root mean square error.

Comparison with other QSAR studies:

In the previous studies they used less number of compounds with limited cores, but in our study we used large number of VEGFR-2 compounds (192) with variety of cores to have a general QSAR model counting for the inhibitory activity in which we depend on large pool of descriptors (1497) not a particular class of them. MLR was applied and gave us good results with R^2 = 0.758**.** To confirm our results, PC-ANN was applied too, in which R^2= 0.681 which may be explained that the relation between the inhibitory activity (pIC_{50}) and the structures is linear. In order to be convinced with this linear relation, PLS was applied and the results are even better than MLR in which R^2= 0.741 and R^2cv = 0.670. Regarding previous studies, in 2009 comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) were performed on a series of selective inhibitors of VEGFR-2 (82 compounds). The best CoMFA and CoMSIA models gave a cross-validated coefficient R^2cv of 0.546 and 0.715 respectively (6).

In 2010 Simultaneous optimization has been proposed and evaluated on a set of novel VEGFR-2 inhibitors including naphthalene and indazole-based compounds (61 compounds). The final support vector regression model was constructed on an optimal set of six descriptors. The results are R^2 (0.908, 0.837) for 45 training and 16 test samples. It is clear that number of compounds is small in comparison to our study (7).

In 2011 3D-QSAR technique was carried out on diaryl acylsulfonamide derivatives as VEGFR-2 inhibitors by using Comparative Molecular Field Analysis (CoMFA) studies to find relations between biological activities of inhibitors and their structures. 3DQSAR technique was applied to a set of fifty ligands in order to facilitate the design of more potent inhibitors. The maximum cross-validated correlation coefficient value was found to be 0.417 (36).

In 2011 two 3D-QSAR models were built using CoMFA and CoMSIA methods, using 80 compounds of indolocarbazole series as VEGFR-2 inhibitors. The two QSAR models with highest predictabilities obtained were (CoMFA model: R^2cv = 0.823, R^2 = 0.979; CoMSIA model: R^2cv = 0.804, R^2 = 0.967). (37)

Another study (38) developed a nonlinear model for the inhibition activities for a set of pyrazine-pyridine biheteroaryls, inhibitors of Vascular Endothelial Growth Factor Receptor-2 (VEGFR-2) based on Least Squares Support Vector Machines (LSSVMs). Five relevant

descriptors selected by heuristic method were used to build linear and nonlinear QSAR models using MLR and LS-SVMs. The nonlinear LS-SVMs model gave the correlation coefficients of 0.921 and the MSE of 0.046 for the training set. The corresponding correlation coefficient and MSE for the test set are 0.877 and 0.041, respectively.

These results have been accepted for publication in the journal: Current Pharmaceutical Design (39).

Chapter Four

Conclusion

To study the inhibitory activity of vascular endothelial growth factor receptor-2 (VEGFR-2), the structures of 192 compounds were built using Hyperchem software. Those structures were optimized using AM1 semi-imperical method. We calculated different groups of descriptors, some descriptors were calculated using Hyperchem software and others were calculated using Dragon software.

SPSS software was used to get the multiple linear QSAR equations. We obtained eight multiple linear equations with good statistical qualities and predictive power. Those equations were used as input to PC-ANN modeling. The best four PC-ANN models were used for hidden nodes optimization. The cross validation results of PC-ANN models were not better than MLR as expected, so we decided to perform PLS analysis to confirm the linearity of the relationship. The eight MLR equations were used as PLS input. The PLS gives improved regression models with better prediction ability compared with PC-ANN.

The most important descriptors in this relation are G2s and G2u which reflect the molecular geometrical coordinates of the compounds; G2s is inversely proportional to the inhibitory activity of the compounds while G2u is directly proportional to the inhibitory activity of the compounds. PLS based models are quite good to describe the QSAR of the VEGFR-2 for the data set in this investigation. A 0.866 and 0.860 correlation coefficients were obtained for the best two models using PLS with 4 LV′s.

References

1. Mousa S. A., (2000), Angiogenesis Inhibitors and Stimulators: Potential Therapeutic Implications, Landes Bioscience Gerogetown, Texas U.S.A.
2. Karamysheva A. F. (2008), Mechanisms of Angiogenesis, Biochemistry (Moscow), 2008, 73 (7): pp. 751-762.

3. Carmeliet B. (2003), angiogenesis in health and disease, Nature medicine, 9 (6): pp 653-660.
4. Young, D.Y. (2001), Computational Chemistry, a practical Guide for applying Techniques to Real- Word problem, John Wiley & sons, Inc. New York.

5. Jensen F. (2007), Introduction to Computational Chemistry, 2ed Ed, John Wiley & Sons, Inc. New York.

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

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

8. Dudek A. Z., Arodz T. and Galvez J. (2006), Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review, Combinatorial Chemistry & High Throughput Screening, 9: pp 213-228 213

9. Selassie C. D., (2003), History of Quantitative Structure-Activity Relationships: Burger's Medicinal Chemistry and Drug Discovery, 6th Ed, Vol 1, John Wiley&Sons, Inc. New York.

10. Deeb O., Hemmateenejad B., Jaber A., Garduno-Juarez R. and Miri R. (2007) " 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 67(11): 2122-2130.

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

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

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

14. Verma J., Khedkar V. M. and Coutinho E. C. (2010), 3D-QSAR in Drug Design - A Review, Current Topics in Medicinal Chemistry, 10: pp 95-115.

15. Dudek A. Z., Arodz T., and Galvez J. (2006), Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review, Combinatorial Chemistry & High Throughput Screening, 9: pp213-228.

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

17. Wang J. C. (2011), Linear Regression: Multiple Linear Regression.

18. Tobias, R. (1995), An Introduction to Partial Least Squares Regression, in Proceedings of the Twentieth Annual SAS Users Group International Conference, Cary, NC: SAS Institute Inc., pp 1250 -1257.

19. Partial Least Squares Regression (PLSR). http://www.vcclab.org/lab/pls/m_description.html#Top

20. Dondeti S., Kannan K., and Manavalan R. (2005), Principal Component Artificial Neural Network Calibration Models for Simultaneous Spectrophotometric Estimation of Phenobarbitone and Phenytoin Sodium in Tablets, Acta Chim. Slov. 52: pp138–144.

21. Deeb O., Khadikar P. V., and Goodarzi M. (2010), QSPR Modeling of Bioconcentration Factors of Nonionic Organic Compounds, Environmental Health Insights, 4: pp 33–47.

22. Kuo G. H., Prouty C., Wang A., Emanuel S., DeAngelis A., Zhang Y., Song F., Beall L., Connolly P. J., Karnachi P., Chen X., Gruninger R. H., Sechler J., Pesquera A. F., Middleton S. A., Jolliffe L., and Murray W. V. (2005), Synthesis and Structure-Activity Relationships of Pyrazine-Pyridine Biheteroaryls as Novel, Potent, and Selective Vascular Endothelial Growth Factor Receptor-2 Inhibitors, J. Med. Chem. 48: pp 4892-4909.

23. Sun L., Liang C., Shirazian S., Zhou Y., Miller T., J. Cui, Fukuda J. Y., Chu J. Y., Nematalla A., Wang X., Chen H., Sistla A., Luu T. C., Flora Tang, Wei J., and Tang C. (2003), Discovery of 5-[5-Fluoro-2-oxo-1,2-dihydroindol-(3Z) ylidenemethyl]-2,4-dimethyl-1H-pyrrole-3-carboxylic Acid (2-Diethylaminoethyl)amide, a Novel Tyrosine Kinase Inhibitor Targeting Vascular Endothelial and Platelet-Derived Growth Factor Receptor Tyrosine Kinase, J. Med. Chem. 46: pp 1116-1119.

24. Sun L., Tran N., Liang C., Tang F., Rice A., Schreck R., Waltz K., Shawver L. K., McMahon G., and Tang C. (1999) Design, Synthesis, and Evaluations of Substituted 3-[(3- or 4-Carboxyethylpyrrol-2-yl)methylidenyl]indolin-2-ones as Inhibitors of VEGF, FGF, and PDGF Receptor Tyrosine Kinases, J. Med. Chem. 42: pp 5120-5130.

25. Bhide R. S., Cai Z. W., Zhang Y. Z., Qian L., Wei D., Barbosa S., Lombardo L. J., Borzilleri R. M., Zheng X., Wu L. I., Barrish J. C., Kim S. H., Leavitt K., Mathur A., Leith L., Chao S., Wautlet B., Mortillo S., Sr. R. J., Kukral D., Hunt J. T., Kamath A., Fura A., Vyas V., Marathe P., D'Arienzo C., G. Derbin, and Fargnoli J. Discovery and Preclinical Studies of (*R*)-1-(4-(4-Fluoro-2-methyl-1*H*-indol-5-yloxy)-5-

methylpyrrolo[2,1-*f*][1,2,4]triazin-6-yloxy)propan- 2-ol (BMS-540215), an In Vivo Active Potent VEGFR-2 Inhibitor 49, pp: 2143-2126.

26. Bilodeau M. T., Balitza A. E., Koester T. J., Manley P. J., Rodman L. D., Doepner C. B., Coll K. E., Fernandes C., Gibbs J. B., Heimbrook D. C., Huckle W. R., Kohl N., Lynch J. J., Mao X., McFall R. C., McLoughlin D., Stein C. M. M., Rickert K. W., Lorenzino L. S., Shipman J. M., Subramanian R., Thomas K. A., Wong B. K., Yu S., and Hartman G. D. (2004) Potent N-(1,3 Thiazol-2-yl)pyridin-2-amine Vascular Endothelial Growth Factor Receptor Tyrosine Kinase Inhibitors with Excellent Pharmacokinetics and Low Affinity for the hERG Ion Channel, J. Med. Chem. 47: pp 6363-6372.

27. Borzilleri R. M., Cai Z. w., Ellis C., Fargnoli J., Fura A., Gerhardt T., Goyal B., Hunt J. T., Mortillo S., Qian L., Tokarski J., Vyas V., Wautlet B., Zheng X., and Bhide R. S. (2005) Synthesis and SAR of 4-(3 hydroxyphenylamino)pyrrolo- [2,1-*f*][1,2,4]triazine based VEGFR-2 kinase inhibitors, Bioorganic & Medicinal Chemistry Letters, 15: pp 1429–1433.

28. McBride C. M., Renhowe P. A., Heise C., Jansen J. M., Lapointe G., Ma S., Pineda R., Vora J., Wiesmann M. and Shafer C. M. (2006) Design and structure–activity relationship of 3-benzimidazol-2-yl-1H-indazoles as inhibitors of receptor tyrosine kinases, Bioorganic & Medicinal Chemistry Letters, 16 : pp 3595–3599

29. Underiner T. L., Ruggeri B., Aimone L., Albom M., Angeles T., Chang H., Hudkins R. L., Hunter K., Josef K., Robinson C., Weinberg L., Yang S. and Zulli A. (2008) , TIE-2/VEGF-R2 SAR and in vitro activity of C3-acyl dihydroindazolo[5,4-a]pyrrolo[3,4-c]carbazole analogs, Bioorganic & Medicinal Chemistry Letters, 18: pp 2368–2372.

30. Gracias V., Ji Z., Zanze I. A., Zapatero C. A., Huth J. R., Song D., Hajduk P. J., Johnson E. F., Glaser K. B., Marcotte P. A., Pease L., Soni N. B., Stewart K. D., Davidsen S. K., Michaelides M. R., and Djuric S.W. (2008) , Scaffold oriented synthesis. Part 2: Design, synthesis and biological evaluation of pyrimido-diazepines as receptor tyrosine kinase inhibitors, Bioorganic & Medicinal Chemistry Letters, 18: pp2691–2695.

31. Ruel R., Thibeault C., LHeureux A., Martel A., Cai Z. W., Wei D., Qian L., Barrish J. C., Mathur A., DArienzo C., J. T. Hunt, Kamath A., Marathe P., Zhang Y., Derbin G., Wautlet B., Mortillo S., Jeyaseelan R., Sr., Henley B., Tejwani R., Rajeev S. Bhide, George L. Trainor, Fargnoli J., and Lombardo L. J. (2008), Discovery and preclinical studies of 5-isopropyl-6-(5-methyl- 1,3,4-oxadiazol-2-yl)-N-(2-methyl-1H-pyrrolo[2,3-b]pyridin- 5-yl)pyrrolo[2,1-f][1,2,4]triazin-4-amine (BMS 645737), an in vivo active potent VEGFR-2 inhibitor, Bioorganic & Medicinal Chemistry Letters, 18: pp 2985–2989.

32. Richon A. B., An Introduction to QSAR Methodology, http://www.netsci.org/Science/Compchem/feature19.html

33. Katritzky A.R., Karelson M., and Petrukhin R. (2005), Comprehensive DEscriptors for Structural and Statistical Analysis. http://www.codessa-pro.com/index.htm

34. Mitraa I., Sahab A., and Roya K. (2010), Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants, Molecular Simulation, 36 (13) : pp 1067–1079.

35. Golbraikh A, Tropsha A. (2002), Beware of q2!. J Mol Graphics Modelling, 20 (4): pp 269-276.

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

37. Tian Y, Xu J, Li Z, Zhu Z, Zhang J, Wu S.(2011), Combined 3D-QSAR and Docking Modelling Study on Indolocarbazole Series Compounds as Tie-2 Inhibitors, Int. J. Mol. Sci. 12: pp 5080-5097.

38. Jiazhong Li, Jin Qin, Huanxiang Liu, Xiaojun Yao, Mancang Liu and Zhide Hu. (2008), In Silico Prediction of Inhibition Activity of Pyrazine-Pyridine Biheteroaryls as VEGFR-2 Inhibitors Based on Least Squares Support Vector Machines. QSAR Comb Sci. 27(2): pp 157-164.

39. Deeb O., Jawabreh S. and Goodarzi M. (2013), Exploring QSARs of Vascular Endothelial Growth Factor Receptor-2 (VEGFR-2) Tyrosine Kinase Inhibitors by MLR, PLS and PC-ANN, Current Pharmaceutical Design, 19: in press.

دراسة العلاقة بين الصيغة البنائية والفاعلية باستخدام طريقة MLR,PLS, PC-ANN لمثبطات مستقبلات عامل نمو بطانة الأوعية الدموية.

مقدمة من:

سناء محمود جوابرة

بكالوريس كيمياء جامعة القدس فلسطين

بإشراف : د. عمر ديب

قدمت هذه الرسالة استكمالا لمتطلبات درجة الماجستير في

الكيمياء الصناعية والتطبيقية

دائرة الكيمياء والكيمياء الصناعية

برنامج الدراسات العليا في التكنولوجيا التطبيقية والصناعية

كلية العلوم والتكنولوجيا

جامعة القدس

1434/2012