Deanship of Graduate Studies

Al-Quds University

جامعة القدس
Al-Quds University

Exploring Quantitative Structure-Activity Relationships (QSARs) of Cyclooxygenase-2 (COX-2) Inhibitors by MLR, PLS and PC-ANN.

Nael Yahya Mohammad Zatari

M.SC. Thesis

Jerusalem-Palestine

1436 / 2014

Deanship of Graduate Studies

Al-Quds University

جامعة القدس
Al-Quds University

Exploring Quantitative Structure-Activity Relationships (QSARs) of Cyclooxygenase-2 (COX-2) Inhibitors by MLR, PLS and PC-ANN.

Nael Yahya Mohammad Zatari

M.SC. Thesis

Jerusalem-Palestine

1436 / 2014

Exploring Quantitative Structure-Activity Relationships (QSARs) of Cyclooxygenase-2 (COX-2) Inhibitors by MLR, PLS and PC-ANN.

Prepared By:

NaelYahya Mohammad Zatari

Bachelor of Pharmacy

Applied science university (Jordan)

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

1436/2014

Al-Quds University

Deanship of Graduate Studies

-Applied and Industrial Technology

-Faculty of science and Technology

Thesis Approval

Exploring Quantitative Structure-Activity Relationships (QSARs) of Cyclooxygenase-2 (COX-2) Inhibitors by MLR, PLS and PC-ANN.

Prepared by:

Student name: NaelYahya Mohammad Zatari

Registration number: 21110060

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: Signature:……..…....

3- External Examiner: Signature:……

Dedication:

To my parents and my wife for their love, support and encouragement to complete my thesis, and also to my brothers, sisters and to my friends for their generous help, assistance and back up.

Truly, this thesis would not be possible without them.

Nael Yahya Mohammad Zatari

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 (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Signed…….

Nael Yahya Mohammad Zatari

Date: ……..

Acknowledgements

I would like to thank Dr. Omar Deeb for his kind interest and for helping me to finish my thesis. I would also like to thank Al-Quds University, especially faculty of science and faculty of pharmacy and the teachers who open the way to do this research.

Abstract

Nonsteroidal anti-inflammatory drugs (NSAIDs) or Cyclooxygenase-2 enzyme inhibitors are a mainstay in the treatment of inflammatory disease and are among the most widely used drugs worldwide. In this thesis, Quantitative structure–activity relationship study using principal component artificial neural network (PC-ANN) methodology was performed to predict the inhibitory activities expressed as pIC_{50} of 121 cyclooxygenase-2 (COX-2) inhibitors. We divided these compounds to two parts according to chemical structure as tricyclics (part 1) which has 48 chemical compounds and non-tricyclics (part 2) that has 73 chemical compounds. The results for each part obtained by PC-ANN give advanced regression models with good prediction ability. Part 1: the two optimal artificial neural network models obtained have correlation coefficients of 0.937 and 0.924. The lowest prediction sum of squares (PRESS) value obtained for the prediction set is 3.947 which accounts for better predictability of the model. Part 2: the two optimal artificial neural network models obtained have correlation coefficients of 0.823 and 0.757. The lowest prediction sum of squares (PRESS) value obtained for the prediction set is 4.727. Artificial neural networks provide improved models for heterogeneous data sets. Both the external and cross-validation methods are used to validate the performances of the resulting models. Randomization test is employed to check the suitability of the models.

الملخص:

مثبطات إنزيم COX-2 الانتقائية والتي تنتمي إلى فصيلة الأدوية المسماة مضادات الالتهاب اللاستيروئيدية (NSAIDs) هي الأكثر استخداماً في العالم لعلاج أمراض متعددة منها التهاب المفصل الروماتويدي وتساعد أيضاً على تقليل الألم والتورم في المفاصل وفي العضلات، وفي دراستنا هذه ؛ جمعنا 121 مركب كيميائي يعمل كمثبط لهذا الإنزيم لكل مركب صيغة بنائية معروفة وتقدر فاعلية كل مركب على تثبيط الإنزيم بقيمة محسوبة مخبرياً ويعبر عنها ب (IC_{50}) . ومن ثم تم تقسيم هذه المركبات بالاعتماد على الصيغة البنائية لكل مركب إلى قسمين (مركبات ذات الحلقات الثلاثية ومركبات لا تمتلك حلقات ثلاثية).وقد قمنا في هذه الدراسة ببناء علاقات كمية خطية ما بين هذه المركبات وفاعليتها باستخدام (MLR) ، كما تم استخدام (PC-ANN) لبناء علاقات كمية غير خطية. إن النتائج التي تم الحصول عليها باستخدام العلاقات المختلفة التي قمنا بها توضح أن العلاقة الغير خطية قد أعطت نتائج أفضل من الخطية. حيث أن معامل الارتباط للعلاقات الأفضل التي تم الحصول عليها للقسم الأول من المركبات 0.937 و 0.924 مع أقل قيمة (PRESS) 3.947 . أما في القسم الثاني فان معامل الارتباط للعلاقات الأفضل التي تم الحصول عليها 0.823 و 0.757 مع أقل قيمة (PRESS) 4.727 .وبعد ذلك تم فحص هذه العلاقات بطريقة اختبار العشوائية وتبين لنا قدرة هذه العلاقات على التنبؤ بفاعلية مركبات أخرى لم تستخدم في بناء هذه العلاقة.

Table of Contents:-

List of tables:

List of figures:

List of software:

1) HyperChem (version 8.0 HyperChem, Inc.) will be used to perform geometry optimization.
(http://www.hyper.com/)

2) Dragon ((version 2.1, Todeschini, R., Milano Chemometrics and QSAR Group, will be used to calculate the different types of descriptors.
(http://www.talete.mi.it/products/dragon_description.htm/)

3) SPSS software (version 11.50, SPSS Inc.)
(http://www-01.ibm.com/software/analytics/spss/)

4) MATLAB (version 6.50, Mathworks Inc.).
(http://www.mathworks.com/products/matlab/)

List of abbreviations:

Abbreviation	Meaning
NSAIDs	Non steroidal antiinflamatory drugs
COX-2	Cyclooxygenase-2 enzyme
COXIBs	Selective Cyclooxygenase-2 inhibitors
PG_S	Prostaglandins
AA	Arachidonic acid
PGG2	Prostaglandins G2
PGHS	prostaglandin H2 synthase
TXA2	Thromboxane A2
WHIM	Weighted Holistic Invariant Molecular
BCUT	Burden eigenvalues
3D MoRSE	Molecular representation of structures based on electron diffraction
GETAWAY	Geometry, topology, and Atom-Weights Assembly
QSARs	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
PCR	Principle 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 errors
CoMFA	Comparative molecular field analysis
CoMSIA	Comparative molecular similarity indices analysis

- Chapter one

Introduction

1.1 Cyclooxygenase-2 Inhibitors

Nonsteroidal anti-inflammatory drugs (NSAIDs) or Cyclooxygenase enzyme inhibitors are a mainstay in the treatment of inflammatory disease and are among the most widely used drugs worldwide[1]. Vane in 1971 showed the pharmacological action of aspirin that is the first NSAID with therapeutic benefits[2] , which has now been used for more than 100 years as a NSAID.[1]

NSAIDs or Cyclooxygenase enzyme inhibitors are anti-inflammatory, antipyretic, and analgesic. They are prescribed as first choice for the treatment of rheumatic disorders as well as relieving the pains of everyday life.[1-3]

A large number of epidemiological studies have indicated that the use of NSAIDs may prevent or delay the clinical features of Alzheimer's disease and the users of (NSAIDs) could be of benefit against the development and growth of malignancies. In contrast, the main limitation in using NSAIDs consists in their side effects, including bronchospasm and gastrointestinal ulcerogenic activity including complications such as bleeding and perforation.[3-5]

Cyclooxygenase enzyme (COX enzyme) is available in two isoforms, COX-1 and COX-2; however the traditional non-steroidal anti-inflammatory drugs (NSAIDs) inhibit both enzymes. Whereas, a new class of COX-2 selective inhibitors (COXIBs) preferentially inhibit the COX-2 enzyme.[6]

Cyclooxygenase (COX) is the key enzyme required for the conversion of arachidonic acid to prostaglandins that are involved in physiological functions such as protection of the stomach mucosa, aggregation of platelets and regulation of kidney function. They also have pathological functions such as their involvement in inflammation, fever and pain.[7-8]

Cyclooxygenase (COX) or prostaglandin H2 synthase (PGHS) is the enzyme that catalyzes the first two steps in the biosynthesis of the prostaglandins (PGs) from the substrate arachidonic acid (AA). These are the oxidation of AA to the hydroperoxy endoperoxide PGG2 and its subsequent reduction to the hydroxyl endoperoxide PGH2. The PGH2 is transformed by a range of enzymes and nonenzymic mechanisms into the primary prostanoids, PGE2, PGF2α, PGD2, PGI2, and TXA2 .[9] As shown in the following figure (1.1).

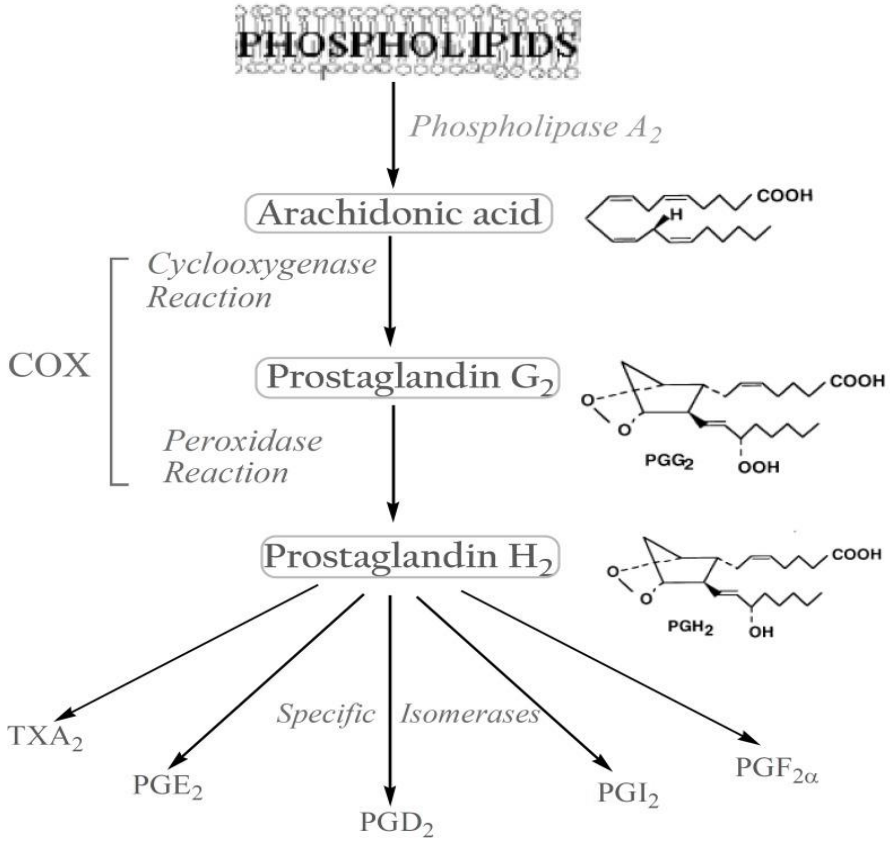


Figure (1.1): Schematic representation of cyclooxygenase pathway.

Inhibition of the PGHSs with NSAIDs acutely reduces inflammation, pain, and fever, and long-term use of these drugs reduces fatal thrombotic events, as well as the development of colon cancer and Alzheimer's disease. [10]

COX-1 is referred to as a constitutive isoform and is constitutively expressed throughout the gastrointestinal system, the kidneys, the vascular smooth muscle and platelets. COX-1 is presumably involved in the housekeeping functions of PGs, such as the cytoprotective effects in the gastric mucosa, the integrity of platelet function and the maintenance of renal perfusion. Conversely, COX-2 is undetectable in most tissues, but its expression can be induced by a variety of stimuli related to inflammatory response. COX-2 is, therefore, commonly referred to as the inducible COX isoform because, like other immediate-early genes, it can be rapidly up-regulated in during various conditions response to growth factors and cytokines.[11-12]

The differences between COX-1 and COX-2 represent in their chemical structure, three amino acid differences result in a larger (about 20%) and more accessible channel, in COX-2. The exchange of a valine at position of 523 in COX-2 for a relatively bulky isoleucine (Ile) residue in COX-1 at the same position of the active site of the enzyme causes a structural modification. This modification in the COX-2 enzyme allows the access to an additional side pocket, which is a pre-requisite for COX-2 drug selectivity. Access to this side pocket is

restricted in the case of COX-1. In addition, the exchange of Ile-434 for a valine in COX-2 allows a neighboring residue phenylalanine-518 (Phe-518) to swing out of the way, increasing further access to the side cavity. There is another essential amino acid difference between the two isoforms, which does not alter the shape of the drug-binding site but rather changes its chemical environment. Within the side pocket of COX-2 is an arginine in place of histidine-513 (His-513) in COX-1, which can interact with polar moieties. But also the cyclooxygenase active site is created by a long hydrophobic channel that is the site of non-steroidal anti-inflammatory drug binding. [13-14]

The undesirable side-effects of NSAIDs are thought to be due to the inhibition of COX-1 (constitutive isoform), whereas the beneficial effects are related to the inhibition of COX-2 (inducible isoform) but increasing selectivity for COX-2 also increased toxicity, since the anti-thrombotic prostacyclin is formed by COX-2 and inhibiting its synthesis precipitated heart attacks. The problem of this side action has not yet been resolved.[1, 15]

The discovery of cyclooxygenase-2 and the establishment of its structure led to the development of selective inhibitors of this enzyme, such as celecoxib and rofecoxib, with potent anti-inflammatory actions but with reduced gastrotoxic effects.[15]

Within the last two decades, the volume of literature on the structural types introduced as selective COX-2 inhibitors is enormous. In this review[13], they have chosen to focus on the structure activity relationship (SAR) and also various structural families of compounds, which have emerged within the last years. Contrary to the classic NSAIDs, this new class of enzyme inhibitors is lacking a carboxylic group, thus effecting COX-2 affinity by a different orientation within the enzyme without formation of a salt bridge in the hydrophobic channel of the enzyme.[13]

In general classification, selective COX-2 inhibitors belong to two major structural classes:

1) Tricyclics and 2) Non-tricyclics.

Tricyclics:

All of the compounds in this class possess 1,2-diarylsubstitution on a central hetero or carbocyclic ring system with a characteristic methanesulfonyl, sulfonamido, azido, methanesulfonamide or pharmacophore-based tetrazole group on one of the aryl rings that plays a crucial role on COX-2 selectivity. Coxibs such as Celecoxib, Rofecoxib, Valdecoxib and etc, belong to this common structural class.

Non-tricyclics “lack the cyclic central core”.

In addition to the classical tricyclic COX-2 inhibitors such as Coxib family, there are several non-classical structures which we here classify as non-tricyclics. These series of compounds lack the cyclic central core. Instead, they possess acyclic central systems such as olefinic, iminic, azo, acetylenic and α,β-unsaturated ketone structures. The central acyclic core may contain a two-membered or three-membered chain structure which is the basic point for sub classification of these compounds.[13]

1.2 COMPUTATIONAL CHEMISTRY

The term computational chemistry is generally used when a mathematical method is sufficiently well developed that it can be automated for implementation on a computer. Note that the words ``exact" and ``perfect" do not appear in these definitions. Very few aspects of chemistry can be computed exactly, but almost every aspect of chemistry has been described in a qualitative or approximately quantitative computational scheme. The challenge in computational chemistry is to simplify the calculation enough to be solvable, but still accurate enough to predict the desired physical quantity.[16]

Molecular descriptors have been applied extensively in, for example, bioinformatics, network biology structure-oriented drug design, medicinal chemistry, chemometrics, chemical graph theory, and mathematical chemistry. Also, their positive impact in quantitative structure–activity relationship/quantitative structure–property relationship (QSAR/QSPR) has been demonstrated and important subgroups of descriptors such as topological indices have been explored. [17]

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. [18]

Scientists mainly use three different methods to perform numerical computation:

1) Ab initio calculation:

“ab initio” is Latin for “from the beginning” We begin with fundamental physical properties, and we calculate how electrons and nuclei interact. Most often this requires solving approximations to the time-independent Schödinger equation. Occasionally we need to solve the time-dependent Schödinger equation. [16]

Schrödinger equation:

Given the importance of the ability to calculate the electronic structure of a molecule in computational chemistry, it is important to outline, albeit briefly, the underlying theory that both the commonly used semi-empirical and density functional methods attempt to solve. The mathematics is complex and will be kept to an absolute minimum, the aim being to set the scene concerning the various components that must be dealt with if quantum mechanics is to be utilized to help understand the electronic structure of chemicals. The subsequent sections dealing with the commonly used semi-empirical and density functional approaches will highlight how each of these methods approximates these important mathematical components.[19]

The starting point of any discussion into quantum mechanics is always the time-independent Schrödinger equation.

Schrödinger equation (1.1);

$$H\psi = E\psi \qquad (1.1)$$

Where H is the Hamiltonian operator,

E is the energy of the molecule and;

Ψ is the wave function which is a function of the position of the electrons and nuclei within the molecule.

An essential part of solving the Schrödinger equation is the Born–Oppenheimer approximation, where the coupling between the nuclei and electronic motion is neglected. This allows the electronic part to be solved with the nuclear positions as parameters, and the resulting potential energy surface (PES) forms the basis for solving the nuclear motion. The major computational effort is in solving the electronic Schrödinger equation for a given set of nuclear coordinates.[20]

2) Semi-empirical techniques:

Semiempirical quantum chemistry attempts to address two limitations, namely slow speed and low accuracy, of the Hartree-Fock (HF) calculation by omitting or parameterizing certain integrals based on experimental data, such as ionization energies of atoms, or dipole moments of molecules. As a result, semiempirical methods are very fast, applicable to large molecules, and may give accurate results when applied to molecules that are similar to the molecules used

for parameterization. On the downside, accuracy of semiempirical methods is erratic on many systems.

The success of semi-empirical methods relies on turning the remaining integrals into parameters, and fitting these to experimental data, especially molecular energies and geometries. Such methods are computationally much more efficient than the ab initio HF method, but are limited to systems for which parameters exist. [20-21]

Molecular mechanics, the molecular mechanics energy expression consists of a simple algebraic equation for the energy of a compound. It does not use a wave function or total electron density. The constants in this equation are obtained either from spectroscopic data or ab initio calculations. A set of equations with their associated constants is called a force field. The fundamental assumption of the molecular mechanics method is the transferability of parameters. In other words, the energy penalty associated with a particular molecular motion, say, the stretching of a carbon-carbon single bond, will be the same from one molecule to the next. This gives a very simple calculation that can be applied to very large molecular systems.[16]

1.3 Quantitative structure activity relationship (QSAR)

QSARs, or quantitative structure–property relationships (QSPRs), are mathematical models that attempt to relate the structure-derived features of a compound to its biological or physicochemical activity. This method has predictive and diagnostic abilities. They can be used to predict the biological activity (e.g., IC50) or class (e.g., inhibitor versus non inhibitors) of compounds before the actual biological testing. They can also be used in the analysis of structural characteristics that can give rise to the properties of interest.[22]

There are many practical purposes of a QSAR, includes the following:-

• To predict biological activity and physico-chemical properties by rational means.

• To comprehend and rationalize the mechanisms of action within a series of chemicals.

Underlying these aims, the reasons for wishing to develop these models include

1) Savings in the cost of product development (e.g. in the pharmaceutical, pesticide, personal products, etc. areas).

2) Predictions could reduce the requirement for lengthy and expensive animal tests.

3) Reduction (and even, in some cases, replacement) of animal tests, thus reducing animal use and obviously pain and discomfort to animals.

4) Other areas of promoting green and greener chemistry to increase efficiency and eliminate waste by not following leads unlikely to be successful. [23]

TECHNIQUES AND TOOLS OF QSAR:-

The process of QSAR model development can be generally divided into three stages: data preparation, data analysis, and model validation. The first stage includes selection of a molecular dataset for QSAR studies, calculation of molecular descriptors, and selection of a QSAR (statistical analysis and correlation) method.

Data preparation starts by selection of the data set to be used; this may simply be the extraction of data from a database or may need additional experimental studies. There are two steps to complete data preparation: geometry optimization and descriptors calculation. Geometry optimization or minimization is finding the coordinates that represents the potential energy minimum. Theoretical molecular descriptor is a value that describes the molecular structure numerically. These descriptors can be simple such as molecular weight or complex such as geometrical descriptors.

In data analysis, the first step is to decide which techniques for statistical analysis and correlation to be used. If our correlation models to be built are linear then we use multilinear regression (MLR) or non linear then we use artificial neural network (ANN). Model validation is the final part of the model development process, the predictive power of the model is tested on an independent set of compounds, generally predictive power is the most important characteristics of the model and model predictivity is the ability of the model to predict accurately the target activity of a compound that was not used for model development.

In model validation step, most of validation processes implement the leave one out (LOO) and leave many out (LMO) cross-validation procedures. The most common outcome parameters resulted from cross-validation procedures are cross-validated determination coefficient q2(R^2cv) and root mean squares error (RMSE). High R^2cv and low RMSE values is a result of good and more predictive model and that lead to better description of the observed data.

Finally and the most important advantage of QSAR is that we can use QSAR resultant models outside the range of the data set; the model can be used to design new drugs depending on the most effective descriptors.[24]

The QSAR models employ descriptors and statistical approaches to provide an estimation of the desired property and the molecular descriptors play a fundamental role in developing models for chemistry, pharmaceutical sciences, environmental protection policy, toxicology, health research, and quality control. Evidence of the interest of the scientific community in molecular descriptors is provided by the huge number of descriptors that have been proposed: more than 5000 descriptors, derived from different theories and approaches are defined and computable by using dedicated software of chemical structure. [25]

Examples of different types of descriptors:

Constitutional Descriptors: Molecular weight, Number of atoms of various elements, Number of bonds of various orders, Number of rings

Topological Descriptors: Weiner index, Randic indices, Kier and Hall indices, Information content, Connectivity index, Balaban index

Quantum Chemical Descriptors: highest occupied molecular orbital energy (HOMO) and lowest unoccupied molecular orbital energy (LUMO), reactivity indices, Refractivity, Total energy, Ionization potential, Electron affinity, Energy of protonation, Orbital populations, Frontier orbital densities.

Chemical Descriptors: Octanol-water partition coefficient (LogP), Surface area, refractivity, volume, polarizability.[16]

Developing QSAR models starts with the collection of data for the property of interest while taking into consideration the quality of the data. It is necessary to exclude low-quality data as they will lower the quality of the model. Following that, representation of the collected molecules is done through the use of features, namely molecular descriptors, which describe important information of the molecules. There are many types of molecular descriptors but not all will be useful for a particular modeling task. Thus, uninformative or redundant molecular descriptors should be removed before the modeling process. Subsequently, for tuning and validation of the QSAR model, the full data set is divided into a training set and a testing set prior to learning. Various modeling methods like multiple linear regression, logistic regression, and machine learning methods are used to build models that describe the empirical relationship between the structure and property of interest. The optimal model is obtained by searching for the optimal modeling parameters and feature subset simultaneously. This finalized model built from the optimal parameters will then undergo validation with a testing set to ensure that the model is appropriate and useful.[22]

In our lab, several QSAR studies have been applied for many topics that got good results over last years to predict compounds' properties, including biological activity, physical property and even toxicity.[26-30]

Statistical methods

Sometimes QSAR statistical methods are also classified into the following two categories, depending upon the type of correlation technique employed to establish a relationship between structural properties and biological activity:

•Linear methods including linear regression (LR), multiple linear regression (MLR), partial least-squares (PLS), and principal component analysis/regression (PCA/ PCR).

•Non-linear methods consisting of artificial neural networks (ANN), k-nearest neighbors (kNN), and Bayesian neural nets. [31]

1.4.1 Multiple Linear Regressions

Multiple linear regressions (MLR) is a method used to model the relationship between two or more explanatory variables (x_1, x2,….,x_p) and a response variable "dependent variable" (Y), and the relationship between them is represented by the following equation (1.2):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i \qquad (1.2)$$

Where:

β_0 is the constant term and β_1 to β_p are the coefficients relating the independent variables (descriptors) to the variable of interest.

$e_{i:}$ is an error term.

Multiple linear regression finds a correlation between molecular structures and their corresponding property (biological activity) through a linear combination of structural descriptors, and the quality of the obtained model is estimated by the correlation coefficient R between the observed values of the investigated property (y) and those predicted by Equation (1.2), and the important thing; the molecular descriptors in the model should be independent of each other and the number of instances for model building should be at least five times the number of descriptors used. [32-34]

In a regression analysis we study the relationship, called the regression function, between one variable (predicted values) y, called the dependent variable, and several others xi (descriptors),

called the independent variables then the use and interpretation of multiple regression models often depend on the estimates of individual regression coefficient. The predictor variables in a regression model are considered orthogonal when they are not linearly related. But, when the regressors are nearly perfectly related, the regression coefficients tend to be unstable and the inferences based on the regression model can be misleading and erroneous, this condition is known as multicollinearity.[34-36]

Construction of a model that describes the relationship of the highly correlated X-data with a property y (biological activity) is then problematic when applying the classic multiple linear regression (MLR) approach, since the regression coefficients cannot be calculated. A possible remedy for this problem is to select several orthogonal variables either using some preliminary knowledge or using a variable selection scheme. Another more general and efficient strategy to deal with the multicollinearity in X-data is to obtain a few orthogonal variables that describe the covariance between X-data and y. The partial least squares (PLS) regression has proved to be a successful tool for this purpose.[37]

In fitting an MLR model, the goal is to find the "best" estimates of the coefficients (β) that minimize the differences between all of the observed responses (y_i) and the corresponding model prediction ($\hat{y}_i$). In the same manner as for simple linear regression, the coefficients are found by minimizing the sum of the squares of the errors (ε_i) (that is, minimize $\sum \varepsilon_i = \sum (y_i - \hat{y}_i)^2$) in a least squares regression analysis.[38]

1.4.2 Partial least squares

Partial least squares (PLS) is a method for constructing predictive models when the factors are many and highly collinear.

In principle, MLR can be used with very 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.[39]

The PLS regression model may be written as the following equation (1.3):

$$Y = b_0 + \sum_{i=1}^{N} b_i x_i, \qquad (1.3)$$

Where Y is an appropriate activity, bi are the PLS regression coefficients, xi is the ith descriptor value, and N is the total number of descriptors. This is not apparently different from MLR, except that the values of the coefficients b are calculated using PLS. However, the assumptions underlying PLS are radically different from those of MLR. In PLS one assumes the x-variables to be collinear and PLS estimates the covariance structure in terms of a limited number of weights and loadings. In this way, PLS can analyze any number of x-variables relating to the number of objects (N). [40]

A partial least squares (PLS) algorithm is used for this type of fitting. This method starts with matrices of field data and activity data. These matrices are then used to derive two new matrices containing a description of the system and the residual noise in the data. Earlier studies used a similar technique, called principal component analysis (PCA). PLS is generally considered to be superior.[16]

PLS has been applied on various QSAR studies like inhibitors of vascular endothelial growth factor receptor-2 (VEGFR-2) tyrosine kinase.[30]

1.4.3 Artificial Neural Networks (ANNs)

The Artificial Neural Networks (ANNs) are a type of mathematical model that simulates the biological nervous system and draws on analogues of adaptive biological neurons. A biological neuron receives inputs from many external resources, combines them, performs a non-linear operation, and then makes a decision based on the final results.

ANNs are known to be a powerful tool to simulate various non linear systems and have been applied to numerous problems of considerable complexity in many field including pharmaceutical research, engineering, psychology and medicinal chemistry; hence ANNs have been shown to be an effective tool to establish this type of relationship and predict the activities of new compounds. In addition, ANNs are certainly very useful in the preformulation design and would help reduce the cost and length of preformulation study.[41]

Artificial neural network consists of input layer, one or more hidden layers and one output layer; the input layer provides data from the external source. The mapping of the input data occurs by neural network hidden layers, and then the final representative signal is generated

by the output layer. The ability of neural networks to classify information depends on hidden layers, which are fully connected by the synapses to the neighboring layers. In each hidden layer and output layer, the processing sums up its input from previous layer by the sigmoidal function to compute the output to the following layer. As shown in the figure (1.2)

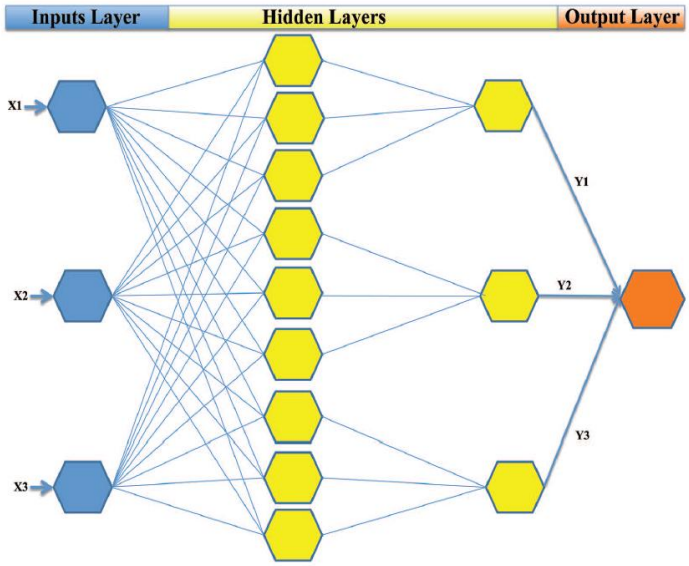


Figure (1.2): A schematic of four layered artificial network. Input layer units (in blue) receive input signals (X_1, X_2, X_3) and transfer the signal to the hidden layers. Output layer receive the signals provides the representative output signal.

The ANN learns an approximate nonlinear relationship by a procedure called training, which is the search process for the optimized set of weight values to minimize the squared error between the estimation and experimental data of units in the output layer. Most commonly used methods is back-propagation method, which requires three simple steps—network design, learning or training, and usage. In the network design stage the number of connections and layers is selected based on the type of application. Then, the training stage requires of selection of training set of data and remodeling of the network to minimize the error. And lastly, following the training ANN is suitable to use.

Number of hidden layers is essential to the purpose and function of an ANN as it influences the number of connections in the network and, thus, its performance. A very common approach to select the optimal number of hidden nodes is by trial and error method.[42]

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.

Principal component analysis (PCA) and more specifically factor analysis (FA) 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.

PCA was used to classify the molecules into training, validation and prediction sets. 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.[43]

1.5 QSAR modeling software:

There are many commercial or free software available for QSAR development. These include specialized software for drawing chemical structures, generating 2D structures, calculating chemical descriptors, developing QSAR models, and general-purpose software that have all the necessary components for QSAR development.

1.5.1 HyperChem (http://www.hyper.com/)

HyperChem is a sophisticated molecular modeling program and simulation program that is known for its quality, flexibility, and ease of use. And it offers many types of molecular and quantum mechanics calculations. Furthermore, HyperChem calculates some of QSAR descriptors, some of structural properties, studying dynamic behavior, etc.

HyperChem lets you build and display molecules easily. Since HyperChem contains a graphical interface, you can monitor the construction of molecules. And by using the Drawing tool, you can draw a two-dimensional (2D) representation of a molecule, and then use the Model Builder to generate a three-dimensional (3D) structure. HyperChem tool icons are shown in the following figure (1.3):

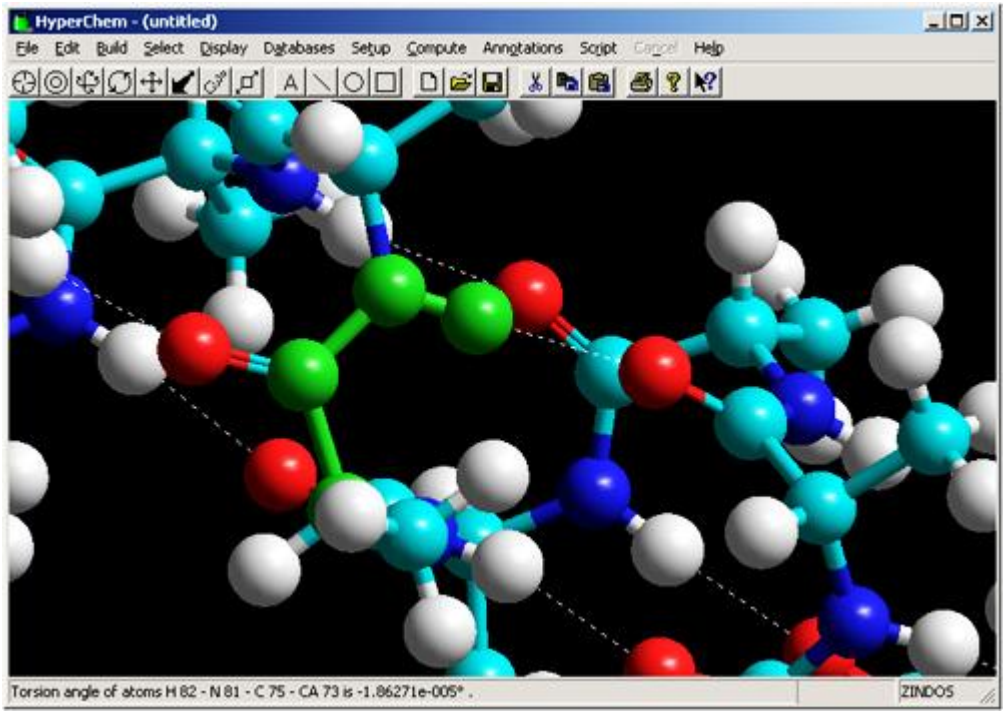


Figure (1.3): HyperChem window and tool icons

To calculate the properties of a molecule, you need to generate a well-defined structure. A calculation often requires a structure that represents a minimum on a potential energy surface. HyperChem contains several geometry optimizers to do this. You can then calculate single point properties of a molecule or use the optimized structure as a starting point for subsequent calculations.

1.5.2-Dragon (http://www.talete.mi.it/products/dragon_description.htm)

Dragon is commercial software for the computation of molecular descriptors. Dragon version 5.5 can compute 3224 molecular descriptors which are divided into 22 blocks. These blocks include constitutional or topological descriptors, walk and path counts, connectivity or information indices, 2D autocorrelations, BCUT descriptors, topological charges indices, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, functional group counts, 2D frequency fingerprints and so on. Dragon can work in both Windows and Linux, and it also has simple functions for conducting preliminary graphical and statistical analysis of descriptors, for example, histograms, Pareto plots, and 2D and 3D scatter plots.

1.5.3 SPSS software (http://www-01.ibm.com/software/analytics/spss/)

The software name stands for Statistical Package for the Social Sciences (SPSS), is software for managing data and calculating a wide variety of statistics. But in our study we use SPSS software to perform multiple linear regression analysis.

SPSS consists of two windows: Data Editor and Data Views

1) Data editor has two views: "Data View" and "Variable View". The Data Editor window (As shown in figure (1.4)) has two views that can be selected from the lower left hand side of the screen. Data View is where you see the data you are using. Variable View is where you can specify the format of your data when you are creating a file or where you can check the format of a pre-existing file. The data in the Data Editor is saved in a file as shown in the following figure:

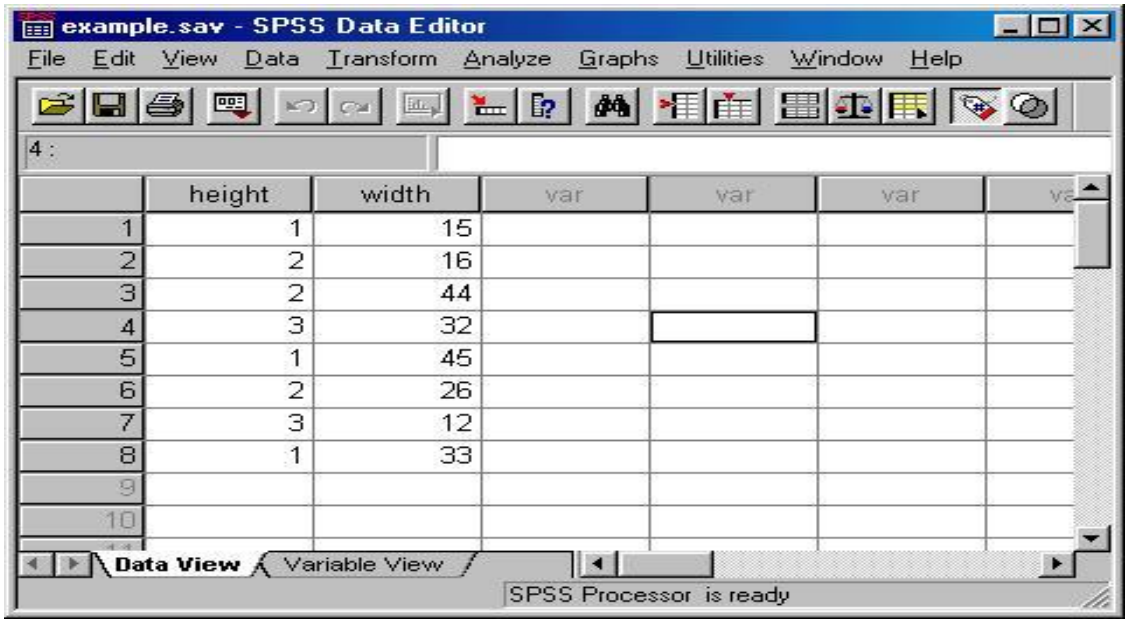


Figure (1.4): Data editor window.

2) The Output Viewer (As shown in figure (1.5)) collects your statistical tables and graphs, and gives you the opportunity to edit them before you save or print them. The Output Viewer is divided into two main sections, an outline pane on the left, and a tables pane on the right. When you print your output, it is the tables pane that is printed. As shown in the following figure:

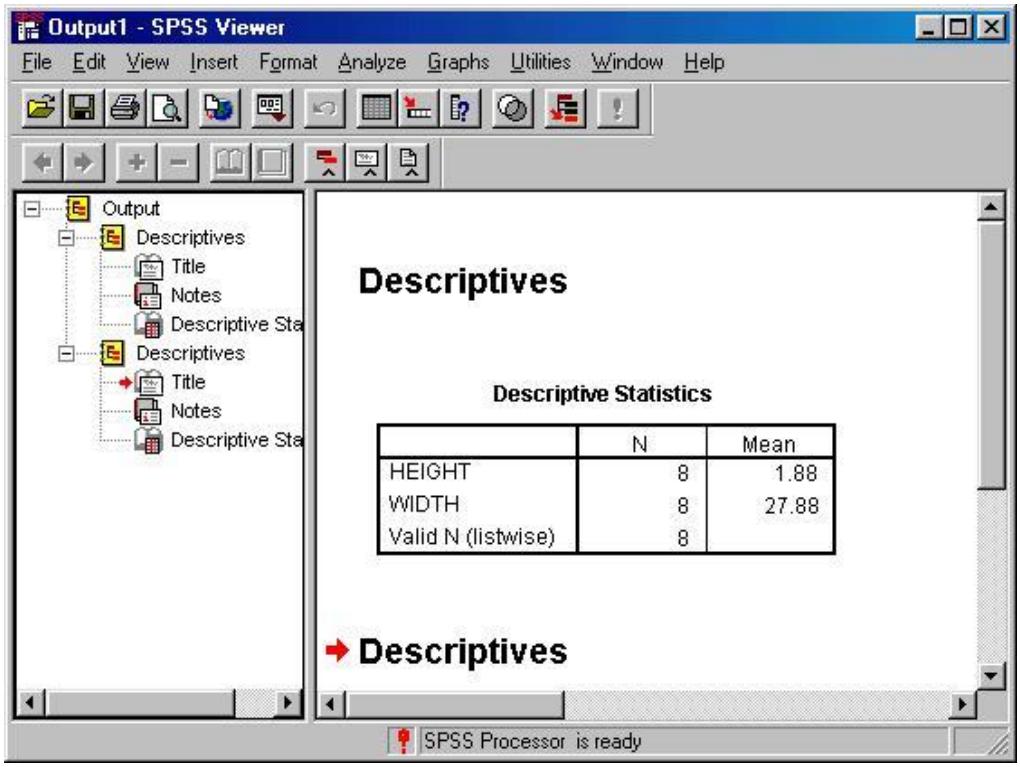


Figure (1.5): Output viewer window.

1.5.4-MATLAB (http://www.mathworks.com/products/matlab/)

MATLAB is commercial software that provides an interactive system for algorithm development, data visualization, data analysis, and numeric computation with wide application in image processing, financial analysis, computational biology, and so on. Data can be analyzed easily with ready-to-use functions, but users are also allowed to customize some of these tools or add their own algorithms for use. It also has functions to integrate MATLAB-based algorithms with external applications and languages such as Microsoft Excel, Java, and C++ .This enables developed QSAR models to be easily distributed as stand-alone programs or software modules.

1.6 Objective

The objective of this study is to develop QSAR models for inhibition activity of 121 chemical compounds of cyclooxygenase-2 inhibitors, we divided 121 chemical compounds into two major structural parts: Tricyclics with 48 chemical compounds and Non-tricyclics with 73 chemical compounds, by applying different statistical methods such as MLR, PLS and PC-ANN these models will be used to design new COX-2 inhibitors.

-Chapter two:

Methodology

The objective of this study was to build QSAR model that can predict the activity of new chemical compound as COX-2 inhibitor, using several statistical methods performed by several types of software.

This QSAR model is achieved by the following steps:-

- Data preparation
- Extracting the descriptors
- Choosing the informative descriptors
- Modeling the descriptors

2.1 Data preparation:

2.1.1 Data selection

A data set of 121 chemical structures of Cyclooxygenase-2 inhibitors and their biological activity (pIC50) were obtained from the literature (44-57), and we divided this data set to two classes according to chemical structure:

- Tricyclic chemical structures (Part 1) from the literature (44-49): These compounds have a tricyclic structure with variety of cores as hetero or carboxylic ring system or olefin.
- Non-tricyclic (Part 2) from the literature (50-57): These compounds lack the cyclic central core, and have monocyclic structures or bicyclic structures.

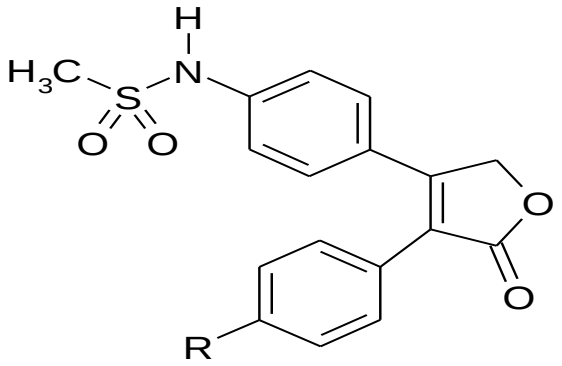
The chemical structures and the biological activities of each class (Tricyclics and Non-tricyclics) are summarized in tables (2.1) and (2.2), respectively.

The biological activity of each compound is expressed as **pIC_{50}**, it's the negative logarithm of the IC50 value in molar, which means negative logarithm of the concentration of drug required for half-maximal inhibition of COX-2 enzyme inhibitors, and it is used to measure the effectiveness of a substance in inhibiting cyclooxygenase-2 enzymes.

This is the formula for nanomolar conversion of IC50 values to pIC50 values

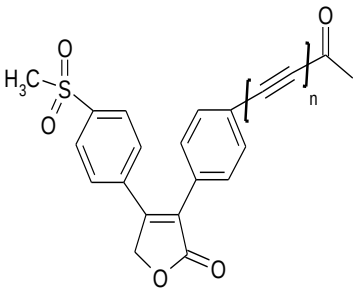
$$pIC50 = -\log (IC50*10^{-9})$$

Table (2.1): Molecular structures and observed inhibitory activities of the 48 COX-2 inhibitors expressed as pIC_{50}. (Part 1)

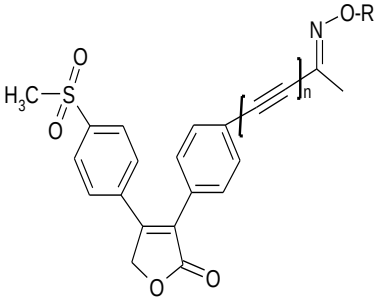


Compound number	Index *	R 1	pIC_{50}
1[P]	11a	H	5.770
2[V]	11b	F	5.495
3[V]	11c	Cl	6.097
4[V]	11d	Br	5.495
5[P]	11e	Me	6.046
6[V]	11f	OMe	4.502

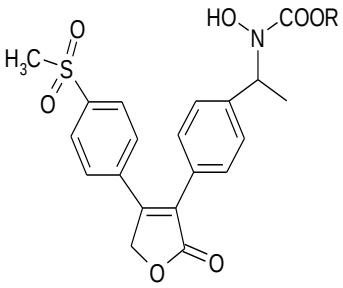
*Ref 44



15 , 16



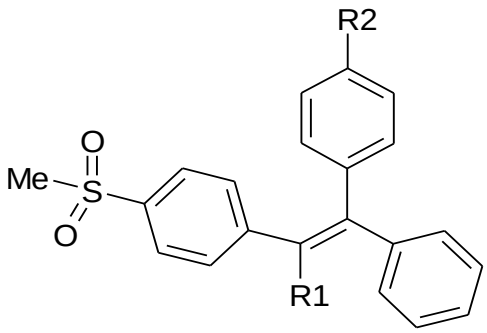
17 , 18



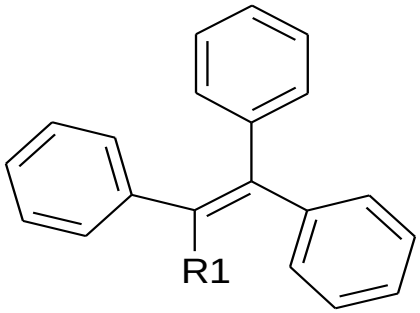
20

Compound number	**Index ***	**n**	**R**	**pIC_{50}**
7^{C}	15	0	---	5.921
8^{C}	16	1	---	5.456
9^{P}	17a	0	H	5.854
10^{C}	17b	0	CH_3	5.538
11^{V}	18a	1	H	5.569
12^{C}	18b	1	CH_3	4.955
13^{C}	20a	---	$CH(Me)_2$	5.180
14^{C}	20c	---	Ph	5.959
15^{C}	20d	---	CH_2Ph	4.921

*Ref 45



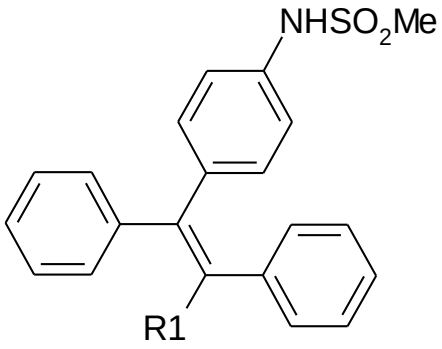
2 a-g or 12a-b or 13a-e



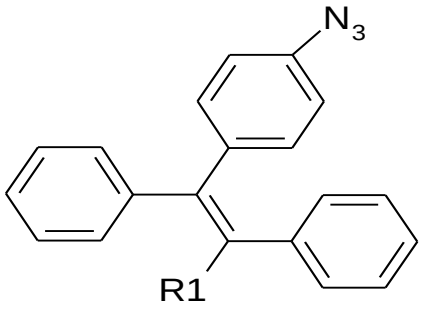
14

Compound number	**Index ***	**R 1**	**R2**	**pIC_{50}**
16^{C}	9a	Me	H	6.201
17^{V}	9b	Et	H	5.921
18^{P}	9c	n-C_4H_9	H	7.854
19^{P}	9d	n-C_6H_{13}	H	7.523
20^{C}	9e	n-C_7H_{15}	H	6.824
21^{C}	9f	n-C_9H1_9	H	5.959
22^{C}	12a	H	H	5.745
23^{P}	(Z)-12b	Et	OH	5.721
24^{C}	(Z)-13b	Et	OAc	7.523
25^{P}	(Z)-13c	n-C_4H_9	OAc	4.500
26^{C}	(Z)-13d	n-C_7H_{15}	OAc	5.102
27^{C}	14	n-C_4H_9	----	5.102

*Ref 46



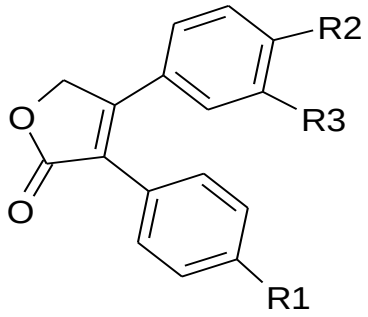
8



9

Compound number	**Index ***	**R 1**	**pIC_{50}**
28[C]	(Z)-8a	Me	4.480
29[V]	(Z)-8b	Et	5.745
30[P]	(Z)-8c	n-Butyl	6.495
31[C]	(Z)-8d	n-Hexyl	7.523
32[C]	(Z)-8f	n-Octyl	5.222
33[C]	(Z)-9a	Et	6.553
34[C]	(Z)-9b	n-Butyl	6.000
35[C]	(Z)-9c	n-Hexyl	6.959
36[C]	(Z)-9d	n-Octyl	7.854

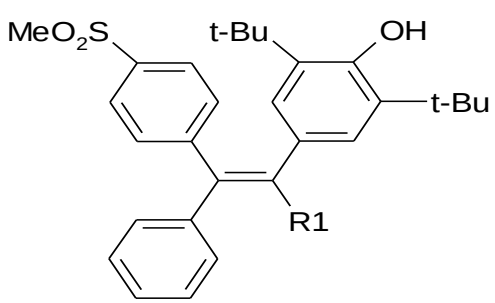
*Ref 47



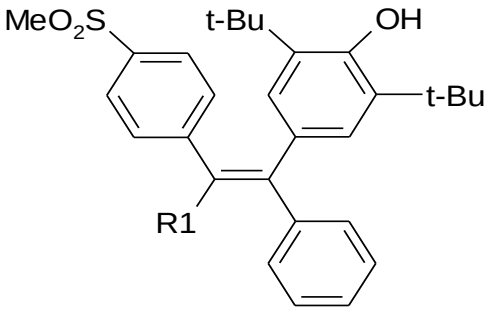
8a-e

Compound number	**Index ***	**R1**	**R2**	**R3**	**pIC_{50}**
37[V]	8a	H	$SO_2NHCOMe$	H	6.495
38[P]	8b	F	$SO_2NHCOMe$	H	5.987
39[C]	8c	Cl	$SO_2NHCOMe$	H	5.349
40[C]	8d	Me	H	$SO_2NHCOMe$	5.314
41[C]	8e	OMe	H	$SO_2NHCOMe$	5.001

* Ref 48



6



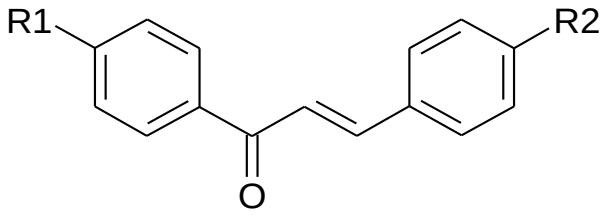
9

Compound Number	**Index ***	**R1**	**pIC_{50}**
42[C]	6a	Me	5.444
43[C]	6b	Et	4.495
44[C]	6c	n-Butyl	5.481
45[C]	6d	n-heptyl	6.000
46[V]	9b	Et	5.752
47[C]	9c	n-Butyl	6.444
48[C]	9d	n-Heptyl	7.000

- Ref 49

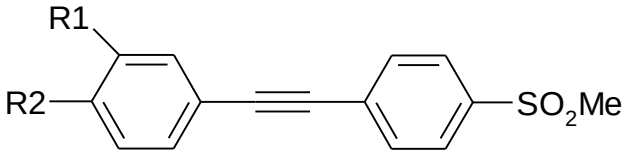
C Compounds classified in the training or calibration set, **P** compounds classified in the external test set (prediction set), **V** compounds classified in the validation set.

Table (2.2) Molecular structures and observed inhibitory activities of the 73 COX-2 inhibitors expressed as pIC_{50}. (Part 2)

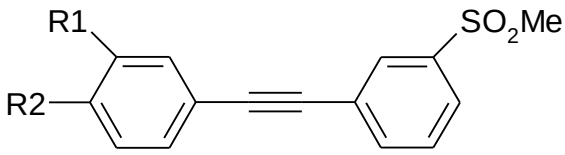


Compound number	Index *	R1	R2	pIC_{50}
1^{C}	9a	H	SO_2Me	6.097
2^{C}	9b	Me	SO_2Me	6.523
3^{V}	9c	F	SO_2Me	5.000
4^{C}	9d	OMe	SO_2Me	5.310
5^{C}	9e	SO_2Me	H	6.000
6^{C}	9f	SO_2Me	Me	6.523
7^{C}	9g	SO_2Me	F	6.222
8^{C}	9h	SO_2Me	OMe	5.495

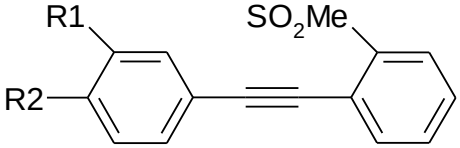
- **Ref 50**



11 a-f



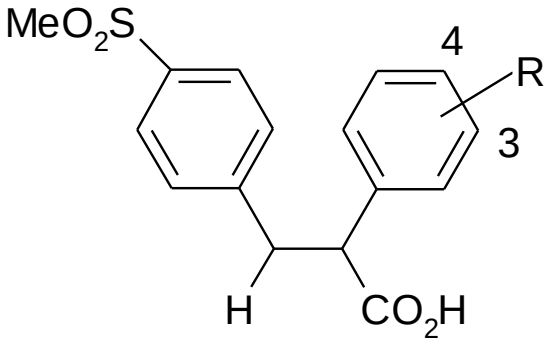
12 a-f



13 a-f

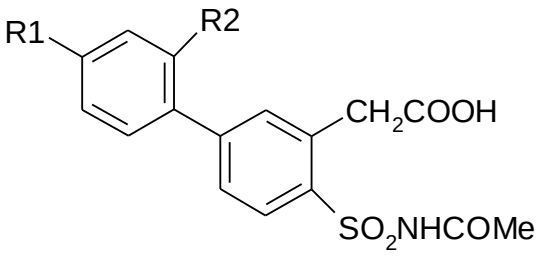
Compound number	**Index ***	**R1**	**R2**	**pIC$_{50}$**
9[V]	11a	H	H	6.051
10[V]	11b	F	H	5.222
11[P]	11d	H	Me	6.495
12[V]	11e	OH	H	6.678
13[C]	11f	OAc	H	7.222
14[C]	12a	H	H	5.495
15[C]	12b	F	H	5.721
16[C]	12c	OMe	H	5.469
17[C]	12d	H	Me	6.495
18[V]	12e	OH	H	6.495
19[P]	12f	OAc	H	7.301
20[C]	13b	F	H	6.495
21[P]	13c	OMe	H	5.319
22[V]	13d	H	Me	4.500
23[C]	13e	OH	H	5.456
24[C]	13f	OAc	H	6.854

***Ref 51**

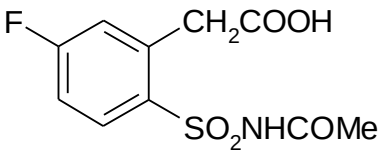


Compound number	**Index ***	**R**	**pIC_{50}**
25^{P}	9a	4-H	5.523
26^{V}	9b	4-Br	5.444
27^{P}	9c	4-F	4.444
28^{V}	9d	4-OH	5.276
29^{C}	9e	4-OMe	5.721
30^{C}	9f	4-OAc	5.538
31^{C}	9g	4-NHAc	5.602
32^{C}	9h	3-Br	6.509

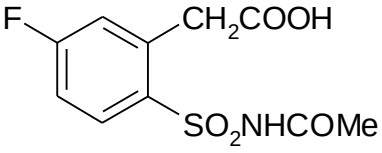
***Ref 52**



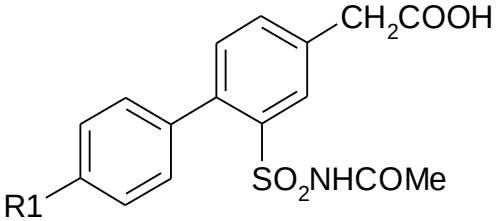
17 a-d, 19



12



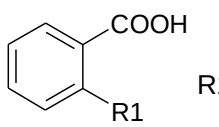
12



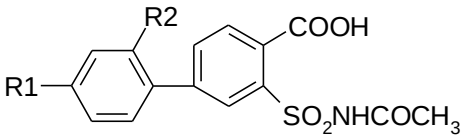
20a-c, e

Compound number	**Index ***	**R 1**	**R2**	**pIC_{50}**
33^{C}	12	---	----	6.010
34^{C}	14	---	---	7.929
35^{C}	17a	H	H	6.081
36^{C}	17c	F	F	6.000
37^{Out}	17d	$OCH(CH_3)_2$	H	5.500
38^{C}	19	SO_2CH_3	H	4.502
39^{C}	20a	H	---	5.754
40^{V}	20b	F	---	5.818
41^{Out}	20c	$OCH(CH_3)_2$	---	6.824
42^{C}	20e	SO_2CH_3	---	5.943

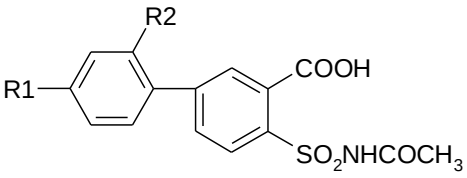
***Ref 53**



11



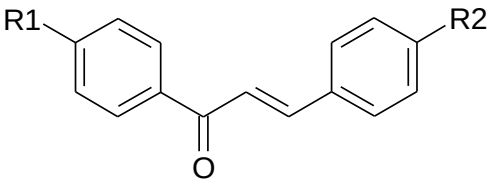
20 a-d



19 a,c

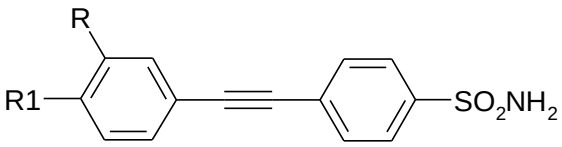
Compound number	**Index ***	**R 1**	**R2**	**R3**	**pIC_{50}**
43^{V}	11	$SO_2NHCOCH_3$	----	----	6.602
44^{C}	19a	H	H	----	7.523
45^{V}	19c	F	F	----	7.060
46^{P}	20a	H	H	H	5.921
47^{C}	20b	F	H	H	5.420
48^{P}	20c	F	F	H	6.114
49^{C}	20d	SO_2CH_3	H	H	6.523

***Ref 54**

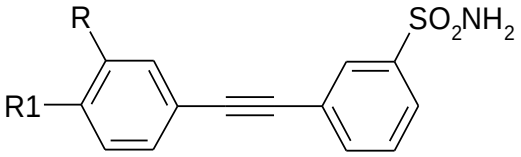


Compound number	**Index ***	**R 1**	**R2**	**pIC_{50}**
50^{P}	7a	$NHSO_2Me$	H	6.495
51^{C}	7b	$NHSO_2Me$	Me	6.000
52^{C}	7d	$NHSO_2Me$	O Me	5.000

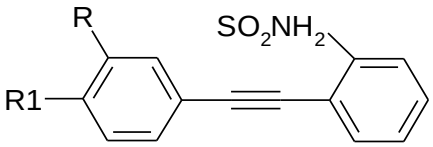
- **Ref 55**



9 c-d



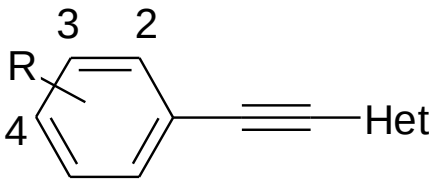
10 a-d



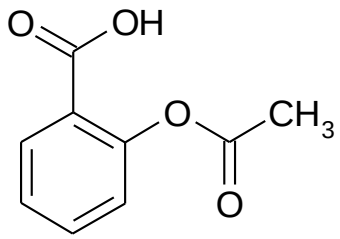
11 b

Compound number	**Index ***	**R**	**R1**	**pIC$_{50}$**
53[V]	9c	OMe	H	5.167
54[P]	9d	OH	H	4.487
55[P]	9e	F	H	6.523
56[V]	10a	H	H	6.347
57[V]	10b	H	Me	5.495
58[P]	10c	OMe	H	5.699
59[C]	10d	F	H	5.167
60[P]	11b	H	Me	4.495

***Ref 56**



20-31



Aspirin

Compound Number	**Index ***	**R**	**Het**	**pIC_{50}**
61[C]	20	4-SO_2NH_2	2-Pyridyl	6.523
62[C]	21	4-SO_2NH_2	4-Pyridyl	4.572
63[P]	22	4-SO_2NH_2	3-Me-2-Pyridyl	7.155
64[C]	23	2-SO_2CH_3	2-Pyridyl	6.678
65[C]	24	2-SO_2CH_3	3-Pyridyl	4.500
66[P]	25	2-SO_2CH_3	4-Pyridyl	6.959
67[C]	26	2-SO_2CH_3	3-Me-2-Pyridyl	6.377
68[C]	27	3-SO_2CH_3	2-Pyridyl	6.699
69[C]	28	3-SO_2CH_3	3-Pyridyl	4.496
70[C]	29	3-SO_2CH_3	4-Pyridyl	6.495
71[C]	30	4-SO_2CH_3	2-Pyridyl	6.481
72[C]	31	4-SO_2CH_3	3-Pyridyl	7.398
73[C]	(Aspirin)		---	5.620

***Ref 57**

[C] Compounds classified in the training or calibration set, **[P]** compounds classified in the external test set (prediction set), **[V]** compounds classified in the validation set and [out] compounds classified as outliers.

2.1.1 Structure drawing and optimization

All chemical structures in the last tables were drawn by HyperChem software for 121 compounds which were taken from the (44-57).

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

1) After drawing the structure using the drawing tools, we choose "add H and model building" from build menu to have 3D structure. as shown in following figures (2.1).

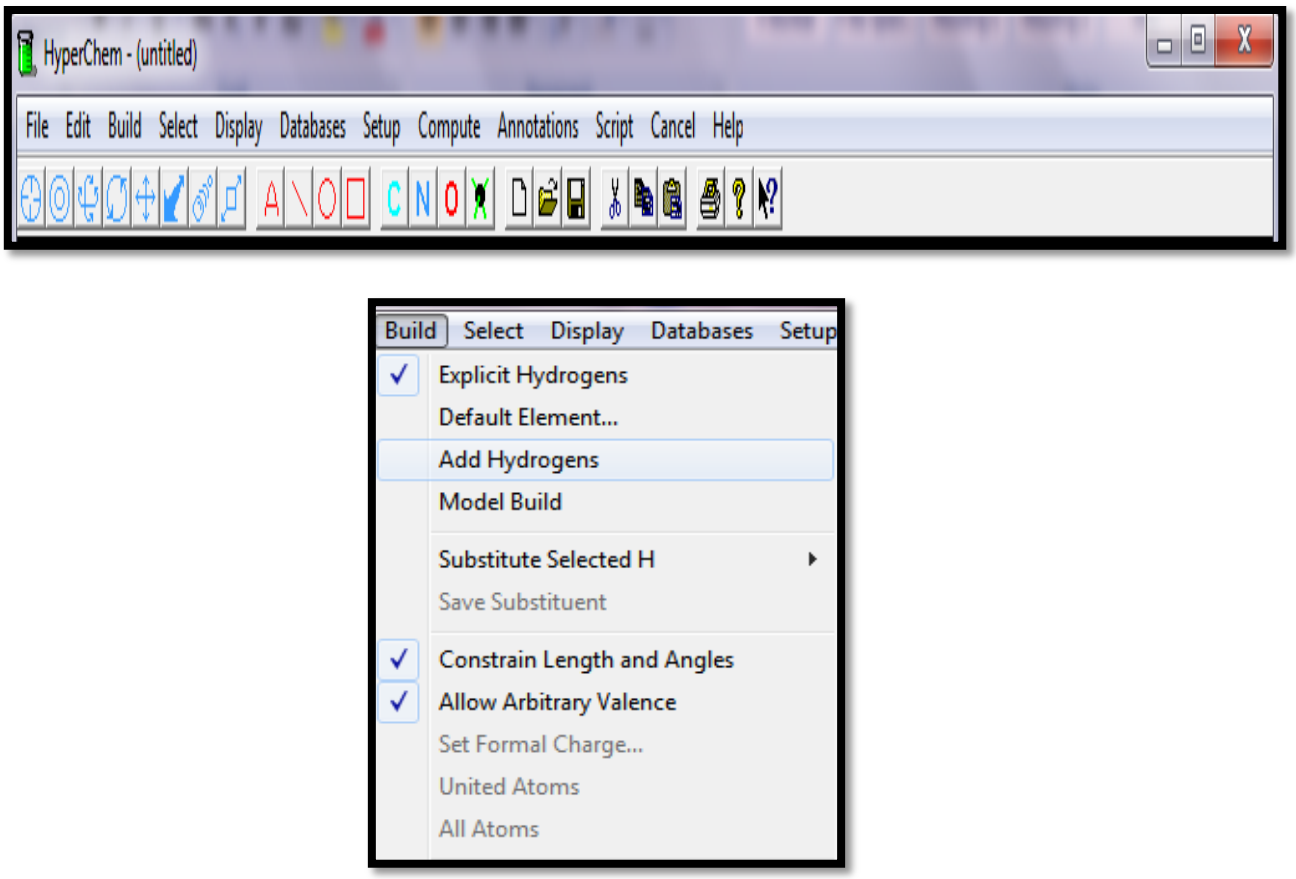


Figure (2.1): menu bar and build menu

2) Then Click on "start log" in the file menu to give it a name, and choose a directory to save it. as shown in following figure.

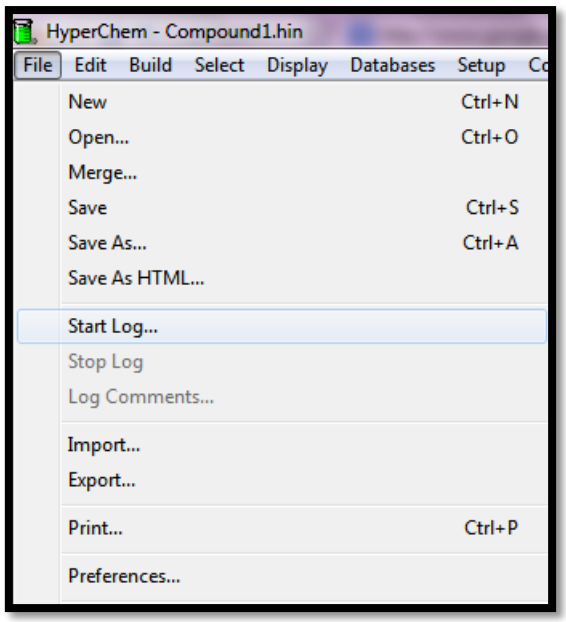


Figure (2.2) file menu

3) From the setup menu choose "semi-empirical" method of calculation and then select "AM1" from the semi-empirical window as shown in following figures.

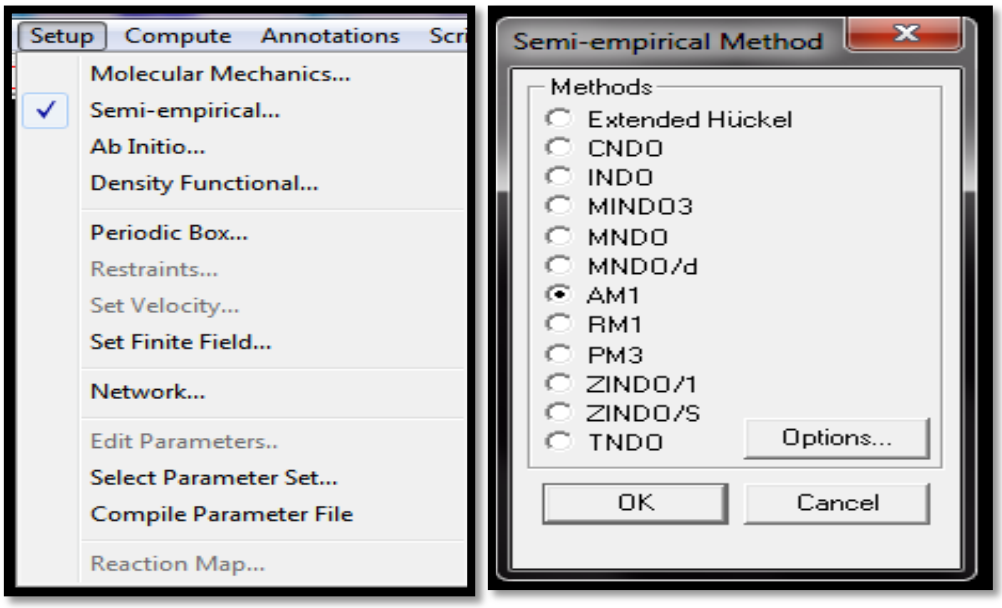


Figure (2.3): setup menu and semi-empirical method

4) Click on the option 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 select accelerate convergence. As shown in following figures (2.4):

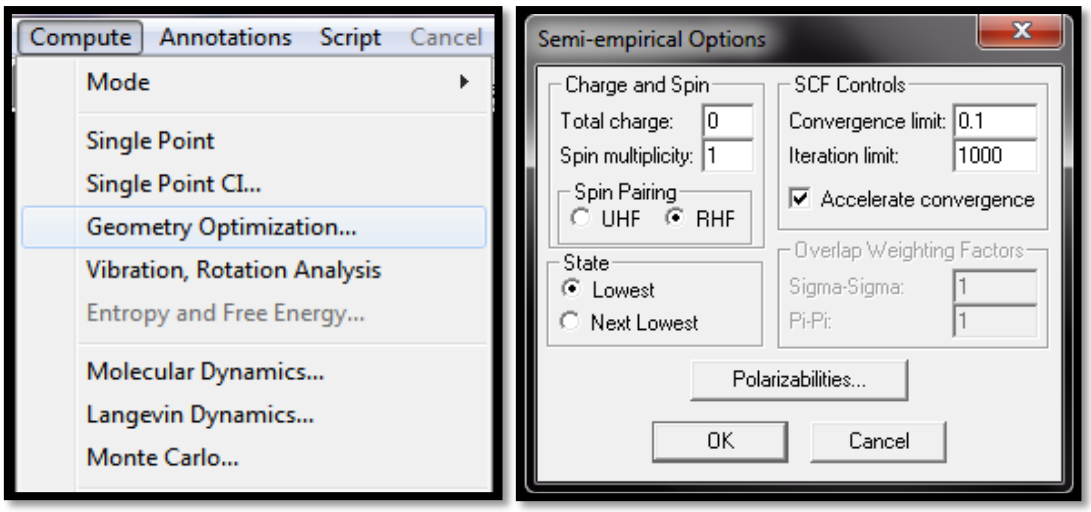


Figure (2.4): compute menu and semi-empirical options

5) Click OK to close the semi-empirical options dialog box, and then click OK to close the semi-empirical method dialog box.
6) From compute menu, Choose "geometry optimization", the semi-empirical optimization dialog box will open as shown in following figure.

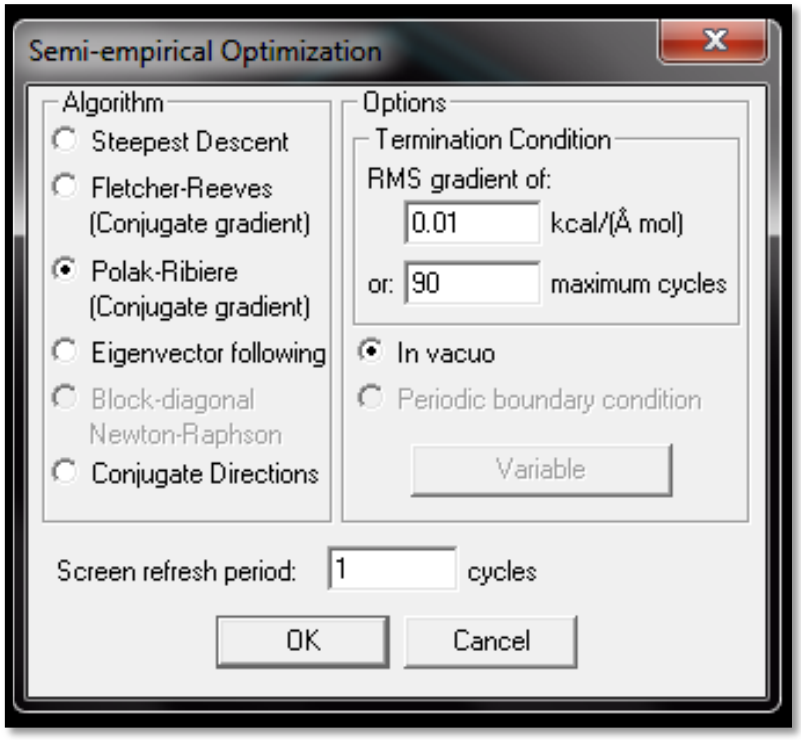


Figure (2.5): semi-empirical optimization.

7) Select Polak-Ribiere as algorithm method, then 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. We can increase the maximum cycles 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 as shown in following figure.

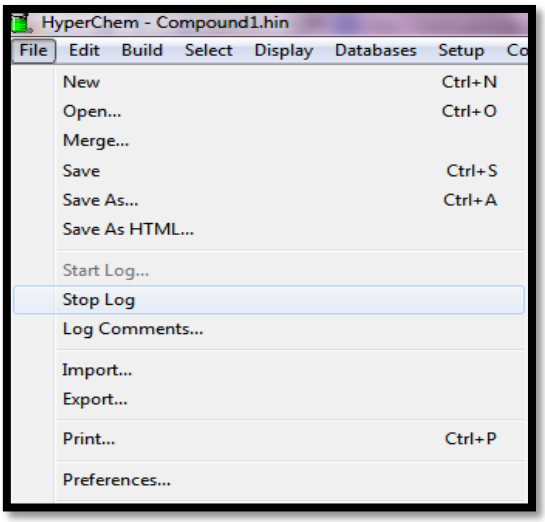


Figure (2.6): file menu

After finishing from these steps, we had 121 HIN files that represent the optimized chemical structures of the compounds, which is the input of Dragon software. And 121 log files that represent the calculation output divided to two parts: part 1 and part 2 with 48 log files and 73 log files, respectively.

2.1.3 Descriptor extraction

Sixteen groups of descriptors were calculated directly and indirectly by using HyperChem and Dragon software.

2.1.3.1 Descriptors calculated by HyperChem

We calculated one group of descriptors called quantum descriptors from the output file (Log file) of HyperChem calculation to obtain the following descriptors:

- Highest occupied molecular orbital energy (E_{HOMO}) and Lowest unoccupied molecular orbital energy (L_{UMO}), Molecular dipole moment (DM), and Heat of formation.

And we calculated other descriptors by HyperChem using the optimized HIN file of chemical structure, then we select the QSAR properties from compute menu after that we can calculate the following descriptors:

- Surface area (Approx), Surface Area (Grid), Volume, Mass, Hydration Energy, Octanol-Water partition coefficient (Log P), Refractivity and Polarizability.

2.1.3.2 Descriptors calculated manually

Four quantum descriptors were calculated manually by using excel software according to the equation below:

$$\text{Electronigativity (x)} = -\frac{(E_{HOMO} + E_{LUMO})}{2}$$

$$\text{Hardness } (\eta) = \frac{(E_{LUMO} - E_{HOMO})}{2}$$

$$\text{Softness (S)} = \frac{1}{\eta}$$

$$\text{Electrophilicity}(\omega) = \frac{x^2}{2\eta}$$

2.1.3.3 Descriptors calculated by Dragon software.

Fifteen groups of descriptors were calculated by Dragon software by applying the following steps:

1) All HyperChem outputs (HIN files) of each compound will be used in Dragon software.
2) Open Dragon software, the program will open a window as in the following figure.(2.7):

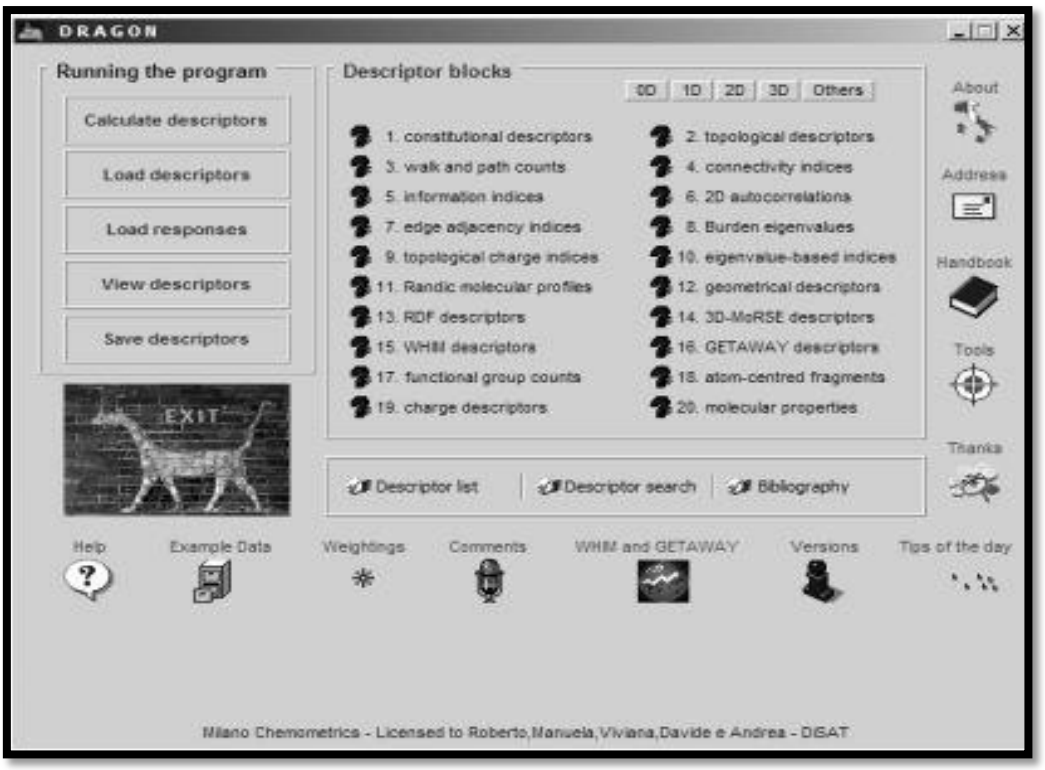


Figure (2.7): Dragon software window.

3) Choose Calculate descriptors
4) Then open the compounds folder, and select all the HIN files.
5) Click on descriptor selection button
6) Now open the descriptors group/ groups " we calculated each group of descriptors alone for all the compounds"
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.

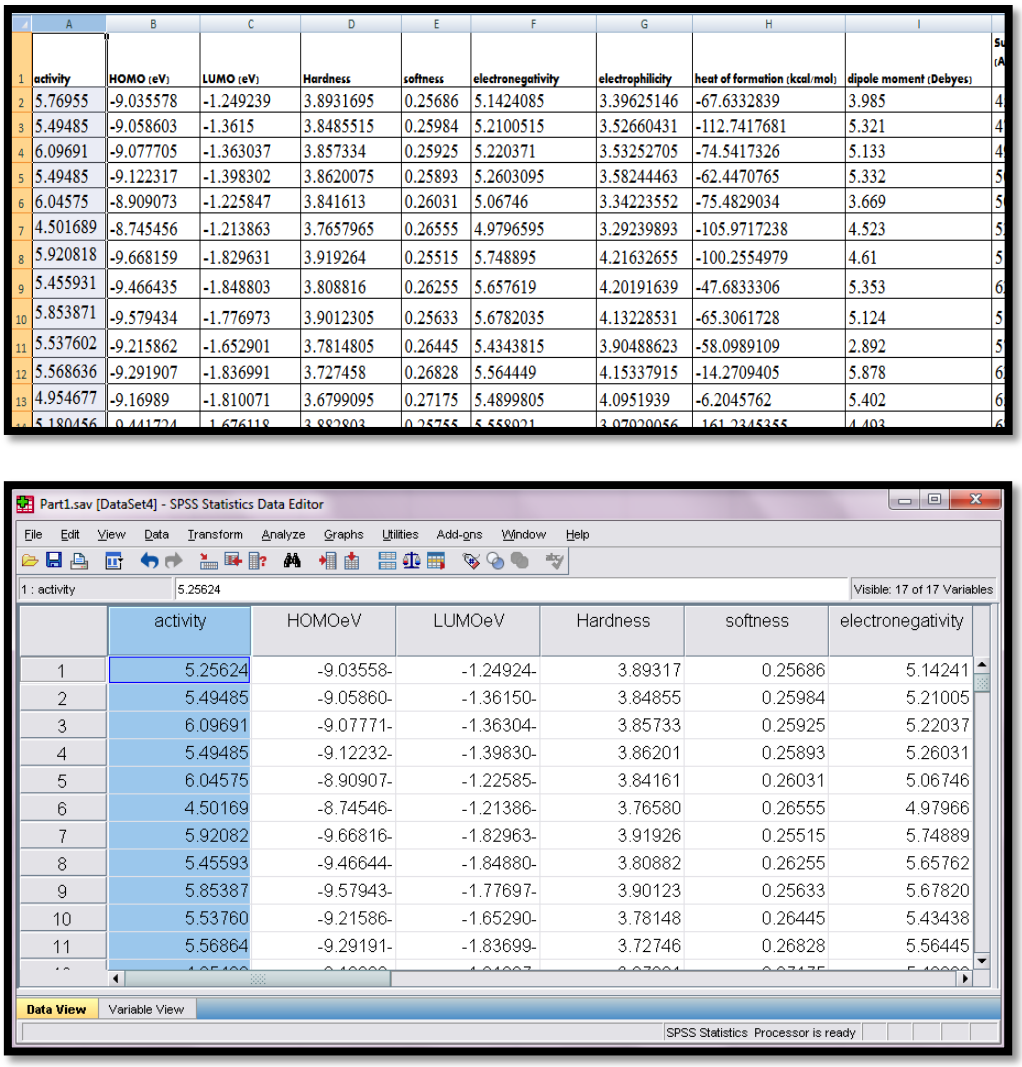
At the end of this work, we had separate file for each group of descriptors ready to be used in the next step.

2.2 Data analysis

Many of descriptors were calculated by Dragon and HyperChem software, and some of these descriptors that can provide the best model are chosen by SPSS software to predict the biological activity.

We collected all descriptors for each Part in the same excel file, and perform multiple linear regression (MLR) analysis to obtain the best model , this process is applied according to the following steps*:*

1) Prepare the files (part 1 and 2) 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. as shown in the following figure (2.8):



	A	B	C	D	E	F	G	H	I	
1	activity	HOMO (eV)	LUMO (eV)	Hardness	softness	electronegativity	electrophilicity	heat of formation (kcal/mol)	dipole moment (Debyes)	Su (A
2	5.76955	-9.035578	-1.249239	3.8931695	0.25686	5.1424085	3.39625146	-67.6332839	3.985	4
3	5.49485	-9.058603	-1.3615	3.8485515	0.25984	5.2100515	3.52660431	-112.7417681	5.321	4
4	6.09691	-9.077705	-1.363037	3.857334	0.25925	5.220371	3.53252705	-74.5417326	5.133	4
5	5.49485	-9.122317	-1.398302	3.8620075	0.25893	5.2603095	3.58244463	-62.4470765	5.332	5
6	6.04575	-8.909073	-1.225847	3.841613	0.26031	5.06746	3.34223552	-75.4829034	3.669	5
7	4.501689	-8.745456	-1.213863	3.7657965	0.26555	4.9796595	3.29239893	-105.9717238	4.523	5
8	5.920818	-9.668159	-1.829631	3.919264	0.25515	5.748895	4.21632655	-100.2554979	4.61	5
9	5.455931	-9.466435	-1.848803	3.808816	0.26255	5.657619	4.20191639	-47.6833306	5.353	6
10	5.853871	-9.579434	-1.776973	3.9012305	0.25633	5.6782035	4.13228531	-65.3061728	5.124	5
11	5.537602	-9.215862	-1.652901	3.7814805	0.26445	5.4343815	3.90488623	-58.0989109	2.892	5
12	5.568636	-9.291907	-1.836991	3.727458	0.26828	5.564449	4.15337915	-14.2709405	5.878	6
13	4.954677	-9.16989	-1.810071	3.6799095	0.27175	5.4899805	4.0951939	-6.2045762	5.402	6
	5.180456	-9.441724	-1.676118	3.882803	0.25755	5.558021	3.97029056	-161.2345355	4.403	6

Figure (2.8): SPSS software window.

2) Open the file containing the dependent variable and independent variables using SPSS, then analyze menu and choose regression and select linear as in the following figure (2.9):

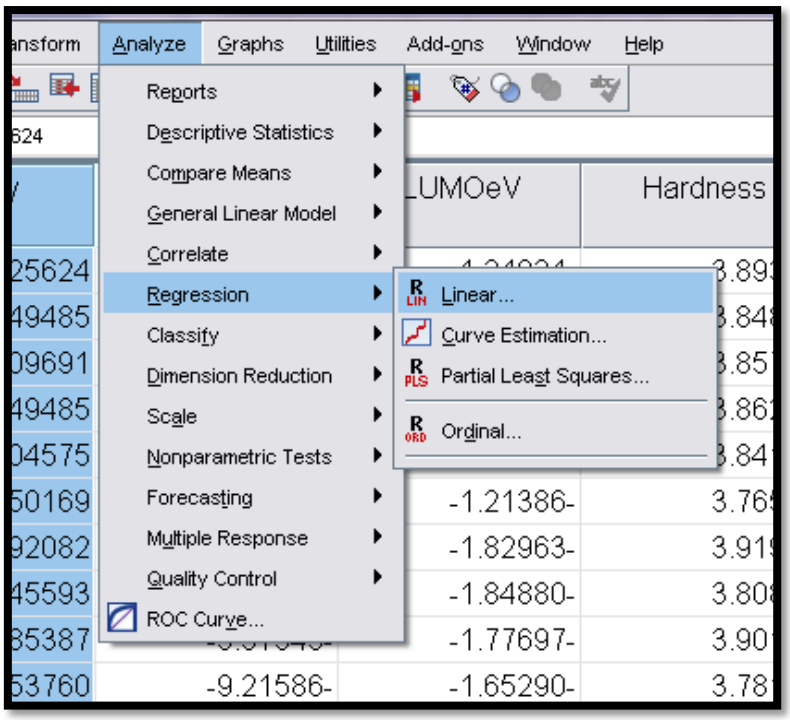


Figure (2.9): Analyze menu.

3) Set activity as the dependent variable and set the descriptors in the input file as the independent variable in the linear regression dialog box, and then press on the options button of the same dialog box as shown in the following figure (2.10):

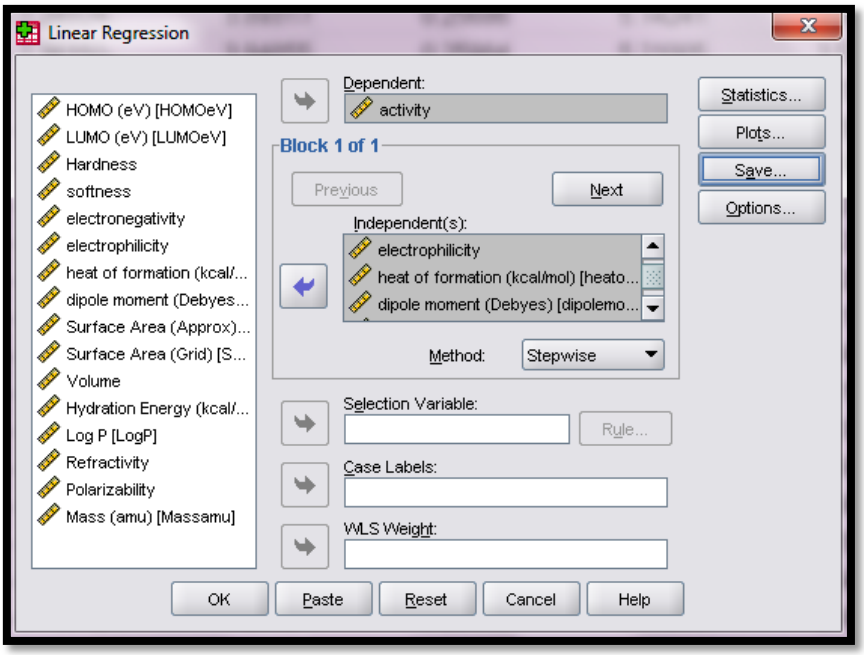


Figure (2.10): Linear regression dialog.

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. As shown in the following figure (2.11):

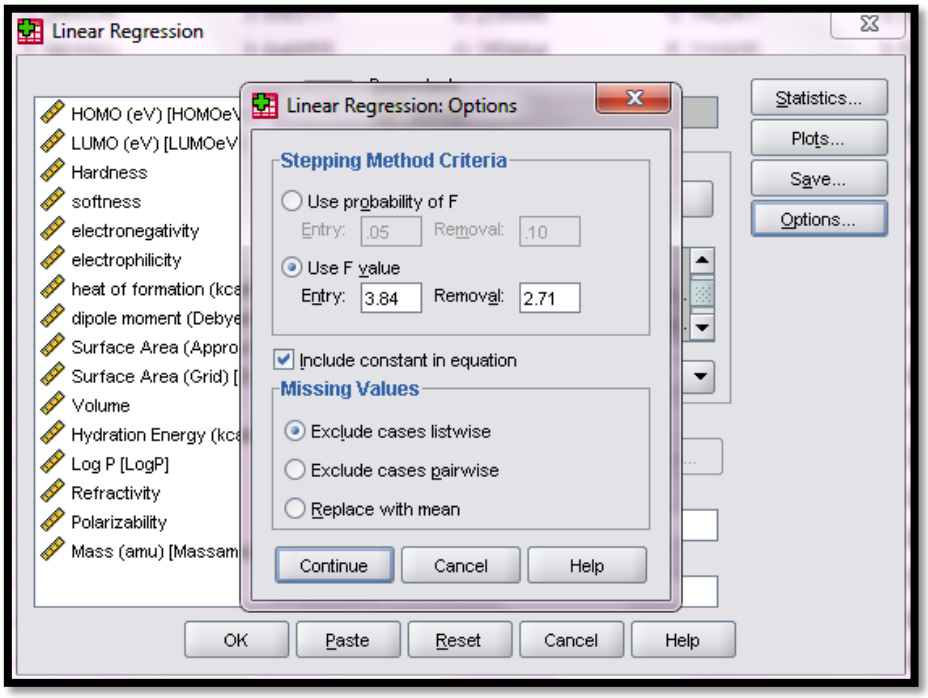


Figure (2.11): linear regression options.

5) Choose the method to be stepwise, click save to store results back to the input sheet, choose the predicted values to be unstandardized and then click continue. As shown in the following figure (2.12):

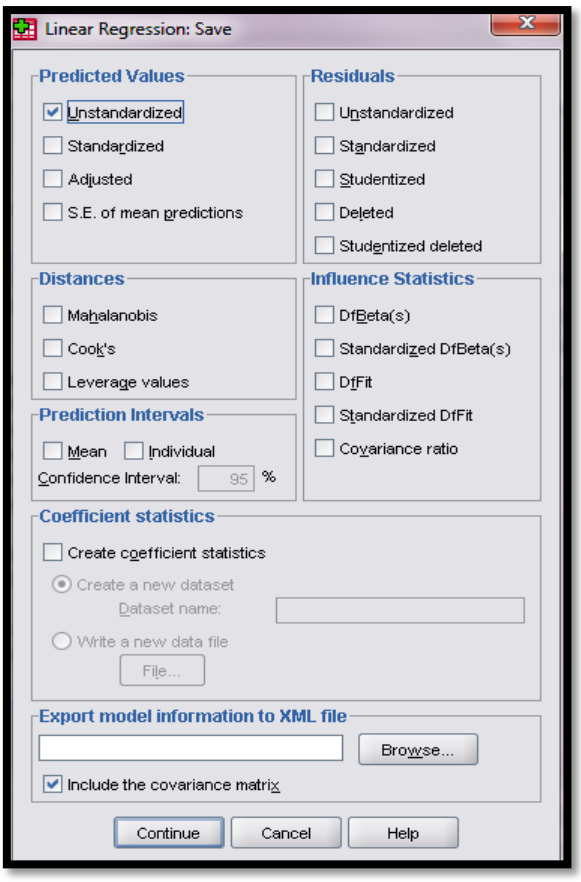


Figure (2.12): Linear regression save.

6) Click statistics to generate additional statistics for variables then click continue, and finally click OK in the linear regression dialog box. As shown in figure (2.10).

By finishing these steps, we will have linear model for each part of data set, which contain the best informative descriptors from each part. So these steps must be performed again on a file that has the whole descriptors chosen by the MLR models of each part of descriptors (final MLR) to get the final model that has the best informative descriptors from the whole descriptors.

2.3 Model Validation

2.3.1 MLR validation

In the previous method, the best models for part 1 and 2 were obtained from the output of SPSS software, and then these models relate the activity with the descriptors linearity, so the

models should be validated. And the statistical fit of a QSAR can be assessed in many easily available statistical terms as correlation coefficient R.

Correlation coefficient (R) gives a quantitative measure of how well each descriptor describes the activity and also it is used for indicating goodness of fit. The values are between 0 and 1, with 0 denoting that model does not explain any variation and 1 denoting that it perfectly explains the observed variation.

R^2 is calculated by SPSS program while building the MLR model.

2.3.1.1 Leave One Out Cross validation using MATLAB

As the name suggests, leave-one-out cross validation (LOOCV) involves using a single observation from the original sample as the validation datum and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data.

Usually, one compound of the set is extracted each time, and then the model is recalculated using as training set the n-1 (where n is number of compounds) remaining compounds, so that the biological activity value for the extracted compound is predicted once for all compounds. This process is repeated n times for all the compounds of the initial set, thus obtaining a prediction for each object. This process referred as leave-one-out (LOO) method.

The idea behind this method is to predict the property value for a compound from the data set, which is in turn predicted from the regression equation calculated from the data for all other compounds. For evaluation, predicted values can be used for PRESS, RMSPE, and squared correlation coefficient criteria (r^2_{cv}).

This type of validation for both parts 1 and 2 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:\Matlab1\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 massage says “model number”, where the next line contains number which 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 directory (C:\Matlab1\work) or the directory you are working in.

2.3.1.2 Leave Many Out Cross validation using MATLAB

In the 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. Also an alternative method can be defined when leaving out more than a compound of the data set at each time.

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 file name 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_LOG.dat", appears in the same directory in addition to printing cross validation parameters on the screen.

2.3.2 Principle Component Artificial Neural Networks (PC-ANN)

We perform PC-ANN analysis when the relation between the activity and the chemical descriptors is more complicated than a linear relationship, to obtain the better model than MLR models. After using SPSS software to perform MLR analysis, we finished up with several good 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:

2.3.2.1 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 set.

1) Prepare two excel files (for part 1 and part 2) 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 "calcaplot.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 will appears. Select the training, validation and test sets molecules from these data points so they span the same space of the entire data.

2.3.2.2 Performing ANN model

1) Prepare an excel file for each model for both part 1 and part 2 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 of ANN analysis , you should edit the Matlab script "ann_ext_test_4loop.m" to indicate the training, validation, and 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 the 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 the 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 the 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 model will be compared with 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 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 the hidden nodes for these models. To do so, you have to choose a range of hidden nodes numbers (in order 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 for each part, you choose the best of them according to the cross validation parameters.

2.3.2.3 Randomization (Chance correlation).

For further validation, run chance correlation test for the optimal models for each part. 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 of 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 produce by chance.

2.3.3 Partial Least Squares (PLS)

Partial least squares is another linear regression method other than MLR which build a linear relation between the activity and the descriptors.

The descriptors which have correlation will be gathered in one latent variable which indicates them and that is avoid the intercorrelation that may be happen in MLR.

PLS is performed using Matlab software by the following steps:

1) In this method, we divide the data set into two sets training set (80% of the data set) for the PLS work and test set (20 % of 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 the following:

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

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

Xtest: that has the column 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 xtest.txt and >> load ytest.txt

These commands to load the files that have the data, then type the following commands in Matlab sheet, and then click enter.

>> [p, q, w, b, t, u, x, y, l] = pls(xcal, ycal, 10);

>> plspsr = plspress(xcal, ycal, p, q, w, b, 10);

Then type the command: >> plot(plsprs,’*’)

Those commands typed to run the needed script.

5) At the 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.

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)

Befor clicking enter, you should change the number 2 in the command with the point obtained from the step 5 and then click enter.

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: >> predtest=c: then click enter
10) From the workspace menu, you can find the predicted results in the training set.

PLS analysis with cross validation can be used to advanced investigation of the linear relationships of the obtained regression models and the PLS results are close to MLR.

On the other hand, the results of MLR are good for all used models, so the PLS method was neglected.

-Chapter three

Results and Discussion

In this study we developed MLR-QSAR model that relates the activity of 121 cyclooxygenase -2 enzyme inhibitors to their structures using their theoretical descriptors as structure indicators. The work was done by successive steps to build the linear and non linear models and perform their validations. The results are discussed in this chapter.

Descriptors calculation

From the literature (44-57), we take the structures and experimental activities values of our 121 compounds of cyclooxygenase-2 enzyme inhibitors.

The compounds were gathered as two parts according to the type of chemical structure. Each part has the compounds structures and their activities as pIC_{50} and these structures and their activities are summarized in the table (2.1) and (2.2).

We built the structure of each compound by HyperChem software and AM1 semi-empirical method was used to optimize the chemical structures. On the other hand, we used HyperChem to calculate some of descriptors (quantum chemical descriptors).

The other groups of descriptors were calculated by Dragon software for each part. We neglected the constant or near constant descriptors because they cannot differentiate between the different compounds.

All Dragon output were fifteen files and each one has the results of one group of the descriptors which are topological, constitutional, BUCT, molecular walk count, Galvez topological and charge indices, charge indices, charge descriptors, 2D autocorrelation, randic molecular profiles, geometrical, RDF, 3D-MoRSE, WHIM, getaway, functional groups and atom-centered fragment descriptors. All output of each group of descriptors were gathered (as one group) in two Excel files to each part.

Finally, Excel files (part1 and part2) were prepared to the next step.

MLR

We used the previous Excel files (part1 and part2) to perform MLR analysis by using SPSS software by stepwise regression method. And the results of performing MLR regression for each part are summarized in table (3.1) and (3.8). The MLR analysis was done by applying the individual method[58], in which we apply MLR to each group of descriptors alone and then perform final MLR on the gathered best descriptors of each group.

By using SPSS software, we tried to reach a model with the highest correlation coefficient (R), and in the same time the lowest number of descriptors for each part to reduce the inter correlation between descriptors in the model as much as we can.

Part 1: Results and discussion.

Table (3.1): The final MLR models for Part1

Model No.*	R	R^2	Adjusted R^2	SE	Descriptors
1	0.594	0.352	0.338	0.703	Mor11p
2	0.638	0.407	0.381	0.680	Mor11p, P1m
3	0.681	0.464	0.428	0.653	Mor11p, P1m, nSO2N
4	0.720	0.518	0.473	0.627	Mor11p, P1m, nSO2N, Km
5	0.747	0.558	0.506	0.608	Mor11p, P1m, nSO2N, Km, RDF150m
6	0.782	0.611	0.554	0.577	Mor11p, P1m, nSO2N, Km, RDF150m, Mor12v
7	0.803	0.645	0.583	0.558	Mor11p, P1m, nSO2N, Km, RDF150m, Mor12v, Mor03u
8	0.831	0.690	0.627	0.528	Mor11p, P1m, nSO2N, Km, RDF150m, Mor12v, Mor03u, RDF135p
9	0.850	0.723	0.657	0.505	Mor11p, P1m, nSO2N, Km, RDF150m, Mor12v, Mor03u, RDF135p, ISH
10	0.873	0.763	0.698	0.474	Mor11p, P1m, nSO2N, Km, RDF150m, Mor12v, Mor03u, RDF135p, ISH, G3s

*Model No. refers to model number as SPSS output, R refers to correlation coefficient, R^2 refers to coefficient of determination, R^2_{adj} refers to adjusted R^2.

Model 10 that has the highest R^2 and R_{adj} is the best model for part 1, and the following equation represents the best MLR model:

(Equation Part 1)

$\mathbf{pIC_{50}}$ = -14.401 (±8.149) – 1.854 (±0.283) × **"Mor11p"** + 13.791 (±3.032) × **"P1m"** - 0.457 (±0.173) × **"nSO2N"** – 9.808 (±2.614) × **"Km"** - 1.185 (±0.252) × **"RDF150m"** -1.815 (±0.445) × **"Mor12v"** + 0.221 (±0. .061) × **"Mor03u"** +0 .346 (±0.107) × **"RDF135p"** + 21.882 (±8.388) × **"ISH"** - 30.921 (±12.446) × **"G3s"**

According to the above equations Part 1, the most important descriptors of this equation are **G3s** and **ISH** which reflect the molecular geometrical coordinates of the compounds; **G3s** is inversely proportional to the inhibitory activity of the compounds while **ISH** is directly proportional to the inhibitory activity of the compounds.

By Matlab software, we applied LOO cross validation on the MLR models for each model that has R^2 more than 0.6.[59] and the results are summarized in table 3.2 below:

Table (3.2) LOO cross validation parameters for the final MLR models 6-10 of Part 1.

Model	PRESS	SPRESS	SST	R^2_{CV}	PRESS/SST	PSE	RSEP
6	13.6238	0.5764	21.4294	0.3642	0.6358	0.5328	8.9637
7	12.4541	0.5580	22.5991	0.4489	0.5511	0.5094	8.5703
8	10.8606	0.5277	24.1926	0.5511	0.4489	0.4757	8.0033
9	9.7088	0.5055	25.3443	0.6169	0.3831	0.4497	7.5670
10	8.3209	0.4742	26.7324	0.6887	0.3113	0.4164	7.0053

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), STT (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 cross validation for part 1 in tables (3.2), we can see that PRESS values always less than SST values and this means that the model predicting ability better than chance.

The best models of part 1 that have the highest values of R^2_{CV} and the lowest values of PSE, the models (9 and 10) are the best for part 1. So we picked those models for part 1 as the best models to be the candidate to next step, as inputs to PC-ANN.

PC-ANN

The first step before running ANN we applied PCA on the data to get rid of the inter correlation between descriptors, divide the compounds 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. This process applied for part 1.

The results of PCA application is shown in the figure (3.1) for part 1.

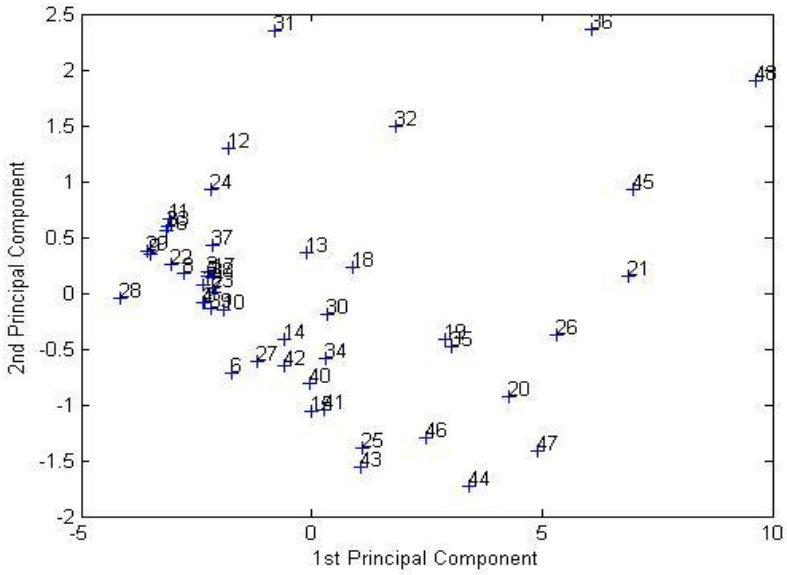


Figure (3.1): Correlation between 1st and 2nd principle components for part 1.

In the previous figure (3.1) of PCA for part 1, the 48 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. And in the next step, 48 compounds were used as data points to ANN for part 1.

PC-ANN models were built using the proper Matlab script. We applied the script on each one of the picked models for part 1(Models 9 and 10) with constant hidden nodes for all the models.

The table (3.3) shows a summary of the cross validation parameters of the models for part 1.

Table (3.3): correlation coefficient and cross validation results for ANN models 9 and 10 for part 1.

Model No.*	hn. No.**	nPCs ***	R_train	PRESS_train	R^2_{CV}_train	R_test	PRESS_test	RSEP_test	R_val	PRESS_val	RSEP_val
9	6	5	0.842	6.898	0.511	0.847	3.861	10.452	0.735	1.288	6.639
10	6	5	0.781	9.784	0.106	0.874	3.529	9.993	0.714	1.255	6.555

*Model No. (Model number),** Hn-No. (number of hidden nodes), *** nPCs (number of principle components).

All results in table (3.3) are close to each other, for this reason we chosen models 9 and 10 of part 1 to apply the next step in ANN work.

Each one of these 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 part 1 are summarized in the tables (3.4), (3.5) for Model 9 and model 10, respectively.

Table (3.4): Correlation coefficient and cross validation parameters for optimizing number of hidden nodes for model 9 for Part 1.

No. Hn.*	nPCs **	R_train	PRESS_train	R^2_{CV}_train	R_test	PRESS_test	RSEP_test	R_val	PRESS_val	RSEP_val
3	**5**	**0.751**	**9.007**	**0.137**	**0.803**	**5.550**	**12.532**	**0.826**	**2.967**	**9.759**
4	**5**	**0.813**	**7.395**	**0.193**	**0.759**	**5.679**	**12.676**	**0.706**	**3.718**	**10.923**
5	**5**	**0.789**	**8.386**	**0.481**	**0.829**	**3.200**	**9.516**	**0.727**	**2.706**	**9.320**
6	**5**	**0.798**	**7.475**	**0.400**	**0.856**	**4.783**	**11.633**	**0.795**	**2.979**	**9.778**
7	**5**	**0.810**	**8.466**	**0.263**	**0.871**	**2.758**	**8.834**	**0.756**	**2.636**	**9.197**
8	**5**	**0.828**	**6.766**	**0.481**	**0.791**	**4.256**	**10.975**	**0.738**	**2.945**	**9.722**
9	**5**	**0.846**	**6.028**	**0.553**	**0.784**	**4.563**	**11.363**	**0.702**	**3.182**	**10.105**
10	**5**	**0.781**	**9.026**	**0.562**	**0.824**	**4.452**	**11.224**	**0.734**	**3.393**	**10.436**
11	**5**	**0.866**	**5.830**	**0.601**	**0.868**	**3.428**	**9.850**	**0.721**	**3.211**	**10.152**
12	**5**	**0.846**	**6.240**	**0.625**	**0.937**	**4.140**	**10.824**	**0.749**	**2.723**	**9.348**
13	**5**	**0.820**	**7.043**	**0.597**	**0.772**	**5.177**	**12.104**	**0.750**	**3.681**	**10.869**
14	**5**	**0.845**	**5.911**	**0.606**	**0.797**	**6.215**	**13.261**	**0.717**	**4.720**	**12.307**
15	**5**	**0.891**	**4.511**	**0.660**	**0.839**	**4.532**	**11.324**	**0.810**	**3.238**	**10.195**
16	**5**	**0.874**	**4.947**	**0.648**	**0.834**	**4.816**	**11.674**	**0.707**	**4.328**	**11.786**
17	**5**	**0.786**	**8.554**	**0.552**	**0.764**	**5.919**	**12.942**	**0.873**	**2.686**	**9.285**
18	**5**	**0.870**	**5.264**	**0.655**	**0.773**	**4.081**	**10.746**	**0.707**	**4.113**	**11.490**
19	**5**	**0.811**	**7.047**	**0.505**	**0.816**	**5.150**	**12.072**	**0.909**	**3.460**	**10.538**
20	**5**	**0.864**	**5.234**	**0.634**	**0.906**	**3.518**	**9.977**	**0.711**	**3.643**	**10.812**

*Hn-No. (number of hidden nodes), ** nPCs (number of principle components).

Table (3.5): Correlation coefficient and cross validation parameters for optimizing number of hidden nodes for model 10 for part 1.

Hn-	nPCs	R_train	PRESS_train	R2CV_train	R_test	PRESS_test	RSEP_test	R_val	PRESS_val	RSEP_val
3	5	0.764	10.641	0.192	0.789	6.786	13.857	0.701	4.768	12.37
4	5	0.75	9.149	0.349	0.8	5.077	11.985	0.704	3.638	10.806
5	5	0.783	8.267	0.27	0.823	4.712	11.547	0.756	3.921	11.217
6	5	0.763	10.047	-0.855	0.883	5.92	12.943	0.709	3.529	10.643
7	5	0.774	9.446	-0.263	0.807	5.535	12.515	0.712	3.834	11.092
8	5	0.753	10.219	0.204	0.844	4.175	10.869	0.729	3.081	9.944
9	5	0.84	6.247	0.521	0.764	5.726	12.729	0.707	3.466	10.547
10	5	0.816	6.937	0.568	0.924	3.947	10.569	0.729	3.665	10.845
11	5	0.83	6.66	0.486	0.838	4.376	11.128	0.725	3.731	10.943
12	5	0.814	7.202	0.527	0.856	4.638	11.456	0.749	4.344	11.807
13	5	0.774	9.024	-0.209	0.782	5.873	12.891	0.818	3.682	10.871
14	5	0.771	9.198	0.073	0.78	4.95	11.835	0.746	3.233	10.186
15	5	0.869	5.051	0.654	0.896	5.292	12.237	0.72	3.769	10.998
16	5	0.777	8.512	0.421	0.915	5.956	12.982	0.747	2.961	9.749
17	5	0.825	6.568	0.505	0.828	5.87	12.888	0.815	3.763	10.989
18	5	0.883	4.55	0.699	0.918	3.26	9.604	0.756	2.71	9.325
19	5	0.866	5.161	0.637	0.904	5.601	12.589	0.745	4.25	11.679
20	5	0.863	5.408	0.564	0.866	5.711	12.712	0.709	3.665	10.845

*Model No. (Model number), Hn-No. (number of hidden nodes), nPCs (number of principle components).

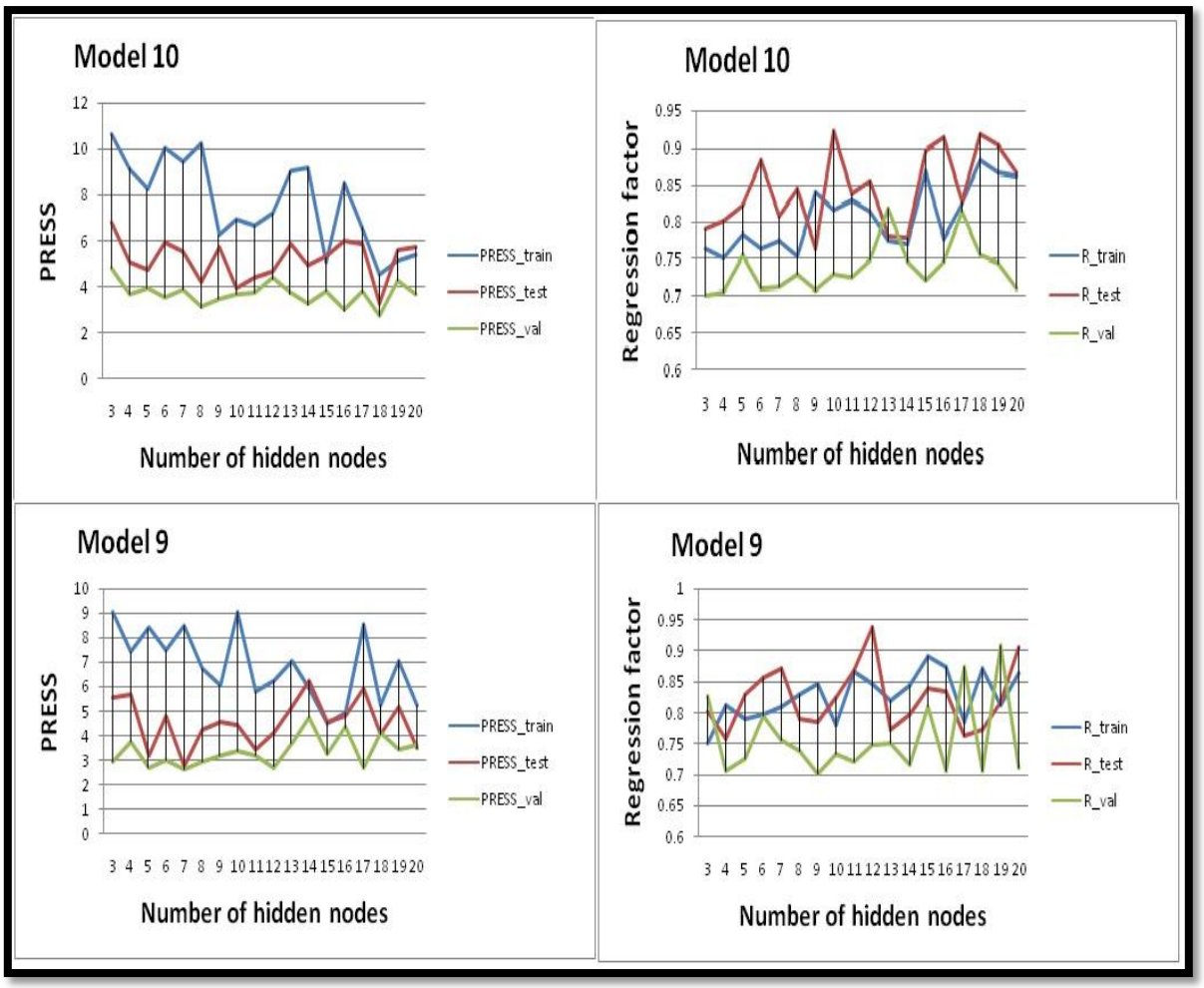


Figure (3.2): PRESS against number of hidden nodes as well as regression factor against number of hidden nodes for model 9 and 10.

Figure (3.2) shows the PRESS values against the number of hidden nodes as well as the regression factor against number of hidden nodes for models **9 and 10 for part 1**. This figure shows that the lowest PRESS value (4.14) is obtained when using 12 hidden nodes for model **9** with regression coefficient for the test set of 0.937. For model **10**, the lowest PRESS (3.947) is obtained when using 10 hidden nodes with regression coefficient for the test set of 0.924.

Both models model 9 and model 10 of part 1 were examined to inspect the presence of outliers that may affect models validity. By inspecting the residuals of these models, the residue equals the difference between the predicted and observed one, there were no outliers as it is shown in figures (3.3).

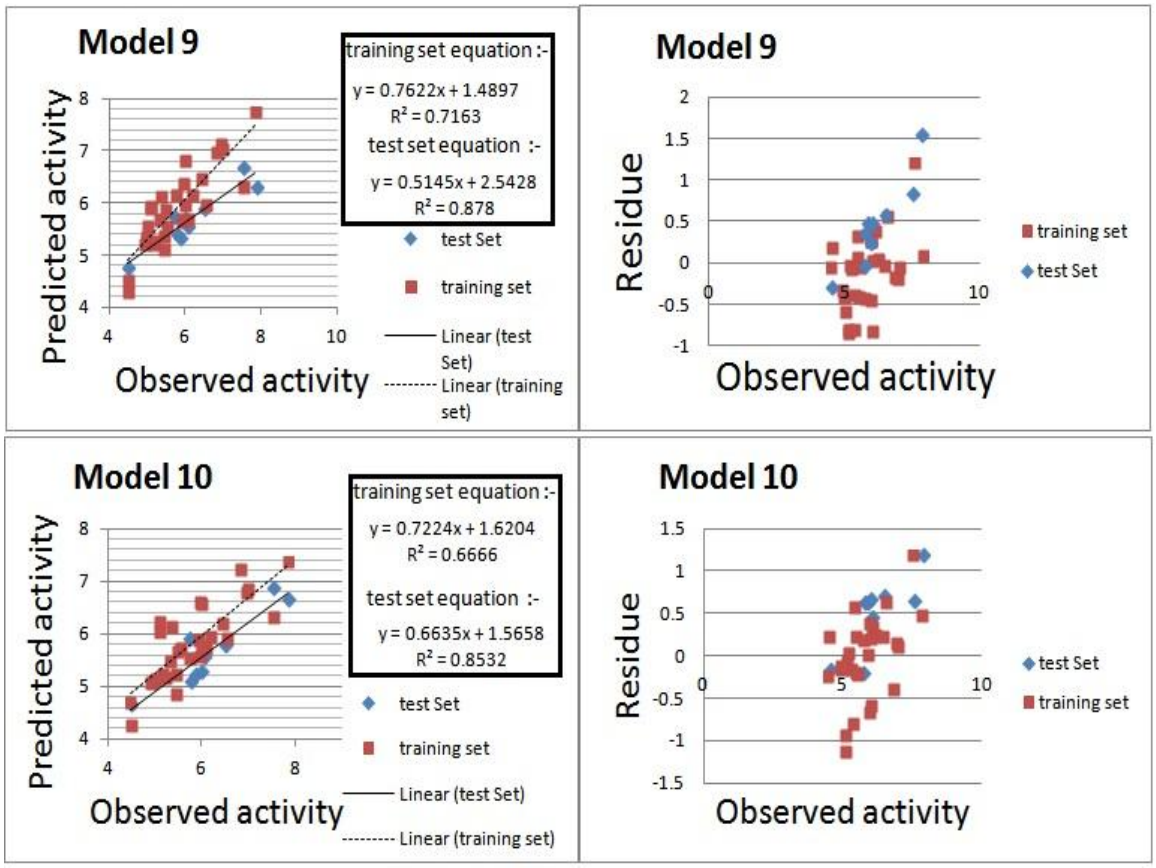


Figure (3.3): Predictive activity against observed one as well as their residue for part 1 models 9 and 10 using 12 and 10 hidden nodes, respectively.

The correlation between calculated and observed pIC_{50} for the training set of model **9** is given by:

Calculated pIC_{50} = 0.762 Observed pIC_{50} + 1.49

And for the test set of this model is given by:

Calculated pIC_{50} = 0.515 Observed pIC_{50} + 2.543

While the Correlation between calculated and observed pIC_{50} for the training set of model **10** is given by:

Calculated pIC_{50} = 0.722 Observed pIC_{50} + 1.62

And for the test set of this model is given by:

Calculated pIC_{50} = 0.664 Observed pIC_{50} + 1.566

Randomization

Randomization test is performed to investigate the probability of chance correlation for the optimal models (models **9** and **10** with 12 and 10 hidden nodes in the network, respectively, of part 1. Chance correlation was done using the same configuration parameters and the same activation functions of all our ANN models. The results of chance correlation of part 1 for models **9** (using 12 hidden nodes) and **10** (using 10 hidden nodes) are summarized in the following tables (3.6) and (3.7) respectively. These tables show that the coefficients of determination obtained by chance are low in general while the PRESS values are high. This indicates that the models obtained from ANN are better than those obtained by chance.

As we can see, our models were validated by calculating different statistical parameters, using external test set and finally performing randomization test.

Table (3.6): Statistical parameters of chance correlation of model 9 with 12 hidden nodes. (Part 1)

Trial No.	nPCs	R_train	PRESS_train	R2CV_train	R_test	PRESS_test	R_val	PRESS_val	R2CV_val
1	5	-0.268	30.745	-9.399	0.189	9.757	-0.14	3.177	-9.438
2	5	-0.268	30.745	-9.399	0.189	9.757	-0.14	3.177	-9.438
3	5	0.097	32.019	-1.683	0.035	11.424	-0.138	7.159	-2.697
4	5	0.033	28.149	-4.082	-0.011	11.395	-0.208	4.146	-2.952
5	5	-0.179	48.25	-1.81	-0.058	17.888	-0.256	5.046	-2.302
6	5	-0.19	36.222	-3.875	0.021	9.534	0.134	3.803	-1.77
7	5	0.097	32.019	-1.683	0.035	11.424	-0.138	7.159	-2.697
8	5	0.241	23.612	-2.988	0.025	11.469	-0.079	2.831	-9.74
9	5	-0.098	37.479	-2.439	-0.179	14.053	0.102	5.523	-3.314
10	5	0.105	26.033	-4.479	-0.096	10.358	0.005	5.74	-12.847

Table (3.7): Statistical parameters of chance correlation of model 10 with 10 hidden nodes.

Trial No.	nPCs	R_train	PRESS _train	R2CV _train	R_test	PRESS _test	R_val	PRESS _val	R2CV _val
1	5	0.172	21.113	-13.051	0.166	8.656	0.012	5.517	-60.904
2	5	0.067	32.604	-2.16	-0.07	12.306	-0.247	10.97	-2.343
3	5	0.054	25.692	-3.011	-0.179	12.554	0.054	6.615	-3.402
4	5	-0.3	36.968	-4.131	-0.087	12.714	-0.177	10.50	-2.009
5	5	-0.094	33.831	-2.436	0.024	9.301	-0.289	13.60	-4.045
6	5	0.279	26.425	-0.656	0.26	9.957	0.236	5.423	-3.844
7	5	0.157	25.195	-2.158	-0.129	12.497	0.032	6.788	-4.322
8	5	0.247	22.6	-1.6	-0.01	10.101	-0.016	5.921	-14.981
9	5	-0.049	37.462	-2.043	0.01	10.186	-0.152	7.241	-5.764
10	5	0.042	32.375	-1.444	-0.232	15.194	0.095	7.077	-3.096

Both the external and cross-validation methods are used to validate the performances of the resulting models. Employed randomization test indicates that the models obtained from ANN are better than those obtained by chance.

Part 2 : Results and discussion:

Table (3.8): The final MLR models for (Part 2).

Model No.	R	R^2	Adjusted R^2	SE	Descriptors
1	0.331	0.109	0.097	0.776	MATS3e
2	0.429	0.184	0.161	0.748	MATS3e, E1v
3	0.533	0.284	0.253	0.706	MATS3e, E1v, E3s
4	0.631	0.398	0.363	0.652	MATS3e, E1v, E3s, Me
5	0.676	0.457	0.417	0.624	MATS3e, E1v, E3s, Me, C-028
6	0.707	0.499	0.454	0.604	MATS3e, E1v, E3s, Me, C-028, G3p
7	0.726	0.527	0.476	0.591	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3
8	0.755	0.57	0.516	0.568	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u
9	0.77	0.593	0.535	0.557	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u
10	0.785	0.616	0.554	0.545	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, MATS8e
11	0.799	0.639	0.573	0.534	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, MATS8e, dipole moment (Debyes)
12	0.822	0.675	0.611	0.51	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, MATS8e, dipole moment (Debyes), G3m
13	0.837	0.7	0.634	0.494	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, MATS8e, dipole moment (Debyes), G3m, H7m
14	0.849	0.72	0.652	0.482	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, MATS8e, dipole moment (Debyes), G3m, H7m, R6m
15	0.848	0.719	0.657	0.478	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, dipole moment (Debyes), G3m, H7m, R6m

*Model No. refers to model number as SPSS output, R refers to correlation coefficient, R^2 refers to coefficient of determination, R^2_{adj} refers to adjusted R^2.

Model 14 that has the highest R^2 and R_{adj} is the best model for part 2, and the following equation represents the final MLR model:

(Equation Part 2)

p IC $_{50}$ = -7.938 (±11.143) - 2.987 (±1.328) × **"MATS3e"** + 7.043 (±1.705) × **"E1v"** + 2.967 (±0.727) × **E3s** -17.669 (±8.819) × **"Me"** + 0.424 (±0 .167) × **"C-028"** + 21.414 (±5.236) × **"G3p"** + 7.092 (±1.299) × **"BELm3"** + 0 .182 (±0.095) × **"Mor03u"** + 28.858 (±8.360) × "**G2u"** + 0 .179 (±0.057) × **"dipole moment (Debyes)"** + 46.973 (±12.190) × **"G3m"** - 4.845 (±1.268) × **"H7m"** + 1.783 (±0.705) × "**R6m"**

According to the above equations Part 2, the most important descriptors of this equation are **G3m**, **G2u and G3p** which reflect the molecular geometrical coordinates of the compounds and they are directly proportional to the inhibitory activity of the compounds.

By Matlab software, we applied LOO cross validation on the MLR models for each model that has R^2 more than 0.6.[59] and the results are summarized in table 3.9 below:

Table (3.9) LOO cross validation parameters for the final MLR models 10-15 of part 2

Model	PRESS	SPRESS	SST	R^2_{CV}	PRESS/SST	PSE	RSEP
10	18.4383	0.5453	29.6084	0.3773	0.6227	0.5026	8.36
11	17.3663	0.5336	30.6803	0.434	0.566	0.4877	8.1134
12	15.592	0.5098	32.4547	0.5196	0.4804	0.4622	7.6877
13	14.3906	0.4939	33.656	0.5724	0.4276	0.444	7.3856
14	13.5021	0.4825	34.5446	0.6091	0.3909	0.4301	7.154
15	13.5021	0.4784	34.5446	0.6091	0.3909	0.4301	7.154

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), STT (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 cross validation for part 2 in tables (3.9), we can see that PRESS values always less than SST values and this means that the model predicting ability better than chance.

The best models of part 2 that have the highest values of R^2_{CV} and the lowest values of PSE, the models (14 and 15) are the best for part 2. So we picked those models for part 2 as the best models to be the candidate to next step, as inputs to PC-ANN.

PC-ANN

The first step before running ANN we applied PCA on the data to get rid of the inter correlation between descriptors, divide the compounds 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. This process applied for part 2.

The results of PCA application is shown in the figure (3.4) for part 2.

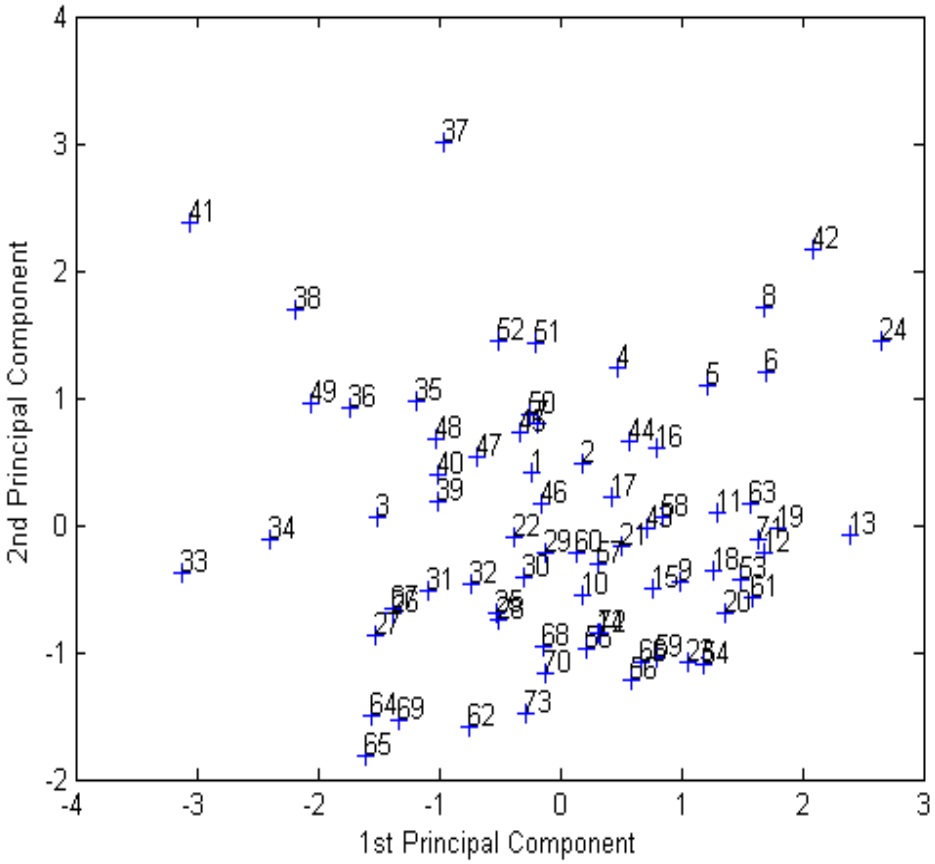


Fig. (3.4): Correlation between 1st and 2nd principle components for part 2.

The previous figure (3.4) of PCA for part 2, the 73 compounds have 2 outliers (compounds 37 and 41), which are lie from the compound cluster, this means that those two compounds act in different way from other compounds with respect to activity and descriptors. Then the 71 compound were partitioned as the previous step to validation set, test set and training set.

And in the next step, 71 compounds were used as data points to ANN.

PC-ANN models were built using the proper Matlab script. We applied the script on each one of the picked models for part 2, and part 2 models 14 and 15 were used by using constant hidden nodes for all the models.

The table (3.10) shows a summary of the cross validation parameters of the models for part 2.

Table (3.10) Correlation coefficient and cross validation results for ANN models 14 and 15 for part2.

Model No.	hn. No	nPCs	R_train	PRESS_train	R^2_{CV}_train	R_test	PRESS_test	RSEP_test	R_val	PRESS_val	RSEP_val
14	7	6	0.706	13.842	-0.533	0.755	6.02	10.933	0.688	4.166	9.337
15	7	6	0.666	14.965	-0.406	0.726	6.294	11.179	0.66	4.378	9.572

*Model No. (Model number), Hn-No. (number of hidden nodes), nPCs (number of principle components).

All results in table (3.10) are close to each other, for this reason we chosen models 14 and 15 of part 2 to apply the next step in ANN work.

Each one of these 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 two parts are summarized in the tables (3.11), (3.12) for Model 14 and model 15 of part 2, respectively.

Table (3.11) Correlation coefficients and cross validation parameters for model optimizing number of hidden nodes for model 14 for part 2.

Hn. No.	nPCs	R_train	PRESS_train	R2CV_train	R_test	PRESS_test	RSEP_test	R_val	PRESS_val	RSEP_val
6	6	0.675	14.846	-0.401	0.656	7.158	11.922	0.662	4.35	9.541
7	6	0.689	14.597	-0.638	0.689	7.184	11.944	0.689	4.086	9.247
8	6	0.654	16.33	-1.656	0.66	7.682	12.351	0.665	4.957	10.185
9	6	0.666	15.166	-0.107	0.654	7.251	11.999	0.655	4.263	9.445
10	6	0.741	12.175	0.074	0.702	6.293	11.178	0.658	4.439	9.639
11	6	0.733	12.833	-0.289	0.738	6.042	10.953	0.658	4.382	9.577
12	6	0.778	10.827	0.141	0.757	5.928	10.849	0.661	4.45	9.651
13	6	0.684	14.324	-0.211	0.733	6.096	11.002	0.662	4.746	9.966
14	6	0.737	13.082	-0.395	0.696	6.775	11.599	0.651	4.323	9.512
15	6	0.744	12.026	0.06	0.652	7.52	12.22	0.69	4.384	9.578
16	6	0.675	14.648	-0.418	0.77	5.602	10.547	0.717	3.892	9.025
17	6	0.756	11.527	0.288	0.718	6.035	10.947	0.686	4.207	9.383
18	6	0.715	13.822	-0.624	0.748	5.573	10.519	0.65	4.329	9.518
19	6	0.665	15.19	-0.64	0.666	6.925	11.726	0.652	4.3	9.486
20	6	0.653	15.859	-0.488	0.728	6.728	11.558	0.653	4.486	9.689

*Model No. (Model number), Hn-No. (number of hidden nodes), nPCs (number of principle components).

With model 14 at 3, 4 and 5 hidden nodes numbers we did not get R of the test set more than 0.6, so the trials were neglected.

Table (3.12) Correlation coefficient and cross validation parameters for model optimizing number of hidden nodes for model 15 for part 2.

hn. NO	nPCs	R_train	PRESS_train	R2CV_train	R_test	PRESS_test	RSEP_test	R_val	PRESS_val	RSEP_val
5	6	0.670	16.033	-1.023	0.706	6.835	12.171	0.706	3.241	8.702
6	6	0.705	14.782	-0.681	0.710	6.616	11.974	0.663	3.538	9.092
7	6	0.730	13.707	-0.392	0.823	4.727	10.122	0.711	3.109	8.522
8	6	0.653	16.133	-0.510	0.772	5.841	11.251	0.689	3.441	8.967
9	6	0.754	12.506	-0.027	0.687	6.617	11.975	0.681	3.246	8.709
10	6	0.692	14.677	-0.108	0.687	6.878	12.209	0.660	3.550	9.108
11	6	0.680	16.342	-1.156	0.707	6.417	11.792	0.700	3.269	8.739
12	6	0.662	16.359	-1.116	0.714	3.570	0.726	0.682	3.553	9.112
13	6	0.659	16.020	-0.609	0.755	5.797	11.209	0.706	3.417	8.936
14	6	0.665	15.791	-0.407	0.818	4.685	10.076	0.688	3.004	8.378
15	6	0.664	15.826	-0.459	0.725	5.867	11.276	0.651	3.745	9.354
16	6	0.717	13.921	-0.185	0.755	5.348	10.766	0.690	3.078	8.481
17	6	0.754	12.343	0.015	0.794	5.175	10.590	0.670	3.464	8.997
18	6	0.756	12.129	0.112	0.723	7.075	12.383	0.656	4.319	10.046
19	6	0.687	15.048	-0.062	0.654	7.020	12.334	0.705	2.940	8.289
20	6	0.70	14.476	-0.116	0.688	6.626	11.98	0.687	3.220	8.674

*Model No. (Model number), Hn-No. (number of hidden nodes), nPCs (number of principle components).

In model 15 the trial with 3 and 4 hidden nodes numbers failed to reach 0.75 as R-test, so the trials were neglected.

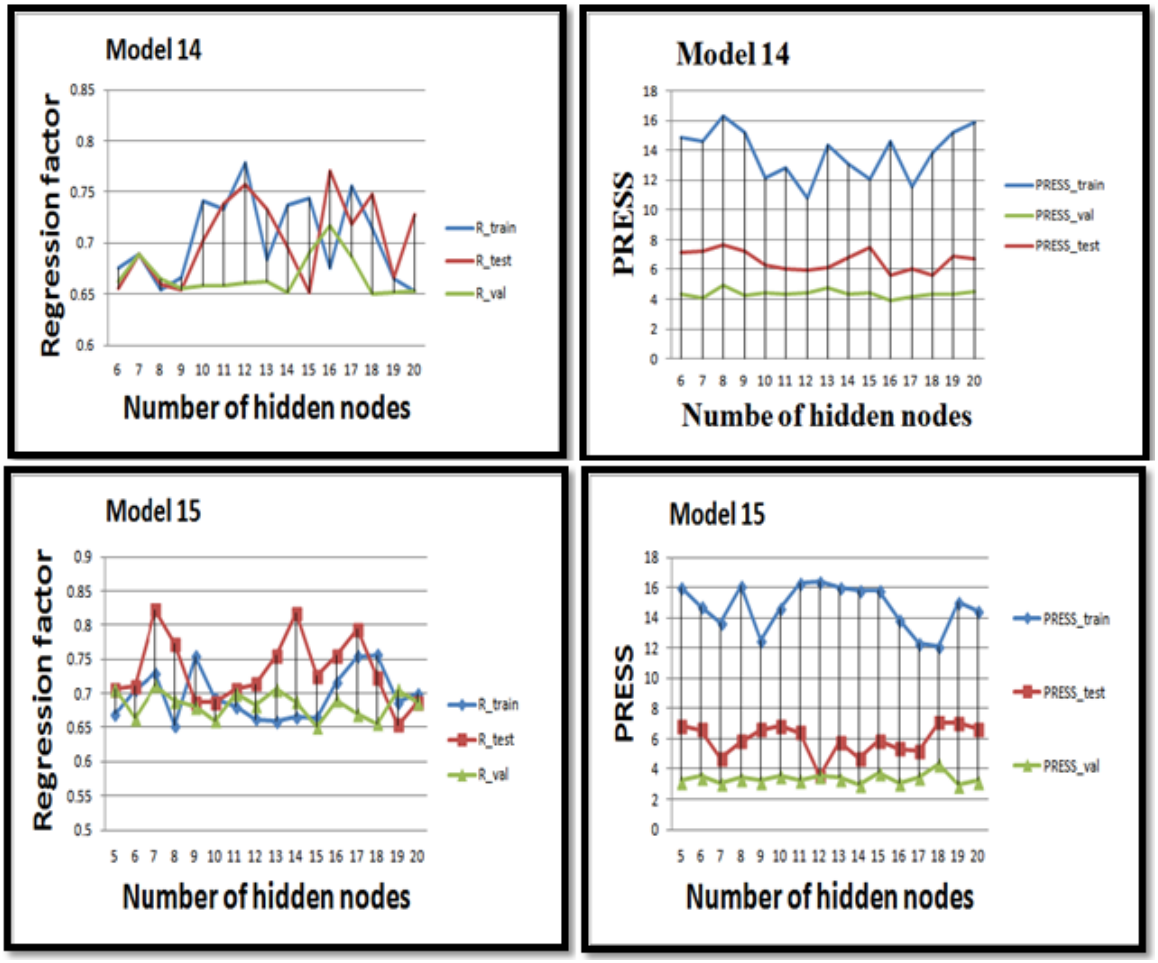


Figure (3.5): PRESS against number of hidden nodes as well as regression factor against number of hidden nodes for model 14 and 15 respectively.

Figure (3.5) shows the PRESS values against the number of hidden nodes as well as the regression factor against number of hidden nodes for models **14 and 15 for part 2**. This figure shows that the lowest PRESS value (5.928) is obtained when using 12 hidden nodes for model **14** with regression coefficient for the test set of 0.757. For model **15**, the lowest PRESS (4.727) is obtained when using 7 hidden nodes with regression coefficient for the highest test set of 0.823.

All models model 14 and model 15 of part 2 were examined to inspect the presence of outliers that may affect models validity. By inspecting the residuals of these models, the residue equals the difference between the predicted and observed one, there were no outliers as it is shown in figures (3.6).

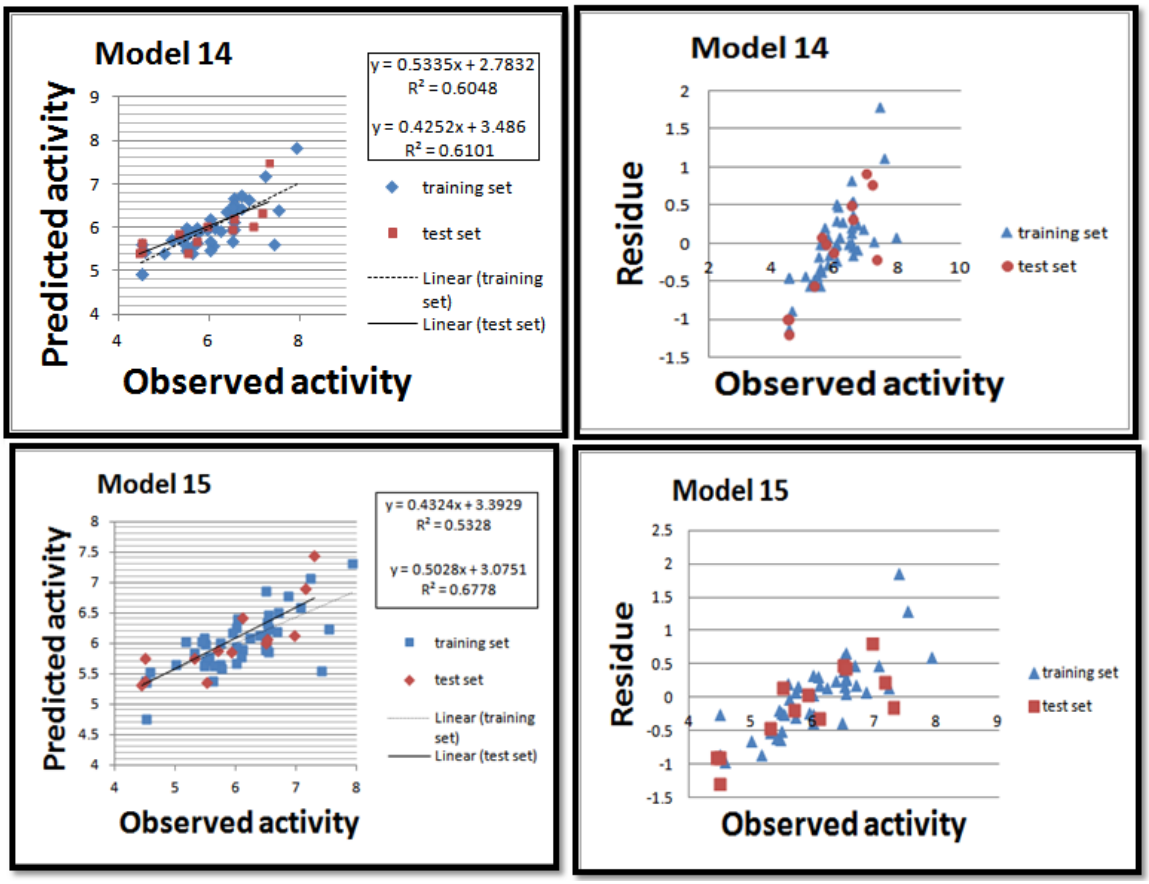


Figure (3.6): Predictive activity against observed one as their residue for part 2 models 14 and 15 using 12 and 7 hidden nodes numbers respectively.

The correlation between calculated and observed pIC_{50} for the training set of model **14** is given by:

Calculated pIC_{50} = 0.534 Observed pIC_{50} + 2.783

And for the test set of this model is given by:

Calculated pIC_{50} = 0.4252 Observed pIC_{50} + 3.486

While the Correlation between calculated and observed pIC_{50} for the training set of model **15** is given by:

Calculated pIC_{50} = 0.432 Observed pIC_{50} + 3.393

And for the test set of this model is given by:

Calculated pIC_{50} = 0.503 Observed pIC_{50} + 3.08

Randomization

Randomization test is performed to investigate the probability of chance correlation for the optimal models (models 14 and 15 with 12 and 7 hidden nodes in the network, respectively, of part 2). Chance correlation was done using the same configuration parameters and the same activation functions of all our ANN models. the results of chance correlation for models **14** (using 12 hidden nodes) and **15** (using 7 hidden nodes) are summarized in the following tables (3.13) and (3.14) respectively. These tables show that the coefficients of determination obtained by chance are low in general while the PRESS values are high. This indicates that the models obtained from ANN are better than those obtained by chance.

As we can see, our models were validated by calculating different statistical parameters, using external test set and finally performing randomization test.

Table (3.13): Statistical parameters of chance correlation of model 14 with 12 hidden nodes. (Part 2)

Trial No.	nPCs	R_train	PRESS_train	R2CV_train	R_test	PRESS_test	R_val	PRESS_val	R2CV_val
1	6	-0.019	62.21	-5.066	-0.116	3.189	-0.006	3.688	-22.013
2	6	0.137	51.255	-3.072	-0.093	2.474	0.157	2.799	-2.102
3	6	-0.061	69.5	-2.482	-0.042	2.849	-0.263	3.043	-7.338
4	6	0.122	51.381	-3.347	-0.047	1.384	0.397	1.398	-20.447
5	6	-0.200	57.773	-8.684	0.249	1.107	-0.036	2.357	-6.273
6	6	-0.176	71.111	-3.402	-0.191	2.456	-0.258	3.464	-1.861
7	6	-0.188	80.24	-3.445	-0.165	2.329	-0.154	2.215	-4.958
8	6	-0.193	58.657	-9.458	0.152	2.347	0.275	2.895	-4.784
9	6	0.218	51.558	-2.354	-0.147	1.647	0.268	1.54	-2.466
10	6	0.193	46.683	-4.436	-0.177	1.37	0.288	1.522	-17.426

Table (3.14): Statistical parameters of chance correlation of model 15 with 7 hidden nodes. (Part 2)

Tri-al No.	nPCs	R_train	PRESS_train	R2CV_train	R_test	PRESS_test	R_val	PRESS_val	R2CV_val
1	6	-0.028	57.208	-4.349	-0.017	1.273	-0.232	2.308	-11.6
2	6	0.076	51.222	-5.105	-0.225	2.609	-0.162	2.143	-5.87
3	6	0.092	110.31	-8.014	-0.112	10.131	-0.043	11.265	-102.793
4	6	0.149	45.089	-11.099	-0.686	1.727	0.155	1.54	-18.123
5	6	0.128	45.614	-13.166	-0.225	1.316	-0.144	1.634	-77.889
6	6	-0.298	57.911	-12.601	-0.228	1.547	0.253	1.652	-34.57
7	6	-0.166	61.472	-5.667	0.142	1.541	0.142	1.561	-11.975
8	6	-0.183	62.8	-6.189	-0.196	2.33	0.283	2.121	-5.291
9	6	-0.102	60.916	-4.952	-0.134	1.407	-0.166	2.858	-20.457
10	6	0.271	45.575	-2.472	0.151	2.093	0.281	2.435	-9.185

Both the external and cross-validation methods are used to validate the performances of the resulting models. Employed randomization test indicates that the models obtained from ANN are better than those obtained by chance.

Comparison with other QSAR studies:

In our study, we developed QSAR models for inhibition activity of 121 chemical compounds of cyclooxygenase-2 inhibitors with various cores and we divided those compounds into two classes tricyclics and non-tricyclics by applying different statistical methods as MLR and PC-ANN. These models will be used to design new COX-2 inhibitors. The numerous QSAR studies in the recent past on COX-2 inhibitors involved small data set of a particular class of compounds with different statistical methods, such as:

- (A sit K. Chakraborti, 2003) In this research, the number of samples was 35 compounds of 1,3-Diarylisoindole and was analyzed using comparative molecular field analysis (CoMFA) and comparative were incorporated to the CoMFA models. The result was (r^2_{cv} =0.536, r^2_{conv} =0.968, SEE=0.222, r=0.6564).[60]

- (S.Prasanna, 2005) In this study, the number of samples was 41 compounds of 2,3-diaryl indoles, in statistically significant linear multiple regression equation with r = 0.942, r^2 = 0.888.[61]

- (M. Khoshneviszadeh, 2008) In this study, the number of samples was 30 compounds (n=30) of 2-Sulfonyl-Phenyl-Indol Derivatives for cox-2 inhibitory activity using chemical, topological, geometrical, and quantum descriptors. Some statistical techniques like stepwise regression, multiple linear regression analysis, and algorithms partial least squares analysis was applied to derive the quantitative structure activity relationship models. The generated equations were statistically validated using cross validation and external test set. The multiple linear regression equation obtained from factor analysis (FA-MLR) as the preprocessing step could predict 77.5% of the variance of the cyclooxygenase-2 inhibitory activity whereas that derived from genetic algorithms partial least squares could predict 84.2% of variances.[62]

- (Shashikant Bhandari, 2009) In this study, 2D and 3D QSAR of series of 80 molecules are containing 4,5 diarylimidazole pharmacophore as selective cyclooxygenase–2 (COX–2) inhibitors. The 3D QSAR studies were performed using two different methods, stepwise variable selection k–nearest neighbor molecular field analysis (SW kNN–MFA) and simulated annealing k–nearest neighbor molecular field analysis (SA kNN–MFA) methods. The 2D QSAR studies were performed using multiple regression, 3D QSAR studies produced cross–validated r^2_{cv} value of 0.688 and 0.733 and conventional r^2 value of 0.912 and 0.794 values using the models SW kNN–MFA and SA kNN–MFA method respectively, whereas the r^2 value in 2D QSAR studies was found to be 0.8943.[63]

- (Girish Kumar, 2012) In this study, the number of compounds was 31 for a series of molecules belonging to tetrasubstituted pyrazoles as COX-II inhibitors. With a correlation coefficient of r^2=0.958, and the squared predictive correlation coefficient of 0.852 was observed between experimental and predicted activity values of test set molecules.[64]

- (Amrita Dwivedi, 2013) The quantitative structure activity relationship (QSAR) study of indole shiff bases to understand the structural features that influence the inhibitory activity toward the cyclooxygenase-2 (COX-2) enzyme. The calculated QSAR results revealed that the drug activity could be modeled by using molecular connectivity indices (^0v,^1v,^2v), wiener index (W) and mean wiener index (WA) parameters. The predictive ability of models was cross validated by evaluating the low residual activity, appreciable cross validated r^2 values (R^2_{cv}) and leave one out (LOO) technique.[65]

-Chapter Four:

Conclusion

Conclusion:

In this study, the QSAR models were built by applying different statistical methods as MLR and PC-ANN for 121 compounds of Cyclooxygenase-2 enzyme inhibitors , we divided those compounds into two parts according to chemical structure as tricyclics (part 1 has 48 chemical compounds) and non-tricyclics (part 2 has 73 chemical compounds).

Each compound was built and optimized by HyperChem software using AM1 semi-empirical method. Then we calculated different groups of descriptors, some of descriptors calculated using HyperChem and the other descriptors by Dragon software.

Multiple linear equations with good statistical qualities and predictive power for both parts were obtained by SPSS software to correlate the activity of each compound with the descriptors. The PC ANN gave better regression models with good prediction ability when we used the MLR equations as inputs for PC-ANN model building. The best PC-ANN models were used for hidden nodes optimization.

The optimal two models of part 1 have prediction coefficients of determinations (R^2) of 0.878 and 0.854. The lowest PRESS obtained is 3.947 and the optimal two models of part 2 have prediction coefficients of determinations (R^2) of 0.677 and 0.573. the lowest PRESS obtained is 4.727. Generally, the models obtained from the ANN analysis are better than those obtained by MLR analysis. But in part 2, the MLR models are slightly better than PC-ANN. Both the external and cross-validation methods are used to validate the performances of the resulting models. Employed randomization test indicates that the models obtained from ANN are better than those obtained by chance.

Part 1 results accepted for publication,[66] while part 2 results submitted for publication.[67]

-References

References:

1) Fiorucci, S., Meli, R., Bucci, M., & Cirino, G. (2001). Dual inhibitors of cyclooxygenase and 5-lipoxygenase. A new avenue in anti-inflammatory therapy? Biochemical Pharmacology, 1433-1438.

2) Vane, J. (1971). Inhibition of Prostaglandin Synthesis as a Mechanism of Action for Aspirin-like Drugs. Nature New Biology, 232-235.

3) Lanas, A. (2001). Cyclo-oxygenase-1/cyclo-oxygenase-2 non-selective non-steroidal anti-inflammatory drugs: Epidemiology of gastrointestinal events. Digestive and Liver Disease, S29-S34.

4) Aisen PS . (2002). Evaluation of selective COX-e inhibitors for the treatment of Alzheimer's disease. J. Pain Symp. Manage23: s35-40.

5) Méric, J., Rottey, S., Olaussen, K., Soria, J., Khayat, D., Rixe, O., & Spano, J. (2006). Cyclooxygenase-2 as a target for anticancer drug development. Critical Reviews in Oncology/Hematology, 51-64.

6) Turini, M., & DuBois, R. (2002). Cyclooxygenase-2: A Therapeutic Target. Annual reviews.

7) Dubois RN, Abramson SB, Crofford L, Gupta RA, Simon LS, Van De Putte LB, Lipsky PE.(1998), Cyclooxygenase in biology and disease, Review., FASEB J.;12(12):1063-73.

8) R.M. Botting, Inhibitors of Cyclooxygenases: Mechanisms, Selectivity and Uses, Journal of Physiology and Pharmacology, 2006, 57, Supp 5, 113-124.

9) Vane, J.R., Bakhle, Y.S., and Botting, R.M. (1998). Cyclooxygenases 1 and 2. Annu. Rev. Pharmacol. Toxicol. 38: 97–120

10) Smith WL, DeWitt DL, Garavito RM, Cyclooxygenases: structural, cellular, and molecular biology, Annual Review of Biochemistry, 2000;69:145-82.

11) Zidar, N., Odar, K., Glavac, D., Jerse, M., Zupanc, T., & Stajer, D. (2009). Cyclooxygenase in normal human tissues - is COX-1 really a constitutive isoform, and COX-2 an inducible isoform? Journal of Cellular and Molecular Medicine, 3753-3763.

12) Morita I. Distinct functions of COX-1 and COX-2. Prostaglandins Other Lipid Mediat 2002; 68-69: 165-175

13) Zarghi A. and Arfaei S., (2011), Selective COX-2 Inhibitors: A Review of Their Structure-Activity Relationships, Iran J Pharm Res. 10(4): 655–683.

14) Picot, D., Loll, P., & Garavito, R. (1994). The X-ray crystal structure of the membrane protein prostaglandin H2 synthase-1. Nature, 243-249.

15) Botting RM. Inhibitors of cyclooxygenases: mechanisms, selectivity and uses. J Physiol Pharmacol.2006;57(Suppl. 5):113–124.

16) Young, D. (2001). Computational chemistry: A practical guide for applying techniques to real world problems. New York: Wiley.

17) Dehmer, M. (2012). Statistical modelling of molecular descriptors in QSAR/QSPR. Weinheim: Wiley-VCH.

18) McDouall, J. (2013). Computational quantum chemistry molecular structure and properties in silico. Cambridge: Royal Society of Chemistry.

19) J. ENOCH, S. (2010). The use of quantum mechanics derived descriptors in computational toxicology. In Recent Advances in QSAR Studies. Springer Netherlands.

20) Jensen, F. (1999). Introduction to computational chemistry. Chichester: Wiley.

21) Dykstra, C. (2005). Semiempirical Quantum-Chemical Methods in Computational Chemistry. In Theory and applications of computational chemistry: The first forty years. Amsterdam: Elsevier.

22) Liew, C., & Yap, C. (2012). Current Modeling Methods Used in QSAR/QSPR. In M. Dehmer (Ed.), Statistical modelling of molecular descriptors in QSAR/QSPR. Weinheim: Wiley-VCH ;.

23) Cronin, M. (2010). Quantitative Structure Activity Relationships (QSARs) – Applications and methodology. In T. Puzyn (Ed.), Recent advances in QSAR studies methods and applications. New York: Springer

24) Deeb, O., & Jawabreh, M. (2012). Exploring QSARs for Inhibitory Activity of Cyclic Urea and Nonpeptide-Cyclic Cyanoguanidine Derivatives HIV-1 Protease Inhibitors by Artificial Neural Network. Advances in Chemical Engineering and Science, 82-100.

25) Consonni, V., & Todeschini, R. (2010). Molecular Descriptors. In T. Puzyn (Ed.), Recent advances in QSAR studies methods and applications (Vol. 8). New York: Springer

26) 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.

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

28) 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.

29) 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.

30) Deeb O. and Sana Jawabreh, Mohammad Goodarzi, (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(12):2237-44.

31) Verma, J., Khedkar, V., & Coutinho, E. (2010). 3D-QSAR In Drug Design - A Review. Current Topics in Medicinal Chemistry, 95-115.

32) Ojha L.K., Chaturvedi A.M., Bhardwaj A., Thakur M. and Thakur A. . (2012). 2D QSAR study of some TIBO Derivatives as an HIV Agent, Asian Journal of Research in Chemistry. 5(3), 377-382.

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

34) Topliss, J.G. and Edwards, R.P. (1979), "Chance factors in studies of quantitative structure-activity relationships", Journal of Medicinal Chemistry, 22, 1238–1244

35) Michael L. Orlov, Multiple Linear Regression Analysis Using Microsoft Excel, http://chemistry.oregonstate.edu/courses/ch361-464/ch464/RegrssnFnl.pdf

36) Montgomery, D. C. and E. A. Peck (1992), Introduction to Linear Regression Analysis, 2nd edition, John Wiley & Sons, New York. (Probability and Statistics Series; 1st edition, 1983).

37) Puzyn, T. (2010). Recent advances in QSAR studies methods and applications. New York: Springer.

38) Rauch Alan.F. -)1997), Appendix D Methods for Multiple Linear Regression Analysis http://scholar.lib.vt.edu/theses/available/etd-219182249741411/unrestricted/Apxd.pdf

39) Randall D. Tobias, An Introduction to Partial Least Squares Regression. http://statistics.ats.ucla.edu/stat/sas/library/pls.pdf

40) Kuz'min, V., Artemenko, A., Muratov, E., Polischuk, P., Ognichenko, L., Liahovsky, A., Varlamova, E. (2010). Virtual Screening And Molecular Design Based On Hierarchical QSAR

Technology. In T. Puzyn (Ed.), Recent advances in QSAR studies methods and applications. New York: Springer.

41) Feng Cheng and Vijaykumar Sutariya, (2012) , Applications of Artificial Neural Network Modeling in Drug Discovery, Clinical and Experimental Pharmacology, Volume 2, Issue 3, 2:3

42) Sutariya, V. (2013). Artificial Neural Network in Drug Delivery and Pharmaceutical Research.The Open Bioinformatics Journal, 49-62.

43) Deeb, O., Khadikar, P., & Goodarzi, M. (2010). QSPR Modeling of Bioconcentration Factors of nonionic Organic compounds. Environmental Health Insights, 33–47-33–47.

44) Zarghi, A.; P. N. Praveen, Rao.; Edward, E. Knaus. (2007). Synthesis and biological evaluation of methanesulfonamide analogues of rofecoxib: Replacement of methanesulfonyl by methanesulfonamido decreases cyclooxygenase-2 selectivity. Bioorg. Med. Chem. 15, 1056-1061.

45) Qiao-Hong Chen, P. N. Praveen Rao, and Edward E. Knaus*, (2006). "Synthesis and biological evaluation of a novel class of rofecoxib analogues as dual inhibitors of cyclooxygenases (COXs) and lipoxygenases (LOXs)" Bioorganic & Medicinal Chemistry, 14: 7898-7909.

46) Md. Jashim, Uddin.; P. N. Praveen, Rao.; Edward E. Knaus. (2004). Design and synthesis of acyclic triaryl (Z)-olefins: a novel class of cyclooxygenase-2 (COX-2) inhibitors. Bioorg. Med. Chem. 12, 5929-5940.

47) Md. Jashim, Uddin.; P. N. Praveen, Rao, Edward, E. Knaus, (2005). Design and synthesis of (Z)-1,2-diphenyl-1-(4-methanesulfonamidophenyl)alk-1-enes and (Z)-1-(4-azidophenyl)-1,2-diphenylalk-1-enes: Novel inhibitors of cyclooxygenase-2 (COX-2) with antiinflammatory and analgesic activity. Bioorg. Med. Chem. 13, 417-424.

48) Zarghi, A.; P. N. Praveen, Rao.; Edward, E. Knaus. (2007), Design and synthesis of new rofecoxib analogs as selective cyclooxygenase-2 (COX-2) inhibitors: Replacement of the methanesulfonyl pharmacophore by a N-acetylsulfonamido bioisostere. J. Pharm. Pharm. Sci. 10, 159-167.

49) Moreau, A.; P. N. Praveen, Rao.; Edward, E. Knaus. (2006), Synthesis and biological evaluation of acyclic triaryl (Z)-olefins possessing a 3,5-di-tert-butyl-4-hydroxyphenyl pharmacophore: Dual inhibitors of cyclooxygenases and lipoxygenases. Bioorg. Med. Chem. 14, 5340-5350.

50) Zarghi, A.; Arfaee, S.; P. N. Praveen, Rao.; Edward, E. Knaus, (2006) Design, synthesis, and biological evaluation of 1,3-diarylprop-2-en-1-ones: A novel class of cyclooxygenase-2 inhibitors. Bioorg. Med. Chem. 14, 2600-2605

51) Qiao-Hong, Chen.; P. N. Praveen, Rao.; Edward, E. Knaus, (2005), Design, synthesis and biological evaluation of linear 1-(4-, 3- or 2-methylsulfonylphenyl)-2-phenylacetylenes: A novel class of cyclooxygenase-2 (COX-2) inhibitors. Bioorg. Med Chem. 13, 6425 -6434

52) Moreau, A.; Qiao-Hong, Chen.; P. N. Praveen, Rao.; Edward, E. Knaus. (2006), Design, synthesis, and biological evaluation of (E)-3-(4-methanesulfonylphenyl)-2-(aryl)acrylic acids as dual inhibitors of cyclooxygenases and lipoxygenases. Bioorg. Med. Chem. 14, 7716-7727

53) Qiao-Hong, Chen.; P. N. Praveen, Rao.; Edward, E. Knaus. (2005) Design, synthesis, and biological evaluation of N-acetyl-2-(or 3-)carboxymethylbenzenesulfonamides as cyclooxygenase isozyme inhibitors. Bioorg. Med. Chem. 13, 4694-4703.

54) Qiao-Hong, Chen.; P. N. Praveen, Rao.; Edward, E. Knaus. (2005), Design, synthesis, and biological evaluation of N-acetyl-2-carboxybenzenesulfonamides: A novel class of cyclooxygenase-2 (COX-2) inhibitors. Bioorg. Med. Chem. 13, 2459-2468

55) Zarghi, A.; Zebardast, T.; Hakimion, F.; Shirazi, F. H.; P. N. Praveen, Rao.; Edward, E. Knaus. (2006), Synthesis and biological evaluation of 1,3-diphenylprop-2-en-1-ones possessing a methanesulfonamido or an azido pharmacophore as cyclooxygenase-1/-2 inhibitors. Bioorg. Med. Chem. 14, 7044-7050

56) Anana, R.; P. N. Praveen, Rao.; Qiao-Hong, Chen.; Edward, E. Knaus. (2006) Synthesis and biological evaluation of linear phenylethynylbenzenesulfonamide regioisomers as cyclooxygenase-1/-2 (COX-1/-2) inhibitors. Bioorg. Med. Chem. 14, 5259-5265.

57) Morshed Alam Chowdhury; Ying Dong, Qiao-Hong Chen, Khaled R A Abdellatif; Edward E Knaus, (2008), Synthesis and cyclooxygenase inhibitory activities of linear 1-(methanesulfonylphenyl or benzenesulfonamido)-2-(pyridyl)acetylene regioisomers. Bioorg. Med. Chem. 16, 1948-1956.

58) O. Deeb, (2010) "Correlation ranking and stepwise regression procedures in PC-ANN modeling and application to predict the toxic activity and HSA binding affinity". Chemometrics and Intelegent Laboratory Systems.; 104181-194.

59) Golbraikh, A., Tropsha, A. (2002), Beware of q2!, Journal of Molecular Graphics and Modelling. 20, 269-276.

60) K. Chakraborti and R. Thilagavathi, (2003), Computer-Aided Design of Non Sulphonyl COX-2 Inhibitors: An Improved Comparative Molecular Field Analysis Incorporating

Additional Descriptors and Comparative Molecular Similarity Indices Analysis of 1,3-Diarylisoindole Derivatives, Bioorganic and Medicinal Chemistry 11 3989-3996,

61) S.Prasanna, E. Manivannan and S. C. Chaturvedi, (2005), Quantitative structure–activity relationship analysis of 2,3-diaryl indoles as selective cyclooxygenase-2 inhibitors, Journal of Enzyme Inhibition and Medicinal Chemistry, 20(5): 455-461

62) M. Khoshneviszadeh, N. Edraki, R. Miri and B. Hemmateenejad, (2008), Exploring QSAR for Substituted 2-Sulfonyl-Phenyl-Indol Derivatives as Potent and Selective COX-2 Inhibitors Using Different Chemometrics Tools, Chem Biol Drug Des 72: 564-574.

63) Shashikant Bhandari, Kailash Bothara,Vidya Pawar, Deepak Lokwani, and Titiksh Devale, (2009), Design of New Chemicals Entities as Selective COX–2 Inhibitors using Structure Optimization by Molecular Modeling Studies, Internet Electronic Journal of Molecular Design, 8, 14–28.

64) Girish Kumar Gupta and Ajay Kumar, (2012), 3D-Qsar Studies Of Some Tetrasubstituted Pyrazoles As COX-II Inhibitors, Acta Poloniae Pharmaceutica - Drug Research, Vol. 69 No. 4 pp. 763-772.

65) Amrita Dwivedi, Ajeet Singh, A.K. Srivastava , (2013), Quantitative structure–activity relationship based modeling of substituted indole Schiff bases as inhibitor of COX-2, Journal of Saudi Chemical Society.

66) Deeb O. and Zatari N., (2014), Exploring Quantitative Structure-Activity Relationships (QSARs) of Cyclooxygenase-2 (COX-2) Inhibitors by MLR and PC-ANN, Journal of Engineering Science and management Education, Special issue on (theoretical / Experimental Aspects of Drug Design). (In Press). **(Part 1)**

67) Deeb O. and Zatari N., (2014), Exploring Quantitative Structure-Activity Relationships (QSARs) of Non-Tri cyclic Cyclooxygenase-2 (COX-2) Inhibitors by MLR and PC-ANN, Journal of Advances in Chemistry. (In Press) **(Part 2)**

دراسة العلاقة بين الصيغ البنائية والفاعلية باستخدام طريقة PC-ANN, MLR, PLS لمثبطات إنزيم COX-2.

مقدمة من :

نائل يحيى محمد زعتري

بكالوريوس صيدلة جامعة العلوم التطبيقية الأردن

بإشراف: د. عمر ديب

قدمت هذه الرسالة استكمالا لمتطلبات درجة الماجستير في

التكنولوجيا التطبيقية والصناعية

كلية العلوم والتكنولوجيا

جامعة القدس

2014/1436