

Deanship of Graduate Studies

Al-Quds University



**Exploring QSARs of some Translocator protein (TSPO)
ligands using MLR and PC-ANN techniques**

Hanaa Saleem Hussein Baniowda

M.Sc. Thesis

Jerusalem-Palestine

1437 / 2016

**Exploring QSARs of some Translocator protein (TSPO)
ligands using MLR and PC-ANN techniques**

Prepared By:

**Hanaa Saleem Hussein Baniowda
B.Sc. Pharmacy Al-Quds University/ Palestine**

Supervisor: Prof. Omar Deeb

**A thesis submitted to the Faculty of the
Graduate Studies of Pharmacy of Al-Quds University
in partial fulfillment of the requirements for the degree
of Master of Pharmaceutical Sciences**

1437/2016

Al-Quds University
Deanship of Graduate Studies
Pharmaceutical Sciences Program



Thesis Approval

Exploring QSARs of some Translocator protein (TSPO) ligands using MLR and PC-ANN techniques

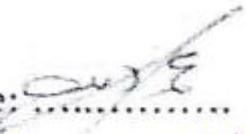
Prepared by: Hanaa Saleem Hussein Baniowda

Registration No.: 21210100

Supervisor: Prof. Omar Deeb

Master thesis Submitted and Accepted Date: 5/1/2016, the names and signatures of the examining committee members are as follows:

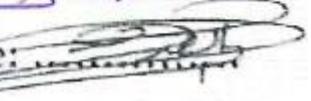
Head of Committee: Prof. Omar Deeb

Signature: 

Internal Examiner: Prof. Rafik Karaman

Signature: 

External Examiner: Dr. Nasr Shriam

Signature: 

Jerusalem–Palestine

1437 / 2016

Dedication

I would like to express my sincere gratitude to my family and my friends who have been a constant source of support and encouragement during the challenges of graduate school and life.

Declaration

I certify that the 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 be submitted for a higher degree to any other university or institution.

Signed:

Hanaa Saleem Hussein Baniowda

Date: 5 /1/2016

Acknowledgment

Foremost, I am highly grateful to God for His blessing that continue to flow into my life, and because of him, I made this through against all challenges.

Also I would like to express my special appreciation and thanks to my academic supervisor Professor Omar Deeb, I would like to thank him for the helpful guidance, continuous support, supervision throughout the different stages of this research, feedback and advice he has provided me with his respective areas of expertise.

Exploring QSARs of some Translocator protein (TSPO) ligands using MLR and PC-ANN techniques

Prepared by: Hanaa Saleem Hussein Baniowda

Supervisor: Prof. Omar Deeb

Abstract

Quantitative structure-activity relationship study was performed to understand the activity of a set of 136 ligands of Translocator protein (TSPO) compounds. QSAR models were developed using multiple linear regression (MLR) as linear method. While principal component - artificial neural networks (PC-ANN) modeling method was used as nonlinear method. The results obtained offer good regression models having good prediction ability.

The MLR resulted with models (12-24) which have coefficient of determination (R^2) > 0.6, the best model (number 24) resulted with correlation coefficient (R) = 0.909, coefficient of determination (R^2) = 0.826, and adjusted coefficient of determination (R^2_{adj}) = 0.788.

Cross Validation leave one out (LOO) and leave many out (LMO) were performed on the resulted MLR models, models 19-24 showed a good predictive power. After that principle component analysis (PCA) performed to divide the data into three data sets, then the ANN performed on the chosen models (19-24) from leave one out (LOO) and leave many out (LMO) validation.

ANN resulted models were validated through randomization test, then the conditions proposed by Golbraikh and Tropsha were applied to conclude that the QSAR models has acceptable prediction power or not. However the best ANN model with a good predictive power was model #24, with R test values 0.832.

Table of Contents

LIST OF TABLES	VI
LIST OF FIGURES	VII
LIST OF ABBREVIATIONS	VIII
1.CHAPTER ONE: INTRODUCTION	2
1.1 OVERVIEW OF COMPUTATIONAL CHEMISTRY	2
1.2 QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIPS (QSAR)	4
1.2.1 QSAR HISTORY	5
1.2.2 QSAR ADVANTAGES AND DISADVANTAGES	6
1.3 QSAR MODEL DEVELOPMENT STEPS.....	7
1.3.1 DATA PREPARATION	7
1.3.2 DATA ANALYSIS.....	8
1.3.2.1 LINEAR MODELS.....	8
1.3.2.2 NONLINEAR MODELS.....	9
1.3.3 MODEL VALIDATION.....	12
1.4 SOFTWARE USED IN QSAR PROCESS.....	15
1.4.1 HYPERCHEM.....	15
1.4.2 DRAGON	16
1.4.3 SPSS	17
1.4.4 MATLAB	18
1.5 TRANSLOCATOR PROTEIN (TSPO) PREVIEW:	19
1.6 RESEARCH OBJECTIVE.....	23
2.CHAPTER TWO: METHODOLOGY.....	25
2.1 DATA PREPARATION	25
2.1.1 DATASET	25
2.1.2 COMPOUNDS OPTIMIZATION	33
2.1.3 DESCRIPTORS CALCULATION	35
2.1.3.1 DESCRIPTORS CALCULATED BY HYPERCHEM	36
2.1.3.2 DESCRIPTORS CALCULATED BY DRAGON.....	38
2.1.3.2.1 BRIEF DESCRIPTION ABOUT DRAGON DESCRIPTORS:.....	38
2.1.3.2.2 STEPS TO PERFORM DESCRIPTORS CALCULATION USING DRAGON SOFTWARE:....	39
2.2 DATA ANALYSIS.....	41
2.2.1 MULTIPLE LINEAR REGRESSION (MLR)	41
2.2.1.1 STEPS TO PERFORM MLR FOR EACH DESCRIPTOR GROUP USING SPSS:	41
2.2.1.2 STEPS TO PERFORM MLR FOR ALL THE DESCRIPTORS RESULTED FROM THE FIRST MLR USING THE SPSS:.....	43

2.2.2 MLR MODEL VALIDATION	44
2.2.2.1 CROSS-VALIDATION.....	44
2.2.2.1.1 STEPS TO PERFORM LEAVE ONE OUT (LOO) USING MATLAB.....	44
2.2.2.1.2 STEPS TO PERFORM LEAVE MANY OUT (LMO) USING MATLAB.....	45
2.2.3 PRINCIPAL COMPONENT ANALYSIS (PCA)	46
2.2.4 ARTIFICIAL NEURAL NETWORKS (ANN).....	47
2.2.4.1 STEPS TO PERFORM ANN FOR EACH MODEL USING MATLAB:	47
2.2.4.2 STEPS TO PERFORM ANN OF THE BEST MODELS WITH RANGE OF HIDDEN NODES (HN) USING MATLAB:.....	48
2.2.5 RANDOMIZATION TEST (CHANCE CORRELATION OR SCRAMBLING MODEL).....	48
3. CHAPTER THREE:RESULTS AND DISCUSSION	50
3.1 DATA PREPARATION RESULTS	50
4. CHAPTER FOUR: CONCLUSIONS	87
REFERENCES	89

LIST OF TABLES

Table 1-1:	DRAGON descriptors blocks	17
Table 2-1:	Dataset, Compounds have activity against TSPO	25
Table 2-2:	Brief description of some of the descriptors used in this study	40
Table 2-3:	The format of the input file in SPSS to perform MLR (The activities of all compounds and their corresponding properties.	41
Table 3-1:	MLR Models resulted from each group of descriptors	52
Table 3-2:	MLR Models resulted from all the groups of descriptors together	56
Table 3-3:	Brief description of the descriptors in the best MLR model equation	59
Table 3-4:	LOO cross validation results	61
Table 3-5:	LMO cross validation results	62
Table 3-6:	Correlation Coefficient and Cross Validation Parameters for ANN Models 19-24	65
Table 3-7:	Correlation Coefficients and Cross Validation Parameters of Number of Hidden Nodes for Model #20	68
Table 3-8:	Correlation Coefficients and Cross Validation Parameters of Number of Hidden Nodes for Model #21	69
Table 3-9:	Correlation Coefficients and Cross Validation Parameters of Number of Hidden Nodes for Model #23	70
Table 3-10:	Correlation Coefficients and Cross Validation Parameters of Number of Hidden Nodes for Model #24	71
Table 3-11:	Summary of the Correlation Coefficients and Cross Validation Parameters of the Optimal Number of Hidden Nodes of Each Model	72
Table 3-12:	Chance Correlation of Model 23 with 5 Hidden Nodes	74
Table 3-13:	Chance Correlation of Model 24 with 7 Hidden Nodes	75

LIST OF FIGURES

Figure 1-1:	The Artificial Neural Network	10
Figure 1-2:	Root Mean Squares Error (RMSE) equation	13
Figure 1-3:	Cross-validation equation	14
Figure 1-4:	HyperChem display screen	16
Figure 1-5:	SPSS display screen	18
Figure 2-1:	Drawing using HyperChem	35
Figure 2-2:	QSAR Properties dialog box in HyperChem software	38
Figure 2-3:	Choosing linear regression analysis using SPSS	42
Figure 2-4:	Dialog box which open after choosing the linear regression analysis	43
Figure 2-5:	MATLAB Command window asking for file name, model number and number of hidden nodes	47
Figure 3-1:	Second and third principal components plot	63
Figure 3-2:	Plots of ANN Predictive Residual Sum of Squares (PRESS) values for the training, test and validation sets versus model number	66
Figure 3-3:	Plots of ANN correlation coefficient (R) values for the training, test and validation sets versus model number	66
Figure 3-4:	Plot of ANN R ² CV (Cross validated correlation coefficient) values for the training, test and validation sets versus model number	67
Figure 3-5:	Plot of the predicted activity against observed one as well as their residues for model 23 using 5 hidden nodes. Training set, validation set, and external test set	76
Figure 3-6:	Plot of the predicted activity against observed one as well as their residues for model 24 using 7 hidden nodes. Training set, validation set, and external test set	77

LIST OF ABBREVIATIONS

AM1	Austin Model 1.
ANN	Artificial Neural Networks.
ANT	Adenine nucleotide translocase.
CBR	Central benzodiazepine receptor.
Hn	Hidden nodes.
HOMO	Highest occupied molecular orbital.
LMO	Leave many out.
LOO	Leave one out.
LUMO	Lowest unoccupied molecular orbital.
MLR	Multiple linear regression.
MPTP	Mitochondrial permeability transition pore.
PBR	Peripheral benzodiazepine receptor.
PCA	Principal component analysis.
pIC₅₀	Activity.
PRESS	Predicted residual sum of squares.
PSE	Predictive Square Errors.
QSAR	Quantitative structure activity relationships.
R	Correlation coefficient.
R²	Coefficient of determination.
R² adj	Adjusted R ² .
R²_{CV} or Q²	Cross-validated coefficient of determination.
RMSE	Root mean-squared error.
RSEP	Relative Standard Error of Prediction.
SPRESS	Uncertainty of prediction
SSE	Error sum of squares.
SST	Total sum of squares.
TSPO	Translocator protein.
VDAC	Voltage dependent anion channel.

χ^2

Chi-squared test.

Chapter one

Introduction

1. Introduction:

1.1 Overview of Computational Chemistry

Recently, there have been ways to approach chemistry problems: non-computational quantum chemistry and computational quantum chemistry.

Non-computational quantum chemistry deals with the formulation of analytical expressions for the properties of molecules and their reactions while computational quantum chemistry is primarily concerned with the numerical computation of molecular electronic structures. Thus In this research the Computational chemistry is used to solve the research problem [1].

Computational chemistry, alternatively sometimes called theoretical chemistry or molecular modeling. It is a field that can be said to be both old and young. It is old in the sense that its foundation was laid with the development of quantum mechanics in the early part of the twentieth century. However It is young, because computer technology has developed in the last 35 years or so [2].

The term computational chemistry is usually used when a mathematical method is sufficiently well developed that it can be automated for implementation on a computer. Thus, computational chemistry is the application of chemical, mathematical and computing skills by using computers in order to generate information such as properties of molecules or simulated experimental results to find the solution of interesting chemical problems [3].

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

This branch of chemistry which generates data to complement experimental data on the structures, properties and reactions of substances [3]. Its calculations are based primarily on Schrödinger's equation (Equation 1-1) [4] and include:

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

$$H\Psi = E\Psi \dots\dots\dots (1-1)$$

Where, H: Hamiltonian operator

Ψ : psi, the wave function

E: total energy of the system

Therefore the helpful applications of computational chemistry have been widely utilized in the medicinal chemistry field. For example, it has allowed researchers to highlight the molecular basis of ligand-receptor interactions; define the pharmacophoric portion of known active ligands and the hindering regions of the inactive ones; build the three-dimensional structure of the unresolved proteins using homology against a known template; design new ligands and predict their binding mode and affinities; and evaluate the crucial properties of compounds for their absorption, distribution, metabolism, and excretion. Indeed, all the steps of a medicinal chemistry workflow could be potentially realized in a virtual mode *in silico*, and if they are performed with competence and profitable criticism, they rationally guide the experimental phases of research and decrease productive costs.

The computational studies can give better results when many experimental data are available, providing a strong background for the calculations. Usually, the methodology has to be chosen on the basis of the amount and type of existing trial results in a certain topic: it could be *ligand-based* if only the information about known ligands and their activities on the target are used in the calculations as in this study or *receptor-based* if the three dimensional structure of the target is utilized to analyze the interaction with different ligands. If both kinds of experimental data are available, a robust computational procedure can be performed combining the *ligand-based* methods (like quantitative structure activity relationship (QSAR), 3DQSAR, and pharmacophoric studies) with *receptor-based* methods (like docking and its applications).

1.2 Quantitative structure activity relationships (QSAR)

QSAR major goal is to formulate mathematical relationship between physico-chemical properties of compounds and their biological response in the system of interest, or with any other endpoint than the biological response such as chemical, physical and pharmaceutical properties. Hansch pioneered this field by demonstrating that the biological activities of drug molecules can be correlated to a few variables (Properties) using simple regression equation (Equation 1-2) [5], and after determination of this correlation there will be two expected outputs of the QSAR modeling; Firstly, enhance understanding of the specifics of drug action. Secondly, provide a theoretical foundation for future leading optimization.

$$\text{Log (1/C)} = a \text{ (lipophilic descriptor)} + b \text{ (Electronic descriptor)} + c \text{ (Steric descriptor)} + d \text{ (other descriptors)} + \text{etc.} \quad \dots\dots\dots \text{(1-2)}$$

Where,

1/C = Measure of biological activity

a, b, c, etc. = Regression coefficients

1.2.1 QSAR History

More than a century ago, Crum-Brown and Fraser expressed the idea that the physiological action of a substance was a function of its chemical composition and constitution [6]. A few decades later, in 1893, Richet showed that the cytotoxicities of a diverse set of simple organic molecules were inversely related to their corresponding water solubility [7]. At the turn of the 20th century, Meyer and Overton independently suggested that the narcotic (depressant) action of a group of organic compounds paralleled their olive oil/water partition coefficients [8, 9]. In 1939 Ferguson introduced a thermodynamic generalization to the correlation of depressant action with the relative saturation of volatile compounds in the vehicle in which they were administered [10]. The extensive work of Albert, and Bell and Roblin established the acids in bacteriostatic activity [11, 12]. Meanwhile on the physical organic front, great strides were being made in the delineation of substituent effects on organic reactions, led by the seminal work of Hammett, which gave rise to “sigma-rho” [13]. Taft devised a way for separating polar, steric, and resonance effects and introducing the first steric parameter, E_s . The contributions of Hammett and Taft together laid the mechanistic basis for the development of the QSAR paradigm by Hansch and Fujita. In 1962 Hansch and Muir published their brilliant study on the structure-activity relationships of plant growth regulators and their dependency on Hammett constants and hydrophobicity [14].

An early example of QSAR in drug design involves a series of 1-(X-phenyl)-3, 3-dialkyl triazenes. These compounds were of interest for their anti-tumor activity, but they also were mutagenic. QSAR was applied to understand how the structure might be modified to reduce the mutagenicity without significantly decreasing the anti-tumor activity. Based on equations it was observed that mutagenicity is more sensitive than anti-tumor activity to the electronic effects of the substituents. Thus, electron-withdrawing substituents were

examined by substituting a sulfonamide group at the para position, the anti-tumor activity was reduced 1.2-fold, whereas the mutagenicity was reduced by about 400-fold [15].

In the last ten years in Al-Quds computational chemistry laboratory, several QSAR studies have been applied to predict compounds properties, including biological activity, physical property, etc. [16-18].

1.2.2 QSAR advantages and disadvantages

QSAR is important in drug development process. It provides quantitative relationship between structure and activity, in which help understanding the effect of structure on activity. It can also be used to help understand the interactions between functional groups in the molecules of greatest activity with those of their target. Besides, it helps in making predictions leading to the synthesis of novel analogues. Thus using QSAR in new drug development process decreases the cost of new drug development.

On the other hand there is chance of false correlation between structure and activity; which may arise firstly because of biological data that came from a considerable experimental error. Secondly because of the dataset size; if it is not large enough, the data collected may not reflect the complete property space.

Consequently, many QSAR results cannot be used to confidently predict the most likely compounds of best activity. However there are many successful applications but do not expect QSAR works all time [19, 20].

1.3 QSAR model development steps

QSAR model development process is typically performed in successive steps divided into three steps; Data preparation, data analysis, and model validation [21]:

1.3.1 Data preparation

Data preparation starts by selection of the data set to be used; which composed of compounds and their certain activity or any other endpoint, and this may simply be extracted from a database or may need additional experimental studies. And after that do a geometry optimization of the data set compounds; which is finding the coordinates that represents the minimum potential energy for the molecular structure in its 3D form, this can be done using software such as HyperChem which will be used in our study.

Computational optimization encompasses a variety of mathematical methods which fall into two broad categories:

- Molecular mechanics—applies the laws of classical physics to molecular nuclei without explicit consideration of electrons.
- Quantum mechanics—relies on the Schrödinger equation to describe a molecule with explicit treatment of electronic structure. It is divided into two methods of calculations:

1. *Ab initio*, the term is Latin for "from scratch". And it was first used by Robert Parr and coworkers. *Ab initio* is a group of methods in which molecular structures can be calculated using nothing but the Schrödinger equation, the values of the fundamental constants and the atomic numbers of the atoms present.
2. Semi-empirical techniques use approximations from empirical (experimental) data to provide the input into the mathematical models. And this method is preferred because it is faster than the *ab initio* method [22].

After geometry optimization using semi-empirical method in our study, the descriptors (properties) should be calculated using HyperChem and Dragon software.

1.3.2 Data analysis

The models building step in which a correlation between the endpoint and certain descriptors is determined. If the correlation models to be built are linear then the multiple linear regression (MLR) is used, however if it is nonlinear then the artificial neural network (ANN) is performed after the MLR.

Among the widely utilized algorithms applied for model construction in QSAR, in our study we will use multiple linear regression (MLR), and Principle Component artificial neural networks (PC-ANN).

1.3.2.1 Linear Models

✓ Multiple linear regression (MLR)

MLR simultaneously considers the relationship between some independent variables and a dependent variable by fitting a linear equation to observed data. Generally, the multiple linear regression model represented in (Equation 1-3) [23]:

$$Y_i = \alpha + \beta_1 X_{i,1} + \dots + \beta_n X_{i,n} + \epsilon_i \dots\dots\dots (1-3)$$

Where:

α : is the intercept

β_1 - β_n : are slopes or coefficients of independent variables.

$X_{i,1}$ - $X_{i,n}$: are independent variables.

ϵ_i : is the error term.

The MLR is the first statistical step that done because of the assumption that there is a linear correlation between the independent variables (descriptors) and the response variable (Y, Activity in our study).

1.3.2.2 Nonlinear Models

✓ Principal component analysis (PCA)

Principal components analysis (PCA) also known as Eigenanalysis, is a statistical technique for analyzing data. Essentially, a set of correlated variables is transformed into a set of uncorrelated variables, which are ordered by reducing variability. The uncorrelated variables are linear combinations of the original variables, and the last of these variables can be removed with minimum loss of real data [24].

- Artificial Neural Networks (ANN)

Artificial neural network (ANN) analysis is a new method of data analysis, which inspired from the nervous system's way of working in processing information [25]. The nervous system as brain has approximately 100 billion neurons, which communicate through electro-chemical signals. The neurons are connected through junctions called synapses. Each neuron receives thousands of connections with other neurons, constantly receiving incoming signals to reach the cell body. If the resulting sum of the signals surpasses a certain threshold, a response is sent through the axon [26]. The ANN attempts to recreate the computational mirror of the biological neural network, although it is not comparable since the number and complexity of neurons which used in a biological neural network is many times more than those in an artificial neutral network.

ANN is composed of a large number of highly interconnected processing elements called artificial neurons (also known as "nodes"), classified into three layers of neurons, input nodes, hidden nodes, and output nodes as seen in (Figure1-1). The neurons work in unison to solve complicated non-linear problems of multivariate systems [27]. Where the input nodes take in information, in the form which can be numerically expressed.

The information is presented as activation values, where each node is given a number, the higher the number, the greater the activation. This information is then passed throughout the network. Based on the connection strengths (weights), inhibition or excitation, and transfer functions, the activation value is passed from node to node. Each of the nodes sums the activation values it receives; it then modifies the value based on its transfer function. The activation flows through the network, through hidden layers, until it reaches the output nodes. The output nodes then reflect the input in a meaningful way to the outside world [28].

As a result of its component the ANN has a remarkable ability to derive meaning from complicated or imprecise data, and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

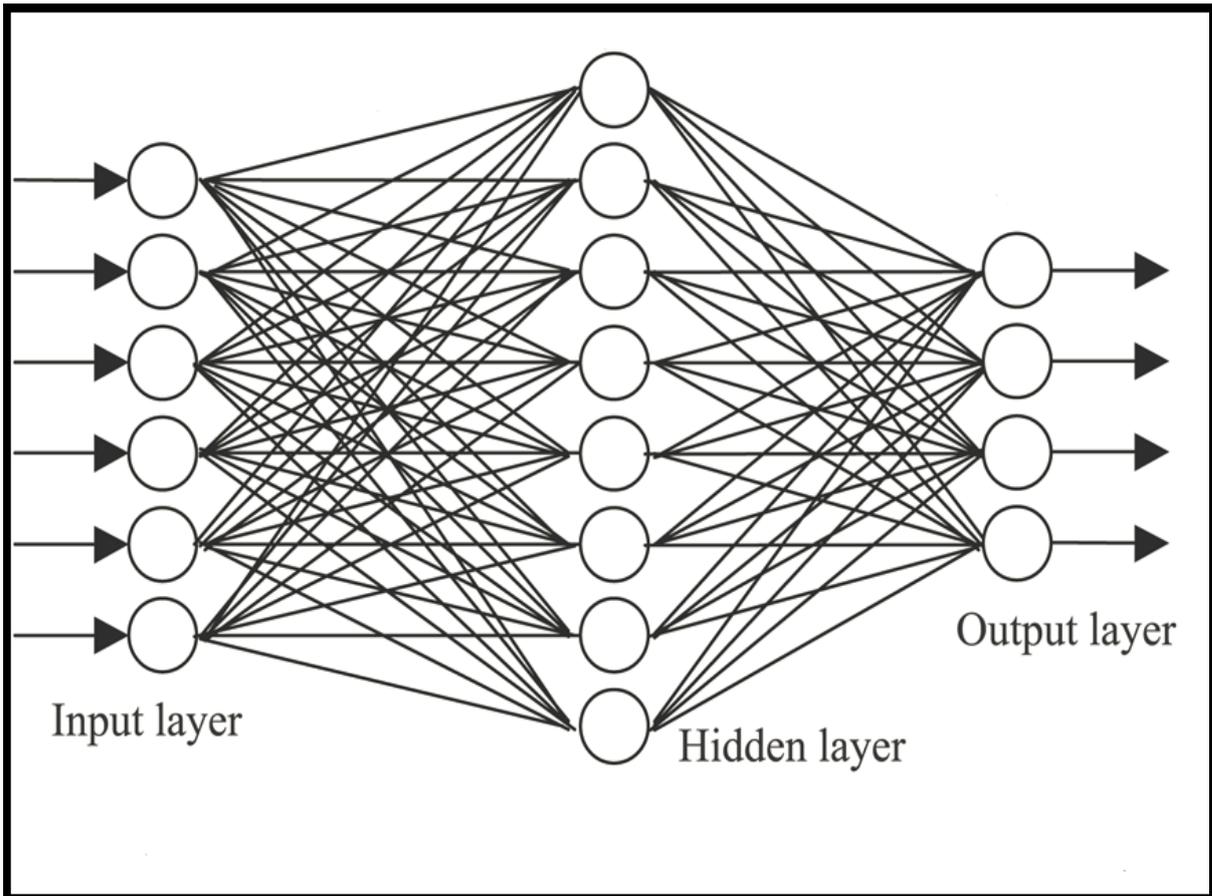


Figure (1-1): The Artificial Neural Network

Neural network simulations appear to be a recent development. However, this field was established before the advent of computers, where the first neuron was produced in 1943 by the neurophysiologist Warren McCulloch and the logician Walter Pitts. Accordingly a wide variety of ANNs are developed and used to model real neural networks, and study behavior and control in animals and machines, but also there are ANNs which are used for engineering purposes, such as pattern recognition, forecasting, and data compression.

Neural networks take a different approach to problem solving than that of conventional computers. Conventional computers use an algorithmic approach i.e. the computer follows a set of instructions in order to solve a problem. Unless the specific steps that the computer needs to follow are known the computer cannot solve the problem. That restricts the problem solving capability of conventional computers to problems that we already

understand and know how to solve. But computers would be so much more useful if they could do things that we don't exactly know how to do. However ANN offer a number of advantages, including requiring less formal statistical training, ability to implicitly detect complex nonlinear relationships between dependent and independent variables, ability to detect all possible interactions between predictor variables, and the availability of multiple training algorithms. Plus all these advantages the ANN is easy to use and understand compared to statistical methods. It is non-parametric model while most of statistical methods are parametric model that need higher background of statistic [29].

In the other hand, because the ANN finds out how to solve the problem by itself, its operation can be unpredictable.

1.3.3 Model validation

Quantitative Structure Activity Relationship (QSAR) is based on the hypothesis that changes in molecular structure reflect changes in the observed response or biological activity. The success of any quantitative structure–activity relationship model depends on the accuracy of the input data, selection of appropriate descriptors, statistical tools and the validation of the developed model. Validation is a crucial aspect of QSAR modeling. Validation is the process by which the reliability and significance of a procedure are established for a specific purpose [30].

QSAR model validation performed either by using the data that created the model (an internal validation) or by using a separate data set (an external validation). The internal validation are: least squares fit (R^2), cross-validation (Q^2) [31, 32], adjusted R^2 (R^2 adj), chi-squared test (χ^2), root mean-squared error (RMSE), bootstrapping and scrambling (Y-Randomization) [33, 34].

The external method is performed by comparing the predicted and observed activities of an (sufficiently large) external test set of compounds that were not used in the model development.

In current research, two internal validation methods have been used; cross-validation and scrambling (Y-Randomization).

Cross-validation

Cross-validation (CV, Q^2 , q^2 , or jack-knifing) is a common method for internal validation of a QSAR model. CV process repeats the regression many times on subsets of data. Usually each molecule is left out once (leave one out, LOO), in turn. Sometimes more than one molecule (leave many out, LMO) is left out at a time.

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 q^2 (R^2_{cv}) and root mean squares error (RMSE), Figure 1-2. High R^2_{cv} and low RMSE values is a result of good and more predictive model and that lead to better description of the observed data. As well as the difference between coefficient of determination (R^2) and Q^2 value should not exceed 0.3 for good predictability.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (X_{obs,i} - X_{model,i})^2}{n}}$$

Figure 1-2: Root Mean Squares Error (RMSE) equation. Where, X_{obs} is the observed values, X_{model} is the mean of the experimental bioactivities, and n is the number of molecules in the set of data being examined.

The cross-validation outcome R^2 (Q^2) equation as seen in (Figure 1-3), which is frequently used as a criterion of both robustness and predictive ability of the model. Many authors consider high Q^2 (for instance, $Q^2 > 0.5$) as an indicator or even as the ultimate proof of the high predictive power of the QSAR model. Therefore if the model have high predictive ability, then there is no need to test the models for their ability to predict the activity of compounds of an external test set [30].

$$Q^2 = 1 - \frac{PRESS}{\sum_{i=1}^N (y_i - y_m)^2}$$
$$PRESS = \sum_{i=1}^N (y_{pred,i} - y_i)^2$$

Figure 1-3: Cross-validation equation. Where PRESS is the predictive residual sum of the squares, y_i is the experimental bioactivity for an individual compound in the training set, and y_m is the mean of the experimental bioactivities

Randomization test (Scrambling model)

Randomization test (Scrambling model) is the second internal validation test performed in this research, which helps to ensure that the model is not due to a chance.

The test performed by randomization of the dependent variables, in which the set of activity values is reassigned randomly to different molecules, and repeating the entire modeling procedure. This process is repeated many times. If the random models activity prediction is comparable to the original equation, then the predictive power of the model is poor and the observations are not sufficient to support the model [30].

1.4 Software used in QSAR process

There are many software available for QSAR development. These include specialized software for drawing chemical structures, interconverting chemical file formats, generating 3D structures, calculating chemical descriptors, developing QSAR models, and general-purpose software that have all the necessary components for QSAR development [35].

In the current research four softwares used; HyperChem (version 8.3 HyperChem, Inc.), Dragon software (version 2.1, Todeschini, R., Milano Chemometrics and QSAR Group. Different statistical packages such as: SPSS software (version 20, SPSS Inc.), MATLAB (version 6.50, Mathworks Inc.).

1.4.1 HyperChem

HyperChem is a sophisticated molecular modeling environment that is flexible, ease of use with high quality (Figure 1-4). As it combines 3D visualization and animation with quantum chemical calculations, molecular mechanics, and dynamics, HyperChem used to draw simple and complex molecular structures, structure optimization, calculate some QSAR properties, and use its output file as input file to another program called "Dragon" to calculate more structure related descriptors. In this research we drew the compounds structures and optimize each one then calculate certain properties.

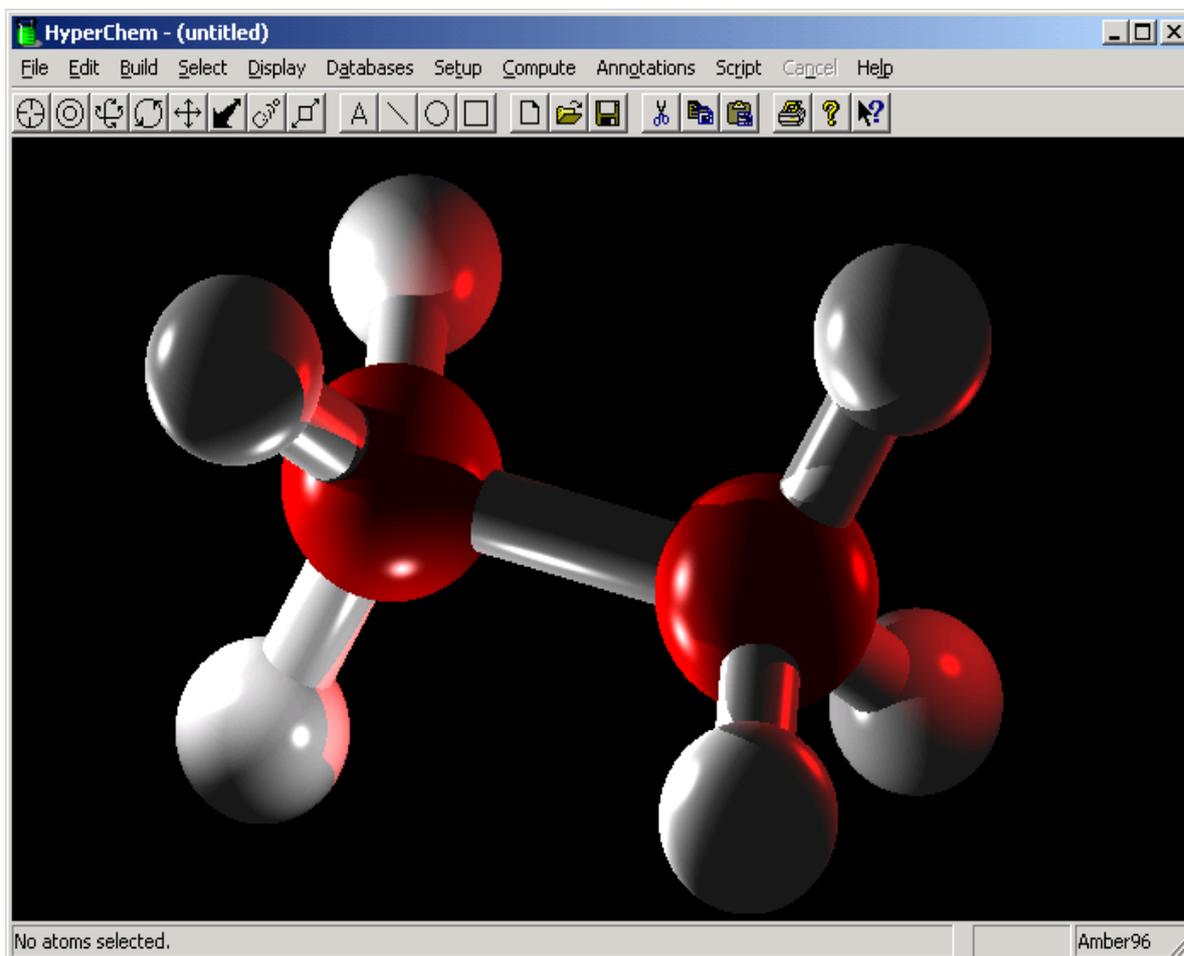


Figure 1-4: HyperChem display screen

1.4.2 Dragon

DRAGON was developed in 1994 by Milano Chemometrics and QSAR Research Group with the name "WHIM/3D QSAR", being specific for the calculation of the WHIM descriptors [36]. Successively, a lot of other descriptors have been implemented leading to a new software, which in 1997 provided about 600 descriptors and was released with the name DRAGON [37].

DRAGON is user-friendly and easy to use software, and able to provide thousands of molecular descriptors that are divided into 18 logical blocks (Table 1-1).

Table 1-1: DRAGON descriptors blocks.

ID	Block description
1	Constitutional descriptors
2	Topological descriptors
3	Molecular walk counts
4	BCUT descriptors
5	Galvez topological charge indices
6	2D autocorrelations
7	Charge descriptors
8	Aromaticity indices
9	Randic molecular profiles
10	Geometrical descriptors
11	RDF descriptors
12	3D-MoRSE descriptors
13	WHIM descriptors
14	GETAWAY descriptors
15	Functional group counts
16	Atom-centred fragments
17	Empirical descriptors
18	Properties

1.4.3 SPSS

SPSS, standing for Statistical Package for the Social Sciences, it is a powerful, user-friendly. SPSS is a software package for the manipulation and statistical analysis of data, (Figure 1-5). It was developed in 1968 by three young men from disparate professional backgrounds Norman H. Nie, C. Hadlai (Tex) Hull and Dale H. Bent [38]. The idea was based on using statistics to turn raw data into information essential to decision-making

SPSS for Windows offers a spreadsheet facility for entering and browsing the working data file — the Data Editor. Output from statistical procedures is displayed in a separate window — the Output Viewer. It takes the form of tables and graphics that can be manipulated interactively and can be copied directly into other applications [39].

SPSS is very common and widely used by social science researchers. Also it is used by market researchers, health researchers, survey companies, government, education researchers, and others. In this research the SPSS will be used to perform MLR analysis.

	Subject	Gender	Weight_Pre	Weight_Post	var	var
1	1.00	Male	78.00	75.00		
2	2.00	Male	92.00	90.00		
3	3.00	Male	85.00	80.00		
4	4.00	Male	78.00	77.00		
5	5.00	Female	68.00	65.00		
6	6.00	Female	69.00	68.00		
7	7.00	Female	65.00	62.00		
8	8.00	Female	80.00	75.00		
9						

Figure 1-5: SPSS display screen

1.4.4 MATLAB

MATLAB stands for MATrix LABoratory and the software is built up around vectors and matrices. This makes the software particularly useful for linear algebra but MATLAB is also a great tool for solving algebraic and differential equations and for numerical integration. MATLAB has powerful graphic tools and can produce nice pictures in both 2D and 3D. It is also a programming language, and is one of the easiest programming languages for writing mathematical programs [40].

The MATLAB mainly used in the current study to perform the cross validation, PCA and ANN.

1.5 Translocator protein (TSPO) preview:

Translocator protein (TSPO), was known as the peripheral benzodiazepine receptor (PBR). First identified in 1977 based on its distinct pharmacology with high affinity binding to benzodiazepines in peripheral tissues [41-44]. The term “peripheral” was used to distinguish it from the plasma membrane “central” benzodiazepine receptor, a complex together with the γ -aminobutyric acid type A receptor that is important for inhibitory neurotransmission in the central nervous system [45, 46]. However it became clear that its density in the brain regions can equal or exceed the density of central benzodiazepine receptor (CBR) in the corresponding regions [47].

TSPO is a protein of 18 kDa consisting of 169 amino acids [48], a five α -helices composed of 21 hydrophobic residues. The N-terminus of the sequence is located in the mitochondrial domain, while the C-terminus is exposed to the cytoplasm. The transmembrane regions are connected by loops rich in hydrophilic residues [49]. TSPO is strictly associated in a trimeric complex with the 32 kDa voltage dependent anion channel (VDAC), and 30 kDa adenine nucleotide translocase (ANT), thus forming the mitochondrial permeability transition pore (MPTP).

TSPO amino acid sequence shows conservation throughout evolution. TSPO in the photosynthetic bacteria *Rhodobacter sphaeroides* shows a 33.5% identity to human TSPO. Both human and mouse TSPO genes translate to a 169-amino acid protein with 81% sequence homology [50, 51]. Relatively the protein sequence of TSPO is conserved from bacteria to humans.

Expression of TSPO has been reported in different tissues including heart, brain, lung, spleen, testis, ovary, adrenal, kidney, bone marrow, salivary gland, adipose tissue, skin,

and liver [52-54]; and within these tissues, TSPO expression is regional and/or cell type specific. Also TSPO is expressed at low levels in other subcellular compartments such as plasma membranes and the nuclear fraction of cells [55].

TSPO binding sites

Although research suggests that there exist multiple TSPO binding sites, the nature of these sites and their functional significance is poorly understood. Two ligands have been essential for characterizing the TSPO: the benzodiazepine Ro 5-4864 and the isoquinoline carboxamide PK11195, both of which are selective for the TSPO and display nanomolar binding affinity. Although these ligands exhibit saturable binding and reciprocal competition in radio ligand binding assays [56]. Furthermore, site-directed mutagenesis studies suggest certain residues in the first putative loop of TSPO are important for the binding of Ro 5-4864 but not PK11195. Thus, it is thought that PK11195 and Ro 5-4864 bind to heterogeneous sites at TSPO, either overlapping or allosterically coupled. Studies also describe PK11195 binding to multiple sites, which contradicts the initial finding that it bound to a single population of saturable sites. Scatchard analysis of 3HPK11195 binding to Ehrlich tumor cells revealed 2 independent binding sites [57].

TSPO Pharmacology

Benzodiazepine Ro5-4864 [4'-chlorodiazepam; 7-Chloro-5-(4-chlorophenyl)-1,3-dihydro-1-methyl-2H-1,4-benzodiazepin-2-one] and a nonbenzodiazepine PK11195 [an isoquinoline carboxamide derivative, 1-(2-Chlorophenyl)-*N*-methyl-*N*-(1-methylpropyl)-3-isoquinolinecarboxamide] were initially established as prototypical TSPO-binding chemicals, because they bind to TSPO but not to γ -aminobutyric acid type A receptor [58, 59]. Based on thermodynamic studies [60], and their opposing effects on neuronal seizures [61], PK11195 was classified as an antagonist and Ro5-4864 as an agonist. This pharmacology has been extensively used in attempts to elucidate the physiological

relevance of TSPO [62]. Although these studies did not readily reveal TSPO function, the ability of these chemicals in detecting TSPO with reasonable accuracy, and the pathological TSPO up-regulation seen at sites of inflammation led to the development of TSPO as a diagnostic target [63]. Radiolabeled forms of these chemicals that bind TSPO could be used to detect inflammatory lesions *in vivo* in a variety of human diseases using positron emission tomography [64, 65]. Clinical trials for different TSPO-binding agents focused on the diagnosis of various pathologies including traumatic brain injury, Alzheimer's disease, Parkinson's disease, multiple sclerosis, encephalopathy, autism, neuroinflammation, neurodegeneration, dementia, and neurocysticercosis.

Human clinical trials to detect cardiac sarcoidosis (NCT02017522), carotid atherosclerosis (NCT00547976), and squamous and basal cell carcinomas (NCT01265472). Thus, remain an area of active research.

TSPO is said to be involved in a variety of biological processes including cholesterol transport, steroidogenesis, calcium homeostasis, lipid metabolism, mitochondrial oxidation, cell growth and differentiation, apoptosis induction, and regulation of immune functions [55].

TSPO in Brain and neurodegenerative diseases

Brain expression of TSPO in physiological conditions is low. In the CNS, TSPO is mainly found in glia and at very low levels in neurons [63, 66, 67]. However TSPO ligands are used for brain imaging of neuroinflammation, since TSPO is upregulated at sites of injury and inflammation, as well as in several neuropathological conditions including stroke and neurodegenerative disorders such as Alzheimer's disease (AD), Parkinson's disease, Huntington's disease, Multiple sclerosis and Amyotrophic lateral sclerosis [68-71]. Under

these conditions, the expression of TSPO is highly enhanced in reactive microglia and astrocytes [72-75].

TSPO is also upregulated in microglia and astrocytes in response to lesions, and its level of upregulation is directly related to the degree of damage. Therefore several studies suggest that TSPO ligands could be used as markers for the state and progression of Traumatic brain injury (TBI). In addition, some studies have addressed the neuroprotective effects of TSPO ligands in experimental models of brain injury [63, 76].

Other several studies have associated psychiatric disorders with a down regulation of TSPO expression in peripheral cells. Decreased TSPO expression has been found in the platelets and lymphocytes of patients with anxiety disorders [77-79], in the platelets of patients suffering from schizophrenia [80] and post-traumatic stress disorder [81] and in a suicidal adolescent population [82]. However, increased TSPO density, measured by distribution volume by positron emission tomography, has been detected in the prefrontal cortex, the anterior cingulate cortex and insula of patients with a major depressive episode [83]. In these patients, greater TSPO density in the anterior cingulate cortex correlated with greater depression severity [83].

Thus TSPO can be exploited as a diagnostic marker to follow disease Progression and therapy efficacy by means of the biomedical imaging technique PET (positron emission tomography) but also as a therapeutic target [84]. Although imaging complications have been encountered as a result of *in vivo* metabolism of these TSPO-binding PET tracers and aberrant signals contributing to nonspecific noise in some cases, new synthetic TSPO-binding chemicals are being developed to tackle these drawbacks [75]. Therefore, diagnostic imaging is probably the primary clinical value that TSPO research has to offer at the present time [85].

1.6 Research objective

The main objective of this study is to develop QSAR models for the activity of 136 chemical compounds of Translocator protein (TSPO) by applying different statistical qualities; MLR and PC-ANN. The resulted models will be used for designing and prediction of the activity of new ligands.

Chapter two

Methodology

2. QSAR process method:

Quantitative structure-activity relationship (QSAR), is an analytical application that is used to interpret the quantitative relationship between the biological activities and particular molecules structures. And to do that the molecular structure and their activity against certain target should be known and experimentally estimated.

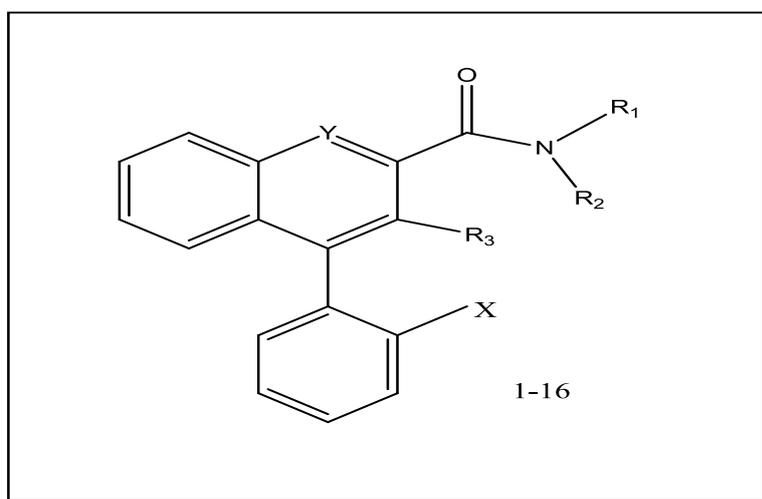
As previously mentioned in Chapter 1; QSAR model development process is typically performed in successive steps divided into three steps; Data preparation, data analysis, and model validation.

2.1 Data preparation

2.1.1 Dataset

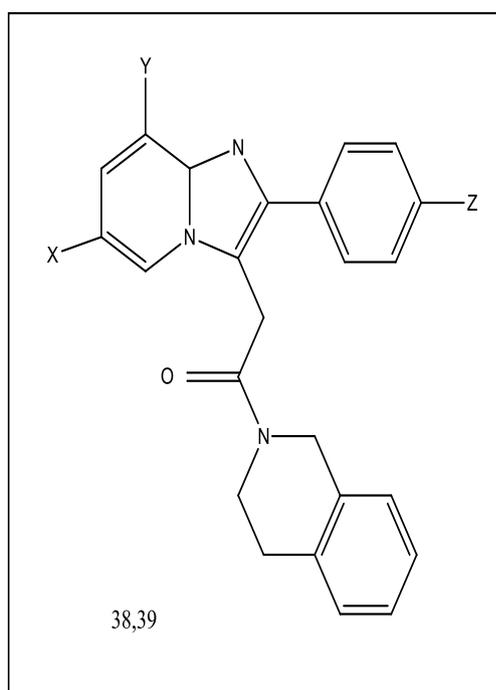
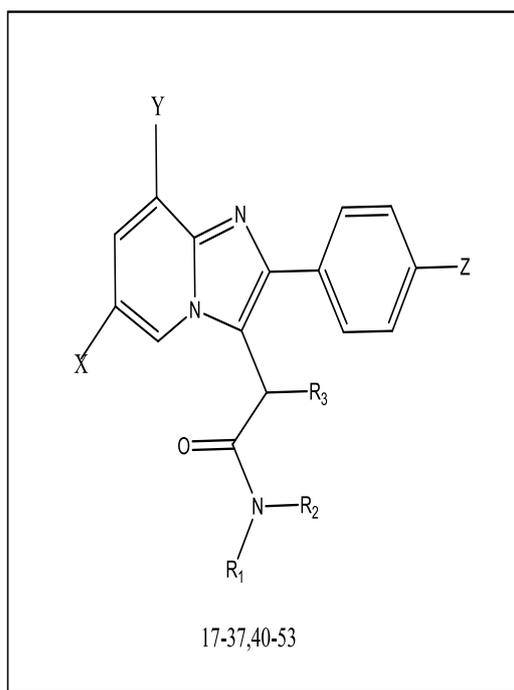
136 compounds and their related observed activity (pIC_{50}) against Translocator protein (TSPO) are carefully taken from references [49, 86-89], which shared the same method of determination of ligands-Receptor activities using rat cortex membrane. The 136 compounds are divided into 18 chemical structure cores as shown in table 2-1.

Table 2-1: Dataset, Compounds have activity against TSPO



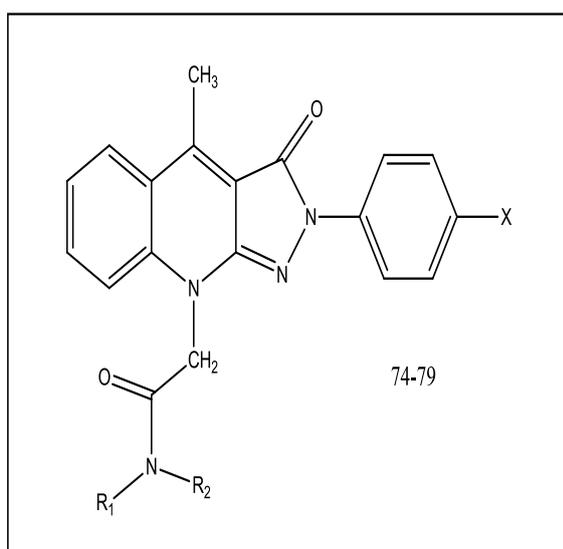
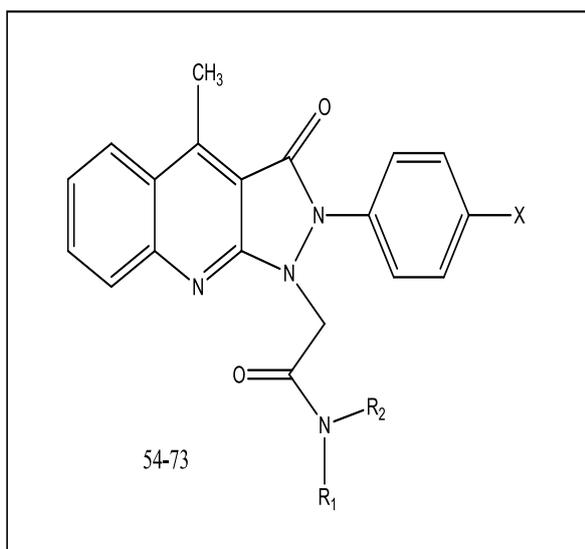
Compounds Number	Index *	X	Y	R1	R2	R3	pIC50
001	3a	H	CH	CH ₂ C ₆ H ₅	H	H	5.569
002	3b	H	CH	CH ₂ C ₆ H ₅	CH ₃	H	7.194
003	3c	H	N	CH ₂ C ₆ H ₅	H	H	7.194
004	3d	H	N	CH ₂ C ₆ H ₅	CH ₃	H	7.420
005	3e	H	CH	CH ₂ C ₆ H ₅	H	CH ₃	5.180
006	3f	H	CH	CH ₂ C ₆ H ₅	CH ₃	CH ₃	8.009
007	3g	H	N	CH ₂ C ₆ H ₅	H	CH ₃	4.988
008	3h	H	N	CH ₂ C ₆ H ₅	CH ₃	CH ₃	8.337
009	3i	F	N	CH ₂ C ₆ H ₅	H	CH ₃	5.827
010	3j	F	N	CH ₂ C ₆ H ₅	CH ₃	CH ₃	8.658
011	3k	H	N	CH ₂ C ₆ H ₅	CH ₂ C ₆ H ₅	CH ₃	7.959
012	3l	H	N	CH ₂ C ₆ H ₅	CH ₃	CH ₂ OH	8.060
013	3m	H	N	CH ₂ C ₆ H ₅	CH ₃	CH ₂ CL	9.347
014	3n	H	N	CH ₂ C ₆ H ₅	CH ₃	CH ₂ N(C ₂ H ₅) ₂	7.921
015	3o	H	N	CH ₂ C ₆ H ₅	CH ₃	CH ₂ N(C ₂ H ₅) CH ₂ C ₆ H ₅	7.886
016	3p	H	N	CH ₂ CCH	CH ₃	CH ₃	7.495

*: Reference [86]



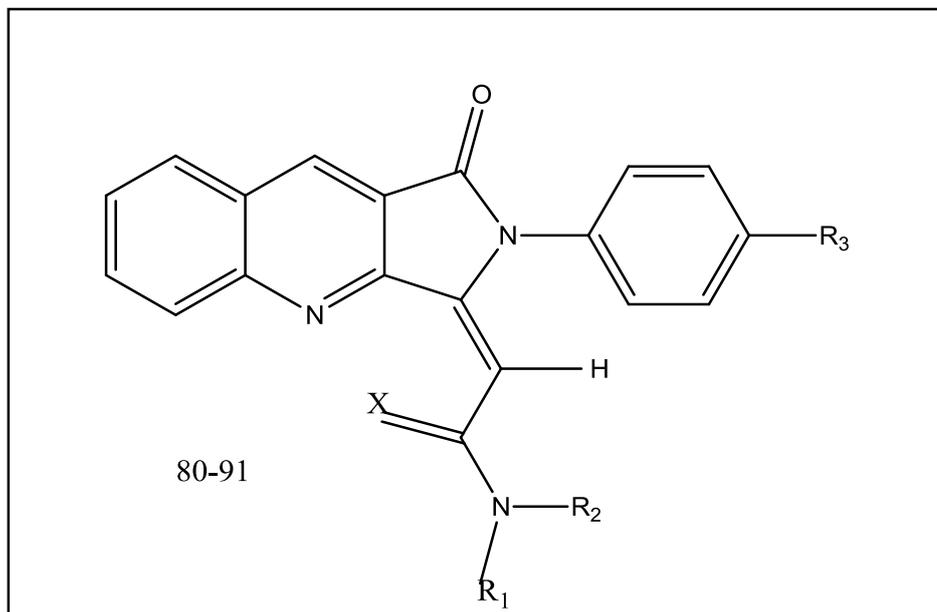
Compounds Number	Index *	X	Y	Z	R1	R2	R3	pIC50
017	1	H	H	Cl	<i>n</i> -C4H9	<i>n</i> -C4H9	H	8.230
018	2	H	Cl	Cl	<i>n</i> -C4H9	<i>n</i> -C4H9	H	8.104
019	3	Cl	Cl	Cl	<i>n</i> -C4H9	<i>n</i> -C4H9	H	8.284
020	4	Cl	Cl	Cl	<i>n</i> -C6H13	<i>n</i> -C6H13	H	6.424
021	5	Cl	H	Cl	<i>n</i> -C4H9	<i>n</i> -C4H9	H	8.485
022	6	Cl	H	Cl	<i>n</i> -C6H13	<i>n</i> -C6H13	H	8.292
023	7	Cl	H	H	<i>n</i> -C4H9	C6H5	H	7.939
024	8	Cl	Cl	Cl	<i>n</i> -C4H9	C6H5	H	7.876
025	9	Cl	H	Cl	<i>n</i> -C4H9	C6H5	H	8.824
026	10	Cl	Cl	H	<i>n</i> -C4H9	CH2C6H6	H	7.616
027	11	Cl	Cl	Cl	<i>tert</i> -C4H9	CH2C6H6	H	5.464
028	12	Cl	Cl	Cl	<i>n</i> -C3H7	4-NO2-CH2C6H5	H	7.566
029	13	Cl	Cl	Cl	C6H5	H	H	7.701
030	14	Cl	Cl	Cl	CH2CHCH2	CH2CHCH2	H	8.092
031	15	Cl	Cl	Cl		-(CH2)4-	H	6.668
032	16	Cl	Cl	H		-(CH2)4-	H	5.907
033	17	Cl	Cl	H		-(CH2)5-	H	6.804
034	18	Cl	Cl	Cl		-(CH2)5-	H	8.301
035	19	Cl	H	Cl		-CH2CH(COOC2H5)(CH2)3-	H	7.454
036	20	Cl	Cl	Cl		-CH2CH(COOC2H5)(CH2)3-	H	6.845
037	21	Cl	Cl	Cl		-(CH2)2N(CH2C6H5)(CH2)2-	H	4.682
038	22	Cl	Cl	H	-	-	-	7.412
039	23	Cl	Cl	Cl	-	-	-	8.313
040	24	Cl	Cl	H	2-pyridylethyl	CH3	H	5.663
041	25	Cl	Cl	Cl	2-pyridylethyl	CH3	H	6.046
042	26	Cl	Cl	H	2-pyridyl	H	H	5.677
043	27	Cl	Cl	Cl	<i>n</i> -C4H9	H	H	6.409
044	28	Cl	Cl	Cl	C6H11	H	H	6.640
045	29	Cl	Cl	H	C6H11	H	H	5.878
046	30	Cl	Cl	Cl	CH2C6H5	H	H	6.772
047	31	Cl	Cl	Cl	<i>n</i> -C3H7	<i>n</i> -C3H7	CH3	5.920
048	32	Cl	Cl	Cl	C6H11	CH3	CH3	5.288
049	33	Cl	Cl	Cl	CH2C6H5	CH3	CH3	5.005
050	34	Cl	Cl	Cl	<i>n</i> -C4H9	CH3	H	9.347
051	35	Cl	Cl	H	<i>n</i> -C4H9	CH3	H	8.456
052	36	Cl	Cl	Cl	C6H5	CH3	H	9.481
053	37	Cl	Cl	Cl	CH2C6H5	CH3	H	8.623

*: Reference [87].



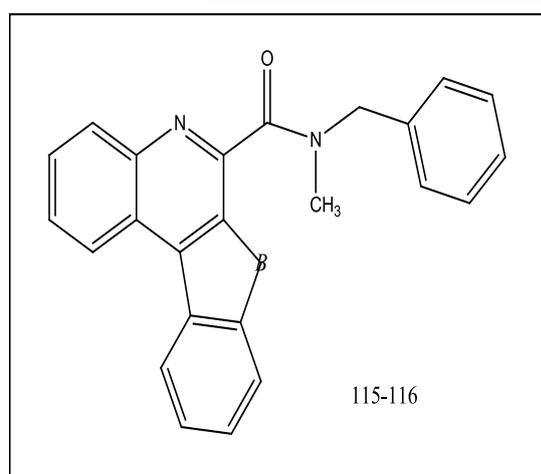
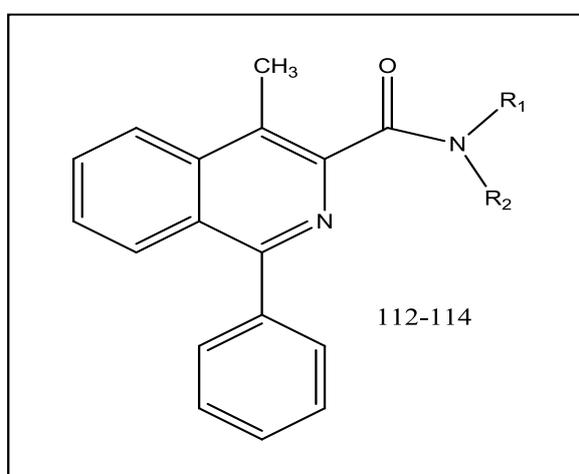
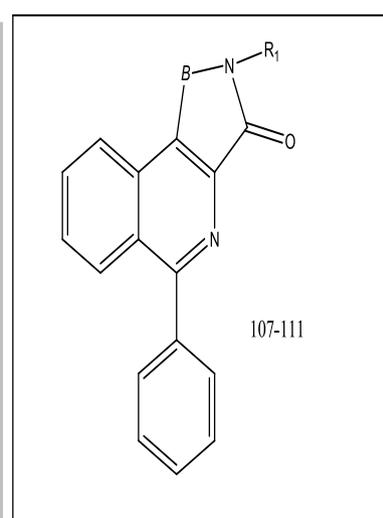
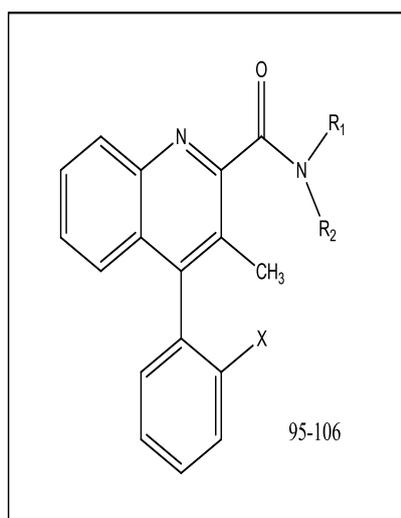
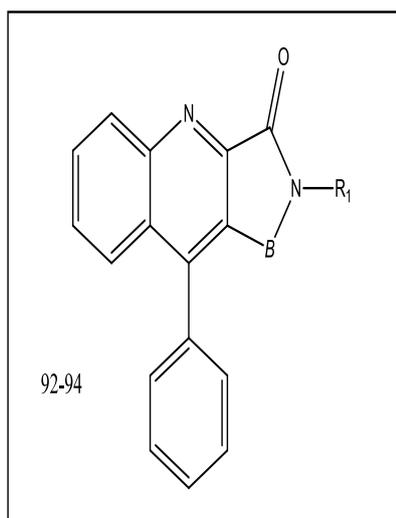
Compounds Number	Index *	R1	R2	X	pIC50
054	10a	CH3	(CH2)3CH3	H	7.187
055	10b	C2H5	C2H5	H	6.745
056	10c	CH(CH3)2	CH(CH3)2	H	5.856
057	10d	(CH2)2CH3	(CH2)2CH3	H	7.046
058	10e	(CH2)3CH3	(CH2)3CH3	H	6.818
059	10f	CH3	C6H5	H	6.932
060	10g	CH3	<i>p</i> -Cl-C6H4	H	8.745
061	10h	CH3	<i>p</i> -CH3OC6H4	H	7.769
062	10i	CH3	CH2C6H5	H	6.055
063	10j	CH2C6H5	C2H5	H	6.658
064	10k	CH2C6H5	CH(CH3)2	H	6.517
065	10l	CH2C6H5	(CH2)3CH3	H	6.157
066	10m	CH2C6H5	CH2C6H5	H	5.460
067	10n	CH3	(CH2)3CH3	F	6.959
068	10o	C2H5	C2H5	F	5.644
069	10p	(CH2)2CH3	(CH2)2CH3	F	7.060
070	10q	CH3	<i>p</i> -Cl-C6H4	F	8.167
071	10r	CH3	(CH2)3CH3	Cl	6.842
072	10s	(CH2)2CH3	(CH2)2CH3	Cl	7.046
073	10t	CH3	<i>p</i> -Cl-C6H4	Cl	7.796
074	11a	CH3	(CH2)3CH3	H	9.854
075	11b	C2H5	C2H5	H	9.081
076	11e	(CH2)3CH3	(CH2)3CH3	H	9.469
077	11g	CH3	<i>p</i> -Cl-C6H4	H	9.886
078	11n	CH3	(CH2)3CH3	F	9.886
079	11q	CH3	<i>p</i> -Cl-C6H4	F	9.553

*: Reference [88].



Compounds Number	Index *	R ₁	R ₂	R ₃	X	pIC ₅₀
080	8a	4-(CH ₃ O)C ₆ H ₄	CH ₃	Cl	=O	9.046
081	8b	4-ClC ₆ H ₄	CH ₃	Cl	=O	9.208
082	8c	CH ₂ C ₆ H ₅	CH ₃	Cl	=O	7.585
083	8d	CH ₂ C ₆ H ₅	C ₂ H ₅	Cl	=O	7.824
084	8e	(CH ₂) ₅ CH ₃	(CH ₂) ₅ CH ₃	Cl	=O	7.678
085	9a	C ₆ H ₅	CH ₃	Cl	=O	7.357
086	9b	4-(CH ₃ O)C ₆ H ₄	H	Cl	=O	7.119
087	9c	4-(OH) C ₆ H ₄	CH ₃	Cl	=O	7.131
088	9d	CH ₂ CCH	CH ₃	Cl	=O	6.495
089	9e	4-(CH ₃ O)C ₆ H ₄	CH ₃	H	=O	7.921
090	9f	4-ClC ₆ H ₄	CH ₃	H	=O	8
091	9g	4-(CH ₃ O)C ₆ H ₄	CH ₃	Cl	=H ₂	5.627

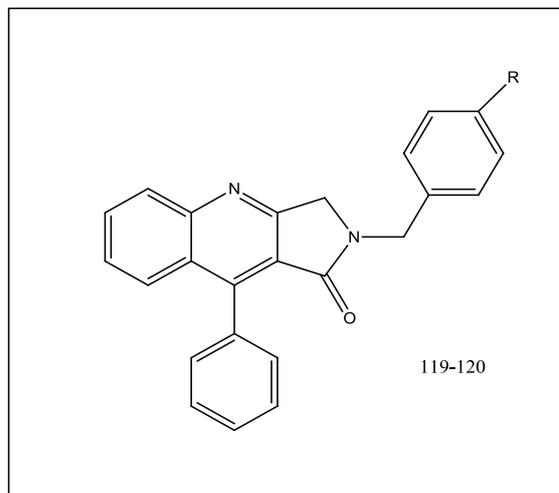
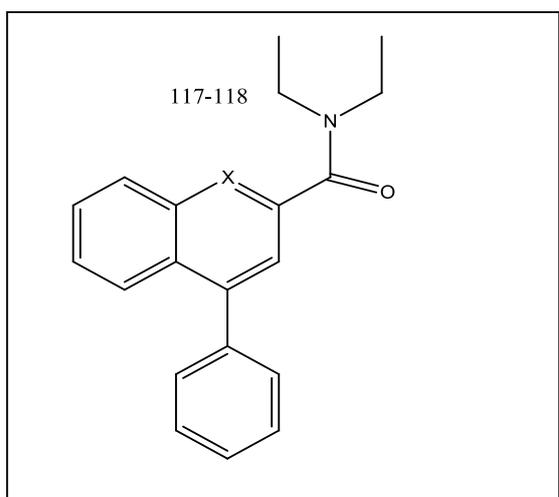
*: Reference [89].



Compounds Number	Index *	Bridge (B)	X	R1	R2	pIC50
092	7a	<i>n</i> -Bu-CH	-	Benzyl	-	6.092
093	7b	CH ₂ -CH ₂ -CH ₂	-	<i>s</i> -Bu	-	5.921
094	7c	CH ₂ -CH ₂ -CH ₂	-	Benzyl	-	6.770
095	8a	-	H	<i>s</i> -Bu	H	6.638
096	8b	-	F	<i>s</i> -Bu	H	7.886
097	8c	-	H	Benzyl	H	5.921
098	8d	-	H	4-Cl-Benzyl	H	5.769
099	8e	-	F	4-Cl-Benzyl	H	6.569
100	8f	-	H	<i>s</i> -Bu	Me	8.678
101	8g	-	F	<i>s</i> -Bu	Me	8.538
102	8h	-	H	Benzyl	Me	8.678
103	8i	-	H	4-Cl-Benzyl	Me	8.009
104	8j	-	F	4-Cl-Benzyl	Me	8.469
105	8k	-	H	4-Cl-Ph	Me	8.194
106	8l	-	H	4-MeO-Ph	Me	8.056

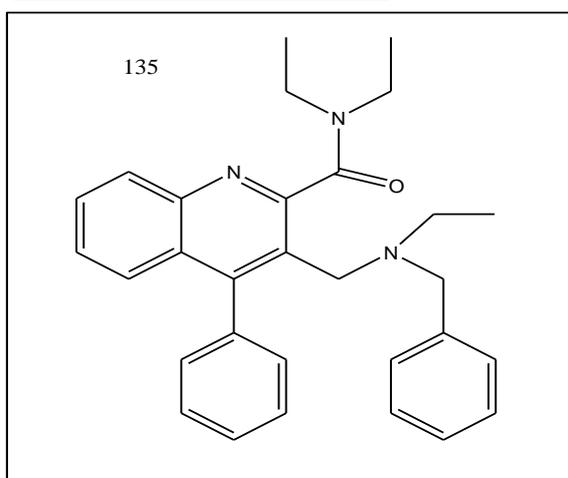
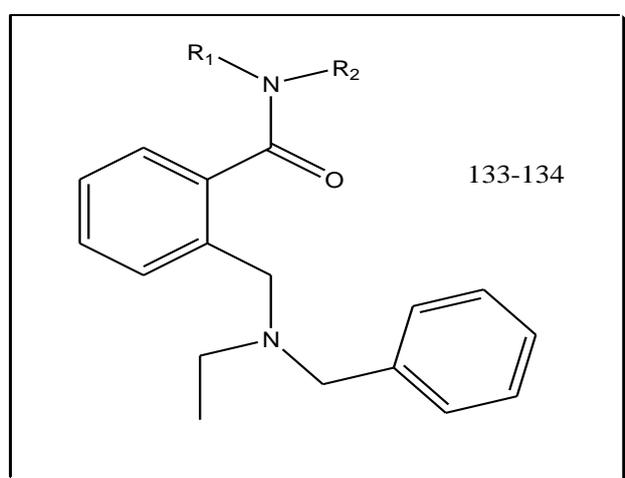
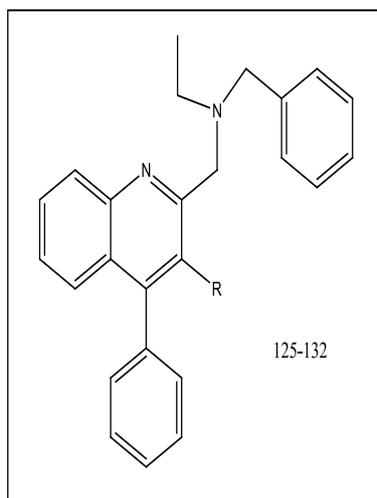
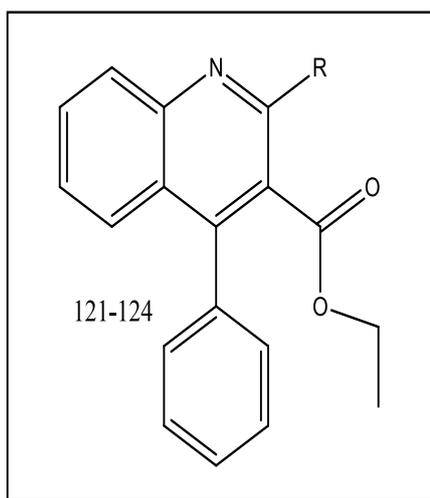
107	9a	CH ₂	-	s-Bu	-	6.208
108	9b	CH ₂	-	Benzyl	-	6.319
109	9c	CH ₂ -CH ₂ -CH ₂	-	s-Bu	-	6.108
110	9d	CH=CH-CH ₂	-	s-Bu	-	6.309
111	9e	CH=CH-CH ₂	-	Benzyl	-	7.347
112	10a	-	-	s-Bu	H	6.259
113	10b	-	-	s-Bu	Me	7.959
114	10c	-	-	Benzyl	Me	8.509
115	11a	CH ₂ -CH ₂	-	-	-	8.051
116	11b	O-CH ₂ -CH ₂	-	-	-	8

*: Reference [49]



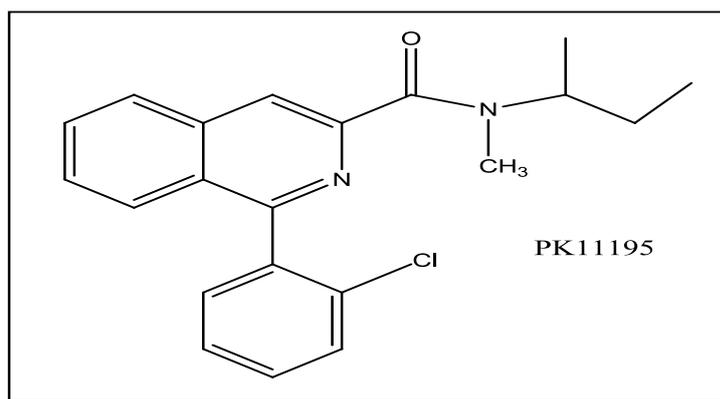
Compounds Number	Index *	X	R	pIC ₅₀
117	4a	CH	-	8.301
118	4b	N	-	7.569
119	5a	-	Me	6.387
120	5b	-	CL	6.678

*: Reference [49]



Compounds Number	Index *	R	R1	R2	pIC50
121	6a	CH ₂ N(Et)Bn	-	-	6.155
122	6b	N(Et)Bn	-	-	6.268
123	12a	Cl	-	-	5.432
124	12b	CH ₂ THIQ	-	-	5.926
125	13a	CONMe ₂	-	-	6.640
126	13b	CONEt ₂	-	-	6.428
127	13c	CON(<i>n</i> -Pr) ₂	-	-	6.341
128	13d	CON(Me)Ph	-	-	5.880
129	13e	CON(Me)4-Cl-Ph	-	-	5.606
130	13f	CON(H) <i>n</i> -Pr	-	-	6.059
131	13g	CON(H)Bn	-	-	5.538
132	13h	H	-	-	5.469
133	15b	-	Et	Et	5.086
134	15c	-	<i>n</i> -Pr	<i>n</i> -Pr	5.052
135	16	-	-	-	8.387

*: Reference [49]



Comp. Number	Reference Compound	pIC50 [86]	pIC50 [87]	pIC50 [88]	pIC50 [89]	pIC50 [49]
136	PK11195	8.075	8.155	8.886	8.677	8.657

2.1.2 Compounds optimization

To calculate the properties of each molecule; a well-defined structure which represents a minimum potential energy surface is needed. Therefore after choosing the 136 compounds, start drawing each compound structure using HyperChem to optimize it.

Steps of optimization using HyperChem:

1. Draw the compound structure on HyperChem Workspace using drawing tools, as seen in figure 2-1.
2. The drawn compound is in two- dimensional (2D) form, therefore a conversion from the 2D compound structure into a 3D structure using the HyperChem Model Builder is needed. So select (Add H and Model build) from the Build menu to convert the 2D form to 3D.
3. Click start log on the File menu to save the new drawn structure, name the file, and choose a directory to save in it,

4. Then in order to perform the optimization of the compound structure; choose the semi-empirical calculation from the setup tab, accordingly a dialog box of types of semi-empirical methods will open. Thus choose from it the AM1 method and after that press on the options button to determine the geometry optimization parameters; total charge = 0, spin multiplicity = 1, spin pairing = RHF (Restricted Hartree-Fock), convergence limit = 0.1.

These parameters mean that the calculation ends when the difference in energy after two consecutive iterations is less than 0.1 kcal/mol. The calculation is performed on the lowest state without special convergence acceleration.

5. Click OK to close the semi-empirical options dialog box, and then click OK to close the semi-empirical method dialog box.
6. To start the optimization process, click on geometry optimization from HyperChem menu. A dialog box of Semi-empirical optimization will open. Set Polak-Ribiere as algorithm method, 0.1 for RMS gradient and keep the defaulted value for the rest of the fields.

Then click OK so the optimization process initiate.

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

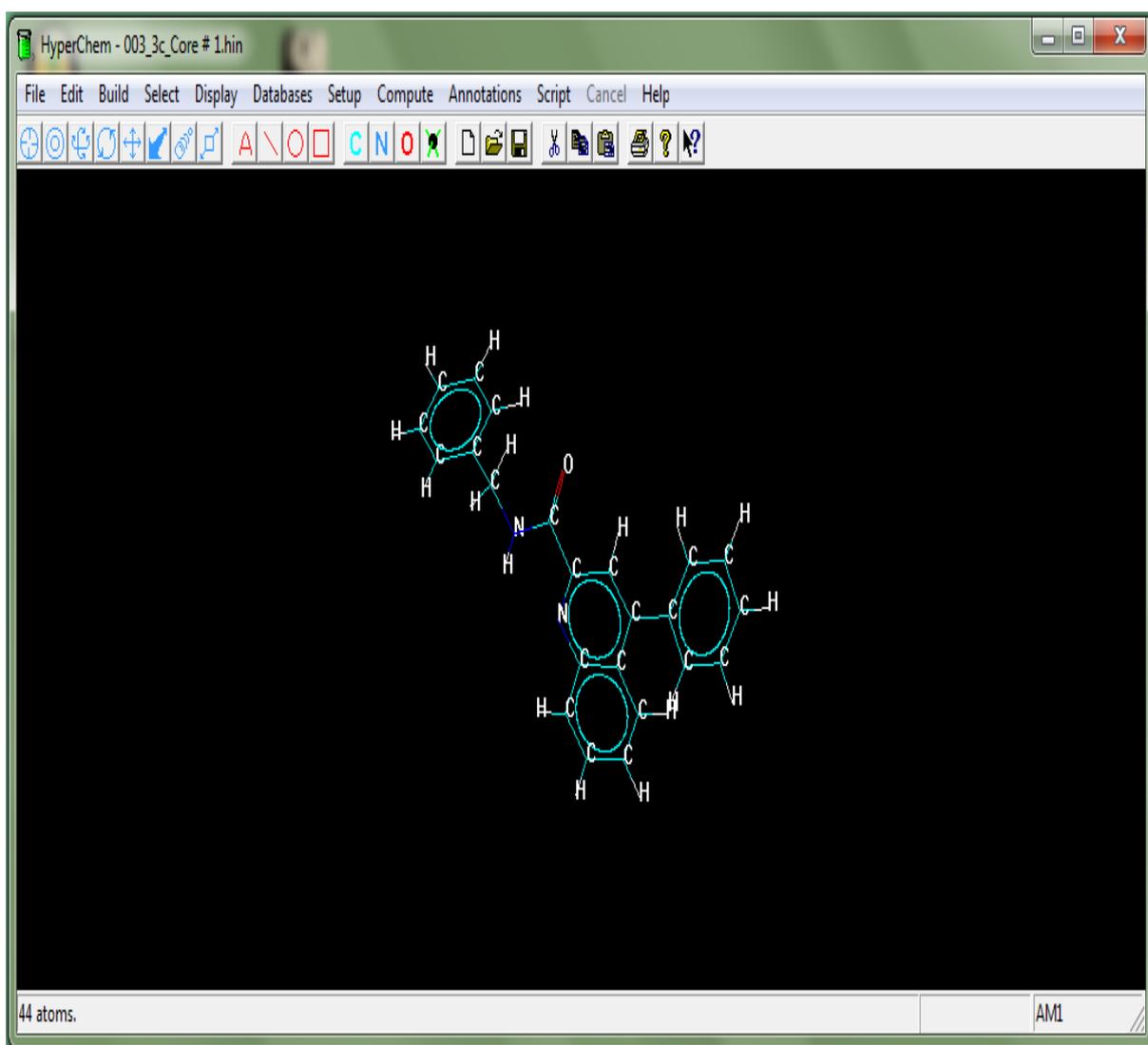


Figure 2-1: Drawing using HyperChem

2.1.3 Descriptors calculation

To establish QSAR we are not able mathematically to link between the chemical structures and their activity directly, thus a numerical factor is needed to link between the chemical structure and the activity. This numerical factor is the chemical structure properties which called Molecular Descriptors

“The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule

into a useful number or the result of some standardized experiment [90] .” Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, health research and quality control. It has been used to predict biological and physicochemical properties of molecules (QSAR/QSPR) and for virtual screening of molecule libraries.

There are simple molecular descriptors derived by counting some atom-types or structural fragments in the molecule, other derived from algorithms applied to a topological representation (molecular graph) and usually called topological or 2D-descriptors, and there are molecular descriptors derived from a geometrical representation called geometrical or 3D-descriptors.

In current research we have been using two software's to calculate different descriptors; HyperChem and Dragon.

2.1.3.1 Descriptors calculated by HyperChem

a. Descriptors extracted from the output log file

The HyperChem calculate the quantum chemical descriptors and more. We open the output log file for each optimized chemical structure and take from it the following values then put the values in excel file:

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

From the HOMO and LUMO values we can calculate the below descriptors:

- ✓ Hardness ($0.5 * (\text{LUMO} - \text{HOMO})$).
- ✓ Softness ($1/\text{Hardness}$).
- ✓ Electronegativity ($-0.5 * (\text{LUMO} + \text{HOMO})$).
- ✓ Electrophilicity ($\text{Electronegativity} * \text{Electronegativity} / (2 * \text{Hardness})$) [91].

b. Descriptors calculated from the HyperChem using the optimized structures

We can calculate certain descriptors by performing the following steps:

1. Open the HyperChem file of the optimized 3D structure of each compound in the dataset.
2. Then choose QSAR properties from the computer tab, thus a dialog box contain the below properties will open (Fig 2-2).
 - ✓ Surface Area (Approx).
 - ✓ Surface Area (Grid).
 - ✓ Volume.
 - ✓ Hydration Energy.
 - ✓ Log P.
 - ✓ Refractivity.
 - ✓ Polarizability.
 - ✓ Mass.

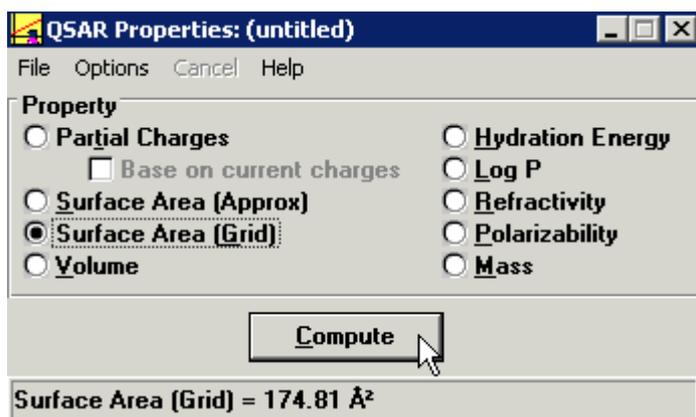


Figure 2-2: QSAR Properties dialog box in HyperChem software.

3. Choose one of the properties in the dialog box and press on Compute button, then copy the result to an excel file. Repeat this step to calculate all the properties one by one for each chemical structure.

2.1.3.2 Descriptors calculated by Dragon

DRAGON 2.1 software provides the calculation of thousands of descriptors which are divided into 18 blocks (groups) of descriptors as seen in table 1- 1, also some of the descriptor groups is mentioned in table 2-2.

2.1.3.2.1 Brief description about Dragon descriptors:

Constitutional descriptors are the most simple and commonly used descriptors, reflecting the composition of a molecule without any geometrical information. Examples of these descriptors are the number of atoms, bonds, rings, specific atom types, rotatable bonds, etc.

The descriptor blocks: topological, walk and path counts, information indices, 2D autocorrelation, and charge indices contain topological and topographic descriptors. Topological descriptors are based on a graph representation of the molecule. They are numerical quantifiers of molecular topology obtained by the application of algebraic operators to matrices representing molecular graphs. They can be sensitive to one or more

structural features of the molecule such as size, shape, symmetry, branching and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. Topographic indices are derived from the graph representation of molecules in the same way as the topological indices, but using the geometric distances between atoms instead of the topological distances.

The blocks: geometrical, RDF, 3D-MoRSE, WHIM, and GETAWAY descriptors include descriptors derived from the knowledge of the 3D structure of the molecule. Some of the Molecular properties block derived from literature models, such as Moriguchi logP, Ghose-Crippen logP, Lipinski rule-of-five, etc.

However the descriptors groups are divided into four types:

0D: Constitutional descriptors.

1D: Empirical, Functional groups, Properties, Atom-centred fragments descriptors.

2D: Autocorrelations, Topological, Molecular walk counts, Galvez topological charge indices, BCUT descriptors.

3D: Geometrical, Randic molecular profiles, WHIM, GETAWAY, RDF, 3D-MoRSE, Charge descriptors.

2.1.3.2.2 Steps to perform descriptors calculation using DRAGON software:

1. After starting DRAGON, press on calculate descriptors button from the left side list of program interface. A dialog box will open, to select the files for calculations.
2. Select the output files resulted from the HyperChem structures optimization process and choose the type of the file to be in (.hin) format then choose the type of descriptor group to be calculated, then press run.

3. Save the output file in notepad format once the calculation of input compound file for certain group of descriptors is done.
4. Change the format file from the notepad to excel format. So we can use it as input file for SPSS and other analysis softwares.
5. Accordingly repeat these steps for all compounds each time calculate one group of descriptors, till all descriptors groups for each compound is calculated.

Table 2-2: Brief description of some of the descriptors used in this study

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

2.2 Data analysis

2.2.1 Multiple linear regression (MLR)

MLR simultaneously considers the relationship between dependent variables (biological activity) and an independent variable (theoretical molecular descriptors) by fitting a linear equation to observed data using SPSS software. The MLR is the first statistical step that is done because of the assumption that there is a linear correlation between the independent variables (descriptors) and the response variable (Y, Activity in our study).

2.2.1.1 Steps to perform MLR for each descriptor group using SPSS:

1. Import to SPSS one of the output files which resulted from the descriptors calculation on HyperChem and DRAGON (e.g. Constitutional descriptors file for all the compounds). The file will contain the activities of all compounds and each compound related calculated descriptors as seen in table 2-3.

Table 2-3: The format of the input file in SPSS to perform MLR (The activities of all compounds and their corresponding properties).

Activity (IC50)	Molecular weight (MW)	Sum of atomic van der Waals volumes (sv)	Sum of atomic Sanderson electronegativities (Se)	Sum of atomic polarizabilities (Sp)
5.569	337.44	30.89	44.43	32.28
7.194	351.47	32.48	47.32	34.04
7.194	338.43	30.28	43.65	31.53
7.420	352.46	31.88	46.54	33.29
5.180	351.47	32.48	47.32	34.04
8.009	365.5	34.08	50.2	35.8
4.988	352.46	31.88	46.54	33.29
8.337	366.49	33.48	49.42	35.04
5.827	370.45	31.99	47.05	33.22
8.658	384.48	33.59	49.94	34.98

2. After importing the file data to the SPSS, press on Analyze tab, choose regression ► Linear as seen in figure 2-3. Thus a dialog box of linear regression as seen in figure 2-4 will open to set the below fields:
 - ✓ First set IC50 as a dependent variable and the descriptors as independent variables. And also choose stepwise method for analysis method.
 - ✓ Then press on options button. Another dialog box will open to set the F value (F entry and F removal). Keep changing the F value till getting a convincing results.
 - ✓ And then press on Statistics to click on the field of Estimate Regression coefficient.
 - ✓ After that press on save button, and click on the unstandardized predicted values field.
3. Click Ok button on the linear regression dialog box, in order to start the MLR.
4. Repeat the previous steps on each descriptors group file.
5. Choose the best model from each descriptor's group output. The best model which has higher R value and minimum number of descriptors among the results of each group.

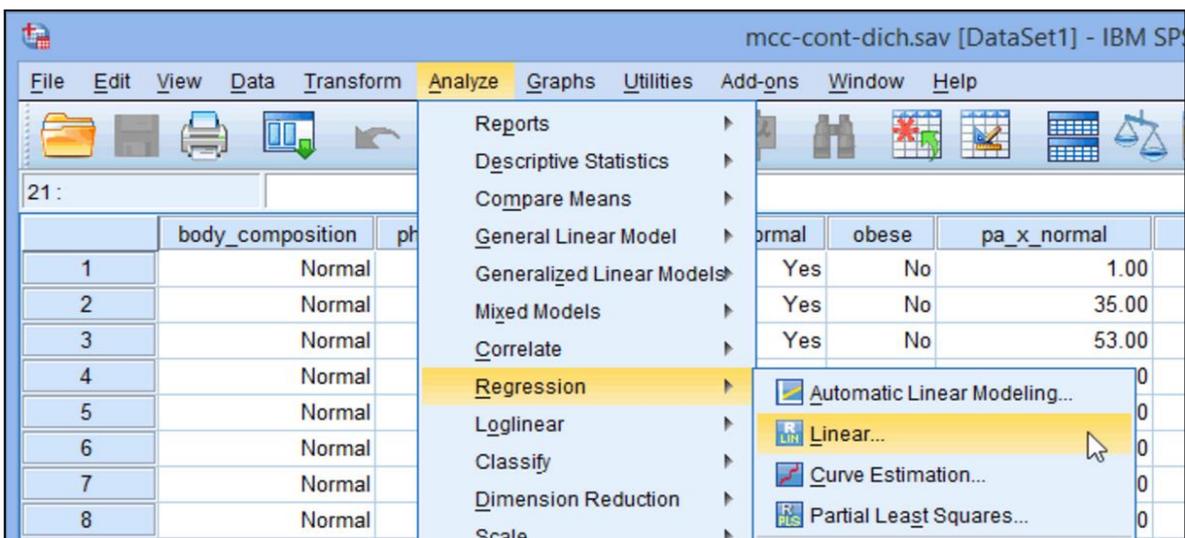


Figure 2-3: Choosing linear regression analysis using SPSS.

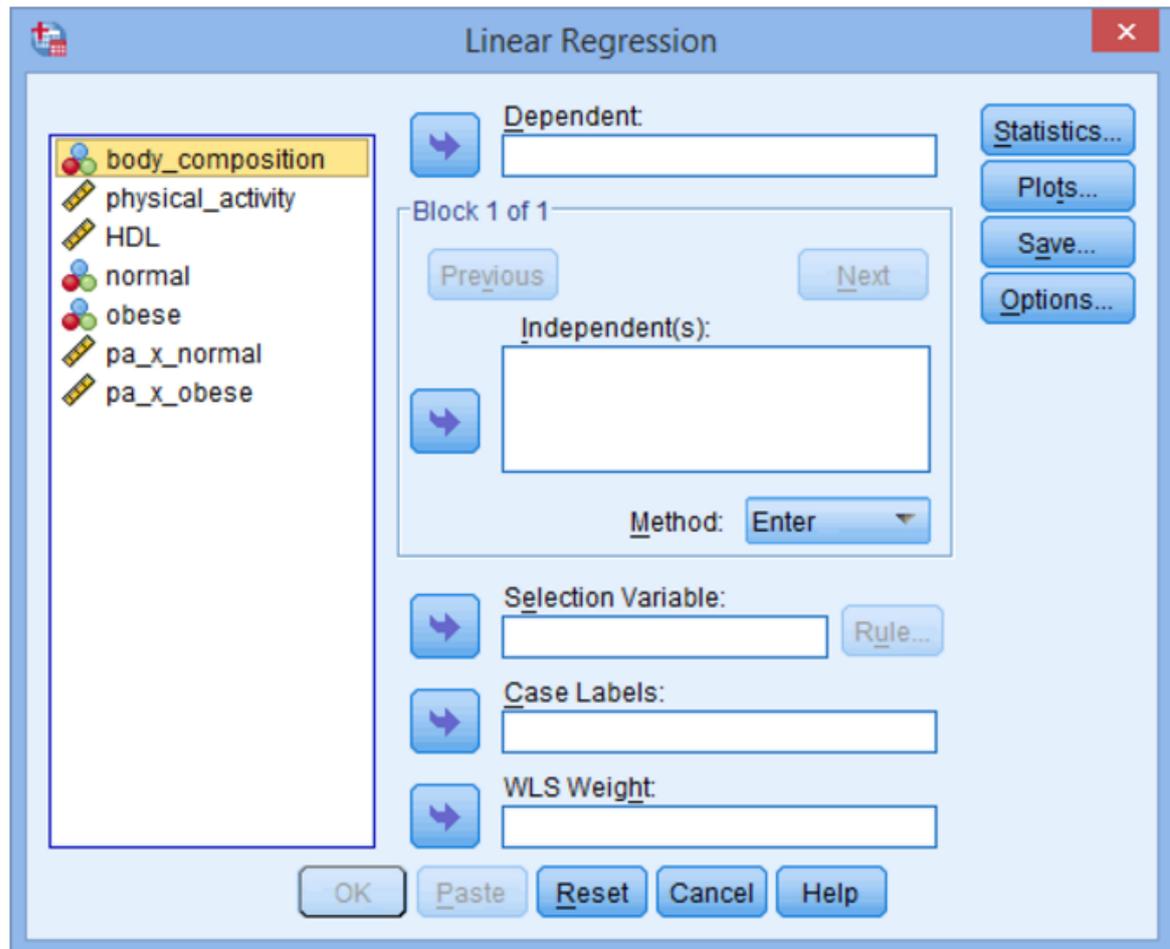


Figure 2-4: Dialog box which open after choosing the linear regression analysis.

2.2.1.2 Steps to perform MLR for all the descriptors resulted from the first MLR using the SPSS:

1. After choosing the best model for each descriptor's group, gather all the descriptors which mentioned in the best models in one file.
2. Then import the prepared file into the SPSS and follow the steps of MLR as in section 2.2.1.1.
3. From the resulted models; choose all the models having $R^2 \geq 0.6$ [92].

2.2.2 MLR Model validation

Model validation has been the subject of much recent debate in the scientific and regulatory communities. It was considered important to develop an internationally recognized set of principles for QSAR validation, to provide regulatory bodies with a scientific basis for making decisions on the acceptability of QSAR estimates of regulatory endpoints, and to promote the mutual acceptance of QSAR model. In current research, two internal validation methods have been performed to validate the MLR and the ANN resulted models; cross-validation and scrambling (Y-Randomization) respectively.

2.2.2.1 Cross-validation

The cross validation is used to validate the models resulted from the MLR. And it's divided into two types of procedures: leave one out (LOO) and leave many out (LMO) cross-validation.

2.2.2.1.1 Steps to perform leave one out (LOO) using MATLAB

1. Prepare a file which contains in the first column the observed activity and then comes the predicted activities of each model resulted from the MLR with $R^2 \geq 0.6$. Where the predicted activities value taken from SPSS.
2. By running a special MATLAB script to perform LOO and entering the file name, the MATLAB will ask for the model number and after that will ask to enter the number of descriptors for the model of interest.

3. A proper output file should look like:

<i>Model</i>	<i>PRESS</i>	<i>SPRESS</i>	<i>SST</i>	<i>R2CV</i>	<i>PRESS/SST</i>	<i>PSE</i>	<i>RSEP</i>

1	25.0874	0.7155	16.9432	-0.4807	1.4807	0.6693	61.1801

4. Choose the models which have PRESS/SST value < 0.4 , and compare it with the LMO results, and continue with the chosen models to the PCA and ANN.

2.2.2.1.2 Steps to perform leave many out (LMO) using MATLAB

1. Prepare a file of each model alone containing the observed activity and the descriptors of the model.
2. Run certain MATLAB script, and choose the data file.
3. A proper output file should look like:

<i>PRESS</i>	<i>SPRESS</i>	<i>SST</i>	<i>R2CV</i>	<i>PRESS/SST</i>	<i>PSE</i>	<i>RSEP</i>

45.7577	1.0198	42.6208	-0.0736	1.0736	0.9292	44.1042

4. Repeat the previous steps for each prepared model file.
5. Choose the models which have PRESS/SST value < 0.4 , and compare it with the LOO results, and continue with the chosen models to the PCA and ANN.

2.2.3 Principal component analysis (PCA)

Principal components analysis (PCA), It is a way of identifying patterns in data, and expressing the data in order to highlight their similarities and differences. Therefore the PCA is used to divide the dataset into three groups; training, validation and test set. As dividing the data should not be done randomly, instead, use the factor spaces of the descriptors and activity data. To do so, gather the descriptors and the activities in a single matrix (X). Perform principal component analysis (PCA) on X and then plot the first score against the second. You will obtain a scatter distribution of data (molecules) in the two first factor spaces. Select the training set molecules from these data points so they span the same space of the entire data. Data division should be done as to have 60% of the data in the training set and 20% for each of the validation and test sets.

Steps to perform PCA using MATLAB:

1. Open MATLAB, and Run special MATLAB script to plot the first two PCs.
2. Then the MATLAB will ask for the data excel file name. Thus enter the name of the file that contains the activities and all descriptors of models which were chosen after the second MLR validation.
3. A figure of first two PCs will produced, from plotting (new data (:,x),new data (:,y),'+').

Where: x label ('xth Principal Component'); y label ('yth Principal Component').

4. Using the data distribution from the figure produced for the first two PCs, select the training, validation and test sets molecules.

Hint: If the first two PCs were not enough to describe the data distribution, plotting the third PC can be helpful.

2.2.4 Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) is used either when the linear method of analysis is not producing good predicting models or when more evidence that the linear method of analysis is good predicting method.

2.2.4.1 Steps to perform ANN for each model using MATLAB:

1. Use the same models which used in PCA to divide the dataset, prepare excel file of activity and descriptors for each model.
2. Open MATLAB, and Run special MATLAB script for ANN.
3. Then the MATLAB will ask for the data excel file name. Thus enter the name of the file.
4. After that the MATLAB will ask for model number and the number of hidden nodes, see figure 2-5.
5. Choose the best models which have high R value for test set and low PRESS and RESP values.

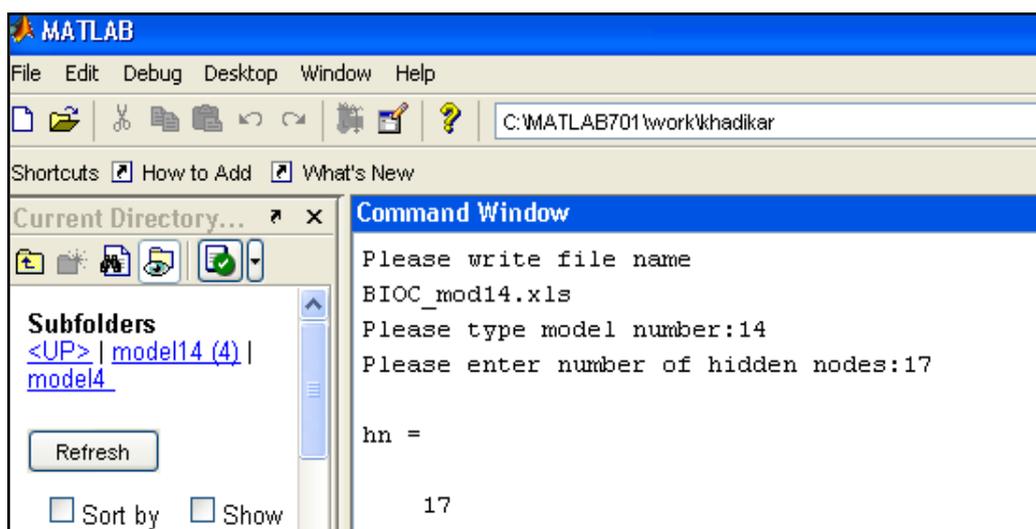


Figure 2-5: MATLAB Command window asking for file name, model number and number of hidden nodes.

2.2.4.2 Steps to perform ANN of the best models with range of hidden nodes (Hn) using MATLAB:

1. After choosing the optimal models based on the first ANN round, perform the ANN again for each model with range of hidden nodes starting from 5 to 20 by repeating the same steps in section 2.2.4.1.
2. Pick best models having high R value for test set, low PRESS, low RESP values, and small number of hidden nodes.

2.2.5 Randomization test (chance correlation or scrambling model)

Randomization test is performed in this research to ensure that the ANN resulted model is not due to a chance.

Steps to perform Randomization test using MATLAB:

1. Prepare file for each model resulted from the ANN, the file content is similar to the content of the LMO data files.
2. Run special MATLAB script, then enter the data file name and the number of trail when asked to.
3. Repeat the test for each model more than ten times.

Summary of QSAR process:

- ✓ Dataset preparation (Chemical structure's and their pIC50)
- ✓ Geometry optimization through semi-empirical quantum mechanics using HyperChem.
- ✓ Descriptors calculation using HyperChem and Dragon.
- ✓ MLR Model building using SPSS as well as validation of these models
- ✓ PC-ANN statistical model using MATLAB as well as validation of the models

Chapter three

Results and Discussion

3. Results and Discussion:

QSAR models were developed as a result of this study using the 136 compounds and their related observed activity (pIC_{50}) against Translocator protein (TSPO).

3.2 Data preparation results

- ✓ Compounds optimization using HyperChem resulted with optimized 136 compounds, through semi-empirical AM1 (Austin Model 1) method. Where the semi-empirical method is used because it is very fast compared with *ab initio* method, applicable to large molecules, and give accurate results.

As well AM1 prove that it's more reliable method than other semi-empirical methods (e.g. MNDO).

- ✓ Descriptors calculation using HyperChem; allow to calculate one group of descriptors called G-16 quantum chemical descriptors. All the descriptors calculated through HyperChem is mentioned below and in section 2.1.3.1.
 - HOMO (highest occupied molecular orbital).
 - LUMO (Lowest occupied molecular orbital).
 - Heat of formation (kcal/mol).
 - Dipole moment (Debyes).
 - Log P.
 - Hardness ($0.5 * (LUMO - HOMO)$).
 - Softness ($1/Hardness$).
 - Electronegativity ($-0.5 * (LUMO + HOMO)$).
 - Electrophilicity ($Electronegativity * Electronegativity / (2 * Hardness)$) [91].

- Surface Area (Approx).
 - Surface Area (Grid).
 - Volume.
 - Hydration Energy.
 - Refractivity.
 - Polarizability.
 - Mass.
- ✓ Descriptors calculation using Dragon; 1235 descriptors have been calculated, in which represented through 18 groups. The results are explained below:
- Two groups (Empirical and Properties descriptors) were constant or near constant. Thus Dragon software discard these groups of descriptors because they are correlated with each other and with activity at the same time.
 - Other groups of descriptors; constitutional, topological, molecular walk counts, BCUT, Galvez topological charge indices, 2D autocorrelations, charge descriptors, aromaticity indices, Randic molecular profiles, geometrical, RDF, 3D-MoRSe, WHIM, GETAWAY, functional, and atom-centered fragments were calculated with non-constant descriptors.
- Examples: A 37 descriptors were calculated within the constitutional group, a 225 descriptors within the topological group and so on...
- ✓ Performing the first MLR using SPSS, in which an MLR for each group of descriptors performed separately, except for the groups which contain small number of descriptors such as charge descriptors, aromaticity indices and G-16 quantum chemical were gathered in one input file for the MLR.

Results of first MLR is summarized in table 3-1, where (No.) refers to group number, (R) refers to correlation coefficient, (R^2) refers to coefficient of determination, ($R^2_{adj.}$) refers to adjusted R^2 , and selected descriptors refer to chosen descriptors by MLR model.

Table 3-1: MLR Models resulted from each group of descriptors.

No.	Group name	# of calculated descriptors	R	R^2	$R^2_{adj.}$	Standard Error of estimation	Selected descriptors
1	Constitutional	37	0.484	0.234	0.176	1.189	Ms, nR10, RBF, nCL, nH, nTB, nR11, nR09, nAB, nCIC, nDB, Mp, AMW, nR07, nCIR, Mv, RBN, Ss, nSK, nX, Sp, Me, nR05
2	Topological	225	0.806	0.649	0.585	0.798	HNar, X4Av, piPC10, SPI, piPC09, D/Dr10, piPC07, D/Dr06, T(O..Cl), CIC2, X5v, T(Cl..Cl), Jhetp, piPC06, piPC03, MPC06, X2sol, SEige, X2Av, D/Dr11, D/Dr07
3	Molecular walk	19	0.532	0.283	0.213	1.098	MWC08, MWC06, SRW08, SRW05, SRW10, MWC05, MWC04, SRW04, MWC09, MWC01, MWC07, SRW07

No.	Group name	# of calculated descriptors	R	R ²	R ² adj.	Standard Error of estimation	Selected descriptors
4	BCUT	64	0.781	0.609	0.521	0.857	BELm2, BELm3, BELe4, BEHe5, BEHv8, BELp1, BEHm4, BEHm8, BELv3, BEHv2, BELm5, BELv5, BEHm5, BELm1, BELe2, BELp3, BEHp4, BEHe4, BELp8, BELm8, BELe5, BEHv4, BEHe6, BELv7, BELm7
5	Galvez topological charge indices	21	0.708	0.501	0.409	0.952	JGI2, JGI6, JGT, JGI3, JGI4, JGI10, GGI6, GGI2, GGI8, GGI1, JGI1, JGI8, GGI3, GGI10, JGI9, GGI7, JGI7, JGI5, GGI4, GGI5, GGI9
6	2D autocorrelations	96	0.729	0.532	0.495	0.880	MATS4e, MATS1v, GATS6e, GATS7v, ATS4p, MATS5p, MATS5m, GATS8e, GATS7p, MATS7m
7+8+17	Charge+ Aromaticity+ G16	35	0.648	0.419	0.300	1.036	dipole moment (Debyes), Hardness, TE1, Hydration Energy (kcal/mol), Qneg2, Log P, qpos, qneg, Mass (amu), Polarizability, LDip, Surface Area (Grid), Volume, Qmean, Refractivity, TE2, heat of formation (kcal/mol), PCWTe, electrophilicity, HOMO (eV), Q2, RPCG, softness

No.	Group name	# of calculated descriptors	R	R ²	R ² adj.	Standard Error of estimation	Selected descriptors
9	Randic molecular profiles	41	0.367	0.135	0.073	1.19163	DP20, SHP2, DP01, DP04, SP02, SP15, SP01, SP13, DP17
10	Geometrical	31	0.539	0.290	0.129	1.15519	DELS, SPH, PJI3, G(N..N), TIE, SPAN, H3D, SPAM, SEig, MEcc, MAXDP, MAXDN, W3D, G(O..Cl), G(F..Cl), J3D, G(Cl..Cl), G2, FDI, ADDD, G1, AGDD, G(N..O), ASP, L/Bw
11	RDF	150	0.658	0.433	0.372	0.98087	RDF125m, RDF030m, RDF035m, RDF075v, RDF090e, RDF070u, RDF125v, RDF130e, RDF120v, RDF105m, RDF090m, RDF015u, RDF050v
12	3D-MorSE	160	0.673	0.453	0.374	0.97902	Mor10u, Mor15e, Mor04p, Mor22e, Mor28v, Mor31u, Mor02u, Mor32e, Mor11e, Mor05m, Mor10v, Mor11u, Mor15u, Mor10p, Mor19p, Mor12u, Mor08m

No.	Group name	# of calculated descriptors	R	R ²	R ² adj.	Standard Error of estimation	Selected descriptors
13	WHIM	99	0.613	0.376	0.279	1.051	L2v, G1v, P2m, G1e, L2p, L2e, Vv, Vp, Ve, E2s, Vu, L2u, Vm, L2m, Ds, E3m, Dm, G3s
14	GETAWAY	196	0.784	0.615	0.544	0.836	R8e+, R1p+, H5e, R4e+, R4u+, R4m+, H3u, HGM, HATS1e, R7m, ITH, R1m+, H5m, R6v+, HATS6p, H5u, HATSv, HATS2p, R8u, H6p, R8m
15	Functional	24	0.694	0.481	0.385	0.970	nCONR2, nN-N, nCONR2Ph, nC=NPh, nCrH2, nNR2, nCONHRPh, n#CH, nCs, nCaH, nCt, n=CHR, nPhX, nCaR, nRORPh, nCp, nCrHR, nC=N, nCONHR, nNHR, nHDon
16	Atom-centered fragments	38	0.795	0.632	0.553	0.828	C-005, C-043, C-027, N-068, N-071, H-050, C-003, C-031, C-040, C-001, C-008, H-054, C-028, C-021, O-059, N-075, C-026, C-006, N-073, H-051, O-060, H-047, C-024, C-016

- ✓ Performing the second MLR using SPSS, in which MLR applied on the groups of descriptors resulted from the first MLR together.

Results of second MLR is summarized in table 3-2, where only the models having $R^2 \geq 0.6$ were taken to continue into the next step. So models 12- 24 which have $R^2 \geq 0.6$ were taken for cross validation (leave one out and leave many out) [92].

Table 3-2: MLR Models resulted from all the groups of descriptors together

Model No.	No. of descriptors	R	R^2	R^2_{adj}	Selected descriptors
12	12	0.787	0.620	0.583	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4
13	13	0.803	0.645	0.607	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e
14	14	0.815	0.665	0.626	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, BELm3
15	15	0.828	0.686	0.647	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, BELm3, nHDon
16	16	0.842	0.709	0.670	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, BELm3, nHDon, C-003

Model No.	No. of descriptors	R	R ²	R ² adj.	Selected descriptors
17	17	0.851	0.725	0.685	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, BELm3, nHDon, C-003, MATS4e
18	18	0.862	0.742	0.703	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, BELm3, nHDon, C-003, MATS4e, BEHm4
19	19	0.875	0.765	0.727	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, BELm3, nHDon, C-003, MATS4e, BEHm4, Hydration Energy (kcal/mol)
20	20	0.886	0.784	0.747	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, BELm3, nHDon, C-003, MATS4e, BEHm4, Hydration Energy (kcal/mol), G(O..Cl)
21	21	0.893	0.797	0.760	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, BELm3, nHDon, C-003, MATS4e, BEHm4, Hydration Energy (kcal/mol), G(O..Cl), electrophilicity
22	22	0.897	0.805	0.767	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, BELm3, nHDon, C-003, MATS4e, BEHm4, Hydration Energy (kcal/mol), G(O..Cl), electrophilicity, Mor22e

Model No.	No. of descriptors	R	R ²	R ² adj.	Selected descriptors
23	23	0.902	0.813	0.775	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, BELm3, nHDon, C-003, MATS4e, BEHm4, Hydration Energy (kcal/mol), G(O..Cl), electrophilicity, Mor22e, Mor08m
24	24	0.909	0.826	0.788	JGI2, Mor10u, C-005, R8e+, nN-N, nR10, Mor19p, RDF035m, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, BELm3, nHDon, C-003, MATS4e, BEHm4, Hydration Energy (kcal/mol), G(O..Cl), electrophilicity, Mor22e, Mor08m, Mor11e

The below equation represents the best MLR model number 24;

$$\begin{aligned}
 \text{pIC}_{50} = & 32.804 (\pm 5.902) - 8.386 (\pm 6.190) \text{JGI2} - 0.159 (\pm 0.122) \text{Mor10u} + 0.249 (\pm 0.120) \text{C-005} \\
 & - 6.653 (\pm 3.382) \text{R8e}^+ - 2.222 (\pm 0.258) \text{nN-N} + 0.661 (\pm 0.192) \text{nR10} \\
 & + 0.537 (\pm 0.191) \text{Mor19p} + 0.106 (\pm 0.022) \text{RDF035m} - 0.043 (\pm 0.031) \text{RDF030m} - \\
 & 1.288 (\pm 0.348) \text{nCONHRPh} - 205.791 (\pm 27.444) \text{X4Av} - 12.548 (\pm 1.485) \text{BEHe4} - \\
 & 36.414 (\pm 15.927) \text{G1e} + 8.988 (\pm 1.817) \text{BELm3} - 0.106 (\pm 0.0947) \text{nHDon} - 2.671 \\
 & (\pm 0.382) \text{C-003} - 4.090 (\pm 0.913) \text{MATS4e} + 5.928 (\pm 1.338) \text{BEHm4} + 0.255 \\
 & (\pm 0.057) \text{Hydration Energy} + 0.032 (\pm 0.007) \text{G(O..Cl)} - 1.366 (\pm 0.471) \\
 & \text{electrophilicity} - 0.525 (\pm 0.181) \text{Mor22e} - 0.360 (\pm 0.102) \text{Mor08m} - 0.288 (\pm 0.103) \\
 & \text{Mor11e}.
 \end{aligned}$$

Where R=0.909, R² = 0.826, R²adj =0.788, and the STD error of the estimate = 0.5700.

Each descriptor in model 24 equation is mentioned with brief description and its group in table 3-3.

Table 3-3: Brief description of the descriptors in the best MLR model equation.

Name	Description	Block (group)
JGI2	Mean topological charge index of order 2	Galvez topol. Charge indices
Mor10u	Signal 10 / unweighted	3D-MoRSE descriptors
C-005	CH3X	Atom-centred fragments
R8e+	R maximal autocorrelation of lag 8 / weighted by Sanderson electronegativity	GETAWAY descriptors
nN-N	Number of N hydrazines	Functional group counts
nR10	Number of 10-membered rings	Ring descriptors
Mor19p	Signal 19 / weighted by polarizability	3D-MoRSE descriptors
RDF035m	Radial Distribution Function - 035 / weighted by mass	RDF descriptors
RDF030m	Radial Distribution Function - 030 / weighted by mass	RDF descriptors
nCONHRPh	Number of secondary amides (aromatic)	Functional group counts
X4Av	Average valence connectivity index of order 4	Connectivity indices
BEHe4	Highest eigenvalue n. 4 of Burden matrix / weighted by atomic Sanderson electronegativities	BCUT
G1e	1st component symmetry directional WHIM index / weighted by Sanderson electronegativity	WHIM descriptors
BELm3	Lowest eigenvalue n. 3 of Burden matrix / weighted by atomic masses	BCUT
nHDon	Number of donor atoms for H-bonds (N and O)	Functional group counts
C-003	CHR3	Atom-centred fragments

Name	Description	Block (group)
MATS4e	Moran autocorrelation of lag 4 weighted by Sanderson electronegativity	2D autocorrelations
BEHm4	Highest eigenvalue n. 4 of Burden matrix / weighted by atomic masses	BCUT
Hydration Energy (kcal/mol)	Hydration Energy (kcal/mol)	G16-quantum-chemical
G(O..Cl)	Sum of geometrical distances between O..Cl	Geometrical descriptors
Electrophilicity	Electrophilicity	G16-quantum-chemical
Mor22e	Signal 22 / weighted by Sanderson electronegativity	3D-MoRSE descriptors
Mor08m	Signal 08 / weighted by mass	3D-MoRSE descriptors
Mor11e	Signal 11 / weighted by Sanderson electronegativity	3D-MoRSE descriptors

Based on the equation of the best MLR model, the following descriptors have a positive effect on the compounds activity:

C-005, nR10, Mor19p, RDF035m, BELm3, BEHm4, Hydration Energy, G(O..Cl).

While the below descriptors have a negative effect on the compounds activity,

JGI2, Mor10u, R8e⁺, nN-N, RDF030m, nCONHRPh, X4Av, BEHe4, G1e, nHDon, C-003, MATS4e, electrophilicity, Mor22e, Mor08m, Mor11e.

- ✓ Cross validation performed on the MLR resulted models (12-24), using MATLAB software. The results of cross validation LOO and LMO are summarized in table (3-4) and (3-5) respectively. Where: PRESS (Predictive residual sum of squares) which also called SSE (Error sum of squares). PRESS is standard index to measure the accuracy of the model, SST (Total sum of squares), R^2_{CV} or Q^2 (Cross validated correlation coefficient), SPRESS (uncertainty of prediction), PSE (Predictive Square

Errors) which also called RMSE (Root Mean Square Error), and RSEP (Relative Standard Error of Prediction).

Table 3-4 and 3-5 show a good predictive power for models 19-24 because of having high R^2_{cv} and PRESS/SST less than 0.4. Thus, models 19-24 were chosen for ANN analysis.

Table 3-4: LOO cross validation results.

model	No. desc.	PRESS	SPRESS	SST	R^2_{cv}	PRESS/SST	PSE	RSEP
12	12	87.824	0.845	125.942	0.302	0.697	0.803	11.004
13	13	87.431	0.846	125.303	0.302	0.697	0.801	10.980
14	14	78.757	0.806	132.741	0.406	0.593	0.760	10.421
15	15	74.705	0.789	136.882	0.454	0.545	0.741	10.149
16	16	68.680	0.759	142.558	0.518	0.481	0.710	9.731
17	17	67.408	0.755	144.202	0.532	0.467	0.704	9.641
18	18	63.928	0.739	147.714	0.567	0.432	0.685	9.389
19	19	56.146	0.695	154.679	0.637	0.362	0.642	8.799
20	20	54.708	0.689	156.560	0.650	0.349	0.634	8.685
21	21	52.413	0.678	158.942	0.670	0.329	0.620	8.501
22	22	50.644	0.669	160.676	0.684	0.315	0.610	8.356
23	23	49.162	0.662	162.211	0.696	0.303	0.601	8.233
24	24	46.108	0.644	165.143	0.720	0.279	0.5822	7.973

Table 3-5: LMO cross validation results.

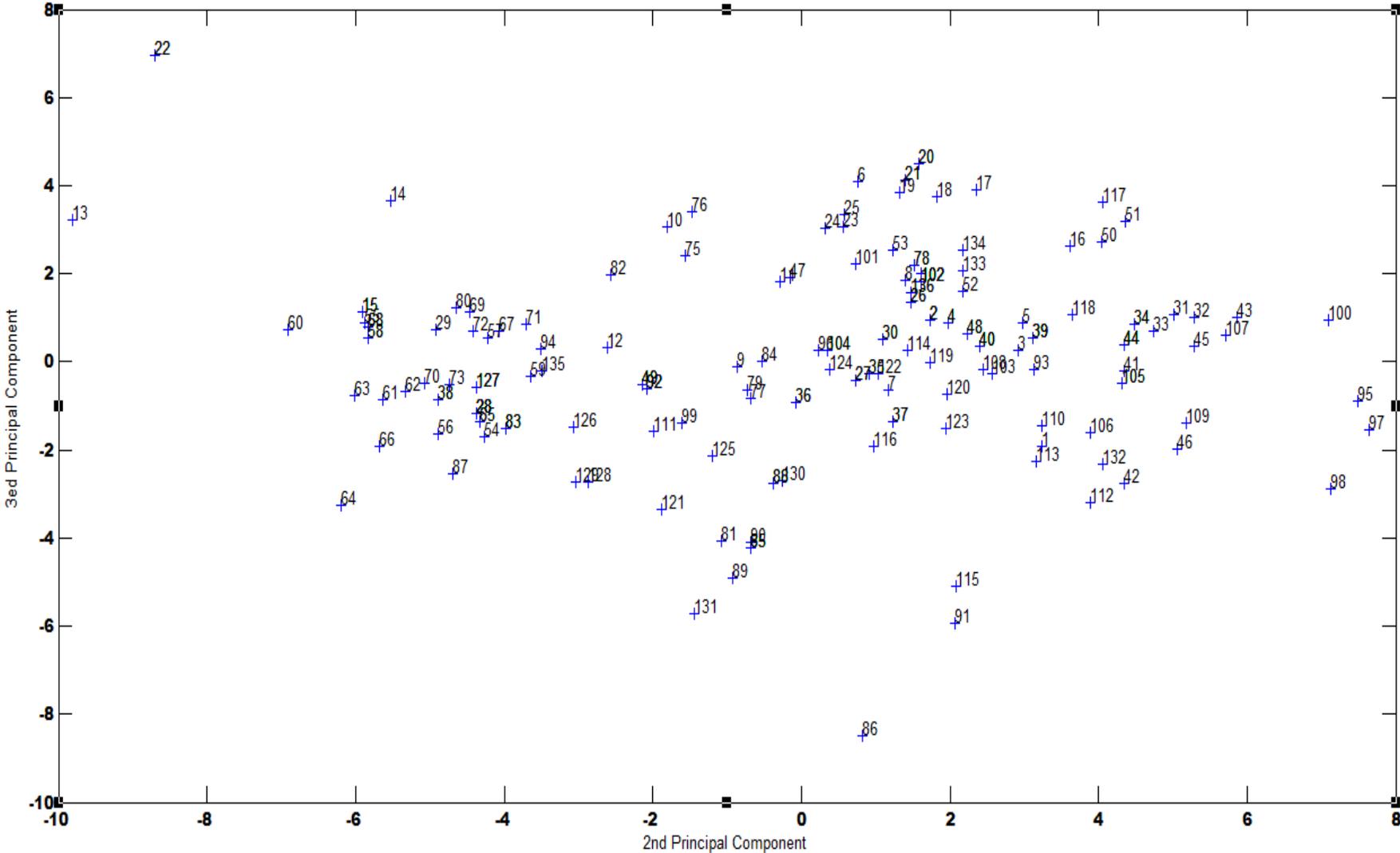
Model	No. desc.	PRESS	SPRESS	SST	R ² _{CV}	PRESS/SST	PSE	RSEP
12	12	102.289	0.911	145.424	0.296	0.703	0.867	11.795
13	13	95.909	0.886	152.582	0.371	0.628	0.839	11.421
14	14	89.930	0.862	154.158	0.416	0.583	0.813	11.060
15	15	84.282	0.838	157.975	0.466	0.533	0.787	10.707
16	16	83.151	0.835	162.574	0.488	0.511	0.781	10.635
17	17	82.525	0.836	172.536	0.521	0.478	0.779	10.595
18	18	77.411	0.813	174.643	0.556	0.443	0.754	10.261
19	19	69.566	0.774	177.321	0.607	0.392	0.715	9.727
20	20	64.308	0.747	179.401	0.641	0.358	0.687	9.352
21	21	62.574	0.740	189.644	0.67	0.33	0.678	9.225
22	22	59.976	0.728	191.991	0.687	0.312	0.664	9.032
23	23	58.036	0.719	195.521	0.703	0.296	0.653	8.885
24	24	52.387	0.687	195.586	0.732	0.267	0.620	8.441

✓ The PCA was performed to divide the molecules into training, validation, and prediction (test) sets. Performing PCA on the whole data of 136 compounds, 24 descriptors and plotting the first and second principals, first and third principals, and second and third principals. The data division into 60% training, 20% test and 20% validation, should be in equal manner in which picking one compound from each zone to each set.

The first and second principals and first and third principals plots were having a condensed data towards the X axis, however second and third principals plot have the data distributed in a good way in comparison with the other plots.

Therefore relying on the second and third principals plot, it shows compounds 13, 22 and 86 as outliers (Figure 3-1). Although these three compounds don't differ structurally in comparison with other compounds. But they behave in a different manner, therefore these compounds removed from the data in the next analysis. And so the data divided after removing the outliers into 60% (81 compounds) training group, 20% (26 compounds) of each test and validation groups.

Figure 3-1: Second and third principal components plot.



✓ First Artificial Neural Networks (ANN), Performed on the chosen models (19-24) from LOO and LMO validation. Apply the ANN on each model with 7 hidden nodes.

The results of first ANN is in table 3-6, the table shows that model 24 has the highest correlation coefficient for the test set (0.850) indicating its high predictive power and the one after it is model 21.

Figure 3-2 shows the relation of PRESS values for the training, test and validation sets versus model number. This figure shows that the minimum PRESS of the training set is obtained for model 21 the one after it is model 20. While the minimum PRESS of the test sets is obtained for model 24 the one after it is model 21 then 23.

Figure 3-3 shows the relation of correlation coefficient (R) values for the training, test and validation sets versus model number. This figure shows that the highest (R) value of the training set is obtained for model 21 then 20. While the highest (R) value of the test set is obtained for model 24 then model 21 then model 23.

Figure 3-4 shows the relation of R^2_{CV} (Cross validated correlation coefficient) values for the training, test and validation sets versus model number. This figure shows that the highest (R^2_{CV}) value of the training set is obtained for model 21 then 20. While the highest (R^2_{CV}) value of the test set is obtained for model 21 then model 24 then model 23.

Accordingly, models 20, 21, 23, and 24 were subjected for further analysis by optimizing the number of hidden nodes, because these models have the highest R, R^2_{CV} and low PRESS values for test set.

Table 3-6: Correlation Coefficient and Cross Validation Parameters for ANN Models 19-24.

Mo.#	hn	nPCs	R_tr	PRESS_tr	R ² _{CV_tr}	R_test	PRESS_test	R ² _{CV_test}	R_val	PRESS_val	R ² _{CV_val}
19	7	7	0.905	19.605	0.753	0.805	15.511	0.401	0.718	26.234	0.210
20	7	7	0.927	15.319	0.824	0.812	15.064	0.437	0.726	26.259	0.298
21	7	7	0.932	14.085	0.840	0.832	14.034	0.607	0.740	25.528	0.358
22	7	6	0.911	18.292	0.789	0.802	15.934	0.335	0.686	28.443	-0.047
23	7	6	0.906	19.118	0.783	0.818	14.526	0.502	0.655	30.734	-0.133
24	7	6	0.909	18.701	0.785	0.850	12.206	0.578	0.685	28.852	0.108

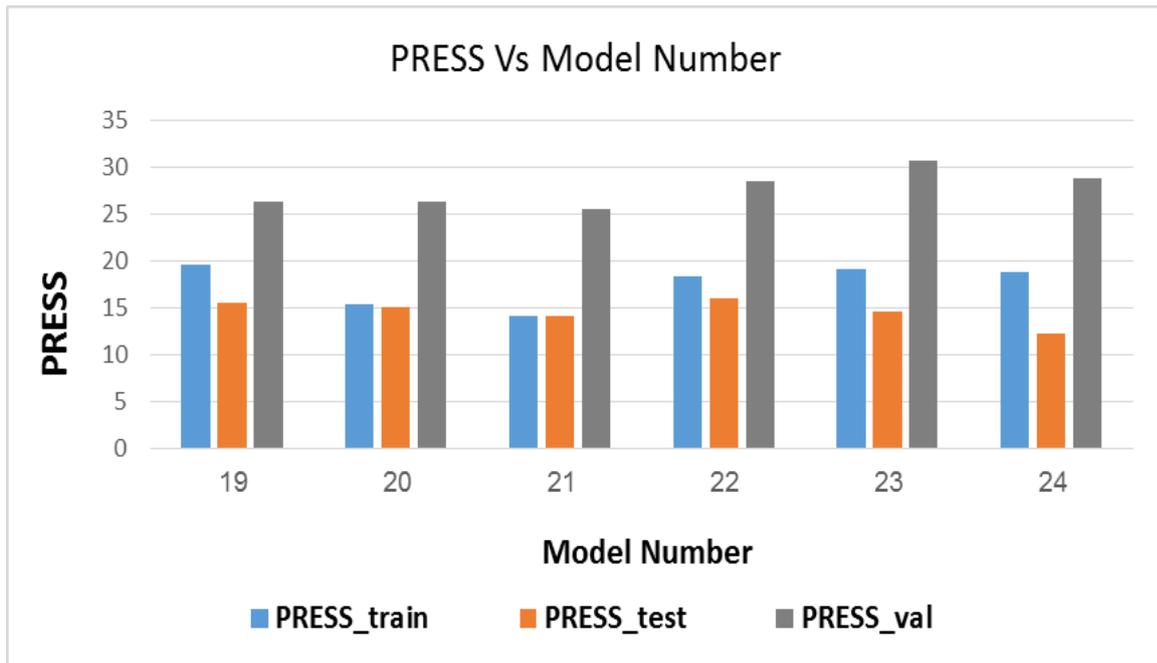


Figure 3-2: Plots of ANN Predictive Residual Sum of Squares (PRESS) values for the training, test and validation sets versus model number.

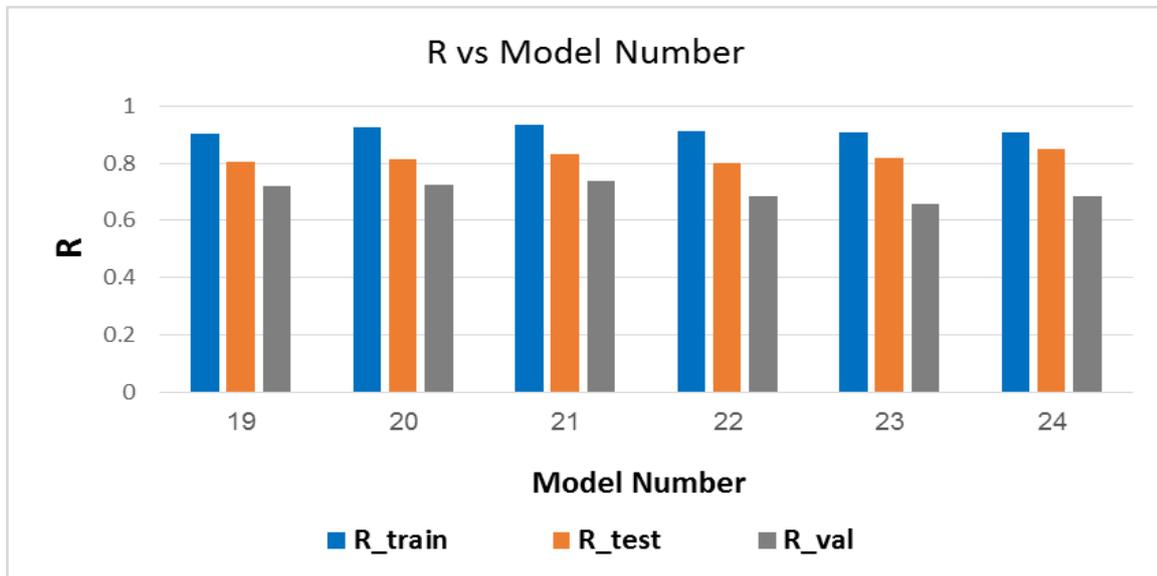


Figure 3-3: Plots of ANN correlation coefficient (R) values for the training, test and validation sets versus model number.

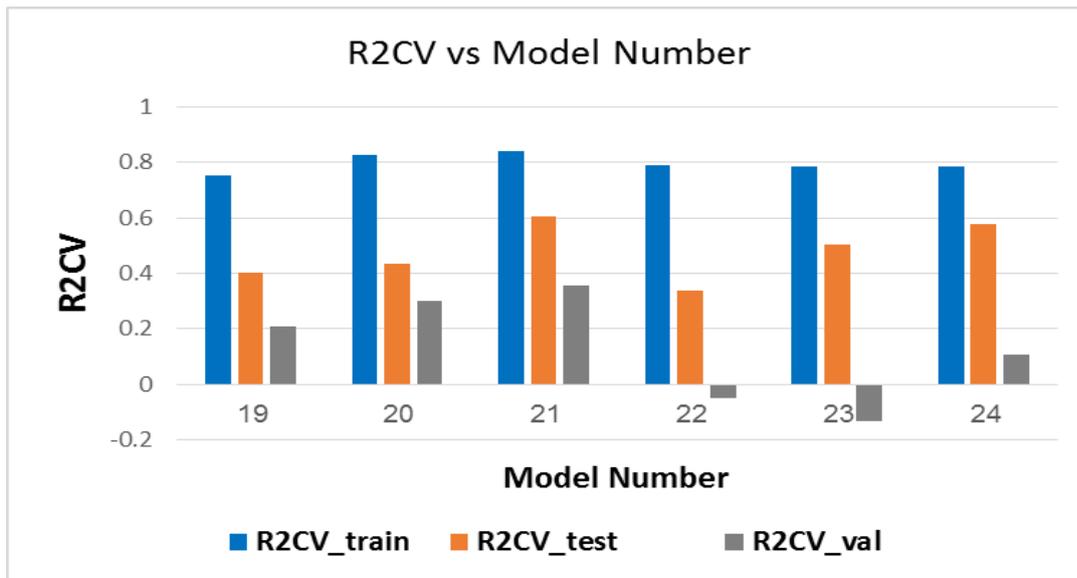


Figure 3-4: Plots of ANN R_{CV}^2 (Cross validated correlation coefficient) values for the training, test and validation sets versus model number.

- ✓ Second ANN performed on the chosen models 20, 21, 23, 24, each model with a range of hidden nodes starting from 5 to 20. The results are shown in tables; 3-7, 3-8, 3-9 and 3-10 respectively.

According to the results tables; model 20 with 10 hidden nodes, model 21 with 7 hidden nodes, model 23 with 5 hidden nodes, and model 24 with 7 hidden nodes were chosen as the best models with the optimal hidden nodes because they have high prediction power (R), minimum PRESS value of the test group, and minimum number of hidden nodes.

- ✓ Table 3-11 summarize the correlation coefficients and cross validation parameters for the optimal number of hidden nodes for each one of the chosen models, where Models 23 and 24 chosen as best models to continue to randomization test.

Table 3-7: Correlation Coefficients and Cross Validation Parameters of Number of Hidden Nodes for Model #20.

hn	nPCs	R_tr	PRESS_tr	R ² _{CV_tr}	RSEP_tr	R_test	PRESS_test	R ² _{CV_test}	RSEP_test	R_val	PRESS_val	R ² _{CV_val}	RSEP_val
5	7	0.903	19.757	0.768	6.694	0.842	13.437	0.483	9.496	0.807	20.161	0.471	12.207
6	7	0.907	19.034	0.787	6.571	0.808	15.277	0.515	10.125	0.728	25.439	0.259	13.712
7	7	0.927	15.319	0.82	5.895	0.812	15.064	0.437	10.054	0.726	26.259	0.299	13.931
8	7	0.922	16.072	0.814	6.038	0.815	15.467	0.520	10.188	0.708	26.973	0.158	14.119
9	7	0.912	18.143	0.784	6.415	0.807	15.290	0.416	10.129	0.671	29.785	-0.030	14.837
10	7	0.938	13.007	0.859	5.432	0.847	13.384	0.483	9.477	0.736	26.058	0.361	13.877
11	7	0.921	16.289	0.820	6.079	0.817	15.042	0.359	10.047	0.688	29.179	0.042	14.684
12	7	0.908	18.927	0.774	6.552	0.807	15.419	0.366	10.172	0.669	30.706	-0.199	15.064
13	7	0.903	20.615	0.739	6.838	0.804	15.704	0.517	10.266	0.679	29.911	0.081	14.868
14	7	0.917	17.037	0.811	6.217	0.809	15.277	0.377	10.125	0.695	29.035	0.254	14.649
15	7	0.922	16.162	0.812	6.055	0.818	15.358	0.614	10.152	0.678	32.198	0.174	15.426
16	7	0.908	18.879	0.781	6.544	0.847	12.532	0.644	9.170	0.693	28.242	0.063	14.447
17	7	0.913	17.872	0.793	6.367	0.817	14.673	0.464	9.923	0.722	27.241	0.338	14.189
18	7	0.928	15.041	0.823	5.841	0.809	15.818	0.265	10.303	0.707	28.550	0.123	14.526
19	7	0.909	18.727	0.783	6.518	0.801	16.211	0.315	10.429	0.748	24.144	0.293	13.358
20	7	0.926	15.574	0.812	5.944	0.811	15.109	0.469	10.069	0.757	24.439	0.328	13.439

Table 3-8: Correlation Coefficients and Cross Validation Parameters of Number of Hidden Nodes for Model #21.

hn	nPCs	R_tr	PRESS_tr	R ² _{CV_tr}	RSEP_tr	R_test	PRESS_test	R ² _{CV_test}	RSEP_test	R_val	PRESS_val	R ² _{CV_val}	RSEP_val
5	7	0.902	20.061	0.773	6.746	0.839	14.090	0.498	9.723	0.756	25.788	0.291	13.81
6	7	0.903	19.896	0.766	6.718	0.813	14.874	0.494	9.991	0.694	28.482	0.185	14.509
7	7	0.932	14.085	0.840	5.652	0.832	14.034	0.607	9.704	0.739	25.528	0.358	13.736
8	7	0.901	20.916	0.716	6.888	0.829	14.141	0.431	9.742	0.735	26.559	0.139	14.010
9	7	0.901	20.259	0.763	6.779	0.807	16.340	0.471	10.472	0.729	27.239	0.187	14.188
10	7	0.912	18.574	0.759	6.491	0.838	13.134	0.539	9.388	0.672	29.733	-0.314	14.824
11	7	0.926	15.351	0.834	5.901	0.811	16.364	0.579	10.479	0.732	26.434	0.124	13.98
12	7	0.917	17.107	0.807	6.229	0.800	17.048	0.571	10.696	0.686	30.444	0.244	14.999
13	7	0.923	15.824	0.823	5.991	0.810	15.321	0.368	10.14	0.716	27.400	0.159	14.230
14	7	0.900	20.319	0.768	6.789	0.829	13.876	0.596	9.649	0.654	32.595	0.029	15.521
15	7	0.905	19.758	0.746	6.695	0.813	15.798	0.241	10.296	0.669	32.587	0.106	15.519
16	7	0.901	20.166	0.769	6.763	0.804	15.683	0.431	10.259	0.728	25.551	0.099	13.742
17	7	0.919	17.195	0.778	6.245	0.817	14.738	0.441	9.945	0.689	29.422	0.077	14.746
18	7	0.929	14.557	0.841	5.746	0.856	11.967	0.544	8.961	0.724	25.729	0.1445	13.789
19	7	0.908	18.974	0.801	6.560	0.804	16.337	0.561	10.470	0.704	29.388	0.205	14.737
20	7	0.910	18.534	0.781	6.484	0.820	14.726	0.361	9.941	0.716	26.183	0.100	13.911

Table 3-9: Correlation Coefficients and Cross Validation Parameters of Number of Hidden Nodes for Model #23.

hn	nPCs	R_tr	PRESS_tr	R ² _{CV_tr}	RSEP_tr	R_test	PRESS_test	R ² _{CV_test}	RSEP_test	R_val	PRESS_val	R ² _{CV_val}	RSEP_val
5	6	0.905	19.511	0.769	6.653	0.832	14.019	0.468	9.699	0.680	30.185	0.183	14.936
6	6	0.909	18.783	0.789	6.527	0.804	16.982	-0.017	10.675	0.679	29.667	0.000	14.807
7	6	0.906	19.118	0.782	6.585	0.818	14.526	0.502	9.873	0.655	30.734	-0.132	15.071
8	6	0.928	15.063	0.824	5.845	0.812	15.349	0.445	10.149	0.722	26.162	0.266	13.905
9	6	0.931	14.482	0.833	5.731	0.803	15.675	0.372	10.256	0.785	20.716	0.435	12.373
10	6	0.903	19.740	0.781	6.692	0.801	16.609	0.137	10.558	0.714	27.067	0.227	14.143
11	6	0.908	19.090	0.759	6.580	0.800	15.915	0.376	10.334	0.618	34.389	-0.129	15.942
12	6	0.910	18.548	0.772	6.486	0.831	13.888	0.424	9.654	0.767	22.356	0.189	12.854
13	6	0.910	18.485	0.779	6.475	0.849	12.798	0.518	9.267	0.662	30.394	-0.05	14.988
14	6	0.917	17.426	0.783	6.287	0.801	16.244	0.362	10.441	0.625	34.708	-0.458	16.016
15	6	0.901	20.441	0.775	6.809	0.818	15.041	0.603	10.047	0.704	27.055	0.038	14.140
16	6	0.939	12.585	0.862	5.343	0.871	10.868	0.603	8.540	0.673	31.389	0.123	15.230
17	6	0.902	19.957	0.769	6.728	0.803	16.809	0.075	10.621	0.659	31.511	-0.273	15.260
18	6	0.926	15.324	0.841	5.896	0.804	16.482	0.571	10.517	0.654	33.083	0.179	15.636
19	6	0.918	16.946	0.815	6.199	0.802	16.063	0.443	10.382	0.642	33.224	0.045	15.669
20	6	0.935	13.616	0.851	5.557	0.811	15.441	0.423	10.179	0.612	34.747	-0.115	16.025

Table 3-10: Correlation Coefficients and Cross Validation Parameters of Number of Hidden Nodes for Model #24.

hn	nPCs	R_tr	PRESS_tr	R ² _{CV_tr}	RSEP_tr	R_test	PRESS_test	R ² _{CV_test}	RSEP_test	R_val	PRESS_val	R ² _{CV_val}	RSEP_val
5	6	0.905	19.461	0.773	6.644	0.836	14.17	0.447	9.751	0.733	24.915	0.188	13.569
6	6	0.916	17.311	0.809	6.266	0.864	13.813	0.171	9.628	0.728	26.033	-0.108	13.871
7	6	0.909	18.701	0.785	6.513	0.850	12.206	0.578	9.050	0.685	28.852	0.108	14.602
8	6	0.939	12.766	0.861	5.381	0.802	15.810	0.518	10.30	0.659	32.616	0.076	15.526
9	6	0.901	20.272	0.763	6.781	0.815	15.077	0.346	10.059	0.626	32.803	-0.28	15.570
10	6	0.904	19.657	0.779	6.678	0.816	14.657	0.487	9.918	0.643	32.636	-0.042	15.530
11	6	0.918	17.379	0.783	6.279	0.831	14.127	0.418	9.737	0.655	33.441	-0.359	15.720
12	6	0.901	20.287	0.755	6.784	0.847	12.705	0.572	9.233	0.644	32.928	0.046	15.599
13	6	0.919	16.803	0.795	6.174	0.850	13.344	0.364	9.463	0.648	31.485	-0.169	15.254
14	6	0.931	14.464	0.836	5.728	0.846	12.668	0.626	9.219	0.726	26.054	0.185	13.876
15	6	0.919	16.688	0.808	6.153	0.837	13.353	0.554	9.466	0.703	28.696	-0.469	14.563
16	6	0.905	19.644	0.757	6.675	0.814	15.279	0.321	10.128	0.673	29.618	-0.385	14.795
17	6	0.912	18.087	0.782	6.405	0.841	13.534	0.413	9.530	0.773	21.857	0.231	12.709
18	6	0.939	12.555	0.873	5.337	0.823	14.911	0.415	10.002	0.691	29.843	-0.054	14.851
19	6	0.938	12.968	0.853	5.424	0.810	15.186	0.459	10.095	0.612	35.254	-0.257	16.141
20	6	0.942	12.428	0.881	5.309	0.839	13.094	0.621	9.374	0.647	33.343	-0.063	15.698

Table 3-11: Summary of the Correlation Coefficients and Cross Validation Parameters of the Optimal Number of Hidden Nodes of Each Model

Mo. #	hn	nPCs	R_tr	PRESS_tr	R ² _{CV_tr}	RSEP_tr	R_test	PRESS_test	R ² _{CV_test}	RSEP_test	R_val	PRESS_val	R ² _{CV_val}	RSEP_val
20	10	7	0.938	13.007	0.859	5.432	0.847	13.384	0.482	9.477	0.736	26.058	0.361	13.877
21	7	7	0.932	14.085	0.84	5.652	0.832	14.034	0.607	9.704	0.739	25.528	0.358	13.736
23	5	6	0.905	19.511	0.769	6.653	0.832	14.019	0.468	9.699	0.68	30.185	0.183	14.936
24	7	6	0.909	18.701	0.785	6.513	0.85	12.206	0.578	9.05	0.685	28.852	0.108	14.602

- ✓ ANN resulted model validation through randomization test, to ensure that the ANN resulted model is not due to a chance. Results of model 23 with 5 Hn and model 24 with 7 hn are shown in tables 3-12 and 3-13 respectively. These tables show that the Correlation coefficients obtained by chance are low in general while PRESS values are high. This indicates that models 23 and 24 which obtained from PCA-ANN are better than those obtained by chance and they are not due to chance.

Figures 3-5 and 3-6 show regressions between observed and predicted activity as well as their residuals for the training, validation, and test sets for these two models.

Table 3-12: Chance Correlation of Model 23 with 5 Hidden Nodes

Trial No.	nPCs	R_tr	PRESS_tr	R ² _{CV_tr}	R_test	PRESS_test	R ² _{CV_test}	R_val	PRESS_val	R ² _{CV_val}
1	6	0.155	186.679	-15.323	-0.248	11.546	-90.928	0.636	5.875	-32.209
2	6	0.151	199.228	-4.803	0.261	12.329	-219.405	-0.223	14.905	-82.514
3	6	0.065	204.245	-8.594	-0.37	11.879	-109.392	0.117	6.576	-46.601
4	6	0.130	189.856	-13.569	-0.348	11.517	-842.777	-0.785	9.682	-72.311
5	6	0.105	191.083	-18.737	-0.138	11.052	-377.239	-0.748	7.378	-317.342
6	6	-0.245	214.801	-31.479	0.311	10.846	-591.173	0.582	5.568	-728.723
7	6	0.116	374.285	-5.095	0.024	11.339	-39.912	-0.783	8.307	-11.770
8	6	0.037	213.845	-6.671	-0.027	11.011	-307.729	-0.630	6.815	-221.652
9	6	0.248	180.384	-6.080	0.198	15.225	-389.789	-0.6916	19.562	-106.548
10	6	-0.043	219.869	-7.936	0.358	9.702	-5.176	0.964	3.538	-8.333

Table 3-13: Chance Correlation of Model 24 with 7 Hidden Nodes

Trial No.	nPCs	R_tr	PRESS_tr	R ² _{CV_tr}	R_test	PRESS_test	R ² _{CV_test}	R_val	PRESS_val	R ² _{CV_val}
1	6	0.224	183.467	-6.709	0.189	12.548	-270.109	-0.257	11.440	-69.747
2	6	0.292	175.064	-10.017	-0.454	14.204	-1085.815	-0.108	13.204	-292.240
3	6	0.077	195.875	-12.219	-0.358	13.879	-63.759	0.354	3.999	-23.686
4	6	-0.277	229.782	-16.547	-0.381	13.326	-502.793	0.652	3.759	-51.454
5	6	0.076	200.859	-13.897	-0.345	11.929	-301.221	-0.808	11.332	-74.118
6	6	0.061	223.055	-4.677	-0.161	11.580	-19590.4	0.639	9.807	-1928.422
7	6	-0.173	283.432	-3.888	0.365	10.382	-27.965	-0.728	19.075	-8.395
8	6	0.047	257.630	-2.261	-0.454	13.189	-35.861	0.699	3.071	-6.059
9	6	-0.179	268.564	-4.842	-0.428	16.450	-19.304	0.443	3.942	-5.322
10	6	0.167	203.664	-3.476	0.009	11.528	-367.780	-0.692	4.978	-832.566

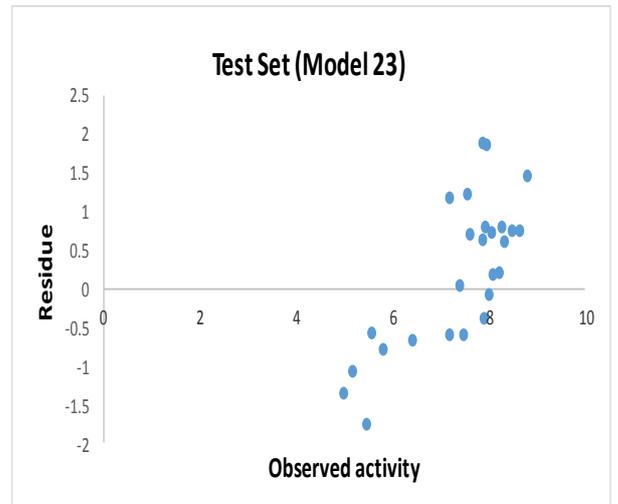
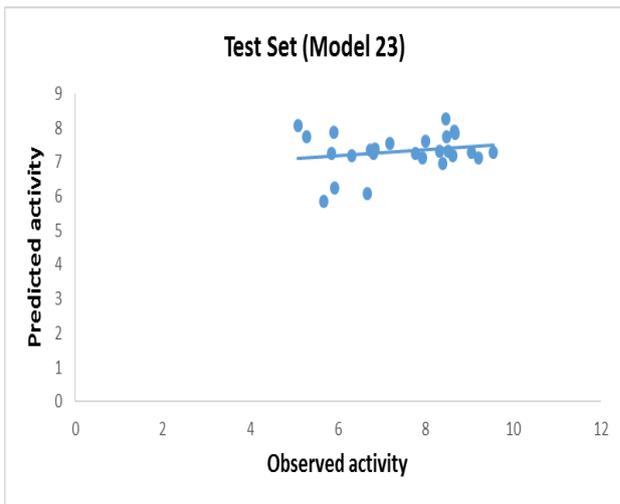
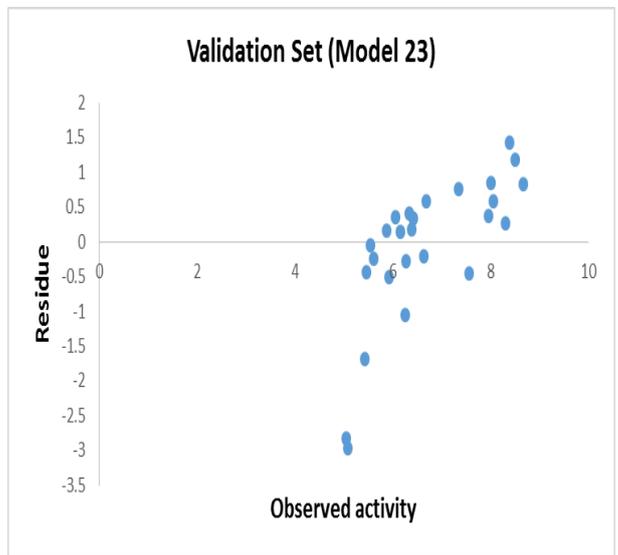
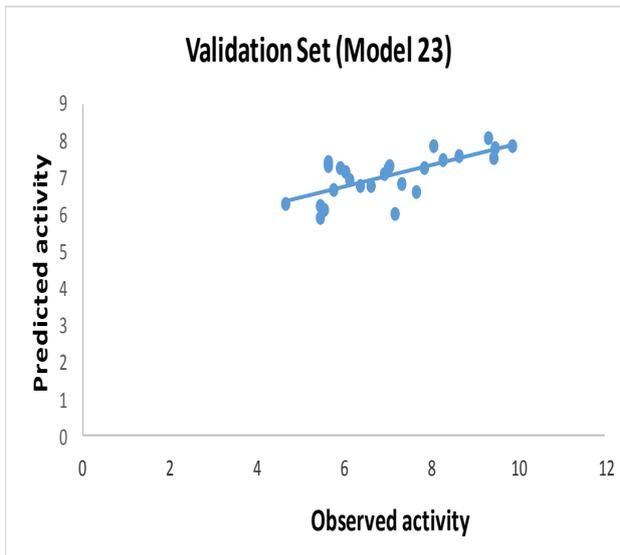
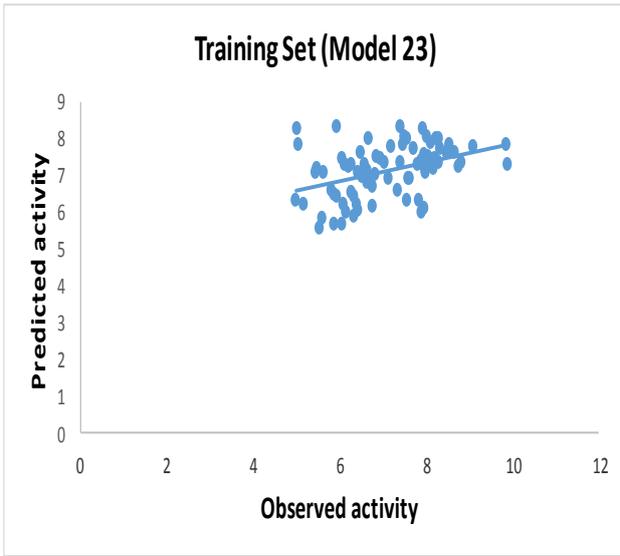


Figure 3-5: Plot of the predicted activity against observed one as well as their residues for model 23 using 5 hidden nodes. Training set, validation set, and external test set.



Figure 3-6: Plot of the predicted activity against observed one as well as their residues for model 24 using 7 hidden nodes. Training set, validation set, and external test set.

The following conditions proposed by Golbraikh and Tropsha [92] were applied to conclude that the QSAR model has acceptable prediction power if:

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

$$(2) R^2 > 0.6$$

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

Or

$$(R^2 - R^2_0) / R^2 < 0.1 \text{ and } 0.85 < k' < 1.15$$

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

$$(4) |R^2_0 - R'^2_0| < 0.3$$

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

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

Model 24 validated according to these criteria, and shows to have acceptable prediction power.

Structure-Activity Relationships of the Dataset:

✓ Compounds 1 to 16 in table 2-1, SAR [86]:

1- Position 3:

The introduction of substituents in position 3 of the quinoline nucleus (Compound 4) increased the TSPO affinity in variable degree depending on the stereoelectronic properties of the substituent involved.

The introduction of a methyl group produced an affinity enhancement of about an order of magnitude (compare 8 with 4), while an affinity increase of about 2 orders of magnitude was observed when a chloromethyl substituent was involved (compare 13 with 4). The introduction of a hydroxymethyl or differently substituted aminomethyl groups (compounds 14,15) had less dramatic effects on TSPO affinity, and the comparison of the affinities shown by 4,8,9-15 suggests that the presence in 3-position of substituents showing a wide range of stereoelectronic properties is compatible with a productive binding to TSPO.

2- Position 2:

Favorable effect of the introduction of a fluorine atom in position 2 of the pendent phenyl group (compare 10 vs. 8).

3- Tolerance showed by the receptor in accommodating the second benzyl group on the amide nitrogen (compare 11 vs. 8).

4- Slight superiority of the quinoline bicyclic system with respect to naphthalene (4 vs. 2 and 8 vs. 6).

✓ **Compounds 17 to 53 in table 2-1, SAR [87]:**

These findings are consistent with suggestions from QSAR analysis on the 2 phenylimidazo[1,2-*a*]pyridine derivatives which suggested that a four carbon chain is the optimum length for the alkyl substitution on the carboxamide nitrogen.

✓ **Compounds 80 to 91 in table 2-1, SAR [89]:**

1- Comparison of the results obtained with compounds 80 and 86 confirms the previously observed difference in affinity between secondary and tertiary amides. In fact, secondary amide 86 shows a significantly lower TSPO affinity when compared to its N-methylated counterpart 80.

2- The comparison of the most potent compounds 80, 81 with 85, 87 demonstrates the importance of a lipophilic substituent in para-position of the amide phenyl.

- The replacement of the amide phenyl group of compound 85 with the benzyl of 82 appears to be well tolerated by TSPO, whereas the same substitution with a propargyl moiety is not accepted equally well (compare 88 vs 85).
- The removal of the lipophilic chlorine atom in the pendant phenyl ring of the most active compounds 80, 81 leads to a decrease in TSPO affinity of about one order of magnitude (compounds 89, 90).
- The transformation of the amide carbonyl of 80 into the methylene group of 91 produces a dramatic decrease (2600 times) in the receptor affinity.

✓ **Compounds 92 to 135 in table 2-1, SAR [49]:**

Compounds 92-116 confirm the essential role of the carbonyl function as the primary pharmacophoric element, and the importance of the role of both the amide substituents and the pendant phenyl ring for which a dispersive nature of the interactions with TSPO binding site.

Compounds 117-135 show that the replacement of the ester function of 121 with a secondary (130) or tertiary amides (125- 127) affords compounds with similar micromolar range affinities, when compared with ester 121.

In addition, the environment of the carbonyl amide seems to be relatively sensitive to steric hindrance since the increase in size of the amide substituents results in a progressive decrease in affinity (compare compounds 126-129 with 125).

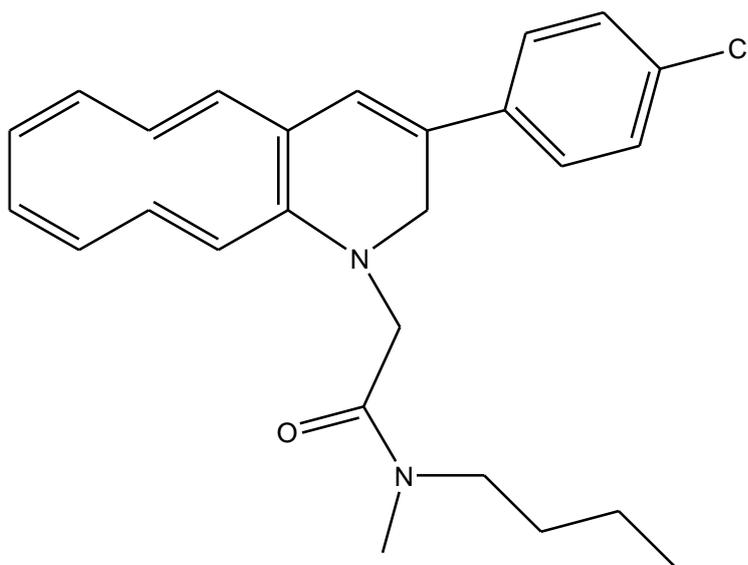
However suitably oriented lipophilic amide group plays a more substantial role and contributes to the binding strength much more than the one in the 3-position of the quinoline nucleus of 125-131. Taken together, these results suggest that the lipophilic amide groups in 3-position of the quinolone nucleus of compounds 125-131 occupy a receptor area different from the one occupied by the lipophilic amide groups of the high affinity TSPO ligands.

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

According to the previous SAR, and based on MLR model, the QSAR for TSPO ligand should have:

- 1- Less number of N hydrazine groups (nN-N).
- 2- Less number of secondary amides (aromatic) (nCONHRPh).
- 3- Less number of donor atoms for H-bonds (N and O).
- 4- Less number of CHR3 groups.
- 5- Less Electrophilicity is required.
- 6- More CH3X groups is useful.
- 7- Increase number of 10-membered rings is useful.

Below suggested compound as TSPO ligand:



According to model 24 equation, the PIC_{50} of the suggested compound is range between 20.7 – 31.9. Where the PIC_{50} of all the compounds used as dataset in this study was ranging between 4.6 – 9.8.

The suggested compound apply Lipinski rule of 5 with molecular weight less than 500, log P less than 5, no more than 5 hydrogen bond donors and no more than 10 hydrogen bond acceptors.

Comparison with previous QSAR studies:

There are seven Quantitative structure-activity relationship (QSAR) studies performed on TSPO ligands for different purposes, however only few of the studies performed on compounds of this study;

- ✓ A linear regression analysis on pCI_{50} values for compounds 17 to 53 listed in table 2-1 showed a good correlation ($R^2 = 0.870$) [87].

- ✓ Study on compounds which have serial numbers from 17 to 39 listed in table 2-1. was performed by Kunal Roy, Toropov and Raska in 2006, which achieved a QSAR modeling of peripheral Versus Central Benzodiazepine Receptor Binding Affinity of 37 compounds 2-Phenylimidazo[1,2-a] pyridineacetamides using Optimal Descriptors Calculated with SMILES (Simplified Molecular Input Line Entry System) [93]. The results indicate promising potential of the optimization of correlation weights based on SMILES notation in modeling studies.

- ✓ A quantitative structure-affinity relationships (QSAR) study performed on compounds 92 to 116 listed in table 2-1. Through comparison of the van der Waals volumes of the different ligands [49].

Other QSAR studies on TSPO ligands:

- ✓ 3D interaction model of endogenous and synthetic peripheral benzodiazepine receptor ligands was developed. Two lipophilic regions and one electrostatic interaction site are essential features for high affinity ligand binding, while a further lipophilic region plays an important modulator role. A comparative molecular field analysis, performed over 130 PBR ligands by means of the GRID/GOLPE methodology, led to a PLS model with both high fitting and predictive values ($r^2 = 0.898$, $Q^2 = 0.761$). The outcome from the 3D QSAR model and the GRID interaction fields computed on the putative endogenous PBR ligands DBI (Diazepam Binding Inhibitor) and TTN (Tetracontatetraneuropeptide) was used to identify the amino acids most probably involved in PBR binding. Three amino acids, bearing lipophilic side chains, were detected in DBI (Phe49, Leu47 and Met46) and in TTN (Phe33, Leu31 and Met30) as likely residues underlying receptor binding [94].

- ✓ Kunal and Sengupta in 2002 performed QSAR study for the binding affinities of 31 compounds of [2-phenylimidazo[1,2-a]pyridin derivatives with central benzodiazepine and peripheral benzodiazepine (TSPO) receptors using physico-chemical parameters. Attempt has been made to explore the structural and/or physico-chemical requirements of the compounds that are responsible for the selective action against peripheral benzodiazepine receptors over central ones [95].

- ✓ Dalai, Leonard & Kunal [96] performed a QSAR for TSPO binding affinity in 2006, with 35 compounds of 2-phenylpyrazolo(1,5-a)pyrimidin-3-yl-acetamides using topological and physicochemical descriptors and resulted with six models with

average $R^2=0.7$. The calculated hydrophobicity, $\log P_{\text{calc}}$, shows a parabolic relation with the TSPO receptor binding affinity, which suggests that the binding affinity increases with the increase in the partition coefficient of the compounds until it reaches the critical value after which the affinity decreases. The range of the optimum values of $\log P_{\text{calc}}$ is between 5.423-5.819 as found from different equations.

- ✓ Roy Kunal and Dalai performed a QSAR study in 2007 to explore the structural and physicochemical requirements of ligands N, N-dialkyl-2-phenylindol-3-yl-glyoxylamides for binding with peripheral benzodiazepine receptor (TSPO) by using 27 compounds. The calculated partition coefficient values show parabolic relations with the TSPO binding affinity, suggesting that the binding affinity increases with increase in the partition coefficient of the compounds until it reaches the critical value after which the affinity decreases. The critical value of $\log P$ is within range of 6.052-6.410 [97].

The disadvantage of the previous QSAR studies on TSPO ligand was that the number of used data set is small (e.g. 29 or 37 etc.) and thus affect the real prediction power of the resulted models. While in the current study 136 compound is used. Also all the previous QSAR studies performed to study certain group of descriptors (properties) such as: studying the physico-chemical parameters or the partition coefficient effect on the compounds activity. While in the current study all the possible properties were calculated for all the compounds and treated to build a predictive MLR model. .

Also the methods used in the current study are MLR and PC-ANN, while in the previous studies either MLR alone or other methods which are having less powerful and prediction capabilities.

Chapter Four

Conclusions

CONCLUSIONS:

A quantitative-structural activity relationship analysis has been conducted on the activity of a set of 136 ligand for Translocator protein (TSPO), by using MLR and principal component-artificial neural networks (PC-ANN) modeling methods, where the strength and the predictive performance of the proposed models was verified using internal (cross-validation and Y-scrambling).

The results obtained by MLR was a number of models (Models 12- 24) which have a good predictive power (R^2) > 0.6 , the best model was model number 24 which includes 24 descriptors, and resulted with $R= 0.909$, $R^2=.826$, and $R^2_{adj.}= 0.788$.

Cross Validation LOO and LMO were performed on the resulted MLR models, models 19-24 showed a good predictive power because of having high R^2_{CV} and PRESS/SST less than 0.4. Thus, models 19-24 were chosen for ANN analysis.

PCA performed to divide the data into three data sets, then the ANN performed on the chosen models (19-24) from LOO and LMO validation.

The results shows that model 24 has the highest correlation coefficient for the test set (0.85016) indicating its high predictive power. While also there are other good predictive models (As model # 20, 21, 23), which chosen to continue ANN to find the optimal number of hidden nodes for each one of these models

According to the results; model 20 with 10 hidden nodes, model 21 with 7 hidden nodes, model 23 with 5 hidden nodes, and model 24 with 7 hidden nodes were chosen as the best models with the optimal hidden nodes because they have high prediction power (R), minimum PRESS value of the test group, and minimum number of hidden nodes.

ANN resulted model were validated through randomization test, then the conditions proposed by Golbraikh and Tropsha were applied to conclude that the QSAR models have acceptable prediction power or not. However the best ANN model with a good predictive power was model #24.

A new suggested compound with predicted PIC_{50} ranging between 20.7 – 31.9.

References:

1. McDouall, J.J., *Computational quantum chemistry: molecular structure and properties in silico* 2013: Royal Society of Chemistry.
2. Cramer, C.J., *Essentials of computational chemistry: theories and models* 2013: John Wiley & Sons.
3. The Shodor Education Foundation, I. *Overview of Computational Chemistry*. Available from: <https://www.shodor.org/chemviz/overview/ccbasics.html>.
4. *Schrodinger Equation*. Available from: <http://hyperphysics.phy-astr.gsu.edu/hbase/quantum/schr.html#c1>.
5. Clark, D.E., ed. *Evolutionary Algorithms in Molecular Design*. Methods and Principles in Medicinal Chemistry, Vol. 8. 2000, Wiley-VCH.
6. Crum-Brown, A. and T. Fraser, *On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia*. Trans. R. Soc. Edinburgh, 1868. **25**: p. 151-203.
7. Selassie, C. and R.P. Verma, *History of quantitative structure–activity relationships*. Burger's Medicinal Chemistry and Drug Discovery, 2003.
8. Meyer, H., *Welche Eigenschaft der Anaesthetica bedingt ihre narkotische Wirkung*. Arch Exp Pathol Pharmacol (Naunyn-Schmiedebergs), 1899. **42**: p. 109-118.
9. Overton, E., *Studien uber die Narkose Fischer*. Jena, Germany, 1901.
10. J. Ferguson, P.R.S.B., . London Ser. , 1939.
11. A. Albert, ed., Chapman and and L. Hall, *Selective Toxicity: The Physicochemical Bases of Therapy*, 1985.
12. Bell, P.H. and R.O. Roblin Jr, *Studies in Chemotherapy. VII. A Theory of the Relation of Structure to Activity of Sulfanilamide Type Compounds*1. Journal of the American Chemical Society, 1942. **64**(12): p. 2905-2917.
13. Hammett, L.P., *Some Relations between Reaction Rates and Equilibrium Constants*. Chemical Reviews, 1935. **17**(1): p. 125-136.
14. SELASSIE, C., *History of Quantitative Structure-Activity Relationships*.
15. Sood, A., *Computational Chemistry Book and Applications* 2010: Drug design by computers.
16. Deeb, O., S. Jawabreh, and M. Goodarzi, *Exploring QSARs of vascular endothelial growth factor receptor-2 (VEGFR-2) tyrosine kinase inhibitors by MLR, PLS and PC-ANN*. Current pharmaceutical design, 2013. **19**(12): p. 2237-2244.

17. Deeb, O., B. Shaik, and V.K. Agrawal, *Exploring QSARs of the interaction of flavonoids with GABA (A) receptor using MLR, ANN and SVM techniques*. Journal of enzyme inhibition and medicinal chemistry, 2014. **29**(5): p. 670-676.
18. Ramírez-Galicia, G., et al., *Exploring QSAR of antiamebic agents of isolated natural products by MLR, ANN, and RTO*. Medicinal Chemistry Research, 2012. **21**(9): p. 2501-2516.
19. Thiochem. *Quantitative Structure-Activity Relationship (QSAR)* Available from: <https://www.theochem.kth.se/courses/molmod/lectures/lecture14.pdf>.
20. *Advantages/Disadvantages of Qsar*. Available from: <http://hydra.vcp.monash.edu.au/modules/mod4/qsarwebp2ad.html>.
21. Deeb, O. and M. Jawabreh *Exploring QSARs for Inhibitory Activity of Cyclic Urea and Nonpeptide-Cyclic Cyanoguanidine Derivatives HIV-1 Protease Inhibitors by Artificial Neural Network*. 2012. DOI: 10.4236.
22. Feynman, R.P. and A.R. Hibbs, *Quantum mechanics and path integrals*. Vol. 2. 1965: McGraw-Hill New York.
23. Goudarzi, N., M. Chamjangali , and P. Kalhor *Linear and Nonlinear QSAR Study of N2 and O6 Substituted Guanine Derivatives as Cyclin-Dependent Kinase 2 Inhibitors*. ISRN Analytical Chemistry, 2013. **2013**.
24. Weitzel, L. and R. Martins *Merging principal component analysis and artificial neural network as a tool for predicting meningitis*. Available from: http://www.ufpa.br/campusmaraba/index/cache/publicacoes/leila_facom_resumo_3.pdf.
25. Buciński, A., et al., *Clinical data analysis using artificial neural networks (ANN) and principal component analysis (PCA) of patients with breast cancer after mastectomy*. Rep Pract Oncol Radiother, 2007.
26. Gershenson, C., *Artificial neural networks for beginners*. arXiv preprint cs/0308031, 2003.
27. Srattaphut, L., et al. *Principal Component Analysis Coupled with Artificial Neural Networks for therapeutic Indication Prediction of Thai Herbal Formulae*. 2012. **7**.
28. *Artificial Neural Network*. Available from: http://www.saedsayad.com/artificial_neural_network.htm.
29. Tu, J.V., *Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes*. Journal of clinical epidemiology, 1996. **49**(11): p. 1225-1231.
30. Veerasamy, R., et al., *Validation of QSAR models-strategies and importance*. International Journal of Drug Design & Discovery, 2011. **3**: p. 511-519.
31. Leach, A.R., *Molecular modelling: principles and applications*2001: Pearson education.
32. Shao, J., *Linear model selection by cross-validation*. Journal of the American statistical Association, 1993. **88**(422): p. 486-494.

33. Wold, S., M. Sjöström, and L. Eriksson, *Partial least squares projections to latent structures (PLS) in chemistry*. Encyclopedia of computational chemistry, 1998.
34. Yasri, A. and D. Hartsough, *Toward an optimal procedure for variable selection and QSAR model building*. Journal of chemical information and computer sciences, 2001. **41**(5): p. 1218-1227.
35. Yee, L.C. and Y.C. Wei, *Current modeling methods used in QSAR/QSPR*. Statistical Modeling of Molecular Descriptors in QSAR/QSPR, 2012. **1**.
36. Todeschini, R., M. Lasagni, and E. Marengo, *New molecular descriptors for 2D and 3D structures. Theory*. Journal of chemometrics, 1994. **8**(4): p. 263-272.
37. Mauri, A., et al., *Dragon software: An easy approach to molecular descriptor calculations*. Match, 2006. **56**(2): p. 237-248.
38. Nie, N.H., D.H. Bent, and C.H. Hull, *SPSS: Statistical package for the social sciences*. Vol. 227. 1975: McGraw-Hill New York.
39. Schneider, W., *Micro Experimental Laboratory: An integrated system for IBM PC compatibles*. Behavior Research Methods, Instruments, & Computers, 1988. **20**(2): p. 206-217.
40. Martinez, W.L. and A.R. Martinez, *Computational statistics handbook with MATLAB2007*: CRC press.
41. Braestrup, C., R. Albrechtsen, and R. Squires, *High densities of benzodiazepine receptors in human cortical areas*. 1977.
42. Braestrup, C. and R.F. Squires, *Specific benzodiazepine receptors in rat brain characterized by high-affinity (3H) diazepam binding*. Proceedings of the National Academy of Sciences, 1977. **74**(9): p. 3805-3809.
43. Regan, J.W., et al., *High affinity renal [3 H] flunitrazepam binding: Characterization, localization, and alteration in hypertension*. Life sciences, 1981. **28**(9): p. 991-998.
44. Davies, L.P. and V. Huston, *Peripheral benzodiazepine binding sites in heart and their interaction with dipyridamole*. European journal of pharmacology, 1981. **73**(2): p. 209-211.
45. Takahashi, H., A. Nagashima, and C. Koshino, *Effect of gamma-aminobutyryl-choline upon the electrical activity of the cerebral cortex*. Nature, 1958. **182**: p. 1443-1444.
46. MacDonald, R. and J.L. BARKER, *Benzodiazepines specifically modulate GABA-mediated postsynaptic inhibition in cultured mammalian neurones*. 1978.
47. Zhang, M.R., et al., *[18F]FMDAA1106 and [18F]FEDAA1106: two positron-emitter labeled ligands for peripheral benzodiazepine receptor (PBR)*. Bioorg Med Chem Lett, 2003. **13**(2): p. 201-4.

48. Papadopoulos, V., et al., *Translocator protein (18kDa): new nomenclature for the peripheral-type benzodiazepine receptor based on its structure and molecular function*. Trends Pharmacol Sci, 2006. **27**(8): p. 402-9.
49. Anzini, M., et al., *Mapping and fitting the peripheral benzodiazepine receptor binding site by carboxamide derivatives. Comparison of different approaches to quantitative ligand-receptor interaction modeling*. J Med Chem, 2001. **44**(8): p. 1134-50.
50. RIOND, J., et al., *Molecular cloning and chromosomal localization of a human peripheral-type benzodiazepine receptor*. European journal of biochemistry, 1991. **195**(2): p. 305-311.
51. Bućan, M., et al., *Comparative mapping of 9 human chromosome 22q loci in the laboratory mouse*. Human molecular genetics, 1993. **2**(8): p. 1245-1252.
52. Anholt, R., et al., *Peripheral-type benzodiazepine receptors: autoradiographic localization in whole-body sections of neonatal rats*. Journal of Pharmacology and Experimental Therapeutics, 1985. **233**(2): p. 517-526.
53. Wang, H.-J., J. Fan, and V. Papadopoulos, *Translocator protein (Tspo) gene promoter-driven green fluorescent protein synthesis in transgenic mice: an in vivo model to study Tspo transcription*. Cell and tissue research, 2012. **350**(2): p. 261-275.
54. Tu, L.N., et al., *Peripheral benzodiazepine receptor/translocator protein global knock-out mice are viable with no effects on steroid hormone biosynthesis*. Journal of Biological Chemistry, 2014. **289**(40): p. 27444-27454.
55. Selvaraj, V., D.M. Stocco, and L.N. Tu, *Minireview: Translocator Protein (TSPO) and Steroidogenesis: A Reappraisal*. Molecular Endocrinology, 2015.
56. Scarf, A.M., L.M. Ittner, and M. Kassiou, *The translocator protein (18 kDa): central nervous system disease and drug design*. J Med Chem, 2009. **52**(3): p. 581-592.
57. Sakai, M., et al., *Translocator protein (18kDa) mediates the pro-growth effects of diazepam on Ehrlich tumor cells in vivo*. European journal of pharmacology, 2010. **626**(2): p. 131-138.
58. Le Fur, G., et al., *Peripheral benzodiazepine binding sites: Effect of PK 11195, 1-(2-chlorophenyl)-n-methyl-n-(1-methylpropyl)-3-isoquinolinecarboxamide: I. In vitro studies*. Life sciences, 1983. **32**(16): p. 1839-1847.
59. Benavides, J., et al., *"Peripheral type" benzodiazepine binding sites in rat adrenals: binding studies with [3H] PK 11195 and autoradiographic localization*. Archives internationales de pharmacodynamie et de therapie, 1983. **266**(1): p. 38-49.
60. Le Fur, G., et al., *Differentiation between two ligands for peripheral benzodiazepine binding sites, [3 H] R05-4864 and [3 H] PK 11195, by thermodynamic studies*. Life sciences, 1983. **33**(5): p. 449-457.
61. Benavides, J., et al., *Opposite effects of an agonist, R05-4864, and an antagonist, PK 11195, of the peripheral type benzodiazepine binding sites on audiogenic seizures in DBA/2J mice*. Life sciences, 1984. **34**(26): p. 2613-2620.

62. Gavish, M., et al., *Enigma of the peripheral benzodiazepine receptor*. Pharmacological reviews, 1999. **51**(4): p. 629-650.
63. Chen, M.-K. and T.R. Guilarte, *Translocator protein 18 kDa (TSPO): molecular sensor of brain injury and repair*. Pharmacology & therapeutics, 2008. **118**(1): p. 1-17.
64. Banati, R., et al., *The peripheral benzodiazepine binding site in the brain in multiple sclerosis*. Brain, 2000. **123**(11): p. 2321-2337.
65. Chauveau, F., et al., *Nuclear imaging of neuroinflammation: a comprehensive review of [11C] PK11195 challengers*. European journal of nuclear medicine and molecular imaging, 2008. **35**(12): p. 2304-2319.
66. Veenman, L., V. Papadopoulos, and M. Gavish, *Channel-like functions of the 18-kDa translocator protein (TSPO): regulation of apoptosis and steroidogenesis as part of the host-defense response*. Current pharmaceutical design, 2007. **13**(23): p. 2385-2405.
67. Papadopoulos, V. and L. Lecanu, *Translocator protein (18 kDa) TSPO: an emerging therapeutic target in neurotrauma*. Experimental neurology, 2009. **219**(1): p. 53-57.
68. Batarseh, A. and V. Papadopoulos, *Regulation of translocator protein 18kDa (TSPO) expression in health and disease states*. Molecular and cellular endocrinology, 2010. **327**(1): p. 1-12.
69. Colasanti, A., et al., *In vivo assessment of brain white matter inflammation in multiple sclerosis with 18F-PBR111 PET*. Journal of Nuclear Medicine, 2014. **55**(7): p. 1112-1118.
70. Rissanen, E., et al., *In vivo detection of diffuse inflammation in secondary progressive multiple sclerosis using PET imaging and the radioligand 11C-PK11195*. Journal of Nuclear Medicine, 2014. **55**(6): p. 939-944.
71. Zürcher, N.R., et al., *Increased in vivo glial activation in patients with amyotrophic lateral sclerosis: Assessed with [11 C]-PBR28*. NeuroImage: Clinical, 2015. **7**: p. 409-414.
72. Liu, J., M.B. Rone, and V. Papadopoulos, *Protein-protein interactions mediate mitochondrial cholesterol transport and steroid biosynthesis*. Journal of Biological Chemistry, 2006. **281**(50): p. 38879-38893.
73. Maeda, J., et al., *Phase-dependent roles of reactive microglia and astrocytes in nervous system injury as delineated by imaging of peripheral benzodiazepine receptor*. Brain research, 2007. **1157**: p. 100-111.
74. Karlstetter, M., et al., *Translocator protein (18 kDa)(TSPO) is expressed in reactive retinal microglia and modulates microglial inflammation and phagocytosis*. Journal of neuroinflammation, 2014. **11**(1): p. 3.
75. Liu, G.J., et al., *The 18 kDa translocator protein, microglia and neuroinflammation*. Brain Pathology, 2014. **24**(6): p. 631-653.
76. Venneti, S., et al., *The high affinity peripheral benzodiazepine receptor ligand DAA1106 binds specifically to microglia in a rat model of traumatic brain injury: implications for PET imaging*. Experimental neurology, 2007. **207**(1): p. 118-127.

77. Johnson, M.R., et al., *Abnormal peripheral benzodiazepine receptor density associated with generalized social phobia*. Biological psychiatry, 1998. **43**(4): p. 306-309.
78. Rocca, P., et al., *Peripheral benzodiazepine receptor messenger RNA is decreased in lymphocytes of generalized anxiety disorder patients*. Biological psychiatry, 1998. **43**(10): p. 767-773.
79. Nudmamud, S., et al., *Stress, anxiety and peripheral benzodiazepine receptor mRNA levels in human lymphocytes*. Life sciences, 2000. **67**(18): p. 2221-2231.
80. Ritsner, M., et al., *Decreased platelet peripheral-type benzodiazepine receptors in persistently violent schizophrenia patients*. Journal of psychiatric research, 2003. **37**(6): p. 549-556.
81. Gavish, M., et al., *Altered platelet peripheral-type benzodiazepine receptor in posttraumatic stress disorder*. Neuropsychopharmacology, 1996. **14**(3): p. 181-186.
82. Soreni, N., et al., *Decreased platelet peripheral-type benzodiazepine receptors in adolescent inpatients with repeated suicide attempts*. Biological psychiatry, 1999. **46**(4): p. 484-488.
83. Setiawan, E., et al., *Role of translocator protein density, a marker of neuroinflammation, in the brain during major depressive episodes*. JAMA Psychiatry, 2015. **72**(3): p. 268-275.
84. Taliani, S., et al., *Novel irreversible fluorescent probes targeting the 18 kDa translocator protein: synthesis and biological characterization*. J Med Chem, 2010. **53**(10): p. 4085-93.
85. Arbo, B., et al., *Therapeutic actions of translocator protein (18kDa) ligands in experimental models of psychiatric disorders and neurodegenerative diseases*. The Journal of steroid biochemistry and molecular biology, 2015. **154**: p. 68-74.
86. Cappelli, A., et al., *Structure-activity relationships in carboxamide derivatives based on the targeted delivery of radionuclides and boron atoms by means of peripheral benzodiazepine receptor ligands*. J Med Chem, 2003. **46**(17): p. 3568-71.
87. Trapani, G., et al., *Structure-activity relationships and effects on neuroactive steroid synthesis in a series of 2-phenylimidazo[1,2-a]pyridineacetamide peripheral benzodiazepine receptors ligands*. J Med Chem, 2005. **48**(1): p. 292-305.
88. Cappelli, A., et al., *Synthesis and structure-activity relationship studies in translocator protein ligands based on a pyrazolo[3,4-b]quinoline scaffold*. J Med Chem, 2011. **54**(20): p. 7165-75.
89. Cappelli, A., et al., *Synthesis and structure-activity relationship studies in peripheral benzodiazepine receptor ligands related to alpidem*. Bioorg Med Chem, 2008. **16**(6): p. 3428-37.
90. Todeschini, R. and V. Consonni, *Handbook of molecular descriptors*. 2000.
91. Parr, R.G. and R.G. Pearson, *Absolute hardness: companion parameter to absolute electronegativity*. Journal of the American Chemical Society, 1983. **105**(26): p. 7512-7516.

92. Golbraikh, A. and A. Tropsha, *Beware of q^2 !* Journal of Molecular Graphics and Modelling, 2002. **20**(4): p. 269-276.
93. Roy, K., A. Toropov, and I. Raska, *QSAR Modeling of Peripheral Versus Central Benzodiazepine Receptor Binding Affinity of 2-Phenylimidazo [1, 2-a] pyridineacetamides using Optimal Descriptors Calculated with SMILES.* QSAR & Combinatorial Science, 2007. **26**(4): p. 460-468.
94. Cinone, N., H.-D. Höltje, and A. Carotti, *Development of a unique 3D interaction model of endogenous and synthetic peripheral benzodiazepine receptor ligands.* Journal of Computer-Aided Molecular Design, 2000. **14**(8): p. 753-768.
95. Roy, K., A.U. De, and C. Sengupta, *QSAR of peripheral benzodiazepine receptor ligand 2-phenylimidazo-[1,2-a]pyridine derivatives with physico-chemical parameters.* Indian J Biochem Biophys, 2003. **40**(3): p. 203-8.
96. Dalai, M.K., J.T. Leonard, and K. Roy *Exploring QSAR of peripheral benzodiazepine receptor binding affinity of 2-phenylpyrazolo[1,5-a]pyrimidin-3-yl-acetamides using topological and physicochemical descriptors.* 2006. **45B**.
97. Roy, K. and M.K. Dalai, *Exploring QSAR of peripheral benzodiazepine receptor binding affinity of N,N-dialkyl-2-phenylindol-3-yl-glyoxylamides using physico-chemical descriptors.* Indian J Biochem Biophys, 2007. **44**(2): p. 114-21.

دراسة العلاقة الكمية بين الفاعلية والصيغة البنائية باستخدام طريقتي (MLR و PC-ANN) لبعض المركبات التي لها فعالية على بروتين (TSPO) Translocator

إعداد: هناء سليم بني عودة

إشراف: أ.د. عمر ديب

المُلخَص:

يتناول موضوع هذا البحث دراسة العلاقة الكمية بين فعالية 136 مركب وصيغها البنائية على بروتين يسمى Translocator. وقد وضعت نماذج QSAR باستخدام الانحدار الخطي المتعدد (MLR) كطريقة خطية . بينما تم استخدام الشبكات العصبية الاصطناعية (PC- ANN) كطريقة غير خطية . النتائج التي تم الحصول عليها هي نماذج ذات قدرة تنبؤ جيدة . النماذج التي نتجت عن MLR و التي حصلت على معامل ارتباط اعلى من 0.6 هم النماذج من 12-24 وكان الافضل بينها نموذج رقم 24 مع معامل ارتباط يساوي 0.909، وتم التحقق من قدرة النماذج على التنبؤ عن طريق استخدام LMO و LOO و اظهرت النماذج 19-24 افضل نتائج، ثم تم توزيع المركبات الى ثلاث مجموعات عن طريق PCA. وتم استخدام النماذج 19-24 في ANN

ومن ثم تم التحقق من قوة وأداء كل النماذج المقترحة من ANN باستخدام (randomization test)

وتطبيق الظروف التي اقترحتها Golbraikh و Tropsha ، وقد وجد ان النموذج رقم 24 هو

الافضل مع معامل ارتباط 0.832.