ISSN 2321-807X

Exploring Quantitative Structure-Activity Relationships (QSARs) of Non-Tri cyclic Cyclooxygenase-2 (COX-2) Inhibitors by MLR and PC-ANN

O. Deeb[1,*] and N. Zatari[1]

[1]Faculty of Pharmacy, Al-Quds University,

PO Box 20002, Jerusalem, Palestine

deeb.omar@gmail.com

ABSTRACT

Quantitative structure–activity relationship study using principal component artificial neural network (PC-ANN) methodology was conducted to predict the inhibitory activities expressed as pIC_{50} of 73 non-tri cyclic cyclooxygenase-2 (COX-2) inhibitors. The results obtained by MLR shows that the best two models are close to each other with regression coefficient of 0.85. These optimal models were further analyzed by PC-ANN and the best model obtained was with regression coefficient of 0.823 for the test set. The lowest prediction sum of squares (PRESS) value obtained for the prediction set is 4.727 which accounts for predictability of the model. Artificial neural networks provide improved models for heterogeneous data sets without splitting them into families. Both the external and cross-validation methods are used to validate the performances of the resulting models. Randomization test is employed to check the suitability of the models.

Indexing terms/Keywords

QSAR; MLR; PC- ANN; Inhibitory activity; Non-tri cyclic cyclooxygenase-2 (COX-2) inhibitors.

Academic Discipline And Sub-Disciplines

QSAR and Drug Design .

SUBJECT CLASSIFICATION

Paper

TYPE (METHOD/APPROACH)

QSAR of Non-tricyclic COX-2 Inhibitors by MLR and PC-ANN techniques

Council for Innovative Research

Peer Review Research Publishing System

Journal: Journal of Advances in Chemistry

Vol. 11, No. 1

editorjaconline@gmail.com

www.cirjac.com

1. INTRODUCTION

Cyclooxygenase-2 (COX-2) inhibition has been one of the most investigated areas of research in the most recent decade owing to its essential role in relieving pain and other inflammatory conditions. Non-steroidal anti-inflammatory drugs (NSAIDs) are deeply used in the treatment of wide variety of inflammatory conditions. NSAIDs are anti-pyretic, analgesic activities and are prescribed as first choice in the treatment of arthritis, rheumatisms and other degenerative of inflammatory joint disease as well as reliving the pains of everyday life. From a historical point of view, the first NSAID with therapeutic benefits was aspirin, which has now been used for more than 100 years as a NSAID. However, these drugs are associated with high risk of gastrointestinal and renal adverse effects. NSAIDs act by inhibition of cyclooxygenase (COX), the enzyme involved in the biosynthesis of prostaglandins and prostacyclins from arachidonic acid. Prostaglandins are involved in physiological functions such as protection of the stomach mucosa, aggregation of platelets and regulation of kidney function. They also have pathological functions such as their involvement in inflammation, fever and pain.

Cyclooxygenase exists in at least two isoforms, namely, the constitutive cyclooxygenase-1 (COX-1) and the inducible cyclooxygenase-2 (COX-2). Inhibition of COX-1 is accountable for the adverse gastrointestinal and renal effects of NSAIDs while the inhibition of COX-2 accounts for NSAIDs therapeutic effects. All classical NSAIDs, such as aspirin and ibuprofen can inhibit both COX-1 and COX-2, but bind more strongly to COX-1. Selective COX-2 inhibitors have the same anti-inflammatory, anti-pyretic, and analgesic activities as do nonselective NSAIDs but without causing gastric ulceration, bleeding and perforation.

Increasing selectivity for COX-2 also increased toxicity, since the anti-thrombotic prostacyclin is formed by COX-2 and inhibiting its synthesis precipitated heart attacks. The problem of this side action has not yet been resolved. Is it possible to obtain the benefit of low gastrotoxicity and avoid the danger of a heart attack? Perhaps lumiracoxib has provided the solution. However, it now appears that by taking any NSAID, patients risk experiencing a heart attack [1-6]. The review [7] provides the various structural classes of selective COX-2 inhibitors with special emphasis on their structure activity relationships. In this study, we concentrate on Non-Tricyclic compounds which lack the cyclic central core and have monocyclic or bicyclic structures.

The importance of developing selective COX-2 inhibitors is manifested by the deep efforts dedicated in this field that resulted in the synthesis of hundreds of compounds, which displayed activity against COX-2. In this study, we are concerned in designing a new set of COX-2-selective inhibitors based on simple but statistically sound Quantitative Structure Activity Relationship (QSAR) models whose parameters can be easily obtained by means of commonly available and less costly computational programs.

Quantitative structure activity relationship (QSAR) is the quantitative correlation of structural properties of a compound with its chemical, physical, pharmaceutical, or biological effect. Based on this assumption, many trials were made to correlate various physicochemical properties of a set of molecules with their experimentally known biological activity, and so QSAR goals are: (1) Prediction of the activity of untested molecules, depending on models developed using a series of molecules and (2) Constructing ideas about mechanism of action of a group of compounds leading to a design of new compounds of better activity and less toxicity. QSAR model development process is typically divided into three steps: data preparation, data analysis and model validation.

Data preparation starts by selection of the data set to be used; this may simply be the extraction of data from a database or may need additional experimental studies. There are two steps to complete data preparation: geometry optimization and descriptors calculation. Geometry optimization or minimization is finding the coordinates that represents the potential energy minimum for the molecular structure in its 3D form. Theoretical molecular descriptor is a value that describes the molecular structure numerically. These descriptors can be simple such as molecular weight or complex such as geometrical descriptors.

In data analysis, the first step is to decide which techniques for statistical analysis and correlation to be used. If our correlation models to be built are linear then we use multilinear regression (MLR) or non linear then we use artificial neural network (ANN).

Model validation is the final part of the model development process, the predictive power of the model is tested on an independent set of compounds, generally predictive power is the most important characteristics of the model and model predictivity is the ability of the model to predict accurately the target activity of a compound that was not used for model development.

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

Multilinear regression (MLR) is multivariate statistical technique to examine the linear relationship between the single dependent variable (activity) and two or more independent variables (molecular descriptors). Collinearity, which often exists between independent variables, generates a severe problem in certain types of mathematical handling such as matrix inversion [8]. As it was recently reviewed by Schneider and Wrede [9], the flexibility of ANN for finding out relationships that are more complex allows this method to be widely applied in QSAR studies. Both linear and nonlinear mapping functions can be modeled by configuring the network properly. To obtain powerful and accurate ANN models, one should train a subset of descriptors instead of all generated descriptors [10–15].

Billones et al. [16] performed QSAR study of COX-2 inhibitors belonging to nine chemical classes using semi-empirical (AM1) computed quantum mechanical descriptors and electrotopological sate (E-state) indices. Another study was performed by Gupta et al, [17] related to 3D-QSAR of some tetrasubstituted pyrazoles as COX-II inhibitors, a six point pharmacophore with 3 hydrogen bond acceptors, one hydrophobic group and two aromatic rings as pharmacophoric feature was developed.

This study aims to predict the inhibitory activity pIC_{50} of the data set in reference [18-25] as one group without splitting them into categorizes. This is achieved by applying ANN to develop new statistically validated QSAR models utilizing different types of descriptors. The strength and the predictive performance of the proposed models were verified using cross validation, chance correlation and external test set. Therefore, the motivation of this work is to provide QSAR models that will be used to predict inhibitory activity of unknown compounds and also these models may be used to design new drugs.

2. MATERIALS AND METHODS

2.1 Software

Geometry optimizations were performed using HyperChem (Version 7.5; Hypercube, Inc, USA, http://www.hyper.com) at the AM1 level of theory. An AM1 optimization was chosen because it was developed and parameterized for common organic structures. Descriptors were calculated using HyperChem and DRAGON (Milano Chemometrics and QSAR Group, USA, evaluation version 5.0, http://www.disat.unimib.it/vhml) software. SPSS software (version 13.0, SPSS, Inc.) was used for the simple MLR analysis while ANN analysis was performed using MATLAB (Version 7.0.1 (R14), http://www.mathworks.com).

2.2 Chemical data and descriptors

A data set of 73non-tricyclic COX-2 inhibitors and their inhibitory activity (pIC_{50}) were obtained from reference [18-25] and used in this study. These inhibitors and their inhibitory activities are included in Table S1 in the supporting information.

The structures of the compounds are drawn by hyperchem software. The resultant structures are 2D then we convert them to 3D. HyperChem software was used to optimize the different compound structures using AM1 semi-empirical level. The optimization was preceded by the Polak-Rebiere algorithm. To be sure that we reached global minima, geometry optimization was run multiple times with different starting points for each molecule.

In this study, a pool of 1481 descriptors classified into 18 different groups was calculated using Dragon software. The constant or nearly constant descriptors for all the 73 compounds were discarded from further analysis. Furthermore, chemical descriptors such as HOMO, LUMO and polarizability were calculated using HyperChem software. Depending on the HOMO and LUMO values, electrophylicity, electronegativity, hardness, and softness descriptors were calculated. Other descriptors such as surface area approximate, surface area grid, volume, mass, polarizability, hydration energy, octanol-water partition coefficient (logP), and refractivity were calculated. Discarding highly inter-correlated ($r>0.95$) descriptors and following the procedure described in the next section, this number of descriptors was declined to 14 descriptors in the "final" MLR regression model (model 14 in Table 1).

2.3 Multiple linear regression (MLR) analysis

Multiple linear regression analysis with stepwise selection and elimination of variables was employed to model the inhibitory activity (pIC_{50}) relationships with each group of descriptors separately. $Log1/IC_{50}$ is the dependent variable and the set of descriptors as independent variables. Then, the “optimal” descriptors for each group were selected and gathered in one group to perform new MLR analysis.

2.4 Principal components analysis (PCA)

Collinear descriptors add redundancy to the input data matrix and consequently the performances of the models obtained by using these descriptors would be degraded. PCA and more specifically factor analysis, groups together variables that are collinear to form a composite indicator capable of capturing as much of common information of those indicators as possible. Each factor reveals the set of variables with the highest relationship. The idea under this approach is to explain the highest possible variation in the indicators set using the smallest possible number of factors. Consequently, the index no longer depends upon the dimensionality of the data set but it is rather based on the 'statistical' dimensions of the data. Application of PCA on a descriptor data matrix results in a loading matrix containing factors or PCs, which are orthogonal and therefore have no correlation with each other.

The PC’s were calculated by singular value decomposition (SVD) method in MATLAB environment (MathWork Inc. Version 7.0.1 (R14)). Due to the quality of data, a previous treatment of the data is essential before applying the multivariate analysis methods. Scaling and centering is one of the pre-processing methods needed before performing the regression methods joint with feature extraction. Projection methods results depend on the normalization of the data. Descriptors with small absolute values have a small contribution to overall variances leading to biased PC’s caused by the presence of other descriptors with higher values. In order to have the focus on the important variables in the model, equal weights are assigned to each descriptor, with appropriate scaling. Furthermore, descriptors were standardized to unit variance and zero mean (autoscaling) to give all variables the same importance. Then, the data matrix containing the entire set of descriptors and activity were simultaneously subjected to PCA.

2.5 Principal component-artificial neural network (PC-ANN) analysis

ANNs are computer-based models in which a number of nodes, also called neurons are interconnected by links forming netlike structure ‘‘layers.’’ A variable value is assigned to every neuron.

There are three kinds of neurons: (a) the input neurons which receive their values from independent variables and constitute the input layer, (b) the hidden neurons which collect values from other neurons, giving a result that is passed to a successor neuron, (c) the output neurons which take values from other units and correspond to different dependent variables, forming the output layer. In this sense, network architecture is commonly represented as I–H–O, where I, H, and O are the number of neurons in the input, hidden, and output layers, respectively.

The weights are links between units that condition the values assigned to the neurons. The weights are adjusted through a training process in order to minimize network error. For this, a non-linear transfer function relates the input parameters with the outputs. Commonly neural networks are adjusted, or trained, so that a particular input leads to a specific target output.

In PC-ANN analysis, as a preliminary treatment, the input data (i.e., molecular descriptors) were normalized to have zero mean and unity variance, and then were subjected to PCA before being introduced into the neural network. It should be illustrated that for each MLR resulted model, separate ANN models were developed so that the input's descriptors were the subsets selected by the stepwise MLR methods. In the case of each MLR model, a feed-forward neural network with back-propagation of error algorithm was constructed to model the activity–structure relationships between the descriptors on one hand and inhibitory activity on the other hand. The model development in ANN and the network architecture is fully described by us [13] and others [14]. The data set was divided into three subsets: training, validation and external test sets. The training and the validation sets are the norm in all model training processes. The test set is used to test the trend of the prediction precision of the model trained at some point of the training evolution. The extracted PC’s for each MLR model were classified homogenously, based on the factors space of the descriptors, into training set (60%), validation set (20%) and external test set (20%) according to the PCA and the first two PC’s were plotted against each other (see Figure 1). Afterward, the training set was used to optimize the network performance. The regression between the network output and the observed activity was calculated for each set individually. The training function 'trainscg' was used to train the network. To find models with lower errors, the ANN algorithm was run many times, with different geometry and initial weights each time.

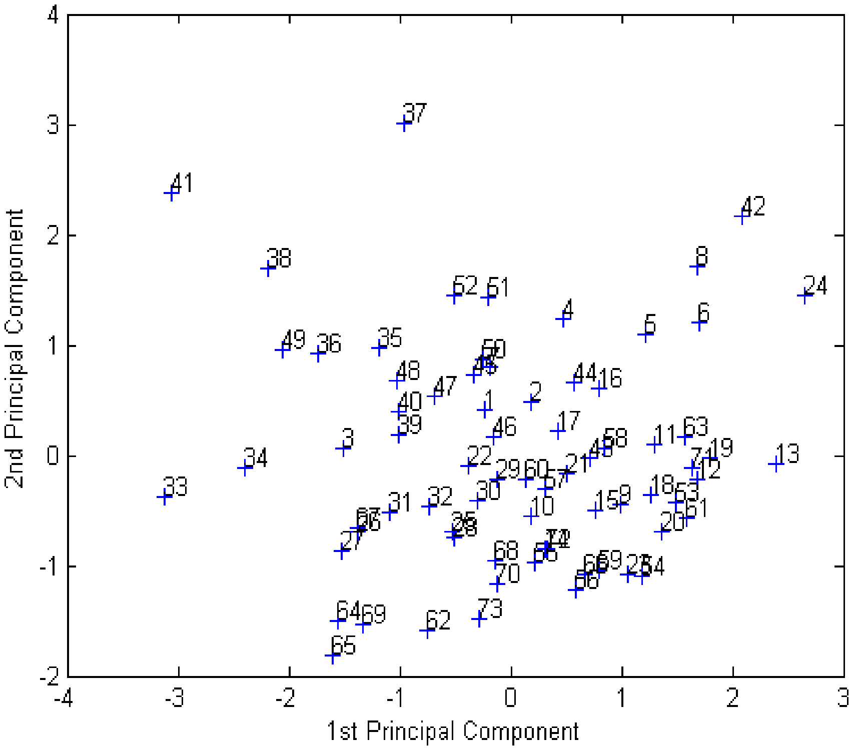


Figure 1. First and second principal components for the factor spaces of the descriptors and non-tricyclic COX-2 inibitory activity data.

3. RESULTS AND DISCUSSION

3.1 MLR analysis

In continuation to recent QSAR studies [26-29] done using similar methods, we developed an ANN-QSAR model that describes the inhibitory activity of a series of compounds using large number of different descriptors. MLR were performed on each one of the groups of descriptors individually (individual approach described in Ref. [30] by Deeb) where pIC_{50} is the dependent variable. Stepwise method is used to develop multilinear equation by correlating dependent variable (activity) and the best independent variables.

Next, a new or “final” MLR analysis was performed by correlating the dependent variable (activity) and the optimal descriptors selected from the individual MLR models. Table 1 shows the regression models suggested from the “final” MLR analysis. The number of descriptors in these models is varied between 1 and 14. Model 14 and 15 are close to each other with one descriptor less. The highest coefficient of determination (R^2) obtained, is 0.720 for a regression model with 14 descriptors (model **14**). Table 2 shows a key for the different descriptors used in the final MLR model.

Table 1. Final MLR model summary.

Model No.	R	R^2	Adjusted R^2	SE	Descriptors
1	0.331	0.109	0.097	0.776	MATS3e
2	0.429	0.184	0.161	0.748	MATS3e, E1v
3	0.533	0.284	0.253	0.705	MATS3e, E1v, E3s
4	0.631	0.398	0.363	0.652	MATS3e, E1v, E3s, Me
5	0.676	0.457	0.417	0.624	MATS3e, E1v, E3s, Me, C-028
6	0.707	0.499	0.454	0.604	MATS3e, E1v, E3s, Me, C-028, G3p
7	0.726	0.527	0.476	0.591	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3
8	0.755	0.570	0.516	0.568	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u
9	0.770	0.593	0.535	0.557	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u
10	0.785	0.616	0.554	0.545	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, MATS8e
11	0.799	0.639	0.573	0.534	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, MATS8e, dipole moment (Debyes)
12	0.822	0.675	0.611	0.51	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, MATS8e, dipole moment (Debyes), G3m
13	0.837	0.700	0.634	0.494	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, MATS8e, dipole moment (Debyes), G3m, H7m
14	0.849	0.720	0.652	0.482	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, MATS8e, dipole moment (Debyes), G3m, H7m, R6m
15	0.848	0.719	0.657	0.478	MATS3e, E1v, E3s, Me, C-028, G3p, BELm3, Mor03u, G2u, dipole moment (Debyes), G3m, H7m, R6m

The following equation represents the best MLR model (model 15) with 13 descriptors:

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

According to the above equation, the most important descriptor in this equation is G3m which is related to the 3rd component symmetry directional WHIM index / weighted by mass; it is directly proportional to the activity of the compounds. The second important descriptor is G2u which is related to 2nd component symmetry directional WHIM index / unweighted; it is directly proportional to the activity of the compounds.

Table 2. Key for the different descriptors used in the final MLR model.

Descriptor symbol	Description
MATS3e	Moran autocorrelation of lag 3 weighted by Sanderson electronegativity
E1v	1st component accessibility directional WHIM index / weighted by van der waals volume
E3s	3rd component accessibility directional WHIM index / weighted by I-state
Me	Mean atomic Sanderson electronegativity (scaled on Carbon atom)
C-028	R—CR—X (Atom-centred fragment)
G3p	3rd component symmetry directional WHIM index / weighted by polarizability
BELm3	lowest eigenvalue n. 3 of Burden matrix / weighted by atomic masses
Mor03u	Signal 03 / unweighted (3D-MoRSE descriptors)
G2u	2nd component symmetry directional WHIM index / unweighted
MATS8e	Moran autocorrelation of lag 8 weighted by Sanderson electronegativity
dipole moment (Debyes)	To determine bond angles and the degree of polarity of covalent bonds
G3m	3rd component symmetry directional WHIM index / weighted by mass
H7m	H autocorrelation of lag 7 weighted by mass
R6m	R autocorrelation of lag 6/weighted by atomic masses

Then, leave one out (LOO) cross validation was performed on models **10**-**15** since these models have coefficients of determination larger than 0.6 [31]. The results of LOO cross validation are summarized in Table S2 in the supporting information. This table shows that the cross-validation coefficient of determination (R^2_{CV}) has positive values starting from model **10** to model **15**. Table S2 shows also those models **14** and **15**, have the highest R^2 and R^2_{cv} values as well as the lowest root mean square error (RMSE) values. Also, PRESS/SST is less than 0.4 for these models. Thus, models **14** and **15** were chosen for further analysis with ANN.

3.2 PCA

The inputs of the ANN were the subset of the descriptors used in different MLR models (Table 1). First, PCA was performed to classify the molecules into training (60%), validation (20%) and test (20%) sets. Figure 1 shows the first and second PC's for the factor spaces of the descriptors and COX-2 inhibitory activity data. According to the pattern of the distribution of the data in factor spaces (Figure 1), the training, validation and test sets molecules were selected homogenously so that molecules in different zones of Figure 1 belong to the three subsets. As we can also see from figure 1 that compounds number 37 and 41 are outliers, which mean that those two compounds behave in a different way in comparison to the rest of other compounds with respect to activity and descriptors. Therefore, the total number of compounds now is 71 compounds. The molecules subjected to the preliminary treatment mentioned previously, the classified data were used as an input for the ANN.

3.3 ANN

In this study, a three-layered feed-forward ANN model with back propagation learning algorithm [32] was employed. At first, non-linear relationship between the subset of descriptors selected by stepwise selection-based MLR and COX-2 inhibitory activity was preceded by ANN models with similar structure. The number of hidden layer's nodes was set to 7 for all models, and the number of nodes in the input layer was the number of descriptors.

The correlation coefficients and cross-validation parameters of ANN analysis for ANN model numbers **14** and **15** are given in Table S3 in the supporting information. This table shows that the results of the two models are close to each other. However, model **14** seems to be better than model **15** since it has higher correlation coefficient for the test set as well as lower relative standard error of prediction.

To optimize the performance of the ANN models **14** and **15**, these models were trained using different number of hidden nodes up to 20. Choosing the best model was based on cross-validation parameters and determination of minimum prediction error [33]. For the evaluation of the predictive ability of a multivariate calibration model, $RMSE_P$ is an important statistical parameter to find the best number of hidden nodes. Moreover, because large numbers of hidden nodes often draw attention to the risk of overfitting [34], considering models with low prediction error is avoided if a large number of hidden nodes are used in their network training.

The results of optimizing the number of hidden nodes for models **14** and **15** are summarized in table S4 and table S5 in the supporting information respectively.

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

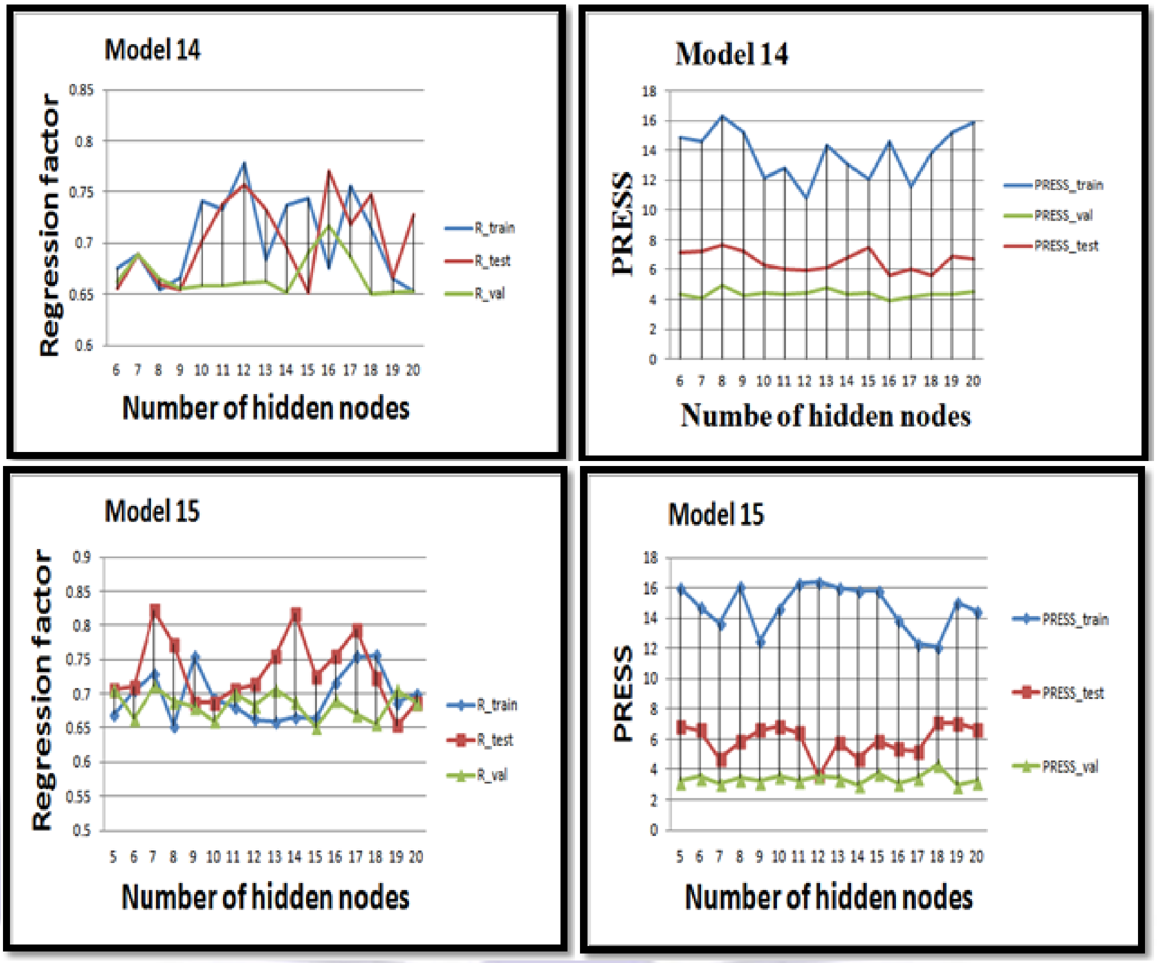


Figure 2. PRESS against number of hidden nodes as well as regression factor against number of hidden nodes for model 14 and 15 respectively.

Randomization test is performed to investigate the probability of chance correlation for the optimal models (models **14** and **15** with 12 and 7 hidden nodes in the network, respectively). Chance correlation was done using the same configuration parameters and the same activation functions of all our ANN models. The results of chance correlation for models **14** (using 12 hidden nodes) and **15** (using 7 hidden nodes) are summarized in Tables S6 and S7 in the supporting information, respectively. These tables show that the coefficients of determination obtained by chance are low in general while the RMSE values are high. This indicates that the models obtained from ANN are better than those obtained by chance. As we can see, our models were validated by calculating different statistical parameters, using external test set and finally performing randomization test.

Figure 3 shows plot of the predicted activity against observed ones for the training and test sets compounds as well as their residuals for models **14** and **15**

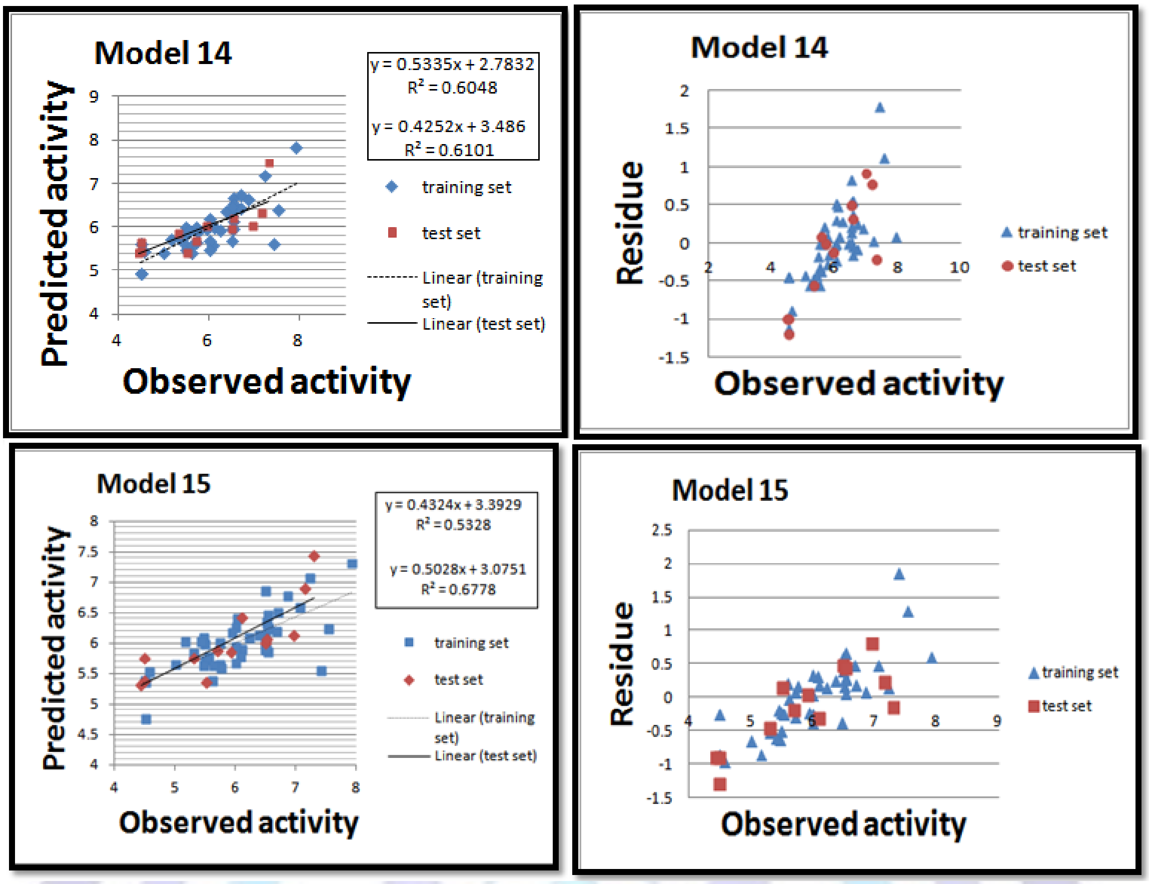


Figure 3. Predicted inhibitory activities against observed ones and their residuals for model 14 and 15 respectively.

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

Calculated pIC_{50} = 0.534 Observed pIC_{50} + 2.783 (1)

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

Calculated pIC_{50} = 0.4252 Observed pIC_{50} + 3.486 (2)

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

Calculated pIC_{50} = 0.432 Observed pIC_{50} + 3.393 (3)

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

Calculated pIC_{50} = 0.503 Observed pIC_{50} + 3.80 (4)

To check the presence of more outliers in a model, for the training and test sets, the standard deviation of the observed activity data was calculated. The residue which is equal to the difference between the predicted and observed one were calculated also. Finally, if the value of the residue is larger than two times the standard deviation of the observed activity, then this point is considered as an outlier. We found that there was no outlier in our data.

4. COMPARISON WITH OTHER QSAR STUDIES

Few QSAR studies [35-40] have been performed on COX-2 inhibitors in the literature. But the data set of these studies was smaller than the data set used in our study. In these studies they used one core while in our study we used different cores and perform QSAR on the whole data without splitting them into cores or families. The previous studies used 3D-QSAR and others used topological indicies while in our study we performed QSAR by applying PC-ANN method and using pool of descriptors. Finally, our results indicate that the proposed models have better predictivity than the models proposed by other studies. Our models may be used to design new COX-2 inhibitors.

Recently [41], we carried out a QSAR study of 48 tricyclic COX-2 inhinitors using MLR and PC-ANN. The results obtained by PC-ANN give advanced regression models with good prediction ability.

5. CONCLUSIONS

MLR as well as ANN modeling method combined with the individual [30] factor selection approach is applied to predict the COX-2 inhibitory activity of a set of 73 non-tricyclic compounds. The results obtained by MLR shows that the best two models are close to each other with regression coefficient of 0.85. These optimal models were further analyzed by PC-ANN and the best model obtained was with regression coefficient of 0.823 for the test set. The lowest prediction sum of squares (PRESS) value obtained for the prediction set is 4.727 which accounts for predictability of the model.. ANN provides improved models for heterogeneous data sets without splitting them into families and gives good regression models with good prediction ability.

Generally, the models obtained from MLR analysis are better than those obtained by ANN analysis which may account for linear relation between the inhibitory activity of these 73 non-tricyclic inhibitors and descriptors. Both the external and cross-validation methods are used to validate the performances of the resulting models. Employed randomization test indicates that the models obtained from ANN are better than those obtained by chance.

ACKNOWLEDGMENTS

The authors thank Al-Quds University for providing computer time and software to perform this work.

References

[1] Fiorucci S, Meli R, Bucci M and Cirino G. Dual inhibitors of cyclooxygenase and 5-lipoxygenase: A new avenue in anti-inflammatory therapy. *Biochem. Pharmacol.* 62 (2001), : 1433-1438.

[2] Griswold DE and Adams JL. Constitutive cyclooxygenase (COX-1) and inducible cyclooxygenase (COX-2): Rationale for selective inhibition and progress to date. *Med. Res. Rev.* 16. (1996): 181-206.

[3] Vane JR, Bakhle YS and Botting RM. Cyclooxygenases 1 and 2. *Annu. Rev. Pharmacol. Toxicol* 38. (1998): 97-120.

[4] Charlier C and Michaux C. Dual inhibition of cyclooxygenase-2 (COX-2) and 5-lipoxygenase (5-LOX) as a new strategy to provide safer non-steroidal anti-inflammatory drugs. *Eur. J. Med. Chem*. 38 (2003): 645-659.

[5] Dannhardt G and Kiefer W. Cyclooxygenase inhibitors-currunt status and future prospects. *Eur. J. Med. Chem*. 36 (2001): 109-126

[6] Konturek PC, Kania J, Burnat G, Hahn EG and Konturek SJ. Prostaglandins as mediators of COX-2 derived carcinogenesis in gastrointestinal tract. *J. Physiol. Pharmacol* 56 (2005): S57-73.

[7] Afshin Zarghi and Sara Arfaei. "Selective COX-2 Inhibitors: A Review of Their Structure-Activity Relationships", Iranian Journal of Pharmaceutical Research 10 (4) (2011): 655-683

[8 Montgomery D.C , Peck E.A 1982, , "Introduction to Linear Regression Analysis", Wiley, New York.

[9] Schneider G. , Wrede P., "Artificial neural networks for computer-based molecular design", Prog. Biophys. Mol. Biol. 70(1998) 175–222.

[10] Gemperline P.J., Long J.R., Gregoriou G. , "Nonlinear multivariate calibration using principal components regression and artificial neural networks", Anal. Chem. 63 (1991) 2313–2323.

[11] Vendrame R. , Braga R.S., Takahata Y. , Galvao D.S. , "Structure–activity relationship studies of carcinogenic activity of polycyclic aromatic hydrocarbons using calculated molecular descriptors with principal component analysis and neural network methods", J. Chem. Inf. Comput. Sci. 39(1999) 1094–1104.

[12] Hemmateenejad B. , Akhond M., Miri R., Shamsipur M., "Genetic algorithm applied to the selection of factors in principle component-artificial neural networks: application to QSAR study of calcium channel antagonist activity of 1, 4-dihydropyridines (nifedipine analogous)", J. Chem. Inf. Comput. Sci. 43(2003) 1328–1334.

[13] Deeb O. , Hemmateenejad B.. "ANN-QSAR model of drug-binding to human serum albumin", Chem. Biol. Drug Des. 70(2007) 19–29.

[14] Hemmateenejad B., Safarpour M. A.., Miri R., Nesari N., " Toward an optimal procedure for PC-ANN model building: prediction of the carcinogenic activity of a large set of drugs", J. Chem. Inf. Model. 45(2005) 190–199.

[15] Ramírez-Galicia G., Garduño-Juárez R. , Deeb O., Hemmateenejad B., "PCR-ANN and RTO approach to L-opioid receptor-binding affinity. pooling data from different sources", Chem. Biol. Drug Des. 71(2008) 260–270.

[16] Billones J. B. and Buenaobra S. M. “Quantitative structure activity relationship (QSAR) study of cyclooxygenase-2 (COX-2) inhibitors”, Philippine Journal of Science, 140(2) (2011)125–132.

[17] Gupta G. K. and Kumar A. , “3D-QSAR Studies Of Some Tetrasubstituted Pyrazoles As COX-II Inhibitors”, Acta Poloniae Pharmaceutica - Drug Research, 69(4) (2012) 763-772.

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

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

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

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

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

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

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

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

[26] Deeb O. and Drabh M., "Exploring QSARs of Some Analgesic Compounds by PC-ANN", Chem Biol Drug Des; 76(2010) 255–262.

[27] Khadikar P.V., Deeb O., Jaber A., Singh J., Agrawal V.K., Singh S. and Lakhwani M.. "Development of Quantitative Structure-Activity Relationship for a set of Carbonic Anhydrase Inhibitors : Use of Quantum and Chemical Descriptors". Letters in Drug Design & Discovery; 3(9) (2006) 622-635.

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

[29] Deeb O., Youssef K.M. and Hemmateenejad B., "QSAR of Novel Hydroxyphenylureas as Antioxidant Agents". QSAR and Combinatorial Sciences; 27(4) (2008) 417-424.

[30] Deeb O., "Correlation ranking and stepwise regression procedures in PC-ANN modeling and application to predict the toxic activity and HSA binding affinity". Chemometrics and Intelegent Laboratory Systems.; 104(2010) 181-194.

[31] Golbraikh A., Tropsha A.. "Beware of q2!". J Mol Graph Model; 20(2002) 269–276.

[32] Rumelhart D. E. , Hinton G. E., Williams R. J. . "Learning representations by back-propagating errors". Nature; 323(1986) 33–536.

[33] Martens H., Naes T. "Multivariate Calibration". Chichester:John Wiley, 1989 [34] E.P.P.A. Derks, L.M.C. Buydens. "Aspects of network training and validation on noisy data: part 1. Training aspects". Chemom Intell Lab Syst 41(1998) 171 -184.

[35] Gautam R. Desiraju , Bulusu Gopalakrishnan, Three-Dimensional Quantitative Structural Activity Relationship (3D-QSAR) Studies of Some 1,5-Diarylpyrazoles: Analogue Based Design of Selective Cyclooxygenase-2 Inhibitors, Molecules, 5 (2000), 945-955.

[36] A sit K. Chakraborti and Thilagavathi R.,Computer-Aided Design of Non Sulphonyl COX-2 Inhibitors: An Improved Comparative Molecular Field Analysis Incorporating Additional Descriptors and Comparative Molecular Similarity Indices Analysis of 1,3-Diarylisoindole Derivatives, Bioorganic and Medicinal Chemistry 11 (2003), 3989-3996.

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

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

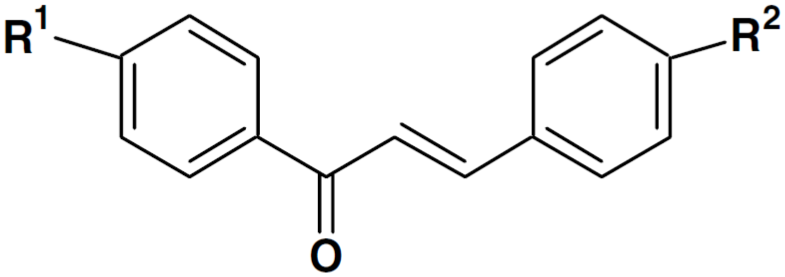
[39] Gupta G. K. and Kumar A., 3D-Qsar Studies Of Some Tetrasubstituted Pyrazoles As COX-II Inhibitors, Acta Poloniae Pharmaceutica - Drug Research, 69(4) (2012), 763-772.

[40] Dwivedi A., Singh A., Srivastava A.K., Quantitative structure–activity relationship based modeling of substituted indole Schiff bases as inhibitor of COX-2, Journal of Saudi Chemical Society, (in Press), (2013)

[41] Deeb O. and Zatari N., Exploring Quantitative Structure-Activity Relationships (QSARs) of Cyclooxygenase-2 (COX-2) Inhibitors by MLR and PC-ANN. Journal of Engineering, Science & Management Education, special issue on (Theoretical/Experimental Aspects of Drug Design), 7(III) (2014): 248-255.

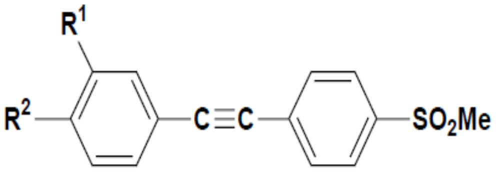
Supplementary material

Table S1. Molecular structures and observed inhibitory activities of the 73 non-tricyclic COX-2 inhibitors expressed as pIC50.

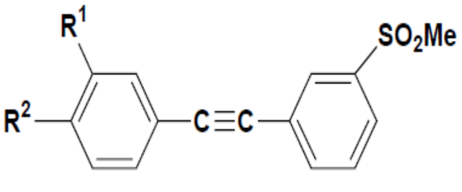


Compound number	Index *	R1	R2	p IC_{50}
1^{C}	9a	H	SO_2Me	6.09691
2^{C}	9b	Me	SO_2Me	6.52287
3^{V}	9c	F	SO_2Me	5
4^{C}	9d	OMe	SO_2Me	5.30980
5^{C}	9e	SO_2Me	H	6
6^{C}	9f	SO_2Me	Me	6.52287
7^{C}	9g	SO_2Me	F	6.22184
8^{C}	9h	SO_2Me	OMe	5.49485

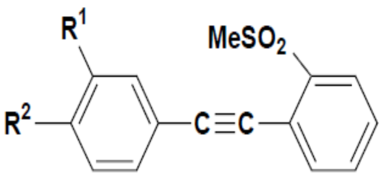
- **Ref 18**



11a-f



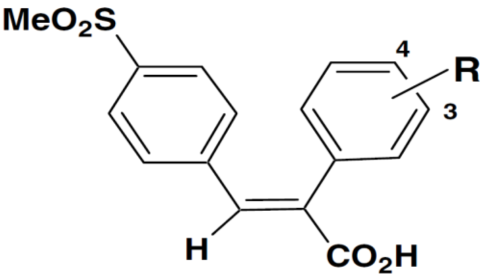
12a-f



13a-f

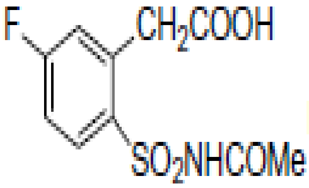
Compound number	Index *	R1	R2	p IC_{50}
9^{V}	11a	H	H	6.05060
10^{V}	11b	F	H	5.22184
11^{P}	11d	H	Me	6.49485
12^{V}	11e	OH	H	6.67778
13^{C}	11f	OAc	H	7.22184
14^{C}	12a	H	H	5.49485
15^{C}	12b	F	H	5.72124
16^{C}	12c	OMe	H	5.46852
17^{C}	12d	H	Me	6.49485
18^{V}	12e	OH	H	6.49485
19^{p}	12f	OAc	H	7.30102
20^{C}	13b	F	H	6.49485
21^{P}	13c	OMe	H	5.31875
22^{V}	13d	H	Me	4.50031
23^{C}	13e	OH	H	5.45593
24^{C}	13f	OAc	H	6.85387

***Ref 19**

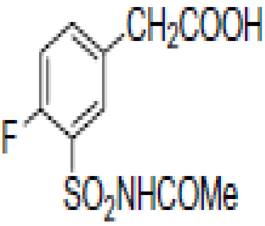


Compound Number	Index *	R	p IC_{50}
25^{P}	9a	4-H	5.52287
26^{V}	9b	4-Br	5.44369
27^{P}	9c	4-F	4.44369
28^{V}	9d	4-OH	5.27572
29^{C}	9e	4-OMe	5.72124
30^{C}	9f	4-OAc	5.53760
31^{C}	9g	4-NHAc	5.60205
32^{C}	9h	3-Br	6.50863

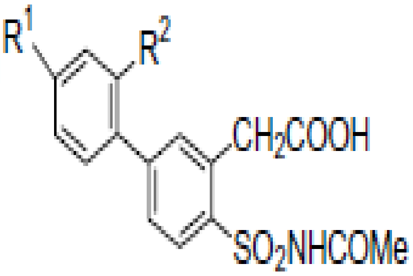
***Ref 20**



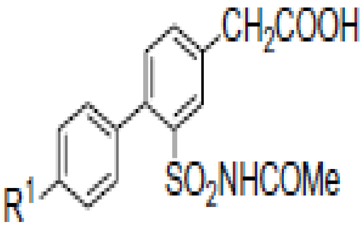
12



14



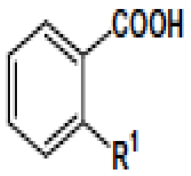
17a-d, 19



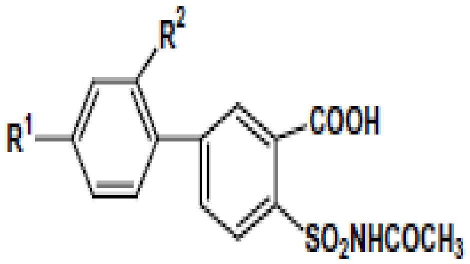
20a-c, e

Compound Number	**Index ***	**R 1**	**R2**	**p IC_{50}**
33^{C}	12	---	----	6.00966
34^{C}	14	---	---	7.92081
35^{C}	17a	H	H	6.08092
36^{C}	17c	F	F	6
37^{out}	17d	$OCH(CH_3)_2$	H	5.50031
38^{C}	19	SO_2CH_3	H	4.50168
39^{C}	20a	H	---	5.75448
40^{V}	20b	F	---	5.81815
41^{out}	20c	$OCH(CH_3)_2$	---	6.82390
42^{C}	20e	SO_2CH_3	---	5.94309

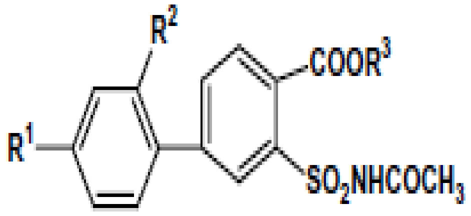
***Ref 21**



1 (Aspirin)
11



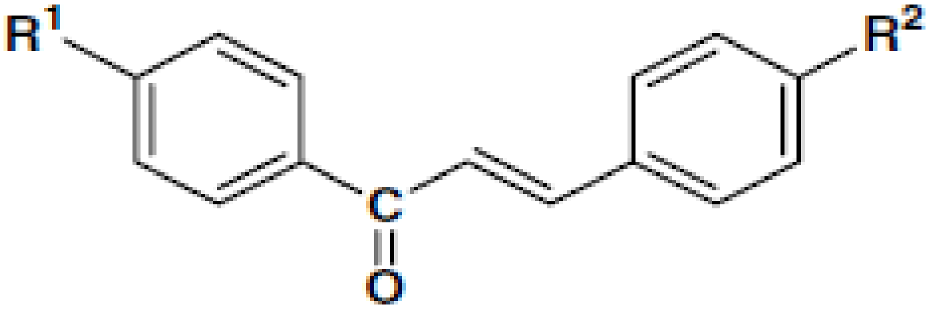
19a-c, e



20a-e, 21

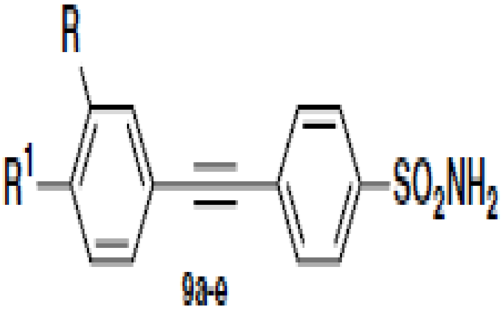
Compound Number	Index *	R 1	R2	R3	p IC_{50}
43^{V}	11	$SO_2NHCOCH_3$	----	----	6.60205
44^{C}	19a	H	H	----	7.52287
45^{V}	19c	F	F	----	7.06048
46^{P}	20a	H	H	H	5.92081
47^{C}	20b	F	H	H	5.42021
48^{P}	20c	F	F	H	6.11350
49^{C}	20d	SO_2CH_3	H	H	6.52287

***Ref 22**

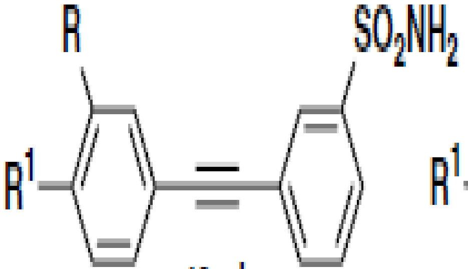


Compound Number	Index *	R 1	R2	p IC_{50}
50P	7a	$NHSO_2Me$	H	6.49485
51C	7b	$NHSO_2Me$	Me	6
52C	7d	$NHSO_2Me$	O Me	5

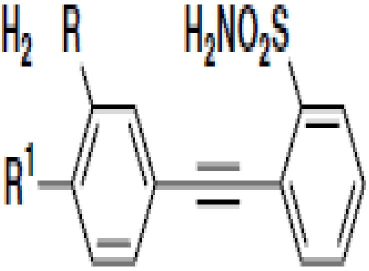
- **Ref 23**



9a-e



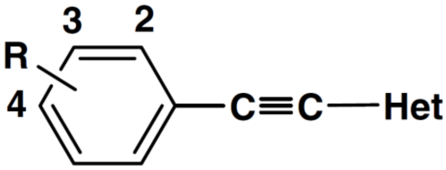
10a-d



11a-c

Compound number	Index *	R	R1	p IC_{50}
53V	9c	OMe	H	5.16749
54P	9d	OH	H	4.48678
55P	9e	F	H	6.52287
56V	10a	H	H	6.34678
57V	10b	H	Me	5.49485
58P	10c	OMe	H	5.69897
59C	10d	F	H	5.16749
60P	11b	H	Me	4.49485

***Ref 24**



20-32

Compound Number	**Index ***	**R**	**Het**	**p IC_{50}**
61^{C}	20	4-SO_2NH_2	2-Pyridyl	6.52287
62^{C}	21	4-SO_2NH_2	4-Pyridyl	4.57186
63^{P}	22	4-SO_2NH_2	3-Me-2-Pyridyl	7.154901
64^{C}	23	2-SO_2CH_3	2-Pyridyl	6.67778
65^{C}	24	2-SO_2CH_3	3-Pyridyl	4.50031
66^{P}	25	2-SO_2CH_3	4-Pyridyl	6.95860
67^{C}	26	2-SO_2CH_3	3-Me-2-Pyridyl	6.37675
68^{C}	27	3-SO_2CH_3	2-Pyridyl	6.69897
69^{C}	28	3-SO_2CH_3	3-Pyridyl	4.49620
70^{C}	29	3-SO_2CH_3	4-Pyridyl	6.49485
71^{C}	30	4-SO_2CH_3	2-Pyridyl	6.48148
72^{C}	31	4-SO_2CH_3	3-Pyridyl	7.39794
73^{C}	(Aspirin)		---	5.61978

***Ref 25**

^{C} Compounds classified in the training or calibration set, ^{P} compounds classified in the external test set (prediction set), ^{V} compounds classified in the validation set. ^{Out} compounds classified as outliers.

Table S2. LOO cross validation parameters for the final MLR models 10-15

Model	PRESS	SPRESS	SST	R^2_{CV}	PRESS/SST	PSE	RSEP
10	18.438	0.545	29.608	0.377	0.623	0.503	8.36
11	17.366	0.534	30.680	0.434	0.566	0.488	8.113
12	15.592	0.51	32.455	0.52	0.480	0.462	7.688
13	14.391	0.494	33.656	0.572	0.428	0.444	7.386
14	13.502	0.483	34.545	0.609	0.391	0.43	7.154
15	13.502	0.478	34.545	0.609	0.391	0.43	7.154

R^2_{cv} is cross-validated correlation coefficient

RSEP is relative standard error of prediction

PRESS is predictive residual sum of squares

SPRESS is uncertainty of prediction

Table S3. Correlation coefficients and cross validation parameters for ANN models 14 and 15.

Model No.	hn.	nPCs	R_train	PRESS_train	SSR_train	R2CV_train	R_test	PRESS_test	RSEP_test	R_val	PRESS_val	RSEP_val
14	7	6	0.706	13.842	9.028	-0.533	0.755	6.020	10.933	0.688	4.166	9.337
15	7	6	0.666	14.965	10.646	-0.406	0.726	6.294	11.179	0.660	4.378	9.572

Table S4. Correlation coefficients and cross validation parameters for optimizing number of hidden nodes for model 14.

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

18	6	0.715	13.822	-0.624	0.748	5.573	10.519	0.65	4.329	9.518
19	6	0.665	15.19	-0.64	0.666	6.925	11.726	0.652	4.3	9.486
20	6	0.653	15.859	-0.488	0.728	6.728	11.558	0.653	4.486	9.689

Table S5. Correlation coefficients and cross validation parameters for optimizing number of hidden nodes for model 15.

hn. NO	**nPCs**	**R_train**	**PRESS_train**	**R2CV_train**	**R_test**	**PRESS _test**	**RSEP_test**	**R_val**	**PRESS_ val**	**RSEP_ val**
5	6	0.670	16.033	-1.023	0.706	6.835	12.171	0.706	3.241	8.702
6	6	0.705	14.782	-0.681	0.710	6.616	11.974	0.663	3.538	9.092
7	**6**	**0.730**	**13.707**	**-0.392**	**0.823**	**4.727**	**10.122**	**0.711**	**3.109**	**8.522**
8	6	0.653	16.133	-0.510	0.772	5.841	11.251	0.689	3.441	8.967
9	6	0.754	12.506	-0.027	0.687	6.617	11.975	0.681	3.246	8.709
10	6	0.692	14.677	-0.108	0.687	6.878	12.209	0.660	3.550	9.108
11	6	0.680	16.342	-1.156	0.707	6.417	11.792	0.700	3.269	8.739
12	6	0.662	16.359	-1.116	0.714	3.570	0.726	0.682	3.553	9.112
13	6	0.659	16.020	-0.609	0.755	5.797	11.209	0.706	3.417	8.936
14	6	0.665	15.791	-0.407	0.818	4.685	10.076	0.688	3.004	8.378
15	6	0.664	15.826	-0.459	0.725	5.867	11.276	0.651	3.745	9.354
16	6	0.717	13.921	-0.185	0.755	5.348	10.766	0.690	3.078	8.481
17	6	0.754	12.343	0.015	0.794	5.175	10.590	0.670	3.464	8.997
18	6	0.756	12.129	0.112	0.723	7.075	12.383	0.656	4.319	10.046
19	6	0.687	15.048	-0.062	0.654	7.020	12.334	0.705	2.940	8.289
20	6	0.700	14.476	-0.116	0.688	6.626	11.984	0.687	3.220	8.674

Table S6. Correlation coefficients and cross validation parameters for chance correlation results for model 14 with 12 hidden nodes.

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

Table S7. Correlation coefficients and cross validation parameters for chance correlation results for model 15 with 7 hidden nodes.

Trial No.	nPCs	R_train	PRESS_train	R2CV_train	R_test	PRESS_test	R_val	PRESS_val	R2CV_val
1	6	-0.028	57.208	-4.349	-0.017	1.273	-0.232	2.308	-11.600
2	6	0.076	51.222	-5.105	-0.225	2.609	-0.162	2.143	-5.870
3	6	0.092	110.310	-8.014	-0.112	10.131	-0.043	11.265	-102.793
4	6	0.149	45.089	-11.099	-0.686	1.727	0.155	1.540	-18.123
5	6	0.128	45.614	-13.166	-0.225	1.316	-0.144	1.634	-77.889
6	6	-0.298	57.911	-12.601	-0.228	1.547	0.253	1.652	-34.570
7	6	-0.166	61.472	-5.667	0.142	1.541	0.142	1.561	-11.975
8	6	-0.183	62.800	-6.189	-0.196	2.330	0.283	2.121	-5.291
9	6	-0.102	60.916	-4.952	-0.134	1.407	-0.166	2.858	-20.457
10	6	0.271	45.575	-2.472	0.151	2.093	0.281	2.435	-9.185