# Quantitative Structure Activity Relationship for Drug Discovery

**Uppaluru Himaja Sree, Dandabathula Kumudhini, Vuyyuru Sruthi Laya, Guttikonda Geetha**

***Abstract*: Quantitative structure-activity relationship (QSAR), gives useful information for drug design and medicinal chemistry. QSAR is a method used to anticipate the organic reaction of a molecule by developing equations which use descriptors calculated from its compounds. The molecular descriptors vary in complexity. A time consuming and expensive process for pharmaceutical industries is drug discovery. An inspiration driving these QSAR models is to help revive the revelation of molecular drug candidates through minimized test work and to bring a drug to market faster. To obtain sorted features principal component analysis is used. The biological activities of the test set are determined by training the neural network using training set. By predicting the activities it can be known whether the drug is close to the target or not.***

***Index Terms*: biological activity, *descriptors, neural networks, Quantitative- structure activity relationship.***

## I. INTRODUCTION

Molecular design is a process that involves iteration. The phases associated with molecular processes are design, synthesis, and test. Evaluation of the results from one iteration gives information and knowledge that enables the subsequent cycle of discovery to be started and further change to be fulfilled.[1] The common feature in the research is the development of a model which validates observed activities to be related to their molecular descriptors. Such types of models are described as Quantitative Structure-Activity Relationships. Data that is especially important for medication and medicinal chemistry will be given by Quantitative Structure-Activity Relationship, known as one of the most main areas of chemistry. Biological activity of compounds is to be predicted using QSAR. Biological activities can be predicted by formulating equations or models utilizing physiochemical properties determined from their molecular structures [2].

### *A. Molecular descriptors:*

Chemical information of a molecular structure that is encoded via mathematical procedures can be expressed in the numerical form called molecular descriptors [3].

### *B. Biological Activity:*

Biological activity can be displayed numerically as the concentration of a molecule expected to show a particular biological reaction [4].

The mathematical form of QSAR is: Activity = f (physiochemical and/or structural properties)

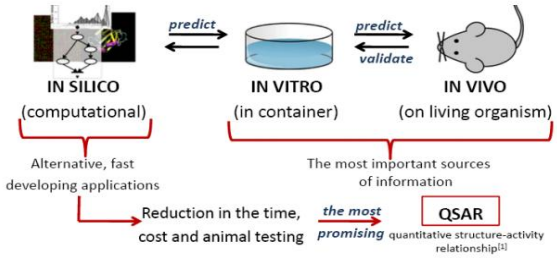


**Figure 1.1 : Approaches to QSAR**

From Figure 1.1 , By utilizing QSAR anyone can interrupt IN VITRO and IN VIVO process, the secondary path comes into existence by choosing QSAR technique

## II. LITERATURE SURVEY

The paper "Neural networks in building QSAR models" analyses a portion of the necessary methods being performed for structuring models similar to the quantitative structure-activity relationship (QSAR) using the artificial neural networks (ANNs).[5] It demonstrates that multi-layer ANN's can be utilized to follow the problems by regression. This paper expands the ability of ANNs, the explainability of the following representations, the concerns of speculation and retention, the training components, and the utilization of neural network groups. [6].

Support vector machines (SVM) address to a standout amongst several promising machine learning (ML) tools that can be related to building up a perceptive description of a quantitative structure-activity relationship (QSAR) utilizing sub-atomic descriptors. In the paper "QSAR models for expectation investigation of HIV protease inhibitors utilizing support vector machines, neural systems and regression" many methods like multiple linear regression (MLR) and artificial neural systems (ANNs) were besides employed to develop quantitative linear and non-linear models to distinguish the outcomes acquired by SVM. The expected results are in great concurrence with the test estimation of HIV activity; likewise, the outcomes reveal the predominance of the SVM over MLR and ANN display [7].

Hydroxyl benzoic esters are add on supplements, being comprehensively used in sustenance, medication, and beauty care items. To explore the association between the sub-atomic structure and antibacterial activity of these mixes and envision the blends with near

**Revised Manuscript Received on June 15, 2019.**

**Uppaluru Himaja Sree**, Information Technology, VR Siddhartha,Vijayawada, India.
**Dandabathula Kumudhini**,, Information Technology, VR Siddhartha,Vijayawada, India.
**Vuyyuru Sruthi Laya**, Information Technology, VR Siddhartha,Vijayawada, India.
**Guttikonda Geetha**, Assistant Professor, Information Technology, VR Siddhartha,Vijayawada, India.

structures, Quantitative Structure-Activity Relationship (QSAR) models of 25 sorts of hydroxyl benzoic esters with the quantum synthetic parameters and sub-atomic accessibility indexes are developed reliant on support vector machine (SVM) by using the R language [8].

Chemo informatics clustering algorithms are critical topics for medication revelation process. Along these lines, there are many algorithms that are accessible for examining vast substance informational collections of medium and high dimensionality. The nature of these calculations relies upon the dataset and the score required by the application. The utilizations algorithms in the medication revelation process are Quantitative Structure-Activity Relationship (QSAR) examination.[9] In view of Structure-Activity Relationship (SAR) demonstration, mixes with comparable structure have same organic action. Bisecting K-Means and Ward calculations are best for extensive number of groups for homogeneous and heterogeneous sets of data in term of standard deviation, yet bisecting K-Means calculation is best in terms of time[10].

.

## III. METHODOLOGY

To detect molecules it is highly crucial to detect molecules that are highly active towards their intended targets than other targets.

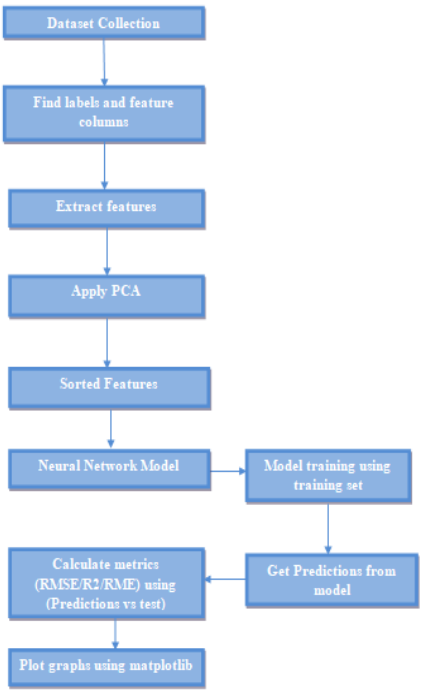


**Fig 3.1 Flow of project**

The biological activities of different molecules need to be predicted whether the molecules are on target or off target. The prediction will be based on molecular descriptors which are generated via mathematical procedures.

The flow of the project is shown in fig 3.1. The Merk's dataset that are utilized for medication disclosure are considered in this paper. The dataset is split into training and testing sets. Only the biologically relevant data needs to be identified from this data set. The dataset contains descriptors that are derived from the structure of molecules. Each row in this dataset represents a molecule. Descriptors and activities are given for training set and just descri0ptors are given for testing set.

Pre-processing is used as a process to safeguard the integrity of data. Labels are the dimensions of the data. Features are the information about molecules that are called as molecular descriptors which are encoded via mathematical procedures. Both the features and labels need to be identified.. Principal Component Analysis (PCA) needs to be done for dimensionality reduction. Only the sorted features will be obtained by doing PCA. In this work, a neural network model called Deep Neural Network (DNN) is developed and the model will be trained utilizing the training set. A model will be created and layers are added using the activation functions. The loss function and the optimizers need to be described. Metrics need to be selected and the model needs to be compiled. The predictions of the model will be obtained and the metrics will be calculated for each epoch.

## IV. ALGORITHM

*Deep Neural Network:*

Deep Neural Network(DNN) is an artificial neural network(ANN) which contains numerous layers between input and output.

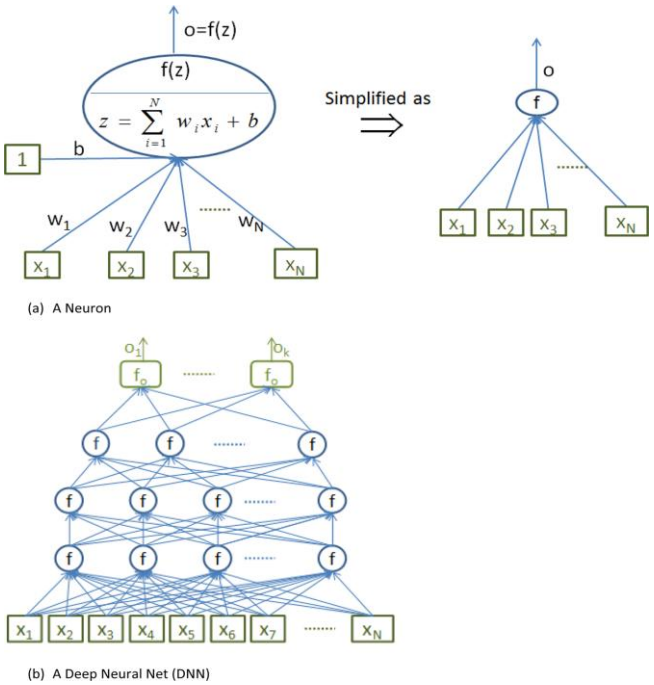


**Fig 4.1 Deep Neural Net**

A network comprises of neurons which are referred to as nodes as shown in fig 4.1 a. Every neuron is associated with activation function as shown in fig 4.1 b. The activation function and the optimizer that is used in this paper is relu and Nadam. There exist mainly three kinds of layers in Deep Neural Networks. They are:

1. Input Layer
2. Output Layer
3. Hidden Layers

In a deep neural network, there will be more than one hidden layer.

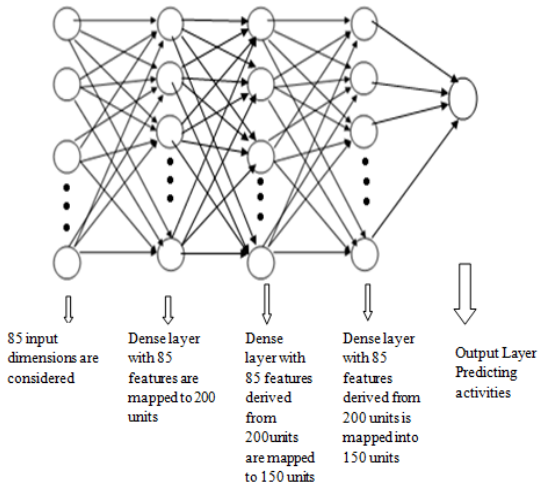


**Fig 4.2 Fully Connected Network**

The fully connected neural network is described as shown in fig 4.2 A sequential model which is a linear stack of layers is created using keras. The bottom layer is the input layer. The information regarding molecules which is represented in numerical form i.e molecular descriptors are given as input to this layer For the input layer, all the features that are obtained after doing Principal Component Analysis are to be taken. Dense layers ate used which are fully connected. The dense layer consists of two outputs a dot product and an addition. They determine the output based on activation function and bias. Here for the input layer, 85 dimensions which are the features obtained after applying PCA are taken. All the features that are taken from the input layer are mapped into 200 units. All the units are the hidden units. All the features that are derived from 200 units are mapped into 150 units. The output layer can be considered as the top layer. The activities which is based on molecular descriptors will be predicted in this layer. The network ends with a Dense layer of size 1 since only a single column is to be predicted. Here five dense layers are used.

| Layer (type) | Output Shape | Param # |
|---|---|---|
| batch_normalization_1 (Batch | (None, 85) | 340 |
| dense_1 (Dense) | (None, 300) | 25800 |
| dense_2 (Dense) | (None, 150) | 45150 |
| dense_3 (Dense) | (None, 200) | 30200 |
| dense_4 (Dense) | (None, 150) | 30150 |
| dense_5 (Dense) | (None, 1) | 151 |

Total params: 131,791
Trainable params: 131,621
Non-trainable params: 170

**Fig 4.3 Information Regarding Layers**

Information regarding layers can be obtained as shown in fig 4.3. The parameters are the weights for the connections. In the first dense layer, there is a total of 25800 parameters that are formed. They are formed by calculating the nodes of the previous layer i.e the inputs with the current layer neurons. There are 85 nodes in the input layer and there are 300 nodes in the present layer. So the parameters formed are 25500. Here for every layer there will be a bias value, so a bias value called as 30 is added. In the second layer 45150 parameters are formed. These parameters are formed by (150*300)+150. 300 neurons from the previous layer are multiplied with 150 neurons and a bias value 150 is added. In the third dense layer the parameters formed by multiplying the previous layers neurons with the neurons present in the current layer and a bias value called as 200 is added. In the next layer a bias value 150 is added. The output layer contains a single neuron. The parameters formed are 151.The output layer contains a single neuron as the activity column needs to be predicted.

## V. RESULTS

In this research work to predict the activity which is a function of molecular descriptors a tool called python is used. Principal Component Analysis (PCA) is utilized to decrease the dimensions i.e features of the dataset.

```
Features Before PCA:2830, After PCA:85
```

**Fig 5.1 Features of dataset**

Fig 5.1 describes the features before and after applying PCA. Using PCA the sorted features will be obtained. Dense layers are added. Dense layers are the fully connected layers. The activation function is calculated for every layer. The optimizer that is used is nadam. The output of one layer will be passed as input to the next layer.

Dense layers implement the following operation: output = activation(dot(input, kernel) + bias). Metrics needs to be specified. This research work includes regression problem. The metrics will be RMSE(Root Mean Square Error), MAE(Mean Absolute Error).

The model is trained to utilize a set called training and the activities of the testing data are determined and metrics are computed for each epoch as shown in fig 5.2.

```
Train on 1815 samples, validate on 598 samples
Epoch 1/10
 - 14s - loss: 2.7780 - mean_squared_error: 2.7780 - rmse: 1.2379 -
- val_r_square: -3.2337e+04
Epoch 2/10
 - 14s - loss: 0.5780 - mean_squared_error: 0.5780 - rmse: 0.6021 -
val_r_square: -1.6271e+04
Epoch 3/10
 - 13s - loss: 0.3677 - mean_squared_error: 0.3677 - rmse: 0.4827 -
val_r_square: -3.6920e+03
```

**Fig 5.2 Metrics for each epoch**

Fig 5.3 describes the value of RMSE for every epoch of the testing set.

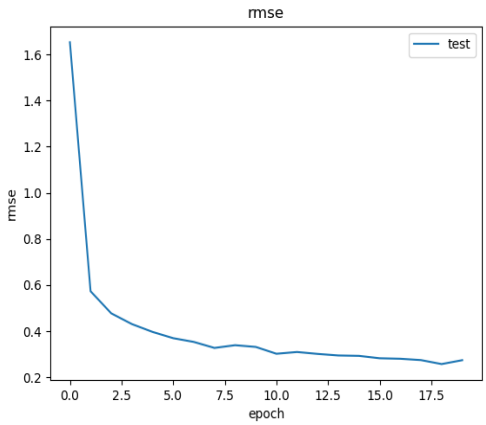


**Fig 5.3** RMSE for each epoch

The predicted and experimented activities are plotted using matplotlib as shown in fig 5.4.

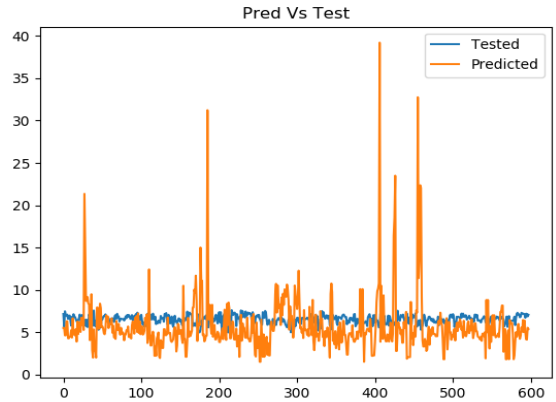


**Fig 5.4** Predicted vs experimented activities

From the fig 5.4 the correlated activities can be known.

## VI. CONCLUSION

The previous couple of decades have seen numerous advances in the improvement of computational models for the expectation of a wide range of organic and chemical activities that are helpful for screening promising molecules with powerful properties. Utilizing QSAR a decision support system is built to relate molecular descriptors with their biological activities. By implementing PCA we get only less number of features using which activities can be predicted quickly. Using QSAR examination a relationship is made to relate molecular descriptors of a compound to its biological activity. The construed association between descriptors and activity is used to measure the property of various molecules or possibly to find the parameters impacting the organic activity. QSAR gives ease apparatuses to the choice of novel "hits" and for "lead" advancement during medication revelation and improvement..Atomic data can likewise associate with physiochemical properties which were named as QSPR (Quantitative Structure-Property Relationship). QSPR can also be found.

## REFERENCES

1. Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, and on Computational Methods in Drug Discovery
2. gor I. Baskin , Vladimir A. Palyulin , and Nikolai S. Zefirov on Neural Networks in Building QSAR Models ,2008 Research Gate
3. Rachid Darnaga, BrahimMinaouia, MohamedFakir on QSAR models for prediction study of HIV protease inhibitors using support vector machines, neural networks and multiple linear regression, Arabian Journal Of Chemistry, Volume 10, Supplement 1, February 2017
4. Li Wen, Qing Li, Wei Li, Qiao Cai, and Yong-Ming Cai on A QSAR Study Based on SVM for the Compound of Hydroxyl Benzoic Esters , BioInorganic Chemistry and Applications, Volume 2017, Article ID 4914272, 10 pages
5. Mohamed G. Malhat ; Hamdy M. Mousa ; Ashraf B. El-Sisi on Clustering of chemical data sets for drug discovery ,2014 9th International Conference on Informatics and Systems
6. Breiman, L. Random forests. Machine Learning 2001, 45, 5−32.
7. Cortes, C.; Vapnik, V. N. Support-vector networks. Machine Learning 1995, 20, 273−297.
8. Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: an ensemble learning tool
for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 2005, 45, 786−799.
9. Bruce, C. L.; Melville, J. L.; Picket, S. D.; Hirst, J. D. Contemporary QSAR classifiers compared. J. Chem. Inf. Model. 2007, 47, 219−227.
10. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput.Sci. 2003, 43, 1947−1958.
11. Fernandez-Delgado, M.; Cernades, E.; Barro, S.; Amorim, D. A. Do we need hundreds of classifiers to solve real world problems? J. Machine. Learning. Res. 2014, 15, 3133−3181.
12. Burden, F. R. Quantitative structure-activity relationship studies using Gaussian Processes. J. Chem.Inf. Comput. Sci. 2001, 41, 830− 835.

## AUTHORS PROFILE

**Uppaluru Himaja Sree,** Student, Velagapudi Rama Krishna Siddhartha Engineering College.

**Dandabathula Kumudhini**, Student , Velagapudi Rama Krishna Siddhartha Engineering College.

**Vuyyuru Sruthi Laya,** Student , Velagapudi Rama Krishna Siddhartha Engineering College.

**Guttikonda Geetha,** Assistant Professor, Velagapudi Rama Krishna Siddhartha Engineering College.