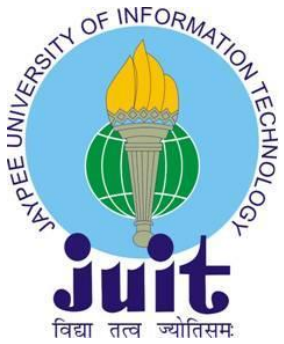
Applications of statistical and docking approach for the prediction of lead molecule against *Isocitrate lyase*, an enzyme involved in the persistence of *Mycobacterium tuberculosis.*



May 2017

Enrollment Number: 131509, 131510

Name of student: Bharti Sharma, Astha Khanduri

Name of Project Guide: Dr. Chittaranjan Rout

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION AND TECHNOLOGY

WAKNAGHAT, Himachal Pradesh

CERTIFICATE

This is to certify that the project report entitled "Applications of statistical and docking approaches for the prediction of lead molecule against *Isocitrate lyase*, an enzyme involved in persistence of *Mycobacterium tuberculosis*", submitted by Bharti Sharma and Astha Khanduri in partial fulfilment for the award of degree of Bachelor of Technology in Bioinformatics to Jaypee University of Information Technology, Waknaghat has been carried out under supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date: **Dr.Chittaranjan Rout**

ACKNOWLEDGEMENT

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them. We are highly indebted to our guide **Dr. Chittaranjan Rout**, for his guidance and constant supervision as well as also for his support in completing the project.

We would like to express our gratitude towards **Ms. Nupur S. Munjal** for providing necessary information regarding the project for their kind co-operation and encouragement which helped us in the completion of this project.

Our thanks and appreciations also go to everyone who has willingly helped us out with their abilities in developing the project.

Thank you

TABLE OF CONTENTS

CHAPTER 2 - INTRODUCTION

2.1 Introduction about MTB

Tuberculosis (TB) is a disease caused by a bacterium called *Mycobacterium tuberculosis*. The bacteria usually attack the lungs, but TB bacteria can attack any part of the body such as the kidney, spine, brain, etc. If it is not treated properly, TB can cause death. TB can also pass from a mother to her unborn child before and after birth. TB is more common in children and people with weak immune systems and people with HIV.

Mycobacterium tuberculosis is a pathogen that hides in the macrophages of its host. Because macrophage is important for removing the infection, the interaction of pathogen and host cell reveals a continuous clash for control.

2.2 *Isocitrate lyase*

Isocitrate lyase (*ICL*) plays a central role in the persistence of *Mycobacterium tuberculosis*. ICL of Mycobacterium tuberculosis (MTB) picks up the most attention because it is linked with tuberculosis. MTB ICL is most troublesome organism/enzyme to contemplate in light of the fact that it develops gradually and is likewise connected with the higher danger of disease. At present, accessible information demonstrated that ICL is steady as a dimer however it will only be functional in a tetrameric form.The enzyme permits net carbon picks up by redirecting acetyl-CoA from B-oxidation of fatty acids into the glyoxylate shunt pathway.

Bacteria and some species of higher plants can get a net increment in malate or oxaloacetatethrough the expression of enzymes of the glyoxylate cycle or glyoxylate shunt. *Isocitrate Lyase* and *Malate Synthase* are the two enzymes that permit the glyoxylate shunt, which convert isocitrate to succinate or to malate via glyoxylate.

Three clinically used TB drugs, isoniazid (INH), rifampicin (RIF) and streptomycin (STREP). Each of the three drug treatments surprisingly triggers atypical example of metabolic modificationscharacteristicof an activation of MTB's ICLs, enzymes usually related with the assimilation of acetyl CoA and renewal of tricarboxylic acid (TCA) cycle metabolites.

Isocitrate lyase is otherwise called isocitritase, isocitrase, isocitrate glyoxylate-lyase and isocitratase. ICL is present in bacteria, archaea, nematodes, fungi, plants, and protists. Overall, ICL has a crucial role in microbial pathogenicity and seed germination of higher plants. An alternate pathway to produce energy when tricarboxylic acid cycle (TCA cycle or Krebs cycle) is downregulated upon oxygen and nutrient

depletion is the glyoxylate cycle. When nearly all of the TCA enzymes are suppressed, glyoxylate enzymes will be unregulated. The initial phase of glyoxylate cycle is similar to the TCA cycle. Whereas, when the cycle is utilized, few beta oxidation steps in TCA cycle are avoided. The acetyl-CoA is the only substrate for cycles. Source of the precursor, acetyl-CoA, is different for the respective cycle. Carbohydrate undertakes glycolysis to generate the acetyl-CoA while lipid undergoes beta-oxidation to generate acetyl-CoA. The beginning point of differentiation for these two cycles occurs when acetyl CoA gets converted to isocitrate. In glyoxylate cycle, two vital enzymes namely, isocitrate lyase (ICL) and malate synthase (MS) are required. The basic purpose of ICL is to cleave isocitrate to glyoxylate and succinate while glyoxylate is converted into malate by adding an acetyl group through MS. It is shown that during downregulation of TCA cycle, the inhibition of ICL is lethal for MTB.[1]

- Isocitrate ⇨ Glyoxylate + Succinate

The structure of ICL1 bound to pyruvate and succinate shows that the active site can house the additional methyl group of 2-methylisocitrate and catalyze the MCL reaction. Hence, they concluded that the supposed substrate specificity of MCLs and ICLs is due to conserved differences in three residues that form a hydrophobic depression in the active site. These residues are Thr347, Trp283 and Phe345 in *M. tuberculosis* ICL1.

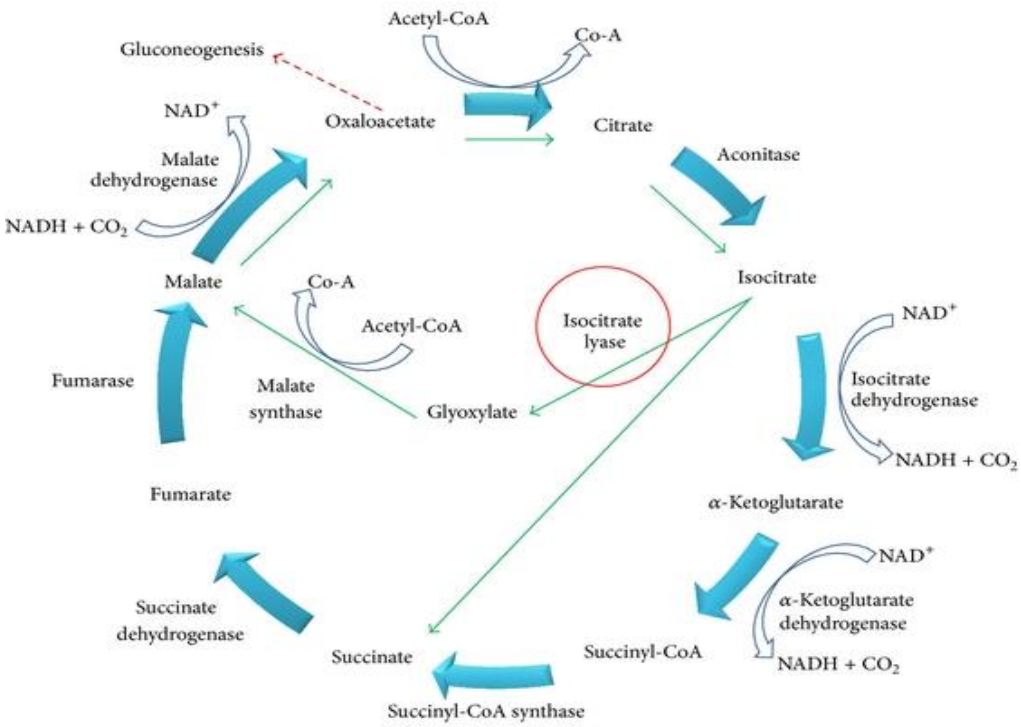


Figure 2.1 TCA Cycle (with blue arrows) and Glyoxylate cycle(with green arrows)

Active MTB operates TCA cycle, mistreatment sugar as main carbon supply to get energy. The activity of MTB by scavenger cell results in O and nutrient depletion, imposing MTB to enter its dormant part. This

causes a colossal metabolic shunt and down regulates TCA enzymes. The viability of MTB is taken care by glyoxylate enzymes which are up-regulated in order to continue generating energy from another carbon supply known as lipids. [2]

2.4 ICL Structure

The structure of MTB ICL is present in PDB with ids: ***1F8I, 1F8M, 1F61***. Currently, obtainable data show that ICL is stable as a dimer but it will only be functional in a tetramer form. It has α/β barrel with eight α-helixes and β-strands in each of its subunits with an additional α-helix and its two succeeding helices projected out from the barrel. These succeeding helices involve in the interaction with neighboring subunit. An essential domain with a number of active side residues i.e. small β-domain is present at the top of the barrel. [3]

1F61 is a ligand-free ICL structure which has its active site in open conformation. Whereas, 1F8M bind pyruvate and 1F81 bind succinate/glyoxylate. Binding of ligand accounts for conformational change which prompts active site to shift into a close conformation. According to the mechanism, the first glyoxylate binds to ICL then succinate is buried deeper than glyoxylate. As per the study, it is proposed thatC-C bond of isocitrate is cleaved by means of Claisen condensation. Further, the authors have also put forward that ICL of MTB has twin role in both glyoxylate cycle and methyl citrate cycle. Methyl citrate cycle removes a toxic by-product of lipid beta oxidation named as propionyl-CoA. Citrate synthase, 2methyl isocitrate lyase (MCL) and methyl citrate dehydrogenase are the enzymes concerned with methyl citrate cycle. Though it was found that MTB does not produce MCL, ICL1 carries MCL's function making it more important than anticipated. On the other hand, the cleavage information regarding Isocitrate-ICL complex is needed which is unavailable presently to prove the above mentioned hypothesis. [4]

The potency ICL as a drug target has been confirmed by numerous studies. As ICL gene is not found in mammals according to Dunn et al., as a result, it is safe to administer any drug against ICL in human. Mu˜nozEl´ıas and Mc Kinney demonstrated that for the survival of MTB, the two isoforms of ICL gene (ICL1: prokaryotic-like isoform and ICL2: eukaryotic-like isoform) coded by *icl* gene (ICL1) and *aceA* gene (ICL2)are necessary jointly. They illustrated that if there is an absence of any of the one isoforms, survival of MTB won't be harmed, however, the absence of both isoforms will cause MTB to be eliminated from the host lungs. Current research concerning ICL is centrally focused on ICL1 (including the solved structure of ICL). Another isoform, the *ace A* gene, is yet less active compared to *icl* gene and is not expressed in all mycobacterium strain.[5]

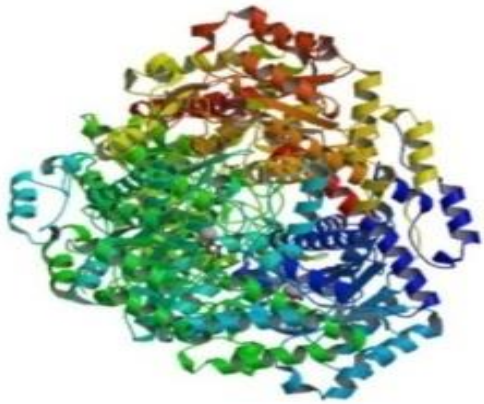


Figure2.2:The above figure is of ICL Structure PDB ID: 1F8I

2.5Role in persistent stage

The ability of *Mycobacterium tuberculosis* to persist in its human host in spite of extensive chemotherapy is thought to be built on subpopulations of non-replicating phenotypically drug-resistant bacilli. Nutrient-starved bacilli lacking the glyoxylate shunt enzyme isocitrate lyase failed to reduce their intracellular ATP level and died, thus establishing a link between ATP control and intermediary metabolism. This reasons that ICL assumes an essential part in the persistent phase of MTB.

2.6Challenges against ICL as a drug target

Having active site of small size, ICL becomes a computationally challenging target. It confines the scope of inhibitor design and only inhibitors with molecular weight< 200 are able to dock properly in the active site. One alsoneeds to identify which groups of the target site are interacting more favorably and the part which is not interacting with the inhibitors. As the compounds molecular weight increases due to the addition of the groups its activity decreases so for the better understanding of the interaction of the compounds with the target site we need to access which parts of the compounds, are the keyto interaction. Hence these challenges against the ICL make the study more challenging.

2.7Current study done (in vitro and in vivo)

Several analogs of 5-nitro-2-furoic acid hydrazones and fourteen derivatives of 5-nitro-2, 6-dioxohexahydro-4-pyrimidinecarboxamides have been synthesized according to the current studies. These derivatives were subjected toin vitro activity against multidrug resistant (MDR) of Mycobacterium tuberculosis. Similarly, from phthalic anhydride, twenty four novel 2-[3-(4-bromo-2-fluorobenzyl)-4-oxo-3,4-dihydro-1-phthalazinyl] acetic acid amides were synthesized and evaluated for in vitro and in vivo evaluation against starved and log phase of mycobacterial species and *Mycobacterium tuberculosis* isocitrate lyase enzyme

inhibition studies.Moreover,novel2-[3-(4-bromo-2-fluorobenzyl)-4-oxo-3,4-dihydro-1-phthalazinyl]acetic acid hydrazones were synthesized from phthalic anhydride and assessed for in vivoandin vitro activities against eight mycobacterial species and *Mycobacterium tuberculosis* isocitrate lyase enzyme inhibition studies. Like this, several compounds were formed who failed to get the desired activity in vivo though they were active in vitro. These compounds were extremely toxic. Some known compounds that are also active in vivo do not target ICL.

2.8Objective of the project

The motive of our project is to generate a QSAR model through statistical such that a new lead molecule can also fit into it. The calculated descriptors will be used as the independent variables while the activity of the structure as the dependent variable for the generation of the model.

The data will be collected from the literature. These inhibitors which are active in vivo will be optimized. After docking the receptor i.e. ICL with these optimized inhibitors, the addition of some of the groups is required to increase the efficacy as addition leads to the increase in the activity. Also, the toxicity is increased with the addition of the groups, as the molecular weight of the compound increases. After analyzing the active site of ICL which is very small, one needs to identify which groups interact more favorably and the part which is not interacting with the active site. The second step will be to find the conserved active site residues in ICL sequences from various M.tuberculosis sequences and also mycobacterium sp. The third step will include the interactions study of top most ligands with the conserved residues. Finally, find out the ADME properties. We will find out the structure-activity relationships of all inhibitors so that a new potential inhibitor against ICL can be tested by fitting in this model. Following is the basic outline of the methodology which will be used in the project.

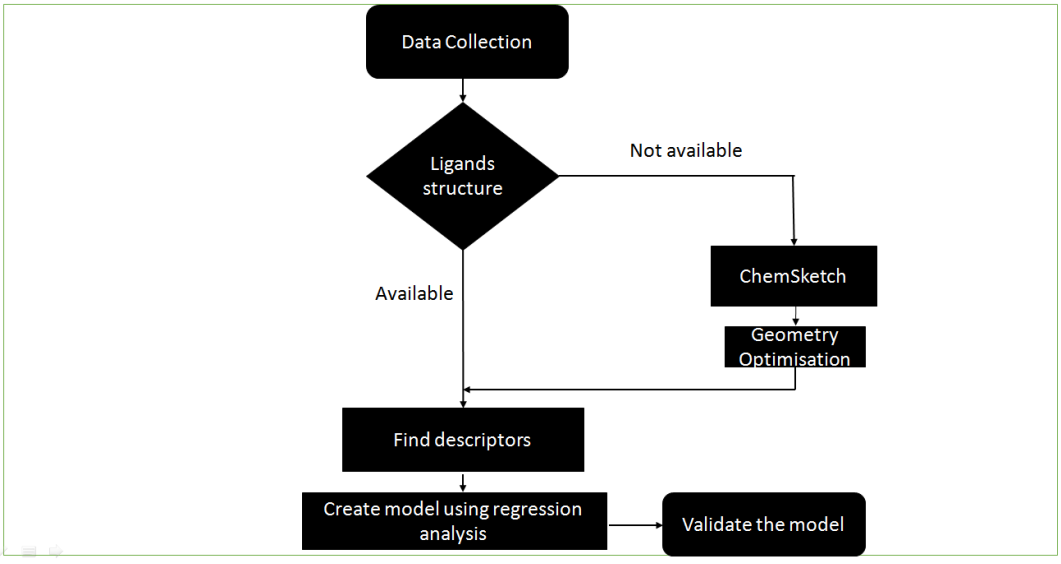


Figure 2: Outline of methodology of the project

CHAPTER 3 - MATERIAL AND METHODS

3.1 Materials

PDB- Protein Data Bank is a database which contains data for three-dimensional structures of large biological molecules like proteins and nucleic acids. The information acquired by NMR spectroscopy or X-ray crystallography is submitted by biochemists and biologists from around the globe. It is an important resource in areas like structural genomics.

ChemSketch: A molecular modeling program for sketching and modifying chemical structures. It makes the understanding of functional groups, their nature and chemical bonds easy by displaying both 2D and 3D molecular structures. [6]

Webqc: An online portal for converting file formats. This interface aims at creating several quantum chemistry programs systems. In addition to it, this software also provides 3D visualization of molecular geometry configuration.

MOPAC:A semi-empirical molecular orbital package which stands for Molecular Orbital PACkage. It is used to study molecular structures, solid state and reactions. Optimization done by the software is broadly used in drug discovery, QSAR/SAR and docking. [7]

Maestro: A platform with highly significant usability which was used for docking in this project conveys clarity and simplicity to an array of molecular systems. It runs on Windows, Linux and Macintosh. [8]

Dragon7: It is an application for the calculation of molecular descriptors. These descriptors can be utilized to assess structure-property relationships or activity of molecular structures. Besides this, they can use for high-throughput screening of molecule databases and similarity analysis. It is extensively used in parts of QSAR suites. It consists of more than 5000 descriptors.

Weka: It is used here for generating equations. It is a collection of machine learning algorithms.

RStudio: An open source providing an integrated development environment (IDE) for R. R is a programming language that's used for statistical analysis like regression, graphics, correlation, etc.

3.2 Methods

3.2.1 Data Collection

- Structure data was collected from the literature.
  Also, activity data and other biological features of the inhibitors (structures) were collected from the literature.

ChemSketch - It was used to draw 2D structures of the inhibitors. Steps:

1. Structures were drawn in 2D space of the software.
2. Then, it was viewed in 3D viewer.
3. Tools >> 3D optimization.
4. The structure was saved in .mol format.

3.2.2Conversion of file format

Discovery Studio:

It was used for converting our 3D .mol files into 3D mol2 format.

Webqc:

Using this online portal, xyz files were made.

1. The content of mol2 files was copied.
2. It was pasted in the input area of webqc portal.
3. Input file type was mentioned as mol2 –Sybyl Mol2 format and the output file type as xyz—XYZ Cartesian coordinates format.
4. The file was converted and the result was copied in the new file with .xyz as extension.

3.2.3Geometry Optimization

MOPAC:

1. Input files were prepared by adding the following command in the beginning of the xyz files**.**
2. They were saved in .mop file format.

Command:

PM6 PRECISE EF T=96H GEO-OK PDBOUT
Name of the file

3. On Linux**,** the mop files were copied and accessed using the terminal.

4**.** Following command was used to optimize the files using MOPAC.

mopac input_filename.mop output_filename.pdb

5. Three files were obtained as output - .arc, .out and .pdb files.

Maestro: It was further used for valences verification and geometry optimization.

3.2.4Feature Extraction

Dragon7:

1. Open Mol2 files after optimizing them.
2. Go to Calculate >> Descriptors.
3. Save them in file/block format.
4. 30 .txt files (descriptors) for every structure were obtained as output.
5. 30 independent excel files of every descriptor were prepared for compiling all structures in it.

Weka:

It was used for generating equations.

1. Click on knowledge explorer.
2. Open excel(.csv) file.
3. In Weka Knowledge Explorers choose classify.
4. Select function in classifier
5. Select linear regression.
6. Then set cross-validation folds to 3, 4, and 5.

3.2.5 Docking

Maestro:

3.2.5.1 Protein preparation

1. Open protein preparation wizard.
2. Import the protein with PDB id: 1F8I
3. Preprocess it under the Import and Process tab
4. Under the Refine tab optimize the structure.
5. Run restrained minimization of the protein structure.
6. Default parameters were taken.
7. Protein is prepared.

3.2.5.2 Ligand preparation

1. Open LigPrep panel in Applications and import the ligand.
2. Keep all the parameters as it is.
3. Save your job in maestro format.
4. Click run and the ligands are prepared.

3.2.5.3 Receptor Grid Generation

1. Go to Applications >> Glide >> Receptor Grid Generation
2. Open Site Tab
3. Select Centroid Of Selected Residues >> Specify Residues
4. Active Site Residues Panel opens.
5. The following residues were picked from the receptor- 283 Trp, 313 Asn, His 193, Ser 315, Ser 317, Phe 345 and Thr 347
6. All of the residues from chain A were selected.
7. The receptor in the workspace will now have an enclosing box as a set of coordinate axes of bright green color which will define the active site of receptor.

3.2.5.4 Ligand Docking

Requirements:

- Receptor Grid (.zip file) generated after the Glide Grid Generation
- Ligprep-out.maegz after ligand preparation

1. Go to Application >> Glide >> Ligand Docking
2. Glide Docking Panel opens
3. Browse the Receptor Grid
4. Browse the Ligand-out.maegz
5. Set the Precision mode of docking as XP Docking mode
6. Activate write XP Descriptors
7. Click run

3.2.6Statistical Analysis

RStudio

1. Two codes were used for the elimination of descriptors that were less correlated with each other.
2. We used two functions in this code for elimination.

First Code:

```
rm(list=ls())
drg<-read.delim("C:/user/bharti/Desktop/DRG/Drg1.txt")
source('C:/user/bharti/Desktop/DRG/vif_func.R', encoding = 'UTF-8')
library("MASS")
library("clusterGeneration")
library("VIF")
library("fmsb")
library("usdm")
r<-nrow(drg)
c<-ncol(drg)-2
y<-matrix( 0, nrow=r,ncol=1, byrow=TRUE )
x<-matrix( 0, nrow=r,ncol=c, byrow=TRUE )
for (i in 1:r )
{
y[i,1]<-drg[i,2]
for (j in 1:c ) {

x[i,j]<-drg[i,j+2]

  }}

x<-data.frame(x)
x
dim(x)
v1<-round((ncol(x)+1)/2)
#View(x)
#vif(x)
v<-matrix( 0, nrow=1,ncol=c, byrow=TRUE )
id1<-matrix( 0, nrow=1,ncol=c, byrow=TRUE )
for (i in 1:c ){
v[i]<-var(x[,i])
if(v[i]==0)

{

id1[i]<-i
}}
id1
x0<-x
if(sum(id1)==0){x}else{x[,id1]}

x

v1<-round((ncol(x)+1)/2)
drlm<-matrix(0, nrow=v1, ncol=5, byrow=TRUE)
pvhchv<-rep(0,v1)
pvhclv<-rep(0,v1)
phchv<-rep(0,v1)
phclv<-rep(0,v1)
pmchv<-rep(0,v1)
pmclv<-rep(0,v1)
pzchv<-rep(0,v1)
pzclv<-rep(0,v1)
nzchv<-rep(0,v1)
nzclv<-rep(0,v1)
```

```
nmchv<-rep(0,v1)
nmclv<-rep(0,v1)
nhchv<-rep(0,v1)
nhclv<-rep(0,v1)
nvhchv<-rep(0,v1)
nvhclv<-rep(0,v1)

cm<-ncol(x)
cm
cmx<-matrix(0, nrow=cm, ncol=cm, byrow=TRUE)
cmx
for(k in 1:v1)
{
        for(i in 1:(cm-1))
        {
        for(j in (i+1):cm)
        {
        if(var(x[,i])*var(x[,j])!=0){cmx[i,j]<-cor(x[,i],x[,j])}else{cmx[i,j]<- -2}
        }}
t(which(cmx==max(cmx), arr.ind=T))[1:2]
         id2<-t(which(cmx==max(cmx), arr.ind=T))[1:2]
        id2
drlm[k,1]<-id2[1]
drlm[k,2]<-id2[2]
drlm[k,3]<-cor(x[,id2[1]],x[,id2[2]])
drlm[k,4]<-var(x[,id2[1]])
print(drlm)
if(drlm[k,4]<drlm[k,5])
{
drlm[k,1]<-id2[1]
drlm[k,2]<-id2[2]
drlm[k,4]<-var(x[,id2[1]])
drlm[k,5]<-var(x[,id2[2]])
}
else
{
drlm[k,1]<-id2[2]
drlm[k,2]<-id2[1]

drlm[k,4]<-var(x[,id2[2]])

drlm[k,5]<-var(x[,id2[1]])

}

print(drlm)
id2
x[id2]<-0
}
x0
drlm
pvhchv<-drlm[drlm[,3]>0.999&drlm[,3]<=1.0,1,drop=FALSE]
pvhclv<-drlm[drlm[,3]>0.999&drlm[,3]<=1.0,2,drop=FALSE]
phchv<-drlm[drlm[,3]>0.9&drlm[,3]<=0.999,1,drop=FALSE]
phclv<-drlm[drlm[,3]>0.9&drlm[,3]<=0.999,2,drop=FALSE]
```

```
pmchv<-drlm[drlm[,3]>0.35&drlm[,3]<=0.9,1,drop=FALSE]
pmclv<-drlm[drlm[,3]>0.35&drlm[,3]<=0.9,2,drop=FALSE]
zchv<-drlm[drlm[,3]>= -0.5&drlm[,3]<=0.35,1,drop=FALSE]
zclv<-drlm[drlm[,3]>= -0.35&drlm[,3]<=0.35,2,drop=FALSE]
nmchv<-drlm[drlm[,3]>= -0.9&drlm[,3]< -0.35,1,drop=FALSE]
nmclv<-drlm[drlm[,3]>= -0.9&drlm[,3]< -0.35,2,drop=FALSE]
nhchv<-drlm[drlm[,3]>= -0.999&drlm[,3]< -0.9,1,drop=FALSE]
nhclv<-drlm[drlm[,3]>= -0.999&drlm[,3]< -0.9,2,drop=FALSE]
nvhchv<-drlm[drlm[,3]> -1&drlm[,3]<=-0.999,1,drop=FALSE]
nvhclv<-drlm[drlm[,3]> -1&drlm[,3]<=-0.999,1,drop=FALSE]
t(pvhchv)#Positive Very High Correlation, High Variance
t(pvhclv)#Positive Very High Correlation, Low Variance
t(phchv)#Positive High Correlation, High Variance
t(phclv)#Positive High Correlation, Low Variance
t(pmchv)#Positive Moderate Correlation, High Variance
t(pmclv)#Positive Moderate Correlation, Low Variance
t(zchv)#Positive No Correlation, High Variance
t(zclv)#Positive No Correlation, Low Variance
t(nmchv)#Negative Moderate Correlation, High Variance
t(nmclv)#Negative Moderate Correlation, Low Variance
t(nhchv)#Negative High Correlation, High Variance
t(nhclv)#Negative High Correlation, Low Variance
t(nvhchv)#Negative Very  High Correlation, High Variance
t(nvhclv)#Negative Very High Correlation, Low Variance
```

Second Code :

```
rm(list=ls())
drg<-read.delim("C:/Users/om/Desktop/A/2 D/Drg 7.txt")
source('C:/Users/om/Desktop/PM 6/vif_func.R', encoding = 'UTF-8')
library("MASS")
library("clusterGeneration")
library("VIF")
library("fmsb")
drg
r<-nrow(drg)
c<-ncol(drg)-4
#View(drg)
y<-matrix( 0, nrow=r,ncol=1, byrow=TRUE )
x<-matrix( 0, nrow=r,ncol=c, byrow=TRUE )
for (i in 1:r )
{
y[i,1]<-drg[i,4]
for (j in 1:c )   {
x[i,j]<-drg[i,j+4]
  }}
y
x<-data.frame(x)
```

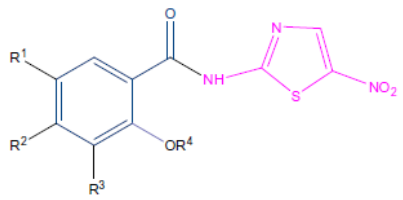
```
x
id<-c(103, 191, 209, 184, 172, 100, 168, 162, 163,  97 , 167,  156,  159,  146,  173,169,  143,
176,  104,  101,  148,  115,  207,  102,  149,  147,  152)
c1<-ncol(xc)
c1
v<-matrix( 0, nrow=1,ncol=c1, byrow=TRUE )
v
id1<-matrix( 0, nrow=1,ncol=c1, byrow=TRUE )
id1
c1
for (i in 1:c1 )
{
v[i]<-var(xc[,i])
if(v[i]==0)
{ id1[i]<-i
  }}
id1
xm1<-xc
if(sum(id1)>0)
{
  xm1<-xc[, -id1]
}
xm1
c2=ncol(xm1)
id2<-matrix( 0, nrow=1,ncol=c2, byrow=TRUE )
id2
crp<-matrix( 0, nrow=1,ncol=c2, byrow=TRUE )
crp
summary(lm(y~xm1[,1]))$r.squared
for (i in 1:c2)
  {
crp[1,i]<-summary(lm(y~xm1[,i]))$r.squared{
id2[i]<-i
   }}
id2
xm2<-xm1
id2
if(sum(id2)>0)
{
  xm2<-xm1[, -id2]}
xm2
# If number of column in xm2 is zero then stop here.
dat1 <- cbind(data.frame(Y=y),as.data.frame(xm2))
fit1 <- lm( Y~. , data=dat1 )
summary(fit1)
summary(fit1)$coefficient[,1]
summary(fit1)$coefficient[,4]
summary.aov(fit1)
fit2<-step(fit1, direction="backward")
summary(fit2)
summary.aov(fit2)
#step(fit2, direction="backward")
anova(fit2,fit1)
```

```
#%%%%%%%%%%
#create mannually
idp<-c(8,9,10,48,49,69,70,109,110,150,151,168,186,187,205,222,223)
xm3<-x[, idp ]
dat2 <- cbind(data.frame(Y=y),as.data.frame(xm3) )
fit3 <- lm( Y~. , data=dat2)
summary(fit3 )
summary.aov(fit3 )
#vif(xm2)
fit4<-step(fit3, direction="backward" )
summary(fit4 )
summary.aov(fit4 )
anova(fit4,fit3,fit2,fit1 )
#create mannually
idp1<-c(1,4,6,12,19,23,26,34,37,45,48,56,72)
xm4<-x[, idp1 ]
dat3 <- cbind(data.frame(Y=y),as.data.frame(xm4) )
fit5 <- lm( Y~. , data=dat3)
summary(fit5 )
summary.aov(fit5 )
vif(xm4 )
fit6<-step(fit5, direction="backward" )
summary(fit6)
summary.aov(fit6)
```

CHAPTER 4 - RESULTS AND OBSERVATIONS

4.1. Data Collection

Core Group for the first set of structures:



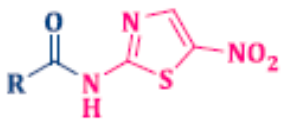
1-18

Fig 4.1aconstitutes the core structure

Compound	R1	R2	R3	R4
1	H	Cl	H	-OH
2	H	Cl	H	-COCH3
3	H	Cl	H	-COC6H5
4	Cl	H	H	-OH
5	Cl	H	H	-COCH3
6	Cl	H	H	-COC6H5
7	Cl	H	Cl-	-OH
8	Cl	H	Cl-	-COCH3
9	Cl	H	Cl-	-COC6H5
10	CH3O-	H	H	-OH
11	CH3O-	H	H	-COCH3
12	CH3O-	H	H	-COC6H5
13	NO2-	H	NO2-	-OH
14	NO2-	H	NO2-	-COCH3
15	NO2-	H	NO2-	-COC6H5
16	(CH3)2CH-	H	(CH3)2CH-	-OH
17	(CH3)2CH-	H	(CH3)2CH-	-COCH3
18	(CH3)2CH-	H	(CH3)2CH-	-COC6H5

Table 4.1 a Contains R group, this R_1, R_2, R_3, R_4 is substituted in the core structure

Core Group for the second set of structures:



19-27

Fig 4.1 b constitutes the core structure

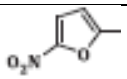
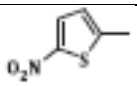
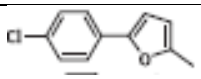
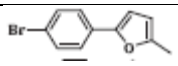
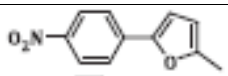
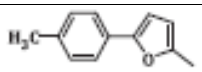
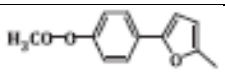
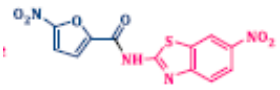
Column	R
19	
20	
21	
22	
23	
24	
25	
26	
27	
28	

Table 4.1 b Contains R group, this R is substituted in the core structure

Core Group for the third set of structures:

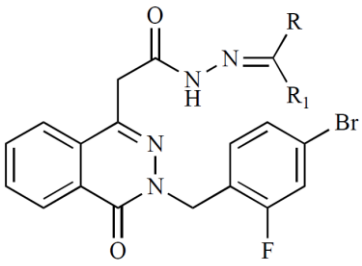
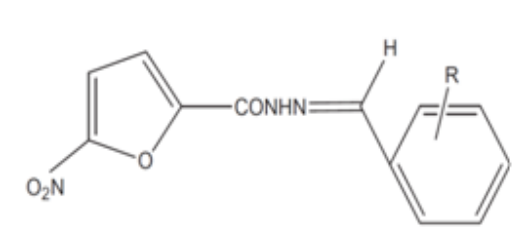


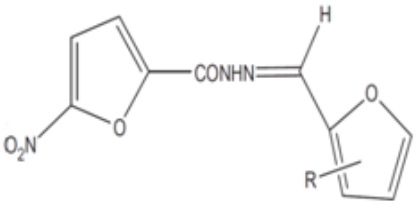
Fig 4.1 c constitutes the core structure

Comp No.	R	R1	IC50	MTB(um)
7a	H	Phenyl	NT	6.73
7b	H	4-Flurophenyl	>129.3	0.81
7c	H	3-Flurophenyl	>129.3	1.61
7d	H	2-Trifluromethylphenyl	117.2	0.36
7e	H	2,6-Dichlorophenyl	>117.0	1.46
7f	H	3-Bromophenyl	>114.9	1.43
7g	H	4-Bromophenyl	>114.9	0.72
7h	H	2-Nitrophenyl	>122.5	0.76
7i	H	3-Nitrophenyl	>122.5	0.37
7j	H	4-Nitrophenyl	>122.5	0.18
7k	H	5-Nitrofuran-2-yl	NT	12.49
7l	H	2-Hydroxyphenyl	NT	6.5
7m	H	4-Methoxyphenyl	NT	6.32
7n	H	4-Hydroxy-3-methoxyphenyl	>122.2	3.05
7o	H	2-Methylphenyl	NT	13.04
7p	H	4-Methylphenyl	NT	6.53
7q	H	4-Dimethylaminophenyl	122.9	3.07
7r	H	4-Benzyloxyphenyl	NT	21.88
7s	CH3	Phenyl	NT	13.04
7t	CH3	4-Flurophenyl	125.7	1.57
7u	CH3	4-Bromophenyl	111.9	2.79
7v	CH3	2-Hydroxyphenyl	NT	6.32
7w	CH3	4-Hydroxyphenyl	NT	6.32
7x	CH3	4-Methylphenyl	>126.7	3.16
7y	CH3	Phenyl	115.4	0.35
7z	CH3	4-Bromophenyl	100.8	2.52

Table 4.1 c Antimycobacterial Activities and Cytotoxicities of Phthalazinyl Hydrazones

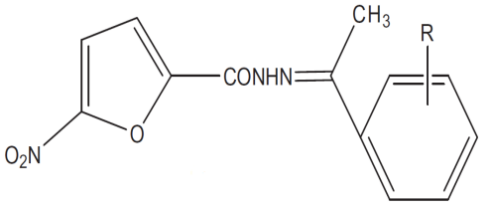
Core Group for the fourth set of structures:





4n-o

4a-m



4p-u

Fig 4.1 d) constitutes the core structure

Comp.No.	R
4a	H
4b	4-NO2
4c	4-Cl
4d	4-Br
4e	4-F
4f	4-OH
4g	4-CH3
4h	4-OCH3
4i	3-Br
4j	4-F
4k	2-CF3
4l	4-OH, 3-OH3
4m	2,6-(Cl2)
4n	H
4o	5-NO2
4p	H
4q	4-NO2
4r	4-Br
4s	4-F
4t	4-CH3
4u	4-OH

Table 4.1 d Physical constants and in vitro antimycobacterial and cytotoxicity of the compounds

Some of the 2D and 3D structures of other inhibitors are shown below respectively.

- 3-Bromopyruvate [11, 16]

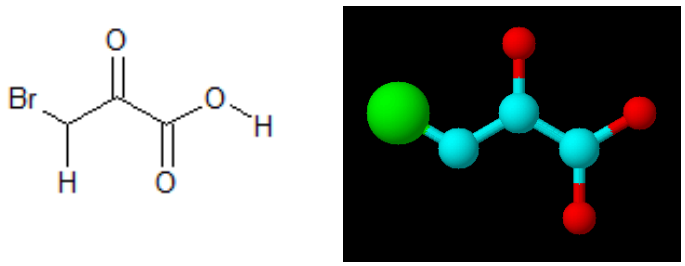


Fig 4.1 e

- 3-Nitropropionate [11, 15]

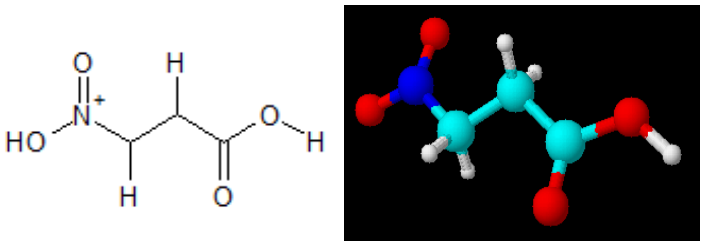


Fig 4.1 f

- 5-Nitro-2,6-dioxohexahydro-4pyrimidinecarboxamides[38]

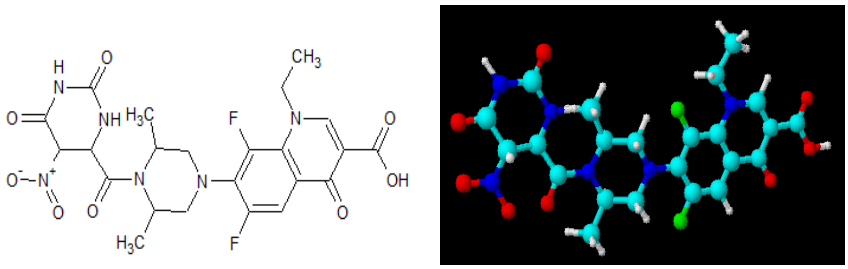


Fig 4.1 g

- Tris-aromaticfuranones[59]

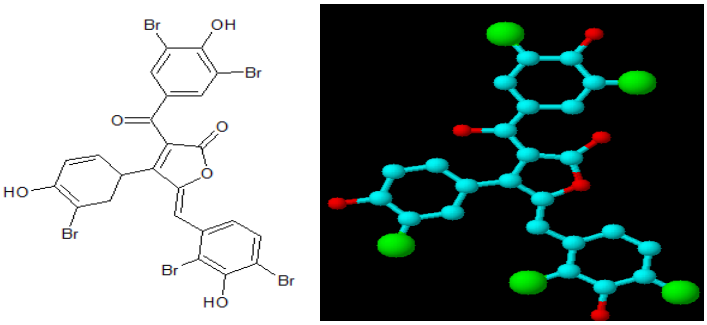


Fig 4.1 h

- Salicylanilidederivatives[67]

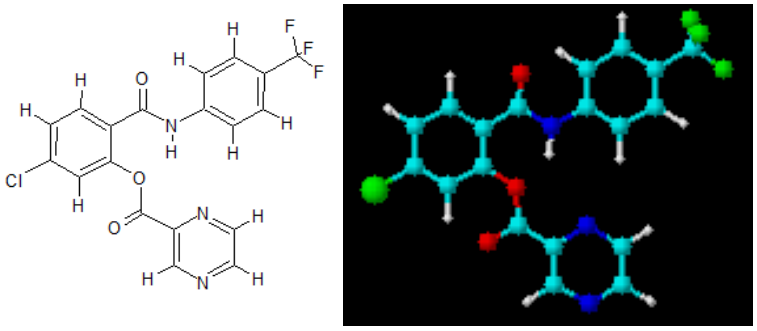


Fig 4.1 i

- Ydcm67

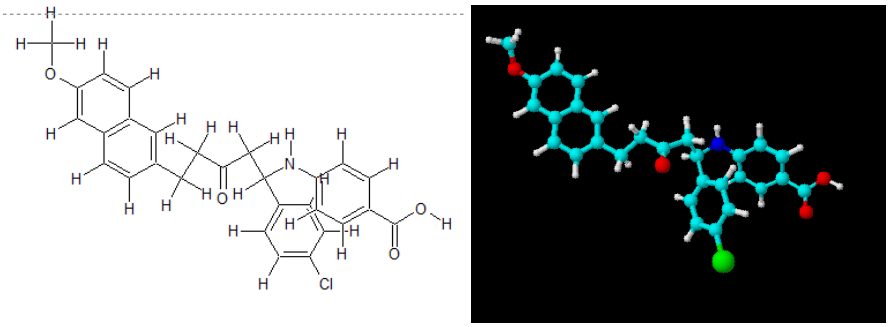


Fig 4.1 j

- 2-(3-Methylbut-2-enyl) benzene-1,4-diol

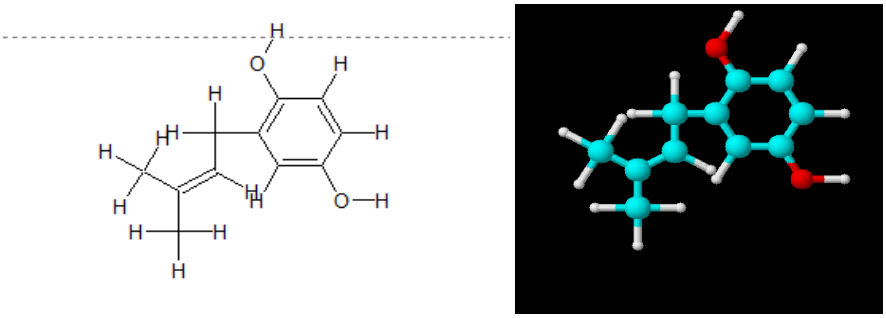


Fig 4.1 k

- (E)-2-(4-Hydroxy-3-methylbut-2-enyl)benzene

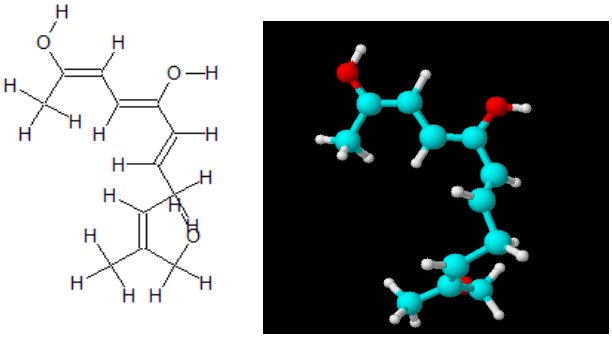


Fig 4.1

(E)-2-(3-Methylnon-2-enyl) benzene-1,4-diol

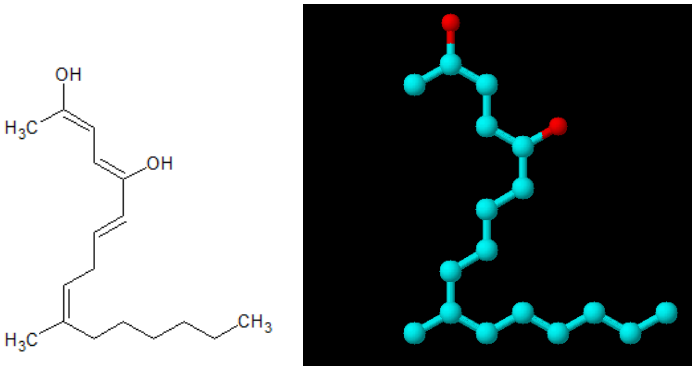


Fig 4.1 m

- (E)-2-(9-Hydroxy-3-methylnon-2-enyl)benzene-1,4-diol

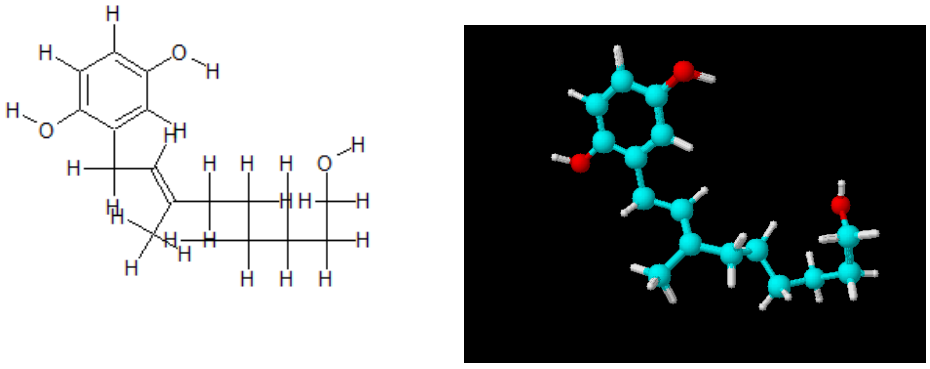


Fig 4.1 n

Fig 4.1e – 4.1 n In the above figures, structures are displayed in 3D viewer in Balls and Sticks format.

4.2 Conversion of file format

Webqc:

- Mol2 was copied and pasted in the input area and the output is shown below.

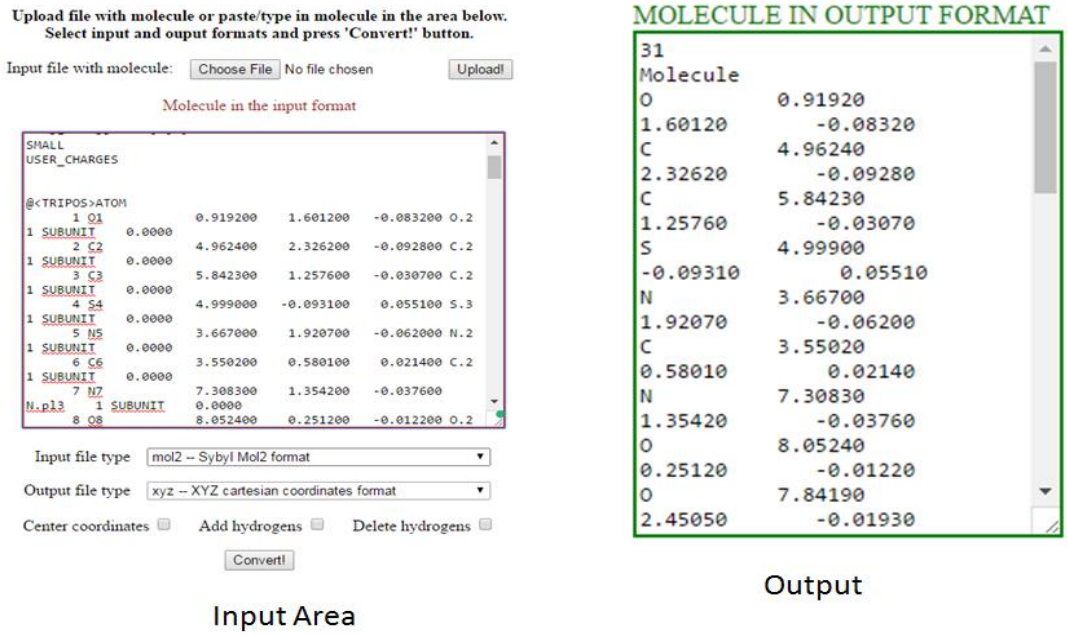


Fig 4.o Example of input and output data in Webqc

4.3 Geometry Optimization:

MOPAC

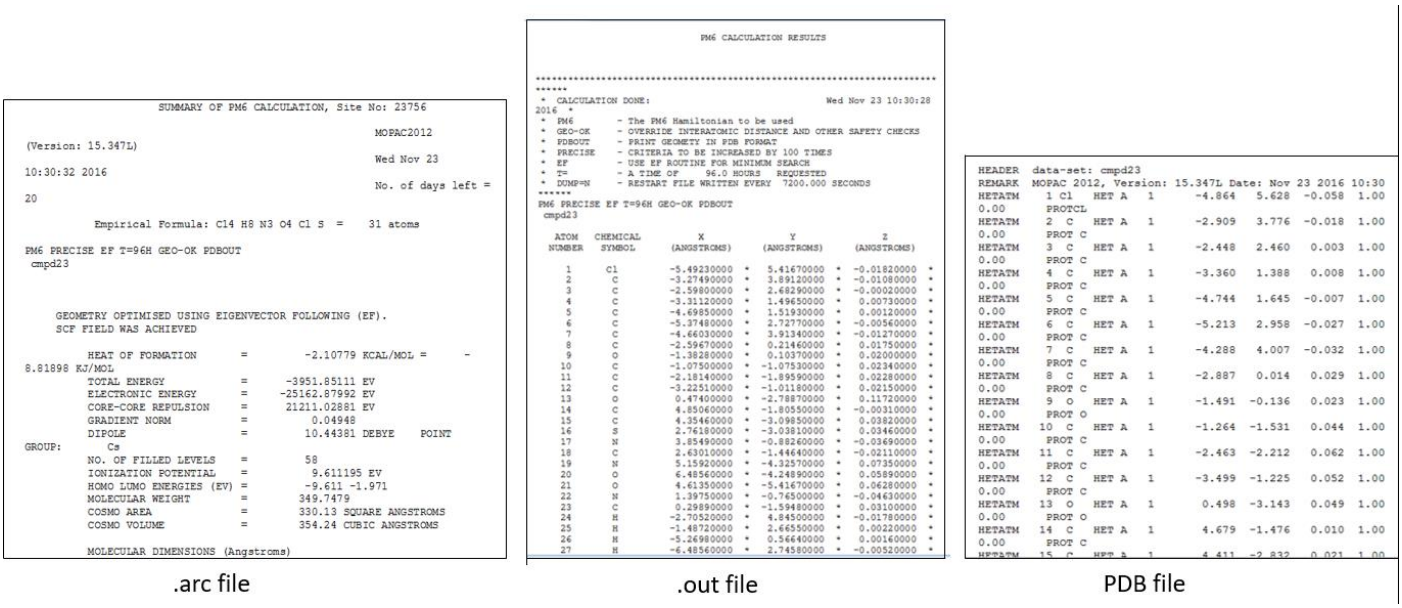
- mop files appeared like this after editing of the xyz files.

```
PM6 PRECISE EF T=96H GEO-OK PDBOUT
cmpd23

Cl          -5.49230         5.41670        -0.01820
C           -3.27490         3.89120        -0.01080
C           -2.59800         2.68290        -0.00020
C           -3.31120         1.49650         0.00730
C           -4.69850         1.51930         0.00120
C           -5.37480         2.72770        -0.00560
C           -4.66030         3.91340        -0.01270
C           -2.59670         0.21460         0.01750
O           -1.38280         0.10370         0.02000
C           -1.07500        -1.07530         0.02340
C           -2.18140        -1.89590         0.02280
C           -3.22510        -1.01180         0.02150
O            0.47400        -2.78870         0.11720
C            4.85060        -1.80550        -0.00310
C            4.35460        -3.09850         0.03820
S            2.76180        -3.03810         0.03460
N            3.85490        -0.88260        -0.03690
C            2.63010        -1.44640        -0.02110
N            5.15920        -4.32570         0.07350
O            6.48560        -4.24890         0.05890
O            4.61350        -5.41670         0.06280
N            1.39750        -0.76500        -0.04630
C            0.29890        -1.59480         0.03100
H           -2.70520         4.84500        -0.01780
H           -1.48720         2.66550         0.00220
H           -5.26980         0.56640         0.00160
H           -6.48560         2.74580        -0.00520
```

Figure 4.3 a.mop file

- Three files were produced of every compound after running mop command on Linux.



```
                    SUMMARY OF PM6 CALCULATION, Site No: 23756

                                                          MOPAC2012
(Version: 15.347L)
                                                          Wed Nov 23
10:30:32 2016
                                                          No. of days left =
20

          Empirical Formula: C14 H8 N3 O4 Cl S  =    31 atoms

PM6 PRECISE EF T=96H GEO-OK PDBOUT
 cmpd23



     GEOMETRY OPTIMISED USING EIGENVECTOR FOLLOWING (EF).
     SCF FIELD WAS ACHIEVED

          HEAT OF FORMATION       =         -2.10779 KCAL/MOL =       -
8.81898 KJ/MOL
          TOTAL ENERGY            =     -3951.85111 EV
          ELECTRONIC ENERGY       =    -25162.87992 EV
          CORE-CORE REPULSION     =     21211.02881 EV
          GRADIENT NORM           =         0.04948
          DIPOLE                  =        10.44381 DEBYE     POINT
GROUP:      Cs
          NO. OF FILLED LEVELS    =        58
          IONIZATION POTENTIAL    =         9.611195 EV
          HOMO LUMO ENERGIES (EV) =        -9.611 -1.971
          MOLECULAR WEIGHT        =       349.7479
          COSMO AREA              =       330.13 SQUARE ANGSTROMS
          COSMO VOLUME            =       354.24 CUBIC ANGSTROMS

          MOLECULAR DIMENSIONS (Angstroms)
```

.arc file

```
                    PM6 CALCULATION RESULTS

*************************************************************************
******
 * CALCULATION DONE:                          Wed Nov 23 10:30:28
2016 *
 *  PM6      - The PM6 Hamiltonian to be used
 *  GEO-OK   - OVERRIDE INTERATOMIC DISTANCE AND OTHER SAFETY CHECKS
 *  PDBOUT   - PRINT GEOMETY IN PDB FORMAT
 *  PRECISE  - CRITERIA TO BE INCREASED BY 100 TIMES
 *  EF       - USE EF ROUTINE FOR MINIMUM SEARCH
 *  T=       - A TIME OF     96.0 HOURS  REQUESTED
 *  DUMP=H   - RESTART FILE WRITTEN EVERY  7200.000 SECONDS
******
PM6 PRECISE EF T=96H GEO-OK PDBOUT
 cmpd23

  ATOM   CHEMICAL          X               Y               Z
 NUMBER   SYMBOL      (ANGSTROMS)     (ANGSTROMS)     (ANGSTROMS)

     1      Cl       -5.49230000  *   5.41670000  *  -0.01820000  *
     2       C       -3.27490000  *   3.89120000  *  -0.01080000  *
     3       C       -2.59800000  *   2.68290000  *  -0.00020000  *
     4       C       -3.31120000  *   1.49650000  *   0.00730000  *
     5       C       -4.69850000  *   1.51930000  *   0.00120000  *
     6       C       -5.37480000  *   2.72770000  *  -0.00560000  *
     7       C       -4.66030000  *   3.91340000  *  -0.01270000  *
     8       C       -2.59670000  *   0.21460000  *   0.01750000  *
     9       O       -1.38280000  *   0.10370000  *   0.02000000  *
    10       C       -1.07500000  *  -1.07530000  *   0.02340000  *
    11       C       -2.18140000  *  -1.89590000  *   0.02280000  *
    12       C       -3.22510000  *  -1.01180000  *   0.02150000  *
    13       O        0.47400000  *  -2.78870000  *   0.11720000  *
    14       C        4.85060000  *  -1.80550000  *  -0.00310000  *
    15       C        4.35460000  *  -3.09850000  *   0.03820000  *
    16       S        2.76180000  *  -3.03810000  *   0.03460000  *
    17       N        3.85490000  *  -0.88260000  *  -0.03690000  *
    18       C        2.63010000  *  -1.44640000  *  -0.02110000  *
    19       N        5.15920000  *  -4.32570000  *   0.07350000  *
    20       O        6.48560000  *  -4.24890000  *   0.05890000  *
    21       O        4.61350000  *  -5.41670000  *   0.06280000  *
    22       N        1.39750000  *  -0.76500000  *  -0.04630000  *
    23       C        0.29890000  *  -1.59480000  *   0.03100000  *
    24       H       -2.70520000  *   4.84500000  *  -0.01780000  *
    25       H       -1.48720000  *   2.66550000  *   0.00220000  *
    26       H       -5.26980000  *   0.56640000  *   0.00160000  *
    27       H       -6.48560000  *   2.74580000  *  -0.00520000  *
```

.out file

```
HEADER  data-set: cmpd23
REMARK  MOPAC 2012, Version: 15.347L Date: Nov 23 2016 10:30
HETATM    1 Cl   HET A   1      -4.864   5.628  -0.058  1.00
0.00      PROTCL
HETATM    2  C   HET A   1      -2.909   3.776  -0.018  1.00
0.00      PROT C
HETATM    3  C   HET A   1      -2.448   2.460   0.003  1.00
0.00      PROT C
HETATM    4  C   HET A   1      -3.360   1.388   0.008  1.00
0.00      PROT C
HETATM    5  C   HET A   1      -4.744   1.645  -0.007  1.00
0.00      PROT C
HETATM    6  C   HET A   1      -5.213   2.958  -0.027  1.00
0.00      PROT C
HETATM    7  C   HET A   1      -4.288   4.007  -0.032  1.00
0.00      PROT C
HETATM    8  C   HET A   1      -2.887   0.014   0.029  1.00
0.00      PROT C
HETATM    9  O   HET A   1      -1.491  -0.136   0.023  1.00
0.00      PROT O
HETATM   10  C   HET A   1      -1.264  -1.531   0.044  1.00
0.00      PROT C
HETATM   11  C   HET A   1      -2.463  -2.212   0.062  1.00
0.00      PROT C
HETATM   12  C   HET A   1      -3.499  -1.225   0.052  1.00
0.00      PROT C
HETATM   13  O   HET A   1       0.498  -3.143   0.049  1.00
0.00      PROT O
HETATM   14  C   HET A   1       4.679  -1.476   0.010  1.00
0.00      PROT C
HETATM   15  C   HET A   1       4.411  -2.832   0.021  1.00
```

PDB file

Figure 4.3b).arc .out and pdb files are shown of the above compound with the molecular formulaC14H8ClN3O4

4.4Docking

Scoring

ligand	GScore	DockScore	HBond	Electro	Sitemap	Activity
cmpd1	-5.57	-5.56	-1.63	-0.25	0	-5.57
cmpd2	-5.66	-5.46	-1.7	-0.43	-0.23	-5.66
cmpd3	-6.38	-6.35	0	-0.13	-0.7	-6.38
cmpd4	-5.61	-5.58	-0.96	-0.28	-0.17	-5.61
cmpd5	-5.19	-5.12	-0.62	-0.21	-0.16	-5.19
cmpd6	-6.02	-5.94	0	-0.09	-0.39	-6.02
cmpd7	-6.07	-5.96	-1.41	-0.22	0	-6.07
cmpd8	-5.76	-5.51	0	-0.2	-0.56	-5.76
cmpd9	-5.94	-5.65	0	-0.04	-0.22	-5.94
cmpd10	-5.5	-5.49	-0.96	-0.33	0	-5.5
cmpd11	-5.46	-5.44	-1.24	-0.19	0	-5.46
cmpd12	-6.07	-6.04	-0.46	-0.04	0	-6.07
cmpd13	-5.47	-4.84	-0.74	-0.23	-0.27	-5.47
cmpd14	-4.86	-4.75	-0.65	-0.2	-0.12	-4.86
cmpd15	-6.45	-5.27	-0.7	-0.16	-0.17	-6.45
cmpd16	-6.34	-6.34	-0.96	-0.36	-0.23	-6.34
cmpd17	-5.47	-5.46	-1.35	-0.5	-0.11	-5.47

cmpd18	-5.91	-5.9	0	-0.14	-0.21	-5.91
cmpd19	-4.36	-4.11	-0.55	-0.17	-0.34	-4.36
cmpd20	-4.61	-2.99	-0.7	-0.3	0	-4.61
cmpd21	-3.72	-3.7	-0.34	-0.12	0	-3.72
cmpd22	-4.44	-4.03	-0.63	-0.19	0	-4.44
cmpd23	-6.2	-5.88	-1.28	-0.36	0	-6.2
cmpd24	-6.43	-6.11	-1.33	-0.44	0	-6.43
cmpd25	-5.97	-5.53	-0.7	-0.32	0	-5.97
cmpd26	-5.32	-5.11	-0.68	-0.17	0	-5.32
cmpd27	-5.63	-5.36	0	-0.03	-0.21	-5.63
ydcm	-8.16	-8.16	-0.37	-0.07	0	-8.16
cmpdiv	-4.69	-4.69	-1.08	-0.29	0	-4.69
cmpdv	-4.67	-4.67	-1.17	-0.47	0	-4.67
cmpdvi	-5.56	-5.55	-1.19	-0.41	0	-5.56
cmpdvii	-6.28	-6.28	-1.33	-0.74	0	-6.28

Table 4.4 Scoring of docked ligands

Ligand Interaction Diagrams:

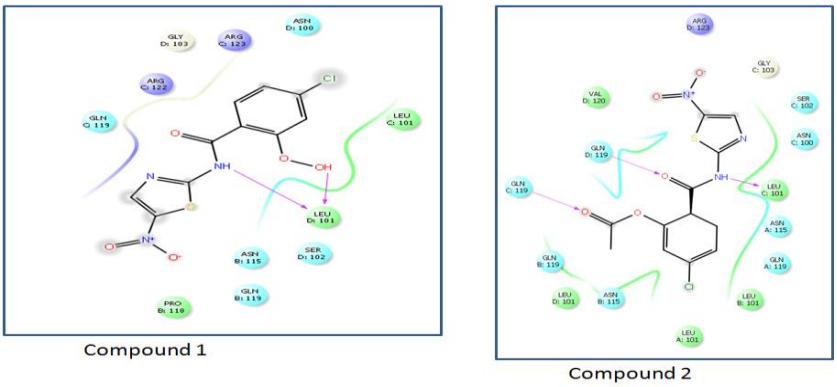


Fig 4.4 a

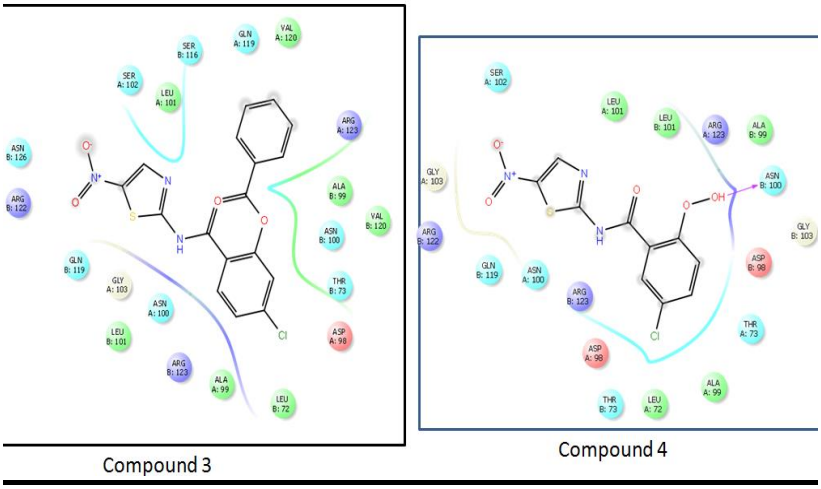
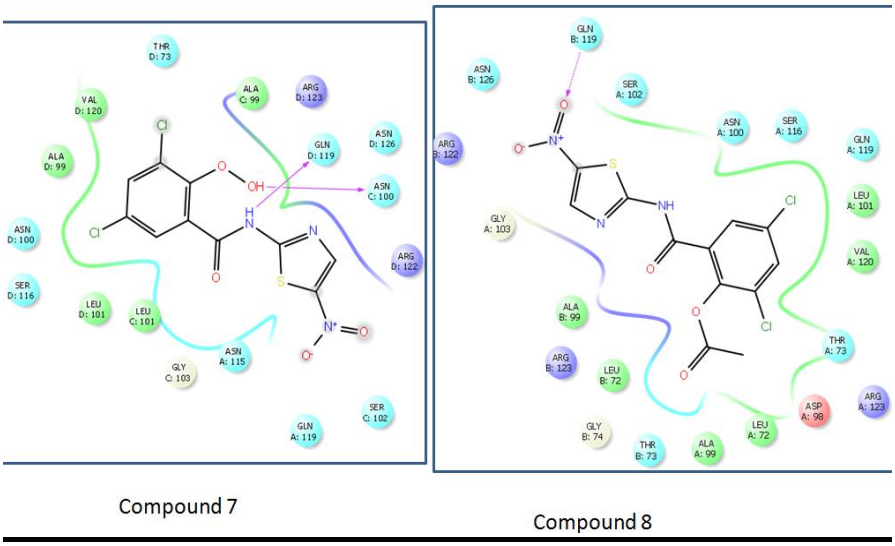
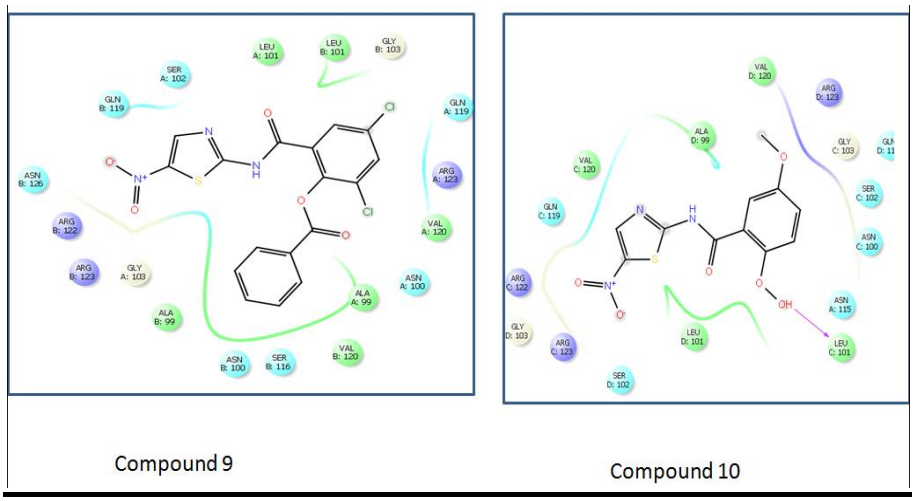


Fig 4.4 b



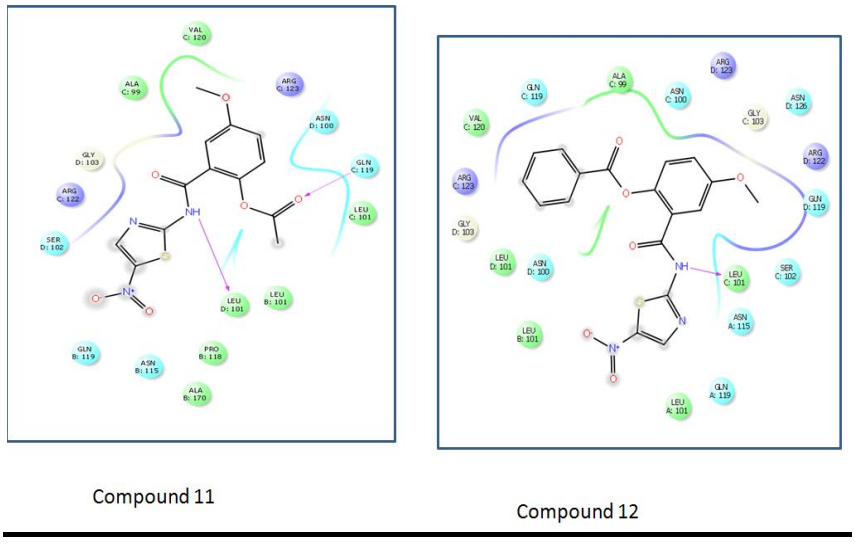
Compound 7 Compound 8

Fig 4.4 c



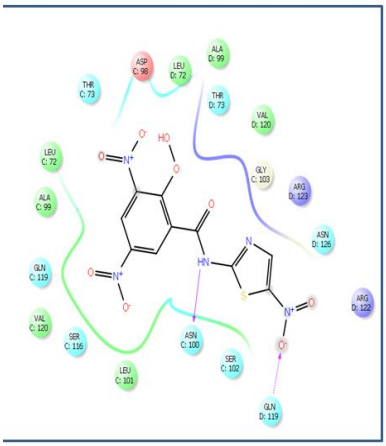
Compound 9 Compound 10

Fig 4.4 d

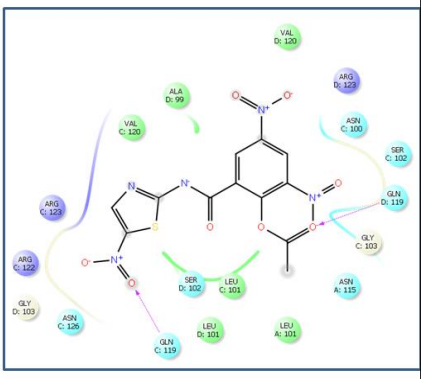


Compound 11 Compound 12

Fig 4.4 e

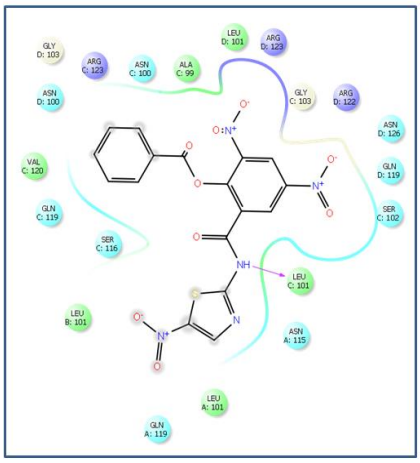


Compound 13

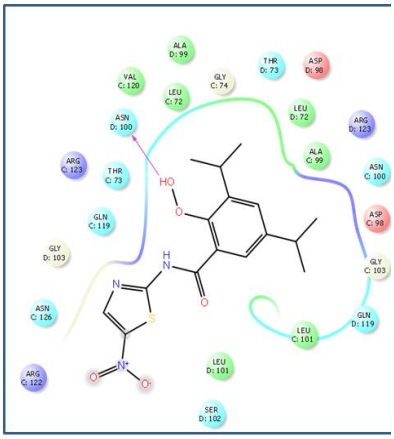


Compound 14

Fig 4.4 f

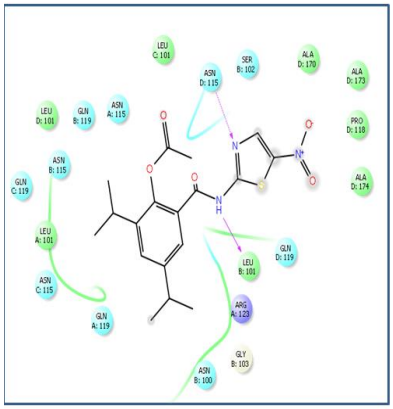


Compound 15

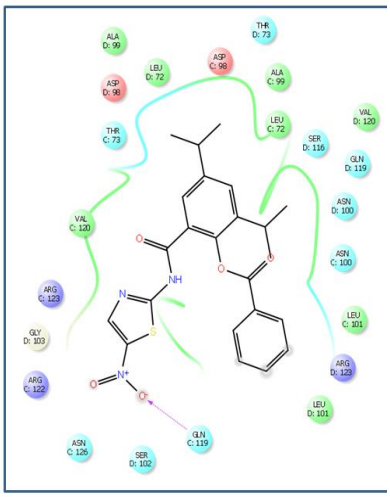


Compound 16

Fig 4.4g



Compound 17



Compound 18

Fig 4.4 h

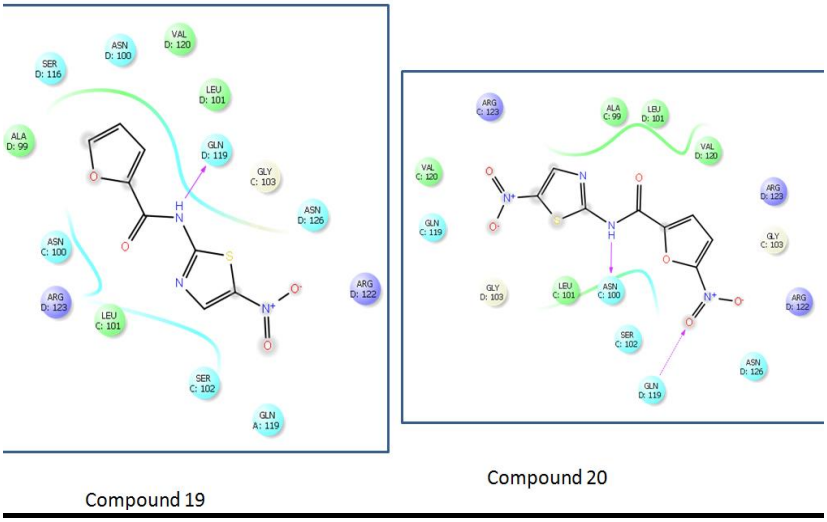


Fig 4.4 i

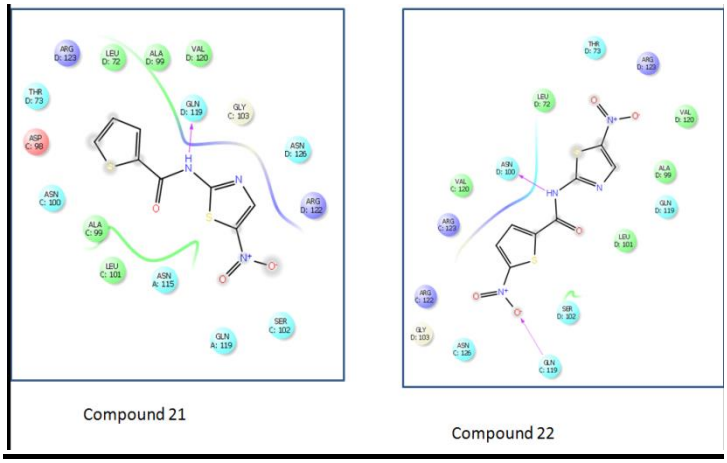


Fig 4.4 j

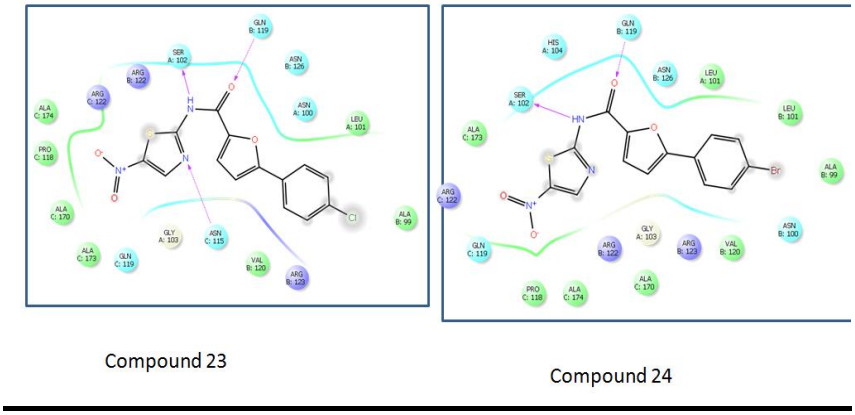
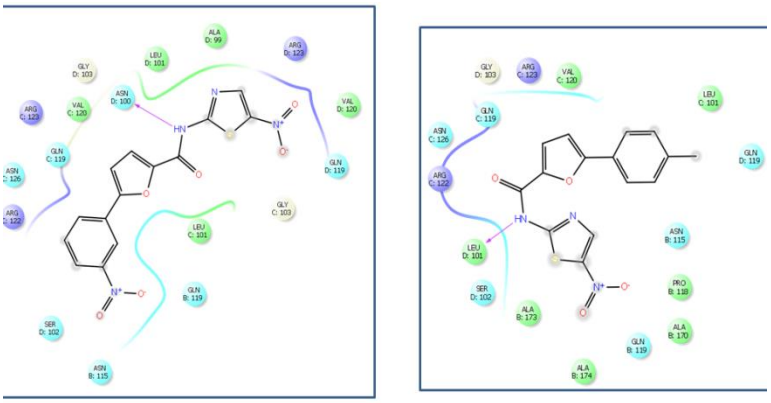
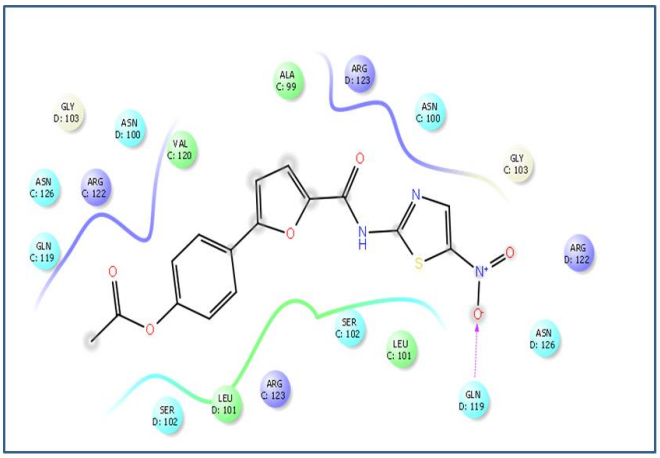


Fig 4.4k



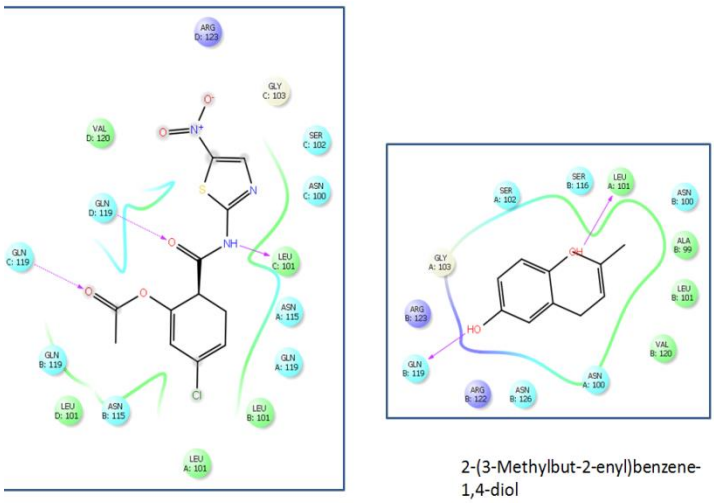
Compound 25 Compound 26

Fig 4.4 l



YDCM

Fig 4.4m



2-Allylbenzene-1,4-diol 2-(3-Methylbut-2-enyl)benzene-1,4-diol

Fig 4.4 n

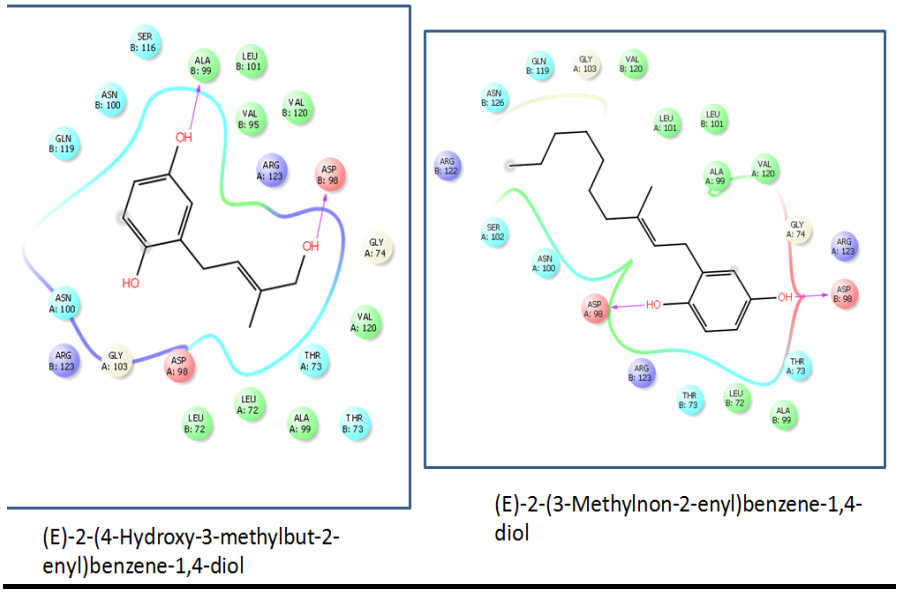


Fig 4.4 o

Fig 4.4a-4.4o Ligand interaction diagrams of docked ligands

Significance of signs, symbols and colors:

4.5 Feature Extraction:

Dragon7

Following are the molecular descriptors generated.

Block ID	Block Descriptor	Desc. No.
1	Constitutional descriptors	47
2	Ring descriptors	32
3	Topological indices	75
4	Walk and path counts	46
5	Connectivity indices	37
6	Information indices	48
7	2D matrix-based descriptors	550
8	2D autocorrelations	213
9	Burden eigenvalues	96
10	P_VSA-like descriptors	45

11	ETA indices	23
12	Edge adjacency indices	324
13	Geometrical descriptors	38
14	3D matrix-based descriptors	90
15	3D autocorrelations	80
16	RDF descriptors	210
17	3D-MoRSE descriptors	224
18	WHIM descriptors	114
19	GETAWAY descriptors	273
20	Randic molecular profiles	41
21	Functional group counts	154
22	Atom-centred fragments	115
23	Atom-type E-state indices	170
24	CATS 2D	150
25	2D Atom Pairs	1596
26	3D Atom Pairs	36
27	Charge descriptors	15
28	Molecular properties	20
29	Drug-like indices	27
30	CATS 3D	300

Table 4.5 a: Descriptor blocks with the number of descriptors calculated for each block

Statistical analysis

1. By performing statistical analysis through Rstudio there were thirty blocks of descriptors.
2. Then all the blocks (>5000) were grouped in 8 sub files, as shown below in the table

Positive Very High Correlation, High Variance
Positive Very High Correlation, Low Variance
Positive High Correlation, High Variance
Positive High Correlation, Low Variance
Positive Moderate Correlation, High Variance
Positive Moderate Correlation, Low Variance
Positive No Correlation, High Variance
Positive No Correlation, Low Variance

For e.g.: This is the first block (i.e. Constitutional indices)

Drg1

>t(pvhchv)#Positive Very High Correlation, High Variance

12 **13** **4**

>t(pvhclv)#Positive Very High Correlation, Low Variance

15 16 **6**

>t(phchv)#Positive High Correlation, High Variance

3 39 22 18 26 32 7

>t(phclv)#Positive High Correlation, Low Variance

5 **24** 17 1 36 37 2

>t(pmchv)#Positive Moderate Correlation, High Variance

9 20 19 29 10 14 **38** 25

>t(pmclv)#Positive Moderate Correlation, Low Variance

30 35 **23** 34 8 31 **33** 21

>t(zchv)#Positive No Correlation, High Variance

11

>t(zclv)#Positive No Correlation, Low Variance

28

3. Classification was done on the basis of 1D, 2D, 3D.
4. Combine all the pvhchv, pvhclv, phclv, phchv on the basis of classification of descriptor block into an excel sheet.

NAME	x1	x2	x3	x4	x5	x6	x7	x8	x9	log of activity
4g	0	0	1	8.368	1	1	1	-2.242	-0.651	5.432585987
4j	0	0	0	5.976	1	1	1	-2.803	-0.817	5.325786609
4k	0	0	0	6.593	1	1	1	-2.27	-0.932	5.252325783
4o	0	0	0	4.729	1	1	1	-3.921	-0.853	5.359036102
4p	1	3	1	8.922	2	0	0	-3.317	-0.034	5.432585987
4q	1	3	1	7.986	2	0	0	-4.044	-0.353	5.280764207
4r	1	3	1	20.635	2	0	0	-3.567	-0.059	5.178970609
4s	1	3	1	8.728	2	0	0	-3.567	-0.269	5.368962206
4t	1	3	2	9.868	2	0	0	-3.567	-0.029	5.382520638
7b	1	2	0	18.326	1	1	1	-1.654	-0.477	4.862135286
7c	1	2	0	18.304	1	1	1	-1.783	-0.511	4.862135286
7d	1	2	0	18.597	1	1	1	-2.617	-0.89	4.763881877
7e	1	2	0	15.858	1	1	1	-2.196	-0.484	4.762173935
7f	1	2	0	18.522	1	1	1	-1.783	-0.346	4.744062185
7g	1	2	0	15.129	1	1	1	-1.654	-0.342	4.744062185
7h	1	2	0	18.468	1	1	1	-2.285	-0.657	4.80811103

7i	1	2	0	17.748	1	1	1	-1.943	-0.596	4.80811103
7j	1	2	0	18.769	1	1	1	-1.652	-0.551	4.80811103
7n	1	2	0	24.389	2	1	1	-1.959	-0.479	4.805659047
7q	1	2	0	23.131	1	1	1	-1.652	-0.337	4.811371017
7t	2	5	1	24.526	2	0	0	-2.036	0.035	4.833898116
7u	2	5	1	35.171	2	0	0	-2.036	0.316	4.717605615
7x	2	5	2	26.186	2	0	0	-2.036	0.356	4.841822087
7y	2	2	0	30.992	3	0	0	-4.495	0.256	4.748404354
7z	2	2	0	28.154	3	0	0	-4.742	0.222	4.613138356

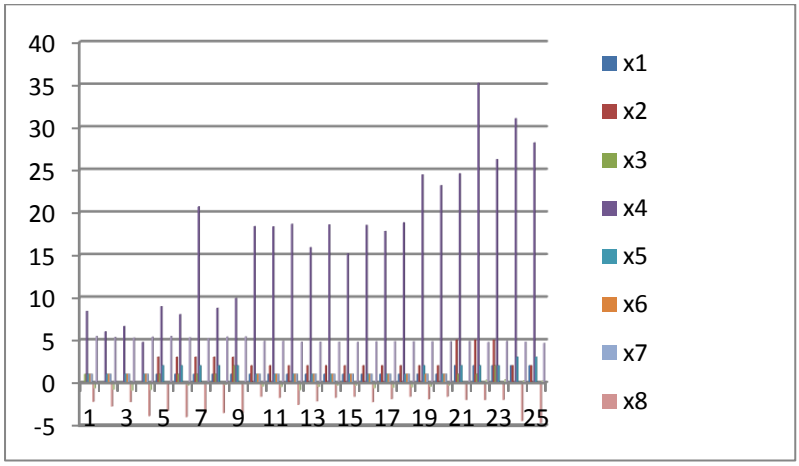
Table 4.5 b Classified descriptors

5. An excel sheet with final four descriptor values for all the molecules with the preceding column of activity data (IC_{50} μm) for 47 molecules.
6. Calculated log of activity for all performing data analysis in excel regression and coefficient taken x as log of activity and y as all descriptors.

NAME	**IC_{50}(μm)**	**x1**	**x2**	**x3**	**x4**	**log of activity**
4g	>228.74	1	8.368	1	-2.242	5.43258599
4j	205.57	0	5.976	1	-2.803	5.32578661
4k	>191.01	0	6.593	1	-2.27	5.25232578
4o	212.52	0	4.729	1	-3.921	5.3590361
4p	>228.74	1	8.922	2	-3.317	5.43258599
4q	>196.52	1	7.986	2	-4.044	5.28076421
4r	>177.50	1	20.635	2	-3.567	5.17897061
4s	214.64	1	8.728	2	-3.567	5.36896221
4t	>217.57	2	9.868	2	-3.567	5.38252064
7b	> 129.3	0	18.326	1	-1.654	4.86213529
7c	> 129.3	0	18.304	1	-1.783	4.86213529
7d	117.2	0	18.597	1	-2.617	4.76388188
7e	> 117.0	0	15.858	1	-2.196	4.76217394
7f	>114.9	0	18.522	1	-1.783	4.74406219
7g	>114.9	0	15.129	1	-1.654	4.74406219
7h	>122.5	0	18.468	1	-2.285	4.80811103
7i	>122.5	0	17.748	1	-1.943	4.80811103
7j	>122.5	0	18.769	1	-1.652	4.80811103
7n	> 122.2	0	24.389	2	-1.959	4.80565905

7q	122.9	0	23.131	1	-1.652	4.81137102
7t	125.7	1	24.526	2	-2.036	4.83389812
7u	111.9	1	35.171	2	-2.036	4.71760562
7x	>126.7	2	26.186	2	-2.036	4.84182209
7y	115.4	0	30.992	3	-4.495	4.74840435
7z	100.8	0	28.154	3	-4.742	4.61313836

Table 4.5 c: Grouped final four descriptors



Regression Statistics	
Multiple R	0.983144
R Square	0.966573
Adjusted R Square	0.88736
Standard Error	0.062481
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	9	1.806129	0.200681	57.83188	5.07E-10
Residual	16	0.062461	0.003904		
Total	25	1.86859			

Intercept	***Coefficients***	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***	***Lower 95.0%***	***Upper 95.0%***
	5.78912513	0.24848035	23.2981207	8.9865E-14	5.26237033	6.31587994	5.26237033	6.31587994
x1	-0.4515987	0.09651036	-4.679277	0.0002513	-0.6561915	-0.2470059	-0.6561915	-0.2470059
x2	-0.00308	0.04177646	-0.0737251	0.94214296	-0.0916421	0.08548217	-0.0916421	0.08548217
x3	0.07133866	0.04539168	1.57162423	0.13560118	-0.0248874	0.16756473	-0.0248874	0.16756473
x4	-0.007824	0.00417978	-1.8718597	0.0796226	-0.0166847	0.00103678	-0.0166847	0.00103678
x5	0.05590036	0.06938087	0.80570278	0.43222619	-0.0911805	0.20298125	-0.0911805	0.20298125

x6	0	0	65535	#NUM!	0	0	0	0
x7	-0.3873374	0.15084893	-2.5677175	0.02064952	-0.7071229	-0.067552	-0.7071229	-0.067552
x8	0.02947373	0.03918681	0.75213381	0.46289268	-0.0535986	0.11254605	-0.0535986	0.11254605
x9	0.00336768	0.11713743	0.02874986	0.97741965	-0.2449526	0.25168793	-0.2449526	0.25168793

Table 4.5 c Regression analysis

7. Further there was an analysis for individual descriptors, in the same way through Rsudio.

log of activity	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
5.43258599	1	2.8	4.5	0.192	5.622	5.32	6.7	-1.32	0.17	0.17
5.32578661	0	2.6	2.9	0	5.76	3.21	6.2	-2.09	0.17	0.62
5.25232578	1	3.1	3.1	0	6.282	4.94	6.6	-2.87	0.17	0.68
5.3590361	0	2.1	1.9	0	4.819	3.28	5.3	-2.61	0.17	0.68
5.43258599	1	3.2	4.5	0.175	4.272	3.21	4.7	-0.68	0.17	0.21
5.28076421	1	3.2	4.1	0.167	4.258	2.12	4.7	-0.44	0.17	0.2
5.17897061	1	4.1	4.2	0.125	4.443	2.66	4.9	-0.85	0.17	0.21
5.36896221	1	3.3	4.1	0.081	4.857	1.75	5.4	-0.63	0.17	0.2
5.38252064	2	3.6	5.6	0.218	5.622	5.32	6.7	-0.21	0.16	0.16
4.86213529	2	8.3	7.4	0.315	11.261	2.22	15	-0.78	0.19	0.37
4.86213529	2	8.4	7.4	0.353	11.352	8.36	13	-2.21	0.16	0.74
4.76388188	3	9	7.5	0.371	13.548	6.43	18	-2.58	0.17	0.32
4.76217394	2	8.5	6.8	0.362	9.59	2.83	13	-0.87	0.16	0.54
4.74406219	2	8.7	7.3	0.36	12.061	9.45	14	-2.26	0.16	0.66
4.74406219	2	8.4	7.3	0.345	9.778	3.13	13	-0.94	0.16	0.41
4.80811103	2	8.5	7.3	0.367	17.938	4.68	21	-1.7	0.16	0.56
4.80811103	2	8.2	7.3	0.366	9.899	6.31	12	-1.25	0.15	0.53
4.80811103	2	8.5	7.3	0.364	6.579	5.37	7.5	-0.89	0.16	0.35
4.80565905	3	9.1	8.4	0.577	19.935	7.84	24	-0.19	0.15	0.42
4.81137102	4	9.1	9.9	0.68	19.189	13.2	24	1.218	0.15	0.48
4.83389812	3	9.7	8.6	0.679	16.152	2.68	19	-0.33	0.15	0.52
4.71760562	3	11	8.6	0.679	15.478	2.5	18	-0.56	0.16	0.57
4.84182209	4	9.9	10	0.679	18.093	14.2	23	-0.25	0.16	0.25
4.74840435	2	13	9.7	0.68	22.538	16.6	28	0.116	0.15	0.25
4.61313836	2	13	9.2	0.68	19.691	16.6	24	-0.18	0.15	0.3

Table 4.5 d Individual descriptors

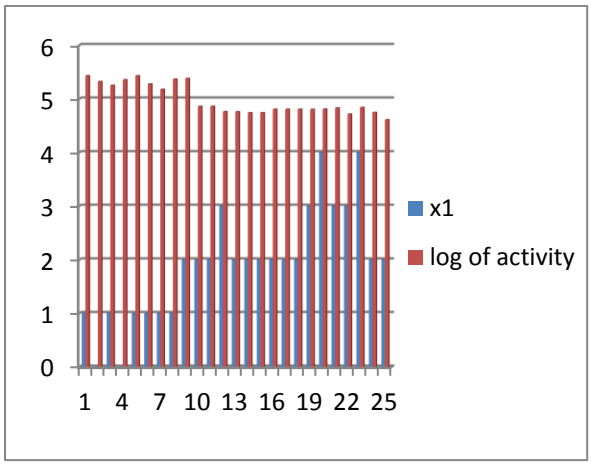


Fig 4.5 a

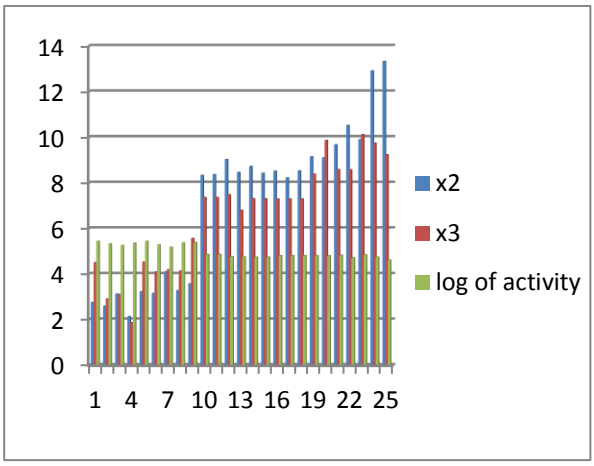


Fig 4.5 b

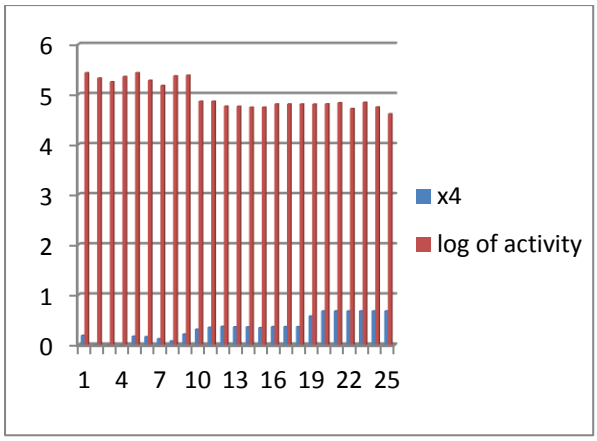


Fig 4.5 c

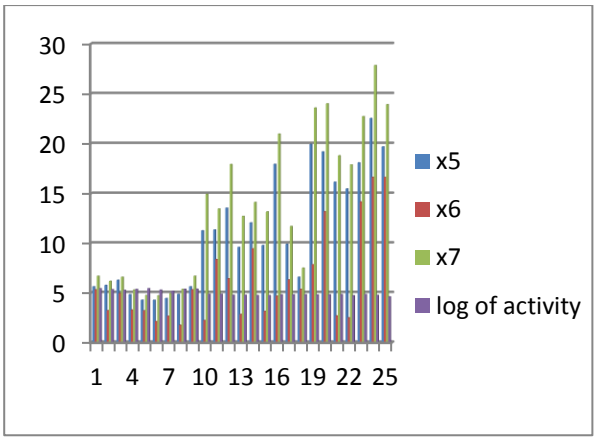


Fig 4.5 d

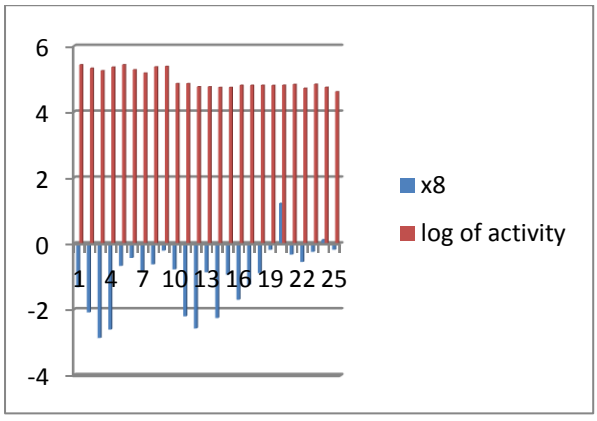


Fig 4.5 e

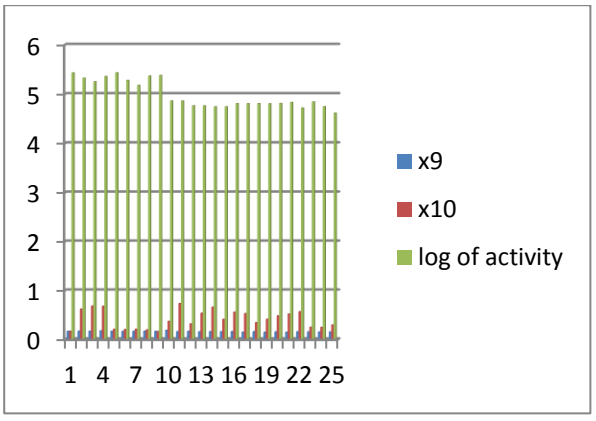


Fig 4.5 f

Figure: 4.5 a- 4.5 f Graph between activity and descriptors obtained from the data analysis

Compound Name	Number of Descriptors	Descriptors after	R^2
Constitutional indices	1 (nCsp3)	1 (nCsp3)	0.502258
Information indices	2 (ATSC7p, ATSC7m)	1 (ATSC7p)	0.899815
Burden eigenvalues	1 (SpMin6_Bh(s))	1 (SpMin6_Bh(s))	0.634452
RDF descriptors	3(RDF100e, RDF125i, RDF105i)	1 (RDF105i)	0.664638
3D-MoRSE descriptors	1(Mor09i)	1(Mor09i)	0.042647
WHIM descriptors	2(G1m, E3S)	1 (G1m)	0.36461

Table 4.5 e: R^2 calculated from regression

Weka

- Out of all 4 folds cross-validation gave good Q^2.

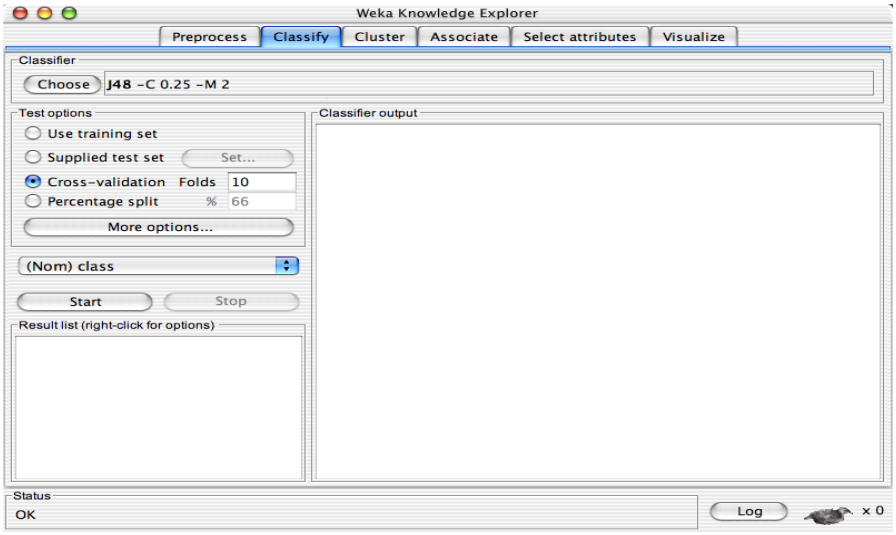


Figure 4.5 g Window of Weka Knowledge Explorer

Compound Name	**Number of Descriptors**	**Descriptors after**	**Q^2**	**R^2**	**Equation**
Constitutional indices	1 (nCsp3)	1 (nCsp3)	0.598	0.502258	-0.19*x1+5.35
Information indices	2 (ATSC7p, ATSC7m)	1 (ATSC7p)	0.9231	0.899815	-0.08*x2+5.54
Burden eigenvalues	1 (SpMin6_Bh(s))	1 (SpMin6_Bh(s))	0.7049	0.634452	-0.95*x4+5.32
RDF descriptors	3(RDF100e, RDF125i, RDF105i)	1 (RDF105i)	0.7336	0.664638	-0.03*x7+5.38
3D-MoRSE descriptors	1(Mor09i)	1(Mor09i)	-0.1562	0.042647	-0.06*x8+4.92
WHIM descriptors	2(G1m, E3S)	1 (G1m)	0.3369	0.36461	17.29*x9+2.19
5250 (grouped)	9	1(ATSC7m)	0.7804	0.966573	-0.03*x4+5.46

Table 4.5 f Equations generated

CHAPTER 5 – CONCLUSION

The aim of our project was to create a QSAR model such that a new lead molecule can also fit into it. We started with collection of data from literature and their optimization through Webqc and MOPAC. Further, the ligands which were high in activity were docked with ICL and showed significant results. On the other side, equations were generated by the calculation of descriptors from Dragon. Feature selection through statistical measures helped in the QSAR model generation.

CHAPTER 6 – REFERENCES

1. Potential Inhibitors for Isocitrate Lyase of Mycobacterium tuberculosis and Non-M. tuberculosis: A Summary Yie-VernLee,1 HabibahA.Wahab,2,3 andYeeSiewChoong1,4
2. L.G.WayneandK.Y.Lin,"GlyoxylatemetabolismandadaptationofMycobacteriumtuberculosistosurviva lunderanaerobic conditions," Infection and Immunity, vol. 37, no. 3, pp. 1042– 1049,1982.
3. M. F. Dunn, J. A. Ram´ırez-Trujillo, and I. Hern´andez-Lucas, "Major roles of isocitrate lyase and malate synthase in bacterial and fungal pathogenesis," Microbiology, vol. 155, no. 10, pp. 3166–3175, 2009.
4. E. J. Mu˜noz-El´ıas and J. D. McKinney, "Mycobacterium tuberculosis isocitrate lyases 1 and 2 are jointly required for in vivo growthandvirulence,"Nature Medicine, vol.11, no.6, pp.638– 644,2005.
5. K.H.Z.Bentrup, A.Miczak, D.L.Swenson, and D.G.Russell, "Characterization of activity and expression of isocitrate lyase in Mycobacterium avium and Mycobacterium tuberculosis," Journal of Bacteriology, vol.181, no.23, pp.7161–7167, 1999.
6. Mauri, Andrea, et al. "Dragon software: An easy approach to molecular descriptor calculations." *Match* 56.2 (2006): 237-248.
7. https://en.wikipedia.org/wiki/Chemsketch
8. Moskovsky, Alexander A., et al. "A new universal Web interface to quantum chemistry programs." *International journal of quantum chemistry* 100.4 (2004): 645-648.