



ISSN NO. 2320-5407

Journal homepage: <http://www.journalijar.com>  
Journal DOI: [10.21474/IJAR01](https://doi.org/10.21474/IJAR01)

INTERNATIONAL JOURNAL  
OF ADVANCED RESEARCH

## RESEARCH ARTICLE

## CHEMOMETRIC AND SIMILARITY BASED ANALYSIS OF DGAT-1 INHIBITORS.

Priya Ojha\*, Pooja Mishra, Seema Kesar, Sneha Singh.

Department of Pharmacy, Banasthali University, Banasthali- 304022, Rajasthan, India.

**Manuscript Info****Manuscript History:**

Received: 12 February 2016  
Final Accepted: 23 March 2016  
Published Online: April 2016

**Key words:**

DGAT-1, QSAR, TSAR, MLR,  
PLS, FFNN, Similarity.

**\*Corresponding Author**

Priya Ojha.

**Abstract**

QSAR (Quantitative structure activity relationship) is a powerful and mathematical technique to set off the correlation in between chemical structure to their biological activity. It was performed on a series of amide-oxadiazole-aniline derivative with activity against DGAT-1 employing various physiochemical parameters like topological, lipophilic and electronic. The best model was generated and shows good correlative and predictive ability with values  $S = 0.33$ ,  $F = 41.91$ ,  $r = 0.94$ ,  $r^2 = 0.88$ ,  $r^2_{(cv)} = 0.84$  was developed using stepwise MLR and a comparable PLS and FFNN model with  $r^2_{(cv)} = 0.89$ ,  $0.88$  and  $0.86$  respectively. After the data reduction, five promising descriptors left were total dipole moment, Log P, VAMP total energy, VAMP LUMO and VAMP HOMO. In addition of QSAR modeling, Lipinski's rule of five was also employed that check the pharmacokinetic profile of the model. The similarity (CARBO and HODGKIN) analysis was also done on the same series which positively support the previous results. The QSAR study reported in the present study provide important structural situation, related to anti-diabetic activity. Present study enlightens the path of determining the potent lead compounds of DGAT-1 antagonist.

Copy Right, IJAR, 2016,. All rights reserved.

**Introduction:-**

Diabetes mellitus is a chronic metabolic disease characterized by hyperglycemia, hyperlipidemia, hyperaminoacidemia, and hypoinsulinaemia.<sup>1</sup> Type II diabetes is a more common form of diabetes constituting 90% of the diabetic population moreover, it is a polygenic disease that results from a complex interplay between genetic predisposition and environmental factors such as diet, degree of physical activity and age.<sup>2</sup> Triacylglycerol (TG) is a highly efficient energy storage form critical for surviving periods of starvation and extended physical activity.<sup>3</sup> Diacylglycerolacyltransferase (DGAT) enzymes catalyze the formation of an ester linkage between a fatty acyl-CoA and the free hydroxyl group of diacylglycerol this action take place in two pathway Glycerol Phosphate and monoacylglycerol. DGAT possesses two isoforms DGAT-1 and DGAT-2. DGAT-1 catalyses the last step of triacylglyceride biosynthesis, transforming diacylglycerol and acyl-CoA into triglyceides.<sup>4, 5</sup> Inhibiting of DGAT-1 might represent a novel approach for that improvement of insulin sensitivity.<sup>6</sup>

There are numerous examples in the literature for the successful use of classical descriptors in QSAR.<sup>7,8</sup> In the view of this, we decided to developed models from classical QSAR descriptor using MLR, PLS and FFNN method to establish the individual and common structural requirement for effective binding of DGAT-1 antagonist.

**Material and methods:-**

**Data set and Biological activity:** The data set containing 48, amide oxadiazole aniline<sup>9</sup> with anti-diabetic activities (Table 1) were taken for present studies in view of high structural diversity and sufficient variation in biological activities. Experimentally determined  $IC_{50}$  values ( $\mu M$ ) of the derivatives were converted into the negative logarithm ( $\text{Log } IC_{50}$ ).

**Generation of structure:** All the chemical structures (anti-diabetic activity) were sketched with the help of Accelrys (Discovery studio version 2.0) and imported into the worksheet of TSAR 3.3 software as .mol files.<sup>10</sup>

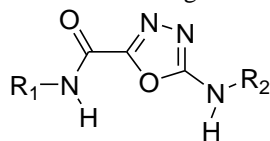
**Defining substituents and Energy optimized structure building:** The series has two major substituent's ( $R_1$  and  $R_2$ ) that were defined using "define substituent" option in the TSAR worksheet toolbar. All the loaded structures and their substituent's were converted into three-dimensional (3D) molecular structures by using Cornia make 3D option and further subjected to optimization using Cosmic – optimize 3D option, which includes valence terms as bond potential, bond angles and non-bonded terms as electrostatic interaction and Vanderwaals interaction. The force field supplied by "Cosmic" for energy calculation during a flexible optimization ensures that only the energetically realistic conformations are considered.<sup>11</sup>

**Calculation of Descriptors and Data reduction:** Initially more than 250 descriptors were calculated for both whole molecule and substituent's separately in TSAR software program. TSAR is an integrated analysis package for the interactive investigation of quantitative structure-activity relationships. It automatically calculates numerical descriptors of molecular structure. The calculated descriptors included molecular attributes, molecular indices, atom count and VAMP parameters.<sup>12</sup> The 48 molecules of the series were randomly divided into training set (32 molecules) and test set (11 molecules). Molecules in a training set further used for multiple linear regressions (MLR), partial least square (PLS) and feed forward neural network (FFNN) model development and test set consisting of 11 molecules which were kept on the other hand for future use to check the predictive power of the development model. There is a significant requirement of data reduction to eliminate the chance of correlation. Correlation matrix was used to reduce the number of descriptors and to identify the best subset of with minimum inter-correlation, than checked the other two descriptors. Pair-wise correlation coefficient was calculated for all the paired descriptors. If the inter-correlation coefficient  $>0.5$  was detected, then the descriptor with high correlation with biological activity was kept and others were discarded. This was done with the intention to get the descriptors which are less correlated to each other (independent in the true sense) and highly correlated to the biological activity.<sup>13,14</sup> Thus, finally five independent molecular descriptors, total dipole moment (subst. 2), log P (whole molecule), VAMP total energy (whole molecule), VAMP LUMO (whole molecule) and VAMP HOMO (whole molecule) were fetched and all the descriptors shows high correlation to the biological activity but did not have any correlation among each other.

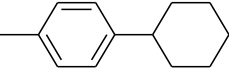
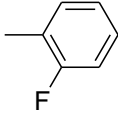
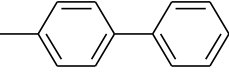
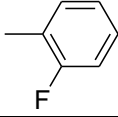
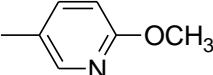
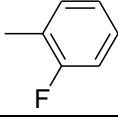
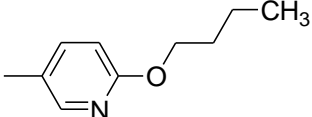
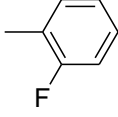
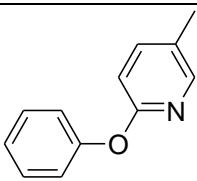
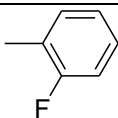
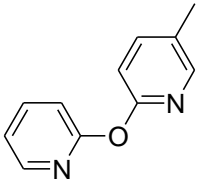
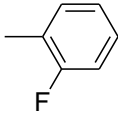
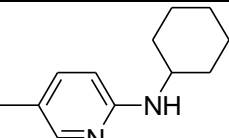
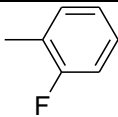
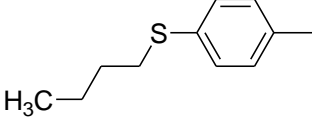
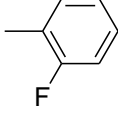
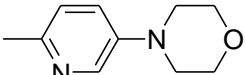
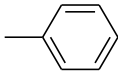
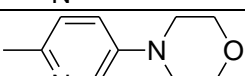
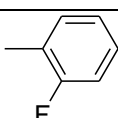
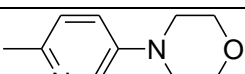
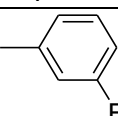
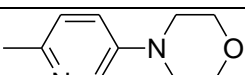

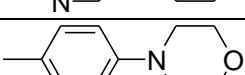

### Model development:-

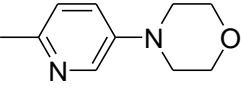
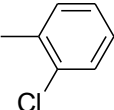
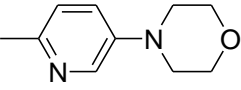
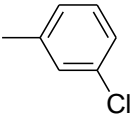
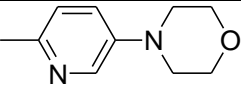
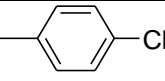
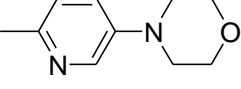
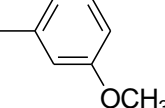
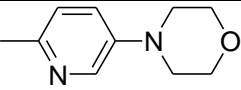
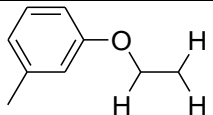
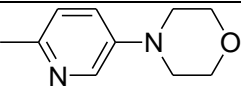
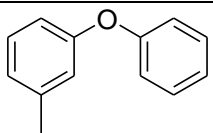
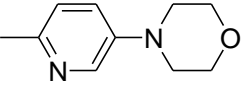
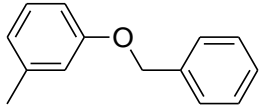
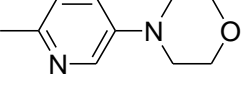
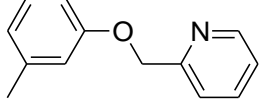
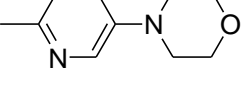
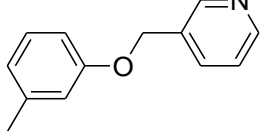
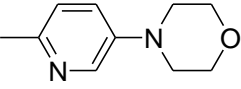
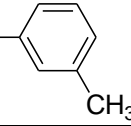
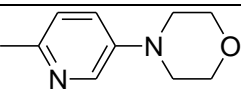
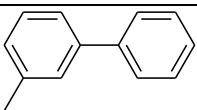
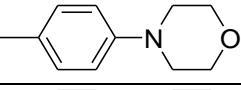
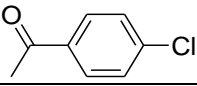
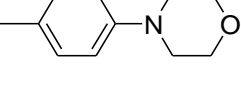
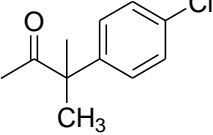
**Linear Regression Analysis:** The relationship between the selected descriptors and the biological activity was quantified by the use of multiple linear regressions (MLR) and partial least squares (PLS) using TSAR. MLR models were generated using biological activity data as dependent variable and selected descriptors as the independent variables. These models establish the relationship between dependent and independent variables. The cross-validation analysis was performed using the leave-one-out (LOO) method where one compound is removed from the data set and its activity is calculated using the model derived from the rest of data set. Statistical significance of the model were tested on the basis of conventional regression coefficient ( $r^2$ ), correlation coefficient ( $r$ ), Fisher's ratio ( $F$ ), and the standard error of estimate ( $s$ ). The PLS regression is described as a predictive method which can handle more than one dependent variable and is not critically influenced by correlations between independent variables.<sup>15</sup> PLS has been recommended as an alternative approach to enlarge the information contained in each model, and avoid the danger of overfitting.<sup>16</sup> To check the robustness and the predictability of the models, multiple linear regression (MLR) and partial least square (PLS) analysis was performed on the same training set of compounds similar to the cross-validation-method used in MLR.<sup>17</sup>

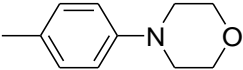
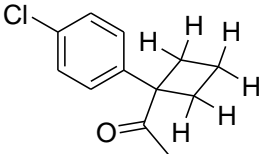
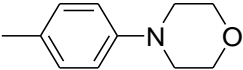
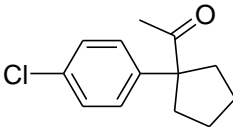
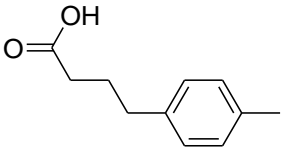
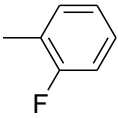
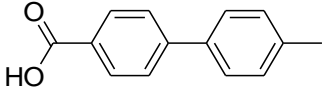
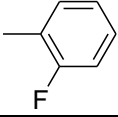
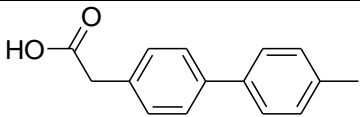
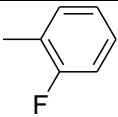
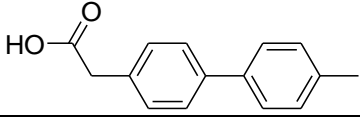
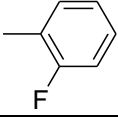
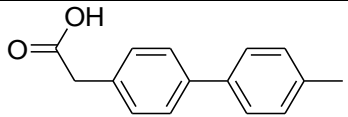
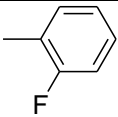
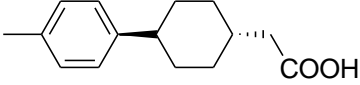
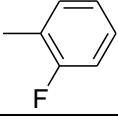
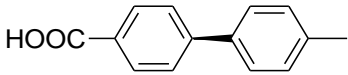
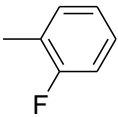
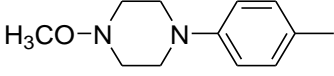
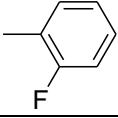
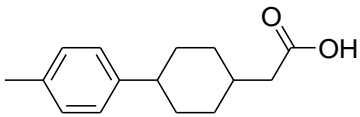
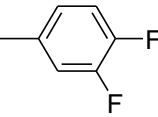
**Nonlinear regression analysis:** Feed forward neural network (FFNN) is a networking process, that used in the designing of plot dependency of final remaining descriptors and no. of plot dependency graphs depend upon on no. of descriptors was left. In this model having closer value of test RMS fit and best RMS fit of the training was obtained. The neural net configuration was modified by changing the number of nodes in the hidden layer. To ascertain that these inhibitors possess suitable pharmacokinetic properties, Lipinski's rule of five was applied to be drug like, a candidate should have less than 5 hydrogen bond donors (HBD), less than 10 hydrogen bond acceptors (HBA), a molecular weight of less than 500 dalton, and a partition coefficient (log p) of less than 5 and 10 or less than rotatable bonds. This rule describes the molecular properties related with pharmacokinetics of molecules.<sup>18</sup> The violation of above rule was analyzed by calculating the parameter for all the molecules. The results clearly indicate that there is zero violation of Lipinski's rule and all the designed compound will have favorable pharmacokinetic profiles summarized in Table 2.

**Table1:** Structure and biological activity data of DGAT1 antagonist used in QSAR analysis.

Comp. Name	R <sub>1</sub>	R <sub>2</sub>	DGAT1 IC <sub>50</sub> (μM)
1.			0.52
2.			0.46
3.			0.35
4.			0.13
5.			11
6.			0.19
7.			0.023
8.			3.7
9.			0.46
10.			0.13
11.			0.62

12.			0.31
13.			0.12
14.			8.0
15.			0.060
16.			0.0067
17.			0.41
18.			0.060
19.			0.016
20.			0.62
21.			0.13
22.			0.15
23.			0.84
24.			0.34

25.			1.3
26.			0.092
27.			0.19
28.			0.73
29.			0.032
30.			0.0066
31.			0.013
32.			0.54
33.			2.1
34.			0.17
35.			0.0057
36.			0.37
37.			0.52

38.			0.32
39.			0.18
40.			3.3
41.			0.21
42.			0.07
43.			0.08
44.			1.3
45.			0.0044
46.			0.035
47.			0.54
48.			0.0006

**Table 2:** Values of calculated parameters for Lipinski's rule of five.

Comp. Name	ADME (Molecular weight)	ADME (H-bond acceptors)	ADME (H-bond donors)	ADME (Log P)	ADME Rotatable bond	ADME Violations
1	383.42	5	2	2.373	5	0
2	401.41	5	2	2.512	5	0
3	397.45	5	2	2.840	5	0
4	384.41	6	2	2.224	5	0
5	384.41	6	2	1.759	5	0
6	381.45	4	2	3.437	5	0
7	400.47	5	2	2.667	5	0
8	416.47	6	2	1.530	5	0
9	397.46	6	2	2.368	5	0
10	384.41	6	2	2.224	5	0
11	354.42	4	2	4.579	7	0
12	380.46	4	2	4.802	5	0
13	374.4	4	2	4.608	5	0
14	329.32	6	2	2.522	5	0
15	371.41	6	2	3.729	8	0
16	391.39	6	2	4.204	6	0
17	392.38	7	2	3.357	6	0
18	390.41	5	3	3.887	6	0
19	386.48	4	2	4.222	8	0
20	366.42	6	2	2.085	5	0
21	384.41	6	2	2.225	5	0
22	384.41	6	2	2.225	5	0
23	384.41	6	2	2.225	5	0
24	402.4	6	2	2.364	5	0
25	400.86	6	2	2.603	5	0
26	400.86	6	2	2.603	5	0
27	400.86	6	2	2.603	5	0
28	396.45	7	2	1.832	6	0
29	410.48	7	2	2.175	7	0
30	458.52	7	2	3.514	7	0
31	472.55	7	2	3.609	8	0
32	473.54	8	2	2.696	8	0
33	473.54	8	2	2.762	8	0
34	380.45	6	2	2.552	5	0
35	442.52	6	2	3.769	6	0
36	427.88	6	2	2.401	5	0
37	469.97	6	2	3.561	6	0
38	481.98	6	2	3.453	6	0
39	496.01	6	2	3.849	6	0
40	384.4	6	3	3.347	8	0
41	418.41	6	3	4.306	6	0
42	432.44	6	3	4.239	7	0
43	432.44	6	3	4.239	7	0
44	438.5	6	3	4.006	7	0
45	438.5	6	3	4.296	7	0
46	424.47	6	3	4.132	6	0
47	424.48	5	2	1.820	5	0
48	456.49	6	3	4.436	7	0

## Result and discussions:-

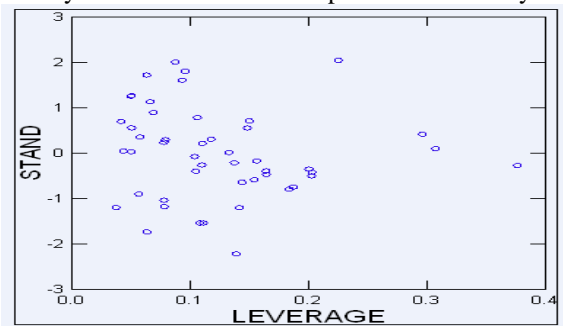
### Linear regression analysis:

The finally five descriptors total dipole moment (subst.2), log P (whole molecule), VAMP total energy (whole molecule), VAMP LUMO (whole molecule), VAMP HOMO (whole molecule) were left in the TSAR worksheet that showed very poor internal predictive ability of the developed model described in Table 3.

**Table3:** Statistical test and their values obtained for whole data sheet.

Statistical test	Values
S value	0.595
F value	11.538
Regression coefficient	0.760
$r^2$	0.579
Cross validation $r^2_{(cv)}$	0.344

In order to improve the statistical quality of the model, five compounds were identified (**47**, **34**, **46**, **22** and **13**) and behaved as outliers which further removed from the model. The deletion of these outliers satisfied all the statistical criteria of a robust model. So the model were generated with S value = **0.485**, F value = **17.967**, regression coefficient (r) = **0.841**,  $r^2 = 0.708$ ,  $r^2_{(cv)} = 0.617$ . These five outliers shows the high residual value instead of high leverage value and therefore were deleted. Applicability domain was also performed for the calculation of leverage using the systat software on the developed model. It is the physico-chemical, structural or biological space, knowledge or information which is applicable to make predictions for new compounds. The graph calculated leverage versus stand residual values in Fig. 1. Finally five parameters highly correlated with activity were used to generate regression equation and analyzed for their relative impact on the activity of the compound in Table 4.



**Figure1:** William's plot (graph of AD)

**Table4:** Correlation matrix showing correlation between biological activity and parameter used.

	Total dipole Moment(subst.2)	Log P (whole molecule)	VAMP total energy	VAMP LUMO	VAMPHOMO	-Log IC <sub>50</sub>
Total dipole Moment (subst. 2)	1	0.105	-0.068	-0.027	0.161	-0.279
Log P (whole molecule)	0.105	1	-0.126	0.206	-0.126	0.543
VAMP total energy	-0.068	-0.126	1	0.321	0.052	-0.437
VAMP LUMO	-0.027	0.206	0.052	1	0.212	-0.214
VAMP HOMO	0.161	-0.126	0.321	0.212	1	0.0993
-Log IC <sub>50</sub>	-0.279	0.543	-0.437	-0.214	0.0993	1

So the best models generated using MLR analysis of this data had  $r^2 = 0.889$  and  $r^2_{(cv)} = 0.842$  values. The final statistical values are given in Table 5.



**Table5:** Statistical tests and their values obtained after performing MLR analysis.

Statistical tests	Values
S value	0.331
F value	41.912
Regression coefficient	0.943
$r^2$	0.889
Cross validation, $r^2_{(cv)}$	0.842

The final regression equations obtained using MLR analysis is represented as (Equation 1).

$$Y = -0.537 \times X1 + 0.639 \times X2 - 0.002 \times X3 - 5.409 \times X4 + 1.856 \times X5 + 4.549 \quad (\text{Equation 1})$$

The PLS analysis was also performed using the same data set, the resulted  $r^2_{(cv)}$  value (**0.801**), statistical significance (**0.847**) and the fraction of variance (**0.887**) clearly demonstrates the high predictive ability of the developed PLS model (Equation 2).

$$Y = -0.545 \times X1 + 0.663 \times X2 - 0.001 \times X3 + 5.051 \times X4 + 1.839 \times X5 + 5.002 \quad (\text{Equation 2})$$

Where X1 = total dipole moment (subst. 2), X2 = log P (whole molecule), X3 = VAMP total energy (whole molecule), X4 = VAMP LUMO (whole molecule), X5 = VAMP HOMO (whole molecule).

For a well-defined problem, both MLR (**0.842**) and PLS (**0.801**) should have comparable results.<sup>13, 14</sup> The  $r^2$  values of training and test set were **0.889**, **0.887** and **0.738**, **0.752** for MLR and PLS models respectively. The experimentally determined Log IC<sub>50</sub> values for the compounds of training and test set with their predicted and actual values are shown in table 6-7 and the graphs of MLR, PLS and FFNN of training and test set were plotted in Fig. 2-7 respectively. Observation of these data suggests that the experimentally observed values and QSAR derived values are in agreement. Therefore, all the t-value, Jackknife SE and Covariance SE values are mentioned in table 8 were significant for best model that confirms the importance of each descriptor.

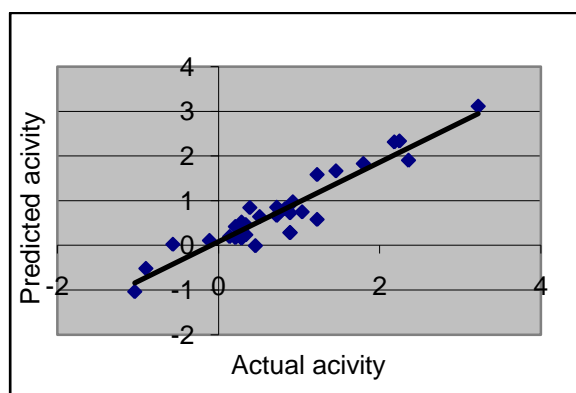
**Table6:** Actual activity versus predicted activity and corresponding residual for the training set of compound.

Comp. Name	Actual activity (Log IC <sub>50</sub> )	Predicted activity		
		MLR	PLS	FFNN
1	0.284	0.166	0.165	0.156
7	0.337	0.237	0.175	0.138
11	-1.041	-1.035	-1.015	-0.609
12	0.721	0.669	0.723	0.619
14	-0.568	0.024	-0.067	0.031
15	0.337	0.463	0.435	0.421
17	0.207	0.174	0.270	0.085
18	0.508	0.643	0.728	0.488
19	0.921	0.973	1.040	0.872
20	-0.903	-0.516	-0.512	-0.332
23	0.387	0.845	0.808	0.734
24	1.222	1.584	1.596	1.610
25	1.796	1.829	1.885	1.912
26	0.207	0.421	0.453	0.417
27	0.824	0.819	0.774	0.793
31	1.036	0.749	0.727	0.708
32	0.721	0.853	0.822	0.822
33	0.137	0.196	0.192	0.185
35	2.180	2.313	2.296	2.274
39	0.769	0.767	0.806	0.738
40	2.244	2.334	2.354	2.302
43	0.284	0.520	0.492	0.381
50	-0.114	0.109	0.174	-0.063
51	2.356	1.903	1.920	1.818
52	1.456	1.664	1.677	1.575
16	0.268	0.448	0.402	0.405
53	3.222	3.113	3.038	2.764
8	0.456	-0.008	-0.011	-0.041

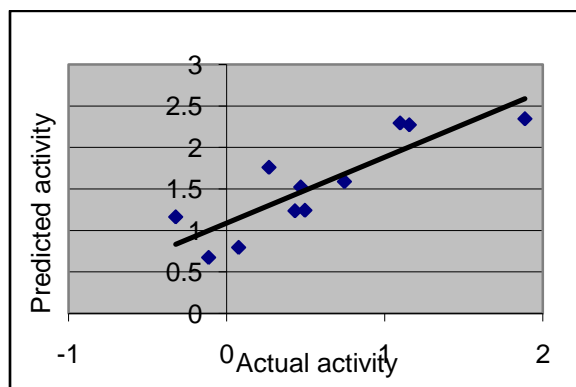
10	0.886	0.283	0.243	0.254
10a	0.886	0.288	0.249	0.259
21	1.222	0.580	0.603	0.473
10b	0.886	0.727	0.694	0.694

**Table7:** Actual activity versus predicted activity for the test set of compound.

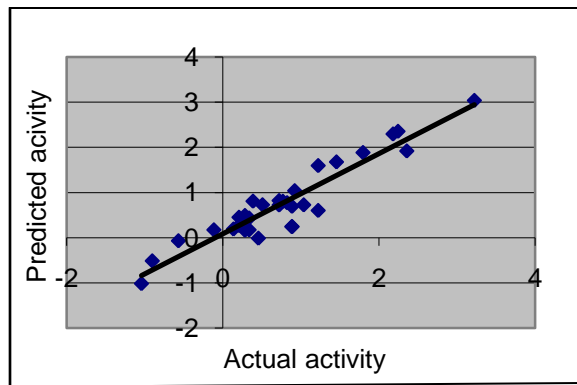
Comp. Name	Actual activity (Log IC <sub>50</sub> )	Predicted activity		
		MLR	PLS	FFNN
38	-0.322	1.161	1.088	1.069
48	1.155	2.272	2.268	2.210
49	1.097	2.294	2.292	2.232
37	0.268	1.759	1.706	1.744
30	-0.114	0.672	0.659	0.626
28	0.076	0.793	0.748	0.765
29	0.468	1.522	1.396	1.548
36	1.886	2.345	2.329	2.287
42	0.431	1.234	1.138	1.261
44	0.495	1.241	1.224	1.143
45	0.744	1.587	1.579	1.502



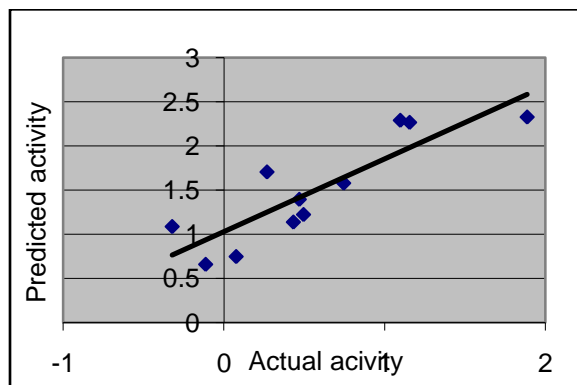
**Figure 2:** Plot of actual activity versus predicted activity for the training set of compound derived from MLR analysis.



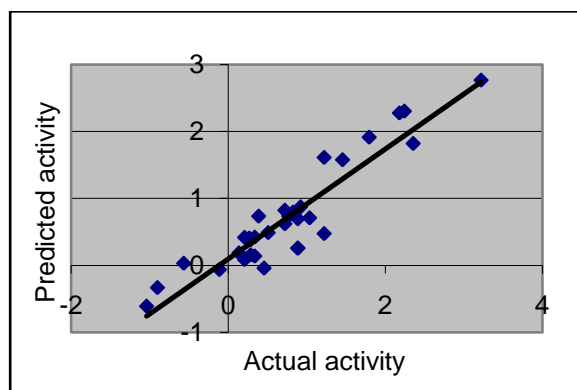
**Figure 3:** Plot of actual activity versus Predicted activity for the test set of compound derived from MLR analysis.



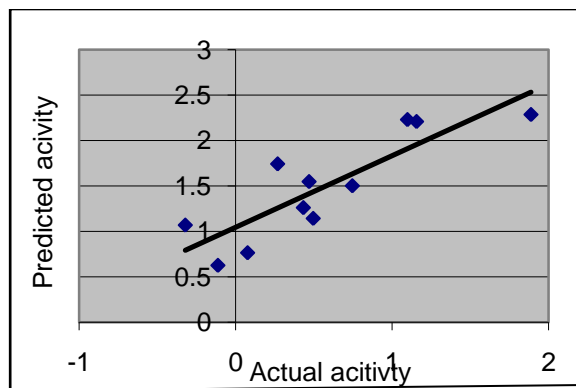
**Figure 4:** Plot of actual activity versus predicted activity for the training set of compound derived from PLS analysis.



**Figure 5:** Plot of actual activity versus predicted activity for the test set of compound derived from PLS analysis.



**Figure 6:** Plot of actual activity versus predicted activity for the training set of compound derived from FFNN analysis.



**Figure 7:** Plot of actual activity versus predicted activity for the test set of compound derived from FFNN analysis.

**Non Linear Regression analysis:** The feed forward neural network (FFNN) analysis was performed using the same data set, by Net Configuration **5-1-1**, and amount of data excluded for testing was set to **25%**, which was used for cross-validation to assess the performance of the trained net. The best model having closer values of test **RMS fit = 0.146** and best **RMS fit = 0.033** and  **$r^2 = 0.864$**  of the training and  **$r^2 = 0.734$**  of the test set was obtained. Dependency plots were drawn to analyze the influence of each independent parameter versus biological activity.

**Table8:** t-test values, Jackknife SE and Covariance SE for the selected descriptors.

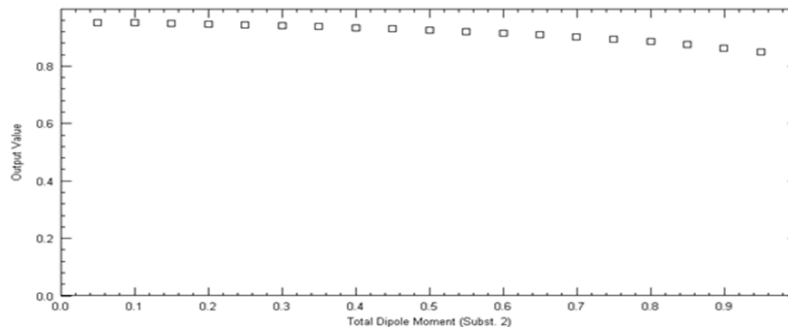
Descriptors	t-value	Jackknife SE	Covariance SE
Total dipole moment (Subst. 2)	-7.213	0.075	0.074
Log P (whole molecule)	10.02	0.059	0.063
VAMP total energy (whole molecule)	-7.606	0.0001	0.0002
VAMP LUMO (whole molecule)	-6.577	0.864	0.822
VAMP HOMO (whole molecule)	7.279	0.249	0.255

### Analysis of descriptors entered:-

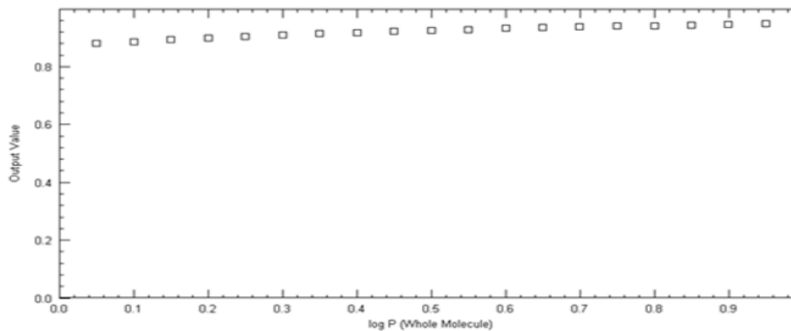
Total dipole moment (subst.2) is electrostatic descriptor which explains the charge distribution, strength and orientation behavior in molecules.<sup>19</sup> It shows negative correlation with biological activity, which indicates lead compound by substituting such groups at 2<sup>nd</sup> position led to decrease in the polarity of the molecule, owing to increase in the biological activity of DGAT-1 antagonist derivatives. This is further supported by the FFNN dependency graph shown in Fig. 8.

Log P (whole molecules) measure the hydrophobic interaction. The hydrophobic effect can be quantified by measuring the partition coefficients of non-polar molecules between polar and non-polar solvents. Hydrophobicity increases with increasing number of carbon atoms in the hydrocarbon chain, positive correlation of log P with the biological activity increases the hydrophobicity of whole molecules as shown in Fig. 9.

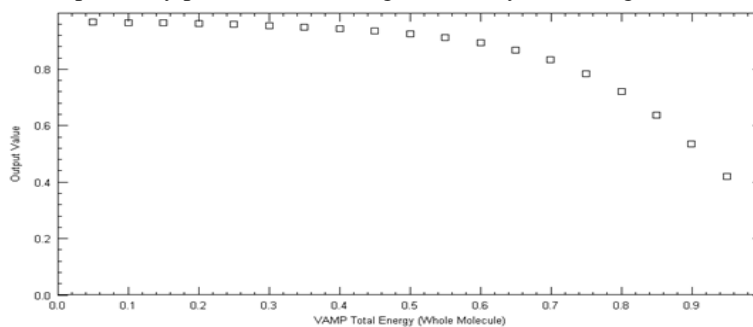
VAMP (total energy, LUMO, HOMO) is electronic parameters. VAMP is a semi-empirical molecular orbital package in TSAR Version 3.3, and is used to calculate the electrostatic properties.<sup>19</sup> As both VAMP (total energy) and VAMP LUMO (lowest unoccupied molecular orbital) parameters negative correlates with the biological activity in the regression equation, its means electron withdrawing group add into the ring increases the biological activity. VAMP HOMO is energy of highest occupied molecular orbital, with “nucleophilicity” properties, so it is positively correlated to the biological activity; it means addition of electron donating group in the ring leads to an increase in the biological activity of DGAT-1 antagonist derivatives (Fig. 10, 11 and 12).



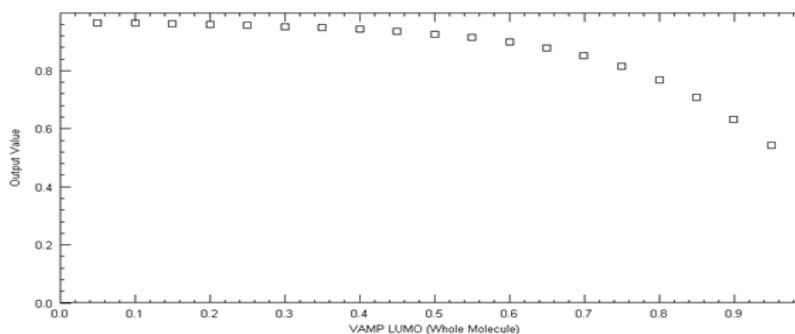
**Figure 8:** Dependency plot between biological activity versus total dipole moment (subst. 2).



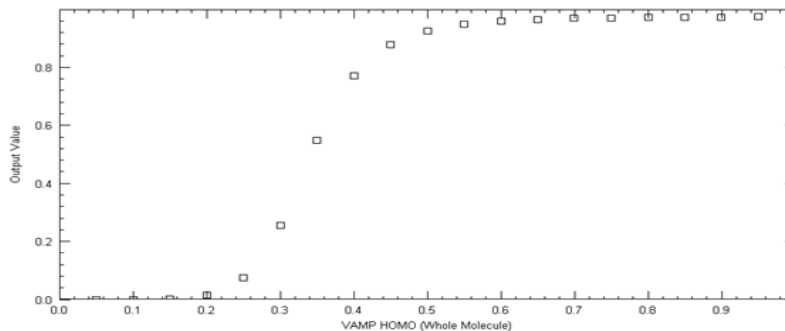
**Figure 9:** Dependency plot between biological activity versus log P (whole molecule).



**Figure 10:** Dependency plot between biological activity versus VAMP Total Energy (whole molecule).



**Figure 11:** Dependency plot between biological activity versus VAMP LUMO (whole molecule).



**Figure 12:** Dependency plot between biological activity versus VAMP HOMO (whole molecule).

### Similarity based regression analysis:-

Structurally similar molecules have similar biological activities observed in studies related to medicinal chemistry. If the training set being investigated comprises of  $N$  compounds, then full pair-wise compound similarity comparison results  $N \times N$  matrix in which each matrix entry is a measure of similarity between the corresponding pair of molecules. Analysis of full matrix introduces some of the location dependent similarity parameters used within methodology and give correlation with binding data.<sup>20</sup>

### Automated Similarity Package (Asp) similarity:-

Asp is used to calculate the similarity of whole molecules in terms of atomic potential charge, shape, lipophilicity or refractivity.

The two types of calculation methods

- Carbo index
- Hodgkin index
- Carbo index

### Material and methods:-

**Data set preparation:** The same series of compounds as in case of physicochemical descriptor based QSAR study were randomly divided into training set and test set. 33 molecules were included in the training set and used to develop a regression model while 12 molecules were used in test set for the predictions of biological activity.

**Descriptor calculation:** The Carbo method was applied to calculate the similarity descriptors including shape, lipophilicity, charge, combined and refractivity similarity indices, by  $N \times N$  method for whole molecules through TSAR 3.3 software.

**Data reduction:** The correlation matrix was generated to study the data patterns and to reduce it. The term close to 1 indicates high co-linearity while below 0.5 indicates that no co-linearity exists between two parameters. Whosoever descriptors causes low productivity and over-fitting of data was discarded. Among the highly correlated parameters, the one that showed low correlation with the biological activity ( $\text{Log IC}_{50}$ ) was excluded while the other was kept. This process was repeated for each and every set of two consecutive parameters and finally 5 descriptors were attained namely Charge similarity vs. molecule **4**, Combined similarity vs. molecule **13**, Lipophilicity similarity vs. molecule **23**, Refractivity similarity vs. molecule **4**, Charge similarity vs. molecule **48** that shows highly correlation to the biological activity but did not have any correlation with each other.

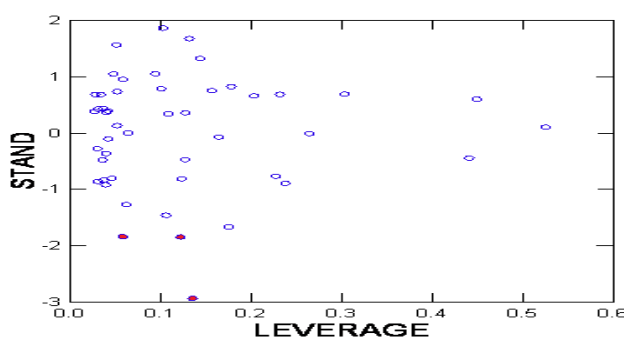
### Result and discussion:-

**Linear regression analysis:** For the data set of Diacylglycerolacyltransferase 1 inhibitor, the best model was generated using MLR analysis after deleting 3 potential outliers includes 5 independent variables, as summarized in Table 9.

**Table 9:** Statistical Value obtained from after MLR analysis.

Statistical tests	Values
S value	0.360
F value	36.334
Regression coefficient	0.933
$r^2$	0.870
Cross validation, $r^2_{(cv)}$	0.768

These above 3 outliers were deleted as shown in the plot of applicability domain Fig.13.

**Figure 13:** Applicability Domain graph between Stand versus Leverage

Finally five parameters, highly correlated with biological activity were used to generate regression equation and analyzed for their relative impact on the activity of the compound in Table 10.

**Table 10:** Correlation matrix showing correlation between biological activity and parameter used.

Descriptors	Charge similarity vs. molecule (4)	Combined similarity vs. molecule (13)	Lipophilicity similarity vs. molecule (23)	Refractivity similarity vs. molecule(32)	Charge similarity vs. molecule (48)	-Log IC <sub>50</sub>
Charge similarity vs. molecule (4)	1	-0.099	-0.192	-0.247	0.349	-0.293
Combined similarity vs. molecule (13)	-0.099	1	-0.246	-0.234	-0.466	0.129
Lipophilicity similarity vs. molecule (23)	-0.192	-0.246	1	0.584	0.276	0.367
Refractivity similarity vs. molecule (32)	-0.247	-0.234	0.584	1	0.0701	-0.234
Charge similarity vs. molecule (48)	0.349	-0.466	0.276	0.0701	1	0.436
-Log IC <sub>50</sub>	-0.293	0.129	0.367	-0.234	0.436	1

The final regression equations obtained using MLR analysis is represented as Equation 1.

$$Y = -2.083 \times X_1 + 2.985 \times X_2 + 9.277 \times X_3 - 119.527 \times X_4 + 2.403 \times X_5 + 108.167 \quad (\text{Equation 1})$$

The PLS analysis was also performed using the same data set, the resulted  $r^2_{(cv)}$  value and statistical significance of clearly demonstrates the high predictive ability of the developed PLS model (Equation 2), mentioned in Table.11.

$$Y = -2.075 \times X_1 + 3.102 \times X_2 + 9.174 \times X_3 - 117.041 \times X_4 + 2.432 \times X_5 + 105.736 \quad (\text{Equation 2})$$

Where, X1=Charge similarity vs. molecule 4, X2=Combined similarity vs. molecule 13, X3=Lipophilicity similarity vs. molecule 23, X4=Refractivity similarity vs. molecule 32, X5=Charge similarity vs. molecule 48.

**Table 11.** Statistical Value obtained from after PLS analysis

Statistical tests	Values
Cross validation, $r^2_{(cv)}$	0.775
Statistical significance	0.908
Fraction of variance explained	0.870
E statistic	0.347
Residual Sum of Squares	4.148
Predictive Sum of Squares	7.174

Validation through external test set was also performed to check the predictive ability of the developed model. A good  $r^2$  value of the test set **0.797** and **0.791** was obtained, by MLR and PLS analysis.

**Non Linear Regression analysis:** The feed forward neural network (FFNN) analysis was performed using the same data set in Table 12.

**Table 12:** Summary of feed forward neural network analysis.

Net Configuration	5-5-1
Excluded for testing was set to	30%,
Test RMS fit	0.180
Best RMS fit	0.073
Training set ( $r^2$ )	0.798
Test set ( $r^2$ )	0.644

The MLR, PLS and FFNN graphs plotted between the actual and predicted activities of training set as well as test set of compound, as shown Fig. 14-19. The actual and predicted activities of training set and test set of compounds are given in Table 13 and 14 respectively. Dependencies plots were drawn to analyze the influence of each independent parameter versus biological activity Fig 20-24.

**Table 13:** Actual activity versus predicted activity for the training set of compound.

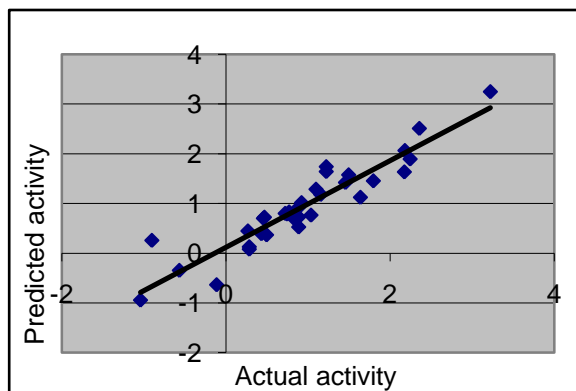
Comp. Name	Actual activity (Log IC <sub>50</sub> )	Predicted activity		
		MLR	PLS	FFNN
1	0.284	0.079	0.081	-0.662
8	0.456	0.701	0.704	0.471
11	-1.041	-0.945	-0.942	-1.186
14	-0.568	-0.347	-0.342	-0.576
19	0.921	1.015	1.015	1.004
24	1.22	1.739	1.773	1.635
25	1.796	1.454	1.466	1.682
27	0.824	0.707	0.707	0.751
29	0.468	0.722	0.719	0.649
31	1.036	0.764	0.763	0.849
32	0.721	0.798	0.795	0.582
34	1.495	1.575	1.568	1.697
35	2.180	2.063	2.038	2.080
39	0.769	0.827	0.827	0.795
40	2.244	1.893	1.876	2.007
42	0.432	0.391	0.387	-0.151
43	0.284	0.127	0.123	0.161
44	0.495	0.367	0.339	0.553
48	1.155	1.179	1.171	1.299
51	2.357	2.508	2.515	1.827
52	1.456	1.421	1.423	1.451
16	0.268	0.446	0.453	0.169
45	0.745	0.785	0.747	0.873
49	1.097	1.286	1.276	1.453



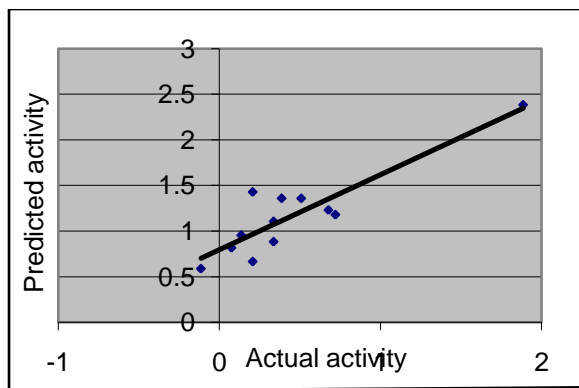
10	0.886	0.525	0.523	0.062
10a	0.886	0.922	0.925	0.895
21	1.222	1.641	1.641	1.668
10b	0.886	0.725	0.727	0.453
22	2.174	1.633	1.658	1.641
13	1.638	1.123	1.123	1.286
53	3.222	3.247	3.259	1.769
20	-0.903	0.257	0.259	-0.337
50	-0.114	-0.636	-0.605	-0.124

**Table 14:** Actual activity versus predicted activity for the test set of compound.

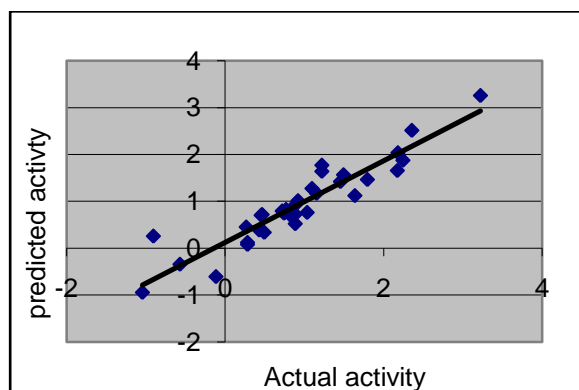
Comp. Name	Actual activity (Log IC <sub>50</sub> )	Predicted activity		
		MLR	PLS	FFNN
7	0.337	1.108	1.115	0.814
15	0.337	0.885	0.885	0.559
17	0.208	1.429	1.414	1.434
18	0.509	1.359	1.363	1.255
23	0.387	1.360	1.381	1.221
26	0.208	0.668	0.669	0.203
28	0.076	0.818	0.819	0.497
33	0.137	0.956	0.955	0.366
36	1.886	2.384	2.349	2.104
30	-0.114	0.589	0.589	-0.140
47	0.678	1.232	1.218	1.131
12	0.721	1.181	1.184	0.918



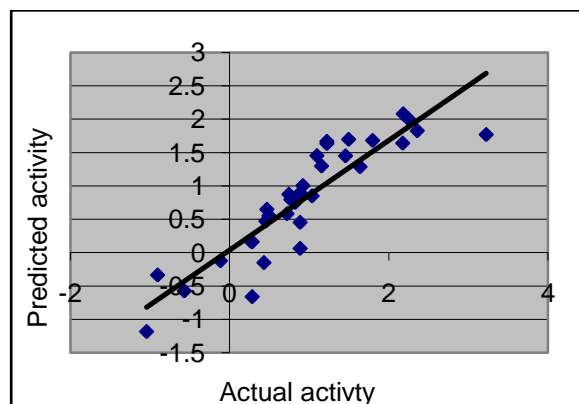
**Figure 14:** Plot of actual activity versus predicted activity for the training set of compound derived from MLR analysis.



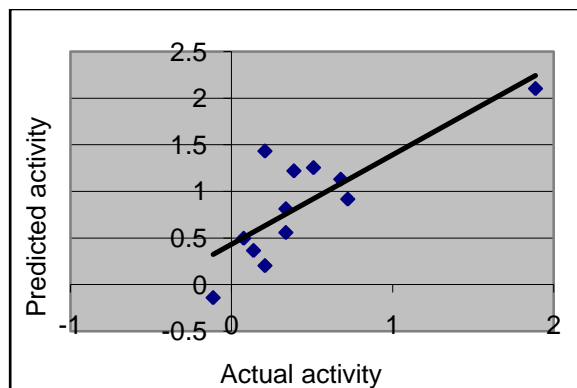
**Figure 15:** Plot of actual activity versus Predicted activity for the test set of compound derived from MLR analysis.



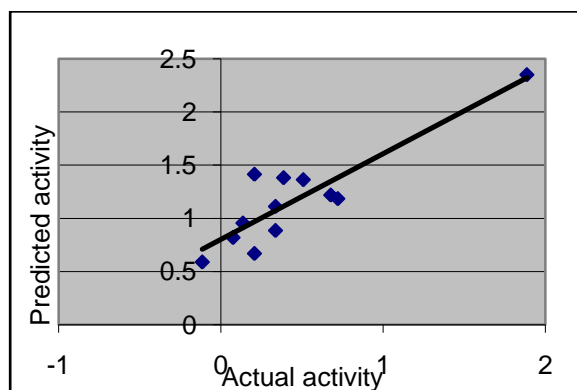
**Figure 16:** Plot of actual activity versus predicted activity for the training set of compound derived from PLS analysis.



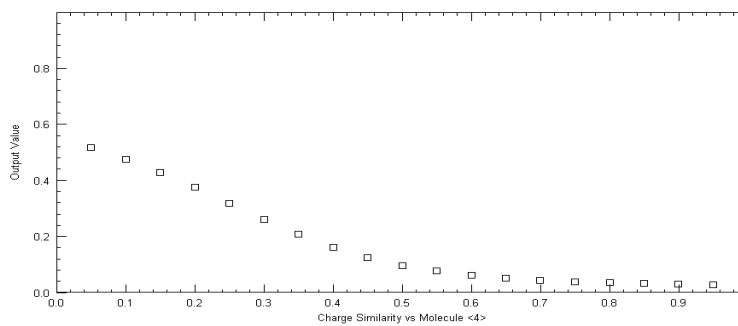
**Figure 17:** Plot of actual activity versus predicted activity for the training set of compound derived from FFNN analysis.



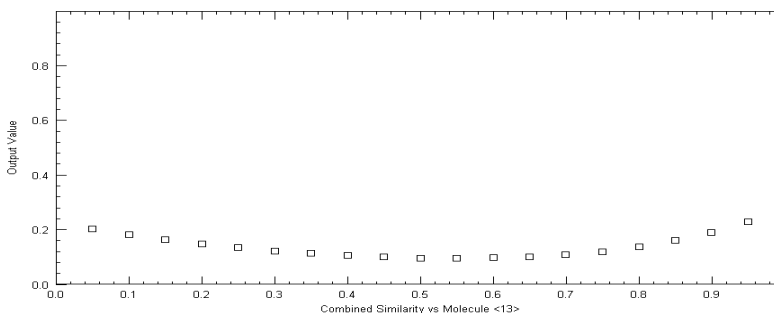
**Figure 18:** Plot of actual activity versus predicted activity for the test set of compound derived from FFNN analysis.



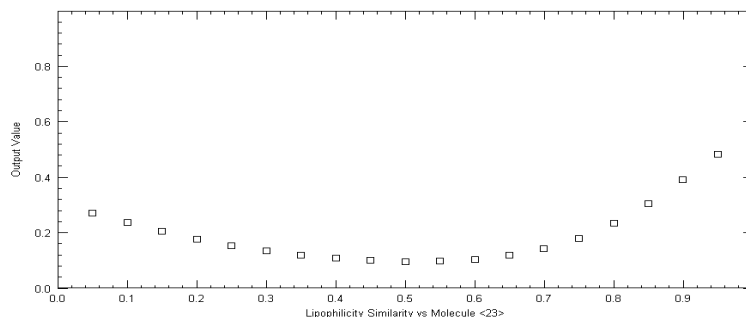
**Figure 19:** Plot of actual activity versus predicted activity for the test set of compound derived from PLS analysis.



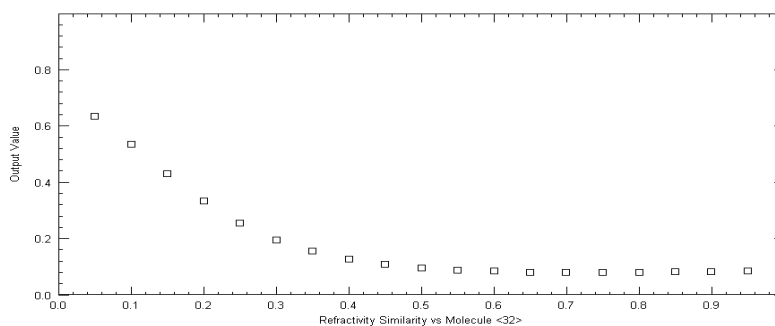
**Figure 20:** Dependency plot between biological activity versus Charge similarity vs. molecule 4.



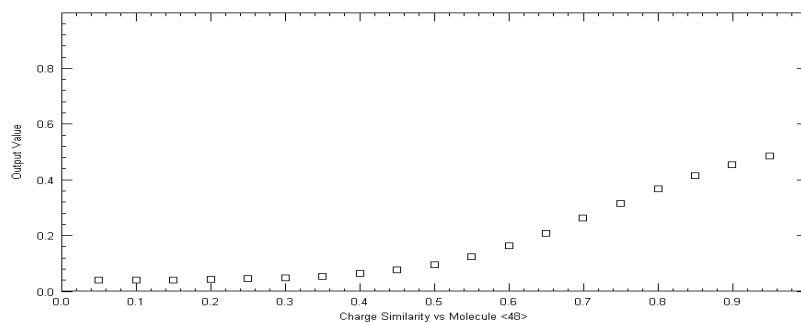
**Figure 21:** Dependency plot between biological activity versus Combined similarity vs. molecule 13.



**Figure 22:** Dependency plot between biological activity versus Lipophilicity similarity vs. molecule 23.



**Figure 23:** Dependency plot between biological activity versus Refractivity similarity vs. molecule 32.



**Figure 24:** Dependency plot between biological activity versus Charge similarity vs. molecule 48.

#### Analysis of descriptors entered:-

Charge molecular descriptors are directly related to the energy of the electrostatic interaction. According to the contemporary theory of molecular structure, all chemical interaction by nature are either electrostatic (polar) or orbital (covalent). The electrical charge in the molecule is the driving force for electrostatic interaction. The charge descriptor has been widely employed as chemical reactivity to measure intermolecular interaction in Asp QSAR study.

The Charge of molecule **4** is negatively correlated with biological activity. The negative correlation of this parameter is further supported by FFNN dependency plot. The Charge of whole compounds should not be similar to it that will lead to an increase in biological activity.

Combined similarity descriptor is combination of charge, shape, lipophilicity and refractivity properties of molecules. The combined similarity vs. molecule **13** positively correlates with biological activity. The positive correlation of this parameter was already shown in Fig. 21. Combined similarity of whole compound should be similar to the molecule which increases the biological activity.

Likewise combined similarity descriptor, lipophilicity of molecule **23** is also positively correlated with biological activity, which increased the lipid solubility and the dependency plot was shown earlier in Fig. 22. The compound

which has good lipophilic property but they have little capacity to form hydrogen bonds then the lipid solubility of whole compounds should be similar to it.

Molecular Refractivity is an additive property, its value increase with molecular weight and volume. Refractivity of molecule **32** is negatively correlated with biological activity. The negative correlation of this parameter is further supported by FFNN dependency plot shown in Fig. 23. The refractivity of whole compounds was not be similar to (molecule **32**), because molecule which is a good electron donor comes in contact with a molecule possesses good electron acceptor capacity, the donor may transfer some of its charge to the acceptor. The Charge of molecule **48** is also positively correlated with biological activity and shown in Fig. 24, and the Charge of it should be similar to the whole series of compounds.

#### Hodgkin index:-

The Hodgkin method was applied to calculate the similarity descriptors including shape, lipophilicity, charge, combined and refractivity similarity indices, by N×N method for whole molecules using TSAR 3.3 software.

#### Material and method:-

- **Data set preparation**

The process of data set preparation was similar as described above, and finally 3 descriptors namely Charge similarity vs. molecule **22**, Charge similarity vs. molecule **46** Combined similarity vs. molecule **48**, were found that are highly correlated to the biological activity but did not have any correlation with each other.

#### Result and discussion:-

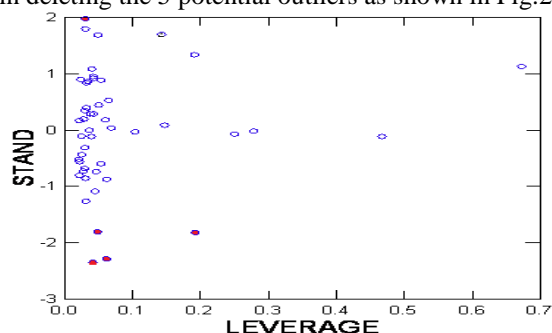
- **Linear regression analysis:**

The best MLR model was generated after deleting 5 potential outliers which include 3 independent variables Table 15.

**Table 15:** Statistical Value obtained after performing MLR analysis.

Statistical tests	Values
S value	0.399
F value	35.954
Regression coefficient	0.891
$r^2$	0.794
Cross validation, $r^2_{(cv)}$	0.733

The applicability domain helps in deleting the 5 potential outliers as shown in Fig.25.



**Figure 25:** Applicability Domain graph between Stand versus Leverage.

The correlation matrix of final three descriptors shows high correlation with biological activity as summarized in Table 16.

**Table 16:** Correlation matrix showing correlation between biological activity and parameter used.

	Charge similarity vs. molecule (22)	Charge similarity vs. molecule (46)	Combined similarity vs. molecule (48)	-Log IC <sub>50</sub>
Charge similarity vs. molecule (22)	1	-0.117	-0.041	0.606
Charge similarity vs. molecule (46)	-0.117	1	0.677	0.003
Combined similarity vs. molecule (48)	-0.041	0.677	1	0.503
-Log IC <sub>50</sub>	0.606	0.003	0.503	1

The final regression equations of MLR were obtained.

$$Y = 2.377 \times X_1 - 3.185 \times X_2 + 6.293 \times X_3 - 2.791 \quad \text{(Equation 1)}$$

Similarly, the PLS equation was also derived and statistical values of the PLS model were given in Table 17.

$$Y = 2.377 \times X_1 - 3.185 \times X_2 + 6.293 \times X_3 - 2.791 \quad \text{(Equation 2)}$$

Where, X<sub>1</sub> = Charge similarity vs. molecule 22, X<sub>2</sub> = Charge similarity vs. molecule 46, X<sub>3</sub> = Combined similarity vs. molecule 48.

**Table 17:** Statistical Value obtained from after PLS analysis.

Statistical tests	Values
Cross validation, $r^2_{(cv)}$	0.751
Statistical significance	0.969
Fraction of variance explained	0.793
E statistic	0.831
Residual Sum of Squares	6.388
Predictive Sum of Squares	7.708

External test set was also performed to check the predictive ability and validate the developed model. A good  $r^2$  value (test set) of **0.767** and **0.767** was obtained through MLR and PLS analysis.

- **Non Linear Regression analysis:** The same data set was performed by FFNN analysis summarized in Table 18.

**Table 18:** Summary of feed forward neural network analysis.

Net Configuration	3-1-1
Excluded for testing was set to	25%
Test RMS fit	0.107
Best RMS fit	0.077
Training set $r^2$	0.776
Test set $r^2$	0.741

In the Fig. 26-31, various MLR, PLS and FFNN graphs were plotted. Similarly, the actual and predicted activities of training set and test set of compounds are also given in Table 19 and 20 respectively.

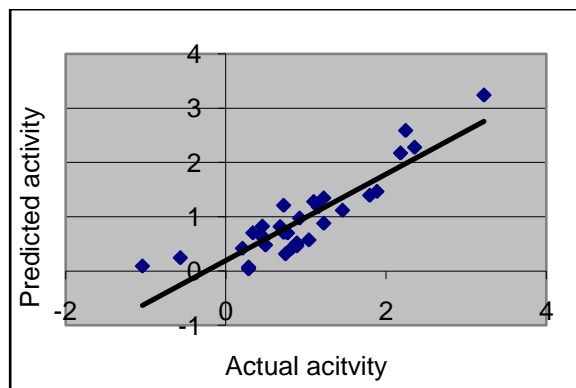
**Table 19:** Actual activity versus predicted activity for the training set of compound.

Comp. Name	Actual activity (Log IC <sub>50</sub> )	Predicted activity		
		MLR	PLS	FFNN
1	0.284	0.038	0.038	-0.273
8	0.456	0.823	0.823	0.764
11	-1.041	0.091	0.091	-0.231
12	0.721	1.208	1.208	1.454
14	-0.568	0.244	0.244	0.115
15	0.337	0.702	0.702	0.739
19	0.921	0.974	0.973	1.122
24	1.222	1.344	1.343	1.680

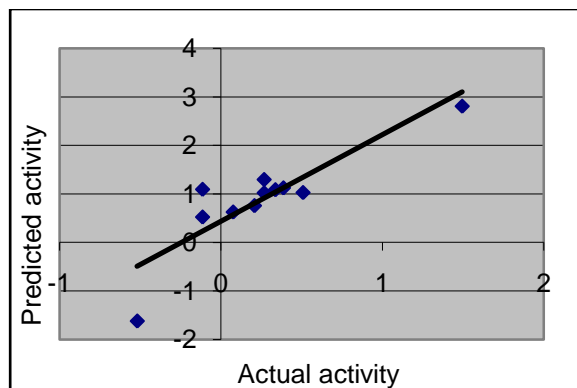
25	1.796	1.392	1.392	1.694
26	0.208	0.417	0.417	0.333
27	0.824	0.422	0.422	0.331
29	0.468	0.605	0.605	0.582
31	1.036	0.572	0.572	0.529
32	0.721	0.709	0.709	0.741
35	2.180	2.170	2.170	2.615
36	1.886	1.463	1.463	1.931
39	0.769	0.699	0.699	0.722
40	2.244	2.586	2.586	2.829
42	0.432	0.674	0.674	0.779
43	0.284	0.072	0.072	-0.128
44	0.495	0.478	0.478	0.421
48	1.155	1.179	1.179	1.388
51	2.357	2.276	2.276	2.426
52	1.456	1.117	1.117	1.125
45	0.745	0.316	0.316	0.199
49	1.097	1.277	1.277	1.519
10	0.886	0.462	0.462	0.312
10a	0.886	0.516	0.516	0.479
21	1.222	0.878	0.878	1.009
10b	0.886	0.500	0.500	0.446
47	0.678	0.817	0.817	0.862
53	3.222	3.237	3.237	2.906

**Table 20:** Actual activity versus predicted activity for the test set of compound.

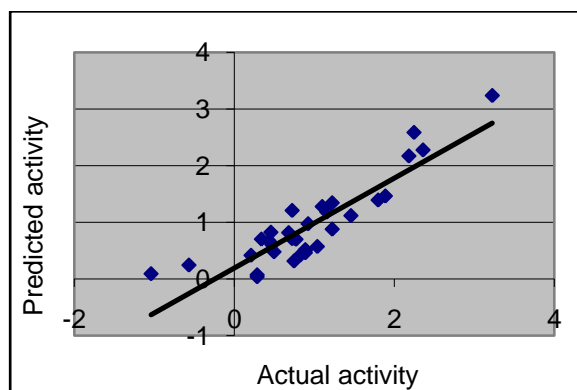
Comp. Name	Actual activity (Log IC <sub>50</sub> )	Predicted activity		
		MLR	PLS	FFNN
7	0.337	1.088	1.088	1.313
17	0.208	0.759	0.759	0.848
18	0.509	1.029	1.029	1.214
23	0.387	1.125	1.125	1.385
26	0.076	0.628	0.628	0.632
34	1.495	2.808	2.808	2.927
50	-0.114	1.095	1.095	1.416
16	0.268	1.023	1.023	1.245
37	0.268	1.297	1.297	1.732
30	-0.114	0.525	0.525	0.467
46	-0.518	-1.618	-1.618	-1.321



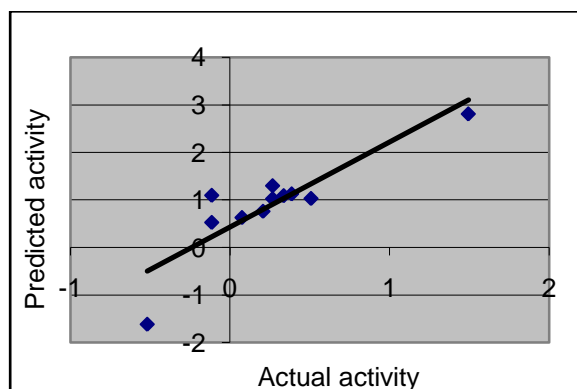
**Figure 26:** Plot of actual activity versus predicted activity for the training set of compound derived from MLR analysis.



**Figure 27:** Plot of actual activity versus predicted activity for the test set of compound derived from MLR analysis.

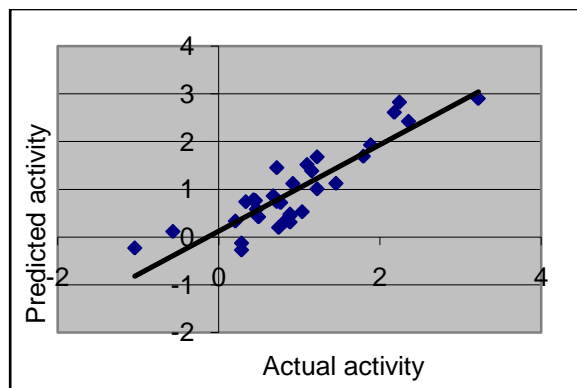


**Figure 28:** Plot of actual activity versus predicted activity for the training set of compound derived from PLS analysis.

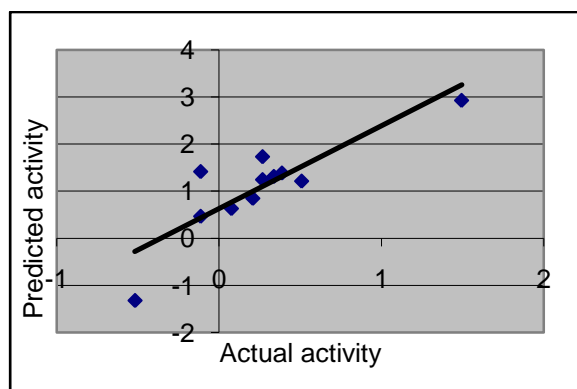


**Figure 29:** Plot of actual activity versus predicted activity for the test set of compound derived from PLS analysis.



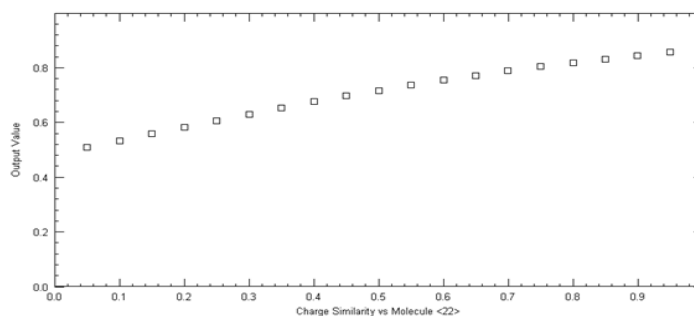


**Figure 30:** Plot of actual activity versus predicted activity for the training set of compound derived from FFNN analysis.

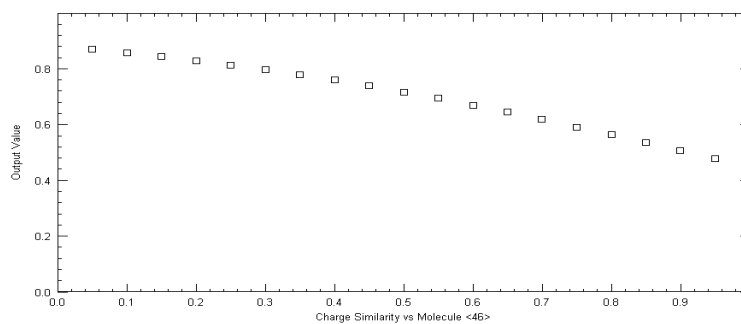


**Figure 31:** Plot of actual activity versus predicted activity for the test set of compound derived from FFNN analysis.

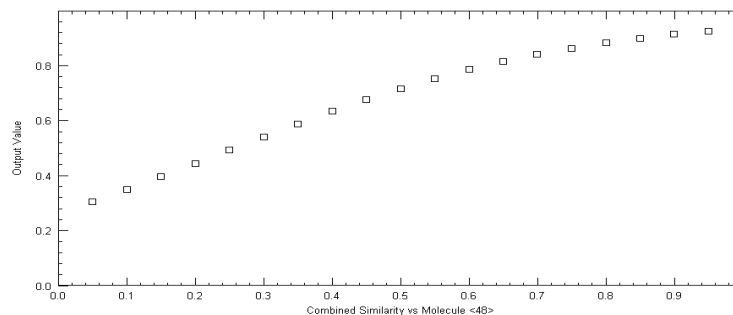
Dependencies plots were drawn in between each independent parameter versus biological activity **fig 32-34**.



**Figure 32:** Dependency plot between biological activity versus Charge similarity vs. molecule 22.



**Figure 33:** Dependency plot between biological activity versus Charge similarity vs. molecule 46.



**Figure 34:** Dependency plot between biological activity versus Combined similarity vs. molecule 48.

#### **Analysis of descriptors entered:-**

The Charge of molecule **22** and **46** are positively correlated with biological activity and these correlations were shown in the FFNN dependency plot Fig. 32-33 respectively. Consequently, the Combined molecule **48** is also positively correlated with biological activity, resulting in increase in the biological activity of the whole series as shown in Fig. 34.

#### **Conclusion:-**

All the results were discussed above, indicate that the using MLR, PLS and FFNN analysis with molecular descriptors belongs to amide-oxadiazole aniline series and generated highly robust QSAR models. Molecular parameters such as shape, lipophilic, and electronic architecture of amide-oxadiazole aniline analogues are considered to be important contributors to their biological properties and it can also be used for the designing of further new anti-diabetic compounds with more potency and reduced mechanism based side effect of traditional anti-diabetic agents. The results obtained from same series in similarity analysis (charge, combined, refractivity and combined) also support the MLR, PLS and FFNN results. This information about the 2D-requirement of the compound is of great value for the effective design of new DGAT derivatives of pharmaceutical importance. So the enhancement in the energy of the molecule will be the strategy to improve the activity.

#### **Acknowledgment:-**

Computational resources were provided by Banasthali University, and the authors are thankful to the Vice Chancellor, for providing all the facilities.

#### **Conflict of interest:-**

The authors report no conflict of interest. Only the authors are responsible for the contents and writing of paper.

#### **References:-**

1. Abdulfatai, B. O., Olusegun, A. O. and Lateefat, B. O. (2012): Type 2 Diabetes Mellitus: A Review of Current Trends. *Oman. Med. J.*, 27(4): 269-273.
2. Ross, S. A., Gulve, E. A., and Wang M. (2004): Chemistry and biochemistry of Type 2 diabetes. *Chem. Rev.*, 104 (3): 1255-1282.
3. Haslam, D. W., James, W. P. (2005): Obesity. *Lancet.*, 366(9492): 1197-1209.
4. Eric Yen, C.-L., Scot, J. S., Koliwad, S., Harris, C., Robert, V. F. (2008): DGAT enzymes and triacylglycerol biosynthesis. *J. Lipid Res.*, 49(11):2283-2301.
5. Mougenot, P., Namane, C., Fett, E., Camy, F. and Rommel, D-F. (2012): Thiadiazoles as new inhibitors of Diacylglycerolacyltransferase type 1. *Bioorg. Med. Chem. Lett.*, 22(7):2497-2502.
6. Qian, Y., Stanley, J. W., Ahmad, M., Wai-Hing, C. A. and Firooznia, F. (2009): Discovery of Orally Active Carboxylic Acid Derivatives of 2-Phenyl-5-trifluoromethyloxazole-4-carboxamide as Potent Diacylglycerol Acyltransferase-1 Inhibitors for the Potential Treatment of Obesity and Diabetes, *J. Med. Chem.*, 52(6): 2433-2446.
7. Topliss, J.G. (1993): Some observation on classical QSAR. *Perspect. Drug. Discov. Des.*, 1(2): 253-268.
8. Hanch, C., Muir, R.M., Fujita, T., Maloney, P. and Geiger, E. (1963): Correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficient. *J. Am. Chem. Soc.*, 85(18):2817-2824.

9. William, M., Matthew, S. A., Alan, M. B., Susan, B. and Linda, K. B. (2012): Identification, optimization and *in vivo* evaluation of oxadiazole DGAT-1 inhibitors for the treatment of obesity and diabetes. *Bioorg. Med. Chem. Letts.*, 22(12): 3843-4206.
10. Paliwal, S., Seth, D., Yadav, D., Yadav, R. and Paliwal, S. (2011): Quantitative structure activity relationship analysis of pyrrolidine analogs for dipeptidyl peptidase IV antagonist. *J. Of Enzyme Inhibition and Med. Chem.*, 26(1): 129-130.
11. Acharya, C., Coop, A. and mackerel, A. D. (2011): Recent advances ion ligand based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr. Comput. Aided. Drug. Des.*, 7(1):10-22.
12. Paliwal, S., Das, S., Yadav, D., Saxena, M. and Paliwal, S. (2011): Quantitative structure activity relationship analysis of N6-substituted adenosine receptor agonist as potential antihypertensive agents. *Med. Chem. Res.*, 20: 1643-1649.
13. Paliwal, S., Narayan, A. and Paliwal, S. (2009): Quantitative structure activity relationship analysis of dicationic-diphenylisoxazole as potent anti-trypanosomal agents. *QSAR Comb. Sci.*, 28(11-12): 1367-1375.
14. Paliwal, S. K., Pal, M. and Siddiqui, A. A. (2010): Qunatitative structure activity relationship analysis of angiotensin II AT<sub>1</sub> receptor antagonists. *Med. Chem. Res.*, 19: 475-489.
15. Spanier, A. M., Okai, H. and Tamura, M. (1993): Food Flavour and Safety, Molecular Analysis and Design. ACS. Symposius Series., 528: 103-104.
16. Kubinyi, H. (1997): QSAR and 3D-QSAR in drug design part 1: methodology. *Drug Discov. Today.*, 2: 457-467.
17. Cramer, R. D. (1993): Partial Least Squares (PLS): its strength and limitations. *Perspect. Drug. Discov. Des.*, 1: 269-278.
18. Lipinski, C. A., Lombardo, F., Dominy, D. W., Feeney, P. J. (2001): Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.*, 46(1-3):3-26.
19. Karelson, M. (2000): Molecular descriptors in QSAR/QSPR. John Wiley and Sons Ltd, first edition.
20. Zdrazil, B., Kaiser, D., Kopp, S., Chiba, P., and Ecker, G. F. (2007): Similarity-based descriptors (SIBAR) as tool for QSAR studies on PGlycoprotein inhibitors: Influence of the reference set. *QSAR Comb. Sci.*, 26: 669–678.