4. RESULTS AND DISCUSSION

To fulfil the aims and objectives two different QSAR models were developed and systematically validated using different validation techniques to understand the structural requirements for ligands as aurora A kinase inhibitors. First model involved the use of 51 imidazo[1,2-*a*]pyrazine derivatives reportrd previously by Merck Research Laboratory[76-78] for generation of 3D-QSAR model using a dock based conformation for alignment of the ligands. Second model pertained to the development of a 4D-QSAR model for a series of 31 benzo[*e*]pyrimido[5,4-*b*][1,4]diazepin-6(11*H*)-one derivatives previously reported[79] in the literature as aurora A kinase Inhibitors. These developed models were validated in a systematic way by using different validation parameters and the obtained statistical values were indicative of highly predictive models. The resultant models offered vital information to understand the structural features required to modify and develop new potential Aurora kinase inhibitors.

The results obtained from the two developed QSAR models are discussed here in two sub-sections.

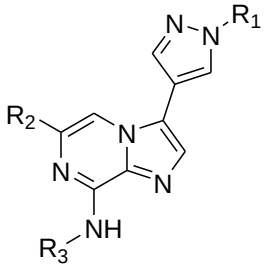
4.1. Identification of structural requirements for imidazo[1,2-*a*]pyrazine derivatives as aurora A kinase inhibitors by using docking based conformation and systematic validation of the developed model.

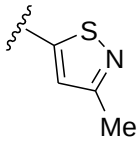
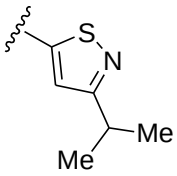
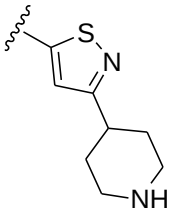
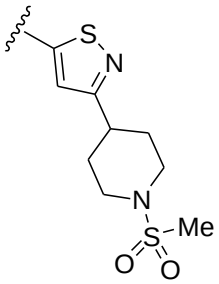
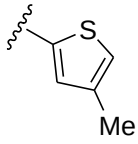
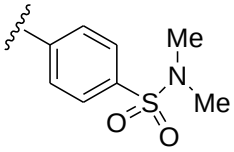
In the present study, structural requirements for a series of imidazo[1,2-*a*]pyrazine derivatives as inhibitor of aurora A kinase have been characterized using 3D-QSAR technique from the data taken from three different communications reported previously by Merck Research Lab.[76-78] All the structures used for this study are shown in **Table 1**. The entire study was carried out with the aid of modules from AutoDock[80] and Tripos-Sybyl 7.0 (SYBYL 7.0; Tripos Inc.).[81] AutoDock was used to find the bioactive conformation of the most active compound (**34**) for the alignment of other molecules from the series under study. Comparative molecular field analysis (CoMFA) and Comparative molecular similarity indices analysis (CoMSIA) models were developed with the aid of Sybyl 7.0.

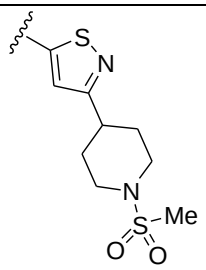
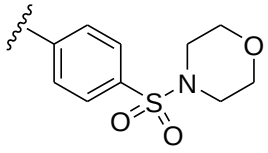
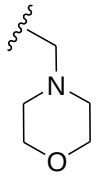
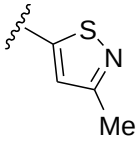
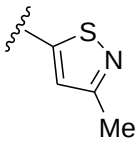
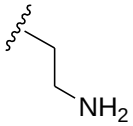
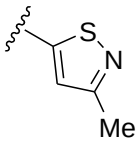
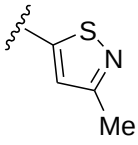
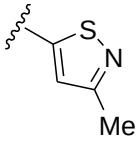
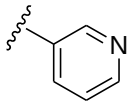
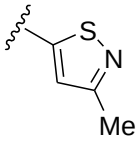
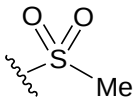
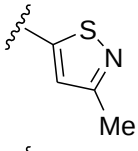
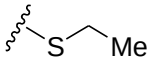
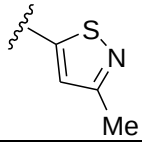
Minimum energy conformation may not be the bioactive one, therefore to get bioactive conformations the most active compound (**34**) was docked with aurora A kinase (PDB: 3MYG)[82]

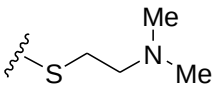
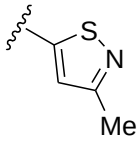
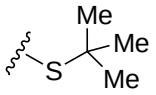
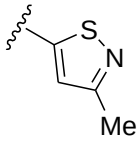
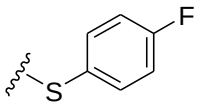
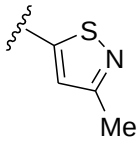
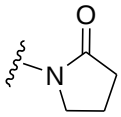
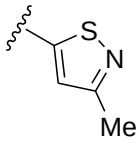
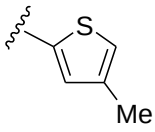
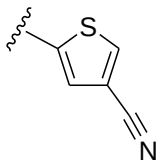
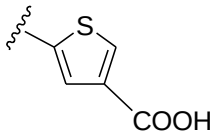
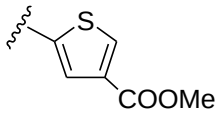
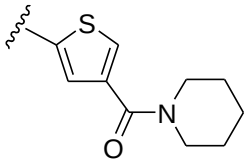
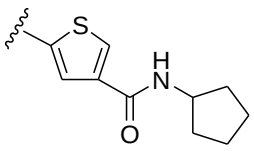
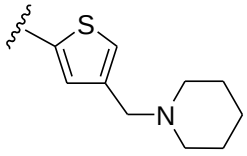
in AutoDock4. The binding energy for the most stable complex was found to be -9.46 kcal/mol. The resultant lowest energy conformation, based on this validated docking study, was selected as

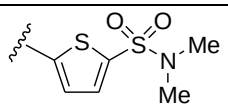
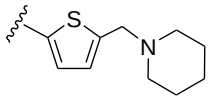
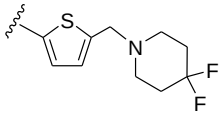
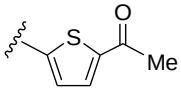
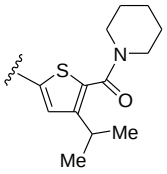
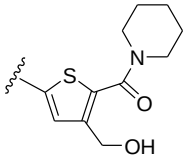
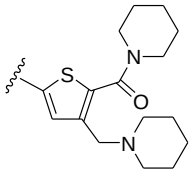
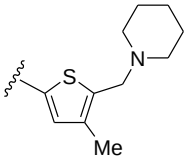
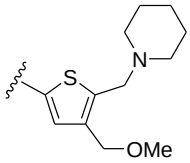
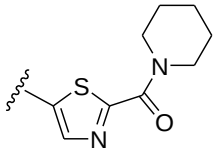
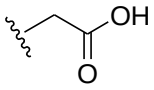
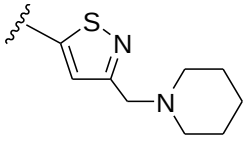
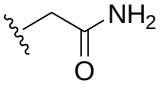
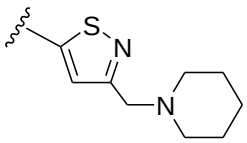
Table 1: Observed and predicted activities of training and test set compounds with their structures.

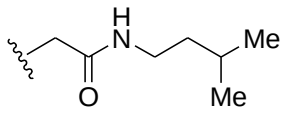
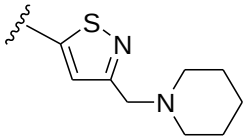
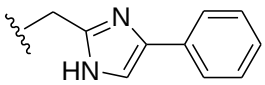
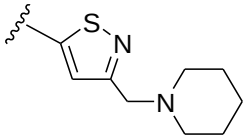
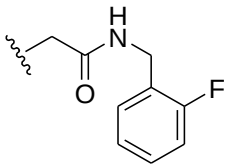
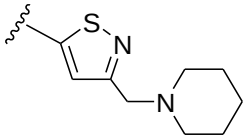
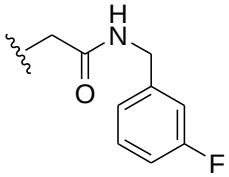
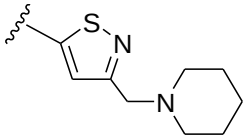
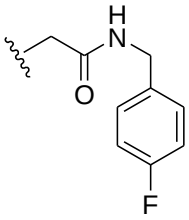
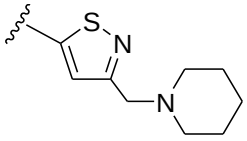
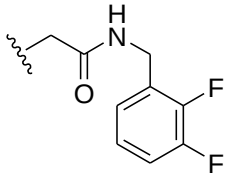
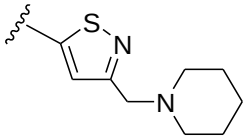
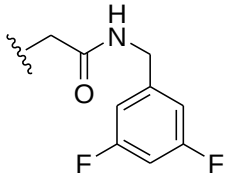
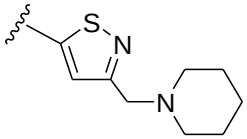
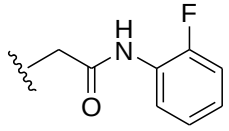
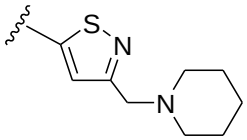
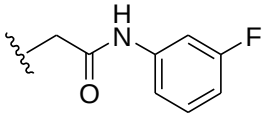
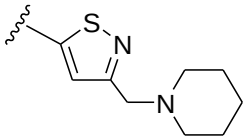
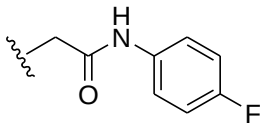
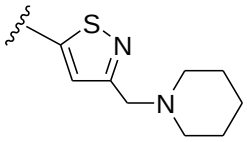


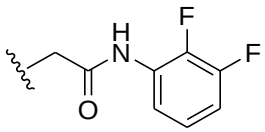
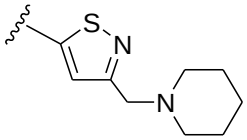
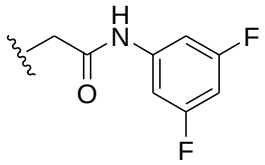
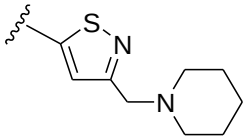
Sr. No.	R1	R2	R3	pIC_{50} (Actual)	pIC_{50} (Predicted)	
					CoMFA	CoMSIA (SEHD)
1	-Me	-H		6.24	6.13	6.12
2*	-Me	-H		6.19	5.91	6.17
3	-Me	-H		5.80	5.44	5.74
4	-Me	-H		6.77	7.24	7.06
5	-Me	-H		5.60	5.71	5.76
6	-Me	-H		5.43	5.41	5.42

7	-H	-H		8.22	7.94	8.29
8	-H	-H		7.09	7.09	7.05
9*	-H			7.30	7.39	7.54
10	-H	$-CF_3$		7.57	7.55	7.70
11	-H			7.08	7.01	7.10
12	-H			7.77	7.50	7.63
13	-H			7.34	7.28	7.75
14	-H			8.10	7.86	7.78
15	-H			6.22	6.23	6.49
16	-H			7.66	7.25	7.21

17	-H			7.09	7.24	6.91
18	-H			6.98	7.23	7.24
19	-H			7.07	7.27	7.11
20	-H			7.85	7.90	7.96
21*	-H	-Me		7.52	6.58	7.10
22*	-H	-Me		7.13	6.37	6.51
23	-H	-Me		5.52	5.54	5.58
24	-H	-Me		5.97	6.08	6.10
25	-H	-Me		6.02	6.03	5.81
26*	-H	-Me		5.52	6.49	5.87
27*	-H	-Me		8.30	7.43	8.00

28	-H	-Me		8.30	8.28	7.72
29*	-H	-Me		7.85	7.98	7.92
30	-H	-Me		7.80	8.00	8.01
31	-H	-Me		7.89	7.72	7.59
32*	-H	-Me		7.40	7.04	7.88
33	-H	-Me		7.08	7.03	7.31
34	-H	-Me		8.70	8.65	8.73
35	-H	-Me		7.85	7.84	7.74
36	-H	-Me		7.51	7.57	7.41
37	-H	-Me		6.75	7.07	7.03
38		-Me		5.74	5.89	5.80
39		-Me		5.80	5.76	5.78

40		-Me		5.75	5.66	5.68
41		-Me		6.23	6.23	6.11
42		-Me		7.36	7.47	7.51
43		-Me		7.52	7.47	7.45
44		-Me		7.34	7.41	7.45
45		-Me		7.35	7.44	7.50
46		-Me		7.66	7.57	7.53
47		-Me		6.70	6.61	6.57
48*		-Me		6.47	6.81	7.10
49		-Me		7.28	6.70	6.67

50*		-Me		6.91	6.93	7.11
51		-Me		5.91	6.54	6.44

* = test set compounds

the bioactive conformation. From the docking study, it was clearly observed that compound (**34**) showed good fitting into the active site hinge region of aurora A kinase. The pyrazole N-H showed H-bonding interaction with Asp-274, whereas the 8-amino N-H interacted with carboxylic –C=O and the N-1 of imidazo scaffold with the N-H of Ala-213 by hydrogen bonding, imparting high stability to the ligand receptor complex (**Figure 5**).

This bioactive conformation was considered for the alignment of all of the remaining ligands from the series under study for the development of a 3D-QSAR model (**Figure 6**).

The developed 3D-QSAR model was systematically validated using different validation parameters. According to the validation criteria proposed by various authors, q^2, r^2_{ncv} and other validation parameters like r^2_{pred}, R_m^2 (overall), R_m^2 (LOO) and external R_m^2 (Test), Model Y-randomization, R^2_p (test) based on Y-randomization,[83] $\{[(r^2-r^2_0)/r^2] < 0.1$ or $[(r^2-r'^2_0)/r^2] < 0.1\}$, k and k',[84,85] sensitivity, specificity, accuracy, PPV or precision, NPV and MCC values[86] were calculated for the best developed model (CoMSIA-SEHD). All the parameters were found to be within the acceptable limits. These resultant values along with individual contributions of steric (S), electrostatic (E), hydrophobic (H), and hydrogen bond donor (D) features are summarized in **Table 2**. The predicted activity by both CoMFA and best CoMSIA (SEHD) model is tabulated in **Table 1**. The graphs of the actual versus predicted activities with intercept and without intercept for the training set and test set compounds are shown in **Figure 7**.

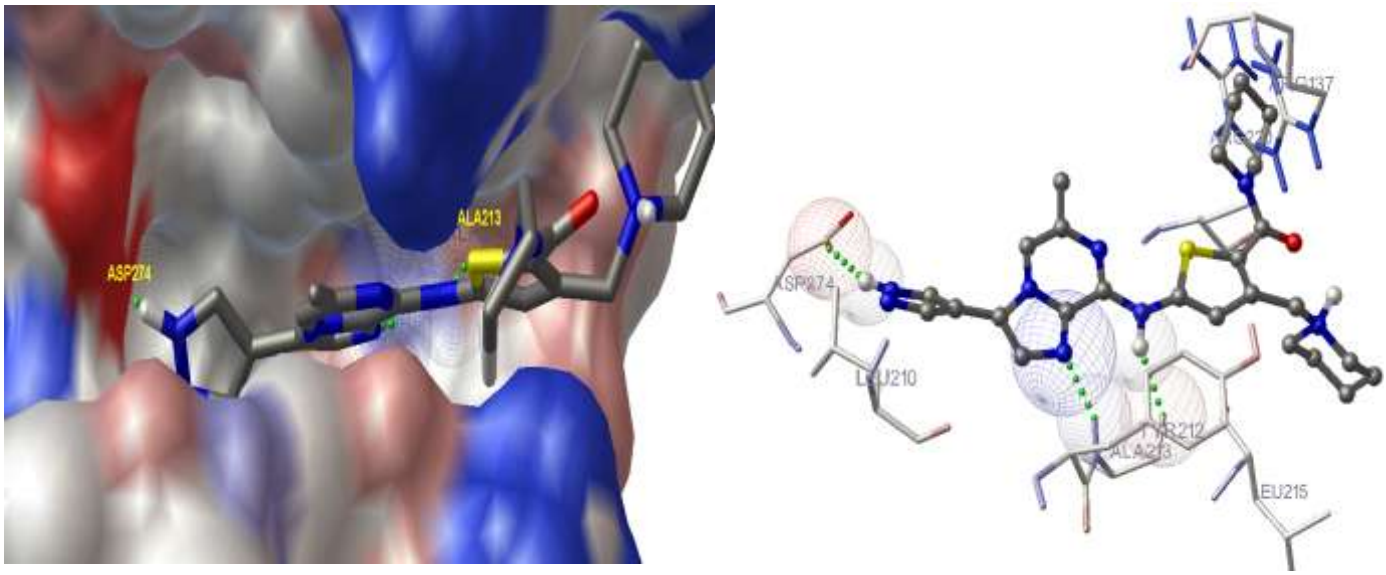


Figure 5: Orientation and binding mode of the most active compound (**34**) within the active site of aurora A kinase (PDB Code: **3MYG**).

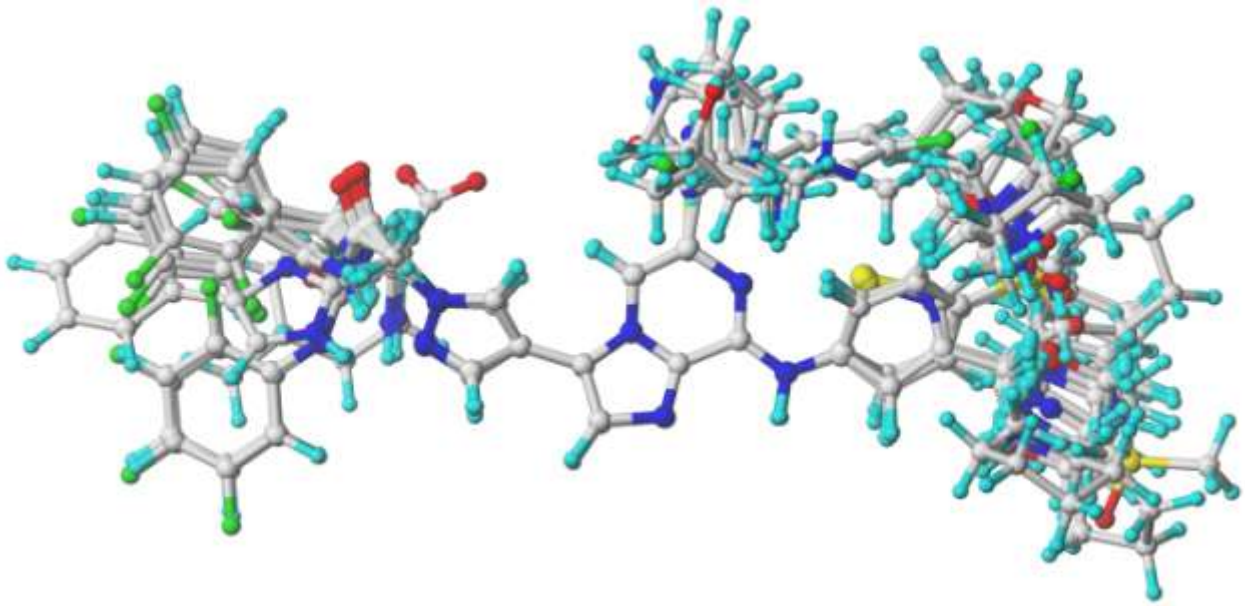


Figure 6: 3D alignment of compounds in the dataset.

Table 2: Summary of CoMFA and CoMSIA results with the results of statistical validation.

Parameter	CoMFA	CoMSIA	Parameter	CoMFA	CoMSIA (SEHD)
q^2	0.537	0.586	**Model Y-randomization (10)**		
ONC	6	5	Range of r^2_{ncv} (randomized)	-----	0.186 to 0.422
r^2_{ncv}	0.939	0.926	Range of r^2_{cv} (randomized)	-----	-0.58 to 0.236
SEE	0.236	0.255	R^2_p	-----	0.742
F_{value}	86.797	88.081	k	-----	1.005
r^2_{bs}	0.959	0.955	k'	-----	0.991
SD_{bs}	0.020	0.018	Sensitivity	-----	0.931
r^2_{pred}	0.426	0.752	Specificity	-----	0.773
	Contributions		Accuracy	-----	0.863
S	0.548	0.125	PPV	-----	0.843
E	0.452	0.362	NPV	-----	0.895
H	-----	0.269	MCC	-----	0.721
D	-----	0.244	$(r^2-r^2_0)/r^2$	-----	0.074
r^2_m (test)	-----	0.595	$(r^2-r'^2_0)/r^2$	-----	0.001
r^2_m (LOO)	-----	0.700			
r^2_m	-----	0.807			

q^2: LOO cross validated squared correlation coefficient; r^2_{ncv}: non-cross validated squared correlation coefficient; SEE: standard error of estimate; F_{value}: fischer test ratio; r^2_{bs}: bootstrapping coefficient of determination; SD_{bs}: bootstrapping standard deviation; k and k': slope for regression line without intercept for actual vs predicted and predicted vs actual; PPV: positive prediction value; NPV: negative prediction value; MCC: Matthew's correlation coefficient.

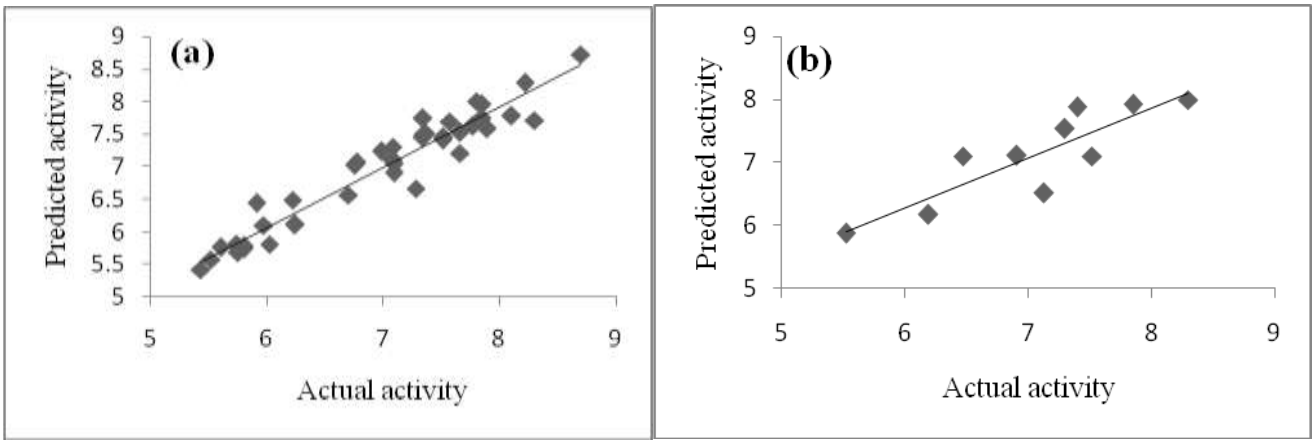


Figure 7: Graph of the actual versus predicted activities for training set (a); and test set (b) from the best predictive CoMSIA (SEHD) model.

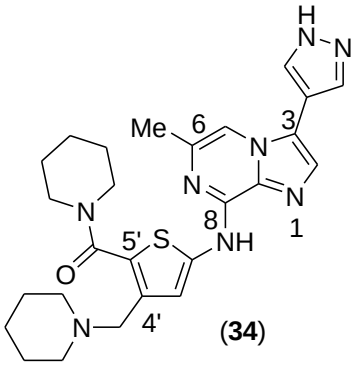
For the analysis of the best developed model (CoMSIA-SEHD) statistical parameters like, 'sensitivity' for describing part of the compounds which were correctly predicted as 'actives' out of the total active compounds, 'specificity' for recounting part of the compounds which were correctly predicted as 'non-actives' out of total non-active compounds, 'accuracy'

for indicating that the compounds were correctly divided and predicted as true actives or true inactives, positive prediction value (PPV) or 'precision', negative prediction value (NPV) and Matthew's correlation coefficient (MCC) were determined. For the assessment of sensitivity, specificity, accuracy, PPV, NPV and MCC, the dataset was classified into inactives and actives with a threshold of pIC_{50} value at 7.0. Compounds with $pIC_{50} \leq 7$ were considered as inactive and compounds with $pIC_{50} > 7$ were considered as active against aurora A kinase. The threshold value of pIC_{50} may vary from data to data and here the value of 7.0 is limited to this particular data set only. To apply the above parameters correctly in the study, the used data set had to be divided into two parts with almost equal number of compounds in each sub-set. One has to consider one part as actives and the other as inactives. By dividing the dataset in this way it was to be observed whether the predicted activities were true positives/true negatives/false positives/false negatives, matching the experimental biological activities. Dividing the data into equal halves is the best approach to avoid bias observations. Many experiments were performed and it was concluded that dividing the data set with equal number of molecules in each set gave the most accurate and unbiased results. If one changes the proportion of compounds in actives or inactives, then the percentage of sensitivity, specificity, accuracy, PPV, NPV and MCC would vary with every change, that means one can obtain good results intentionally regarding sensitivity, specificity, accuracy, PPV, NPV and MCC by dividing the data set unequally. Therefore equal division is the best approach that gives unbiased results. This division is for the purpose of validation of the predicted activity, whether the activity gets predicted truly or falsely i.e. whether the activity prediction is true or false. As per the division mentioned above compounds with $pIC_{50} \leq 7$ were considered as inactives and compounds with $pIC_{50} > 7$ were considered as actives against aurora A kinase enzyme. The condition of $\{[(r^2-r^2_0)/r^2] < 0.1$ or $[(r^2-r'^2_0)/r^2] < 0.1\}$ were also determined and were found to be within limits. Here r^2_0 and r'^2_0 are squared correlation coefficients for the regression line without intercept for actual versus predicted and predicted versus actual activity respectively.

According to the division on the basis of threshold value, accuracy of the best model indicated that 86.3% compounds were correctly classified and predicted as 'true active' or 'true inactive'; sensitivity indicated that 93.1% compounds were correctly predicted 'actives' out of the total 'actives' whereas, specificity indicated that 77.3% compounds were correctly predicted as 'non-actives' out of the total 'inactives'. MCC was used as a measure of quality of binary

classification. It measures the quality with the scale of +1, 0,-1 indicating perfect prediction, no better than random prediction and total disagreement between prediction and observations. Here, MCC value of 0.721 indicated a very good prediction. All the statistical values complied well with the requirement of a robust predictive model.

For the development of the best model, both of the CoMFA and CoMSIA techniques with different combinations of S (steric), E (electrostatic), H (hydrophobic), D (hydrogen bond donor) and A (hydrogen bond acceptor) parameters were tried and their predictive abilities were evaluated. In CoMSIA, it was found that, omitting either hydrophobic or hydrogen bond donor parameters or both resulted in poor predictive models. Exclusion of hydrogen bond acceptor parameter showed improvement in the predictive ability of the model. Presence of steric and electrostatic parameters supported the predictive model. Among all the combinations, CoMSIA-SEHD model was found to be a highly predictive model that has been discussed here. The 3D contour maps generally provide plenty of understanding of essential structural features required for biological activity. **Figure 8** (**a** and **b**) shows CoMSIA steric and electrostatic, where as **Figure 8**(**c** and **d**) describes hydrophobic and hydrogen bond donor contour maps respectively. Compound (**34**) being the most active derivative served as the reference molecule.



(**34**)

In **Figure 8**(**a**), sterically favored green contour and sterically disfavored yellow contours over the 5' and near 4' positions of thiophene ring of compound (**34**) explained its high activity. Orientation of the bulky group on position 4' of thiophene ring responsible for higher or lower biological activity was depicted well. The existence of green as well as yellow contours near this position suggests the sterically favored and disfavored substituents. The yellow contours at this position are due to the steric protrusion of fragments of amino acids Arg-137, Lys-224 and Pro-214 of the receptor. In the active compounds (**7**, **27** and **34**) the bulky grouping at 4' position was

facing the green contour whereas in the inactive compounds (**24, 25** and **26**) the 4' bulky groups were oriented towards yellow contours due to steric interference. The presence of yellow contour near pyrazole ring explained the lower activity of compounds (**1, 5, 38** and **51**) having steric groups which could interfere with amino acids Asp-274, Lys-162 and gatekeeper residue Leu-210. Further, compounds (**7**, **27** and **34**) were found to be more active because of absence of steric groups on pyrazole ring. The appearance of yellow contour at position 6 described the presence of less bulky steric groups that improved the activity of compounds (7, **27** and **34**). **Figure 8**(**b**) described the electrostatic contour maps, where blue contours indicated the presence of electropositive groups and red contours the presence of electronegative groups on the ligand, for improved activity. High activity of compound (**34**) was well explained by the existence of blue contour over quaternary nitrogen of 4' methyl piperidinyl group and red contour over the carbonyl linker between thiophene and 5' piperidinyl group. Here, C=O group of Pro-214 and oxygen of Tyr-212 were observed at blue contour as counterpart on the receptor, whereas the existence of red contours were very well supported by the presence of positively charged N of Arg-137 and Lys-224. Poor activity of compounds (**24, 25** and **26**) were explained by the orientation of carbonyl group in electropositive blue contour which might interfere with negative part of Pro-214 and Tyr-212, and thus were found to be less active.

Figure 8(**c**) explains the hydrophobic contour maps, where magenta color represents the presence of favored hydrophobic groups and white color contours represent the presence of favorable hydrophilic groups. In **34**, the orientation of 5' carbonyl group towards hydrophilic contour and 4' $-CH_2-$ group towards hydrophobic contours explained the high activity of compound (**34**). Here, hydrophobic contour in the CoMSIA model was explained well due to the presence of hydrophobically favoured counterparts of Leu-215 and Lys-224 amino acids on the receptor. The amide and ester groups in compounds (**24, 25** and **26**) found to be aligned with hydrophobic region of the receptor and hence showed less activity. Whereas, the presence of white contour near the 6 position explained the requirement of hydrophilic or less hydrophobic small groups at this position for activity as this part faces the solvent accessible area of the receptor. **Figure 8**(**d**) represents the hydrogen bond donor contour map, where cyan color explains the presence of hydrogen bond donor group favorable for activity and yellow color for unfavorable. Compounds (**7**, **27** and **34**) were active and showed hydrogen bond donor 1*H*-pyrazole group near the cyan contour. Docking results of compound (**34**) with aurora A kinase in

AutoDock supported this observation, where N-H group showed hydrogen bonding with Asp-274, whereas, the less active compounds (**5, 38** and **40**) failed to show this feature.

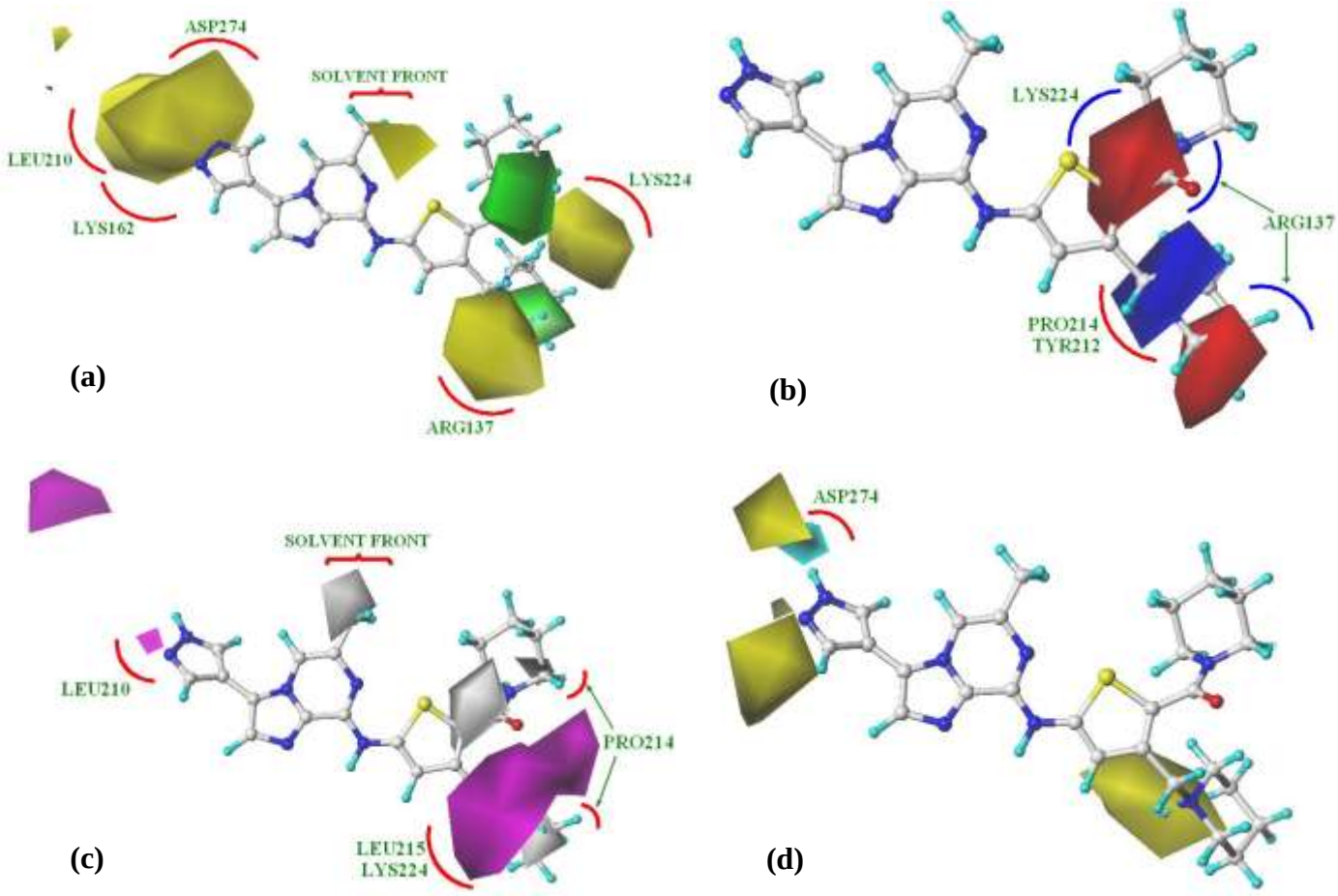


Figure 8: Comparative molecular similarity indices analysis contour maps (SEHD) with compound (**34**). (**a**) sterically favored green and disfavored yellow, (**b**) blue electropositive and red electronegative; (**c**) hydrophobic favored magenta and disfavored white, (**d**) cyan favored as hydrogen bond donor and yellow as disfavored.

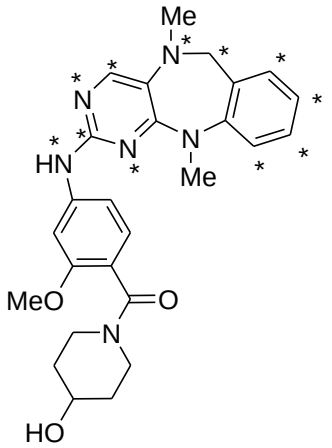
4.2. Development of 4D-QSAR model for the investigation of essential structural requirements for Benzo[*e*]pyrimido[5,4-*b*][1,4]diazepin-6(11*H*)-one derivatives as aurora A kinase inhibitors and methodical validation of the developed model.

Till date very few studies using 4D-QSAR technique have been reported and to the best of our knowledge none of these is on aurora kinase inhibitors. This work pertains to the development of a 4D-QSAR model for a series of benzo[*e*]pyrimido[5,4-*b*][1,4]diazepin-6(11*H*)-one derivatives as aurora A kinase inhibitors and its methodical validation. The extent of validation utilized here for a 4D-QSAR model remains unreported till date. The 4D-QSAR approach is a grid-based technique first proposed by Hopfinger *et al*.[87] As compared to conventional 3D-QSAR, this method uses a fourth dimension i.e. conformational flexibility.

In this work, a 4-D-QSAR model using an LQTA-QSAR approach[88] with previously reported 31 derivatives of benzo[*e*]pyrimido[5,4-*b*][1,4]diazepin-6(11*H*)-one[79] as potent aurora A kinase inhibitors has been created. Instead of single conformation, the conformational ensemble profile (CEP) generated for each ligand by using trajectories and topology information retrieved from molecular dynamics (MD) simulations from GROMACS package[89] were aligned and used for the calculation of intermolecular interaction energies at each grid point. The descriptors generated on the basis of these Coulomb and Lennard-Jones potentials as independent variables were used to perform a PLS analysis using biological activity as dependent variable. A good predictive model was generated with 9 field descriptors and 5 latent variables (LV). The model showed $Q_{LOO}^2 = 0.718$; $R^2 = 0.915$ and $R_{pred}^2 = 0.839$. This model was further validated systematically by using different validation parameters. In this work we used receptor independent LQTA (*Laboratório de Quimiometria Teórica e Aplicada*)-QSAR approach to develop 4D-QSAR model and report interesting findings which could be helpful for the design of new active derivatives.

LQTA-QSAR is based on making of conformational ensemble profile (CEP) for each compound instead of using a single conformation, and calculating the 3D descriptors for a given set of compounds using Coulomb and Lennard-Jones potentials. The geometry of individual 3D models of the compounds from the data set was energy minimized by using HF/3-21G basic set

level and the electrostatic partial atomic charges (ChelpG) were calculated in Gaussian[90] and used for the determination of Coulombic interaction energy descriptors. The output files from the Gaussian were converted into .mol2 files using Open Babel 2.3.1,[91] and were then submitted to the online server PRODRG[92] to generate geometry (.gro) and topology (.top) files. Thus, the resulting files were used as input data for GROMACS for molecular dynamics simulation to generate CEP. Considering the explicit aqueous medium (SPC/E water models), MD simulations were performed. In MD each molecule was optimized using steepest descent and conjugate gradient where the energy minimization convergence criterion was 50 N of the maximum force applied to the atoms in the identified systems, and the volume was balanced using stepwise heating of the system. The heating system followed steps of 50, 100, 200, and 350 K for 100 ps in each step. The system was subsequently cooled to 300 K and MD simulation for 500 ps was carried out. Trajectory files were saved every 2 ps of simulation. The conformations generated for each compound after a simulation were arranged in the same .gro file, all the conformations were aligned using atom based alignment and these data were used for the generation of QSAR models. All molecules were aligned on compound (**22**) and here, * indicates the atoms selected for alignment.



(**22**)

AutoDock4 was used to generate the coordinates of a virtual cubic grid. For the present study, a grid of 28x28x28 $Å^3$ size was created for the generation of Coulomb and Lennard-Jones descriptors. This grid box was large enough to occupy all the conformations of the compounds in the dataset. Each grid point in this case had a spacing of 1 Å. As 4D-QSAR is a conformation/alignment independent method, all the generated conformations of each ligand should be occupied within the generated grid. Therefore, care was taken while sizing the grid by

using the GUI in AutoDock, and the distance between the CEP coordinates and 3D lattice border was kept at least 5 Å to ensure the measurement of correct interactions. Different types of atoms, ions or groups can be used as probes to determine interactions with the compound from the lattice points. The selected probe computes the steric and electrostatic 3D properties, also called descriptors/variables, for each individual grid point on the basis of Lennard-Jones and Coulomb potential functions. Here, for the generation of descriptors we used NH_3^+ as probe in LQTA grid with the coordinates from AutoDock generating 43,904 descriptors in all. Probes are different types of atoms, ions or functional groups. The probes reported for LQTA grid are based on the ff43a1 force field parameterization. Several reports are available to justify the use of single probe instead of multiple probes to generate good results. This also reduces the generation of unwanted excess data. The NH_3^+ probe used here corresponds to the protonated free amino terminal of peptides.

The generated descriptors from the LQTA grid were used to build a data matrix using MATLAB.[93] For the reduction of variables, initially the descriptors were divided into two groups, Lennard-Jones and Coulomb potentials. A 30 kcal/mol cut-off was applied to the data for data reduction. As no significant statistical information could be obtained from descriptors with coefficient values lower than 0.3, they were eliminated by calculating and comparing the correlation coefficient values in between each variable and biological activity. This resulted into 4,114 variables for Lenard-Jones and 19,111 for Coulomb potentials. Furthermore, for the purpose of data reduction independent variables were plotted against the dependent variables to get an idea about how they dispersed in the model space. Those variables that were not uniformly distributed with respect to the biological activity were eliminated from the variable set. At this point, 23 independent Lenard-Jones variables and 11,321 Coulomb potential variables were obtained. The 11,321 Coulomb potentials were subjected to ordered predictor selection (OPS) algorithm[94] which reduced the data to columns of 40 variables. In the end, the data were preprocessed using autoscaling technique to generate a QSAR model.

The remaining descriptors after systematic data reduction and optimization were considered as independent variables and biological activity ($p\text{IC}_{50}$) as the dependent variable. The data set was divided randomly into training set (25 compounds) and test set (06 compounds) over the entire range of biological activity considering that the test set should include low,

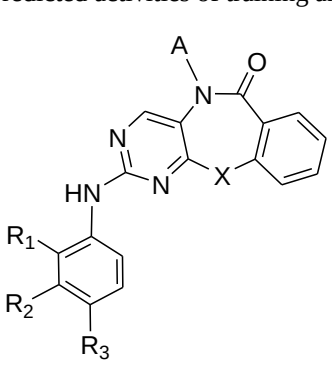
moderate and high activity compounds. QSAR models were generated using partial least squares (PLS) method[95,96] to determine cross validated coefficients. To ensure the quality of the developed model and its ability to predict the activity, different validation methods were implemented. The models were internally validated by means of coefficient of determination (R^2), standard error of calibration (SEC), F-ratio test, Q_{LOO}^2 and SEV. Further, the robustness of the model was analyzed with the help of Leave-N-Out (LNO)[97] cross validation (N = 1,2,...,7; ~30% of training set) and the N was repeated five times. Model Y-randomization[98] was carried out to rule out the chance correlation. Q^2 and R^2 values on Y-randomization ideally should not offer good values. Y vector randomization for 25 times yielded poor Q^2 and R^2 values in our case. R_p^2 based on Y-randomization was also applied to the generated model which applied a penalty to the models' coefficient of variance for the dissimilarity between mean coefficient of determination (R_r^2) of randomized models and coefficient of determination (R^2) of the non-randomized model. In order to have a better predictive potential of the model, a modified R^2 (R_m^2) was also determined.[83]

Further, the model was externally validated by predicting the activity of external set or test set of compounds and evaluated by means of coefficient of determination (R_{pred}^2), standard error of external prediction (SEP), and the average relative error (ARE_{pred}).[99] Additionally, different external validation techniques were implemented. Slopes for regression line with no intercept by exchanging X and Y axis (k, k')[84,85] were determined. Statistical techniques mentioned by Saha and Raghava[86] were also adopted, such as 'sensitivity value' to explain part of the compounds which are appropriately predicted as 'actives' out of the total active compounds, a 'specificity value' to recount part of the compounds which are correctly predicted as 'non-actives' out of the total inactive compounds, an 'accuracy value' to find out whether the compounds are correctly divided and predicted as true actives or true inactives, a 'positive prediction value' (PPV) or precision, a 'negative prediction value' (NPV), and a 'Matthew's correlation coefficient' (MCC) value for the generated model. To assess the sensitivity, specificity, accuracy, PPV, NPV and MCC, the dataset was classified into inactives and actives keeping the threshold of $p\text{IC}_{50}$ value at 7.8. Compounds with $p\text{IC}_{50} < 7.8$ were measured as inactive and compounds with $p\text{IC}_{50} \geq 7.8$ were taken as actives against aurora A kinase. The condition of $\{[(R^2-R^2_0)/R^2] < 0.1$ or $[(R^2-R'^2_0)/R^2] < 0.1\}$ were also implemented. Here R^2_0 and

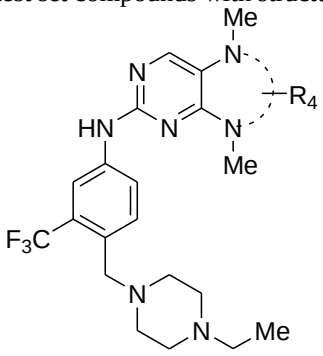
R'^2_0 are squared correlation coefficients for the regression lines without intercept for actual versus predicted and predicted versus actual activity respectively.

The *n*D-QSAR approaches are good enough to interpret the descriptors into activity prediction models. Here, the final model was obtained based on 9 descriptors and 5 latent variables (LV). These 5 LV accumulated around 85.70% of the information which was good enough to explain 91.54% of the variance and generate the model with low standard error of calibration (SEC = 0.22). F-ratio test was carried out to recognize the model which best fits the population from which the data was sampled. The F-value obtained here for this model was 63.37. At 95% confidence level, this F-value was much higher than the critical or tabulated value (F = 63.37 > 2.62). This value was also adequate enough to explain 71.8% of variance. Here, we observed a very low standard error of validation (SEV = 0.40). The predictive residual sum of squares of cross validated ($PRESS_{val}$) procedure was 1.20 which was much smaller than the sum of squares of the experimental *p*IC_{50}, suggesting that the model was real and not due to a chance correlation. Overfitting and chance correlation were further checked by the model Y-randomization which was carried out for the developed model wherein the dependent variable (biological activity) was randomized 25 times and the randomized R^2 values were found to be in the range of 0.13 to 0.47. The randomized Q_{LOO}^2 was in the range of -5.29 to -0.21. Here, the intercept for $R(y,y_{rand})R^2$ = 0.24 was less than 0.3 and the intercept $R(y,y_{rand})Q_{LOO}^2$ = -2.9 was less than 0.05, thus indicating the absence of chance correlation (**Figure 9**). The LNO validation was performed for ~30% of the data of training set, i.e. N = 1 to 7 and was repeated five times. The deviation from Q_{LNO}^2 for each N was within 0.1. Here, Q_{LNO}^2 was 0.687 and the difference between Q_{LOO}^2 and Q_{LNO}^2 was only 0.03 suggesting that the model was reliable and robust (**Figure 10**). Further, some more validation parameters were used in this model, such as R_m^2, R_p^2 based on model Y-randomization, ARE_{pred}, $\{[(R^2-R^2_0)/R^2] < 0.1$ or $[(R^2-R'^2_0)/R^2] < 0.1\}$, k and k', sensitivity, specificity, accuracy, PPV or precision, NPV and MCC values and all the results were obtained within acceptable limits and are summarised in **Table 4**. The external/test set coefficient of determination for the model (R_{pred}^2 = 0.839) was high enough to suggest that the model has good prediction accuracy. The actual and predicted activities for the dataset compounds are mentioned in **Table 3**.

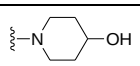
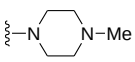
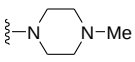
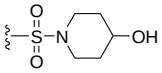
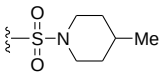
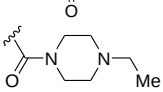
Table 3: Observed and predicted activities of training and test set compounds with structures of all compounds.

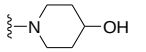
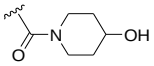
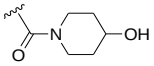
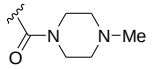
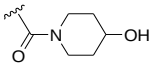
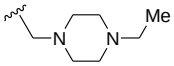
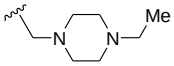
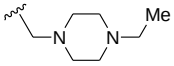
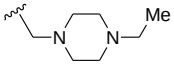
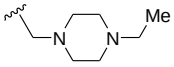
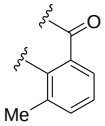
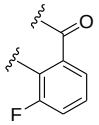
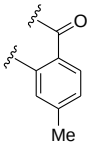
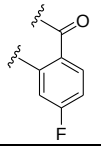


1 - 27



28 - 31

Sr. No.	X	A	R1	R2	R3	R4	pIC_{50}		Residuals
							Actual	Predicted	
1						---	8.2518	8.3967	-0.1449
2						---	5.9172	6.1796	-0.2624
3*						---	8.1426	8.3687	-0.2261
4						---	8.0555	7.9983	0.0572
5						---	7.7746	7.9668	-0.1922
6						---	8.4685	8.3631	0.1054
7						---	8.2291	8.2938	-0.0647
8						---	8.2146	7.7813	0.4333
9*						---	7.7351	7.5019	0.2332
10*						---	6.4907	6.3635	0.1272
11						---	7.0762	7.1226	-0.0464
12						---	8.0457	7.9554	0.0903
13						---	7.0614	7.7065	-0.6451
14						---	7.2740	7.1276	0.1464
15*						---	7.8268	8.5513	-0.7245
16						---	7.9706	8.0315	-0.0609
17						---	7.9244	8.0587	-0.1343

18		—H	—H	—H		---	7.0065	6.8154	0.1911
19		—Me	—H	—CF_3		---	8.2924	8.2700	0.0224
20*		—Me	—H	—Cl		---	8.4948	8.3506	0.1442
21		—Me	—H	—Me		---	8.4317	8.1779	0.2538
22		—Me	—H	—OMe		---	8.5850	8.1710	0.4140
23		—Me	—H	—CF_3		---	7.7746	7.9996	-0.2250
24*		—Me	—H	—CF_3		---	7.0877	7.0943	-0.0066
25		—Me	—H	—CF_3		---	6.9956	6.9787	0.0169
26	O	—Me	—H	—CF_3		---	6.3400	6.3626	-0.0226
27	S	—Me	—H	—CF_3		---	6.2716	6.1206	0.1510
28	---	---	---	---	---		7.9355	7.8300	0.1055
29	---	---	---	---	---		7.8728	8.0768	-0.2040
30	---	---	---	---	---		6.8356	6.8719	-0.0363
31	---	---	---	---	---		6.5257	6.4742	0.0515

*Test set compounds

As per the division of the dataset on the basis of threshold values (pIC_{50} < 7.8 were measured as inactives and compounds with $pIC_{50} \geq 7.8$ were taken as actives) the performance of the developed model was measured. The rationale behind this threshold value was to keep almost equal number of compounds from the data set into active and inactive sets as explained in the 3D-QSAR model. For the developed model, a 0.9032 accuracy was measured which indicated that 90.32% of the compounds were accurately classified and predicted as 'true actives' or 'true inactives'. The sensitivity value of 0.9375 indicated that 93.75% of compounds were correctly

predicted as 'actives' from the total 'actives' while, the specificity pointed out that 86.67% of compounds were exactly predicted as 'non-actives' out of the total 'inactives'. A value of 0.8824

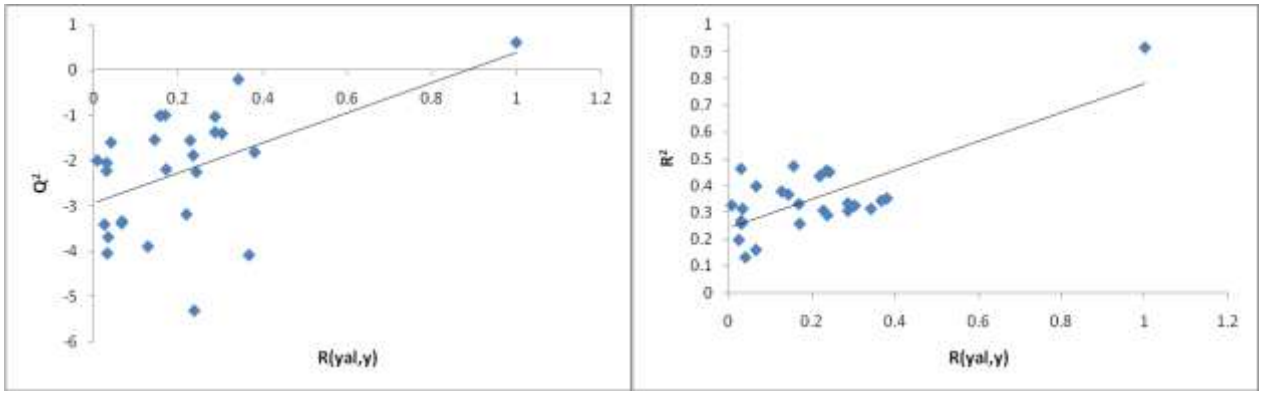


Figure 9: Plots for Q^2 and R^2 for 25 y-randomization.

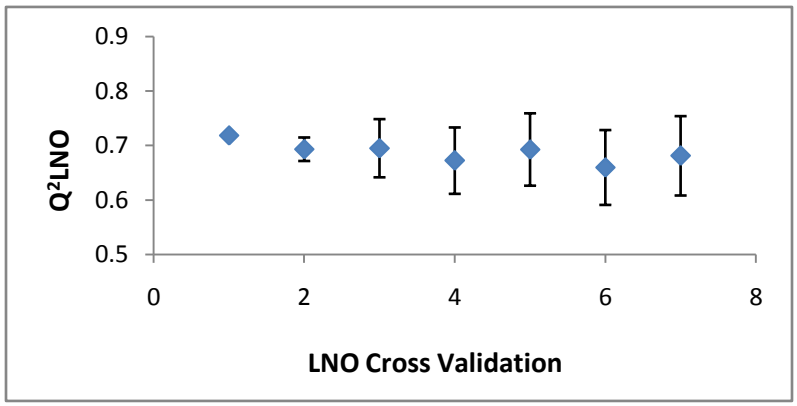


Figure 10: Plot of the LNO results for the dataset (N = 1,2,…,7. Five repetitions).

for precision was observed for the developed model. MCC was used to assess the excellence of binary classification. This measures the quality of the classification and prediction of the model on a scale of +1, 0, -1 which indicates a perfect prediction, no better than random prediction and total disagreement between prediction and observations, respectively. The MCC value for the present model was 0.8076 which showed that the model had very good prediction ability. All these results from various statistical methods point towards a robust predictive model. Graphs of actual versus predicted activity for the training set and test set molecules are shown in **Figure 11**. The equation obtained from the developed QSAR model is given below:

$$pIC_{50} = 0.196\ (22.22.12\ LJ{+}NH_3^+) + 0.259\ (19.24.13\ LJ{+}NH_3^+) - 0.013\ (18.25.14\ LJ{-}NH_3^+) - 0.172\ (26.21.15\ LJ{-}NH_3^+) - 0.185(25.21.16\ LJ{-}NH_3^+) - 0.501(20.25.13\ LJ{-}NH_3^+) - 0.013\ (19.26.13\ LJ{-}NH_3^+) - 0.265\ (20.21.12\ LJ{-}NH_3^+) - 0.584\ (17.16.15\ LJ{-}NH_3^+) + 31.93.$$

Table 4: Summary of 4D-QSAR results with the results of statistical validation and Autoscaled coefficients.

Validation Parameter	Results	Validation Parameter	Results
Q^2	0.718	k	0.99
F-value	63.37	k'	1.01
SEC	0.22	Sensitivity	0.937
SEV	0.40	Specificity	0.867
LV	5	Accuracy	0.903
LV accumulate	85.70	PPV	0.882
R^2	0.915	NPV	0.929
Q^2_{LNO}	0.687	MCC	0.808
R^2_{pred}	0.839	$(R^2-R^2_0)/R^2$	0.0059
SEP	0.334	$(R^2-R'^2_0)/R^2$	0.08
Average Relative Error (ARE_{pred})	2.118%	**Autoscaled Coefficients Data**	
R^2_m (test)	0.780	**Descriptors**	**Autoscaled Coefficients**
R^2_m (LOO)	0.833	17.16.15 LJ – NH_3^+	-0.827
R^2_m (overall)	0.803	20.21.12 LJ – NH_3^+	-0.429
Model Y-randomization (25 trials)		19.26.13 LJ – NH_3^+	-0.303
Range of Q^2 (randomized)	-5.29 to -0.21	20.25.13 LJ – NH_3^+	-0.426
Intercept	-2.9<0.05	25.21.16 LJ – NH_3^+	-0.270
Range of R^2 (randomized)	0.13 to 0.47	26.21.15 LJ – NH_3^+	-0.359
Intercept	0.24<0.3	18.25.14 LJ – NH_3^+	-0.324
R^2_p	0.700	19.24.13 LJ + NH_3^+	0.473
		22.22.12 LJ + NH_3^+	0.219

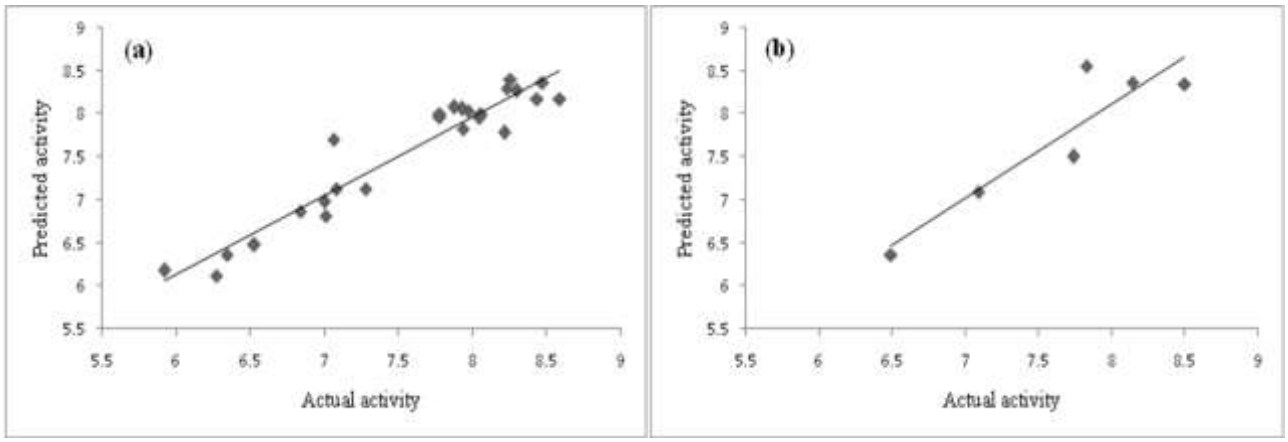


Figure 11: Graph of actual versus predicted activities for training set (**a**); and test set (**b**) from the best predictive model.

The contour representation and analysis of the developed model in relation to active and inactive compounds is explained here. The contour maps for the obtained coordinates for the interaction points are visualized by using UCSF chimera. The variables have been standardized by autoscaling and the autoscaled coefficients values are mentioned in **Table 4**. In **Figure 12** the green regions explain the steric interactions with respect to positive PLS regression coefficient,

whereas red region represents negative regression coefficient with steric interactions. The contour maps usually provide ample understanding of necessary structural features required for biological activity. Compound (**22**), being the most active, and compound (**2**), being the least active among the dataset compounds, are used as the reference molecules to explain the contours (**Figure 12**). The red contours (LJ– 18.25.14; 19.26.13 and 20.25.13) around the benzo ring of the nucleus explain that this region is sterically unfavourable, i.e. presence of bulky groups here leads to lesser active molecules. This was also supplemented by negative autoscaled coefficients (-0.324, -0.303, -0.426). The absence of such a group in compound (**22**) explains its better activity, while in compounds (**30** and **31**) the presence of methyl and fluorine respectively in this region (LJ– 18.25.14; 19.26.13 and 20.25.13) make them less active. The alignment of contours over the receptor active site (**Figure 13**) supported this observation. The regions of these red contours on the receptor showed steric hindrance because of Lys162 and Leu164. Whereas, in compounds (**28** and **29**) the 10-methyl or fluorine at position 10 of benzo part of the nucleus are away from the red contours (LJ– 18.25.14; 19.26.13 and 20.25.13) and thus explained the presence of lower bulky groups showing good activity. Further, these contours explained the reason for lesser activity for compound (**2**), where the presence of methoxy group on the 1st position of the phenylamino group of the structure entered the red contour (LJ– 17.16.15) (negative sign autoscaled coefficient -0.827) and was thus found to be very less active, while the absence of this group at 1st position in compound (**3**) showed improved activity, i.e. presence of bulkier group impacts the activity negatively. In case of the most active compound, the presence of methoxy group on 2nd position of the phenylamino part of the structure is away from the red contour. This was also true in the alignment of ligand contours over the receptor active site. Here, the red contour represents steric hindrance because of Phe275. In this representation, the 2nd position of the phenylamino moiety finds the free space in the active site, whereas for the 1st position there is steric hindrance. This supports the higher activity of compound (**22**) and low activity of compound (**2**) and strongly supports the validation and accuracy of the model.

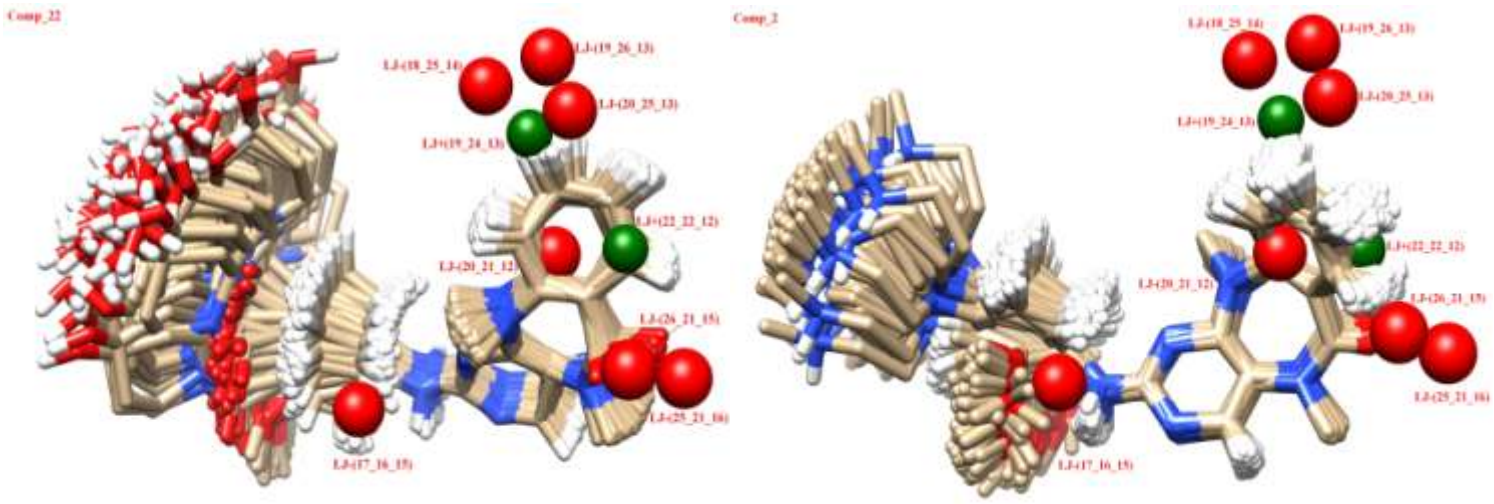


Figure 12: Contour map representation with regression sign of the most active compound (**22**) and the least active compound (**2**); here, Red contours explain sterically unfavourable regions whereas, Green contours explain sterically favorable regions.

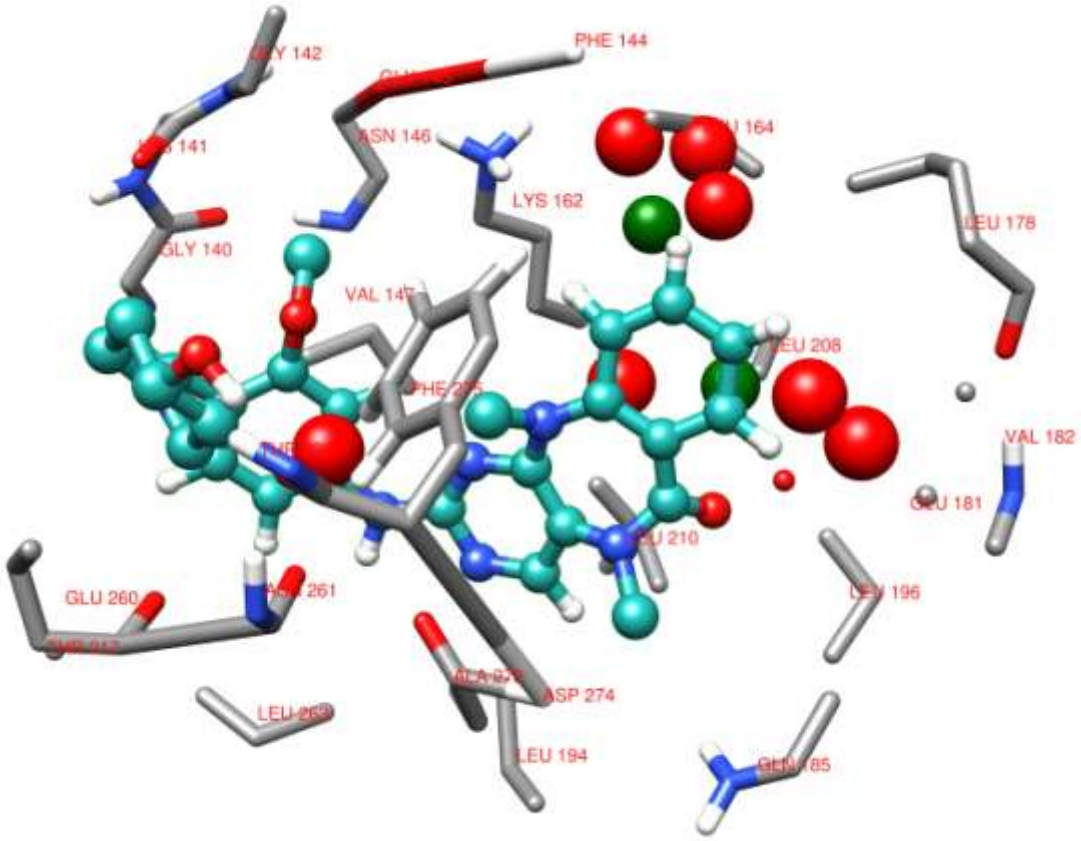


Figure 13: The most active compound (**22**) in the active site of the target receptor along with aligned contours. (Compound (**22**) ball and stick model Cyan color; Residue shown in stick format).