5. EXPERIMENTAL

This part deals with the development of 3D-QSAR and 4D-QSAR models using two different series of compounds as aurora A kinase inhibitors. Towards the successful development of these models, the followed methodology is described in this section.

Method followed for successful completion of 3D-QSAR model i.e. "identification of structural requirements for imidazo[1,2-*a*]pyrazine derivatives as aurora A kinase inhibitors by using docking based conformation to develop 3D-QSAR model and systematic validation of the developed model" is described in first half of the experimental section. In the second part, the method followed to develop 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 validation of the developed model is described. In the last part the validation techniques followed for both the models are summarized.

5.1. 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.

Docking studies

The Docking studies were performed by using AutoDock4. To validate the docking study, the co-crystallized ligand within the 3D structure of aurora A kinase (PDB Code: 3MYG) was re-docked into the active site of the enzyme. In the validation by re-docking, similar type of interactions between the ligand and the receptor should be observed as that of ligand receptor complex obtained by X-ray technique, and the re-docked ligand should not deviate from the original co-crystallized ligand by more than 2 Å root mean square deviation value (RMSD). If these two conditions are satisfied, then one can consider the generated grid for docking is accurate and can be used for docking of other ligands under study. In this study comparable interactions were observed between the re-docked ligand and the enzyme as was observed in the original co-crystallized structure i.e. related orientations of the groups and binding interactions with ALA213 and ASP274. The RMSD between the predicted conformation and the original conformation of compound as existed in the X-ray crystallographic structure was found to be 0.26 Å. The docking

of the most active ligand (**34**) in the active site of the target enzyme (PDB: 3MYG)[82] was performed using AutoDock4. For the ligand receptor complex, ten docking experiments were performed using Lamarckian genetic algorithm. The maximum number of energy evaluations of 25 million was applied for each docking experiment. The resultant lowest energy conformation based on this study, was selected as the bioactive conformation for the generation of 3D-QSAR model.

Data set and molecular alignment

A data set of 51 inhibitors imidazo[1,2-*a*]pyrazine derivatives as aurora A kinase inhibitors imidazo[1,2-*a*]pyrazine derivatives were retrieved from previous reports by Merck Research Laboratories.[76-78] Chiral compounds with undefined stereochemistry in the given data and compounds with non-discrete quantitative biological activity were dropped from the current study. The reported compounds showed wide structural variations and a broad range of activity. The activity values reported in IC_{50} (μM / nM) formats were converted to molar values and finally to pIC_{50} (-log IC_{50}) values and used as dependent variable in the QSAR model development. For all the structures, ionization status was determined at physiological *p*H. The dataset of compounds so generated at physiological *p*H were considered for model development.

All the molecular structures were constructed on the coordinates of the docked conformation of compound (**34**), considering their ionisation state, in Tripos-Sybyl7.0 [81] and prepared by using Gasteiger-Huckel charges. Compounds were aligned using atom fit function on 8-aminoimidazo[1,2-*a*]pyrazine part of the dock-based conformation. The dataset of 51 compounds was divided as 41 compounds in training set and 10 compounds in the test set on the basis of structural diversity and a broad range of biological activity.

3D QSAR model

For the development of CoMFA and CoMSIA models on the aligned database, the field energies were calculated in a grid with 2Å spacing and sp^3 C^+ as probe atom. For CoMFA, steric and electrostatic energy values a default value of 30 kcal/mol was kept; whereas, for CoMSIA the attenuation factor was held constant at 0.3. The parameters generated for CoMFA and CoMSIA were taken as independent variables and regression analysis was performed using partial least

squares (PLS) method to determine cross validated coefficient (q^2). The models were internally evaluated by leave-one-out (LOO) cross validation. The number of components was set at 6 and optimal number of components (ONC) out of it were determined. With this ONC the final non-cross validated PLS models were generated to determine non-cross-validated squared correlation coefficient (r^2_{ncv}), standard error of estimate (SEE) and F-test value to describe explained versus unexplained variables. To assess the predictive ability outside the model, the test set of 10 compounds was used to determine predictive squared correlation coefficient (r^2_{pred}).

5.2. 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 validation of the developed model.

For the development of 4D-QSAR model, a data set of 31 benzo[*e*]pyrimido[5,4-*b*][1,4]diazepin-6(11*H*)-ones was selected from the literature.[79] The nanomolar range of the biological activity for aurora A kinase in the form of IC_{50} was converted into molar concentration and then to negative log IC_{50} (pIC_{50}) and used as the dependent variable. The values so obtained are well spread in the range of almost four log units from 5.91 to 8.58.

LQTA-QSAR[88] 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.

Gaussian[90] was used to energy optimize the compounds. Open Babel[91] was used to convert the extension of files and PRODRG online server[92] was used to generate the geometry and topology files. GROMACS[89] the molecular dynamics tool was used to generate CEP for each ligand under study.

AutoDock4 was used to generate the coordinates of virtual cubic grid.[80] The probes reported for LQTA grid are based on the ff43a1 force field parameterization. The NH_3^+ probe was used here which corresponds to the protonated free amino terminal of peptides. The LQTA grid with coordinates from AutoDock was used to generate the descriptors that were used as independent parameters in QSAR development.

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. Further filters were applied to reduce the data and are discussed in Result and Discussion section to get better understanding of the model.

The data set was divided into training set (25 compounds) and test set (06 compounds) randomly over the entire biological activity range considering that the test set should include compounds with low, moderate and high biological activity. QSAR models were generated using partial least–squares (PLS) method 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 experimental procedure followed is discussed in more detail under Result and Discussion section to get better understanding of the developed 4D-QSAR model.

5.3. Additional validation parameters used to validate the developed model

Validation of any technique is essential to determine its versatile applicability. Here, different validation parameters adopted for validation of the developed models are described.

It is widely known that the R^2_{pred} predictive coefficient is used to explain the predictive ability of the developed model. The values of R^2_{pred} are mainly controlled by sum of squared differences between the observed values of the test set and the mean observed values of the training data set. This is inadequate to assess the reliability of a model for prediction of new compounds and thus may not be enough to explain external predictive accuracy of the model. Even a good coefficient of determination (R^2) between the observed and predicted values may not be actually good. The intercorrelation may give good numerical value, but the actual residual differences may be numerically large. Thus, to avoid the above mentioned two serious problems and to get better and accurate external predictive potential, an additional validation parameter termed as R_m^2 i.e. modified R^2 has been introduced.[83]

$$R_m^2\,(test) = R^2 \times (1 - \sqrt{R^2 - R_0^2})$$

Here R^2_0 is squared correlation coefficient value with intercept zero. The value of R_m^2 above 0.5 or nearer to the predicted R^2 value indicates that the model is truly predictive. Along with the test set, it can also be applied to the training set considering the data of training set obtained from Leave One Out (LOO) method. It is preferable to apply this method to the overall data set i.e. training and test set together, to suggest the best predictive model among all the other developed comparable models.

Another widely used validation technique is Y-randomization.[83] This technique is of two types: model randomization and process randomization. In model Y randomization, the Y entries are scrambled and new QSAR models are developed using same set of variables as that of unrandomized model. Whereas, in process Y randomization, the Y entries are scrambled and the variable selection is done again afresh from the entire data matrix. Ideally the coefficient values of randomized models should be less than the value of non-randomized model and it is always preferable if these values are below 0.5. The intercept on randomized values should be less than 0.3. One can follow randomization for checking $Q^2_{(LOO)}$ as well, here the intercept of randomized model should be less than 0.05.

To avoid the chance correlation or false positive observations one more validation technique has been introduced which uses the data obtained from Y randomization. This parameter is termed as R_p^2. It penalizes the model R^2 for the difference between squared correlation coefficient of non-randomized model and squared mean correlation coefficient from randomized models.

$$R_p^2 = R^2 \times \sqrt{R^2 - R_r^2}$$

Here, R_r^2 is squared mean correlation coefficient from randomized models. It is assumed that the R^2_p value above 0.5 is must for an acceptable model.

As it is well established that the q^2 is the internal predictive ability of a model determined most commonly by using leave one out (LOO) method. Though its value must be higher than 0.5, it is not necessary that this value is sufficient enough to have a high predictive power model. Some external validation parameter like determination of slope k and k' are useful in this case. Let us consider that y_i and $\tilde{y}_i$ are the actual and predicted activities of data set of the model. If we

plot y verses $\tilde{y}$ for a QSAR model, the regression line will bisect the axes in positive direction and can be expressed as, $y^r = a\tilde{y} + b$.

Here, $a = \frac{\sum(y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sum(\tilde{y}_i - \bar{\tilde{y}})^2}$ and; $b = \bar{y} - a\bar{\tilde{y}}$.

The bar line indicates the average (mean) value. For an ideal model the value for slope *a*, should be 1 and intercept *b* must be 0. A QSAR model may have high predictive ability if the value is close to ideal one. This means that the correlation coefficient *R* between the actual *y* and predicted $\tilde{y}$ activities should be near to 1 and regressions of *y* against $\tilde{y}$ or $\tilde{y}$ against *y* from the origin, i.e. $y^{r0} = k\tilde{y}$ and $\tilde{y}^{r0} = k'y$, respectively, must be differentiated by at least either *k* or *k*' close to 1. The slopes *k* and *k*' are calculated by using the following equations:

$$k = \frac{\sum y_i \tilde{y}_i}{\tilde{y}_i{}^2} \qquad\qquad k' = \frac{\sum y_i \tilde{y}_i}{y_i{}^2}$$

The $\{[(R^2 - R^2_0)/R^2] < 0.1$ or $[(R^2 - R'^2_0)/R^2] < 0.1\}$ criterion also supports the accuracy of the developed model. Here R^2_0 is coefficient value with intercept set to zero and R'^2_0 is coefficient value with intercept set to zero by altering x and y axis on the graph.[84,85]

Prediction of the activity truly is a very important requirement for development of an accurate QSAR model. In order to fulfil this requirement, Saha and Raghava[86] have reported some parameters like the 'sensitivity' value for describing part of the compounds which are correctly predicted as 'actives' out of the total active compounds; 'specificity' value for recounting part of the compounds which are correctly predicted as 'non-actives' out of the total non-active compounds; 'accuracy' value 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) values. To apply the above parameters correctly in the study, the used data set has to be divided into two parts with almost equal number of compounds in each sub-set. One has to consider compounds with higher activity part as 'active' and the lower activity part as 'inactive'. By dividing the dataset in this way it is to be observed whether the predicted activities are true positives (TP)/true negatives (TN)/false positives (FP)/false negatives (FN), matching the experimental biological activities. If the predicted activity of a compound from active set falls in the same set, then it is assigned as

TP. If the actual activity of a compound is in the inactive set and observed value falls in active set value, then it is assigned as FP. In similar way the TN and FN are assigned to the data set compounds and number (n) of TP/TN/FP/FN are calculated and used in the following formulae to determine the said parameters of validation.

$$\text{Sensitivity} = \frac{\text{n(TP)}}{\text{n(TP)} + \text{n(FN)}}$$

$$\text{Specificity} = \frac{\text{n(TN)}}{\text{n(TN)} + \text{n(FP)}}$$

$$\text{Accuracy} = \frac{\text{n(TP)} + \text{n(TN)}}{\text{n(TP)} + \text{n(FN)} + \text{n(TN)} + \text{n(FN)}}$$

One can convert these values into percentage values by multiplying with 100.

$$\text{PPV} = \frac{\text{n(TP)}}{\text{n(TP)} + \text{n(FP)}}$$

$$\text{NPV} = \frac{\text{n(TN)}}{\text{n(TN)} + \text{n(FN)}}$$

$$\text{MCC} = \frac{\text{n(TP)} \times \text{n(TN)} + \text{n(FP)} \times \text{n(FN)}}{\sqrt{[\text{n(TP)} + \text{n(FP)}][\text{n(TP)} + \text{n(FN)}][\text{n(TN)} + \text{n(FP)}][\text{n(TN)} + \text{n(FN)}]}}$$

To apply these parameters correctly in the study, the used data set has to be divided into two parts with almost equal number of compounds in each sub-set. One has to consider one part as 'active' and the other as 'inactive'. Dividing the data into equal halves is the best approach to avoid bias observations. We performed many experiments and came to the conclusion that dividing the data set with equal number of molecules in each set gives the most accurate results that are unbiased. If one changes the proportion of compounds in favour of actives or inactives, then the percentage of sensitivity, specificity, accuracy, PPV, NPV and MCC varies 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 to one's advantage. Therefore equal division is the best approach that gives unbiased results.