Life Science Informatics Publications
Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences
Journal Home page http://www.rjlbpcs.com/
RJLBPCS
ISSN 2454-6348

**Original Research Article** **DOI: 10.26479/2023.0903.01**

# A QUANTITATIVE-STRUCTURE-ACTIVITY-RELATIONSHIP (QSAR) STUDY OF SOME IRREVERSIBLE INHIBITORS OF HUMAN CATHEPSIN B

**Indrani sarkar[1]*, Sudeshna Sarkar[2]**

1. Department of Basic Science and Humanities (Physics), Narula Institute of Technology, 81, Nilgunj Road, Agarpara, Kolkata 700109, West Bengal, India
2. Department of Tropical Medicine, Calcutta School of Tropical Medicine, 108, Chittaranjan Avenue, Calcutta Medical College, College Square, Kolkata 700073, West Bengal, India .

**ABSTRACT:** Cathepsin B belongs to a family of lysosomal cysteine proteases. It contains a highly reactive cysteine residue at the active site. The elevated level of cathepsin B causes neuromuscular dysfunction resulting in muscular dystrophy. Cathepsin B also seems to cause memory loss and neuronal cell death resulting in Alzheimer's symptoms. It has an important role during viral infection and replication in human cells. Inhibitors of cathepsin B are considered therapeutic targets for these diseases. In this report, a mathematical model of Multiple Linear Regression, for ordinary least squares is developed using a genetic algorithm for the selection of variables. The work is carried out using QSARINS software. The model is extensively validated according to OECD standards, and its robustness, stability, low correlation of descriptors, and good predictive power are also checked. It is found that the model fit is not the product of a chance correlation. Four possible outliers are identified in the model application domain, but in the molecular docking study, they seem to bind properly in the protease active site.

**Keywords:** Cathepsin B; Quantitative Structure-Activity Relationship; molecular descriptors; Multiple Linear Regression

**Article History: Received: May 06, 2023; Revised: May 20, 2023; Accepted: June 05, 2023.**

**Corresponding Author: Dr. Indrani Sarkar *** PhD.
Department of Basic Science and Humanities (Physics), Narula Institute of Technology, 81, Nilgunj Road, Agarpara, Kolkata 700109, West Bengal, India. Email Address: indrani.sarkar@nit.ac.in

## 1. INTRODUCTION

Cysteine proteases contain a highly reactive cysteine residue at the active site. Cysteine proteases are abundant in the human body and they play many important roles like intracellular proteolysis and tumor cell proliferation. Cathepsin B belongs to a family of lysosomal cysteine proteases [1, 2,17,18]. It influences the activity of matrix metalloproteinase and cathepsin D. It causes neuromuscular dysfunction resulting in muscular dystrophy. Cathepsin B also seems to cause memory loss and neuronal cell death resulting in Alzheimer's symptoms. It has an important role during viral infection and replication in human cells. So, Cathepsin B has been chosen as a target for developing drugs. Most of the potent inhibitors of cathepsin B form an irreversible covalent bond with the cysteine residue and bind in the active site by hydrogen bonding and hydrophobic interactions, thus disabling the catalytic activity of the protease. In this report, Multiple Linear Regression (MLR) models for ordinary least squares are developed. The selection of variables is done using a genetic algorithm. The model is developed using QSARINS software, with appropriate fitting parameters. The model is validated for its stability and ability to predict new compounds. The descriptors used to build the model show low correlation. The molecular docking method studies possible outliers identified in the model application domain.

## 2. MATERIALS AND METHODS

The Quantitative Structure-Activity Relationship (QSAR) Method is used for screening chemical compounds without experimental data. The properties of a molecule depend on its structure. Quantitative structure-activity research aims to determine the correlation between molecular structures and their biological activities. Models are built by regression analysis using a variety of molecular properties also known as descriptors [9, 10]. QSAR models are being used to search for new molecules with improved biological activity.

### 2.1 **Preparation of data set**

The three-dimensional structures of 35 small molecules are taken from PubChem Database. Biological activity data is collected from PubChem (PubChem bioassay accession 523 and 820) (Table 1). The molecules are subjected to energy minimization using MMFF94 force field, and 500 steps of steepest descents are carried out till the RMSD of potential energy is less than 0.001. The corresponding descriptor values are generated using PaDEL software. The descriptors are considered as independent variables X and the biological activity ($IC_{50}$) is taken as dependent variable Y. Highly correlated or identical descriptors with a correlation greater than 95% are discarded to avoid redundancy of descriptors. This is done by pairwise correlation or by calculating correlation among all pairs of descriptors. Similarly, descriptors with zero values are deleted. Descriptors having the same value for 80% of the compounds are also deleted. 397 data are excluded and 574 data are used for the computation. Fig1 a and b shows the variable profile and distribution of the experimental data for the descriptor ALogP.

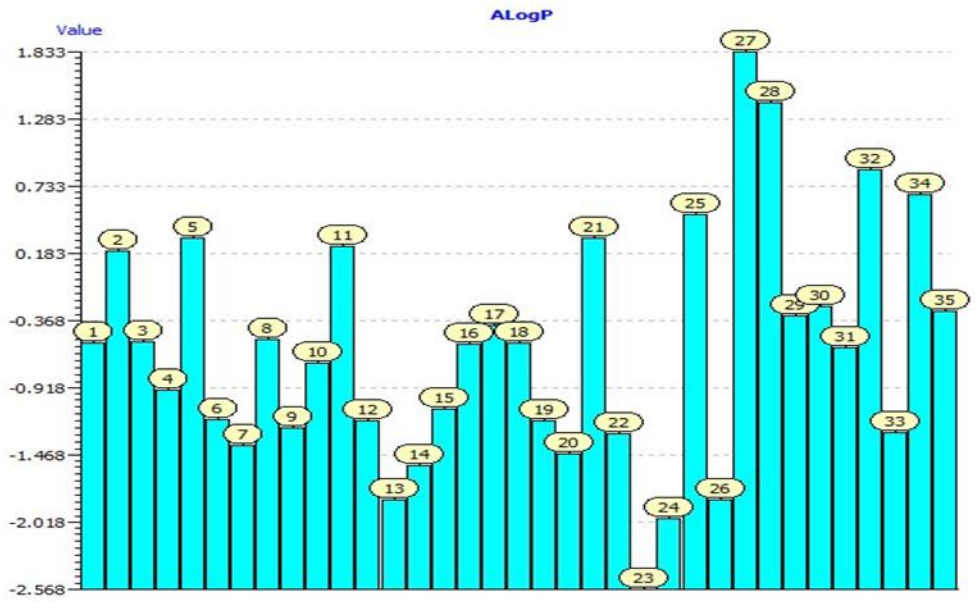



**Fig 1 a) Variable profile**

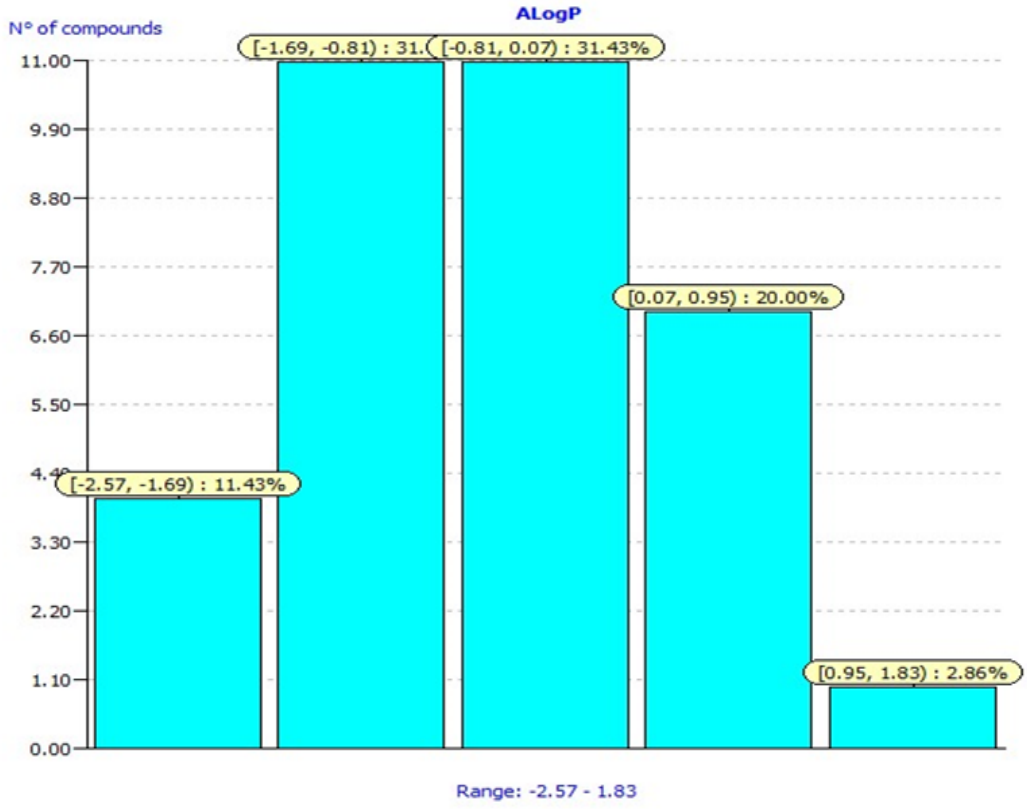



**Fig 1 b) Distribution of the experimental data**

The reduced set of descriptors with activity data of the compounds are subjected to variable selection procedure. The data is divided into training and test set randomly (70% training and 30% test). The Genetic-Algorithm-Variable Selection (GA-VSS) procedure is adopted to select the most significant descriptors variables.

**Table 1: List of 35 small molecules taken from pubchem database**

| Sl no | PubChem CID | $IC_{50}$ (μM) | $logIC_{50}$ | IUPAC Name | Molecular Formula |
|---|---|---|---|---|---|
| 1 | 11834381 | 2.82 | −5.550 | [5-amino-1-(benzenesulfonyl)pyrazol-3-yl] 4-methylthiophene-2-carboxylate | $C_{15}H_{13}N_3O_4S_2$ |
| 2 | 11834389 | 33.1 | −4.480 | [1-(4-methylphenyl)sulfonylpyrazol-3-yl] thiophene-2-carboxylate | $C_{15}H_{12}N_2O_4S_2$ |
| 3 | 11834392 | 3.23 | −5.490 | [1-(4-methoxyphenyl)sulfonyl-5-(thiophene-2-carbonylamino)pyrazol-3-yl] thiophene-2-carboxylate | $C_{20}H_{15}N_3O_6S_3$ |
| 4 | 1506381 | 45.97 | −4.338 | 4-[1-[(3-methoxyphenyl)methyl]benzimidazol-2-yl]-1,2,5-oxadiazol-3-amine | $C_{17}H_{15}N_5O_2$ |
| 5 | 2212050 | 7.11 | −5.148 | benzotriazol-1-yl-(2-ethylsulfanylphenyl)methanone | $C_{15}H_{13}N_3OS$ |
| 6 | 286532 | 11.46 | −4.941 | [4-(4-methoxybenzoyl)-5-oxido-1,2,5-oxadiazol-5-ium-3-yl]-(4-methoxyphenyl)methanone | $C_{18}H_{14}N_2O_6$ |
| 7 | 2998380 | 9.39 | −5.027 | *N*-[2-(benzenesulfonyl)-3,6-dihydrothiazin-1-ylidene]benzenesulfonamide | $C_{16}H_{16}N_2O_4S_3$ |
| 8 | 3236798 | 1.19 | −5.926 | 1,3-dimethyl-5-phenyl-6-(1,2,4-triazol-4-yl)pyrrolo[3,4-d]pyrimidine-2,4-dione | $C_{16}H_{14}N_6O_2$ |
| 9 | 3240114 | 0.69 | −6.160 | [5-amino-1-(4-methoxyphenyl)sulfonylpyrazol-3-yl] thiophene-2-carboxylate | $C_{15}H_{13}N_3O_5S_2$ |
| 10 | 3241895 | 0.44 | −6.362 | [5-amino-1-(4-fluorophenyl)sulfonylpyrazol-3-yl] thiophene-2-carboxylate | $C_{14}H_{10}FN_3O_4S_2$ |
| 11 | 3243025 | 0.85 | −6.073 | [3-oxo-2-(trifluoromethyl)-4*H*-1,4-benzoxazin-2-yl] acetate | $C_{11}H_8F_3NO_4$ |
| 12 | 3243128 | 0.26 | −6.608 | [5-amino-1-(benzenesulfonyl)pyrazol-3-yl] thiophene-2-carboxylate | $C_{14}H_{11}N_3O_4S_2$ |
| 13 | 3243168 | 8.56 | −5.067 | (1,3-dioxoisoindol-2-yl)methyl 2-(furan-2-carbonylamino)acetate | $C_{16}H_{12}N_2O_6$ |

| | | | | | |
|---|---|---|---|---|---|
| 14 | 3250046 | 18.35 | −4.736 | [2-(4-methoxyphenyl)-2-oxo-1-phenylethyl] 2-(furan-2-carbonylamino)acetate | $C_{22}H_{19}NO_6$ |
| 15 | 3685806 | 22.28 | −4.652 | (5-amino-1-methylsulfonylpyrazol-3-yl) thiophene-2-carboxylate | $C_9H_9N_3O_4S_2$ |
| 16 | 5293426 | 2.25 | −5.648 | 2-[[5,6-bis(furan-2-yl)-1,2,4-triazin-3-yl]sulfanyl]-*N*-phenylacetamide | $C_{19}H_{14}N_4O_3S$ |
| 17 | 573353 | 33.86 | −4.470 | 4-[1-[(4-fluorophenyl)methyl]benzimidazol-2-yl]-1,2,5-oxadiazol-3-amine | $C_{16}H_{12}FN_5O$ |
| 18 | 646525 | 1.99 | −5.701 | [5-amino-1-(4-methylphenyl)sulfonylpyrazol-3-yl] thiophene-2-carboxylate | $C_{15}H_{13}N_3O_4S_2$ |
| 19 | 646749 | 12.27 | −4.911 | [5-amino-1-(4-methoxyphenyl)sulfonylpyrazol-3-yl] 4-methylbenzoate | $C_{18}H_{17}N_3O_5S$ |
| 20 | 647599 | 1.26 | −5.899 | [5-amino-1-(4-fluorophenyl)sulfonylpyrazol-3-yl] furan-2-carboxylate | $C_{14}H_{10}FN_3O_5S$ |
| 21 | 648315 | 6.36 | −5.197 | [2-oxo-1-pyridin-2-yl-2-(thiophen-2-ylmethylamino)ethyl] thiophene-2-carboxylate | $C_{17}H_{14}N_2O_3S_2$ |
| 22 | 651936 | 1.75 | −5.757 | [5-amino-1-(4-methylphenyl)sulfonylpyrazol-3-yl] furan-2-carboxylate | $C_{15}H_{13}N_3O_5S$ |
| 23 | 653316 | 44.58 | −4.351 | 2-[2-(4-amino-1,2,5-oxadiazol-3-yl)benzimidazol-1-yl]-1-piperidin-1-ylethanone | $C_{16}H_{18}N_6O_2$ |
| 24 | 653862 | 0.92 | −6.035 | [5-amino-1-(4-methoxyphenyl)sulfonylpyrazol-3-yl] furan-2-carboxylate | $C_{15}H_{13}N_3O_6S$ |
| 25 | 654815 | 2.12 | −5.674 | *N*-[(3,4-dichloro-5-oxo-2*H*-furan-2-yl)carbamoyl]acetamide | $C_7H_6Cl_2N_2O_4$ |
| 26 | 655490 | 9.56 | −5.019 | [5-amino-1-(4-methoxyphenyl)sulfonylpyrazol-3-yl] benzoate | $C_{17}H_{15}N_3O_5S$ |
| 27 | 658111 | 6.72 | −5.173 | methyl (5-cyano-3,3-dimethyl-8-morpholin-4-yl-1,4-dihydropyrano[3,4-c]pyridin-6-yl)sulfanylformate | $C_{17}H_{21}N_3O_4S$ |
| 28 | 658152 | 19.69 | −4.706 | diethyl 2-[cyano-[4-(dimethylamino)-6-methylsulfanyl-1,3,5-triazin-2-yl]amino]propanedioate | $C_{14}H_{20}N_6O_4S$ |
| 29 | 658724 | 8.93 | −5.049 | [4-[(2-methoxyphenyl)iminomethyl]-2-phenyl-1,3-oxazol-5-yl] acetate | $C_{19}H_{16}N_2O_4$ |

| | | | | | |
|---|---|---|---|---|---|
| 30 | 658964 | 39.99 | −4.398 | [4-[(2-methoxyphenyl)iminomethyl]-2-phenyl-1,3-oxazol-5-yl] propanoate | $C_{20}H_{18}N_2O_4$ |
| 31 | 660829 | 38.47 | −4.415 | [2-(furan-2-yl)-4-(phenyliminomethyl)-1,3-oxazol-5-yl] furan-2-carboxylate | $C_{19}H_{12}N_2O_5$ |
| 32 | 665480 | 2.09 | −5.680 | *tert*-butyl *N*-[(1*R*)-2-methyl-1-[5-[(3-methylphenyl)methylsulfonyl]-1,3,4-oxadiazol-2-yl]butyl]carbamate | $C_{20}H_{29}N_3O_5S$ |
| 33 | 714967 | 14.2 | −4.848 | 2-[cyano-(4-methoxy-6-pyrrolidin-1-yl-1,3,5-triazin-2-yl)amino]acetamide | $C_{11}H_{15}N_7O_2$ |
| 34 | 794694 | 4.17 | −5.380 | 2-(4-chlorophenyl)sulfonyl-4,5-dimethyl-3,6-dihydrothiazine 1-oxide | $C_{12}H_{14}ClNO_3S_2$ |
| 35 | 971438 | 37.19 | −4.430 | 4-[1-[(5-methoxy-2-methylphenyl)methyl]benzimidazol-2-yl]-1,2,5-oxadiazol-3-amine | $C_{18}H_{17}N_5O_2$ |

### 2.2 Software

For the calculation of molecular descriptors, PaDEL Software [5] is used. The QSARINS (QSAR-Insubria) software developed at the University of Insubria is used for model building. To reduce the computation time, only a small number of descriptors are used per model and all combinations are explored using the all-subset technique. Next genetic algorithm (GA) method is applied to develop models with larger number of descriptors.

### 2.3 Multiple Linear Regression Model

MLR model gives a linear relationship between the biological activity (here half maximal inhibitory concentration, IC50) and the molecular descriptors of the compounds. Ordinary least squares (OLS) algorithm is used in the process [3, 11]. The optimum models are ordered by the software according to $R^2$.

### 2.4 Fitting Criteria

This contains the following criteria $R^2$, $R^2$adj, $R^2$- $R^2$adj, LOF [7], $k_{xx}$ (inter correlation among descriptors), delta k (difference of correlation among the descriptors $k_x$ and descriptors plus the responses $k_{xy}$), RMSE (training), MAE (training), RSS (training), CCC (training) and S and F vales [7]. $R^2$ (regression coefficient) evaluates fitness of a particular model. It should be closer to zero for the model being good. $R^2$ greater than 0.6 is acceptable for QSAR model generation. As the number of descriptors increases, $R^2$ value improves, but to avoid statistical incompatibility value of $R^2$adj is noted. Adding useless variables to the model will decrease $R^2$adj. Similarly adding useful variables $R^2$adj will increase. $R^2$adj will always be less than or equal to $R^2$. LOF (Lack of Fit) should have near zero value and not greater than 0.4 to have a model with less error. F (Fischer criteria) should have higher values. This signifies that the model is significant and is not obtained by chance. $K_{xx}$

denotes the total correlation among the block of descriptors [21, 22]. It should have low value. $K_{xy}$ denotes the correlation among the descriptors plus the responses. The model makes sense if $K_{xy} - K_{xx} < \delta_x$, where $\delta_x$ is a user defined threshold value. MEA or mean absolute error in fitting calculated on training set should be small. MAE tr (Mean Absolute Error in fitting) is calculated using the training set. RMSEtr gives Root Mean Square Error in the training set. RSStr means Residual Sum of Squares in fitting for training set. CCCtr is the Concordance correlation coefficients on the training set [3,4]. CCC values should be high and near 1. RMSE training, RMSE validation and s values should be close for model statistics where s value denotes the standard error of estimate.

**2.5 Internal validation**

The model robustness is checked by iterated cross validations. The corresponding cross validated (CV) correlation coefficient ($Q^2_{LOO}$, $Q^2_{LMO}$) are calculated by Leave-One-Out (LOO) and Leave-Many-out (LMO) methods. This is done iteratively by excluding one compound from the descriptor dataset (LOO), and then computing a model with the remaining compounds. The model then makes a prediction for the excluded one. The value of $Q^2_{LOO}$ should be greater than $R^2$ and the model can be considered to be robust. Leave-More (or Many)-Out (LMO) technique checks the behavior of the model when many compounds are excluded. 30% of compounds are excluded randomly and the model is calculated using the remaining data. Then the model performance is verified by making prediction with the excluded compounds. The model becomes stable if the averages ($R^2_{LMO}$ and $Q^2_{LMO}$) are close to the $R^2$ and $Q^2_{LOO}$ values of the model. $Q^2_{LOO}$ and $Q^2_{LMO}$ should be greater than 0.6. RMSEcv and MAE should be less than 0.5. One of the most important conditions is that RMSEtr should be less than RMSEcv, RMSEtr, RMSEcv and standard error s values should be close. To confirm that the model was not obtained by chance correlation, the Y-scrambling procedure was applied. Responses or the experimental data are placed randomly so that there is no correlation of responses with the descriptors. Hence the performance of the successive models should decay drastically. For a well validated model $R^2$ and $Q^2$ values after each iteration and their averages ($R^2_{YS}$ and $Q^2_{YS}$) become less than the corresponding values of the model.

**2.6 External validation**

After internal validation the model is used predict new compounds. The model equation from the training set is applied on the excluded compounds never used before in model calculation. The performance of the model is measured by different criteria, such as: $RMSE_{EXT}$,$Q^2_{F1}$,$Q^2_{F2}$,$Q^2_{F3}$,$r^2_m$ plus $\Delta r^2_m$, CCC, and the Golbraikh and Tropsha [8] method. $Q^2_{F1}$, $Q^2_{F2}$, $Q^2_{F3}$ values should be greater than 0.7, $CCC_{ext} > 0.85$, $R^2_{ext} > 0.6$ and $r^2_m > 0.6$. Here $RMSE_{ext}$ should be less and comparable with RMSE and overall error of the model. k and k' are slopes of the regression line which is within the cut off value (0.85 and 1.15).

## 3. RESULTS AND DISCUSSION

The data (574 descriptors) are processed using the QSARINS software [9,10,11]. Several MLR models are developed (Table 3) with low multicollinearity between descriptors (Table 2). The mean values of $R^2$ and $Q^2_{LOO}$ versus the number of variables of the generated models are plotted (Fig.2). This presents the performance of the models against their size. The values of $R^2$ and $Q^2_{LOO}$ rose with new addition of descriptors So the added descriptor does not make any improvement. The models with five variables had very similar descriptor combinations. Based on the statistical results model MLR1 with five variables was chosen and was used to predict the new inhibitors outside this dataset. The best MLR model obtained is shown below with its statistical parameters (Table 3)

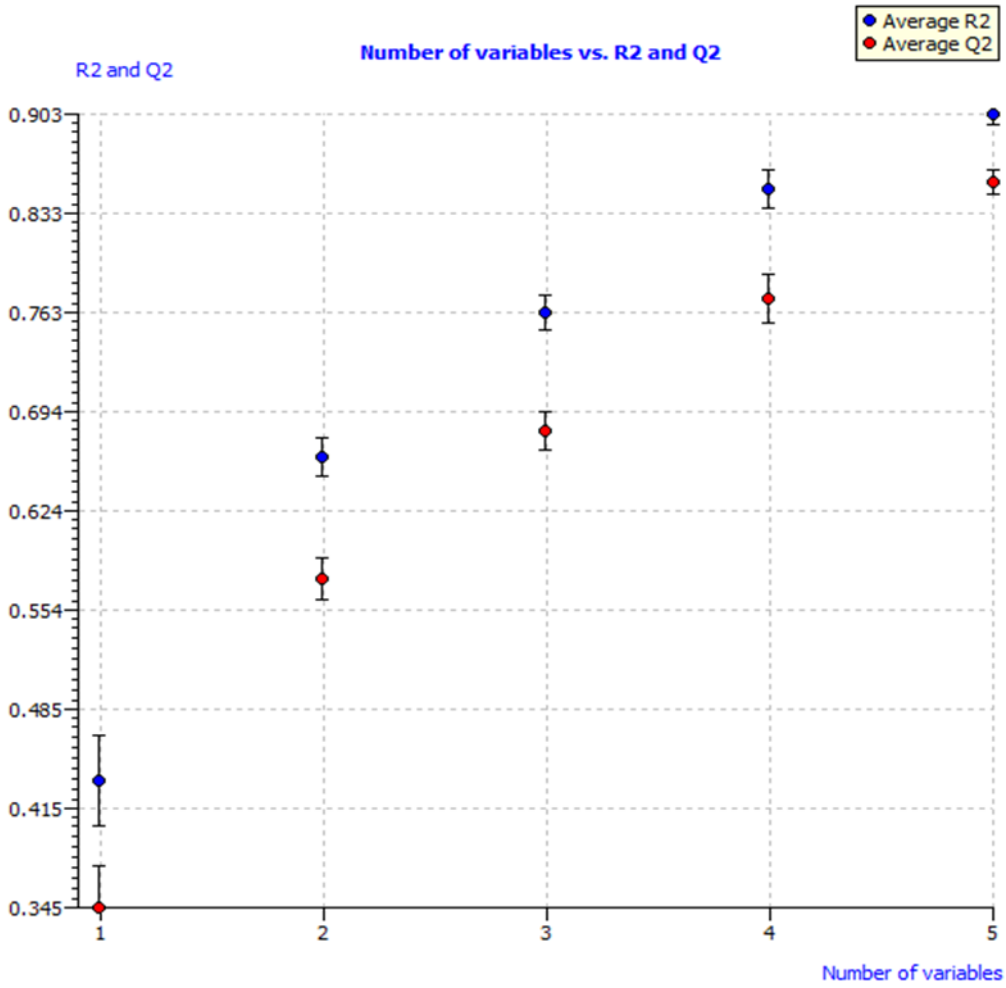



**Fig 2: Average models $R^2$ vs $Q^2$**

**Table 2. Correlation matrix shows that the descriptors have little correlation**

| | AATS4i | ATSC8e | MATS5m | minaaS | JGI6 |
|---|---|---|---|---|---|
| AATS4i | 1.00 | | | | |
| ATSC8e | 0.236 | 1.000 | | | |
| MATS5m | -0.176 | -0.085 | 1.000 | | |
| minaaS | 0.188 | 0.110 | 0.232 | 1.000 | |
| JGI6 | 0.341 | 0.0258 | -0.234 | -0.257 | 1.000 |

**Table 3. Variables with their coefficients for MLR model**

| Variable | Coeff. | Std. coeff | Std. err. | (+/-) Co. int. 95% | p-value |
|---|---|---|---|---|---|
| Intercept | 9.2540 | | 1.5446 | 3.2330 | 0.0000 |
| AATS4i | -0.0358 | -0.2771 | 0.0097 | 0.0202 | 0.0011 |
| ATSC8e | 0.2121 | 0.3271 | 0.0461 | 0.0965 | 0.0001 |
| MATS5m | -2.6781 | -0.4978 | 0.3648 | 0.7635 | 0.0000 |
| minaaS | -0.4373 | -0.3300 | 0.0924 | 0.1933 | 0.0001 |
| JGI6 | 101.6992 | 0.6552 | 10.7189 | 22.4349 | 0.0000 |

**Fitting criteria**

**$R^2$: 0.9174 $R^2$adj: 0.8957 $R^2$-$R^2$adj: 0.0217 LOF: 0.0876**

**Kxx: 0.1971 Delta K: 0.1325 RMSE tr: 0.1776 MAE tr: 0.1346**

**RSS tr: 0.7887 CCC tr: 0.9569 s: 0.2037 F: 42.2305**

**Internal validation criteria**

**$Q^2_{LOO}$: 0.8637 $R^2$-$Q^2_{LOO}$: 0.0537 RMSE cv: 0.2282 MAE cv: 0.1762**

**PRESS cv: 1.3018 CCC cv: 0.9291**

**$Q^2_{LMO}$: 0.8443 $R^2$Yscr: 0.2093 $Q^2$Yscr: -0.4138 $RMSE_{AV}$ Yscr: 0.5481**

**(External validation criteria)**

**RMSE ext: 1.0095 MAE ext: 0.6936 PRESS ext: 10.1915 $R^2$ext: 0.0247**

**$Q^2$-$F_1$: -1.3910 $Q^2$-$F^2$: -1.7808 $Q^2$-$F^3$: -1.6669 CCC ext: 0.1250**

**r2m aver.: 0.0038 r2m delta: 0.0074**

**Calculated external data regression angle from diagonal: -33.7626°**

The model statistics shows an $R^2$ (0.9174) and $R^2_{adj}$ 0.8957. This indicates that a new descriptor can be added to the model. The low value of the LOF parameter (0.0876) indicated that no overfitting was found in the model and it has a good fit with minimum number of descriptors. The small value of $K_{xx}$ (0.1971) shows that the correlation among the model descriptors is low. Delta K parameter (0.1325) shows that the model has good correlation between descriptors and the experimental data (Log IC50). Other estimated parameters show small error in the calculation of training set ($RMSE_{tr}$ = 0.1776; $MAE_{tr}$ = 0.1346; s = 0.2037). Yellow dots in the scatter plot (Fig. 3a) represent the values predicted by the model equation versus the experimental values. Fig 3b shows scatter plot by LOO method. The yellow points (training set) are calculated using the LOO method and the blue points

(test set) are calculated using the model equation. The scatter plot also shows strong outliers in the data.

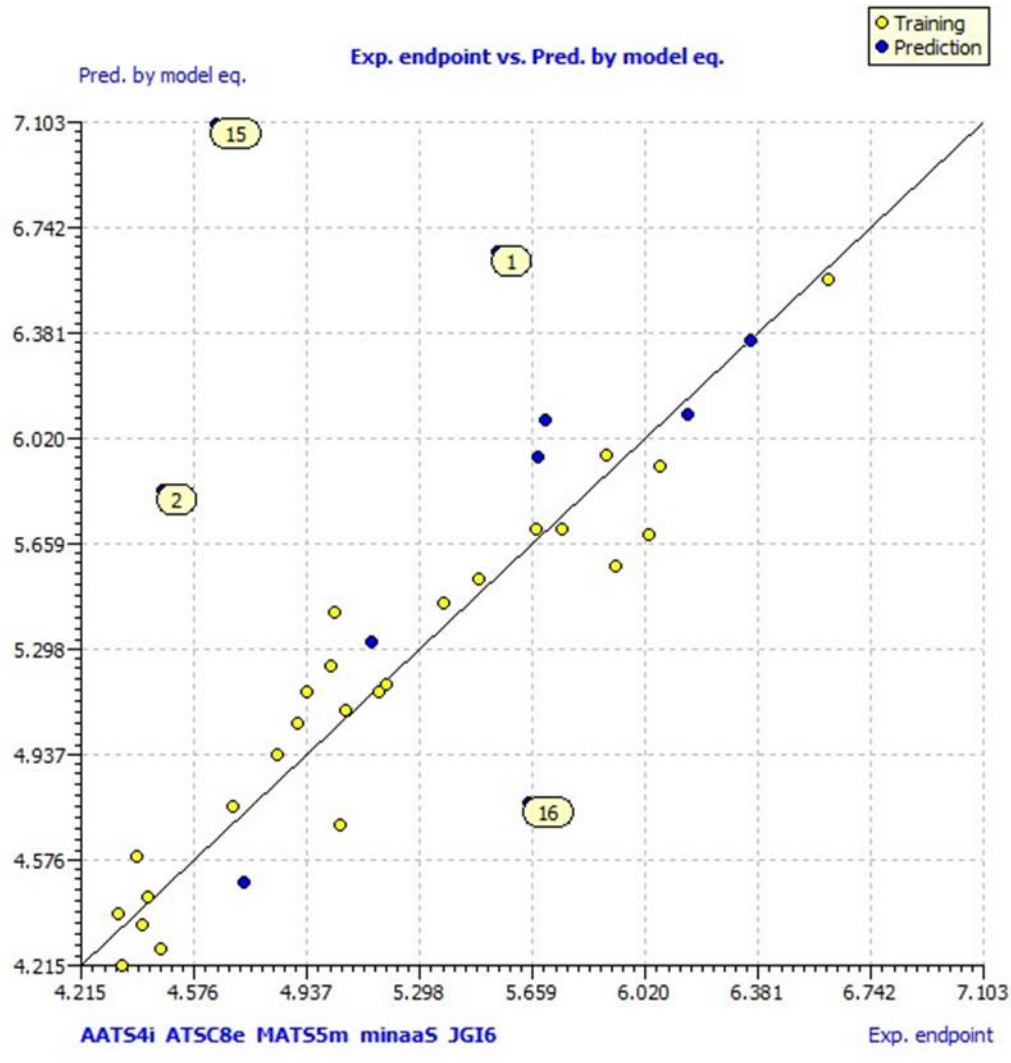



**Fig 3a. Scatter plot of Experimental endpoint vs. Predictions by the model equation**

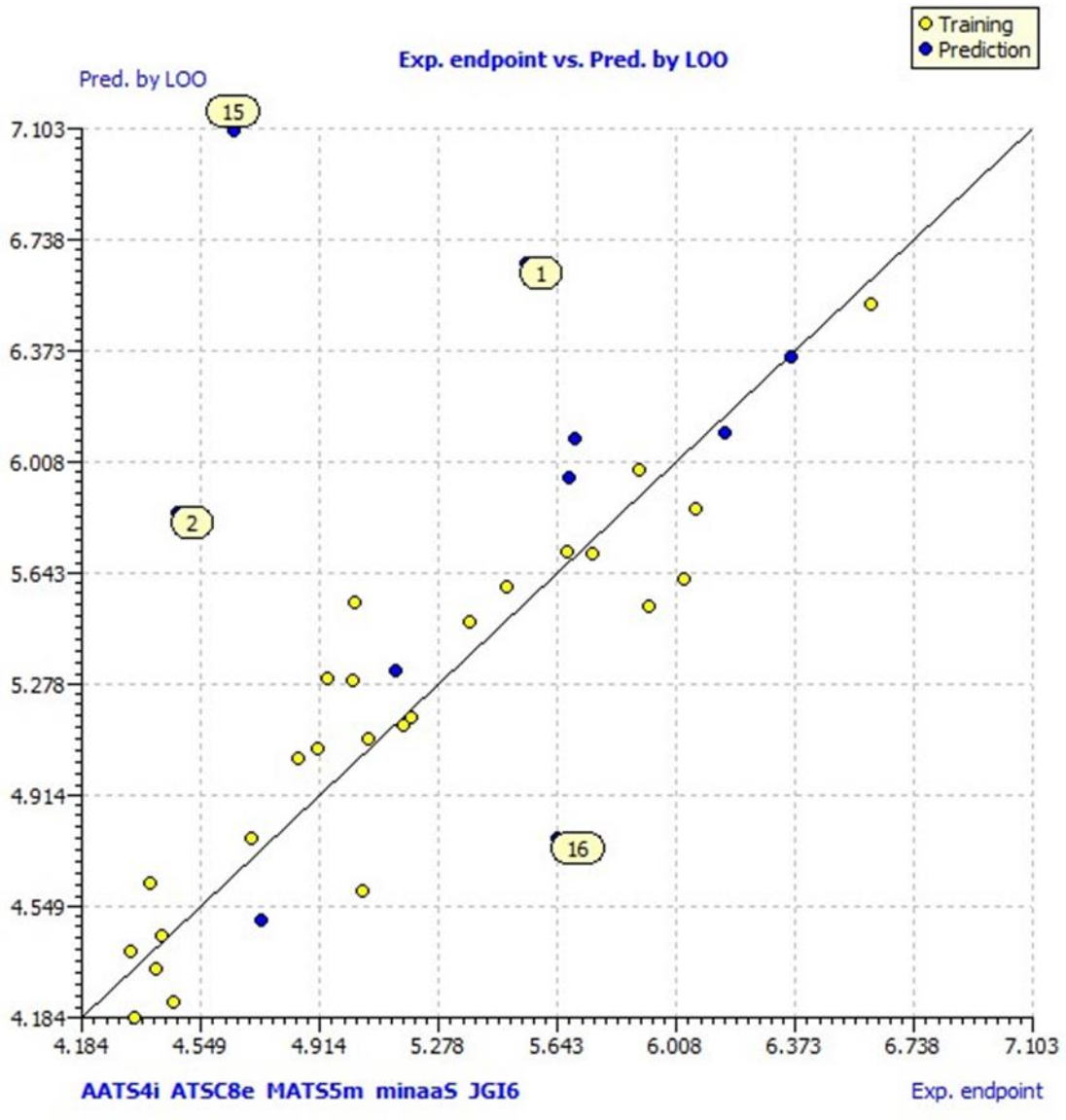



**Fig 3b Scatter plot of experimental endpoint data a vs. LOO predictions**

The applicability of the models is checked by the fitting, stability in CV and its efficiency in predicting unknown compounds. The stability of the model is evaluated using internal validation. The result shows that the variance found in the prediction by LOO ($Q^2_{LOO} = 0.8637$) is comparable with $R^2 = 0.9174$. So, the prediction from internal validation procedure is good. The error in the predictions ($RMSE_{cv} = 0.2282$ and $MAE_{cv} = 0.1762$) is small and so the model can be considered as internally stable. Fig. (4) shows the residual plot of experimental endpoints vs. residuals from the LOO predictions. . Leaving Many-Out (LMO) uses the technique of leaving out the 30% of the dataset to study the performance of the model; $R^2 = 0.9174$ and $Q^2_{LMO} = 0.8443$ values are comparable, so the model can be considered to be stable. Value of $Q^2_{LMO}$ (0.8443) is comparable with $Q^2_{LOO}$ (0.8637) of the model. The $Q^2_{LMO}$ versus $K_{xy}$ plot (Fig.5) presents scatter plot of LMO models compared to the QSAR model. QSAR model is labeled as "model $Q^2$" by the blue point and the LMO models performances ($Q^2$ as red points) are reported on the ordinate axes. The

performance of LMO models is almost similar to the original model. The similarity in the values of $Q^2_{LMO}$, which confirms the robustness of the model. To rule out the possibility of chance correlation, the Y-scrambling procedure is used. Values of $R^2$ Y-scr and $Q^2$ Y-scr are 0.2093 -0.4138 respectively. The $R^2$ Y-scr and $Q^2$ Y-scr values against $R^2$ and $Q^2$ of the model are shown in Fig.6. The values of $R^2$ and $Q^2$ of the model are found to be very far from the values obtained for these parameters in the Y-scrambling procedure. This signifies that the model is not obtained by chance. External validation technique is used to check the predictive ability of the model. Their parameters are at per with the model ($R^2_{ext} = 0.0247$), $RMSE_{ext} = 1.0095$, $MAE_{ext} = 0.6936$, $PRESS_{ext} = 10.1915$, $Q^2$ -$F_1$ = -1.3910, $Q^2$ -$F_2$ = -1.7808, $Q^2$ - $F_3$ = -1.6669, $CCC_{ext}$ =0.1250, $r^2_{m\_aver} = 0.0038$, $r^2_{m\_delta} = 0.0074$. Here $R^2_{ext}$ is the coefficient of determination in the external validation procedure [8], $RMSE_{ext}$ measures the Root Mean Square Error , $MAE_{ext}$ is the Mean Absolute Error ; $PRESS_{ext}$ is the Predictive Residual Sum of Squares , $Q^2$ -$F_1$ [19] $Q^2$ -$F_2$ [20], and $Q^2$ -$F_3$ [21,22] measure the variances given in external validation; $CCC_{ext}$ is the Concordance Correlation Coefficient [3,4] $r^2_{m\_aver}$ and $r^2_{m\_delta}$ are the Roy criteria: average and delta [16]. Fig 7 present q-q plot of experimental values vs. residuals from the LOO predictions. Values of the theoretical quantiles (Z values) are plotted on the abscissa and the values of the residuals of the predictions on the ordinate. The yellow points (training set) are values predicted by LOO while the blue points are calculated using the model equation. William graph of the model (Fig 8) shows that most of the compounds are within the applicability domain of the model (within the critical leverage h*= 0.720). The HAT values of the diagonal elements are plotted on the abscissa and the predicted residuals are on the ordinate. Yellow points represent training set for the LOO and the blue points are prediction set calculated using the model equation. The dashed horizontal ones are the user defined threshold for Y-outliers. HAT values higher than the cutoff value $h^* = 3p'/n$, where $p'$ is the number of model variables plus one and n is the number of the objects used to calculate the model, are considered as outliers Probable outlier compounds are also identified. The Insubria graph (Fig 9) provided by QSARINS also displays applicability domain. Here HAT diagonal values are reported on the abscissa and the predicted data are reported on the ordinate one. Yellow (training set) and blue (prediction set) data points and are data predicted by model equation when experimental value is known. Red points represent unknown molecules outside training set predicted by the model equation. The model equation is built using five molecular descriptors. (Table 3) [13,14]. 12 compounds were taken from PubChem (Table 4) and were tested with the model equation and the result is shown in Table 5.

**Table 4: 12 compounds taken from pubchem and used for validation**

| PubChem ID | IUPAC name of the compounds |
|---|---|
| 123664 | (2S,3S)-3-[[(2S)-4-methyl-1-(3-methylbutylamino)-1-oxopentan-2-yl]carbamoyl]oxirane-2-carboxylic acid |
| 2142225 | 3-anilino-2H-isoquinolin-1-one |
| 2883004 | 4-[1-(2-ethenoxyethyl)benzimidazol-2-yl]-1,2,5-oxadiazol-3-amine |
| 2997948 | 1,8-diamino-3,6-dipyrrolidin-1-yl-2,7-naphthyridine-4-carbonitrile |
| 2997975 | (2-oxo-1,2-diphenylethyl) 2-(furan-2-carbonylamino)acetate |
| 561320 | 4-(1-benzylbenzimidazol-2-yl)-1,2,5-oxadiazol-3-amine |
| 647501 | 1-ethyl-6-methyl-3-phenylpyrimido[5,4-e][1,2,4]triazine-5,7-dione |
| 653297 | 1-ethyl-6-methyl-3-thiophen-2-ylpyrimido[5,4-e][1,2,4]triazine-5,7-dione |
| 719048 | spiro[5,6,7,8-tetrahydro-4H-[1,2,4]triazolo[5,1-b]quinazoline-9,1'-cyclohexane]-2-amine |
| 738531 | 4-(1H-benzimidazol-2-yl)-1,2,5-oxadiazol-3-amine |
| 787437 | 3-amino-7-benzyl-1-sulfanylidene-6,8-dihydro-5H-thiopyrano[3,4-c]pyridine-4-carbonitrile |
| 439487 | (2R,3R)-3-[[(2S)-1-[4-(diaminomethylideneamino)butylamino]-4-methyl-1-oxopentan-2-yl]carbamoyl]oxirane-2-carboxylic acid |

**Table 5: Comparison of experimental values of $pic_{50}$ and values obtained from model equation of MLR1, MLR2, MLR3, MLR4, MLR5**

| Compound ID | MLR1 | MLR2 | MLR3 | MLR4 | MLR5 | experimental data pIC50 | experimental data $IC_{50}(\mu M)$ |
|---|---|---|---|---|---|---|---|
| 123664 | 5.290 | 4.828 | 5.479 | 4.906 | 4.795 | 8.816 | 0.001526 |
| 2142225 | 5.636 | 5.839 | 5.341 | 5.836 | 6.056 | 4.755 | 17.574 |
| 2883004 | 4.584 | 4.828 | 4.818 | 4.815 | 4.905 | 4.526 | 29.7412 |
| 2997948 | 5.652 | 5.811 | 5.422 | 5.759 | 6.431 | 5. 498 | 3.17087 |
| 2997975 | 4.478 | 4.163 | 4.471 | 4.026 | 4.167 | 4.721 | 18.9995 |
| 561320 | 4.006 | 4.310 | 4.074 | 4.158 | 4.365 | 4.649 | 22.4328 |
| 647501 | 5.492 | 5.321 | 5.433 | 5.378 | 5.535 | 7.148 | 0.071007 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 653297 | 5.393 | 5.298 | 5.181 | 5.316 | 5.947 | 7.141 | 0.072233 |
| 719048 | 5.467 | 5.401 | 5.464 | 5.538 | 5.514 | 4.887 | 12.9471 |
| 738531 | 4.480 | 4.930 | 4.841 | 4.674 | 4.976 | 4.336 | 46.1134 |
| 787437 | 5.814 | 5.726 | 5.118 | 5.899 | 6.226 | 5.785 | 1.63948 |
| 439487 | 5.279 | 4.487 | 5.593 | 4.537 | 4.443 | 7.911 | 0.012283 |

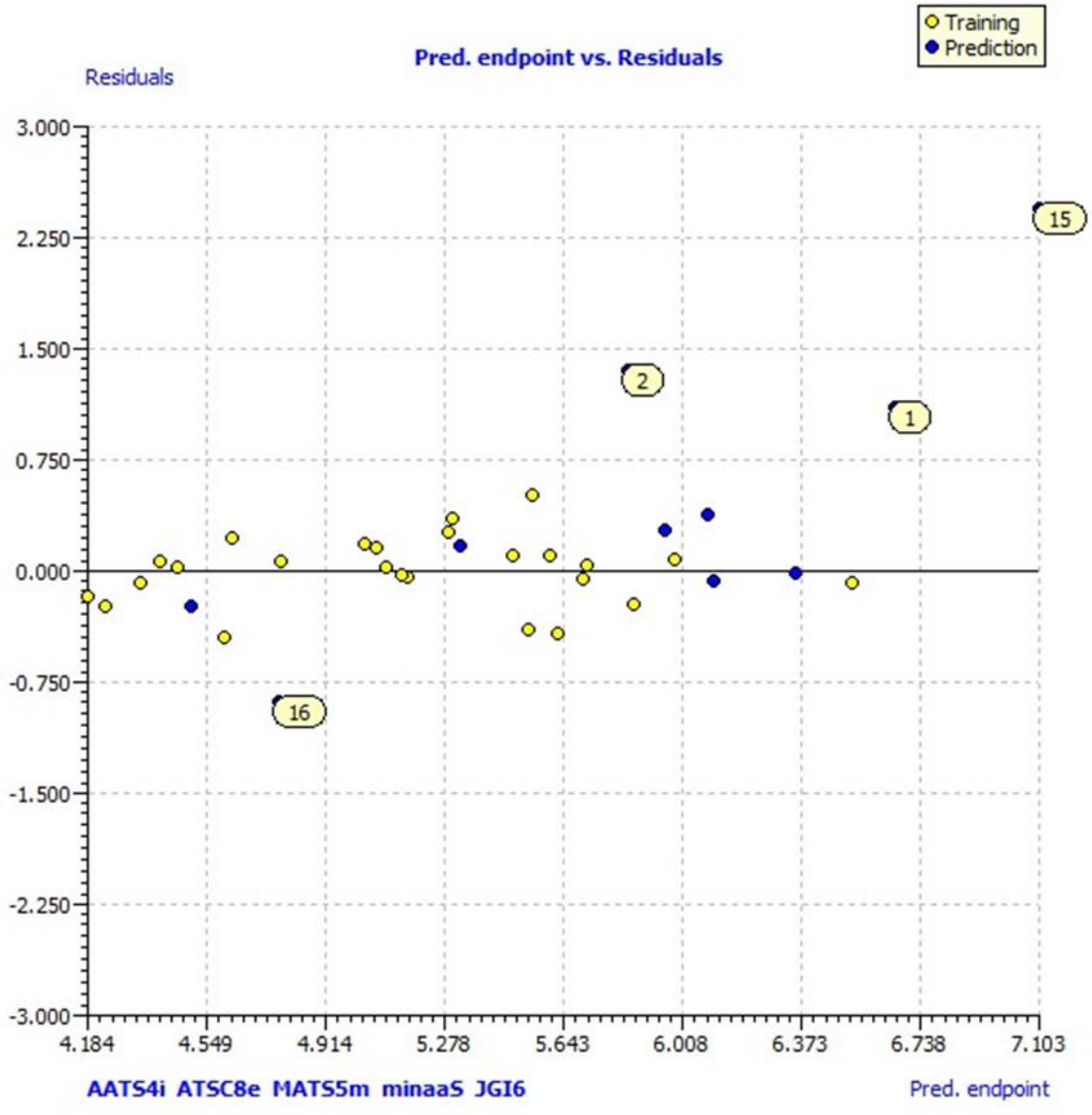



**Fig 4. Residual plot of Experimental values vs. residuals from the LOO predictions. On the abscissa axes the values of the experimental values are reported, while on the ordinate the values of the residuals of the predictions are reported.**

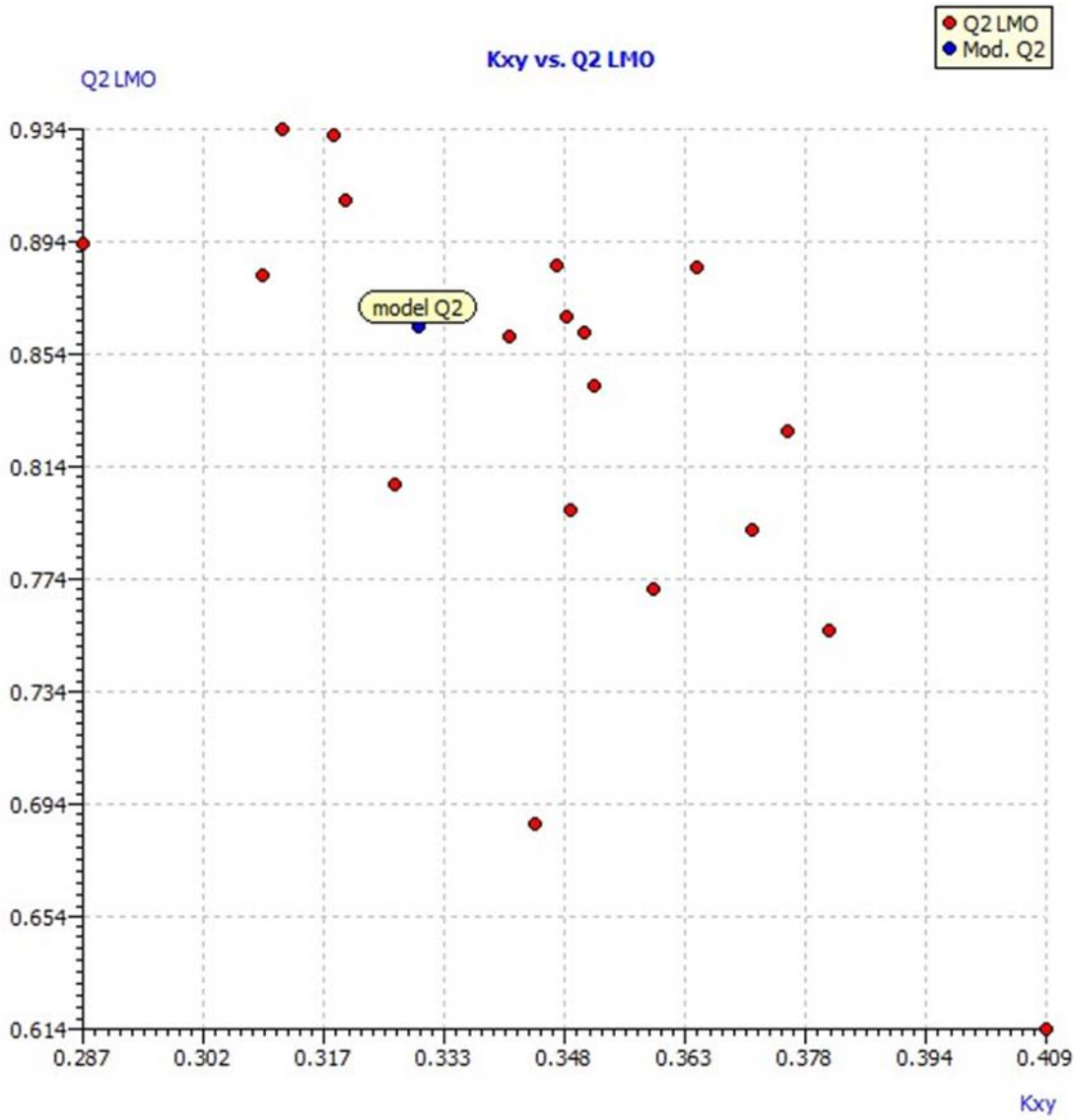



**Fig 5. Plot of LMO models compared with the original QSAR model**

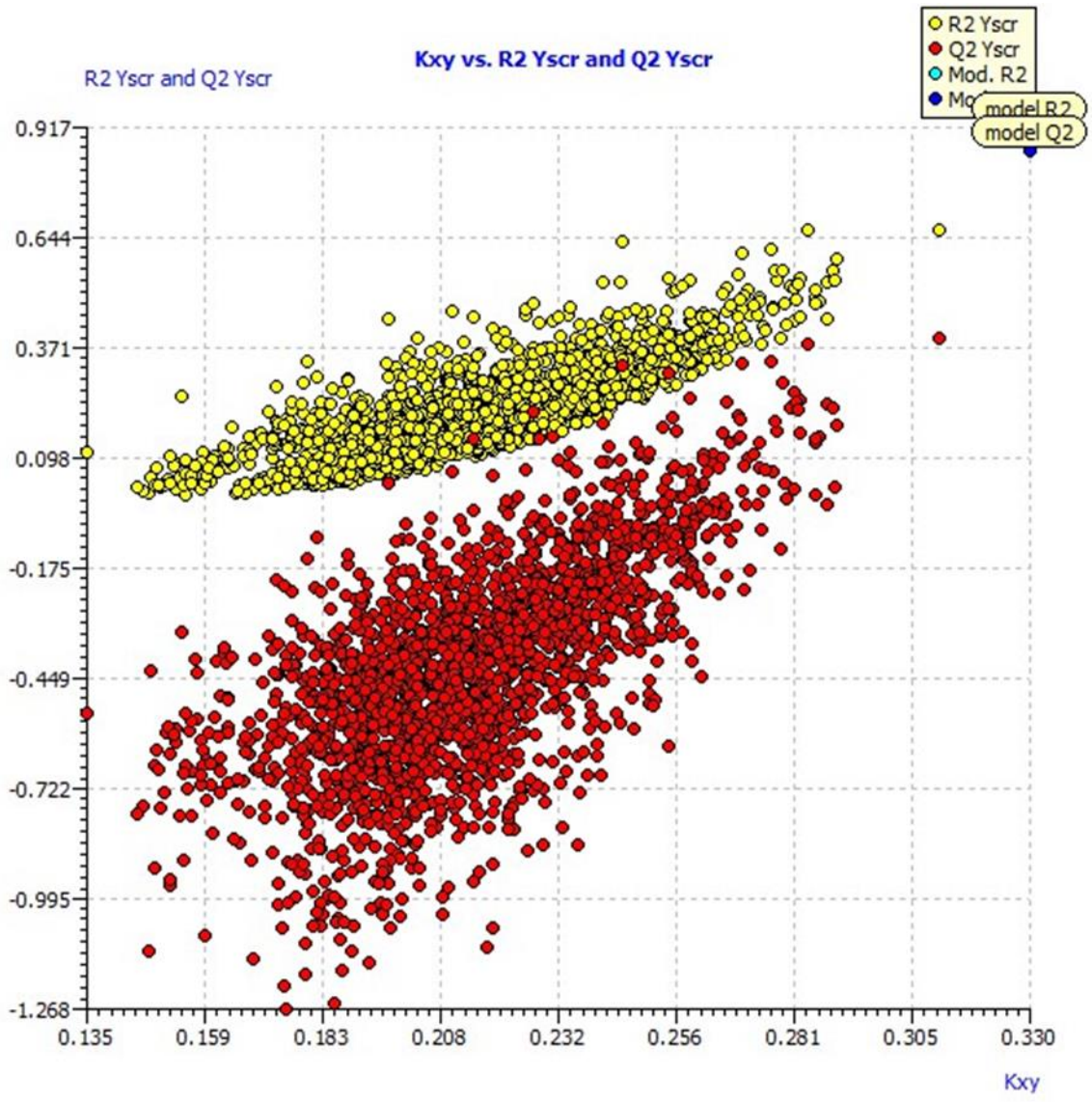



**Fig 6. Scatter plot of Y-scrambled models compared to the original QSAR model**

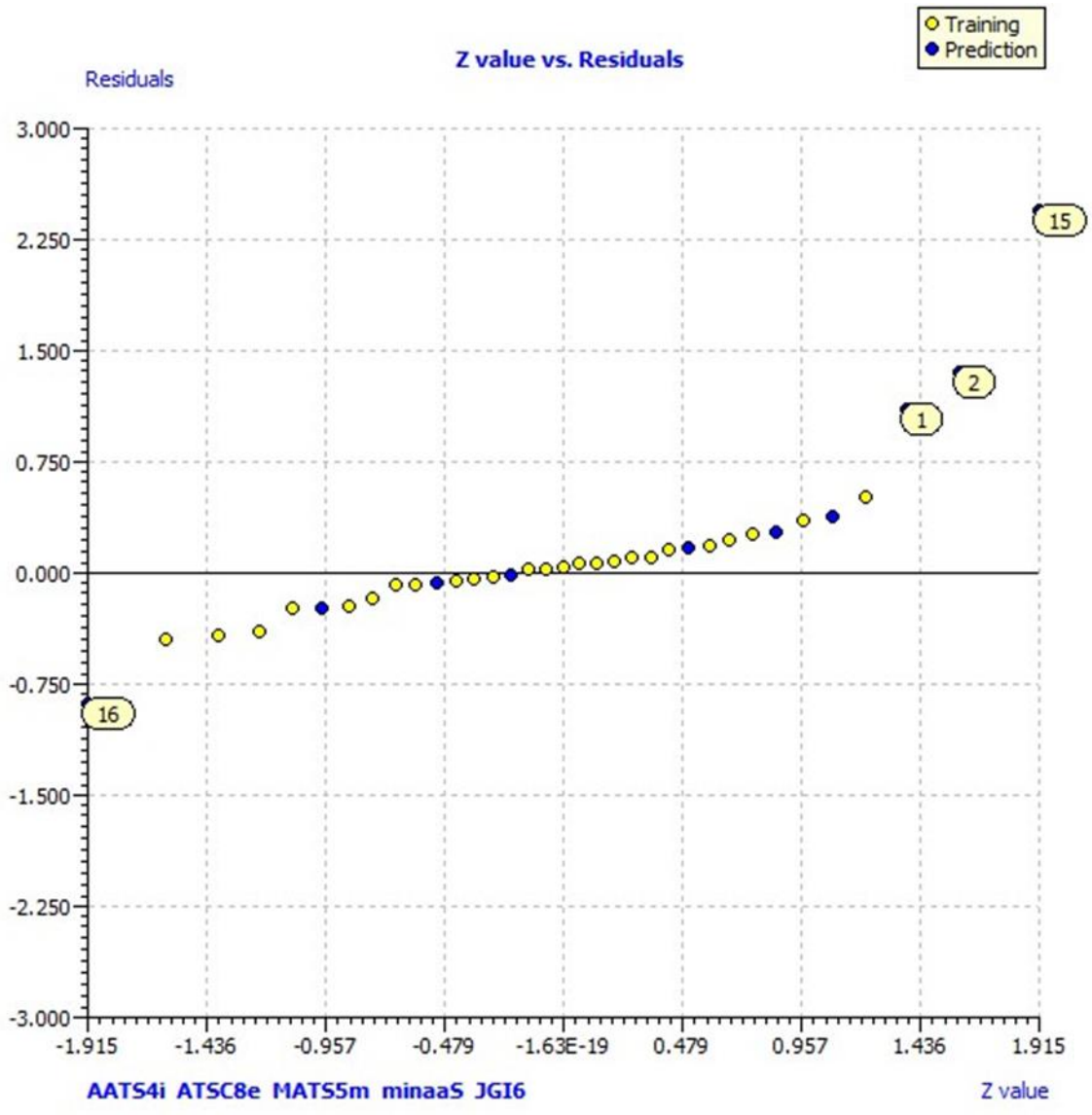



**Fig 7 q-q plot of experimental endpoint vs. residuals from the predictions by model equation.**

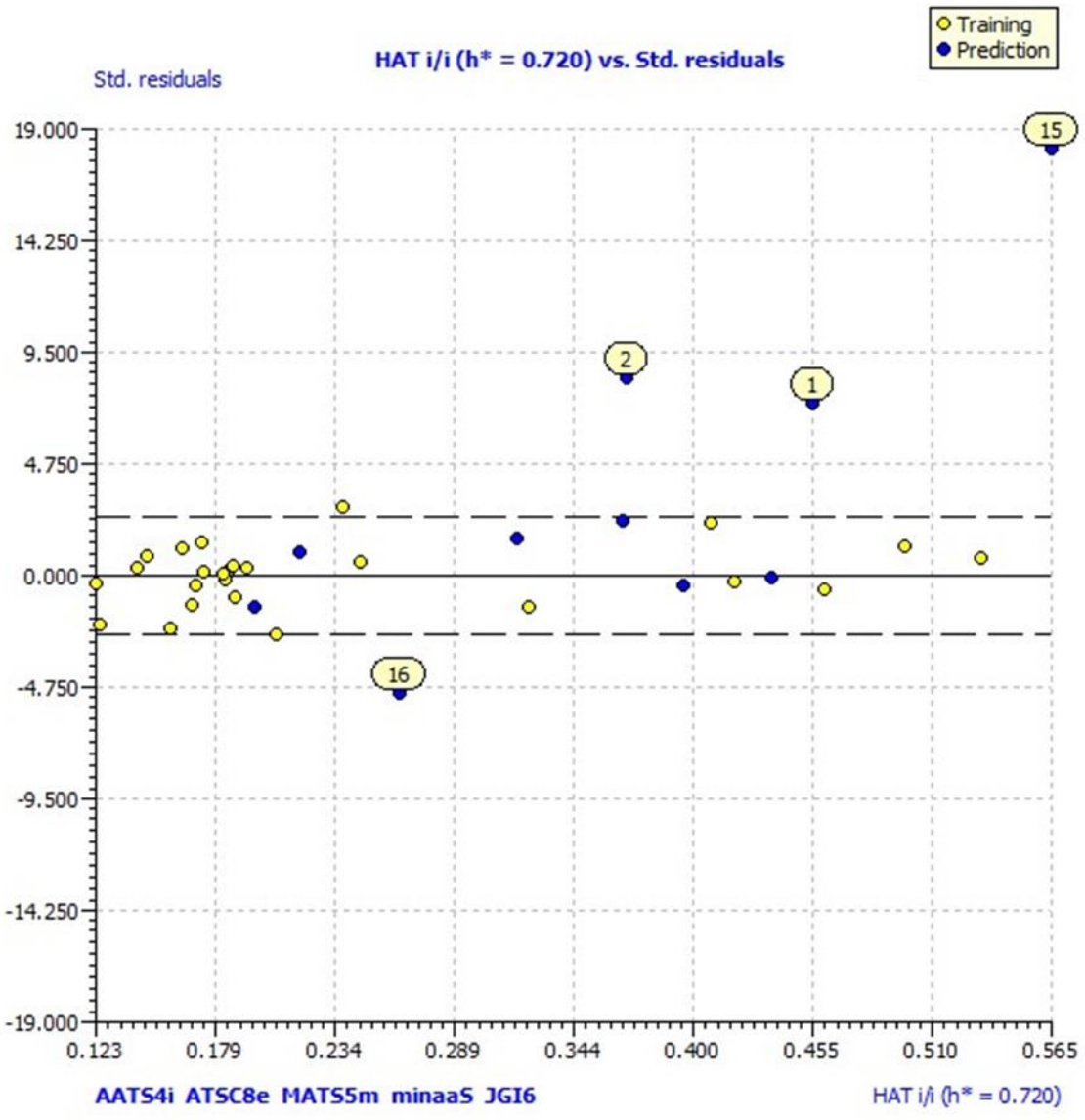



**Fig 8 Williams plot of diagonal hat elements vs. standardized residual predictions by LOO**

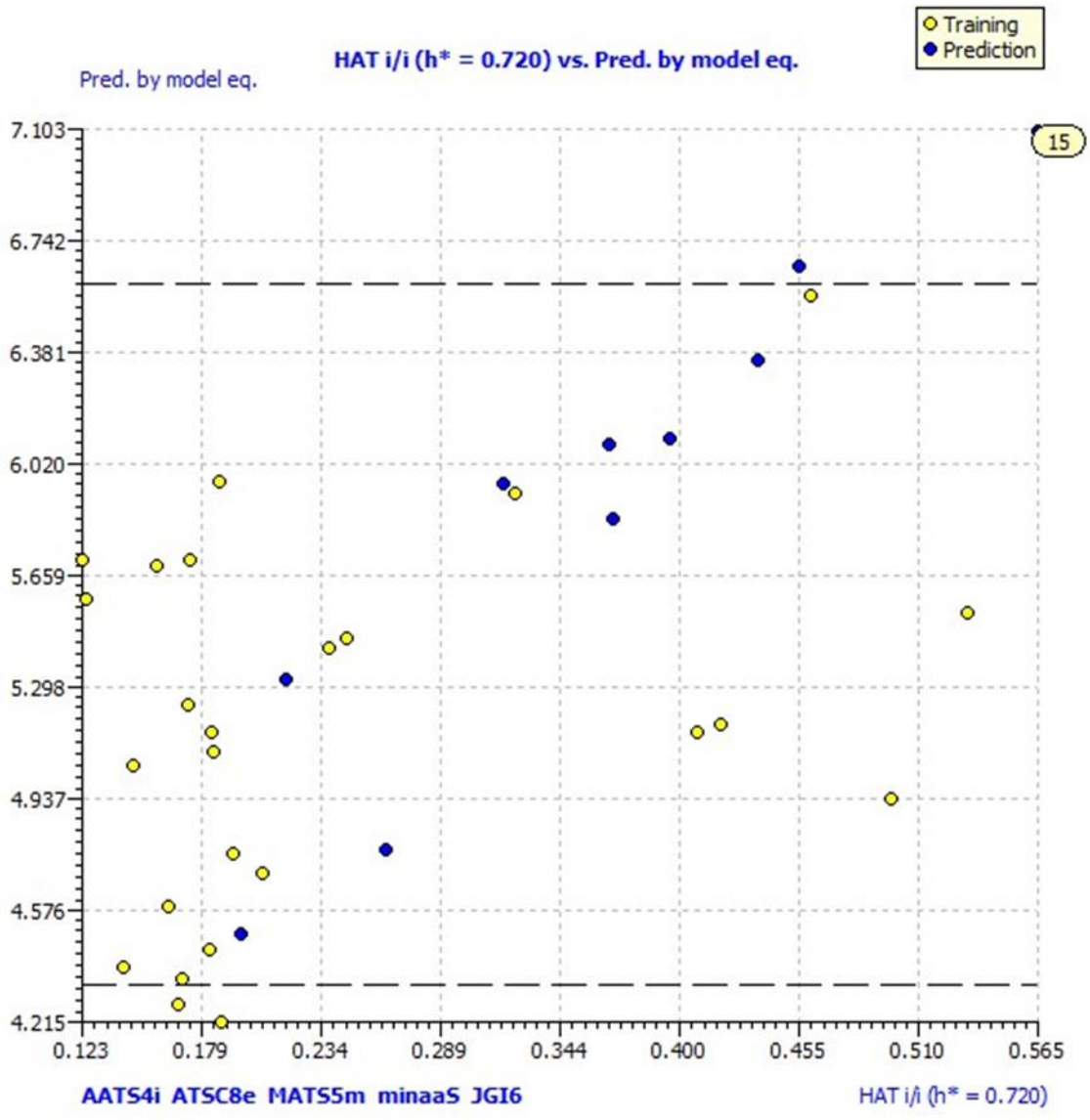



**Fig 9 Insubria graph: diagonal hat elements vs. predictions by model equation**

Using the PCA, a list of models is selected, to calculate the average their performances by means of a combined modeling. Selected models are:

MLR1 model: 9.2540 -0.0358**AATS4i** +0.2121**ATSC8e** -2.6781**MATS5m** -0.4373**minaaS** +101.6992**JGI6**

MLR 2 model: 14.040 -0.0447 **AATS4i** -2.7438**MATS5m** -1.2151Sp**Max6_Bhm** -0.3713**minaaS** +98.0854 **JGI6**

MLR3 model: 3.7234 + 0.2068**ATSC8e** +0.0094**ATSC3i** -2.4719**MATS5m** -0.6140**minaaS** +90.9740**JGI6**

MLR4 model: 12.0392 -0.0507**AATS4i** -2.5642**MATS5m** -0.4891**minaaS** +107.5669**JGI6** -

0.0005**WPATH**

MLR5 model: 14.7568 -0.0501**AATS4i** -2.8345**MATS5m** -1.2330**SpMax6_Bhm** + 2.4780**VCH-5** + 104.7325JGI

Fitting criteria for the combined model is given below:

$R^2$ ACM= 0.9348, $R^2$ WCM= 0.9424, MAEtr= 0.1194, RMSE $_{tr}$= 0.1587, $CCC_{tr}$= 0.9650

External validation criteria of the combined model is

MAE $_{ext}$: 0.7254, RMSE $_{ext}$: 1.0768, CCC $_{ext}$: 0.0836

$Q^2$-$F_1$: -1.7201 $Q^2$-$F_2$: -2.1635 $Q^2$-$F_3$: -2.0340

Calculated external data (ACM) regression angle from diagonal: -36.7936°

Calculated external data (WCM) regression angle from diagonal: -37.1370°

the Average Combined Prediction (ACM) is calculated averaging the predictions of the molecules separately from every model. Insubria graph of average hat diagonal elements vs. ACM predicted by model eq shown in Fig (10). Fig (11and 12) present William's plot predicted by combined model equation using ACM and WCM respectively.

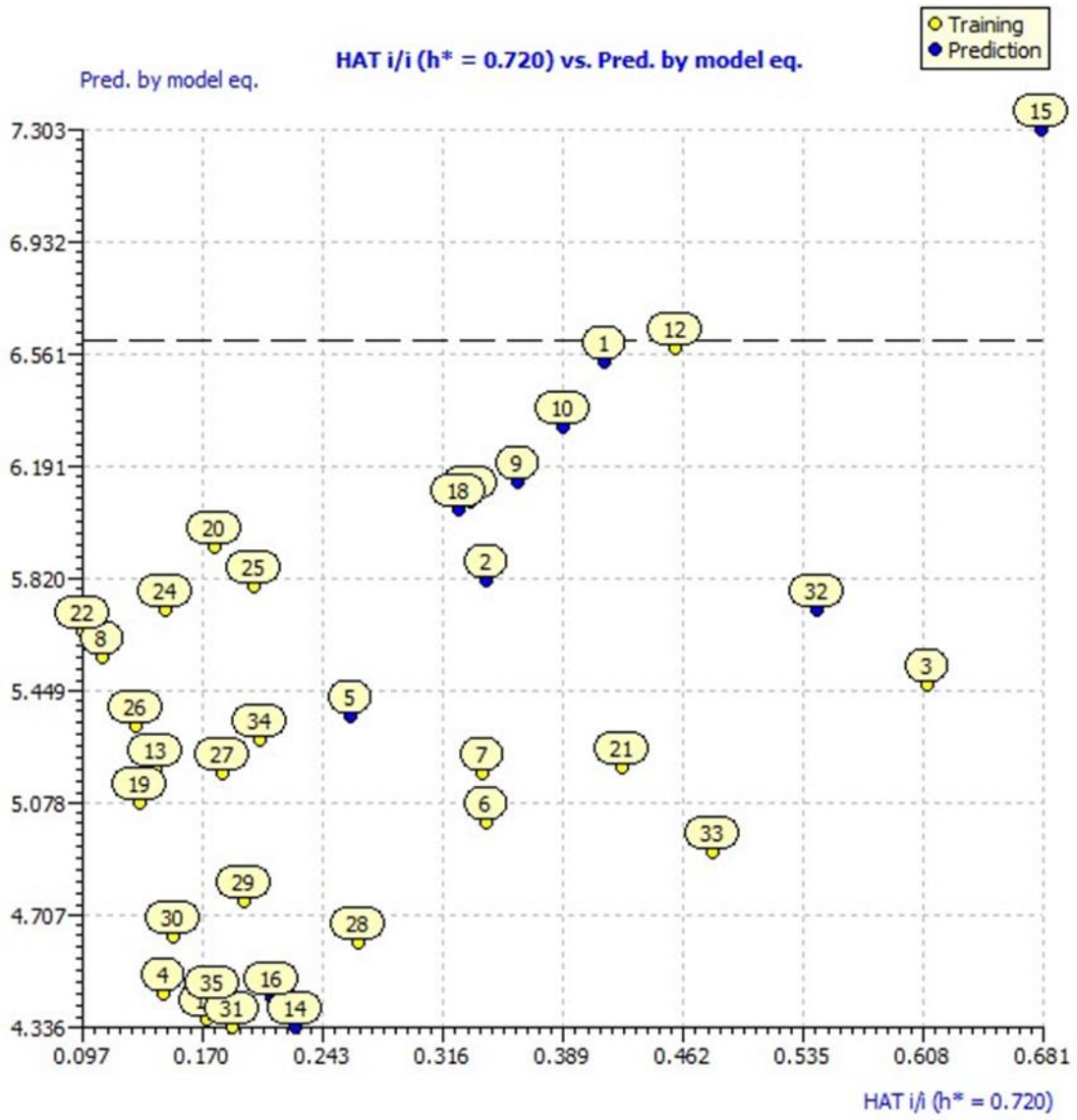



**Fig 10 Insubria graph diagonal hat elements vs. predictions by the combined model equation**

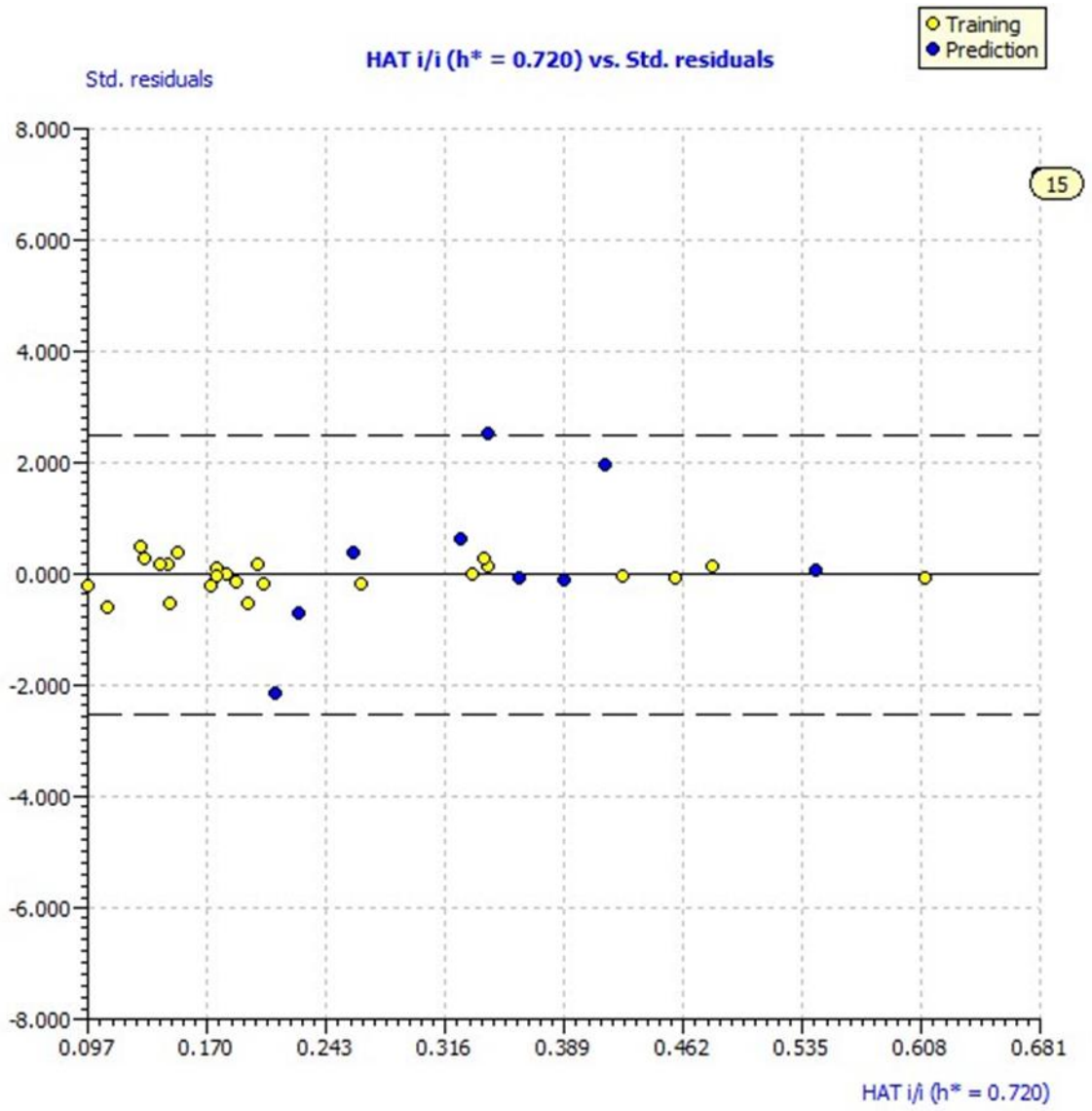



**Fig 11 Williams plot predicted by combined model equation using ACM**

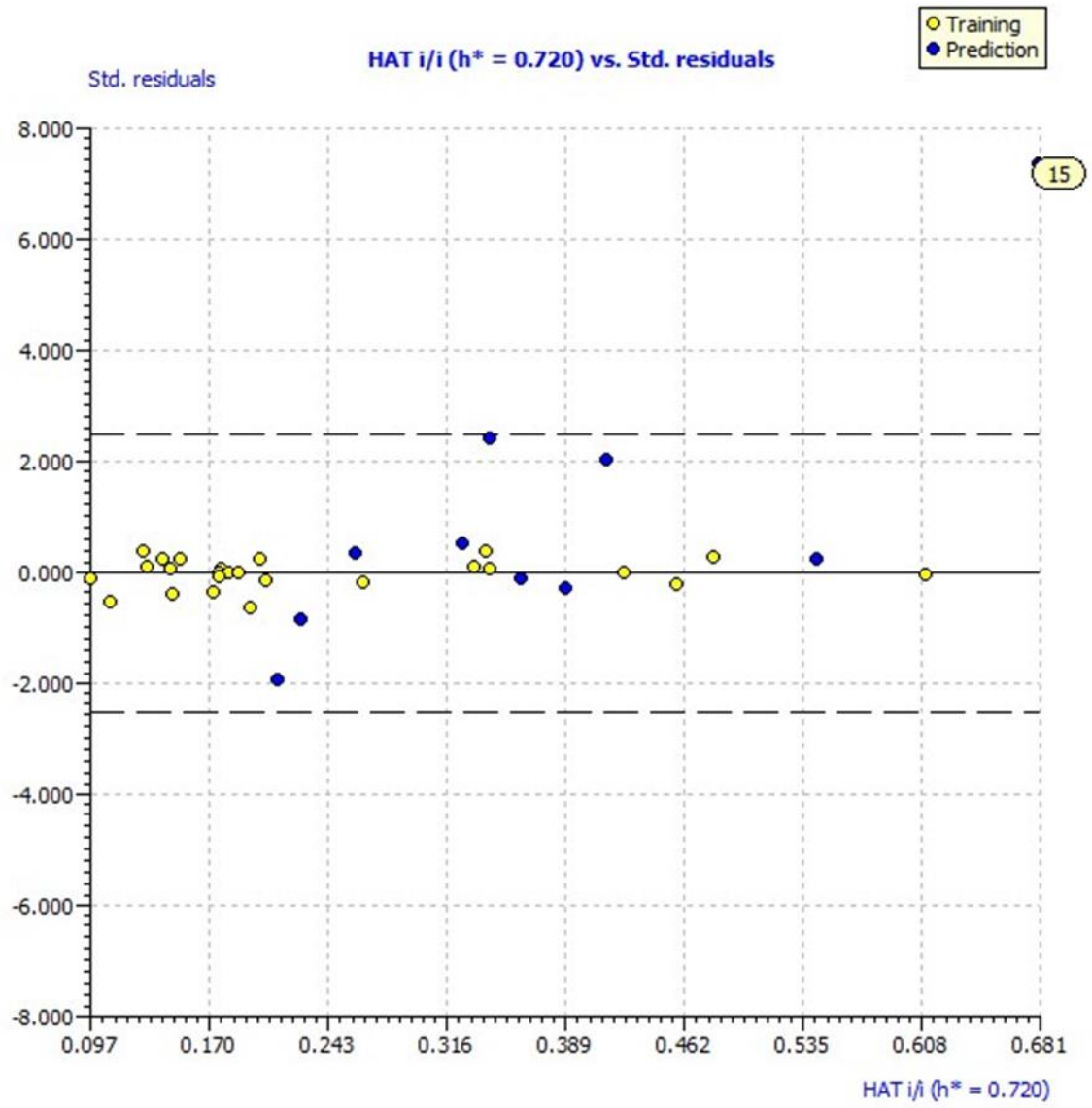
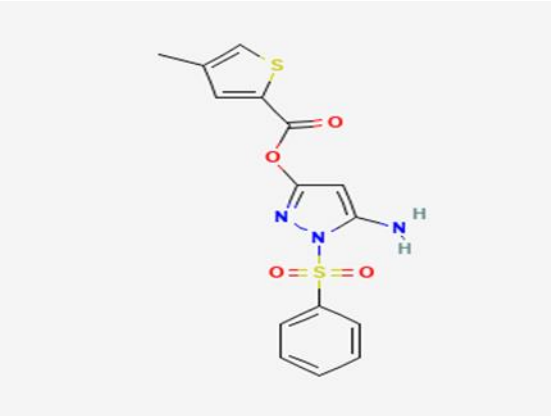


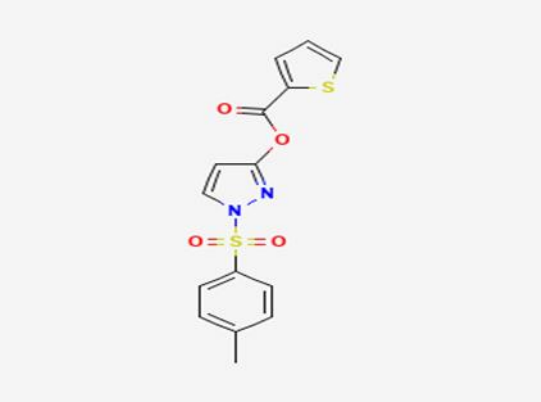

**Fig 12 Williams plot predicted by combined model equation using WCM**

The three-dimensional structural coordinates of cathepsin B bound with an irreversible inhibitor Dipeptidyl Nitrile (DPN) is obtained from Protein Databank (PDB code: 1GMY) [12]. Cathepsin B contains three chains A, B and C. Chain A is used for docking studies. Hydrogen atoms and partial charges are added to the protein and energy minimization using an OPLS force field is performed to avoid short contacts. The minimized protein structure is used to dock compounds 1,2 ,15 and 16 (identified as outliers) into the DPN binding site using Auto Dock Vina program [6,23] (Fig 13). In previous molecular docking studies of E64 derivatives with cathepsin B shows that all inhibitors bind to S subsites of cathepsin B [15]. The computer simulation study shows the presence of a large hydrophobic pocket around the active thiol group of the active site of cathepsin B. The inhibitors form a covalent bond with the active site of cathepsin B. The docked molecules 1,2,15 and 16 are found to stabilize in the active site by formation of hydrogen bond, Van der Walls and pi pi stacking

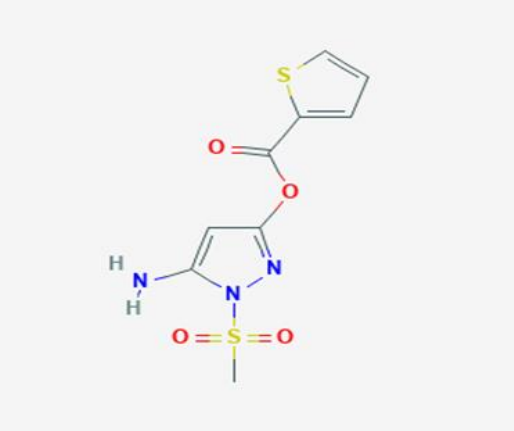
interactions (Fig 14,15,16,17). Fig 18 shows the superposition of docked compounds 1,2,15 and 16 in the active site of cathepsin B. These compounds are also active inhibitors as found from their biological activity data. (Table 1).



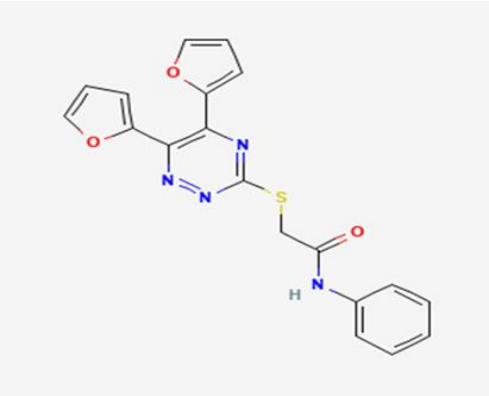
Compound 1 : ID 11834381



Compound 2:ID 11834389



Compound 15 :ID 3685806



Compound 16: ID 5293426

**Fig13: 2D structures of compounds 1,2,15,16**

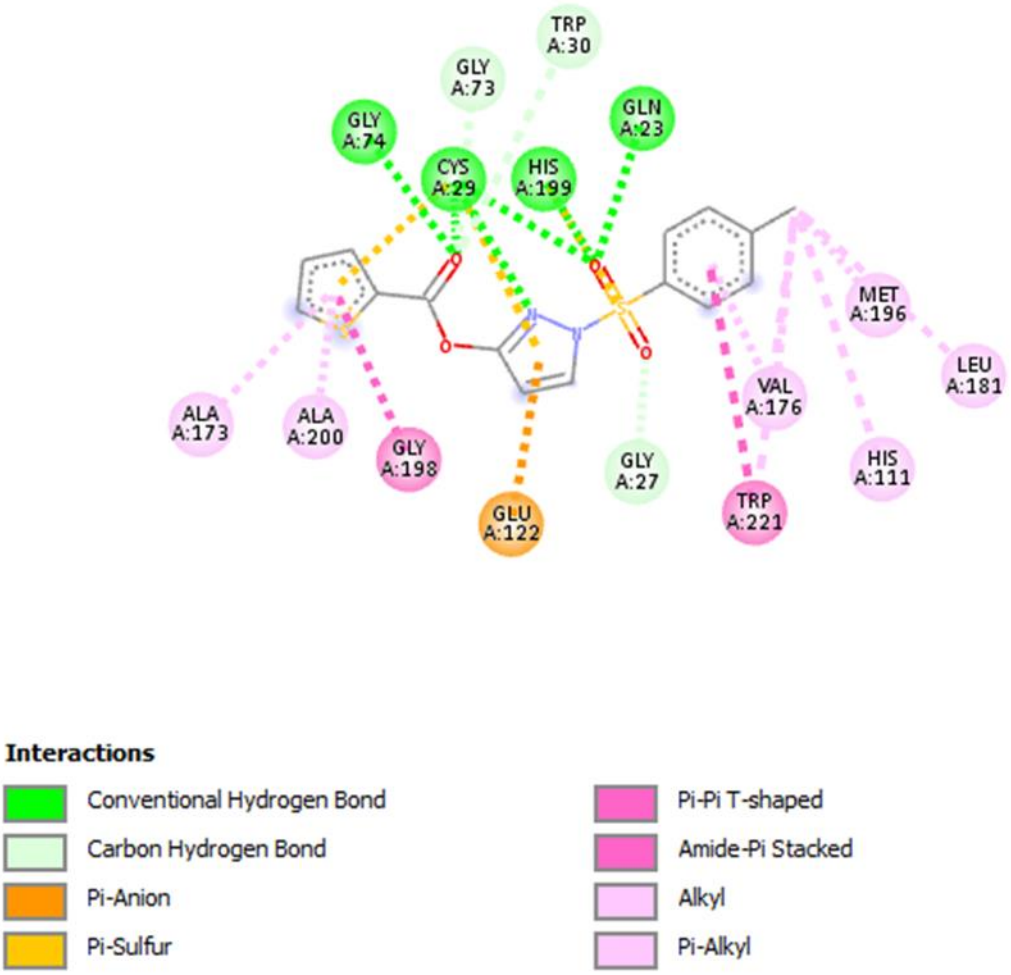



**Fig 14. Interaction with Molecule 1 with cathepsin B (PDB code: 1GMY)**

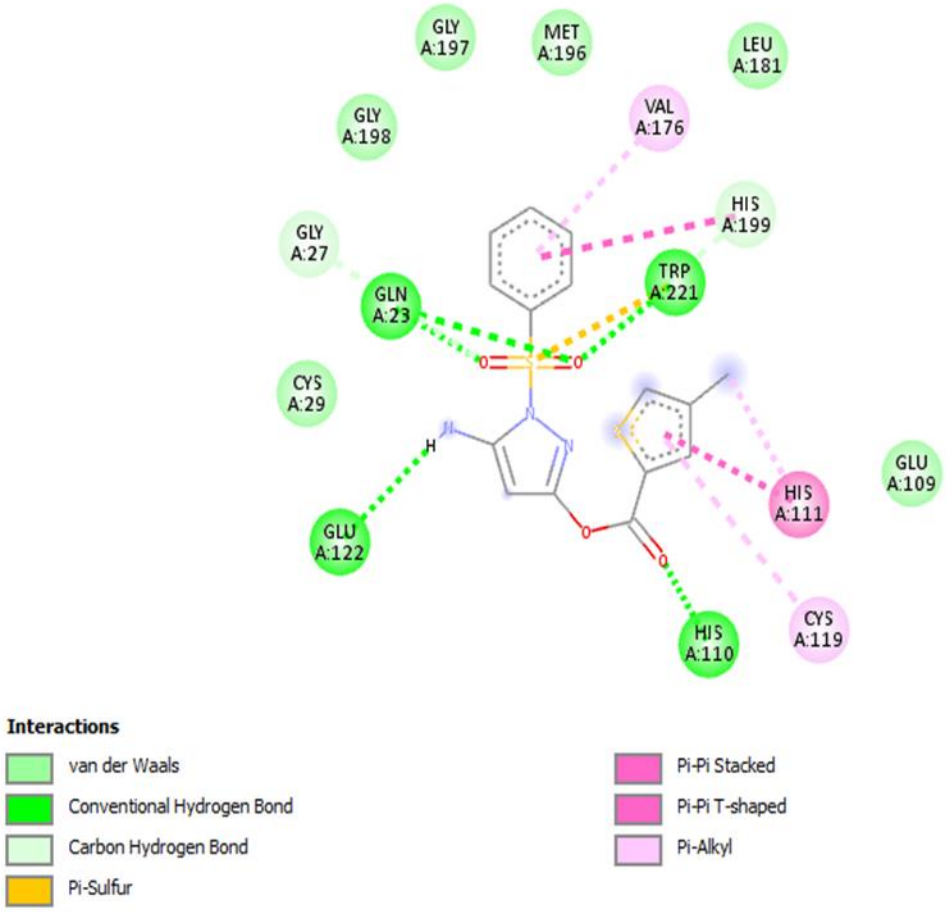



**Fig 15. Interaction with Molecule 2 with cathepsin B (PDB code: 1GMY)**

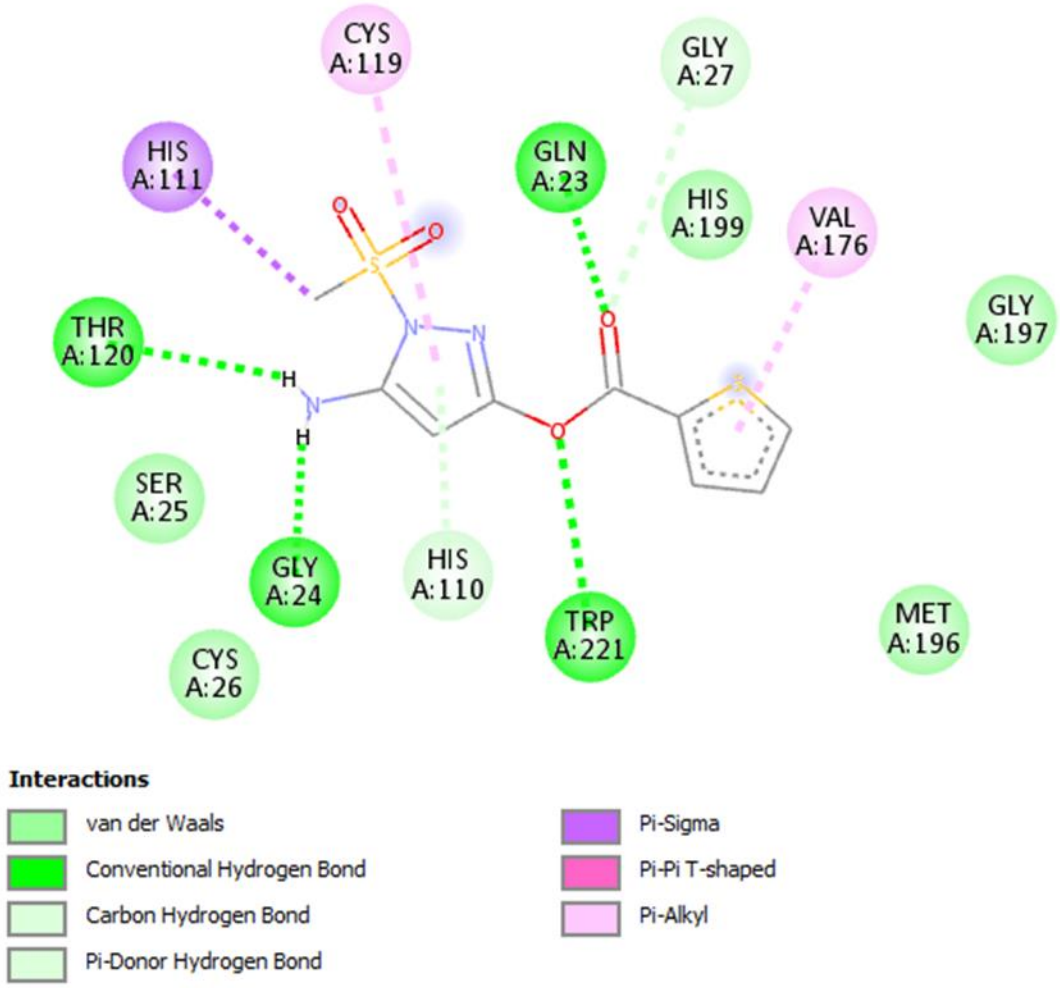



**Fig 16. Interaction with Molecule 15 with cathepsin B (PDB code: 1GMY)**

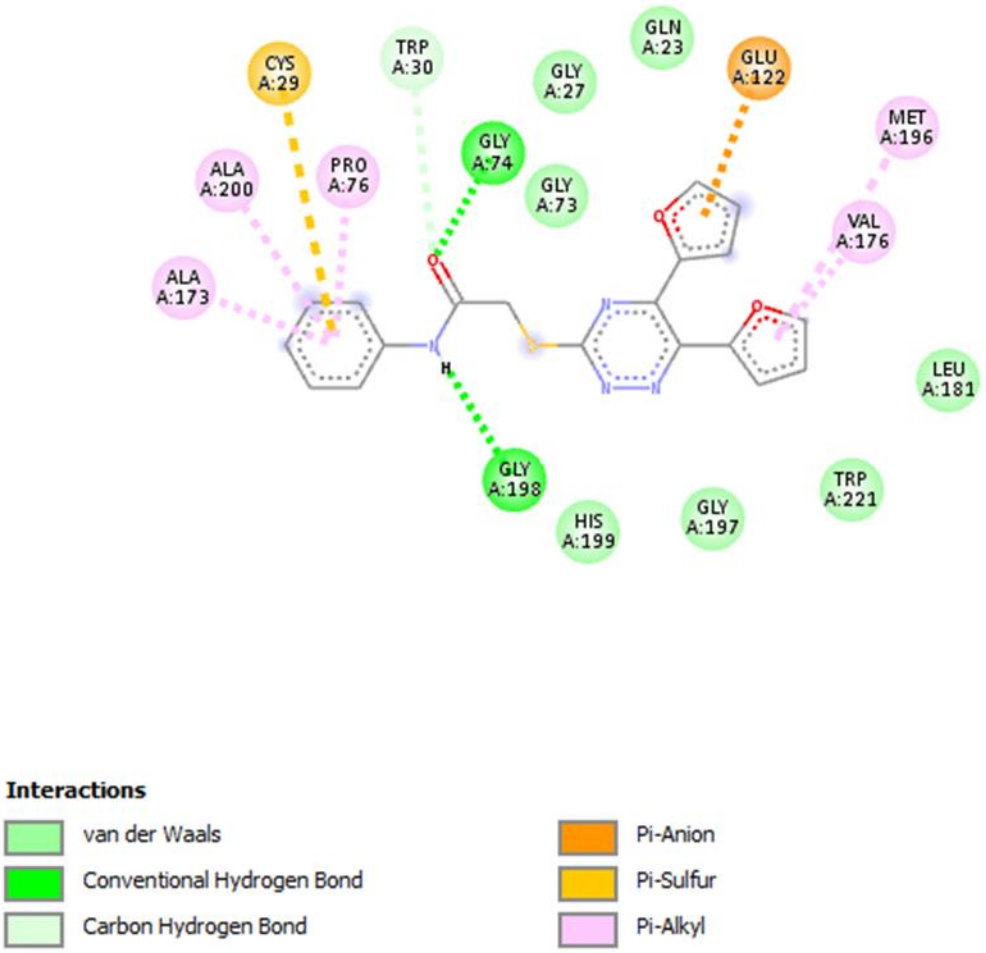



**Fig 17. Interaction with Molecule 16 with cathepsin B (PDB code: 1GMY)**

**Fig 18 Superposition of docked compounds 1,2,15 and 16 in the active site of cathepsin B ( courtesy: BIOVIA Discovery Studio Visualizer 2021)**

## 4. CONCLUSION

In the present article, a QSAR-MLR model for ordinary least squares is developed with a series of inhibitors of human cathepsin B using QSARINS software. Molecular descriptors are calculated using PaDEL software. Descriptor selection is done using genetic algorithm. This model fulfills all regulatory principles stated by the OECD. The robustness of the model is tested by internal validation (LOO, LMO and Y-scrambling) procedure and the predictability is determined by external validation. Four possible outliers are identified in the model application domain but, in the molecular docking study, these compounds are found to fit well inside the active site. The experimental bio activity data (IC50) of these outliers also show that they are good irreversible inhibitors.

## ACKNOWLEDGEMENT

The authors thank Prof. Paola Gramatica for the free license of QSARINS software. (QSARINS software web: www.qsar.it). The authors are grateful to The University Grant Commission, Government of India for funding.

**ETHICS APPROVAL AND CONSENT TO PARTICIPATE**

Not applicable.

**HUMAN AND ANIMAL RIGHTS**

No Animals/Humans were used for studies that are base of this research.

**CONSENT FOR PUBLICATION**

Not applicable.

**FUNDING**

None.

**CONFLICT OF INTEREST**

The authors have no conflict of interest.

**REFERENCES**

1. Barrett AJ and Kirschke H.; Cathepsin B, Cathepsin H, and cathepsin L.; Methods Enzymol 1981; 80 Pt C 535–561. [PubMed: 7043200]
2. Chapman HA Jr, Munger JS, Shi GP. The role of thiol proteases in tissue injury and remodeling. Am J Respir Crit Care Med 1994; 150: S 155–159. [PubMed: 7952652]
3. Chirico, N., Gramatica, P., Real external predictivity of QSAR models: how to evaluate it? comparison of different validation criteria and proposal of using the concordance correlation coefficient. J. Chem. Inf. Model. 2011; 51 2320.
4. Chirico, N., Gramatica, P. 2012 Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. J. Chem. Inf. Model. 2012; 52 2044
5. Chun Wei Yap PaDEL-descriptor: An open-source software to calculate molecular descriptors and fingerprints, Journal of Computational Chemistry; 17 December 2010 (http://padel.nus.edu.sg/software/padeldescriptor)
6. Eberhardt J., Santos-Martins D., Tillack A. F., and Forli S. 2021 Auto Dock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. Journal of Chemical Information and Modelling Aug 23;61(8) 3891-3898 Doi: 10.1021/acs.jcim.1c00203. Epub 2021 Jul 19
7. Friedman, J.H. Multivariate Adaptive Regression Splines. Annals of Statistics 1991;19 1-67.
8. Golbraikh, A. and Tropsha, A. Beware of $q^2$. J. Mol. Graphics Model. 2002; 20 269-276.
9. Gramatica P., Chirico N., Papa E., Cassani S, Kovarich S. QSARINS: a new software for the development, analysis and validation of QSAR MLR models J. Comput. Chem. 2013; 34 2121-2132
10. Gramatica, P., Chirico, N., Papa, E., Kovarich, S., Cassani, S. QSARINS: A New Software for the Development, Analysis, and Validation of QSAR MLR Models. Journal of Computational Chemistry, Software news and updates 2013; 34 2121-2132.

11. Gramatica, P., Cassani, S., Chirico, N. QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS. Journal of Computational Chemistry, Software news and updates 2014; 35 1036–1044.
12. Greenspan PD, Clark KL, Tommasi RA, Cowen SD, McQuire LW, Farley DL, van Duzer JH, Goldberg RL, Zhou H, Du Z, Fitt JJ, Coppa DE, Fang Z, Macchia W, Zhu L, Capparelli MP, Goldstein R, Wigg AM, Doughty JR, Bohacek RS, Knap AK;Identification of dipeptidyl nitriles as potent and selective inhibitors of cathepsin B through structure-based drug design. J Med Chem. 2001; 44 4524– 4534. [PubMed: 11741472]
13. Hall, L. H., and Kier, L. B. Electro topological state indices for atom types: A novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 1995; 35 1039-1045.
14. Kier, L. B. Molecular connectivity in chemistry and drug research (New York: Academic Press) 1976.
15. Matsumoto K., Mizoue K., Kitamura K., Tse W C, Huber C P , Ishida T ;Structural basis of inhibition of cysteine proteases by E-64 and its derivatives Biopolymers 1999 ;51(1) 99-107
16. Ojha, P.K., Mitra, I., Das, R.N., Roy, K.; Further exploring rm 2 metrics for validation of QSPR models. Chemom. Intell. Lab. Syst. 2011;107 194–205.
17. Rooprai HK and McCormick D.; Proteases and their inhibitors in human brain tumours: a review. Anticancer Res 1997; 17 4151–4162. [PubMed: 9428349]
18. Roberts LR, Adjei PN, Gores GJ.; Cathepsins as effector proteases in hepatocyte apoptosis. Cell Biochem Biophys 1999;30 71–88. [PubMed: 10099823]
19. Shi, L.M., Fang, H., Tong, W., Wu, J., Perkins, R., Blair, R.M., Branham, W.S., Dial, S.L., Moland, C.L., Sheehan, D.M., QSAR models using a large diverse set of estrogens. J. Chem. Inf. Comput. Sci. 2001;41 186–195.
20. Schüürmann , G., Ebert, R., Chen, J., Wang, B., Kühne, R.; External validation and prediction employing the predictive squared correlation coefficient - test set activity mean vs training set activity mean. J. Chem. Inf. Model. 2008; 48 2140–2145.
21. Todeschini, R. and Consonni, V.; Molecular descriptors for chemo informatics, (Weinheim: Wiley VCH) 2009; pg. 875-882, pg. 678-681.
22. Todeschini, R. and Consonni, V.; Molecular descriptors for chemo informatics, (Weinheim: Wiley VCH) 2009;27-37.
23. Trott O and Olson A. J.;Auto Dock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, Journal of Computational Chemistry 2010; 31, 455-461.