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: 10.26479/2024.1001.02**

QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS (QSAR) AND MOLECULAR MODELLING STUDIES OF A SERIES OF INHIBITORS FOR BONE MORPHOGENETIC PROTEIN 1 (BMP1)

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: Bone Morphogenetic Protein 1 (BMP1) inhibition is a potential method for treating fibrosis. BMP1, a member of the zinc metalloprotease family, is required to convert pro-collagen to collagen. A novel class of reverse hydroxamate BMP1 inhibitors was discovered, and cocrystal structures with BMP1 were obtained. This study builds an ordinary least squares Multiple Linear Regression model using a range of BMP1 inhibitors. A genetic algorithm is used to select a set of descriptors with very low correlation to characterize the biological activity IC50. The task is completed using software called QSARINS. The OECD has thoroughly validated the model. Its excellent predictive abilities, stability, and robustness are also studied. There is no chance correlation in the model fit. Six possible outliers are identified in the model application domain; however, they appear to bind appropriately in the protease active site according to molecular docking experiments.

Keywords: Multiple Linear Regression, Molecular Descriptors, Quantitative Structure-Activity Relationship, QSARINS, Bone morphogenetic protein1.

Article History: Received: Jan 02, 2024; Revised: Jan 08, 2024; Accepted: Jan 18, 2024.

Corresponding Author: Dr. Indrani Sarkar Ph.D.
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

Bone morphogenic protein1 (BMP1)/tolloid metalloproteinases also known as Procollagen -C proteinases are a small subgroup of the astacin family in vertebrates [11,12,14,15]. The BMP1 group of endopeptidases are multidomain, secreted, zinc endopeptidases. These proteinases are different from other members of the astacin family in having a cysteine rich loop region, Pro61-Cys-Gly-Cys-Cys-Ser66 that corresponds to the edge β-strand of matrix metalloproteases. An additional disulphide bond between cys62-cys65 is suggested in addition to the two conserved disulphide bonds present in astacin. This cysteine rich region acts as a flap above the active site helping to clamp the substrate into the active site of the proteinase. The substrate specificity of BMP1 differs from astacin markedly. Four types of bone morphogenic protein 1 (BMP1) proteinases have been identified in mammals. It includes the BMP1, mTLL1 and mTLL2 (mammalian tolloid like 1 and 2), in that order Mammalian tolloid (mTLD) is the longer splice variation of BMP1 [17,18]. In vertebrates, the cleavage of procollagen's solubilizing C terminal globular domain results in the synthesis of insoluble fibrillar collagen, which is catalyzed by the BMP1/Procollagen -C Proteinase. By biosynthetically digesting different precursor proteins into mature functioning enzymes, structural proteins, and proteins involved in starting the mineralization of the extracellular matrix (ECM) of hard tissues, this group of metalloproteinases plays important roles in controlling the creation of the ECM [20,23,24]. The production of extracellular matrix (ECM) is a tightly regulated process required for bone morphogenesis, normal wound healing and repair of bone fractures in adult. Excessive accumulation of ECM, particularly fibrillar collagen results in a variety of chronic fibrotic conditions including pulmonary, renal and liver fibrosis and scleroderma. Inhibition of Procollagen –C Proteinases/BMP1 may interfere in the progression of fibrosis and are thus important druggable targets as excessive fibrillogenesis of collagen can lead to a number of diseases. Some small molecules functioning as inhibitors of PCP have been reported. Hydroxamate derivatives of several diamino acids were described as

potent inhibitors of PCP. This report uses a series of inhibitors to develop a Multiple Linear Regression model for ordinary least squares.

2. MATERIALS AND METHODS

Structure of a molecule determines its characteristics. The goal of quantitative structure-activity research, or QSAR, is to link molecular structures to their biological roles. Several molecular characteristics, sometimes referred to as descriptors, are used to construct MLR models [2,3]. Models of QSAR are used to predict novel compounds with higher biological activity.

2.1 Preparation of data set

The ChEMBL database (ChEMBL ID: CHEMBL 3898; Uniprot Accession: P13497; Type: Single Protein; Organism: Homo sapiens) is the source of biological activity (IC50) data. The PubChem Database is used to compile the 3D structures of 261 inhibitors. The molecules are energy minimized to 500 steps of steepest descents using the MMFF94 force field, until the RMSD of potential energy is less than 0. 001. PaDEL program is used to get the descriptor values for the molecules. The descriptors are viewed as independent variables X, while the biological activity (IC50) is considered the dependent variable Y. For every pair of descriptors, the correlation between them is computed. Descriptors that have a correlation of greater than 95% are deemed highly correlated and are removed. In a similar way, descriptors with 0 values are eliminated. Descriptors with identical values for 80% of the compounds are removed. After 457 data are excluded from the computation, 563 data are used. Figures 1 and 2 show the experimental data distribution and variable profile for the descriptor ALogP. Using biological activity data, variable selection is applied to the reduced collection of descriptors. Thirty percent are utilized for testing, and seventy percent are used for training. The most significant descriptor variables are identified using the GA-VSS method (genetic algorithms with variable selection).

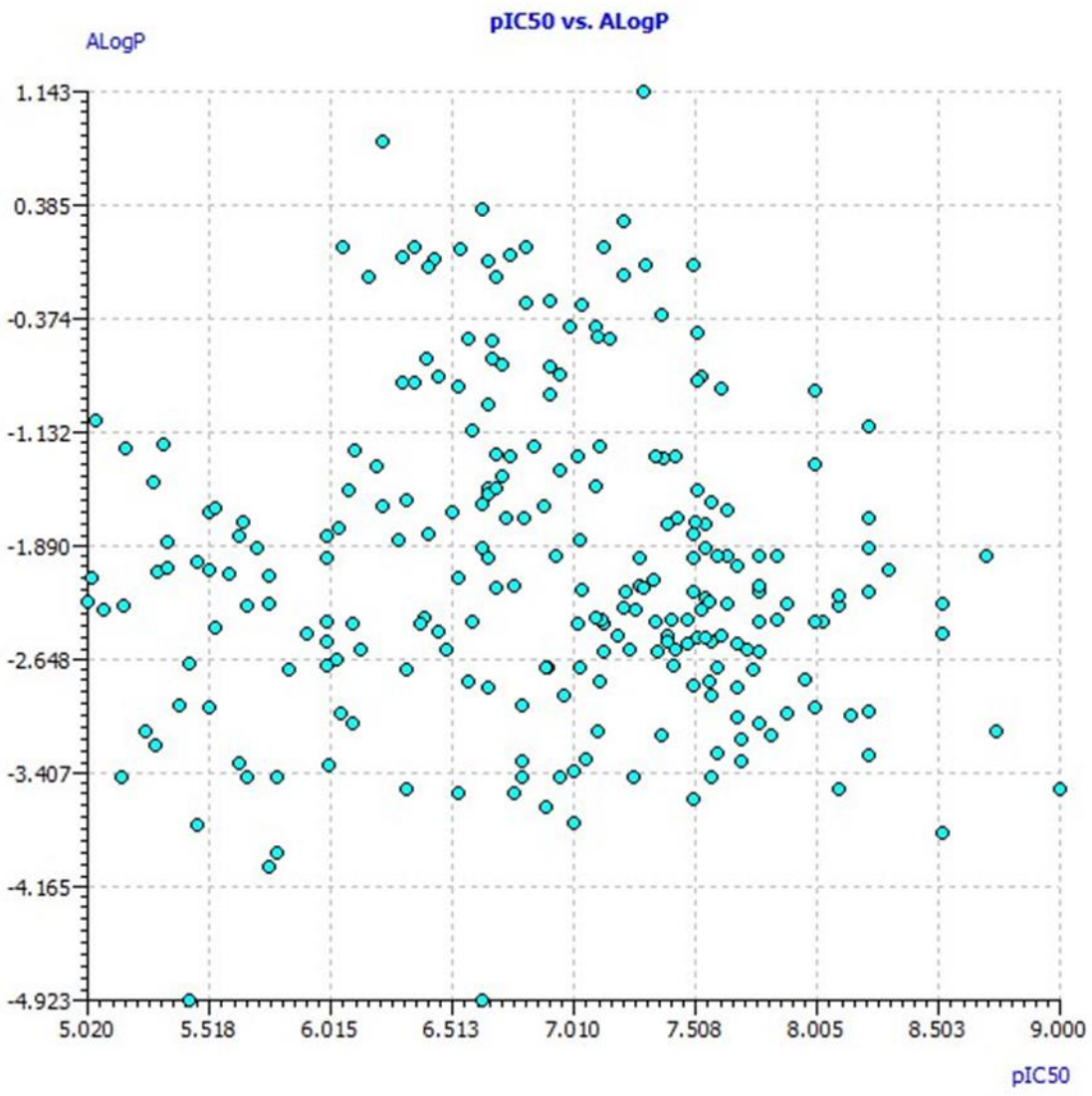

Fig 1 Variable profile

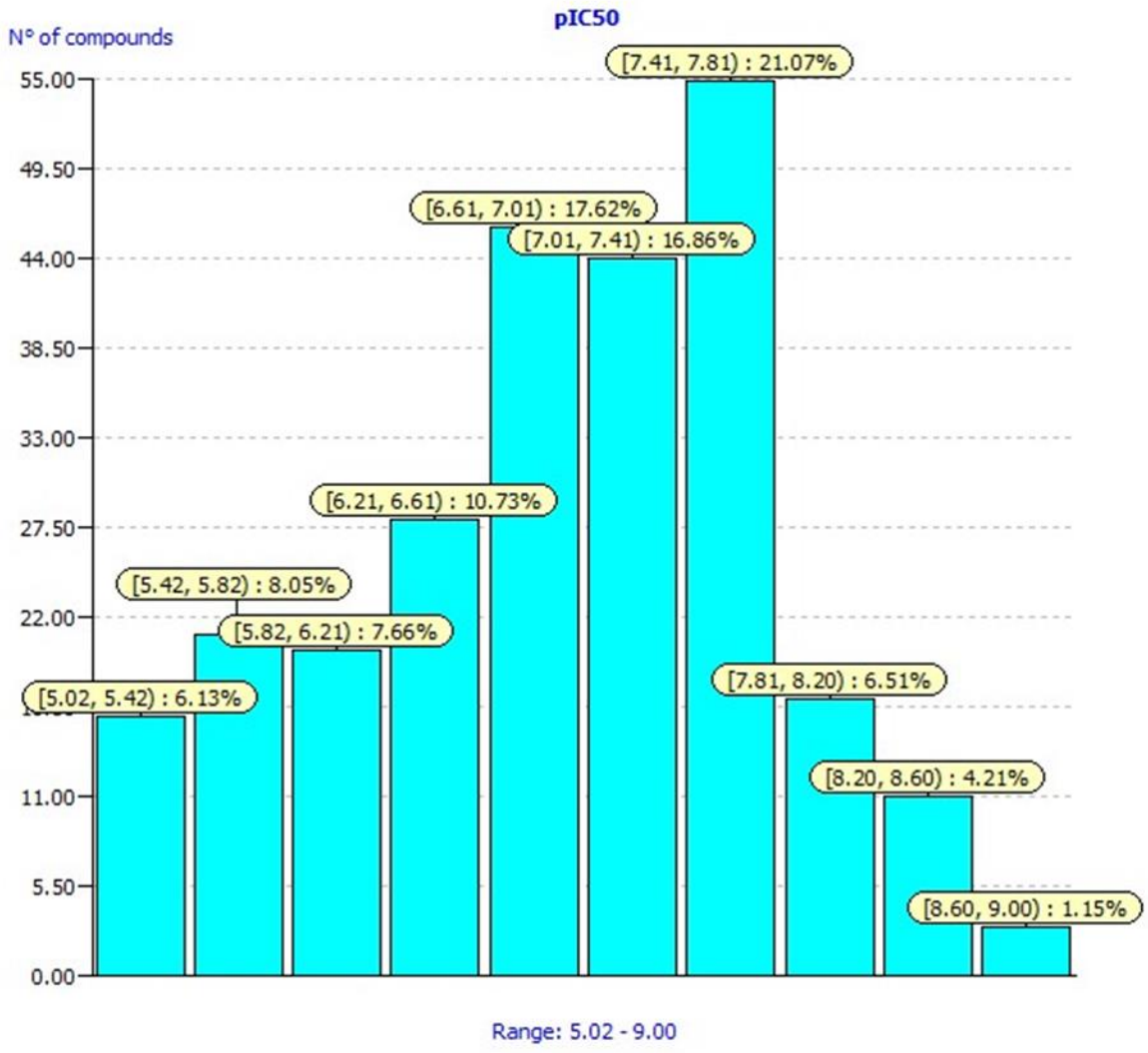

Fig 2 Distribution of the experimental data

2.2 Software

The computation of molecular descriptors is done using PaDEL Software [4]. QSARINS (QSAR-Insubria) software is used to build models [8,9]. A few descriptors are combined in each model. All possible combinations are examined using the all-subset approach. The genetic algorithm (GA) method is used to build models with more descriptors. There were several models made, with 1, 2, 3, 4, 5, and 6 descriptors in each.

2.3 Multiple Linear Regression Model

The MLR model displays a linear relationship between the molecular attributes or descriptors, and their biological activity, or IC50 (half maximum inhibitory concentration). The method makes use of Ordinary Least Squares (OLS) [8]. Based on R^2, the best models are ranked.

2.4 Fitting Criteria

Among the variables included in this are the Friedman lack of fit criterion, 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 [6]. The regression coefficient, or R^2, is used to evaluate the fitness of the model. It ought to be nearer 0 if the model is sound. More than 0.6 is an appropriate R^2 value for the QSAR model. R^2_{adj} value is indicated to avoid statistical incompatibility, whereas R^2 value increases as the number of descriptors increases. The model will become less accurate if extraneous variables are included. In a similar vein, adding important variables raises R^2_{adj}. R^2_{adj} will always be equal to or less than R^2. A model's LOF (Lack of Fit) should be around zero and not greater than 0.4 in order to have a smaller error. F (Fischer criteria) values should be high. This demonstrates that the model has purpose and was not produced randomly. K_{xx} [25,26] displays the whole correlation between the block of descriptors. It should have small value. The correlation between the responses and the descriptions is represented by the symbol K_{xy}. The model makes sense if $K_{xy} - K_{xx} < \delta_x$, where δ_x is a user-set threshold value. The mean absolute error, or MEA, in a fitting should be minimal. The mean absolute error in fitting, or MAE_{tr}, is calculated using the training set. The Root Mean Square Error is presented by $RMSE_{tr}$ in the training set. Residual Sum of Squares in the training set is known as RSS_{tr}. Using the training set, the Concordance Correlation Coefficients (CCC_{tr}) should be close to 1 and have a high value. The standard error of estimate, or s values, for RMSE training and validation should be near in model statistics.

2.5 Internal validation

If the averages of the R^2 and Q^2_{LOO} values are within a certain range, the model is said to be stable. Both Q^2_{LOO} and Q^2_{LMO} ought to be higher than 0.6. MAE and $RMSE_{cv}$ should both be under 0.5. $RMSE_{tr}$ should be smaller than $RMSE_{cv}$. Standard error s, $RMSE_{tr}$, and $RMSE_{cv}$ values need to be close. The Y-scrambling approach is used to verify that the model was not created by accidental correlation. There is no association between the responses and the descriptors since the experimental data or responses are distributed at random. As a result, the functioning of subsequent models should degrade rapidly. The model's reliability is assessed using iterated cross

validations. The cross-validated (CV) correlation coefficient is obtained using the equivalent Leave-One-Out (LOO) and Leave-Many-out (LMO) techniques (Q^2_{LOO}, Q^2_{LMO}). One compound (LOO) is removed from the descriptor collection, and the other compounds are then used iteratively to calculate a model. The model then makes a prediction for the one that was left out. If the value of Q^2_{LOO} is greater than R^2, the model may be considered appropriate. The Leave-More (or Many)-Out (LMO) method looks at the behaviour of the model when some compounds are omitted. The model is calculated using the remaining compounds after thirty percent of them are randomly excluded. Next, predictions are generated to evaluate the model's performance using compounds that are left out of the model. The model is considered stable if the averages of the R^2 and Q^2_{LOO} values fall within a specific range. Q^2_{LOO} and Q^2_{LMO} should both be more than 0.6. Both MAE_{cv} and $RMSE_{cv}$ must be less than 0.5. $RMSE_{tr}$ ought to be less than $RMSE_{cv}$. The values of $RMSE_{tr}$, $RMSE_{cv}$, and standard error s must be around. Y-scrambling method is employed to confirm that the model was not formed by chance correlation. Since the experimental data and replies are dispersed randomly, there is no correlation between the responses and the descriptors. Consequently, the performance of later models ought to deteriorate quickly.

2.6 External validation

After internal validation, the model is tested for its ability to predict new compounds. The model equation is applied to the excluded substances that have never before been used in a model computation. The model's performance is evaluated using a variety of metrics, such as $RMSE_{ext}$, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, r^2_m plus Δr^2_m, CCC_{ext}, and the Golbraikh and Tropsha [7] method. Anticipated values are for Q^2_{F1}, Q^2_{F2} and $Q^2_{F3} > 0.7$, for $CCC_{ext} > 0.85$, $R^2_{ext} > 0.6$, and $r^2_m > 0.6$. In this case, $RMSE_{ext}$ should be less than RMSE and equal to the overall error. The regression lines with slopes, k and k', lie between the 0.85 and 1.15 cutoff levels.

3. RESULTS AND DISCUSSION

QSARINS software is used to process the data (678 descriptors). Several MLR models are constructed using low multicollinearity (Table 1) between descriptors (Table 3). In Figure 3, the average R^2 and Q^2_{LOO} values are plotted versus the total number of variables. This demonstrates the models' efficacy in relation to their size. There was no benefit from the addition of the additional descriptor, as evident by the increased R^2 and Q^2_{LOO} values. The combinations of descriptors shared by the five-variable models are numerous. To predict the novel inhibitors outside of this dataset, the best MLR model (Model MLR1) with six descriptors is selected (Table

2). Statistical result of MLR1 is shown below:

(**Fitting criteria**)

R^2: 0.6444 R^2_{adj}: 0.6323 R^2-R^2_{adj}: 0.0121 LOF: 0.3006

K_{xx}: 0.3658 Delta K: 0.0197 RMSE $_{tr}$: 0.5125 MAE $_{tr}$: 0.4007

RSS $_{tr}$: 48.3378 CCC $_{tr}$: 0.7837 s: 0.5226 F: 53.4521

(Internal validation criteria)

Q^2_{loo}: 0.6131 R^2-Q^2_{loo}: 0.0313 RMSE $_{cv}$: 0.5346 MAE_{cv}: 0.4173

PRESS $_{cv}$: 52.5947 CCC $_{cv}$: 0.7652

Q^2_{LMO}: 0.6045 R^2_{Yscr}: 0.0332 Q^2_{Yscr}: -0.0455 $RMSE_{AV\ Yscr}$: 0.8450

(External validation criteria)

RMSE $_{ext}$: 0.7094 MAE $_{ext}$: 0.5411 PRESS $_{ext}$: 38.7555 R^2_{ext}:0.3252

Q^2-F_1: 0.2348 Q^2-F_2: 0.2280 Q^2-F_3: 0.3187 CCC $_{ext}$: 0.5410 r^2m aver.: 0.1761 r^2m delta: 0.1633

Calc. external data regression angle from diagonal: -20.5338°

Table 1. Correlation matrix shows that the descriptors have little correlation

	GATS8s	SpMin4_Bhi	VP-3	nHBint7	MAXDP	nHBAcc2
GATS8s	1.000					
SpMin4_Bhi	0.101	1.000				
VP-3	-0.101	0.529	1.000			
nHBint7	0.016	0.287	-0.110	1.000		
MAXDP	0.360	0.631	0.438	0.130	1.000	
nHBAcc2	-0.275	0.343	0.184	0.356	0.200	1.000

Table 2. Variables with their coefficients for mlr1 model

Variable	Coeff.	Std. coeff	Std. err.	(+/-) Co. int. 95%	p-value
GATS8s	1.4061	0.3263	0.2244	0.4429	0.0000
SpMin4_Bhi	-2.5667	-0.3155	0.5457	1.0769	0.0000
VP-3	0.4905	0.4756	0.0604	0.1193	0.0000
nHBint7	-0.1515	-0.4128	0.0190	0.0375	0.0000
MAXDP	-0.6626	-0.5804	0.0728	0.1437	0.0000
nHBAcc2	0.2765	0.4553	0.0323	0.0637	0.0000
Intercept	7.9356		0.5926	1.1695	0.0000

Table 3: 30 compounds taken from PubChem and used for validation

PubChem ID	IUPAC name of the compounds
44274338	(2S)-2-[1,3-benzodioxol-5-ylmethyl-(4-methoxyphenyl)sulfonylamino]-3-[(4-methoxyphenyl)methylsulfanyl]propanoic acid
44306397	(2R)-N-hydroxy-2-[(4-methoxyphenyl)sulfonylamino]-5-(2-thiophen-2-ylethylcarbamoylamino)pentanamide
44274123	(2R)-2-[(3-bromophenyl)methyl-(4-methoxyphenyl)sulfonylamino]-N-hydroxy-3-thiophen-2-ylpropanamide
44274346	(2R)-N-hydroxy-2-[(4-methoxyphenyl)sulfonyl-[(3-methyl-2-nitrophenyl)methyl]amino]propanamide
44306466	N-[(4R)-5-(hydroxyamino)-4-[(4-methoxyphenyl)sulfonylamino]-5-oxopentyl]cyclopropanecarboxamide
44306031	(2R)-5-[(3-fluorophenyl)carbamoylamino]-N-hydroxy-2-[(4-methoxyphenyl)sulfonylamino]pentanamide
44274266	(2R)-2-[(3-bromophenyl)methyl-(4-methoxyphenyl)sulfonylamino]-N-hydroxypropanamide
44274296	(2R)-N-hydroxy-2-[(4-methoxyphenyl)sulfonyl-[(4-methylsulfanylphenyl)methyl]amino]propanamide
44274143	(2S)-2-[1,3-benzodioxol-5-ylmethyl-(4-methoxyphenyl)sulfonylamino]-N-hydroxy-3-(4-methylphenyl)sulfanylpropanamide
44274141	(2R)-2-[(3,4-dichlorophenyl)methyl-(4-methoxyphenyl)sulfonylamino]-N-hydroxy-3-thiophen-2-ylpropanamide
44274308	(2R)-N-hydroxy-2-[(4-methoxyphenyl)sulfonyl-[[2-(trifluoromethyl)phenyl]methyl]amino]propanamide
44274114	(2R)-N-hydroxy-2-[(4-methoxy-3-methylphenyl)methyl-(4-methoxyphenyl)sulfonylamino]-3-thiophen-2-ylpropanamide
44274122	(2R)-N-hydroxy-2-[(4-methoxyphenyl)sulfonyl-[(2,4,5-trifluorophenyl)methyl]amino]-3-thiophen-2-ylpropanamide
44274203	(2R)-N-hydroxy-2-[(4-methoxyphenyl)sulfonyl-[(4-methylphenyl)methyl]amino]-3-thiophen-2-ylpropanamide
44305976	(2R)-N-hydroxy-2-[(4-methoxyphenyl)sulfonylamino]-5-[[(1R,2S)-2-phenylcyclopropyl]carbamoylamino]pentanamide
11857267	3-[[4-[(4-chlorophenyl)carbamoylamino]phenyl]sulfonyl-[2-(4-methoxyphenyl)ethyl]amino]-N-hydroxypropanamide
9954697	N-hydroxy-3-[[4-[(E)-N'-hydroxycarbamimidoyl]phenyl]sulfonyl-[2-(4-methoxyphenyl)ethyl]amino]propanamide
44274344	(2R)-2-[(3,5-difluorophenyl)methyl-(4-methoxyphenyl)sulfonylamino]-N-hydroxypropanamide
44318564	(2-bromophenyl)methyl N-[(2R,3R)-1-[[(2R)-1-(hydroxyamino)-1-oxo-3-(1,3-thiazol-4-yl)propan-2-yl]amino]-3-methyl-1-oxopentan-2-yl]carbamate
10001331	N-[[[(2R)-2-[[formyl(hydroxy)amino]methyl]heptanoyl]amino]methyl]-7-methoxy-1-benzofuran-2-carboxamide
118226146	N-[[[(2R)-2-[[formyl(hydroxy)amino]methyl]heptanoyl]amino]methyl]-5-phenylfuran-2-carboxamide
44318623	benzyl N-[(2R,3R)-1-[[(2R)-1-(hydroxyamino)-1-oxo-3-pyridin-2-ylpropan-2-yl]amino]-3-methyl-1-oxopentan-2-yl]carbamate
10117482	5-[(3R)-6-cyclohexyl-1-(hydroxyamino)-1-oxohexan-3-yl]-N-(4-methoxyphenyl)sulfonyl-1,2,4-oxadiazole-3-carboxamide

10205950	5-[(3R)-6-cyclohexyl-1-(hydroxyamino)-1-oxohexan-3-yl]-N-(4-methylphenyl)sulfonyl-1,2,4-oxadiazole-3-carboxamide
9890632	N-(benzenesulfonyl)-5-[(3R)-6-cyclohexyl-1-(hydroxyamino)-1-oxohexan-3-yl]-1,2,4-oxadiazole-3-carboxamide
44425141	(3R)-6-cyclohexyl-N-hydroxy-3-[3-[[2-(methanesulfonamido)ethylamino]methyl]-1,2,4-oxadiazol-5-yl]hexanamide
145976911	N-[[[(2R)-2-[(1R)-1-[formyl(hydroxy)amino]propyl]heptanoyl]amino]methyl]-7-methoxy-1-benzofuran-2-carboxamide
118225789	N-[[[(2R)-2-[(1R)-1-[formyl(hydroxy)amino]propyl]heptanoyl]amino]methyl]-5-phenylfuran-2-carboxamide
118225963	4-[5-[[[(2R)-2-[[formyl(hydroxy)amino]methyl]heptanoyl]amino]methylcarbamoyl]furan-2-yl]benzoic acid
118226041	3-[5-[[[(2R)-2-[[formyl(hydroxy)amino]methyl]heptanoyl]amino]methylcarbamoyl]furan-2-yl]benzoic acid

The model statistics show that R^2 is 0.6444 and R^2_{adj} is 0.6323. This implies that a new descriptor can be added to the model. The model had a good fit with the fewest number of descriptors and no overfitting, as indicated by the low value of the LOF parameter (0.3006). The low value of K_{xx} (0.3658) suggests that the model's properties don't really relate to one another. The model's delta K parameter (0.0197) shows a strong correlation between the descriptors and Log IC50. Additional estimated values (s = 0.5226; MAE_{tr} = 0.4007; $RMSE_{tr}$ = 0.5125) show a small computation error in the training set. Yellow dots in the scatter plot (Fig. 3) indicate the values predicted by the model equation compared to the experimental values.

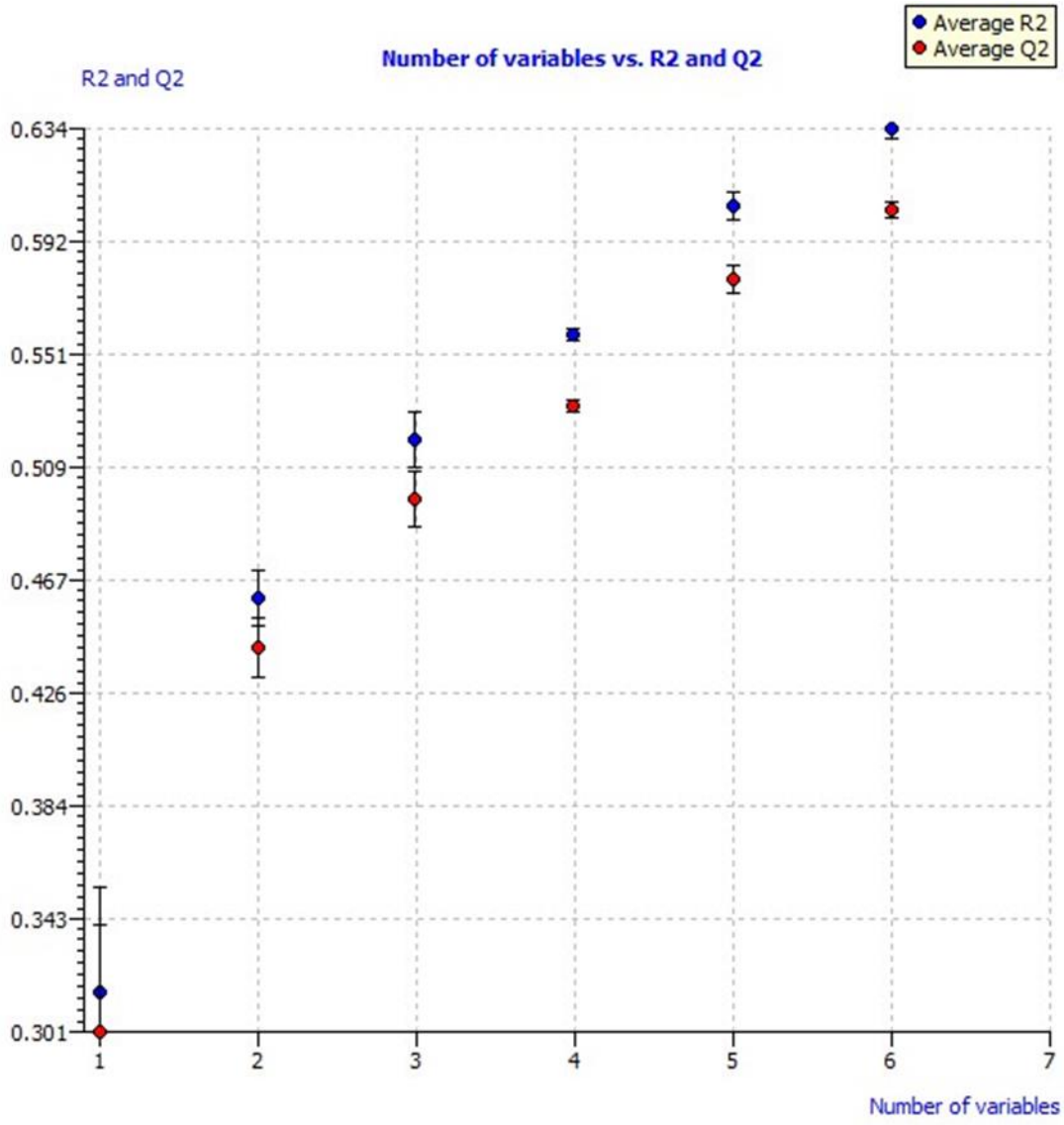

Fig 3: Average models R^2 vs Q^2

The blue points (test set) are calculated using the model equation, and the yellow points (training set) are obtained using the LOO approach. The scatter plot displays data outliers as well (Fig 4). Through internal validation, the model's resilience is evaluated. R^2 = 0.6444 and the variance discovered by LOO in its prediction (Q^2 LOO = 0.6131) are similar. As a result, the prediction from internal validation is accurate (Fig 5). Because of the low prediction error ($RMSE_{cv}$ = 0.5346 and MAE_{cv} = 0.4173), the model can be regarded as internally stable. Thirty percent of the dataset is excluded when using the Leaving Many-Out (LMO) technique. Given that Q^2_{LMO} (0.6045) and R^2 (0.6444) have similar values, the model might be regarded as stable. Q^2_{LOO}

(0.6131) and Q^2_{LMO} (0.6045) values are comparable. The Q^2_{LMO} vs K_{xy} plot (Fig. 6) shows a scatter plot of LMO models.

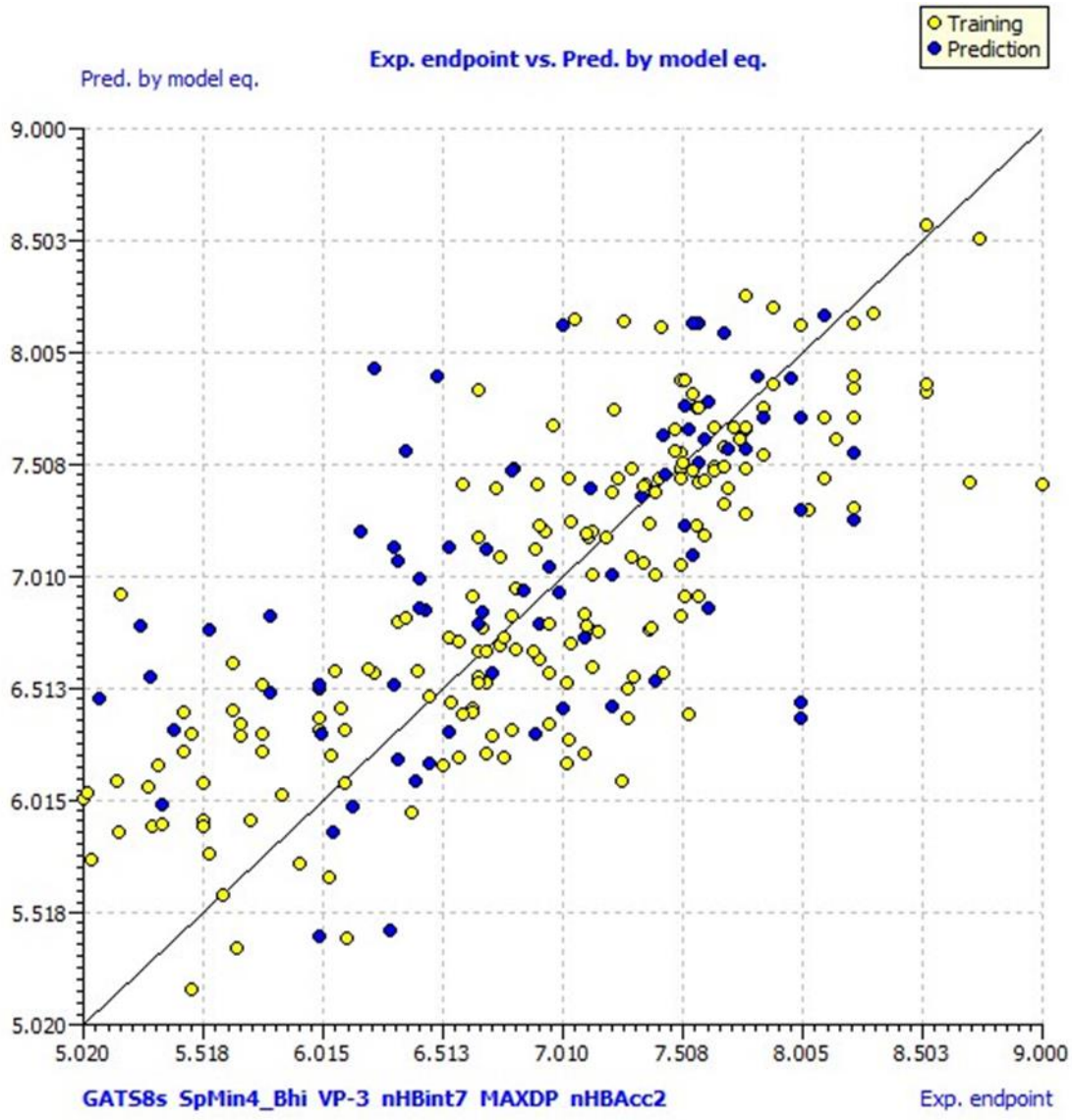

Fig 4. Scatter plot of Experimental endpoint vs. Predictions by the model equation

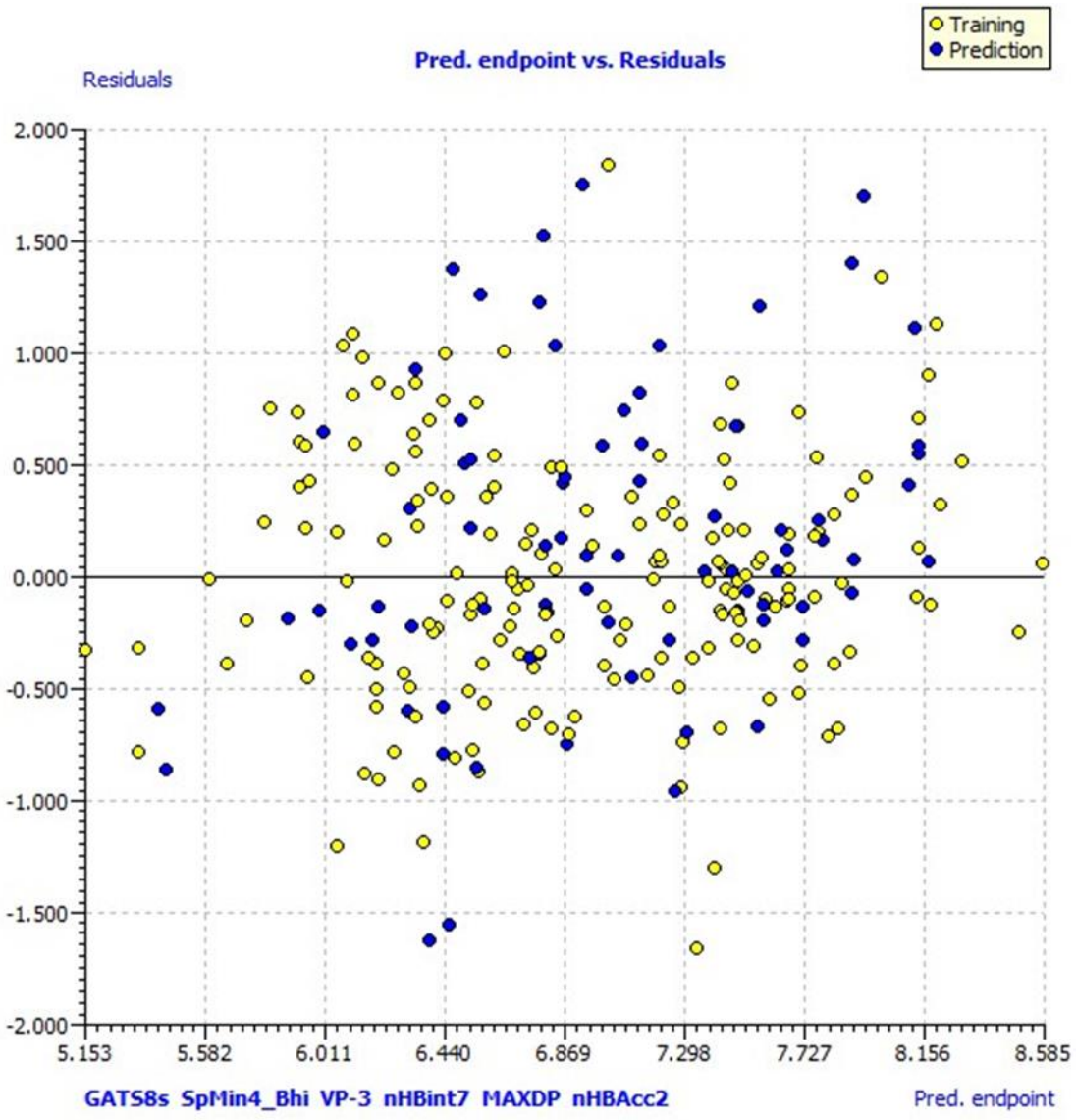

Fig 5. Residual plot of Experimental values vs. residuals from the LOO predictions. On the abscissa axes the values of the experimental values are reported, while on the ordinate the values of the residuals of the predictions are reported

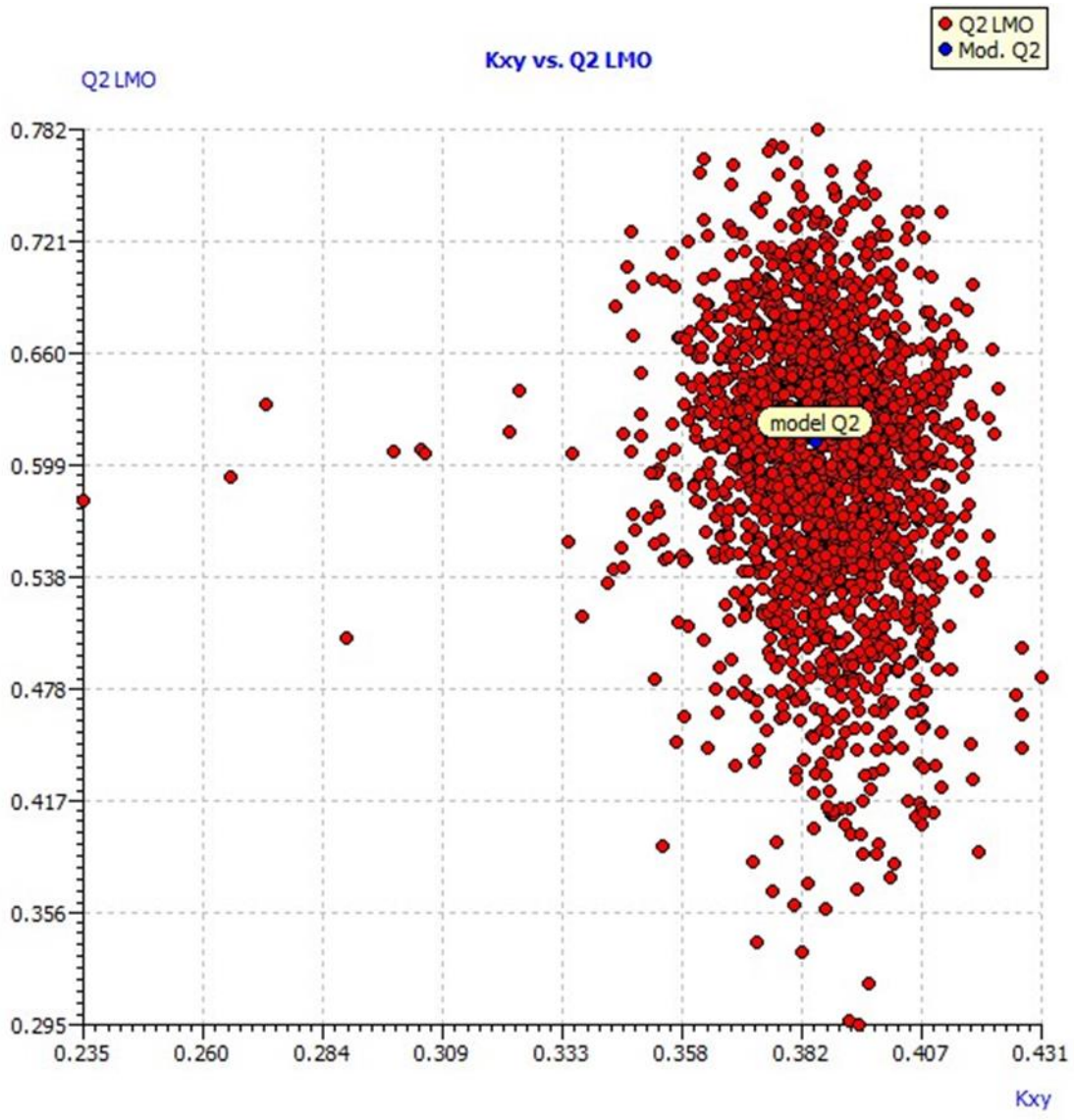

Fig 6. Plot of LMO models compared with the original QSAR model

The performance of the LMO model is indicated by the red points on the ordinate axes, whereas the blue point indicates the QSAR model as "model Q^2." Performance of the LMO variant is comparable to that of the original model. Y-scrambling is used to remove the possibility of random association. Q^2 Y-scr is -0.0455 and R^2 Y-scr is 0.0332. The R^2 Y-scr and Q^2 Y-scr values are plotted against R^2 and Q^2 of the model (Fig.7). It is found that the values obtained for these parameters using the Y-scrambling technique differ significantly from the values for R^2 and Q^2 in the model. This implies that the correlation in the model is not random. thought to be steady.

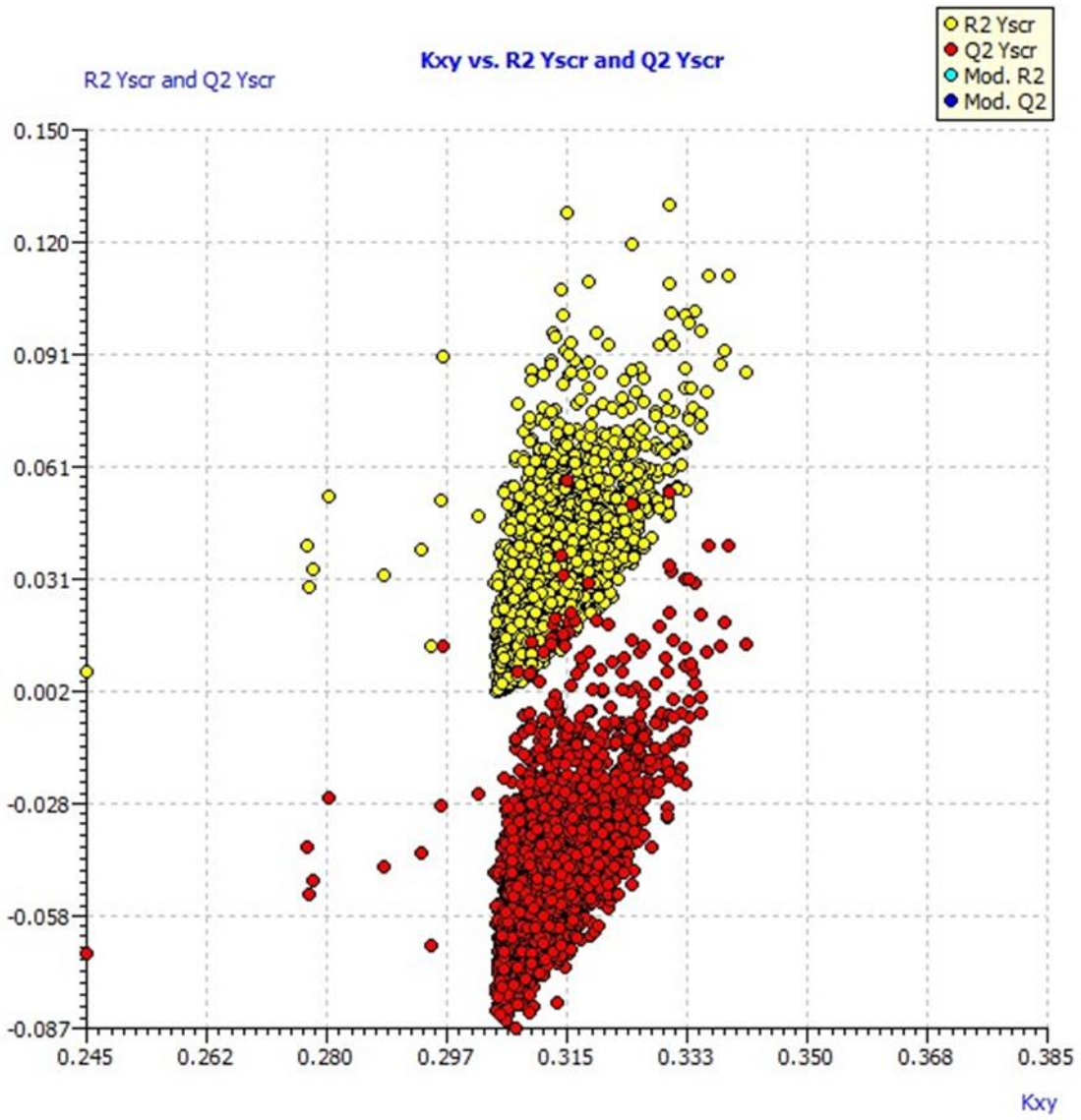

Fig 7. Scatter plot of Y-scrambled models compared to the original QSAR model

External validation is employed in order to assess the prediction potential of the model. The model's parameters R^2_{ext} = 0.3252, $RMSE_{ext}$ = 0.7094, MAE_{ext} = 0.5411, $PRESS_{ext}$ = 38.7555, Q^2 -F_1 = 0.2348, Q^2 -F_2 = 0.2280, Q^2 - F_3 = 0.3187, CCC_{ext} = 0.5410, r^2_m_aver = 0.1761, r^2_m_delta =0.1633) are in agreement with the results. Here, R^2_{ext} stands for the coefficient of determination of the external validation process [7]. Q^2 -F_1 [21], Q^2 -F_2 [22], and Q^2 -F_3 [25,26] measure the variances provided in external validation; CCC_{ext} is the Concordance Correlation Coefficient [2,3], and r^2_m_aver and r^2_m_delta are the Roy criteria average and delta [19]. MAEext is the Mean Absolute Error, while MSE_{ext} measures the Root Mean Square Error. $PRESS_{ext}$ stands for

Predictive Residual Sum of Squares. A q-q plot of experimental vs residual values from LOO is shown in Fig. 8.

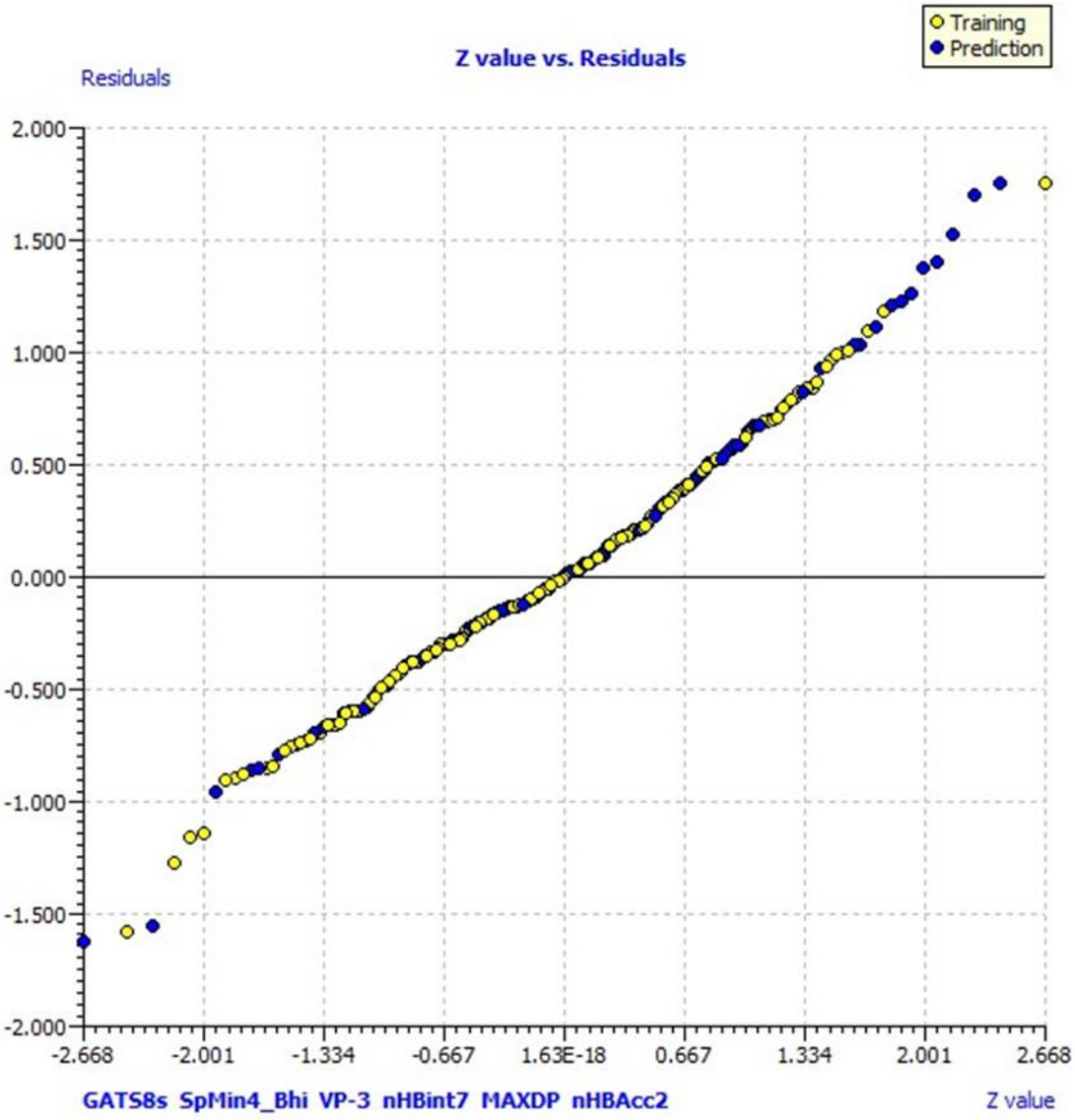

Fig 8. q-q plot of experimental endpoint vs. residuals from the predictions by model equation

The values of the theoretical quantiles (Z values) are plotted on the abscissa, while the values of the residuals from the predictions are represented on the ordinate. While the yellow points (training set) indicate values predicted by LOO, the blue points are obtained using the model equation. By averaging the individual predictions made by each model for every molecule (MLR1 to MLR6), the Average Combined Prediction (ACM) is produced. The William graph of the combined model (Fig. 9) shows that most of the compounds are inside the application area of the model

(within the critical leverage h*= 0.114). The ordinate represents the projected residuals, and the abscissa represents the HAT values of the diagonal elements. The combined model equation yields the prediction set represented by the blue dots, while the LOO's training set is represented by the yellow points. The horizontal dashed lines indicate the user-defined threshold for Y-outliers. HAT readings that exceed the cutoff value, h* = 3p/n, are considered outliers.

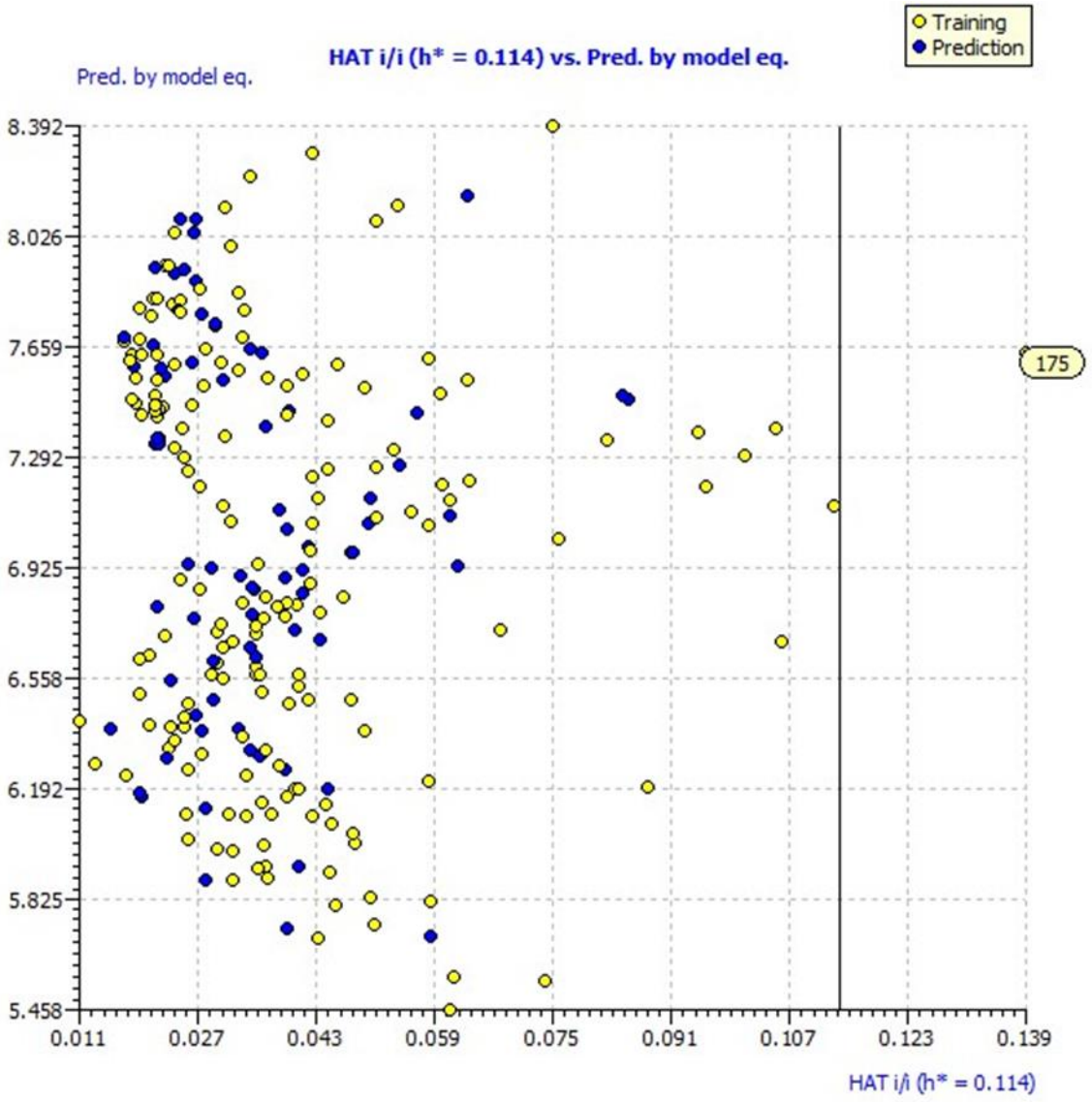

Fig 9 Insubria graph diagonal hat elements vs. predictions by the combined model equation

In this case, p is the number of model variables plus one, and n is the number of objects. The QSARINS-provided Insubria graph (Fig. 10) additionally shows the applicability domain [9]. In this case, the abscissa displays the HAT diagonal values, while the ordinate displays the projected

data. When the experimental value is known, the combined model predicts the data points in the blue (prediction set) and yellow (training set), respectively [13, 10]. The 30 compounds that were examined using the model equation are listed in Table 3. The outcomes are shown in Table 4. In order to determine the average performance of all the models through combined modelling, a list of models is selected using PCA. Among the chosen models are:

Model 1: 7.9356 + (1.4061) GATS8s +(-2.5667) SpMin4_Bhi +(0.4905) VP-3 +(-0.1515) nHBint7 +(-0.6626) MAXDP + (0.2765) nHBAcc2

Model 2: 5.8222 + (0.8574) GATS8s + (-3.2286) SpMin4_Bhi + (0.4845) VP-3 + (1.0203) naaO + (0.0453) minHBint5 + (0.1728) nHBAcc2

Model 3: 6.2781 + (0.9409) GATS8s + (-3.3886) SpMin4_Bhi + (0.4192) VP-3 + (-0.0660) nHBint7 + (0.9097) naaO + (0.2098) nHBAcc2

Model 4: 9.2678+ (0.2033) GATS8s + (-5.0378) SpMin4_Bhi + (0.3910) VP-3 + (1.0173) naaO + (0.0587) minHBint5 + (0.0004) WPATH

Model 5: 5.4598+ (1.5839) GATS8s+ (-3.3420) SpMin4_Bhi + (0.5040) C3SP3 + (0.4194) VP-3 + (-0.1437) nHBint5 + (0.2231) nHBAcc2

Model 6: 6.3611+ (1.0569) GATS8s +(-3.9611) SpMin4_Bhi +(0.4085) VP-3 +(0.1426) CrippenLogP +(0.9491) naaO+(0.2185) nHBAcc2

Model 7: 6.8281+(0.7500) GATS8s+(-3.5204) SpMin4_Bhi +(0.3979) VP-3 +(1.1891) naaO +(-1.8318) hmin +(0.1767) nHBAcc2

Model 8: 6.0899+(1.0625) GATS8s+(-3.6517) SpMin4_Bhi+ (0.4767) VP-3 +(-0.0641) nHBint9+(1.0019) naaO +(0.2206) nHBAcc2

Model 9: 6.0065+(0.9000) GATS8s +(-4.0116) SpMin4_Bhi +(0.3022) VP-3 +(1.1248) naaO +(0.0630) minHBint5+(0.0342) TIC0

Model 10: 9.0420+(0.9838) GATS8s +(-5.2311) SpMin4_Bhi +(0.3296) VP-2+(1.0257) naaO +(0.0672) minHBint5 +(0.0003) WPATH

(Fitting criteria)

$R2_{ACM}$: 0.6632 R^2_{WCM}: 0.6852 MAE_{tr}: 0.3947 $RMSE_{tr}$: 0.4997 CCC_{tr}: 0.7895

(External validation criteria)

MAE_{ext}: 0.5213 $RMSE_{ext}$: 0.6723 CCC_{ext}: 0.5718

Q^2-F1: 0.3128 Q^2-F2: 0.3067 Q^2-F3: 0.3881

Calc. external data (ACM) regression angle from diagonal: -20.1616°

Calc. external data (WCM) regression angle from diagonal: -19.7391°

Table 4: Comparison of experimental values of PIC_{50} and values obtained from model equation of MLR1, MLR2, MLR3, MLR4, MLR5, MLR6

Compound ID	MLR1	MLR2	MLR3	MLR4	MLR5	MLR6	experimental data pIC50
44274338	6.7	6.4	6.0	6.3	5.9	5.8	6.6
44306397	7.2	6.9	7.5	7.4	6.8	6.9	7.11
44274123	6.6	6.7	6.8	6.9	6.4	5.9	6.58
44274346	6.7	6.5	6.4	6.2	6.3	6.6	6.92
44306466	6.3	5.9	6.4	6.7	5.8	5.6	6.11
44306031	6.6	6.5	6.1	6.4	6.2	5.9	6.8
44274266	7.2	6.8	7.2	7.6	7.5	7.8	7.03
44274296	7.3	6.7	7.5	7.6	7.5	7.8	7.1
44274143	7.0	6.9	7.2	7.6	7.5	7.9	7.35
44274141	6.5	6.7	6.8	6.2	6.9	5.9	6.36
44274308	8.1	8.4	8.5	7.9	7.8	7.7	8.22
44274114	6.7	6.5	6.4	6.9	5.9	5.8	6.31
44274122	6.9	6.5	6.4	6.2	6.3	6.6	6.75
44274203	7.1	6.5	6.4	6.2	6.3	5.9	6.96
44305976	6.8	6.5	6.4	6.2	6.3	5.8	6.77
11857267	6.7	6.5	6.4	6.2	6.3	7.1	6.52
9954697	7.3	6.8	7.7	7.6	7.5	6.8	7.1
44274344	6.7	6.5	6.4	6.2	6.3	5.8	6.85
44318564	7.0	6.8	7.2	7.6	7.5	6.7	7.11
10001331	6.7	6.5	6.4	6.2	5.9	5.8	6.8
118226146	7.8	7.0	7.2	7.9	7.5	6.9	7.6
44318623	6.7	6.5	6.4	6.2	6.3	6.6	6.66
10117482	7.0	6.8	7.2	7.6	7.5	7.8	7.77
10205950	7.0	6.8	7.2	7.6	7.5	7.8	7.85
9890632	8.1	8.4	8.5	7.9	7.8	7.7	8.15
44425141	7.0	6.8	7.2	7.6	7.5	7.8	7.7
145976911	6.7	6.5	6.4	6.2	6.3	7.3	6.9
118225789	7.1	6.8	7.2	7.6	7.5	7.8	7.5
118225963	8.8	8.7	8.5	8.5	8.6	9.4	9
118226041	8.3	8.4	8.5	7.9	7.8	7.5	8.1

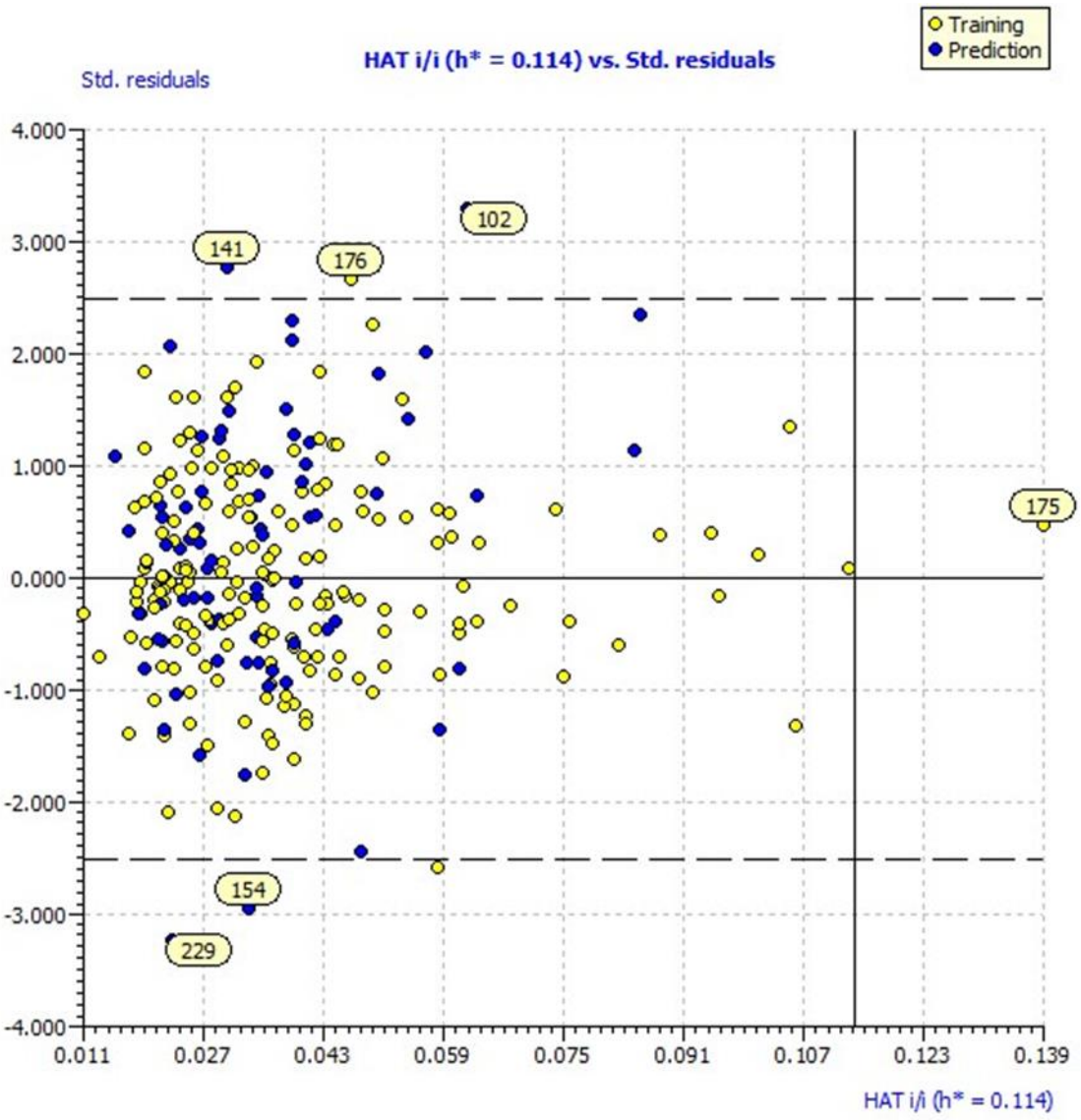

Fig 10. Williams's plot predicted by combined model equation using ACM

BMP-1, or bone morphogenetic protein, is a member of the astacin-like zinc metalloproteinases tolloid subgroup. The procollagen C-pro peptides are broken down by this procollagen C-proteinase (PCP). In each of the three procollagen chains, cleavage takes place between an invariant aspartic acid residue and either a particular alanine or glycine residue, depending on the procollagen chain. For BMP-1's PCP activity, the glutamic acid residue (Glu94) in the consensus sequence HEXXH is crucial. The residues 170–184 of the S1' loop form one side of the BMP-1 S1' pocket. Lys176 is present in this loop. The BMP-1 S1' pocket is defined by the Lys87 and Lys176-. In the P1' position of the procollagen chains, the positively charged side chains of these lysyl residues connect to the acidic side chain of aspartic acid. The 3-D structural coordinates of

BMP1 (Classification-Hydrolase, Organism-Homo Sapiens) complexed with a reverse hydroxamate inhibitor at a resolution of 1.45 A [1] is taken from Protein Databank (PDB code: 6bsl) (Fig 11). There are two chains A and B in the structure of 6bsl.

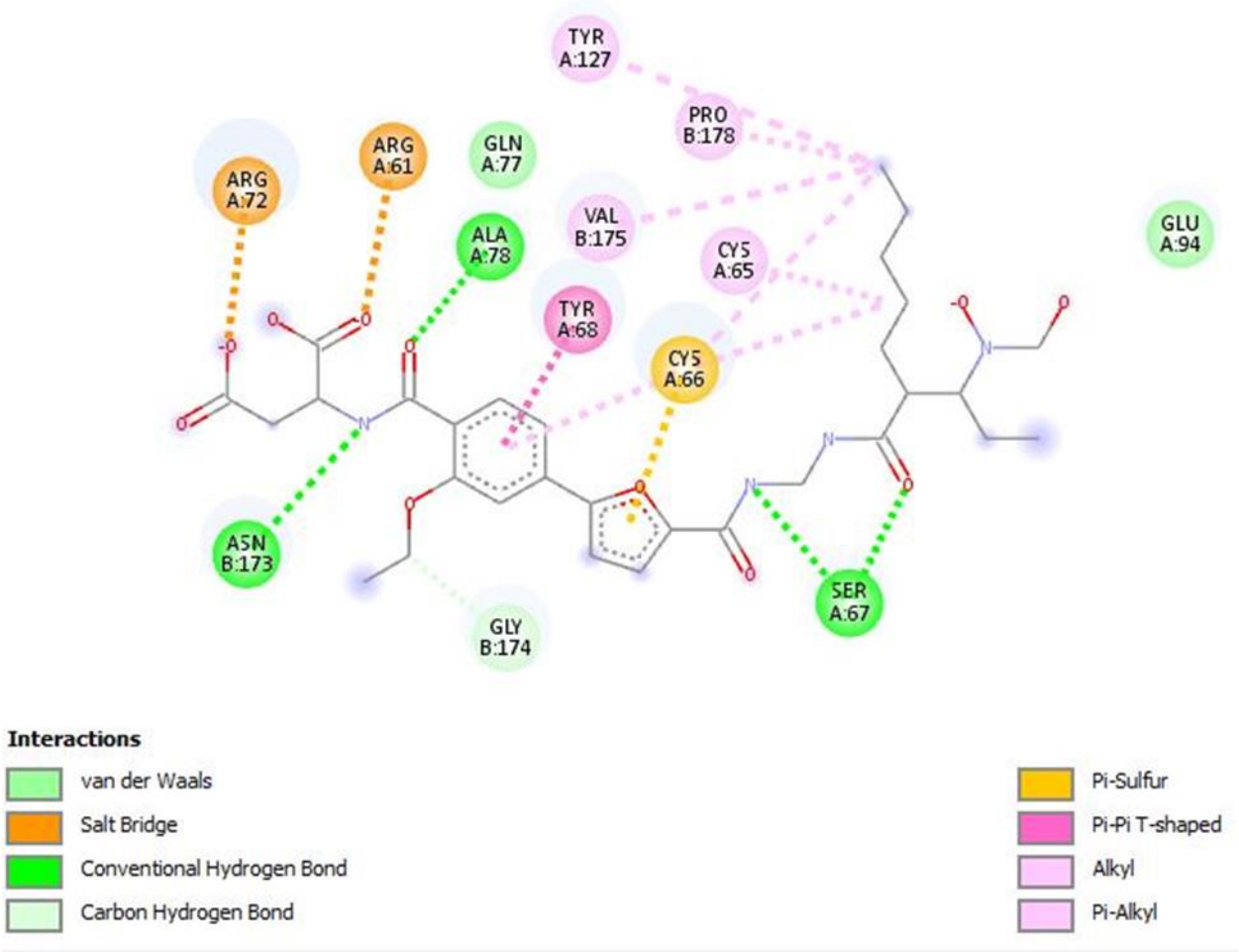

Fig11: BMP1 (Classification-Hydrolase, Organism-Homo Sapiens) complexed with a reverse hydroxamate inhibitor at a resolution of 1.45 A (PDB code: 6bsl)

Docking uses the 202 amino acids in Chain A. within BMP-1. In order to avoid short interactions, partial charges and hydrogen atoms are added to the protein structure (6bsl), and energy minimization is done with an OPLS force field. Using the minimized protein structure, the Auto Dock Vina software [5, 27] docks the compounds (x and y outliers) into the binding site. The outliers binding at the protein's active site resembles that in the crystal structure, based on the computer modelling approach. It is shown that by creating hydrogen bonds and

pi pi stacking interactions, the docked molecules stabilize in the active site (Fig 12, 13, 14, 15, 16, 17, 18, 19). The optimal docked poses of the outliers are superimposed in the 1RT2 active site (Fig. 18). Their experimental IC50 data reveal that they are likewise potent inhibitors (CHEMBL).

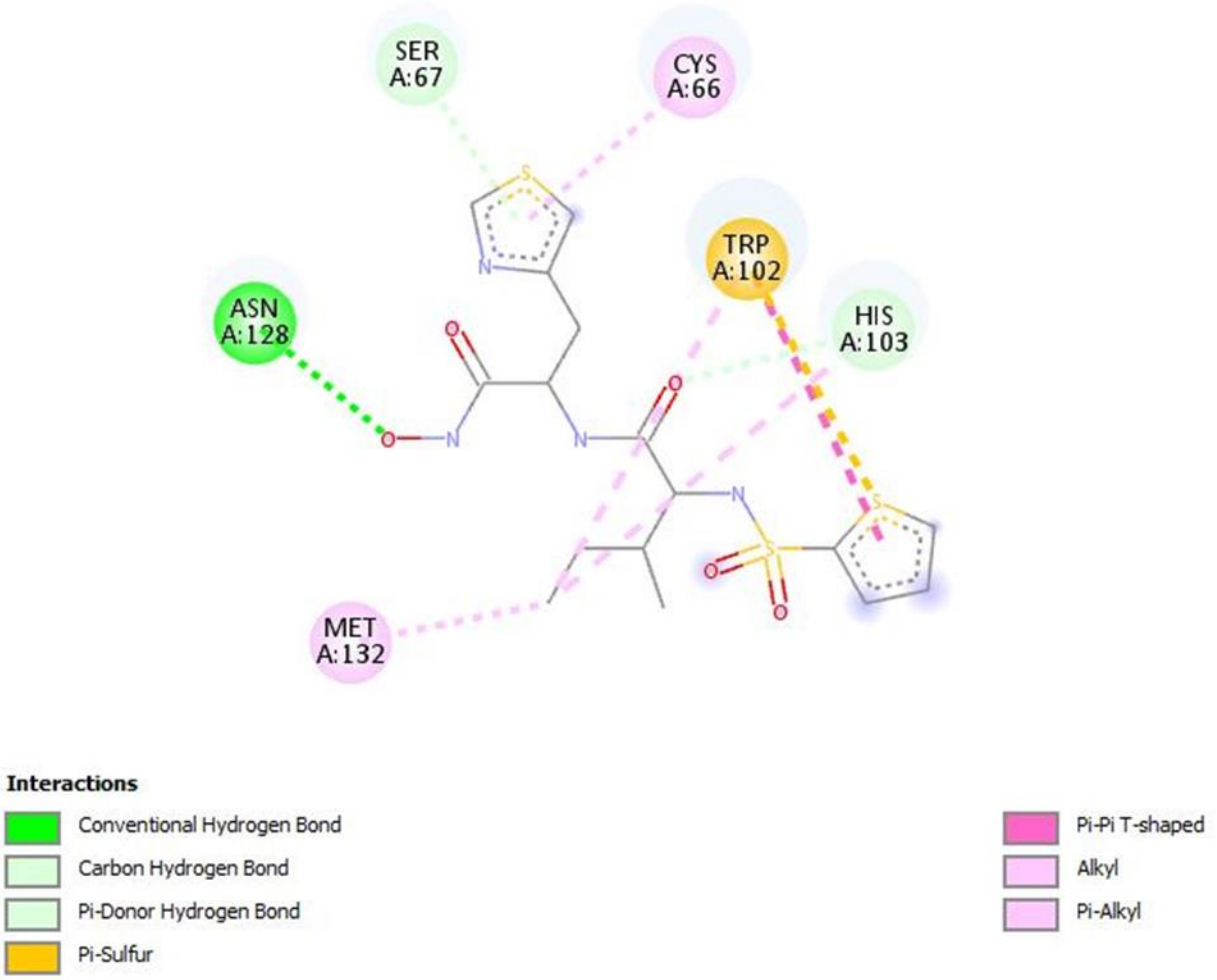

Fig 12. Interaction with Molecule 175 with BMP1

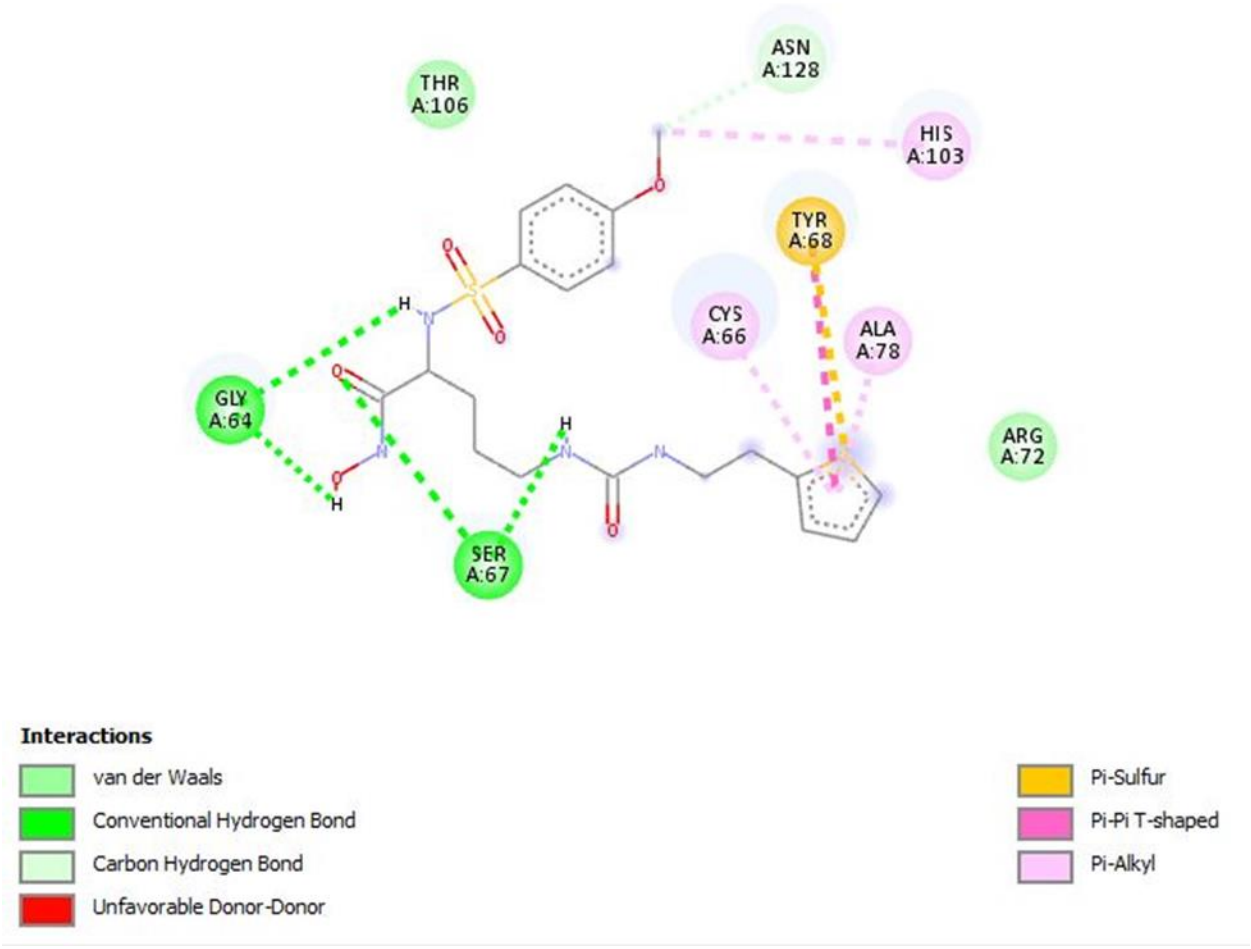

Fig 13. Interaction with Molecule 141 with BMP1

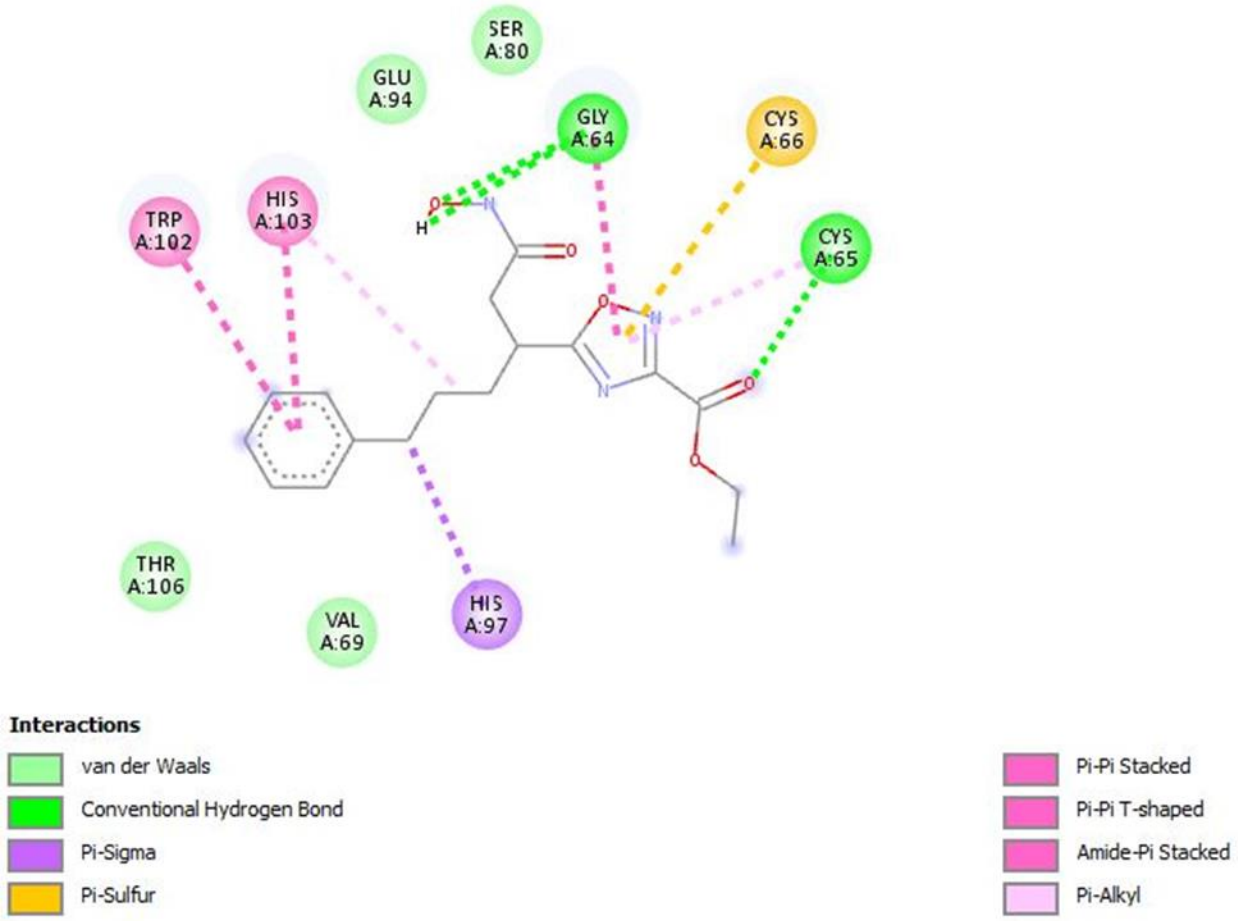

Fig 14. Interaction with Molecule 176 with BMP1

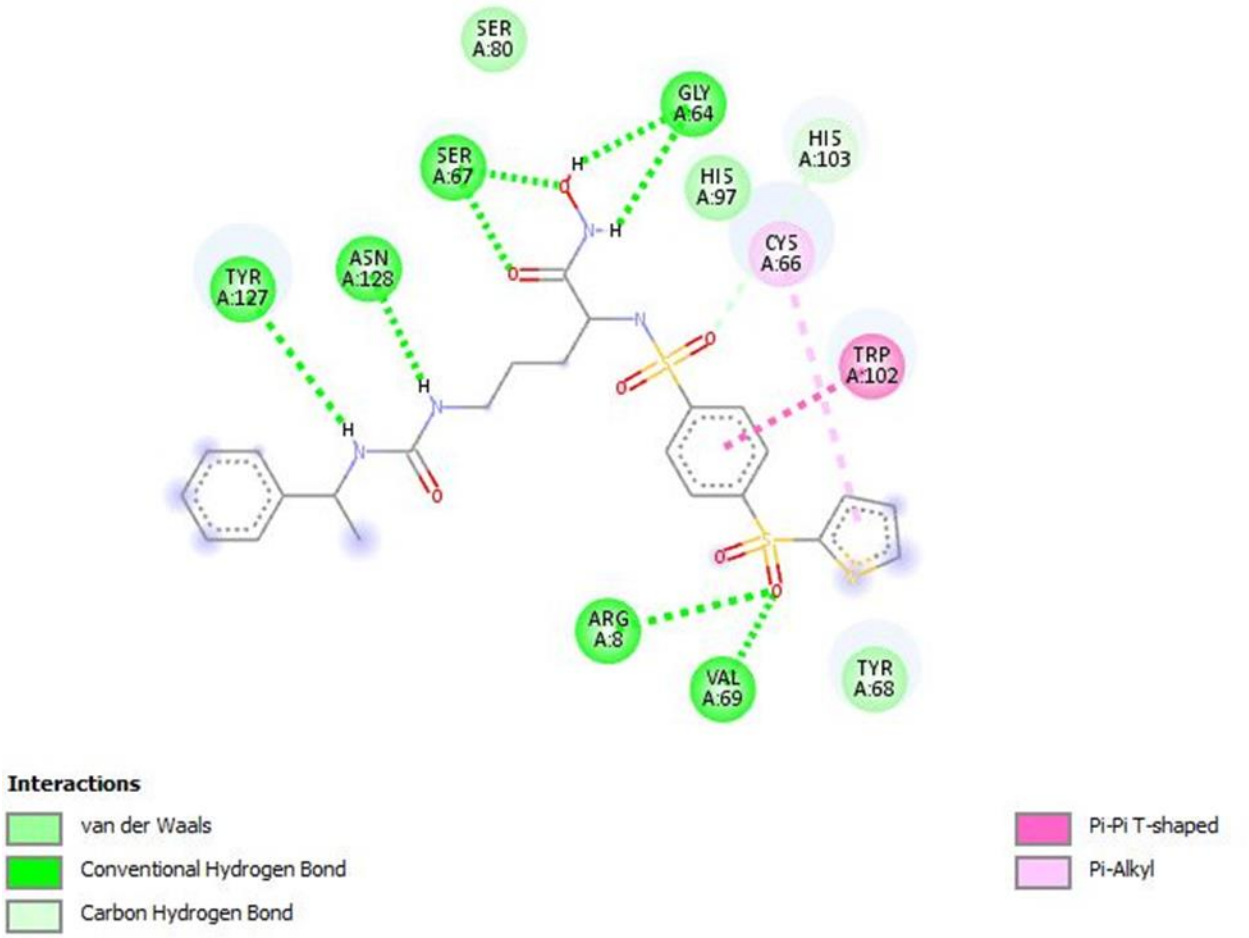

Fig 15. Interaction with Molecule 102 with BMP1

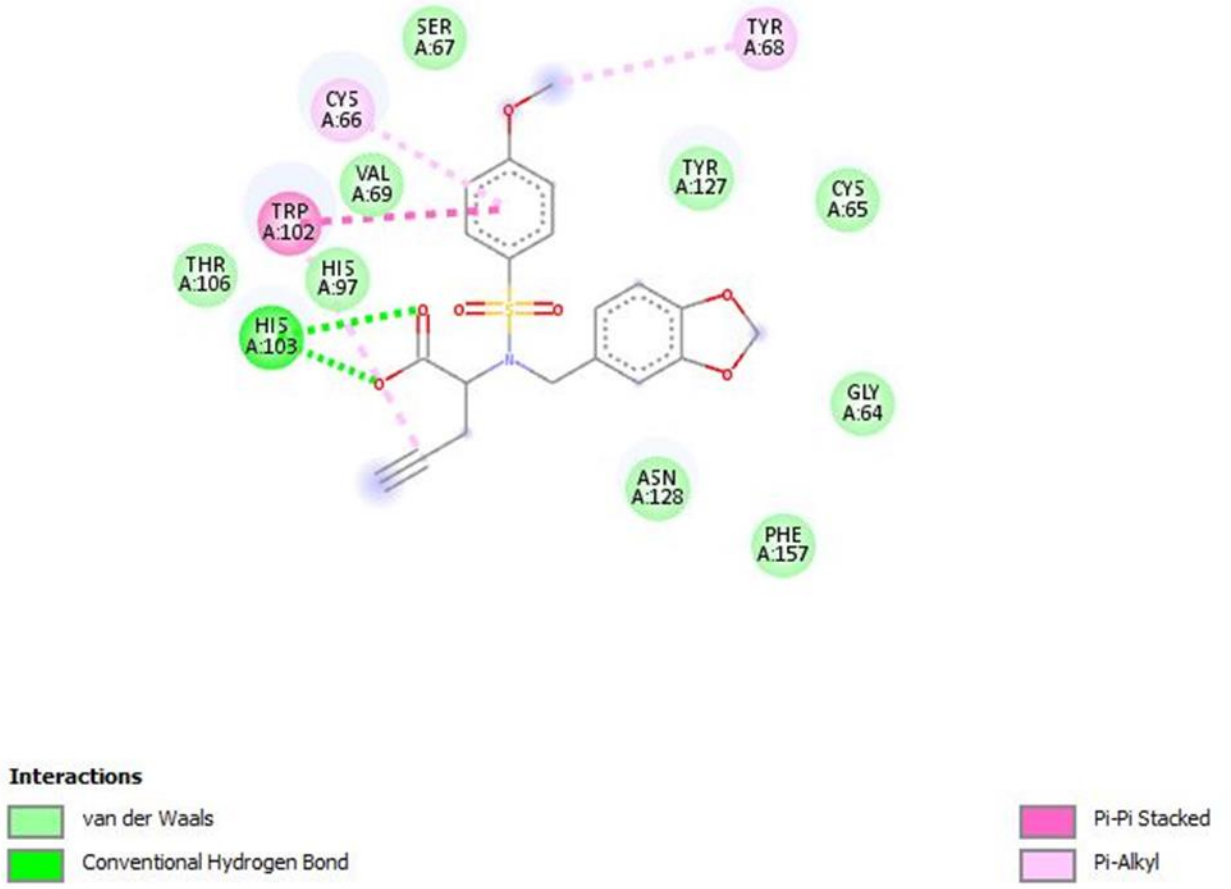

Fig 16. Interaction with Molecule 229 with BMP1

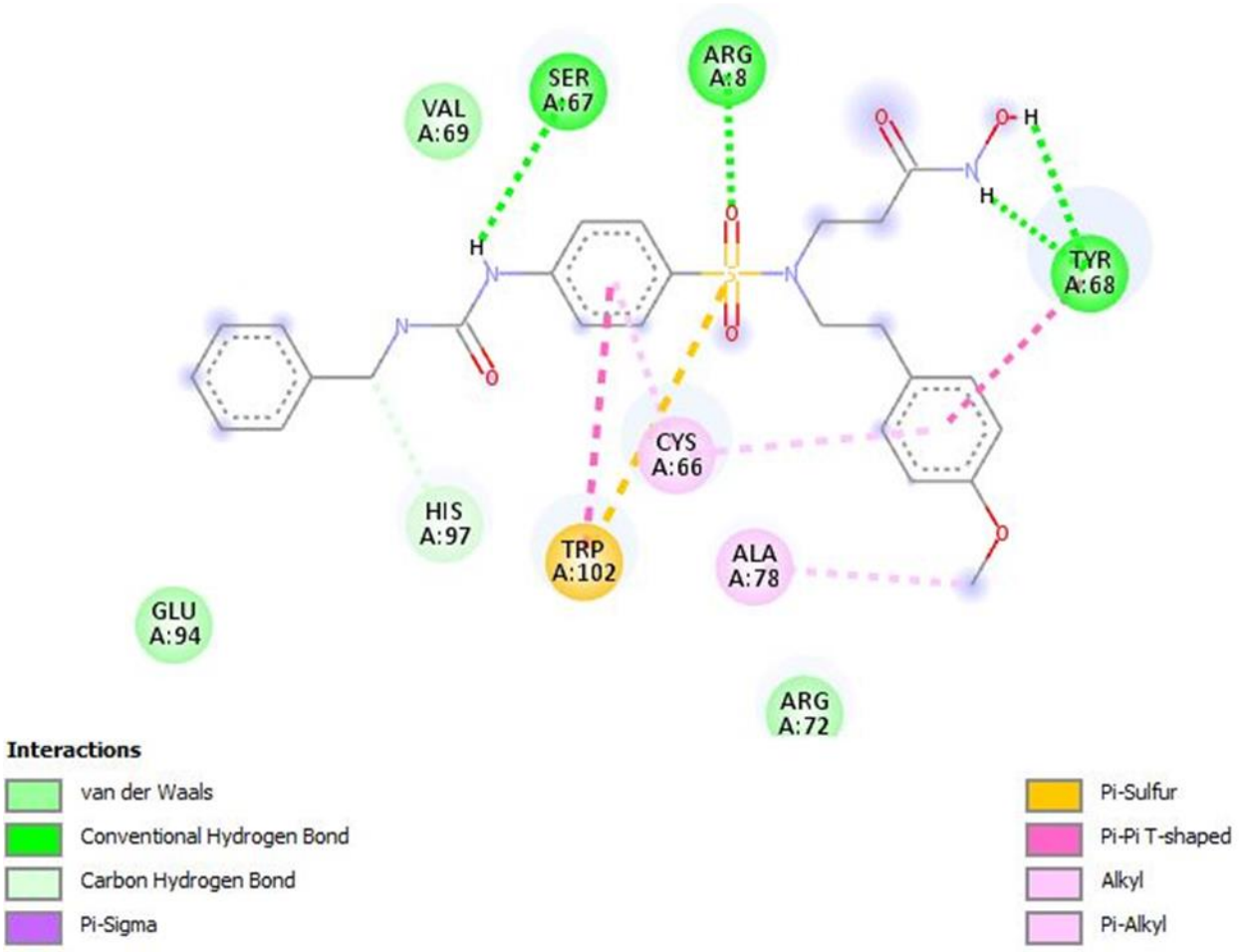

Fig 17. Interaction with Molecule 154 with BMP1

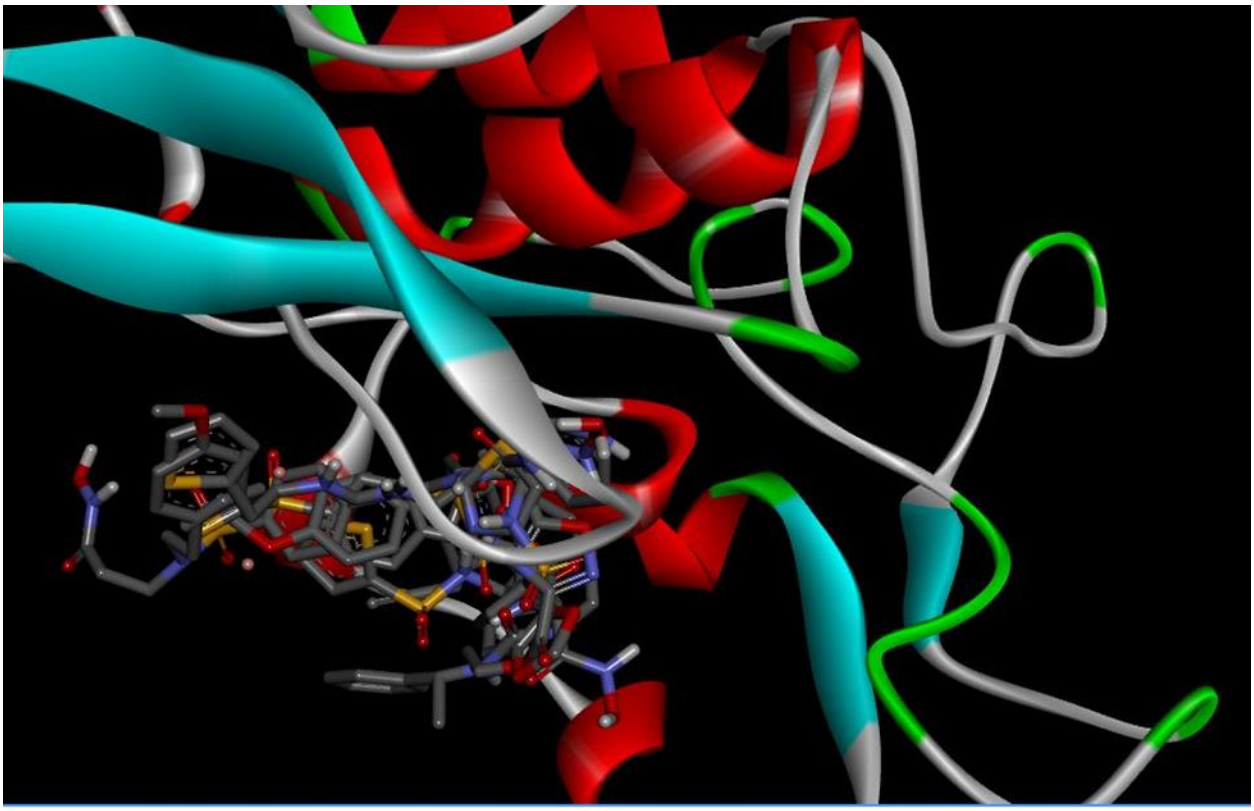

Fig 18 Superposition of best poses of the docked compounds in the active site of BMP1 (courtesy: BIOVIA Discovery Studio Visualizer 2021)

4. CONCLUSION

Since bone morphogenetic protein 1 (BMP1), a zinc metalloprotease, is necessary for the conversion of pro-collagen to collagen, blocking BMP1 may be a useful treatment for fibrosis. Cocrystal structures including BMP1 were obtained, along with the discovery of a novel family of reverse hydroxamate BMP1 inhibitors. The tiny molecule occupies the nonprime side of the metalloprotease pocket in the reported binding mode, which makes it distinctive and allows for the development of metalloprotease selectivity [16]. In the current study, a series of inhibitors are used to create a QSAR-MLR model for ordinary least squares utilizing QSARINS software. PaDEL software is used to calculate molecular descriptors. Using genetic algorithm, meaningful descriptors are chosen. This model complies with all of the OECD- declared regulatory principles. Internal validation (LOO, LMO, and Y-scrambling) and external validation both assess the model's robustness and determine its capacity to predict novel chemicals. The model application domain identifies a few potential outliers (Fig.9); however, the molecular docking investigation reveals

that these substances fit quite well inside the active site. These outliers are potent irreversible inhibitors, as evidenced by the experimental bioactivity data (IC50) for them. Significant active site residues are shown in Fig 11 (BMP1 complexed with a reverse hydroxamate inhibitor). The inhibitor is stabilized via hydrogen bonding and electrostatic interactions.

ACKNOWLEDGEMENT

The author thanks Prof. Paola Gramatica for a free license for the QSARINS software. (Website for QSARINS software: www.qsar.it). The author expresses gratitude to the University Grant Commission, Government of India for financing.

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

There is no conflict of interest.

REFERENCES

1. Amano S., Scott I.C., Takahara K. et al. Bone morphogenetic protein 1 is an extracellular processing enzyme of the laminin 5 gamma 2 chain. J. Biol. Chem. 2000; 275 (30): 22728–35.
2. 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.
3. Chirico, N., Gramatica, P. 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.
4. Chun Wei Yap. PaDEL-descriptor: An open-source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry. 2011 May;32(7):1466-74.
5. Eberhardt J., Santos-Martins D., Tillack A. F., and Forli S. Auto Dock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. Journal of Chemical Information and Modelling. Aug 23 2021; 61(8):3891-3898.

6. Friedman, J.H. Multivariate Adaptive Regression Splines. Annals of Statistics. 1991;19: 1-67.
7. Golbraikh, A. and Tropsha, A. Beware of q^2. J. Mol. Graphics Model. 2002; 20**:** 269-276.
8. 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.
9. 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.
10. 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.
11. Hsieh Y., Tung S., Pan H., Yen C., Xu H., Deng Y. Upregulation of bone morphogenetic protein 1 is associated with poor prognosis of late-stage gastric Cancer patients. BMC Cancer. 2018; 18: 508.
12. Kessler E., Takahara K., Biniaminov L., et al. Bone morphogenetic protein-1: the type I procollagen C-proteinase. Science. 1996;271 (5247): 360–2.
13. Kier, L. B., Hall L. H. Molecular connectivity in chemistry and drug research. Medicinal Chemistry.1976;14:1-257.
14. Kallander L.S., Washburn D., Hilfiker M.A., Eidam H.S, Lawhorn B.G, Prendergast J et al; *Reverse Hydroxamate Inhibitors of Bone Morphogenetic Protein* 1, ACS Med Chem Lett., 2018; 9: 736-740.
15. Lowery J.W. and Rosen V. Bone Morphogenetic Protein–Based Therapeutic Approaches. Cold Spring Harb Perspect Biol. 2018 Apr; 10(4): a022327.
16. Laure Garrigue-Antar, Catherine Barker, and Karl E. Kadler, Identification of Amino Acid Residues in Bone Morphogenetic Protein-1 Important for Procollagen C-proteinase Activity, The Journal of Biological Chemistry; July 13, 2001; 276, (28):26237-2624.
17. Mi Bai, Juan Lei, Shuqin Wang, Dan Ding, Xiaowen Yu, Yan Guo. BMP1 inhibitor UK383,367 attenuates renal fibrosis and inflammation in CKD. Am J Physiol Renal Physiol. 2019 Dec 1;317(6): F1430-F1438.

18. N'diaye E., Cook R., Wang H, Wu P., LaCanna R., Wu C. et al. Extracellular BMP1 is the major proteinase for COOH-terminal proteolysis of type I procollagen in lung fibroblasts. Am J Physiol Cell Physiol. 2021;320(2):C162–74.
19. 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.
20. Rattenholl A., Pappano W.N., Koch M., et al. Proteinases of the bone morphogenetic protein-1 family convert procollagen VII to mature anchoring fibril collagen. J. Biol. Chem. 2002; 277 (29): 26372–8.
21. 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.
22. Schùùrmann, G., Ebert, R., Chen, J., Wang, B., Kùhne, R.; External validation and prediction employing the predictive squared correlation coefficient - test set activity means vs training set activity mean. J. Chem. Inf. Model. 2008; 48 :2140–2145.
23. Tabas JA, Zasloff M, Wasmuth JJ, et al. Bone morphogenetic protein: chromosomal localization of human genes for BMP1, BMP2A, and BMP3. Genomics.1991; 9 (2): 283–9.
24. Takahara K, Lyons GE, Greenspan D.S. Bone morphogenetic protein-1 and a mammalian tolloid homologue (mTld) are encoded by alternatively spliced transcripts which are differentially expressed in some tissues. J. Biol. Chem. 1991; 269 (51): 32572–8.
25. Todeschini, R. and Consonni, V. Molecular descriptors for chemo informatics, Weinheim: Wiley VCH. 2009;27-37
26. Todeschini, R. and Consonni, V. Molecular descriptors for chemo informatics, Weinheim: Wiley VCH. 2009; 875-882.
27. 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.

Peer review under responsibility of Life Science Informatics Publications
2024 Jan – Feb RJLBPCS 10(1) Page No.1