International Journal of Drug Design and Discovery

***Volume* 4 • *Issue* 3 • *July–September* 2013. 1134-1148**

3D-QSAR Studies Combined with Virtual Screening to Identify Novel Inhibitors of N-Acetyl Glucosamine 1-Phosphate Uridyltransferase from *Mycobacterium Tuberculosis*

J.T. Patrisha, Madu Battu, D. Sriram and P. Yogeeswari*

Computer Aided Drug Design Lab,Department of Pharmacy, Birla Institute of Technology and Sciences (BITS-Pilani), Hyderabad.

ABSTRACT: The emergence of drug resistant tuberculosis poses a serious concern globally and researchers are in rigorous search for new drugs to fight against the disease. Recently, the bacterial GlmU protein (N-acetylglucosamine-1-phosphate uridyltransferase), involved in peptidoglycan, lipopolysaccharide and techoic acid synthesis, has been identified to be crucial for cell wall synthesis and survival of bacteria. The absence of this enzyme in humans makes it an important and attractive target to develop drugs for the treatment of this lethal disease. In the present study, pharmacophore mapping studies followed by 3D-QSAR analyses were undertaken for a set of 27 molecules reported to be inhibitors of GlmU. A three point pharmacophore with two hydrogen bond acceptors (A) and one hydrogen bond donor (D) as pharmacophoric features were developed. The generated model showed reasonable predictive power, with a correlation coefficient Q^2=0.5701. The external validation indicated that our 3D-QSAR model possessed high predictive powers with R_0^2 value of 0.9607 and $r_m^2$and r_p^2 values of 0.5100 and 0.5893 respectively. This model was then employed as 3D search query to virtually screen against public compound libraries (May bridge chemical libraries) in order to identify new scaffolds. We have identified six diverse non-peptidic scaffolds which were based on the pharmacophore fitness, docking score, interacting amino acids and ADME properties for drug-likeness that can be experimentally validated.

KEYWORDS: 3D-QSAR, N-acetyl glucosamine 1-phosphate uridyltransferase (GlmU), PHASE.

Introduction

Drug resistance has become a major stumbling block to overcome diseases and thus there is always a need to find new drugs and new pathways. Studies suggest that the prevalence of multi- drug resistant tuberculosis (MDR-TB) range from 6.7% for three drugs to 34% for four drugs and has caused anannual loss of around $4 - $5 billion. At present very few drugs are available in the market for treatment of MTB infection as evolution of drug-resistant strain has resulted in little efficacy and some of them have shown undesired side-effects in host[1-7]. New targets and pathways are being discovered recently for tuberculosis due to resistance or for inhibiting dormant survival state of the bacteria. GlmU is one such target which is essential for the survival of the pathogen[5]. Recent studies on the mycobacterial proteome using in-silico analysis suggested GlmU to be a potential drug target[6]. This protein is a bifunctional enzyme that catalyzes conversion of glucosamine-1-phosphate to N-acetyl glucosamine-1-phosphate at the C-terminal domain followed by conversion of N-acetyl glucosamine-1-phosphate to UDP-GluNAc at the N-terminal domain[8,9]. Though the second step is also present in humans, the first step happens only in the prokaryotes[5]. The absence of the first step in human makes it suitable for designing non-toxic inhibitors. The acetyltransferase active site and the uridyltransferase active site. Kinetic and structural studies demonstrated that the two active sites residue on two distinct protein domains[8]. The acetyltransferase reaction is carried out within the C-terminal (acetyltransferase)domain, while the rate-limiting uridyltransferase reaction occurs within the N-terminal (uridyltransferase) domain[9-11].

The identification of inhibitors using wet lab techniques is an expensive and time consuming work. Thus, there is a need to develop theoretical models for predicting inhibitors against this new target. In the present study ligand-based approaches have been used to provide insights into the structural and binding environment based on the key interactions shown by reported pubchem inhibitors (PubChem Bioassay AID-1376)[16,17]. Several other computational approaches have also been employed for the design and development of the GlmU inhibitors like three dimensional quantitative structure activity relationship (3DQSAR)[12-15], pharmacophoremap generation, docking studies and so on.

In the present work we have attempted to develop 3D-QSAR models for the prediction of new GlmU inhibitors.

* *For correspondence:* Yogeeswari

Tel: +91-40-66303515; Fax: +91-40-66303998

Email: pyogee@hyderabad.bits-pilani.ac.in

A predictive model based on 3D-QSAR approach has been developed using PHASE module of Schrödinger suite for the same set of inhibitors as reported in pubchem library. Docking studies have been done using GLIDE module of Schrödinger suit. Partial Least Square (PLS) statistical analysis and the contour maps generated were useful in the analysis of variation in activity based on structural features.

Methods

Computational details

All computations were carried out on an Intel Core 2 Duo E7400 2.80GHz capacity processor with a memory of 2GB RAM running with the RHEL 5.2 operating system. PHASE 3.3 implemented in the Maestro 9.2 software package (Schrodinger, LLC)[12-15] was used to generate pharmacophore and 3D QSAR models. Each structure was represented by a set of points in 3D-space, which coincides with the various chemical properties that may make comfortable non-covalent binding among the ligand and its binding pocket[13-15].

Generation of datasets

We chose 27GlmU inhibitors from PubChem Bioassay AID-1376[16,17] with known IC_{50}values. These inhibitors showed a wide range of activity (1.79μM-1073μM) and structural diversity. The LigPrep 2.5[18] application from Schrodinger software package was utilized to build and energetically minimize structures and to add hydrogens and generate stereoisomers at neutral pH 7 using ionizer subprogram. Canvas 1.4[20] cheminformatics package that cluster molecules based on tanimoto similarities between a set of linear fingerprint descriptors was used to determine the structural diversity among the compounds and to select representative molecules from each clusters. Clustered molecules with structural diversity were utilized for 3D QSAR development.

The IC_{50} values were converted to $pIC_{50}$to get the linear relationship in the QSAR equation, using the following formula:

$$pIC_{50} = -\log IC_{50}$$

where IC_{50} is the concentration of the antagonist producing 50% inhibition of GlmU enzyme. The dataset consisted of some highly active and inactive molecules with few molecules as moderately active.

Molecular alignment and development of pharmacophore model

To develop 3D-QSAR model, pharmacophore models and statistical analyses were performed using PHASE[13-15]. PHASE provides a built-in set of six pharmacophore features, hydrogen bond acceptor (A), hydrogen bond donor (D), hydrophobic group (H), negatively ionizable (N), positively ionizable(P) and aromatic ring (R). Conformers were generated using a Macromodel torsion angle search approach followed by minimization of each generated structure using OPLS-2005[19] as force field with implicit distance dependent dielectric solvent model. A maximum of 1000 conformers were generated per structure using Macro Model search method (ConfGen) in the preprocess minimization of 100 steps, including post process minimization of 50 steps. Each generated conformers were further filtered using a relative energy scale of 10.0 kCal/mol and a RMSD of 1Aº.

After conformers generation step there was an essential step of creating pharmacophore sites on each ligand structure responsible to facilitate non-covalent binding interaction between the ligand and the receptor. The threshold range of the active and inactive pIC_{50} was 5.1 and 3.65 respectively. Pharmacophore features were created using the clean minimized structure. Last step in pharmacophore generation was 'scoring hypothesis' in which hypotheses were ranked to make rational choices among the hypotheses and the most appropriate one for further exploration. Common pharmacophores were examined by a scoring protocol to identify the pharmacophore from each surviving n-dimensional box that yielded the best alignment of the active set ligands. The inactive molecules were scored to observe the alignment of these molecules with respect to the Pharmacophore hypothesis[14,15] to enable selection of the hypothesis. Larger the difference between the scores of active and inactives better is the hypothesis in distinguishing the actives from inactives. The 27 ligands were aligned (**Fig 1**) with the pharmacophore template of compound with high active score.

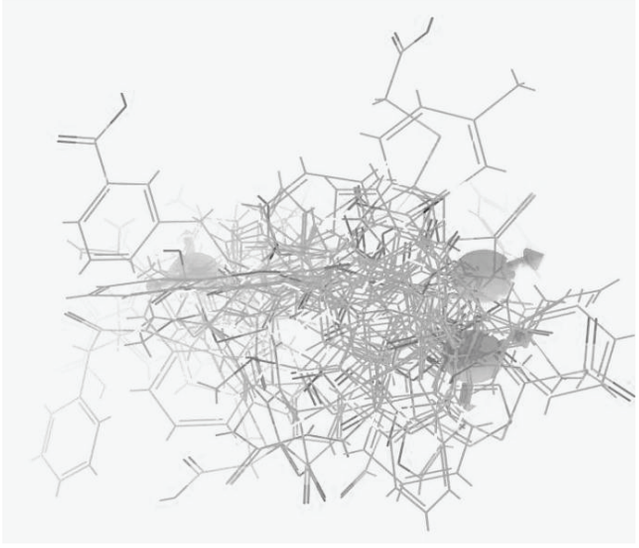


Fig. 1 All 27 compounds aligned to the best selected pharmacophore AAD.

3D-QSAR modeling

Pharmacophore dataset comprising of 27 molecules were divided randomly as 22 for training set and 5 for test set (Tables 1-4), by using the method "Automated Random Selection" option present in the PHASE module[13-15]. PHASE provided the means to build QSAR models using the activities of the ligands that match a given hypothesis. PHASE QSAR models were based on PLS regression, applied to a large set of binary valued variables which were individually derived from a regular grid of cubic volume elements, with each cubic element is represented by a set of bit values (0 or 1) to account for the different type of pharmacophore features in the training set[22]. The independent variables in the QSAR model were derived from a regular grid of cubic volume elements that span the space occupied by the training set ligands. The ligand set used were diverse structures and hence pharmacophore based QSAR models were generated for AAD hypothesis using the 22-member training set and a grid spacing of 1.00 Angstrom. QSAR-models containing one to three PLS factor were generated. A model with PLS factor onewas considered as the best statistical model. This model was validated by predicting activities of test set molecules[23-27].

Table 1 Compounds for 3D-QSAR study with their experimental and predicted activity.

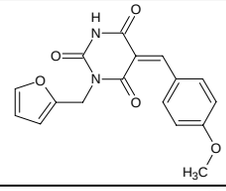
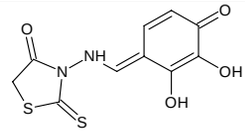
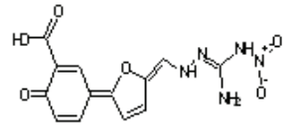
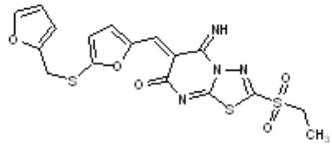
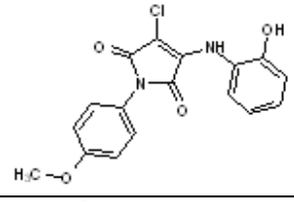
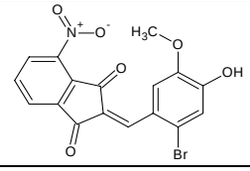
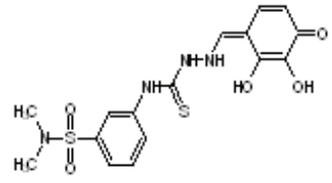
S.No.	Compounds	IC50 (μM)	Experimental pIC50	Predicted pIC50	Residual error	Data set
1.		1.79	5.747	5.210	-0.537	Training
2.		7.79	5.108	4.850	-0.258	Training
3.		28.41	4.547	4.390	-0.157	Training
4.		49.4	4.306	4.390	0.084	Training
5.		53.6	4.271	4.010	-0.261	Training
6.		60.9	4.215	4.340	0.125	Training
7.		68.4	4.165	4.310	0.145	Training

Table 1 *Contd…*

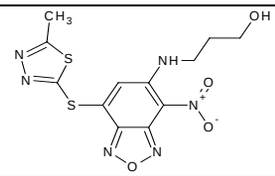
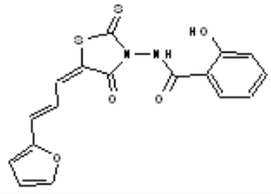
S.No	Compounds	Experimental IC50(μM)	Experimental pIC50	Predicted pIC50	Residual error	Dataset
8.		77.98	4.108	3.990	-0.118	Training
9.		80.62	4.094	4.150	0.056	Training

Table 2 Compounds for 3D-QSAR study with their experimental and predicted activity.

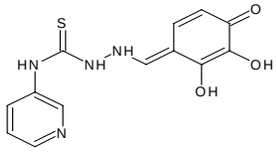
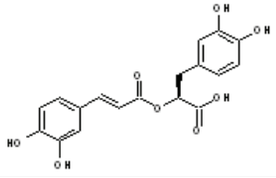
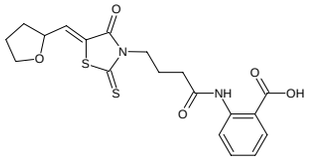
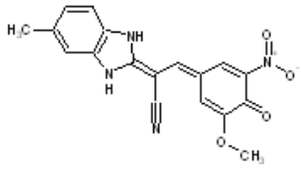
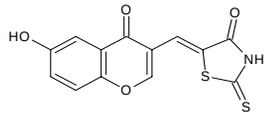
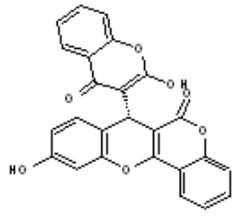
S.No	Compounds	Experimental IC50(μM)	Experimental pIC50	Predicted pIC50	Residual error	Dataset
10.		82.97	4.081	4.280	0.199	Training
11.		84.94	4.071	4.390	0.319	Training
12.		86.02	4.065	3.920	-0.145	Training
13.		94.27	4.026	3.990	-0.036	Training
14.		111.69	3.952	4.200	0.248	Training
15.		117.18	3.931	4.020	0.089	Training

Table 2 *Contd…*

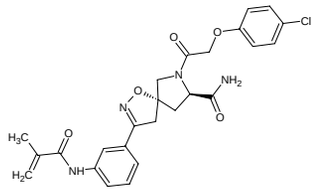
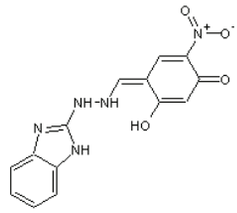
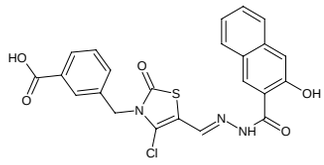
S.No	Compounds	Experimental IC50(μM)	Experimental pIC50	Predicted pIC50	Residual error	Dataset
16.		118.06	3.928	4.010	0.082	Training
17.		124.94	3.903	3.970	0.067	Training
18.		134.17	3.872	3.910	0.038	Training

Table 3 Compounds for 3D-QSAR study with their experimental and predicted activity.

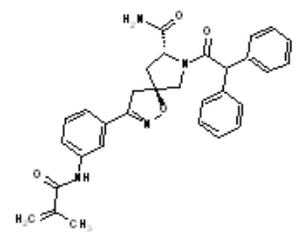
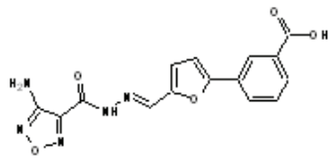
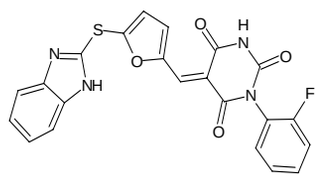
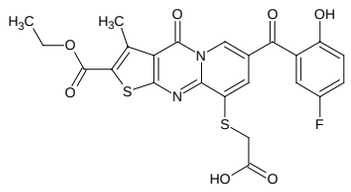
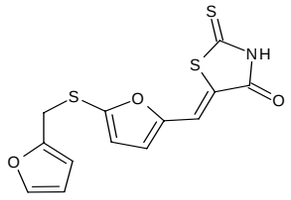
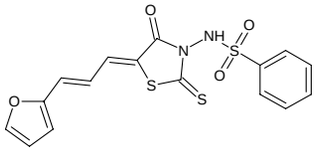
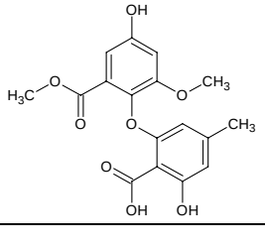
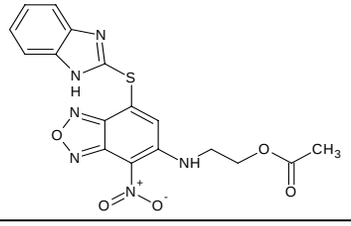
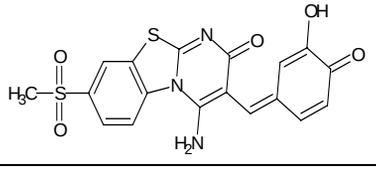
S.No.	Compounds	Experimental IC50	Experimental pIC50	Predicted pIC50	Residual error	Data set
19.		166.77	3.778	3.550	-0.228	Training
20.		188.13	3.726	3.930	0.204	Training
21.		219.12	3.659	4.060	0.401	Training
22.		1073.41	2.969	2.640	-0.329	Training

Table 4 Compounds for 3D-QSAR study with their experimental and predicted activity.

S.No.	Compounds	IC50(µM)	Experimental pIC50	Predicted pIC50	Residual error	Data set
23.		17.74	4.751	4.400	-0.351	Test
24.		49.58	4.305	4.100	-0.205	Test
25.		82.7	4.082	4.190	0.108	Test
26.		116.16	3.935	4.170	0.235	Test
27.		159.09	3.798	3.860	0.062	Test

Pharmacophore validation

To prove the specificity and selectivity of a pharmacophore hypothesis we validated the best three hypotheses with enrichment factor calculation. Ligand decoy sets were available for download (http://www.schrodinger.com/glide_decoy_set). We generated a decoy[21] database by using Generate Phase Database sub application window from PHASE application. A decoy set consists of 1027 molecules which includes 27 active molecules of GlmU inhibitors. Decoy set was used to check how well the hypothesis was able to discriminate the active Glmu inhibitor compounds from other molecules, based on parameters such as total number of compounds in the hit list (H_t), number of active percent of yields (%Y), percent ratio of actives in the hit list (%A), Enrichment factor (EF) and goodness of fit (GH) were calculated using the Ligand pharmacophore mapping protocol. Equations (1) and (2) were used to calculate the EF and GH[21]:

$$EF = \frac{(H_a \times D)}{(H_t \times A)} \quad \text{.....(1)}$$

$$GH = \left(\left(\frac{H_a}{4H_tA}\right) \times (3A + H_t)\right) \times \left(1 - \left(\frac{H_t - H_a}{D - A}\right)\right) \quad \text{.....(2)}$$

Where in 'H_t' is total number of compounds in the hit list, 'H_a' is the total number of actives molecules in the hit list, 'A' is the total number of actives in the decoy set and 'D' is the total number of molecules in the decoy set. These EF and GH based Validated pharmacophores were further validated by building a 3D QSAR model and by external statistical validation.

PLS analysis and External statistical validation of QSAR models

All 3D QSAR models were generated by using significant statistical method of partial least square analysis. The cross validation analysis was performed using the leave one out (LOO) method which evaluates the predictive ability of QSAR model. The cross validated coefficient, r^2_{cv} (also called as LOO-q^2) was calculated using the following equation:

Formula:

$$r^2_{cv} = 1 - \frac{\Sigma(Y_{obs} - Y_{pred})^2}{\Sigma(Y_{obs} - Y_{mean})^2} \quad \text{.....(3)}$$

Where Y_{pred}, Y_{obs} and Y_{mean} are the predicted,observed and mean values of the target property (pIC_{50}) respectively. $(Y_{obs}-Y_{mean})^2$ is the predictive residual sum of squares (PRESS).

The predictive correlation coefficient (r^2_{pre}), based on molecules of test set and is defined by,

$$r^2_{pre} = \frac{SD - PRESS}{SD} \quad \text{.....(4)}$$

Where SD is the sum of the squared deviations between the biological activities of the test set and mean activities of the training set molecules. PRESS is the sum of squared deviation between predicted and actual activity values for every molecule in the test set[32,33].

Based on the reported external validation methods of 3D QSAR, we evaluated the true predictive abilities of the generated models of GlmU inhibitors. To validate the true predictivity of the established models, it is crucial to perform external validation. According to literature[34-37], 3D-QSAR models were accepted if they satisfy all of these following conditions:

$r^2_{cv} > 0.5$, $r^2 > 0.6$, R^2_0 or R'^2_0 close to r^2, i.e $[[(r^2 - R^2_0)/r^2]] < 0.1$, or $r^2 - R'^2_0 r^2 < 0.1$, $0.85 \leq k \leq 1.15$, and $r^2m > 0.5$ .....(5)

The r^2 value can be calculated using the following formula:

$$R = \frac{\Sigma(Y_i - \overline{Y}_o)(\overline{Y}_i - \overline{Y}_p)}{\sqrt{\Sigma(Y_i - \overline{Y}_o)^2 \Sigma(\overline{Y}_i - \overline{Y}_p)^2}} \quad \text{.....(6)}$$

In these above equation, $\bar{y}$ and $\tilde{\bar{y}}$ are the observed and predicted activity, $\overline{Y}_0$ and $\overline{Y}_p$ are the average values of the observed and predicted pIC_{50} values of the test set molecules.

If we plot y (observed activity)versus $\bar{y}$ (predicted activity) for the ideal QSAR model, the regression line will bisect the angle formed by positive directions of the orthogonal axes $\bar{y}_i$ and y_i. The regression line is expressed by $y^r = a\bar{y} + b$,

Where

$$a = \frac{\Sigma(y_i - \bar{y})(\bar{y}_i - \bar{\tilde{y}})}{\sqrt{\Sigma(\bar{Y}_i - \bar{\tilde{y}})^2}} \quad \text{.....(7a)}$$

and

$$b = (\bar{y} - a\bar{\tilde{y}}) \quad \text{..... (7b)}$$

In the above equation (7a) and (7b), $\bar{y}$ and $\bar{\tilde{y}}$ are the average values of observed and predicted activities respectively, and the summation indicates the overall n compounds in the test set.

For the ideal QSAR model, the slope of regression (a) is equal to 1, the intercept (b) is 0.A real QSAR model may have a high predictivity, if it is close to the ideal model. The correlation coefficient R between the actual y and predicted $\tilde{y}$ activities must be close to 1, and regression of y against $\tilde{y}$ or $\tilde{y}$ against y through the origin, that is $y^{ro} = k\tilde{y}$ and $\tilde{y}^{ro} = k'y$, respectively, should be validated by at least either k or k' close to 1.

The slopes k and k'[36] were calculated by the following equations:

$$k = \frac{\in y_i \tilde{y}_i}{\in \tilde{y}^2} \quad \text{.....(8a)}$$

$$k' = \frac{\in y_i \tilde{y}_i}{\in \tilde{y}_i^2} \quad \text{.....(8b)}$$

To add value to thepredictivity of our QSAR model, regression lines which is passing through the origin defined by $y^{ro} = k\tilde{y}$ and $\tilde{y}^{ro} = k'y$ which are not close to the optimum regression lines

$y^r = a\bar{y} + b$ and by $\tilde{y}^r = a'\bar{y} + b'$. Correlation coefficients for these lines are R^2_0 and R'^2_0 have different values which can be calculated using the following formulae:

$$R^2_0 = 1 - \frac{\Sigma(-y)^2}{\Sigma(-\bar{y})^2} \quad \text{.....(9a)}$$

$$R^2_0 = 1 - \frac{\Sigma(y-)^2}{\Sigma(y-y)^2} \quad \text{.....(9b)}$$

The summations were for overall n compounds in the test set.

For better external predictive potential of the model, a parameter of modified r^2 [r^2_m] is very important factor, which can be used for the whole set considering LOO-predicted values for the training set and predicted values of the test set compounds. The r^2_m statistic equation for overall

test and training set values, generally used for selection of the best predictive models from among comparable models[38]. Substantiation of the particular QSAR models with value can be defined by the following equation:

$$r_m^2 = r^2\left(1-\sqrt{\left|r^2 - R_0^2\right|}\right) \qquad \text{…..(10)}$$

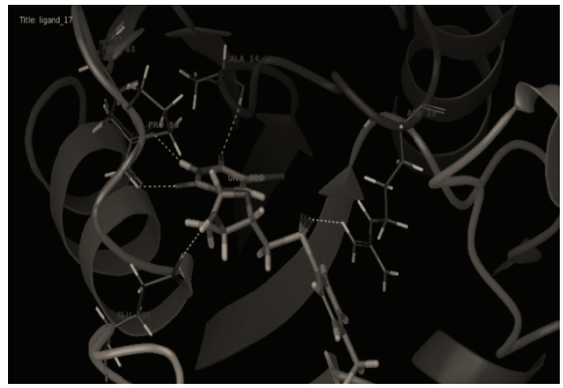
3D QSAR model with Good predictive ability and its respective pharmacophore were selected among the top three pharmacophore models and used to find for newer-leads from compound libraries like Maybridge chemical libraries by using 'Find matches' in Phase module of Schrödinger suit.

Docking studies and ADME prediction

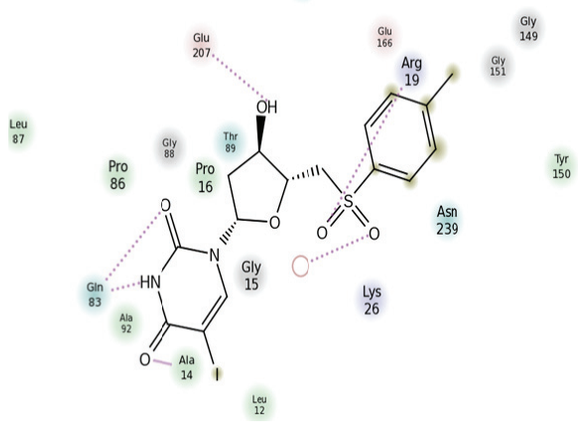
Virtual screening of the compound library was carried out by using Glide[28,29] and QikProp[41] modules of Schrodinger, LLC, 2011. We used Find matches in Phase module of Schrödinger suitfor the best pharmacophore we were selected. Matching value with the pharmacophore will be given in the form of fitness score. We selected compounds with fitness score above 2. We performed docking studies against GlmU protein of *M. tuberculosis* using GLIDE (Grid based Ligand Docking with Energetics)[28-30] module of Schrödinger suite. We used 3D8V protein from Protein Databank (PDB) for our docking studies. In 3D8V protein few atoms and few loops were missing. We filled the missing atoms and loops with the same 3D8V $GlmU_{mtb}$ using PRIME[31] module of Schrödinger suite.

We prepared our protein using protein preparation wizard by adding hydrogens, optimization and impref mini-

using OPLS2005 as force field, removed water molecules and finally refined the protein structure. We used this prepared structure for the preparation of the grid using GLIDE receptor grid generation module of Schrödinger suite. We used the substrate binding pocket to generate the grid. We got the grid center values as 28.6459, -32.5961, 37.5428. Then we docked the substrate in the grid. The docking score was -12.389. To validate our docking protocol RMSD between crystal structure and docked structure 0.3229 Å was calculated. Then we used May bridge database compounds for screening. Prepared protein (with Protein preparation wizard) and prepared ligands (with LigPrep application) were used for the different levels of 'docking based virtual screening methods'[30] such as HTVS (high throughput virtual screening), SP (standard precision), and XP (extra precision). Step wise docking and post docking energy minimization revealed the ligand interactive nature as Glide score. The interaction between the highest docking score lead and the protein is shown in **Fig 2A** and **2B**. We got 27 leads as our output. The compounds were taken further to check the ADME properties using QikProp module which usually predict physically significant descriptors and pharmaceutically relevant properties of all organic molecules.



(5A)



(5B)

Fig. 2 Virtual screening Results: Highest docking score lead's interaction.

Results and Discussion

Determination of bestpharmacophore and its validation

PHASE QSAR models may be either atom based or Pharmacophore based, the difference being whether all atoms were taken into account, or merely the pharmacophore sites that can be matched to the hypothesis. The training set molecules were sufficiently rigid and congeneric, and hence our3D QSAR approach involved the generation of a common pharmacophore hypothesis built on the principle of identification and alignment of pharmacophoric features of the chemical structures. We divided our data set into actives, inactives and moderately actives. Twenty three hypotheses were produced after all scoring functions were calculated (survival active, survival inactive, post-hoc). Out of these we selected top three pharmacophore hypotheses based on good survival activity, vector, volume, energy scores, best active alignment and number of matches (Table 5). The hypothesis 1 has 2 hydrogen bond acceptors and 1 hydrogen bond donor, hypothesis 2 has 3 hydrogen bond acceptors and hypothesis 3 has 2 hydrogen bond acceptors and 1 hydrophobic as features. These top three hypotheses were validated using decoys set[21] of 1027 com-

pounds in which 27 were actives and 1000 were compounds with unknown activity. Using the Find matches in Phase module of Schrödinger suite, with the total number of molecules in the database (D) 1027, 45 compounds were obtained as hits (Ht) for the hypothesis 1, in which 46.67% were active yields (%Y), 77.78% ratio of actives were retrieved in the hit lists (%A), and the values of EF (8.64) and GH (0.52) indicated a good sign of the high efficiency of Hypothesis 1 (Table 6). From the overall validations, we were assured that hypothesis 1 can predict most of the experimentally active kind of molecules in the same scale compared to the remaining 2 hypotheses. Hence hypothesis 1, a three point model AAD, had two hydrogen bond acceptors (A) and one hydrogen bond donor (D) with good scores of EF, GH, % of actives and other parameters was shortlisted for further studies (Table 6). 3D QSAR models are then developed for the pharmacophore hypothesis using the training set structures that match the pharmacophore on three sites. However, we utilized the three models for the 3D QSAR studies by generation of pharmaophore based 3D QSAR models and PLS analysis[37-40].

Table 5 Top three hypotheses selected based on their scores.

S.NO.	Hypothesis	Survival Score	Survival inactive Score	Vector Score	Volume Score	Site Score
1.	AAD	3.254	2.616	0.826	0.462	0.87
2.	AAA	3.031	2.211	0.815	0.442	0.63
3.	AAH	2.720	1.756	0.513	0.442	0.67

Table 6 Statistical parameters followed on best three hypothesis after screening of the decoy sets of molecules and QSAR internal validation.

Top 3 Hypotheses	AAD	AAA	AAH
[a] H_t	45	55	60
[b] H_a	21	19	13
[c] %Y	46.67	34.55	21.67
[d] % A	77.78	70.37	48.15
[e] EF	8.64◊	6.39	4.01
[f] Fn	6	8	14
[g] Fp	24	36	47
[h] GH	0.52◊	0.40	0.26

[a] Total number of hit molecules from the database

[b] Total number of active molecules in hit list

[c] Yield of actives =[(H_a/H_t) × 100]

Ratio of actives =[(H_a/A) × 100]

[e] Enrichment factor using formula

[f] False negatives =[A - H_a]

[g] False Positives = [H_t - H_a]

[h] Goodness of fit score using formla

◊ best EF and GH scores among three

3D-QSAR models generation and PLS analysis.

To develop superlative 3D QSAR models which are meant to exhibit reliable predictions it necessitates internal and external statistical validation. Models which are capable of fulfilling statistical validation parameter boundaries can display more reliable predictions[42]. Randomly chose 22 compounds in the training set and 5 compounds in the test to develop the 3D QSAR. Important parameters obtained based on LOO method, (Table 7) favored the internal statistical validation by PLS analysis. Among the best three models, hypothesis 1 showed good external predictive ability for each combination as compared to others. Hypothesis 1 showed a good R^2 value for the training set of 0.8101, good predictive power with Q^2 of 0.5701 for the test sets, with SD of 0.2369, and F value of 85.3. Further the integrity of the model was predicted by r^2_{pred} for test set with the value of 0.5893 (Table 7). The accepted LOO-cross validated value of training set (R^2) should be greater than 0.6, LOO cross validated value for test set (Q^2) should show a value greater than 0.55 to attain good predictive capacity, and standard deviation (SD) below 0.3, with minimum root mean square error (RMSE), and high value of variance ratio (F) to provide conventional QSAR validation limits. And the predictive correlation coefficient (r^2_{pred}) value generated based on molecules of test set demonstrated real predictive capacity and robustness of the QSAR model[35-40].

Table 7 PHASE 3d-QSAR result summary: PLS statistics results.

Statistical parameters	AAD	AAA	AAH
Number of molecules in Training set	22	22	22
Number of molecules in Test set	5	5	5
R^2	0.8101	0.7384	0.7543
Q^2	0.5701	- 1.143	-0.0095
SD	0.2369	0.2383	0.205
F-value	85.3	79	150.4
Pearson-R	0.8155	-0.7726	0.2232
RMSE	0.2189	0.5857	0.2941
r^2_{pred}	0.5893	0.5097	0.1156

PLS statistic parameters:

SD - Standard deviation of the regression.

R^2- for the regression.

F -variance ratio.

r^2_{pred} - predictive correlation coefficient value.

RMSE-root mean square error.

Q-squared (Q^2)value of Q^2 for the predicted activities.

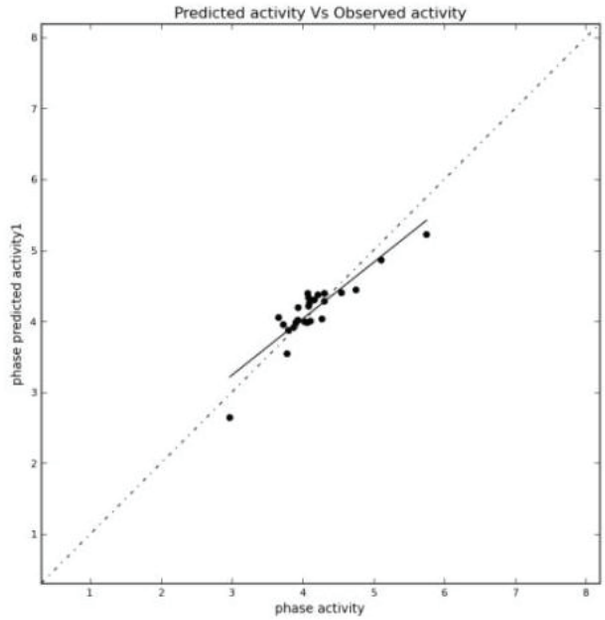
Pearson R -correlation between the predicted and observed activity for the test set.

External Statistical Validation

For a QSAR model, internal validation of LOO cross validated Q^2 is commonly used to assess predictive ability, where a high value of Q^2 is necessary and important but Q^2 alone is not sufficient condition for a model to have a high predictive power. A reliable model should also be characterized by a high correlation coefficient R (or r^2) between the predicted and observed activities of compounds from an external test set[22]. In the present study the best predictive ability of the model was characterized by correlation coefficient R = 0.8219 (r^2 = 0.6755). High slope of regression lines through the origin k value of 1.008and k′ value of 0.9892 (either k or k′ should be close to 1)[36] gave the substantial values of R_0^2 value 0.9607 and the R'^2_0 value 0.9816, which were obtained by calculating correlation coefficient of regression lines of the scatter plot obtained by means of actual activity versus predicted activity and predicted activity versus actual activity plots respectively (**Figure 3**). The calculated relation between r^2, R_0^2 and R'^2_0 gave ($r^2 - R_0^2 / r^2$) values of -0.4222 and second relation ($r^2 - R'^2_0 / r^2$) of value -0.4532 showed optimum values within the statistical limits (Table 7). Yet, our established QSAR model from Hypothesis 1 (finalized after PLS analysis), gave r_{cv}^2 value of 0.5893. A parameter of modified r^2 [r_m^2][38] was considered as a better external predictive potential for the whole set of compounds which was of 0.5100 (>0.5) defined through scatter plot best fit line values (Figure 1). This appeared to be truly predictive by fulfilling the requirements of every parameter in the external validation (Table 3). Truly, we considered this model as statistically significant model[36-40]. Besides, we resumed further steps to predict the activities of new leads from the compound libraries by using Hypothesis 1. Plots of predicted vs. actual pIC_{50} for training and test set areshown in Figure 3.

Virtual screening and docking

Virtual screening studies of the commercial database are fruitful resource for initial lead identification[30]. The idea based on the 3D QSAR and its corresponding pharmacophore were utilized as support to find important features for the inhibition of GlmU, to help in designing of the lead molecules. The best validated pharmacophore model (hypothesis 1) was used to screen against the databases of 5000 compounds (Maybridge) as presented in the flowchart (Figure 4). Fitness more than 2.0 was taken as limit for the HTVS (High throughput virtual screening) in which 503 ligands were docked to substrate active site of the GlmU PDB ID-3D8V[11], and finally we got 224 ligand molecules as hits from HTVS(High throughput virtual screening). These hits were further docked using Glide SP (standard precision) docking module and 107 ligand molecules were selected based on the docking score and visual occupancy of ligand into the pocket. Finally, we subjected the Glide SP 50 filtered ligands to Glide XP (extra precision) docking simulation. Virtual screening workflow is given in **Fig 4**. Top 19 ligand molecules more than a docking score of -2.00 kcal/mole were visually inspected for the pose and important binding residues. Based on these top six ligands (**Fig 6**) with diverged structural scaffolds, matching all the three pharmacophoric features of hypothesis 1 were selected. Predicted activities, Glide scores, fitness, H-bond data are presented in the Table 9.



Best fit line Y=0.79X+0.86 (R^2=0.8101)

Fig. 3 Scatter plot plotted between Observed vs. predicted activity of GlmU inhibition by the best model obtained using compounds 22 as the training set and validated using compounds 5 as the test set.

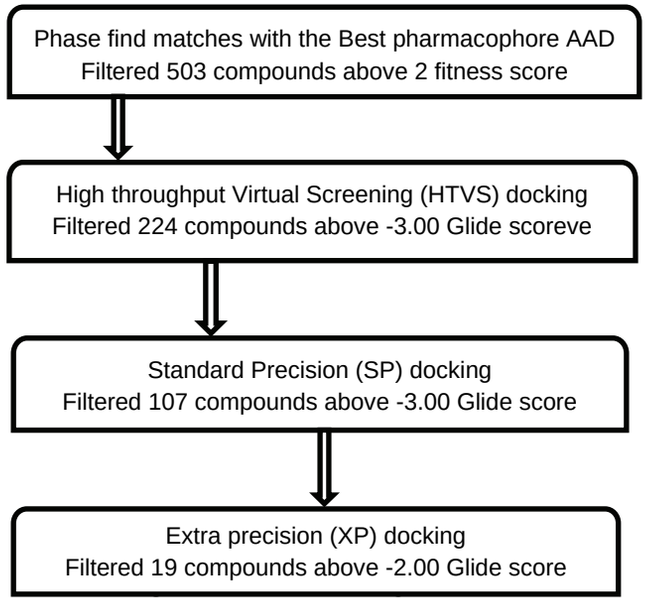


Fig. 4 Virtual screening workflow.

Table 8 External statistical validation results of quantitative structure activity relationship (QSAR) result for the Hypothesis 1 (AAD. 1) common pharmacophore hypothesis.

External validation	Parameter calculated	Limitations
r^2_{cv}	0.5893	$r^2_{cv} > 0.5$
R	0.8219	Must close to 1
r^2	0.6755	$r^2 > 0.6$
k value	1.0083	$0.85 \leq k \leq 1.15$
k' value	0.9892	$0.85 \leq k' \leq 1.15$
R^2_0	0.9607	R^2_0 or R'^2_0 close to r^2
R'^2_0	0.9816	R^2_0 or R'^2_0 close to r^2
$[(r^2 - R^2_0) / r^2]$	-0.4222	$[(r^2 - R^2_0) / r^2] < 0.1$
$[(r^2 - R'^2_0) / r^2]$	-0.4532	$[(r^2 - R'^2_0) / r^2] < 0.1$
r^2_m	0.5100	$r^2_m > 0.5$

r^2_{cv} - cross validated coefficient

R (or r^2) - correlation coefficient between the actual and predicted activities

k and k' - slope values of regression lines

R^2_0 and R'^2_0 - correlation coefficients for the regression lines through the origin $[(r^2 - R^2_0) / r^2]$ and

$[(r^2 - R^2_0 / r^2]$ - to calculate relation between r^2, R^2_0 and R'^2_0

r^2_m - modified squared correlation coefficient.

Table 9 Lead compounds with their docking score.

Name	Docking Score	Ligand Interaction	Fitness	H-bond	Predicted Activity
Lead 1	-7.017	Gln83, Ala14, Arg19, Glu207	2.25	6	4.388
Lead 2	-4.862	Gln83	2.291	1	4.409
Lead 3	-3.216	Asn181, Gly88	2.229	2	4.392
Lead 4	-2.771	Glu207	2.196	1	4.358
Lead 5	-2.084	Ala14, Lys26	2.231	2	4.344
Lead 6	-3.599	Gly113, Leu12, Lys26	2.286	3	4.318

ADME predictions

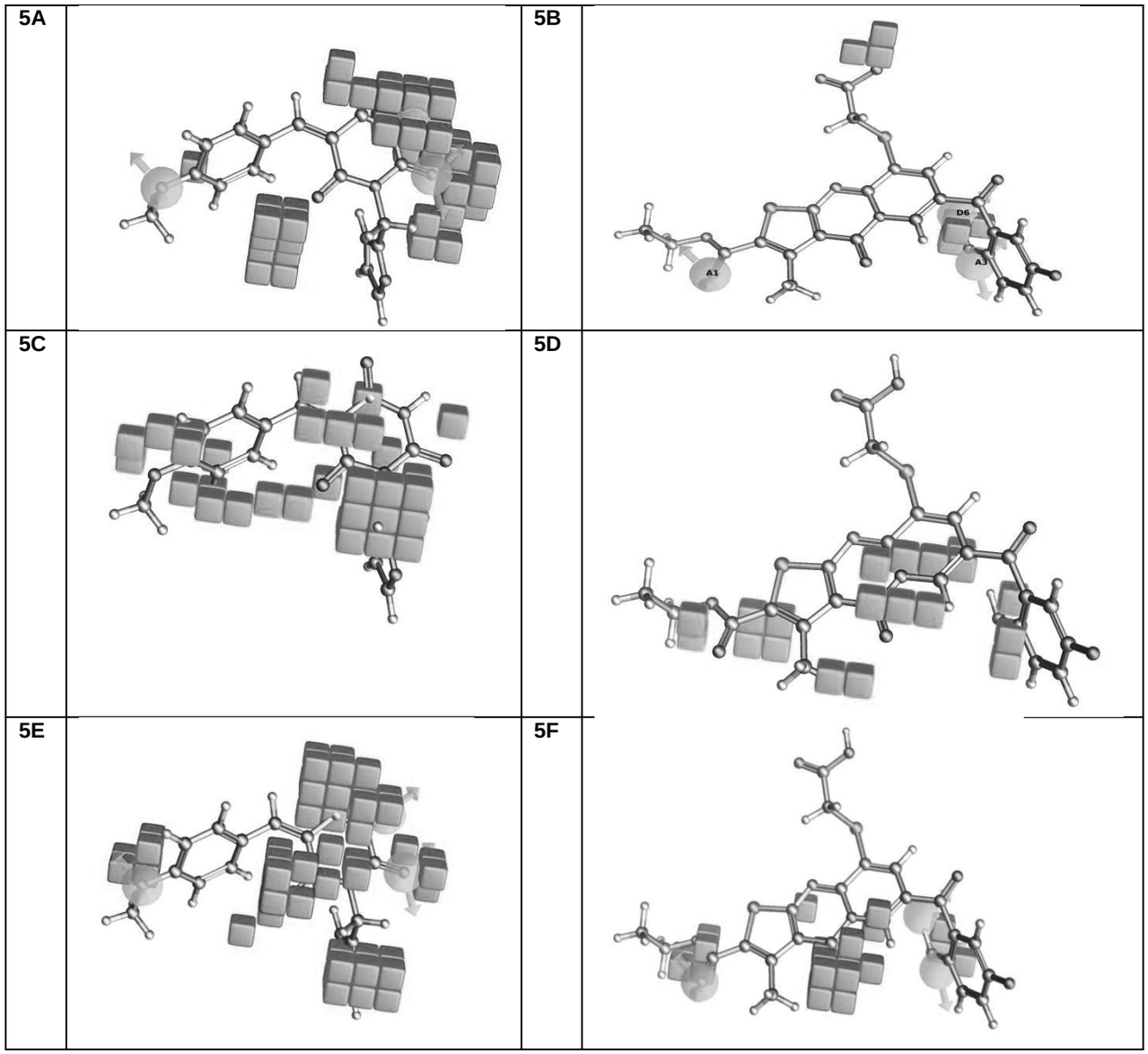
We finally evaluated the 6 lead compounds for the pharmaceutically relevant properties to check drug likeness and predictions for drug's pharmacokinetics in the human body including its ADME. QikProp module was used for evaluation of drug-like behavior through analysis of pharmacokinetic parameters required for absorption, distribution, metabolism and excretion (ADME). All the six lead compounds showed good partition coefficient (QPlogPo/w) values which were critical for understanding of absorption and distribution of drugs, to range from 0.026 to 1.554. Factor QPPCaco indicating permeability of the six lead compounds ranged from 39.236 to 239.548, where QPPCaco was a predicted apparent Caco-2 cell permeability in nm/sec value a key factor for estimation of cell permeability in biological membranes and its metabolism. All the six lead compounds passed the entire pharmacokinetic requirement for a drug-like compound and were within the acceptable range defined for human use. Additional parameters such as molecular weight, H-bond donors, H-bond acceptors, and human oral absorption according to Lipinski's rule of 5 etc. were also evaluated for their drug-like behavior and are represented in Table 10. Thus, compounds with binding interaction and good predicted pharmacokinetic properties were finalized.

Counter maps

The final validated hypothesis 1 obtained from 3D QSAR was used to generate counter maps. These counter maps were important to identify the positions of the substitutions or replacements of atoms to enhance bioactivity. Inhibitory activity can be gained by visualizing and understanding the maps against most active (**1**) and least active (**22**) compounds. This could help in discovering novel scaffolds with good biological activity. The most and least active ligand counter maps were generated and are shown in **Fig 5**. Counter maps indicated H-bond donor effect on the most active ligand[(1)] and least active ligand[(22)] **(Fig 5A** and **5B),** the hydrophobic effect of the ligands (**Fig 5C** and **5D**) and the electron withdrawing nature (**Fig 5E** and **5F**) of both ligands represented in the figure are discussed further.

The hydrogen bond donor nature for the most active compound **1** and the least active compound **22** when compared showed their most favorable region blue color and unfavorable regions red color (**Fig 5A** and **5B**). Hydrogen bond donor mapping revealed that favor regions lied near the nitrogens and oxygen of pyrimidinetrione indicating their importance for activity compared to the least active compound **22.** Therefore the presence of pyrimidinetrione in the scaffold backbone is very much needed for the activity.

Fig 5C and **5D** when compared for their hydrophobic nature for the most active compound **1** and least active compound **22** revealed that favored green color region around the furyl, benzyl rings showed that the terminal hydrophobic rings were very much needed for the activity of the compound and unfavorable region yellow color on methoxy moiety revealed that increase in the carbon chain could increase the activity. In **Fig 5E**, the favored red color regions were observed near hydrogen bond acceptors along with respective acceptor hypothesis features of most active compound which indicated that these features were crucial for the activity and these groups should be un substituted when further lead modifications indicated. In the least active compound as in **Fig 5F** the unfavorable region blue color surrounded the napthyl ring moiety which indicated that a decrease in the ring size could increase the biological activity of the compound.



(a) H-bond donor effect: Most active; **(b)** Least active (Blue- favorable, Red- unfavorable);
(c) Hydrophobic effect: Most active; **(d)** Least active (Green- favorable, Yellow- unfavorable);
(e) Electron with-drawing effect: Most active; **(f)** Least active (Red-favorable, Blue –unfavorable)

Fig. 5 Contour maps of the most active (1) and least active (22) compounds.

Conclusion

New inhibitors design for GlmU as one of the potential new antimicrobial targets for combatting *Mycobacterium tuberculosis*, was the objective of our present work. The pharmacophore and 3D QSAR studies using 27 diverse GlmU inhibitors was explored. 3D QSAR model was developed using these 27 selected compounds and validated by both LOO and external validation methods. A three point features pharmacophore with two hydrogen bond acceptors and one hydrogen bond donor was developed. Hypothesis 1 was selected as the best based on its $R^2 = 0.8101$ and $Q^2 = 0.5701$. Using 3D QSAR counter map visualization, SAR studies were done that could help to yield an insight for the further design of newer leads. Virtual screening of Maybridge database using the best pharmacophore yielded 503 hits which were filtered by three consecutive docking runs (HTVS, Glide SP and Glide XP) to finally identify 6 top ranked hits shown in **Fig 6**. Furthermore, refinement based on the ADME predictions resulted in six valuable hits with good binding scores and pharmacokinetic properties. Thus the present work revealed new diverse GlmU inhibitors as valuable leads for further biological assays.

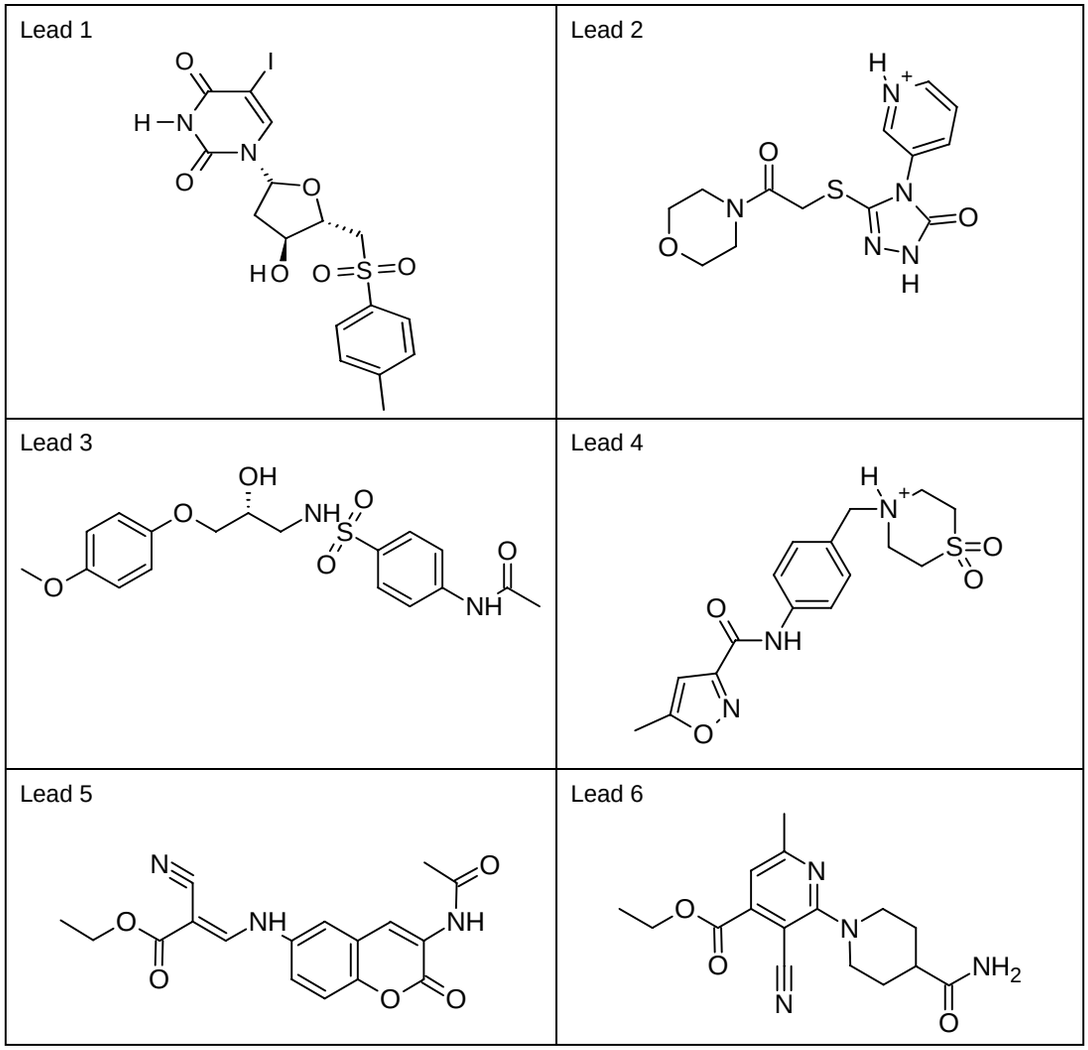


Fig. 6 Hits identified by 3D QSAR and virtual screening and docking.

Acknowledgements

The authors wish to thank the OSDD-CSIR, Government of India, NewDelhi for funding the project. One of the authors J.T.Patrisha acknowledges OSDD-CSIR, Government of India for providing Research Fellowship for this work.

References

[1] Anurag M, Dash D (2009). Unraveling the potential of intrinsically disordered proteins as drug targets:application to Mycobacterium tuberculosis. MolBiosyst 5:1752-1757.

[2] Barreteau H, Kovac A, Boniface A, Sova M, Gobec S, Blanot D (2008). Cytoplasmic steps of peptidoglycan biosynthesis.FEMS MicrobiolLett 32:168-207.

[3] Basu A, Jasu K, Jayaprakash V, Mishra N, Ojha P, Bhattacharya S (2009). Development of CoMFA and CoMSIA models of cytotoxicity data of anti-HIV-1-phenylamino-1H-imidazole derivatives. Eur J Med Chem 44:2400–2407.

[4] Canvas, version 1.4, 2011, Schrodinger, LLC, New York, NY.

[5] Consonni V, Ballabio D, Todeschini R (2009). Comments on the definition of the Q^2 parameter for QSAR validation. JChemInf Model 49: 1669-1678.

[6] Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, Dye C (2003). The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. Int Arch Med 163: 1009-1021.

[7] Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA (2006). PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. J Comput Aided Mol Des 20:647 – 671.

[8] Dixon SL, Smondyrev AM, Rao SN(2006). PHASE: a novel approach to pharmacophore modeling and 3D database searching. ChemBiol Drug Des 67: 370-372.

[9] Frieden TR, Munsiff SS(2005). The DOTS strategy for controlling the global tuberculosis epidemic. Clin Chest Med 26: 197-205.

[10] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knol EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J MedChem47:1739–1749.

[11] Gehring AM, Lees WJ, Mindiola DJ, Walsh CT, Brown ED(1996). Acetyltransfer precedes uridylyl transfer in the formation of UDP-N-acetyl glucosamine in separable active sites of the bifunctional GlmU protein of Escherichia coli. Biochemistry 35: 579-585.

[12] Glide, version 5.7, 2011, Schrodinger, LLC, New York, NY.

[13] Global tuberculosis control. World Health Organization report. 2011.

[14] Golbraikh A, Tropsha A, 2002, Beware of q2. JMol GraphModel 20: 269–276.

[15] Johnson SR(2008). The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). JChemInf Model 48: 25-26.

[16] Kawatkar S, Wang H, Czerminski R, Joseph-McCarthy D (2009). Virtual fragment screening: an exploration of various docking and scoring protocols for fragments using Glide. J ComputAided Mol Des 23: 527-539

[17] Lewis RA (2005). A general method for exploiting QSAR models in lead optimization. J Med Chem48: 1638-1648.

[18] Li Y, Wang Y, Zhang F (2010). Pharmacophore modeling and 3D-QSAR analysis of phosphoinositide 3-kinase p110 αinhibitors. JMolModel16:1449–1460.

[19] LigPrep, version 2.5, 2011, Schrodinger, LLC, New York, NY.

[20] Lu P, Wei X, Zhang R (2010) CoMFA and CoMSIA 3D-QSAR studies on quionolonecaroxylic acid derivatives inhibitors of HIV-1 integrase. Eur JMedChem 45:3413–3419.

[21] Macro Model, version 9.9, 2011, Schrödinger, LLC, New York, NY.

[22] Mochalkin I, Lightle S, Narasimhan L, Bornemeier D, Melnick M, Vanderroest S, McDowell L(2008). Structure of a small-molecule inhibitor complexed with GlmU from Haemophilusinfluenzae reveals an allosteric binding site. Protein Sci 17:577-582.

[23] Pan X, Tan N, Zeng G, Huang H, Yan H (2010). 3D QSAR studies on ketoamides of human cathepsin K inhibitors based on two different alignment methods. Eur J Med Chem 45: 667–681.

[24] Phase, version 3.3, 2011, Schrodinger, LLC, New York, NY.

[25] Prime, version 3.1, 2011, Schrodinger, LLC, New York, NY.

[26] QikProp, version 3.4, 2011, Schrodinger, LLC, New York, NY.

[27] Robert CG, Kevin VP, Barbara EL (2007). The evolution of extensively drug resistant Tuberculosis (XDR-TB): History, status and issues for global control. Infect DisordDrug Targets 7(2),73:91.

[28] Roy K, Paul S (2008). Exploring 2D and 3D QSARs of 2,4-diphenyl-1,3-oxazolines for ovicidal activity against Tetranychusurticae. QSAR CombSci 28:406-425.

[29] Roy K, Roy PP (2009). Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. Eur J Med Chem 44:2913-2922.

[30] Roy PP, Paul S, Mitra I, Roy K (2009). On two novel parameters for validation of predictive QSAR models. Molecules 14: 1660-1701.

[31] Roy PP, Roy K (2008). On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci27: 302-313.

[32] Sakkiah S, Meganathan C, Sohn YS, Namadevan S, Lee KW (2012). Identification of important chemical features of 11β-hydroxysteroid dehydrogenase type1 inhibitors: Application of ligand based virtual screening and density functional theory. Int J MolSci 13:5138-5162.

[33] Schrödinger, L.L.C., New York, www.schrodinger.com.

[34] Schüürmann G, Ebert RU, Chen J, Wang B, Kühne R (2008). External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs. training set activity mean. J ChemInf Model 48: 2140-2145.

[35] Sharma MC, Smita Sharma (2010). 3D-Quantitative Structure-Activity Relationship Analysis of Some 2-Substituted Halogenbenzimidazoles Analogues with Antimycobacterial activity. Int J ChemTech Res 2:606-614.

[36] Singla D, Anurag M, Dash D, Raghava GP(2011). A web server for predicting inhibitors aginst bacterial target GlmU protein. BMC Pharmacol 11:5.

[37] Stanton DT(2003). On the physical interpretation of QSAR models. J ChemInfComputSci 43: 1423-1433.

[38] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003). Random Forest: A classification and regression tool for compound classification and QSAR modeling. J ChemInfComputSci 43: 1947-1958.

[39] Verma SK, Jaiswal M, Kumar N, Parikh A, Nandicoori VK, Prakasha B (2009). Structure of N-acetylglucosamine-1-phosphate uridyltransferase (GlmU) from Mycobacterium tuberculosis in a cubic space group. ActaCrystallogr Sect F StructBiolCrystCommun 65: 435-439.

[40] Wang Y, Bolton E, Dracheva S, Karapetyan K, Shoemaker BA, Suzek TO, Wang J, Xiao J, Zhang J, Bryant SH (2010). An overview of the PubChemBioAssay resource. Nucleic Acids Res D255–266.

[41] Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009). PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res W623–633.

[42] Zhang Z, Bulloch EM, Bunker RD, Baker EN, Squire CJ (2009). Structure and function of GlmU from Mycobacterium tuberculosis. Actacrystallogr D BiolCrystallogr 65:275-283.