

# **Deciphering the molecular interaction and bioactivity guided chemical space analysis of small molecule inhibitors of *Mycobacterium carbonic anhydrase* (*Mtb-CA*) through the implementation of WebApp-based generative machine learning models**

- **Principle Investigator:** Dr. Ashok Aspatwar
- **ML Model Generation and Optimization/Software Development:** Ratul Bhowmik/Ajay Manaithiya
- **Chemical Data Curator:** Ajay Manaithiya, Ranajit Nath

-5.000

5.000

# PRIMARY OBJECTIVES

- Generation of a robust multi-molecular signature implemented machine learning-assisted quantitative structural activity relationship (**ML-QSAR**) model for predicting bioactivity (**K<sub>i</sub>**) against ***Mtb-CA***
- Development of a web application using the development ML-QSAR model
- Screening of large drug-like molecule libraries using the developed web application to identify new ***Mtb-CAIs***
- Molecular docking, pharmacokinetic analysis, and molecular dynamics simulation studies on the newly identified ***Mtb-CAIs*** to identify the best-hit compound

## ML MODEL FOR BIOACTIVITY PREDICTION OF Mtb-CA INHIBITORS



Curation of Mtb-CA inhibitors with Ki value

Total inhibitors: 267

Molecular Features

Prediction Parameter:  
Bioactivity Class ('pKi')

Dataset Splitting into training and test set

ML Algorithms (Random Forest Regressor)

Final number of descriptors, number of molecules in training and test

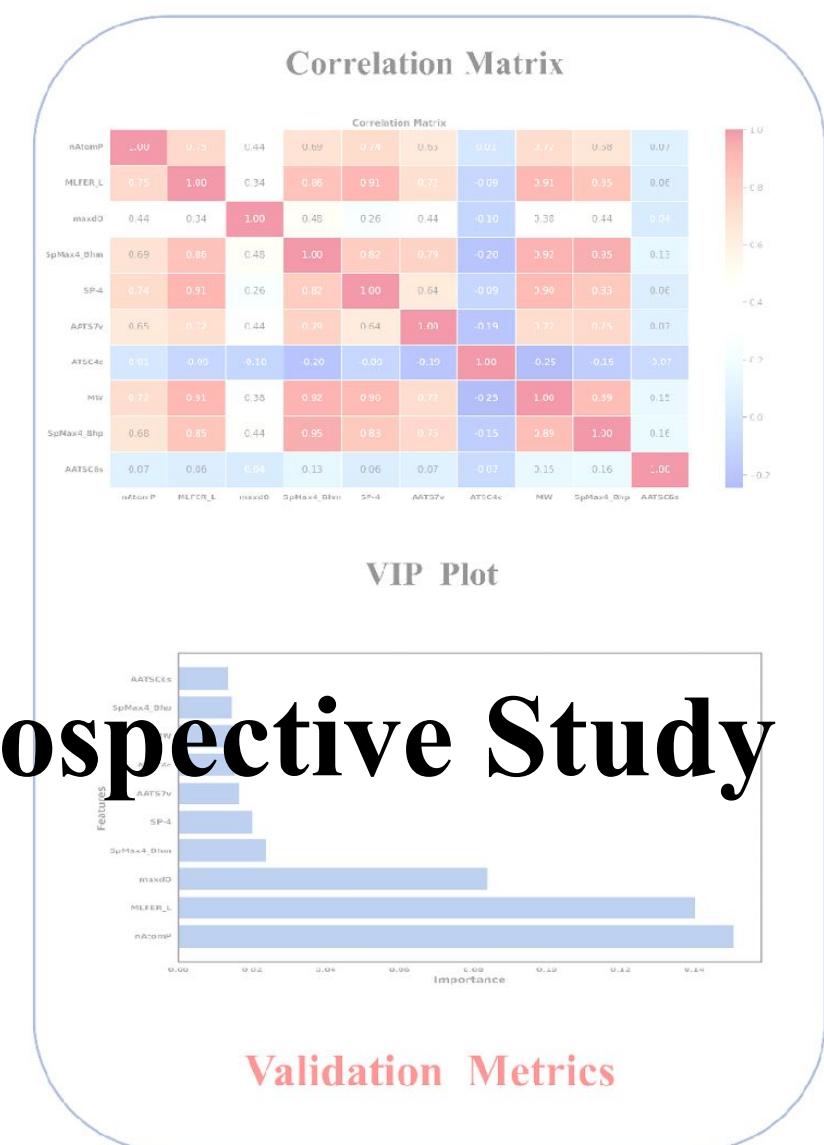
Initial Model Performance

**Model Metrics**  
Correlation Coefficient (Train, Test)  
RMSE(Train, Test)  
MAE (Train, Test)

False Positive Removal

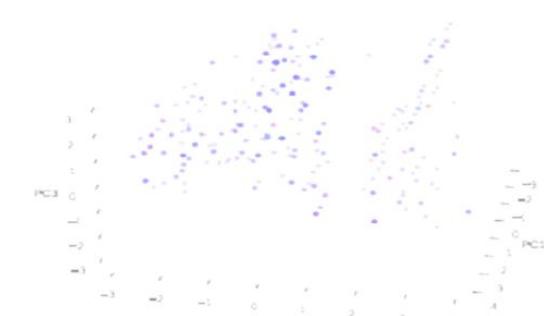
False Negative Removal

Applicability Domain Analysis

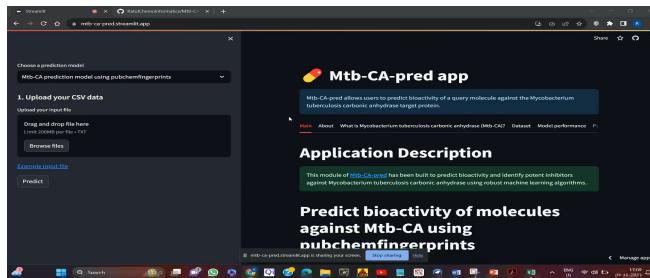
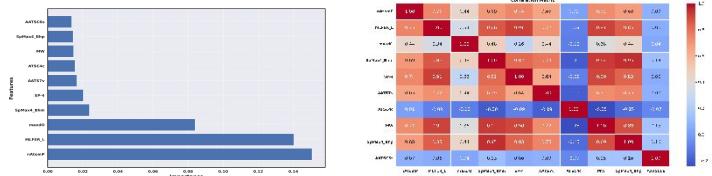
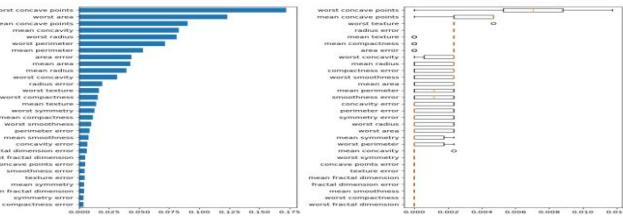
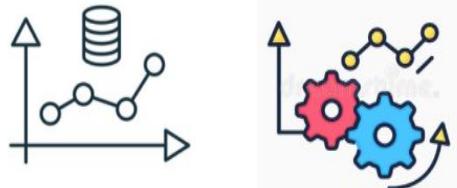


Final Model Performance

The final machine-learning model



**Model Metrics**  
Correlation Coefficient (Train, Test)  
RMSE(Train, Test)  
MAE (Train, Test)



Initial machine learning models deployment using 3 different molecular signatures with respect to Mtb-CAIs against their respective pKi values)  
(P: 881; S: 307 ; D: 1444)

**PubChem Fingerprints; Substructure Fingerprints; 1D & 2D molecular descriptors**

**Machine Learning model generation and statistical analysis  
(P:186; S:17; D:760)**

Variance Importance Plot (VIP) generation for all the molecular features of previously discovered Mtb-CAIs against their pKi value. Selection of only the significant molecular features after removal of highly-correlated and constant molecular features

Improvement in model robustness and statistical metrics using only the significant molecular features in accordance with the feature importance score of VIP plot.

**Machine Learning model generation with significant molecular signatures**

**Web-app deployment using final machine learning model using Streamlit AI-interface**

Web app development

**Machine Learning assisted web application for bioactivity prediction for Mtb-CAIs**

**P: PubChem Fingerprints**

**S: Substructure Fingerprints**

**D: 1D & 2D Descriptors**

# ML-QSAR MODEL FOR BIOACTIVITY PREDICTION OF *Mtb-CA* INHIBITORS USING PUBCHEM FINGERPRINTS



## Curation of *Mtb-CA* inhibitors with Ki value

Total inhibitors: 267

PubChem fingerprints: 881

Prediction Parameter: Bioactivity Class ('pKi')

Dataset Splitting into training and test set

Final number of descriptors(186), number of molecules in training(214) and test(53)

ML Algorithms (Random Forest Regressor)

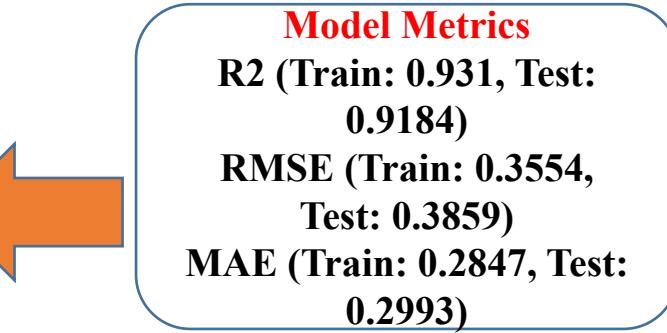
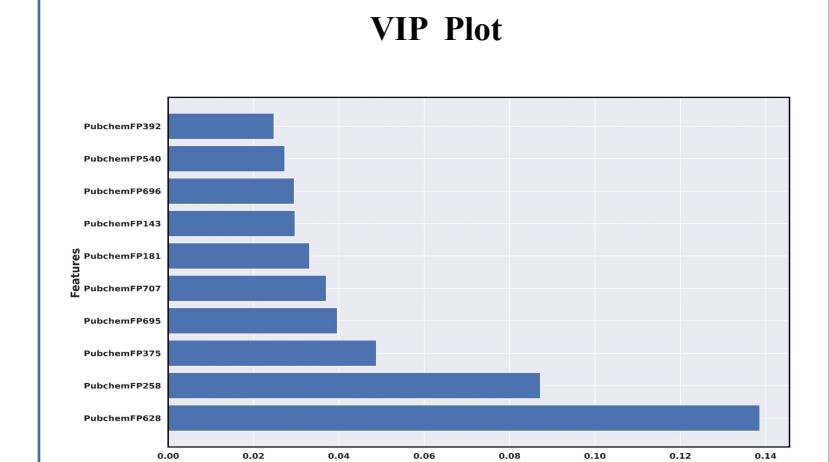
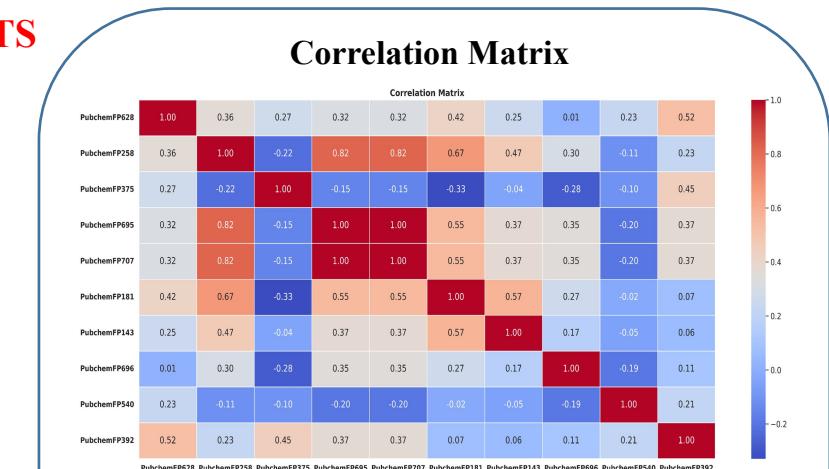
Best algorithm selection(Random Forest Classifiers)

Initial Model Performance

**Model Metrics**  
R2 (Train: 0.8603, Test: 0.8079)  
RMSE (Train: 0.5455, Test: 0.62)  
MAE (Train: 0.3951, Test: 0.4648)

FP and FN Removal

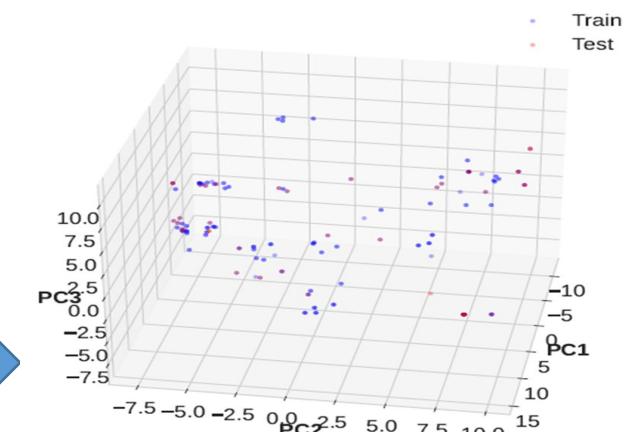
Mention molecules removed (40)

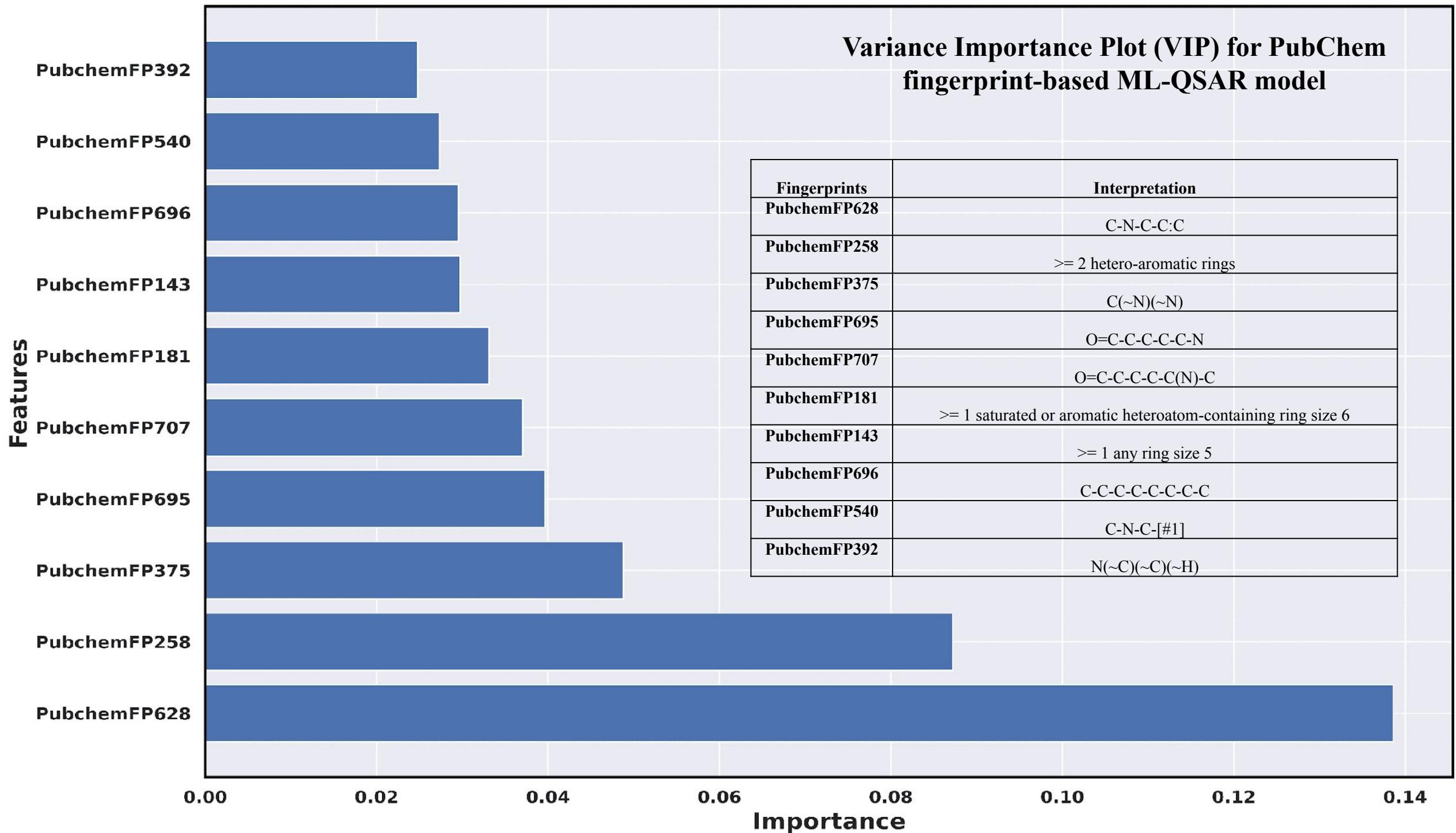


Final Model Performance

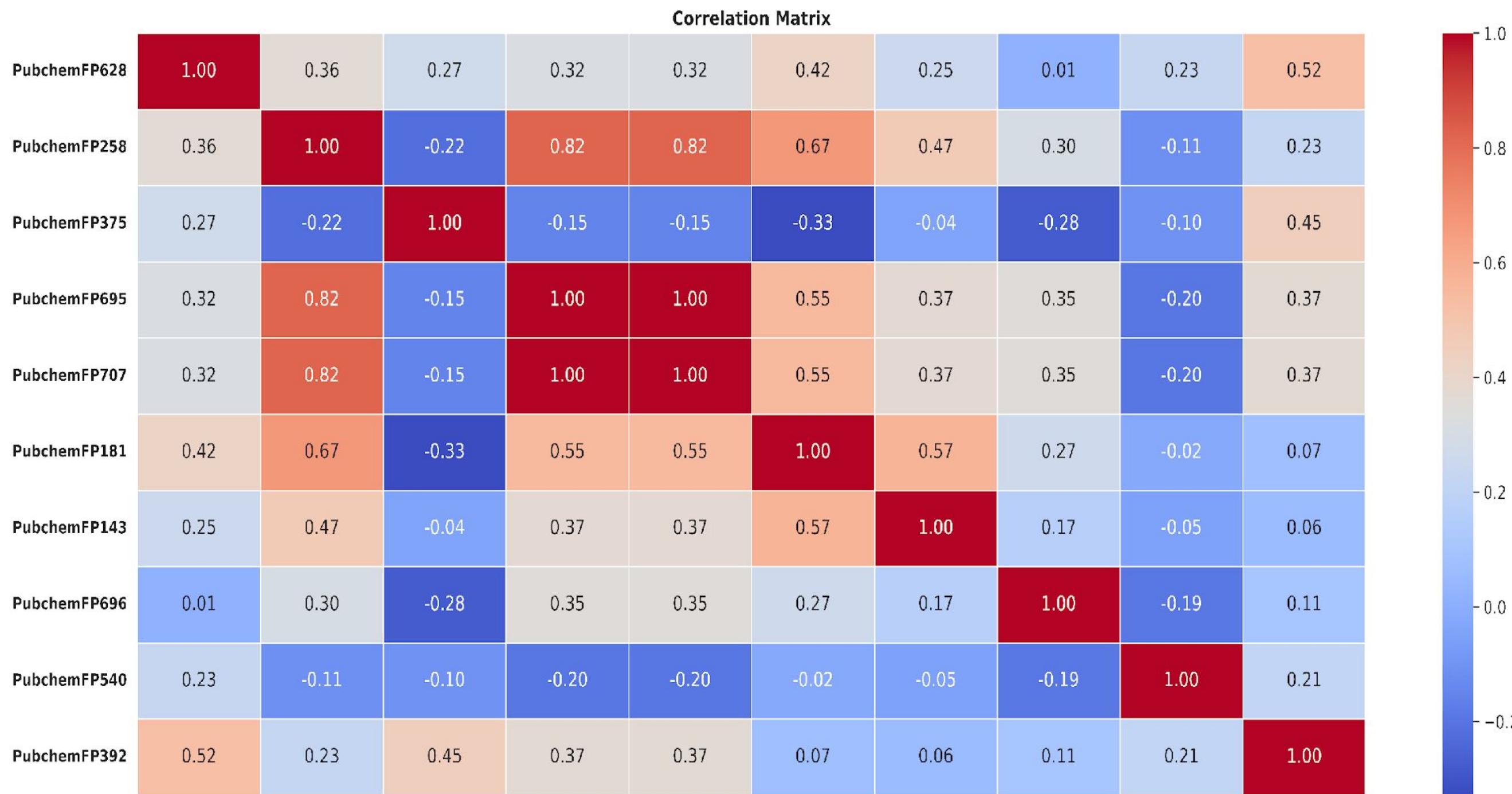
The final machine-learning model

number of molecules in training(186) and test(41)



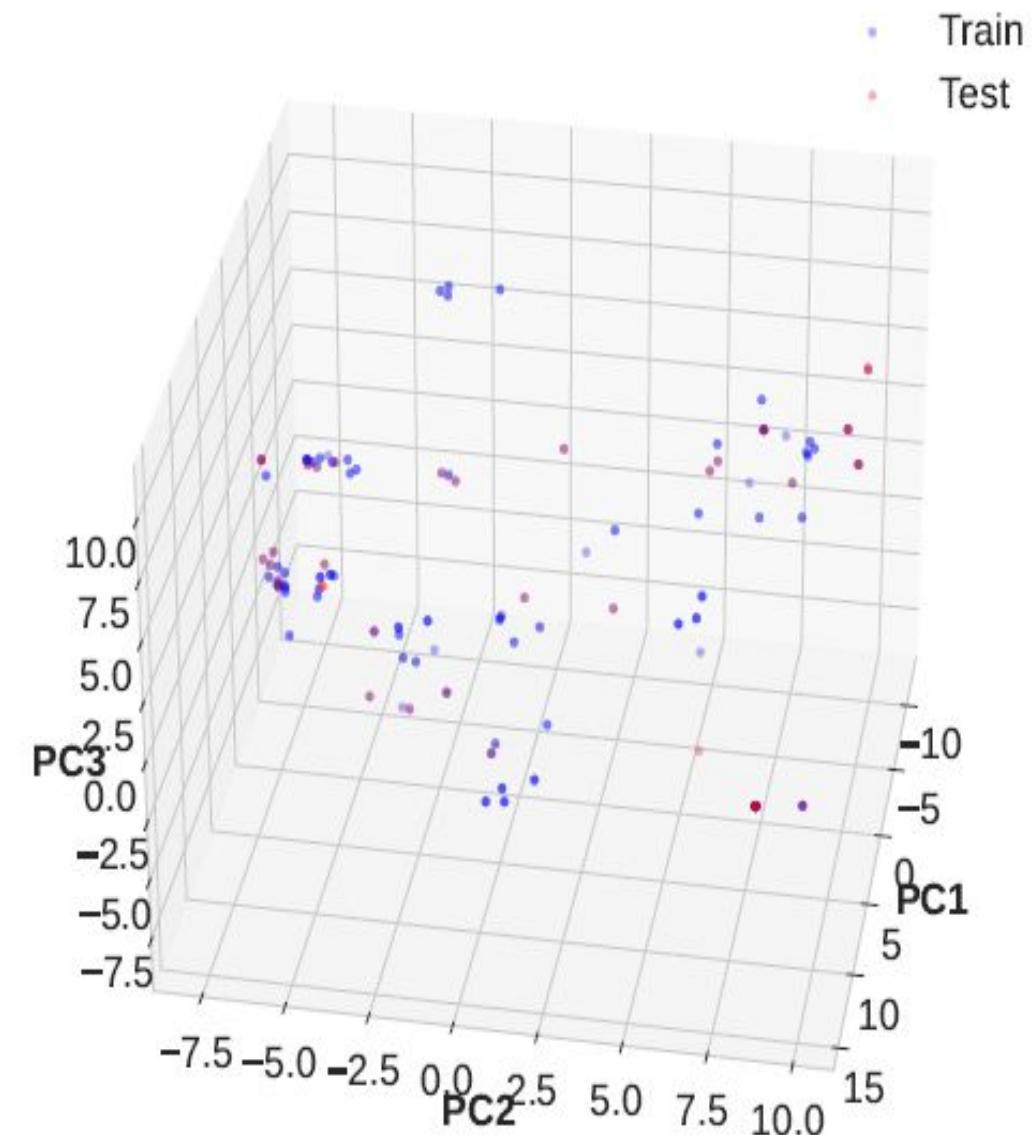
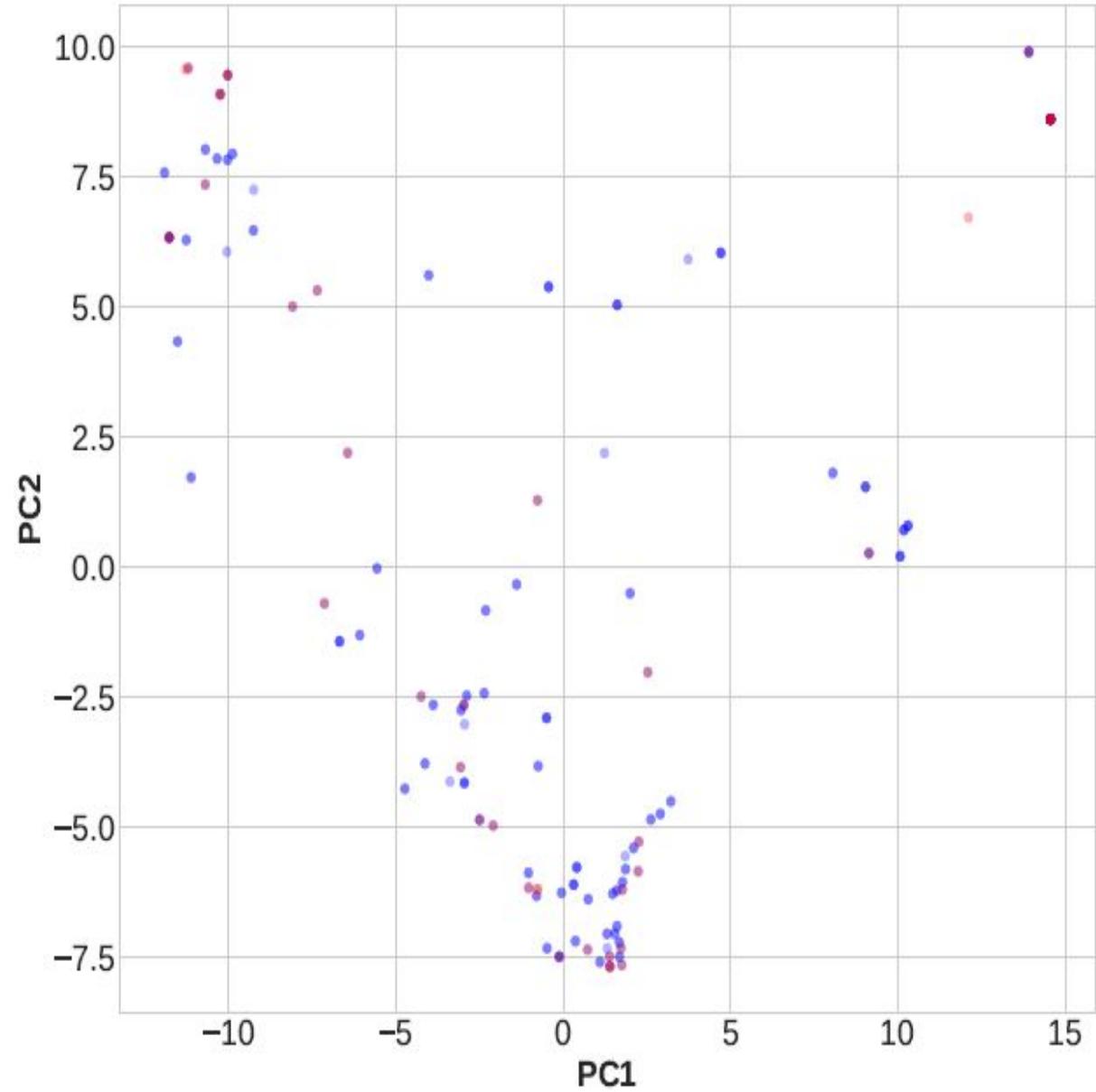


# Correlation Matrix for PubChem fingerprint-based ML-QSAR model



PubchemFP628 PubchemFP258 PubchemFP375 PubchemFP695 PubchemFP707 PubchemFP181 PubchemFP143 PubchemFP696 PubchemFP540 PubchemFP392

# Applicability Domain (AD)/Principle Component Analysis (PCA) for PubChem fingerprint-based ML-QSAR model



# ML-QSAR MODEL FOR BIOACTIVITY PREDICTION OF *Mtb-CA* INHIBITORS USING SUBSTRUCTURE FINGERPRINTS



## Curation of *Mtb-CA* inhibitors with Ki value

Total inhibitors: 267

Substructure fingerprints: 307

Prediction Parameter: Bioactivity Class ('pKi')

Dataset Splitting into training and test set

Final number of descriptors(17), number of molecules in training(214) and test(53)

ML Algorithms (Random Forest Regressor)

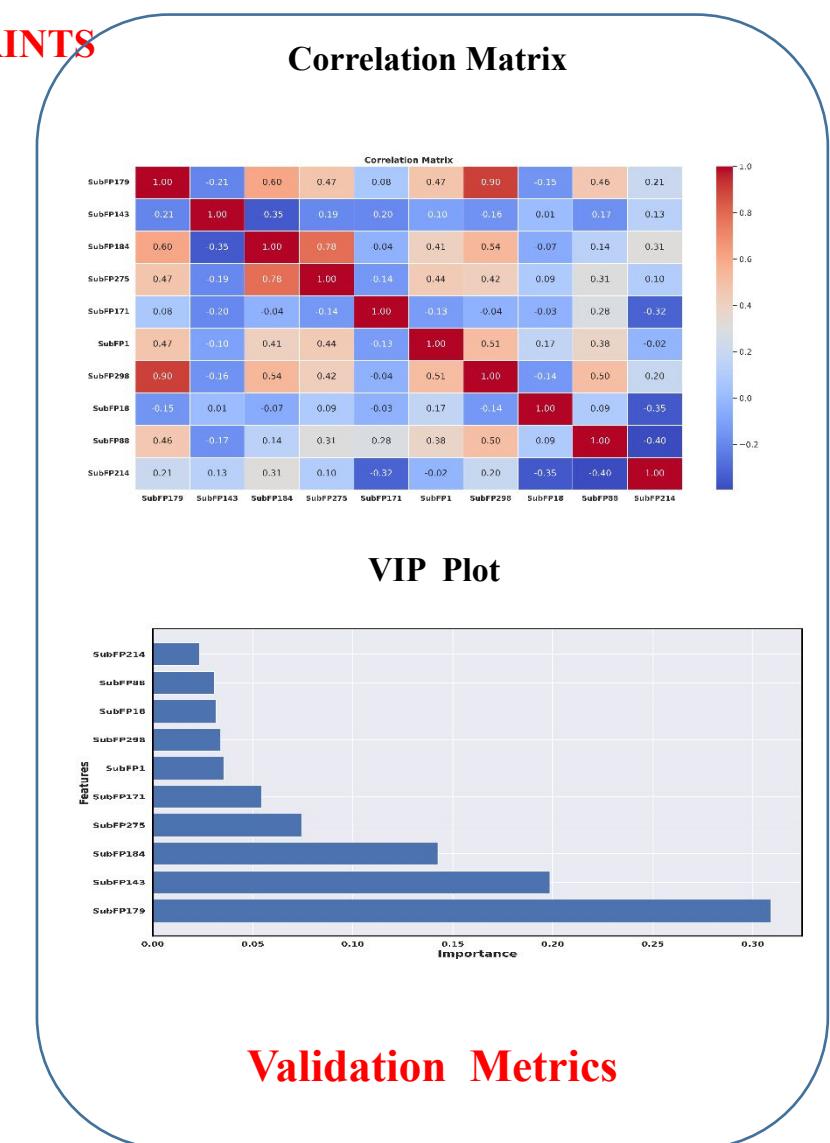
Best algorithm selection(Random Forest Classifiers)

Initial Model Performance

**Model Metrics**  
R2 (Train: 0.8277, Test: 0.6364)  
RMSE (Train: 0.5952, Test: 0.8444)  
MAE (Train: 0.4573, Test: 0.6187)

FP and FN Removal

Mention molecules removed (53)

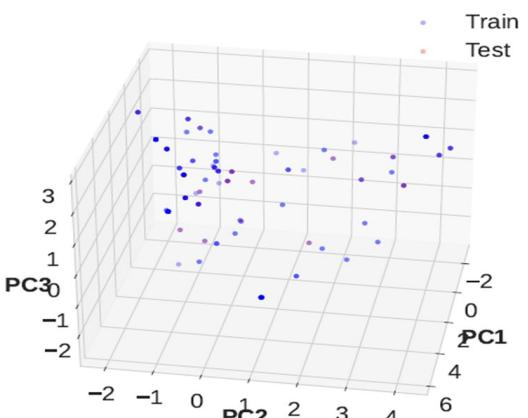


**Model Metrics**  
R2 (Train: 0.9227, Test: 0.9126)  
RMSE (Train: 0.373, Test: 0.347)  
MAE (Train: 0.3082, Test: 0.2681)

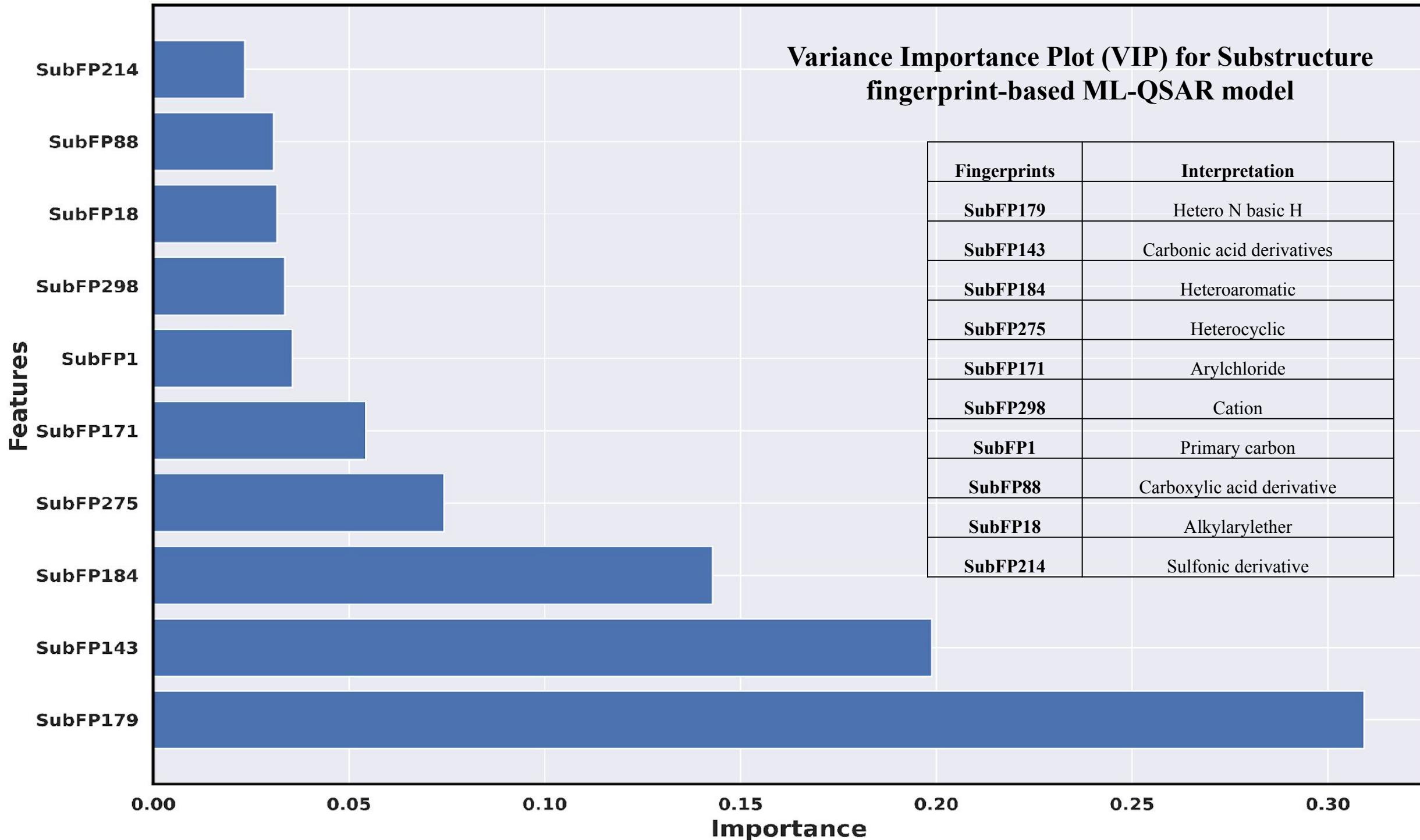
Final Model Performance

The final machine-learning model

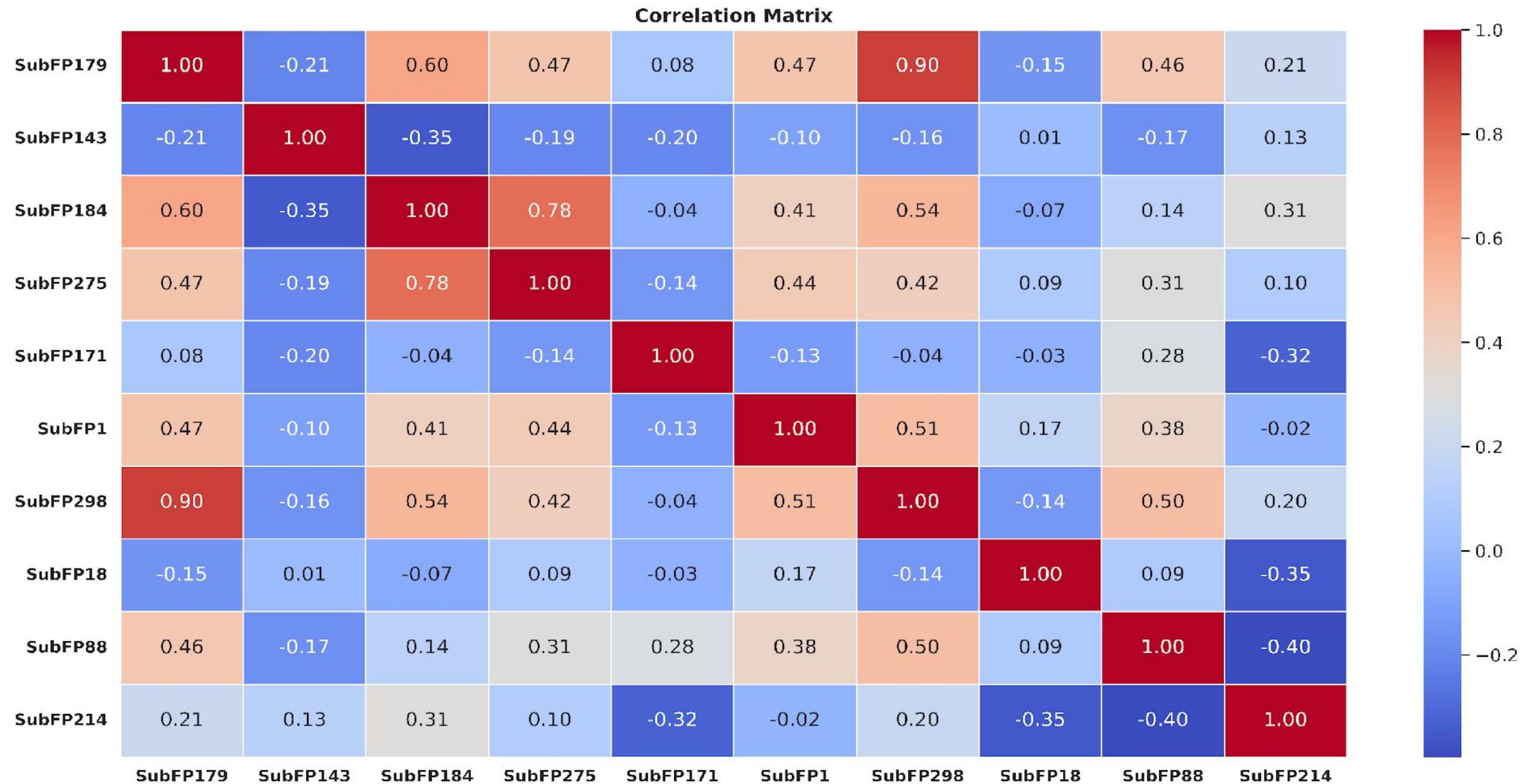
number of molecules in training(178) and test(36)



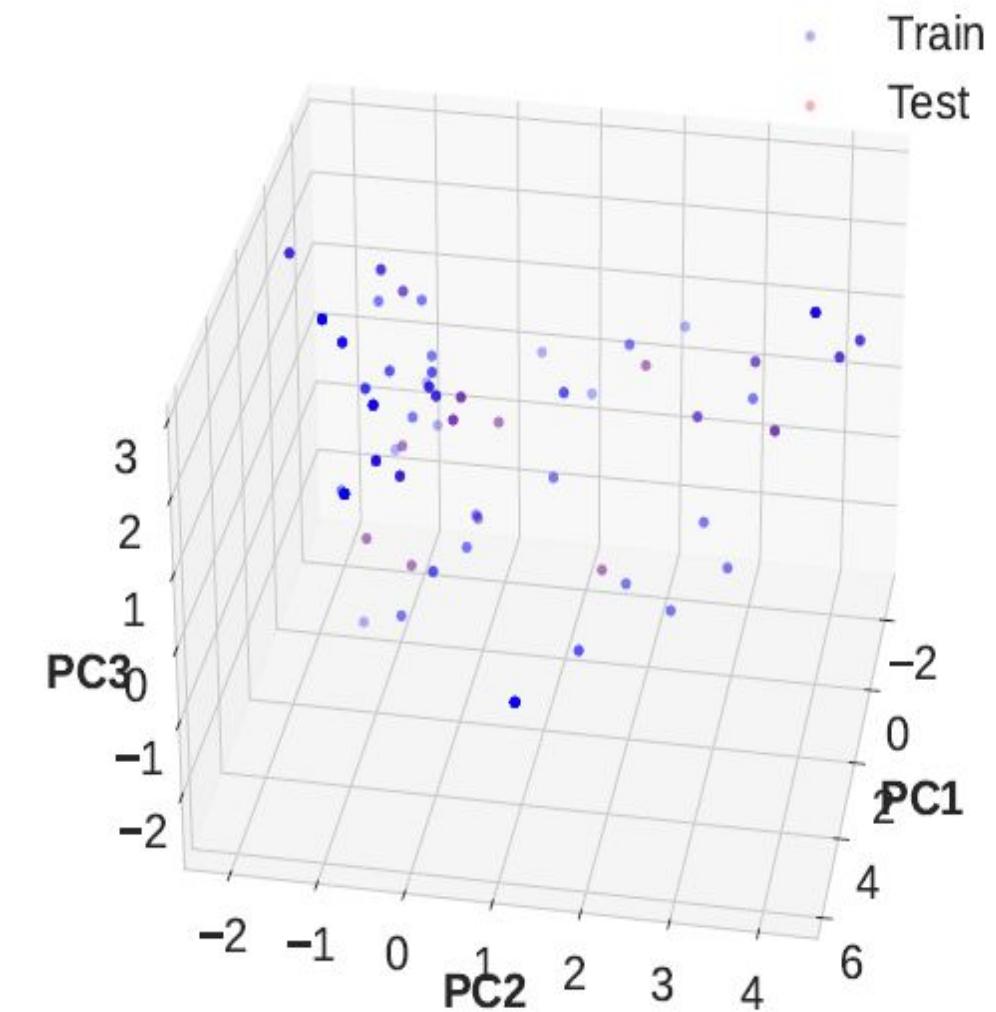
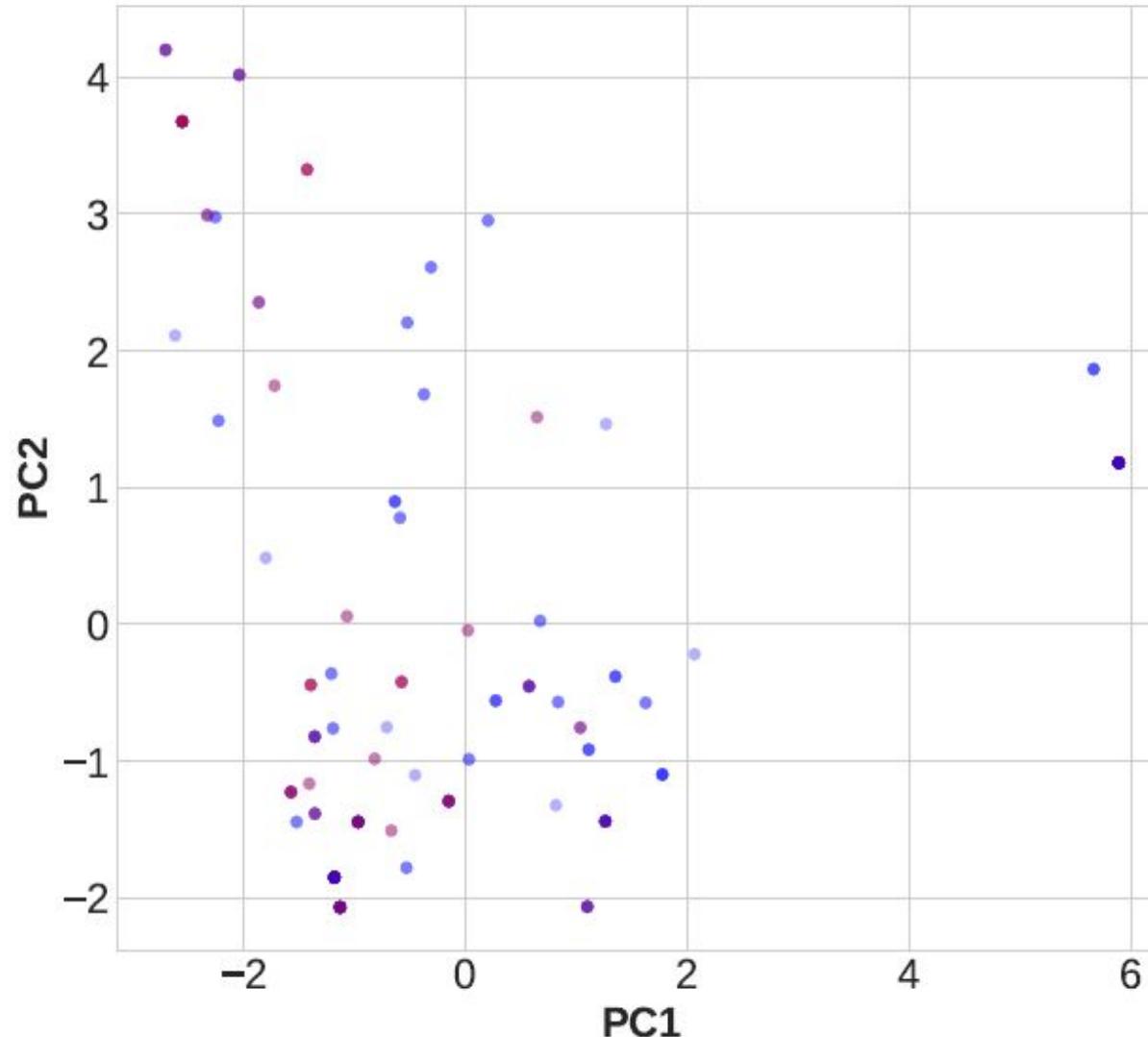
Applicability Domain Analysis



# Correlation Matrix for Substructure fingerprint-based ML-QSAR model



# Applicability Domain (AD)/Principle Component Analysis (PCA) for Substructure fingerprint-based ML-QSAR model



# ML-QSAR MODEL FOR BIOACTIVITY PREDICTION OF *Mtb-CA* INHIBITORS USING 1D & 2D MOLECULAR DESCRIPTORS



## Curation of *Mtb-CA* inhibitors with Ki value

Total inhibitors: 267

Molecular Descriptors: 1444

Prediction Parameter: Bioactivity Class ('pKi')

Dataset Splitting into training and test set

Final number of descriptors(760), number of molecules in training(214) and test(53)

ML Algorithms (Random Forest Regressor)

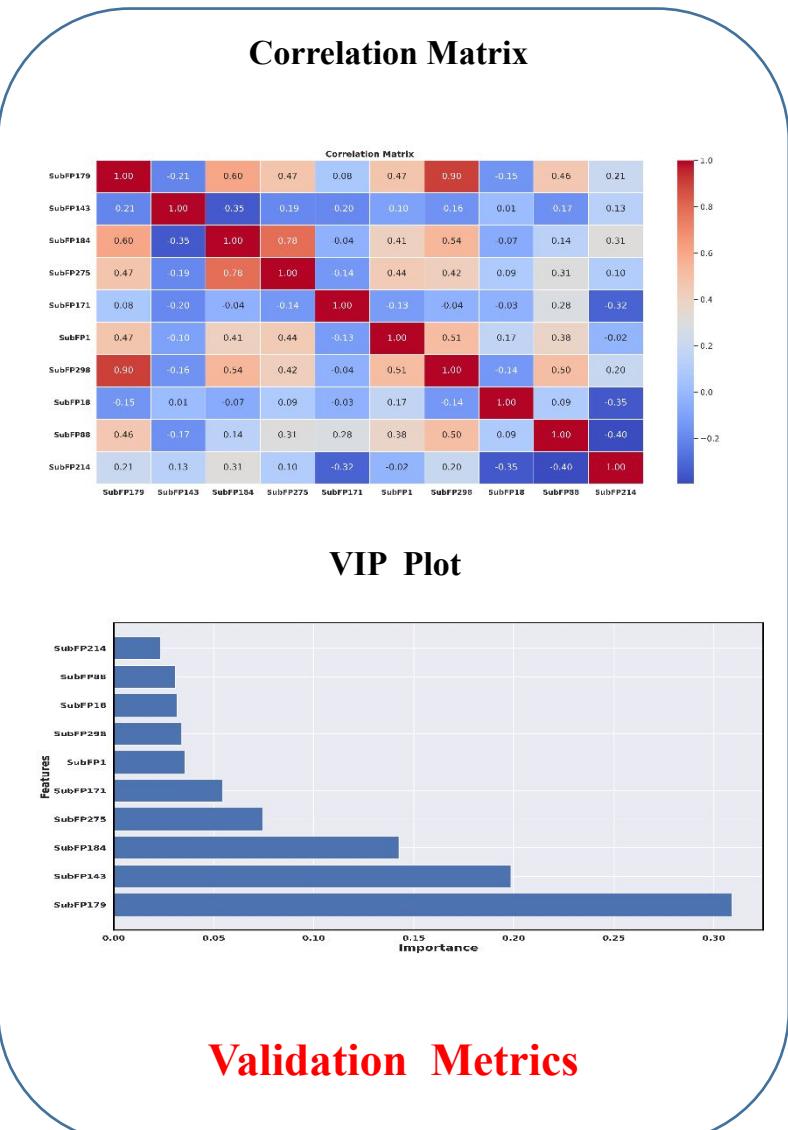
Best algorithm selection(Random Forest Classifiers)

Initial Model Performance

**Model Metrics**  
R2 (Train: 0.8789, Test: 0.8124)  
RMSE (Train: 0.5056, Test: 0.648)  
MAE (Train: 0.3677, Test: 0.5261)

FP and FN Removal

Mention molecules removed (32)

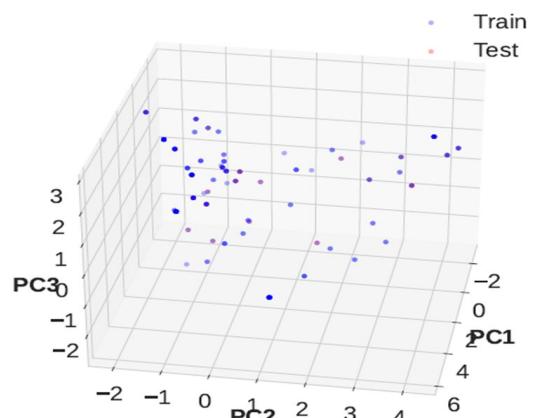


**Model Metrics**  
R2 (Train: 0.9447, Test: 0.8748)  
RMSE (Train: 0.3278, Test: 0.5289)  
MAE (Train: 0.2588, Test: 0.4314)

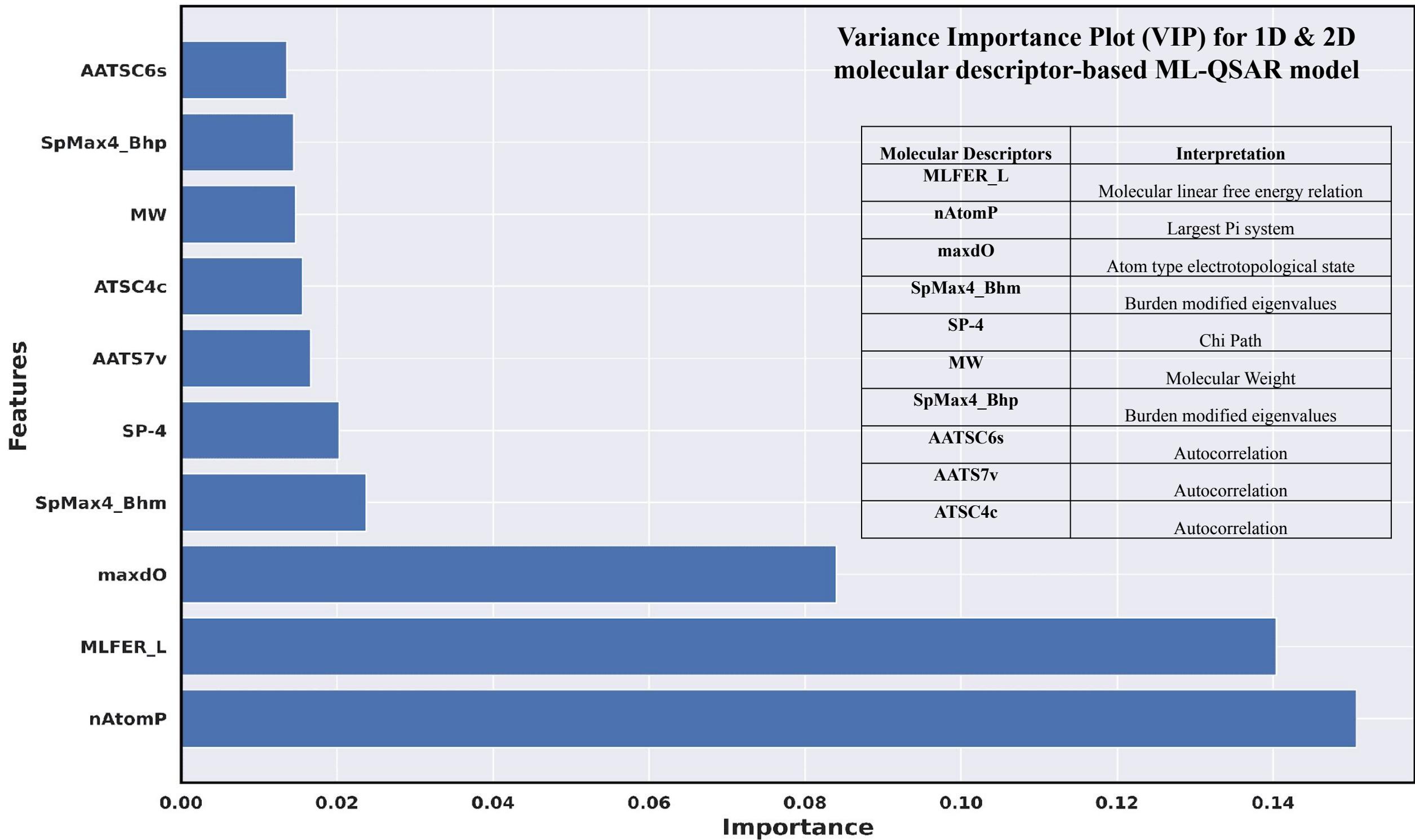
Final Model Performance

The final machine-learning model

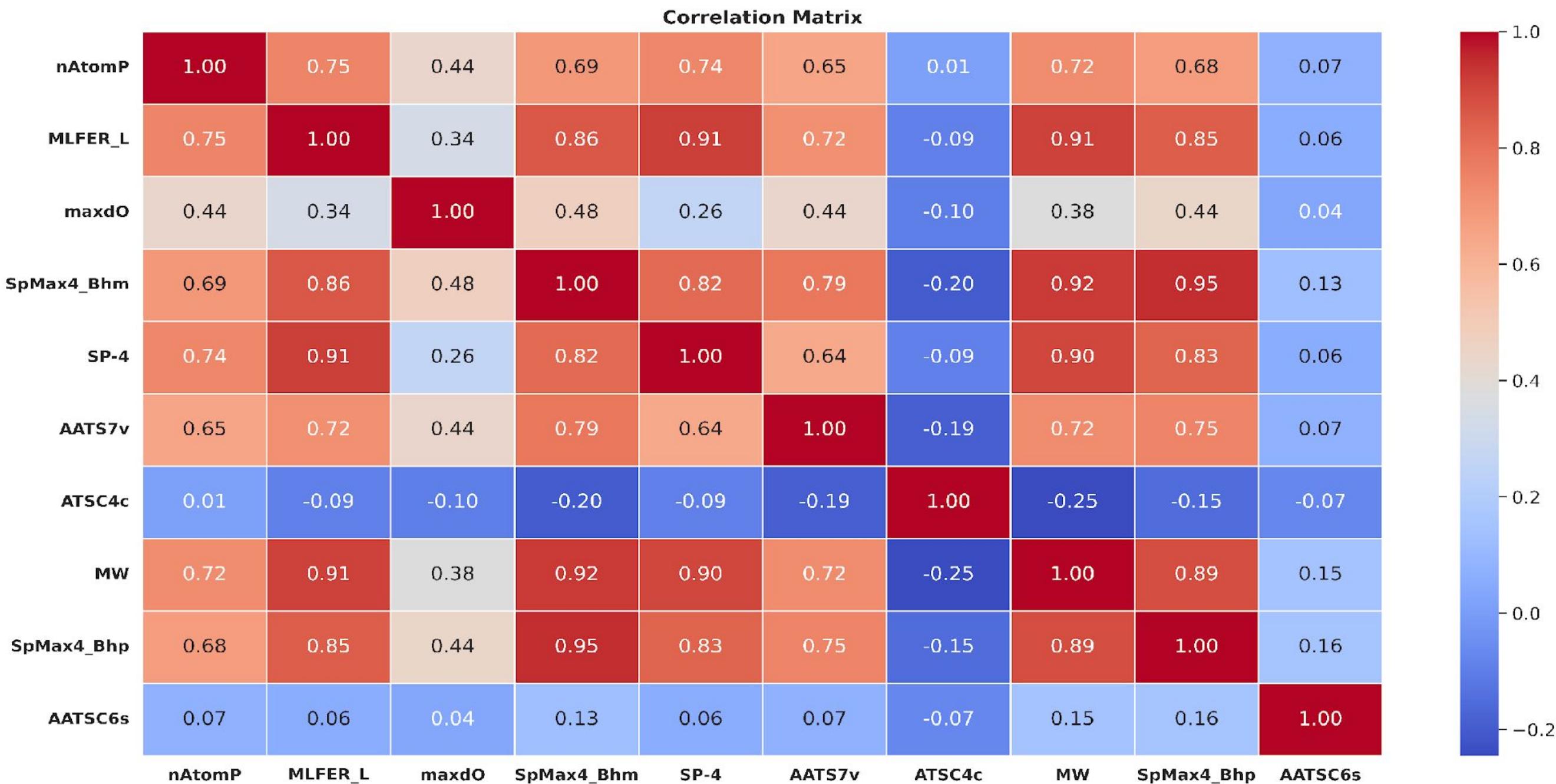
number of molecules in training(191) and test(44)



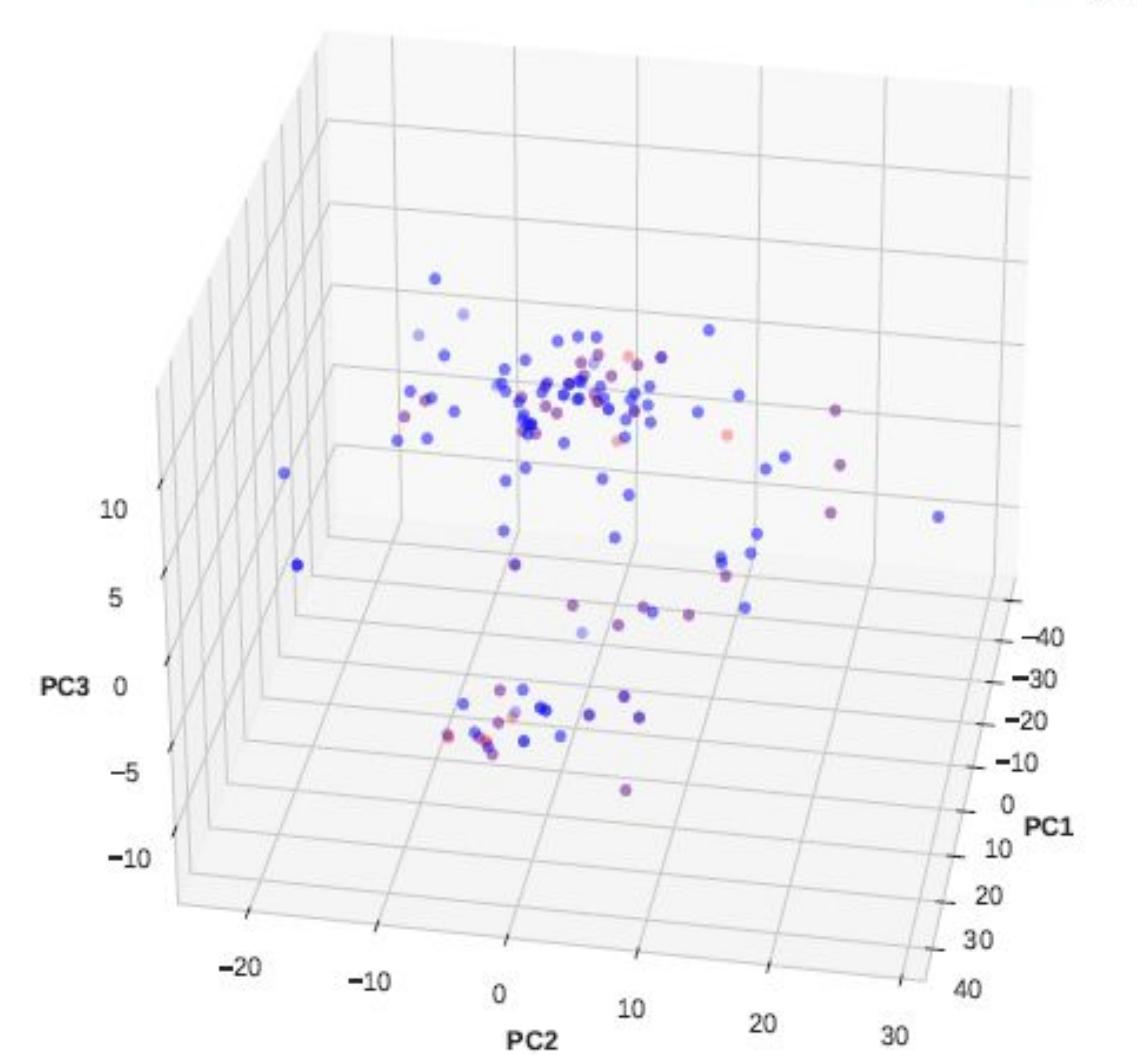
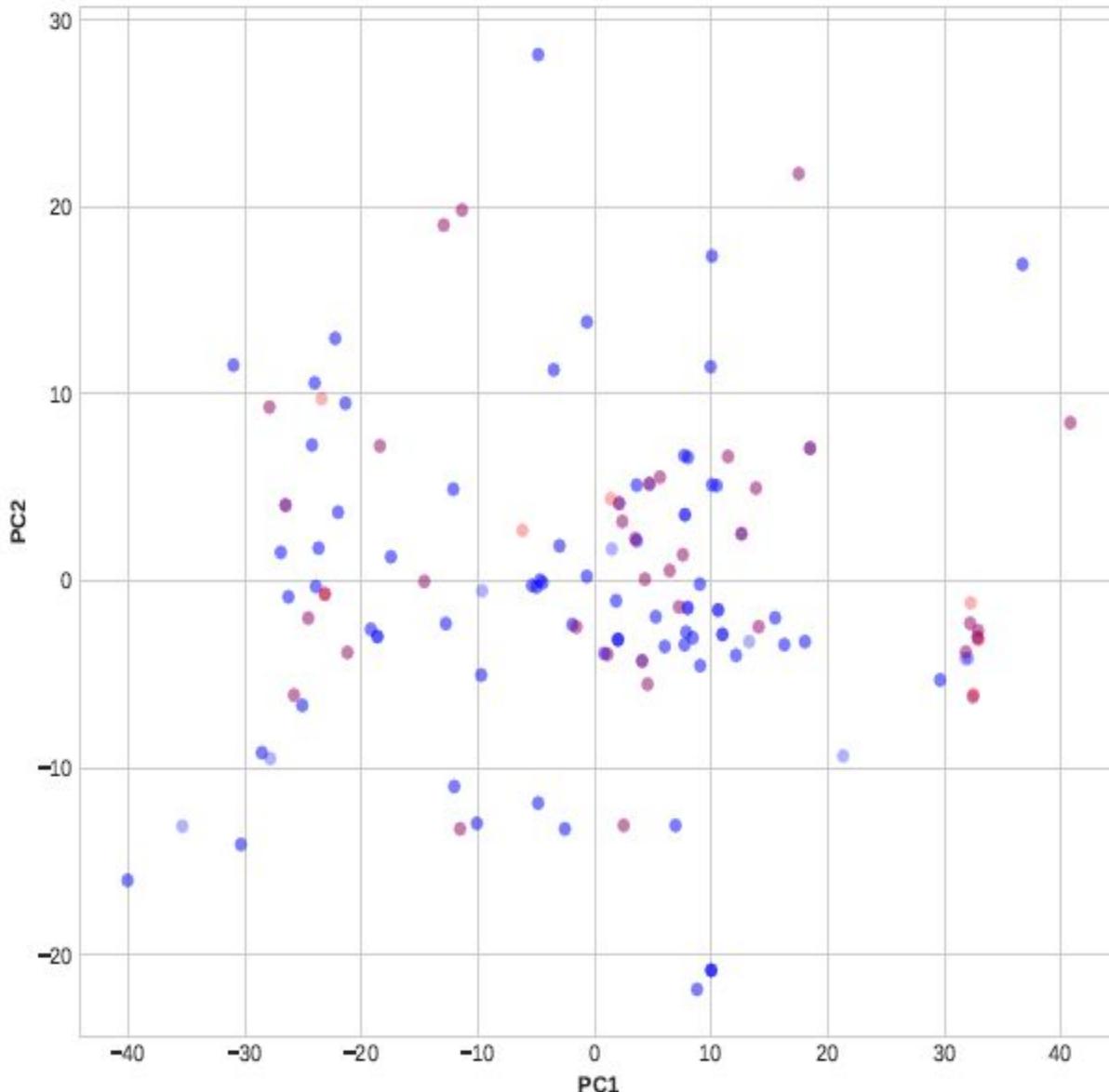
Applicability Domain Analysis



# Correlation Matrix for 1D and 2D molecular descriptor-based ML-QSAR model



# Applicability Domain (AD)/Principle Component Analysis (PCA) for 1D & 2D molecular descriptor-based ML-QSAR model





Choose a prediction model

Mtb-CA prediction model using pubchemfingerprints

**1. Upload your CSV data**

Upload your input file

Drag and drop file here

Limit 200MB per file • TXT

[Browse files](#)[Example input file](#)[Predict](#)

# Mtb-CA-pred app

Mtb-CA-pred allows users to predict bioactivity of a query molecule against the Mycobacterium tuberculosis carbonic anhydrase target protein.

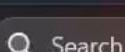
[Main](#) [About](#) [What is Mycobacterium tuberculosis carbonic anhydrase \(Mtb-CA\)?](#) [Dataset](#) [Model performance](#) [Py](#)

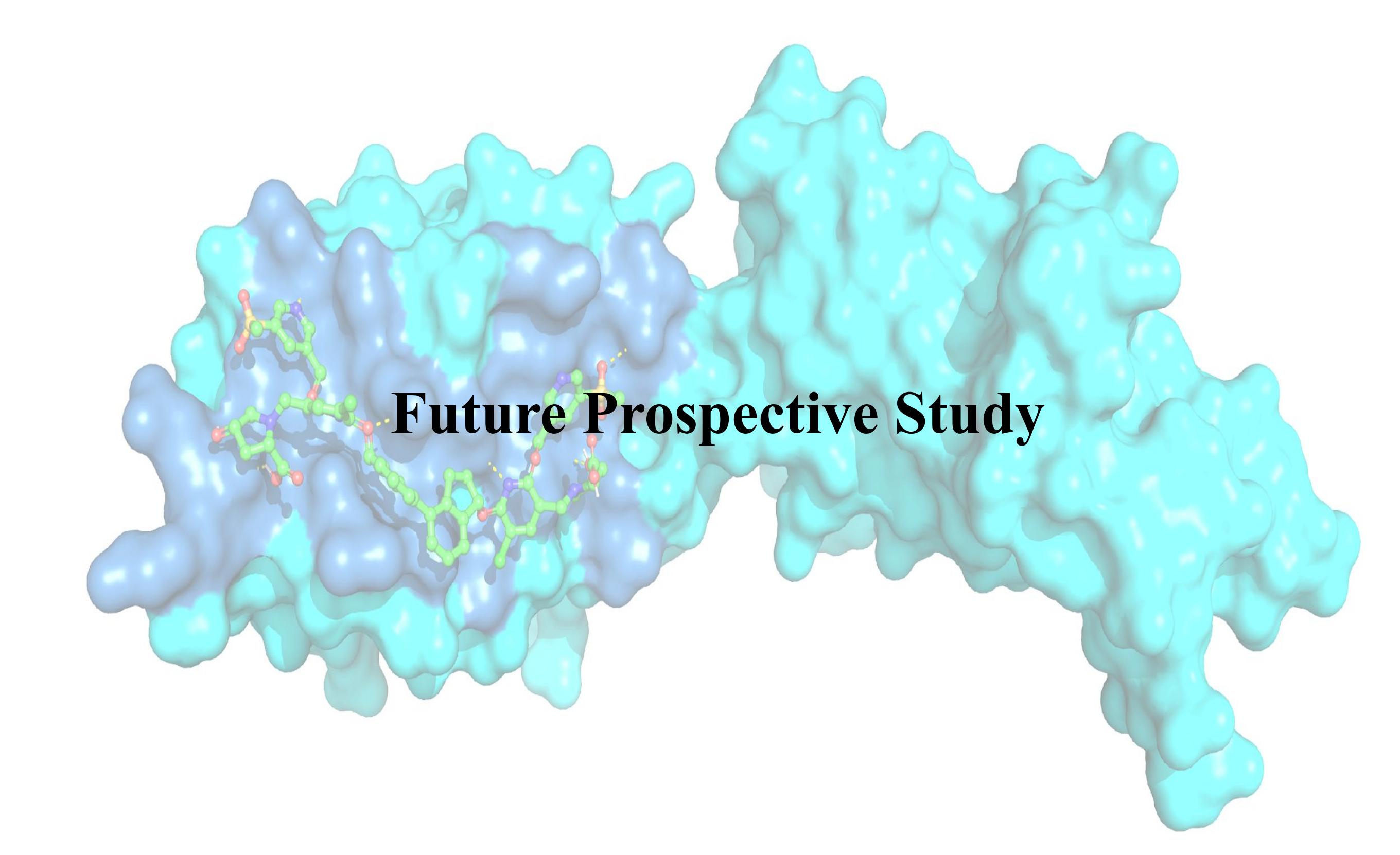
## Application Description

This module of [Mtb-CA-pred](#) has been built to predict bioactivity and identify potent inhibitors against Mycobacterium tuberculosis carbonic anhydrase using robust machine learning algorithms.

# Predict bioactivity of molecules against Mtb-CA using pubchemfingerprints

II mtb-ca-pred.streamlit.app is sharing your screen.

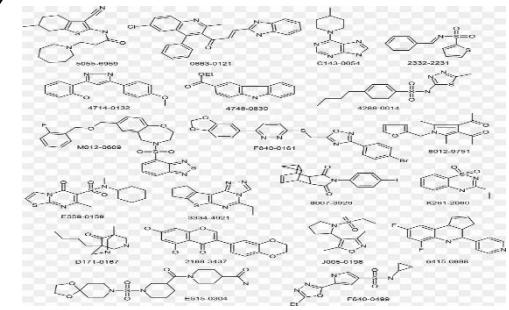
[Stop sharing](#)[Hide](#)[Manage app](#)17:09  
11-Nov-2023



A 3D surface model of a protein binding site is shown, colored with a gradient from light blue to cyan. A small, rigid ligand molecule is bound within the cavity. The ligand has a complex structure with various functional groups, including a carbonyl group, a phosphate group, and aromatic rings. It is shown in stick representation with green carbon atoms, red oxygen atoms, and purple nitrogen atoms. Yellow dashed lines represent hydrogen bonds between the ligand and the surrounding protein residues.

**Future Prospective Study**

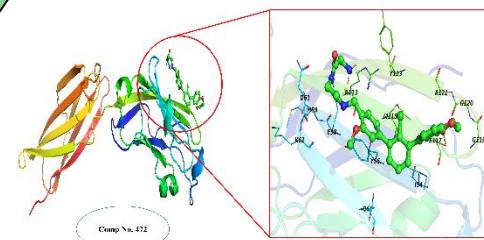
## Retrieval of large drug-like molecules chemical libraries from different databases



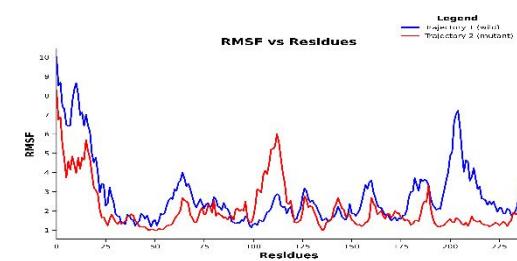
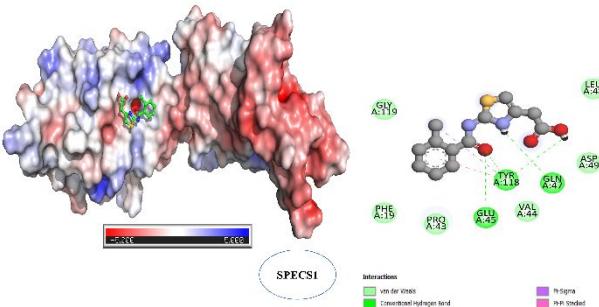
## Screening of the retrieved libraries by using Lipinski's Rule of Five and ML-QSAR based web application

A screenshot of the Mtb-CA-pred app web interface. The interface includes a header with the app name and a sub-header 'Application Description'. Below this, there is a section titled 'Predict bioactivity of molecules against Mtb-CA using pubchemfingerprints' with a sub-section 'RMSF vs Residues' containing a line graph and a legend.

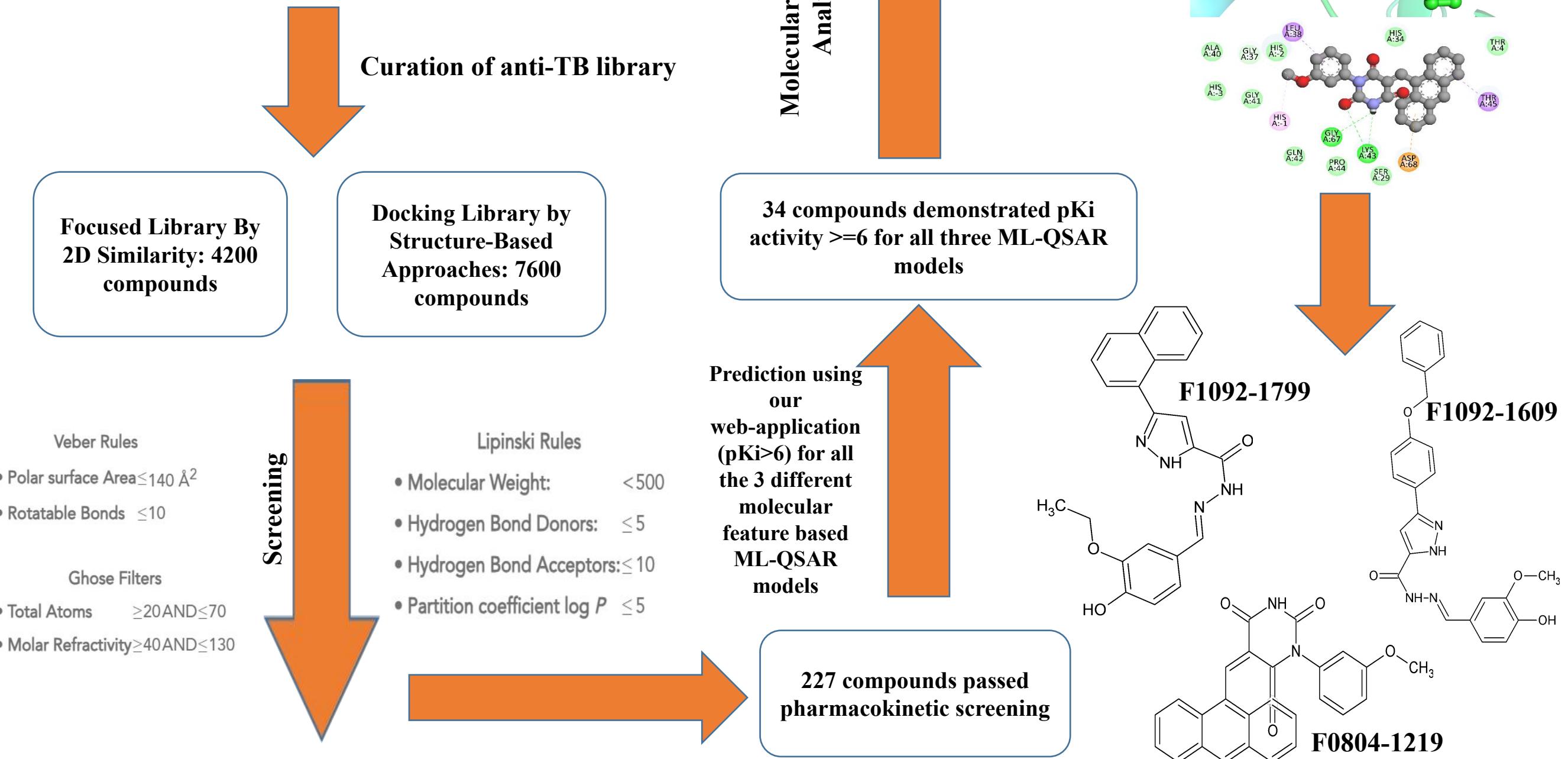
## Molecular Docking, ADMET analysis, Molecular Dynamics Simulation studies



## Hit Compounds



# Virtual Screening and Molecular Docking Workflow

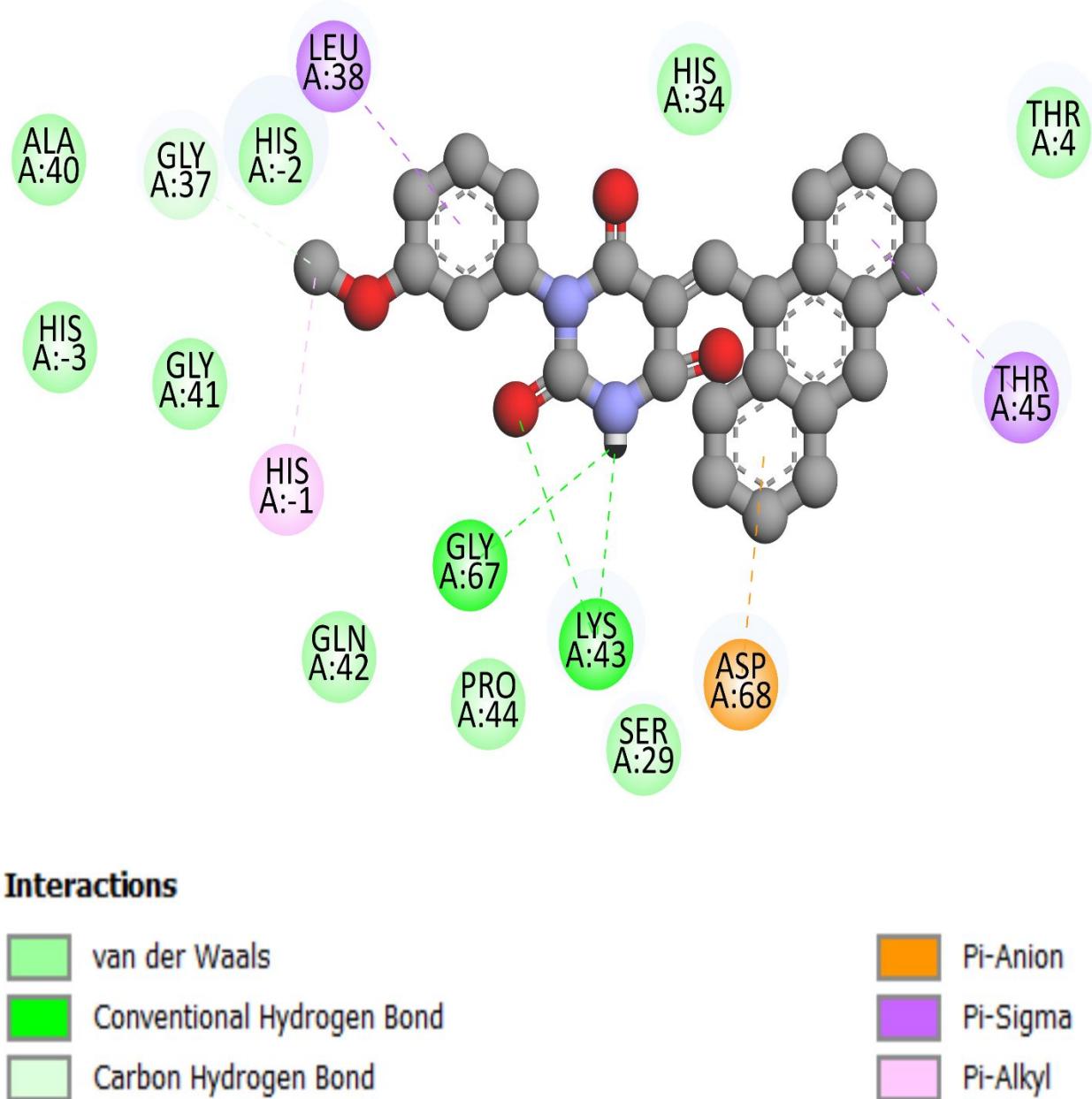
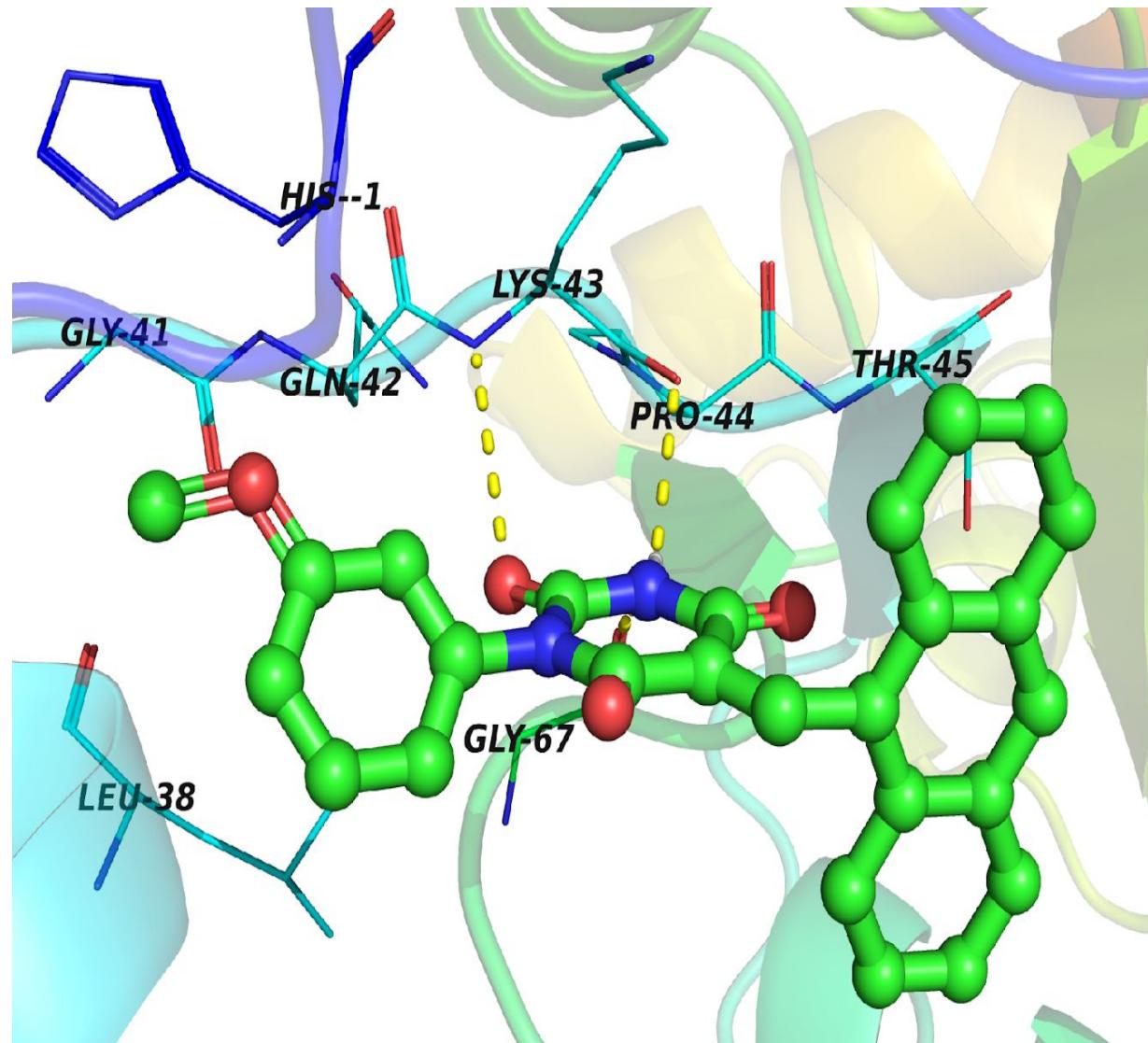


# Comparative analysis of predicted bioactivity and binding affinity of small molecules against *Mtb carbonic anhydrase (mtbca2 and mtbca1)*

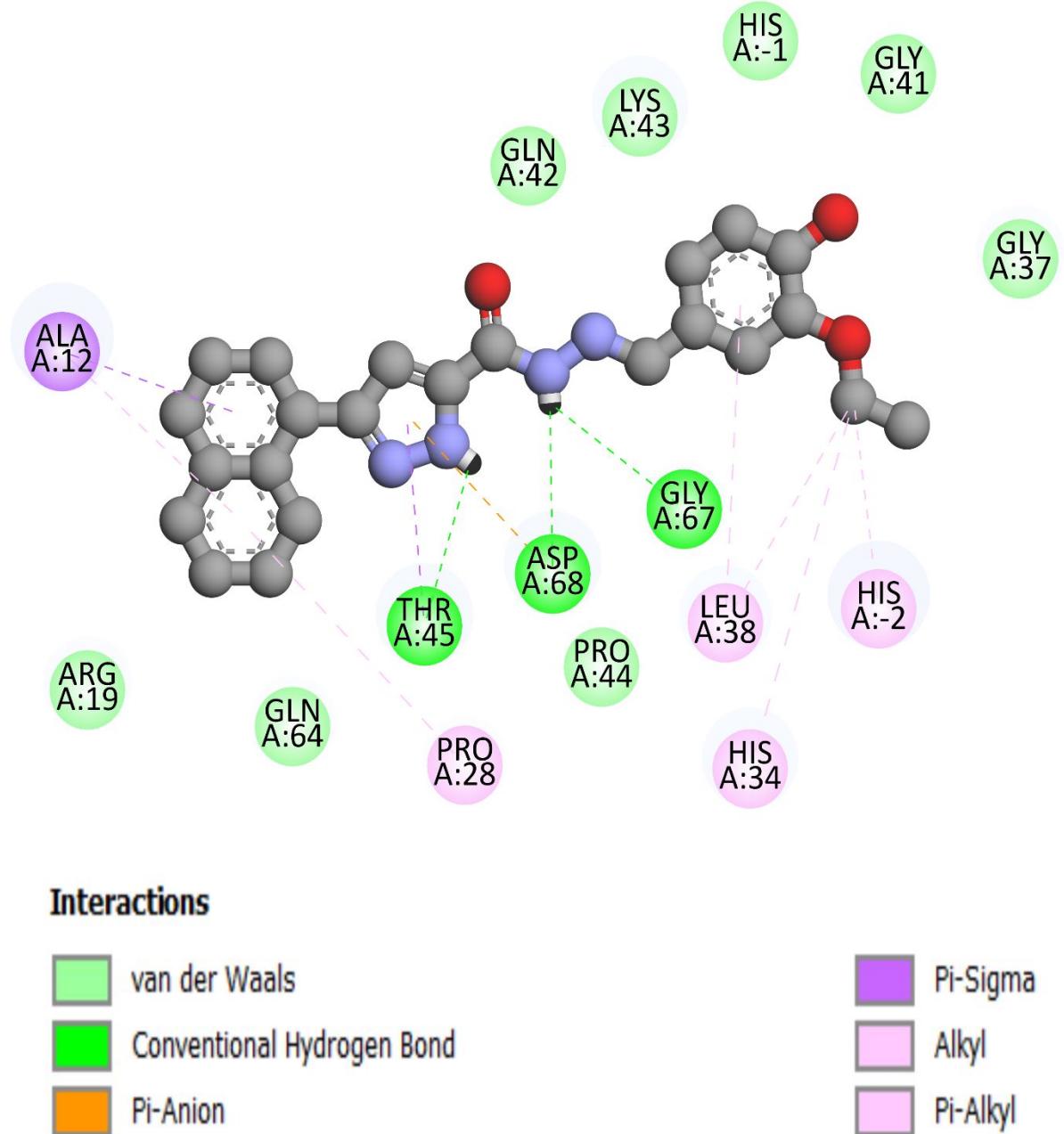
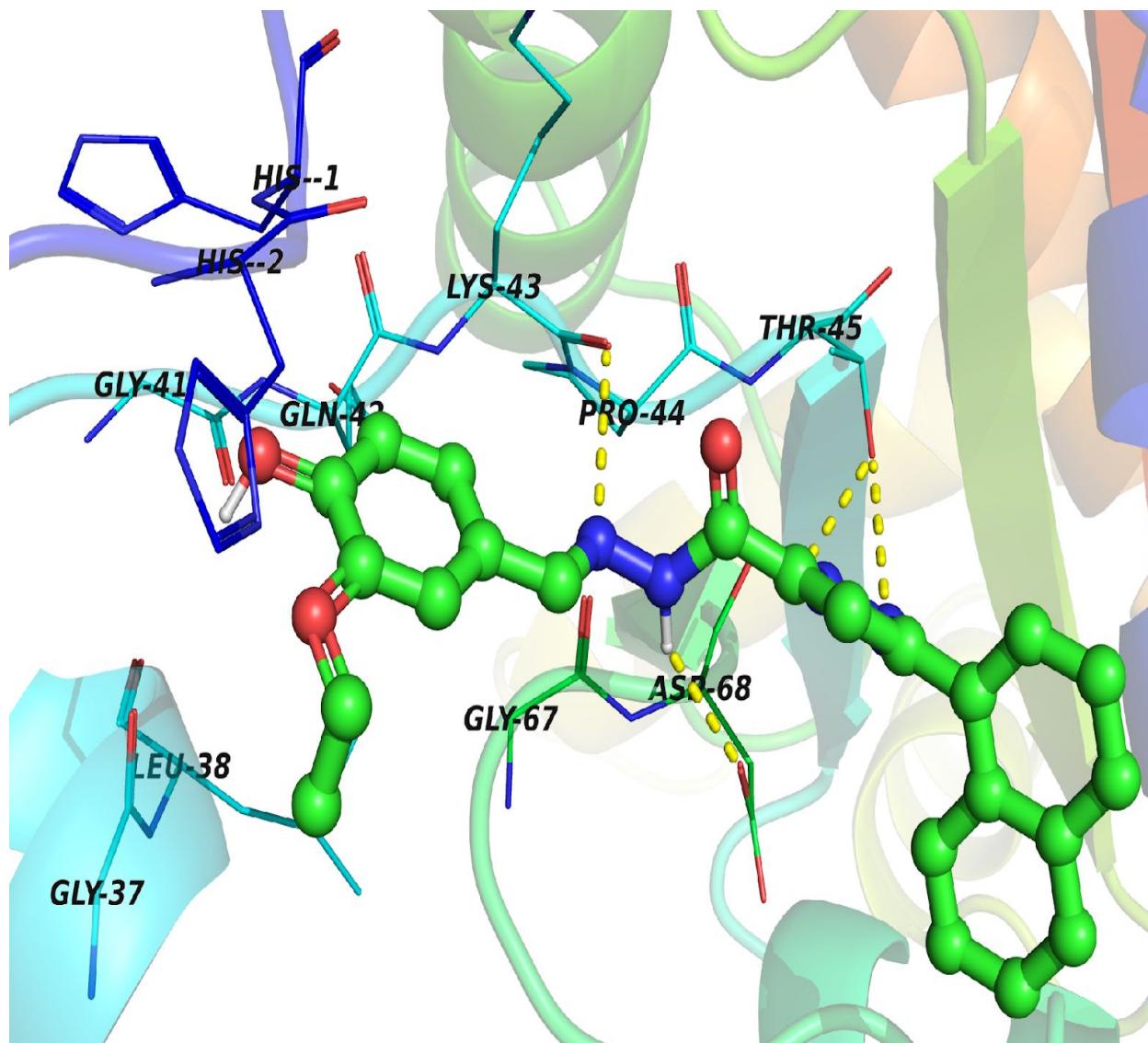
Molecule	F0804-1219	F1092-1799	F1092-1609	Rifamycin	Bedaquiline
<b>predictedpKi_2Ddescriptors</b>	6.461 (345.94 nM)	6.635 (231.74 nM)	6.556 (277.97 nM)	6.297 (504.66 nM)	6.051 (889.20 nM)
<b>predictedpKi_PubChemFPs</b>	6.341 (456.04 nM)	6.037 (918.33 nM)	6.04 (912.01 nM)	6.605 (248.31 nM)	5.846 (1425.61 nM)
<b>predictedpKi_SubstructureFPs</b>	6.069 (853 nM)	6.637 (230.67 nM)	6.692 (202.23 nM)	5.546 (2844.46 nM)	5.817 (1524.05 nM)
<b>Cumulative pKi</b>	<b>6.29 (512.87 nM)</b>	<b>6.47 (338.85 nM)</b>	<b>6.43 (371.53 nM)</b>	<b>6.15 (707.95 nM)</b>	5.90 (1258.93 nM)
<b>Binding affinity (mtbca1) (kcal/mol)</b>	<b>-7.7</b>	<b>-7.5</b>	<b>-7.9</b>	-7.4	-7.1
<b>Binding affinity (mtbca2) (kcal/mol)</b>	<b>-8.3</b>	<b>-8.1</b>	-7.3	-7.3	-6.6

\*Satisfactory binding affinity/ good bioactivity

# Molecular Interaction Analysis of F0804-1219 molecule with *Mtb carbonic anhydrase* (*mtbca2*)



# Molecular Interaction Analysis of F1092-1799 molecule with *Mtb carbonic anhydrase (mtbca2)*



## Interactions

van der Waals

Conventional Hydrogen Bond

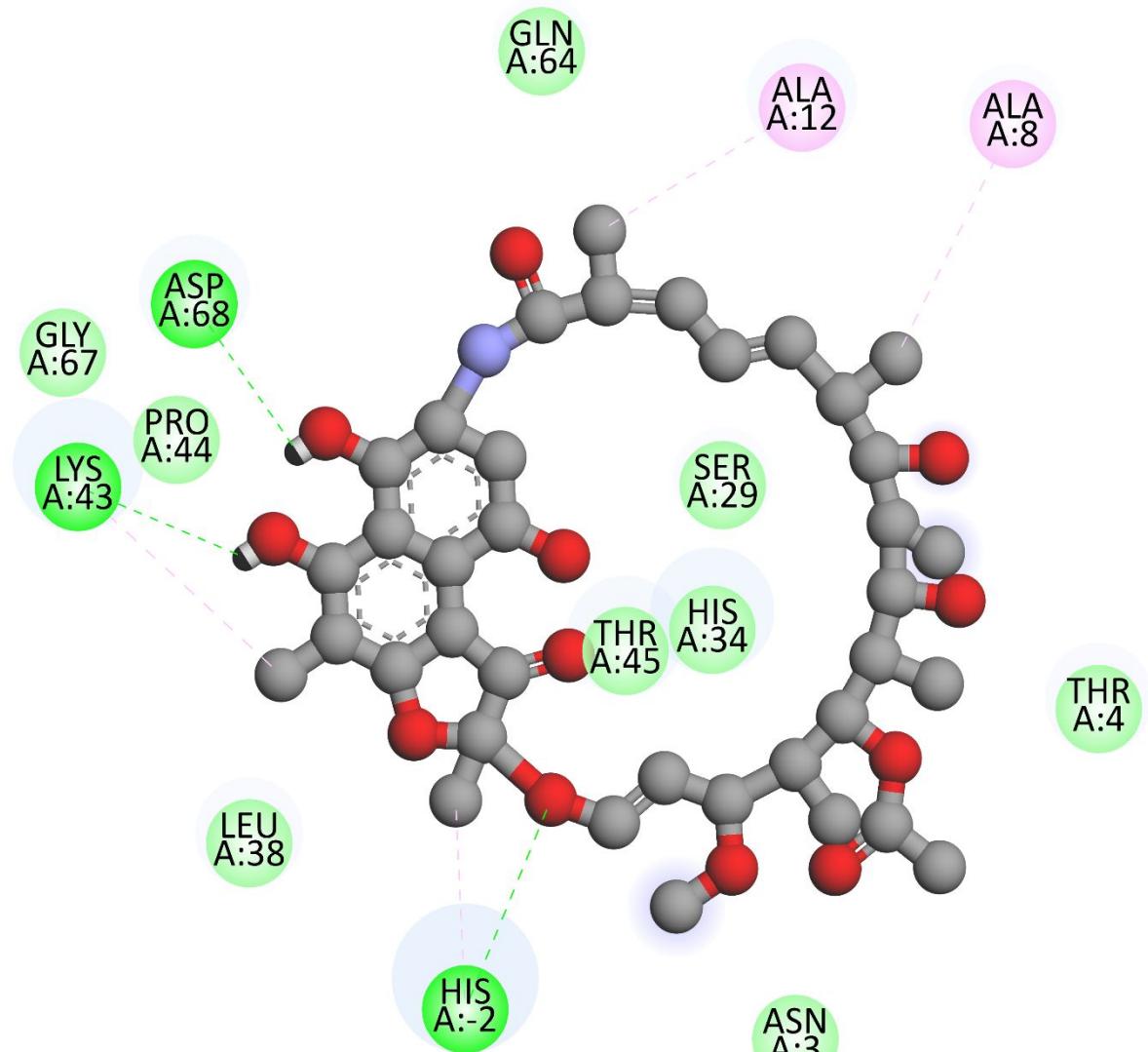
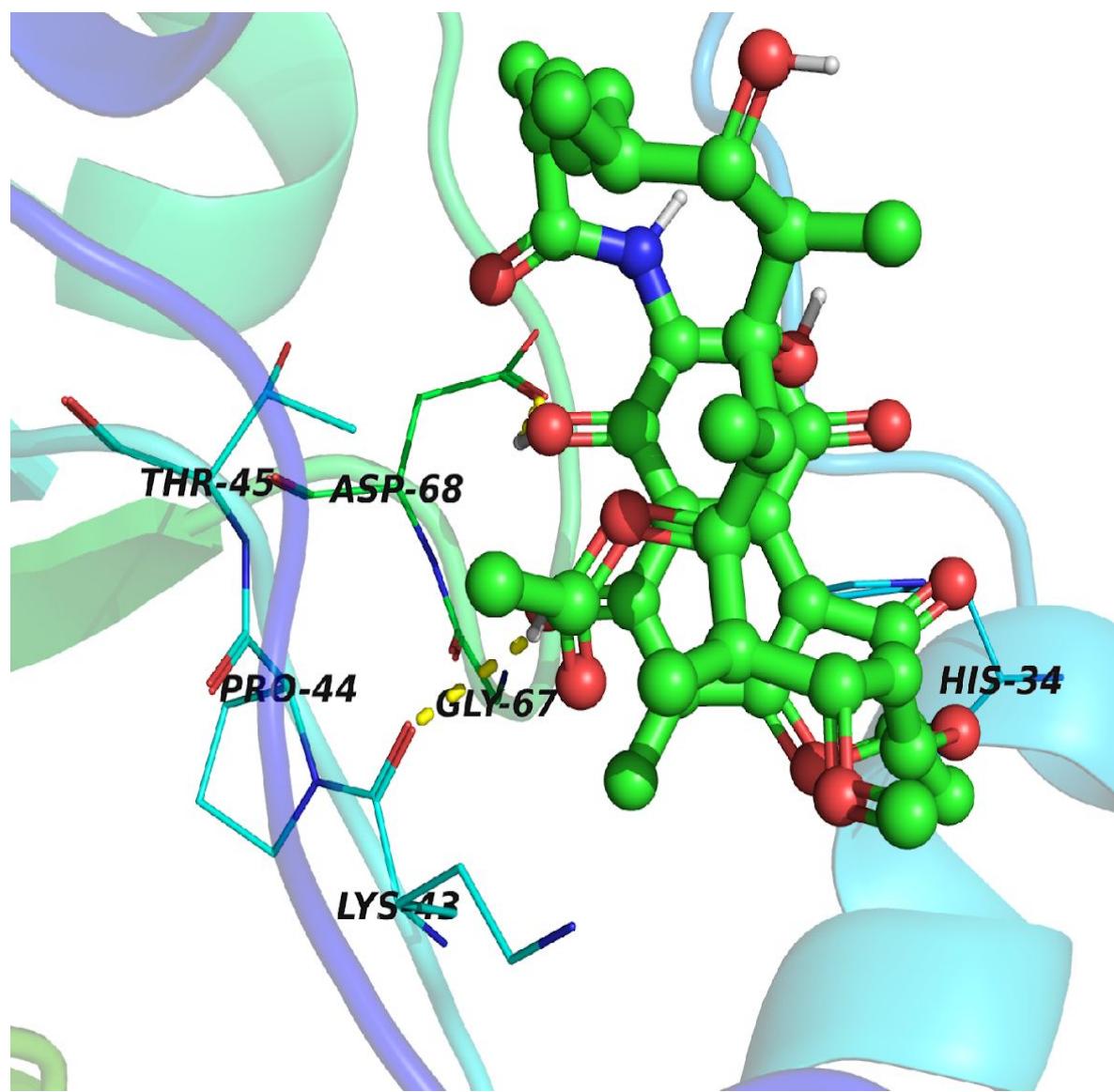
Pi-Anion

Pi-Sigma

Alkyl

Pi-Alkyl

# Molecular Interaction Analysis of Rifamycin with *Mtb* carbonic anhydrase (*mtbca2*)



## Interactions

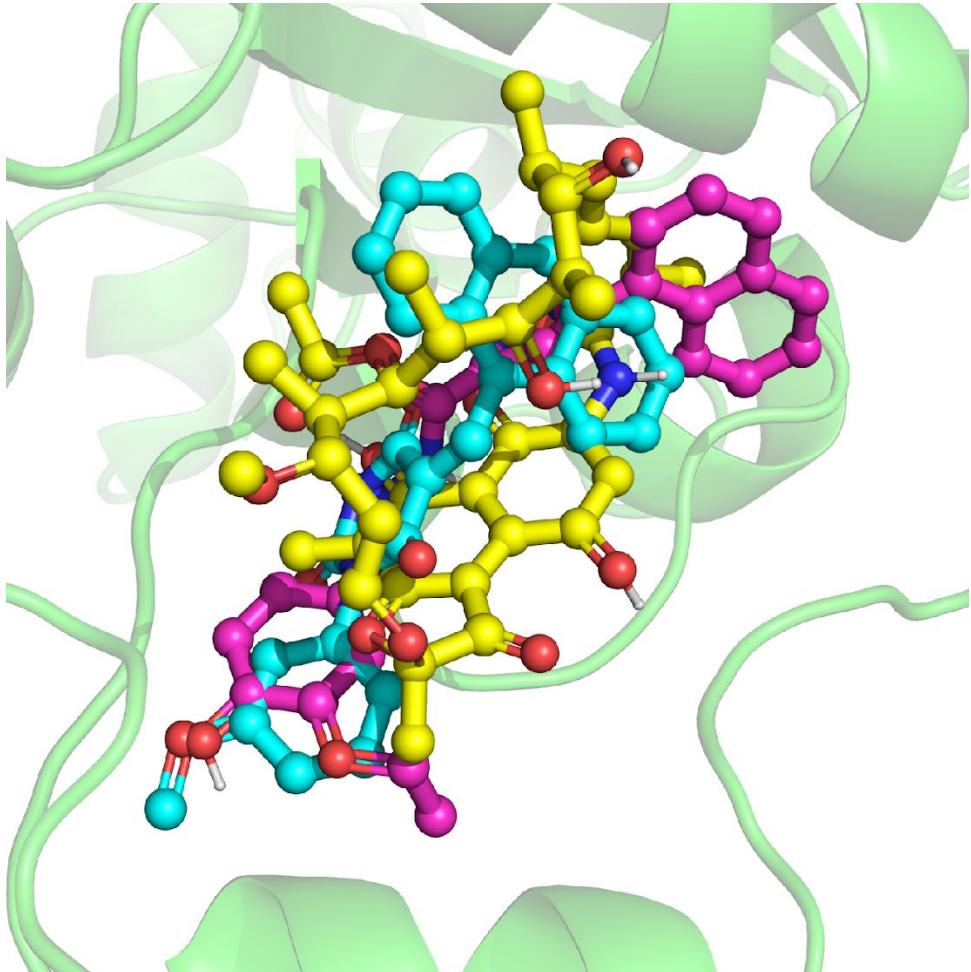
van der Waals

Conventional Hydrogen Bond

Alkyl

Pi-Alkyl

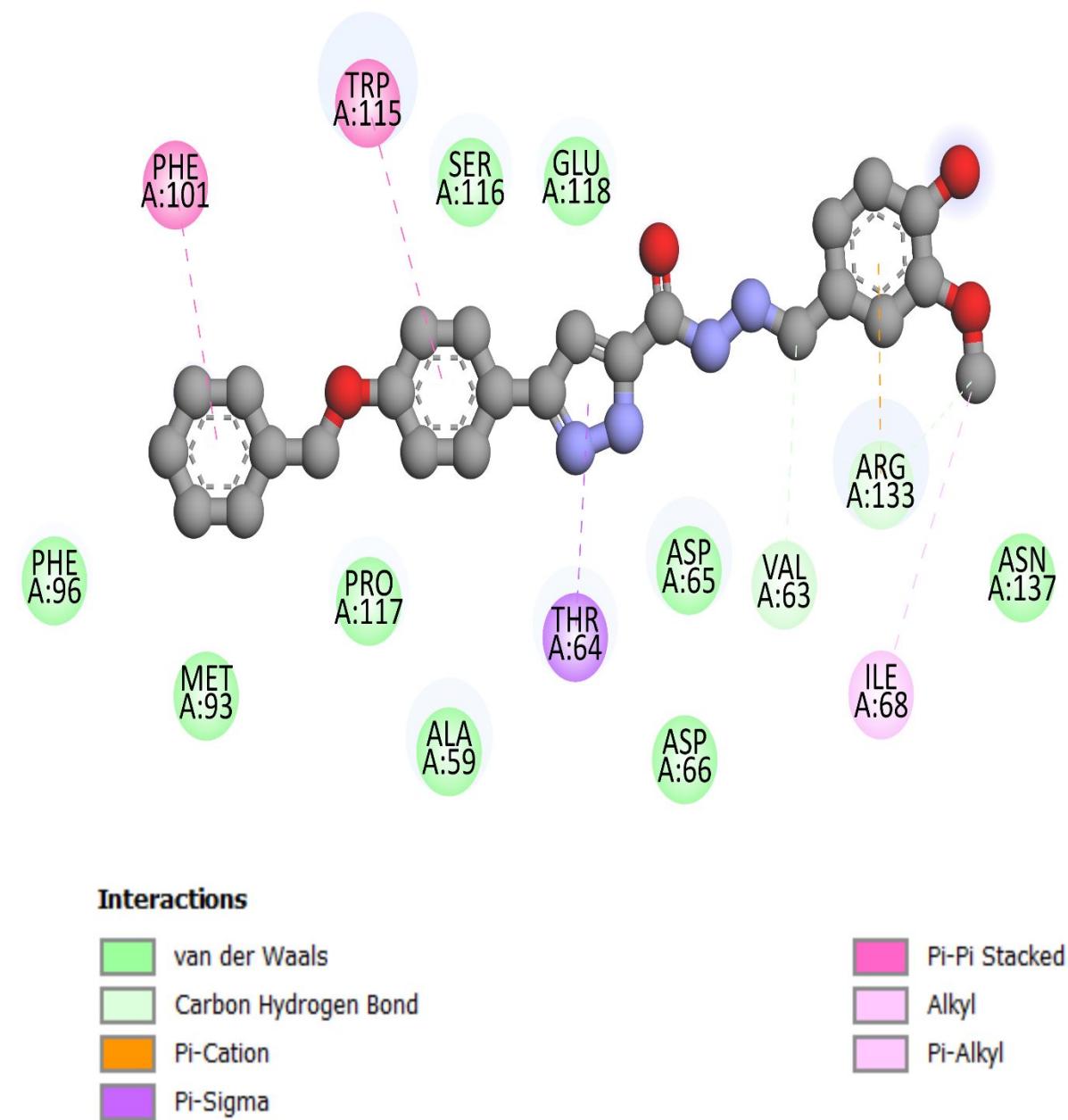
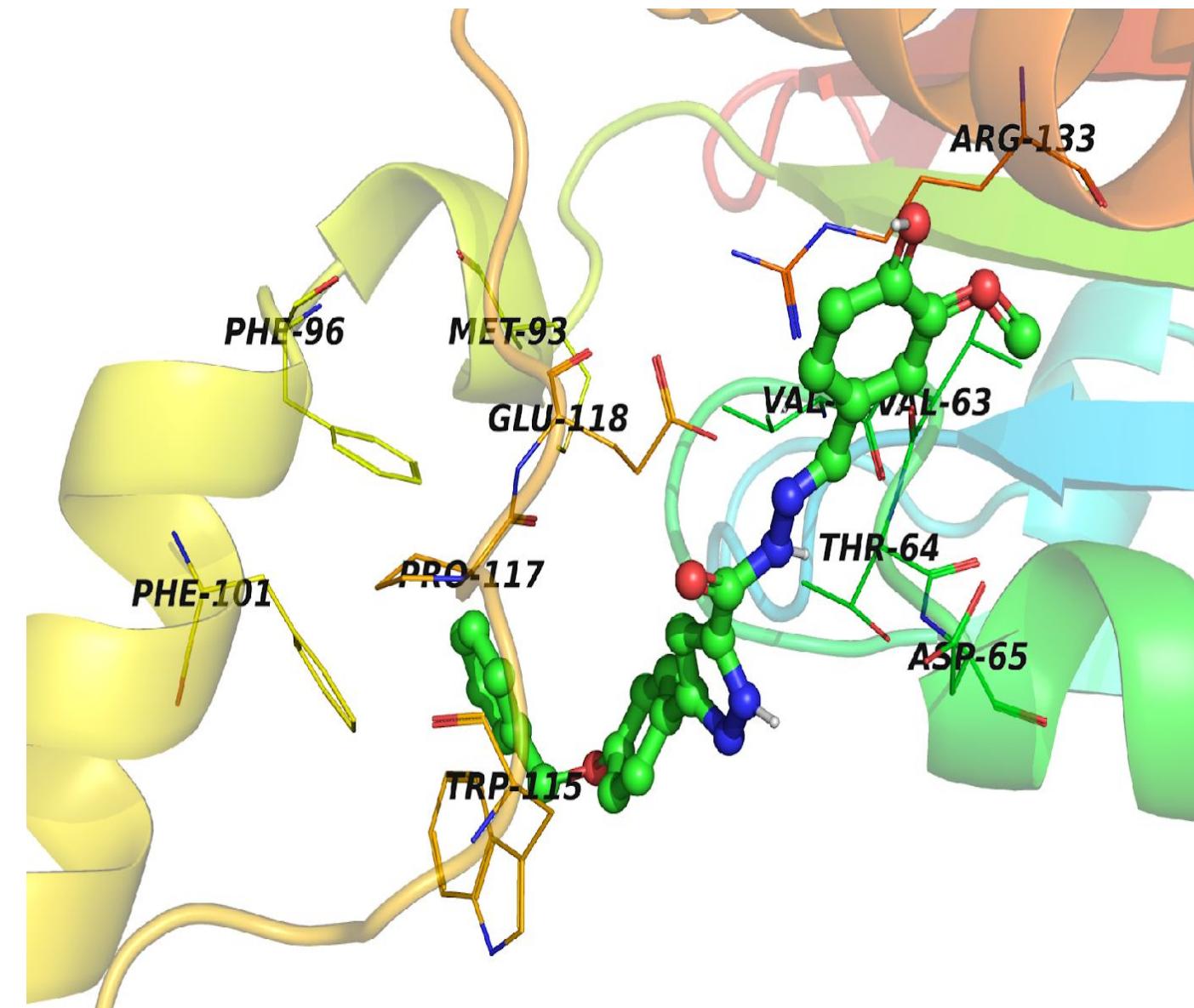
# Common Molecular Interaction Analysis with *Mtb carbonic anhydrase (mtbca2)*



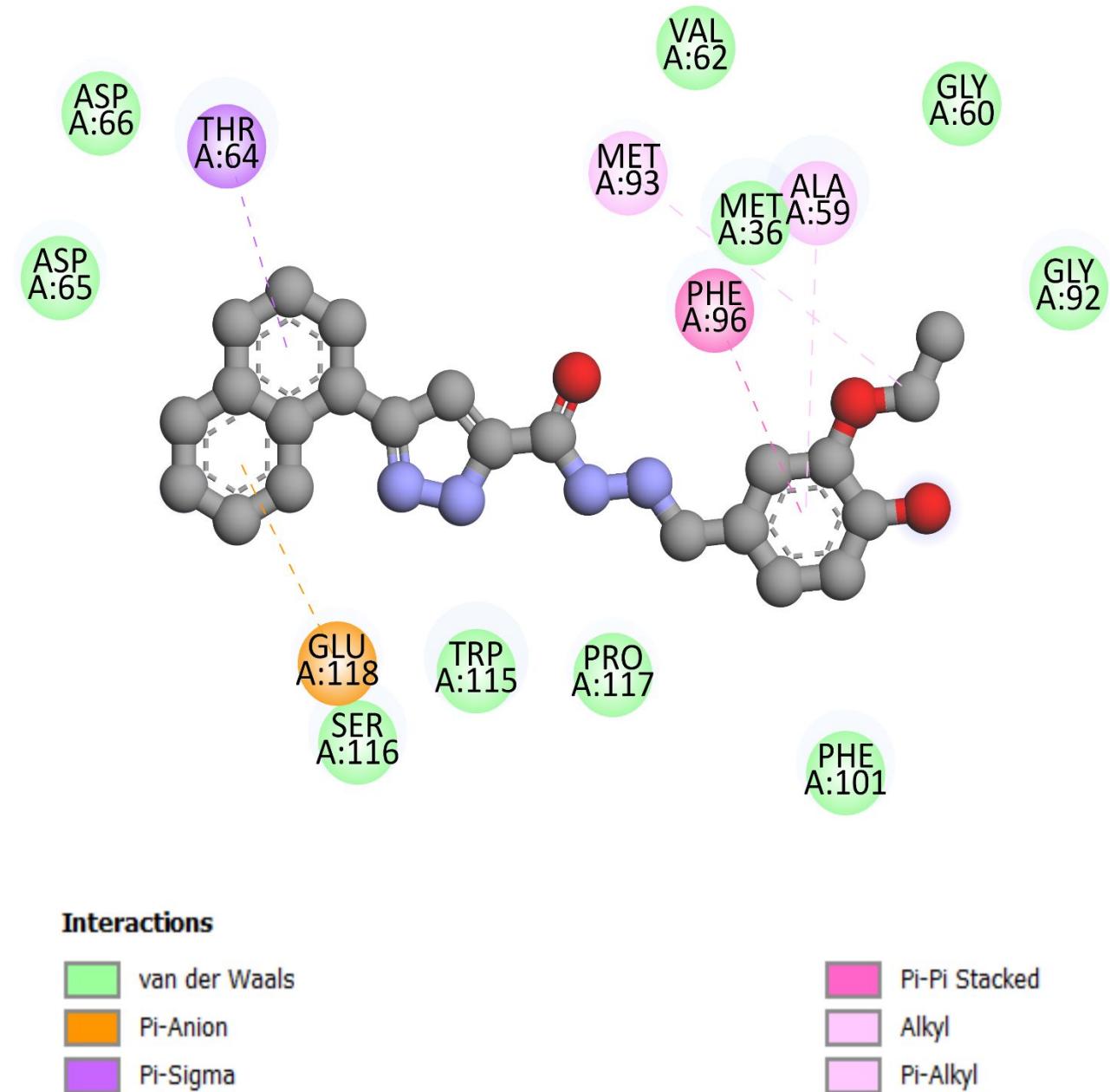
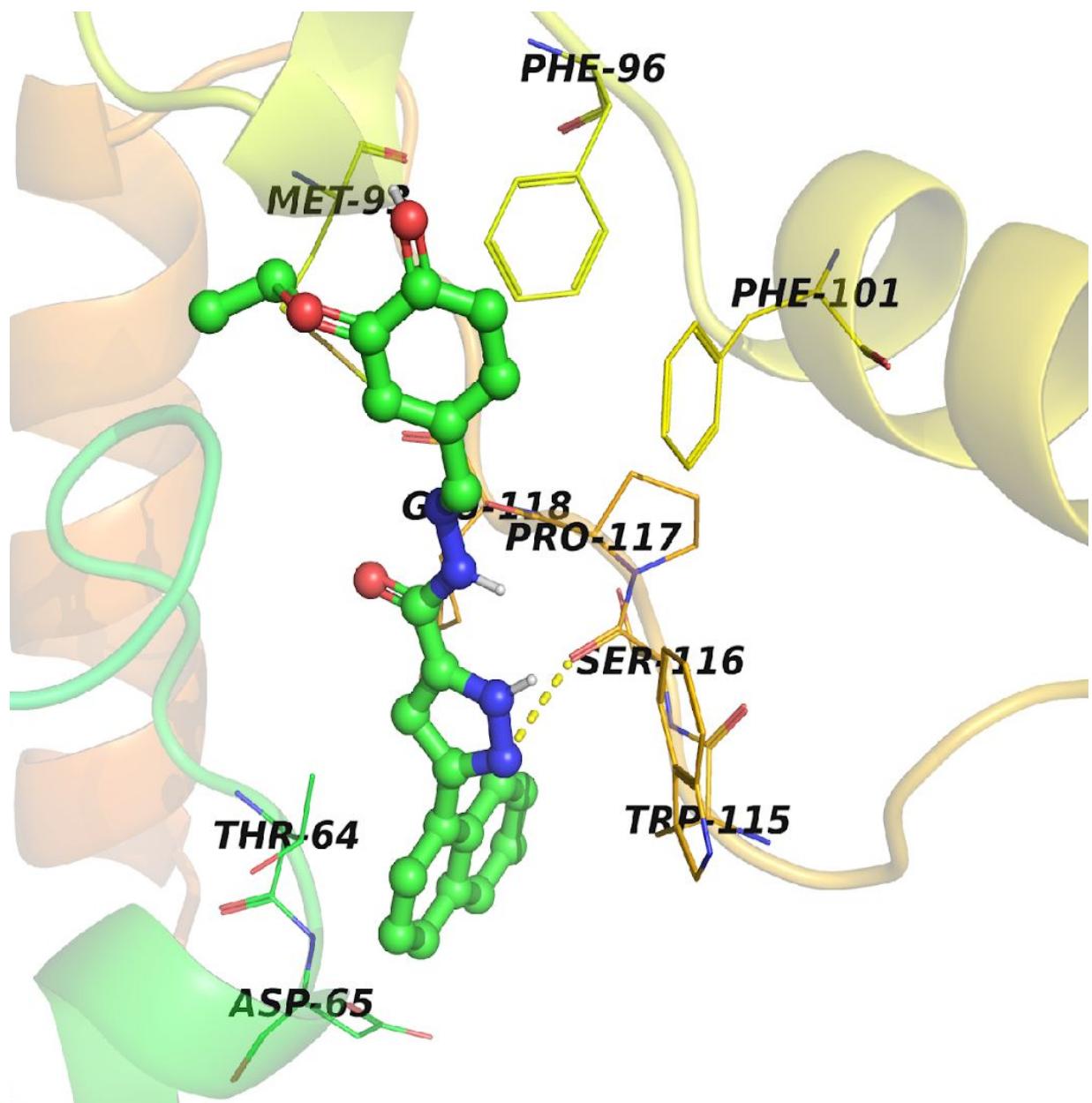
Molecule	Amino Acid Interactions ( <i>mtbca2</i> )
F0804-1219	LEU38, HIS-2, GLY37, ALA40, HIS-3, GLY41, HIS-1, <b>GLY67</b> , LYS43, PRO44, GLN42, SER29, <b>ASP68</b> , THR45, THR4, <b>HIS34</b>
F1092-1799	ALA12, ARG19, GLN64, <b>THR45</b> , <b>ASP68</b> , <b>GLY67</b> , PRO28, PRO44, LEU38, HIS-2, HIS34, GLY37, GLY41, HIS-1, <b>LYS43</b> , GLN42
Rifamycin	ASP68, <b>GLY67</b> , PRO44, LYS43, LEU38, HIS-2, ASN3, <b>THR45</b> , THR4, <b>HIS34</b> , SER29, GLN64, ALA12, ALA8

\*Common Residues/Proposed Active site for small molecule binding

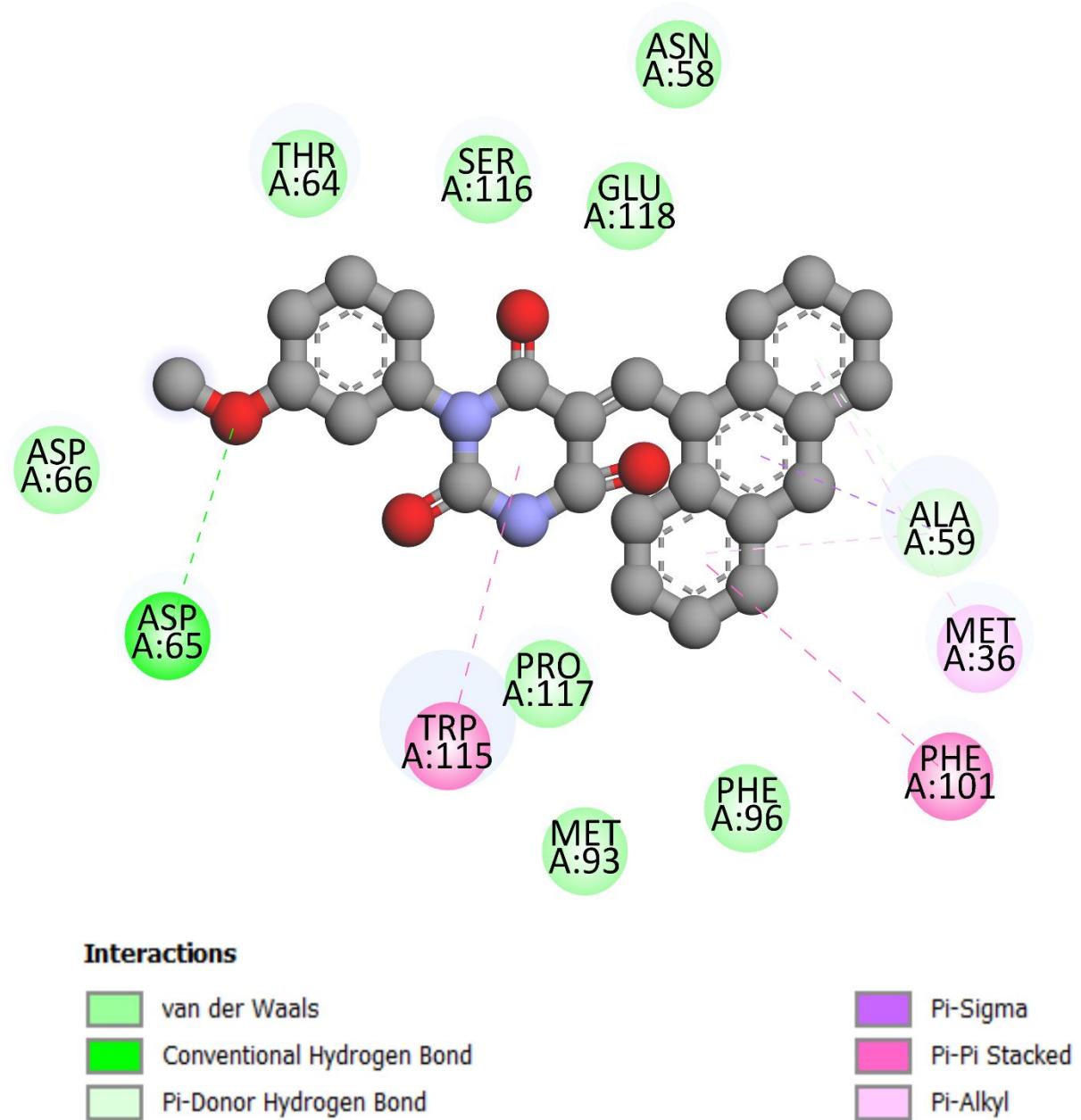
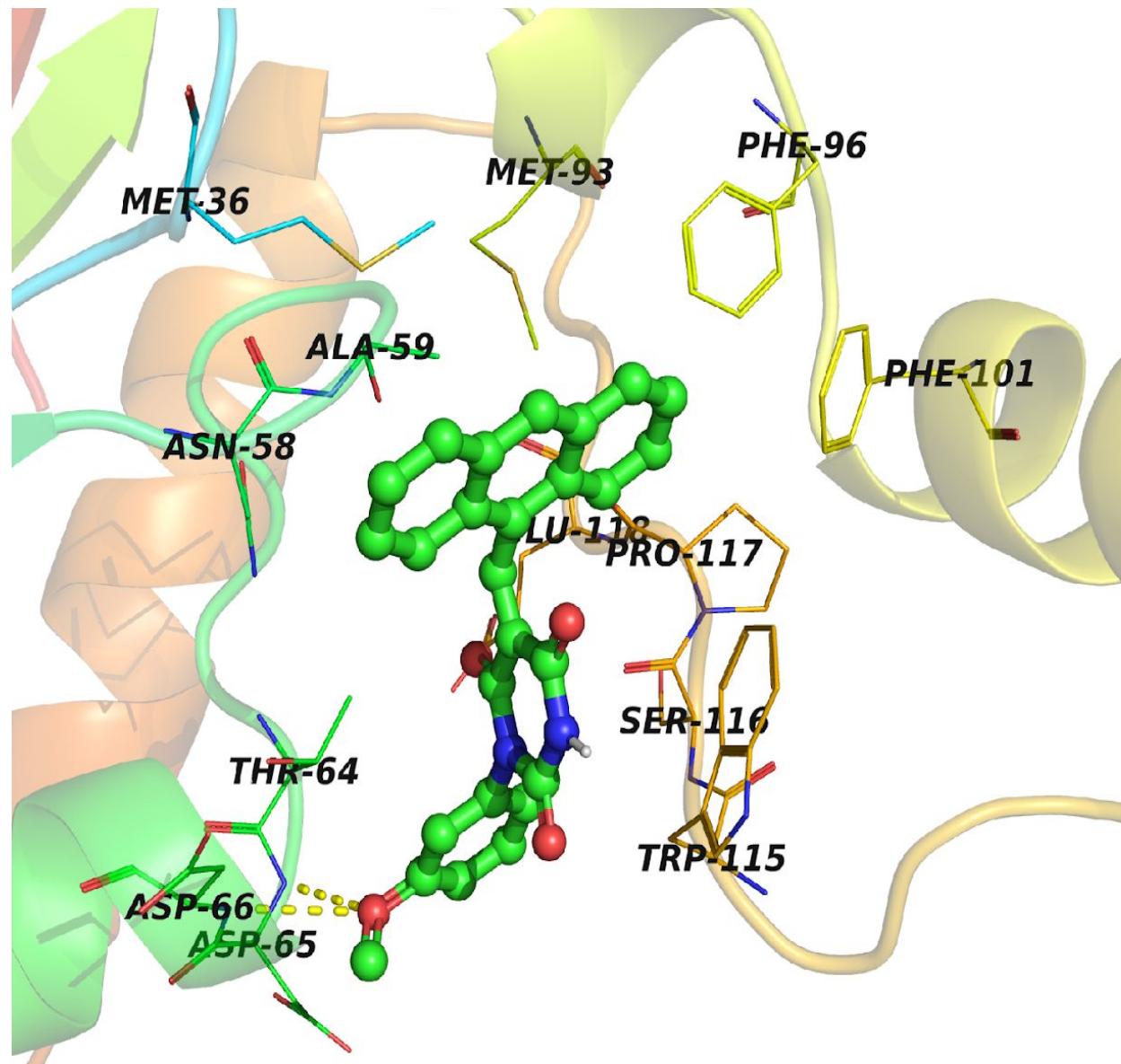
# Molecular Interaction Analysis of F1092-1609 molecule with *Mtb carbonic anhydrase (mtbca1)*



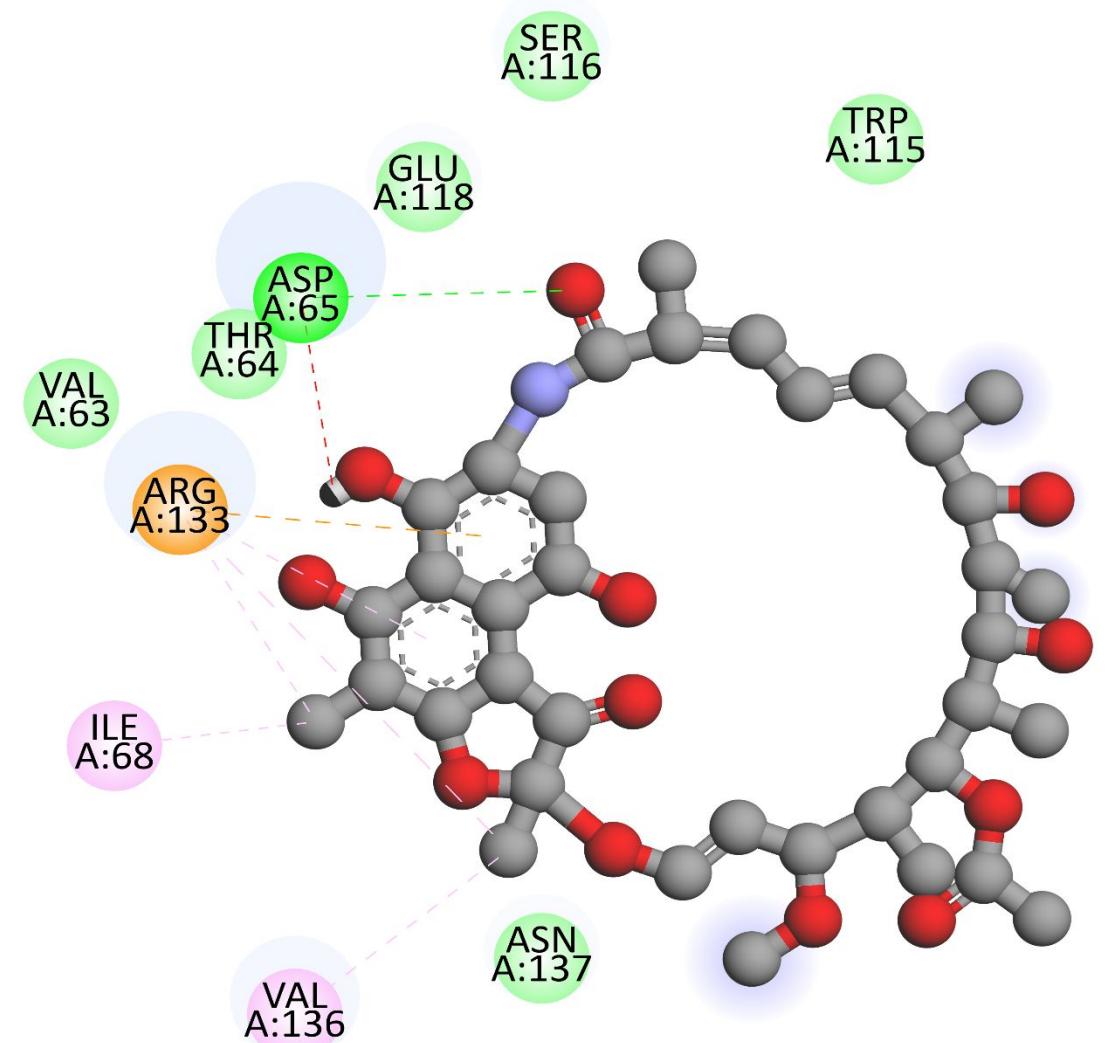
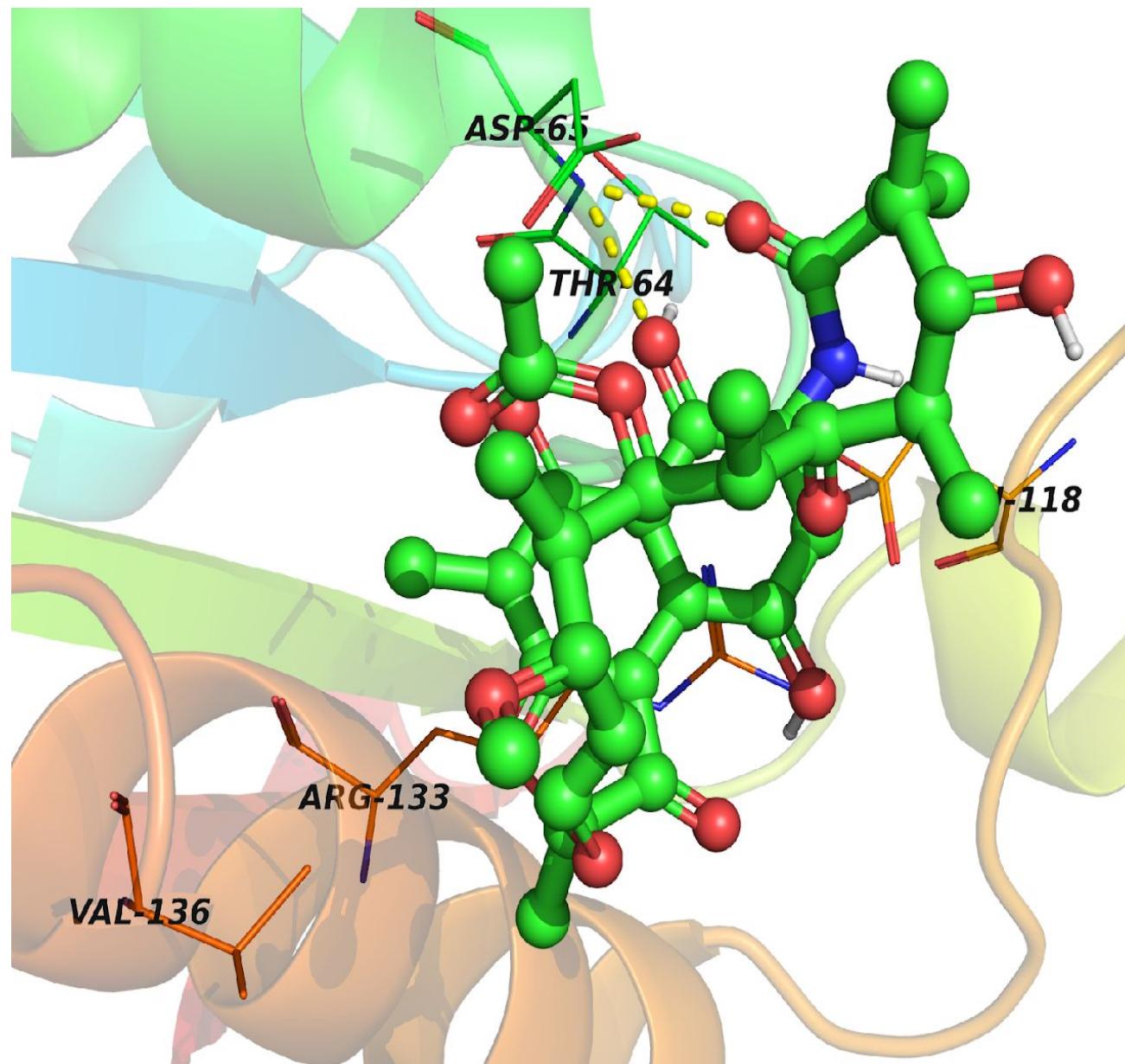
# Molecular Interaction Analysis of F1092-1799 molecule with *Mtb carbonic anhydrase (mtbca1)*



# Molecular Interaction Analysis of F0804-1219 molecule with *Mtb carbonic anhydrase (mtbca1)*



# Molecular Interaction Analysis of Rifamycin with *Mtb* carbonic anhydrase (*mtbca1*)

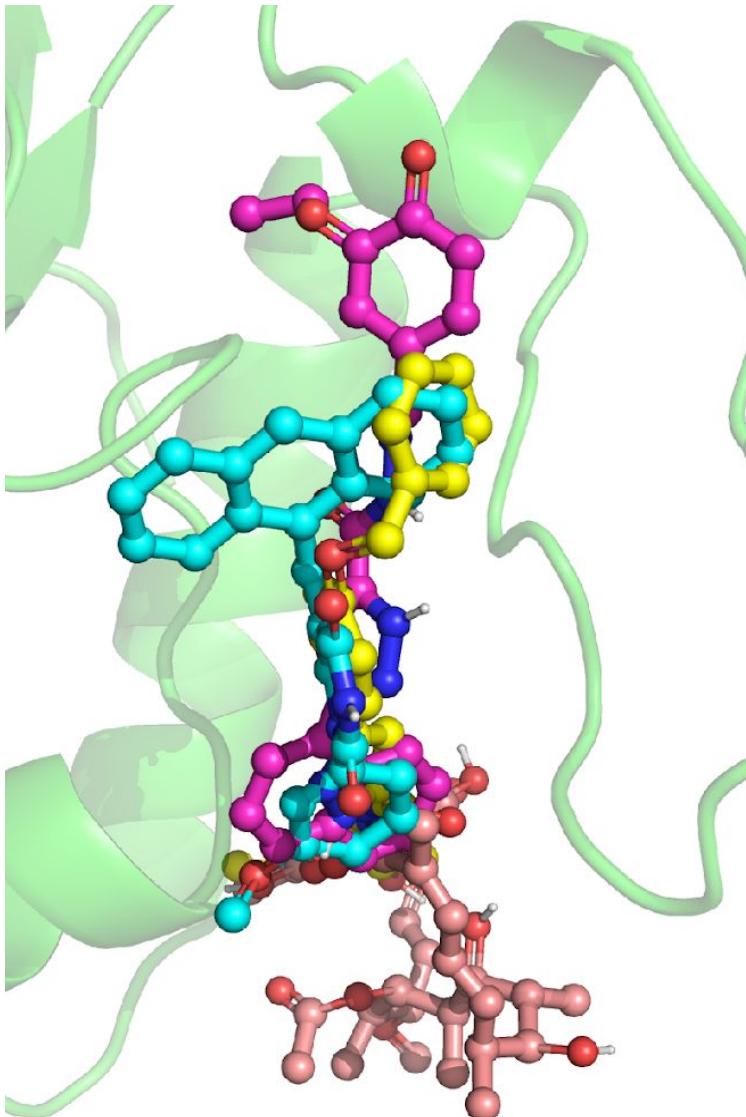


## Interactions

- van der Waals
- Conventional Hydrogen Bond
- Unfavorable Donor-Donor

- Pi-Cation
- Alkyl
- Pi-Alkyl

# Common Molecular Interaction Analysis with *Mtb carbonic anhydrase (mtbca1)*



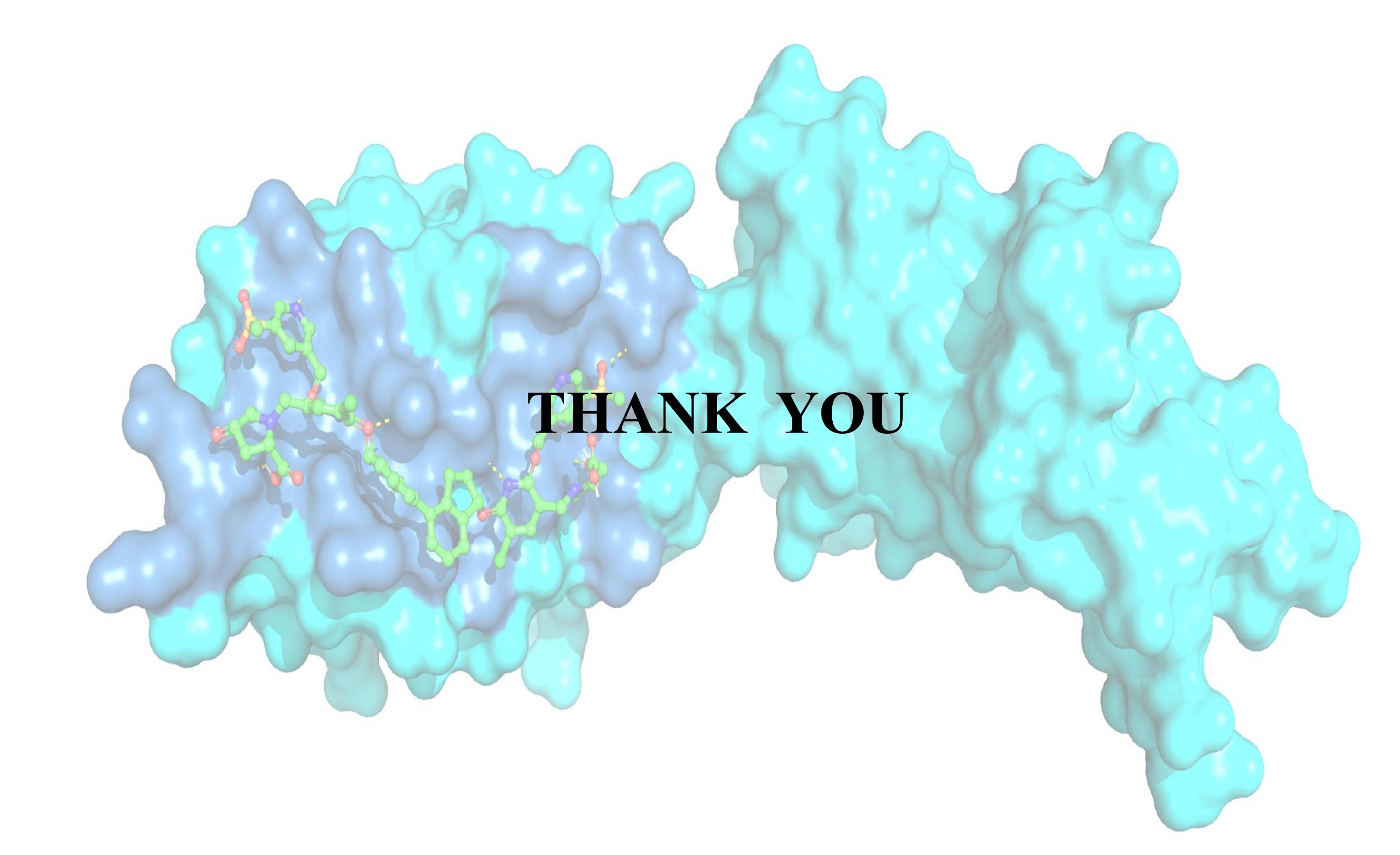
Molecule	Amino Acid Interactions (mtbca1)
<b>F0804-1219</b>	ASP66, <b>ASP65, TRP115, PRO117, MET93, PHE96, PHE101, MET36, ALA59, THR64, SER116, GLU118, ASN58</b>
<b>F1092-1799</b>	<b>THR64, ASP66, ASP65, GLU118, SER116, TRP115, PRO117, PHE101, VAL62, MET93, MET36, ALA59, PHE96, GLY60, GLY92</b>
<b>F1092-1609</b>	PHE96, <b>MET93, PRO117, ALA59, THR64, ASP65, ASP66, VAL63, ARG133, ILE68, ASN137, GLU118, SER116, TRP115, PHE101</b>
<b>Rifamycin</b>	<b>TRP115, SER116, GLU118, ASP65, THR64, VAL63, ARG133, ILE68, VAL136, ASN137</b>

\*Additional common residues between discovered small molecules

\*Common Residues/Proposed Active site for small molecule binding

# Comparative toxicity analysis of discovered *mtbca* inhibitors with respect previously discovered *anti-TB* drugs

Molecule	F0804-1219	F1092-1799	F1092-1609	Rifamycin	Bedaquiline
<b>algae_at</b>	0.0134913	0.00748598	0.00549825	0.000101664	0.00167536
<b>Ames_test</b>	mutagen	mutagen	mutagen	non-mutagen	non-mutagen
<b>Carcino_Mouse</b>	negative	negative	negative	positive	positive
<b>Carcino_Rat</b>	negative	negative	positive	negative	positive
<b>daphnia_at</b>	0.00614039	0.00923821	0.00781777	0.0168088	0.00092812
<b>hERG_inhibition</b>	medium_risk	high_risk	high_risk	ambiguous	low_risk
<b>medaka_at</b>	0.00012039	0.00023018	0.00016632	0.000941901	3.6221E-06
<b>minnow_at</b>	0.00144448	0.00110825	0.0007795	0.00115342	3.1658E-05
<b>TA100_10RLI</b>	positive	negative	negative	negative	negative
<b>TA100_NA</b>	negative	negative	negative	negative	negative
<b>TA1535_10RLI</b>	negative	negative	negative	negative	negative
<b>TA1535_NA</b>	negative	negative	negative	negative	negative



**THANK YOU**