Mechanistic modelling of *Mycobacterium Tuberculosis* beta carbonic anhydrase 3 inhibitors

Ajay Manaithiya[1#], Ratul Bhowmik[1#], Rajarshi Ray[1#], Fabrizio Carta[2], Claudiu T Supuran[2], Seppo Parkkila[1,3], and Ashok Aspatwar[1]

[1]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

ajay.manaithiya@tuni.fi, ratul.bhowmik@tuni.fi, rayrajarshi9@gmail.com, seppo.parkkila@tuni.fi

[2]Department of NEUROFARBA, Section of Pharmaceutical and Nutraceutical Sciences, University of Florence, Firenze, Italy fabrizio.carta@unifi.it, claudiu.supuran@unifi.it

[3]Fimlab Ltd., Tampere University Hospital, Tampere, Finland

Abstract

This study demonstrated a novel bioinformatics pipeline to unravel a mechanistic understanding of the physiological and biochemical mechanism of *Mycobacterium tuberculosis* β-carbonic anhydrase 3, and its role in the key biological processes and metabolic pathways associated with its pathogenesis cycle. We initiated the study by constructing the 3D protein structure of β-CA3 from its sequence with the aid of homology modeling. To thoroughly investigate the essential properties of β-CA3, we further conducted computational genome sequencing, and proteomic analysis to reveal the key biological and structural information. The construction and development of a Python-based custom web server, SEQEclipse (https://seqeclipse.streamlit.app/), which is equipped with integrated work packages, was performed to study and gain insights into the sequence structure-function relationship of β-CA3 through analysis of its physicochemical properties, identification of phosphorylated sites and disordered regions, determination of the functional class of the sequence, BLAST and MSA analysis, determination of the protein-protein interaction mechanism, and development of an ab initio modeling strategy. We also employed a quantitative structural activity relationship (QSAR) model to understand the inhibition mechanism of small molecule modulators of β-CA3 concerning their diverse physiochemical properties. We further extended the QSAR strategy to mechanistic systems biology approach to unravel the key biological processes and metabolic pathways associated with the top β-CA3 modulators and their underlying mechanism of action to halt the pathogenesis cycle of Mtb. Our study outlines the biological significance of β-CA3 functionality and provides a mechanistic understanding of phenotype-based design of small molecule modulators as anti-TB drugs for future prospects.

Keywords: β-CA3, *Mycobacterium Tuberculosis*, SEQ-Eclipse web server, QSAR, System biology

***Corresponding author**
Dr. Ashok Aspatwar, PhD
Email Id: ashok.aspatwar@tuni.fi

#These authors contributed equally: Ajay Manaithiya, Ratul Bhowmik, and Rajarshi Ray

Highlights

- Homology modeling was applied to decode the intricate dual-domain architecture of β-CA3, which is crucial for pioneering targeted therapies against *Mycobacterium tuberculosis*.
- SEQEclipse, a Python-based web server, was developed to centralize bioinformatics analyses, enhancing the drug design workflow.
- Machine learning QSAR models were utilized to pinpoint potent compounds targeting β-CA3, with descriptors such as 'nHeavyAtom' Atsm1, Atsm2, 'apol', 'SubFP287', 'SubFP184', 'SubFP214', and 'SubFP275, showing high correlation and interaction potential.
- Systems biology was used to illustrate how identified active and inactive compounds interfere with essential metabolic pathways in *Mycobacterium tuberculosis*, suggesting novel therapeutic approaches.
- The need for validation of computational findings to confirm the efficacy and safety of identified compounds has been emphasized.

1. Introduction

Tuberculosis (TB) is a highly infectious disease caused by *Mycobacterium tuberculosis* (Mtb) [1]. It has been shown that one out of three people among the global population are infected with Mtb and 5-10% of infected people develop active TB in their lifetime [2]. In terms of localization for Mtb infection, the lungs are mostly targeted, causing a pulmonary infection that may worsen over time. However, TB can also spread to other regions, such as the lymph nodes, urinary tract, and brain. According to the Global Tuberculosis Report 2023 in 2023, the total number of TB cases was approximately around 9.9 million, among which approximately 5.6 million (57%) were male whereas approximately 3.2 million (32%) were female and approximately 1.1 million (11%) were children [3,4][5]. Another highly concerning fact about Mtb infection is that multi-drug resistant (MDR) TB and extended drug-resistant (XDR) TB are difficult-to-treat due to their ability to evade currently used drugs. Thus, to address difficult-to-treat TB, virtual analysis of the target protein using bioinformatics tools analysis of the base sequence needs to be performed to study the internal properties and molecular parameters that can be exploited to respond to drug action and use

cheminformatics to virtually create and screen molecular compounds that can respond positively to Mtb infection [6].

To address the integral properties of Mtb β-CA3, genome sequencing and proteomics approaches were used. These workflows provided a mechanistic understanding of the biological and structural attributes that are fundamental to designing new anti-TB drugs. The use of mechanistic QSAR models, along with bioactivity prediction of small molecule β-CA3 modulators, promises a novel rational for designing anti-TB drugs based on pathways and phenotype associated with Mtb's pathogenesis or survival. Furthermore, an integrated system biology approach in conjunction with QSAR modelling approach also provided the insight about the functionality of β-CA3 functionality, and its role in the pathogenesis of Mtb. Additionally, our system biology methodology also mechanistically outlined the mechanism of action of β-CA3 small molecule modulators and their role in inhibition of β-CA3, which ultimately disrupts the Mtb pathogenesis process [7]. In TB research, *Mycobacterium marinum* (*M. marinum*), a close relative of Mtb and a model bacterium, has been used to study the physiological and biological properties of Mycobacterium species and understand their evolutionary characteristics [8]. This process of gene sequencing coupled with proteomics and transcriptome data can provide information regarding changes in gene expression and protein production levels in response to various molecular compounds, which can be an integral solution for addressing the properties of drug-resistant Mtb [9]. Among the various target proteins that are analyzed in the above mentioned workflows, carbonic anhydrases (CAs, EC 4.2.1.1) are the focus of the study.

CAs are enzymes that are classified into eight various unrelated gene families, namely α-, β-, γ-, δ-, ζ-, η-, θ-, and ι-CAs[10]. CAs are associated with various integral physiological functions within living organisms [4,6,10–19]. First, vertebrates are mostly categorized into the α-CA gene family, which consists of various molecular conformations of the target enzyme[10]. The main biological function of CA includes the reversible hydration of carbon dioxide to bicarbonate and hydrogen ions to maintain the proper pH homeostasis and its involvement in biosynthetic reactions such as gluconeogenesis and ureagenesis in mammal[4,10]. However, various scientific studies have shown that dysregulations or mutations within the enzyme result in the malfunctioning of the target, leading to various diseases, such as glaucoma, epilepsy, and cancer [10]. In addition, in pathogenic bacteria and parasites, CAs play a crucial role in disease pathogenesis and survival [20][21]. Hence, by utilizing CAs as targets and leveraging drug development workflows, various drugs that can counter disease effects can be created for proper disease treatment[14,22]. On the other hand, organisms belonging to bacterial and pathogenic parasitic categories are known to contain the β-CA family, which has been scientifically studied because it is also associated with the progression of various diseases. These organisms are also analyzed using molecular assessment processes, and the obtained information is fed to previously mentioned drug development pipelines to screen molecules or compounds that can have inhibitory effects on disease-causing proteins. Hence, using proper parameters and inhibition constants, compounds can be designed that can be docked onto the target protein to reduce or completely suspend the pathogenic effects when expressed on the host organism. Some of the inhibitors that

have already been analyzed in the available literature and are clinically used are sulfonamides and their bioisosteres [2,5–8].

For this project, the primary research goal was focused on computational modeling of Mtb CAs (especially β-CA3) belonging to the class β-CA gene family. β-CAs are encoded by the genes, Rv1284, Rv3588c, and Rv3273, which are also conventionally labeled β-CA1, β-CA2, and β-CA3, respectively[6,8,14]. The mentioned Mtb β-CAs are enzymes that are more specifically classified as metalloenzymes that contain zinc as a cofactor and have similar biological and physiological properties to many other bacterial β-CAs that contain various conserved amino acid domains that are responsible for catalytic processes, such as the residues that are present as the binding residues for zinc ions: Cys42, Asp44, His97, and Cys101 [6,8]. The abovementioned β-CAs were genetically cloned and purified through analytical techniques, and biokinetic analysis was performed to assess the properties of the proteins. Although only β-CA1 and β-CA2 have been crystallized through X-ray crystallography, structural information about β-CA3 can be derived through homology modeling from the protein sequence of β-CA Can2 from *Cryptococcus neoformans* as a template (PDB code: 2W3N) [4,8,13,14,23–26].

For this project, we implemented advanced computational strategies to model the β-CA3 structure and its biological functions by curating information from NCBI and UniProt databases for performing molecular modeling approaches and to decipher a mechanistic understanding of ontological information about gene and their expression in different biological systems. β-CA3 is known to have a sulfate transporter domain at the N-terminal end along the amino acid residue range (121-414) and a β-CA domain along the amino acid residue range (571-741)[27–29]. This β-CA3 derived from the Rv3273 gene was obtained from the Mtb strain H37Rv, which is a transmembrane protein with a hydrophobic N-terminus [30]. We utilized various bioinformatics analysis tools and servers to compute the physiochemical properties of sample protein sequences from different databases in order to derive evolutionary relationships and phylogenetic understanding through the analysis of protein-protein and protein chemical interactions with respect to β-CA3 protein as the main target sequence derived from *Cryptococcus neoformans* and a sequence obtained from wet lab experimentation.

We also constructed the 3D protein structure of β-CA3 using homology modeling approach and validated it using Ramachandran Plot analysis. Furthermore, a Python-based custom web server, SEQEclipse, was constructed to analyze the sequence structure-function relationship of β-CA3. The server was programmed with basic biopython work packages and was used to study the physicochemical properties of the β-CA3 sequence, identify phosphorylated sites and disordered regions, determine the functional class of the sequence, perform BLAST and MSA analyses, investigate the protein-protein interaction mechanism, and employ an ab initio modeling strategy for β-CA3. We further developed a quantitative structure-activity relationship (QSAR) model to provide a detailed mechanistic understanding of small molecule β-CA3 modulators and their mechanism of action to inhibit the β-CA3 enzyme. The objective of our QSAR model was to identify crucial molecular characteristics that strongly influence the inhibitory mechanism against β-CA3, which might effectively halt the progression of TB and offer useful insights for the design and

discovery of anti-TB drugs. We expanded the QSAR technique to a mechanistic systems biology approach to uncover the fundamental biological processes and metabolic pathways linked to the most important β-CA3 small molecule modulators, as well as their underlying mechanism of action to stop the progression of TB.

Our overall computational study provides a detailed mechanistic understanding of the biological significance and regulatory mechanism of β-CA3, and its significant role in the pathogenesis of Mtb. Our study also outlines the rational design of anti-TB β-CA3 modulators (inhibitors/drugs) through a detail understanding of their mechanism of action concerning their gene-pathway-phenotype interaction.

2. Methods and Materials

We designed the initial computational methodology of the study to decipher a mechanistic understanding of the biochemical and physiological mechanism of β-CA3 and its associated genes, pathways and phenotypes that play an important part in the pathogenesis cycle of Mtb.

2.1. Extracting and Analysing Physicochemical Properties of CA Proteins from UniProt and PDB Using ProtParam

We initially extracted biological data for β-CA3 and its analogous CAs from the UniProt[11] and PDB databases[31] in FASTA format. We initiate our bioinformatics workflow by calculating the various physico-chemical properties of the amino acid samples of the retrieved CA sequences obtained from the databases mentioned previously. Using the Expasy ProtParam Server[32] we further calculated several properties namely number of amino acids present in the retrieved sequences, the quantity of negatively charged amino acids and positively charged amino acids respectively which can elucidate a mechanistic understanding about the overall charge of the protein sample for respective retrieved CA sequences. We also obtained the molecular weight of the entire sequence along with the pH value at isoelectric point (PI) which is an integral factor required during protein quantification through SDS PAGE methods. We also obtained the instability index value which determines the overall stability of the protein sample used which can further determine the structural dynamics of the protein. A value of instability index less than 40 are classified as stable conformation structure of proteins whereas value above 40 are termed as unstable conformations which may require Molecular Dynamics methods to minimise and stabilise it. Moreover, we also get to know the aliphatic index of the protein sequence sample which mainly is defined by the number of aliphatic amino acid side chains present which overall determines the thermostability and stability of the protein structure which can be an integral factor into classifying them into heat labile compounds. We also obtained the value for Grand average of hydropathicity (GRAVY) which determines whether the protein sequence for the CA belongs to hydrophobic or hydrophilic class. A negative range of GRAVY classifies the sequence into hydrophilic class whereas a positive score refers to hydrophobic class. Hence this factor can give an important insight into understanding the bioactivity and pharmokinetic properties of drug molecules when developed against β-CA3.

Analyzing Phosphorylation Sites and Disordered Regions in CA Protein Sequences

Following a detailed comparative analysis for the the physico-chemical data of β-CA3 and its analogous CA sequences, we further studied the phosphorylation sites of the retrieved protein sequences which are mainly occurring during post translational modifications and cellular localisations in addition with the kinase enzymes responsible for each respective phosphorylation. Using the NetPhos 3.1 server[33], we predicted a detailed information about the respective subset of amino acid along with the respective Kinase enzyme which is responsible for phosphorylation of that specific subset of amino acids. The graphical representation provided by the server further provided a mechanistic interpretation regarding the overall phosphorylation potential across the amino acid sequences, where spikes at specific amino acids represent higher intensity of Kinase activity leading to increased phosphorylation activity. Likely, we also analysed the disordered regions of the amino acid sequence of the retrieved CA sequences to find out the unstable regions of the protein which are variable in nature and can be used as potential binding sites for small molecule β-CA3 modulators. Using the IUPred2A server[34], we further obtained the graph consisting of the disorder probability scores of the protein sequence across each amino acid position where higher peaks represent the higher intensity of disorder regions present. This analysis determines the variability of sequences due to unstable conformations and can be studied in detail to gain an understanding of the β-CA3 small molecule modulators binding sites.

Determination of the functional class of the protein sequence

For the following analysis we further used the VICMPred Server[35,36] to determine the biological process and categories namely virulence factors, information molecule, cellular process and metabolism molecule. These factors provide information of the overall activity and processes the amino acid sequence of the specific retrieved CA sequences is responsible for. This server takes in the amino acid sequence in FASTA format and based on its trained model having Support Vector Machine (SVM) algorithm as its backbone returns the classification score based on the probability among the four classes as mentioned. The class with the most score is generally the biological function the amino acid sequence protein sample is responsible for. Virulence factors refer to the involvement of the amino acid sequence in virulence mechanism , Information molecule refers to the protein structure being associated with genetic based information creating a genomic map , Cellular processes categories refer to the respective sequence being involved in cellular localisation processes post translation and Metabolism molecule refers to the protein sequence being involved in metabolic processes or cycles to produce metabolites or bioactive products for maintaining overall biological function of the living system.

2.2. Performing BLAST and sequence alignment to construct phylogenetic trees for evolutionary analysis

For the next analysis we performed the BLASTP from NCBI database [37] on the sample amino acid sequences to find the similar sequences across the databases to perform Multiple Sequence Alignment (MSA) and Phylogenetic Tree analysis to derive evolutionary characteristics of the β-CA3 against the similar proteins across the protein database. We analyzed the similar sequences using MEGA 11[38]

software to execute MSA and phylogenetic tree analysis using the Neighbor Joining method coupled with bootstrap scores to provide confidence scores regarding evolutionary similarity. Bootstrap scores below 60% were filtered out to remove outliers and focus on sequences with evolutionary similarity, facilitating the analysis of their biological activities in a unified manner.

2.3. Analysis of protein-protein interactions and biological activity potential using the STRING server

Protein sequence data were initially retrieved from the UniProt and PDB database, we also curated the official gene symbols associated with the β-CA3 and its analogous CA sequences. Using the STRING Server[39], we further analyzed a biological network of the β-CA3 and its modulators to obtain descriptive information regarding the Biological processes and molecular functions associated with the phenotypic or Gene Ontology(GO) terminologies in addition with the strength score of the processes representing the intensity of biological activity with respect to β-CA3. We also gained detailed insight regarding the metabolic pathway through KEGG databases in which β-CA3 is associated with the production of metabolites in order to maintain the biological balance. Ultimately, we also obtained a mechanistic understanding regarding the protein domains and cellular localisations where β-CA3 is associated with. Overall, the STRING server allows us to perform a mechanistic understanding into the biological network analysis of β-CA3 with other proteins in order to delve more into the functional enrichment analysis involving biological activity and confidence of metabolism action to maintain the biological balance with respect to Mtb's survival.

2.4. Determination of tertiary protein structure through ab initio modeling approaches through the I TASSER Server.

The 3D protein structure for β-CA3 was obtained by performing homology modeling of the β-CA3 sequence comparison across the protein database, where the structure was modeled based on the available information from similar proteins with respect to β-CA3. We further sequenced our sample to obtain the amino acid sequence of β-CA3 protein and performed ab initio modeling to obtain the tertiary structure and conformational features, including the physicochemical and biological information of our protein. The I-TASSER server was further used to analyze the β-CA3 sequence to get a mechanistic sequence-structure function understanding[40–42] along with information regarding the structural information possible for our amino acid sequence to conformationally fold into. The I-TASSER server also provided the essential biological and structural details with graphical representation demonstrating the predicted secondary structure of each amino acid sequence along with the confidence score of that prediction between the range of 0-9, where 0 corresponded to the least confidence factor while 9 corresponded to the highest confidence of prediction of the amino acid being classified into that particular secondary structure. Information regarding solvent accessibility was also accessed from the server with a graphical representation exhibiting confidence scores in range of 0-9, which were assigned with respect to whether the specific amino acid at

the particular position was buried or exposed to the structural surface pertaining to hydrophobicity or hydrophilicity respectively. Solvent accessibility score of 0 corresponds to amino acid being buried or away from the solvent environment whereas higher scores and score 9 corresponds to the amino acid being heavily exposed to the solvent environment. Moreover, a graphical representation providing information regarding the Normalised B Factor was also provided across each amino acid residue positions which further provided information related to thermal stability of protein structure from its native structure. Ultimately, we deduced possible hits for the models of the generated tertiary structure generated from our query sequence along with the PDB codes of the similar matched proteins from which the structural and crystallised information was gathered from to generate the predictive model structure for our query. We also analyzed the TM score and RMSD values from I-TASSER server for our possible hits against our query to select the best model based on the high confidence scores and deviation scores from the provided metrices. Besides this, we also deduced information regarding the possible small molecules β-CA3 modulators (ligands/inhibitors) which can bind with the β-CA3 at specific binding site residues and positions which can be highly utilised for protein-drug docking approaches. Information regarding Gene Ontology (GO) terms connecting to phenotypes namely molecular functions, cellular localisations and biological processes were also deduced, which can be further utilised to perform detailed study on the generated protein functions.

Verifying the protein structure produced through ab initio modeling from the ITASSER server

We further performed validation of the predicted β-CA3 structure to detect the regions of errors and anomalies produced from the ab initio modeling strategy. We further utilized the SAVES v6.0 server and its active tools, like ERRAT[43], for analysis of the errors of the modeled structure against the available information from the biological databases to produce the overall quality of non-bonded atomic interactions in the form of a quality score into a graphical format, VERIFY 3D[44] which compares the 3D modeled sequence against the provided amino acid 1D sequence to check for the overall quality and model factor score considering the structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc) and comparing the results to optimal structures. Other toolkits within SAVES v6.0 included WHAT CHECK and PRO CHECK[45], which were also utilized to check the overall stereochemical properties of the modeled structure against the linear sequence and analyze a residue-residue based geometry to understand the intramolecular interactions, bonds, molecular information, conformational quality, bond lengths, bond angles, and other geometric parameters of β-CA3. Hence this analysis provided us with the errors and validation parameters associated with the virtually modelled sequence of β-CA3 and quality parameterisation of mentioned molecular and atomic characters which are overall associated with the stability and integrity of its structure.

SEQEclipse: Comprehensive Web Server for Bioinformatics Analysis

We also developed a Python programming based web server called SEQeclipse, which utilises the backbone of Streamlit (https://seqeclipse.streamlit.app/), a python library for web development using specific bioinformatics work packages utilizing[46] STRING server API[39], ESMSFold API[47] by Meta and

Py3DMol[48] which is an interactive Python module to incorporate molecular visualisations into our constructed web application. The web application functions by accepting the FASTA file of the amino acid sequence of any protein sample from the user as input. The bioinformatics work packages programmed into the server involve different protein parameter calculations using the Expasy ProtParam tool provided by Biopython which calculates the various physico-chemical properties of the given sequence. Besides this, we also included a BLASTP interface, which will take in input the sequence and provide the user with all the similar sequences across the protein database based on the E Value provided by the user manually. We also included a protein structure visualizer functionality using the PDB module from Biopython and Py3DMol which will take input from the user the amino acid sequence and provide a user-interactive live 3D model for the linear amino acid sequence. We have also utilized the STRING API in our web application. which will enable the user to perform similar operations of STRING into SEQeclipse server to compute the protein-protein or protein-chemical interactions along with the gene ontology and functional enrichment analysis results. Lastly, we have also programmed a homology modeling feature into our web server, which utilizes the ESMFold API functionality to utilize its features where the user will be able to provide a linear sequence of amino acids and our server will return them with the predicted 3D structure of the protein by sequence comparison across the databases. Hence, we present our SEQeclipse server to perform most of the important bioinformatics analysis in a single place in a user-friendly manner.

2.5. Mechanistic QSAR model for the β-CA3 inhibition mechanism

We constructed a QSAR model to elucidate the mechanism underlying the inhibitory effect of small molecule modulators on β-CA3. Our goal through the QSAR model was to identify key molecular signatures that demonstrated direction correlation with the inhibition mechanism against β-CA3, which could further stop the pathogenesis of Mtb, thus providing valuable insights with respect to anti-TB drug design and discovery.

2.5.1. Database selection and screening process

A mechanistic QSAR approach methodology was initiated by retrieving 101 small molecule β-CA3 inhibitors from the ChEMBL database and their biological activity in terms of Ki_{50} value. The principle of uniformity was served using a normalizing procedure, where the original Ki_{50} was transformed to pKi_{50} ($-\log10(pKi_{50})$). Following this, we also utilized the rdkit package to calculate Lipinski's rule of five descriptors, namely molecular weight, LogP, hydrogen bond acceptor, and donor for all the 101 small molecule β-CA3 inhibitors. Consequently, we performed a chemical space analysis to explore the relationships between the Lipinski rule of five descriptors and the pKi_{50} values of the small molecule inhibitors. We also performed clustering of these small molecules based on their Ki50 value. Among the 101 β-CA3 inhibitors, 61 molecules were clustered as active with Ki50 value in the range of 100-1000 nanomolar, 13 molecules were clustered as inactive with Ki50 value greater than 10000 nanomolar, and 27 molecules were clustered as intermediate with Ki50 value ranging in between 1000-10000 nanomolar.

2.5.2. Calculation of chemical descriptors, dataset division, and QSAR model generation

The choice of the right molecular features is one of the crucial steps during the construction of any QSAR model. For the dataset of 101 β-CA3 inhibitors, we utilized the PaDELPy tool to calculate 1D 2D physico-chemical molecular descriptors and binary substructure fingerprints[49]. This process contributed to the generation of 1444 1D 2D physico-chemical molecular descriptors and 307 binary substructure fingerprints. We further performed feature selection strategy to remove the low variance constant and highly correlated molecular descriptors and substructure fingerprints. The final dataset was then divided into training and test subsets in an 80:20 ratio. The employment of machine learning algorithms on these train and test subsets was performed to develop robust QSAR models which can mechanistically derive conceptions and correlation among molecular features (both structural and physico-chemical) and the biological activity (pKi_{50} value).

In the present study, we employed a Random Forest (RF) ML algorithm to generate each of our multi-diverse molecular feature-based ML-QSAR models (substructure fingerprints and 1D 2D physico-chemical molecular descriptors) to correlate classification-based molecular fingerprints and numerical value-assigned phyico-chemical molecular descriptors with biological activity (pKi). Simple decision tree predictors are gathered together to form an ensemble in a supervised machine-learning process called random forest. To create a single model that is more stable and accurate than a collection of individual decision trees, which may not always produce accurate predictions, it is necessary to integrate many decision trees into the model. Notably, RF attempts to address the overfitting problem that decision trees are known to exhibit. Moreover, the bootstrap aggregation or bagging approach is used to train RF. In bagging, training data subsets are randomly sampled (with replacement), a model is fitted to the updated training sets, and the predictions are aggregated.

2.5.3. Performance Evaluation of the ML Algorithm

The accuracy and predictive value of each ML-QSAR model were further assessed by performing both 5-fold and 10-fold cross-validations on the training set. The cross-validation method was performed by splitting the training set into five and ten similar-sized subsets. The model was constructed using a subset of the segments, and the accuracy of the prediction was assessed on the remaining set. This recurrent cycle made every segment serve as the validation set. First, we tested the external dataset before performing the optimal hyperparameterization tuning for all the generated QSAR models in order to meet the generalized requirement. A performance evaluation of the models was performed statistically using statistical metrices such as Q2, Kendall's tau (τ), MALE, MAPE, RMSPE, RMSLE, Spherman's rho, MAE, RMSE, RAE, and RRSE (see equations in the **supplementary file**). The above mentioned key parameter statistical indicators have created a strong foundation for assessing model's efficiency [50–52].

The correlation coefficient (Q) is a measure used to estimate the predictive performance when modeling data for unforeseen conditions. This is an indication of the generalization capability of a model because it can predict outcomes that may lie beyond the domain of training data. A different set of techniques, namely Kendall Tau and Spearman's Rho, tests the determination of how well data points, such as those found in

training sets, were ordered during predictions. Various metrics, including the MAE, RMSE, RAE, and RRSE, which are used to illustrate the size of the errors, have been used to quantitatively measure changes in real property values. For example, the mean absolute logarithmic error (MALE) and root mean squared logarithmic error (RMSLE) may be more appropriate error metrics when dealing with data with a skew or overestimation is more crucial, or when underestimations are to be minimized. Absolute error measures such as MAPE and RMSPE are based on the difference between the measured value and the actual value. These measures show the error magnitudes. These parameters present an exhaustive assessment of model proficiency, hence assisting in obtaining an understanding of the precision and exactness of predictive modelling. Similarly, VIP analysis determines the main driving variables that are connected to each model category and hence, is useful for the interpretation of predictive capability.

2.6. Inhibition mechanism of β-CA3 modulators through a mechanistic systems biology approach

We further extended the QSAR strategy to a mechanistic systems biology approach to unravel the key biological processes and metabolic pathways associated with the top β-CA3 small molecule modulators and their underlying mechanism of action to halt the pathogenesis of Mtb.

2.6.1. Acquisition of chemical compounds and targets

Ten β-CA3 small molecule modulators (5 active and 5 inactive molecules) were subjected to a mechanistic system biology analysis. Predictions of 903 biological targets for the small molecule modulators were done by the Swiss Target Prediction tool (http://www.swisstargetprediction.ch/) (**Table s1**). With a threshold > 0.1 in probability analysis, 284 targets were selected (**Table s1**). All the genes associated with TB were retrieved from GeneCard and OMIM databases. GeneCard (https://www.genecards.org/) showed that there are 3 target genes, while OMIM (https://www.omim.org/) displayed 191 (**Table s2**). Gene targets associated with biofilm formation in Mtb were curated from GeneCard and OMIM databases, revealing 208 and 195 target genes respectively **(Table s3 and Table s4)**. A Venn Diagram analysis of β-CA3 small molecule modulators and target genes was performed to identify common genes. This mechanistic approach analyzed that these 21 genes were conserved among small molecule modulators and the target genes. (**Table s5**).

2.6.2. Gene ontology /KEGG analysis

Functional enrichment analysis, covering biological pathways, molecular functions, cellular components, and biological processes, was conducted using the go:profiler server (https://biit.cs.ut.ee/gprofiler/)[53] and David server (https://david.ncifcrf.gov/home.jsp). KEGG data was obtained from the STRING database (https://string-db.org/). Heatmap analysis was utilized to visualize gene pathways and their associated p-values, using data from the go:profiler server.

2.6.3. Compounds-Targets-Pathways and Protein-Protein Interactions

The Compound-Targets-Pathways (C-T-P) network, which investigates connections among compounds, targets, and pathways, is an important resource that helps in studies regarding the relationships between them. Similarly, the PPI network displays the mechanisms underlying the diseases through the identification of proper targets. Both networks were constructed using Cytoscape 3.10.2 software (https://cytoscape.org//index.html). The degree of the compounds was assessed based on their importance. The main nodes within the PPI network were outlined by using the cytoHubba plugin in Cytoscape 3.10.2.

3. Results and Discussions

The initial part of our research study demonstrates the biological significance and functionality of β-CA3 with respect to the pathogenesis and survival of Mtb. The study utilizes different bioinformatics approaches to unveil the genetic modulators, key biological processes, and the metabolic pathways associated with β-CA3. This study is also aimed to provide a mechanistic understanding of how these pathways and phenotypes associated with the pathogenesis of Mtb concerning genetic regulation of β-CA3 can be used as a muti-prospective strategy to initiate phenotype-based anti-TB drug design. We further hypothesized a phenotypic-based drug design concept through an integrated system biology QSAR approach by correlating molecular signatures and mechanism of action of small molecule modulators with inhibition activity against β-CA3.

3.1. Analysis β-CA3 Physicochemical Properties with ProtParam and SEQEclipse

The physicochemical characteristics of the β-CA3 amino acid sequence revealed several essential features. The sequence was represented by 216 amino acids with 30 negatively charged residues (aspartate and glutamate) and 13 positively charged residues (arginine and lysine). The β-CA3 molecule also demonstrated a molecular weight of 23,302.07 daltons and a theoretical isoelectric point is 4.62. The aliphatic index was shown to be 94.40, which is a structural formula typical of alkanes. There were only 3,233 atoms located within the structure. The measured GRAVY is -0.071, which indicates that this protein has a hydrophilic character in its amino acid sequence. These properties place β-CA3 on the negative side of the protein spectrum, making it of great interest for further investigations of its molecular mechanisms. The results showed that the average PI of the β-CA3 protein is negative because the number of negatively charged amino acid residues is greater than the number of positively charged residues. A stability index of less than 40 and a negative GRAVY score indicate the presence of a hydrophilic protein chain. These findings can be used as a basis for more advanced practical experiments.

Analysis of Phosphorylation and Disorder of β-CA3 Protein

Analysis of the β-CA3 protein sequence using the NetPhos server yielded significant findings. Setting the score filter threshold to 0.5, allowed for the identification of key phosphorylation sites within the protein structure. The output from the NetPhos server was shown graphically, and it revealed that for different regions of the protein sequence, the likelihoods of phosphorylation varied, with high-points on the graph

corresponding to higher probabilities of phosphorylation at specific residues, indicating areas of heavy kinase-activity (**Figure s1**). The highest peaks that cross above the threshold line might indicate those amino acids that are most prone to phosphorylation. The distribution trend of these peaks indicates that β-CA3 may have several regions that are not only functionally important but also may be regulated by phosphorylation. The serine residues, as represented by the blue lines, seem to have a greater number of possible phosphorylation sites than threonine and tyrosine (**Figure 1**). Such a map of phosphorylation potential is essential for revealing the regulatory mechanisms by which this protein functions and highlights molecules for further validation. In addition, these phosphorylation sites might be significant in terms of protein-protein interactions and the overall activity and stability of the protein.

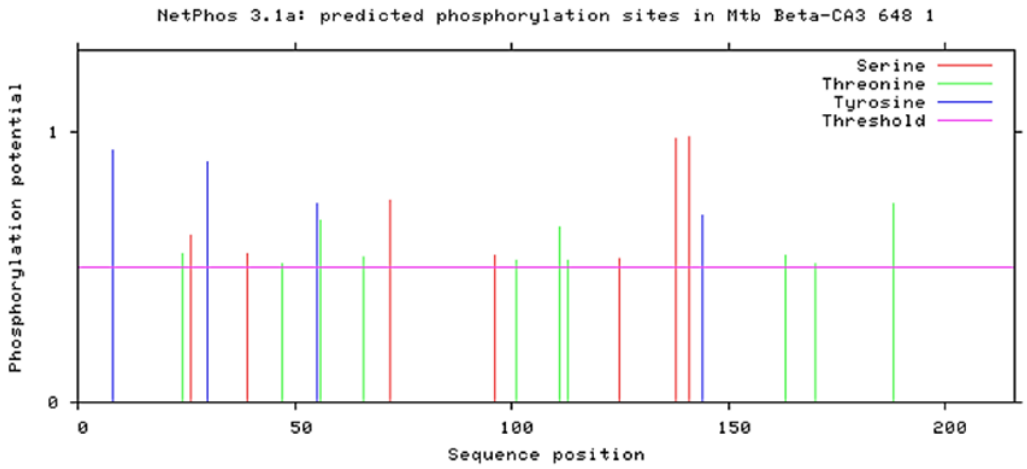


Fig. 1. Graphical representation of phosphorylation potential across specific amino acid residues from the NetPhos server on our analyzed protein sequence of β-CA3.

Through the IUPred2A server, the disordered regions in the β-CA3 protein sequence were identified. The graph shows only disorder and high points for instability. Such an abnormal location of the protein might alter the protein's function and interactions. The regions identified by the ANCHOR2 prediction as red lines represent binding sites that can interact with proteins or signaling sites. The structured areas that were below the threshold are not shown in order (**Figure 2**). The variations in the IUPred2A disorder prediction highlighted by the blue line might indicate that these flexible zones play a vital role in protein activity or control. This disordered pattern suggests that β-CA3 has cellular functions, such as interactions and modifications.

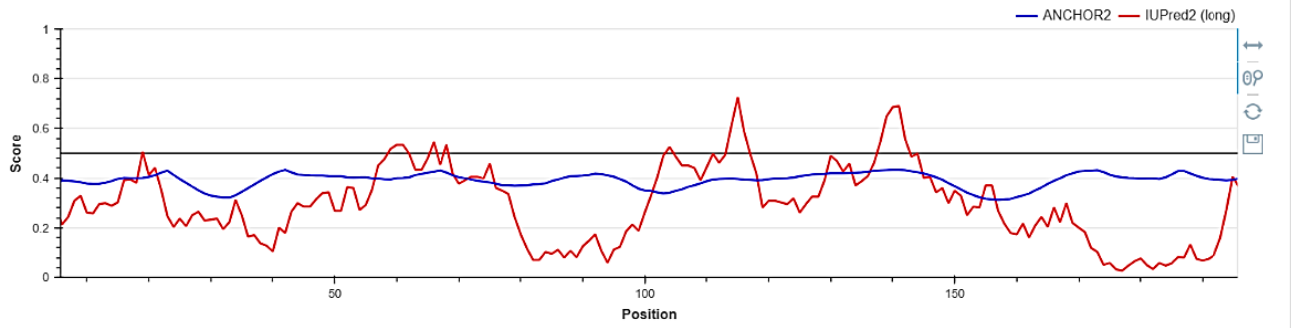


Fig.2. Graphical representation of disordered regions of the analyzed protein sequence of β-CA3 using the IUPred2A server

3.2. Classifying β-CA3 Biological Activity with the VICMPred Server

The β-CA3 sequence, when analyzed by the VICMPred server, showed the best performance among the other classifications in the area of cellular processes. This means that the basic biological influence of β-CA3 might be its cellular function. The protein may function by participating in cell signaling, transport mobility, or other cellular actions. The output of the VICMPred server clearly showed that the β-CA3 protein occupied cell processes according to a score that was much greater than that for the information molecules, metabolism, and virulence factor groups. The server pattern-based prediction method assigned scores to all classes, with cellular processes at 2.6626162, information molecules at 4.6333137, metabolism at 0.82377832, and virulence factors at -1.7326721 (**Figure s2**). Disruption of β-CA3 function may be relevant for cellular processes such as division, differentiation, and response to stimuli. A lower metabolic rate and associations with information molecules by β-CA3 cells mean that the cell does not play a primary role in these functions, but it most likely plays an important role. The negative score of virulence factors shows that β-CA3 is unlikely to be involved in pathogenic mechanisms.

3.3. Phylogenetic insights and evolutionary relationships of β-CA3 from BLAST and MSA

A BLASTP search was carried out to identify proteins with similarity to β-CA3, with 100 similar sequences obtained (**Figure s3**). The sequences were aligned using the MEGA 11 software, which is known to be pivotal prior to phylogenetic tree construction. This multiple sequence alignment revealed regions of conserved and variable sequences among β-CA3 homologs (**Figure s4**). The neighbor-joining method using the phylogenetic tree approach was used to evaluate the evolutionary relationships among β-CA3 strains via bootstrap analysis with a threshold value of 55. The bootstrap scores added a sort of assurance to the connections plotted on the tree. These observations from the phylogenetic analysis will provide insight into which evolutionary trajectories and how the β-CA3 proteins are interrelated (**Figure s5**). This finding indicated significant conservation across various organisms, with the bootstrap scores highlighting the robustness of these evolutionary links. These data are very important for understanding the functional evolution of β- CA3 which has some effects on biological processes in various species.

3.4. Protein-protein interactions and functional analysis of β-CA3 via the STRING server

We used the STRING server to interpret the biological associations of β- CA3 with other Mtb proteins and their significance in the pathogenesis of Mtb. The result of the STRING server analysis showed that the transmembrane function of the β-CA3 protein is one of its key characteristics and that the protein functions as a part of the sulfate transport system. The protein complex was composed of a network of proteins displaying multiple interactions, where the width of connecting lines signified the reliability levels of all the

interactions (**Figure 3 and Figure s6**). The confidence scores ranged from 0 to 1 depicting the chances of interacting biomolecules. Through functional enrichment analysis, it became clear that metabolic processes such as carbon utilization and fatty acid biosynthetic processes participated, which were key biological process for the survival of the Mtb (**Figure s7)**. Molecular functions such as carbonate dehydratase activity and fatty acid synthase activity were found in β-CA3 and are also correlated with the nitrogen metabolism pathway. The GO terms provide detailed explanations of the molecular functions, biological processes, and subcellular localizations, and we can conclude that β-CA3 plays a key role in the cellular activities of the *M. tuberculosis* H37Rv strain (**Table 1)**. In addition, a cross-database search in the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database revealed several features, such as carbonic anhydrase activity and involvement in the polyketide synthase complex, suggesting that β-CA3 may play a promising role in the bacterial metabolism process and may be pathogenic. The fact that β-CA3 is linked to nitrogen metabolism by the KEGG pathway shows the significance of β-CA3 in bacterial survival (**Table 1**). Additionally, the fact that it is located in the PKS complex is indicative of its role in the synthesis of sophisticated organic compounds. In addition, the protein domains were shown to be correlated with the enzymatic activity demonstrated by the bacteria, such as carbonic anhydrase and polyketide synthesis, suggesting their physiological relevance. This comprehensive analysis supports the hypothesis that β-CA3 not only is integral to sulfate transport but also may play diverse roles in bacterial cells, potentially making it a target for therapeutic interventions.

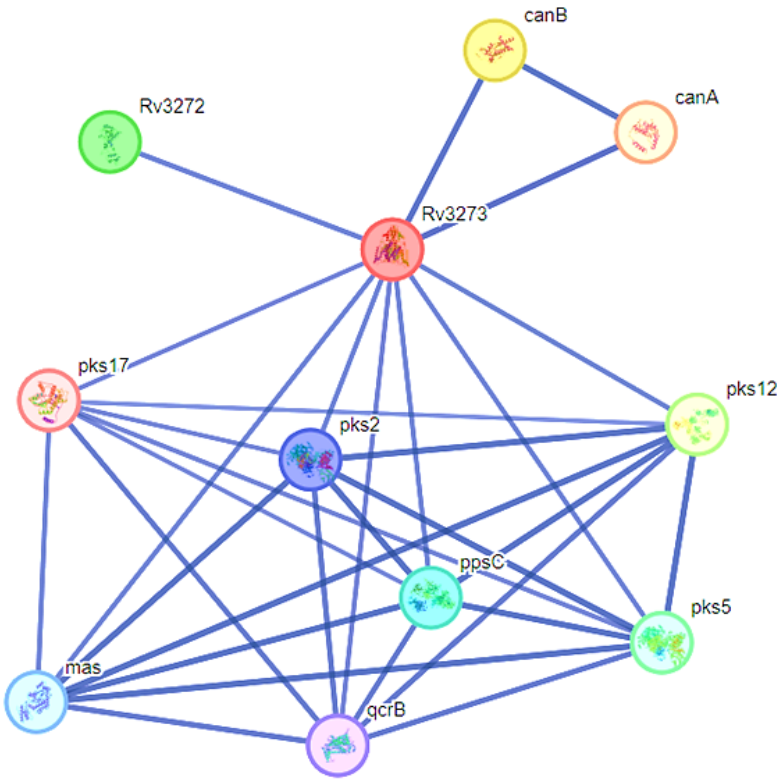


Fig.3. The STRING server biological network showing all the interactions between the β-CA3 protein (denoted by the gene name Rv3273). The thickness of the lines indicates the confidence of interactions.

Table 1: Details regarding GO terms for functional enrichment and cellular localization with metabolic pathways for our query β-CA3 protein (Rv3273).

Category	ID	Description	Count in Network	Strength	False Discovery Rate
Biological Process (GO)	GO:0019776	Carbon utilization	2 of 29	2.86	0.0048
Biological Process (GO)	GO:0005975	Carbohydrate catabolic process	6 of 29	1.58	2.06E-07
Biological Process (GO)	GO:0006633	Fatty acid biosynthetic process	6 of 93	1.37	4.19E-05
Molecular Function (GO)	GO:0004393	Carbonate dehydratase activity	3 of 3	2.56	6.67E-07
Molecular Function (GO)	GO:0043815	[Acyl-carrier-protein] synthase activity	5 of 19	1.98	6.46E-05
Molecular Function (GO)	GO:0031177	Phosphopantetheine binding	6 of 24	1.96	6.92E-08
Molecular Function (GO)	GO:0004312	Fatty acid synthase activity	6 of 31	1.85	8.96E-08
KEGG Pathway	mtv: M00190	Nitrogen metabolism	3 of 22	1.7	0.0004
Subcellular Localization	GOCC:0043081	Polyketide synthase complex	6 of 23	1.98	1.09E-08
Protein Domain (Pfam)	PF00550	Phosphopantetheine attachment site	3 of 15	1.86	0.0068
Protein Domains (InterPro)	IPR036874	Carbonic anhydrase superfamily	3 of 3	2.55	1.22E-05
Protein Domains (InterPro)	IPR001765	Carbonic anhydrase	2 of 3	2.56	0.0024
Protein Domains (InterPro)	IPR018492	Carbonic anhydrase, prokaryotic-like, conserved site	2 of 2	2.56	0.0024
Protein Domains (InterPro)	IPR036398	Polyketide synthase, ketoreductase domain	6 of 13	2.23	3.62E-08
Protein Domains (InterPro)	IPR041204	Polyketide synthase, dehydratase domain superfamily	5 of 11	2.22	0.00039
Protein Domains (SMART)	SM00927	Carbonic anhydrase	5 of 9	2.36	7.26E-06
Protein Domains (SMART)	SM00849	Enolase	3 of 3	2.51	6.19E-09
Protein Domains (SMART)	SM00822	PKS_DH	5 of 11	2.3	1.14E-10
Protein Domains (SMART)	SM00827	Acyl-transferase in polyketide synthase (PKS) enzyme	5 of 17	2.23	4.04E-08
Protein Domains (SMART)	SM00825	Beta-ketoacyl synthase	3 of 11	2	0.00011
Protein Domains (SMART)	SM00823	Phosphopantetheine attachment site	3 of 12	1.96	0.00012

3.5. Inferring the Tertiary Structure of the β-CA3 Protein via I-TASSER Server Predictions

The absence of the crystallized structure of β-CA3 led to the use of the I-TASSER server for ab initio modeling instead. This approach, in turn, resulted in a sequence-based protein model that compared the query protein with a database to find structural and functional analogs (**Table 2**). The server identified the

secondary structure, the solvent accessibility, and the normalized B factor of the amino acid sequence, providing clues on the protein's stability and interaction with soluble molecules (**Figure s8)**. The best hits, having the highest similarity scores, are taken as reference structures illustrating the possible form of β-CA3 and its bioactivity of which has been implicated (**Figure 4**).

Fig.4. The modeled structure of the query β-CA3 protein was determined using ITASSER with a confidence score of 0.79 or 79% with respect to the best similarity hit.

Table 2: Top 10 PDB Structures with the Greatest Structural Similarity to the Query β-CA3 Protein Alongside Their Corresponding Ligands and Binding Sites

Top 10 PDB Structures with the Highest Structural Similarity to the Query β-CA3 Protein					
Rank	**PDB Hit**	**TM-score**	**RMSD[a]**	**IDEN[a]**	**Cov**
1	5wcA	0.954	0.89	0.232	0.977
2	1ekjG	0.858	1.58	0.212	0.903
3	3ucjA	0.835	1.84	0.259	0.912
4	7coiA	0.834	1.71	0.253	0.898
5	4xyA	0.827	1.99	0.24	0.907
6	4oj1A	0.819	1.69	0.262	0.884
7	6gwuA	0.819	1.81	0.233	0.893
8	4o1kA	0.818	1.92	0.246	0.903
9	5cokA	0.809	2.06	0.262	0.898
10	1ddzA	0.806	2.33	0.221	0.907

Corresponding Ligands and Binding Sites for Structurally Similar Hits to β-CA3 Protein					
Rank	**Cluster size**	**PDB Hit**	**Lig Name**	**Download Complex**	**Ligand Binding Site Residues**
1	0.36	58	3e1wZNP	Rep. Mult	36,38,94,97
2	0.28	33	5xch1KT	Rep. Mult	33,35,41,42,43,46,58,100
3	0.27	127	1ekjG	ACT	36,38,60,97,98,99
4	0.05	9	3ucjA	AZM	27,55,79
5	0.05	9	3eqdP	SO4	3,84,04,19,41,84,186

3.6. Validation and Quality Control of the β-CA3 Protein Structural Model Using SAVES v6.0

The assessment of the modeled structure of the β-CA3 protein is important to avoid inevitable mistakes and determination of accuracy. As presented in **Fig. 5**, ERRAT server analysis revealed a high-quality predicted structure, yielding a 94.63% accuracy score. Fortunately, a few density areas colored red in the graphical representation surpassed the 99% rejection level, so these regions pointed at the local model errors. In contrast to that, **Fig. 5's** VERIFY 3D, plot shows the protein structure where the model ranked low against the protein sequence. This shows that these results may not be representative of reality and, thus, require more advanced refinements, such as molecular dynamics simulations or experiments in which the compounds crystallize. This approach is intended to resolve the discrepancies and contribute to their stabilization. The responses from the WHATCHECK and PROCHECK tools, which are positive and constructive regarding structure. The presence of unwanted deviations which may be due to poor bond formation or bond interactions between atoms emphasizes the need for more optimization routines (**Figure s9**). The VERIFY program assesses a 3D structure (the atomic model) against its 1D sequence of amino acids, and it breaks down the score of each residue into the environmental factors of the molecule. A score of 0.02 and above corresponds to an environment where the residue experiences a normal environment, whereas scores below the threshold represent probable problems.

ERRAT helps analyze the results of nonpolar interaction prediction between different atoms in the protein model. The graph shows the error value corresponding to each residue window that has a certain number of residue spans. This graph shows that the majority of the structures are below the commonly used 95% confidence level, which suggests that there is an acceptable level of quality in this structure. On the other hand, the distribution is asymmetric, indicated by the region in red, which is the most significant at 99% being hostile. Thus, one may infer that these regions do not fall within the expected range of nonbonded interactions using a dataset of trustable structures derived from experimental observations (**Figure 5)**. The high peak in the graph indicates a potential error hotspot that could be due to incorrect folding or an error in modeling. The trend across the VERIFY 3D graph on the sequence reveals that some β-CA3 protein model parts do not have the optimum conformation **(Figure 5)**. Those areas for which the average score is below the threshold line are investigated more closely to determine whether they are true structural features or modeling errors. The Ramachandran plot shows the phi (φ)and psi (ψ) torsion angles for the amino acid sequence in the protein structure. Ideally, most of the residues should lie within a sufficient region (accepted as dark red), while only a few of them are allowed (shown as lighter yellow) and none are allowed in the disallowed region. The main Ramachandran plot shows that most of the described residues fall into the favored regions, indicating a good protein structure model (**Figure 5**). The existence of any residue in disallowed regions is an indication of the need to reevaluate the model or the regions of this protein that are flexible or disordered in vivo.

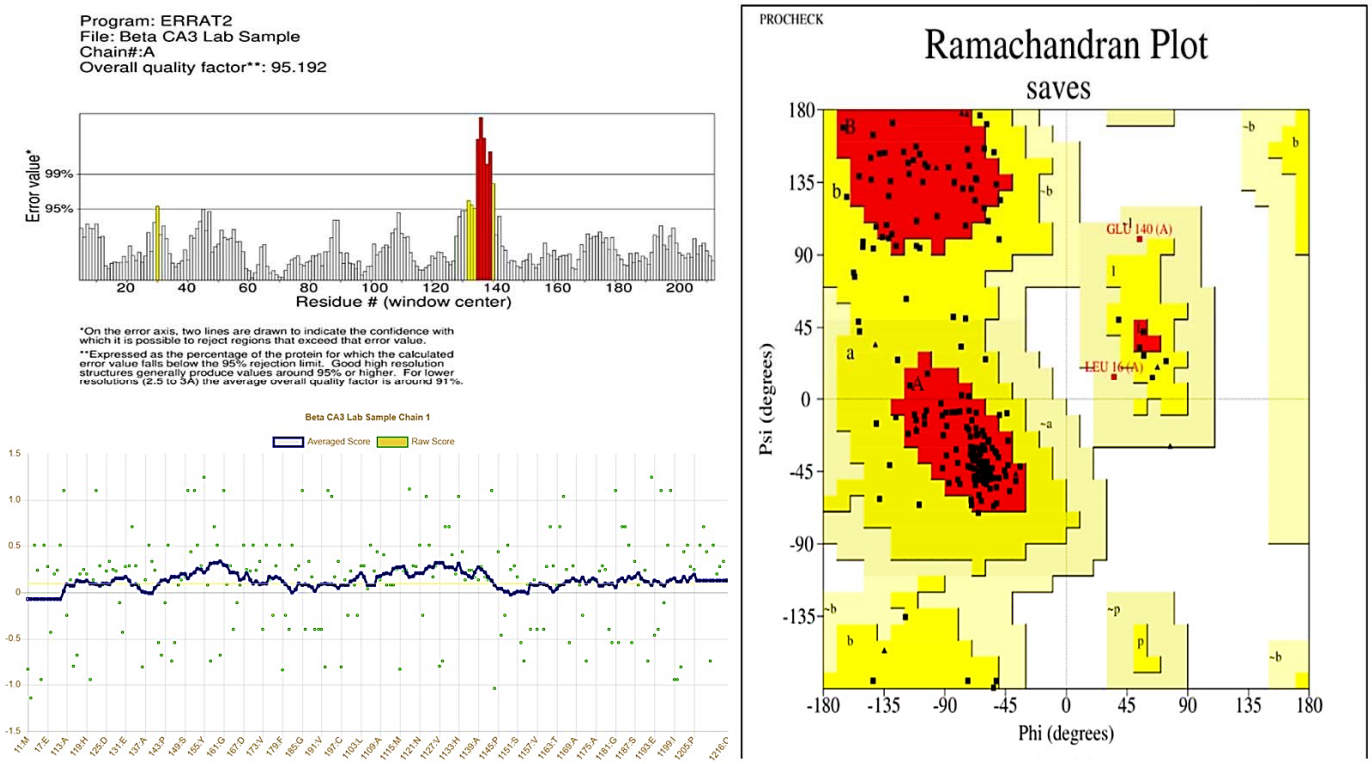


Fig.5. Validation of the Modeled β-CA3 Protein with Computational Tools; The top shows the ERRAT analysis of structural errors, the bottom panel shows VERIFY 3D sequence alignment, and the right panel shows the Ramachandran plot from PROCHECK.

Figure 5 shows certain positive feedback regarding the overall structural and functional information of the modeled structure from the ITASSER server. On the other hand, there are certain anomalies and negative feedback which are also due to the poorly optimized structure as previously mentioned. The anomalies are caused by improper interatomic interactions and bond formations which can be fixed using stabilization approaches.

3.7. From Sequence to Function: Deciphering the β-CA3 Protein through SEQ Eclipse

The SEQ Eclipse web server (https://seqeclipse.streamlit.app/), custom-developed and accessible at the provided link, offers a unified platform for conducting bioinformatics analyses, and was used to examine the β-CA3 protein sequence (**Figure 6**). With SEQ Eclipse, the β-CA3 protein sequence was solved providing information that revealed its biological and structural features. However, the specific outputs of the Supplementary Figures (**Figures s10-s15**) were not directly viewed, hence their representation of the server's ability to define the protein's physicochemical characteristics, its interactions with other proteins and genes, and its role in cellular processes are assumed. Our study included a thorough bioinformatics analysis of β-CA3, providing valuable insight into intricate issues, such as physicochemical parameters, biological functions, and molecular interactions. Such in silico analyses are laying the road for elucidating the protein behavior and its physiological mechanism in cellular conditions.

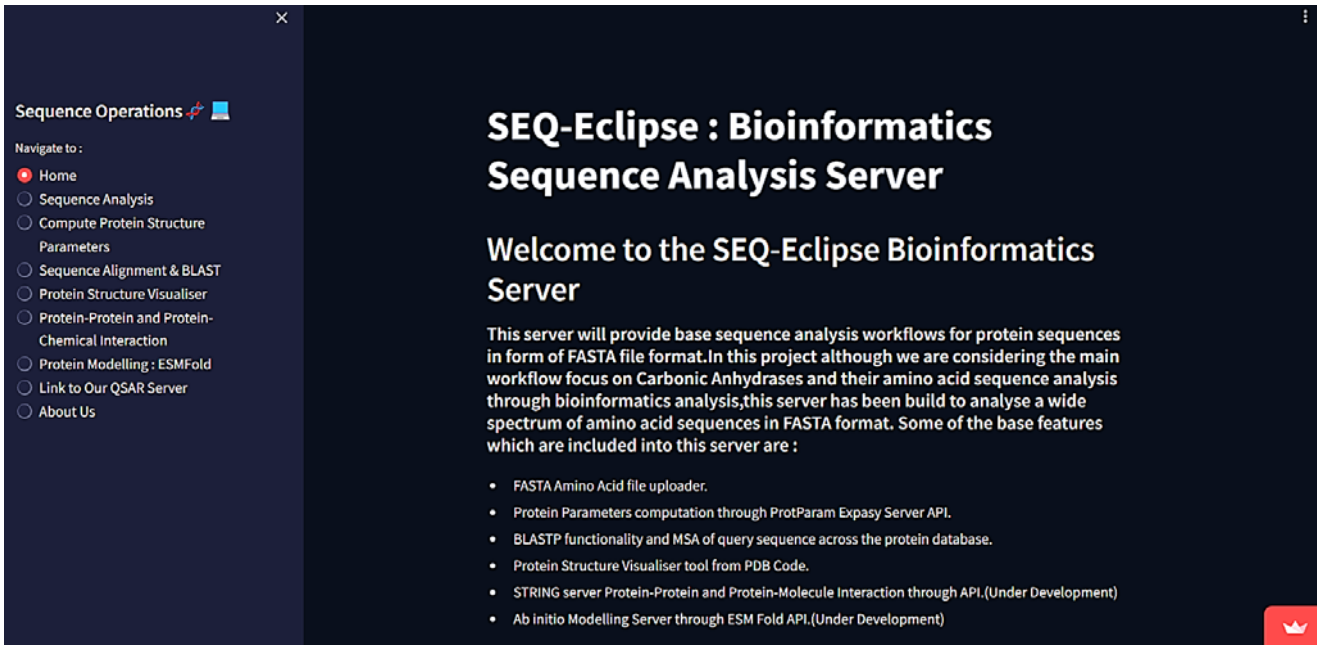


Fig.6. Home Page of SEQ Eclipse.

Hence, in conclusion, we present a detailed bioinformatics analysis of β-CA3 to determine its important physicochemical properties, biological functions and molecular interactions with a multitude of other genes and proteins that are overall responsible for maintaining its metabolism and functionality, which are of significant importance in regards to the pathogenesis of Mtb.

3.8. Machine learning QSAR model prediction

We assessed machine learning models to identify and filter out any unusual data points. In a model utilizing random forests to predict chemical fingerprint statistics, we employed principal component analysis (PCA) to detect and eliminate structurally irrelevant molecules. Correlation coefficients and other statistical parameters were computed for each fingerprint dataset (**Table s1**). Classification analysis plots were generated to visualize the predictive performance of the machine learning models. For example, for the 1D/2D descriptor model and substructure fingerprint model, the plots showed strong agreement between the predicted and experimental pKi_{50} values. Most data points were on top of the identity line, which means that the models have very robust predictive ability. In the theoretical molecular 1D and 2D descriptor datasets, the training set correlation coefficient was approximately 97.90%, while for the test set, it was 89.40%. The root mean squared error (RMSE) for the training set was approximately 21.40%, and for the test set, it was 31.60%. Additionally, 5-fold and 10-fold cross-validations of the training set data yielded correlation coefficients of 75.09% and 73.40%, respectively. Similarly, in the substructure fingerprint dataset, the training set exhibited a correlation coefficient of approximately 96.40%, while for the test set, it was 79.40%. The RMSE for the training set was around 24.40%, and for the test set, it was approximately

coefficients of 77.40% and 77.70%, respectively (**Table 3**). Based on the correlation coefficient between the training and test sets, we inferred the high reliability of the QSAR models. Furthermore, the 10-fold cross-validation results for both models were exceptionally satisfactory. Principal component analysis (PCA) was employed to identify and eliminate outliers from the dataset. PCA revealed certain compounds positioned far from the main cluster of data points, indicating potential outliers with atypical features not representative of the dataset (**Figures 7 and 9**).

Table 3: Statistical Evaluation and Cross-Validation of Multi-Feature ML-QSAR Models.

Metric	1D/2D Descriptor		Substructure Fingerprint	
	Training	**Test**	**Training**	**Test**
Correlation coefficient	0.979	0.894	0.964	0.794
Kendall's tau	0.812	0.424	0.786	0.389
Mean absolute logarithmic error	0.023	0.03	0.023	0.04
Mean absolute percentage error	0.027	0.034	0.027	0.045
Root mean square logarithmic error	0.031	0.04	0.033	0.057
Root mean square percentage error	0.038	0.045	0.038	0.063
Spearman's rho	0.951	0.566	0.934	0.438
Mean absolute error	0.161	0.233	0.173	0.308
Root mean squared error	0.214	0.316	0.24	0.442
Relative absolute error	23.34%	32.04%	25.63%	45.60%
Root relative squared error	24.68%	39.13%	28.59%	57.55%
Total Number of Instances	56	12	55	11
Cross-Validation statistical evaluation of Multi-Feature ML-QSAR				
Metric	**10 Folds CV**	**5 Folds CV**	**10 Folds CV**	**5 Folds CV**
Correlation coefficient	0.734	0.759	0.777	0.774
Kendall's tau	0.507	0.554	0.557	0.56
Mean absolute logarithmic error	0.062	0.061	0.053	0.054
Mean absolute percentage error	0.074	0.074	0.062	0.063
Root mean square logarithmic error	0.085	0.084	0.072	0.073
Root mean square percentage error	0.105	0.105	0.085	0.085
Spearman's rho	0.688	0.739	0.759	0.779
Mean absolute error	0.439	0.433	0.392	0.397
Root mean squared error	0.591	0.578	0.531	0.535
Relative absolute error	62.31%	61.78%	57.35%	58.15%
Root relative squared error	66.63%	65.68%	62.31%	62.68%
Total Number of Instances	56	56	55	55

A study of the QSAR for each model was undertaken to determine the significance of each descriptor and fingerprint and to identify how they influence bioactivity. Correlation matrices with both 1D/2D property representations and substructure fingerprints were used to determine the interdependencies among molecular datasets. Some descriptors displayed strong positive correlations, indicating potential collinearity issues, while others showed negative correlations or minimal linear relationships. This knowledge is essential for understanding the influence of descriptors and optimizing the feature set to improve model performance.

The RF regressor approach was utilized to identify the top 10 molecular fingerprints of the QSAR models. Based on feature engineering analysis using VIP plots for both models, a comparative study was conducted based on the structural insights of four molecules with the best activity for both models, compared to well-known approved anti-bacterial inhibitors, for proper interpretation of QSAR models.

In the feature plot of selected active and inactive compounds for 1D and 2D descriptor fingerprint-based models, 'ATS2m' and 'nAromBond' emerged as the most significant molecular fingerprints, while 'AMR', 'apol', 'natom', and 'nHeavyAtom' demonstrated notable contributions to the QSAR model for both active and inactive compounds. Additionally, in the top 10 VIP plots, overall variance importance plots (VIPs) for the QSAR database revealed good correlation of 1D and 2D descriptors. Descriptors such as 'nHeavyAtom', 'ATS2m', 'apol', 'ALogp2', 'nc', 'ALogP', 'nAtom', 'ATS1m', 'nAromBond', and 'AMR' exhibited the best variance importance plots and significantly contributed to the QSAR model (**Figure 8**).

Among these three features, ATS1m', AMR, nHeavyAtom, and apol were deemed crucial with significant importance, while ATS2m and nAromBond also made noteworthy contributions. We further scrutinized the conservation of these significant molecular fingerprints among three active and three inactive molecules in the dataset. The active molecules, namely CHEMBL6633, CHEMBL467897, and CHEMBL77517, targeted β-CA3, exhibiting experimental and predicted pKi_{50} values of 7.48, 7.42, and 7.82, respectively. They demonstrated the presence of 'ATS1m', AMR, 'nHeavyAtom', and apol. Conversely, the inactive molecules, CHEMBL277836, CHEMBL14060, and CHEMBL379206, displayed predicted and experimental pKi_{50} values of 4.51, 4.10, and 4.90, respectively, attributed to the absence of ATS1m, apol, and AMR, specifically in terms of atomic mass and atomic polarizability. The molecular weight can be varied by increasing or decreasing the number of heavy atoms (in addition to hydrogen atoms) in a molecule thus, the size, complexity, and interactions with biological targets might also be different. Metabolites with a greater number of heavy elements can show different structural recognitions and thus various biological activities. The molar refractivity (AMR) characterizes aqueous molecules' overall polarizability and size, potentially linking compounds with formally similar AMR values with similar biological activities, resulting in an interaction with the same target. By using autocorrelation descriptors (ATS1m and ATS2m), one can analyze the spatial arrangements of atoms in a manner that provides better insights into the association of molecules with biological targets and interactions at receptor sites. The number of C atoms (nC) influences the hydrophobicity, shape, flexibility, and capability of a molecule to bind to specific biological entities. The LogP determines the distribution coefficient between n-octanol and water, which parallels the lipophilicity of a chemical that helps to increase membrane permeability and ability of the compound to interact with biological targets. ALogP2 is a measure accounting for nonlinearity between lipophilicity and biological activity, and is particularly useful in structure-activity modeling. The atomic polarizability (apol) can enhance dipole-dipole interactions with targets, which can result in stronger interactions. The number of atoms (nAtom) affects molecule's size, shape, and complexity, which in turn affects the ligand's binding affinity. The presence of a double bond between two atoms indicates aromaticity (nAromBond). It is

associated with bioactive molecules that are capable of forming interactions with aromatic residues in the binding pocket, which in turn affect the activity (**Figure 8)**.

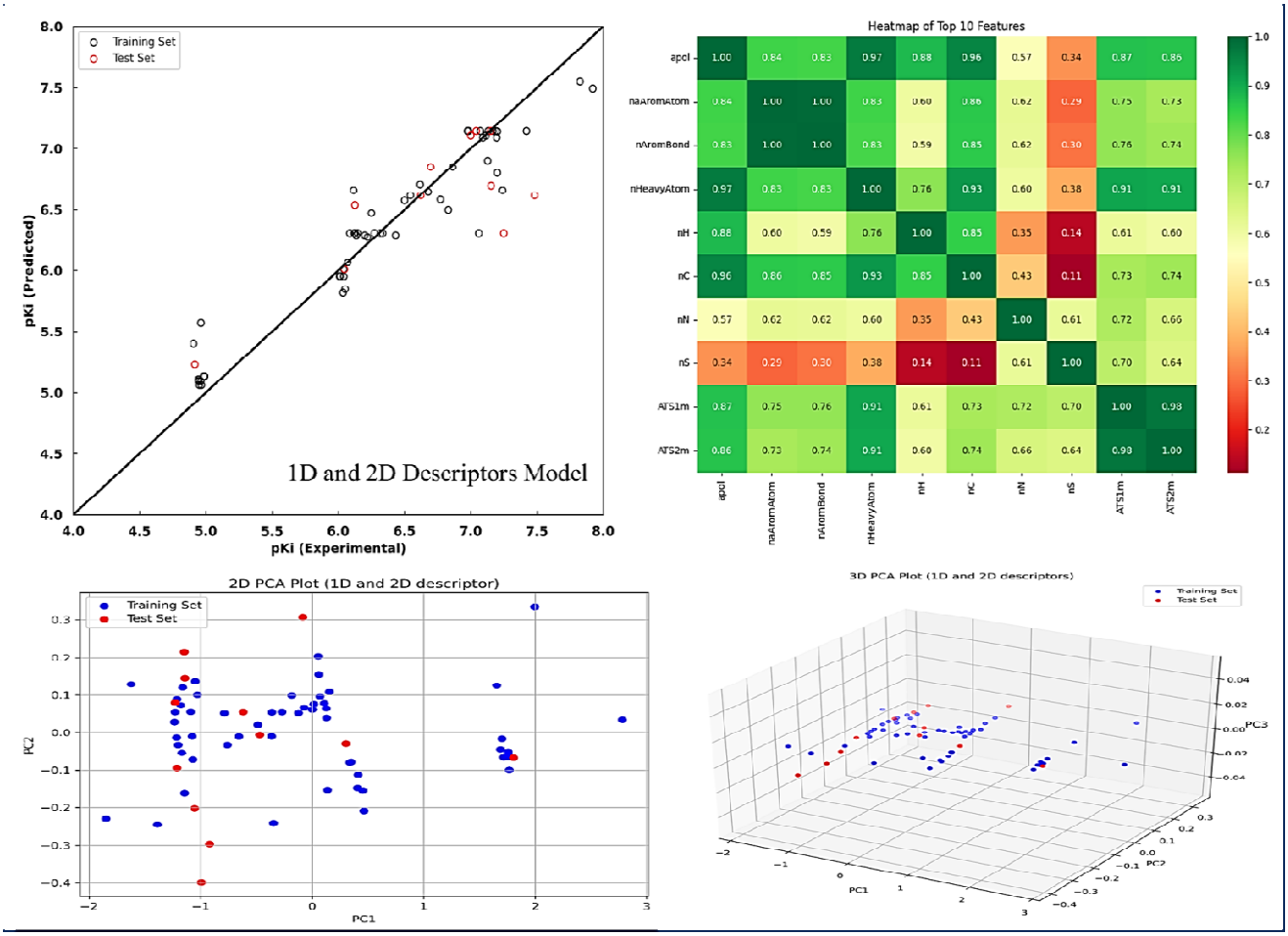
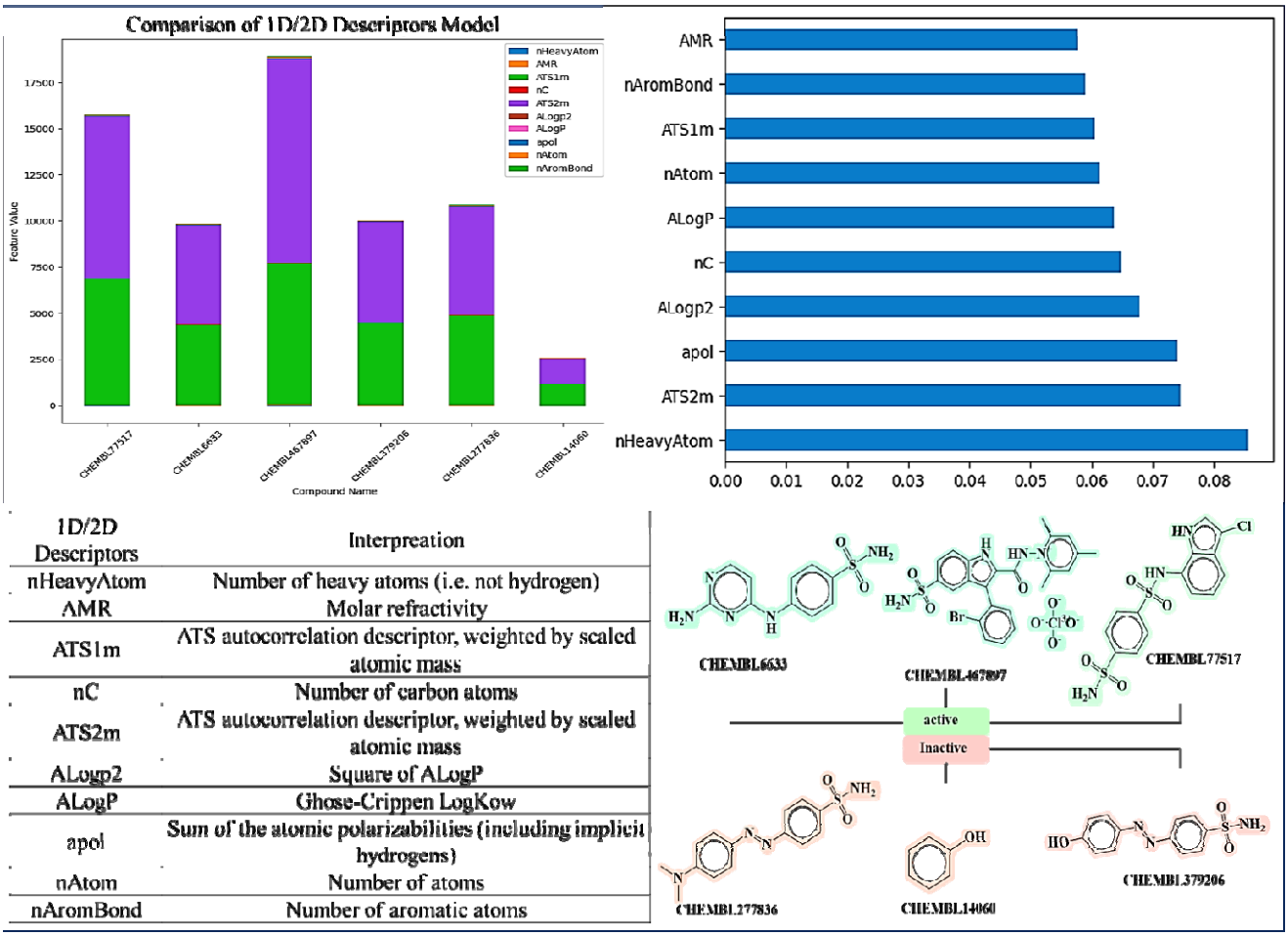


Figure 7: The upper left panel depicts regression plots for various ML-QSAR predictive models. The upper right panel displays a heatmap correlation matrix for 1D/2D descriptors, while the lower left and right panels present 2D and 3D PCA plots, respectively, visualizing the data distribution of 1D/2D descriptors.



1D/2D Descriptors	Interpreation
nHeavyAtom	Number of heavy atoms (i.e. not hydrogen)
AMR	Molar refractivity
ATS1m	ATS autocorrelation descriptor, weighted by scaled atomic mass
nC	Number of carbon atoms
ATS2m	ATS autocorrelation descriptor, weighted by scaled atomic mass
ALogp2	Square of ALogP
ALogP	Ghose-Crippen LogKow
apol	Sum of the atomic polarizabilities (including implicit hydrogens)
nAtom	Number of atoms
nAromBond	Number of aromatic atoms

Figure 8: Comparison of the structures of the top six CheMBL molecules with highlighted feature values (upper left) and variable importance in projection (VIP) plots (upper right). The lower right section displays the extraction of CheMBL compounds from the QSAR model set using 1D/2D fingerprints, while the table in the lower left presents interpretations of the top ten descriptors for the QSAR model.

On the other hand, for the substructure fingerprint-based model, the feature values of the selected active and inactive compounds identified 'SubFP287' and 'SubFP275' as the most significant features, along with other substructure fingerprints such as 'SubFP214', 'SubFP88', and 'SubFP137'. Furthermore, 'SubFP169' potentially decreased the feature value of selected inactive compounds, and common substructures such as 'SubFP214', 'SubFP137', 'SubFP287', and 'SubFP287' displayed fluctuations in features for the QSAR model dataset of selected compounds. Additionally, in the top 10 VIP plots, 'SubFP296' emerged as the most crucial contributor, while 'SubFP298' and 'SubFP179' demonstrated significant importance, along with other substructure fingerprints such as 'SubFP184', 'SubFP181', and 'SubFP137', which exhibited average importance in QSAR design. 'SubFP275', 'SubFP2', 'SubFP88', and 'SubFP1' were identified as compounds of lower importance for the QSAR model dataset (**Figure 10)**. The substructure fingerprints provide valuable descriptors for use in the design of QSAR models. They provide the means for specific chemical features to be observed within the molecules, implying the identification of the structures whose motifs play

underscores their influence on both reactivity and biological activity, while SubFP169's detection of phenol moieties highlights their well-known antioxidant and antimicrobial properties. Similarly, SubFP184 contains heteroaromatic rings, with unique chemical and biological characteristics, and SubFP214 contains of sulfonic acid derivatives that affect key physicochemical properties. Additionally, the detection of heterocyclic rings by SubFP275 underscores its prevalence among pharmacologically active compounds. SubFP287's ability recognize conjugated double bonds and SubFP137's ability to recognize vinylogous ester moieties provide further insights into its molecular structure and potential reactivity. Among the analyzed substructure fingerprints, 'SubFP287', 'SubFP184', 'SubFP214', and 'SubFP275' were identified as pivotal, demonstrating significant importance in the analysis. Additionally, 'SubFP88' and 'SubFP137' also exhibited noteworthy contributions. Further analysis revealed the presence of these significant molecular fingerprints among both active and inactive molecules within our dataset. The active molecules, namely CHEMBL512633, CHEMBL20, and CHEMBL77517, which target β-CA3, showcasing experimental and predicted pKi_{50} values of 8.14, 7.92, and 7.82, respectively, and exhibit the presence of 'SubFP287', 'SubFP184', 'SubFP214', and 'SubFP275'. Conversely, the inactive molecules, CHEMBL592384, CHEMBL53566, and CHEMBL379206, displayed predicted and experimental pKi_{50} values of 4.94, 4.91, and 4.90, respectively, attributed to the presence of 'SubFP169', specifically in terms of the phenolic group **(Figure 10)**.

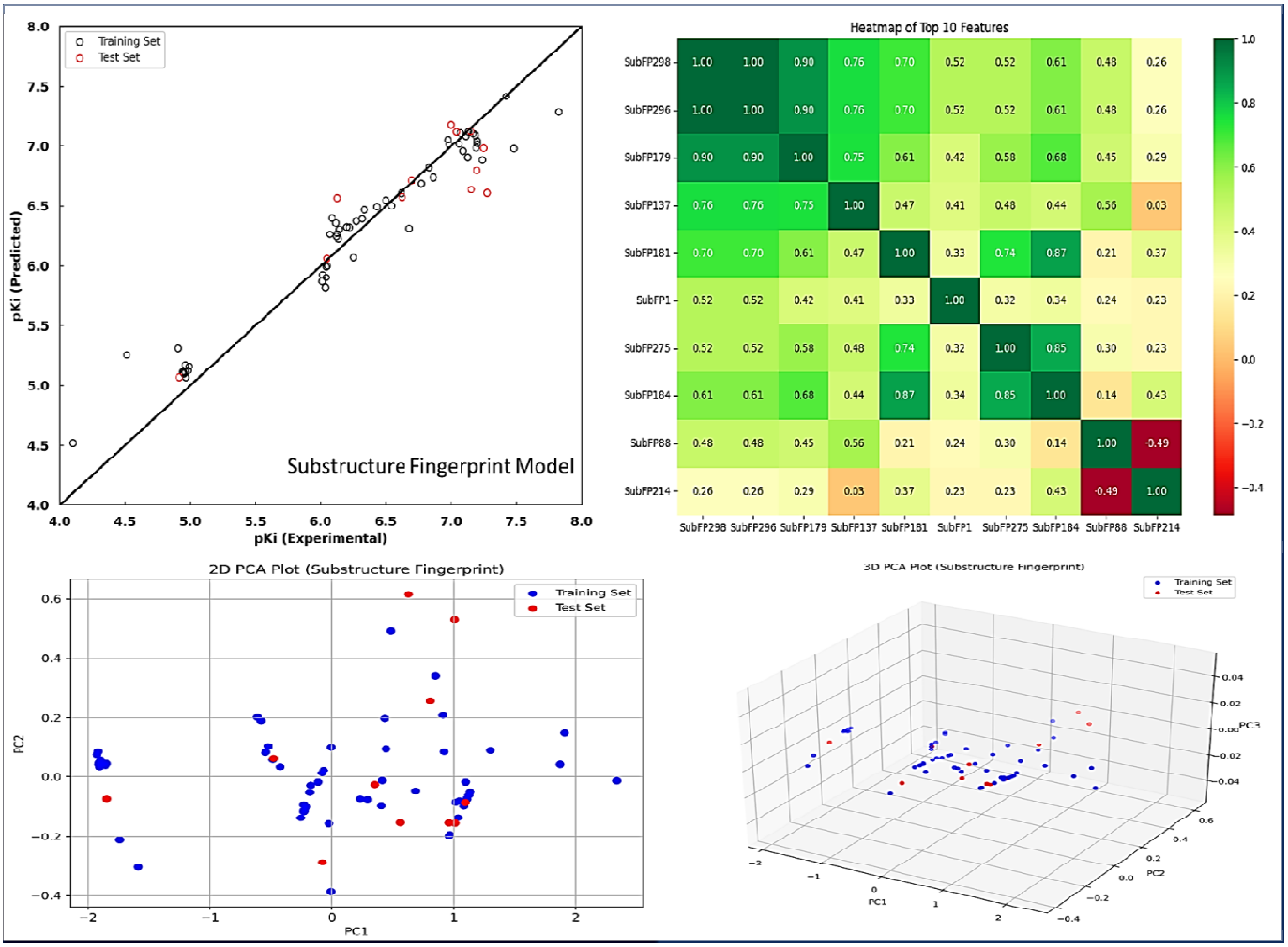
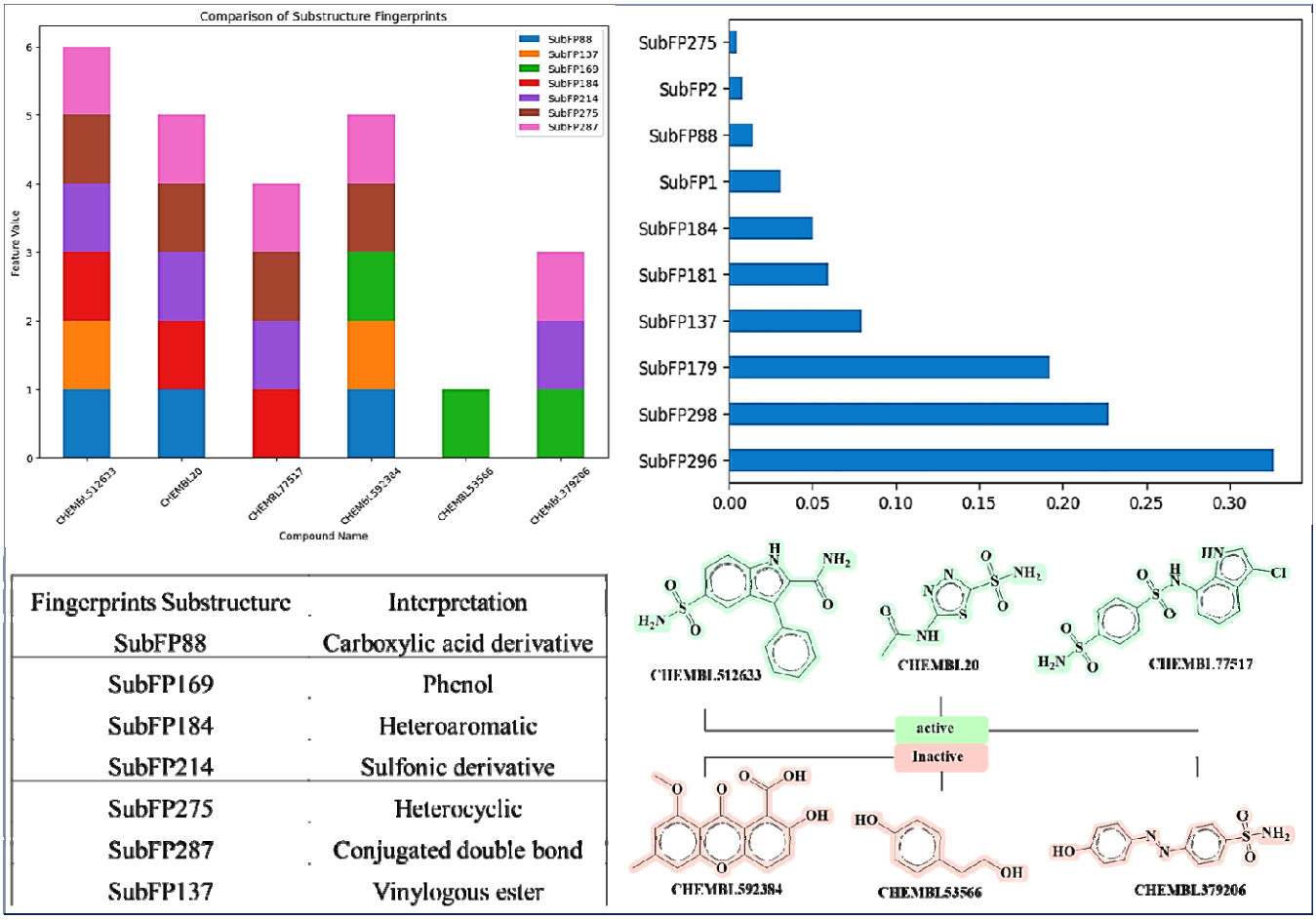


Figure 9: The upper left panel depicts regression plots for various ML-QSAR predictive models. The upper right panel displays a heatmap correlation matrix for substructure fingerprints, while the lower left and right panels present 2D and 3D PCA plots, respectively, visualizing the data distribution for substructure fingerprints.



Fingerprints Substructure	Interpretation
SubFP88	Carboxylic acid derivative
SubFP169	Phenol
SubFP184	Heteroaromatic
SubFP214	Sulfonic derivative
SubFP275	Heterocyclic
SubFP287	Conjugated double bond
SubFP137	Vinylogous ester

Figure 10: Comparison of the structures of the top six ChEMBL molecules with highlighted feature values (upper left) and variable importance in the projection (VIP) plot (upper right). The lower right section displays the extraction of ChEMBL compounds from the QSAR model set using substructure fingerprints, while the table in the lower left presents interpretations of the top ten descriptors for the QSAR model.

3.9. Mechanistic System Biology Approach

3.9.1. Phenotype analysis of β-CA3 and its small molecule modulators

The Gene Ontology (GO) study provides detailed information about the biological processes, cellular components, and molecular functions of Mtb, while also providing false discovery rate (FDR) scores. These FDR scores specifically highlight the most profoundly enriched GO terms, and highlight which processes probably have a substantial effect in regard to their implications for the biology of this species. Biological reactions, such as fatty acid and lipid biosynthetic pathways, were significantly enriched with FDR values of $1.73 * 10^{-23}$ and $1.06 * 10^{-15}$, respectively, which indicate the high activity of these processes in the organism. This phenomenon might be important for cell wall synthesis and membrane regulation and is tolerated by Mtb under stress conditions. Other highly represented processes include DIM/DIP cell wall layer assembly and respiration under aerobic conditions, most likely because of the preservation of cell integrity and

and the plasma membrane respiratory chain complex 1 with low FDR scores means that those processes are important regardless of the outcomes of the two tests. In terms of molecular function, among several other functions, phosphopantetheine binding, and 3-oxoacyl-[acyl-carrier-protein] synthase activities are important, and interestingly, they are among the most enriched activities, showing how Mtb relies on a complex network of biochemical interactions for its metabolic functions (**Figure 11**).

A functional protein interaction network (PPI) model was constructed using the STRING database and the Rv3273 UniProt ID (**Figure s16**). Twenty highly connected 20 genes at the center of metabolic functions of Mtb have been extracted from the network (**Figure 11**). Among the most pertinent pathways highlighted by the next KEGG pathway analysis were the oxidation of carbohydrates and their subsequent use in energy generation during cellular respiration, the synthesis and secretion of siderophore in heavy metal availability regulation, and the metabolism of nitrogenous compounds. Oxidative phosphorylation was strongly associated with 14 genes, with a p-value 8.59×10^{-11}, which is supported by a small (0.05) false discovery rate (FDR) score, thus highlighting the fundamental role of this pathway in bacterial energy generation. Regardless of the fact that it is not directly shown in the diagram, β-CA3 is regarded as the most significant factor responsible for maintaining the cellular pH of a given bacterium. β-CA3 may control over these pathways by playing a regulatory role. Another example is enrichment of the siderophore peptide biosynthesis pathway, which could indirectly indicate the involvement of the β-CA3 enzyme in iron metabolism while not being a part of the siderophore biosynthesis pathway. Moreover, 'metabolic pathways' remained as one of the six annotations where 26 genes from the PPI network were based, which suggests a strong implication of the impact of metabolic systems impact on Mtb. The nitrogen metabolism pathway, though second only to oxidative phosphorylation in terms of its extent, still plays an important role in the body's metabolism. This fact shows that the gene is actively associated with an assimilation and utilization process for nitrogen, thus digitalizing infection and survival in the new host. The relationship between the CO_2 regulating function of β-CA3 and complex metabolic processes not only clarifies the need for additional experimental studies, but also helps to explore what regulatory mechanisms and roles β-CA3 may be responsible for within that system. This information is particularly important for understanding the metabolism of Mtb, which is significantly dependent on β-CA3, and could be exploited for anti-TB strategy development.

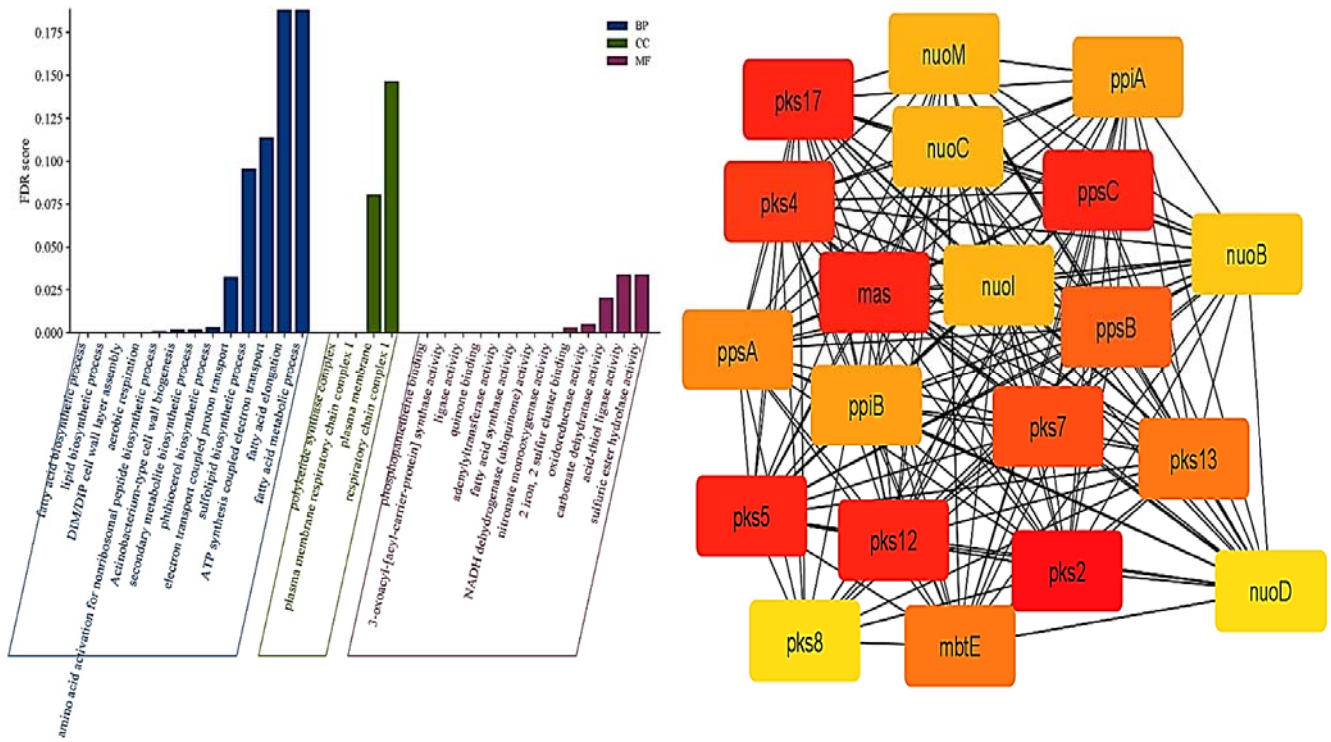


Fig. 11: Gene Ontology analysis of *Mycobacterium tuberculosis* β-CA3; right panel: common target gene visualization.

3.9.2. Identification of Hub genes

We curated a set of 21 common genes from both β-CA3 small molecule modulators and target genes specific to β-CA3 in Mtb. Additionally, a network diagram was constructed to visualize the complex interplay between these genes. Notably, emphasis was placed on the top 5 genes CA2, CA5A, CA9, CA12, and CA14 as hub genes due to their substantial connectivity (**Figure 12**), indicating their potential pivotal role in modulating the pathogenesis of Mtb. Four genes, namely NOS2, CTSG, ABCB1, and JAK1, were identified as common targets among β-CA3 small modulators, biofilm formation, and other interactive genes associated with β-CA3 in Mtb (**Figure s17 and Table s6**).

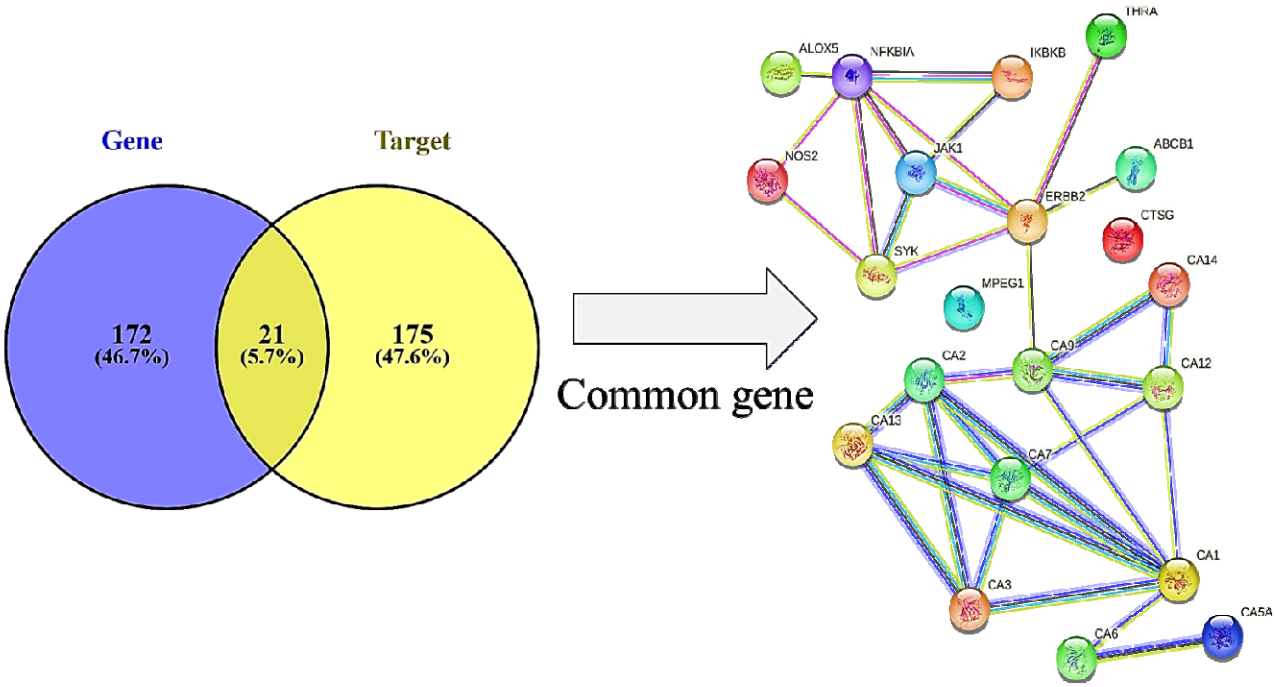


Fig. 12. The right panel shows a network of common genes, while the left panel displays a Venn plot diagram highlighting overlapping targets between the potential compounds and disease-related genes.

3.9.3. Gene Ontology and KEGG Analyses

We screened of 21 genes through GO and KEGG enrichment analyses, and ranked them by ascending p-value. KEGG analysis revealed 81 significantly enriched pathways ($P<0.05$), aiming to uncover the mechanisms underlying the inhibitory effects of potential small molecule modulators on the β-CA3. A functional enrichment analysis was carried out on the top 21 target genes using g:profiler. Gene Ontology analysis highlighted the critical role of β-CA3 in the survival and virulence of Mtb, indicated by the significant increase in carbonate dehydratase activity (GO:0004089), with a profound adjusted p-value of 1.069×10^{-25} **(Table 4)**. This is one of the essential processes that Mtb uses for pH regulation by reversibly absorbing carbon dioxide because of the behavior of the host cell environment. Such an exceptionally low p-value reflects a very high level of statistical confidence, indicated that this enzyme's activity is a potential focal point for targeted drug intervention. The transition metal ion binding function (GO:0046914) with a padj of 4.816 x 10^-8, which is also relevant to the function of β-CA3 as a metalloenzyme, implies that this activity is significantly associated with the survival of M.tb, highlighting the importance of metal ions in the catalytic activity of β-CA3 (**Figure 13** and **Figure s18**) (**Table 4**). Genes associated with carbonate hydratase activity such as CA1, CA2, CA3, and CA9 others, are essential. Additionally, genes involved in transition metal ion binding (GO:0046914), including ALOX5 and THRA were significantly enriched (padj of 4.82E-08), emphasizing the importance of metalloenzymes and zinc-binding proteins such as β-CA3 in Mtb physiology, which are crucial for its enzymatic functions. β-CA3 regulation is directly connected to the

acidic environment of macrophages. When the padj value is comparatively higher, it still represents a biologically meaningful enrichment that could influence the bacterium's ability to maintain pH homeostasis.

The cytosolic localization (GO:0005829) of β-CA3, supported by genes such as JAK1, IKBKB, SYK, ALOX5, CTSG, NFKBIA, NOS2, ERBB2, and THRA, is integral to the intracellular survival of Mtb, emphasizing the need to target this enzyme within its cellular milieu to effectively inhibit its function. This cellular component's significance, with an adjusted p-value of 0.001619, suggested a strong association with the bacterium's metabolic processes. This finding reinforces the targeting of the enzyme in its native cellular environment to disrupt its function effectively. The identification of genes in the "myelin sheath" (GO:0043209) also raises considerations for neurological implications and the host's immune interactions.

KEGG pathway analysis revealed enrichment of the metabolic pathways (KEGG:01100), further highlighting the vulnerability of Mtb's metabolic network, which could be disrupted to curtail the energy production and biosynthesis capabilities of the bacterium. Furthermore, processes such as the one-carbon metabolic process (GO:0006730) and nitrogen metabolism (KEGG:00910), with exceedingly low padj values of 5.523 x 10^-20 and 4.257 x 10^-22 respectively, are crucial for the growth and function of Mtb. One-carbon metabolism is essential for the synthesis of nucleic acids and amino acids, and the nitrogen metabolism pathway is vital for the biosynthesis of the building blocks required by the bacterium. The significance of the carbonate dehydratase activity associated with β-CA3 cannot be overstated, as indicated by the extremely low padj value. This finding suggests the strong potential of the for selective inhibition of β-CA3 as a strategy for antitubercular therapy. The association with the most enriched metabolic pathways further highlights the importance of this target, as disruptions to these pathways could critically affect the survival and virulence of Mtb. The selective targeting β-CA3 has a dual benefit: it not only does it directly disturbs the essential enzymatic activity of Mtb, but also indirectly disturbs its metabolic flexibility. By further elucidating the function of β-CA3 in these essential pathways we will be able to devise a drug development strategy that will be more effective at treating tuberculosis.

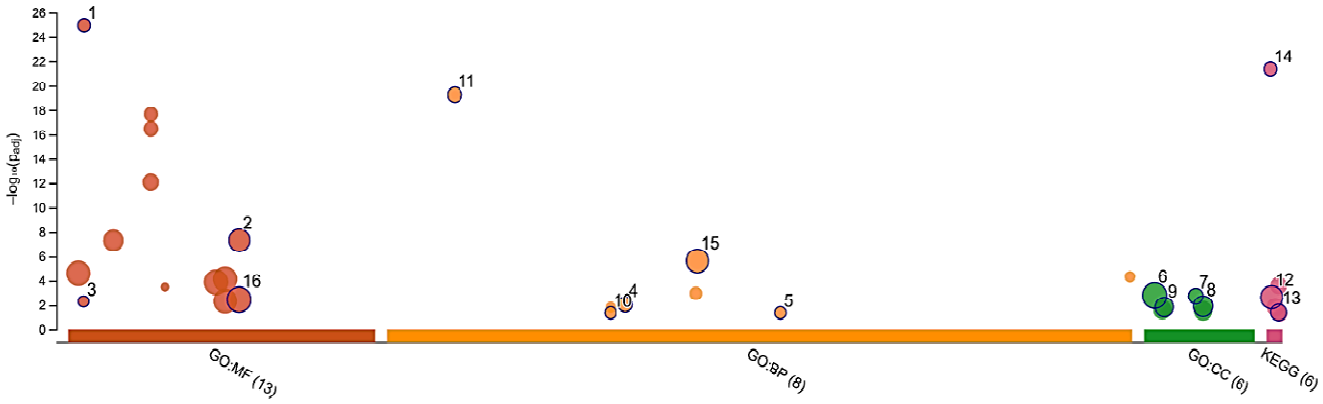


Fig.13. Gene Ontology Categorization and Pathway Enrichment Scatterplot.

Table 4: Analysis of Gene Ontology and KEGG Pathways for Common Genes

ID	Source	Term ID	Term Name	padj (query_1)
1	GO:MF	GO:0046914	transition metal ion binding	$4.82 * 10^{-08}$
2	GO:MF	GO:0004064	arylesterase activity	0.004868
3	GO:BP	GO:2001225	regulation of chloride transport	$4.68 * 10^{-05}$
4	GO:BP	GO:0034110	regulation of homotypic cell-cell adhesion	0.008596
5	GO:BP	GO:0032849	positive regulation of cellular pH reduction	0.01481
6	GO:BP	GO:0051712	positive regulation of killing of cells of another organism	0.0414
7	GO:CC	GO:0005829	cytosol	0.001619
8	GO:CC	GO:0043209	myelin sheath	0.001789
9	GO:CC	GO:0045177	apical part of the cell	0.01179
10	GO:CC	GO:0016323	basolateral plasma membrane	0.01394
11	GO:BP	GO:0006730	one-carbon metabolic process	$5.52 * 10^{-20}$
12	KEGG	KEGG:00910	Nitrogen metabolism	$4.26 * 10^{-22}$
13	GO:MF	GO:0004089	carbonate dehydratase activity	$1.07 * 10^{-25}$
14	GO:MF	GO:0016836	hydro-lyase activity	$2.19 * 10^{-18}$
15	GO:MF	GO:0016835	carbon-oxygen lyase activity	$3.30 * 10^{-17}$

3.9.4. Compound-target-pathways analysis

We addressed the relationships between β-CA3 small molecule modulators, targets, and pathways that we constructed with Cytoscape 3.10.2 on a network basis. As **Figure 14a** suggests, the C-T-P network reflects the C-T interactions between pathways. The complexity of the compound-gene-pathway network underscores the mechanism of action against β-CA3, which may introduce different therapeutic interventions. The tool Cytoscape permitted the design of a detailed diagram to display the mechanisms, including binding, especially between drug candidates (β-CA3 small molecule modulators) from the ChEMBL repository and their targeting sites within the TB bacteria. This extensive map concentrates on analyzing the mechanistic connections of these small molecule modulators with β-CA3). The map had 42

how they, all works together. The network's density score is equal to 0.0, meaning that it has few point-to-point connections and assumes that some connections are more preferred than others. The unique shapes and colors used in the network illustrate different types of biological components: molecules presented as octagonal, experimentally determined genes are circles with different numbers of connections shown in yellow, and the other genes are circles with brown colors; shapes of triangles representing biofilm formation genes in tuberculosis and biological processes are rectangles with pink. The clustering coefficient is 0.197. In practical, while there are some clusters where the points are more closely connected, the network does not have tight clusters. Network analysis revealed a recursive structure in the form of four distinct groups or connected components, suggesting that the network consists of isolated clusters, in which only close contact occurs. There is typically an average of seven nodes connected to each node. The structure of the network is thus of a great interactive nature. Connected with an average shortest path length of 1.538 and illustrated by a network radius of 1 and diameter of 4, that topology shows the system to be an intimate one where all the components are in close proximity. Interestingly, however, there are only 273 direct paths, which amount to 15% of the total possibilities of direct connections, suggesting that nodes are indirectly connected via intermediaries rather than direct routes. The compliance of all the nodes to each other that is, the uniform connectivity implied by a homogeneity score of 1, could mean that each node has a degree of influence that is the same as that of the other nodes, which can explain a balanced network where no single node is too powerful in the network. In addition, multiple connections exist between 10 of these node pairs, which could mean that some interactions are notably stronger or perhaps quite important among others for network integrity. **Figure 14b** reveals the top 10 most interactive pathways within the network, showing more genetic connections with the targets and pathways than with any other route. These paths are crucial and will be the focus of further study.

The output C-T-P network was not only in the static form but also dynamic model that can represent the characteristics of multiplexing and multi targets directly and could be applied to systems biology (**Figure 14a**). The network could reveal the mechanisms of action of these compounds, showing how they may lead those compounds to exhibit polypharmacology. Thus, these molecules could interfere with the activity of more than one target at the same time, which could enhance their bioactivity against the β-carbonic anhydrases-3. Furthermore, the degrees of freedom of each common target are shown in the **Table s7**.

The compounds CHEMBL77517, CHEMBL512633, and CHEMBL379206 had significant interactions in the network, each targeting different genes frequently linked to vital metabolic and immune pathways, which can be exploited in the discovery of efficient drugs for Mtb particularly those targeting β-CAs and its associated genetic regulators. CHEMBL77517 has been proven to interact with CA2, NFKBIA, CA9, ERBB2, and CA12. This compound might inhibit the activity of the bacterial carbonic anhydrase (CA2 and CA9) and become less effective in maintaining the ion balance and pH homeostasis inside the bacterial cell, which might cause the homeostatic state of the bacterium to be disrupted (**Figure 14a**). In this respect, the influence of NFKBIA has attracted special interest because this gene participates part in regulating inflammatory reactions [54]. CHEMBL77517 might possibly tap the role of NFKBIA resulting in the host's

ability to evade Mtb. ERBB2, is not the main element of the Mtb mechanism of action, but it can influence cell signaling, in this way, host communication can change. Like CA9 and CA2, CA12 has the potential to play a role in pH control and ion movement. The compound CHEMBL512633 performs its actions by targeting several genes, including SYK, which is involved in the control of signaling from immune cells as well as biofilm formation, which is important because Mtb is known to create biofilms during infections. The targeting of CA2 and CA9 by CHEMBL512633, similar to CHEMBL77517, points to potential disruption of the bacterium's metabolic processes, while its interaction with CA12 may further affect carbonic anhydrase activity. **CHEMBL379206** interacts with genes such as CA2, ERBB2, CA9, and CA12. The double interaction with ERBB2 suggests a potential off-target effect or a yet unidentified role of ERBB2 in the disease process. Like for the other two compounds, the targeting of CA2 and CA9 by CHEMBL379206 indicates possible interference with the bacterial carbonic anhydrase system. Interactions with CA12 could have an additional impact on the bacterial metabolism.

The metabolic pathway highlighted in the graph is directly connected to several genes: These three regions are: CA2, CA9, and CA12. The genes interact with the molecules CHEMBL77517, CHEMBL512633, and CHEMBL379206 in particular, which a network of therapeutic targets (**Figure 14b)**. These genes, which encode carbonic anhydrases, play a pivotal role in pH regulation and are essential for the metabolic functioning of Mtb. The interactions of the CA2 and CA9 genes with metabolic pathways suggest a direct influence on the ability of Mtb to regulate its internal environment. Compounds targeting these genes could impede the bacterium’s metabolic reactions, potentially leading to a weakened state that could increase susceptibility to treatment. As a result, the state of the bacterium may be weakened increasing its vulnerability to treatment. Therefore, compounds affecting CA12 could have indirect effects on metabolism of Mtb. Given that the SYK pathway is not a redundant pathway, targeting this pathway could reinforce the susceptible immune responses mediated by Mtb, which, in the end, may be involved in controlling infection. Since the SYK pathway is nonredundant, targeting it could strengthen the immune system's specific responses against Mtb, potentially leading to better control of the infection. SYK plays a role in the formation of biofilms, which are complex communities of microorganisms that adhere to surfaces and can be resistant to antibiotic treatment [55,56]. Biofilms are significant in the context of tuberculosis because Mtb can reside within biofilms in the human body, contributing to the chronic nature of the disease and its resistance to treatment [57,58]. The network constructed and confirmed that the SYK gene has complex interactions with the chemical CHEMBL512633 (**Figure 14b)**. This adjacent process implies that the compound can be efficient in switching on and off SYK, which eventually results in a reduction in biofilm formation and increase in the immune response. If CHEMBL512633 can modulate SYK activity, it may affect the formation and maintenance of biofilms. Disrupting biofilms could expose Mtb to the immune system and antibiotics, potentially making infections easier to clear. The interpretation of the interactions within the network provides insight into the multifaceted way that these compounds could combat Mtb. Through the selective targeting of genes involved in metabolic control and the immune response, these compounds might represent a strategy that not only inhibits bacterial activities but also supports the body’s

innate mechanisms to eliminate infection. The network shows that inhibiting these genes would disturb the metabolic stability of Mtb thereby limiting its potential to infect the host and survive.

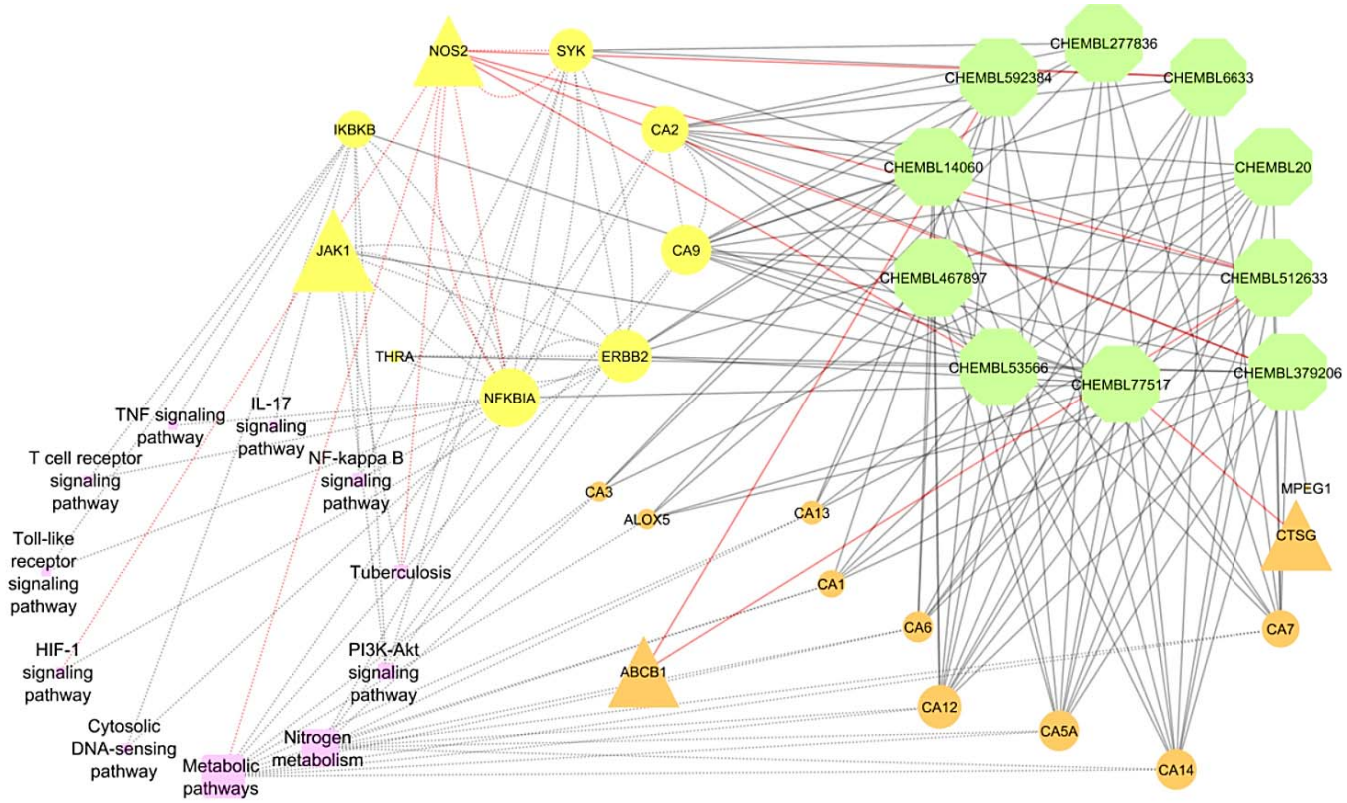


Figure 14a: Interactive network map displaying gene and compound associations with key biological pathways. Highlighted yellow colors show an experimentally determined gene brown color shows no experimentally determined gene, an octagon shape of green color shows the compound, and a square of pink color shows pathways interacting with the compound and genes with the highest connectivity within the network. The dashed line shows the pathway interconnected with the gene and target, and the solid line shows that the gene and target are interconnected. The yellow triangle color indicates a gene associated with biofilm formation associated gene that is connected to the target gene.

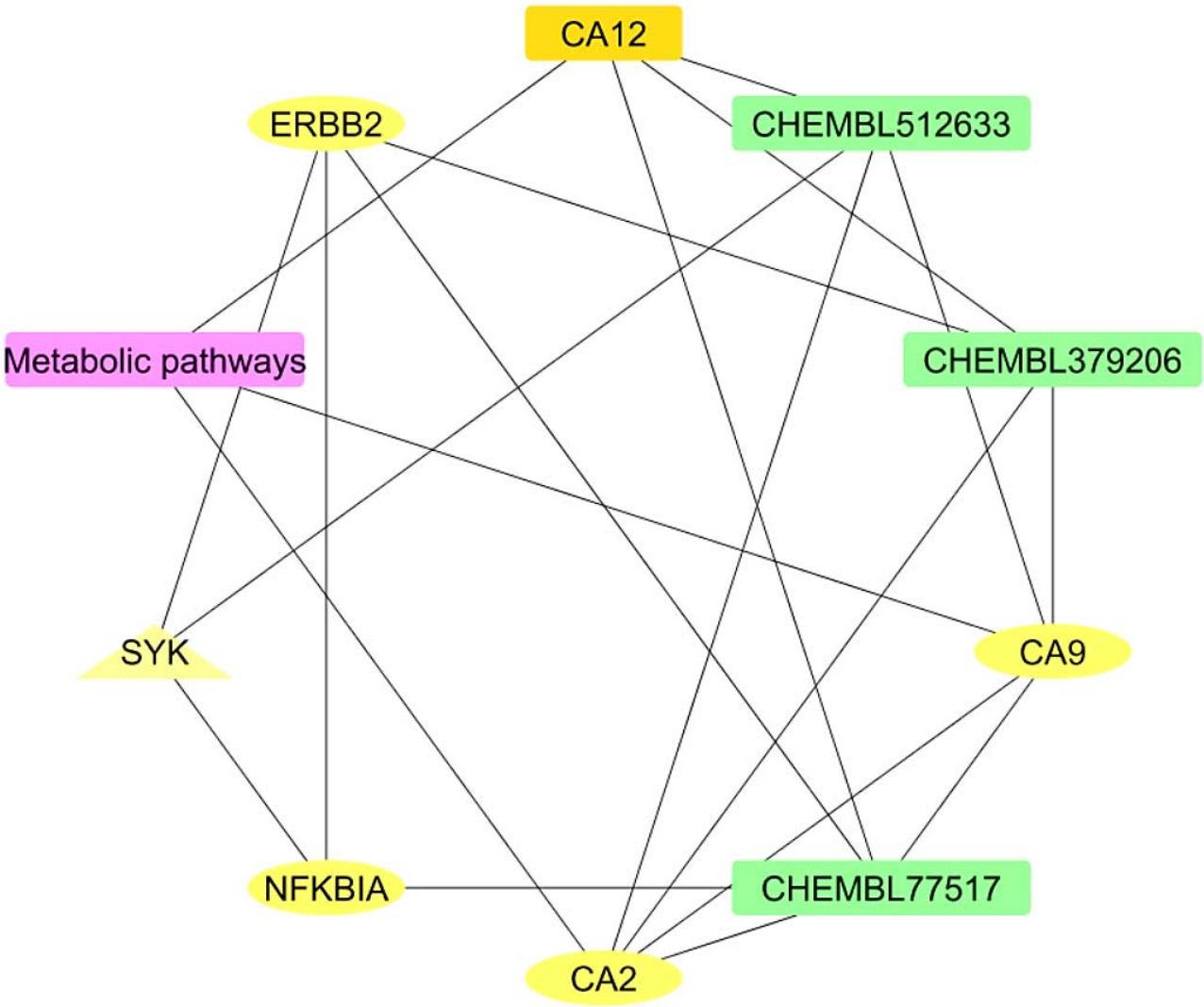


Figure 14b: An image of the top 10 genes by degree of connectivity, including three compounds (highlighted in green) that demonstrate the most frequent interactions with these genes and pathways.

Conclusion

This study extrapolates the biochemical and physiological significance of β-CA3, and its key role in Mtb's survival and pathogenesis. Employing a multidisciplinary approach, we utilized homology modeling, proteomic analyses, and advanced bioinformatics studies through our constructed SEQEclipse platform to reveal the unique structure and intricate physiological biochemical roles of β-CA3 in Mtb pathogenesis. Moreover, evolutionary insights from phylogenetic analyses emphasize the conservation of β-CA3 across species, suggesting broader implications for the use of β-CA3 inhibitors (β-CA3 small molecule modulators) in tuberculosis treatment across different bacterial strains. The potential of these inhibitors to disrupt crucial metabolic and physiological pathways in Mtb positions them as promising candidates for novel anti-TB therapies. Our findings confirmed the dual-domain nature of the enzyme and highlighted its critical involvement in cellular processes and metabolic pathways. The use of mechanistic QSAR models allowed

providing a basis for future drug design efforts aimed at combatting tuberculosis. The QSAR analysis identified potent compounds targeting β-CA3, where models demonstrated strong correlation coefficients exceeding 0.85 for key descriptors such as 'nHeavyAtom', which relates to molecular size and reactivity, and 'apol', indicative of polarizability, ‘ATS1m and ATS2m’ for spatial arrangements of atoms, which are crucial for interaction with the enzyme's active site. The application of systems biology further elucidated the complex interactions between these small molecule modulators and β-CA3, offering insights into their potential mechanisms of action within the pathogen’s metabolic network. Notably, the interaction network mapped using Cytoscape illustrated that active compounds such as CHEMBL77517 and CHEMBL512633 target pathways involved in nitrogen metabolism and fatty acid biosynthesis pathways are essential for Mtb virulence and survival. Despite the interest in this topic, the results are still preliminary and need to be confirmed in experiments. The in-silico predictions, although promising for generating robust biological effects, should be complemented by wet laboratory experiments to understand the diverse activities of these β-CA3 and validate any therapeutic potential they may have. Future research should focus on experimental validation of these computational findings and explore the clinical applicability of these promising compounds.

Acknowledgments: The authors extend their sincere appreciation to the following organizations for their invaluable support and funding: 1) the Finnish Cultural Foundation, 2) the Tampere Tuberculosis Foundation, (Ashok Aspatwar); 3) the Academy of Finland, and 4) the Jane and Aatos Erkko Foundation (Seppo Parkkila).

Competing interests: The authors declare no competing interests.

References

[1] P. Glaziou, K. Floyd, M.C. Raviglione, Global Epidemiology of Tuberculosis, Semin. Respir. Crit. Care Med. 39 (2018) 271–285. https://doi.org/10.1055/S-0038-1651492.

[2] A.A. Agyeman, R. Ofori-Asenso, Tuberculosis—an overview, J. Public Heal. Emerg. 1 (2017) . https://doi.org/10.21037/JPHE.2016.12.08.

[3] R. Bhowmik, A. Manaithiya, B. Vyas, R. Nath, K.A. Qureshi, S. Parkkila, A. Aspatwar, Navigating bioactivity space in anti-tubercular drug discovery through the deployment of advanced machine learning models and cheminformatics tools: a molecular modeling based retrospective study, Front. Pharmacol. 14 (2023) 1265573. https://doi.org/10.3389/fphar.2023.1265573.

[4] A. Aspatwar, M.E.E. Tolvanen, C. Ortutay, S. Parkkila, Carbonic anhydrase related proteins: molecular biology and evolution, Subcell. Biochem. 75 (2014) 135–156. https://doi.org/10.1007/978-94-007-7359-2_8.

[5] Global tuberculosis report 2023. Geneva: World Health Organization, January. (2023). https://iris.who.int/. (accessed April 30, 2024).

[6] A. Aspatwar, V. Kairys, S. Rala, M. Parikka, M. Bozdag, F. Carta, C.T. Supuran, S. Parkkila, Mycobacterium tuberculosis β-Carbonic Anhydrases: Novel Targets for Developing Antituberculosis Drugs, Int. J. Mol. Sci. 20 (2019) 5153. https://doi.org/10.3390/IJMS20205153.

[7] T.P. Stinear, T. Seemann, P.F. Harrison, G.A. Jenkin, J.K. Davies, P.D.R. Johnson, Z. Abdellah, C. Arrowsmith, T. Chillingworth, C. Churcher, K. Clarke, A. Cronin, P. Davis, I. Goodhead, N. Holroyd, K. Jagels, A. Lord, S. Moule, K. Mungall, H. Norbertczak, M.A. Quail, E. Rabbinowitsch, D. Walker, B. White, S. Whitehead, P.L.C. Small, R. Brosch, L. Ramakrishnan, M.A. Fischbach, J. Parkhill, S.T. Cole, Insights from the complete genome sequence of Mycobacterium marinum on the evolution of Mycobacterium tuberculosis (Genome Research (2008) 18 (729-741)), Genome Res. 18 (2008).

[8] S.T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M.A. Quail, M.A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J.E. Sulston, K. Taylor, S. Whitehead, B.G. Barrell, Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence, Nature. 393 (1998) 537–544. https://doi.org/10.1038/31159.

[9] T. Garnier, K. Eiglmeier, J.C. Camus, N. Medina, H. Mansoor, M. Pryor, S. Duthoy, S. Grondin, C. Lacroix, C. Monsempe, S. Simon, B. Harris, R. Atkin, J. Doggett, R. Mayes, L. Keating, P.R. Wheeler, J. Parkhill, B.G. Barrell, S.T. Cole, S. V. Gordon, R.G. Hewinson, The complete genome sequence of Mycobacterium bovis, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 7877–7882. https://doi.org/10.1073/PNAS.1130426100.

[10] A. Aspatwar, M.E.E. Tolvanen, H. Barker, L. Syrjanen, S. Valanne, S. Purmonen, A. Waheed, W.S. Sly, S. Parkkila, Carbonic anhydrases in metazoan model organisms: molecules, mechanisms, and physiology, Physiol. Rev. 102 (2022) 1327–1383. https://doi.org/10.1152/PHYSREV.00018.2021.

[11] A. Bateman, M.J. Martin, S. Orchard, M. Magrane, S. Ahmad, E. Alpi, E.H. Bowler-Barnett, R. Britto, H. Bye-A-Jee, A. Cukura, P. Denny, T. Dogan, T.G. Ebenezer, J. Fan, P. Garmiri, L.J. da Costa Gonzales, E. Hatton-Ellis, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, V. Joshi, D. Jyothi, S. Kandasaamy, A. Lock, A. Luciani, M. Lugaric, J. Luo, Y. Lussi, A. MacDougall, F. Madeira, M. Mahmoudy, A. Mishra, K. Moulang, A. Nightingale, S. Pundir, G. Qi, S. Raj, P. Raposo, D.L. Rice, R. Saidi, R. Santos, E. Speretta, J. Stephenson, P. Totoo, E. Turner, N. Tyagi, P. Vasudev, K. Warner, X. Watkins, R. Zaru, H. Zellner, A.J. Bridge, L. Aimo, G. Argoud-Puy, A.H. Auchincloss, K.B. Axelsen, P. Bansal, D. Baratin, T.M. Batista Neto, M.C. Blatter, J.T. Bolleman, E. Boutet, L. Breuza, B.C. Gil, C. Casals-Casas, K.C. Echioukh, E. Coudert, B. Cuche, E. de Castro, A. Estreicher, M.L. Famiglietti, M. Feuermann, E. Gasteiger, P. Gaudet, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, A. Kerhornou, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, V. Muthukrishnan, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, S. Poux, M.

Pozzato, M. Pruess, N. Redaschi, C. Rivoire, C.J.A. Sigrist, K. Sonesson, S. Sundaram, C.H. Wu, C.N. Arighi, L. Arminski, C. Chen, Y. Chen, H. Huang, K. Laiho, P. McGarvey, D.A. Natale, K. Ross, C.R. Vinayaka, Q. Wang, Y. Wang, J. Zhang, UniProt: the Universal Protein Knowledgebase in 2023, Nucleic Acids Res. 51 (2023) D523–D531. https://doi.org/10.1093/NAR/GKAC1052.

[12] M.S. Patrikainen, M.E.E. Tolvanen, A. Aspatwar, H.R. Barker, C. Ortutay, J. Jänis, M. Laitaoja, V.P. Hytönen, L. Azizi, P. Manandhar, E. Jáger, D. Vullo, S. Kukkurainen, M. Hilvo, C.T. Supuran, S. Parkkila, Identification and characterization of a novel zebrafish (Danio rerio) pentraxin-carbonic anhydrase, PeerJ. 5 (2017) e4128 . https://doi.org/10.7717/PEERJ.4128.

[13] A. Aspatwar, S. Haapanen, S. Parkkila, An Update on the Metabolic Roles of Carbonic Anhydrases in the Model Alga Chlamydomonas reinhardtii, Metabolites. 8 (2018) 22. https://doi.org/10.3390/METABO8010022.

[14] A. Aspatwar, J.Y. Winum, F. Carta, C.T. Supuran, M. Hammaren, M. Parikka, S. Parkkila, Carbonic Anhydrase Inhibitors as Novel Drugs against Mycobacterial β-Carbonic Anhydrases: An Update on In Vitro and In Vivo Studies, Molecules. 23 (2018) 2911. https://doi.org/10.3390/MOLECULES23112911.

[15] A. Aspatwar, H. Barker, H. Aisala, K. Zueva, M. Kuuslahti, M. Tolvanen, C.R. Primmer, J. Lumme, A. Bonardi, A. Tripathi, S. Parkkila, C.T. Supuran, Cloning, purification, kinetic and anion inhibition studies of a recombinant β-carbonic anhydrase from the Atlantic salmon parasite platyhelminth Gyrodactylus salaris, J. Enzyme Inhib. Med. Chem. 37 (2022) 1577–1586. https://doi.org/10.1080/14756366.2022.2080818.

[16] A. Aspatwar, C.T. Supuran, A. Waheed, W.S. Sly, S. Parkkila, Mitochondrial carbonic anhydrase VA and VB: properties and roles in health and disease, J. Physiol. 601 (2023) 257–274. https://doi.org/10.1113/JP283579.

[17] A. Aspatwar, L. Syrjänen, S. Parkkila, Roles of Carbonic Anhydrases and Carbonic Anhydrase Related Proteins in Zebrafish, Int. J. Mol. Sci. 23 (2022) 4342. https://doi.org/10.3390/IJMS23084342.

[18] A. Aspatwar, M.E.E. Tolvanen, E. Jokitalo, M. Parikka, C. Ortutay, S.K.E. Harjula, M. Rämet, M. Vihinen, S. Parkkila, Abnormal cerebellar development and ataxia in CARP VIII morphant zebrafish., Hum. Mol. Genet. 22 (2012) 417–432. https://doi.org/10.1093/HMG/DDS438.

[19] A. Aspatwar, M.E.E. Tolvanen, S. Parkkila, Phylogeny and expression of carbonic anhydrase-related proteins, BMC Mol. Biol. 11 (2010). https://doi.org/10.1186/1471-2199-11-25.

[20] A. Aspatwar, K.A. Qureshi, A. Parvez, J. Parkkinen, and S. Parkkila, Bicarbonate and bacteria: Friends or foes?" PloS Pathogens (2024). https://doi.org/10.1371/journal.ppat.1012029.

[21] A. Aspatwar, H. Barker, M. Tolvanen, R.Z. Emameh, S. Parkkila, Carbonic anhydrases from pathogens: protozoan CAs and related inhibitors as potential antiprotozoal agents, Carbon. Anhydrases Biochem. Pharmacol. an Evergr. Pharm. Target. (2019) 449–475. https://doi.org/10.1016/B978-0-12-816476-1.00020-4.

[22] A. Aspatwar, V. Kairys, S. Rala, M. Parikka, M. Bozdag, F. Carta, C.T. Supuran, S. Parkkila, Mycobacterium tuberculosis β-Carbonic Anhydrases: Novel Targets for Developing Antituberculosis Drugs, Int. J. Mol. Sci. 20 (2019) 5153. https://doi.org/10.3390/IJMS20205153.

[23] A.S. Covarrubias, T. Bergfors, T.A. Jones, M. Högbom, Structural Mechanics of the pH-dependent Activity of β-Carbonic Anhydrase from Mycobacterium tuberculosis, J. Biol. Chem. 281 (2006) 4993–4999. https://doi.org/10.1074/JBC.M510756200.

[24] S. Bienert, A. Waterhouse, T.A.P. De Beer, G. Tauriello, G. Studer, L. Bordoli, T. Schwede, The SWISS-MODEL Repository-new features and functionality, Nucleic Acids Res. 45 (2017) D313–D319. https://doi.org/10.1093/NAR/GKW1132.

[25] R.S. Rowlett, Structure and catalytic mechanism of the β-carbonic anhydrases, Biochim. Biophys. Acta - Proteins Proteomics. 1804 (2010) 362–373. https://doi.org/10.1016/J.BBAPAP.2009.08.002.

[26] T. Garnier, K. Eiglmeier, J.C. Camus, N. Medina, H. Mansoor, M. Pryor, S. Duthoy, S. Grondin, C. Lacroix, C. Monsempe, S. Simon, B. Harris, R. Atkin, J. Doggett, R. Mayes, L. Keating, P.R. Wheeler, J. Parkhill, B.G. Barrell, S.T. Cole, S. V. Gordon, R.G. Hewinson, The complete genome sequence of Mycobacterium bovis, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 7877–7882. https://doi.org/10.1073/pnas.1130426100.

[27] S. Gu, J. Chen, K.M. Dobos, E.M. Bradbury, J.T. Belisle, X. Chen, Comprehensive proteomic profiling of the membrane constituents of a Mycobacterium tuberculosis strain., Mol. Cell. Proteomics. 2 (2003) 1284–1296. https://doi.org/10.1074/mcp.M300060-MCP200.

[28] Y. Xiong, M.J. Chalmers, F.P. Gao, T.A. Cross, A.G. Marshall, Identification of Mycobacterium tuberculosis H37Rv integral membrane proteins by one-dimensional gel electrophoresis and liquid chromatography electrospray ionization tandem mass spectrometry, J. Proteome Res. 4 (2005) 855–861. https://doi.org/10.1021/pr0500049.

[29] K.G. Mawuenyega, C. V. Forst, K.M. Dobos, J.T. Belisle, J. Chen, E.M. Bradbury, A.R.M. Bradbury, X. Chen, Mycobacterium tuberculosis functional network analysis by global subcellular protein profiling, Mol. Biol. Cell. 16 (2005) 396–404. https://doi.org/10.1091/MBC.E04-04-0329.

[30] J.C. Camus, M.J. Pryor, C. Médigue, S.T. Cole, Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv, Microbiology. 148 (2002) 2967–2973. https://doi.org/10.1099/00221287-148-10-2967.

[31] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242. https://doi.org/10.1093/NAR/28.1.235.

[32] M.R. Wilkins, E. Gasteiger, A. Bairoch, J.C. Sanchez, K.L. Williams, R.D. Appel, D.F. Hochstrasser, Protein Identification and Analysis Tools in the ExPASy Server, Methods Mol. Biol. 112 (1999) 531–552. https://doi.org/10.1385/1-59259-584-7:531.

[33] N. Blom, S. Gammeltoft, S. Brunak, Sequence and structure-based prediction of eukaryotic protein phosphorylation sites, J. Mol. Biol. 294 (1999) 1351–1362. https://doi.org/10.1006/JMBI.1999.3310.

[34] B. Mészáros, G. Erdös, Z. Dosztányi, IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic Acids Res. 46 (2018) W329–W337. https://doi.org/10.1093/NAR/GKY384.

[35] S. Saha, G.P.S. Raghava, Predicting Virulence Factors of Immunological Interest, Methods Mol. Biol. 409 (2007) 407–415. https://doi.org/10.1007/978-1-60327-118-9_31.

[36] S. Saha, G.P.S. Raghava, VICMpred: An SVM-based Method for the Prediction of Functional Proteins of Gram-negative Bacteria Using Amino Acid Patterns and Composition, Genomics. Proteomics Bioinformatics. 4 (2006) 42–47. https://doi.org/10.1016/S1672-0229(06)60015-6.

[37] E.W. Sayers, E.E. Bolton, J.R. Brister, K. Canese, J. Chan, D.C. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, T. Madej, A. Marchler-Bauer, C. Lanczycki, S. Lathrop, Z. Lu, F. Thibaud-Nissen, T. Murphy, L. Phan, Y. Skripchenko, T. Tse, J. Wang, R. Williams, B.W. Trawick, K.D. Pruitt, S.T. Sherry, Database resources of the national center for biotechnology information, Nucleic Acids Res. 50 (2022) D20–D26. https://doi.org/10.1093/NAR/GKAB1112.

[38] K. Tamura, G. Stecher, S. Kumar, MEGA11: Molecular Evolutionary Genetics Analysis Version 11, Mol. Biol. Evol. 38 (2021) 3022–3027. https://doi.org/10.1093/MOLBEV/MSAB120.

[39] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A.L. Gable, T. Fang, N.T. Doncheva, S. Pyysalo, P. Bork, L.J. Jensen, C. Von Mering, The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest, Nucleic Acids Res. 51 (2023) D638–D646. https://doi.org/10.1093/NAR/GKAC1000.

[40] X. Zhou, W. Zheng, Y. Li, R. Pearce, C. Zhang, E.W. Bell, G. Zhang, Y. Zhang, I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction, Nat. Protoc. 17 (2022) 2326–2353. https://doi.org/10.1038/s41596-022-00728-0.

[41] W. Zheng, C. Zhang, Y. Li, R. Pearce, E.W. Bell, Y. Zhang, Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations, Cell Reports Methods. 1 (2021) 100014. https://doi.org/10.1016/j.crmeth.2021.100014.

[42] J. Yang, Y. Zhang, I-TASSER server: new development for protein structure and function predictions, Nucleic Acids Res. 43 (2015) W174–W181. https://doi.org/10.1093/NAR/GKV342.

[43] C. Colovos, T.O. Yeates, Verification of protein structures: patterns of nonbonded atomic interactions, Protein Sci. 2 (1993) 1511–1519. https://doi.org/10.1002/PRO.5560020916.

[44] R. Lüthy, J.U. Bowie, D. Eisenberg, Assessment of protein models with three-dimensional profiles, Nature. 356 (1992) 83–85. https://doi.org/10.1038/356083A0.

[45] R.A. Laskowski, M.W. MacArthur, D.S. Moss, J.M. Thornton, PROCHECK: a program to check the stereochemical quality of protein structures, J. Appl. Crystallogr. 26 (1993) 283–291. https://doi.org/10.1107/S0021889892009944.

[46] P.J.A. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M.J.L. De Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics, Bioinformatics. 25 (2009) 1422–1423.

https://doi.org/10.1093/BIOINFORMATICS/BTP163.

[47] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (2023) 1123–1130. https://doi.org/10.1126/science.ade2574.

[48] N. Rego, D. Koes, 3Dmol.js: molecular visualization with WebGL, Bioinformatics. 31 (2015) 1322–1324. https://doi.org/10.1093/BIOINFORMATICS/BTU829.

[49] C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474. https://doi.org/10.1002/JCC.21707.

[50] D. Valencia, R.E. Lillo, · Juan Romo, A Kendall correlation coefficient between functional data, Adv. Data Anal. Classif. 13 (2019) 1083–1103. https://doi.org/10.1007/s11634-019-00360-z.

[51] J. Spiegel, H. Senderowitz, Evaluation of QSAR Equations for Virtual Screening, Int. J. Mol. Sci. 21 (2020) 7828. https://doi.org/10.3390/IJMS21217828.

[52] Evaluation Metric for Regression Models - Analytics Vidhya, (2021). https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/ (accessed April 30, 2024).

[53] L. Kolberg, U. Raudvere, I. Kuzmin, P. Adler, J. Vilo, H. Peterson, g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update), Nucleic Acids Res. 51 (2023) W207–W212. https://doi.org/10.1093/NAR/GKAD347.

[54] M. Kolesnichenko, N. Mikuda, U.E. Höpken, E. Kärgel, B. Uyar, A.B. Tufan, M. Milanovic, W. Sun, I. Krahn, K. Schleich, L. von Hoff, M. Hinz, M. Willenbrock, S. Jungmann, A. Akalin, S. Lee, R. Schmidt□Ullrich, C.A. Schmitt, C. Scheidereit, Transcriptional repression of NFKBIA triggers constitutive IKK□ and proteasome□independent p65/RelA activation in senescence , EMBO J. 40 (2021) e104296. https://doi.org/10.15252/embj.2019104296.

[55] A. Mócsai, J. Ruland, V.L.J. Tybulewicz, The SYK tyrosine kinase: a crucial player in diverse biological functions, Nat. Rev. Immunol. 10 (2010) 387–402. https://doi.org/10.1038/nri2765.

[56] M. Wagener, J.C. Hoving, H. Ndlovu, M.J. Marakalala, Dectin-1-Syk-CARD9 signaling pathway in TB immunity, Front. Immunol. 9 (2018) 225. https://doi.org/10.3389/FIMMU.2018.00225/BIBTEX.

[57] A. Solokhina, D. Brückner, G. Bonkat, O. Braissant, Metabolic activity of mature biofilms of Mycobacterium tuberculosis and other non-tuberculous mycobacteria, Sci. Reports 7 (2017) 9225. https://doi.org/10.1038/s41598-017-10019-4.

[58] P. Chakraborty, S. Bajeli, D. Kaushal, B.D. Radotra, A. Kumar, Biofilm formation in the lung contributes to virulence and drug tolerance of Mycobacterium tuberculosis, Nat. Commun. 12 (2021) 1606. https://doi.org/10.1038/s41467-021-21748-6.