Deciphering the Biochemical and Physiological Mechanisms of Mycobacterium Tuberculosis β-Carbonic Anhydrase 3 and Its Modulators through Mechanistic QSAR Model and Integrated Systems Biology Approaches

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

[2]Department of NEUROFARBA, Section of Pharmaceutical and Nutraceutical Sciences, University of Florence, Firenze, Italy

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

Abstract

This study in particular demonstrates a novel bioinformatics pipeline to highlight the physiological and biochemical mechanism of β-CA3, a β-carbonic anhydrase encoded by *Mycobacterium tuberculosis* (Mtb) of Rv3273, which possesses a unique two-domain structure, with a sulfate transporter domain at the N-terminal and β-CA domain at the C-terminal. 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 strategies to uncover the key biological and structural insights. Construction and development of a Python-based custom web server, SEQEclipse (https://seqeclipse.streamlit.app/) equipped with basic biopython work packages was done 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 functional class of the sequence, BLAST and MSA analysis, protein-protein interaction mechanism, and 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 a mechanistic system 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 overall study outlines the biological significance of β-CA3 functionality along with a mechanistic understanding of designing small molecule modulators as anti-TB drugs for future prospectives.

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

Corresponding author:
***Dr. Ashok Aspatwar**

PhD (BioMediTech)
Adjunct Professor
Faculty of Medicine and Health Technology,
Tampere University,
Tampere, Finland
Email Id: ashok.aspatwar@tuni.fi
Orcid ID: https://orcid.org/0000-0002-6938-7835

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

Highlights

- Applied homology modeling to decode the intricate dual-domain architecture of β-CA3, crucial for pioneering targeted therapies against *Mycobacterium tuberculosis*.
- Developed SEQEclipse, a Python-based web server, to centralize bioinformatics analyses, enhancing the drug design workflow.
- Utilized machine learning QSAR models to pinpoint potent compounds targeting β-CA3, with descriptors like 'nHeavyAtom' Atsm1, Atsm2, 'apol', 'SubFP287', 'SubFP184', 'SubFP214', and 'SubFP275, showing high correlation and interaction potential.
- Employed systems biology to illustrate how identified active and inactive compounds interfere with essential metabolic pathways in *Mycobacterium tuberculosis*, suggesting novel therapeutic angles.
- Emphasized the need for validation of computational findings to confirm the efficacy and safety of identified compounds.

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 the infected people develop an active TB in their lifetime [2]. In terms of localization for Mtb infection, the lungs are mostly targeted, causing a pulmonary infection that may become worse over time. However, it has been found that TB can also spread to other regions such as lymph nodes, urinary tract, and brain. The Global Tuberculosis Report 2023 showed that in 2023 the total number of TB cases were around 9.9 million among which around 5.6 million (57%) consisted of the male population whereas around 3.2 million (32%) consisted the female population and around 1.1 million (11%) of them were children [3,4][5]. Another highly concerning fact about Mtb infection is the multi-drug resistant TB and extended drug-resistant TB which belongs to difficult-to-treat variants of TB due to its ability to evade drug actions due to genetic mutations which make major of existing drugs ineffective against Mtb infection. Thus to deal with this problem statement virtual analysis of the target protein using bioinformatics analysis of the base sequence needs to be performed to study the internal properties and molecular parameters which can be exploited to

respond to drug action and use cheminformatics to create and screen molecular compounds virtually which can respond positively to Mtb infection[6].

To address the integral properties of mycobacterium species, genome sequencing and proteomics are used. These workflows dissect the biological and structural nuances that are fundamental to designing the new anti-TB drugs. The use of ML-curated QSAR models, along with activity prediction, makes the drug-designing process increasingly smooth. Besides, system biology approaches reveal mycobacterial connections, offering a probable drug development mechanism for TB treatment against drug-resistant Mtb [7]. Fore 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 its evolutionary characteristics [8]. This process of gene sequencing coupled with proteomics and transcriptome data can provide information regarding gene expression and protein production level changes on action with various molecular compounds which can be an integral solution to deal with the drug-resistant Mtb property [9]. Among the various target proteins that are analyzed in the mentioned workflows, the ones we will be specifically focussing on are known as carbonic anhydrases (CAs, EC 4.2.1.1).

The CAs are enzymes that are classified into eight various unrelated gene families namely (α-, β-, γ-, δ-, ζ-, η-, θ-, and ι-CAs) [10,11]. CAs are associated with the performance of various integral physiological functions within living organisms. Firstly, organisms belonging to the taxonomic class of vertebrates are mostly categorized into the α-CA gene family which consists of various molecular conformations of the target enzyme [4,6,10–19]. The main biological function of CA includes the reversible hydration of carbon dioxide to bicarbonate and hydrogen ions for maintaining the proper pH homeostasis and their involvement in biosynthetic reactions like gluconeogenesis and ureagenesis in mammals[4]. However, it has been observed through various scientific studies that dysregulations or mutations within the enzyme result in malfunctioning of the target leading to various diseases glaucoma, epilepsy, and cancer[17]. In addition, in pathogenic bacteria and parasites, CAs play a crucial role in disease pathogenesis and survival in the host [20][21]. Hence utilizing these targets and leveraging drug development workflows various drugs that can counter the disease effects can be created to allow proper treatment of diseases[1]. On the other hand, organisms belonging to bacterial and pathogenic parasitic categories are observed to contain the β-CA family, which are scientifically studied to observe that they are also associated with the progression of various diseases, which are also analyzed using molecular assessment processes and obtained information is fed to priorly mentioned drug development pipelines to screen molecules or compounds which can have inhibition properties against the disease-causing protein. 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 onto the host organism. Some of the inhibitors that are already analyzed in the available literature and clinically used are sulphonamides and their bio isosteres [2,5–8].

For this project, the primary focus will be concentrated on Mtb CAs belonging to the class β-CA gene family. The β-CAs are encoded by the genes, Rv1284, Rv3588c, and Rv3273 which are also conventionally

labeled as β-CA1, β-CA2, and β-CA3 respectively [6,8]. The mentioned Mtb β-CAs are the enzymes that are more specifically classified in metalloenzymes that contain zinc as a cofactor and have similar biological and physiological properties like many other bacterial β-CAs that contain various conserved amino acid domains that are responsible for catalytic processes like the residues which are present as the binding residues for zinc ion: Cys42, Asp44, His97, and Cys101[6,8]. The above-mentioned β-CAs are genetically cloned and purified through analytical techniques and bio-kinetic analysis has been performed to assess the properties of protein. Although only β-CA1 and β-CA2 have been crystallized through X-ray crystallography, the structural information about β-CA3 can be derived through homology modeling from the protein sequence of β-CA Can2 from Cryptococcus neoformans organism as a template (PDB code: 2W3N) [4,8,10,11,22–25]. In this project, we will put more focus on the β-CA3 structure and functions which involves information gathering from NCBI and UniProt databases which will give an overall idea about performing molecular modeling processes and gathering ontological information about gene products and expression analysis. 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)[26–28]. This β-CA3 derived from the gene Rv3273 was obtained from the Mtb strain of H37Rv which is analyzed to be a transmembrane protein with a hydrophobic N-terminal[29]. We have utilized various bioinformatics analysis tools and servers to assess the physio-chemical properties of the sample protein sequence obtained from databases, derive evolutionary information and phylogenetic studies, and analyzing protein-protein and protein-chemical interactions with β-CA3 protein as the main target sequence derived from Cryptococcus neoformans and a sequence obtained from wet lab experimentation. The experiment will conclude with creating a homology model of the sequence and performing validation and error analysis to check the viability of the protein modelled. 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 analysis, investigate the protein-protein interaction mechanism, and employ an ab initio modeling strategy. We further developed a quantitative structure-activity relationship (QSAR) model to provide a detailed knowledge of how small molecule modulators block the β-CA3 enzyme. The objective of our QSAR model was to find crucial molecular characteristics that showed a strong link with 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.

2. Methods and Materials

This work has involved the use of different bioinformatic databases and servers for the acquisition and analysis of the specified biological data for CAs. The results of the study revealed the biological, physicochemical, and structural/chemical characteristics of the CAs, as well as their enrichment function and cellular locations of them.

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

Initially, biological data for every CA were extracted from the UniProt[12] and PDB databases[30] in FASTA format, which is typically used by bioinformatic servers. The bioinformatics process was commenced by calculating the various physicochemical properties of amino acid samples from these CAs. The Expasy ProtParam Server[31] is used to calculate several properties of the two protein samples including the number of amino acids, and the quantities of negatively and positively charged amino acids which provide insights into their overall charge. Furthermore, the molecular weight, and the pH at the isoelectric point (PI), are crucial for protein quantification via SDS PAGE methods are also determined. The stability index was also calculated, which showed the stability of the protein; the values under 40 indicated a stable conformation, while the values above 40 pointed to the instability, which would potentially require Molecular Dynamics methods for stabilization. The aliphatic index was calculated as total aliphatic groups on the side chain of amino acids increase protein thermostability and molecular stability. The result was obtained by calculating the grand average of hydropathicity (GRAVY) which placed the protein in either the hydrophobic positive or hydrophilic categories if the outcome was negative. This data is fundamental in establishing the bioactivity or interoperability and ADMET of the protein that binds with this protein if any. Returned data from the ProtParam server can be utilized in comparing CA samples.

2.2. Analyzing Phosphorylation Sites and Disordered Regions in CA Protein Sequences

This data was used for phosphorylation site detection as well as kinase enzyme identification for protein sequences. The NetPhos 3.1 server[32] generated tabular and graphics outputs of amino acid sequences of CAs with the aid of the provided FASTA file. The table provided the list of amino acids and kinases (enzymes that add a phosphate group) that contribute to phosphorylation while the graph emphasized phosphorylation potential which had sharp peaks pointing to areas of intense activity. Besides, different charges of the CA sample were also taken into consideration and were presented using the IUPred2A server[33]. A web-based interface presented disorder scores, with their relation to high variability highlighted (regions with high variability could serve as ligand binding sites).

2.3.Determination of the functional class of the protein sequence

The VICMPred Server[34,35] was used to analyze biological processes and categories varied from virulence factors, information molecules, and cellular processes to metabolism molecules of the given amino acid sequence. The sequence was then converted into the FASTA format that was subjected to classification using the SVM (Support Vector Machine) algorithm which gave scores based on the probability among four possible classes. The highest grade in this category provided the most certain answer to the primary

biological function of the protein. Definitions established needed to include the following terms: virulence factors as they are part of the virulence mechanism, information molecules as they generate genomic maps, cellular processes as part of processes involved in post-translational localization, and metabolism molecules as they create metabolites or bioactive products.

2.4.Performing BLAST and Sequence Alignment to feed into phylogenetic trees for evolutionary analysis

BLASTP from NCBI[36] was used to run the sample amino acid sequences similar to the dynamic databases for Multiple Sequence Alignment (MSA) and Phylogenetic Tree analysis. This process served to deduce the evolutionary characteristics of the query protein compared to those of similar proteins present in the database. Once the similar sequences were obtained, they were saved in a text document and uploaded into MEGA 11[37] software. There, MSA was performed and a phylogenetic tree was created 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.

2.5.Analysis of Protein-Protein interactions and biological activity potential using STRING server

Protein Sequence samples were obtained from UniProt and PDB databases along with their gene names and functional annotations about specific CA proteins. These sequences of amino acids or gene names, along with the organism from where the parent was taken, were all typed into the STRING Online database[38]. The result of STRING visualized a network of interactions in the center revolving around the searched protein/gene, with its biological and functional interactions with other proteins, genes, and chemicals. Nodes represented proteins and chemical and connecting lines potential interaction shown. This network permitted attenuation of view interaction amount or type. The tabular format showed the details of interaction and their respective confidence scores. The data analysis tools provided biological processes and molecular functions using GO terms and metabolic pathways of the human protein or gene from the KEGG database. Besides protein domains and cellular locations, there was relevant data. Using the STRING database, the molecular interaction network and functional enrichment associated with the target protein could be identified in a focused way.

2.6.Determination of tertiary protein structure through ab-initio modelling approaches through I TASSER Server.

The introduction section noted that exact information on β-CA3 was unresolved due to the lack of a fully crystallized structure and tertiary conformations. The structure in the database was obtained through homology modeling based on similar proteins. From the laboratory-extracted sequence of β-CA3, ab-initio modeling was performed to derive the tertiary structure and conformational features including physico-chemical and biological information. The I TASSER provided for the was employed to input the β-CA3 amino acid sequence in FASTA format[39–41], with different structural details produced. L I-TASSER's

outcomes showed line graphs of the determined secondary structure for each amino sequence acid with confidence levels ranging from 0 (lowest confidence) to 9 (highest confidence). It also presented graphs that showed solvent accessibility with acescores from 0 to 9 that represent the relative exposure level of amino acids to the solvent environment where 0 indicates completely buried and 9% extremely exposed. Moreover, the Thermal Stability profile of the protein structure was modeled based on NBF (Normalized B-Factor)'s result and thus, it is known which residue position is stable and which is unstable. The results of the modeling include an ensemble of models for the tertiary structure with its corresponding PDB code, TM score, and RMSD value to guide the best model selection using confidence and deviation scores. Another type of output were potential small molecules and ligands that could bind with protein together and binding site clarification that would be useful for docking studies. With the available Gene Ontology (GO) terms linked to molecular functions, cellular locations, as well as biological processes, gene function was able to be studied further.

2.7.Verifying the protein structure produced through ab-initio modelling from the ITASSER server

The ITASSER server offered an atomic model for the protein of β-CA3 based on the database information from the laboratory. The structural predictions provided by bioinformatics servers on databases should thus only serve as references to allow modeling virtually these structures similarly and, hence, the detection of potential errors. To check the predicted structure, compare it to the error regions, and validate its excellent quality, SAVES v6.0 server was used. Those tools, such as ERRAT[42], were introduced that determined the quality of the model by studying non-interaction interactions among atoms, resulting in an overall quality score. The VERIFY 3D[43] study compared the 3D model with the 1D amino acid sequence to check the completion of quality structure and disclose the structural class. The final tools WERE WHAT CHECK and PRO CHECK[44], which added the stereochemical properties of the model. It concerned the residue-residue geometry, molecular interactions, bond length, and angles. These information quality parameters and a Reliable Error detector or safety mechanism for the stability and integrity of the vehicle. Corrected levels of errors will likely be mitigated by Molecular Dynamics simulation and the inclusion of cofactors that will stabilize protein structure.

2.8. SEQEclipse: Comprehensive Web Server for Bioinformatics Analysis

The SEQEclipse is an in-house-developed server, built around the Streamlit structural design, which integrates Biopython for bioinformatics toolkit[45], STRING server API[38], ESMSFold API[46], for molecular visualizations, and Py3DMol[47]. SEQEclipse includes the feature of posting FASTA sequences of the samples of protein and automatically detecting sequences and names and it is easy to copy sequences for further operations. It combines with the Expasy ProtParam through the Biopython module for physicochemical properties of sequences, the BLASTP interface for finding similar sequences based on user-specified E values, and the PDB module from Biopython and Py3DMol for protein structures modeling, which enables live visualization in 3D. Protein-protein, protein-chemical interactions Gene ontology, and Enhanced features were performed using STRING API. We also integrate the homology modeling using the

ESMFold API, to predict 3D protein structures by comparing sequences. The one unique aspect of SEQEclipse is that it brings different key bioinformatics analyses into a centralized user-friendly interface.

2.9.Mechanistic QSAR model for β-CA3 inhibition mechanism

We constructed a QSAR model to gain a mechanistic understanding of the inhibition mechanism of β-CA3 through small molecule modulators. Our goal through 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.

2.9.1. Database selection and screening process

In the first stage, the QSAR approach retrieved 101 compounds inhibiting the β-Carbonic anhydrase 3 enzyme from the ChEMBL database. The principle of uniformity was served using a normalizing procedure, where the original Ki_{50} was transformed to pKi_{50} ($-\log10(pKi_{50})$). Then using the rdkit package sphecies were computed and all inhibitors were compared using the rule of five proposed by Lipinski. Consequently, we used descriptors to calculate them, and an analysis of this dataset was conducted using them to explore the relationships between the descriptors and the pKi_{50} values. The visualization was conducted via matplotlib and seaborn packages to discover elements of structure-activity relationships. Postprocessing named the data set refinement that led to a result of 101 β-Carbonic anhydrases-3 inhibitors separated into the following clusters: 61 active compounds, 13 inactive compounds, and 27 compounds showing intermediate activity.

2.9.2. Calculation of Chemical Descriptors and dataset division in QSAR Modeling

The choice of the right descriptors is one of the crucial issues that can be used in the construction of the QSAR model strength. In the dataset of 101 molecules, the PaDELPy tool contributed significantly to the experimental molecular descriptors and binary substructure fingerprint calculations[48]. This process contributed results of 1017 1D and 2D feature descriptors, and also 336 binary substructure fingerprints. The aspects that either were not changed much or were equal were removed. On the other hand, the continuous features were corrected to keep the model credible. Given this, feature selection was performed then to find some significant molecular attributes that provide a binary classification mechanism for these features. The dataset was then divided into training and test subsets in an 80:20 ratio. The area of employment of machine learning algorithms on these subsets is to develop the energetic stability of ML-QSAR bioactivity prediction models. These models build up the conceptions among molecular features (both structural and numerical) and the biological activity (pKi_{50} value). The Random Forest Algorithm (RFA) being the backbone of the ML-QSAR models for the β-Carbonic anhydrases-3 inhibitors (CAI-3) contributed significantly towards the development of predictive models.

2.9.3. Generation of Machine Learning Model

RF algorithm is the most powerful machine learning technique which brings together a good number of tree-based models to form a consistent and robust group. The goal of this approach is, in contrast with that of standalone decision trees, to find a more stable and accurate model. This new approach aims to overcome inconsistency in prediction eventually. The bootstrap aggregation technique is an effective way for the random forests to avoid the over fitting issue often faced by the decision trees. Such sampling strategies apply random sampling to subsets of training data with replacement and predictions are aggregated afterwards. In this way, ML-QSAR approaches (such as random forest-based nonlinear prediction models) have gotten along with the emergence of, being able to forecast the activity of new molecules against β-Carbonic anhydrases-3 by means various of molecular properties.

2.9.4. Performance Evaluation of the ML Algorithm

The accuracy and predictive value of each ML-QSAR model are assessed using both 5-fold and 10-fold cross-validations on the training set. The method was 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 then assessed on the remaining set. This recurrent cycle made every segment serve as the validation set. First, we did a test with the external data set right before we decided on the optimal hyperparameters for all three models to meet the generalization requirement. A performance evaluation of the models was done statistically using indicators such as Q2, Kendall's tau (τ), MALE, MAPE, RMSPE, RMSLE, Spherman's rho, MAE, RMSE, RAE, and RRSE (See equations in **supplementary file**). Such KPIs have created a strong foundation in the assessment of the model's efficiency[49–51].

The correlation coefficient (Q) is a measure used to estimate the predictive performance when modeling data for un-foreseen conditions. It is an indication of the generalization capability of a model as it can predict the outcomes that may lie beyond the domain of training data. A different set of techniques, namely, Kendall Tau and Spearman's Rho, test the determination of how well data points like those found in training sets were ordered during predictions. Various metrics, including MAE, RMSE, RAE, and RRSE, which are worked out to illustrate the size of the errors, have been used to put a quantitative measure of changes from the real property value. For example, MALE (Mean Absolute Logarithmic Error) and RMSLE (Root Mean Squared Logarithmic Error) may be more appropriate error metrics when dealing with data with a skew or overestimation is more crucial or underestimations are to be minimized. Absolute error measures like 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 to get an understanding of the precision and exactness of predictive modeling. Similarly, VIP analysis determines the main driving variables that are connected to each model category and hence, useful for the interpretation of predictive capability.

2.10. Inhibition mechanism of β-CA3 modulators through mechanistic system biology approach

We further extended the QSAR strategy to a mechanistic system 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.

2.10.1. Acquisition of Chemical Compounds and Targets

The chosen final molecules with the best QSAR model (prepared of 5 active and 5 inactive molecules) were subjected to the next system biology analysis. Predictions for 903 targets of the compound 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**). The selected gene for β-Carbonic anhydrases-3 was 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 *Mycobacterium tuberculosis* were sourced from GeneCard and OMIM databases, revealing 208 and 195 target genes respectively **(Table s3 and Table s4)**. A Venn Diagram analysis of compounds and target genes was the method used to identify common genes. This analysis has shown that these 21 genes are present in common in both the compounds and the target gene (**Table s5**).

2.10.2. Gene ontology /KEGG analysis

Functional enrichment analysis, covering biological pathways, molecular functions, cellular components, and biological processes, was conducted using go:profiler server (https://biit.cs.ut.ee/gprofiler/)[52] and David server (https://david.ncifcrf.gov/home.jsp). KEGG data were obtained from the STRING database (https://string-db.org/) using the gene database. Heatmap analysis was utilized to visualize gene pathways and their associated p-values, using data from the go:profiler server.

2.10.3. Compounds-Targets-Pathways and Protein-Protein Interaction

The Compounds-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. In the same way, 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 on Cytoscape 3.10.2.

3. Results and Discussions

The introductory part spotted the significance of β-CA3 for informational analysis as that concept is still less explained in the biological databases. The study used bioinformatics software to check biological functions and the structural conformation of the given protein sequence. This study aimed to give an understanding of pathways' enrichment and their localization, comprising metabolism and that potentially could be the resource for new-class drugs and bioactive compounds against β-CA. The next section of this study describes various in silico analyses using this and other web servers of bioinformatics. In the next sections, the SEQEclipse web server will be addressed, which is developed for this platform, and a machine learning

QSAR model will be built. This model will be integrated into the systems biology to identify the molecular mechanisms of compounds associated with the target pathways.

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

The outcome of physicochemical characteristics of the β-CA3 amino acid sequence obtained during the laboratory experiments showed some essential features. The sequence is represented by 216 amino acids with 30 negatively charged residues (Aspartate and Glutamate) and 13 positively charged residues (Arginine and Lysine). β-CA3 molecule has a molecular weight of 23,302.07 daltons and its theoretical isoelectric point is 4.62. The aliphatic index is 94.40, giving it a structural formula typical of alkanes. There are only 3,233 atoms located within the structure. The measured GRAVY score is -0.071, which indicates that this protein has a hydrophilic character in its sequence of amino acids. These properties place β-CA3 to be in 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 β-CA3 protein is negative due to the number of negatively charged amino acid residues being larger than positively charged. The stability index under 40 and the negative GRAVY score give an idea of the presence of a hydrophilic protein chain. These findings can be used as a basis for more advanced practical experiments.

3.2.Phosphorylation and Disorder Analysis of β-CA3 Protein

The analysis conducted on the β-CA3 protein sequence using the NetPhos server yielded significant findings. Setting the score filter threshold to 0.5, which is the server guideline, allowed for the identification of a few 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, pointing to 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 higher number of possible phosphorylation sites than Threonine and then Tyrosine (**Figure 1**). Such a phosphorylation potential map is essential in uncovering the regulatory mechanisms by which this protein functions and allows to highlight molecules for further validation. In addition, these phosphorylation sites might be significant in terms of protein-protein interactions and their overall activity and stability characteristics of the protein.

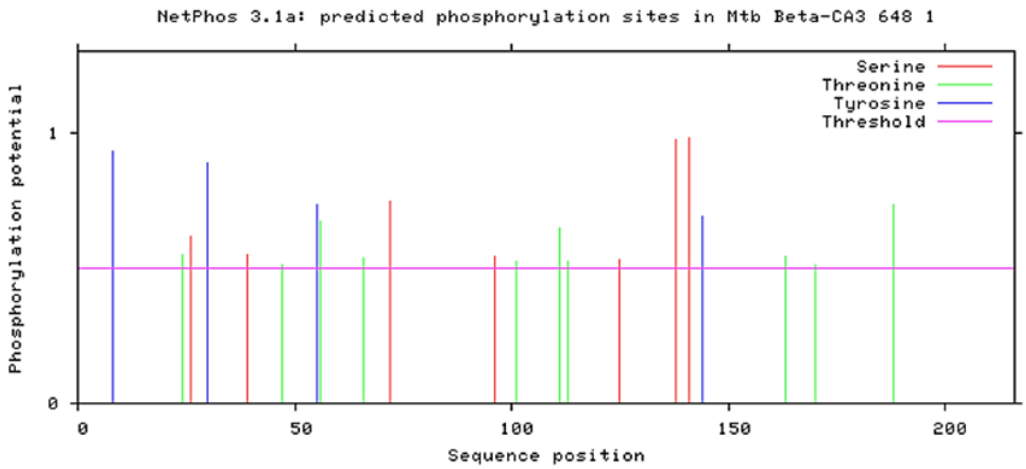


Fig. 1. Graphical representation of Phosphorylation potential across specific amino acid residues from NetPhos server on our analysed protein sequence of β-CA3.

Through IUPred2A servers, the disordered regions in the β-CA3 protein sequence were identified. The graph just showed 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 line represent binding sites that can be a place of interaction for protein or signaling sites. The structured areas that were below the threshold did not show the order (**Figure 2**). The variations in the IUPred2A disorder prediction highlighted by the blue line might be indicative of these flexible zones having a vital role in the protein activity or control. This disorder pattern gives hints of β-CA3 cellular function such as the interaction and modification.

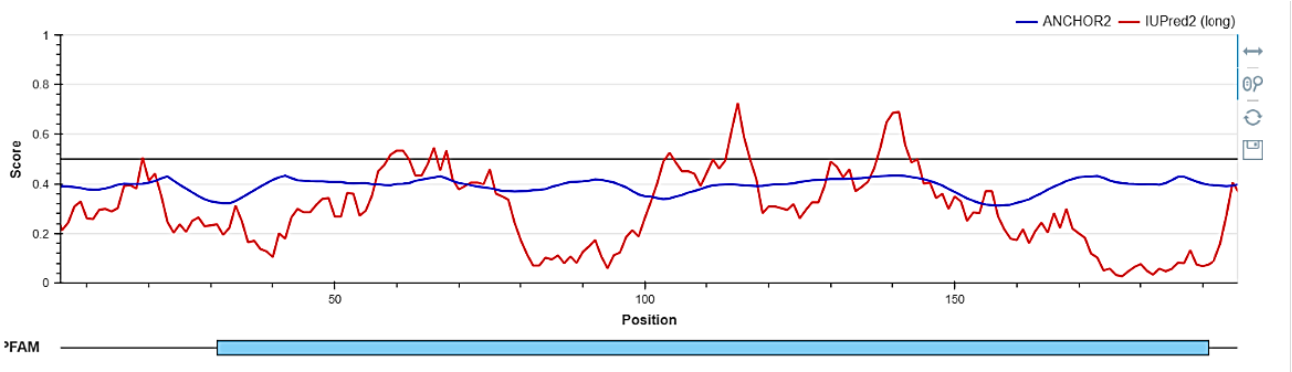


Fig.2. Graphical representation of disordered regions of analysed protein sequence of β-CA3 using the IUPred2A server

3.3.Classifying β-CA3 Biological Activity with VICMPred Server

The β-CA3 protein from the lab data sequences, when analyzed by VICMPred server, showed the best performance among other classifications in the area of cellular processes. This means that the basic biological influence of β-CA3 might be the cellular function. The protein may function by participating in cell signaling, transport mobility, or other cellular actions. The output of the VICMPred server clearly

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**). The disruption of β-CA3 function may have relevance for cellular processes like 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 the primary role in these functions, but it most probably plays an important role. The negative score of virulence factors shows that β-CA3 is not likely to be involved in pathogenic mechanisms.

3.4.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 the phylogenetic tree construction. This Multiple Sequence Alignment revealed the regions of conserved and variable sequences among β-CA3 homologs (**Figure s4**). The neighbor-joining method using the phylogenetic tree approach helped evaluate the evolutionary relationships among β-CA3 strains by bootstrap analysis with a threshold value of 55. To the tree, the bootstrap scores added a sort of assurance to the connections plotted. These observations on the phylogenetic analysis will give us an idea about which evolutionary trajectories and how the β-CA3 proteins are interrelated (**Figure s5**). It indicated significant conservation across various organisms, with the bootstrap scores highlighting the robustness of these evolutionary links. These data are very important for understanding β- CA3 functional evolution that comes along with some effects on biological processes in various species.

3.5.Protein-Protein Interactions and Functional Analysis of β-CA3 via STRING Server

We used the STRING server to input the amino acid sequence in FASTA format and selected the organism as *Mycobacterium tuberculosis* H37Rv (Virulent strain) to obtain the biological network associated with the β-CA3 protein. The result of the STRING server analysis showed the β-CA3 protein transmembrane function to be one of its key characteristics and the protein 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**). Given confidence scores ranged from 0 to 1 depicting the chances of interacting biomolecules. Through the functional enrichment analysis, it became clear that there was participation of metabolic processes such as carbon utilization and fatty acid biosynthetic processes (**Figure s7)**. Molecular functions like carbonate dehydratase activity and fatty acid synthase activity were found in β-CA3 that are also correlated with the nitrogen metabolism pathway. The GO terms provide detailed explanations of the molecular functions, the biological processes, and the subcellular localizations, and we can conclude then that β-CA3 represents a key role in the cellular activities of the M. tuberculosis H37Rv strain (**Table 1)**. In addition to this, cross-database search in STRING showed features such as carbonic anhydrase activity and the status of its involvement in the polyketide synthase complex pointing to the promising role of β-CA3 in the bacterial

metabolism process and those that may be pathogenic. The fact that β-CA3 is linked to nitrogen metabolism by the KEGG pathway shows the significance of β-CA3 in bacteria's survival (**Table 1**). Also, the fact that it is located in the polyketide synthase 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 is not only integral to sulfate transport but may also play diverse roles in the bacterial cell, potentially making it a target for therapeutic interventions.

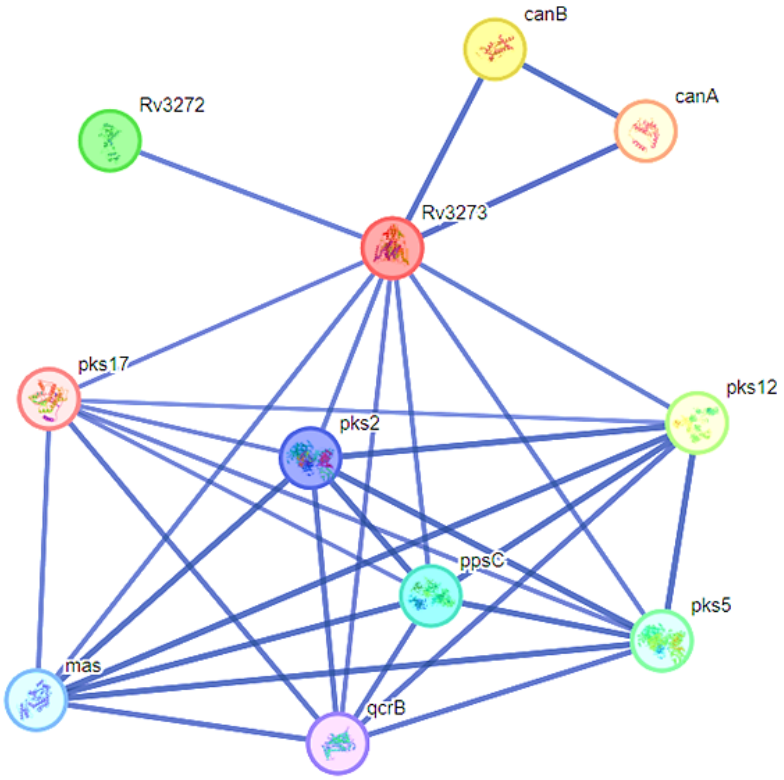


Fig.3. STRING server biological network showing all the interactions between β-CA3 protein(denoted by gene name Rv3273). The thickness of lines indicates the confidence of interactions.

Table 1: Details regarding GO terms for functional enrichment and cellular localisation 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	GO:00043	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

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

The absence of β-CA3's crystallized structure led to the usage 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, its bioactivity being implicated (**Figure 4**).

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

Table 2: Top 10 PDB Structures Displaying 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	5swcA	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

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

The assessment of the modeled structure of β-CA3 protein is important to avoid the inevitable mistakes and determine accuracy. As presented in **Fig. 5**, ERRAT server analysis showed a high quality of the predicted structure, yielding a 94.63% score of its accuracy. Fortunately, a few density areas colored red in the graphical representation surpassed the 99% rejection level, so these regions pointed at the local model's errors. Contrasted to that, **Fig. 5's** VERIFY 3D, plots the protein's structure where the model ranked low

against a protein's 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 is intended to resolve the discrepancies and contribute to their stabilization. Dock 19 and 20 include the response of WHATCHECK and PROCHECK tools which are positive and constructive to structure. The presence of unwanted deviations which may be due to bad bond formation or bond interactions between atoms emphasizes the need for more optimization routine (**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 which has a certain number of residues span. This graph shows that the majority of the structure is below the commonly used 95% confidence measuring, and this would suggest 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 are not falling in the expected range of non-bonded 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 which average score is below the threshold line give rise to closer investigation to disclose whether they are true structural features or modeling errors. The Ramachandran plot demonstrates the phi (φ)and psi (ψ) torsion angles for amino acid sequence in the protein structure. Ideally, most of the residues should lie within the sufficient region (accepted as dark red), while only a few of them are allowed (shown as lighter yellow) and none in the disallowed region. The main of the Ramachandran plot shows that most of the described residues fall into the favored regions, thus being an indicator of 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.

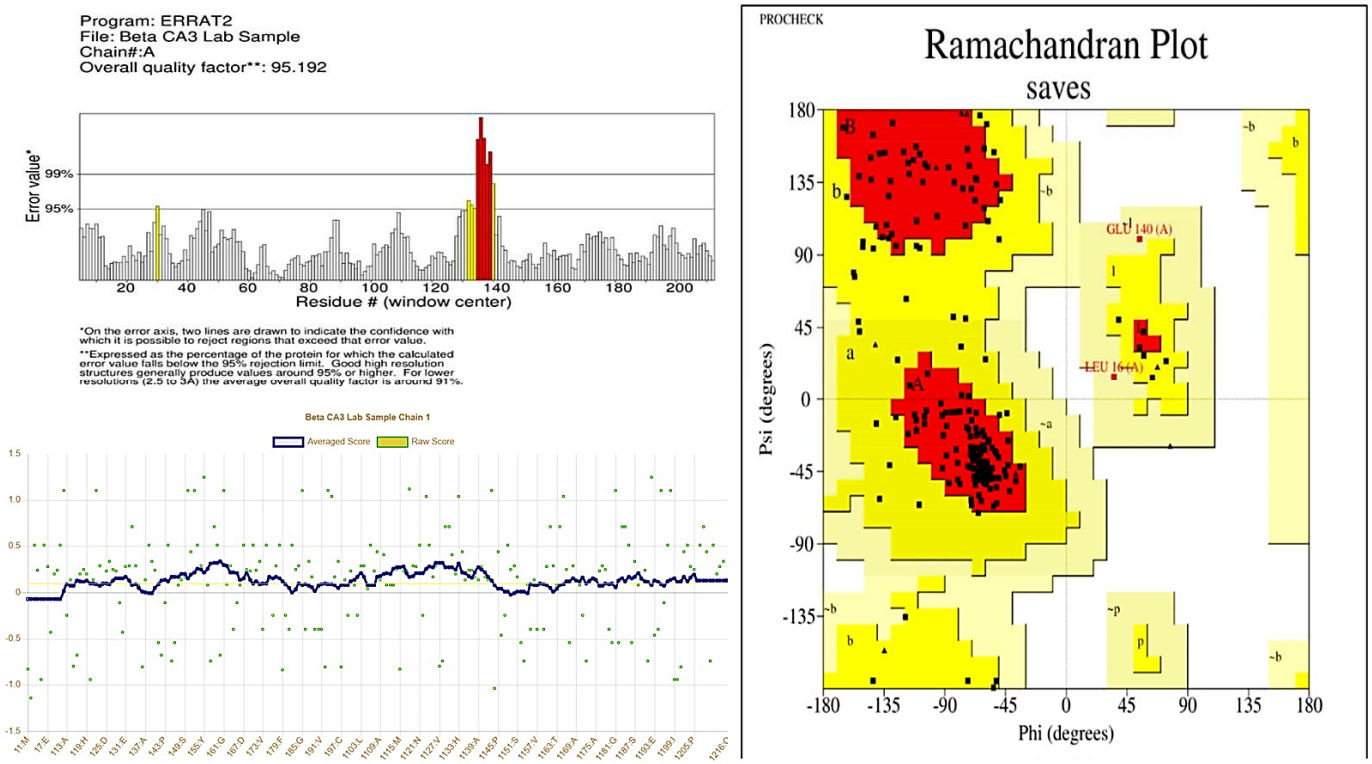


Fig.5. Validation of Modeled β-CA3 Protein with Computational Tools; The Top shows ERRAT analysis of structural errors, the bottom displays VERIFY 3D sequence alignment, and the right illustrates the Ramachandran Plot from PROCHECK.

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

3.8.From Sequence to Function: Deciphering 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, which 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, those 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, as well as its role in cellular processes are assumed. Our study ended with 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 hypothesizing the protein behavior and meaning in cellular conditions. SEQ Eclipse is concluded by the notion that bioinformatics is a tool that provides an initial understanding that needs to be confirmed or

invalidated through experimentation. The true biological activity and the full spectrum of a protein's role cannot be entirely captured through computational means alone.

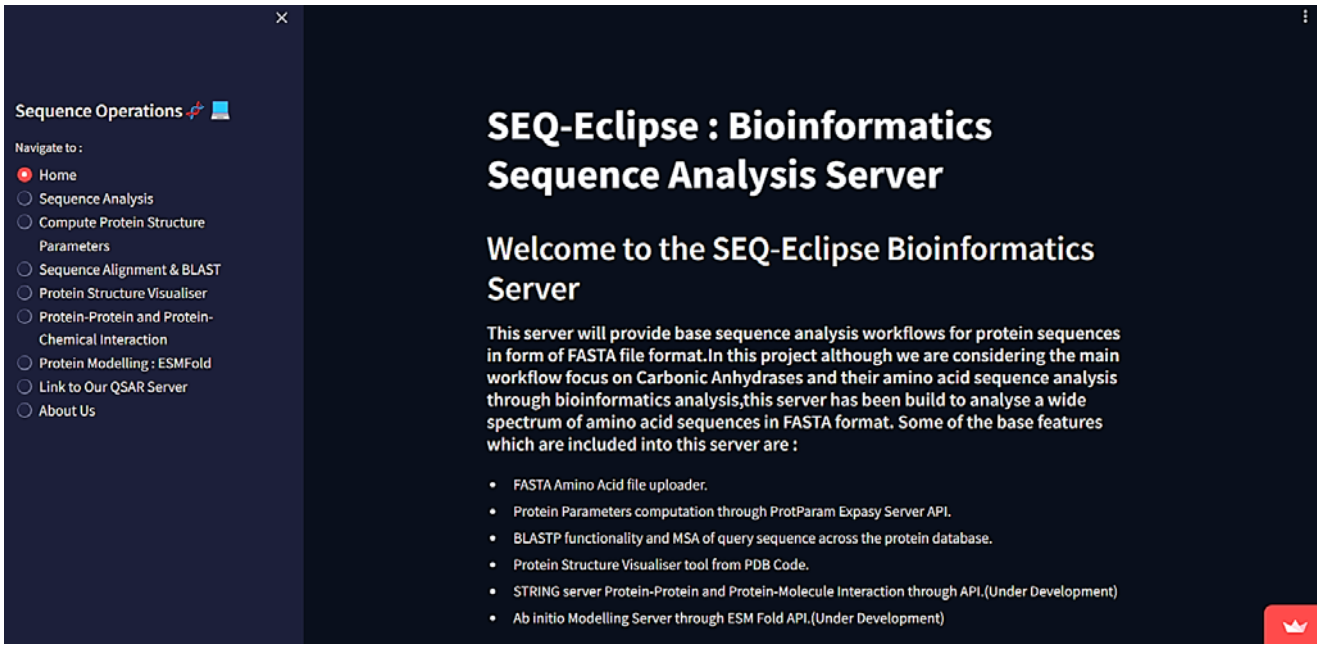


Fig.6. Home Page of SEQ Eclipse.

Hence concluding the study, we present a detailed bioinformatics analysis on β-CA3 to determine its important physico-chemical properties along with its biological functions and molecular interactions against a multitude of other genes and proteins which overall is responsible for maintaining its metabolism and functionality within the living cells. Although bioinformatics analysis is not the complete evidence of interpreting biological activities and various wet lab experimentations are also required to physically validate the data produced in virtual mode, we can get a brief walkthrough of the processes and biological abilities of the carbonic anhydrases and visualize their potential in the biological multiverse.

3.9.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 uncommon chemicals. Correlation coefficients and other statistical parameters were computed for each fingerprint dataset (**Table s1**). Classification analysis plots were utilized to visualize the predictive performance of the machine learning models. Such as the 1D/2D descriptors model and substructure fingerprint model, the plots showed a very strong agreement between predicted and experimental pKi_{50} values. Most data points were on top of the identity line which means that models have very robust predictive ability. In the theoretical molecular 1D and 2D descriptor dataset, the training set correlation coefficient was approximately 97.90%, while for the test set, it was 89.40%. The root mean

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 44.02%. Moreover, 5-fold and 10-fold cross-validations of the training set data produced correlation 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 QSAR for each model was undertaken to know the significance of each descriptor and fingerprint and to identify how they influence the bioactivity. Correlation matrices with both 1D/2D properties representation and substructure fingerprints helped to look for the interdependencies among molecular data sets. 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 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 viral 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 plot, 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 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 the heavy atoms (besides hydrogen atoms) in a molecule thus the size, complexity, and interactions with biological targets might also be different. Metabolites with a raised number of heavy elements can show up in 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) has an influence like molecule's hydrophobicity, shape, flexibility, and finally, its capability 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 the membrane permeability and the compound to interact with the biological targets. ALogP2 is a measure accounting for non-linearity between lipophilicity and biological activity, which 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 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)**.

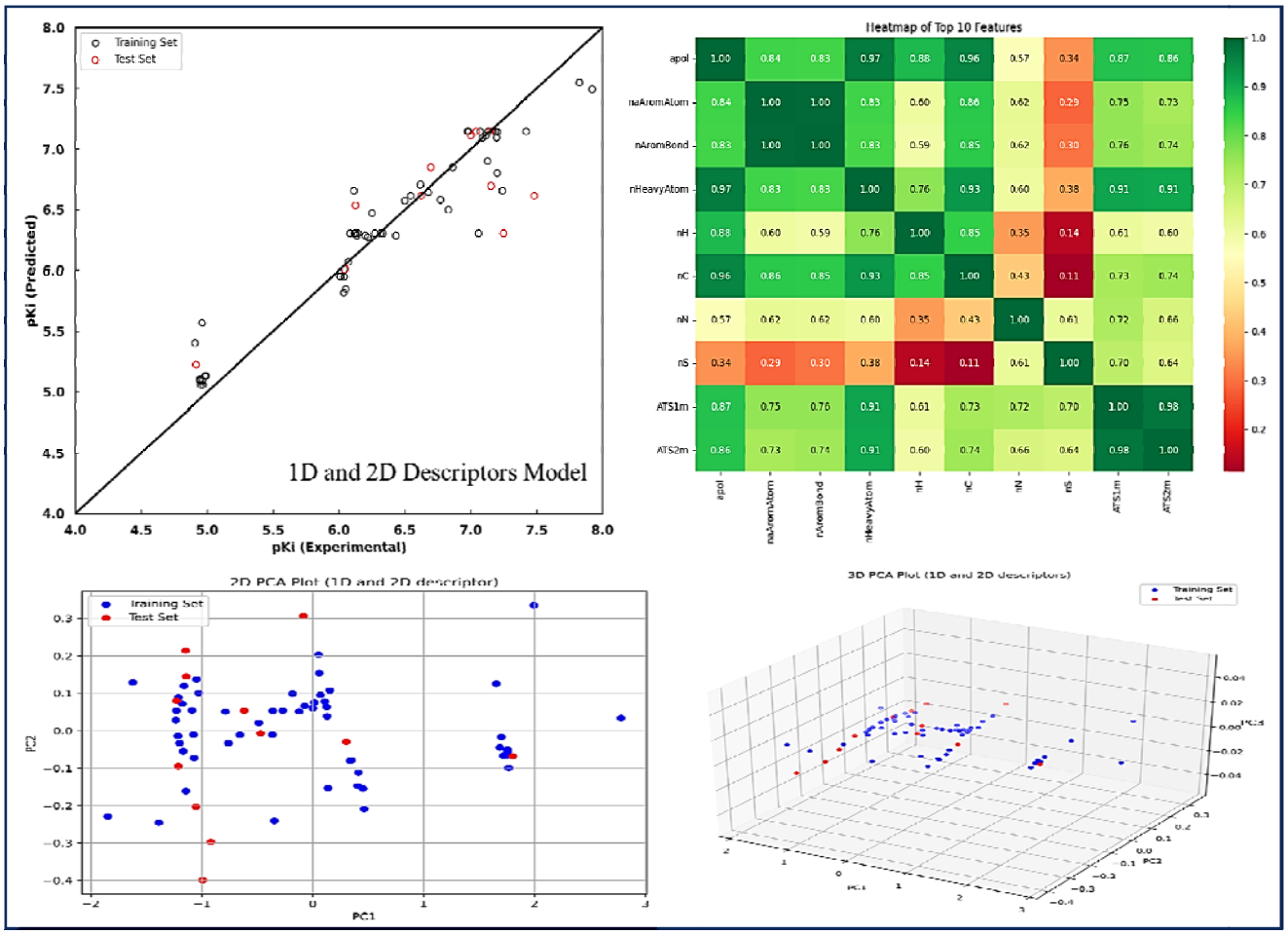
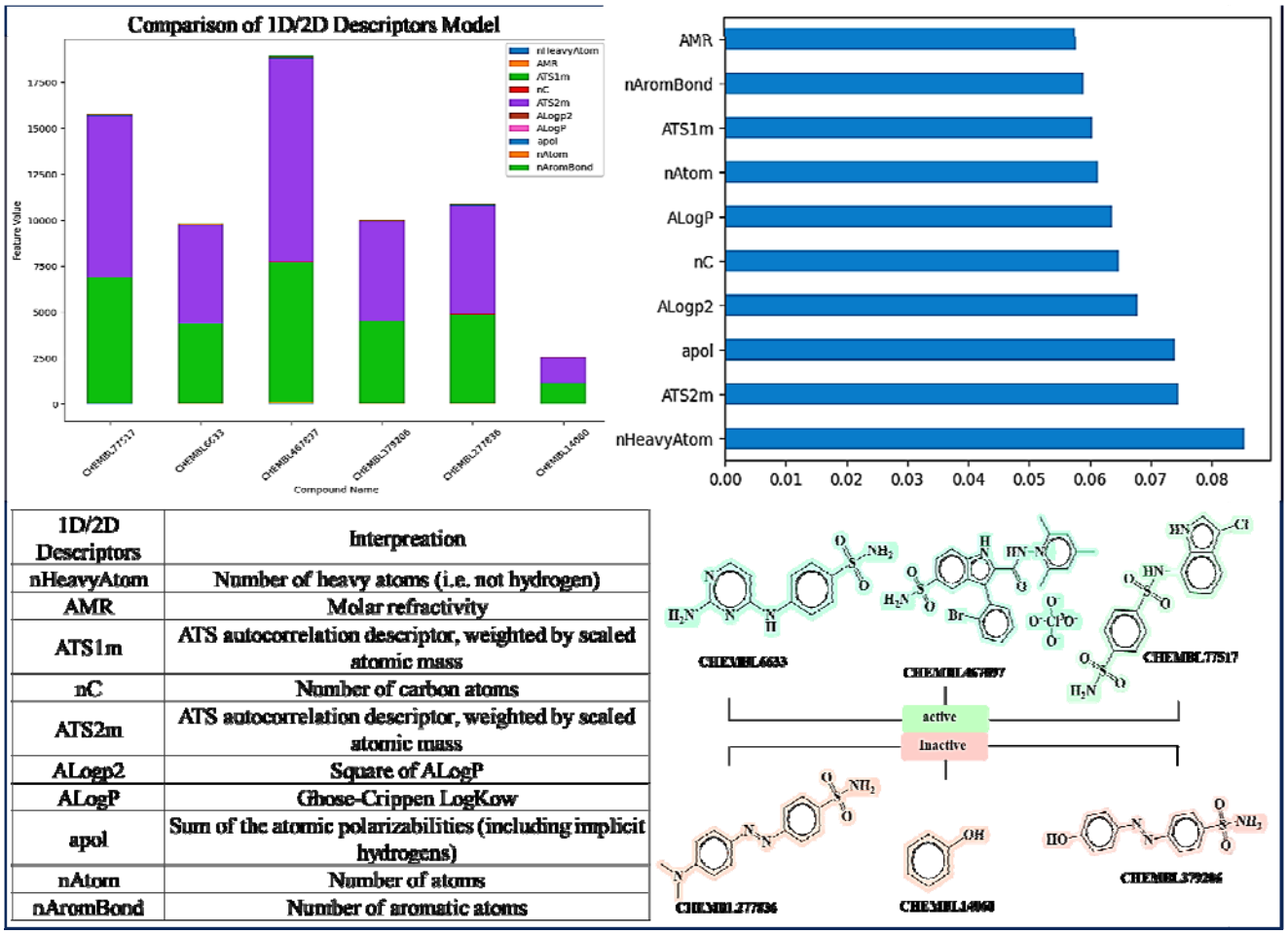


Figure 7: The upper left panel depicts regression plots for various ML-QSAR predictive models. The upper right panel displays a heatmap correlation matrix for 1D/2D descriptors, while the lower left and right panels present 2D and 3D PCA plots, respectively, visualizing the data distribution of 1D/2D 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) plot (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 value of 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 like '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, alongside other substructure fingerprints like '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 analysis of descriptors used in the design of QSAR models. They provide the means for specific chemical features to be observed within the molecules, implying the pinpointing of the structures whose motifs play a role in the

influence on both reactivity and biological activity, while SubFP169's detection of phenol moieties highlights their well-known antioxidant and antimicrobial properties. Similarly, SubFP184 identifies heteroaromatic rings, contributing unique chemical and biological characteristics, and SubFP214 indicates the presence of sulfonic acid derivatives, affecting key physicochemical properties. Additionally, SubFP275's detection of heterocyclic rings underscores their prevalence in pharmacologically active compounds. SubFP287's recognition of conjugated double bonds and SubFP137's revelation of vinylogous ester moieties provide further insights into molecular structures and potential reactivity. Among the analyzed substructure fingerprints, 'SubFP287', 'SubFP184', 'SubFP214', and 'SubFP275' have been identified as pivotal, demonstrating significant importance in the analysis. Additionally, 'SubFP88' and 'SubFP137' also exhibit noteworthy contributions. Further analysis entailed examining the presence of these significant molecular fingerprints among both active and inactive molecules within our dataset. The active molecules, namely CHEMBL512633, CHEMBL20, and CHEMBL77517, targeted β-CA3, showcasing experimental and predicted pKi_{50} values of 8.14, 7.92, and 7.82, respectively, and exhibited 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)**.

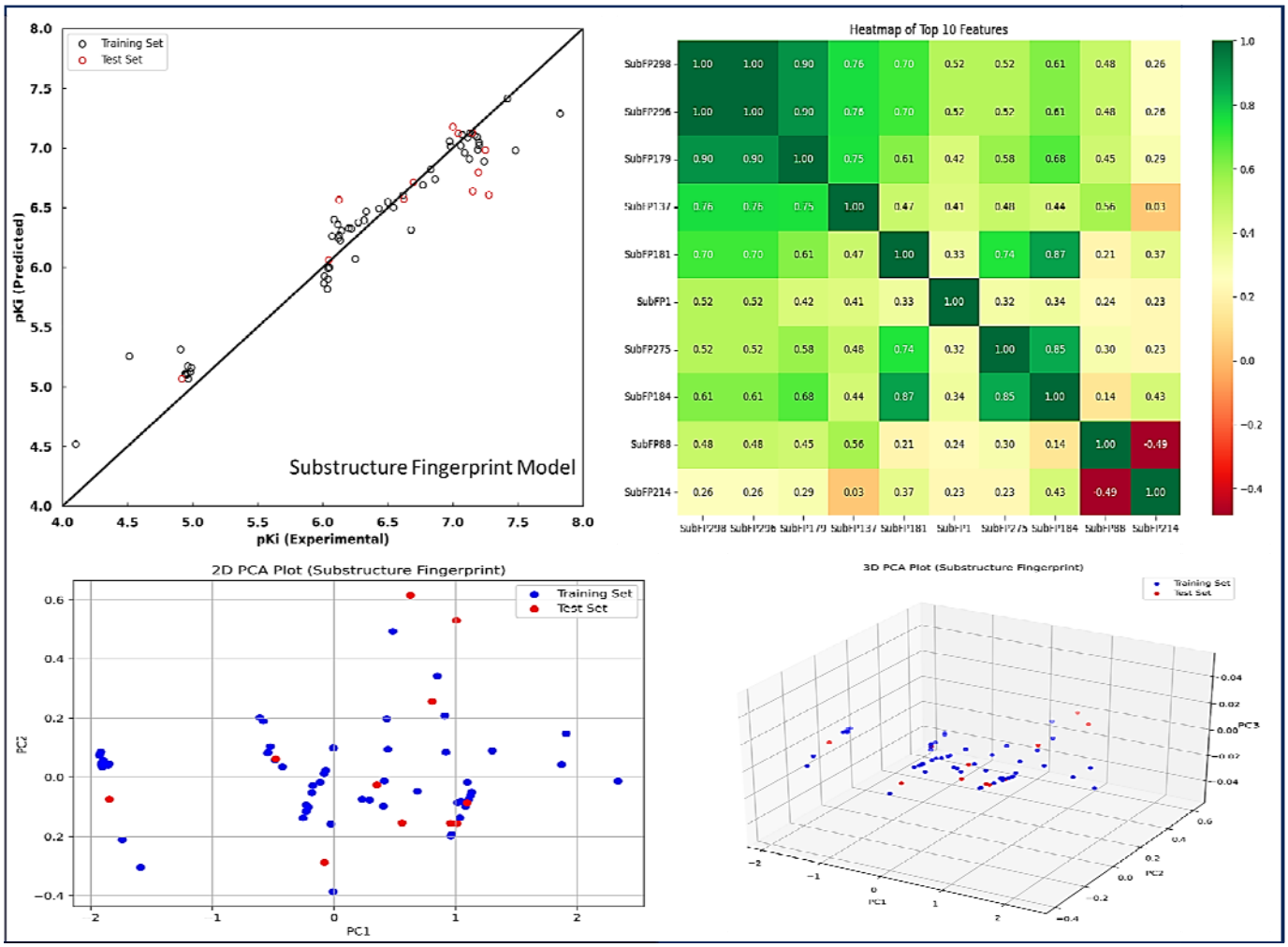
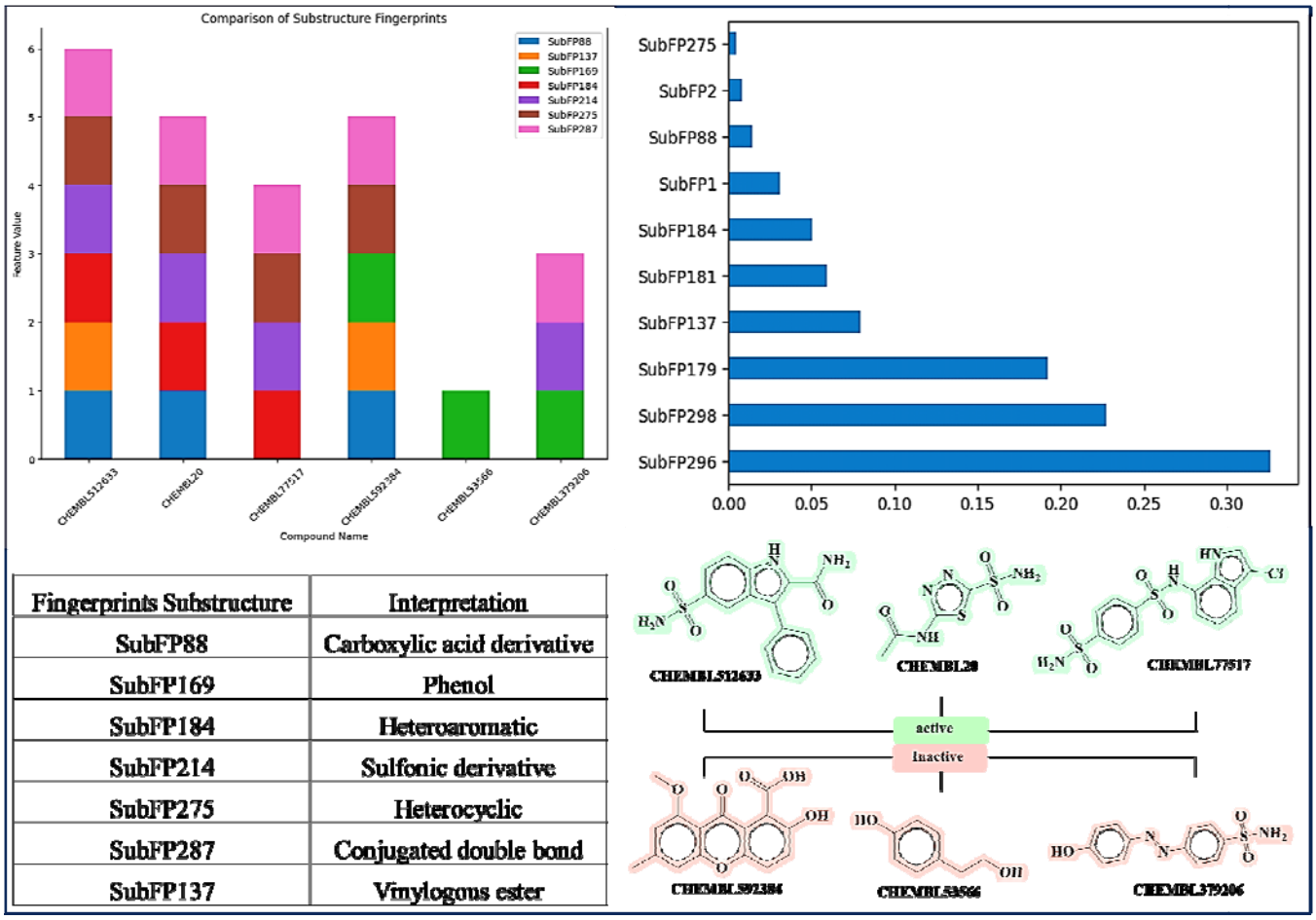


Figure 9: The upper left panel depicts regression plots for various ML-QSAR predictive models. The upper right panel displays a heatmap correlation matrix for substructure fingerprints, while the lower left and right panels present 2D and 3D PCA plots, respectively, visualizing the data distribution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

3.10. System Biology

3.10.1. Functionality analysis of β-CA3 protein and target compound

The Gene Ontology (GO) study has resulted in groups under biological processes, cellular components, and molecular functions for *Mycobacterium tuberculosis* (Mtb), while also providing false discovery rate (FDR) scores. These FDR scores specifically put a spotlight on the most profoundly enriched GO terms, and highlight which processes probably have a substantial effect when it comes to their implication regarding the biology of this species. Biological reactions, such as fatty acids and lipids' biosynthetic pathways, have been significantly enriched with FDR values of 1.73 $*10^{-23}$ and 1.06 $*10^{-15}$ respectively, which indicate high activity of these processes in the organism. This might be important for the process of cell wall synthesis and membrane regulation and has been tolerated by *Mycobacterium tuberculosis* through stress conditions, which can be seen all around. Other processes with a high representation include DIM/DIP cell wall layer assembly and respiration under aerobic conditions; most possibly because of the preservation of the cell

synthase complex 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. Where the molecular function side is concerned, among several others, phosphopantetheine binding, and 3-oxoacyl-[acyl-carrier-protein] synthase activities are important, and interestingly, they also happen to be one of the most enrichment activities, showing how the organism 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**). Highly connective 20 genes at the center of metabolic functions of M. tuberculosis have been extracted from the network (**Figure 11**). Amongst 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 secrete Siderophore peptides in heavy metal availability regulation, and the metabolism of nitrogenous compounds. Oxidative phosphorylation stands out with 14 genes having a strong impact on the process loads of p-value (8.59×10^{-11}), which is supported by a small (0.05) FDR score thus highlighting the pathway's fundamental role in bacterium energy generation. Regardless of the fact it’s not directly shown on the diagram, β-CA3 is regarded as the most significant thing that is responsible for maintaining the cellular pH in the given bacterium. It could be possible that β-CA3 implements controls over these pathways by playing a regulatory role. Another example is enriching 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' remains as one of the six annotations where 26 genes from the PPI network were based, which suggests a strong implication on the metabolic systems impact of M.tb. The nitrogen metabolism road, though only second to oxidative phosphorylation in terms of the extent of its representation, still plays an important function in the body’s metabolism. This fact shows that the gene is actively associated with an assimilation and utilization process for nitrogen, thus digitalizing the infection and survival in the new host. The relation between the CO2 regulating function of β-CA3 and the 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 is particularly important for understanding the metabolism of M.tb which is significantly dependent on β-CA3, which could potentially be exploited for anti-TB strategy development.

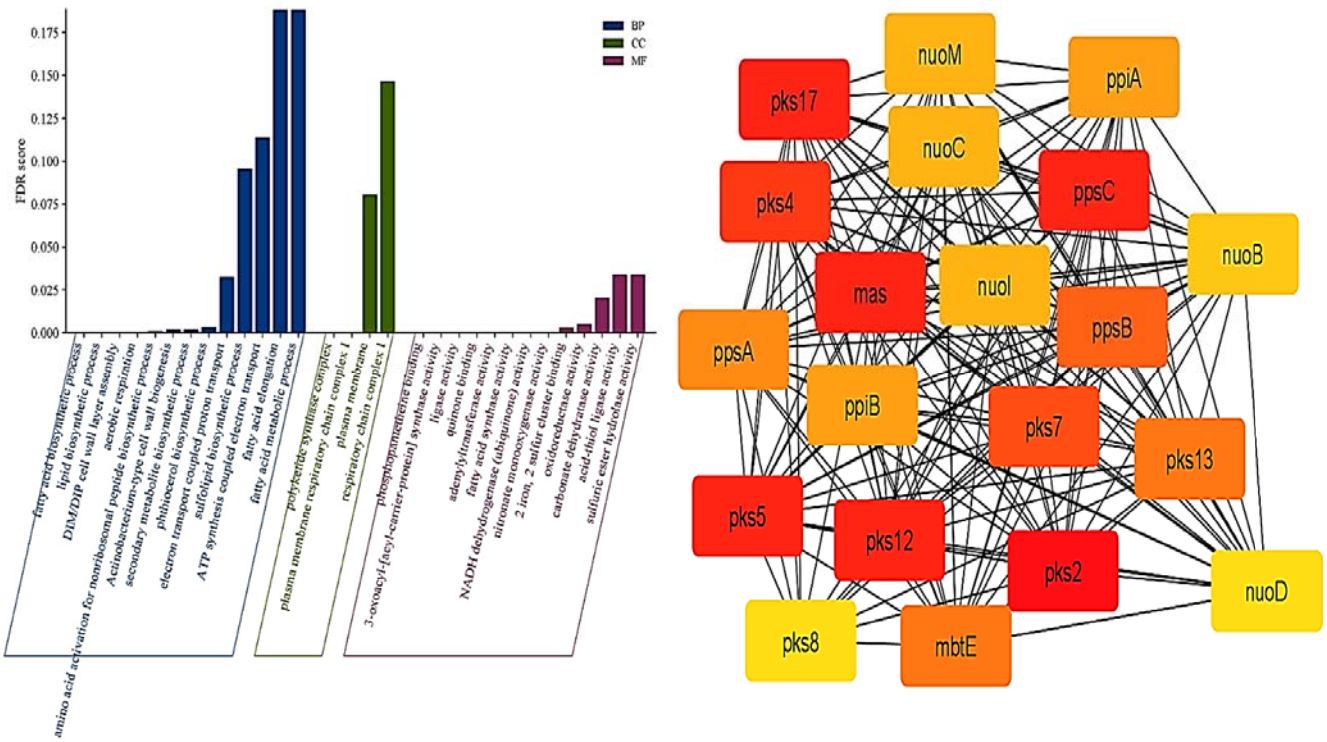


Fig. 11: Gene Ontology Analysis of *Mycobacterium tuberculosis* β-CA3; Right Panel: Common Target Gene Visualization.

3.10.2. Identification of Hub genes

We curated a set of 21 common genes from both the target compounds and target genes specific to the disease β-Carbonic anhydrases-3 in M. Tuberculosis. 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 disease process. Four genes, namely NOS2, CTSG, ABCB1, and JAK1, were identified as common targets among compounds, biofilm formation, and β-Carbonic Anhydrase-3 in M. tuberculosis (**Figure s17 and Table s6**).

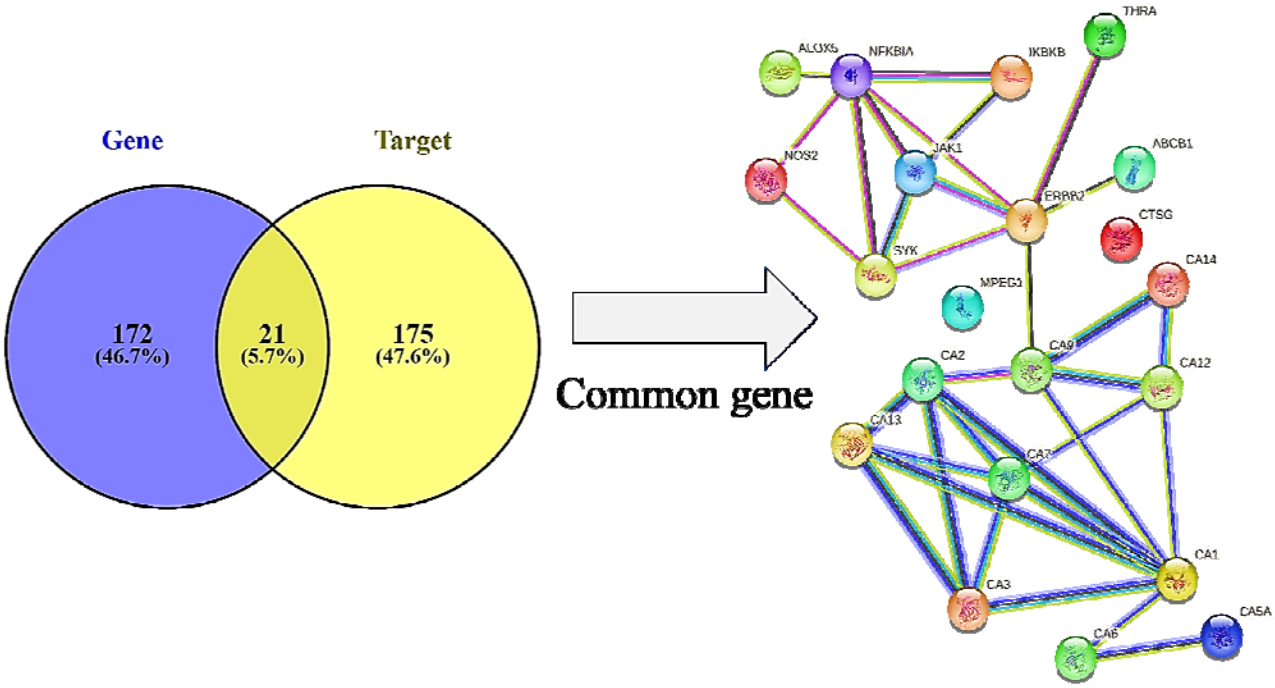


Fig. 12. The right panel shows a network of common genes, while the left panel displays a Venn plot diagram highlighting overlapping targets between the potential compound and disease-related genes.

3.10.3. Insights from Gene Ontology and KEGG Analyses

We conducted a screening of 21 genes through GO and KEGG enrichment analyses, ranking them by ascending p-value. The KEGG analysis revealed 81 significantly enriched pathways ($P<0.05$), aiming to uncover the mechanisms behind the inhibitory effects of potential compounds on the β-Carbonic anhydrases-3. A functional enrichment analysis was carried out on the top 21 target genes using the g:profiler. Gene Ontology analysis has highlighted the critical role of β-carbonic anhydrase 3 (β-CA3) in the survival and virulence of *Mycobacterium tuberculosis* (M.tb), underlined by the most statistically significant enrichment for carbonate dehydratase activity (GO:0004089), with a profound adjusted p-value of 1.069×10^{-25} **(Table 4)**. This is one of the essential things that M.tb uses for pH regulation by reversibly absorbing carbon dioxide resulted because of the behavior of the host cell environment. Such an exceptionally low p-value reflects a very high level of statistical confidence, signaling this enzyme's activity as 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, 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 such as CA1, CA2, CA3, CA9, and others, associated with carbonate dehydratase activity, are essential. Additionally, transition metal ion binding (GO:0046914), with genes including ALOX5 and THRA also implicated, is significantly enriched (padj of 4.82E-08), emphasizing the importance of metalloenzymes and zinc-binding proteins like β-CA3 in M.tb's

balance in the bacterial cellular fluid, which, in turn, has a role in pathogen survival within the acidic environment of macrophage. While 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 M.tb's intracellular survival, 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, suggests a strong association with the bacterium's metabolic processes. This 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 has accentuated the enrichment of the Metabolic pathways (KEGG:01100), further highlighting the vulnerability of M.tb's metabolic network, which could be disrupted to curtail the bacterium's energy production and biosynthesis capabilities. 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 M.tb. 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 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 suggests a strong potential for selective inhibition of β-CA3 as a strategy for anti-tubercular therapy. The association with the most enriched metabolic pathways further amplifies the importance of this target, as disruptions to these pathways could critically affect the survival and virulence of M.tb. Targeting β-CA3 selectively presents a dual benefit: not only does it directly disturbs essential enzymatic activity of M.tb, but indirectly disturb its metabolic flexibility as well. By diving deeper into the function of β-CA3 in those essential pathways we will be able to devise a drug development strategy which will be more effective at tackling tuberculosis.

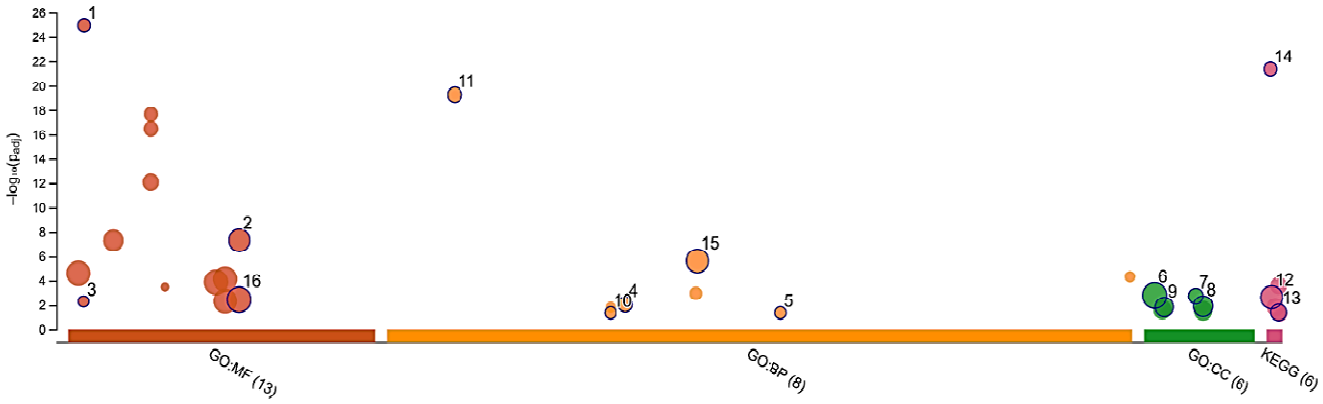


Fig.13. Gene Ontology Categorization and Pathway Enrichment Scatterplot.

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	$3.30 * 10^{-17}$

3.10.4. Compound-Target-Pathways analysis

We addressed the relation between compounds, targets, and pathways that we structured 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 the action against β-Carbonic anhydrases-3, which introduces possible different therapeutic interventions. The tool Cytoscape permitted designing a detailed diagram to display the mechanisms including bindings found especially between drug candidates from the ChEMBL repository and their targeting sites within the TB bacteria. This extensive map is concentrated on analyzing the connections of these compounds with a gene of the bacteria called ß-Carbonic anhydrases-3(ß-CA3). Map had 42 nodes, each node representing

density score is equal to 0.0, meaning that it has few point-to-point connections and assumes 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 ovalites, experimentally determined genes are circles with yellow colors of different numbers of connections, and the other genes are circles with brown colors, shapes of triangles representing biofilm formation gene in tuberculosis and biological processes are rectangles with pink. The clustering coefficient is not 0, but 0.197, the reality is that while there are some clusters where the points are more closely connected, the network does not have tight clusters. Network analysis shows 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 takes place. 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 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 other 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 the other ones, which can explain a balanced network where no single node is too powerful in the network. Besides, between 10 of these node pairs there exist multiple connections, which could mean that some interactions are notably stronger or perhaps quite important among others for the network's integrity. **Figure 14b** reveals the top 10 most interactive routes within the network, showing the genetic connection with the targets and pathways more than any other. 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 the dynamic model that can represent the characteristics of multiplexing and multitargets directly and could be applied to system 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 degree of freedom of each common target was shown in the **Table s7**.

The compounds CHEMBL77517, CHEMBL512633, and CHEMBL379206 have signified 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 on M.Tb particularly target β-Carbonic anhydrases (β-CHEMBL77517 has proven to exhibit interactions 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 probably cause the homeostatic state of the bacterium to be disrupted (**Figure 14a**). In this respect, the influence of NFKBIA attracts special interest as this gene takes a part in regulating inflammatory reactions[53]. CHEMBL77517 might possibly tap the role of NFKBIA resulting in the host's ability to evade the M.tb. The

special relation with ERBB2, which isn't the main element of the M.tb mechanism of action, but it can influence cell signaling, and in this way, the host-pathogen communication can be changed. 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 focusing on several targets among them the gene SYK which is involved in the control of the signaling from the immune cells as well as the biofilm formation, which is of importance, as M. tb is known to create biofilms during infections. The targeting of CA2 and CA9 by CHEMBL512633, like 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 the other two compounds, the targeting of CA2 and CA9 by CHEMBL379206 indicates a possible interference with the bacterial carbonic anhydrase system. The interactions with CA12 could have an additional impact on the bacterium's metabolism.

The metabolic pathway highlighted in the graph is directly connected to several genes: These three regions are namely: CA2, CA9, and CA12. The genes interact with the molecules CHEMBL77517, CHEMBL512633, and CHEMBL379206 in particular which is forming 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 *Mycobacterium tuberculosis*. The interactions of CA2 and CA9 genes with the metabolic pathway suggest a direct influence on M.tb's ability to regulate its internal environment. Compounds targeting these genes could impede the bacterium's metabolic reactions, potentially leading to a weakened state that could be more susceptible to treatment. As a result, the bacterium's state may be weakened, making it more vulnerable to treatment. Compounds affecting CA12 could, therefore, have indirect effects on M.tb's metabolism. Given that the SYK pathway is not a redundant pathway, targeting this pathway could reinforce the susceptible immune responses mediated by the M.tb, which, in the end, may be directed to controlling the infection. Since the SYK pathway is nonredundant, targeting it could strengthen the immune system's specific responses against M.tb, 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[54,55]. Biofilms are significant in the context of tuberculosis because M. tuberculosis can reside within biofilms in the human body, contributing to the chronic nature of the disease and its resistance to treatment[56,57]. The network map is articulated and proves the SYK gene to have complex interactions with the chemical CHEMBL512633 (**Figure 14b)**. This adjacent process implies that the compound can be efficient in switching on and off of SYK, which eventually results in the reduction of the biofilm formation and enhancing the immune response. If CHEMBL512633 can modulate SYK activity, it may affect the formation and maintenance of biofilms. Disrupting biofilms could expose M. tuberculosis to the immune system and antibiotics, potentially making infections easier to clear. The interpretation of the interactions within the network provides an insight into the multi-faceted way that these compounds could be combating

M.tb. Through the selective targeting of genes involved in metabolic control and immune response, these compounds might represent the strategy that not only inhibits bacteria's activities but also supports the body's innate mechanisms to get rid of the infection. The network shows that inhibiting these genes would disturb the metabolic stability of M.tb thereby limiting its potential to infect the host and survive.

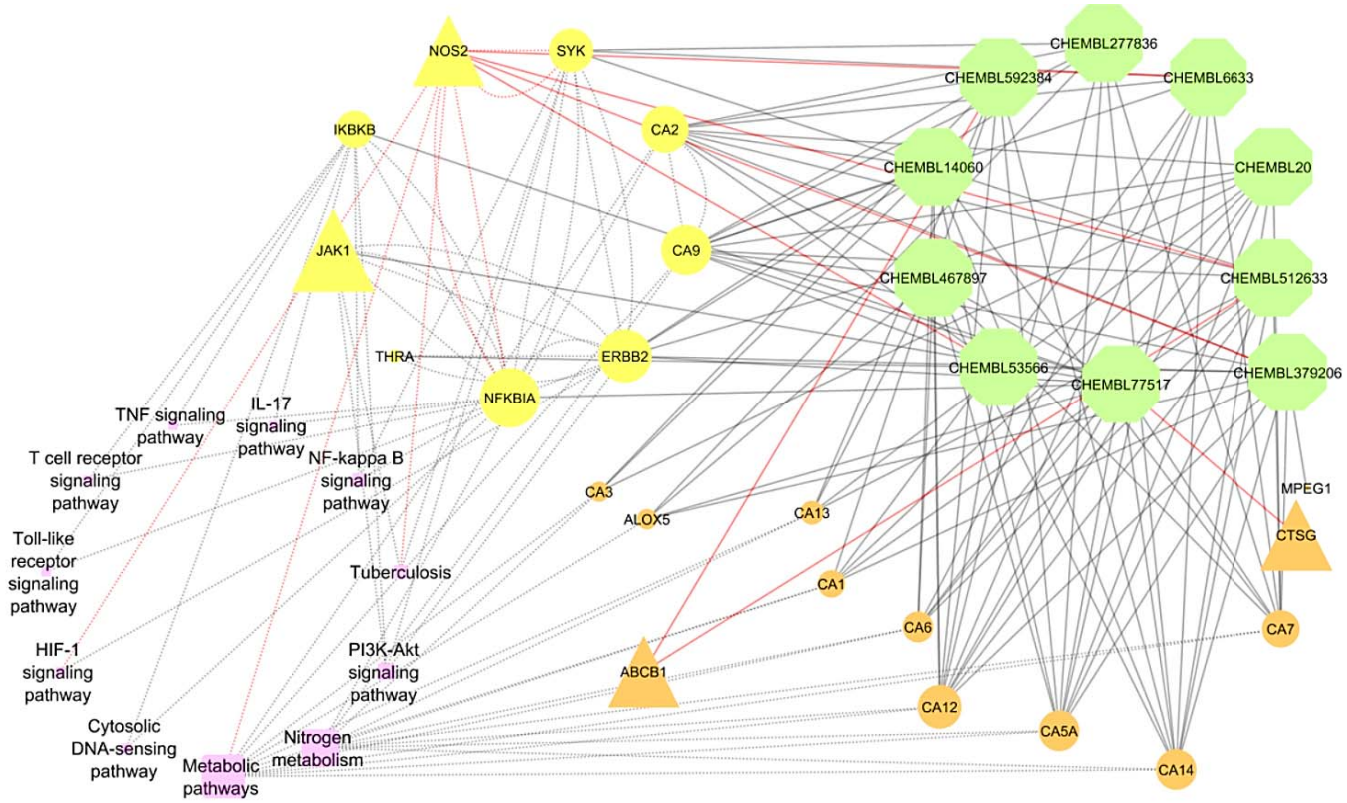


Figure 14a: Interactive network map displaying gene and compound associations with key biological pathways. Highlighted yellow colors show an experimentally determined gene brown color shows no experimental determined gene an octagon shape of green color shows the compound and a square of pink color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

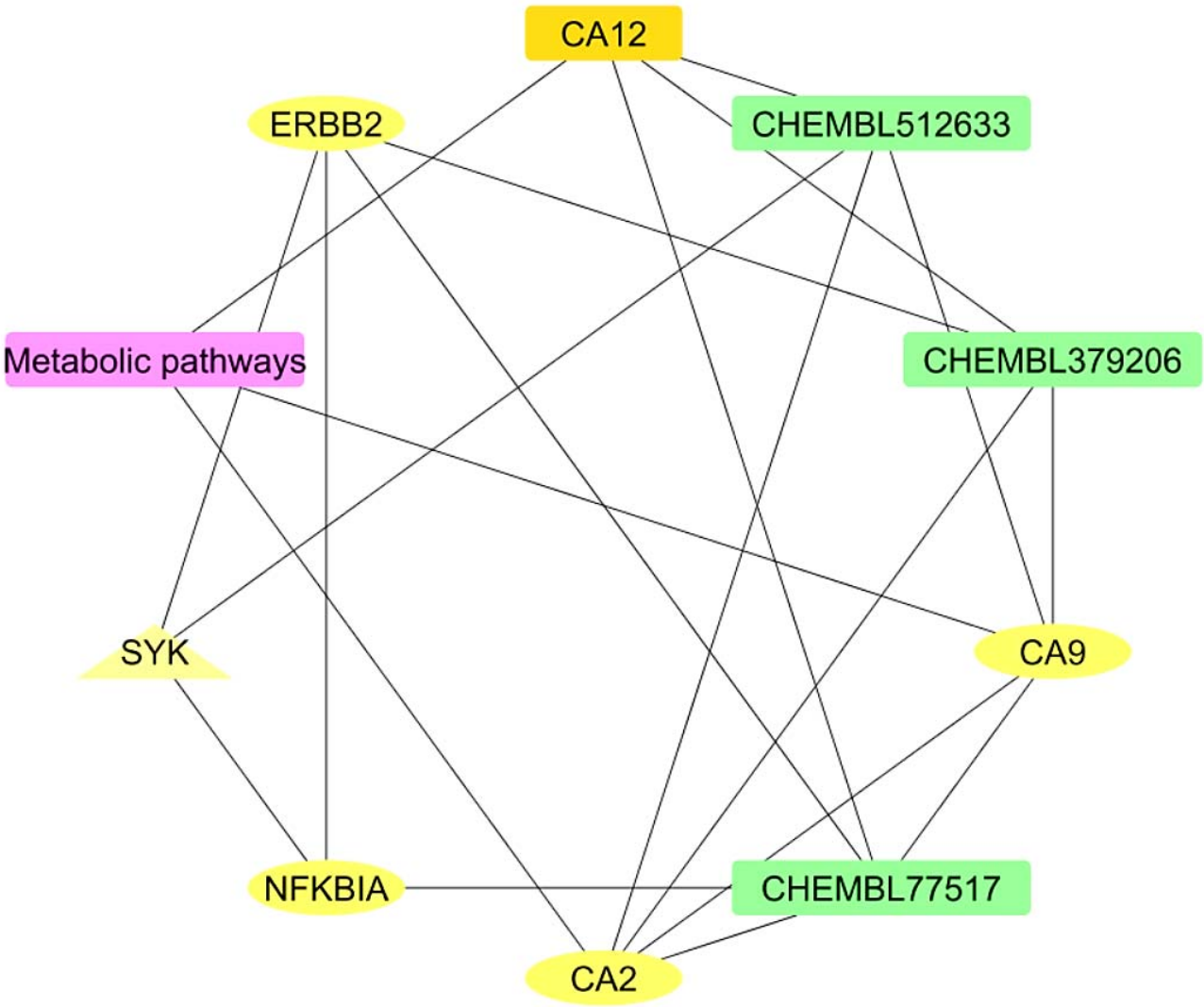


Figure 14b: An Image focuses on the top 10 genes by degree of connectivity, including three compounds (highlighted in green) that demonstrate the most frequent interactions with these genes and pathways.

Conclusion

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

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, 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 essential for M. tuberculosis virulence and survival. Despite the interest gained information, the results are still preliminary and need to be confirmed in experiments. The silico predictions, although promising in generating a robust biological effect, should be complemented by wet lab experiments to understand the diverse activities of the compounds and validate any therapeutic potentials they may have. Future research should focus on experimental validation of these computational findings and explore the clinical applicability of the promising compounds.

Acknowledgment: The authors extend their sincere appreciation to the following organizations for their invaluable support and funding: 1) Finnish Cultural Foundation, 2) Tampere Tuberculosis Foundation, (Ashok Aspatwar); 3) Academy of Finland, and 4) 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) 7–7. 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/BIBTEX.

[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/.

[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. 2019, Vol. 20, Page 5153. 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 18 (2008) 729-741. https://doi.org/10.1101/gr.075069.107.

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

[11] 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). https://doi.org/10.3390/MOLECULES23112911.

[12] 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.

[13] 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). https://doi.org/10.7717/PEERJ.4128.

[14] 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.

[15] 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.

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

[17] 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.

[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] P.J. and P.S. Aspatwar A, Qureshi KA, Parvez A, 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.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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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. 2022 1710. 17 (2022) 2326–2353. https://doi.org/10.1038/s41596-022-00728-0.

[40] 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). https://doi.org/10.1016/J.CRMETH.2021.100014.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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 (80-.). 379 (2023) 1123–1130.

https://doi.org/10.1126/SCIENCE.ADE2574.

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

[48] 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.

[49] 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.

[50] J. Spiegel, H. Senderowitz, Evaluation of QSAR Equations for Virtual Screening, Int. J. Mol. Sci. 21 (2020) 1–20. https://doi.org/10.3390/IJMS21217828.

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

[52] 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.

[53] 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). https://doi.org/10.15252/EMBJ.2019104296.

[54] A. Mócsai, J. Ruland, V.L.J. Tybulewicz, The SYK tyrosine kinase: a crucial player in diverse biological functions, Nat. Rev. Immunol. 2010 106. 10 (2010) 387–402. https://doi.org/10.1038/nri2765.

[55] M. Wagener, J.C. Hoving, H. Ndlovu, M.J. Marakalala, Dectin-1-Syk-CARD9 signaling pathway in TB immunity, Front. Immunol. 9 (2018) 332218. https://doi.org/10.3389/FIMMU.2018.00225.

[56] A. Solokhina, D. Brückner, G. Bonkat, O. Braissant, Metabolic activity of mature biofilms of Mycobacterium tuberculosis and other non-tuberculous mycobacteria, Sci. Reports 2017 71. 7 (2017) 1–7. https://doi.org/10.1038/s41598-017-10019-4.

[57] 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. 2021 121. 12 (2021) 1–17. https://doi.org/10.1038/s41467-021-21748-6.