参赛学生姓名：_Jeffrey Weizheng Li_____

中学：___复旦大学附属中学国际部_____

省份：_____上海市_____

国家/地区：____中国/南方_____

指导老师姓名：__张鹭、赵婷玉_____

指导老师单位：_复旦大学生命科学学院、复旦大学附属中学国际部_____

论文题目：AI-Guided Design and Preliminary Validation of Anti-Tuberculosis Subunit Vaccine

本参赛人声明所提交的论文是在指导老师指导下进行的研究工作和取得的研究成果。以本参赛人所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果。若有不实之处，本人愿意承担一切相关责任。

参赛人：Jeffrey Weizheng Li

指导老师：张鹭、赵婷玉

2024 年9 月14 日

AI-Guided Design and Preliminary Validation of Anti-Tuberculosis Subunit Vaccine

Jeffrey Weizheng Li

Fudan International School

Abstract:

Objective: To use AI technology to evaluate the structural proteins of the ESX-1 secretion system of *Mycobacterium tuberculosis* and experimentally assess the potential of EccA1 as a preventive DNA vaccine against tuberculosis. **Methods:** The immunogenicity of ESX-1 secretion system proteins was predicted using AI online software programs. The structural component antigen EccA1 protein, which showed excellent prediction results, was further evaluated for its biological characteristics. The encoding gene Rv3868 of the EccA1 protein was amplified by PCR from the *Mycobacterium tuberculosis* genome, and a recombinant plasmid pcDNA::Rv3868 was constructed using the eukaryotic expression vector pcDNA3.1+. After gene sequencing, plasmid was extracted using a plasmid extraction kit to obtain endotoxin-free plasmid. The plasmid was transfected into the HEK293T cell line, and the expression of the EccA1 protein was verified by Western-blotting experiments and its cytotoxicity was assessed. The plasmid was injected intramuscularly to immunize zebrafish larvae, and the safety at the animal level of the recombinant plasmid was evaluated by observing the survival of zebrafish. The larvae were infected with *Mycobacterium marinum* 24 hours after immunization, and the change in the weight of the larvae was used as an evaluation index; adult fish were infected with *Mycobacterium marinum* 3 weeks after immunization, and the change in the bacterial load in the adult fish was used as an index to evaluate the preventive protective ability of pcDNA::Rv3868 against tuberculosis. **Results:** The EccA1 protein is predicted to be a secreted antigen. The antigen contains both strong B cell and T cell epitopes. pcDNA::Rv3868 can be highly expressed in eukaryotic cells and is safe for cell lines and zebrafish larvae. The weight of zebrafish infected with *Mycobacterium tuberculosis* after immunization with pcDNA::Rv3868 was higher than that of the negative control group, and the bacterial load in the body

significantly decreased. **Conclusion:** EccA1 is a very promising target for an anti-tuberculosis vaccine candidate.

Keywords: Tuberculosis; AI-guided antigen design; DNA vaccine; EccA1; Zebrafish

Table of Contents

I. Origin of the Study

Seven years ago, after I settled in Shanghai with my parents from the United States and visited my relatives in my hometown in Hunan, I noticed that some middle-aged uncles were constantly coughing even during the hot summer vacation. I learned from my father's conversations with them that they had all suffered from pulmonary tuberculosis. Although they had been cured years ago, they often relapsed a few months after stopping their medication, and there were no better drug treatments available. I often wondered if there was a vaccine as effective as the annual flu shot to prevent tuberculosis?

After entering high school, I learned about vaccines and antibody drugs in my science classes. Occasionally, from the WeChat public accounts my father read, I saw news that the World Health Organization reported a rebound in the global incidence and death rates from tuberculosis three years after the COVID-19 pandemic. I then contacted the laboratory at Fudan University that specializes in tuberculosis vaccine research, hoping to explore a useful treatment method.

II. Background Introduction

Upon reviewing the literature, I found that pulmonary tuberculosis (TB), caused by *Mycobacterium tuberculosis* (M.tb), is an ancient respiratory infectious disease. Patients expel M.tb from their respiratory tract through coughing and sneezing. The aerosols containing M.tb may survive for several hours under the right conditions, and once inhaled, they can settle in the lungs of a new host, leading to tuberculosis. The World Health Organization (WHO) reported that in 2021, there were about 10 million new cases of tuberculosis worldwide, with the most in the Southeast Asia region, about 4.5 million, while the Americas had about 290,000 and Europe only 220,000. Eight countries account for two-thirds of the total number of new cases globally, in descending order: India (28%), Indonesia (9.2%), China (7.4%), Philippines (7.0%), Pakistan (5.8%), Nigeria (4.4%), Bangladesh (3.6%), and the Democratic Republic of Congo (2.9%). In the past two years during the COVID-19 pandemic, the number of tuberculosis deaths has also increased, from 1.4 million in 2019 to 1.6 million in 2021. As one of the high-incidence countries of tuberculosis, China had about 740,000 new

cases of tuberculosis in 2021, although it decreased by 100,000 compared to 2020, it is still a long way to achieve the global goal of ending the tuberculosis epidemic by 2035.

Vaccines are very necessary for the global prevention and control of tuberculosis. The only vaccine currently licensed for the prevention of tuberculosis is the Bacillus Calmette-Guérin (BCG) vaccine. From 1908 to 1920, the Calmette and Guérin team passed the virulent *Mycobacterium bovis* (M. bovis) through 230 generations to obtain the attenuated live bacterial vaccine—BCG. Since its introduction in 1921, BCG has been used for 103 years. China has included BCG vaccination in the national immunization program, and it is generally stipulated that newborns should complete vaccination before the age of three months, hence BCG is also known as the first shot after birth. Studies have found that BCG has good protective effects against childhood miliary tuberculosis and tuberculous meningitis, but with increasing age, its protective effect against adolescent and adult pulmonary tuberculosis also gradually weakens, with a protective effect ranging from 0-80%. Secondly, vaccination with BCG is not safe for immunodeficient patients, such as those positive for HIV. Newborns infected with HIV have a higher risk of developing disseminated BCG disease after vaccination at birth. Therefore, new vaccines are needed to supplement the shortcomings of the existing BCG vaccine.

The main strategy for the development of a new generation of anti-tuberculosis vaccines internationally is to replace the existing BCG vaccine with a recombinant BCG vaccine, or to add subunit vaccines to enhance and extend the protective effect of the recombinant BCG vaccine. Under such a strategy, the development of subunit vaccines has attracted much attention. After nearly 30 years of research by scientists worldwide, nearly 20 new anti-tuberculosis candidate vaccines have entered different stages of clinical development, most of which are subunit vaccines. Subunit vaccines are vaccines in the form of proteins or nucleic acids, with adjuvants or packaged in viruses, which can be used to build preventive and therapeutic vaccines, and can also be used to strengthen the protective effect of existing BCG vaccines. Subunit vaccines are safer, easier to control the dose, and can also induce a longer-lasting and stronger immune response. Which functional antigens are used as the core components of

subunit vaccines will directly affect the anti-tuberculosis effect of the vaccine and are the key to the successful development of new vaccines.

The ESX-I secretion system, also known as the Type VII secretion system, can secrete the unique early secreted antigens ESAT-6 and CFP-10 of M.tb. By comparing the genomes of *Mycobacterium tuberculosis* H37Rv, H37Ra, *Mycobacterium bovis*, and its attenuated strain M. bovis BCG, it was found that all BCG strains lack a 9505bp region, called the Region of Difference 1 (RD1), which includes the ESX-1 secretion system and plays an important role in the pathogenic process of mycobacteria, related to bacterial virulence. Although there is still a lack of a thorough understanding of the ESX-I system, it has been clear that it plays an important role in the interaction between pathogens and hosts, which has laid its important immunological value. The recombinant BCG or subunit vaccine based on the ESX-1 system secreted antigens ESAT-6 and CFP-10 has shown enhanced immunogenicity in animal models.

The ESX-1 system includes up to 25 proteins, among which the conserved components of ESX-1, EccA1, EccB1, EccCa1, EccCb1, EccD1, and EccE1, help ESAT-6 and CFP-10 cross the inner membrane, and then pass through the outer membrane through the channel formed by the self-assembly of EspC, and are secreted outside the cell to function. Relatively, the immunogenicity of these conserved protein components has been little concerned. However, there have been reports in the literature that they may be related to the virulence of tuberculosis bacteria. Their vaccine value needs to be systematically evaluated.

With the rapid development of biotechnology and artificial intelligence (AI), the field of antigen vaccine design is undergoing an unprecedented transformation. AI technology, with its powerful data processing capabilities and learning algorithms, has shown great potential in various aspects of vaccine development, especially in antigen prediction and design, immune simulation and evaluation, vaccine platform integration. Protein antigen epitopes are a key part of vaccine design, directly affecting the immune efficacy of the vaccine. Utilizing AI technology, such as online software like BepiPred Server and SEPPA, allows for the direct prediction of antigen epitopes from protein sequences. These predictive results not only facilitate the rapid screening of potential

antigenic peptides but also guide subsequent vaccine design.

At the outset of my experiment, I employed AI methods like AlphaFold3 and SEPPA to systematically compare and assess all antigens of the ESX-1 system. In addition to the strongly immunogenic components such as ESAT-6 and CFP-10 validated by AI algorithms, other conserved structural components like EccA1, EccB1, EccCa1, EccCb1, EccD1, and EccE1 were also indicated to possess a certain level of immunogenicity. Particularly, EccA1 was predicted to have stronger B and T cell epitopes among these proteins, making it a promising candidate for the development of subunit vaccines worth exploring.

Further literature analysis shows that EccA1 (encoding gene: Rv3868) is widely present in the Mycobacterium genus. In the ESX-1 system-mediated secretion process of virulence-related proteins, the ATPase activity of EccA1 is essential, providing the necessary energy for protein secretion. Surprisingly, after the EccA1-ATPase mutation, the synthesis of virulence lipid mycolic acid in the extracellular membrane of the *Mycobacterium marinum* is defective, and the bacterial virulence is reduced. This suggests that in addition to being an essential ATPase in pathogenic mycobacteria, EccA1 protein also has an important virulence-determining function. Further evaluation using various bioinformatics analysis methods has found that EccA1 may be a secreted immunogen with the potential to become a core component of an anti-tuberculosis vaccine. Therefore, this project further constructs the recombinant plasmid pcDNA::Rv3868 as a candidate DNA vaccine and verifies the expression of the candidate DNA vaccine in the HEK 293T cell line. After successful verification of the construction, the zebrafish infection model is applied to explore the anti-tuberculosis protective effect of the pcDNA::Rv3868 candidate vaccine.

There are many animal models for tuberculosis research, including mouse models, guinea pig models, and zebrafish models. Due to the small size, relatively convenient and low-cost maintenance of mice, and the wide range of immune and genetic backgrounds of mouse strains, the mouse model has become the most commonly used. However, the mouse model finds it difficult to exhibit the significant pathological characteristics of human pulmonary tuberculosis. In contrast, guinea pigs are more

susceptible to M.tb, and the main pathological characteristics after infection, such as pulmonary granuloma structure, central caseous necrosis, cavitation, and calcification, are similar to humans. However, compared to mice, there is a lack of diversified commercial antibodies needed for vaccine evaluation, and knockout strains are also difficult to ensure. The most ideal primate animal model is costly, has the highest requirements for breeding technology and environment, and has issues of long duration and ethics. In comparison, this study uses the zebrafish (Danio rerio) infection model for vaccine evaluation. Similar to the human immune system, zebrafish also have a complex immune system, including innate and adaptive immunity, and are more cost-effective and easier to breed than mice. Currently, the zebrafish model has been widely used in the study of various human diseases. Zebrafish females have a large number of eggs, and the differences among individuals in the same batch are small. Similar to the mouse model, the gene knockout technology in zebrafish is also relatively mature, and the target gene can be knocked out through CRISPR/Cas9 technology to study its role in disease development. Before the age of 14 days, it is possible to achieve transparency of the zebrafish body, which is helpful for studying the host's innate immune mechanism.

Zebrafish can be naturally infected by *Mycobacterium marinum* (M. marinum) and suffer from fish tuberculosis. A stable bacterial load and caseous granulomas are formed in the fish body. The granuloma structure formed after M. marinum infects zebrafish is similar to the granuloma structure formed by M.tb in humans. In recent years, it has been widely used in the research of in vivo anti-tuberculosis drug screening, evaluation of new vaccine effects, and dynamic changes of granulomas in the body.

III. Methods and Materials

1. Experimental Animals and Strains

- Zebrafish Larvae: 3-day post-fertilization (dpf) AB strain wild-type zebrafish (Danio rerio), also known as zebrafish, bred in-house at the laboratory's zebrafish facility.

- M. marinum: Mycobacterium marinum wild-type strain M, used for zebrafish

infection.

- DNA Candidate Vaccine (pcDNA::Rv3868): A recombinant DNA construct self-assembled in this study using the pcDNA3.1+ plasmid as the vector.

- Escherichia coli Trans5a (Beijing Allsheng Jin Technology Co., Ltd.)

-Mycobacterium tuberculosis Genome (Mycobacterium tuberculosis H37Rv, preserved in the laboratory)

2. Experimental Methods

2.1 AI Prediction and Evaluation of Immunogenicity of ESX-I Proteins

The three-dimensional structures of important secreted proteins from *Mycobacterium tuberculosis* (EccA1, EccB1, EccCa1, EccCb1, EccD1, EccE1, EspC) are obtained through two methods: downloading existing protein structures from the PDB website (www.rcsb.org) and predicting protein three-dimensional structures using AlphaFold3 (https://www.alphafoldserver.com/). Input the structure files of the aforementioned proteins into the SEPPA3.0 (http://www.badd-cao.net/seppa3/index.html) prediction website to determine the potential B cell epitopes of the target proteins.

2.2 Bioinformatics Analysis of EccA1 Protein

The Prot-Param software was used to analyze the amino acid sequence and physicochemical properties; Protscale was used to understand the hydrophilicity or hydrophobicity of the protein; SOPMA was used to predict secondary structure and coiled coils (Coils); TMHMM and SignalIP 5.0 were used to predict transmembrane regions and signal peptides; use AlphaFold3 and SWISS-MODEL to simulate tertiary structures; Bepipred Linear Epitope Prediction 2.0, ABCpred, and SEPPA3.0 software were used to predict B cell antigenic epitopes; NetMHCIIpan3.2 Server and NetMHCpan - 4.0 Server were used to measure T cell epitopes.

2.3 Construction of pcDNA::Rv3868 Plasmid

2.3.1 Amplification of Rv3868 Gene Sequence

The nucleotide sequence of Rv3868 (encoding EccA1) was downloaded from the mycobrowser website (mycobrowser.epfl.ch/genes/Rv3868). EccA1 was amplified using PCR with the genomic template of M.tb in our lab. Primer information is shown in Table 1.

Table 1. PCR primer sequences for the Rv3868 gene.

Primers	Sequence information	Enzyme
Rv3868-F	gactcactatagggagacccaagcttGCCACCatgactgatc gcttggccag	Hind III
Rv3868-R	cactggcggccgttactagtggatcctcaCTTATCGTCG TCATCCTTGTAATCttctctcatgttgaggtgtgcg	BamH I

Note: Underlined sections represent restriction enzyme cutting sites, and lowercase letters indicate the vector homology arms.

Table 2. PCR amplification system for the Rv3868 gene.

Items	**Volume**
2 × Phenta Master Mix	25 μL
pCDNA-*Rv3868*-F（10 μM）	2 μL
pCDNA- *Rv3868*-R（10 μM）	2 μL
M.tb genome	0.5μL
ddH_2O	To 50 μL

2.3.2 Recovery of Rv3868 PCR product

The purity of the PCR product was verified by 1% agarose gel electrophoresis. The gene fragment recovery was carried out according to the instructions of the AXYGEN PCR Clean-Up Kit AP-PCR-250G. After measuring the concentration of the product, it was stored at -20°C.

2.3.3 Extraction of pcDNA3.1 plasmid

The plasmid extraction procedure was performed following the instructions of the bacterial plasmid mini prep kit (AXYGEN Corporation, USA). After reviving 50 mL of glycerol stock of DH5α E. coli, the bacterial pellet was collected by centrifugation and the plasmid was extracted. The isolated plasmid DNA was measured for concentration and then stored at -20°C.

2.3.4 Enzymatic digestion

BamH I and Hind III restriction enzymes were used to digest the pcDNA3.1 plasmid

DNA. The digestion system is shown in the table below. The reaction was carried out in a 37°C metal bath for 30 minutes. The digested plasmid DNA was recovered from the agarose gel through 1% agarose gel electrophoresis, and the gel recovery procedure was performed according to the instructions of the DNA Gel Extraction Kit AP-GX-250G (AXYGEN Corporation, USA).

The gel-recovered digested vector DNA and the Rv3868 recovery product were subjected to homologous recombination, following the system in Table 3 for multi-fragment homologous recombination.

Table 3 Homologous recombination reaction system.

Items	Volume
pcDNA3.1	9 μL
Rv3868 template	1 μL
5×CE MultiS Buffer	4 μL
Exnase MultiS	2 μL
ddH_2O	To 20 μL

After preparing the homologous recombination system, the PCR machine is set to react at 37°C for 30 minutes, followed by immediate cooling at 4°C.

2.3.5 Transformation of recombinant plasmid

1) Thaw the DH5α competent cells on ice.

2) Carefully add 10 μL of the recombination product to 100 μL of competent cells on ice, and let it stand for 30 minutes.

3) Heat shock at 42°C in a metal bath for 45 seconds, then immediately cool on ice for 3 minutes.

4) Add 1 mL of antibiotic-free LB liquid medium and culture at 37°C with shaking for 1 hour.

5) Centrifuge at 3000 rpm at room temperature for 3 minutes, discard the supernatant, resuspend the cells in the remaining 100 μL of medium, and spread 50 μL of the suspension onto LB agar plates with ampicillin resistance, then culture upside down at

37°C for 12 hours.

6) Randomly select 3-5 single colonies for PCR verification of the bacterial culture (the system is shown in Table 4).

Table 4 Bacterial cell PCR reaction system.

Items	Volume
2 × T5 Master Mix	10 μL
Bacteria cells	1 μL
pCDNA-*Rv3868*-F（10 μM）	1 μL
pCDNA-*Rv3868*-R（10 μM）	1 μL
ddH_2O	7 μL

7) For samples with positive bacterial cell PCR, the remaining bacterial cell samples are sent for sequencing verification. Strains with the correct gene sequence verified by Snapgene software are successfully constructed and named as pcDNA::Rv3868. The original pcDNA3.1+ without recombinant exogenous genes serves as the negative control for this experiment and is named as pcDNA.

2.4 Intracellular expression of pcDNA::Rv3868

2.4.1 Endotoxin-free plasmid extraction

Follow the TIANGEN Endotoxin-Free Plasmid Large-Scale Extraction Kit method to extract plasmids pCDNA and pcDNA::Rv3868 from recombinant *E. coli* DH5a for cell and animal experiments.

2.4.2 HEK293T cell culture

1) Thaw HEK 293T cells in a 37°C water bath.

2) Transfer the cells to a sterile centrifuge tube, add 1 mL of DMEM cell culture medium, centrifuge at 1000 rpm for 5 min, and discard the supernatant.

3) Resuspend the cells in 1 mL of DMEM cell culture medium. Add 9 mL of DMEM cell culture medium to a 100 mm cell culture dish, add the cell suspension, and culture in a 37°C cell incubator.

4) Discard the supernatant of the culture medium, add 1 mL of 0.25% trypsin to each dish and let it stand for 1 min, then add 1 mL of DMEM cell culture medium to

terminate the digestion process.

5) Collect the cell suspension and place it in a centrifuge tube, centrifuge at 1000 rpm for 5 min.

6) Transfer the cells to 2-3 cell culture dishes for further culture. When the cells reach the third passage, they are officially used for the experiment.

2.4.3 HEK293T Cell DNA transfection

1) DNA transfection. Add 50 μL of Opti-MEM® Medium, 1 μg of DNA to an EP tube, gently mix, and then add 1.6 μL of Lipo8000TM transfection reagent (per well).

2) Replace the cell culture medium, add 50 μL of the prepared plasmid-transfection reagent mixture to each well, and after transfection for different periods, discard the cell culture supernatant and wash the culture dish with PBS.

3) Add 100 μL of RIPA strong cell lysis buffer to each well, fully lyse, collect the lysate, centrifuge at 12000 rpm, 4°C for 30 min, and retain the supernatant for western-blot verification.

2.4.4 Western-Blot verification of pcDNA::Rv3868

1) Taking 20 μg of cell lysate and mix with 5 × protein loading buffer in proportion, boil for 15 min to fully denature the proteins.

2) After centrifugation at 11000 rpm for 10 min, take the supernatant and add it to the SDS-PAGE protein gel, perform protein electrophoresis at 120 v for about 30 min.

3) After the electrophoresis is finished, take out the SDS-PAGE gel and transfer it to a PVDF membrane on an eBlot™ L1 rapid wet transfer device.

4) After the transfer is finished, transfer the membrane to the blocking solution and incubate on a shaker at room temperature for 1 hour.

5) Transfer the membrane to TBST buffer and wash it with the TBS buffer.

6) Mark the Rv3868 protein (EccA1) with anti-Flag antibody, and the internal reference protein GAPDH with anti-GAPDH antibody. Dilute the primary antibody with 5% skim milk at a ratio of 1:2000, transfer the washed membrane to the primary antibody, and incubate slowly on a shaker at room temperature for 1 hour.

7) Wash the membrane as in step 6 to remove the primary antibody from the membrane surface.

8) Dilute the secondary antibody with 5% skim milk at a ratio of 1:5000, transfer the washed membrane to the secondary antibody, and incubate for 45 min.

9) Wash the membrane as in step 6 to remove the secondary antibody from the membrane surface.

10) Dry the buffer on the membrane, evenly add the developing solution, and expose and image in a Las4000 imager.

2.5 Safety assessment of pcDNA::Rv3868 candidate vaccine in eebrafish larvae

1) Zebrafish was used for mating and egg production. Observe and collect the eggs 30 min later, transfer them to a culture dish, and culture in a 30°C incubator. Remove dead eggs and hatched membranes daily.

2) On the third day, the zebrafish larvae were taken out of the culture dish, anesthetized with tricaine, then fixed in agarose gel, and immunized with a microinjection system. 5 μg/larvae of pcDNA::Rv3868, pcDNA was immunized into the dorsal muscle of zebrafish. Then, the zebrafish larvae were continued to culture in a 30°C incubator.

3) The mortality of the larvae was checked twice a day; the larvae fish food was fed every other day.

2.6 Cultivation of *Mycobacterium marinum*

Mycobacterium marinum glycerol stocks were streaked on 7H10 solid medium containing the appropriate antibiotics and 10% OADC, and then incubated at 30°C for about 7 days. Single colonies were cultured in 7H9 liquid medium containing OADC and antibiotics to the logarithmic phase with shaking.

2.7 Evaluation of the protective effect of pcDNA::Rv3868 candidate vaccine against *Mycobacterium marinum* infection in zebrafish larvae

1) 3-day-old zebrafish larvae were taken out as described above, randomly divided them into 3 groups; each group was with 45 fish, namely PBS group, pcDNA::Rv3868, and pcDNA group.

2) Zebrafish larvae was immunized with a microinjection system as described above. 2 μg/larvae of pcDNA::Rv3868 or pcDNA was immunized into the dorsal muscle of zebrafish. An equal volume of PBS into the same part of the PBS group larvae.

3) 24 hours later, zebrafish were fed with different doses of M. marinum at 1*10^5 CFU/mL, 5*10^5 CFU/mL, 2.5*10^6 CFU/mL for 6 hours, then replaced the sterile zebrafish culture water, and continued to culture. Repeat the infection process daily.

4) Observe and record the mortality, feeding, and swimming status of zebrafish daily.

5) On the 7th day of infection, take out the zebrafish, anesthetize and kill with tricaine. Weigh the body weight of zebrafish in groups of 5. Evaluate the anti-tuberculosis protective effect of pcDNA::Rv3868 based on the weight changes of zebrafish in different groups.

2.8 Determination of bacterial load in adult zebrafish

At the planned experimental time points, 4 zebrafish from each group were taken out to determine the number of *Mycobacterium marinum* in vivo. Zebrafish were anesthetized in tricaine for 5-10 minutes until dead, washed with sterile PBS buffer, and placed each fish into a labeled shaking tube (with beads for crushing) containing 1 mL of 10% glycerol PBS solution. Crush in a homogenizer at 5000 rpm for 30 seconds with a 5-minute interval for three times. Dilute the homogenate and plate, count after 7 days of culture. Each fish is plated at three dilution gradients and three replicates to ensure accurate counting. Analyze the counting results with GraphPad 8.0.

IV. Results

4.1 Identification of B-cell Epitopes of Important Secreted Proteins of *Mycobacterium tuberculosis* by Integrating AlphaFold3 and Protein Structure Prediction

The three-dimensional structures of important secreted proteins of *Mycobacterium tuberculosis* (EccA1, EccB1, EccCa1, EccCb1, EccD1, EccE1, EspC) were obtained through two approaches: downloading existing protein three-dimensional structures from the PDB website (www.rcsb.org) and predicting protein three-dimensional structures using AlphaFold3 (https://www.alphafoldserver.com/). The structural files of the aforementioned proteins were input into the SEPPA3.0 (http://www.badd-cao.net/seppa3/index.html) prediction website to determine the potential B-cell epitopes of the target proteins. It was found that the B-cell epitope prediction result for

the EccA1 protein is 147, ranking at the forefront among all predicted conserved component proteins (Table 5).

Table 5 Prediction results of B-cell epitopes for transmembrane channel proteins in the ESX-I System.

Genes	Amino acids number	B cell AA numbers	Epitopes
EccCa1	747	214	28.6%
EccA1	573	147	25.7%
EspA	446	107	24.0%
EccCb1	591	141	23.9%
PE35	99	22	22.2%
EspK	729	131	18.0%
EccD1	511	53	10.4%
EccE1	462	33	7.1%

4.2 Analysis of the biochemical properties of the EccA1 protein

Upon analysis with ProtParam, the EccA1 protein, consisting of 573 amino acids, contains 64 basic amino acids (K, R) and 81 acidic amino acids (D, E). The hydrophobic amino acids (A, V, L, I, P, F, W, M, G) are 300, accounting for 52.3%, while the hydrophilic amino acids (Y, S, T, C, N, Q, K, R, H, D, E) are 273, accounting for 47.7%. The theoretical isoelectric point of 5.07. The instability index is 36.93, classifying the protein as a stable protein. The lipophilicity index is 88.57, the hydrophilicity average is -0.192, and the absorbance value at a wavelength of 280 nm is 1.043. Protscale analysis indicates that the protein possesses amphiphilic properties (Figure 1).

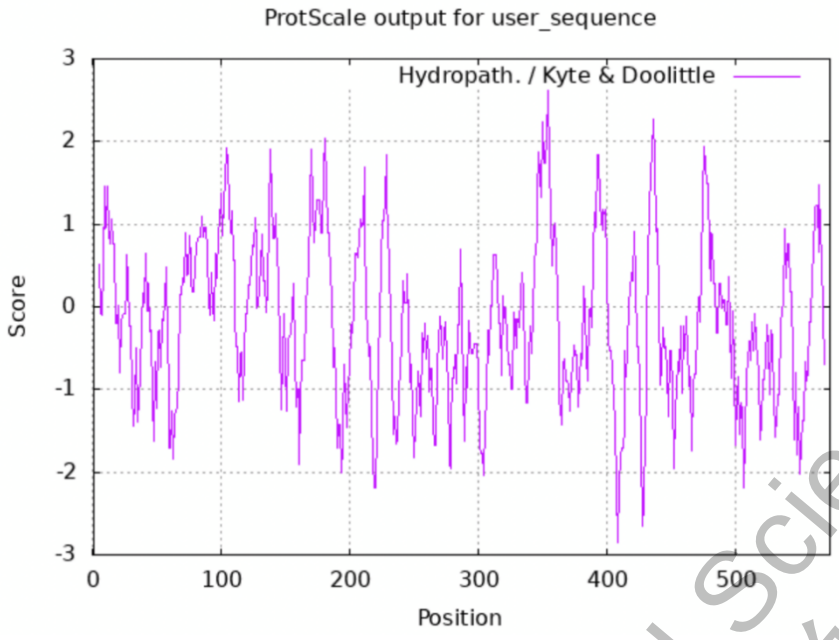

Figure 1. Hydrophilic and hydrophobic profile of M.tb Ecca1 protein

Note: The horizontal axis represents the position of amino acids, and the vertical axis represents the hydrophobicity score (hydrophilic and hydrophobic values are represented by negative and positive values, respectively).

4.3 Secondary and tertiary structure of the EccA1 protein

The secondary structure of EccA1 protein was analyzed with SOPMA, revealing that the random coils, alpha-helices, beta-sheets, and turns account for 23.91%, 65.27%, 0.00%, and 4.71%, respectively. Coils analysis indicated that the protein lacks coiled-coil structures. EccA1 protein has a minimal number of transmembrane regions using TMHMM transmembrane region analysis (Figure 2). The SignalP prediction value is below the threshold of 0.5, which implies that the EccA1 protein does not contain a signal peptide sequence.

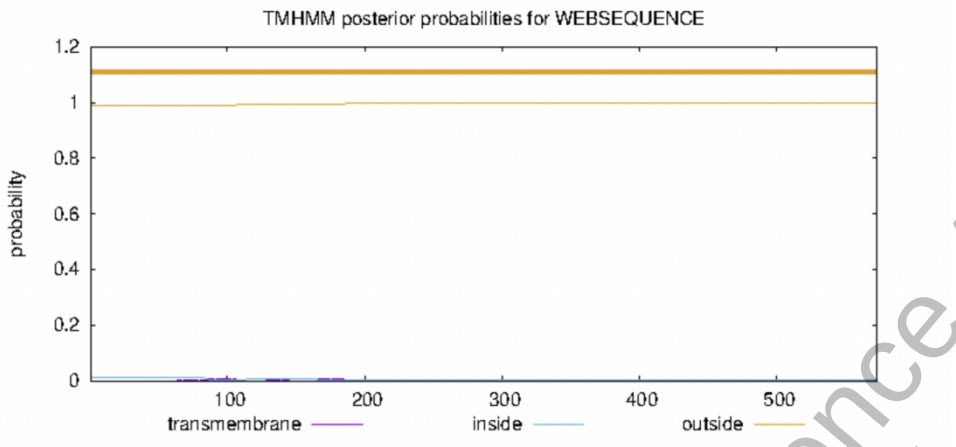

Figure 2. Prediction of the transmembrane domain of Ecca1

AlphaFold3 and SWISS-MODEL were used to predict the tertiary structure of the EccA1 protein. The results indicate a high content of random coils and a relatively compact structure (Figure 3A). A peptide segment composed of amino acids (AA) from positions 11 to 73 contains three alpha-helices, which form a peptidoglycan-binding structural domain (Figure 3B).

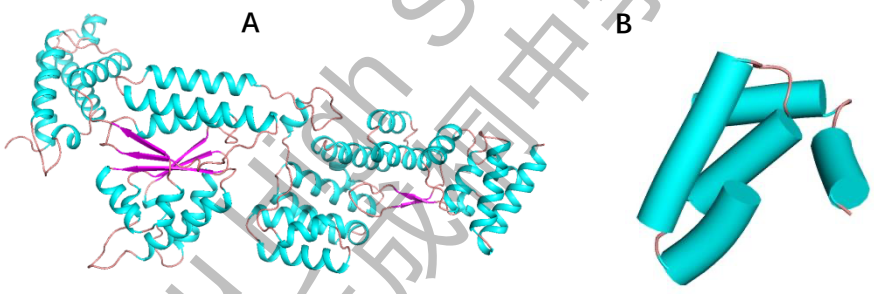

Figure 3. Tertiary structure of M.tb EccA1

A, Tertiary structure of the EccA1 protein; B, Peptidoglycan-binding structural domain composed of amino acids 11-73 of the EccA1 protein.

4.4 B-cell and T-cell Epitopes of the EccA1 protein

Bepipred Linear Epitope Prediction 2.0 and ABCpred, were used to predict the B-cell epitopes. The result showed that EccA1 has a total of 13 B-cell epitopes (threshold: 0.6), specifically at positions 194-201, 247-277, and 408-409. Additionally, by considering the regions of random coils present in the secondary structure of the EccA1 protein, potential B-cell epitopes may be located in the amino acid segments 196-201,

247-255, 260-276, and 408-409.

The $CD4^+$ T-cell epitopes of the EccA1 protein were measured using the NetMHC II pan 3.2 Server software. The length of the peptide segment was set to 15 amino acids, and the prediction was made for human leukocyte antigen (HLA) class II subtypes that have a broad gene distribution (HLA-DR B1-0101, 0301, 0401, 0701, 0801, 0901, 1101, 1302, 1501) to assess the affinity values. The prediction results are shown in Table 6, with a total of 559 peptide segments for the HLA-DR proteins. Except for DRB1-0301, EccA1 has a total of 105 strongly binding peptide segments, with the highest number of strong and weak binding peptide segments for HLA-DRB1-0901. Among the strongly binding peptide segments, WYLAMARRS (210-218) is a binding epitope for DRB1-0801 and DRB1-1101; VRAAKGMKV (315-323) is a binding epitope for DRB1-0701 and DRB1-1501; VRAAKGMKV (315-323) is also a binding epitope for DRB1-0701 and DRB1-0901.

For the prediction of $CD8^+$ T-cell epitopes of the EccA1 protein, the online software NetMHCpan 4.0 Server was utilized. The length of the peptide segment was set to 9 amino acids (AA), and the prediction was conducted for the most common human class I leukocyte antigens: A2 (HLA-A0201) and A3 (HLA-A0301). The results identified 8 strong binding peptide segments for HLA-A0201: SLFESAVSM (7-15), ALLEWLQTT (227-235), ALKDPSYRL (244-252), LLAEAQAEL (282-290), VLFIDEAYA (393-401), FIDEAYALV (395-403), ALDTLLARM (417-425), and RLLETNEGL (445-453); and 6 strong binding peptide segments for HLA-A0301: LMARVRAAK (311-319), RVRAAKGMK (314-322), GMKVAQPSK (320-328), FTGPPGTGK (332-340), RLVVIIAGY (431-439), and AVHAHLNMR (564-572).

Table 6 Prediction results of $CD4^+$ T-cell epitopes of EccA1

HLA-DR	Total peptides	
	Strong-binding peptides	Weak-binding peptides
DRB1-0101	10	45
DRB1-0301	4	34

DRB1-0401	6	39
DRB1-0701	7	46
DRB1-0801	21	30
DRB1-0901	16	52
DRB1-1101	23	21
DRB1-1302	5	29
DRB1-1501	17	24

4.5 Construction and enzymatic verification of pcDNA::Rv3868

Recognizing the potential of the EccA1 protein as a target for anti-tuberculosis vaccines, this project aims to construct a candidate DNA vaccine using the EccA1 protein as an antigen and the pCDNA plasmid as a vector, to explore the anti-tuberculosis potential of the EccA1 protein. Primers were designed based on the M. tuberculosis H37Rv genome, and the encoding gene Rv3868 was obtained through PCR to construct the recombinant pCDNA. The construct was then enzymatically verified using BamHI and HindIII double digestion (Figure 4). The successful construction was named pcDNA::Rv3868. The constructed vector was sequenced by a sequencing company and analyzed using Snapgene software to confirm the absence of errors in the gene sequence before proceeding to the next steps of in-cell expression and animal-level assessment.

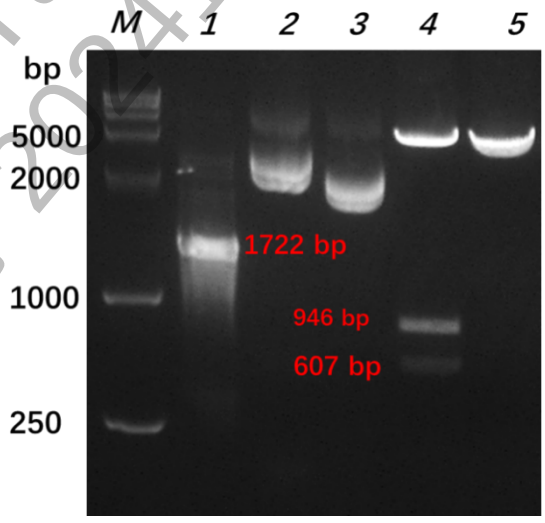

Figure 4. Verification of the pcDNA::Rv3868 vector.

Lane M: DL5000 Marker; Lane 1: PCR product of Rv3868; Lane 2: pcDNA::Rv3868 plasmid; Lane 3: pcDNA plasmid; Lane 4: Double digestion products of pcDNA::Rv3868 with BamHI and HindIII; Lane 5: Double digestion products of pcDNA with BamHI and HindIII.

Due to the presence of multiple BamHI sites within the gene, the gene was cut into several fragments, two of which are larger and labeled in Figure 4. Additionally, because of the absence of the HindIII site, the size of the double-cut recombinant plasmid backbone is slightly larger than that of the original plasmid digest product.

4.6 Expression of pcDNA::Rv3868 in HEK293T Cells

The well-constructed recombinant plasmid was transfected into HEK293T cells to assess its expression in eukaryotic cells. The experimental group consisted of cells transfected with pcDNA::Rv3868, while the negative control was transfected with pcDNA. All plasmids were obtained through an endotoxin-free extraction method. Lipo8000TM was used for cell transfection at a transfection amount of 2.5 μg, with transfection times of 24 hours and 48 hours. After the transfection was complete, cell lysates were collected for western-blot validation. The theoretical size of the target protein is approximately 62.43 kDa, with GAPDH as an internal control protein (36 kDa) (Figure 5). The western-blot validation results indicate that the pcDNA::Rv3868 recombinant plasmid can express in eukaryotic cells, confirming the successful construction of the recombinant plasmid, which can now be used for animal experiments.

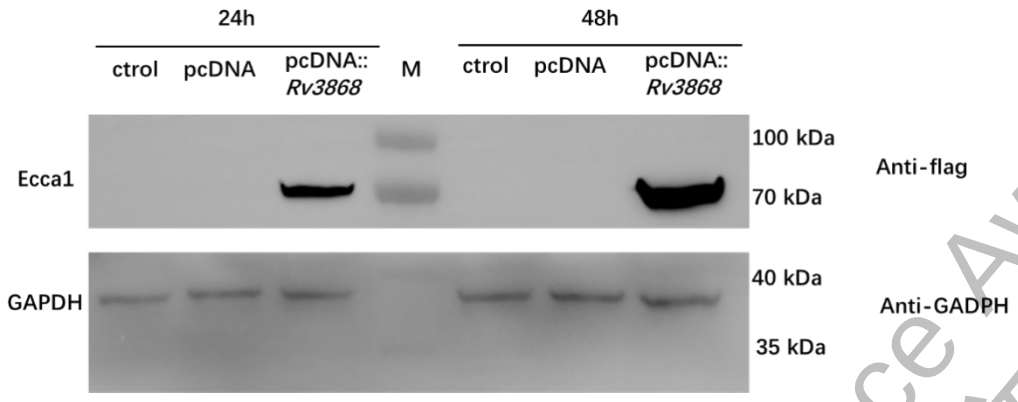

Figure 5. Western-blot result of pcDNA::Rv3868 in HEK 293T Cells. The control lane represents the cell lysate from untransfected cells, the M lane represents the protein marker, and the remaining lanes are as indicated in the figure.

There is no significant difference compared to the control group for the state of the cells after transfection at 24 and 48 hours (Figure 6) (the treatment of the control group was identical in every way except that no plasmid was added during transfection).

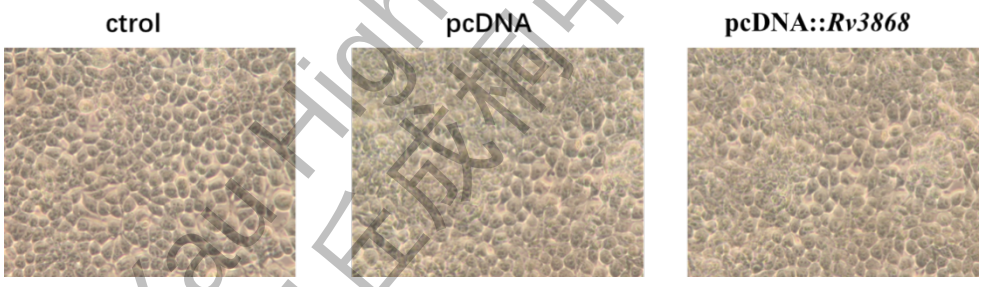

Figure 6. HEK293T cells after 48 hours of transfection with pcDNA::Rv3868. The cells in the image were photographed under a phase-contrast microscope at a magnification of 400X.

4.7 Safety assessment of pcDNA::Rv3868 in zebrafish larvae

Zebrafish larvae are widely used to assess the potential toxicity and safety of drugs or compounds. To investigate the feasibility of the pcDNA::Rv3868 candidate vaccine, it is essential to first determine whether the vaccine is safe at the animal level. We used a 3-day post-fertilization (dpf) zebrafish larvae model and injected a high dose (5 ng)

of plasmid, continuously observing the impact on the survival of zebrafish larvae. The results showed that during the entire 7-day observation period, there were no deaths among the zebrafish, and their growth status was good, consistent with the non-injected plasmid group. This indicates that the pcDNA::Rv3868 candidate vaccine has good safety in the zebrafish animal model selected for this experiment (Figure 7).

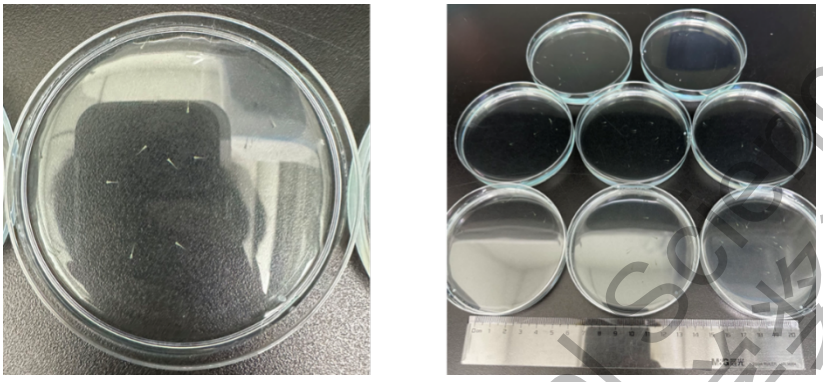

Figure 7. The zebrafish larvae status after high-dose injection of pcDNA::Rv3868 for safety assessment.

4.8 Anti-tuberculosis protective effect of pcDNA::Rv3868 in zebrafish larvae model

The zebrafish model was used to further validate the anti-tuberculosis protective effect of the pcDNA::Rv3868 vaccine. The experimental group was injected with pcDNA::Rv3868, the negative control group was injected with pcDNA, and the blank control group was injected with an equal amount of sterile PBS. 2 μg of DNA injected into the back muscle each time for immunization. 24 hours after immunization, the larvae were soaked in a culture solution containing different concentrations of M. marinum for infection challenge. The mortality of zebrafish was observed daily after challenge, and on the 7th day after infection, the zebrafish were euthanized with tricaine, and their body weight was measured.

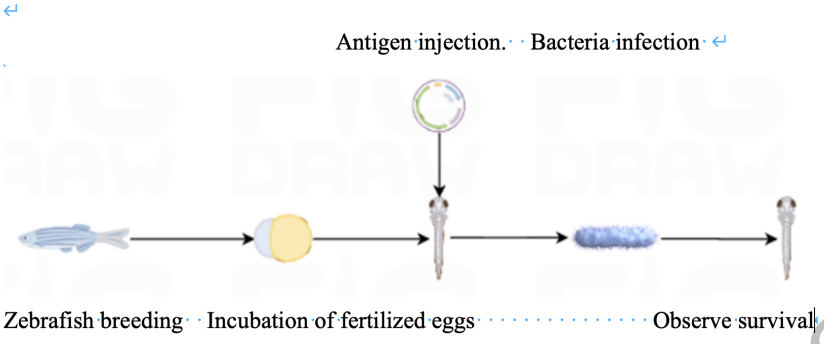

Figure 8. Experimental procedure for evaluating the protective effect of pcDNA::Rv3868 in larvae.

No mortality was observed in the zebrafish after challenge. This situation is related to the infection route, dose, and experimental period of this experiment. At the same time, tuberculosis is a chronic wasting disease, and there are very few cases of sudden death due to high-dose virulence. After *Mycobacterium tuberculosis* infection, an important clinical sign for assessing tuberculosis is the emaciation of the body. Therefore, we chose to compare the changes in body weight of zebrafish under different infection doses. The results are shown in Figure 9. In the high-dose infection group, the body weight of zebrafish immunized with pcDNA::Rv3868 was 8% higher than that of the pcDNA-immunized group, showing a certain protective advantage.

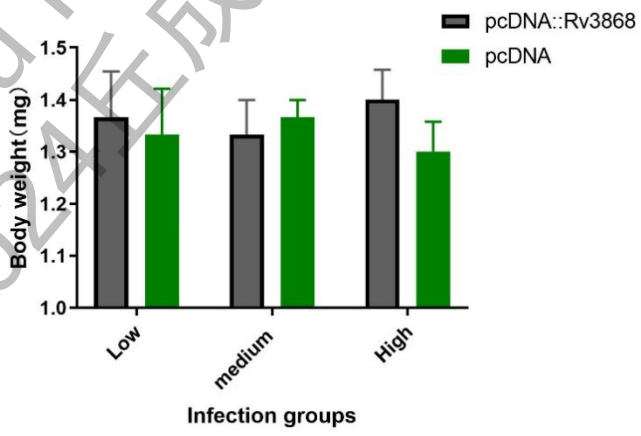

Figure 9. The effect of the pcDNA::Rv3868 candidate vaccine on the juvenile zebrafish (n=15).

4.9 Anti-tuberculosis protective effect of pcDNA::Rv3868 in adult zebrafish model

Adult zebrafish of the AB strain, aged 3-4 months, were randomly divided into 15 groups, with 20 fish per group. The experiment was conducted according to the procedure shown in Figure 10. The pcDNA empty vector immunization group was used as a negative control. After infection with *Mycobacterium marinum* M.m535, the feeding and mortality of zebrafish were observed daily, and three weeks after infection, five fish from each group were taken for total bacterial load counting of M.m535 in the whole fish body to evaluate the vaccine effect.

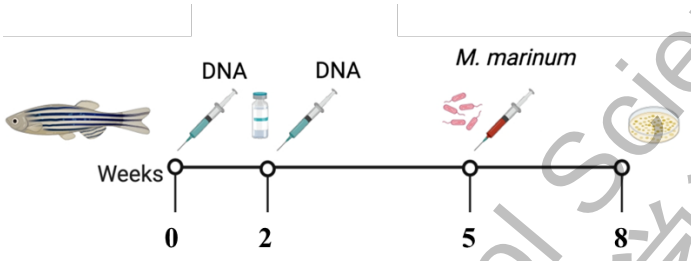

Figure 10. Experimental procedure for evaluating the protective effect of pcDNA::Rv3868 in adult zebrafish.

The results showed (Figure 11), with the non-immunized PBS group and the pcDNA empty vector immunization group as negative controls, immunization with pcDNA::Rv3868 significantly reduced the total bacterial load of Mycobacterium marinum in the whole fish body. Compared with the pcDNA empty vector immunization group, the average bacterial load per fish in the pcDNA::Rv3868 immunization group could be reduced by 0.68log10 CFU (t-test, P-value=0.0169). Compared with the non-immunized PBS group, the average bacterial load per fish in the pcDNA::Rv3868 immunization group could be reduced by 1.07log10 CFU (t-test, P-value=0.0339). This shows that pcDNA::Rv3868 has better potential for vaccine development.

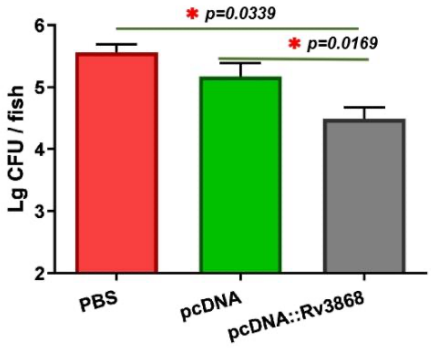

Figure 11. Bacterial load in zebrafish immunized with recombinant DNA plasmid three weeks after *Mycobacterium marinum* infection. Data are represented as Mean ± SEM, with at least three zebrafish per group. Difference analysis: t-test*, P-value<0.05. The non-immunized (PBS) and pcDNA empty vector immunized (pcDNA) groups were used as negative controls. The horizontal axis represents the group, and the vertical axis represents the M. marinum load per fish.

V. Conclusion and Discussion

The multiple shortcomings of BCG in anti-tuberculosis applications have made it impossible to rely solely on BCG to achieve the goal of preventing and eliminating tuberculosis. To further reduce the harm of tuberculosis to human life and health, the development of new anti-tuberculosis vaccines is urgent. Although this research has attracted much attention from researchers, clinical trials are still progressing slowly, and to date, BCG remains the only officially recommended choice for human anti-tuberculosis. Faced with the complex and severe situation of tuberculosis, it is necessary not only to screen and obtain more mycobacterial antigens but also to conduct comprehensive and in-depth research on whether they have the potential for anti-tuberculosis vaccines.

With the rapid development of bioinformatics, protein structure prediction and antigen epitope prediction have greatly facilitated protein function research. This study used bioinformatics methods to predict and discover the physiological characteristics of the EccA1 protein. Although SignalP predicted that the EccA1 protein lacks a signal peptide sequence, SOPMA analysis showed that alpha-helices account for 65.27% of the secondary structure of the EccA1 protein, suggesting a higher possibility of

existence outside the membrane. This indicates that the protein EccA1, although functioning as an ATPase, is likely a secreted antigen. In pathogenic *Mycobacterium marinum*, mutations in the EccA1 protein affect bacterial virulence, indirectly supporting this inference.

The analysis results of B-cell and T-cell epitopes help predict the potential immune responses induced during the immune process. The predicted B-cell and T-cell antigenic epitopes of the EccA1 protein do not cross, but the peptide segment from 311 to 323, as a T-cell epitope, includes sequences that can bind to both HLA-Class I and HLA-Class II receptors, potentially stimulating CD4 and CD8 T-cell immune responses. The peptide segment from 244 to 252 contains both T-cell and B-cell epitopes. This suggests that the EccA1 protein has multiple immune-modulating functions and is a very promising candidate target for anti-tuberculosis.

To verify the above inference, this project used pCDNA as a vector and the EccA1 protein from M. tb as an antigen to construct the candidate DNA vaccine pcDNA::Rv3868. The recombinant plasmid was successfully expressed in HEK293T cells and showed safety at the cellular level under experimental conditions. In the safety evaluation model of zebrafish larvae, high-dose immunization with the candidate DNA vaccine did not affect the survival of zebrafish, still showing safety at the animal level. After immunization with pcDNA::Rv3868, upon infection with three doses of pathogenic *Mycobacterium marinum*, one week later in the high-dose infection group, the body weight of immunized zebrafish larvae was higher than that of the negative control group, showing an anti-tuberculosis protective effect. The protective effect assessment in adult zebrafish was also carried out in parallel, and it was found that on the third week (21 days) of *Mycobacterium marinum* infection, the pcDNA::Rv3868 immunization group could control the effect of mycobacterial infection in the body, with an average mycobacterial load of 4.49 log10 CFU per fish, which was significantly different from the negative control group (t-test, P-value<0.05), further proving the anti-tuberculosis protective potential of the Rv3868 antigen and suggesting its application prospects for the development of anti-tuberculosis vaccines.

In summary, this study validates that the EccA1 protein is a high-quality anti-

tuberculosis vaccine antigen. Its vaccine potential still needs to be comprehensively evaluated in various animal models such as mice and guinea pigs. At the same time, due to the single antigen type of DNA vaccines, to maximize the advantages of the EccA1 antigen, it is possible to further construct subunit vaccines containing multiple antigens on this basis to improve their vaccine effects.

References：

1, WHO. (2022). Global tuberculosis report 2022.

2, Sangeeta Tiwari, et al. Infect and Inject: How Mycobacterium tuberculosis Exploits Its Major VirulenceAssociated Type VII Secretion System, ESX-1. Microbiol Spectrum. 2019. 7(3): BAI-0024-2019.

3, Zhihao Xiang, et al. Immunogenicity and protective efficacy of a fusion protein tuberculosis vaccine combining five Esx family proteins. Frontiers in Cellular and Infection Microbiology. 2017. 7:226.

4, Lu Zhang, et al. Variable virulence and efficacy of BCG vaccine strains and correlation with genome polymorphisms. Molecular Therapy. 2016. 24(2): 398-405.

5, Haili Xu et al., Advances in Tuberculosis Vaccine and Immunization Strategy Research, Chinese Journal of Infection and Chemotherapy. 2023. 659-666

Acknowledgements

I would like to thank Prof. Lu Zhang at School of Life Sciences, Fudan University for her meticulous guidance and valuable suggestions. Prof. Zhang guided me into the field of *Mycobacterium tuberculosis* vaccine research and provided great assistance in experimental design, data analysis, and thesis writing. Also, I would like to thank Meng Zhang and Ziwei Zhou from School of Life Sciences, Fudan University, during the completion of my research project for their help and guidance. In addition, I would like to thank my high school teacher Ms. Tingyu Zhao from FDIS, who taught me a lot of biological knowledge and helped with my research topic. Finally, I would like to thank my parents and sister for their love, unlimited support, and encouragement.

Declaration of Academic integrity

The author declares no conflict of interest.

The project was completed at School of Life Sciences, Fudan University, where the student was enrolled as an internship student.