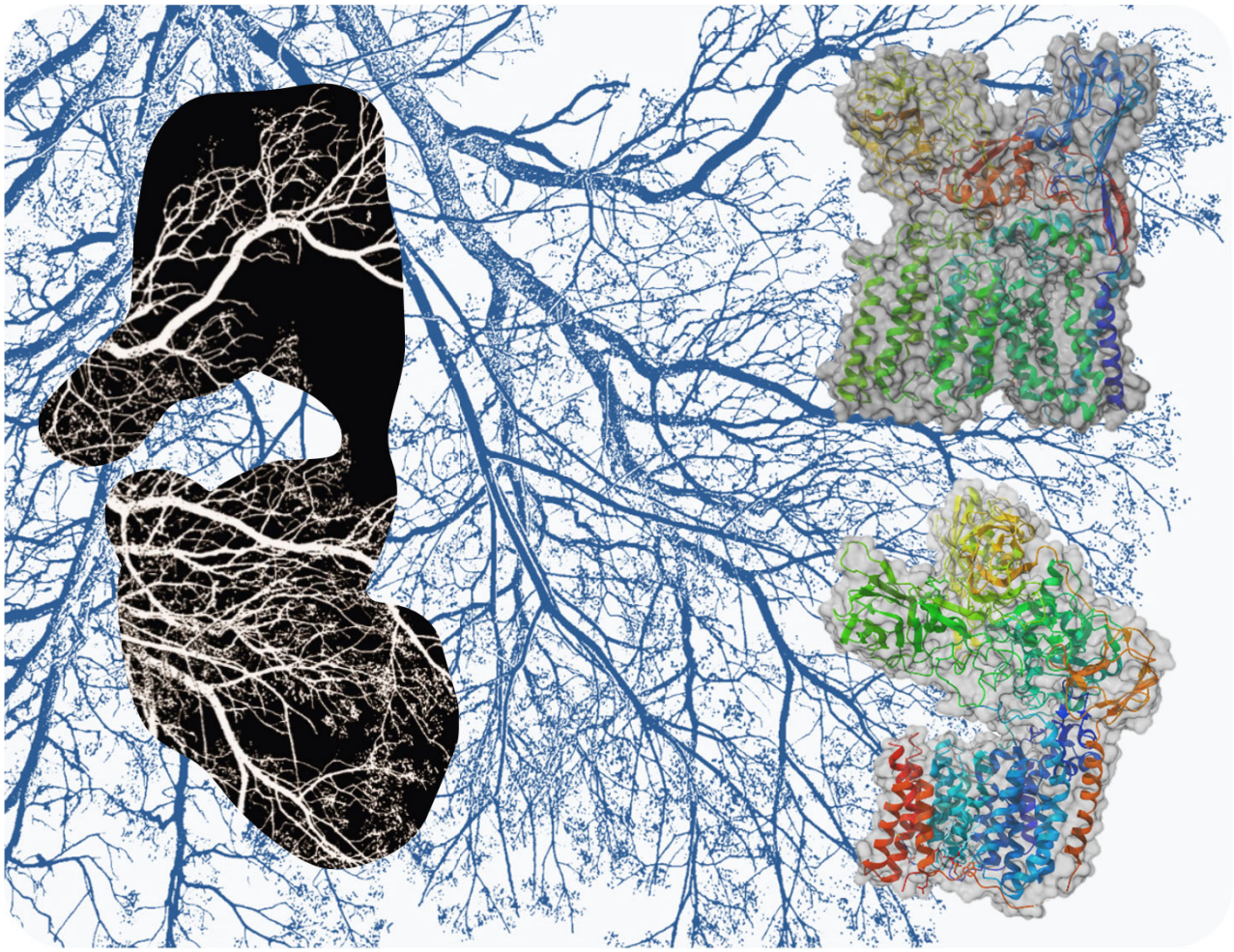
2024

Structural characterization of mycobacterial arabinofuranosyltransferases

José António Vieira Rodrigues



Dissertation presented to obtain the

Ph.D degree in Molecular Biosciences

Speciality Biochemistry

Oeiras, November 2024

INSTITUTO DE TECNOLOGIA QUÍMICA E BIOLÓGICA ANTÓNIO XAVIER | UNIVERSIDADE NOVA DE LISBOA

ITQB NOVA, Avenida da República, 2780-157 Oeiras, Portugal | Tel +351 214 469 100 | itqb.unl.pt

Structural characterization of mycobacterial arabinofuranosyltransferases

José António Vieira Rodrigues

Dissertation presented to obtain the Ph.D. degree in Molecular Biosciences, speciality in Biochemistry

Instituto de Tecnologia Química e Biológica António Xavier

Universidade Nova de Lisboa

Oeiras, November 2024

Supervisor:

Margarida Archer, Ph. D.	Head of the Membrane Protein Crystallography Laboratory Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa

Co-Supervisor:

Filippo Mancia, Ph. D.	Associate Professor Co-Director Graduate Education Department of Physiology and Cellular Biophysics Columbia University, New York, USA

Jury Members

President

Mariana Luísa Tomás Gomes de Pinho, Ph.D.	Aggregate Professor Instituto de Tecnologia Química e Biologics António Xavier, Universidade Nova de Lisboa

Jurors

João Henrique Resende de Oliveira de Morais Cabral, Ph.D.	Principal Researcher i3S - Instituto de Investigação e Inovação em Saúde da Universidade do Porto
Miguel Viveiros Bettencourt, Ph.D.	Cathedratic Professor Instituto de Higiene e Medicina Tropical da Universidade NOVA de Lisboa
Ana Luísa Moreira de Carvalho, Ph.D.	Auxiliary Researcher Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa
Elsa Maria Ribeiro dos Santos Anes, Ph.D.	Cathedratic Professor Faculdade de Farmácia da Universidade de Lisboa

Printed in Portugal

Acknowledgments

I need to thank my supervisor, Dr. Margarida Archer, for many things. First, for sponsoring my application for a research fellowship back in 2015: I still remember the phone call that literally made me move to Oeiras. It is fair to say that it was a life-changing phone call, because it was. That was the start of my journey which would culminate in this PhD, and I am deeply grateful for the scientific guidance, the confidence and for all the support throughout the years, especially during the core years of my PhD work. I cherished the honesty and kindness shown during the most difficult times and will never forget the several leaps of faith taken in my behalf.

I would also like to thank my co-supervisor Dr. Filippo Mancia, for welcoming a somewhat inexperienced student in the lab, back in 2016, allowing me to learn and share a lab bench with some of the most amazing scientists and people. I have gained some of the best tools during my visits to Mancia lab, not only in terms of lab work, but also in terms of scientific insight and personal growth. New York is often a special place for most people who visit, but Mancia lab is definitely a gem that only a select few get to know. I am very glad I got to be one of them.

I would like to thank the members of my PhD thesis committee, Dr. Madalena Pimentel and Dr. João Vicente, for all the useful discussions regarding my PhD work and for all the support given whenever I asked.

I would like to thank Yong Zi Tan for the best collaboration I had during my PhD. I still remember the time when we were all fired up about the promising results from the grids of EmbB, the main target at the time, when I suggested that I might have had an “interesting target that was bigger and cleaner than EmbB”, although with crappy yields. You actually considered my input, took the time to prepare those grids, and we shared the results with Filippo. That target was AftD. Thank you for your trust and friendship.

I would like to thank our collaborators Dr. Madalena Pimentel and Dr. Maria João Catalão, for welcoming me in their lab and teaching me how to work with mycobacteria. I truly appreciate all your kindness and support during my visit to FFUL. It was a great learning experience, and I had a wonderful time there.

I would like to thank the Steyaertlab at the VIB-VUB Center for Structural Biology, in particular to Els Pardon, Alison Lundqvist and Jan Steyaert for providing the nanobodies and megabodies against AftA.

I would like to acknowledge the Merck Satellite Lab team at IBET, in particular, I would like to thank Pedro Sousa and Ana Rita Lemos for all the help with the SPR experiments with AftA; Ana Catarina Paiva and Micael Freitas for the help with the fractionator.

I also would like to thank the Butcher Research Group at the University of Helsinki for all the support with the single particle cryo-EM analysis of AftA, in particular, Aušra Domanska, Pasi Laurinmäki and Sarah Butcher.

A very special thanks goes to all my colleagues at the Membrane Protein Crystallography Lab. It is impossible not to thank each and every one of you for all the shared laughs and shared pain throughout the years.

You guys are all unique and deserve recognition, although most of you did absolutely nothing to make the pain go away. Oh, the resentment… only time will heal.

José Brito, it is shameful that a master crystallographer such as yourself would allow a PhD student from MPX to graduate without ever having produced a single diffracting crystal. You may say that it is on me, I disagree! On a serious note, thank you for all the empathy and guidance you have always shown to all MPX students, and for all the good spirits you always have.

Márcia, thank you for the ever-present eye-rolling judgement of ALL my actions. I had a great time annoying you, I'm sure it was reciprocal! Diogo, your sarcasm, cynicism, and questionable music taste were not always what I needed, but it sure did disrupt the lab routine, mostly in a good way. Anyways, thank you both, you were there with me from the start, and I could count on you to unwind and decompress about everything. I already miss our time together (although I will never publicly admit it)!

Vanessa, you and your calendar planning sheets were a constant reminder of my poor time-management skills. Every late night I had to stay in the lab I wished I would have a tiny fraction of your organizational skills. Thank you for your constant positive attitude towards life, some days it brushed off on me. Also, thanks for all the help with the lab work, you really make it easier for anyone to work with you.

I would also like to acknowledge some of the past MPX members that shared some of their expertise with me. I would like to thank Ana Lúcia Rosário for teaching me all the basic skills to work with membrane proteins, and Federico Issoglio for introducing me to molecular dynamics and for bringing that South American vibe to MPX… just seeing you sipping some mate would relax anyone.

A very special thanks goes to all members of the MX unit, for all the support throughout the years and for all the scientific input given on my work. Your door was always open for me to ask for advice and, mostly, reagents and equipment. I appreciate it!

I would also like to thank all my MolBioS PhD colleagues for all the laughs and support. A special mention goes to Ricardo and Filipe, my podcasting partners.

I would like to also acknowledge the ITQB community: students, staff, researchers… everyone. In this place we are allowed to be curious, we can be apprentices, we can be mentors, we can be conference organizers, or even be pool party planners, and much more…we can find ourselves in here.

The following personal acknowledgments will be addressed in Portuguese.

Muita coisa me faltou durante o doutoramento: tempo, paciência, talvez um microscópio eletrónico mais perto do laboratório… Mas apoio da minha família sempre tive, graças a Deus.

Quero agradecer aos meus pais e irmã pelo seu apoio incondicional. É engraçado perceber que, para eles, o meu trabalho em cada um dos meus ciclos de estudo pôde ser resumido numa palavra-chave: no mestrado a palavra era “Leveduras”, no doutoramento será “Proteínas” … Não estão propriamente perto, mas também não estão errados na escolha da palavra-chave.

Apesar de todas as minhas inquietações e incertezas, sempre tiveram confiança em mim e no meu percurso, mesmo quando eu não tinha. Estarei sempre agradecido por me apoiarem em todas as minhas decisões.

Por último, o meu maior agradecimento vai para a minha esposa Ana Luís. Como muitas vezes me senti em relação ao doutoramento, sinto também que este agradecimento vai ser escasso, mas como sempre, vou dar o meu melhor. Aqui vai!

Por cada sacrifício que tive de fazer durante o doutoramento, tu fizeste outro igual ou maior. Pelo meu sucesso deixaste a estabilidade de um emprego na nossa terra, assim deixando as tuas raízes, o conforto e a qualidade de vida que a serra nos dá. Esperaste sozinha quando eu estava do outro lado do Atlântico, mais do que uma vez, e sabemos bem em que circunstâncias. Aguentaste as minhas ausências cada vez que fiquei mais tarde no laboratório, ou quando saía de casa à noite para ir acabar alguma coisa, e mesmo agora durante a escrita desta tese, não consigo contar as vezes que assumiste

todas as responsabilidades, todo tempo com os nossos filhos, só para que eu pudesse avançar um parágrafo que fosse (o que nem sempre aconteceu, como deves imaginar). Devo-te todo esse tempo, a ti e aos nossos filhos.

Obrigado por seres um verdadeiro pilar para mim, uma constante nesta vida de incertezas e decisões difíceis. A única pessoa que eu quero mesmo deixar orgulhosa com este trabalho, és tu.

E a ironia é que nunca irás ler o trabalho (nem terias de o ler), mas tenho a certeza que não deixarás de me aplaudir, como fazes todos os dias.

E depois daquele dia de defesa em que, uma vez mais, fizeste tudo para que eu estivesse tranquilo e sem preocupações, ainda maior se torna a minha dívida para contigo, uma dívida que planeio pagar ao longo da nossa vida.

Maria e Francisco, obrigado por serem o que de melhor tenho na vida. Sempre.

Funding and Institutional Support

This research was supported by Fundação para a Ciência e Tecnologia (FCT) through my PhD fellowship (PD/BD/128261/2016 and COVID/BD/151615/2021) within the MolBioS program, and by Instituto de Tecnologia Química e Biológica António Xavier (ITQB NOVA).

This research was supported by Instruct-ERIC, through the Instruct Workshop on Nanobodies for Structural Biology and beyond (September 9th – 19th, 2018, Brussels, Belgium).

This research was supported by ProMeTeus – Protein and Membrane Technology Consortium, through the secondment of José Rodrigues to the New York Structural Biology Center COMPPÅ – Center on Membrane Protein Production and Analysis (April 7th, 2019 to July 5th, 2019, New York).

This research was supported by the EU Horizon 2020 Twinning project IMpaCT, through the IMpaCT workshops: “Single Particle Sample Preparation” (October 29th to November 9th, 2021, University of Helsinki), “Single Particle Sample Preparation – Online Workshop" (November 2nd to 5th, 2020, University of Helsinki – remote); and “Data Processing Workshop”, (September 17th to 20th, 2019, ITQB NOVA).

This research was supported by Fundação Luso-Americana para o Desenvolvimento (FLAD) through a fellowship attributed to José Rodrigues (Proj.166/2017) and funding attributed to Dr. Margarida Archer.

Abstract

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis* and is one of the most ancient infectious diseases known to affect mankind. In fact, still to this day, TB is one of the leading causes of death worldwide, as it is estimated that around 1.3 – 1.5 million people die from TB and 10 million new cases appear, every year. In addition, about 370,000 to 450,000 people develop multidrug-resistant TB (MDR-TB), for which the treatment regimens are longer and present lower success rates.

This disease represents a serious burden to mankind and still many challenges need to be overcome to eradicate it. Monitoring, prevention, and treatment of TB are key aspects towards eradication of TB, so research and activities focused on the improvement of diagnostics, development of vaccines and development of new therapeutics are essential to reach the proposed targets for 2030 and 2035 of World Health Organization (WHO) “End TB Strategy”.

The treatment for drug-susceptible TB consists in a daily regimen of isoniazid, rifampicin, pyrazinamide and ethambutol, during 4 – 6 months. This regimen has approximately 85% success rate, despite the deleterious side effects caused by the regimen, such as liver toxicity. Even though TB is a treatable disease, the frequent occurrence of drug resistance to one or more drugs administered to patients is a serious problem. Therefore, it is important to discover new ways to fight *M. tuberculosis*.

One way of targeting *M. tuberculosis* is to inhibit the synthesis of its unique cell wall. In fact, most of the current treatment regimens against TB comprise drugs that target the synthesis of arabinogalactan or mycolic acids, which are main components of the mycobacterial cell wall.

The goal of this work is to perform the structural characterization of a group of membrane proteins, arabinofuranosyltransferases (AraT), involved in the synthesis of arabinan of mycobacterial cell wall. The group is composed by seven different AraTs – AftA, AftB, AftC, AftD, EmbA, EmbB and EmbC – each being involved in different stages of arabinan biosynthesis. Interestingly, EmbA – C are known targets of ethambutol (EMB), a first-line anti-TB drug, while AftA – D are considered new putative targets for the development of novel therapeutics against TB.

Chapter II describes an efficient high-throughput strategy to screen 96 targets of mycobacterial AraTs and to select the best conditions for large-scale protein production, thus enabling further functional and structural studies of selected candidates. We were

able to successfully express in *Escherichia coli* and purify by metal-affinity chromatography 17 AraTs, of which 5 candidates were scaled-up to pursue structural characterization studies.

Chapter III presents the 3D structure of AftD from *Mycobacterium abscessus*, solved by single particle cryo-electron microscopy (cryo-EM) at 2.9 Å resolution. AftD is the largest predicted AraT of *M. tuberculosis* and has a conserved glycosyltransferase type C (GT-C) fold and three carbohydrate-binding modules. Surprisingly, the solved structure of AftD showed a non-covalently bound acyl carrier protein (ACP) from the host organism, *E. coli*. Mutagenesis experiments aimed at disrupting the interaction between AftD and ACP induced conformational changes on AftD active site, suggesting a regulatory role for ACP. In addition, experiments with a conditional knockout constructed in *Mycobacterium smegmatis* confirmed the essentiality of the putative active site and ACP binding for AftD function.

Chapter IV describes the 3D structure of EmbB from *M. smegmatis*, solved by single-particle cryo-EM at 3.3 Å resolution. EmbB also has a conserved GT-C fold and two carbohydrate-binding modules. Several mutations associated with ethambutol resistance are located around the putative active site, suggesting it to be the binding site for ethambutol. In contrast with other reported EmbB structures, our EmbB structure did not show any ACP bound.

Chapter V presents preliminary structural studies of AftA from *Mycobacterium vanbaalenii* in complex with an anti AftA megabody, using single particle cryo-EM. The use of nanobodies and megabodies to form protein complexes with a target protein allows the size of the particles to be increased so to facilitate the workflow of single particle cryo-EM, constituting a useful tool to enable the structure determination of small proteins (<100 kDa).

The work disclosed in this PhD thesis contributed to the understanding of the function of mycobacterial AraTs through the structural characterization of AftD from *M. abscessus* and EmbB from *M. smegmatis*, thereby providing new insight on the role of these proteins in the biosynthesis of the mycobacterial cell wall, as well as providing targets for new drug design studies to search for novel therapeutics against Tuberculosis.

Resumo

A tuberculose (TB) é uma doença infecciosa causada por *Mycobacterium tuberculosis* e é uma das mais antigas doenças infecciosas que afetam a humanidade. De facto, ainda hoje a TB é uma das principais causas de morte em todo o mundo, sendo que se estima que em cada ano cerca de 1.3 – 1.5 milhões de pessoas morram de TB e que surjam 10 milhões de novos casos da doença. Além disso, cerca de 370,000 a 450,000 pessoas desenvolvem TB multirresistente (TB-MDR), para a qual os regimes de tratamento são mais longos e apresentam taxas de sucesso mais baixas.

Esta doença representa um fardo pesado para a humanidade, sendo que muitos desafios precisam ainda de ser superados para erradicá-la. A monitorização, prevenção e o tratamento da TB são aspetos fundamentais para a erradicação da TB, pelo que a investigação e atividades direcionadas na melhoria de diagnóstico, desenvolvimento de vacinas e desenvolvimento de novas terapêuticas são essenciais para atingir as metas propostas para 2030 e 2035 pela Organização Mundial de Saúde (OMS) na estratégia de erradicação da TB – "End TB Strategy".

O tratamento da TB sensível a fármacos consiste num regime diário de isoniazida, rifampicina, pirazinamida e etambutol, durante 4 a 6 meses. Este regime tem uma taxa de sucesso de aproximadamente 85%, apesar dos efeitos colaterais nocivos causados pelo regime, como a toxicidade hepática. Embora a TB seja uma doença tratável, a frequente resistência a um ou mais medicamentos administrados aos pacientes é um problema grave. Portanto, é importante descobrir novas formas de combater o agente causador de TB: *M. tuberculosis*.

Uma forma de atingir a micobactéria *M. tuberculosis* é inibir a síntese da sua distinta parede celular. Na verdade, a maioria dos regimes de tratamento atuais contra a TB compreende medicamentos que visam a síntese de arabinogalactano ou a síntese de ácidos micólicos, sendo estes os principais componentes da parede celular das micobactérias.

O objetivo deste trabalho é realizar a caracterização estrutural de um grupo de proteínas membranares, arabinofuranosiltransferases (AraT), envolvidas na síntese de arabinano da parede celular de micobactérias. O grupo de proteínas membranares é composto por sete AraTs diferentes – AftA, AftB, AftC, AftD, EmbA, EmbB e EmbC – cada uma envolvida em diferentes etapas da biossíntese de arabinano. Curiosamente, as proteínas EmbA – C são alvos já conhecidos do etambutol, um fármaco de primeira linha

anti-TB, as proteínas AftA – D são considerados novos alvos putativos para o desenvolvimento de novos fármacos contra a TB.

O Capítulo II descreve uma estratégia eficiente de triagem de alto rendimento de 96 proteínas AraTs de micobactérias e de seleção das melhores condições para a produção de proteínas AraTs em larga escala, permitindo assim realizar estudos funcionais e estruturais das proteínas selecionadas. Nesse trabalho, conseguimos expressar com sucesso em *Escherichia coli* e purificar por cromatografia de afinidade 17 AraTs. Do conjunto de 17 AraTs selecionadas, a produção de 5 AraTs candidatas foi escalada para permitir estudos de caracterização estrutural.

O Capítulo III apresenta a estrutura 3D da AftD de *Mycobacterium abscessus*, resolvida por crio-microscopia eletrónica (cryo-EM) de partícula única com resolução de 2.9 Å. AftD é a maior AraT prevista de *M. tuberculosis* e possui um *fold* conservado característico de glicosiltransferases de tipo C (GT-C) e possui três módulos de ligação a carboidratos. Surpreendentemente, a estrutura resolvida da AftD revelou uma proteína de transporte de acilo (ACP) ligada de forma não covalente, proveniente do organismo hospedeiro, *E. coli*. Experiências de mutagénese dirigida com o intuito de interromper a interação entre AftD e ACP induziram alterações conformacionais no centro ativo da AftD, sugerindo que ACP terá um papel regulador. Além disso, experiências com um mutante KO condicional desenvolvido em *Mycobacterium smegmatis* confirmaram que o presumível centro ativo da AftD e a interação AftD-ACP são essenciais para a função da AftD.

O Capítulo IV descreve a estrutura 3D da EmbB de *M. smegmatis* resolvida por cryo-EM de partícula única com resolução de 3.3 Å. EmbB também possui um *fold* conservado característico GT-C e dois módulos de ligação a carboidratos. Várias mutações associadas à resistência ao etambutol estão localizadas à volta do presumível centro ativo da proteína, sugerindo que seja este o local de ligação do etambutol à proteína. Em contraste com outras estruturas relatadas de EmbB, a nossa estrutura de EmbB não revelou qualquer ligação da EmbB com ACP.

O Capítulo V apresenta estudos estruturais preliminares de AftA de *Mycobacterium vanbaalenii* em complexo com um anticorpo do tipo *megabody* anti-AftA, usando cryo-EM de partícula única. A utilização de anticorpos do tipo *nanobody* ou *megabody* permite aumentar o tamanho das partículas para facilitar o fluxo de trabalho de cryo-EM de partícula única na determinação da estrutura 3D de proteínas, constituindo uma ferramenta útil para permitir a determinação da estrutura de pequenas proteínas (<100 kDa).

O trabalho desta tese de doutoramento contribuiu para a compreensão da função dos AraTs de micobactérias através da caracterização estrutural de AftD de *M. abscessus* e de EmbB de *M. smegmatis*, proporcionando assim uma nova visão sobre o papel destas proteínas na biossíntese da parede celular de micobactérias, bem como contribuiu para fornecer alvos para novos estudos de desenvolvimento de fármacos para procurar novos tratamentos contra a tuberculose.

List of Abbreviations

ACP	acyl carrier protein
AcpM	meromycolate extension acyl carrier protein
AftA	arabinofuranosyltransferase AftA
AftB	arabinofuranosyltransferase AftB
AftC	arabinofuranosyltransferase AftC
AftD	arabinofuranosyltransferase AftD
AG	arabinogalactan
AIDS	Acquired Immunodeficiencye Deficiency Syndrome
Ara	arabinose
Ara*f*	arabinofuranosyl
AraT	arabinofuranosyltransferases
ATP	adenosine triphosphate
BCE	Before Common Era
BCG	Bacille Calmette-Guerin
BSA	bovine serum albumin
CAD	Computer-Aided Detection
CBM	carbohydrate binding module
CD-HIT	Cluster Database at High Identity with Tolerance
CDR	complementarity-determining region
cKO	conditional knockout
COVID-19	Coronavirus Disease 2019
cryo-EM	cryogenic electron microscopy
CTF	contrast transfer function
CV	column volumes
DDM	n-dodecyl-β-D-maltopyranoside

D-GalN	galactosamine
DNA	deoxyribonucleic acid
DPA	decaprenyl-monophosphoryl-D-arabinose
DPG	diphosphatidylglycerol
DR-TB	Drug-r Resistant Tuberculosis
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
EMB	Ethambutol
EmbA	arabinofuranosyltransferase A
EmbB	arabinofuranosyltransferase B
EmbC	arabinofuranosyltransferase C
f	furanose
FCT	Fundação para a Ciência e Tecnologia
FDR	false discovery rate
FSC	Fourier shell correlations
Gal	galactose
Gal*f*	galactofuranosyl
GDN	diosgenin
GFP	green fluorescent protein
GlcNAc-P	N-acetylglucosamine phosphate
GT	glycosyltransferase
GT-C	glycosyltransferase type C
HEPES	4- (2-hydroxyethyl)-1-piperazineethanesulfonic acid
HFW	horizontal field width
HIV	Human Immunodeficiency Virus
HTP	hHigh- throughput
IDA	information dependent acquisition

IGRA	Interferon-gamma release assay
IPTG	isopropyl β-D-1-thiogalactopyranoside
JCSG	Joint Center for Structural Genomics
JM	juxtamembrane
LAM	lipoarabinomannan
LB medium	Luria-Bertani medium
LC-MS	liquid chromatography-–mass spectrometry
LIC	Ligation Independent Cloning
LM	lipomannan
LOS	lipooligosaccharides
mAGP	mycolyl-arabinogalactan-peptidoglycan
Mb	megabody
MD	Molecular Dynamics
MDR-TB	Multidrug- Rresistant Tuberculosis
mRNA	messenger ribonucleic acid
MS	mass spectrometry
MS/MS	tandem mass spectrometry
Msm	*Mycobacterium Mycolicibacterium smegmatis*
MSP	membrane scaffold protein
MSP-1E3D1	membrane scaffold protein 1E3D1
Mtb	*Mycobacterium tuberculosis*
MurNGly	N-glycolylmuramic acid
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
MW	molecular weight
NAG	N-acetylglucosamine
NAM	N-acetylmuramic acid
nanoLC-MS/MS	Nano-liquid chromatography-tandem mass spectrometry

Nb	nanobody
ND	nanodiscs
Ni^{2+}-NTA	nickel nitrilotriacetate
NIH	National Institutes of Health
NTM	nontuberculous mycobacterial disease
OD	optical density
PBS	phosphate buffered saline
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PE	phosphatidylethanolamine
PEG	polyethylene glycol
PG	Peptidoglycan
PGL	phenolic glycolipids
PGly	phosphatidylglycerol
PI	phosphatidylinositol
PIM	phosphatidyl-myo-inositol mannoside
PMSF	phenylmethylsulfonyl fluoride
POPC	1-palmitoyl-2-oleoyl-glycero-3-phosphocholine
POPG	palmitoyl-2-oleoyl-sn-glycero-3-phospho-(1′-rac-glycerol)
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
RP-LC	reversed-phase chromatography
RR-TB	rifampicin-resistant Tuberculosis
SCF	sampling compensation factor
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC	size exclusion chromatography
SGC	Structural Genomics Consortium
SPR	Surface Plasmon Resonance

TB	Tuberculosis
TCEP	tris(2-carboxyethyl)phosphine hydrochloride
TEM	Transmission Electron Microscope
TEV	tobacco etch virus
TM	transmembrane
WHO	World Health Organization
WT	wild-type
XDR-TB	Extensively Drug ResistantDrug-Resistant Tuberculosis

Table of Contents

List of Figures

List of Tables

CHAPTER ONE

General Introduction

Section One

1.1 Historic significance of Tuberculosis

Tuberculosis (TB) is an old disease that afflicts humankind for thousands of years. The disease is referenced in the Bible by the ancient Hebrew word schachepheth [1] while the ancient Greeks used the word phthisis, meaning “wasting away”. TB was depicted in Egyptian art of 5000 years ago and signs of the illness were found in Egyptian mummies skeletons [2]. In fact, the first evidence of human TB was observed in infected bones from the Neolithic period (3000-5000 BC), showing typical macroscopic changes of the spine [3] (Figure 1.1).

Hippocrates (460 – 370 BCE), extensively described TB in his writings, recognizing the distinctiveness of the disease among other afflictions of that era. Hippocrates stated that “The greatest and most dangerous disease, and the one that proved fatal to the greatest number, was consumption.” (2.3.13, Epidemics) [4], described pre-terminal symptoms of TB, such as diarrhea (5.14, Aphorisms) or hair loss (5.11, Aphorisms) and recognized that “phthisis most commonly occurs between the ages of 18 and 35 years” (5.9, Aphorisms) [5][6].

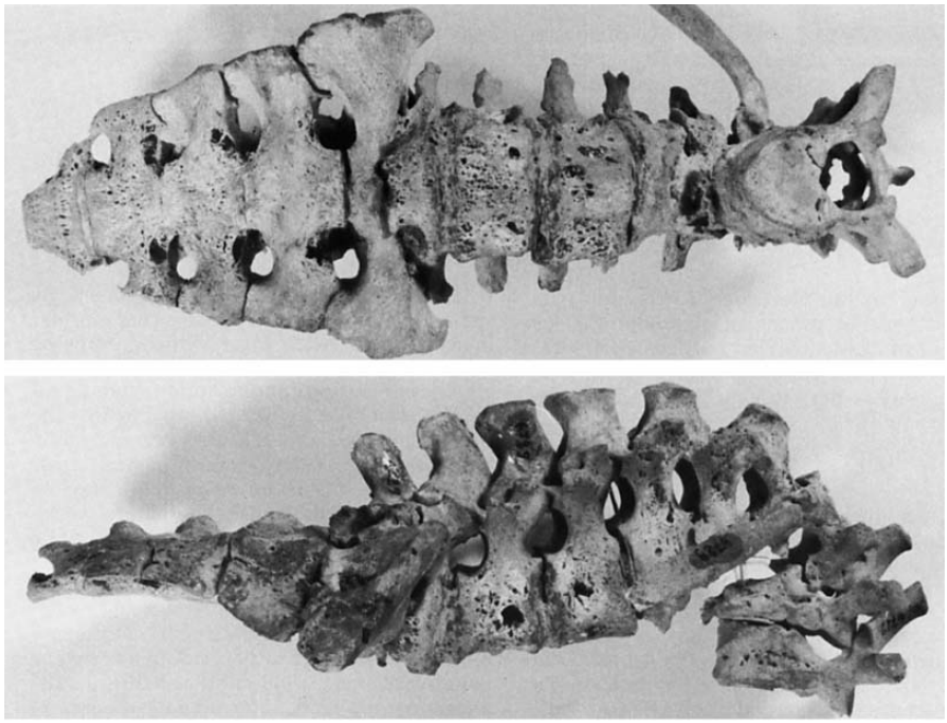


***Figure 1.1** – Spinal bones from the Neolithic period (3000-5000 BC) presenting increased porosity and cavitations, typical traits of skeletal Tuberculosis. [3]*

During the Middle Ages, another form of TB which affected cervical lymph nodes became widely known as scrofula or “King’s Evil”. This form of the disease was believed to be diagnosed and cured by royal touch from the Kings of England and France. For more than 500 years, people suffering from scrofula attended distinct ceremonies to be touched, in the neck or head, by the king or queen to be cured from the disease [7] (Figure 1.2).

There were significant advances in diagnosis and in the study of TB, from late 17th century onwards. The infectious nature of the disease was recognized, and public health measures were adopted in the former Republic of Lucca (Italy), stating that “human health should no longer be endangered by objects remaining after the death of a consumptive. The names of the deceased should be reported to the authorities, and measures undertaken for disinfection.” [8]. But it would be Benjamin Marten who would first suggest that TB was a communicable disease, caused by “wonderfully minute living creatures”, and that extended periods of close contact with a sick person with TB was a requirement to become infected [9].

Figure 1.2 – Painting depicting the Royal Touch practice: Queen Mary I of England touches the neck of a boy to cure scrofula or King’s evil. *Watercolor by M.S. Lapthorn, 1911. [7]*

One very important development was made by a French military surgeon, Jean-Antoine Villemin, who observed that soldiers that were stationed for long periods in close quarters such as barracks, became sick with TB more frequently than those who were on the field [10]. He successfully confirmed the infectious nature of TB in 1865 by inoculating a healthy rabbit with a “purulent liquid from a tuberculous cavity” removed during the autopsy of a person who perished from TB. The rabbit then developed extensive TB as it was observed once it was sacrificed and autopsied after 3 months [2].

From the Villemin demonstration, attempts to isolate the virulent source of TB were performed by Theodor Klebs, who identified mobile bacilli from tuberculous material cultivated in egg whites. These bacilli were able to cause TB in Guinea pigs [11].

The scientific movement started by Villemin defending the infection theory culminated in the discovery of the *Mycobacterium tuberculosis* (*Mtb*) bacillus by Dr. Robert Koch in 1882 [12] (Figure 1.3).

To finally settle the dispute over the infectious nature of TB, Robert Koch applied the knowledge acquired regarding many infectious diseases:

- First, it was necessary to know the elements (*i.e.* microorganisms) that were present in the diseased parts of the body, whilst not belonging to the body.
- If such elements would exist, they must have been removed from the diseased part and got rid of any sign of the disease, to clarify if the element would be the cause of the disease or a mere accompaniment of the disease.
- Finally, once fully isolated, such elements should have the capacity to cause the disease when introduced to healthy organisms.

Each step involved several significant scientific developments, such as improving staining methods to detect *Mtb* bacilli, isolate and cultivate of *Mtb* for several generations using stiffened blood serum as solid media [12].

This pivotal work of Robert Koch was presented on the 24th March 1882 to the Society of Physiology in Berlin, and was further enlarged and compiled in the book “The Aetiology of Tuberculosis” [12].

The discovery of *Mtb* bacillus was the cornerstone for the science-based fight against TB, and Koch’s presentation is commemorated every year on March 24 as World Tuberculosis Day, since 1982, to raise awareness about the disease [13].

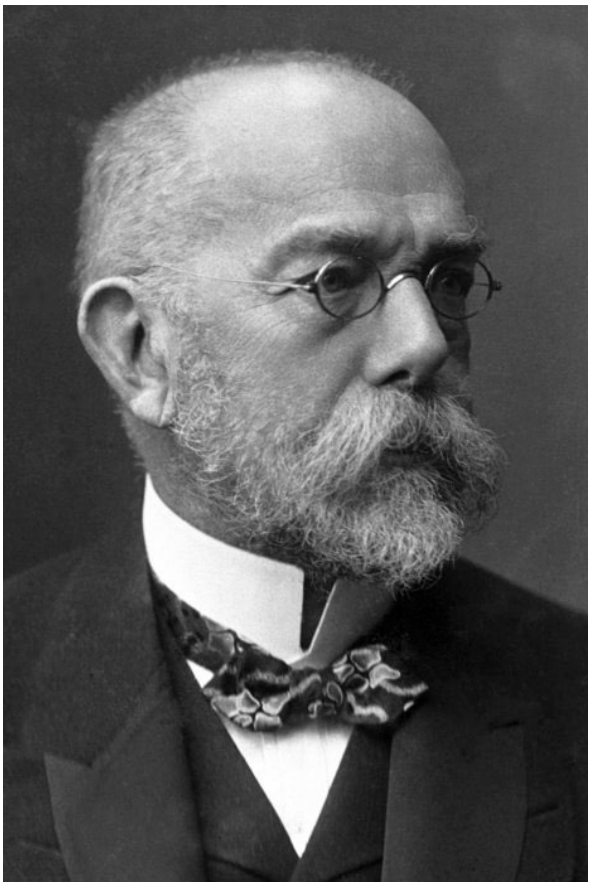


Figure 1.3 – Robert Koch, Nobel laurate in Physiology or Medicine in 1905 "for his investigations and discoveries in relation to tuberculosis".
Source: www.nobelprize.org/images/koch-12845-portrait-mini-2x.jpg

Koch's research paved the way to diagnosis and treatment of TB. In 1890, Koch isolated tuberculin, a substance from *Mtb* first intended to "render harmless the pathogenic bacteria", however it was proved to be ineffective as treatment for TB [2]. Despite not being suited to treat TB, tuberculin was widely used for diagnosis and tuberculin skin tests are still used today [10] (Figure 1.4).

Another major development in the diagnosis of TB was the discovery of X-rays by Wilhelm Konrad von Roentgen in 1985, which allowed the direct observation of the lungs by chest radiography, becoming a powerful tool to diagnose pulmonary TB [9] (Figure 1.5).

The last advance before the discovery of antibiotics against TB in the 1940s was the Bacille Calmette-Guerin (BCG) vaccine, developed by Albert Calmette and Camille Guerin in 1906 [9]. It consisted in a non-virulent strain of *Mycobacterium bovis* formulated into a live attenuated vaccine. It was widely administered and still is to this day, especially in newborns from countries where there is increased risk of TB or leprosy.

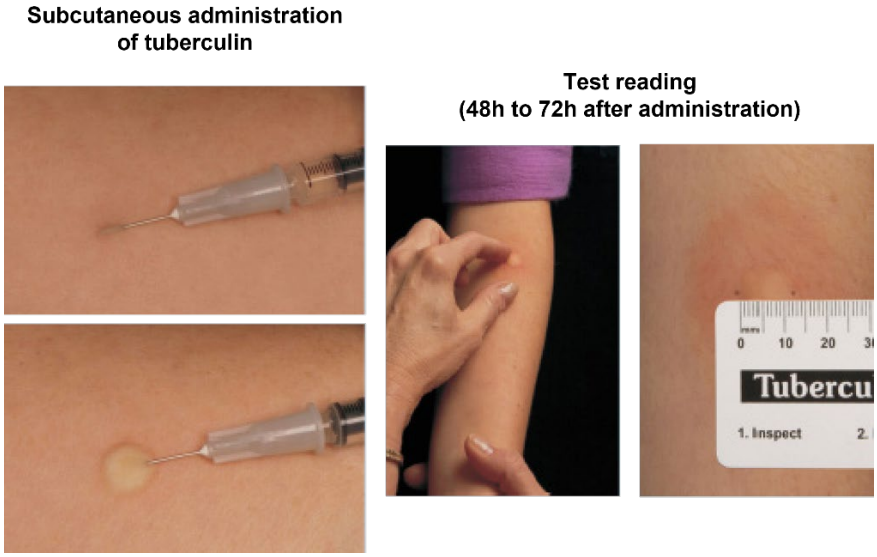


Figure 1.4** – **Administration and reading of Tuberculin tests.
Source: www.cdc.gov/tb/education/mantoux/pdf/Mantoux_TB_Skin_Test.pdf

Figure 1.5** – **Chest X-ray screening for Tuberculosis.
Sources: www.archives.gov.on.ca/en/explore/online/health_promotion/pics/5202-chest-x-ray-520.jpg; www.som.org.uk/occupational-health-learning/images/large/Tuberculosis.jpg

The BCG vaccine is effective in the prevention of extra pulmonary TB up to the age of 15 years, however it does not offer protection against pulmonary TB infections in adults [14].

With the discovery of antibiotics against TB, starting with the isolation of streptomycin from the fungi *Streptomyces griseus* in 1943 by Selman Waksman, it became possible to treat TB without the need of surgical procedures such as pneumothorax, lobectomies and segmentectomies, which often lead to complications [8]. However, with the use of antibiotics came the problem of drug resistance, which stands as a major challenge in modern Medicine, not exclusive to TB. The importance of antibiotics and subsequent drug resistance will be further discussed in later sections.

TB is a disease that shaped human history. It is one of the oldest diseases to afflict mankind and therefore left a mark in historical documents, in visual arts and literature. But the greatest impact of TB was in the scientific progress attained in the fight against it. Several Nobel prizes were awarded for research linked to or widely applied to TB: Wilhelm Konrad von Roentgen in 1901 for the discovery of X-rays, Robert Koch in 1905 for the work in the *Mtb* bacillus and Selman Waksman in 1952 for the discovery of streptomycin.

Even so, the fight against TB continues.

1.2 Tuberculosis etiology

Tuberculosis (TB) has been known by many names since mankind started to document this ailment: consumption, the “White Plague” or even “Captain of Death”, which sounds particularly terrifying.

So, what is this disease, really?

TB is an infectious disease caused mainly by *Mycobacterium tuberculosis* (*Mtb*), although it can also be caused by *Mycobacterium africanum* especially in West Africa and in some parts of East Africa [15]. Other zoonotic *Mycobacterium* species, such as *M. bovis* or *M. caprae*, can also cause TB in humans in rare occasions [16].

TB transmission occurs through the air, which means that if a healthy person has prolonged contact with an infected person with active pulmonary TB, there is a possibility that airborne transmission of *Mtb* occurs. Such transmission does not mean that the disease will develop upon contact with *Mtb*. In fact, the most common outcome is either the total elimination of *Mtb* bacilli by the immune system of the new host; or the occurrence of a latent infection that does not develop into TB.

Nevertheless, there are susceptible people that carry higher risk of developing TB following the acquisition of the infection. There is an increased risk for people infected with HIV and people suffering from malnutrition, but these are not the only conditions associated with higher risk, as there are several other medical conditions that also favor the occurrence of TB [17]. Therefore, after contact with *Mtb* there are three possible scenarios: total eradication of the pathogen, latent TB infection or active TB infection (Figure 1.6), wherein the different risk factors play a role to define the outcome [16].

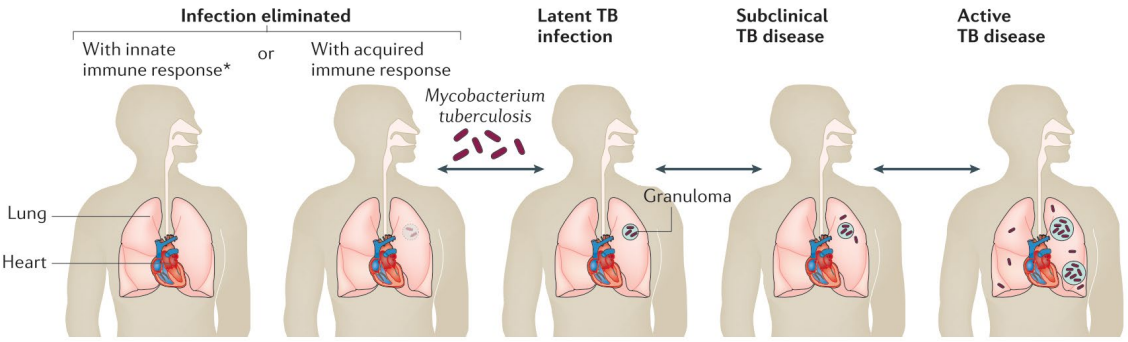


***Figure 1.6** – **Transmission and infection development of Tuberculosis.** Adapted from M. Pai et al. (2016). [16]*

Active and Latent TB Infection

During the airborne transmission to a new host, the bacilli travel through the respiratory tract and reach the alveoli, in the lungs. Once they reach the alveoli, alveolar macrophages, dendritic cells and neutrophils act to eliminate the foreign intruders by phagocytosis, thus internalizing the bacilli to destroy the foreign body.

Once internalized by host cells, mycobacteria start to undermine normal functions of the host cell by reducing the acidity in the phagosome, blocking the fusion of the phagosome with the lysosome and inhibiting apoptosis and autophagy [16][18]. By doing so, mycobacteria can survive inside the macrophage, the infected macrophage can cross host epithelial barriers and further disseminate the pathogen, while the bacteria replicates until they are released upon host cell death. The release of *Mtb* bacilli leads to infection of other immune cells, which in turn recruit other cells to the primary infection site, such as macrophages, neutrophils, lymphocytes and dendritic cells. This accumulation of cells leads to the formation of granulomas, which are hallmarks to TB [19][20].

Granulomas have a basic structure comprising a central necrotic core (caseous) surrounded by macrophages and by an outer layer of lymphocytes comprising B cells or T cells [21] (Figure 1.7). In addition, other cells such as neutrophils, monocytes, dendritic cells can also be found in the granuloma, which influences the structure and type of granuloma: caseous granulomas, fibrotic granulomas and non-necrotizing granulomas [18][20].

On one hand, granulomas are a primary host defense mechanism formed in the site of infection that encapsulate the bacteria and shields the rest of the body from the pathogen, thus acting as a prison for the bacteria. On the other hand, this prison also shields the bacteria from other immune responses, thus the bacteria may stay dormant for long periods of time and re-activate once there is an opportunity. At the granuloma, the bacteria can continue to infect macrophages and to recruit new immune cells, expanding the bacterial niche despite remaining “incarcerated” inside the granuloma [19].

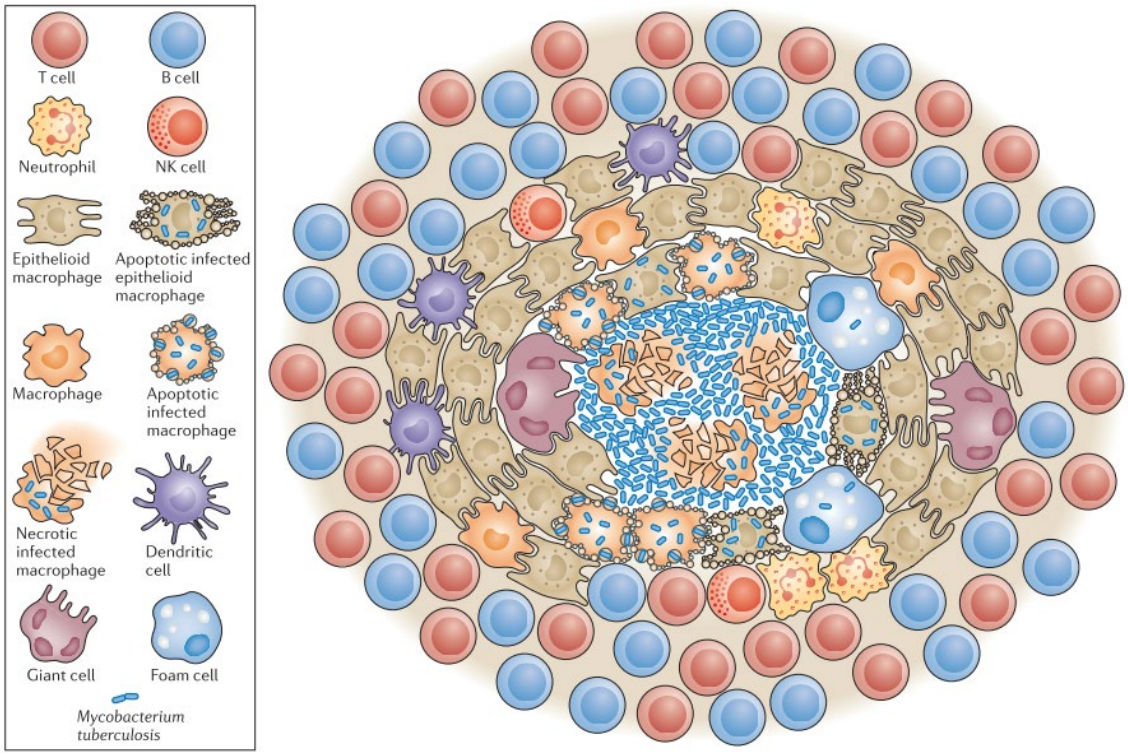


Figure 1.7 – Structure of a tuberculous granuloma and its cellular components. *Adapted from L. Ramakrishnan, et al (2012) [21].*

As the infected macrophages die, bacilli are released along with cell debris, forming heterogeneous aggregates that are internalized by new recruited macrophages. This "snowball" effect leads to a continuous macrophage death and increased *Mtb* growth and replication in dead cells [22].

As long as the granuloma is able to control the growth of bacteria and no dissemination of *Mtb* occurs in other parts of the body, the host remains asymptomatic and not infectious, thus having a latent TB infection. These individuals can receive preventive treatment to either clear the infection or keep it under control without it becoming active. However, if the granulomas cannot control bacterial growth, they rupture and damage the lung parenchyma and disseminate *Mtb* bacilli to other parts of the lung. In this scenario, the TB infection becomes active as the bacteria can spread from the primary infection location and colonize the lung, starting on the peripheral lymph nodes. The host then becomes infectious with the most common form of active TB – Pulmonary TB, or can even develop Extrapulmonary TB, in which the infection reaches other organs: brain – Tuberculous meningitis; bones – Skeletal Tuberculosis, Spinal tuberculosis or Pott's disease; gastrointestinal track; or genitourinary track [23]. Extrapulmonary TB represents roughly 15% of all TB infections and appears in immunocompetent and

immunocompromised persons, wherein people with HIV are particularly more susceptible to develop TB after a primary infection [23][24].

For people with active TB without access to treatment, the mortality rate is high, wherein 70% of individuals with sputum smear-positive pulmonary TB and 20% of individuals with culture-positive/sputum smear-negative pulmonary TB die within 10 years after being diagnosed [25].

Access to diagnosis tools and treatment are key to reduce TB infection and mortality rates and stand as central pillars in the strategy to reduce the global burden of TB.

1.3 Burden of Tuberculosis

TB is an infectious disease that still stands as one of the leading causes of death in the world and the second most deadly disease caused by single infectious agent, ranking above AIDS and only surpassed recently by COVID-19 [26][27].

It is estimated that a quarter of the world's population, around 2 billion people, has been infected with *Mtb.* In 2022, 1.3 million people died from TB, 10.6 million people developed TB (Figure 1.8), the number of people newly diagnosed with TB increased to 7.5 million and the number of cases of drug-resistant TB (DR-TB) are estimated to be 410 000 [27].

Before 2019, the number of deaths and mortality rates were steadily decreasing for more than a decade (since 2005), but the COVID-19 pandemic had an enormous impact on the worldwide fight against TB. In fact, it is estimated that disruptions caused by COVID-19 have resulted in almost half a million excess deaths from TB between 2020 and 2022, compared to the expected number, if pre-pandemic trends had been maintained [27]. In 2020, the number of reported new cases decreased dramatically (Figure 1.9) while the number of deaths and mortality rate increased [26][27]. With the 2023 publication of the WHO Global Tuberculosis Report, it is evident that there is a return to the trend of yearly decrease of deaths and mortality rates, which were increasing between 2019 and 2021 (Figure 1.10) [26][27].

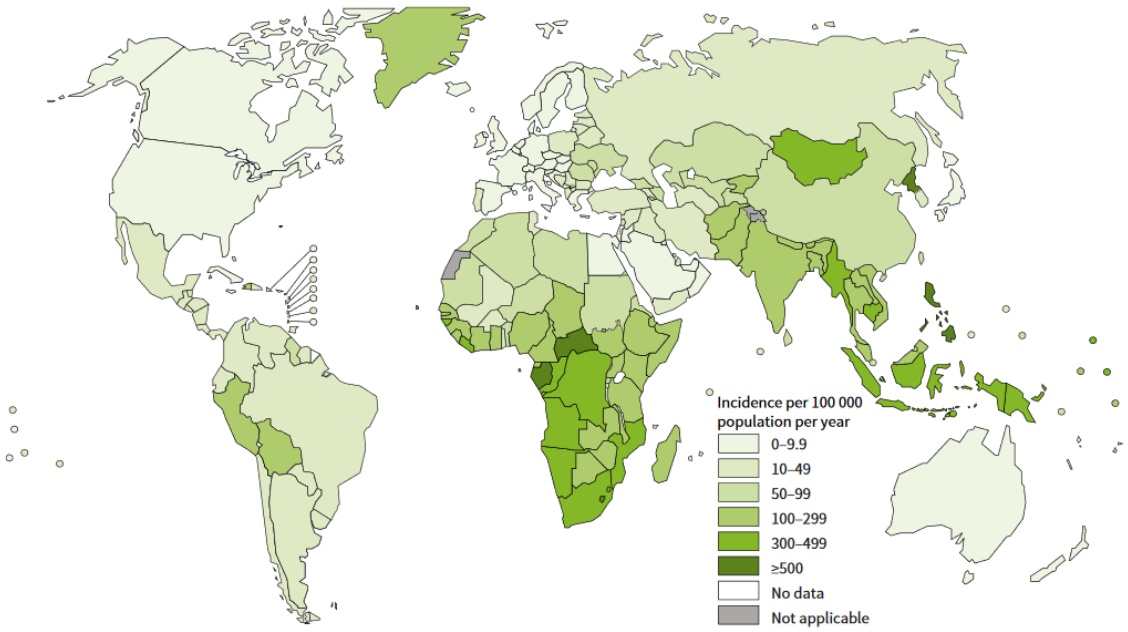


***Figure 1.8** – **Global incidence of Tuberculosis,** WHO Global Tuberculosis Report 2023.*

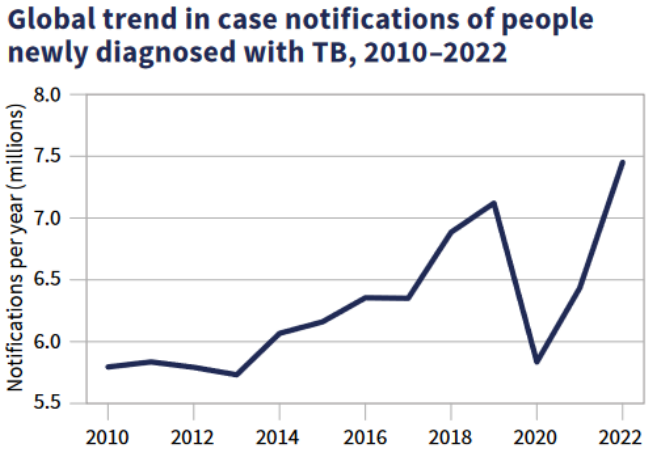


Figure 1.9 *– **Notifications of newly diagnosed cases of Tuberculosis**, WHO Global Tuberculosis Report 2023.*

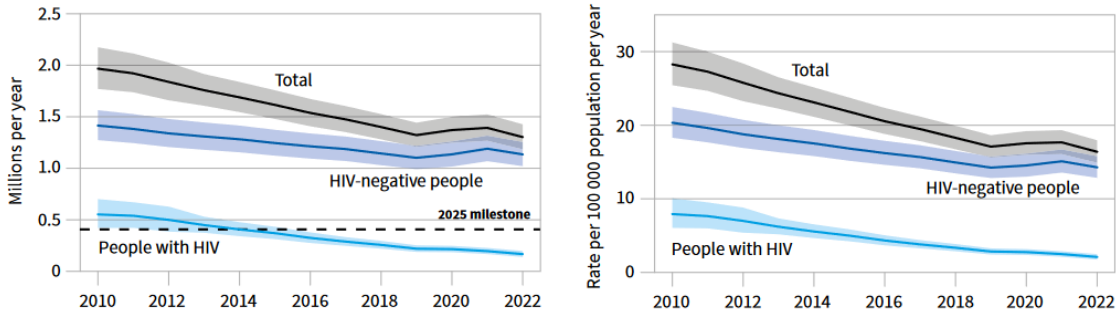


***Figure 1.10 – Global trends in the estimated number of deaths by Tuberculosis (left) and Tuberculosis mortality rate (right)**, WHO Global Tuberculosis Report 2023.*

TB infection can be treated with antibiotics, and the recommended treatment comprises a 6-month administration of four first-line drugs: isoniazid, rifampicin, pyrazinamide and ethambutol, which usually has a success rate of 86%. However, the emergence of drug-resistant TB is a major public health threat, especially for multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB). MDR-TB is a type of TB that is resistant to the two most potent first-line drugs: isoniazid and rifampicin, while XDR-TB is a type of MDR-TB that is also resistant to any fluoroquinolone and at least one of the second-line drugs (levofloxacin, moxifloxacin, bedaquiline and linezoid). Every year, it is estimated that 470 000 people develop MDR-TB and 180 000 people die from it [28].

The road to eradicate TB is sinuous. It takes a continuous global effort in several fronts to decrease gradually the burden of this disease. Events like the COVID-19 pandemic, war and economic crisis have worldly vast impact in the fight against TB, namely in prevention, identification, and treatment. Therefore, both applied and fundamental research play an important role to discover new ways to fight TB, to develop new diagnostic tools, new effective ways to treat TB and MDR-TB and to increase the knowledge of mycobacteria.

Section Two

2.1 Tuberculosis treatment: present and future perspectives

As it is clear from the previous section, TB is a complex disease with several forms of active TB, the most common being Pulmonary TB. In addition, latent TB is a condition similar to having a "ticking-time bomb" inside the body, as the infection can stay dormant for a long period of time but can eventually switch to an active form and wreak havoc. Therefore, treatment strategies to eliminate or control *Mtb* infections are the most important weapons to save people's lives.

Preventive treatment

Preventive treatment is the available health care intervention that reduces the risk of latent TB infection developing to an active infection. This regimen is mostly administered to people with HIV, as they are more susceptible to develop TB disease, but is also administered to people who had contacts with *Mtb*. The preventive treatment regimen options are summarized in Table 1.1 [26][27].

***Table 1.1** – **WHO recommended preventive Tuberculosis treatment regimens.** [26][27]*

Regimen	Drugs Administered		Frequency	Duration
3HP	Rifapentine	Isoniazid	Weekly	3 months
3HR	Rifampicin	Isoniazid	Daily	3 months
1HP	Rifapentine	Isoniazid	Daily	1 month
4R	Rifampicin		Daily	4 months
6H	Isoniazid		Daily	6 months or longer

Treatment of active drug-susceptible TB

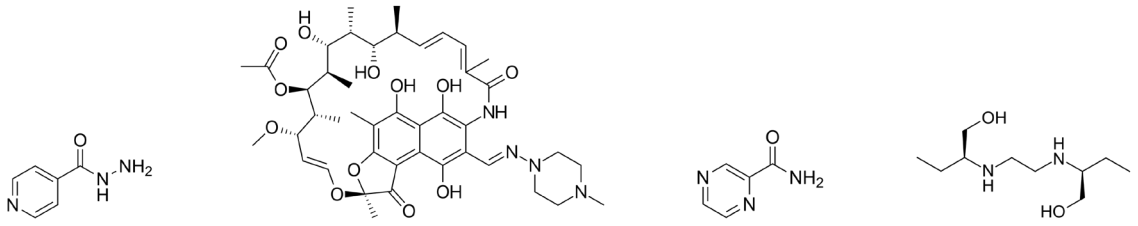
The WHO estimated that 80% of TB cases are drug-susceptible and 20% are resistant to at least one drug. The recommended treatment regimen for drug-susceptible TB consists in an intensive phase of isoniazid, rifampicin, pyrazinamide and ethambutol administration for 2 months, followed by a continuation phase of isoniazid and rifampicin

administration for 4 months (2HRZE/4HR). This regimen has an approximate 85% success rate, however the long duration of the treatment may cause liver toxicity and other complications. Therefore, close monitoring of deleterious effects is recommended.

There are other treatment regimens recommended depending on the age of patients or if there are pre-existing conditions, especially for people with HIV. The recommended treatment regimens for drug susceptible TB are summarized in Table 1.2 [29]. The first-line anti-TB drugs used in the treatment of active drug-susceptible TB are represented in Figure 1.11.

***Table 1.2** – **WHO recommended active Tuberculosis treatment regimens.** [29]*

New patients with pulmonary TB			
Regimen	Drugs Administered	Frequency	Duration
2HRZE/4HR	Isoniazid Rifampicin Pyrazinamide Ethambutol	Daily	6 months
People aged 12 years or older with pulmonary TB			
2HPMZ/2HPM	Isoniazid Rifapentine Moxifloxacin Pyrazinamide	Daily	4 months
Children and adolescents between 3 months and 16 years of age with non-severe TB			
2HRZ(E)/2HR	Isoniazid Rifampicin Pyrazinamide Ethambutol	Daily	4 months



Isoniazid Rifampicin Pyrazinamide Ethambutol

Figure 1.11 – First-line anti TB drugs used to treat active drug-susceptible TB.

Treatment of active drug-resistant TB (DR-TB)

The treatment regimens for drug-resistant TB (DR-TB) vary according to category of drug resistance. There are five categories to classify the cases of DR-TB: isoniazid-resistant TB, rifampicin-resistant TB (RR-TB), multidrug-resistant TB (MDR-TB), extensively drug-resistant TB (XDR-TB) and pre-XDR-TB.

Drug resistance and susceptibility of all TB patients should be confirmed to adjust the treatment regimen accordingly. Depending on the resistance profile of the *Mtb* strain, the drug cocktail composition may vary, and the duration of the treatment may also be shorter or longer. For instance, a 6 month-treatment regimen is recommended for patients with rifampicin-susceptible/isoniazid-resistant TB (Table 1.3).

However, for patients with MDR-TB the treatment regimens are longer and the drugs administered vary according to the resistance profile of the strain. The drugs administered in longer regimens are presented in Table 1.4. The recommended treatment regimens for DR-TB are summarized in Tables 1.3 and 1.5 [30].

Table 1.3 – Treatment regimen recommended for rifampicin-susceptible/isoniazid-resistant TB.
Adapted from WHO consolidated guidelines on tuberculosis [30]

Rifampicin-susceptible / Isoniazid-resistant TB	
Drugs Administered	Duration
Rifampicin Ethambutol Pyrazinamide Levofloxacin	6 months

Table 1.4 – ***Grouping of medicines recommended for use in longer MDR-TB regimens.*** *Adapted from WHO consolidated guidelines on tuberculosis [30]*

Groups	Drugs
A	Levofloxacin / Moxifloxacin
	Bedaquiline
	Linezolid
B	Clofazimine
	Cycloserine / Terizidone
C	Ethambutol
	Delamanid
	Pyrazinamide
	Imipenem–cilastatin / Meropenem
	Amikacin / Streptomycin
	Ethionamide / Prothionamide
	P-aminosalicylic acid

***Table 1.5** – **Recommended treatment regimens for MDR-TB patients.** Adapted from WHO consolidated guidelines on tuberculosis [30]*

Shorter Regimen for MDR-TB			
Eligible patients	**Drugs Administered**		**Duration**
Confirmed MDR-TB not yet exposed to treatment with second-line medicines	Intensive phase (4 to 6 months)	Bedaquiline Levofloxacin/moxifloxacin Ethionamide Ethambutol Isoniazid (high-dose) Pyrazinamide Clofazimine	9 – 12 months
	Continuation phase (5 months)	Levofloxacin/moxifloxacin Clofazimine Ethambutol Pyrazinamide	
Longer Regimen for MDR-TB			
Eligible patients	**Drugs Administered**		**Duration**
Patients with all forms of MDR-TB	At least 4 drugs administered [1]	All Group A drugs + At least 1 Group B	18 – 20 months [2] (duration may vary according to patient's response to treatment)
		2 Group A drugs + All Group B drugs	
		If the regimen cannot be composed with agents from groups A and B alone, group C agents are added to complete it.	

[1] At least 3 drugs administered if Bedaquiline is stopped.

[2] regimen duration may vary according to patient's response to treatment.

2.2 Role of research in the fight against TB

Research and innovation are essential in the fight against TB to reach the proposed targets for 2030 and 2035 of the WHO “End TB Strategy”.

At the moment, research is focused on improving diagnostics, developing new drugs and developing effective vaccines against TB. The most recent research developments are presented in Table 1.6.

Table 1.6 – Recent or current research developments in the fight against TB. [27]

<table>
<tr><th colspan="3">Diagnostics</th></tr>
<tr><td>Molecular tests</td><td colspan="2" rowspan="5">Detection of TB disease</td></tr>
<tr><td>New antigen-based skin tests</td></tr>
<tr><td>Interferon-gamma release assays (IGRAs)</td></tr>
<tr><td>Biomarker-based assays</td></tr>
<tr><td>Aerosol-capture technologies</td></tr>
<tr><td>Computer-aided detection (CAD)</td><td colspan="2">Detection of TB disease (digital chest radiography)</td></tr>
<tr><td>Molecular tests</td><td colspan="2" rowspan="2">Drug resistance</td></tr>
<tr><td>Culture-free targeted-sequencing solutions</td></tr>
<tr><td>Broth microdilution methods</td><td colspan="2">Drug susceptibility</td></tr>
<tr><th colspan="3">Drugs</th></tr>
<tr><td rowspan="4">Phases I, II or III
28 drugs</td><td colspan="2">18 new chemical entities</td></tr>
<tr><td>7 repurposed drugs</td><td>clofazimine, levofloxacin, linezolid, moxifloxacin, rifampicin (high dose), rifapentine and tedizolid</td></tr>
<tr><td>2 drugs with accelerated regulatory approval</td><td>bedaquiline, delamanid</td></tr>
<tr><td>1 drug approved by US FDA</td><td>pretomanid</td></tr>
<tr><td>Clinical Trials
29 drugs</td><td colspan="2"></td></tr>
<tr><th colspan="3">Vaccines</th></tr>
<tr><td rowspan="3">16 vaccine candidates</td><td colspan="2">4 candidates in Phase I</td></tr>
<tr><td colspan="2">8 candidates in Phase II</td></tr>
<tr><td colspan="2">4 candidates in Phase III</td></tr>
</table>

Section Three

3.1 The Cell Envelope of Mycobacteria: the essential barrier

Most of the current treatment regimens against TB comprise drugs that target an essential component of mycobacteria: the cell envelope (Figure 1.12).

In particular, the cell wall is a major part of the cell envelope that protects the bacteria from external harms, allowing the pathogen to survive in harsh environments, to resist the action of several drugs and to persist for long periods of time inside the host. It is significant that a single cellular compartment is targeted by a plurality of inhibitors, thus establishing that the mycobacterial cell wall is essential for the survival of the pathogen.

So, what differentiates the mycobacterial cell wall from other bacterial cell walls?

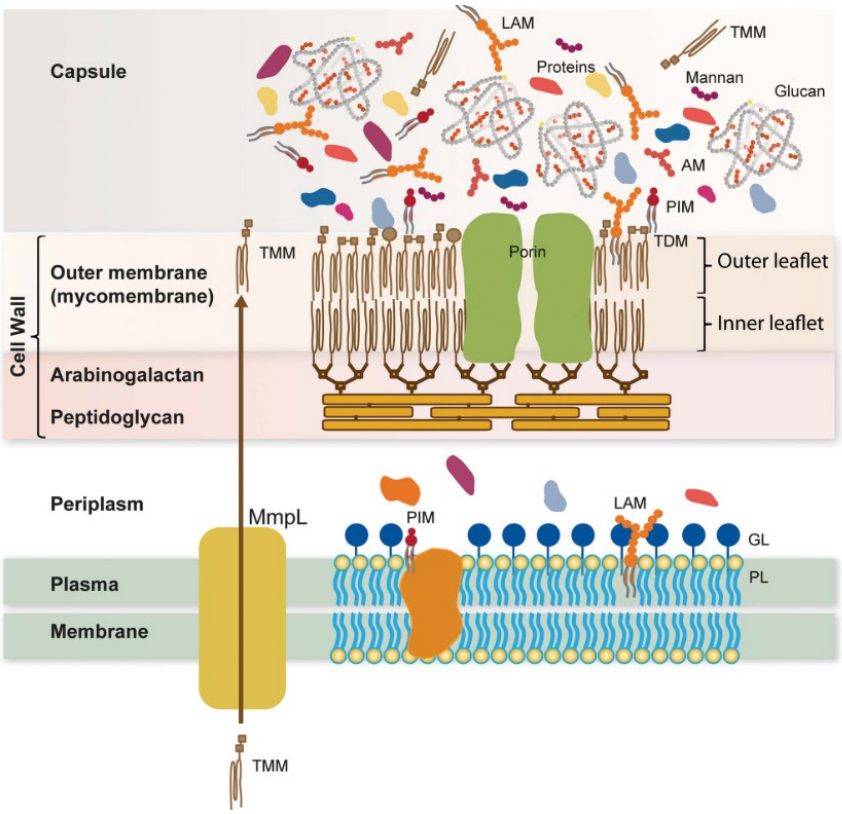


Figure 1.12 – Mycobacterial cellular envelope model. *The cellular envelope comprises a plasma membrane; the periplasmic space; the cell wall complex mAGP comprising peptidoglycan, arabinogalactan and the mycolic acid-rich outer membrane; and a top capsular layer. Adapted from M. Daffé, et al (2019) [31]*

One way to differentiate and classify bacteria is using staining techniques, whereas the Gram-staining technique differentiates bacteria that possess a certain type of cellular envelope: Gram-positive bacteria possess a cell wall with a thicker peptidoglycan layer; Gram-negative bacteria possess an outer membrane and a periplasmic space between both membranes (plasma membrane and outer membrane).

What if the bacteria possess a cellular envelope with both hallmarks of Gram-positive and Gram-negative bacteria, *i.e.* a thick peptidoglycan and an outer membrane?

That is precisely the case for mycobacteria.

Mycobacteria are classified as Gram-positive bacteria, however, the bacilli do not retain the stain as much, resulting in a Gram-neutral or Gram ghost appearance instead [32]. In fact, the mycobacterial cellular envelope is different from both Gram-positive and Gram-negative bacteria, as it is composed by a plasma membrane, a cell wall core, an outer membrane and a thick capsule (Figures 1.12 and 1.13) [33]–[36].

The cell wall core of mycobacteria is an intricate scaffold composed of mycolic acids, arabinogalactan, and peptidoglycan, also designated as mAGP complex (mycolyl-arabinogalactan-peptidoglycan) (Figure 1.14).

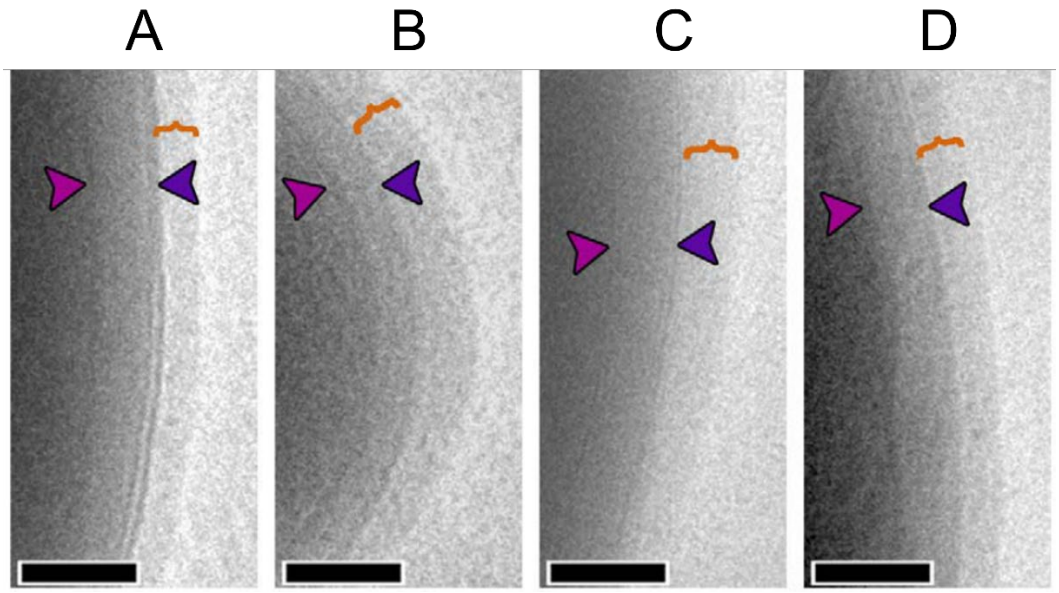


Figure 1.13 – ***Cryo electron micrographs of the mycobacterial envelope.*** *(A) M. smegmatis; (B) M. tuberculosis; (C) M. marinum; and (D) M. bovis BCG. Orange bracket indicates the capsule thickness. Arrows point to plasma membrane (magenta) and to outer membrane (purple). Scale bars: 100 nm. Adapted from M. Sani, et al (2010) [34]*

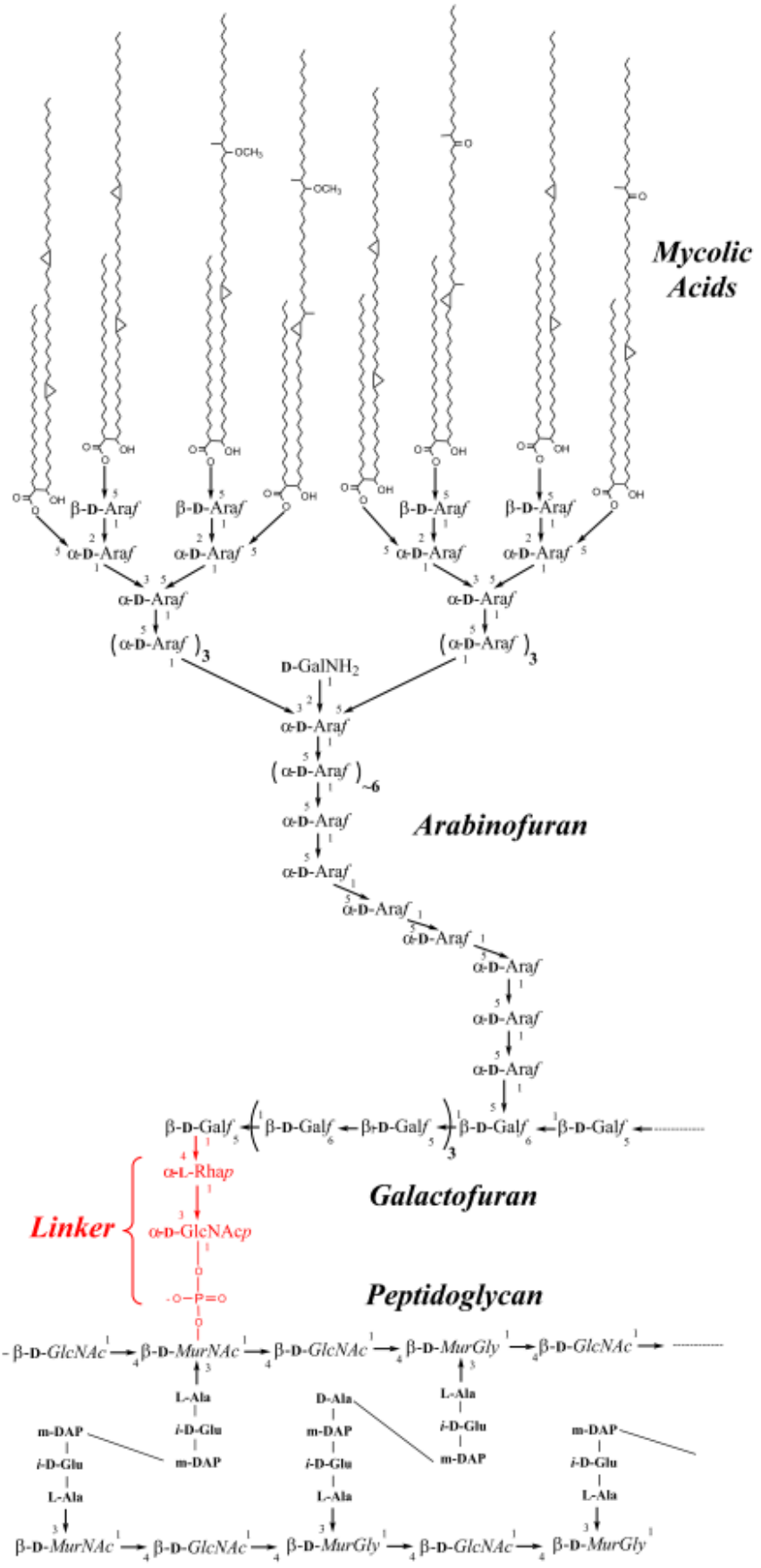


Figure 1.14* – *The cell wall core of M. tuberculosis. *Adapted from A. E. Grzegorzewicz, et al (2016) [37]*

3.2 Composition of the mycobacterial cell envelope

Plasma Membrane

The plasma membrane of mycobacteria consists in a lipid bilayer that comprises phospholipids which are ubiquitous to other plasma membranes, such as phosphatidylethanolamine (PE), phosphatidylinositol (PI) and diphosphatidylglycerol (DPG) [38][39]. In addition, mycobacterial plasma membrane contains several types of phosphatidylinositol mannosides (PIMs) which may contribute to a less fluid plasma membrane and increased resistance to antimicrobial agents' diffusion across the membrane. The low fluidity of the mycobacterial plasma membrane may be attributed to the uneven distribution of PIMs in the inner and outer leaflets of the bilayer [38][40].

Peptidoglycan

Peptidoglycan (PG) is the major component of the cell wall in almost all bacteria, as it is the component that reinforces plasma membrane to be able to withstand the osmotic pressure of the cytoplasm without bursting. The PG scaffold is composed of a repeating disaccharide of N-acetylglucosamine (NAG) and N-acetylmuramic acid (NAM), wherein each NAM unit has a peptide attached responsible for the crosslinking of different strands of the polymer, resulting in a honeycomb structure [41].

In mycobacteria, there is an increased level of crosslinking of the peptides in the PG, which can be around 70 – 80 % cross-linking compared to 30 – 50 % in *Escherichia coli* [42]. Another significant difference regarding the mycobacterial PG is the N-glycolylation of the muramic acid moieties, instead of N-acetylation, which contributes to an increased resistance to β-lactam antibiotics and to lysozyme [43].

The PG of mycobacteria also connects to the galactan core of the arabinogalactan (AG) layer of the cell wall by the addition of a rhamnosyl residue to the 3-position of NAG, which will prime the scaffold for the addition of the first two galactofuranosyl (Gal*f*) residues of the galactan, thus initiating the biosynthesis of AG [37].

Arabinogalactan

Arabinogalactan (AG) is a highly branched heteropolysaccharide that constitutes approximately 35% of the mycobacterial cell wall [44][45]. AG constitutes the middle

layer of the mAGP complex, between the inner PG layer and the top mycolic acid layer, and it can be divided into three distinct constituents: a linker unit; galactan, constituted by galactose (Gal); and arabinan, constituted by arabinose (Ara), whereas both Gal and Ara sugar residues are in the furanose (*f*) ring form, a quite uncommon form in nature [44][46].

The linker unit contains a diglycosylphosphoryl bridge comprised of α-L-Rha*p*-(1 → 3)-α-D-GlcNAc, which constitutes a connection between the linear galactan core of AG to the C6 position of N-glycolylmuramic acid (MurNGly) residues of PG [45][47]. This particular connection is unique to mycobacteria and related actinomycetes, therefore constituting a promising drug target [47].

The galactan core comprises AG in a linear chain of approximately 30 alternating 5- and 6-linked β-D-galactofuranosyl (Gal*f*) residues. Most arabinofuranosyl (Ara*f*) residues of arabinan are α-5-linked to form the backbone, with branching points at the C3 hydroxyl position, resulting in 3, 5-Ara*f*-linked residues [45][47]. Moreover, some inner 3, 5 α-D-Ara*f* residues comprise galactosamine (D-GalN) or succinyl substituents at the C2 position, which may vary in number per AG depending on the type of mycobacteria: one D-GalN residue per AG in slow-growing mycobacteria; three succinyl esters per AG in both slow and fast-growing mycobacteria [46].

At the non-reducing termini of arabinan is found the characteristic hexa-arabinoside motif [β-D-Ara*f*-(1→2)-α-D-Ara*f*]$_2$-3,5-α-D-Ara*f*-(1→5)-α-D-Ara*f*] to which mycolic acids are attached to at least two thirds of the available sites at the end of the arabinan chain [48]. The structure of AG is depicted in Figure 1.15.

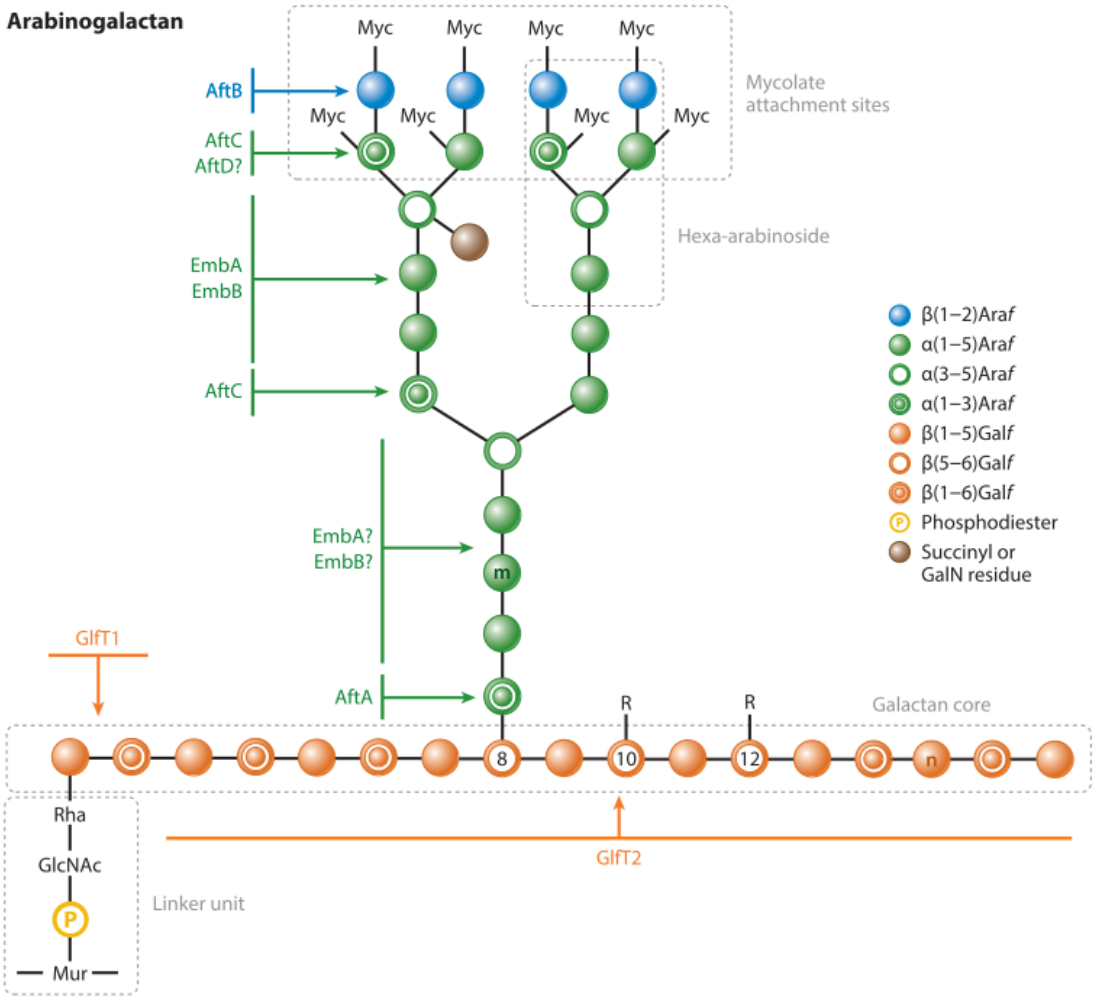


Figure 1.15 – Structure of Arabinogalactan and key enzymes involved in the biosynthesis of the galactan core and arabinan. *Adapted from M. Jankute, et al (2015) [46]*

Mycolic Acids

The last layer of the mAGP complex is composed of mycolic acids. Mycolic acids comprise about 30% of the dry weight of the cell wall [49] and are most associated to the waxy nature of the outer membrane of mycobacteria [50].

Mycolic acids consist of β-hydroxy fatty acids with long α-alkyl side chain, characterized to be very hydrophobic as the fatty acids contain 70 to 90 total carbon atoms. In *M. tuberculosis* there are three structural classes of mycolic acids: α-mycolic acids, methoxy-mycolic acids and keto-mycolic acids; wherein α-mycolic are the most common form (>70%) while methoxy- and keto-mycolic acids comprise 10 to 15% of the total mycolic acids (Figure 1.16) [51]. There are several variations in the structure of mycolic acids, namely in terms of chain length and the spacing between different functional groups: cyclopropane rings and carboxyl group [51]. These variations translate into

different distribution of mycolic acid composition across different species of mycobacteria [52] and impact the cell wall fluidity and permeability as the different conformations of α-mycolic, methoxy- and keto-mycolic acids contribute to a more or less packed lipid layer [38][52][53]. In addition, changes in mycolic acid composition also affect the ability of strains to grow inside macrophages, directly impacting the virulence of mycobacteria [54].

The different structures and conformations of mycolic acids are depicted in Figure 1.16.

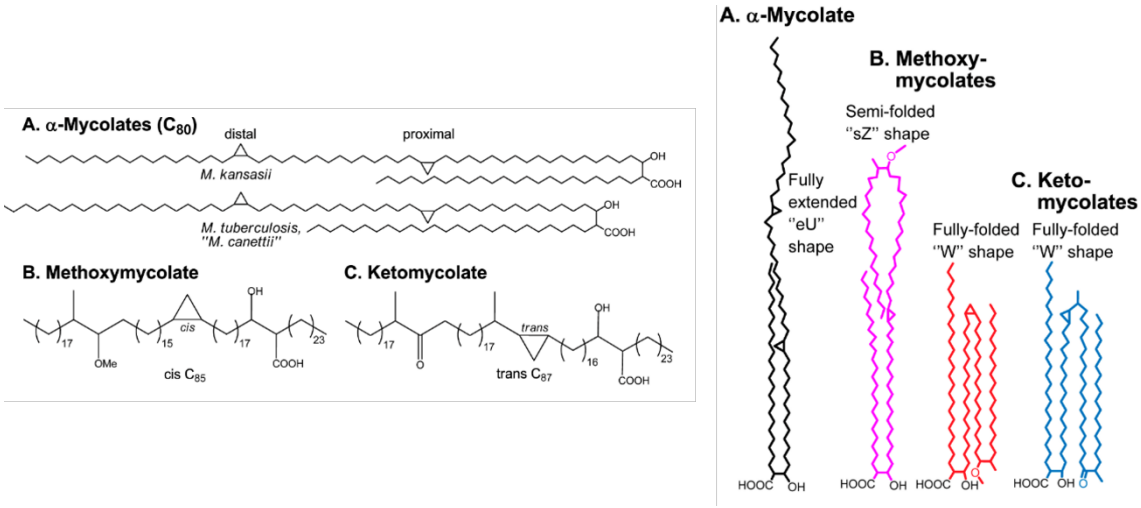


***Figure 1.16** – **Representative structure and conformations of mycolic acids.** Adapted from) D. E. Minnikin, et al (2015) [38]*

Capsule

The outermost layer of the mycobacterial cell wall is a thick capsule (Figure 1.13), which is composed of a matrix of secreted polysaccharides, proteins, and lipids [55]. The different constituents play important roles in interaction between bacteria and host cells. The main capsular polysaccharide α-glucan is involved in phagocytosis, modulation of immune response and persistence of infection, while some secreted proteins are involved in the bacterial translocation from the phagosome into the cytosol [34].

Despite the 30 nm thickness, this matrix is easily removed by mild detergents or by mechanical shear stress, as all of the components are not covalently linked to the cell wall (Figure 1.17) [34][55][56]. Sani and colleagues found that the removal of the capsule of *M. bovis* BCG significantly reduces the binding of the bacteria to macrophages, which suggests that the capsule plays a role in the binding to macrophages. In addition, the presence of the capsule induces a lower amount of pro-inflammatory cytokines from the infected macrophages [34].

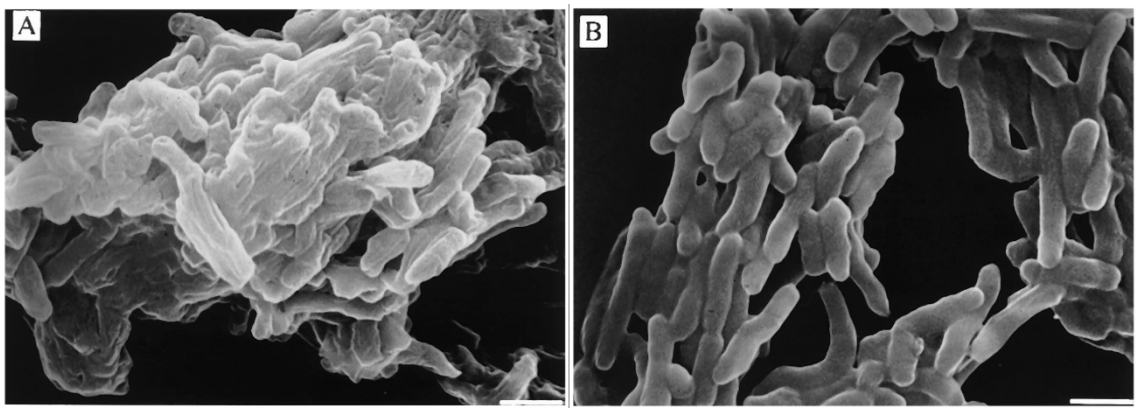


Figure 1.17 – ***Capsular surface of M. tuberculosis before (A) and after (B) treatment with Tween 80 (detergent).*** *Scale bar: 1 µm. Adapted from A. Ortalo-Magné, et al (1996) [56]*

3.3 Arabinogalactan biosynthesis – the role of Arabinofuranosyltransferases

The work presented in this thesis will focus on a set of enzymes – arabinofuranosyltransferases (AraT) – involved in the biosynthesis of arabinogalactan, specifically in the synthesis of arabinan. However, the biosynthesis of AG is a complex process comprising the following major steps: formation of the linker unit, polymerization of galactan and polymerization of arabinan.

The biosynthesis of AG begins in the cytoplasm with the formation of the disaccharide linker unit that connects peptidoglycan and arabinogalactan. A first N-acetylglucosamine phosphate (GlcNAc-P) sugar residue is transferred to decaprenyl-phosphate (C_{50}-P) to form the first glycolipid intermediate C_{50}-P-P-GlcNAc, to which is then transferred a rhamnose residue to form the C_{50}-P-P-GlcNAc-Rha [57][58]. This second glycolipid primes for the galactan chain to be attached, starting with the transfer of two Gal*f* residues by the transferase GlfT1. A second Gal*f* transferase GlfT2 then further extends the galactan chain by adding sequential Gal*f* residues with alternating β-(1→5) and β-(1→6) linkages until the ~30-residue linear chain is completed [59][60].

At this point, the galactan core is ready to receive Ara*f* residues to start the polymerization of arabinan. All Ara*f* residues are transported and delivered to the galactan core using a donor compound, decaprenyl-monophosphoryl-D-arabinose (DPA) [61]. Each Ara*f* residue donated by DPA is transferred to the growing polymer chain by different AraTs, depending on the stage of the arabinan biosynthesis.

AftA adds the first Ara*f* residues to galactan, priming it for further polymerization of arabinan (Figure 1.15) [62]. Once the first Ara*f* residues are attached to the galactan core, two other AraTs – EmbA and EmbB – further extend the arabinan chain with α-(1→5) polymerization of Ara*f* residues and are responsible for the characteristic terminal hexa-arabinofuranoside motif of AG and lipoarabinomannan (LAM) [63].

Since AG is an intricate polysaccharide, there are other AraTs responsible for the α-(1→3) branching of both AG and LAM alike. That is the case of AftC, which is responsible for the addition of branching points for both AG and LAM [64][65]. AftD is also associated with branching of AG and LAM, however the specific arabinofuranosyl transferase activity of AftD remains unclear if it catalyzes a α-(1→3) or α-(1→5) transfer [66][67].

Lastly, AftB primes AG termini for the attachment of mycolic acid, capping the terminal hexa-arabinofuranoside motif of AG with a final the β-(1→2) Ara*f* residue [68]. Besides

the arabinan polysaccharide, succinyl residues are added to the inner regions of arabinan to complete the primary structure of AG (Figure 1.15) [69].

AraTs are not exclusive to AG as they perform the same function in the biosynthesis of LAM, which are lipoglycans that are anchored to both inner and outer membranes of the cell wall. However, there is a particular AraT which is exclusively associated with the biosynthesis of LAM, EmbC, responsible for the addition of α-(1→5) Ara*f* residues to the primed lipomannan core (Figure 1.18) [70].

Considering that several AraTs are necessary for the biosynthesis of essential cell wall components, such as AG or LAM, it is fair to suggest that AraTs have high potential as drug targets. In fact, Emb proteins – EmbA, EmbB and EmbC – correspond to AraTs that are already known to be susceptible to ethambutol (EMB), a first line drug against TB that targets AG and LAM biosynthesis [71][72].

From a structural biology perspective, the three-dimensional structure of the different AraTs can provide insight into the function and mechanisms for the transferase activity of each different protein; lay the basis for novel drug development against AraTs not yet targeted by current therapeutics – AftA, AftB, AftC an AftD; and clarify inhibition and resistance mechanisms to Ethambutol for the AraTs susceptible for this drug.

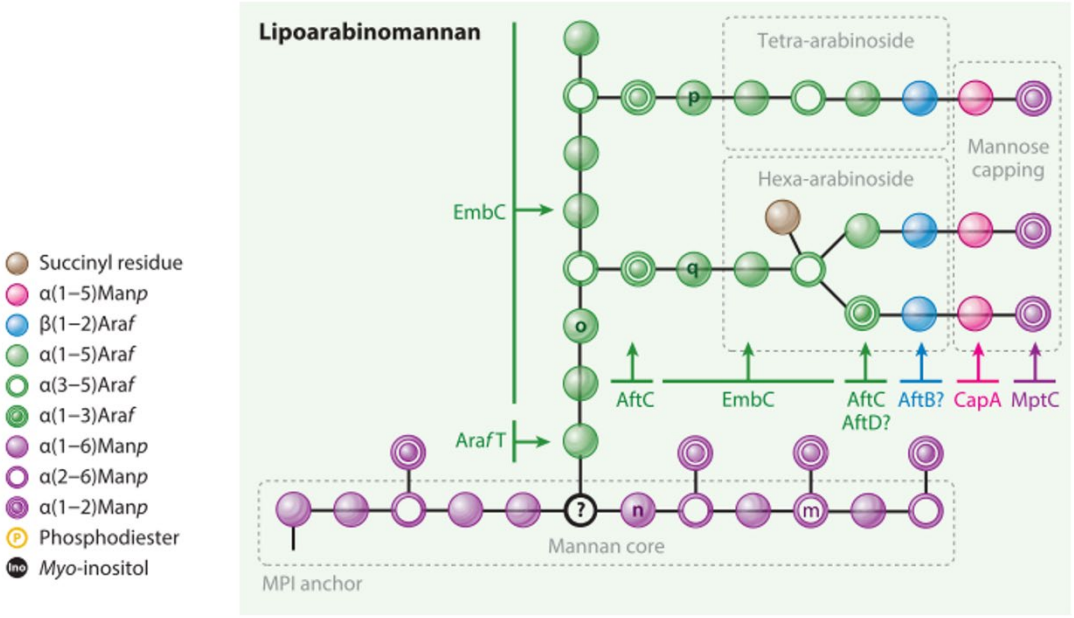


***Figure 1.18** – **Structure of Lipoarabinomannan and key enzymes involved in the biosynthesis.** Adapted from M. Jankute, et al (2015) [46]*

Objectives

The research work presented in this thesis aims to further advance the knowledge of membrane proteins of mycobacteria, in particular mycobacterial AraTs involved in the biosynthesis of arabinan, as some of these proteins are already targeted by a first-line drug against TB, others may represent new therapeutic targets to explore for developing novel drugs against TB.

The main goal of the work is to perform structural and functional characterization of mycobacterial AraTs to assess their potential as novel drug targets against TB and to gain insight on AraT activity and inhibition mechanisms.

References

[1] D. VS and D. TM, “Old Testament biblical references to tuberculosis” *Clin. Infect. Dis.*, vol. 29, no. 6, pp. 1557–1558, 1999

[2] T. M. Daniel, “The history of tuberculosis” *Respir. Med.*, vol. 100, no. 11, pp. 1862–1870, Nov. 2006

[3] V. Formicola, Q. Milanesi, and C. Scarsini, “Evidence of spinal tuberculosis at the beginning of the fourth millennium BC from Arene Candide cave (Liguria, Italy)” *Am. J. Phys. Anthropol.*, vol. 72, no. 1, pp. 1–6, Jan. 1987

[4] Hippocrates, *Of the Epidemics*., section III, 400 B.C.E.; http://classics.mit.edu/Hippocrates/epidemics.2.ii.html.

[5] T. M. Daniel and P. A. Iversen, “Hippocrates and tuberculosis” *Int. J. Tuberc. Lung Dis.*, vol. 19, no. 4, pp. 373–374, Apr. 2015

[6] Hippocrates, *Aphorisms*., Book II, section V, 400 B.C.E.; http://classics.mit.edu/Hippocrates/aphorisms.5.v.html .

[7] J. F. Murray, H. L. Rieder, and A. Finley-Croswhite, “The King’s Evil and the Royal Touch: The medical history of scrofula” *Int. J. Tuberc. Lung Dis.*, vol. 20, no. 6, pp. 713–716, Jun. 2016

[8] Y. Agarwal, R. K. Chopra, D. K. Gupta, and R. S. Sethi, “The Tuberculosis Timeline: Of White Plague, a Birthday Present, and Vignettes of Myriad Hues” 2017

[9] R. Prabhu, V. Singh, R. Prabhu, and V. Singh, “The History of Tuberculosis: Past, Present, and Future” *Adv. Microbiol.*, vol. 9, no. 11, pp. 931–942, Nov. 2019

[10] E. Armocida and M. Martini, “Tuberculosis: a timeless challenge for medicine” *J. Prev. Med. Hyg.*, vol. 61, no. 2, p. E143, Jul. 2020

[11] I. Barberis, N. L. Bragazzi, L. Galluzzo, and M. Martini, “The history of tuberculosis: from the first historical records to the isolation of Koch’s bacillus” *J. Prev. Med. Hyg.*, vol. 58, no. 1, p. E9, 2017

[12] R. Koch, *Aetiology of Tuberculosis*. New York : William R. Jenkins., 1890.

[13] World Health Organization, “World TB Day.” https://www.who.int/europe/campaigns/world-tb-day (accessed Aug. 09, 2022).

[14] A. D. Bendre, P. J. Peters, and J. Kumar, “Tuberculosis: Past, present and future of the treatment and drug discovery research.” *Curr. Res. Pharmacol. drug Discov.*, vol. 2, p. 100037, 2021

[15] M. L. Silva, B. Cá, N. S. Osório, P. N. S. Rodrigues, A. R. Maceiras, and M. Saraiva, “Tuberculosis caused by *Mycobacterium africanum*: Knowns and unknowns” *PLOS Pathog.*, vol. 18, no. 5, p. e1010490, May 2022

[16] M. Pai *et al.*, “Tuberculosis” *Nat. Rev. Dis. Prim.*, vol. 2, Oct. 2016

[17] P. Glaziou, K. Floyd, and M. C. Raviglione, “Global Epidemiology of Tuberculosis” *Semin. Respir. Crit. Care Med.*, vol. 39, no. 3, pp. 271–285, 2018

[18] W. Zhai, F. Wu, Y. Zhang, Y. Fu, and Z. Liu, “The immune escape mechanisms of *Mycobacterium Tuberculosis*” *Int. J. Mol. Sci.*, vol. 20, no. 2, Jan. 2019

[19] C. J. Cambier, S. Falkow, and L. Ramakrishnan, “Host Evasion and Exploitation Schemes of *Mycobacterium tuberculosis*” *Cell*, vol. 159, no. 7, pp. 1497–1509, Dec. 2014

[20] M. Divangahi, N. Khan, and E. Kaufmann, “Beyond Killing *Mycobacterium tuberculosis*: Disease Tolerance” *Front. Immunol.*, vol. 9, p. 2976, Dec. 2018

[21] L. Ramakrishnan, “Revisiting the role of the granuloma in tuberculosis” *Nat. Rev. Immunol. 2012 125*, vol. 12, no. 5, pp. 352–366, Apr. 2012

[22] D. Mahamed *et al.*, “Intracellular growth of *Mycobacterium tuberculosis* after macrophage cell death leads to serial killing of host cells” *Elife*, vol. 6, Jan. 2017,

[23] S. Y. Rodriguez-Takeuchi, M. E. Renjifo, and F. J. Medina, “Extrapulmonary tuberculosis: Pathophysiology and imaging findings” *Radiographics*, vol. 39, no. 7, pp. 2023–2037, Nov. 2019

[24] M. P. Golden, “Extrapulmonary Tuberculosis: An Overview - American Family Physician” vol. 72, no. 9, 2005

[25] E. W. Tiemersma, M. J. van der Werf, M. W. Borgdorff, B. G. Williams, and N. J. D. Nagelkerke, “Natural History of Tuberculosis: Duration and Fatality of Untreated Pulmonary Tuberculosis in HIV Negative Patients: A Systematic Review” *PLoS One*, vol. 6, no. 4, p. e17601, 2011

[26] World Health Organization, “Annual Report of Tuberculosis” *Annu. Glob. TB Rep. WHO*, pp. 1–68, 2022

[27] World Health Organization, “Global tuberculosis report 2023” 2023

[28] World Health Organization, “WHO announces updated definitions of extensively drug-resistant tuberculosis” Jan. 27, 2021. https://www.who.int/news/item/27-01-2021-who-announces-updated-definitions-of-extensively-drug-resistant-tuberculosis

[29] WHO consolidated guidelines on tuberculosis, “Module 4: Treatment Drug-susceptible tuberculosis treatment” p. 1 to 72, 2022

[30] World Health Organization, “WHO consolidated guidelines on tuberculosis. Module 4, Treatment : drug-resistant tuberculosis treatment.” p. 98.

[31] D. M and M. H, “Unraveling the Structure of the Mycobacterial Envelope” *Microbiol. Spectr.*, vol. 7, no. 4, Aug. 2019

[32] S. Trifiro, A. M. Bourgault, F. Lebel, and P. Rene, “Ghost mycobacteria on Gram stain” *J. Clin. Microbiol.*, vol. 28, no. 1, pp. 146–147, 1990

[33] B. Zuber, M. Chami, C. Houssin, J. Dubochet, G. Griffiths, and M. Daffé, “Direct visualization of the outer membrane of mycobacteria and corynebacteria in their native state” *J. Bacteriol.*, vol. 190, no. 16, pp. 5672–5680, Aug. 2008

[34] M. Sani, E. N. G. Houben, J. Geurtsen, J. Pierson, and K. De Punder, “Direct Visualization by Cryo-EM of the Mycobacterial Capsular Layer: A Labile Structure Containing ESX-1-Secreted Proteins” *PLoS Pathog*, vol. 6, no. 3, p. 1000794, 2010

[35] M. Daffé, “The cell envelope of tubercle bacilli” *Tuberculosis*, vol. 95, no. S1, pp. S155–S158, Jun. 2015

[36] M. Daffé and H. Marrakchi, “Unraveling the Structure of the Mycobacterial Envelope” *Microbiol. Spectr.*, vol. 7, no. 4, Jul. 2019

[37] A. E. Grzegorzewicz *et al.*, “Assembling of the *Mycobacterium tuberculosis* cell wall core” *J. Biol. Chem.*, vol. 291, no. 36, pp. 18867–18879, Sep. 2016

[38] D. E. Minnikin *et al.*, “Pathophysiological Implications of Cell Envelope Structure in *Mycobacterium tuberculosis* and Related Taxa” *Tuberc. - Expand. Knowl.*, Jul. 2015

[39] L. Chiaradia *et al.*, “Dissecting the mycobacterial cell envelope and defining the composition of the native mycomembrane” *Sci. Rep.*, vol. 7, no. 1, p. 12807, 2017

[40] R. Bansal-Mutalik and H. Nikaido, “Mycobacterial outer membrane is a lipid bilayer and the inner membrane is unusually rich in diacyl phosphatidylinositol dimannosides” *Proc Natl Acad Sci U S A*, vol. 111, no. 13, pp. 4958–4963, 2014

[41] S. O. Meroueh *et al.*, “Three-dimensional structure of the bacterial cell wall peptidoglycan” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 12, pp. 4404–4409, Mar. 2006

[42] P. Brennan and D. Crick, “The Cell-Wall Core of *Mycobacterium tuberculosis* in the Context of Drug Discovery.” *Curr. Top. Med. Chem.*, vol. 7, no. 5, pp. 475–488, Mar. 2007

[43] J. B. Raymond, S. Mahapatra, D. C. Crick, and M. S. Pavelka, “Identification of the namH gene, encoding the hydroxylase responsible for the N-glycolylation of the mycobacterial peptidoglycan” *J. Biol. Chem.*, vol. 280, no. 1, pp. 326–333, Jan. 2005

[44] M. McNeil, S. J. Wallner, S. W. Hunter, and P. J. Brennan, “Demonstration that the galactosyl and arabinosyl residues in the cell-wall arabinogalactan of *Mycobacterium leprae* and *Myobacterium tuberculosis* are furanoid” *Carbohydr. Res.*, vol. 166, no. 2, pp. 299–308, Sep. 1987

[45] M. Jankute, S. Grover, A. K. Rana, and G. S. Besra, “Arabinogalactan and lipoarabinomannan biosynthesis: structure, biogenesis and their potential as drug targets” *Futur. Microbiol*, vol. 7, no. 1, pp. 129–147, 2011

[46] M. Jankute, J. A. Cox, J. Harrison, and G. S. Besra, “Assembly of the Mycobacterial Cell Wall” *Annu Rev Microbiol*, vol. 69, pp. 405–423, 2015

[47] M. McNeil, M. Daffe, and P. J. Brennan, “Evidence for the nature of the link between the arabinogalactan and peptidoglycan of mycobacterial cell walls.” *J. Biol. Chem.*, vol. 265, no. 30, pp. 18200–18206, Oct. 1990

[48] M. McNeil, M. Daffe, and P. J. Brennan, “Location of the mycolyl ester substituents in the cell walls of mycobacteria” *J. Biol. Chem.*, vol. 266, no. 20, pp. 13217–13223, Jul. 1991

[49] S. M. Batt, D. E. Minnikin, and G. S. Besra, “The thick waxy coat of mycobacteria, a protective layer against antibiotics and the host’s immune system” *Biochem. J.*, vol. 447, no. 10, pp. 1983–2006, May 2020

[50] C. L. Dulberger, E. J. Rubin, and C. C. Boutte, “The mycobacterial cell envelope — a moving target” *Nat. Rev. Microbiol.*, vol. 18, no. 1, pp. 47–59, Jan. 2020

[51] K. Takayama, C. Wang, and G. S. Besra, “Pathway to Synthesis and Processing of Mycolic Acids in *Mycobacterium tuberculosis*” *Clin. Microbiol. Rev.*, vol. 18, no. 1, pp. 81–101, 2005

[52] M. Watanabe *et al.*, “Location of functional groups in mycobacterial meromycolate chains ; the recognition of new structural principles in mycolic acids” *Microbiology*, vol. 148, pp. 1881–1902, 2002.

[53] J. Liu, C. E. Barry, G. S. Besra, and H. Nikaido, “Mycolic acid structure determines the fluidity of the mycobacterial cell wall” *J. Biol. Chem.*, vol. 271, no. 47, pp. 29545–29551, 1996

[54] C. E. Barry, “Interpreting cell wall ‘virulence factors’ of *Mycobacterium tuberculosis*” *Trends Microbiol.*, vol. 9, no. 5, pp. 237–241, May 2001

[55] M. Daffé and G. Etienne, “The capsule of *Mycobacterium tuberculosis* and its implications for pathogenicity” *Tuber. Lung Dis.*, vol. 79, no. 3, pp. 153–169, Jun. 1999

[56] A. Ortalo-Magné *et al.*, “Identification of the Surface-Exposed Lipids on the Cell Envelopes of *Mycobacterium tuberculosis* and Other Mycobacterial Species” *J. Bacteriol.*, vol. 178, no. 2, pp. 456–461, 1996.

[57] K. Mikušová, M. Mikuš, G. S. Besra, I. Hancock, and P. J. Brennan, “Biosynthesis of the linkage region of the mycobacterial cell wall” *J. Biol. Chem.*, vol. 271, no. 13, pp. 7820–7828, Mar. 1996

[58] J. A. Mills *et al.*, “Inactivation of the mycobacterial rhamnosyltransferase, which is needed for the formation of the arabinogalactan-peptidoglycan linker, leads to irreversible loss of viability” *J. Biol. Chem.*, vol. 279, no. 42, pp. 43540–43546, Oct. 2004

[59] K. Mikušová *et al.*, “Identification of a Novel Galactosyl Transferase Involved in Biosynthesis of the Mycobacterial Cell Wall” *J. Bacteriol.*, vol. 188, no. 18, p. 6592, Sep. 2006

[60] L. J. Alderwick *et al.*, “Expression, purification and characterisation of soluble GlfT and the identification of a novel galactofuranosyltransferase Rv3782 involved in priming GlfT-mediated galactan polymerisation in *Mycobacterium tuberculosis*” *Protein Expr. Purif.*, vol. 58, no. 2, pp. 332–341, Apr. 2008

[61] B. A. Wolucka, M. R. McNeil, E. De Hoffmann, T. Chojnacki, and P. J. Brennan, “Recognition of the lipid intermediate for arabinogalactan/arabinomannan

biosynthesis and its relation to the mode of action of ethambutol on mycobacteria" *J. Biol. Chem.*, vol. 269, no. 37, pp. 23328–23335, Sep. 1994

[62] L. J. Alderwick, M. Seidel, H. Sahm, G. S. Besra, and L. Eggeling, "Identification of a novel arabinofuranosyltransferase (AftA) involved in cell wall arabinan biosynthesis in *Mycobacterium tuberculosis*" *J Biol Chem*, vol. 281, no. 23, pp. 15653–15661, 2006

[63] V. E. Escuyer *et al.*, "The role of the embA and embB gene products in the biosynthesis of the terminal hexaarabinofuranosyl motif of *Mycobacterium smegmatis* arabinogalactan" *J Biol Chem*, vol. 276, no. 52, pp. 48854–48862, 2001

[64] H. L. Birch *et al.*, "A truncated lipoglycan from mycobacteria with altered immunological properties" *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 6, pp. 2634–2639, Feb. 2010,

[65] H. L. Birch *et al.*, "Biosynthesis of mycobacterial arabinogalactan: identification of a novel α(1→3) arabinofuranosyltransferase" *Mol. Microbiol.*, vol. 69, no. 5, p. 1191, Sep. 2008

[66] H. Skovierova *et al.*, "AftD, a novel essential arabinofuranosyltransferase from mycobacteria" *Glycobiology*, vol. 19, no. 11, pp. 1235–1247, 2009

[67] L. J. Alderwick, H. L. Birch, K. Krumbach, M. Bott, L. Eggeling, and G. S. Besra, "AftD functions as an alpha1-->5 arabinofuranosyltransferase involved in the biosynthesis of the mycobacterial cell wall core" *Cell Surf*, vol. 1, pp. 2–14, 2018

[68] M. Seidel, L. J. Alderwick, H. L. Birch, H. Sahm, L. Eggeling, and G. S. Besra, "Identification of a novel arabinofuranosyltransferase AftB involved in a terminal step of cell wall arabinan biosynthesis in Corynebacterianeae, such as *Corynebacterium glutamicum* and *Mycobacterium tuberculosis*" *J Biol Chem*, vol. 282, no. 20, pp. 14729–14740, 2007

[69] S. Bhamidi *et al.*, "The Identification and Location of Succinyl Residues and the Characterization of the Interior Arabinan Region Allow for a Model of the Complete Primary Structure of *Mycobacterium tuberculosis* Mycolyl Arabinogalactan" *J. Biol. Chem.*, vol. 283, no. 19, p. 12992, May 2008

[70] L. Shi *et al.*, "The carboxy terminus of EmbC from *Mycobacterium smegmatis* mediates chain length extension of the arabinan in lipoarabinomannan" *J. Biol. Chem.*, vol. 281, no. 28, pp. 19512–19526, Jul. 2006

[71] R. Goude, A. G. Amin, D. Chatterjee, and T. Parish, “The arabinosyltransferase EmbC is inhibited by ethambutol in *Mycobacterium tuberculosis*” *Antimicrob Agents Chemother*, vol. 53, no. 10, pp. 4138–4146, 2009

[72] A. E. Belanger *et al.*, “The embAB genes of *Mycobacterium avium* encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol.” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 93, no. 21, p. 11919, Oct. 1996

CHAPTER TWO

High Throughput Expression Screening of Arabinofuranosyltransferases from Mycobacteria

This chapter contains published data:

Rodrigues, J., Almeida V. T., Rosário A. L., Tan Y. Z., Kloss B., Mancia F., Archer M., "High throughput expression screening of arabinofuranosyltransferases from mycobacteria" *Processes*, vol. 9, no. 4, p. 629, Apr. 2021, doi: 10.3390/pr9040629.

Author Contributions – J. Rodrigues

Conceptualization: M. Archer, F. Mancia, B. Kloss and *J. Rodrigues*,

Methodology: *J. Rodrigues*, V. T. Almeida, A. L. Rosário and Y. Z. Tan

Data analysis: all authors

Original draft preparation: *J. Rodrigues*, V. T. Almeida

Review and editing: all authors

1. Summary

Studies on membrane proteins can help to develop new drug targets and treatments for a variety of diseases. However, membrane proteins continue to be among the most challenging targets in structural biology. This uphill endeavor can be even harder for membrane proteins from *Mycobacterium* species, which are notoriously difficult to express in heterologous systems. Arabinofuranosyltransferases are involved in mycobacterial cell wall synthesis and thus potential targets for antituberculosis drugs. A set of 96 mycobacterial genes coding for Arabinofuranosyltransferases was selected, of which 17 were successfully expressed in *E. coli* and purified by metal-affinity chromatography. We herein present an efficient high-throughput strategy to screen in microplates a large number of targets from *Mycobacteria* and select the best conditions for large-scale protein production to pursue functional and structural studies. This methodology can be applied to other targets, is cost and time effective and can be implemented in common laboratories.

Keywords:

membrane proteins; overexpression in *E. coli*; protein purification; high-throughput protocol; Arabinofuranosyltransferases; *Mycobacteria*

2. Introduction

Membrane proteins represent 20 to 30% of open-reading frames of all genomes sequenced [1][2] and perform essential functions in cells, such as transportation, signal transduction and energy production [3]. They also play important roles in several diseases and, as a result, are attractive therapeutic targets, estimated to represent more than 30% of all marketed drugs [4]–[6]. However, biochemical and structural characterization of membrane proteins have several bottlenecks, namely toxicity by excess of mRNA levels of the target protein [7], toxicity caused by heterologous expression [8], membrane lipid composition [9][10], detergent extraction and solubility [11][12], which ultimately results in low amounts of membrane protein produced.

Many efforts have been devoted on the development of protocols to efficiently produce membrane proteins in *Escherichia coli*. An elegant approach to accelerate this process involves the fusion of green fluorescent protein (GFP) to monitor the expression and purification processes [13][14]. A commonly used strategy consists of varying different parameters simultaneously, such as expression vectors with different tags and promoters, host strains, homologues, or solubilizing detergents [15][16]. High-throughput (HTP) protein production platforms have also been developed by Structural Genomics Consortia, such as JCSG [17], Northeast SGC [18] or the New York Consortium on Membrane Protein Structure (NYCOMPS, New York, NY, USA) [19][20].

Based on the HTP strategy to express prokaryotic membrane proteins previously developed at NYCOMPS, we have devised an accessible protocol to screen mycobacterial membrane proteins, which are difficult to express in heterologous systems [21]. An initial set of 96 target genes was assembled from the genomic sequences coding for Arabinofuranosyltransferases (AraTs) from 14 different *Mycobacterium* species. Even though *Mycobacteria* proteins display low yield of production in *E. coli*, it remains one of preferred hosts for heterologous expression of *Mycobacteria* proteins and is compatible with HTP strategies.

The AraT targets are integral membrane enzymes that play a pivotal role in the synthesis of arabinan, an important component of *Mycobacteria* cell envelope [22][23]. AraTs transfer arabinose, in the furanose conformation (Ara*f*), from a single donor decaprenylphosphoryl-β-D-arabinose (DPA) to the arabinan domain of arabinogalactan (AG) of mycobacterial cell wall [24]. The cell envelope is crucial for growth and virulence of pathogenic *Mycobacteria* [25] and is a major contributor to resistance against common antibiotics [26]. AraT family comprises 7 sub-families: EmbA, EmbB and EmbC, which

are inhibited by the first-line antitubercular drug ethambutol (EMB), and AftA, AftB, AftC and AftD, which are potential novel therapeutic targets against tuberculosis (TB).

In this study, we present a simple and cost-effective methodology to screen the production of AraTs from different *Mycobacterium* species and choose the most promising targets to proceed biochemical and structural studies.

3. Materials and Methods

3.1. High-Throughput Cloning of Arabinofuranosyltransferase Genes

96 genes of AftA, AftB, AftC, AftD, EmbA, EmbB and EmbC were identified from 14 *Mycobacterium* genomes using a bioinformatics approach [19] (Table S2.1). Ligation Independent Cloning (LIC) was used to clone all selected targets, using the protocols previously described by Bruni and Kloss [27], with the following modifications: (1) target sequences were amplified by Polymerase Chain Reaction (PCR) using oligonucleotides that were compatible with LIC-adapted expression vectors (pNYCOMPS-N23 and pNYCOMPS-C23) containing appropriate overhangs with the start codon ATG, instead of the endogenous GTG start codon for some of the target genes; (2) XL10 *E. coli* was used for cloning purposes. All liquid handling was performed by hand, using multichannel pipettes. The resulting constructs were used to transform *E. coli* C41, C43 and BL21 (DE3) pLysS strains. Transformants were selected on Luria-Bertani (LB) medium [10 g/L Tryptone, 5 g/L Yeast Extract and 10 g/L NaCl] plates containing appropriate antibiotics: 100 µg/mL of ampicillin for *E. coli* C41 and C43 cells, 100 µg/mL of ampicillin and 34 µg/mL of chloramphenicol for *E. coli* BL21 (DE3) pLysS cells.

3.2. High-Throughput Expression Screening and Purification

Pre-cultures were grown overnight at 37 °C, 200 rpm in 600 µL of LB medium supplemented with the appropriate antibiotics. For this purpose, 96-well plates (VWR) were used. The overnight cultures were used to inoculate 2.5 mL of 2xYT medium [16 g/L Tryptone, 10 g/L Yeast Extract and 5 g/L NaCl], supplemented with antibiotics, in 24-well plates (UNIPLATE Collection and Analysis Microplate), at an initial optical density (OD_{600}) of 0.03 – 0.08 (see Table S2.2). Cells were grown at 37 °C, 200 rpm, until the cultures reached OD_{600} of 0.4 – 1.2 (2 to 2.5 h) (see Table S2.3), then cultures were cooled to 22 °C and gene expression was induced overnight (~16 h) with 0.25 mM isopropyl-β-D-thiogalactoside (IPTG). Cells were harvested the next day by centrifugation at 3200× *g*, for 20 min at 4 °C. OD_{600} measurements were done in TECAN Spark 10 M (Table S2.4). Cell pellets were re-suspended in 300 µL of lysis buffer (BugBuster supplemented with 0.1 mg/mL Lysozyme, 3 U/mL Benzonase, 2 mM $MgCl_2$ and 0.5 mM phenylmethylsulfonyl fluoride (PMSF)) through 10 min of vigorous shaking,

using a plate shaker, at room temperature. Detergent n-Dodecyl β-D-maltoside (DDM) was added to the lysate at 1% (w/v) final concentration and the plate containing the samples was incubated for 2 h at 4 °C, with gentle agitation. For the separation of the insoluble cell debris, plates were centrifuged at 3200× *g* for 20 min, 4 °C and 250 µL of the supernatants were transferred to a 96-well filter plate containing a bed of 50 µL Ni^{2+}-NTA agarose resin (HisPur™ Ni^{2+}-NTA Spin plate (Thermo Scientific™, Waltham, MA, USA)), previously washed with double distilled water and equilibrated with buffer (20 mM HEPES pH 7.5, 200 mM NaCl, 10 mM Imidazole, 0.1% DDM). Imidazole at 10 mM final concentration was added to each sample, to avoid unspecific binding of contaminants to the Ni^{2+}-NTA resin. Plates were incubated for 15 min in a plate-shaker at 4 °C. The plates were then centrifuged, the flowthrough fractions collected and reloaded to the resin bed, repeating the 15 min incubation with the Ni^{2+}-NTA resin bed. After the second incubation step, plates were centrifuged to remove unbound proteins. The resin was washed three times with 250 µL washing buffer (20 mM HEPES pH 7.5, 200 mM NaCl, 60 mM Imidazole and 0.1% DDM) and finally eluted with 250 µL of elution buffer (20 mM HEPES pH 7.5, 200 mM NaCl, 300 mM Imidazole and 0.05% DDM). Eluted samples were run on SDS-PAGE: 10% polyacrylamide gels were used for targets with molecular weights between 73–149 kDa (AftD, EmbA, EmbB and EmbC) and 12% polyacrylamide gels for 47–75 kDa targets (AftA, AftB and AftC).

3.3. Large Scale Protein Expression and Purification

50 mL cultures of each target in *E. coli* C41 cells were grown overnight in 250 mL flasks at 37 °C, 200 rpm, in LB medium supplemented with 100 µg/mL of ampicillin. The overnight cultures were used to inoculate 4 × 500 mL of 2xYT medium, supplemented with 100 µg/mL of ampicillin, in 2.5 L Thomson's Ultra Yield™ Flasks (Oceanside, CA, USA), at an initial OD_{600} around 0.05. Cells were grown at 37 °C, 200 rpm, until the cultures reached OD_{600} of 0.8 (2 to 2.5 h), then cultures were cooled to 22 °C and gene expression was induced overnight (~16 h) with 0.25 mM IPTG. Cells were harvested in the next day by centrifugation at 4472× *g*, for 15 min at 4 °C. OD_{600} measurements were done in Ultrospec 10 Cell Density Meter. Cell pellets were re-suspended and homogenized in lysis buffer (20 mM HEPES pH 7.5, 200 mM NaCl, 20 mM $MgSO_4$, 1 mM tris(2-carboxyethyl)phosphine (TCEP)), protease inhibitor EDTA-free cocktail (Thermo Scientific™, Waltham, USA; Catalog number: 88266) and 25 U/mL of Benzonase nuclease (Santa Cruz Biotechnology, Dallas, USA; Catalog number: sc-

391121). Cell suspension was passed twice at 15,000 psi on a cell disruptor (Constant Systems Ltd., Daventry, UK). Membranes were collected by ultracentrifugation at 197,215× *g*, for 30 min at 4 °C. Membranes were manually homogenized using a Wheaton® glass homogenizer (DWK Life Sciences Limited, Stoke-on-Trent, UK) in 20 mM HEPES pH 7.5 and 200 mM NaCl, to which DDM was added to a final concentration of 1% (w/v). Membranes were solubilized for 2 h, with gentle agitation, at 4 °C. Soluble membrane fraction was collected by ultracentrifugation at 203,756× *g* for 30 min, at 4 °C. The supernatants were collected and incubated with 2 mL of equilibrated Ni^{2+}-NTA agarose resin for 1.5 h at 4 °C, with gentle agitation. Imidazole was added to each sample to a final concentration of 10 mM, to prevent unspecific binding of contaminants. After incubation, the sample was loaded into a column for elution by gravity flow. The resin bed was washed with 10 column volumes (CV) with washing buffer (20 mM HEPES pH 7.5, 200 mM NaCl, 0.1% DDM, 60 mM Imidazole), and the proteins were eluted with 4 CV of elution buffer (20 mM HEPES pH 7.5, 200 mM NaCl, 0.05% DDM and 300 mM Imidazole). Eluted samples were concentrated and injected into a Superdex 200 column (Cytiva Europe GmbH, Freiburg, Germany) to assess protein dispersity. The collected fractions were run on SDS-PAGE: 10% polyacrylamide for 73–149 kDa targets (AftD, EmbA, EmbB and EmbC) and 12% polyacrylamide for those around 47–75 kDa (AftA, AftB and AftC).

4. Results

4.1. Genomic Expansion and High-Throughput Cloning of Arabinofuranosyltransferases

A set of 96 target genes was assembled from the genomic sequences coding for seven AraTs from *Mtb* (AftA, AftB, AftC, AftD, EmbA, EmbB and EmbC). To each "seed" sequence, a "cluster" of homolog sequences, from 14 different *Mycobacterium* genomes was expanded, coding for proteins likely to have similar structure as the seed protein [28] (Table S2.1). LIC was performed as described previously by Bruni and Kloss [27]. Briefly, all sequences were amplified by PCR, using genomic DNA available from ATCC® (Manassas, VA, USA) [https://www.lgcstandards-atcc.org (accessed on 26 March 2021)] and primer pairs compatible with LIC-adapted expression vectors (pNYCOMPS-N23 and pNYCOMPS-C23) that contained decahistidine affinity tag and Tobacco Etch Virus (TEV) protease cleavage site (ENLYFQS). 56 targets were successfully cloned into pNYCOMPS-N23 and 40 targets into pNYCOMPS-C23. Previous screening experiments had shown no expression for all constructs in pNYCOMPS-C23 vector (data not shown), therefore only the clones in pNYCOMPS-N23 were used for the HTP expression screening approach.

4.2. High-Throughput Expression of Arabinofuranosyltransferases

All 56 positive clones in pNYCOMPS-N23 were transformed into C41, C43 and BL21 (DE3) pLysS *E. coli* strains. 24 deep-well plates were used to grow the positive clones simultaneously. Growth conditions, 2xYT rich medium, 0.25 mM IPTG and overnight post-induction at 22 °C, were established based on the results obtained in previous experiments. This allowed a fast and reliable comparison between different *E. coli* strains, also leaving room for optimization after target selection.

Cell harvesting by centrifugation and lysis were performed in 24 deep-well plates, maintaining a HTP downstream processing of the samples. Extraction of membrane proteins was achieved by adding detergent directly to each well, after cell lysis, incubating the plate at low temperature. The 24 deep-well plate is centrifuged again to clear the solubilized lysate from the cell debris. The solubilized lysate was transferred to a HisPurTM Ni^{2+}-NTA Spin 96 well plate (Thermo Scientific™, Waltham, MA, USA) for

affinity chromatography purification. In this step, the use of adjustable multichannel pipettes to transfer solutions from 24-well plate to 96-well plate was important for sake of speed and reproducibility/reliability, however, standard multichannel pipettes can also be used although not in an optimal manner.

After a single Ni^{2+}-NTA purification step, the amount of eluted target protein was too low to be detected by SDS-PAGE. Considering that the sample solution slowly flows from the filter plate by gravity during the incubation period, a second passage was deemed necessary to increase the contact time between the sample and Ni^{2+}-NTA resin, after which the eluted AraTs could be visualized on the gel. The full pipeline is summarized in Figure 2.1.

In total, 17 out of 96 distinct proteins were produced and purified, resulting in 18% success rate of protein production (Table 2.1). All three *E. coli* host strains were able to produce target proteins: 16 in C41, 6 in C43 and 8 in BL21 (DE3) pLysS (Table 2.2). AftB and EmbC proteins were not detected in any *E. coli* strain using this HTP method, suggesting that different, perhaps more tailored conditions may be needed to successfully produce these proteins.

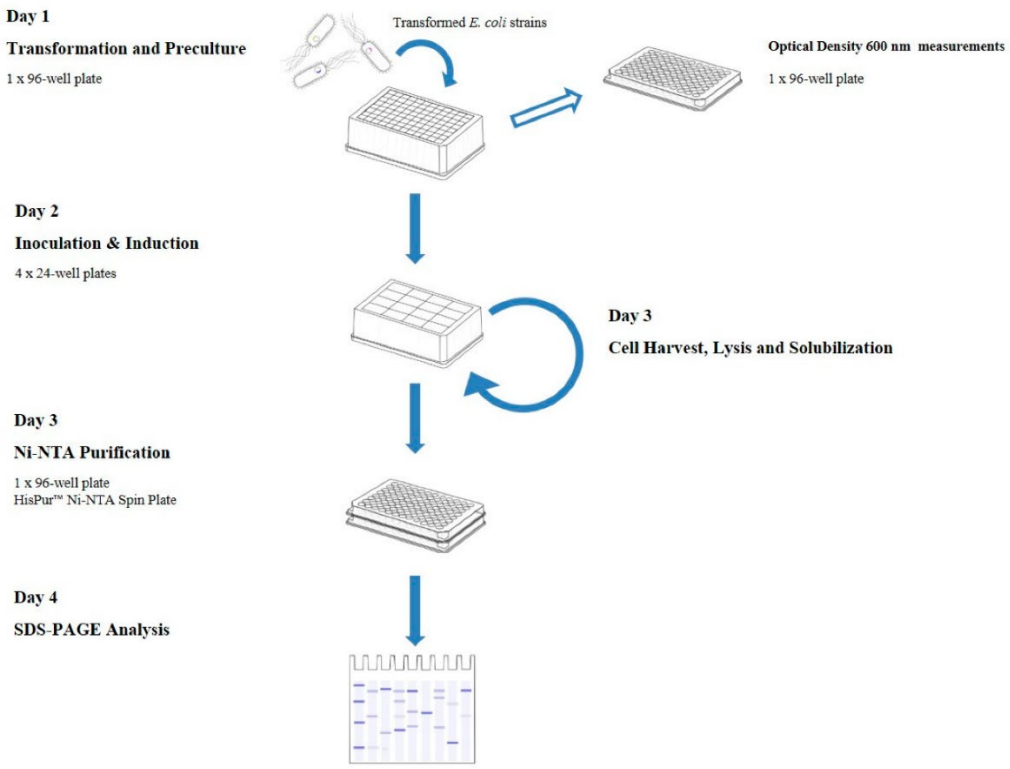


Figure 2.1 – Schematic representation of the high-throughput screening for membrane protein production and purification.

Table 2.1 – Summary of high-throughput screening of membrane protein production.

	Number	Success Rate (%)
Targets	96	-
Positive Clones (pNYCOMPS-N23)	56	58
[1] Positive Clones (pNYCOMPS-C23)	40	42
Proteins purified (total)	17	18
E. coli C41::pNYCOMPS-N23	16	17
E. coli C43::pNYCOMPS-N23	6	6
E. coli BL21 (DE3) pLysS::pNYCOMPS-N23	8	8

[1] Expression of these constructs was not determined in this study.

Table 2.2 – Protein targets used in the high-throughput screening and expression results for each E. coli strain tested.

ID	GenBank ID	Organism	Predicted Protein	C41	C43	BL21 (DE3) pLysS
A1	SIU02450.1	*Mycobacterium bovis* AF2122/97	AftA	−	−	−
A8	EUA63955.1	*Mycobacterium abscessus* 1948	AftA	−	−	+
C2	ABP43658.1	*Mycobacterium gilvum* PYR-GCK	AftA	+	−	−
C9	AFC41461.1	*Mycobacterium intracellulare* ATCC 13950	AftA	−	−	−
F1	ABM16394.1	*Mycobacterium vanbaalenii* PYR-1	AftA	+	−	+
F3	AGP61782.1	*Mycobacterium yongonense* 05-1390	AftA	−	−	−
F7	CDQ43571.1	*Mycobacterium neoaurum*	AftA	+	−	+
F8	AAS02550.1	*Mycobacterium avium subsp. Paratuberculosis* K-10	AftA	−	−	−
A2	SIU02464.1	*Mycobacterium bovis* AF2122/97	AftB	−	−	−
A9	EUA63936.1	*Mycobacterium abscessus* 1948	AftB	−	−	−
C7	CCP46634.1	*Mycobacterium tuberculosis* H37Rv	AftB	−	−	−
D9	AGP61763.1	*Mycobacterium yongonense* 05-1390	AftB	−	−	−
D10	AFC41442.1	*Mycobacterium intracellulare* ATCC 13950	AftB	−	−	−
D11	ABM16411.1	*Mycobacterium vanbaalenii* PYR-1	AftB	−	−	−
E2	AIR19061.1 *	*Mycobacterium kansasii* 662	AftB	−	−	−
E5	CDQ43590.1	*Mycobacterium neoaurum*	AftB	−	−	−
A3	AMC65006.1	*Mycobacterium bovis* AF2122/97	AftC	+	−	+
A10	EUA61591.1	*Mycobacterium abscessus* 1948	AftC	+	−	−
C11	AFC44620.1	*Mycobacterium intracellulare* ATCC 13950	AftC	+	+	−
C12	AGZ53302.1	*Mycobacterium kansasii* ATCC 12478	AftC	+	+	−
D1	ABP46386.1	*Mycobacterium gilvum* PYR-GCK	AftC	+	+	+
D2	CDQ43952.1	*Mycobacterium neoaurum*	AftC	+	+	+
E6	EHB50241.1	*Mycobacterium rhodesiae* JS60	AftC	−	−	−

ID	GenBank ID	Organism	Predicted Protein	C41	C43	BL21 (DE3) pLysS
E7	AFM16967.1	*Mycobacterium chubuense* NBB4	AftC	−	−	−
E8	AAS05110.1	*Mycobacterium avium subsp. Paratuberculosis* K-10	AftC	+	+	−
E9	AGP64972.1	*Mycobacterium yongonense* 05-1390	AftC	+	+	−
F9	ABM13300.1	*Mycobacterium vanbaalenii* PYR-1	AftC	−	−	−
A4	CAB5247947.1	*Mycobacterium bovis* AF2122/97	AftD	−	−	−
D4	AGZ51741.1	*Mycobacterium kansasii* ATCC 12478	AftD	−	−	−
H11	EUA63217.1	*Mycobacterium abscessus* 1948 F5/8	AftD	+	−	+
A5	SIU02452.1	*Mycobacterium bovis* AF2122/97	EmbA	−	−	−
A11	EUA63951.1	*Mycobacterium abscessus* 1948	EmbA	+	−	−
B4	AAC45280.1	*Mycobacterium tuberculosis* H37Rv	EmbA	−	−	−
B8	ACC43760.1	*Mycobacterium marinum* M	EmbA	+	−	−
C5	ABP43656.1	*Mycobacterium gilvum* PYR-GCK	EmbA	−	−	−
G5	AAS02546.1	*Mycobacterium avium subsp. Paratuberculosis* K-10	EmbA	−	−	−
G11	CDQ43576.1	*Mycobacterium neoaurum*	EmbA	−	−	−
H1	ABM16396.1	*Mycobacterium vanbaalenii* PYR-1	EmbA	−	−	−
H2	AGZ51276.1	*Mycobacterium kansasii* ATCC 12478	EmbA	−	−	−
A6	SIU02453.1	*Mycobacterium bovis* AF2122/97	EmbB	−	−	−
A12	EUA63949.1	*Mycobacterium abscessus* 1948	EmbB	+	−	−
B3	AAC45281.1	*Mycobacterium tuberculosis* H37Rv	EmbB	−	−	−
B7	ACC43761.1	*Mycobacterium marinum* M	EmbB	−	−	−
C4	ABP43655.1	*Mycobacterium gilvum* PYR-GCK	EmbB	−	−	−
G3	AFC41455.1	*Mycobacterium intracellulare* ATCC 13950	EmbB	−	−	−
G6	AAS02545.1	*Mycobacterium avium subsp. Paratuberculosis* K-10	EmbB	−	−	−
G9	AGP61776.1	*Mycobacterium yongonense* 05-1390	EmbB	−	−	−
H7	KEP38884.1	*Mycobacterium kansasii*	EmbB	−	−	−

ID	GenBank ID	Organism	Predicted Protein	C41	C43	BL21 (DE3) pLysS
H9	ABM16397.1	*Mycobacterium vanbaalenii* PYR-1	EmbB	+	−	+
B1	EUA63954.1	*Mycobacterium abscessus* 1948	EmbC	−	−	−
C3	ABP43657.1	*Mycobacterium gilvum* PYR-GCK	EmbC	−	−	−
G4	AFC41460.1	*Mycobacterium intracellulare* ATCC 13950	EmbC	−	−	−
G7	AAS02549.1	*Mycobacterium avium subsp. Paratuberculosis* K-10	EmbC	−	−	−
G10	AGP61781.1	*Mycobacterium yongonense* 05-1390	EmbC	−	−	−
H4	CDQ43572.1	*Mycobacterium neoaurum*	EmbC	−	−	−
H5	AGZ51274.1	*Mycobacterium kansasii* ATCC 12478	EmbC	−	−	−

* Record removed.

We found that a single His-tag purification step was not sufficient, considering that persistent contaminants from the host cell were present across all targets (Figure 2.2; see Figure S2.1 and S2.2). Moreover, the production yields for the target proteins herein studied were low in all *E. coli* host strains.

Nevertheless, we were able to successfully identify bands in the SDS-PAGE that could correspond to the target proteins, based on their predicted molecular weight (MW) and considering the gel shifting for membrane proteins in denaturing protein gels [29]. Due to this anomalous migration pattern, bands related to membrane proteins in SDS-PAGE most often appear ~20 – 30% below their predicted MW.

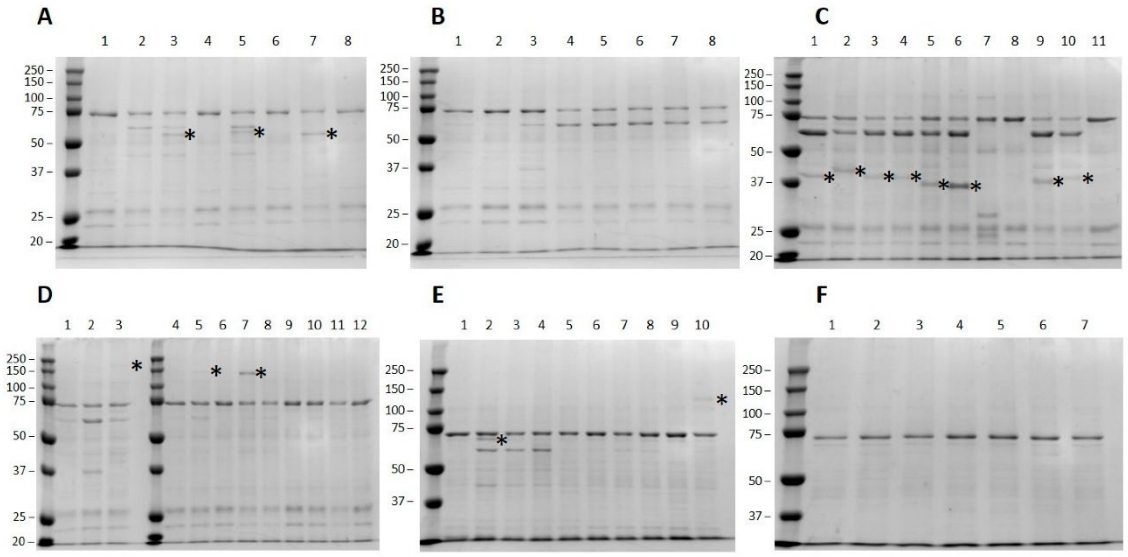


Figure 2.2 – Protein expression results from the HTP screening of 56 AraTs from Mycobacteria overexpressed in E. coli C41 cells. ***(A)*** *AftA proteins: 1 – A1; 2 – A8; 3 – C2; 4 – C9; 5 – F1; 6 – F3; 7 – F7; 8 – F8.* ***(B)*** *AftB proteins 1 – A2; 2 – A9; 3 – C7; 4 – D9; 5 – D10; 6 – D11; 7 – E2; 8 – E5.* ***(C)*** *AftC proteins: 1 – A3; 2 – A10; 3 – C11; 4 – C12; 5 – D1; 6 – D2; 7 – E6; 8 – E7; 9 – E8; 10 – E9; 11 – F9.* ***(D)*** *AftD and EmbA proteins: 1 – A4; 2 – D4; 3 – H11; 4 – A5; 5 – A11; 6 – B4; 7 – B8; 8 – C5; 9 – G5; 10 – G11; 11 – H1; 12 – H2.* ***(E)*** *EmbB proteins: 1 – A6; 2 – A12; 3 – B3; 4 – B7; 5 – C4; 6 – G3; 7 – G6; 8 – G9; 9 – H7; 10 – H9.* ***(F)*** *EmbC proteins: 1 – B1; 2 – C3; 3 – G4; 4 – G7; 5 – G10; 6 – H4; 7 – H5. Asterisks (⁎) indicate protein bands corresponding to the predicted MW of the corresponding target.*

4.3. Validation of HTP Target Selection by Large-Scale Protein Production

Based on SDS-PAGE analysis (Figure 2.2), we selected one target from each cluster for large-scale production: AftA and AftC from *M. neoaurum*, AftD from *M. abscessus* 1948 F5/8, EmbA from *M. marinum* M. and EmbB from *M. vanbaalenii* PYR-1, all produced in *E. coli* C41. Growth conditions were similar to the ones used in the HTP screening, although cell lysis and membrane extraction steps were modified according to the cell mass. Most importantly, the incubation time of solubilized membranes with Ni^{2+}-NTA resin was increased to improve protein binding and purification yield.

All chosen targets were successfully produced in large scale, thus validating the selection made from the HTP screening. Size exclusion chromatography (SEC) was performed after affinity chromatography to further purify the protein and as tool for preliminary biophysical characterization of each protein (Figure 2.3). Although all targets show a SEC profile with a peak right after the void volume of the column, suggesting protein aggregation in the presence of DDM, it was still possible to identify heterogeneous protein populations in most target samples. Upon SDS-PAGE analysis of the SEC elution fractions, we observed that the dominant protein bands correspond to the desired targets, however there were still contaminants present. AftA (Figure 2.3A), EmbA (Figure 2.3D) and EmbB (Figure 2.3E) showed the least amount of contaminant proteins.

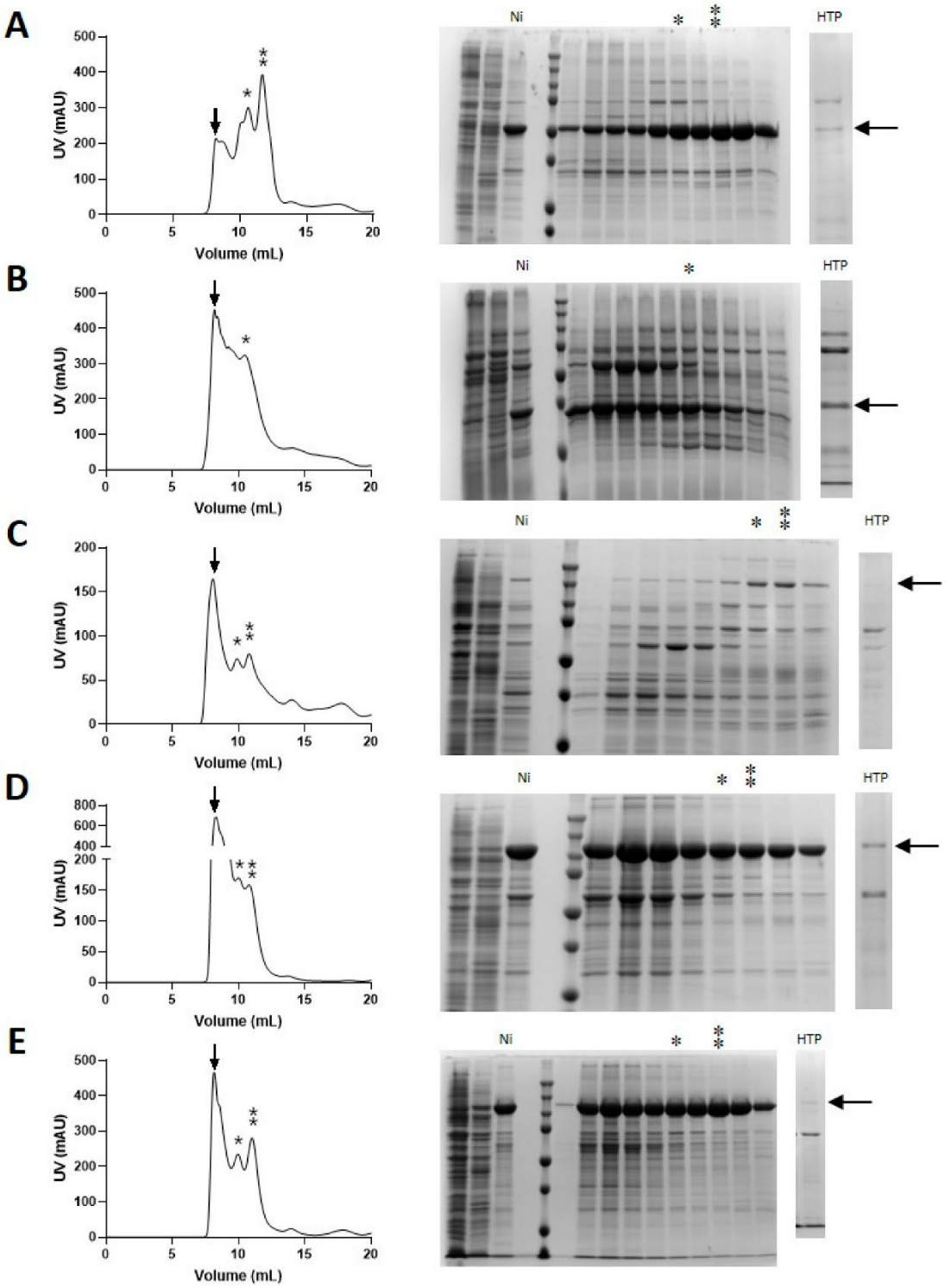


Figure 2.3 – Size-exclusion chromatography (SEC) elution profiles of large-scale expression experiments of AraTs and respective SDS-PAGE. *Vertical arrows indicate void volume; asterisks (*) and (⁂) indicate different populations observed in the SEC elution profiles. Ni: Ni^{2+}-NTA elution samples; HTP: Small-scale HTP results; horizontal arrows indicate the bands corresponding to the target protein of interest.* ***(A)*** *AftA from M. neoaurum,* ***(B)*** *AftC from M. neoaurum,* ***(C)*** *AftD from M. abscessus 1948 F5/8,* ***(D)*** *EmbA from M. marinum* M. *and* ***(E)*** *EmbB from M. vanbaalenii PYR-1.*

For EmbB, a second SEC step was performed (Figure 2.4), since the SEC profile showed two distinct and separate peaks, corresponding to different EmbB populations. In the second SEC step each EmbB population was run separately. We could observe that the high molecular weight EmbB population behaves as a stable monodisperse population (Figure 2.4A, 2.4B), while the low molecular weight EmbB population splits into the same two populations observed in the first SEC run (Figure 2.3E), suggesting that EmbB monomers are prone to form an equilibrium with stable EmbB dimers.

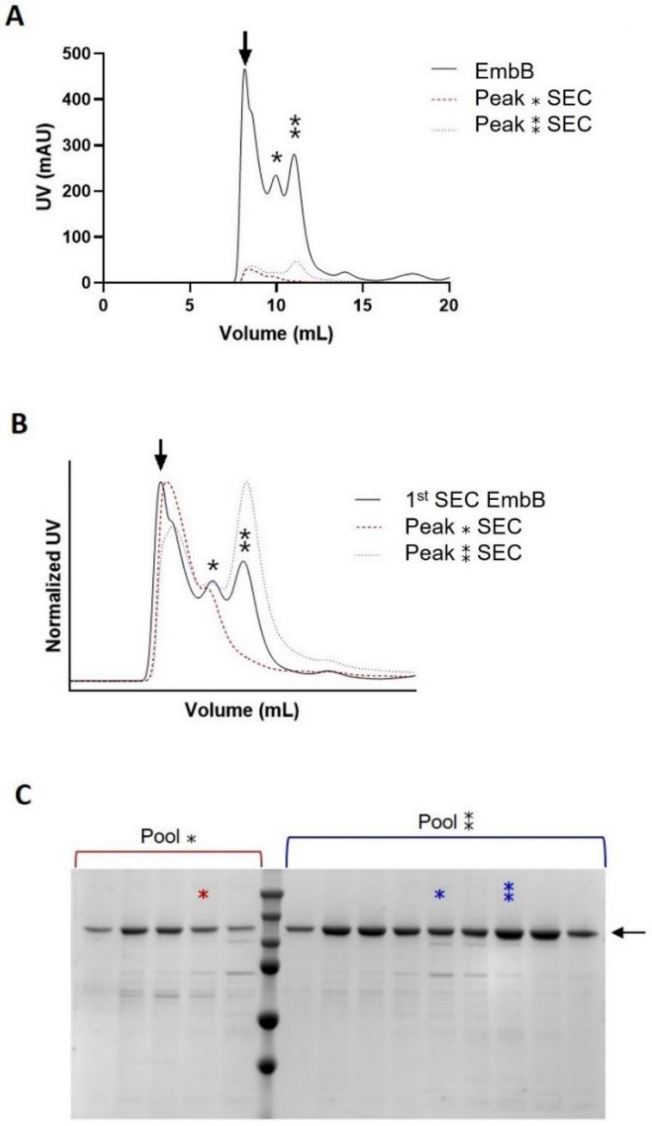


Figure 2.4 – Size-exclusion chromatography elution profiles of EmbB from M. vanbaalenii PYR-1. (A) *SEC elution profiles of EmbB: after Ni^{2+}-NTA (full black line), peak (⁎) after 1st SEC (red dashed lines), peak (⁑) after 1st SEC (spotted blue line).* ***(B)*** *Normalized SEC elution profiles of EmbB: after Ni^{2+}-NTA (full black line), peak (⁎) after 1st SEC (red dashed lines), peak (⁑) after 1st SEC (spotted blue line).* ***(C)*** *SDS-PAGE analysis of peak (⁎) and (⁑) SEC. Vertical arrows indicate void volume. Colored asterisks correspond to the peaks observed in each SEC run, respective to different EmbB populations. Arrow indicates the corresponding band of EmbB (115 kDa).*

5. Discussion

The need to screen the expression of a large number of membrane protein targets, as well as the selection of optimal conditions for production and purification of desired targets, led to the development of several HTP strategies. The strategy used in this study is not novel and was intended to setup a protocol to search for the best candidates to pursue functional and structural studies on AraTs from *Mycobacteria.* Embs are targets of ethambutol, whereas Afts are potential targets to develop new drugs to treat TB. Nevertheless, the protocol herein described can also be applied to evaluate the expression and purification of other membrane proteins.

The methodology involved the selection of 13–14 orthologue genes of each AraT sub-family (EmbA-C, AftA-D) from a variety of host genomes, gene expression with vectors harboring a poly-histidine affinity tag at either N- or C-terminus, transformation into three different *E. coli* strains, membrane extraction and protein solubilization by DDM detergent, and purification by Ni^{2+}-NTA chromatography. By using this simple combinatorial approach, we were able to clone 56 genes at pNYCOMPS-N23 and 40 at pNYCOMPS-C23, and produce 17 proteins out of 96 chosen targets, corresponding to a success rate of 18%. Such rate is not surprising, considering that membrane proteins are often difficult to express and purify [7][30]. Heterologous expression of mycobacterial proteins in *E. coli* has previously been reported not to exceed 40% [30]–[32].

No expression of AraTs cloned into pNYCOMPS-C23 was observed (data not shown). It is well known that type and location of the fused affinity tag has a significant effect at all stages of protein production [33], however it is not possible to know *a priori* the impact caused by tag addition. This unpredictability is somehow the foundation of HTP approaches – try as many conditions as is reasonably possible and assess what works to proceed with further studies.

Concerning the host organism, *M. smegmatis* could be a viable alternative for the heterologous expression of *Mycobacteria* proteins [31], yet we considered it not appropriate for a HTP approach due to its slower growth rate compared to *E. coli*, and mostly due to its waxy surface [34], which promotes clumping, film formation and cell adhesion to surfaces, especially plastic, preventing an optimal use of 96- or 24-well plates for cell growth. Instead, we used *E. coli* C41 and C43 suited for overexpression of toxic and membrane proteins [35], and BL21 (DE3) pLysS for controlled expression [36]. The different expression levels observed among *E. coli* hosts suggests that the type of strain plays a pivotal role in the number of well-expressed AraTs, as also reported for

other target proteins [37]. Indeed, regulation of T7 RNA polymerase expression either by mutations in its promoter (C41 and C43 Walker strains) or by its natural inhibitor T7 lysozyme (T7Lys, pLysS strain) can significantly influence membrane protein overexpression yields [38].

Large scale production (from 2 L of culture) of 5 targets, chosen based on the results of small-scale experiments, yielded purified proteins in milligram amounts, although differences are observed on the intensity of their respective bands (SDS-PAGE) in small and large-scale experiments. The shorter incubation time with the Ni^{2+}-NTA agarose in the small-scale screening (2 × 15 min vs. 1.5 h) could account for this discrepancy. Moreover, the aeration rate related with the size and shape of the growth vessel (24-well plate vs. 2.5 L flasks) may also affect the overexpression levels.

In addition, switching to a cobalt spin plate, instead of nickel, may increase binding specificity of the target protein and thus further improve the results for the small-scale screening. We expect similar results will be obtained on the scale-up production of other targets that showed expression on the HTP screening. Fusions with GFP tag could be advantageous to monitor the various steps of protein production by measuring fluorescence, a very sensitive detection method [13]. However, this methodology is not suitable for membrane proteins with periplasmic C-terminus [13][20][39], which is the case for most of the AraTs herein studied, so it was not considered.

Detergents are required to extract and purify target proteins and their choice is a key parameter on the entire process. We chose to use DDM since it is a mild detergent and one of the most commonly used for this purpose [20][27][33][40]. The aggregation detected in the large-scale production experiments suggests further detergent screens may be needed to select the best detergent formula for each individual target.

We were able to separate two different populations of EmbB from *M. vanbaalenii* PYR-1, likely constituted by monomer and dimer, respectively. Despite the aggregation, all targets showed soluble populations in the SEC elution profiles, which represents a good starting point for optimization towards structural studies. Interestingly, 3D structures of both oligomeric states have been already characterized by single particle cryogenic-electron microscopy (cryo-EM) for EmbB from *M. smegmatis* [41][42]. Noteworthy, different detergents or solubilizing agents may be needed for structural studies. The cryo-EM structures of several AraTs, namely *M. tuberculosis* EmbA–EmbB complex, *M. tuberculosis* EmbB [43], *M. smegmatis* EmbB [42] and *M. abscessus* AftD [44], have been recently characterized and different solubilization agents were used, namely glyco-

diosgenin (GDN) detergent, amphipols or nanodiscs. Interestingly, the structure of EmbC solubilized in DDM has been determined by X-ray crystallography [43].

On one hand, the production of AftB and EmbC targets was not achieved using the HTP workflow with “standard” conditions. Therefore, other parameters must be explored, such as growth media, temperature, incubation time and type of detergent, host strains or expression vectors, which will likely lead to better success rates. On the other hand, AftA and AftC from *M. neoaurum* were expressed and are attractive targets for drug development [45]–[47].

The presented protocol stands as a simplified approach based in previous HTP strategies developed at NYCOMPS [19][20] and by others [15], to identify the best candidates for further biochemical and structural studies in a fast and affordable manner for most laboratories. Most importantly, this methodology delivers similar results to conventional medium throughput approaches and, by reducing the variables throughout, it allows target optimization for large-scale protein production.

6. References

[1] E. Wallin and G. von Heijne, “Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms” *Protein Sci*, vol. 7, no. 4, pp. 1029–1038, 1998

[2] A. Krogh, B. Larsson, G. Von Heijne, and E. L. L. Sonnhammer, “Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes” *J. Mol. Biol.*, vol. 305, no. 3, pp. 567–580, Jan. 2001

[3] G. von Heijne, “The membrane protein universe: what’s out there and why bother?” *J. Intern. Med.*, vol. 261, no. 6, pp. 543–557, Jun. 2007

[4] E. A. Yasi, N. S. Kruyer, and P. Peralta-Yahya, “Advances in G protein-coupled receptor high-throughput screening” *Current Opinion in Biotechnology*, vol. 64. Elsevier Ltd, pp. 210–217, Aug. 01, 2020

[5] Y. Arinaminpathy, E. Khurana, D. M. Engelman, and M. B. Gerstein, “Computational analysis of membrane proteins: the largest class of drug targets” *Drug Discovery Today*, vol. 14, no. 23–24. Elsevier Current Trends, pp. 1130–1135, Dec. 01, 2009

[6] J. Davey, “G-protein-coupled receptors: New approaches to maximise the impact of GPCRs in drug discovery” *Expert Opin. Ther. Targets*, vol. 8, no. 2, pp. 165–170, Apr. 2004

[7] S. Schlegel *et al.*, “Optimizing membrane protein overexpression in the *Escherichia coli* strain Lemo21(DE3)” *J. Mol. Biol.*, vol. 423, no. 4, pp. 648–659, Nov. 2012

[8] F. Gubellini *et al.*, “Physiological response to membrane protein overexpression in *E. coli*” *Mol. Cell. Proteomics*, vol. 10, no. 10, Oct. 2011

[9] C. Hunte, “Specific protein-lipid interactions in membrane proteins” *Biochem. Soc. Trans.*, vol. 33, no. 5, pp. 938–942, Nov. 2005

[10] A. G. Lee, “How lipids affect the activities of integral membrane proteins” *Biochimica et Biophysica Acta - Biomembranes*, vol. 1666, no. 1–2. Elsevier B.V., pp. 62–87, Nov. 03, 2004

[11] D. Hardy, E. Desuzinges Mandon, A. J. Rothnie, and A. Jawhari, “The yin and yang of solubilization and stabilization for wild-type and full-length membrane protein” *Methods*, vol. 147, pp. 118–125, 2018

[12] S. M. Smith, “Strategies for the Purification of Membrane Proteins” *Methods Mol Biol*, vol. 1485, pp. 389–400, 2017

[13] D. Drew, M. Lerch, E. Kunji, D. J. Slotboom, and J. W. de Gier, “Optimization of membrane protein overexpression and purification using GFP fusions” *Nat. Methods*, vol. 3, no. 4, pp. 303–313, Apr. 2006

[14] L. E. Bird *et al.*, “Green fluorescent protein-based expression screening of membrane proteins in *Escherichia coli*” *J. Vis. Exp.*, no. 95, p. 52357, Jan. 2015

[15] S. Eshaghi, “An efficient strategy for high-throughput expression screening of recombinant integral membrane proteins” *Protein Sci.*, vol. 14, no. 3, pp. 676–683, Mar. 2005

[16] P. Ma *et al.*, “An efficient strategy for small-scale screening and production of archaeal membrane transport proteins in *Escherichia coli*” *PLoS One*, vol. 8, no. 10, p. e76913, 2013

[17] M. A. Elsliger *et al.*, “The JCSG high-throughput structural biology pipeline” *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, vol. 66, no. 10, pp. 1137–1142, Oct. 2010

[18] R. Xiao *et al.*, “The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium” *J. Struct. Biol.*, vol. 172, no. 1, pp. 21–33, Oct. 2010

[19] J. Love *et al.*, “The New York Consortium on Membrane Protein Structure (NYCOMPS): a high-throughput platform for structural genomics of integral membrane proteins” *J. Struct. Funct. Genomics*, vol. 11, no. 3, pp. 191–199, 2010

[20] F. Mancia and J. Love, “High-throughput expression and purification of membrane proteins” *J. Struct. Biol.*, vol. 172, no. 1, pp. 85–93, Oct. 2010

[21] M. Bellinzoni and G. Riccardi, “Techniques and Applications: The heterologous expression of *Mycobacterium tuberculosis* genes is an uphill road” *Trends Microbiol.*, vol. 11, no. 8, pp. 351–358, 2003

[22] M. Jankute, S. Grover, A. K. Rana, and G. S. Besra, “Arabinogalactan and lipoarabinomannan biosynthesis: structure, biogenesis and their potential as drug targets” *Futur. Microbiol*, vol. 7, no. 1, pp. 129–147, 2011

[23] F. E. Umesiri, A. K. Sanki, J. Boucau, D. R. Ronning, and S. J. Sucheck, “Recent advances toward the inhibition of mAG and LAM synthesis in *Mycobacterium tuberculosis*” *Med Res Rev*, vol. 30, no. 2, pp. 290–326, 2010

[24] B. A. Wolucka, M. R. McNeil, E. De Hoffmann, T. Chojnacki, and P. J. Brennan,

"Recognition of the lipid intermediate for arabinogalactan/arabinomannan biosynthesis and its relation to the mode of action of ethambutol on mycobacteria" *J. Biol. Chem.*, vol. 269, no. 37, pp. 23328–23335, Sep. 1994

[25] M. Jankute, J. A. Cox, J. Harrison, and G. S. Besra, "Assembly of the Mycobacterial Cell Wall" *Annu Rev Microbiol*, vol. 69, pp. 405–423, 2015

[26] K. A. Abrahams and G. S. Besra, "Mycobacterial cell wall biosynthesis: a multifaceted antibiotic target" *Parasitology*, vol. 145, no. 2, pp. 116–133, Feb. 2018

[27] R. Bruni and B. Kloss, "High-Throughput Cloning and Expression of Integral Membrane Proteins in *Escherichia coli*" *Curr. Protoc. Protein Sci.*, vol. 74, no. 1, pp. 29.6.1-29.6.34, Nov. 2013

[28] M. Punta *et al.*, "Structural genomics target selection for the New York consortium on membrane protein structure" *J. Struct. Funct. Genomics*, vol. 10, no. 4, pp. 255–268, Dec. 2009

[29] A. Rath, M. Glibowicka, V. G. Nadeau, G. Chen, and C. M. Deber, "Detergent binding explains anomalous SDS-PAGE migration of membrane proteins" *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 6, pp. 1760–1765, Feb. 2009

[30] N. Moreland *et al.*, "A flexible and economical medium-throughput strategy for protein production and crystallization" *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 61, no. Pt 10, pp. 1378–1385, Oct. 2005

[31] G. Bashiri and E. N. Baker, "Production of recombinant proteins in *Mycobacterium smegmatis* for structural and functional studies" *Protein Sci*, vol. 24, no. 1, pp. 1–10, 2015

[32] R. M. Goldstone, N. J. Moreland, G. Bashiri, E. N. Baker, and J. Shaun Lott, "A new Gateway® vector and expression protocol for fast and efficient recombinant protein expression in *Mycobacterium smegmatis*" *Protein Expr. Purif.*, vol. 57, no. 1, pp. 81–87, 2008

[33] O. Lewinson, A. T. Lee, and D. C. Rees, "The Funnel Approach to the Precrystallization Production of Membrane Proteins" *J. Mol. Biol.*, vol. 377, no. 1, pp. 62–73, Mar. 2008

[34] K. J. Kieser and E. J. Rubin, "How sisters grow apart: mycobacterial growth and division" *Nat Rev Microbiol*, vol. 12, no. 8, pp. 550–562, 2014

[35] B. Miroux and J. E. Walker, "Over-production of proteins in *Escherichia coli*: Mutant hosts that allow synthesis of some membrane proteins and globular

proteins at high levels" *Journal of Molecular Biology*, vol. 260, no. 3. Academic Press, pp. 289–298, Jul. 19, 1996

[36] F. W. Studier, "Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system" *J. Mol. Biol.*, vol. 219, no. 1, pp. 37–44, May 1991

[37] K. Mathieu *et al.*, "Functionality of membrane proteins overexpressed and purified from *E. coli* is highly dependent upon the strain" *Sci. Rep.*, vol. 9, no. 1, pp. 1–15, Dec. 2019

[38] S. Wagner *et al.*, "Tuning *Escherichia coli* for membrane protein overexpression" *Proc Natl Acad Sci U S A*, vol. 105, no. 38, pp. 14371–14376, 2008

[39] J. Fan *et al.*, "An efficient strategy for high throughput screening of recombinant integral membrane protein expression and stability" *Protein Expr. Purif.*, vol. 78, no. 1, pp. 6–13, Jul. 2011

[40] M. S. Willis and C. M. Koth, "Structural Proteomics of Membrane Proteins: a Survey of Published Techniques and Design of a Rational High Throughput Strategy" in *Methods in molecular biology (Clifton, N.J.)*, vol. 426, 2008, pp. 277–295.

[41] L. Zhang *et al.*, "Cryo-EM snapshots of mycobacterial arabinosyltransferase complex EmbB(2)-AcpM(2)" *Protein Cell*, vol. 11, no. 7, pp. 505–517, 2020

[42] Y. Z. Tan *et al.*, "Cryo-EM structure of arabinosyltransferase EmbB from *Mycobacterium smegmatis*" *Nat. Commun.*, vol. 11, no. 1, p. 3396, 2020

[43] L. Zhang and Y. Zhao, "Structures of cell wall arabinosyltransferases with the anti-tuberculosis drug ethambutol" *Science (80-.).*, vol. 368, no. 6496, pp. 1211–1219, 2020

[44] Y. Z. Tan *et al.*, "Cryo-EM Structures and Regulation of Arabinofuranosyltransferase AftD from Mycobacteria" *Mol. Cell*, vol. 78, no. 4, pp. 683-699.e11, 2020

[45] L. J. Alderwick, M. Seidel, H. Sahm, G. S. Besra, and L. Eggeling, "Identification of a novel arabinofuranosyltransferase (AftA) involved in cell wall arabinan biosynthesis in *Mycobacterium tuberculosis*" *J. Biol. Chem.*, vol. 281, no. 23, pp. 15653–15661, Jun. 2006

[46] J. Zhang *et al.*, "Reconstitution of functional mycobacterial arabinosyltransferase AftC proteoliposome and assessment of decaprenylphosphorylarabinose analogues as arabinofuranosyl donors" *ACS Chem Biol*, vol. 6, no. 8, pp. 819–828, 2011

[47] L. Favrot and D. R. Ronning, “Targeting the mycobacterial envelope for tuberculosis drug development” *Expert Rev Anti Infect Ther*, vol. 10, no. 9, pp. 1023–1036, 2012

7. Acknowledgments

Author Contributions

Conceptualization: M. Archer, F. Mancia, B. Kloss and *J. Rodrigues*,

Methodology: *J. Rodrigues*, V. T. Almeida, A. L. Rosário and Y. Z. Tan

Data analysis: all authors

Original draft preparation: *J. Rodrigues*, V. T. Almeida

Review and editing: all authors

Supervision and funding acquisition: M. Archer and F. Mancia.

We thank Alessio Bortoluzzi, Catarina Paiva and Sandra Santos for advice and help with equipment for high throughput protein expression.

Funding

This research was funded by Fundação para a Ciência e Tecnologia (FCT), Lisbon, Portugal; grants (PTDC/BIA-BQM/30421/2017 and PTDC/BIA-BQM/4056/2020 to M.A. and PD/BD/128261/2016 to J.R.), European Union's Horizon 2020 research and innovation programme under grant agreements: No. 857203 (Twinning), No. 823780 (MSCA-RISE) and No. 731005, Instruct-ULTRA, a project to further develop the services of Instruct-ERIC (M.A, J.R. and V.T.A.). This work was also supported by National Institutes of Health (NIH), Bethesda, MD, USA; grants GM132120 (to F.M.) and GM116799 (to Wayne A. Hendrickson).

8. Supplemental Information

Table S2.1 – List of genes coding for AraTs from several Mycobacterium species selected for cloning and expression screening.

ID	GenBank ID	Organism	Predicted Protein	% Identity (*M. tuberculosis*)	Predicted Molecular Weight (kDa)	Predicted Transmembrane Helices
A1	SIU02450.1	*Mycobacterium bovis* AF2122/97	AftA	100	70	13
A2	SIU02464.1	*Mycobacterium bovis* AF2122/97	AftB	99	69	9
A3	AMC65006.1	*Mycobacterium bovis* AF2122/97	AftC	99	49	8
A4	CAB5247947.1	*Mycobacterium bovis* AF2122/97	AftD	99	146	9
A5	SIU02452.1	*Mycobacterium bovis* AF2122/97	EmbA	99	116	13
A6	SIU02453.1	*Mycobacterium bovis* AF2122/97	EmbB	99	118	13
A7	SIU02451.1	*Mycobacterium bovis* AF2122/97	EmbC	99	118	13
A8	EUA63955.1	*Mycobacterium abscessus* 1948	AftA	65	68	13
A9	EUA63936.1	*Mycobacterium abscessus* 1948	AftB	67	71	10
A10	EUA61591.1	*Mycobacterium abscessus* 1948	AftC	64	47	8
A11	EUA63951.1	*Mycobacterium abscessus* 1948	EmbA	65	114	12
A12	EUA63949.1	*Mycobacterium abscessus* 1948	EmbB	68	73	8
B1	EUA63954.1	*Mycobacterium abscessus* 1948	EmbC	68	117	11
B2	AAC45279.1	*Mycobacterium tuberculosis* H37Rv	EmbC	100	117	13
B3	AAC45281.1	*Mycobacterium tuberculosis* H37Rv	EmbB	100	118	12
B4	AAC45280.1	*Mycobacterium tuberculosis* H37Rv	EmbA	100	116	13
B5	CCP42964.1	*Mycobacterium tuberculosis* H37Rv	AftD	100	146	9
B6	ACC43759.1	*Mycobacterium marinum* M	EmbC	86	117	14
B7	ACC43761.1	*Mycobacterium marinum* M	EmbB	89	116	12
B8	ACC43760.1	*Mycobacterium marinum* M	EmbA	87	118	13
B9	AFM19671.1	*Mycobacterium chubuense* NBB4	EmbB	72	115	13

ID	GenBank ID	Organism	Predicted Protein	% Identity (*M. tuberculosis*)	Predicted Molecular Weight (kDa)	Predicted Transmembrane Helices
B10	AFM19669.1	*Mycobacterium chubuense* NBB4	EmbC	75	116	12
B11	AFM19670.1	*Mycobacterium chubuense* NBB4	EmbA	69	115	13
B12	AFM19668.1	*Mycobacterium chubuense* NBB4	AftA	67	71	10
C1	ACC43758.1	*Mycobacterium marinum* M	AftA	83	70	13
C2	ABP43658.1	*Mycobacterium gilvum* PYR-GCK	AftA	68	67	13
C3	ABP43657.1	*Mycobacterium gilvum* PYR-GCK	EmbC	74	115	14
C4	ABP43655.1	*Mycobacterium gilvum* PYR-GCK	EmbB	70	115	13
C5	ABP43656.1	*Mycobacterium gilvum* PYR-GCK	EmbA	69	114	12
C6	AIU11367.1	*Mycobacterium smegmatis* str. MC2 155	AftA	68	67	12
C7	CCP46634.1	*Mycobacterium tuberculosis* H37Rv	AftB	100	69	9
C8	AAS02532.1	*Mycobacterium avium* subsp. *Paratuberculosis*	AftB	82	70	9
C9	AFC41461.1	*Mycobacterium intracellulare* ATCC 13950	AftA	77	68	13
C10	ABK72123.1	*Mycobacterium smegmatis* str. MC2 155	AftC	70	49	8
C11	AFC44620.1	*Mycobacterium intracellulare* ATCC 13950	AftC	84	50	8
C12	AGZ53302.1	*Mycobacterium kansasii* ATCC 12478	AftC	89	49	8
D1	ABP46386.1	*Mycobacterium gilvum* PYR-GCK	AftC	70	48	9
D2	CDQ43952.1	*Mycobacterium neoaurum*	AftC	70	48	8
D3	ABK71542.1	*Mycobacterium smegmatis* str. MC2 155	AftD	71	148	12
D4	AGZ51741.1	*Mycobacterium kansasii* ATCC 12478	AftD	82	148	13
D5	ACC38960.1	*Mycobacterium marinum*	AftD	80	146	7
D6	AFM15049.1	*Mycobacterium chubuense* NBB4	AftD	70	145	12

ID	GenBank ID	Organism	Predicted Protein	% Identity (*M. tuberculosis*)	Predicted Molecular Weight (kDa)	Predicted Transmembrane Helices
D7	ADT97050.1	*Mycobacterium gilvum* Spyr1	AftD	70	148	13
D8	ABP43645.1	*Mycobacterium gilvum* PYR-GCK	AftB	71	70	9
D9	AGP61763.1	*Mycobacterium yongonense* 05-1390	AftB	82	72	10
D10	AFC41442.1	*Mycobacterium intracellulare* ATCC 13950	AftB	82	72	10
D11	ABM16411.1	*Mycobacterium vanbaalenii* PYR-1	AftB	70	72	9
D12	EHB54870.1	*Mycobacterium rhodesiae* JS60	AftB	68	73	10
E1	CDM79377.1	*Mycobacterium marinum* E11	AftB	85	73	10
E2	AIR19061.1 *	*Mycobacterium kansasii* 662	AftB	86	73	10
E3	ABK75671.1	*Mycobacterium smegmatis* str. MC2 155	AftB	73	70	9
E4	AFM19681.1	*Mycobacterium chubuense* NBB4	AftB	68	72	11
E5	CDQ43590.1	*Mycobacterium neoaurum*	AftB	67	67	11
E6	EHB50241.1	*Mycobacterium rhodesiae* JS60	AftC	71	49	7
E7	AFM16967.1	*Mycobacterium chubuense* NBB4	AftC	71	49	9
E8	AAS05110.1	*Mycobacterium avium* subsp. *Paratuberculosis* K-10	AftC	83	49	8
E9	AGP64972.1	*Mycobacterium yongonense* 05-1390	AftC	83	50	8
E10	ACC40492.1	*Mycobacterium marinum*	AftC	87	50	8
E11	AAS06236.1	*Mycobacterium avium* subsp. *Paratuberculosis* K-10	AftD	79	145	9
E12	EHB55421.1	*Mycobacterium rhodesiae* JS60	AftD	71	149	13
F1	ABM16394.1	*Mycobacterium vanbaalenii* PYR-1	AftA	67	68	13
F2	CDQ42439.1	*Mycobacterium neoaurum* F5/8	AftD	67	148	13
F3	AGP61782.1	*Mycobacterium yongonense* 05-1390	AftA	77	68	13

ID	GenBank ID	Organism	Predicted Protein	% Identity (*M. tuberculosis*)	Predicted Molecular Weight (kDa)	Predicted Transmembrane Helices
F4	EHB54850.1	*Mycobacterium rhodesiae* JS60	AftA	71	66	13
F5	AGZ51273.1	*Mycobacterium kansasii* ATCC 12478	AftA	84	70	13
F6	AGP66385.1	*Mycobacterium yongonense* 05-1390	AftD	80	146	13
F7	CDQ43571.1	*Mycobacterium neoaurum*	AftA	67	67	11
F8	AAS02550.1	*Mycobacterium avium* subsp. *Paratuberculosis* K-10	AftA	79	75	11
F9	ABM13300.1	*Mycobacterium vanbaalenii* PYR-1	AftC	69	48	9
F10	ABM11102.1	*Mycobacterium vanbaalenii* PYR-1	AftD	70	147	10
F11	AFP42646.1	*Mycobacterium smegmatis* str. MC2 155	EmbA	69	117	13
F12	ABK72840.1	*Mycobacterium smegmatis* str. MC2 155	EmbB	69	117	13
G1	ABK72375.1	*Mycobacterium smegmatis* str. MC2 155	EmbC	75	115	10
G2	AFC41456.1	*Mycobacterium intracellulare* ATCC 13950	EmbA	83	115	14
G3	AFC41455.1	*Mycobacterium intracellulare* ATCC 13950	EmbB	85	115	12
G4	AFC41460.1	*Mycobacterium intracellulare* ATCC 13950	EmbC	85	114	13
G5	AAS02546.1	*Mycobacterium avium* subsp. *Paratuberculosis* K-10	EmbA	83	117	14
G6	AAS02545.1	*Mycobacterium avium* subsp. *Paratuberculosis* K-10	EmbB	84	115	12
G7	AAS02549.1	*Mycobacterium avium* subsp. *Paratuberculosis* K-10	EmbC	85	117	13
G8	AGP61777.1	*Mycobacterium yongonense* 05-1390	EmbA	83	116	14
G9	AGP61776.1	*Mycobacterium yongonense* 05-1390	EmbB	85	115	12

ID	GenBank ID	Organism	Predicted Protein	% Identity (*M. tuberculosis*)	Predicted Molecular Weight (kDa)	Predicted Transmembrane Helices
G10	AGP61781.1	*Mycobacterium yongonense* 05-1390	EmbC	85	114	13
G11	CDQ43576.1	*Mycobacterium neoaurum*	EmbA	69	116	13
G12	EHB54852.1	*Mycobacterium rhodesiae* JS60	EmbA	72	116	13
H1	ABM16396.1	*Mycobacterium vanbaalenii* PYR-1	EmbA	70	116	13
H2	AGZ51276.1	*Mycobacterium kansasii* ATCC 12478	EmbA	88	115	13
H3	AEV72559.1	*Mycobacterium rhodesiae* NBB3	EmbC	76	114	14
H4	CDQ43572.1	*Mycobacterium neoaurum*	EmbC	72	115	13
H5	AGZ51274.1	*Mycobacterium kansasii* ATCC 12478	EmbC	88	117	14
H6	ABM16395.1	*Mycobacterium vanbaalenii* PYR-1	EmbC	73	116	10
H7	KEP38884.1	*Mycobacterium kansasii*	EmbB	90	117	11
H8	AEV72557.1	*Mycobacterium rhodesiae* NBB3	EmbB	76	114	13
H9	ABM16397.1	*Mycobacterium vanbaalenii* PYR-1	EmbB	70	115	13
H10	AHC23140.2	*Mycobacterium neoaurum* VKM Ac-1815D	EmbB	69	115	13
H11	EUA63217.1	*Mycobacterium abscessus* 1948 F5/8	AftD	62	149	12
H12	AFC46048.1	*Mycobacterium intracellulare* ATCC 13950	AftD	79	146	13

* Record removed.

Table S2.2 – Optical Density at 600 nm (OD_{600}) of small-scale cultures. In-plate target coordinates.

AftA	A1	A8	C2	C9	F1	F3	F7	F8			
AftB	A2	A9	C7	D9	D10	D11	E2	E5			
AftC	A3	A10	C11	C12	D1	D2	E6	E7	E8	E9	F9
AftD	A4	D4	H11								
EmbA	A5	A11	B4	B8	C5	G5	G11	H1	H2		
EmbB	A6	A12	B3	B7	C4	G3	G6	G9	H7	H9	
EmbC	B1	C3	G4	G7	G10	H4	H5				

Table S2.3 – OD_{600} upon induction of expression (2xYT medium). *Color scheme: green color – OD_{600} interval (0.5 – 1.0), red color – OD_{600} below 0.5 or above 1.0. The black contoured squares show the values obtained for the targets selected for large scale.*

***E. coli* strain**		**OD_{600} upon Induction**										
BL21 (DE3) pLysS	AftA	0.92	0.99	1.04	0.85	1.03	1.05	0.80	1.10			
	AftB	0.82	0.87	1.00	0.75	0.78	0.82	0.58	0.65			
	AftC	1.02	0.64	0.72	0.80	1.05	0.84	0.88	1.07	0.86	0.75	1.23
	AftD	0.93	1.00	0.61								
	EmbA	0.87	0.72	0.79	0.74	0.93	0.93	1.16	1.11	0.89		
	EmbB	1.06	0.91	0.89	0.88	0.97	0.82	0.56	0.90	0.30	0.70	
	EmbC	1.03	0.47	1.12	0.96	1.06	0.46	0.40				
C41	AftA	1.19	1.00	0.77	0.57	0.68	0.63	0.76	0.60			
	AftB	0.43	0.46	0.78	0.75	0.81	0.79	0.72	0.60			
	AftC	1.01	0.55	0.87	0.77	1.14	0.84	0.85	1.06	0.66	0.59	0.86
	AftD	0.55	0.75	0.62								
	EmbA	0.51	0.72	0.64	0.74	0.51	0.52	0.62	0.53	0.58		
	EmbB	0.58	0.56	0.37	0.67	0.40	0.43	0.50	0.53	0.74	0.68	
	EmbC	0.76	0.97	0.70	0.63	0.81	0.63	0.40				
C43	AftA	0.73	0.80	0.88	0.71	0.84	0.81	0.92	0.81			
	AftB	0.81	0.75	0.52	0.83	0.78	0.85	0.79	0.76			
	AftC	0.79	0.76	0.75	0.72	0.82	1.04	0.83	0.87	0.80	0.78	0.83
	AftD	0.66	0.52	0.74								
	EmbA	0.71	0.65	0.62	0.70	0.67	0.61	0.57	0.60	0.56		
	EmbB	0.74	0.66	0.68	0.70	0.48	0.67	0.68	0.71	0.60	0.82	
	EmbC	0.88	0.78	0.72	0.61	0.48	0.63	0.65				

Table S2.4 – Variation of OD_{600} between the induction time and cell harvesting. *A graduated scale of three colors was applied to these values, indicating average values in white, the lower values in red and higher values in green. The black contoured squares show the values obtained for the targets selected for large scale.*

***E. coli* strain**		**OD_{600} Variation**										
BL21(DE3) pLysS	AftA	9.36	6.43	8.44	7.87	3.82	4.13	1.47	8.71			
	AftB	7.94	8.33	8.26	6.76	5.51	7.11	7.80	5.70			
	AftC	8.94	6.73	7.39	1.55	7.17	6.97	6.96	6.82	7.20	7.44	8.25
	AftD	5.15	6.75	5.44								
	EmbA	8.54	6.51	7.85	8.08	7.78	8.30	7.71	7.60	7.58		
	EmbB	6.92	8.33	5.19	5.35	8.20	7.70	7.18	8.37	5.23	1.58	
	EmbC	6.76	7.65	6.41	7.38	7.13	8.14	7.94				
C41	AftA	9.14	8.19	6.40	9.16	8.97	9.07	7.70	8.41			
	AftB	7.43	9.65	10.67	14.87	7.48	9.07	6.12	6.45			
	AftC	9.15	8.89	10.06	9.60	8.45	10.17	7.93	7.38	9.27	9.06	7.52
	AftD	5.62	7.52	5.03								
	EmbA	7.09	6.39	6.57	5.75	8.32	9.60	6.29	3.40	6.45		
	EmbB	6.67	7.67	4.15	5.01	5.28	5.41	4.00	5.22	7.18	6.98	
	EmbC	6.56	8.90	7.96	8.69	9.38	9.45	7.72				
C43	AftA	6.97	7.06	6.96	6.82	6.66	6.76	7.70	7.25			
	AftB	7.31	6.85	12.09	8.64	7.66	7.27	7.22	7.09			
	AftC	7.24	9.82	7.95	11.34	7.62	12.73	7.83	8.02	8.80	7.64	7.38
	AftD	6.61	10.16	7.21								
	EmbA	6.70	10.83	7.38	7.06	7.56	7.82	6.82	7.30	6.60		
	EmbB	7.06	8.08	8.86	7.99	7.43	7.32	7.43	7.62	6.90	8.50	
	EmbC	7.21	7.12	6.87	6.89	6.68	7.46	7.01				

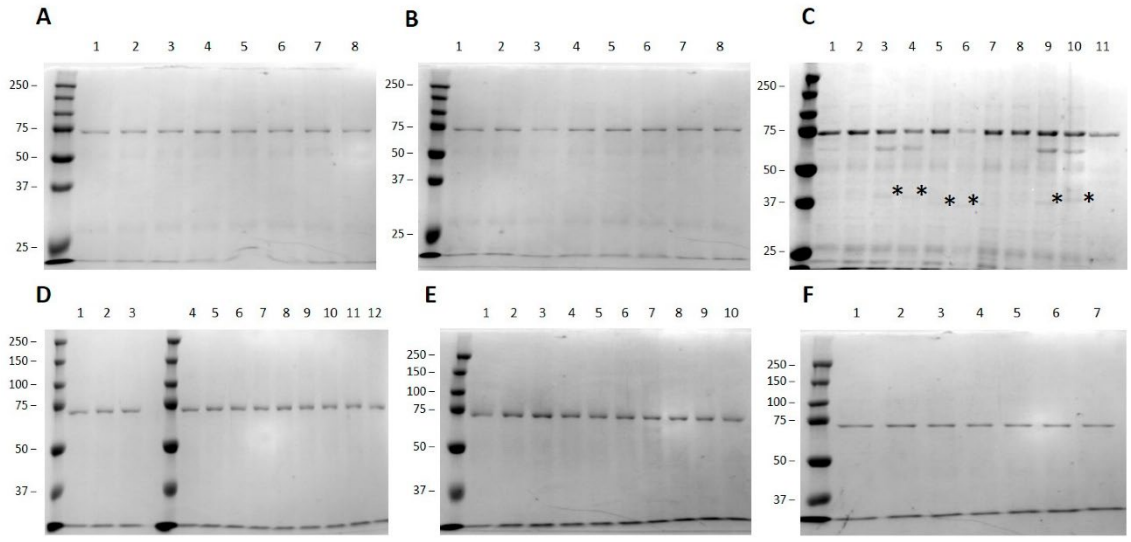


Figure S2.1 – Protein expression results from the HTP screening of 56 AraTs from Mycobacteria, overexpressed in E. coli C43 cells. ***(A)*** *AftA proteins: 1 – A1; 2 – A8; 3 – C2; 4 – C9; 5 – F1; 6 – F3; 7 – F7; 8 – F8.* ***(B)*** *AftB proteins 1 – A2; 2 – A9; 3 – C7; 4 – D9; 5 – D10; 6 – D11; 7 – E2; 8 – E5.* ***(C)*** *AftC proteins: 1 – A3; 2 – A10; 3 – C11; 4 – C12; 5 – D1; 6 – D2; 7 – E6; 8 – E7; 9 – E8; 10 – E9; 11 – F9.* ***(D)*** *AftD and EmbA proteins: 1 – A4; 2 – D4; 3 – H11; 4 – A5; 5 – A11; 6 – B4; 7 – B8; 8 – C5; 9 – G5; 10 – G11; 11 – H1; 12 – H2.* ***(E)*** *EmbB proteins: 1 – A6; 2 – A12; 3 – B3; 4 – B7; 5 – C4; 6 – G3; 7 – G6; 8 – G9; 9 – H7; 10 – H9.* ***(F)*** *EmbC proteins: 1 – B1; 2 – C3; 3 – G4; 4 – G7; 5 – G10; 6 – H4; 7 – H5. Asterisks (∗) indicate protein bands corresponding to the predicted MW of the corresponding target.*

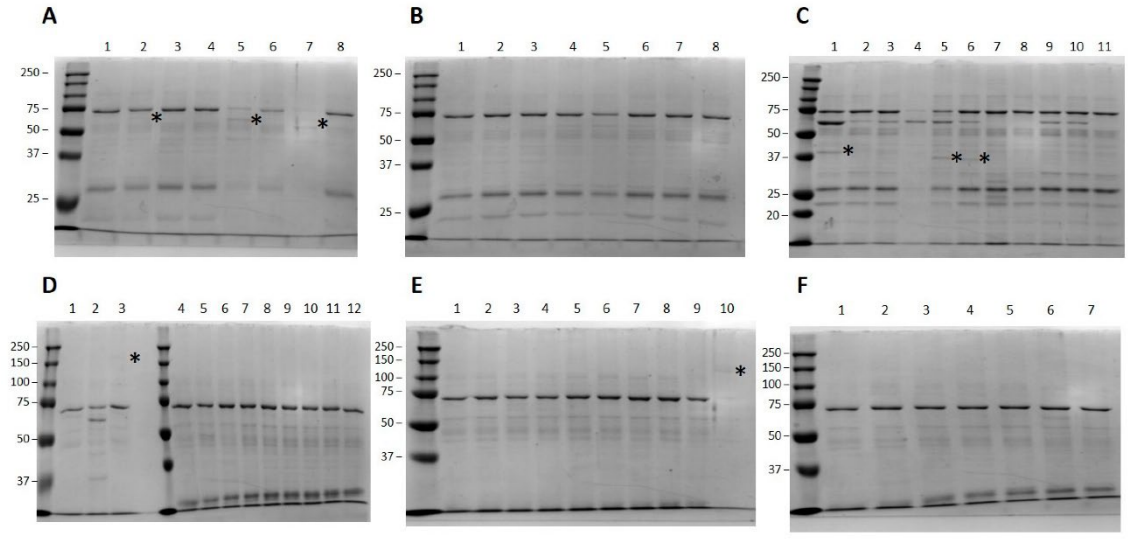


Figure S2.2 – Protein expression results from the HTP screening of 56 AraTs from Mycobacteria, overexpressed in E. coli BL21 (DE3) pLysS cells. ***(A)*** *AftA proteins: 1 – A1; 2 – A8; 3 – C2; 4 – C9; 5 – F1; 6 – F3; 7 – F7; 8 – F8.* ***(B)*** *AftB proteins 1 – A2; 2 – A9; 3 – C7; 4 – D9; 5 – D10; 6 – D11; 7 – E2; 8 – E5.* ***(C)*** *AftC proteins: 1 – A3; 2 – A10; 3 – C11; 4 – C12; 5 – D1; 6 – D2; 7 – E6; 8 – E7; 9 – E8; 10 – E9; 11 – F9.* ***(D)*** *AftD and EmbA proteins: 1 – A4; 2 – D4; 3 – H11; 4 – A5; 5 – A11; 6 – B4; 7 – B8; 8 – C5; 9 – G5; 10 – G11; 11 – H1; 12 – H2.* ***(E)*** *EmbB proteins: 1 – A6; 2 – A12; 3 – B3; 4 – B7; 5 – C4; 6 – G3; 7 – G6; 8 – G9; 9 – H7; 10 – H9.* ***(F)*** *EmbC proteins: 1 – B1; 2 – C3; 3 – G4; 4 – G7; 5 – G10; 6 – H4; 7 – H5. Asterisks (∗) indicate protein bands corresponding to the predicted MW of the corresponding target.*

CHAPTER THREE

Cryo-EM Structures and Regulation of Arabinofuranosyltransferase AftD from Mycobacteria

This chapter contains published data:

Tan Y. Z., **Rodrigues, J.**, Zhang L., Zheng R. B, Giacometti S. I, Rosário A. L., Kloss B., Dandey V. P., Wei H., Brunton R., Raczkowski A. M., Athayde D., Catalão M. J., Pimentel M., Clarke O. B., Lowary T. L., Archer M., Niederweis M., Potter C. S., Carragher B. & Mancia F., "Cryo-EM Structures and Regulation of Arabinofuranosyltransferase AftD from Mycobacteria" *Molecular Cell*, vol. 78, no. 4, pp. 683-699.e11, May 2020, doi: 10.1016/j.molcel.2020.04.014

Author Contributions – J. Rodrigues

Genomics expansion and small-scale screening: A. L. Rosário, B. Kloss and *J. Rodrigues*

Medium-scale expression and purification: *J. Rodrigues*

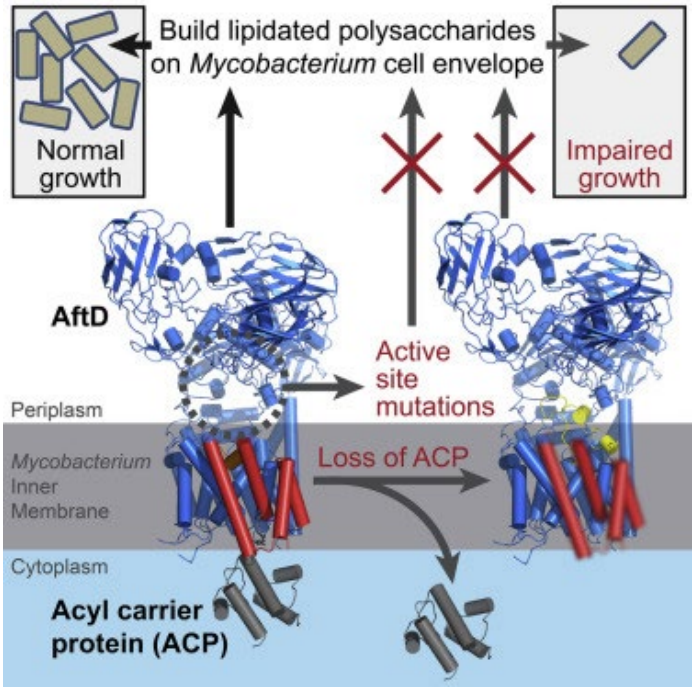
Expression and purification of AftD in *M. smegmatis*: *J. Rodrigues*, D. Athayde

Manuscript preparation: Y. Z. Tan and F. Mancia with input from all authors.

1. Summary

Mycobacterium tuberculosis causes tuberculosis, a disease that kills over 1 million people each year. Its cell envelope is a common antibiotic target and has a unique structure due, in part, to two lipidated polysaccharides – arabinogalactan and lipoarabinomannan. Arabinofuranosyltransferase AftD is an essential enzyme involved in assembling these glycolipids. We present the 2.9 Å resolution structure of *Mycobacterium abscessus* AftD, determined by single-particle cryo-electron microscopy. AftD has a conserved GT-C glycosyltransferase fold and three carbohydrate-binding modules. Glycan array analysis shows that AftD binds complex arabinose glycans. Additionally, AftD is non-covalently complexed with an acyl carrier protein (ACP). 3.4 Å and 3.5 Å structures of a mutant with impaired ACP binding reveal a conformational change, suggesting that ACP may regulate AftD function. Mutagenesis experiments using a conditional knockout constructed in *M. smegmatis* confirm the essentiality of the putative active site and the ACP binding for AftD function.

Keywords: glycosyltransferase, *Mycobacterium tuberculosis*, lipomannan, lipoarabinomannan, arabinofuranose, membrane protein, nanodisc, single-particle cryo-electron microscopy, acyl carrier protein



2. Introduction

Tuberculosis (TB) is one of the leading causes of death from a single infectious agent, outranking HIV/AIDS (WHO, 2018), being recently surpassed by COVID-19 [1]. The causative agent of TB is *Mycobacterium tuberculosis*, which belongs to a genus with a unique cell envelope made of lipids and carbohydrates (Figure 3.1A). The major component of the mycobacterial cell envelope is the mycolyl-arabinogalactan-peptidoglycan (mAGP) complex, a large macromolecule comprised of peptidoglycan (PG), arabinogalactan (AG) and long-chain (70 to 90 carbon atoms long) mycolic acids [2][3]. The mAGP complex is, in turn, decorated by a variety of glycolipids: phosphatidyl-*myo*-inositol mannoside (PIM), lipomannan (LM), and lipoarabinomannan (LAM) (Figure 3.1A). *M. tuberculosis* produces between 2,000 to 10,000 distinct lipids [4], most of which form an outer membrane of very low permeability [5][6]. This cell envelope is crucial for the growth and virulence of *M. tuberculosis* [7][8] and is a major contributor to the resistance of *M. tuberculosis* to common antibiotics. The importance of this unique structure in mycobacteria makes its biosynthetic enzymes attractive drug targets [9][10]. For example, isoniazid and ethambutol, two of the four TB frontline drugs, target mycobacterial cell wall synthesis [11].

Of the enzymes involved in mAGP biosynthesis, arabinofuranosyltransferases (AraTs) are responsible for the addition of D-arabinofuranose sugar moieties to AG and LAM [9]. All mycobacterial AraTs belong to the glycosyltransferase 53 (GT53) family [12][13] and are predicted to be multi-pass transmembrane (TM) proteins that use decaprenylphosphoryl-D-arabinofuranose (DPA) to donate an arabinofuranose residue to mAGP [14]. AraTs add 90 arabinofuranose residues to the galactan core of AG and 55 – 72 arabinofuranose residues to the mannan core of LAM. The best characterized AraTs are EmbA, EmbB, and EmbC [15][16], which are named Emb because they are putative targets of ethambutol [17][18]. Other family members include arabinofuranosyltransferases A (AftA) [19], B (AftB) [20], C (AftC) [21], D (AftD) [22], and possibly other enzymes yet to be identified [23].

AG synthesis starts when AftA adds a single D-arabinofuranose to the C5 of the D-galactofuranosyl residues at positions 8, 10, and 12 of the existing galactan. Thereafter, EmbA and EmbB extend the arabinan chain through α-(1→5) linkages; α-(1→3) branching is then introduced by AftC and, possibly, AftD. AftB is the final enzyme to cap off the AG, with the addition of a terminal β-(1→2)-linked D-arabinofuranose.

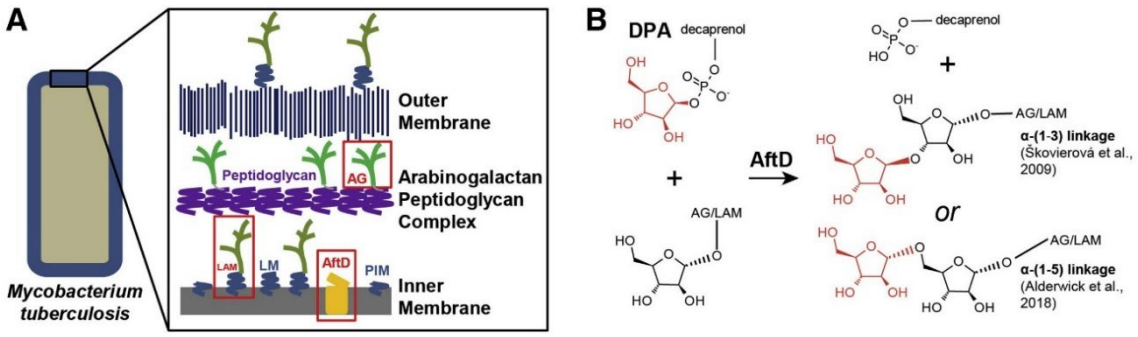


Figure 3.1 – Functional Analysis of the Role of Arabinofuranosyltransferase AftD in the Biosynthesis of Mycobacterial Cell. ***(A)*** *Model of cell envelope of M. tuberculosis. Red boxes highlight the arabinogalactan (AG) and lipoarabinomannan (LAM) components, which are synthesized by AftD (yellow).* ***(B)*** *Reaction catalyzed by AftD enzyme.*

For LAM, an unknown enzyme [23] primes the mannan chain with an α-(1→6) glycosidic bond before being extended through α-(1→5) linkages by EmbC. AftC and, possibly, AftD then catalyzes α-(1→3) branching before the LAM is capped in a β-(1→2) manner by AftB. These pathways are by no means fully characterized [19].

AftD is essential for growth of mycobacteria [22], and with a theoretical molecular mass of 150 kDa, it is the largest predicted glycosyltransferase encoded in the mycobacterial genome. Its precise catalytic activity is yet to be resolved; it has been described as an α-(1→3) [22] and/or α-(1→5) [24] linking enzyme and is thought to add the final few α-linked arabinofuranose residues to AG and LAM [24]; (Figure 3.1B). Additionally, its large size in comparison to the other AraTs (which range from 49 to 118 kDa) has raised the question of whether it could have additional functions beyond catalyzing glycosidic bond formation [22].

To better understand the function of this enzyme, we determined the 2.9 Å structure of recombinant AftD from *M. abscessus* produced in *E. coli* using single-particle cryogenic electron microscopy (cryo-EM). We analyzed its *in vitro* carbohydrate-binding capability via glycan array analysis, and we constructed a conditional *aftD* deletion mutant in *M. smegmatis* to interrogate the functionality of AftD mutants *in vivo*. Unexpectedly, analysis of the structure revealed that AftD is tightly bound to an acyl carrier protein (ACP). The physiological relevance of this interaction was confirmed in *M. smegmatis* through mass spectrometry. Finally, we purified and determined structures of an AftD mutant designed to abrogate ACP binding to 3.4 and 3.5 Å resolution. These structures revealed the ordering of a loop at the putative active site, suggesting that the ACP may regulate AftD function.

3. Materials and Methods

3.1. Phylogenetic and Conservation Analysis

PSI-BLAST [25] was performed using the amino acid sequence of *M. abscessus* AftD until no more additional sequences were found, producing 2385 sequences. Duplicate sequences from the same species and truncated variants of the same protein were removed, resulting in 665 unique sequences. Lengths of these sequences were plotted in Figure S3.1. Membrane topology for these sequences were predicted using the TMHMM server [26].

In order to compare sequences that are more closely related to *Mycobacterium* AftD, sequences shorter than 95% the length of *M. abscessus* AftD (~1400 residues) were excluded. CD-HIT [27] was then used to provide representative sequences and remove biases in the distribution of the data – notably *Mycobacterium* sequences dominate the original sequence list likely because of the medical relevance of this genus. The 289 final sequences were then aligned with MUSCLE [28] and used for conservation analysis.

3.2. Generation of a *M. smegmatis aftD* Conditional Knockout Strain

To construct a parent strain for *aftD* conditional knockout (cKO), a L5 integration vector pML3621 (L5 *integrase, Hyg*^{R}*, loxP-P*$_{smyc}$*::tetR38-loxP*) was first transformed into wild-type *M. smegmatis* mc^{2}155 cell to create ML2253 strain, which can constitutively express a reverse TetR (TetR38). The ML2253 colony was selected on 7H10 plates supplemented with 50 µg/mL hygromycin (Fisher) after 4 days at 37 °C. Subsequently, a temperature-sensitive replication vector pML2714 (pAL5000ts, Kan^{R}, *cre* recombinase expression) was transformed into ML2253 for unmarking. The unmarked colony (named ML2217) was selected on 7H10 plates supplemented with 30 µg/mL kanamycin (Fisher) after 7 days at 32 °C. Next, the unmarking plasmid ML2714 was removed by temperature selection on 7H10 (no antibiotic) plates at 40 °C for 4 days.

To conditionally knockout *aftD* in *M. smegmatis* mc^{2}155, a suicide vector pML4205 containing a TetR38-binding region (tetO-4C5G) and the 5' end (1st to 1089th bp) of *aftD* gene was transformed into the parent strain ML2217 for single cross-over selection on 7H10 plates supplemented with 50 µg/mL hygromycin at 37 °C for 4 days. PCR validation was then performed for the single cross-over colonies. The correct colonies were named

ML2218, the *aftD* conditional KO strain. In the presence of 20 ng/mL tetracycline, TetR38-tetracycline complex binds the upstream region (tetO-4C5G) of *aftD* and represses the transcription of *aftD* gene.

3.3. *M. abscessus aftD* Complementation

For the complementation experiment, the *M. abscessus aftD* gene, as well as the *aftD* variations, were cloned into an Ms6 integration vector pML4214 (Ms6 *integrase*, Ms6 attP attachment site, Kan^R) individually with a 200-bp upstream DNA fragment of *M. abscessus aftD* gene. The pML4214 and the pML4214-derivative plasmids were transformed into ML2218 strain individually. All the colonies were selected on 7H10 plates supplemented with 30 μg/mL kanamycin (Fisher) after 4 days at 37 °C.

3.4. Phenotypic Assays for *M. smegmatis aftD* Conditional Knockout Strain

To observe a clear phenotype of *aftD* conditional knockout mutant, the remaining cellular AftD protein levels were depleted as much as possible. First, the ML2218 seed culture was grown to an OD_{600} of 1.0 at 37 °C with a shaking speed of 200 rpm. For the AftD depletion, ML2218 cells were grown in 10 mL of 7H9/glycerol/tyloxapol liquid media supplemented with 20 ng/mL tetracycline with an initial OD_{600} of 0.002 at 37 °C, 200 rpm for 2 – 3 days. Subsequently, the culture was used for growth curve assay, microplate appearance assay and colony forming assay.

For the growth curve assay, all the AftD-depleted strains were inoculated into the fresh 10 mL of 7H9/glycerol/tyloxapol liquid media supplemented with or without 20 ng/mL tetracycline with an initial OD_{600} of 0.005. The cultures were then grown at 37 °C, 200 rpm and time points taken approximately every 4 – 8 hours until stationary phase. All experiments were done in triplicate.

For the microplate appearance assay, all the AftD-depleted strains were inoculated into 12-well plates (Costar). Each well contains 3 mL of 7H9/glycerol/tyloxapol liquid media supplemented with or without 20 ng/mL tetracycline with an initial OD_{600} of 0.05. The 12-well plates were shaking at 120 rpm at 37 °C for 2 – 3 days.

For the colony forming assay, all the AftD-depleted cells were streaked on 7H10 plates with or without 20 ng/mL of tetracycline, kept at 37 °C for 6 days.

3.5. Scanning Electron Microscopy

10 mL of *M. smegmatis* cells were harvested by spinning cells at 8000 *g* for 1 min with a Sorvall ST 16 R centrifuge (Thermo Scientific) to give around 0.4 – 0.5 g of cell pellet mass. Supernatant was discarded. Cell pellet was then resuspended with 2.5% glutaraldehyde (Tousimis) in a 15 mL Falcon tube and incubated at 4 °C for 2 hours. The suspension was then spun down again at 8000 *g* for 1 min and resuspended with PBS for 2 times, then with an ethanol (Fisher) gradient (30%, 50%, 70%, 100%) 15 min for dehydration. Each step takes 15 minutes and was followed with a spin down and resuspension. After the final dehydration, fixed cells were left to air dry overnight.

The cells were then smeared onto a scanning electron microscope sample holder using a pipette tip for imaging. The entire surface of the specimen was then sputter-coated with a thin layer of gold/palladium. The cells were imaged using secondary electron mode in a FEI Helios Nanolab 650 (FEI). Images were recorded using the scanning electron microscope beam at 2 keV and 100 pA with a working distance of 4.0 ± 0.5 mm. Data acquisition occurred manually using an in-house AutoIt script (https://www.autoitscript.com/site/) written by Bill Rice. The raw images were 3072 pixels by 2048 pixels, with 6.08 nm pixel width, a horizontal field width (HFW) of 18.67 mm and dwell time of 10.00 µs.

To measure the dimensions of cells, Adobe Photoshop was used to select out each cell. The cell length and width were calculated using ImageJ's plugin, Pombe Measurer (www.columbia.edu/~zz2181/Pombe_Measurer.html). Both D'Agostino & Pearson omnibus normality test and Shapiro-Wilk normality test showed that the distribution of lengths and widths were not normal. Hence, one-way ANOVA Kruskal-Wallis parametric test was done. Dunns post-test was then done between all pairs of samples. Statistical tests were done using GraphPad Prism 5.

3.6. Genomic Expansion and Small-Scale Screening

AftD genes were identified from a collection of fourteen *Mycobacterium* genomes using a bioinformatics approach [29]. Ligation independent cloning (LIC) was used to clone these targets from the genomes into five LIC-adapted expression vectors (pNYCOMPS-Nterm, pNYCOMPS-Cterm, pNYCOMPS-N23, pNYCOMPS-C23 and pMCSG7-10x) that contained a Tobacco Etch Virus (TEV) protease cleave site (ENLYFQSYV) and decahistidine affinity tag. Small and medium scale expression was done in a high throughput manner as described in detail in a previous protocol by Bruni and Kloss [30]. Only one construct*, M. abscessus aftD* in the pNYCOMPS-N23 plasmid, was successfully cloned and expressed in small and medium scale. This construct was thus brought forward for further structural studies.

3.7. *M. abscessus* AftD Mutagenesis

Site-directed mutagenesis on the *M. abscessus aftD* gene was done with overlapping primers designed per mutant using the Quick Change Lightning Kit (Agilent). The same pairs of primers were used on *aftD* in pNYCOMPS-N23 plasmid (for *E. coli* expression) and pML4214-Mab Promoter::*aftD M. abscessus* (for *M. smegmatis* expression). Small scale expression tests were carried out for the AftD mutants in the pNYCOMPS-N23 plasmid as described in detail in a previous protocol by Bruni and Kloss [30] (Figure S3.5). No TEV digestion was carried out for the small-scale expression tests.

3.8. *M. abscessus* AftD Expression, Purification and Nanodisc Reconstitution

M. abscessus AftD WT in the pNYCOMPS-N23 plasmid was transformed into BL21 (DE3) pLysS *E. coli* competent cells and plated onto Luria-Bertani (LB) agar (Fisher) plates supplemented with 100 µg/mL ampicillin (Sigma) and 35 µg/mL chloramphenicol (Sigma) and grown overnight at 37 °C. In the next day, a colony was picked and used to inoculate a starter culture containing 150 mL of Terrific broth medium (Fisher) supplemented with 100 µg/mL ampicillin and 35 µg/mL chloramphenicol. The starter culture was grown overnight at 37 °C with shaking (240 rpm) in an incubator shaker (New

Brunswick Scientific). In next day, 12 2-L baffled flasks each with 800 mL of Terrific Broth medium (Fisher) supplemented with 100 mg/mL ampicillin and 35 mg/mL chloramphenicol were inoculated with 10 mL of starter culture. The cultures were then grown at 37 °C with shaking (240 rpm) until OD_{600} reached 1.0 (approximately 3 hours). Temperature was then reduced to 22 °C and AftD expression was induced with 0.2 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) (Fisher). The culture was then incubated overnight with shaking (240 rpm). In the next day, the cells were harvested by centrifugation at 4,000 *g* in H6000A/HBB6 rotor (Sorvall) for 30 min at 4 °C. The supernatant was discarded and the pellet was resuspended in chilled 1x phosphate buffered saline (PBS) and centrifuged again at 4,000 *g* for 30 min at 4 °C. The supernatant was again discarded and the pellet was resuspended in lysis buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 20 mM $MgSO_4$, 10 µg/mL DNase I (Roche), 8 µg/mL RNase A (Roche), 1 mM tris(2-carboxyethyl)phosphine hydrochloride (TCEP), 1 mM PMSF, 1 tablet/1.5 L buffer EDTA-free cOmplete protease inhibitor cocktail (Roche). For 9.6 L of culture, there was ~15 – 20 g of wet cell pellet mass, which was resuspended with ~250 mL of lysis buffer. Cells were lysed by putting the suspension through a chilled Emulsiflex C3 homogenizer (Avestin) three times. To isolate the membrane fractions, the lysate was put through ultracentrifugation at 37,000 *g* in Type 45 Ti Rotor (Beckman Coulter) at 4 °C for 30 min. The supernatant was discarded and the pellet was resuspended in the lysis buffer up to a volume of 240 mL and homogenized using a hand-held glass homogenizer (Konte). The membrane fraction was then stored at -80 °C until the use.

The thawed membrane fraction was solubilized by adding n-dodecyl-β-D-maltopyranoside (DDM) to a final concentration of 1% (w/v) detergent for 2 hours at 4 °C with gentle rotation. Insoluble material was removed by ultracentrifugation at 40,000 *g* in Type 45 Ti Rotor at 4 °C for 30 min. The supernatant was added to six Falcon tubes containing pre-equilibrated Ni^{2+}-NTA resin (QIAGEN) in the presence of 40 mM imidazole and incubated with gentle rotation at 4 °C for 2 hours. The resin was washed with 10 column volumes of wash buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 60 mM Imidazole, 0.1% DDM and eluted with elution buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 200 mM Imidazole, 0.05% DDM. The eluted protein was exchanged into a buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 0.05% DDM using a PD-10 desalting column (GE) and concentrated using a 100-kDa concentrator (Pierce) to ~1 mg/mL.

The protein was then incorporated into lipid nanodisc [31] with the molar ratio 1:360:6 of *E. coli* polar lipid extract (Avanti): membrane scaffold protein 1E3D1 (MSP-1E3D1) and

incubated for 2 hours with gentle agitation at 4 °C. The *E. coli* polar lipid extract was prepared by adding the solid extract to deionized water to a final concentration of 10 mM. The mix was put on ice and then gently sonicated with a tip sonicator (Fisher Scientific) to dissolve the lipids. The lowest power rating was used and sonication was stopped when the mixture turned from cloudy to semi-transparent. No detergent was added to the lipid extract. Reconstitution was initiated by removing detergent with the addition of 150 mg Bio-beads (Bio-Rad) per mL of protein solution for overnight with constant rotation at 4 °C. Bio-beads were removed by passing the protein solution through an Ultrafree centrifugal filter unit (Fisher) at 4,000 *g* in a Centrifuge 5424 R (Eppendorf) at 4 °C for 1 min and the nanodisc reconstitution mixture was re-bound to Ni^{2+}-NTA resin with 15 mM imidazole for 2 hours at 4 °C in order to remove free nanodisc. The resin was washed with 10 column volumes of wash buffer consisting of 20 mM HEPES pH 7.5, 200 mM NaCl and 20 mM Imidazole, followed by 3 column volumes of elution buffer consisting of 20 mM HEPES pH 7.5, 200 mM NaCl and 200 mM Imidazole.

TEV protease [32] (~0.5 mg TEV protease added per pellet equivalent from 800 mL of initial bacterial culture) was added to the eluted protein, which was then injected into a Slide-A-Lyzer dialysis cassette (Thermo Fisher) and dialyzed against 200 mL of 20 mM HEPES pH 7.5 and 200 mM NaCl overnight at 4 °C. The mixture was then re-bound to Ni^{2+}-NTA resin with 15 mM imidazole for 1 hour at 4 °C in order to remove uncleaved AftD, cleaved tag, TEV protease and contaminant proteins. The resin was washed with 3 column volumes of wash buffer consisting of 20 mM HEPES pH 7.5, 200 mM NaCl. The flowthrough and wash fractions were concentrated using a 100-kDa concentrator to under 500 mL and loaded onto a Superdex 200 Increase 10/300 GL size-exclusion column (GE Healthcare Life Sciences) in gel filtration buffer (20 mM HEPES pH 7.5 and 200 mM NaCl). Throughout the entire process of purification, 15 mL of samples were taken and added to 5 mL of 6x reducing Laemmli SDS sample buffer (Bioland Scientific). The samples were then loaded on a 4%–20% Mini-PROTEAN TGX precast protein gel (Bio-rad) for protein gel electrophoresis in a Tris/Glycine/SDS buffer. The gel was developed using InstantBlue (Sigma) protein stain.

For *M. abscessus* AftD-R1389S, the expression and purification protocol is largely the same except for the following. First, AftDR1389S was not able to be cleaved by TEV protease, hence the protein was not put through overnight TEV cleavage. Second, during the second nickel affinity purification after reconstitution into nanodiscs, additional washes of 5 column volumes of 20 mM HEPES pH 7.5, 500 mM NaCl and 20 mM Imidazole and 5 column volumes of 20 mM HEPES pH 7.5, 500 mM NaCl, 5 mM

adenosine triphosphate (ATP) (Sigma) and 20 mM Imidazole were done in order to increase purity of the protein, given TEV cleavage was no longer done.

3.9. Single-Particle Cryo-EM Sample Vitrification

Purified AftD WT was concentrated using a 100-kDa concentrator (Pierce) to between 5 – 20 µL of sample at ~1 mg/mL. Due to the low yields obtained, the samples were vitrified using a semi-automated Spotiton V1.0 robot [33]–[35]. This device, for vitrifying cryo-EM samples, uses piezo-electric dispensing to apply small (50 pL) drops of sample across a 'self-blotting' nanowire grid as it flies past en-route to plunging into liquid ethane. Owing to the small amounts of sample required per vitrified grid, only ~3 µL of sample was required to be aspirated for multiple grids. Homemade carbon support nanowire grids with a regular array of holes were used [36], and were plasma-cleaned (Gatan Solarus) for 15 s using an H_2/O_2 mixture. The Spotiton V1.0 robot was operating at room temperature and moderate humidity.

For purified AftD-R1389S, the protein was concentrated using a 100-kDa concentrator (Pierce) to between ~10 µL of sample at ~2 mg/mL. 2.5 µL of sample was added to a plasma-cleaned (Gatan Solarus) 1.2/1.3 µm holey gold grid (Quantifoil UltrAuFoil) and blotted using filter paper on one side for 2 s using the Leica GP plunger system before plunging immediately into liquid ethane for vitrification. The plunger was operating at 6 °C with > 80% humidity to minimize evaporation and sample degradation. Due to conditions previously optimized using Spotiton, Leica GP plunging could be used to obtain good grids for the mutant without expending too much of the protein sample.

3.10. Data Acquisition

Images were recorded on a Titan Krios electron microscope (FEI) equipped with a K2 summit direct detector with a Quantum energy filter (Gatan) operating at 1.0605 Å per pixel in counting mode using the Leginon software package [37]. Pixel size was calibrated in-house using a proteasome test sample. Energy filter slit width of 15 eV was used during the collection and aligned automatically every hour using Leginon. Data collection was performed using a dose of ~95.72 $e^-/Å^2$ across 90 frames (150 ms per frame) at a dose rate of ~8.2 e^-/pix/sec, using a set defocus range of -1.0 µm to -2.0 µm. 100 µm objective aperture was used. For AftD WT, a total of 7,274 micrographs were

recorded over four sessions using an image beam shift data collection strategy [38]. For AftD-R1389S, the same strategy over a single session of three days yielded 4,886 micrographs. The dose was marginally different at ~96.78 $e^-/Å^2$.

3.11. Data Processing

For *M. abscessus* WT AftD, data from the four sessions were processed separately and combined toward the end of the processing pipeline. Movie frames were aligned using MotionCor2 [39] with 5 by 5 patches and B-factor of 100 through the Appion software package [40]. Micrograph CTF estimation was performed using both CTFFind4 [41] and GCTF [42], and the best estimate based on confidence was selected within the Appion software package. Template free particle picking with Gautomatch (Kai Zhang, personal communication, https://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/) using an extremely lenient threshold (to avoid missing any particles) was used to pick 2,800,102 particles (extracted 256 box size binned by 2) that were transferred into Relion 2.1 [43], [44] for 2D classification. 2D class averages that were ice or showed no features were discarded. An initial model was then generated in CryoSPARC *ab initio* [45] and used for Relion 3D refinement. The Euler angles and shifts were then used to re-extract the particles and re-center them. Relion 3D classification was then performed using 7 classes and only classes with clear structural features were selected.

At this stage, a total of 498,060 selected particles from all four session were combined and re-extracted unbinned. Per-particle CTF using GCTF was then performed before the particles were put through a Relion 3D refinement using the previously obtained initial model. A mask generated around the transmembrane region was then used to performed 3D classification with 8 classes, and particles belonging to classes with high resolution features were selected and re-extracted with a bigger box size of 288. A total of 490,616 particles were left in the final stack.

CryoSPARC non-uniform refinement [46] produced a 3.2 Å map, which was then put through *cis*TEM [47] CTF refinement to obtain better defocus values. The particles with the refined defocus values were then put through another round of CryoSPARC non-uniform refinement to produce a 3.0 Å map. Signal subtraction was then done in Relion to remove the nanodisc and flexible ACP density, and the signal subtracted particles were then put through CryoSPARC non-uniform refinement. The resulting Euler angles and shifts were then used to reconstruct the non-signal subtracted particles to produce a 2.9 Å map, called the AftD global refined map. This final map was sharpened using

phenix.auto_sharpen [48][49], which automatically determined that b_iso sharpening to high-resolution cutoff should be the algorithm to use. Overall b_sharpen applied was -56.04 Å^2 and final b_iso obtained was -44.30 Å^2.

When examining the map by eye, it was noticed that the domain corresponding to CBM3 had a lower resolution (poor density) compared to the rest of the map. In order to improve the resolution at this flexible portion, signal subtraction was performed to isolate just the CBM3 region, and focused classification using ten classes without alignment was done. A total of 50,410 particles were present in the highest resolution class. The resulting Euler angles and shifts were then used to reconstruct the non-signal subtracted particles to produce a 3.0 Å map, called the AftD-CBM3 focused classified map. This final map was sharpened using phenix.auto_sharpen, which automatically determined that b_iso sharpening to high_resolution cutoff should be the algorithm to use. Overall b_sharpen applied was -29.62 Å^2 and final b_iso obtained was -45.97 Å^2.

For *M. abscessus* AftD-R1389S, the same data processing strategy was used except for the following differences. First, 226,478 particles were picked initially and a consensus map of 150,978 particles was obtained. Second, from these 150,978 particles, signal subtraction was done using a spherical mask around the putative active site of AftD where additional density could be seen. Focused 3D classification without alignments was then performed using Relion with five classes. Different T values (10, 20, 30, 40, 50, 60, 70, 80 and 90) were titrated and T = 60 was determined by visual inspection and comparison of classes to give the best classification. Classes 3, 4 and 5 looked noisy and not significantly different. Classes 1 and 2 had high resolution details and had significant differences between them. A CryoSPARC non-uniform refinement was then performed using particles from each of these classes to give final maps at 3.5 Å and 3.4 Å each, respectively. Both final maps were sharpened using phenix.auto_sharpen, using b-factors of -125.09 Å^2 and -129.19 Å^2, respectively.

All conversions between Relion, CryoSPARC, and *cis*TEM were performed using Daniel Asarnow's pyem script (personal communication, https://github.com/asarnow/pyem).

3.12. Model Building and Refinement

In order to aid in model building against a single structure, phenix.combine_focused_maps [48][49] was used to combine the WT AftD global

refined map and WT AftD-CBM3 focused classified map into the WT AftD composite map.

For model building of WT AftD, the Chimera [50] program was first used to segment out the protein components and exclude the nanodisc. *De novo* model building for the cryo-EM map its primary sequence was initiated using Buccaneer [51] within the CCP-EM software suite [52]. After 2 rounds of Buccaneer model building, the partial model was brought into Coot [53] program for manual model building. After the model was built, it was refined against the cryo-EM map utilizing real space refinement in the Phenix program. Restraints for the lipids were generated using phenix.eLBOW and for the metal ions using phenix.ready_set. Thereafter, model adjustment and refinement were performed iteratively in Coot and Phenix, with the statistics being examined using Molprobity [54] until no further improvements were observed. The final map and model were then validated using 1) EMRinger [55] to compare map to model, 2) ResMap [56] to calculate map local resolution, 3) 3DFSC program suite [57] to calculate degree of directional resolution anisotropy through the 3DFSC and 4) SCF program [58] to calculate the sampling compensation factor (SCF), which quantifies how inhomogeneity in Euler angle distributions contributes to attenuation of the FSC. Map-to-model FSCs were also calculated by first converting the model to a map using Chimera molmap function at Nyquist resolution (2.121 Å). A mask was made from this map using Relion (after low-pass filtering to 8 Å, extending by 1 pixel and applying a cosine-edge of 3 pixels), and was then applied to the density map. Map-to-model FSC was calculated using EMAN [59] proc3d between these maps.

For model building of AftD-R1389S, the WT structure was overlaid and Coot was used to manually remove or build residues. Other than that, the model building and refinement strategy mirrored the WT AftD mentioned above.

3.13. Model Analysis

A cavity search using the Solvent Extractor from Voss Volume Voxelator server [60] was performed using an outer probe radius of 5 Å and inner probe radius of 2 Å. In order to search for other PDB structures with similar fold, a Dali server [61] search was performed – first globally and then against the different domains of the model. The Dali server was used to generate the structural conservation figures. Coot SSM superpose was used to align structures of other glycosyltransferases against AftD.

3.14. Screening, Imaging and Data Analysis of the Glycan Array

Glycan array analysis was done with *M. abscessus* AftD WT protein solubilized in DDM. Slides were pre-wetted in buffer A (25 mM Tris-HCl pH 7.8, 0.15 mM NaCl, 2 mM $CaCl_2$ and 0.05% Tween 20) for 5 min, rinsed with buffer B (25 mM Tris-HCl pH 7.8, 0.15 mM NaCl, and 2 mM $CaCl_2$) three times, and blocked overnight with buffer C (1% BSA in 25 mM Tris-HCl pH 7.8, 0.15 mM NaCl, and 2 mM $CaCl_2$) at 4 °C. Aliquots (500 µL) of serial dilutions of protein samples in buffer C were transferred to wells of the slide module immediately after aspiration of the blocking buffer. Wells were sealed with an adhesive seal and incubated for 60 min at 37 °C. Protein was removed by aspiration, and slides were washed 10 times with buffer A and three times with buffer B. Fluorescence was measured directly or after addition of a secondary antibody in buffer C (1:1000 dilution). Slides were incubated with a secondary antibody at room temperature for 40 min before being washed repeatedly with buffer A and deionized water.

Before being scanned, slides were dried by centrifugation. Microarrays were scanned at 5 µm resolution with a GenePix 4000B scanner (Molecular Devices, Sunnyvale, CA). The fluorescent signal was detected at 532 nm for Cy3 or Alexa Fluor 555 and 488 nm for Alexa Fluor 488. The laser power was 100%, and the photomultiplier tube gain was 400. The fluorescent signals were analyzed by quantifying the pixel density (intensity) of each spot using GenePix ProMicroarray Image Analysis Software version 6.1. Fluorescence intensity values for each spot and its background were calculated. The local background signal was automatically subtracted from the signal of each separate spot, and the mean signal intensity of each spot was used for data analysis. Averages of triplicate experiments and standard deviations were calculated using Microsoft Excel.

3.15. Overexpression of *M. abscessus* AftD in *M. smegmatis*

M. abscessus AftD, containing N-terminal FLAG and His_{10} tags, was cloned into pVVAP [62] plasmid for overexpression. This plasmid contains a pAL5000 replication origin, a kanamycin resistance cassette, and allows expression of N-terminally FLAG+His_{10}-tagged AftD. The plasmid was transformed by electroporation into *M. smegmatis* mc^2155 and plated onto Middlebrook 7H10 (Sigma-Aldrich), plates, supplemented with 0.5% glucose and 15 µg mL^{-1} kanamycin, and left at 37 °C until single colonies were visible (2 to 3 days). A single colony was picked to inoculate 30 mL of Middlebrook 7H9 medium

(USBiological) supplemented with 0.5% glucose, 0.05% Tween 80 and 15 mg mL^{-1} kanamycin, to grow overnight at 37 °C with constant orbital shaking at 220 rpm. The starter culture was then used the next day to inoculate 5 L of Middlebrook 7H9 medium (USBiological) supplemented with 0.2% succinate, 0.05% Tween 80 and 15 µg mL^{-1} kanamycin, at an initial OD_{600} of 0.05, and left to grow overnight at 37 °C with constant orbital shaking at 220 rpm. The next morning, AftD expression was induced by adding acetamide to a final concentration of 0.2%, once the culture reached an OD_{600} of 0.8 to 1.0, roughly 16 h after inoculation. The culture was grown for 24 h after induction at 37 °C with constant orbital shaking at 220 rpm, then harvested at 4 °C by centrifugation for 20 min at 4000 rpm (JA-10 rotor, Beckman Coulter). Supernatant was then discarded, and the cell pellet was washed once with PBS, centrifuged again at 4 °C for 20 min at 4000 rpm, then frozen at -20 °C.

The cells were thawed and resuspended in lysis buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 20 mM $MgSO_4$, 10 µg/mL DNase I, 1mM TCEP, 1 tablet/1.5 L buffer EDTA-free cOmplete protease inhibitor cocktail (Roche). For 5 L of culture, there was ~10 g of wet cell pellet mass, which was resuspended in ~60 mL of lysis buffer. Cells were then lysed in a French Press, at 1500 bar, after 4 passages. To isolate the membrane fractions, the lysate was put through ultracentrifugation at 42,000 *g* in Type 45 Ti Rotor (Beckman Coulter) at 4 °C for 45 min. The supernatant was discarded, the pellet was resuspended in the lysis buffer up to a volume of 15 mL and homogenized using a hand-held glass homogenizer (Wheaton) and then stored at -20 °C. The thawed membrane fraction was solubilized by adding DDM to a final concentration of 1% (w/v) detergent for 3 hours at 4 °C with gentle rotation. Insoluble material was removed by ultracentrifugation at 42,000 *g* in Type 45 Ti Rotor at 4 °C for 45 min. The supernatant was added to a Falcon50 tube containing pre-equilibrated ANTI-FLAG® M2 Affinity Gel (Sigma) and incubated with gentle rotation overnight at 4 °C. The gel was washed with 10 column volumes of wash buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 0.05% DDM and incubated at 4 °C with gentle rotation for 1h with elution buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 0.05% DDM, 100 µg/mL FLAG peptide (Sigma), then finally eluted. The eluted fraction was then run on both SDS-PAGE and Native-PAGE gels and sent for subsequent mass spectrometry analysis.

3.16. Mass Spectrometry Analysis and Protein Identification

Samples were subjected to trypsin digestion. Briefly, protein sample was reduced with 10 mM Dithiothreitol (DTT) (Sigma) for 40 min at 56 °C followed by alkylation with iodoacetamide (Sigma) 55 mM for 30 min in the dark. Excessive iodoacetamide was quenched by further incubation with DTT (10 mM for 10 min in the dark). The resulting sample was digested overnight with trypsin (Proteomics grade from Promega) at 37 °C (1:50 protein/trypsin ratio), dried and resuspended in 8 mL LC-MS water 0.1% formic acid (Fisher Chemicals).

Nano-liquid chromatography-tandem mass spectrometry (nanoLC-MS/MS) analysis was performed on an ekspert NanoLC 425 cHiPLC® system coupled with a TripleTOF® 6600 with a NanoSpray® III source (Sciex). Peptides were separated through reversed-phase chromatography (RP-LC) in a trap-and-elute mode. Trapping was performed at 2 µL/min with 100% A (0.1% formic acid in water, Fisher Chemicals, Geel, Belgium), for 10 min, on a Nano cHiPLC Trap column (Sciex 200 µm x 0.5 mm, ChromXP C18-CL, 3 µm, 120 Å). Separation was performed at 300 nL/min, on a Nano cHiPLC column (Sciex 75 µm x 15 cm, ChromXP C18-CL, 3 µm, 120 Å). The gradient was as follows: 0-1 min, 5% B (0.1% formic acid in acetonitrile, Fisher Chemicals); 1-61 min, 5%–30% B; 61-63 min, 30%–80% B; 63-71 min, 80% B; 71-73 min, 80%–5% B; 73-90 min, 5% B.

Peptides were sprayed into the MS through an uncoated fused-silica PicoTip emitter (360 µm O.D., 20 µm I.D., 10 ± 1.0 µm tip I.D., New Objective). The source parameters were set as follows: 15 GS1, 0 GS2, 30 CUR, 2.5 keV ISVF and 100 °C IHT. An information dependent acquisition (IDA) method was set with a TOF-MS survey scan of 400-2000 m/z for 250 ms. The 50 most intense precursors were selected for subsequent fragmentation and the MS/MS were acquired in high sensitivity mode (150-1800 m/z for 40 ms each). The selection criteria for parent ions included a charge state between +2 and +5 and counts above a minimum threshold of 125 counts per second. Ions were excluded from further MS/MS analysis for 12 s. Fragmentation was performed using rolling collision energy with a collision energy spread of 5.

The obtained spectra were processed and analyzed using ProteinPilot software, with the Paragon search engine (version 5.0, Sciex). The following search parameters were set: search against *M. smegmatis* and *E. coli* from Uniprot/SwissProt database, and protein sequence of AftD from *M. abscessus*; Iodoacetamide, as Cys alkylation; Tryspsin, as digestion; TripleTOF 6600, as the Instrument; ID focus as biological modifications and

Amino acid substitutions; search effort as thorough; and a FDR analysis. Only the proteins with Unused Protein Score above 1.3 and 95% confidence were considered.

3.17. Quantification and statistical analysis

For calculations of Fourier shell correlations (FSC), the FSC cut-off criterion of 0.143 [63] was used. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. For calculation of *M. smegmatis* cell length and width, both D'Agostino & Pearson omnibus normality test and Shapiro-Wilk normality test showed that the distribution was not normal. Hence, one-way ANOVA Kruskal-Wallis parametric test was done. Dunns post-test was then done between all pairs of samples.

4. Results

4.1. Phylogenetic Analysis Reveals Existence of AftD Homologs across Actinobacteria

Iterative rounds of PSI-BLAST [25] were used to identify homologs of AftD using the amino-acid sequence of the protein from *M. tuberculosis* as a template. This analysis revealed that AftD is found not just in the *Corynebacteriales* order but also in 11 other orders of the *Actinobacteria* class [64], including *Streptomycetales*, *Propionibacteriales*, and *Frankiales* (Figure S3.1). This is consistent with the fact that members within the *Actinobacteria* class have similar cell-envelope structures [65]. All homologs have a similar predicted topology, computed using the TMHMM server [26], with an N-terminal TM region, followed by a long periplasmic region and a shorter C-terminal TM region. The average length of all AftD proteins is 1,381 amino acids. Most sequences (including the mycobacterial ones) cluster around this length, while AftD homologs in corynebacteria are significantly shorter, with an average length of 1,070 residues (Figure S3.2A). Excluding corynebacteria, the average length of AftD from the remaining genera is 1,422 residues.

4.2. Conditional Deletion of the *M. smegmatis aftD* Gene Results in an Altered Cell Phenotype

As AftD has been reported to be an essential enzyme in *M. tuberculosis* [66] and in *M. smegmatis* [22], a conditional knockout (cKO) was designed to interrogate the function of AftD in *M. smegmatis* and provide a system for structure-based mutagenesis studies. Attempts to obtain a full *aftD* knockout *M. smegmatis* strain were performed but deemed unsuccessful, thus confirming that AftD is essential in *M. smegmatis*.

First, we constructed the *M. smegmatis* ML2217 strain (wild-type; WT), which constitutively produces the reverse Tet repressor TetR38. Then, the chromosomal *aftD* gene in *M. smegmatis* ML2217 was placed under the control of the tetO-4C5G operator [67], which is constitutively transcribed in the absence of tetracycline (Figure S3.2B). In the presence of tetracycline, the TetR38 binds to tetO-4C5G, repressing transcription of the *aftD* gene. To our knowledge, this is the first conditional *aftD* deletion mutant in mycobacteria that can be readily controlled in a temperature-independent fashion.

Phenotypic analyses of WT *M. smegmatis* (ML2217) and the conditional *aftD* knockout mutant (ML2218) were performed in medium with and without tetracycline. When grown in 7H9 medium without tetracycline, both WT and cKO strains had similar growth curves, reaching an OD_{600} of 2.5 after about 32 h at the shaking speed of 200 rpm (Figure 3.2A). When 20 ng/mL tetracycline was added to the growth media, OD_{600} failed to reach 0.5 after 48 h (Figure 3.2A), confirming that *aftD* is, indeed, essential for the growth of *M. smegmatis* under those conditions. In the presence of tetracycline to silence *aftD* expression, cKO cells readily aggregated in liquid culture at the lower shaking speed (120 rpm), suggesting that the properties of the cell surface have changed (Figure 3.2C). Growth of the conditional *aftD* knockout mutant on agar plates with tetracycline was slower compared to that on plates lacking tetracycline after 2 days of incubation. On day 6, the colonies formed by the conditional *aftD* knockout mutant on agar plates with tetracycline were flat and smoother compared to colonies of the control strains (Figure 3.2D). Phenotypes of conditional mutants, which are regulated by small molecules, are, in general, stronger in liquid cultures compared to agar plates, because access and subsequent uptake of small molecules is more efficient in liquid cultures. When imaged by scanning electron microscopy, cells from the cKO strain appeared shorter in the presence of tetracycline (Figure 3.2E). Indeed, this observation was corroborated by measurements of the cell length and width, showing a statistically significant reduction in cell length of the cKO strain in the presence (1.9 ± 0.6 µm) versus the absence (2.5 ± 0.8 µm) of tetracycline, while the widths remained the same (Figure 3.2F).

4.1. Genomic Expansion and Complementation of the *M. abscessus aftD* Gene

To study the AftD function *in vitro* and determine its structure, we adopted a structural genomics approach [29][30][68] to identify homologs that yielded high expression levels in *E. coli* and were stable in detergents compatible with structure determination. High-throughput small-scale expression and purification screens of AftD from 14 mycobacterial species revealed that, under our experimental conditions, only the *M. abscessus* homolog was expressed and purified from *E. coli* in quantities detectable in Coomassie-stained protein gels (work described in Chapter 2: *High Throughput Expression Screening of Arabinofuranosyltransferases from Mycobacteria*). This homolog has a high sequence identity of 62% and 70% with the *M. tuberculosis* and *M. smegmatis* AftDs, respectively.

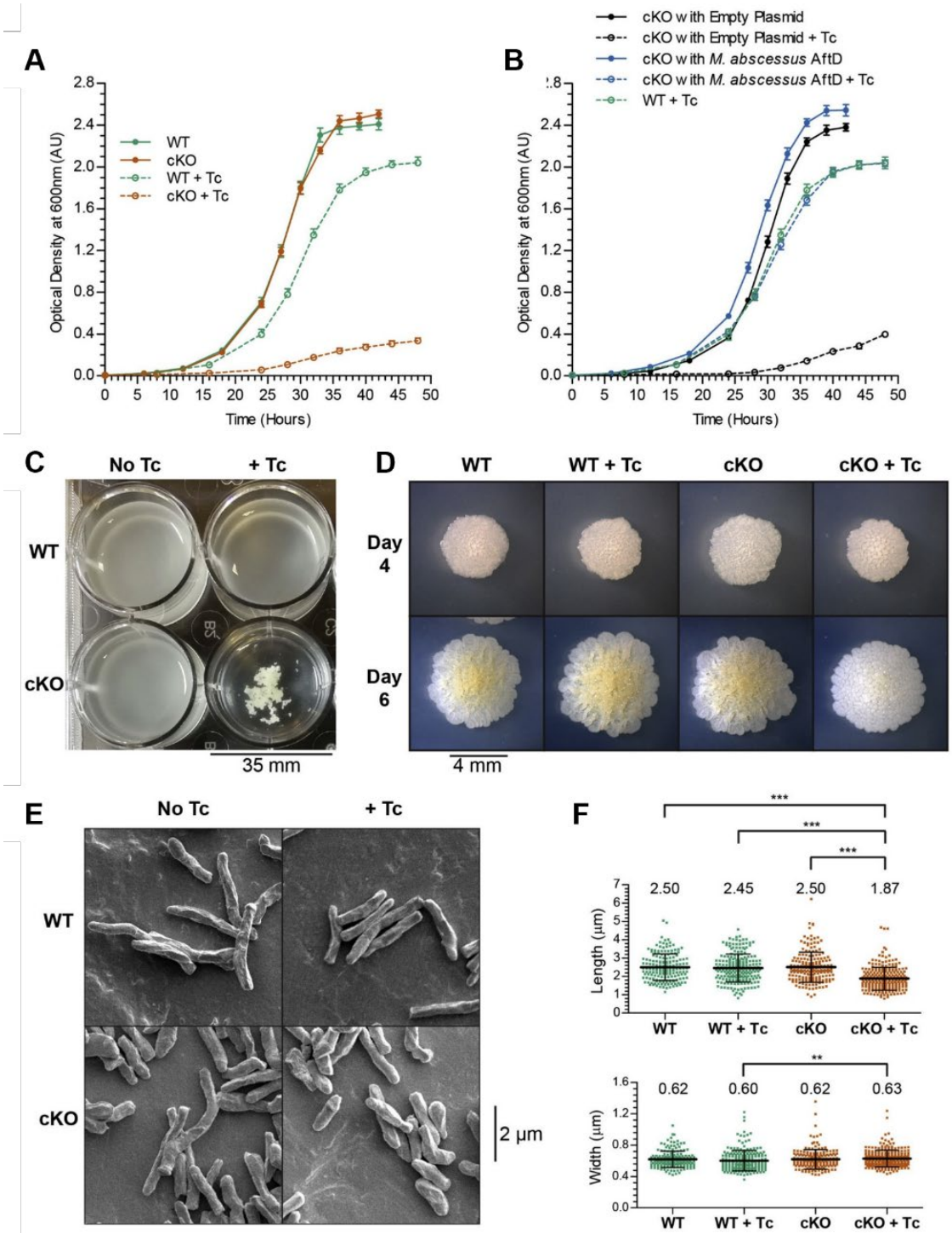


Figure 3.2 – Phenotypical analysis of wild-type M. smegmatis and the conditional aftD knockout mutant ML2218 (cKO). (A) *Growth of WT and the cKO without and with 20 ng/mL tetracycline.* ***(B)*** *Complementation of the cKO with empty plasmid and plasmid carrying the M. abscessus aftD gene. The strains were grown with and without tetracycline. Both growth curves were recorded in biological triplicates. Standard deviations are plotted.* ***(C)*** *Clumping of the cKO cells in the presence of 20 ng/mL tetracycline in liquid culture in microplates.* ***(D)*** *Colony morphology of WT and cKO M. smegmatis without and with 20 ng/mL tetracycline grown at 37°C for 4 and 6 days.* ***(E and F)*** *Scanning electron microscopy micrographs* ***(E)*** *and quantification of the cell length and width* ***(F)*** *from these images. For the cell lengths and widths* ***(F)****, one-way ANOVA Kruskal-Wallis parametric test was done, as the distribution was not normal. Dunn's post-test was done between all pairs of samples.*

To confirm the *in vivo* functionality of *M. abscessus* AftD, we complemented the *M. smegmatis* cKO with the *M. abscessus aftD* gene. Initial experiments using tetracycline or acetamidase promoters to express the *aftD* gene resulted in cell aggregation, suggesting altered cell-surface properties of *M. smegmatis*, probably resulting from *aftD* overexpression. However, an expression plasmid containing the *M. abscessus aftD* gene transcribed from its endogenous promoter (located in a 200-bp upstream fragment) fully restored wild-type growth of the cKO strain in the presence of tetracycline (Figure 3.2B). From these results, we could conclude that *M. abscessus* AftD is functional in *M. smegmatis* and that the phenotype observed in the conditional *aftD* knockout *M. smegmatis* mutant is, indeed, due to the depletion of the *M. smegmatis* AftD protein.

4.2. Structure Determination Using Cryo-EM

We determined the structure of full-length *M. abscessus* apo-state AftD using single-particle cryo-EM. Despite substantial efforts to optimize the overexpression and purification of AftD in *E. coli*, the final yields after purification were low: less than 100 µg protein could be isolated from 9.6 L cell culture. Nevertheless, AftD was successfully incorporated into lipid-filled nanodiscs after a two-step purification in detergent (Figure 3.3). To generate multiple cryo-EM grids of the sample at various conditions for screening and data collection despite the low yield, a picoliter-dispensing robot (Spotiton) was used to vitrify the sample [33]–[35]. Data were collected on a Titan Krios microscope equipped with a Gatan K2 Summit camera fitted with an energy filter. 7,274 micrographs were collected, and 889,985 particles were initially processed (Figure S3.3C; Table S3.1). After data processing, the final map was produced from 490,616 particles, resulting in a reconstruction with an overall resolution of 2.9 Å (Figure S3.3). An atomic model for the entire protein was built, except for both termini (residues 1 and 1,392-1,410) and some disordered loop regions (253-259, 270-286, 330-356, 390-401, 1,281-1,285, and 1,358-1,367) (Figure 3.4).

The region represented by residues 950-1,090 was flexible; signal subtraction followed by focused classification of this part [69] improved the map quality to allow for better model building. The final refined model exhibits good stereochemistry and correlates well with the density map (Table S3.1).

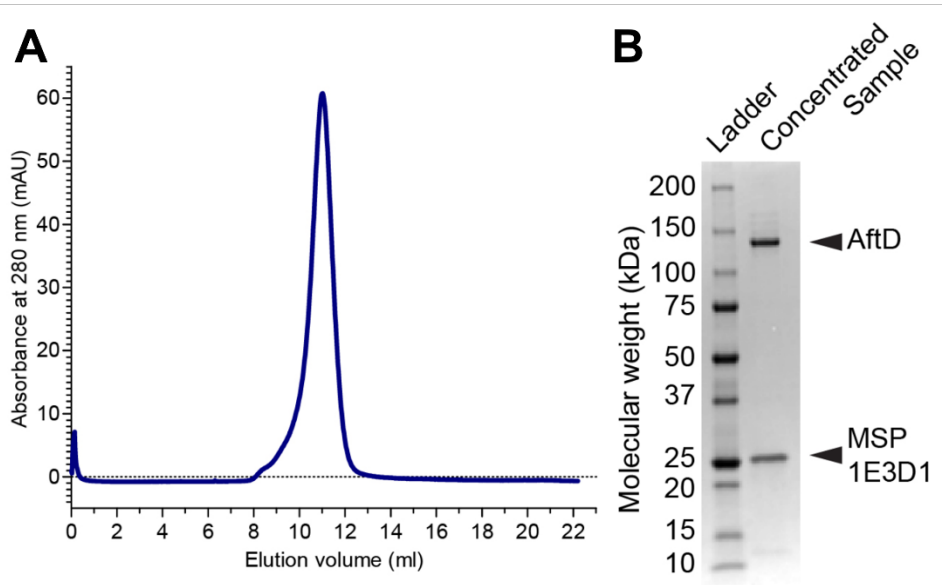


Figure 3.3 – Purification of wild-type M. abscessus AftD. (A) *Representative size-exclusion chromatography (SEC) trace of AftD incorporated into MSP-1E3D1 nanodiscs. Fractions corresponding to the peak were used for cryo-EM analysis.* ***(B)*** *SDS-PAGE gel of the concentrated sample from the peak fractions stained with Coomassie blue. The predicted molecular weight of AftD is 150 kDa, while that for MSP-1E3D1 is 32 kDa.*

4.1. Architecture of AftD

AftD consists of a membrane-embedded portion with 16 TM helices and a soluble portion that resides on the periplasmic side [5] of the inner membrane (Figure 3.4A). Both termini of AftD project into the cytoplasm. Starting from the N-terminus, the first 11 TM helices form a typical GT-C glycosyltransferase fold (Figures 3.4A and 3.4B; [70]). Thereafter, the polypeptide chain enters the periplasm. The soluble, periplasmic portion of AftD is large, accounting for ~60% of the mass of the protein, and is made up of five distinct domains: a domain of mixed α/β folds that is homologous to other solved GT-C structures, three carbohydrate-binding modules (CBMs), and one structural β-sandwich domain (Figure 3.4C). Although these domains are discrete, they have extensive inter-domain interactions that keeps them tightly associated, including some secondary structural elements: β4 from the soluble GT-C domain forms a beta-sheet with β29 and β30 of CBM3, and β7 of CBM1 forms a beta-sheet with β16 of CBM2 (Figure 3.4C). After the soluble domains, the polypeptide chain enters back into the membrane via a single TM helix (TM12) that connects via a long loop to TM13–TM16 and the C-terminus as a separate membrane-bound domain from the N-terminal TM helices (Figure 3.4C).

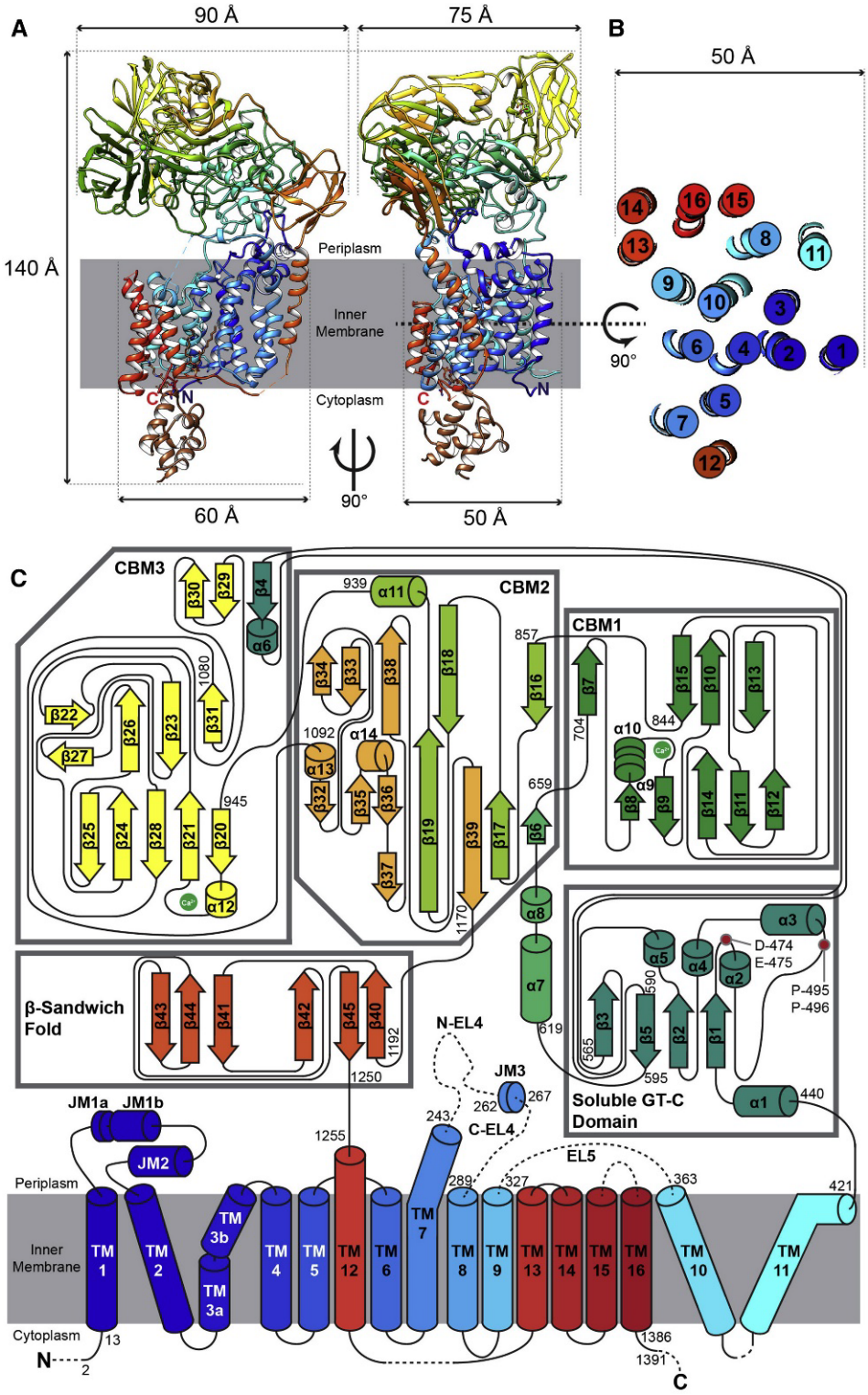


Figure 3.4 – Architecture of AftD Complex. (A) *Single-particle cryo-EM structure of the AftD complex, rendered as a cartoon and indicated in rainbow from the N-terminus (blue) to the C-terminus (red). The complexed E. coli ACP and ligands are indicated in brown. Ca^{2+} ions are rendered as green spheres. Two orthogonal views that are perpendicular to the plane of the membrane are shown. The approximate dimensions of the monomer are 85, 145, and 75 Å (width, height, and depth).* ***(B)*** *TM helix arrangement of AftD, viewed as a slice and magnified.* ***(C)*** *Two-dimensional topological diagram of AftD. The individual soluble domains are enclosed in separate gray boxes. The topology diagram is colored in rainbow from the N-terminus (blue) to the C-terminus (red). Unbuilt parts of the model due to poor map density are indicated by dotted lines. The two bound Ca^{2+} atoms are indicated as green circles, and side chains coordinating them are indicated as sticks. The putative catalytic residues previously hypothesized [22] are indicated as red circles.*

4.2. AftD Has a Conserved Glycosyltransferase GT-C Fold

Using the Dali server [61], atomic models with structural similarity to AftD were retrieved, and top hits corresponded to all full-length GT-C structures deposited in the database: yeast oligosaccharyltransferase STT3 (PDB: 6EZN and PDB: 6C26) [71][72], human oligosaccharyltransferase STT3A (PDB: 6S7O) and STT3B (PDB: 6S7T) [73], *Archaeoglobus fulgidus* oligosaccharyltransferase AglB (PDB: 3WAJ, 5GMY, and 3WAK) [74][75], *Campylobacter lari* oligosaccharyltransferase PglB (PDB: 3RCE, 5OGL, and 6GXC) [76]–[78], *Cupriavidus metallidurans* aminoarabinose-transferase ArnT (PDB: 5EZM and 5F15) [79], and yeast mannosyltransferase Pmt1-Pmt2 (PDB: 6P2R and 6P25) [80]. The first ~600 residues of AftD correspond to this homologous GT-C fold and will subsequently be referred to as the GT-C super-domain of AftD (Figures 3.4C and S3.4C). This GT-C super-domain comprises a total of 11 TM helices (blue in Figure 3.4B), the same as Pmt1-Pmt2 (Figure S3.4D) but two less than the other solved GT-Cs. The two missing TM helices of AftD correspond to TM helix 8 for STT3 (Figure S3.4E), helices 8 and 9 for PglB (Figure S3.4F), and helices 12 and 13 for ArnT (Figure S3.4G). The remaining 11 TM helices are structurally homologous to the other solved GT-Cs in several aspects: (1) they are in the same arrangement in AftD as seen in the other five solved GT-C structures; (2) there are two periplasmic juxtamembrane (JM) helices (JM1 and JM2) between TM1 and TM2; (3) TM3 is kinked by a proline and broken into two smaller helices; and (4) TM7 extends beyond the membrane into the periplasm and is joined by a longer loop to TM8. A short alpha helical JM helix (JM3) is present in the loop (Figure 3.4C).

4.3. Putative Active Site and Substrate Binding Pocket

The postulated substrates for AftD are the donor DPA [81] and a poly-arabinose acceptor that is part of the growing AG or LAM [22] (Figure 3.1B). Although the structure of AftD was solved in the apo state, the high degree of homology as described earlier between AftD's GT-C super-domain and the other solved GT-C glycosyltransferases allows for inference of mechanistic insights. In the GT-C super-domain, a large cavity is present between the periplasmic domain and the periplasm-facing ends of TM4, TM6, TM7, TM9, and TM10, with a volume of ~1,240 $Å^3$ (Figure 3.5A). Using the unique AftD sequences of the *Actinobacteria* class, sequence conservation was mapped onto the structure of AftD to reveal that this cavity contains several highly conserved residues, including E251,

D474, and E475, which are the most plausible catalytic residue candidates. When representative GT39 (Pmt1) and GT66 (PglB) family structures were aligned with the GT-C super-domain, the active sites of both structures result in the same area as that of AftD (Figure 3.5B). Other GT-C glycosyltransferases, such as PglB, require the use of divalent metal cations to function [82], and AftD has a negatively charged pocket that could be used for this purpose (Figures 3.5A and 3.5B, red dotted-line circles). Interestingly, the charge around the putative active site for AftD is slightly positive, compared to slightly negative for Pmt1 and PglB. This is consistent with the different acceptor substrates of these proteins: Pmt1 has an aspartic acid residue that forms electrostatic interactions with serine/threonine on the substrate [80], and PglB has a preference for the negatively charged side chain at the −2 position [77], while AftD is expected to bind to a poly-arabinofuranose glycan moiety (which can form ion-dipole interactions with positively charged residues).

A shallow groove along TM6, TM9, and TM13 of AftD starts at the cytoplasmic side of the membrane and ends at the putative active site and is likely to accommodate the putative DPA donor (Figure S3.4B). A diacyl-phospholipid (modeled as phosphatidylethanolamine) can be resolved in the density map near the cytoplasmic region. There is moderate conservation in this region, and it is also homologous to the polyprenol binding pockets of Pmt1-Pmt2 (Figure S3.4D), STT3 (Figure S3.4E), PglB (Figure S3.4F), and ArnT (Figure S3.4G), making this site the likely binding pocket for the lipidic section of DPA. In comparison with the locations of lipidic donor substrates in the other homologs, it seems that, although the general binding area is conserved, the exact identification of the helices that form the binding site are likely to differ.

There are three long loops (N-EL4 between TM7 and JM 3, C-EL4 between JM 3 and TM8, and EL5 between TM9 and TM10) that do not have clear density present and consist of 7, 17, and 27 residues, respectively (Figures 3.4C and 3.5A). Given that the structure of AftD was solved in the apo state, these unresolved loops could be analogous to those of other glycosyltransferases that span across the active site and become ordered only upon ligand binding. Indeed, superimposition with the 3D structures of Pmt1 (Figure S3.4D), STT3 (Figure S3.4E), PglB (Figure S3.4F), and ArnT (Figure S3.4G) suggests that Pmt1 LL4, STT3 EL5, PglB EL5, and ArnT PL4 loops are analogous to the AftD EL4 loop, as they all extend from analogous TM helices. In contrast, AftD EL5 seems to be unique to this enzyme (Figures S3.4D–S3.4G). A few conserved residues are present on both loops, suggesting that these segments are likely to be important for AftD's enzymatic function; in particular, W272 is 100% conserved, and D351 is moderately conserved (Figure 3.5A).

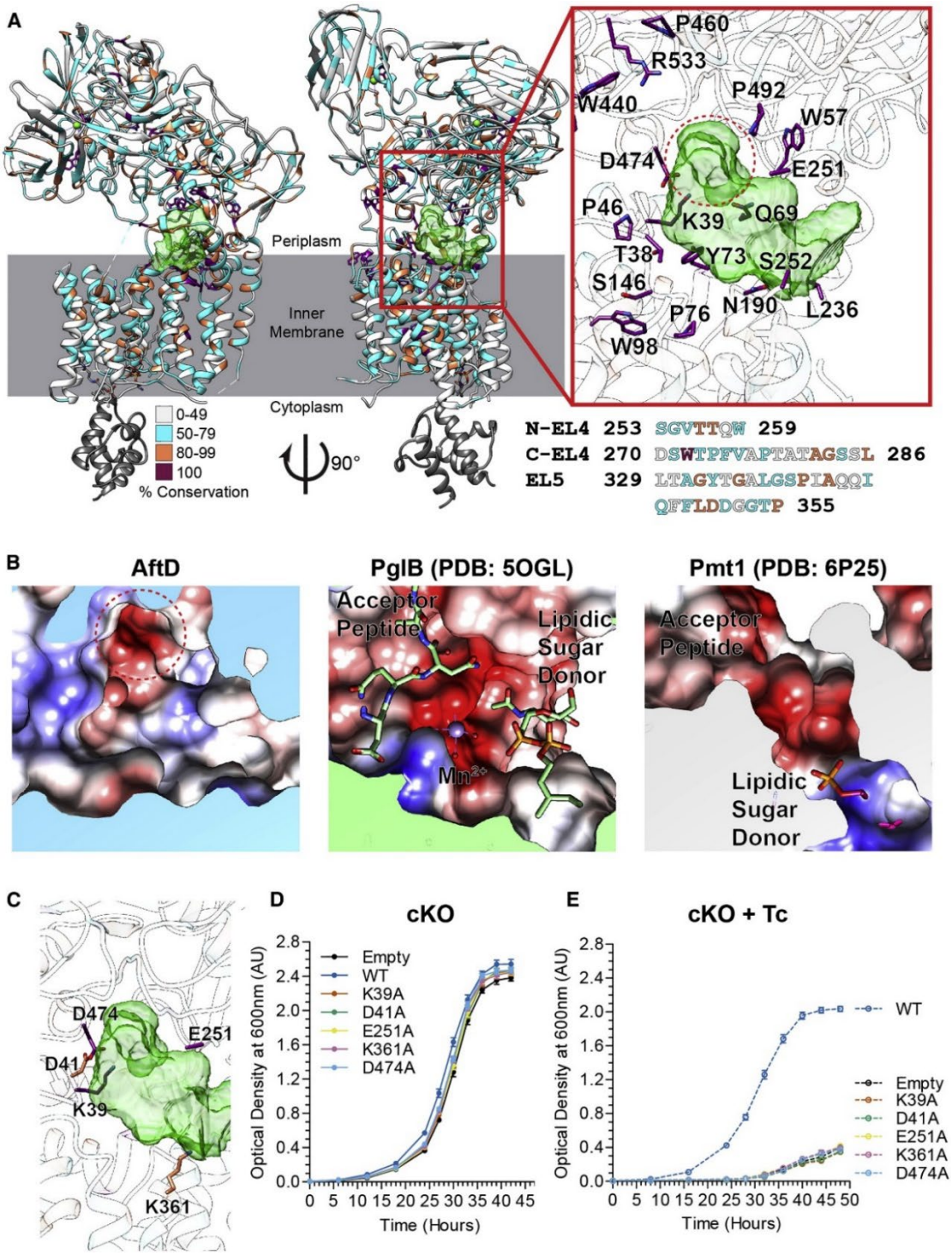


Figure 3.5 – Sequence Conservation and Putative Active Site of AftD. (A) *Structure of AftD, rendered as a cartoon and colored based on sequence conservation. The putative active site cavity, generated by the Voss Volume Voxelator server [60] using probes of 2 and 5 Å radii, is indicated in semi-transparent green. Residues with 100% conservation are indicated in magenta and have their side chains displayed. The box indicates a putative active site cavity with strictly conserved residues around it labeled. The primary sequences of the un-modeled flexible loops around the active site are indicated below the box, colored by conservation.* ***(B)*** *Columbic potential maps of the putative active site region of AftD, aligned PglB and Pmt1. The putative divalent cation binding pocket is circled with a red dotted line and is mirrored in* ***(A)****. The ligands and Mn^{2+} ion in the structure of PglB are indicated as sticks and ball, respectively.* ***(C)*** *Residues that have been mutated in the M. abscessus aftD gene have side chains displayed as sticks, and the model colored by conservation.* ***(D and E)*** *The mutant plasmids were then transformed into the cKO to assess the effect of these mutations on the growth rate of M. smegmatis, in the absence* ***(D)*** *or presence* ***(E)*** *of tetracycline.*

We designed a series of mutants based on the apo structure of AftD to examine structure-based functional hypotheses. Mutations around the putative active site were designed to target highly conserved residues that could be involved in substrate binding or catalysis (Figure 3.5C). These residues (K39, D41, E251, K361, and D474) were mutated to alanine. All mutated AftD proteins were expressed in *E. coli* and purified using non-ionic detergents to yields comparable to that of WT (Figure S3.5). All of these *aftD* mutations were also cloned in the mycobacterial expression vector to examine whether these mutated *aftD* genes restored wild-type growth of cKO *M. smegmatis* mutant with tetracycline. None of these AftD mutants were capable of complementing the growth defect of the cKO *M. smegmatis* mutant under non-permissive conditions. These results suggest that the mutated amino acid residues play essential roles for the proper functioning of AftD (Figures 3.5D and 3.5E).

4.4. CBMs of the AftD Soluble Domain Help Bind Complex Arabinofuranose Chains

In addition to the GT-C super-domain, AftD has a large soluble domain consisting of three CBMs that exhibit β-sandwich folds (Figure 3.6A). There is an unusual structural motif of four consecutive prolines (residues 863–866) on CBM2, which appears to serve a structural role by making hydrophobic interactions with six other neighboring prolines from within CBM2 and also from a region between CBM1 and the GT-C soluble domain (Figure S3.4A). The structure shows that two CBMs (CBM1 and CBM3) bind an ion that is likely calcium, based on comparison with other structural homologs (PDB: 1UYY and 4A45). Notably, CBM3 is more flexible than the other two CBMs, as evidenced by the smearing of the cryo-EM density at CBM3 before focused classification.

Structural homologs for each of the CBMs were identified using the Dali server, and the highest scoring structural homolog with a ligand is shown in Figure 3.4A to illustrate putative glycan binding sites. The putative glycan binding sites of CBM1 and CBM3 have an extensive network of polar (S, T, N, and Q) and charged (D, E, R, and K) residues that can form ion-dipole or hydrogen bonds with the carbohydrate ligand (Figure 3.4B). In addition, CBM1 and CBM3 have aromatic side chains (W757, W818 for CBM1, and F1066 for CBM3) that can facilitate hydrophobic stacking interactions with the face of carbohydrate rings, an important mechanism of binding in CBMs [83].

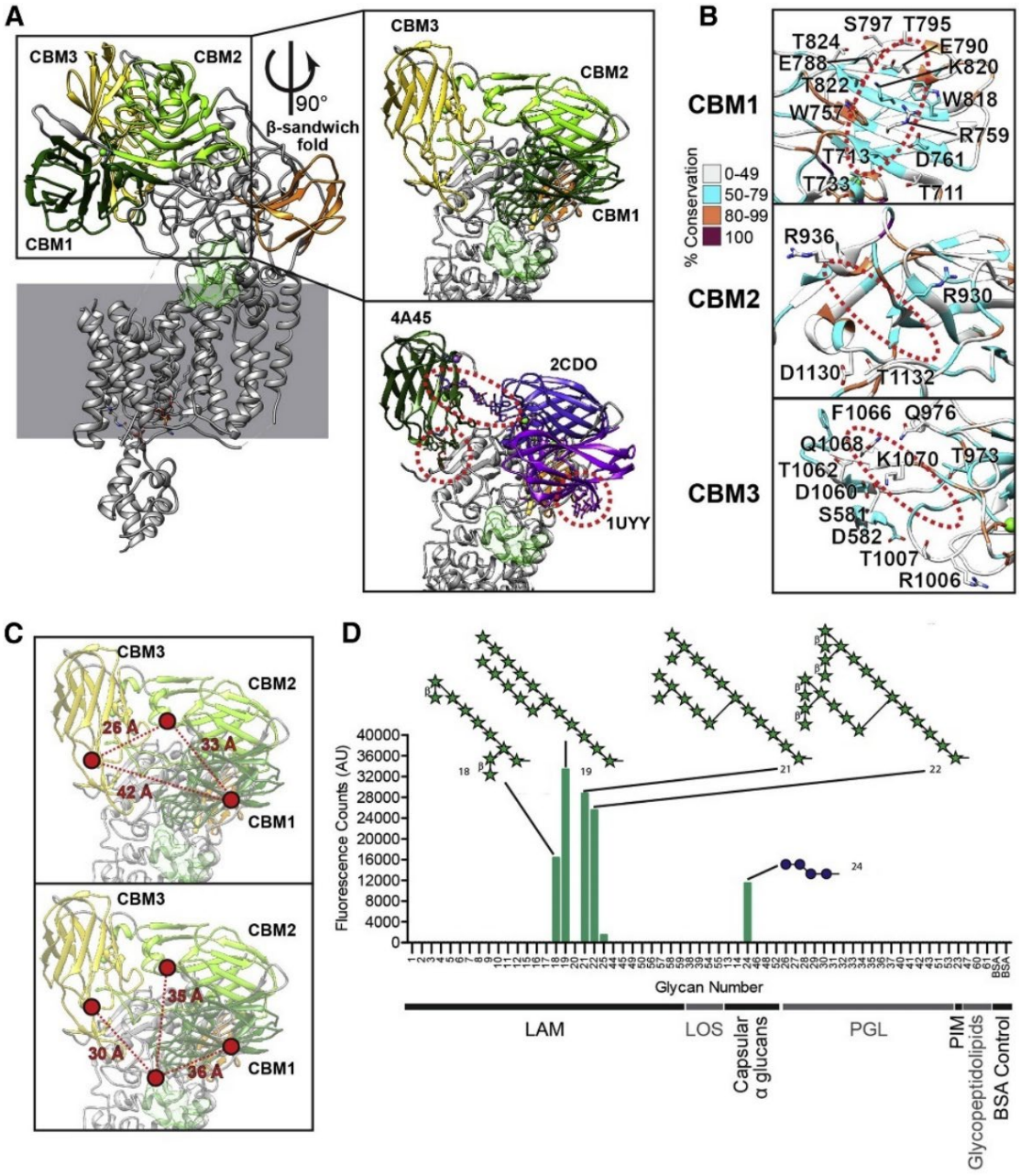


Figure 3.6 – The Carbohydrate-Binding Modules of AftD. (A) *The three soluble CBMs and a β-sandwich fold domain are colored in the AftD structure. The model is rendered as a cartoon. In the first insert on top, the CBMs are zoomed-in, and the putative active site cavity is colored in semi-transparent green, as in Figure 3.3. In the bottom insert, similar protein structures (PDB IDs are indicated) aligned to and replacing each AftD CBM to show putative glycan binding scheme.* ***(B)*** *Putative sugar binding interfaces of the CBMs, with residues that could be involved in binding labeled and the putative location of the sugars circled in red dotted lines. All residues are colored by conservation.* ***(C)*** *Distances among the CBMs (top panel) and to the putative ligand binding cavity are indicated in light green (bottom panel).* ***(D)*** *Glycan array analysis of AftD, with top glycan motif hits indicated, where the green star represents arabinofuranose and the blue sphere represents glucopyranose. The classes of glycans used are: lipoarabinomannan (LAM), trehalose mycolate and lipooligosaccharide (LOS), capsular α glucan, phenolic glycolipid (PGL), phosphatidyl-myo-inositol mannoside (PIM), glycopeptidolipid, and bovine serum albumin (BSA) control. For structures of all glycans on the array see [84].*

These findings suggest that CBM1 and CBM3 are likely CBMs. CBM2, on the other hand, has only four polar/charged and no aromatic residues at the putative binding site, indicating that it binds carbohydrates weakly or not at all.

The relative locations of these three CBMs are quite far from each other, between 26 and 42 Å apart (Figure 3.4C). All three CBMs are, in turn, 30 – 36 Å away from the edge of the putative active site, and they fan out in different directions from it. The length of one arabinofuranose residue in an extended conformation is about 3.6 Å. This arrangement of the CBMs suggests that AftD may have evolved to bind long (>10 U) and possibly nonlinear arabinan chains.

To test the aforementioned hypothesis, we probed the carbohydrate-binding ability of the protein *in vitro* by screening purified AftD against an array of mycobacterial glycan fragments [84].

AftD appears to bind preferentially to long and branched arabinofuranose chains (Figure 3.6D). Notably, the arabinofuranose glycans to which AftD binds have 10 or more arabinofuranose residues. Conversely, AftD seems not to bind to other mycobacterial glycans on the array that lack arabinofuranose residues: glycopeptidolipids, PIMs, phenolic glycolipids (PGLs), trehalose mycolates and lipooligosaccharides (LOSs), and capsular α glucans, with the exception of a glucose tetrassaccharide (Figure 3.6D). These results support both our structural hypothesis and the previously predicted function of AftD acting as a late arabinosyltransferase in the biosynthesis of AG/LAM [22] [24]. The significance of the observed binding of the glucose tetrasaccharide is, at this point, unclear.

4.5. AftD Is Complexed with ACP

At the C-terminus of the soluble domains of AftD, the protein chain re-enters into the membrane via TM12, followed by an 18-residue loop that arches over TM5, TM6, TM7, TM9, and TM10 to reach TM13 (Figures 3.4C and 3.7A). TM13 – TM16 then form a four-helix bundle that interface with TM8 and TM9, which is unique to AftD in comparison to other Aft/Emb enzymes. This four-helix bundle shows a contiguous density adjoining the cytoplasmic side, consistent with the presence of an associated protein (Figure 3.7A). Using mass spectrometry, this additional component was identified as the 10 kDa acyl carrier protein (ACP) from *E. coli*, and the ACP crystal structure (PDB: 1T8K) was readily accommodated in the density (Figure 3.7A). The density map for ACP was of lower

resolution (~4-5 Å; Figure S3.3F) than for AftD, likely due to conformational flexibility. Indeed, focused classification of the ACP portion revealed multiple low-resolution conformational states. Notably, there were no classes without density for ACP, consistent with the tight association of the two proteins during expression and purification.

The negatively charged surface of *E. coli* ACP, which has an isoelectric point around 4, interacts with a positive patch of AftD comprising the arginine residues R310, R315, and R1389 (Figures 3.7B–3.7D). *E. coli* ACP is found in its holo-form with 4'-phosphopantetheine covalently linked via a phosphodiester bond to S36 [85].

This prosthetic group can either be buried within the ACP hydrophobic tunnel-like cavity in the core of the four helices or extended toward its cognate protein [87].

The coulombic potential map shows that the 16 Å long 4'-phosphopantetheine fills the hydrophobic pocket of the four-helix bundle (TM13 – TM16) of AftD (Figures 3.7C and 3.7D). Its C-terminal thiol group points toward the interspace between TM15 and TM16. Moreover, 4'-phosphopantetheine not only forms extensive hydrophobic interactions with AftD residues L1332, V1335, and F1324 but also establishes hydrogen bonds between its O33 and the backbone carbonyl of A1386, and its phosphate group with the side chain of R315, a highly conserved residue among AftDs (Figure 3.7D).

To determine the physiological relevance of the ACP-AftD interaction, we purified *M. abscessus* AftD from a *M. smegmatis* mycobacterial expression system. Mass spectrometry analysis of the purified protein revealed the presence of *M. smegmatis* meromycolate extension ACP AcpM (KEGG: MSMEG_4326; Table S3.2), a 99-residue protein that has 39% sequence identity with *E. coli* ACP. AcpM also has a negatively charged surface (PDB: 1KLP; Figure 3.7B, box) like the *E. coli* ACP [88]. *M. smegmatis* has another ACP (KEGG: MSMEG_2132) that shares around 40% sequence identity with AcpM, and it was not detected via mass spectrometry. These results indicate that the presence of ACP in the AftD structure was not an artifact due to the overexpression system in *E. coli*.

Five mutants in the region of AftD that interacts with ACP were also generated aimed at disrupting ACP binding. These mutants were successfully purified in *E. coli* (Figure S3.5) and cloned into the mycobacterial expression vector. To investigate the functional relevance of the ACP-AftD interaction, we examined whether these mutant genes restored wild-type growth of the cKO *M. smegmatis* mutant with tetracycline.

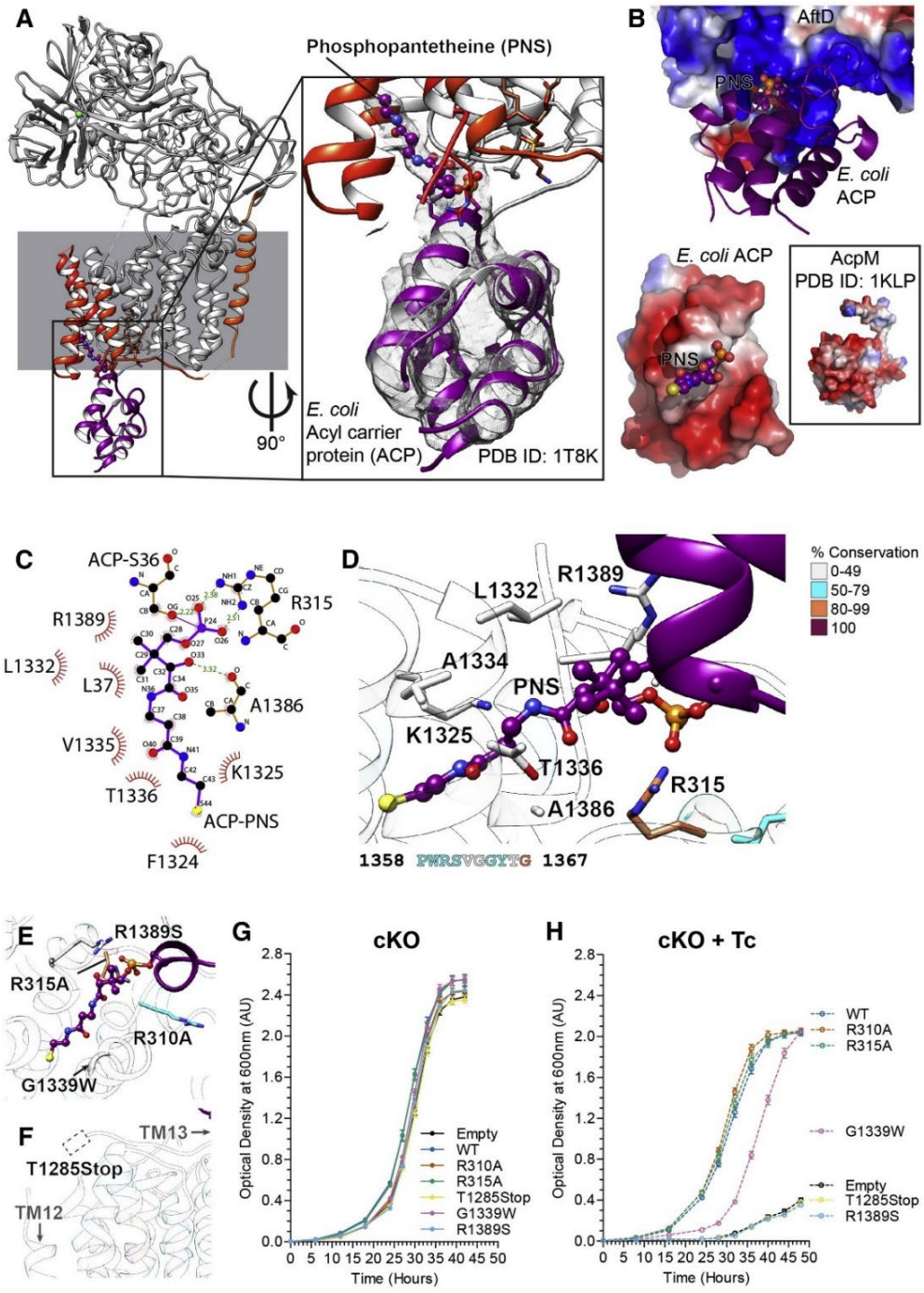


Figure 3.7 – ACP in Complex with AftD. (A) *TM13, TM14, TM15, and TM16 form a hydrophobic groove where the 4'-phosphopantetheine of ACP binds. AftD is rendered as a cartoon with TM12–TM16 being colored in the same colors as in Figure 3.2. ACP is indicated in magenta. The insert shows a zoomed view of ACP and 4'-phosphopantetheine, with the corresponding density map indicated as mesh.* ***(B)*** *Surface electrostatic potential maps of AftD (top), E. coli ACP (below), and M. smegmatis AcpM (below, insert), with red indicating negatively charged and blue indicating positively charged.* ***(C)*** *LigPlot* [86] *on 4'-phosphopantetheine showing its atomic interactions with AftD.* ***(D)*** *Zoom-in on the binding interface between AftD and ACP, with side chains of relevant residues displayed as sticks. Residues have been colored by conservation using the same scheme as in Figures 3.3 and 3.4. The primary sequence of the un-model flexible loop between TM15 and TM16 is shown below the insert, colored by conservation.* ***(E–H)*** *Residues that have been mutated in the M. abscessus AftD protein have side chains displayed as sticks, and the model colored by conservation* ***(E)****. If the residue is not resolved in the structure, its putative position is indicated by a dotted lined box* ***(F)****. These mutations are also indicated in a chart* ***(G)****. These mutant plasmids were then transformed into the cKO to assess their effect on the growth rate of M. smegmatis* ***(H)****, in the absence or presence of tetracycline.*

The insertion of a stop codon at T1285 produced a C-terminal truncation that eliminated the four-helix ACP-binding bundle altogether (Figure 3.7F). This truncated protein resulted in a growth defect akin to the empty plasmid, suggesting that either the four-helix bundle of the enzyme, *per se*, or its interaction with ACP is important for AftD function (Figures 3.7G and 3.7H). Mutations were also made to three arginine residues (R310A, R315A, and R1389S) to abolish (by mutations to alanine) or reduce (by mutations to serine) the charge interactions between the protein part of ACP and AftD (Figure 3.7E). R310A and R315A mutants fully restored growth of the cKO *M. smegmatis* in the presence of tetracycline, indicating that these residues were not important individually for proper AftD function (Figures 3.7G and 3.7H). By contrast, R1389S showed a severe growth defect, consistent with a functional role of this residue (Figures 3.7G and 3.7H). Another mutation was introduced at G1339 to add a bulky tryptophan to prevent the binding of 4'-phosphopantetheine to AftD's four-helix bundle (Figure 3.7E). This mutant only partially rescued growth of the cKO *M. smegmatis* mutant in the presence of tetracycline, suggesting that G1339 is also important for the function of AftD (Figures 3.7G and 3.7H).

To further characterize the two classes of AftD mutants that were made (targeting putative active site versus ACP binding), we performed microplate appearance (Figure 3.8A) and colony morphology assays (Figure 3.8B) on a representative mutant from each class: D474A (putative active site) and R1389S (ACP binding). Both mutants showed phenotypes similar to those of the cKO (Figures 3.2C and 3.2D) and each other, suggesting that disruption of ACP binding may be linked to loss of enzymatic function.

4.6. The Structure of the AftD-R1389S Mutant Shows Ordering of EL5 Loop with Disordering of ACP

To investigate how ACP binding might be correlated to enzyme activity, we determined the structure of the AftD-R1389S mutant using single-particle cryo-EM (Figures 3.8 and S3.6). The structure determination pipeline was similar to that described for WT AftD, with two notable exceptions: (1) tobacco etch virus (TEV) cleavage was not performed, as the mutation resulted in TEV cleavage being very inefficient (Figures S3.6A and S3.6B); and (2) AftD-R1389S was vitrified on gold grids using the Leica GP, with biochemical conditions optimized earlier with Spotiton (Figure S3.6C).

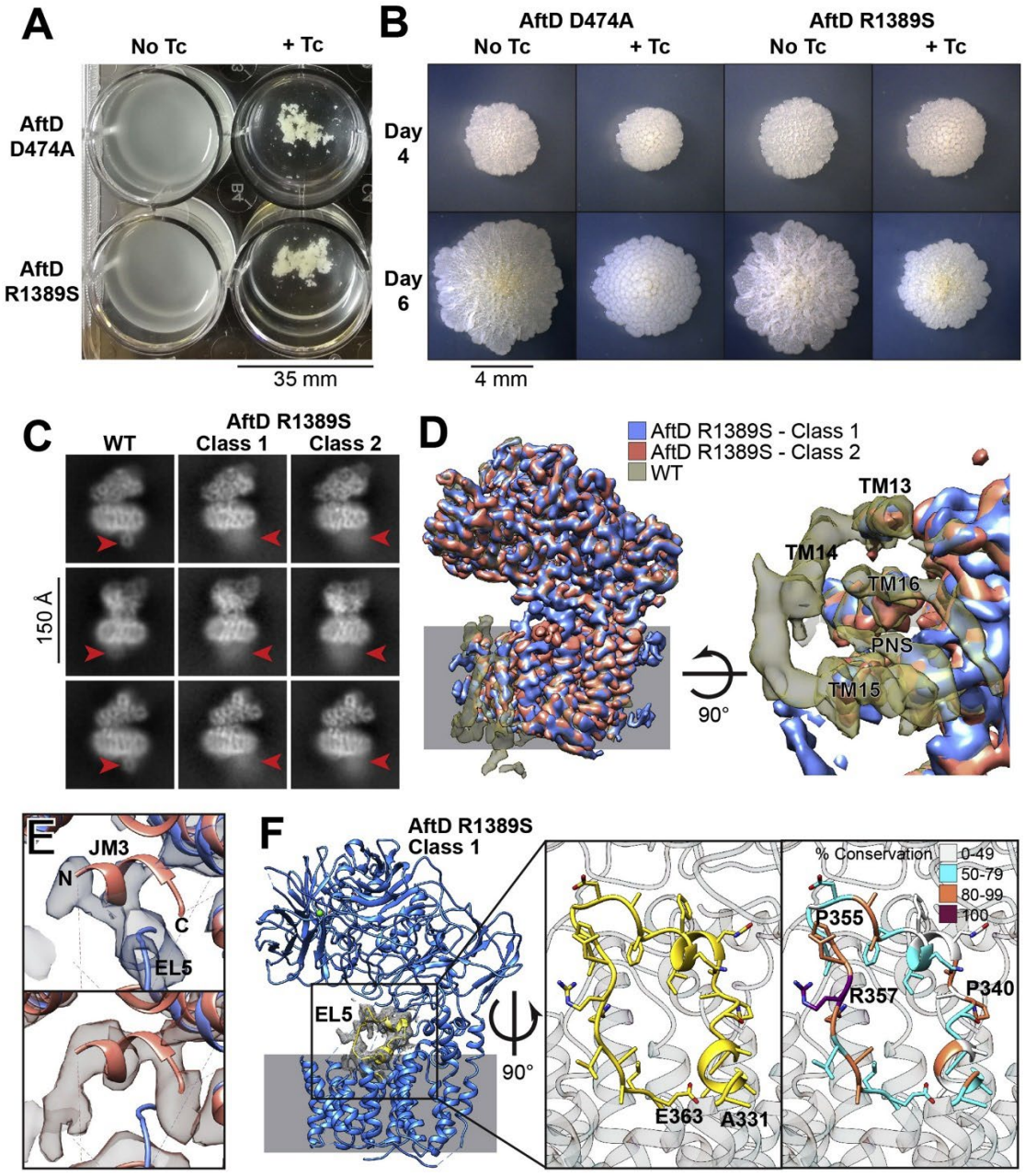


Figure 3.8 – Phenotypical analysis of aftD mutant M. smegmatis (D474A, R1389S) and structural characterization of AftD mutants D474A, R1389S. *Microplate appearance assay* ***(A)*** *and colony-forming assay* ***(B)*** *across 4, 6, and 8 days of AftD mutants D474A and R1389S.* ***(C)*** *2D class averages comparison between the AftD WT particles and mutant R1389S class 1 and 2 particles, with the red arrow pointing to the position of ACP.* ***(D)*** *Superposition of maps from classes 1 and 2 from AftD mutant R1389S with the WT map. The maps have been low-pass filtered to 4 Å and displayed at similar thresholds for comparison.* ***(E)*** *Superposition of classes 1 and 2 models from AftD-R1389S against class 1 map (top) and class 2 map (bottom) at the JM3 region. Colored scheme follows* ***(D)****.* ***(F)*** *The model of class 1 from AftD-R1389S with the ordered loop (EL5) indicated in yellow. The rest of the structure is rendered in blue. The inserts are zoomed around EL5 (residues 331 to 363), colored in yellow and by conservation.*

4,886 micrographs were collected, and 226,478 particles were initially processed following data collection on a Titan Krios equipped with a Gatan K2 Summit camera fitted with an energy filter (Figures S3.6C and S3.6D; Table S3.1).

After data processing, the consensus reconstruction with an overall resolution of 3.4 Å was obtained from 150,978 final particles (Figure S3.6E). 2D class averages of the dataset revealed a fuzzy but identifiable density for the ACP; in comparison, the WT 2D class average showed a very distinct density with sharp features for the ACP (Figure 3.8C). This suggests that the mutation did not fully abrogate ACP binding but caused it to be weakly associated with AftD-R1389S and, hence, highly flexible. This was corroborated by mass spectrometry analysis of the AftD bands from the SDS-PAGE gel (Figure S3.6B). When inspecting the consensus reconstruction of AftDR1389S, a distinct density near the active site that was not seen in the WT structure was observed (Figure S3.6E). Signal subtraction followed by focused classification was performed to resolve the heterogeneity, which produced two high-resolution maps of AftD: class 1 at 3.5 Å and class 2 at 3.4 Å (Figures 3.8D and S3.6E–S3.6G). For both these classes, 2D class averages (Figure 3.6C) had fuzzy density for the ACP, and both of them also showed worse resolution for the last four TM helixes at ~5 Å (Figure S3.6F). When the maps were low-pass filtered to 4 Å and compared at similar display thresholds, TM14 was the most disordered helix, and its density was hardly seen, compared to the other helices (Figure 3.8D). The linker between TM14 and TM15 was also poorly ordered. No density was observed where the ACP binds to the WT structure, and no density for the 4'-phosphopantetheine was apparent in either of the AftDR1389S classes (Figure 3.8D), which suggests that 4'-phosphopantetheine's interaction with AftD might not be crucial for ACP-AftD binding. Between the two classes, the most notable difference was in the putative active site region (Figure 3.8F). Here, the class 2 structure highly resembles the WT structure. However, in class 1, there appears to be a significant conformational change. First, the previously disordered EL5 loop between TM 9 and TM10 is now resolved and physically occludes the putative active site. Second, the previously disordered N-EL4 and C-EL4 loop between TM7 and TM8 still remains poorly ordered, but the JM3 now appears to be disordered with poor map density (Figure 3.8E). In comparison, JM3 is still present in class 2 (Figure 3.8E).

An atomic model was built into the resolved EL5 loop density, resulting in a U-shaped structure connecting TM9 and TM10 (Figure 3.8F). Out of the residues present in the loop, R357 is absolutely conserved. Two prolines, P340 and P355, are responsible for the sharp turns in the loop structure, and both also have a high degree of conservation (85.1% and 93.1% respectively) (Figure 3.8F).

5. Discussion

5.1. The Conditional *aftD* Mutant in *M. smegmatis* Provides a System to Examine Structure – Function Relationships of AftD

In this study, we constructed a tetracycline-inducible conditional *aftD* knockout in *M. smegmatis* and showed that AftD is essential in *M. smegmatis* as reported previously [22], in contrast to closely related corynebacteria, which produce a much shorter AftD protein (Figure S3.2A; [24]).

We observed that AftD depletion increased cell clumping and altered the colony morphology of *M. smegmatis* (Figure 3.2C and 3.2D). Both phenotypes are consistent with a change of the cell-surface properties and the function of AftD in adding to long, branched arabinan as shown in this study (Figure 3.6D). The branched arabinan chain forms an integral part of the AG polymer and of the attached outer membrane [5], which is essential in mycobacteria but not in corynebacteria [89]. Presumably, interfering with the AG synthesis disturbs the proper assembly of the outer membrane and its properties, thus explaining the observed phenotypes. AftD depletion resulted in shorter *M. smegmatis* cells in contrast to the phenotype in a previous report, which utilized a temperature-sensitive rescue plasmid to deplete AftD from *M. smegmatis* at 42 °C [22]. This discrepancy might result from the nonphysiological temperature in the latter experiment, which is known to trigger a heat shock response in mycobacteria, altering the expression of many genes [90] and protein properties. By contrast, our conditional mutant enables us, for the first time, to examine the function of AftD in mycobacteria under physiological conditions.

5.2. The Full-Length Structure of a Mycobacterial Membrane Glycosyltransferase

The GT-C membrane glycosyltransferase family [70] is underrepresented in the PDB, with currently 14 structures of just five unique proteins across 3 GT families [71–75, 77–80]. None of the previous structures is of a mycobacterial GT-C glycosyltransferase or one belonging to the GT53 family. The AftD structure presented herein is the first full-length structure of these two classifications (Figure 3.4). The high degree of structural conservation of AftD with other GT-C structures across all three domains of life indicates

a remarkable retention of the fold across evolutionary time and points to the likely existence of a common ancestral GT-C fold (Figure S3.4C). However, the structural conservation is not uniform: the TM regions are marginally better conserved than the soluble GT-C domain, an observation also predicted previously [91]. The variability of the soluble domain is likely due to variations in substrates of each GT-C: for the five GT-C members where atomic structures are available, their substrates are either lipid-linked glycans and peptides (STT3, PglB, AglB, and Pmt1-Pmt2) or lipid-linked glycans and lipid A (ArnT). AftD differs from these, as its acceptor is thought to be a complex glycan (Figure 3.6D), providing an explanation for the fact that it has three additional CBM domains attached to its soluble GT-C domain.

5.3. Arabinosyltransferase Function of AftD

Structure-based alignment with the other GT-C glycosyltransferases of known structure allowed us to locate the putative active site of AftD in a cavity between the membrane-bound TM and soluble GT-C domain (Figure 3.5). We confirmed, by mutagenesis studies, that five of the conserved residues (K39, D41, E251, K361, and D474) around the putative active site are essential for function; thus, we postulate that these are involved in either catalysis or substrate binding. For substrate binding, these residues need to interact either with the phospho-arabinofuranose moiety of DPA, the terminal arabinofuranose(s) of AG/LAM, or the divalent cation likely required for catalysis [82]. The map shows poorly resolved densities for the flexible EL4 loop around this site for apo-AftD, and we expect this region to become ordered to assist in substrate binding, as is shown to occur in PglB and ArnT [77][79]. Similarly, we suggest that N-EL4 should become ordered to trap DPA (state 2′ in Figure 3.9), being closest to the putative DPA binding site (Figures S3.4B and S3.4D–S3.4G).

On the other hand, C-EL4 is well positioned to become ordered when the growing end of the arabinan acceptor enters the active site (state 2 in Figure 3.9). In our structure, no divalent cation density around the putative active site was observed. In an analogy with PglB, we predict that substrate binding and ordering of the EL4 loop might be required to trap a catalytic cation [77].

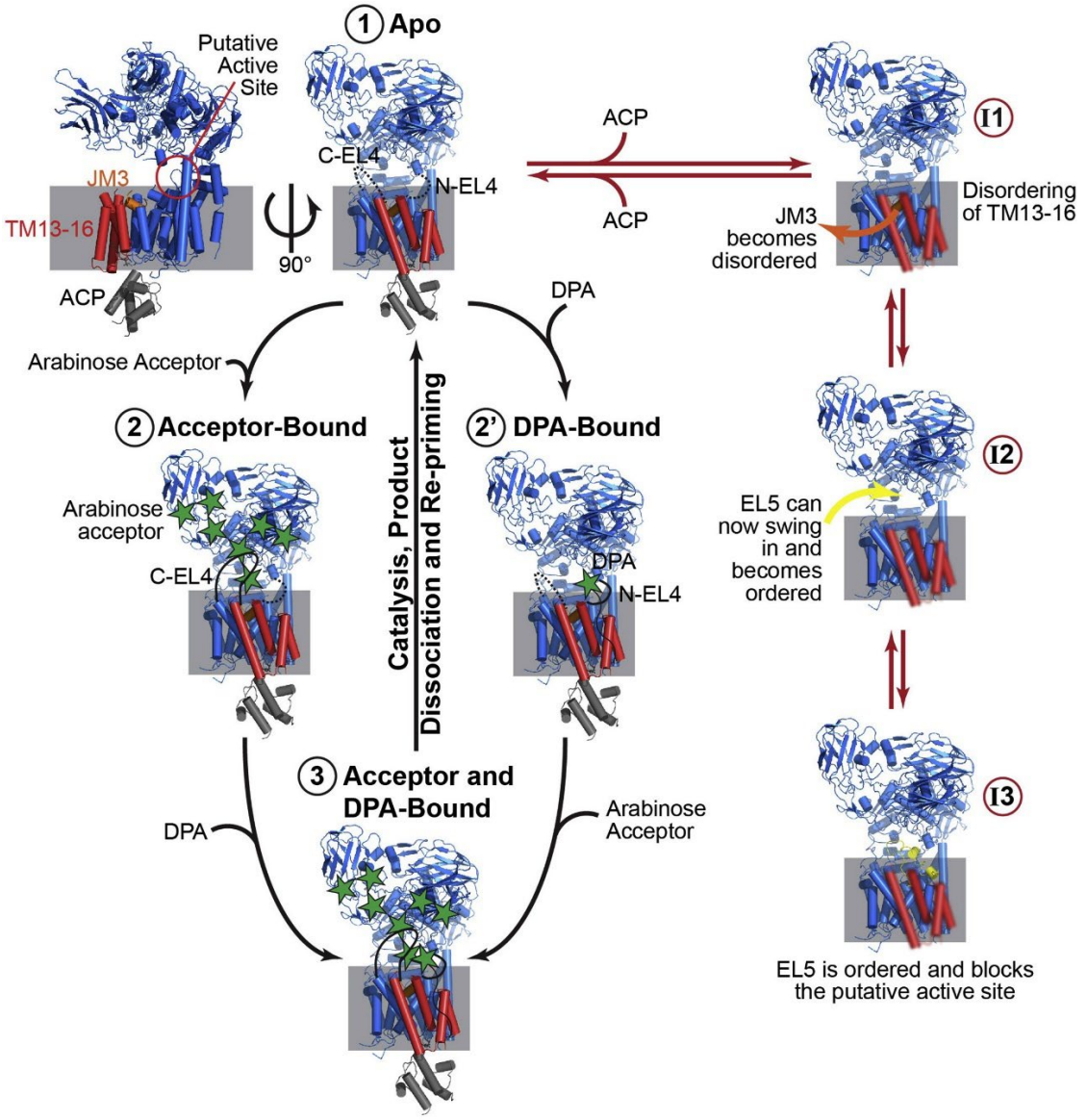


Figure 3.9 – Proposed Mechanism of Action for AftD. *AftD is rendered in blue, orange (JM3), red (TM 13-16), and yellow (ordered EL5); ACP is rendered in gray. The catalytic cycle for AftD is indicated by the black arrows from states 1 through 3. The inhibitory cycle is indicated by red arrows from states I1 to I3. The solved WT AftD structure corresponds to state 1, the AftD-R1389S class 2 structure corresponds to state I2, and the AftD-R1389S class 1 structure corresponds to state I3.*

We identified three putative CBMs in the periplasmic domain and showed the possible glycan binding interfaces and residues by comparison to other structurally homologous CBMs for which ligand-bound structures are available (Figure 3.6). The sizable distances between the AftD CBMs (26 to 42 Å; Figure 3.6C) suggests the binding of a large, complex glycan substrate, which corroborates previous work placing AftD as a late-stage glycosyltransferase that acts near the end of the synthesis pathway for the arabinans in both LAM and AG [22][24]. CBM multivalency carries two advantages: (1) modules can work together cooperatively to result in stronger binding than that for individual CBMs; and (2) each module could recognize different regions of the substrate, thus improving

binding specificity [92]. In AftD, these domains, which rise 70 Å up from the membrane (Figure 3.4A), are likely to bind specifically to different portions of AG/LAM. The glycan array result supports this hypothesis: branched arabinofuranose glycans >10 residues in length, which are too large to be bound by a single CBM, appear to show the most convincing binding profiles (Figure 3.6D). The fact that both elongating AG and LAM are restrained in one dimension, being anchored via a polyprenol-pyrophosphate in the periplasmic membrane [2] and to either the periplasmic membrane through phosphoinositol mannoside [93]–[95] or, possibly, also to the outer mycolate membrane [96], respectively, will likely significantly impact binding affinities.

Combining our results with the knowledge derived from other GT-C membrane glycosyltransferases [77][79] studied both structurally and functionally, we propose the following putative enzymatic cycle (Figure 3.9). The characterized wild type AftD structure represents the apo state, bearing disordered N-EL4 and C-EL4 regions (state 1 in Figure 3.9). Binding of the substrates will result in the ordering of N-EL4 and C-EL4, respectively, stabilization of the substrate-enzyme complex (state 2 or 2′ in Figure 3.9) and priming of the system for catalysis to occur (state 3 in Figure 3.9). Thereafter, the product is released, and the active site is re-primed. This would reset AftD back to the apo-state, readying it for another catalytic cycle.

5.4. Role of the ACP in AftD

In the structure of AftD, an *E. coli* ACP was so tightly bound to the cytoplasmic face of the enzyme at a C-terminal four-helix bundle that it remained associated during purification (Figure 3.7A). The tight binding of the ACP is quite remarkable, because complexes of ACP with partner proteins have typically been difficult to observe due to the transient and weak nature of their interactions [87]. Indeed, some of these complexes could only be obtained via crosslinking experiments [97]. Neither the ACP nor the four-helix bundle it associates with (AftD TMs 11-14) is present in other GT-C glycosyltransferases, for which structural information is available. This interaction was found to be: (1) physiological, as AftD purified from *M. smegmatis* also remained associated with the *M. smegmatis* ACP homolog AcpM (Table S3.2); and (2) likely essential for function, as AftD variants genetically engineered to disrupt AcpM binding did not rescue the growth defect of AftD-depleted *M. smegmatis* (Figure 3.7H).

Structurally, mycobacterial ACP is very similar to its *E. coli* counterpart, with helix II (the recognition helix) being the most conserved [88]. The only notable difference is the

presence of a ~35-residue extension at the C-terminus that is highly flexible and is thought to interact with long-chain intermediates of mycolic biosynthesis or mediate specific protein-protein interactions (Figure 3.7B) [98]. The highly conserved helix II contains serine 36, which is located within a signature motif of DSL/X-hydrophobic residue and is modified by a 4'-phosphopantetheine. The high degree of conservation of ACP across species is the likely reason why we isolated an ACP-AftD complex from a heterologous expression system. In contrast to *E. coli*, the genome of *M. smegmatis* encodes two different ACPs. The AcpM variant is an essential protein [66][99][100] that is involved in both fatty acid synthesis [101] and mycolic acid synthesis [102].

As no further electron density was observed beyond the terminal thiol group of the 4'-phosphopantetheine, it is not clear whether the ACP has no cargo bound, a disordered cargo, or a variety of different cargos in different conformations that average out during cryo-EM analysis. In AftD, the ACP could be serving as a shuttle for starting materials and intermediates throughout the fatty acid biosynthetic pathway. In *E. coli*, ACP interacts with at least 12 enzymes involved in fatty acid biosynthesis, as well as with seven other enzymes from disparate biosynthetic pathways [103]. In mycobacteria, the ACP also functions as a shuttle for mycolic acid biosynthesis [104]. Intriguingly, ACP has been reported to be essential for an *E. coli* enzyme catalyzing the synthesis of β-(1→2)-linked glucan in membrane-derived oligosaccharides from UDP-glucose, although the 4'-phosphopantetheine moiety was not needed [105].

The elucidation of the structures of mutant AftD-R1389S (Figures 3.8 and S3.6) pointed to another hypothesis for the presence of the ACP – as a possible regulator of AftD enzyme activity (Figure 3.9). ACP regulation of catalytic activity has, indeed, been observed in other proteins such as WaaP [106] and, possibly, SpoT-like proteins [107]. When the 4'-phosphopantetheine is removed from the four-helix bundle, the helices become more disordered, as evidenced by the lower resolution, especially for TM14 (Figure 3.8D; state I1 in Figure 3.9). This, in turn, results in JM3 moving away and the EL5 loop becoming ordered (Figure 3.8E; state I2 in Figure 3.9). The ordered EL5 now appears to block the putative active site, possibly preventing substrates from entering and, thus, effectively halting enzymatic function (Figure 3.8F; state I3 in Figure 3.9). Out of the two solved dominant structures of AftD-R1389S, class 2 may represent an intermediate state in this conformational change as the four-helix bundle is disordered but JM3 is still present. Class 1, on the other hand, likely represents the final state of this postulated conformational change, with EL5 ordered and JM3 displaced. The fact that a mutation of the enzyme active site (AftD-D474A) results in *M. smegmatis* that looks phenotypically similar to that of AftD-R1389S is consistent with the observation that both

these mutations could abrogate the same function of AftD – presumably, its glycosyltransferase activity (Figures 3.8A and 3.8B). As the ACP binds the cytoplasmic face of AftD, this could enable the transmission of intracellular signals across the membrane. Given that AftD is the crucial AraT that is expected to add the final α-linked arabinofuranose sugars on AG [24], inactivating this enzyme could prevent the attachment of mycolic acids and block the proper assembly of the entire outer membrane [3]. It is possible that a regulation of AftD by AcpM could link the biosynthesis of AG to the availability of sufficient mycolic acids. Indeed, it has been recently shown that there is a physical interaction between AftD and mycolic acid transporter MmpL3 [108].

5.5. AftD as a Potential Drug Target

Although there are currently no drugs targeting AftD, it is an attractive drug target for three reasons. First, it is an essential gene in mycobacteria and is accessible to drugs on the inner membrane (Figure 3.1A) [109]. Second, AftD is conserved across many other bacteria of the *Actinobacteria* class, which includes other pathogens that cause diseases such as diphtheria (*Corynebacterium diphtheria*), leprosy (*M. leprae* or *M. lepromatosis*), and nocardiosis (various *Nocardia* species). Third, the fact that AftD has two potential sites at which to target drugs (the active site and the ACP binding site) and both appear to be essential for function further contributes to the potential of this protein for structure-based drug design efforts.

6. Conclusions

Here, we present the first full-length mycobacterial GT-C structure, determined by single-particle cryo-EM to 2.9 Å resolution. The structure of wild type AftD reveals a conserved GT-C fold and also pinpoints the location of the putative active site. The residues that are likely involved in substrate binding and catalysis were probed through mutagenesis studies using an engineered *aftD* cKO strain in *M. smegmatis*. Furthermore, the structure and glycan array analysis show that AftD binds specifically to large, highly branched arabinofuranose residues that are found in AG/LAM. Finally, to the best of our knowledge, this is the first report of a glycosyltransferase structure with a tightly associated ACP. This interaction, when disrupted using a specific single mutant (AftD-R1389S), resulted in conformational changes in AftD, as observed in the structures of this mutant. These structural changes regulate the enzymatic activity of AftD and might link AG to mycolic acid synthesis. Such a mechanism, still speculative at this stage, would establish AftD as a key enzyme in the center of two major pathways for cell wall biosynthesis in mycobacteria and other mycolic-acid-producing bacteria in the order *Corynebacteriales*.

7. References

[1] World Health Organization, “Annual Report of Tuberculosis” *Annu. Glob. TB Rep. WHO*, pp. 1–68, 2022

[2] L. J. Alderwick, J. Harrison, G. S. Lloyd, and H. L. Birch, “The Mycobacterial Cell Wall--Peptidoglycan and Arabinogalactan” *Cold Spring Harb. Perspect. Med.*, vol. 5, no. 8, pp. 1–16, Aug. 2015

[3] A. E. Grzegorzewicz *et al.*, “Assembling of the *Mycobacterium tuberculosis* Cell Wall Core” *J. Biol. Chem.*, vol. 291, no. 36, pp. 18867–18879, Sep. 2016

[4] E. Layre, R. Al-Mubarak, J. T. Belisle, and D. B. Moody, “Mycobacterial Lipidomics” *Mol. Genet. Mycobact.*, pp. 341–360, Oct. 2015

[5] C. Hoffmann, A. Leis, M. Niederweis, J. M. Plitzko, and H. Engelhardt, “Disclosure of the mycobacterial outer membrane: cryo-electron tomography and vitreous sections reveal the lipid bilayer structure” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 10, pp. 3963–3967, Mar. 2008

[6] H. Nikaido and V. Jarlier, “Permeability of the mycobacterial cell wall” *Res. Microbiol.*, vol. 142, no. 4, pp. 437–443, 1991

[7] C. E. Barry, “Interpreting cell wall ‘virulence factors’ of *Mycobacterium tuberculosis*” *Trends Microbiol.*, vol. 9, no. 5, pp. 237–241, May 2001

[8] M. Jankute, J. A. G. Cox, J. Harrison, and G. S. Besra, “Assembly of the Mycobacterial Cell Wall” *Annu. Rev. Microbiol.*, vol. 69, no. 1, pp. 405–423, Oct. 2015

[9] K. A. Abrahams and G. S. Besra, “Mycobacterial cell wall biosynthesis: a multifaceted antibiotic target” *Parasitology*, vol. 145, no. 2, pp. 116–133, Feb. 2018

[10] B. CE, C. DC, and M. MR, “Targeting the formation of the cell wall core of *M. tuberculosis*” *Infect. Disord. Drug Targets*, vol. 7, no. 2, pp. 182–202, Apr. 2007

[11] M. M. Islam *et al.*, “Drug resistance mechanisms and novel drug targets for tuberculosis therapy” *J. Genet. Genomics*, vol. 44, no. 1, pp. 21–37, Jan. 2017

[12] J. A. Campbell, G. J. Davies, V. Bulone, and B. Henrissat, “A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence

similarities” *Biochem. J.*, vol. 326 (Pt 3, no. Pt 3, pp. 929–939, 1997

[13] P. M. Coutinho, E. Deleury, G. J. Davies, and B. Henrissat, “An evolving hierarchical family classification for glycosyltransferases” *J. Mol. Biol.*, vol. 328, no. 2, pp. 307–317, Apr. 2003

[14] B. A. Wolucka, M. R. McNeil, E. De Hoffmann, T. Chojnacki, and P. J. Brennan, “Recognition of the lipid intermediate for arabinogalactan/arabinomannan biosynthesis and its relation to the mode of action of ethambutol on mycobacteria.” *J. Biol. Chem.*, vol. 269, no. 37, pp. 23328–23335, Sep. 1994

[15] A. E. Belanger *et al.*, “The embAB genes of *Mycobacterium avium* encode an arabinosyl transferase involved in cell wall arabinan biosynthesis that is the target for the antimycobacterial drug ethambutol” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 93, no. 21, pp. 11919–11924, Oct. 1996

[16] A. Telenti *et al.*, “The emb operon, a gene cluster of *Mycobacterium tuberculosis* involved in resistance to ethambutol” *Nat. Med.*, vol. 3, no. 5, pp. 567–570, 1997

[17] R. Goude, A. G. Amin, D. Chatterjee, and T. Parish, “The arabinosyltransferase EmbC is inhibited by ethambutol in *Mycobacterium tuberculosis*” *Antimicrob. Agents Chemother.*, vol. 53, no. 10, pp. 4138–4146, Oct. 2009

[18] K. Takayama and J. O. Kilburn, “Inhibition of synthesis of arabinogalactan by ethambutol in *Mycobacterium smegmatis*” *Antimicrob. Agents Chemother.*, vol. 33, no. 9, pp. 1493–1499, 1989

[19] L. J. Alderwick, M. Seidel, H. Sahm, G. S. Besra, and L. Eggeling, “Identification of a novel arabinofuranosyltransferase (AftA) involved in cell wall arabinan biosynthesis in *Mycobacterium tuberculosis*” *J. Biol. Chem.*, vol. 281, no. 23, pp. 15653–15661, Jun. 2006

[20] M. Seidel, L. J. Alderwick, H. L. Birch, H. Sahm, L. Eggeling, and G. S. Besra, “Identification of a novel arabinofuranosyltransferase AftB involved in a terminal step of cell wall arabinan biosynthesis in Corynebacterianeae, such as *Corynebacterium glutamicum* and *Mycobacterium tuberculosis*” *J. Biol. Chem.*, vol. 282, no. 20, pp. 14729–14740, May 2007

[21] H. L. Birch *et al.*, “Biosynthesis of mycobacterial arabinogalactan: identification of a novel alpha(1-->3) arabinofuranosyltransferase” *Mol. Microbiol.*, vol. 69, no. 5, pp. 1191–1206, Sep. 2008

[22] H. Škovierová *et al.*, “AftD, a novel essential arabinofuranosyltransferase from

mycobacteria” *Glycobiology*, vol. 19, no. 11, pp. 1235–1247, 2009

[23] S. K. Angala *et al.*, “Identification of a Novel Mycobacterial Arabinosyltransferase Activity Which Adds an Arabinosyl Residue to α-d-Mannosyl Residues” *ACS Chem. Biol.*, vol. 11, no. 6, pp. 1518–1524, Jun. 2016

[24] L. J. Alderwick, H. L. Birch, K. Krumbach, M. Bott, L. Eggeling, and G. S. Besra, “AftD functions as an α1 → 5 arabinofuranosyltransferase involved in the biosynthesis of the mycobacterial cell wall core” *Cell Surf. (Amsterdam, Netherlands)*, vol. 1, pp. 2–14, Mar. 2018

[25] S. F. Altschul and E. V. Koonin, “Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases” *Trends Biochem. Sci.*, vol. 23, no. 11, pp. 444–447, Nov. 1998

[26] A. Krogh, B. Larsson, G. Von Heijne, and E. L. L. Sonnhammer, “Predicting transmembrane protein topology with a hidden markov model: application to complete genomes” *J. Mol. Biol.*, vol. 305, no. 3, pp. 567–580, Jan. 2001

[27] W. Li and A. Godzik, “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences” *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006

[28] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput” *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, Mar. 2004

[29] J. Love *et al.*, “The New York Consortium on Membrane Protein Structure (NYCOMPS): A high-throughput platform for structural genomics of integral membrane proteins” *J. Struct. Funct. Genomics*, vol. 11, no. 3, pp. 191–199, Sep. 2010

[30] R. Bruni and B. Kloss, “High-Throughput Cloning and Expression of Integral Membrane Proteins in *Escherichia coli*” *Curr. Protoc. Protein Sci.*, vol. 74, no. 1, pp. 29.6.1-29.6.34, Nov. 2013

[31] T. H. Bayburt and S. G. Sligar, “Membrane protein assembly into Nanodiscs” *FEBS Lett.*, vol. 584, no. 9, pp. 1721–1727, May 2010

[32] R. B. Kapust, J. Toözseór, T. D. Copeland, and D. S. Waugh, “The P1′ specificity of tobacco etch virus protease” *Biochem. Biophys. Res. Commun.*, vol. 294, no. 5, pp. 949–955, Jun. 2002

[33] V. P. Dandey *et al.*, “Spotiton: New features and applications” *J. Struct. Biol.*, vol.

202, no. 2, pp. 161–169, May 2018

[34] T. Jain, P. Sheehan, J. Crum, B. Carragher, and C. S. Potter, “Spotiton: A prototype for an integrated inkjet dispense and vitrification system for cryo-TEM” *J. Struct. Biol.*, vol. 179, no. 1, pp. 68–75, Jul. 2012

[35] I. Razinkov *et al.*, “A new method for vitrifying samples for cryoEM” *J. Struct. Biol.*, vol. 195, no. 2, pp. 190–198, Aug. 2016

[36] H. Wei *et al.*, “Optimizing ‘self-wicking’ nanowire grids” *J. Struct. Biol.*, vol. 202, no. 2, pp. 170–174, May 2018

[37] C. Suloway *et al.*, “Automated molecular microscopy: The new Leginon system” *J. Struct. Biol.*, vol. 151, no. 1, pp. 41–60, Jul. 2005

[38] A. Cheng *et al.*, “High resolution single particle cryo-electron microscopy using beam-image shift” *J. Struct. Biol.*, vol. 204, no. 2, pp. 270–275, Nov. 2018

[39] S. Q. Zheng, E. Palovcak, J. P. Armache, K. A. Verba, Y. Cheng, and D. A. Agard, “MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy” *Nat. Methods 2017 144*, vol. 14, no. 4, pp. 331–332, Feb. 2017

[40] G. C. Lander *et al.*, “Appion: An integrated, database-driven pipeline to facilitate EM image processing” *J. Struct. Biol.*, vol. 166, no. 1, pp. 95–102, Apr. 2009

[41] A. Rohou and N. Grigorieff, “CTFFIND4: Fast and accurate defocus estimation from electron micrographs” *J. Struct. Biol.*, vol. 192, no. 2, pp. 216–221, Nov. 2015

[42] K. Zhang, “Gctf: Real-time CTF determination and correction” *J. Struct. Biol.*, vol. 193, no. 1, pp. 1–12, Jan. 2016

[43] D. Kimanius, B. O. Forsberg, S. H. W. Scheres, and E. Lindahl, “Accelerated cryo-EM structure determination with parallelisation using GPUS in RELION-2” *Elife*, vol. 5, Nov. 2016

[44] S. H. W. Scheres, “RELION: Implementation of a Bayesian approach to cryo-EM structure determination” *J. Struct. Biol.*, vol. 180, no. 3, pp. 519–530, Dec. 2012

[45] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker, “cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination” *Nat. Methods 2017 143*, vol. 14, no. 3, pp. 290–296, Feb. 2017

[46] A. Punjani, H. Zhang, and D. J. Fleet, “Non-uniform refinement: adaptive

regularization improves single-particle cryo-EM reconstruction" *Nat. Methods 2020 1712*, vol. 17, no. 12, pp. 1214–1221, Nov. 2020

[47] T. Grant, A. Rohou, and N. Grigorieff, "CisTEM, user-friendly software for single-particle image processing" *Elife*, vol. 7, Mar. 2018

[48] P. D. Adams *et al.*, "PHENIX: a comprehensive Python-based system for macromolecular structure solution" *Acta Crystallogr. Sect. D*, vol. 66, no. 2, pp. 213–221, Jan. 2010

[49] P. V. Afonine *et al.*, "Real-space refinement in PHENIX for cryo-EM and crystallography" *Acta Crystallogr. Sect. D*, vol. 74, no. 6, pp. 531–544, May 2018

[50] E. F. Pettersen *et al.*, "UCSF Chimera—A visualization system for exploratory research and analysis" *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004

[51] K. Cowtan and IUCr, "The Buccaneer software for automated model building. 1. Tracing protein chains" *urn:issn:0907-4449*, vol. 62, no. 9, pp. 1002–1011, Aug. 2006

[52] T. Burnley, C. M. Palmer, and M. Winn, "Recent developments in the CCP-EM software suite" *Acta Crystallogr. Sect. D*, vol. 73, no. 6, pp. 469–477, May 2017

[53] P. Emsley and K. Cowtan, "Coot: model-building tools for molecular graphics", vol. 60, no. 12, pp. 2126–2132, Nov. 2004

[54] V. B. Chen *et al.*, "MolProbity: all-atom structure validation for macromolecular crystallography", vol. 66, no. 1, pp. 12–21, Dec. 2009

[55] B. A. Barad *et al.*, "EMRinger: side chain–directed model and map validation for 3D cryo-electron microscopy" *Nat. Methods 2015 1210*, vol. 12, no. 10, pp. 943–946, Aug. 2015

[56] A. Kucukelbir, F. J. Sigworth, and H. D. Tagare, "Quantifying the local resolution of cryo-EM density maps" *Nat. Methods 2013 111*, vol. 11, no. 1, pp. 63–65, Nov. 2013

[57] Y. Zi Tan *et al.*, "Addressing preferred specimen orientation in single-particle cryo-EM through tilting" *Nat. Methods 2017 148*, vol. 14, no. 8, pp. 793–796, Jul. 2017

[58] P. R. Baldwin and D. Lyumkis, "Non-uniformity of projection distributions attenuates resolution in Cryo-EM" *Prog. Biophys. Mol. Biol.*, vol. 150, pp. 160–183, Jan. 2020

[59] S. J. Ludtke, P. R. Baldwin, and W. Chiu, “EMAN: Semiautomated Software for High-Resolution Single-Particle Reconstructions” *J. Struct. Biol.*, vol. 128, no. 1, pp. 82–97, Dec. 1999

[60] N. R. Voss and M. Gerstein, “3V: cavity, channel and cleft volume calculator and extractor” *Nucleic Acids Res.*, vol. 38, no. suppl_2, pp. W555–W562, Jul. 2010

[61] L. Holm and L. M. Laakso, “Dali server update” *Nucleic Acids Res.*, vol. 44, no. W1, pp. W351–W355, Jul. 2016

[62] M. J. Catalão, F. Gil, J. Moniz-Pereira, and M. Pimentel, “The mycobacteriophage Ms6 encodes a chaperone-like protein involved in the endolysin delivery to the peptidoglycan” *Mol. Microbiol.*, vol. 77, no. 3, pp. 672–686, Aug. 2010

[63] P. B. Rosenthal and R. Henderson, “Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy” *J. Mol. Biol.*, vol. 333, no. 4, pp. 721–745, Oct. 2003

[64] W. Ludwig *et al.*, “Road map of the phylum Actinobacteria” *Bergey’s Man. Syst. Archaea Bact.*, pp. 1–37, Sep. 2015

[65] K. C. Rahlwes, I. L. Sparks, and Y. S. Morita, “Cell Walls and Membranes of Actinobacteria” *Subcell. Biochem.*, vol. 92, pp. 417–469, 2019

[66] M. A. Dejesus *et al.*, “Comprehensive essentiality analysis of the M*ycobacterium tuberculosis* genome via saturating transposon mutagenesis” *MBio*, vol. 8, no. 1, Jan. 2017

[67] J. H. Kim *et al.*, “A genetic strategy to identify targets for the development of drugs that prevent bacterial persistence” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 47, pp. 19095–19100, Nov. 2013

[68] F. Mancia and J. Love, “High-throughput expression and purification of membrane proteins” *J. Struct. Biol.*, vol. 172, no. 1, pp. 85–93, Oct. 2010

[69] X. C. Bai, E. Rajendra, G. Yang, Y. Shi, and S. H. W. Scheres, “Sampling the conformational space of the catalytic subunit of human γ-secretase” *Elife*, vol. 4, no. December 2015, Dec. 2015

[70] J. Liu and A. Mushegian, “Three monophyletic superfamilies account for the majority of the known glycosyltransferases” *Protein Sci.*, vol. 12, no. 7, pp. 1418–1431, Jul. 2003

[71] L. Bai, T. Wang, G. Zhao, A. Kovach, and H. Li, “The atomic structure of a

eukaryotic oligosaccharyltransferase complex" *Nat. 2018 5557696*, vol. 555, no. 7696, pp. 328–333, Jan. 2018

[72] R. Wild, J. Kowal, J. Eyring, E. M. Ngwa, M. Aebi, and K. P. Locher, "Structure of the yeast oligosaccharyltransferase complex gives insight into eukaryotic N-glycosylation" *Science (80-.).*, vol. 359, no. 6375, pp. 545–550, Feb. 2018

[73] A. S. Ramírez, J. Kowal, and K. P. Locher, "Cryo-electron microscopy structures of human oligosaccharyltransferase complexes OST-A and OST-B" *Science*, vol. 366, no. 6471, pp. 1372–1375, Dec. 2019

[74] S. Matsumoto, A. Shimada, J. Nyirenda, M. Igura, Y. Kawano, and D. Kohda, "Crystal structures of an archaeal oligosaccharyltransferase provide insights into the catalytic cycle of N-linked protein glycosylation" *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 44, pp. 17868–17873, Oct. 2013

[75] S. Matsumoto, Y. Taguchi, A. Shimada, M. Igura, and D. Kohda, "Tethering an N-glycosylation sequon-containing peptide creates a catalytically competent oligosaccharyltransferase complex" *Biochemistry*, vol. 56, no. 4, pp. 602–611, Jan. 2017

[76] C. Lizak, S. Gerber, S. Numao, M. Aebi, and K. P. Locher, "X-ray structure of a bacterial oligosaccharyltransferase" *Nat. 2011 4747351*, vol. 474, no. 7351, pp. 350–355, Jun. 2011

[77] M. Napiórkowska *et al.*, "Molecular basis of lipid-linked oligosaccharide recognition and processing by bacterial oligosaccharyltransferase" *Nat. Struct. Mol. Biol. 2017 2412*, vol. 24, no. 12, pp. 1100–1106, Oct. 2017

[78] M. Napiórkowska, J. Boilevin, T. Darbre, J. L. Reymond, and K. P. Locher, "Structure of bacterial oligosaccharyltransferase PglB bound to a reactive LLO and an inhibitory peptide" *Sci. Reports*, vol. 8, no. 1, pp. 1–9, Nov. 2018

[79] V. I. Petrou *et al.*, "Structural biology: Structures of aminoarabinose transferase ArnT suggest a molecular basis for lipid A glycosylation" *Science*, vol. 351, no. 6273, pp. 608–612, Feb. 2016

[80] L. Bai, A. Kovach, Q. You, A. Kenny, and H. Li, "Structure of the eukaryotic protein O-mannosyltransferase Pmt1−Pmt2 complex" *Nat. Struct. Mol. Biol.*, vol. 26, no. 8, pp. 704–711, Jul. 2019

[81] L. J. Alderwick, H. L. Birch, A. K. Mishra, L. Eggeling, and G. S. Besra, "Structure, function and biosynthesis of the *Mycobacterium tuberculosis* cell wall:

arabinogalactan and lipoarabinomannan assembly with a view to discovering new drug targets" *Biochem. Soc. Trans.*, vol. 35, no. 5, pp. 1325–1328, Nov. 2007

[82] C. B. Sharma, L. Lehle, and W. Tanner, "N-Glycosylation of Yeast Proteins" *Eur. J. Biochem.*, vol. 116, no. 1, pp. 101–108, May 1981

[83] A. B. Boraston, D. N. Bolam, H. J. Gilbert, and G. J. Davies, "Carbohydrate-binding modules: fine-tuning polysaccharide recognition" *Biochem. J.*, vol. 382, no. 3, pp. 769–781, Sep. 2004

[84] R. B. Zheng *et al.*, "Insights into Interactions of Mycobacteria with the Host Innate Immune System from a Novel Array of Synthetic Mycobacterial Glycans" *ACS Chem. Biol.*, vol. 12, no. 12, pp. 2990–3002, Dec. 2017

[85] P. W. Majerus, A. W. Alberts, and P. R. Vagelos, "Acyl Carrier Protein. IV. The Identification of 4'-phosphopantetheine" *Proc. Natl. Acad. Sci. United States*, vol. 53, no. 2, pp. 410–417, Feb. 1965

[86] A. C. Wallace, R. A. Laskowski, and J. M. Thornton, "LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions" *Protein Eng. Des. Sel.*, vol. 8, no. 2, pp. 127–134, Feb. 1995

[87] J. E. Cronan, "The chain-flipping mechanism of ACP (acyl carrier protein)-dependent enzymes appears universal" *Biochem. J.*, vol. 460, no. 2, pp. 157–163, Jun. 2014

[88] H. C. Wong, G. Liu, Y. M. Zhang, C. O. Rock, and J. Zheng, "The solution structure of acyl carrier protein from *Mycobacterium tuberculosis*" *J. Biol. Chem.*, vol. 277, no. 18, pp. 15874–15880, May 2002

[89] D. Portevin *et al.*, "A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms" *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 1, pp. 314–319, Jan. 2004

[90] G. R. Stewart *et al.*, "The Heat Shock Response of *Mycobacterium tuberculosis*: Linking Gene Expression, Immunology and Pathogenesis" *Int. J. Genomics*, vol. 3, no. 4, pp. 348–351, Aug. 2002

[91] L. L. Lairson, B. Henrissat, G. J. Davies, and S. G. Withers, "Glycosyltransferases: Structures, Functions, and Mechanisms", vol. 77, pp. 521–555, Jun. 2008

[92] A. B. Boraston, E. Kwan, P. Chiu, R. A. J. Warren, and D. G. Kilburn, "Recognition and hydrolysis of noncrystalline cellulose" *J. Biol. Chem.*, vol. 278, no. 8, pp.

6120–6127, Feb. 2003

[93] J. Korduláková *et al.*, “Definition of the First Mannosylation Step in Phosphatidylinositol Mannoside Synthesis” *J. Biol. Chem.*, vol. 277, no. 35, pp. 31335–31344, Aug. 2002

[94] D. J. Lea-Smith *et al.*, “Analysis of a new mannosyltransferase required for the synthesis of phosphatidylinositol mannosides and lipoarbinomannan reveals two lipomannan pools in corynebacterineae” *J. Biol. Chem.*, vol. 283, no. 11, pp. 6773–6782, Mar. 2008

[95] Y. S. Morita *et al.*, “PimE is a polyprenol-phosphate-mannose-dependent mannosyltransferase that transfers the fifth mannose of phosphatidylinositol mannoside in mycobacteria” *J. Biol. Chem.*, vol. 281, no. 35, pp. 25143–25155, Sep. 2006

[96] D. Alsteens, C. Verbelen, E. Dague, D. Raze, A. R. Baulard, and Y. F. Dufrêne, “Organization of the mycobacterial cell wall: A nanoscale view” *Pflugers Arch. Eur. J. Physiol.*, vol. 456, no. 1, pp. 117–125, Apr. 2008

[97] L. Tallorin, K. Finzel, Q. G. Nguyen, J. Beld, J. J. La Clair, and M. D. Burkart, “Trapping of the Enoyl-Acyl Carrier Protein Reductase-Acyl Carrier Protein Interaction” *J. Am. Chem. Soc.*, vol. 138, no. 12, pp. 3962–3965, Apr. 2016

[98] M. L. Schaeffer, G. Agnihotri, H. Kallender, P. J. Brennan, and J. T. Lonsdale, “Expression, purification, and characterization of the *Mycobacterium tuberculosis* acyl carrier protein, AcpM” *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids*, vol. 1532, no. 1–2, pp. 67–78, May 2001

[99] J. E. Griffin, J. D. Gawronski, M. A. DeJesus, T. R. Ioerger, B. J. Akerley, and C. M. Sassetti, “High-Resolution Phenotypic Profiling Defines Genes Essential for Mycobacterial Growth and Cholesterol Catabolism” *PLOS Pathog.*, vol. 7, no. 9, p. e1002251, Sep. 2011

[100] M. Sassetti, D. H. Boyd, E. J. Rubin, C. M. Sassetti, D. H. Boyd, and E. J. Rubin, “Genes required for mycobacterial growth defined by high density mutagenesis” *Mol. Microbiol.*, vol. 48, no. 1, pp. 77–84, Apr. 2003

[101] L. Kremer *et al.*, “Biochemical Characterization of Acyl Carrier Protein (AcpM) and Malonyl-CoA:AcpM Transacylase (mtFabD), Two Major Components of *Mycobacterium tuberculosis* Fatty Acid Synthase II” *J. Biol. Chem.*, vol. 276, no. 30, pp. 27967–27974, Jul. 2001

[102] O. Zimhony *et al.*, “AcpM, the meromycolate extension Acyl carrier protein of *Mycobacterium tuberculosis*, is activated by the 4′-phosphopantetheinyl transferase PptT, a potential target of the multistep mycolic acid biosynthesis” *Biochemistry*, vol. 54, no. 14, pp. 2360–2371, Apr. 2015

[103] C. Nguyen *et al.*, “Trapping the dynamic acyl carrier protein in fatty acid biosynthesis” *Nat. 2013 5057483*, vol. 505, no. 7483, pp. 427–431, Dec. 2013

[104] K. Takayama, C. Wang, and G. S. Besra, “Pathway to synthesis and processing of mycolic acids in *Mycobacterium tuberculosis*” *Clin. Microbiol. Rev.*, vol. 18, no. 1, pp. 81–101, Jan. 2005

[105] H. Therisod and E. P. Kennedy, “The function of acyl carrier protein in the synthesis of membrane-derived oligosaccharides does not require its phosphopantetheine prosthetic group.” *Proc. Natl. Acad. Sci.*, vol. 84, no. 23, pp. 8235–8238, Dec. 1987

[106] N. N. K. Kreamer *et al.*, “Acylated-acyl carrier protein stabilizes the *Pseudomonas aeruginosa* WaaP lipopolysaccharide heptose kinase” *Sci. Reports 2018 81*, vol. 8, no. 1, pp. 1–12, Sep. 2018

[107] A. Battesti and E. Bouveret, “Bacteria possessing two RelA/SpoT-like proteins have evolved a specific stringent response involving The acyl carrier protein-SpoT interaction” *J. Bacteriol.*, vol. 191, no. 2, pp. 616–624, Jan. 2009

[108] J. M. Belardinelli *et al.*, “The MmpL3 interactome reveals a complex crosstalk between cell envelope biosynthesis and cell elongation and division in mycobacteria” *Sci. Reports 2019 91*, vol. 9, no. 1, pp. 1–14, Jul. 2019

[109] L. Favrot and D. R. Ronning, “Targeting the mycobacterial envelope for tuberculosis drug development” *Expert Rev. Anti. Infect. Ther.*, vol. 10, no. 9, pp. 1023–1036, Sep. 2014

8. Acknowledgments

Author Contributions

Conceptualization: F. Mancia and O. B. Clarke

Genomics expansion and small-scale screening: A. L. Rosário, B. Kloss and *J. Rodrigues*

Medium-scale expression and purification: *J. Rodrigues*

Large-scale expression, purification and negative-stain EM: Y. Z. Tan

Grid vitrification: V. P. Dandey, H. Wei and Y. Z. Tan

Cryo-EM data collection, processing and analysis: Y. Z. Tan

Model building: Y. Z. Tan and O. B. Clarke

Phylogenetic analysis: Y. Z. Tan

Construction and characterization of *M. smegmatis* Conditional aftD deletion mutant, AftD mutants' function in *M. smegmatis*: L. Zhang (**Supervision**: M. Niederweis)

Scanning electron micrographs collection: A. M. Raczkowski and Y. Z. Tan

Construction of AftD mutants: L. Zhang, Y. Z. Tan and S. I. Giacometti

Glycan array experiments: R. B. Zheng (**Supervision**: T. L. Lowary)

Expression and purification of AftD in *M. smegmatis*: *J. Rodrigues*, D. Athayde (**Supervision**: M. J. Catalão, M. Pimentel and M. Archer)

EM analysis supervision: B. Carragher and C. S. Potter

Project supervision: F. Mancia

Manuscript preparation: Y. Z. Tan and F. Mancia with input from all authors.

We thank Joel Nott (Iowa State University) and Ricardo Gomes (UniMS, ITQB/IBET, Oeiras, Portugal) for protein mass spectrometry help. We thank Ed Eng, Bill Rice, Laura Kim, Mikhail Kopylov, and Kelsey Jordan (New York Structural Biology Center, Simons

Electron Microscopy Center) for help with microscope setup. We thank Sargis Dallakyan, Carl Negro, Shaker Krit, and Swapnil Bhatkar (New York Structural Biology Center, Simons Electron Microscopy Center) for computation support. We thank Dirk Schnappinger (Weill Cornell Medicine) for providing the components of the reverse TetR system. We thank Jeremie Vendome (Schrödinger) and Dmitry Lyumkis (Salk Institute) for critical reading of the manuscript. We thank Christina Chen for help with nanodisc incorporation, Khuram Ashraf for help with expression optimization, Vasileios Petrou for helpful discussions, and Leora Hamberger for her assistance in managing the Mancia laboratory (Columbia University).

Funding

This work was supported by grants from the NIH (P41 GM103310 to C.S.P. and B.C.; and R01 GM111980, R35 GM132120, and R21 AI119672 to F.M.), the Agency for Science, Technology and Research Singapore (to Y.Z.T.), the University of Alabama at Birmingham (to M.N.), the Fundação para a Ciência e Tecnologia (FCT), Lisbon, Portugal (PD/BD/128261/2016 to J.R.; PTDC/BIA-BQM/30421/2017 and IF/00656/2014 to M.A.; and PTDC/BIA-MIC/31233/2017 to M.J.C.), the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement nos. 731005 and 823780 (to J.R. and M.A.), the Simons Foundation (SF349247 to C.S.P. and B.C.), NYSTAR (to C.S.P. and B.C.), and the Alberta Glycomics Centre (to T.L.L.). Some of the work was performed at the Center for Membrane Protein Production and Analysis (COMPPÅ; NIH P41 GM116799 to Wayne Hendrickson) and at the National Resource for Automated Molecular Microscopy at the Simons Electron Microscopy Center (P41 GM103310), both located at the New York Structural Biology Center. M.A. acknowledges the MostMicro Research Unit (financially supported by LISBOA-01-0145-FEDER-007660 funded by FEDER funds through COMPETE2020 and by national funds through FCT), and iNOVA4Health (LISBOA-01-0145-FEDER-007344, co-funded by FEDER under PT2020).

9. Supplemental Information

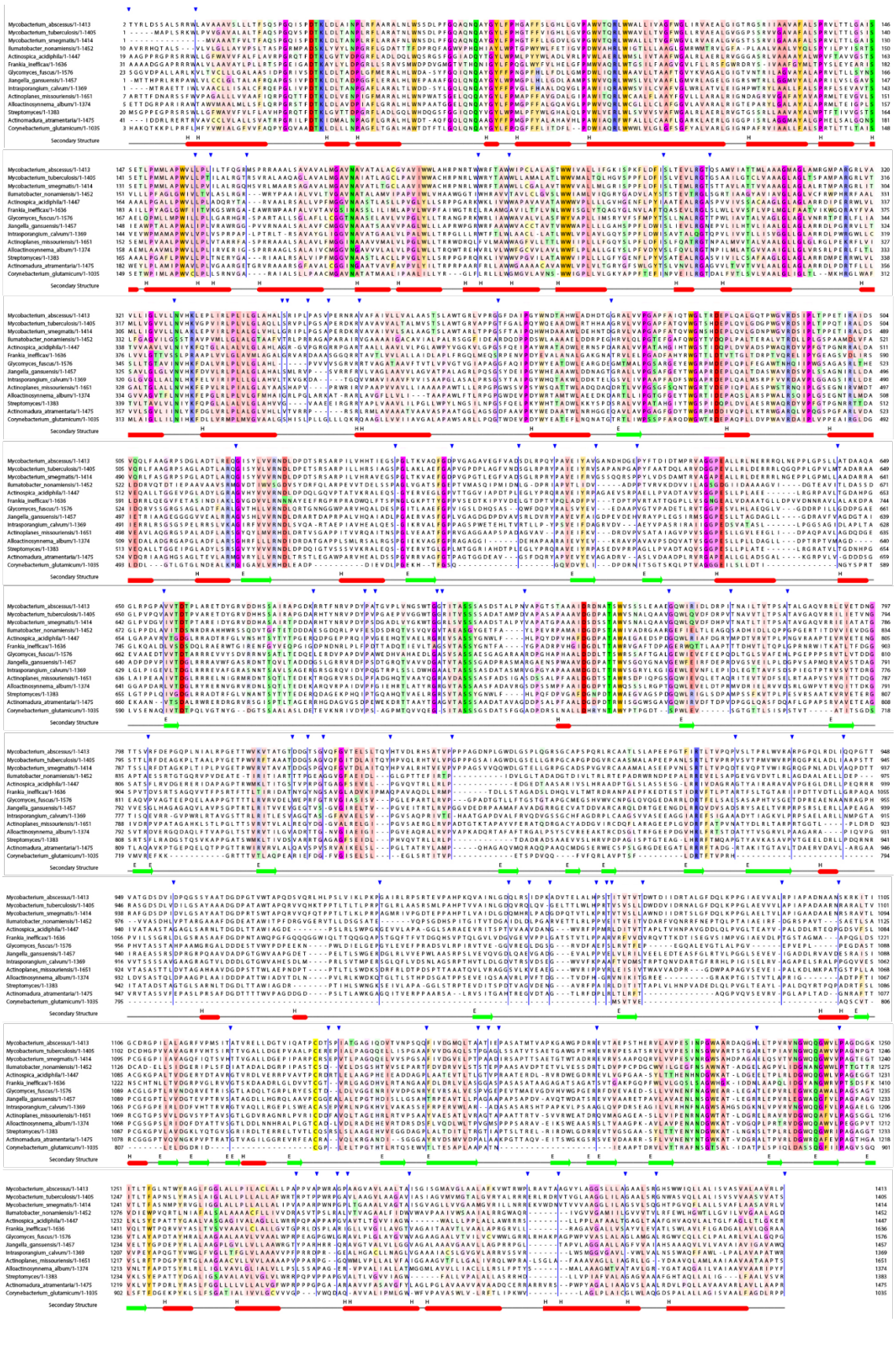


Figure S3.1– Multiple sequence alignment of AftD across the Actinobacteria class. *The sequence alignment was generated using MUSCLE and comprises of the following species (order in brackets): M. abscessus (Corynebacteriales), M. tuberculosis (Corynebacteriales), M. smegmatis (Corynebacteriales), Actinospica acidiphila (Catenulisporales), Streptomyces mirabilis (Streptomycetales), Actinomadura atramentaria (Streptosporangiales), Glycomyces fuscus (Glycomycetales), Actinoplanes missouriensis (Micromonosporales), Alloactinosynnema album (Pseudonocardiales), Jiangella gansuensis (Jiangellales), Intrasporangium calvum (Micrococcales), Kribbella flavida (Propionibacteriales), Ilumatobacter nonamiensis (Acidimicrobiales), Frankia inefficax (Frankiales) and Corynebacterium glutamicum (Corynebacteriales). The residues are colored by their physiochemical properties using the Zappo scheme: Aliphatic/hydrophobic residues are in salmon (ILVAM), aromatic residues in orange (FWY), positively charged residues in blue (KRH), negatively charged residues in red (DE), hydrophilic residues in green (STNQ), conformationally special residues in pink (PG) and cysteines in yellow (C). The intensity of the coloring is proportional to the degree of conservation. Secondary structure elements derived from the M. abscessus structure are labelled and aligned below. Gaps and unaligned parts of the sequence have been shrunk and indicated by the blue arrow.*

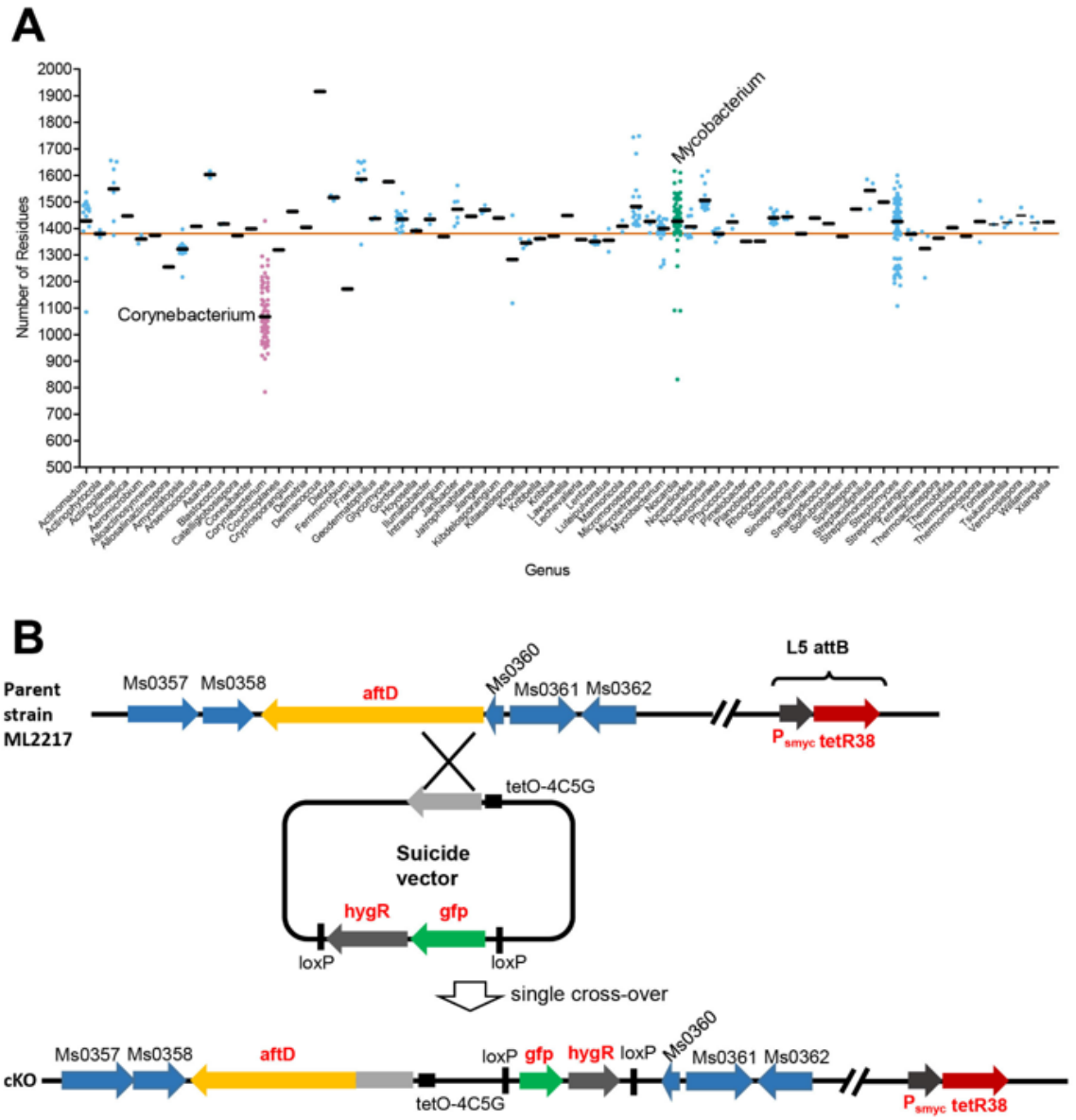
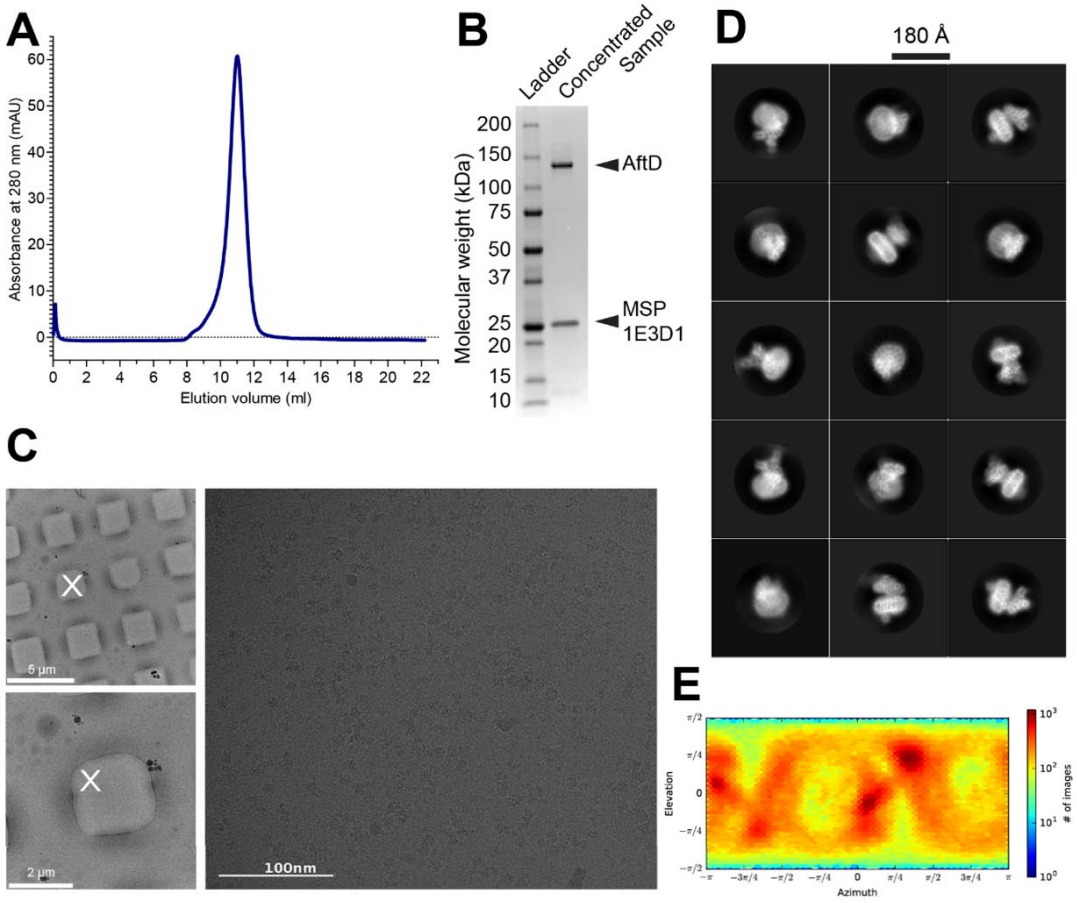
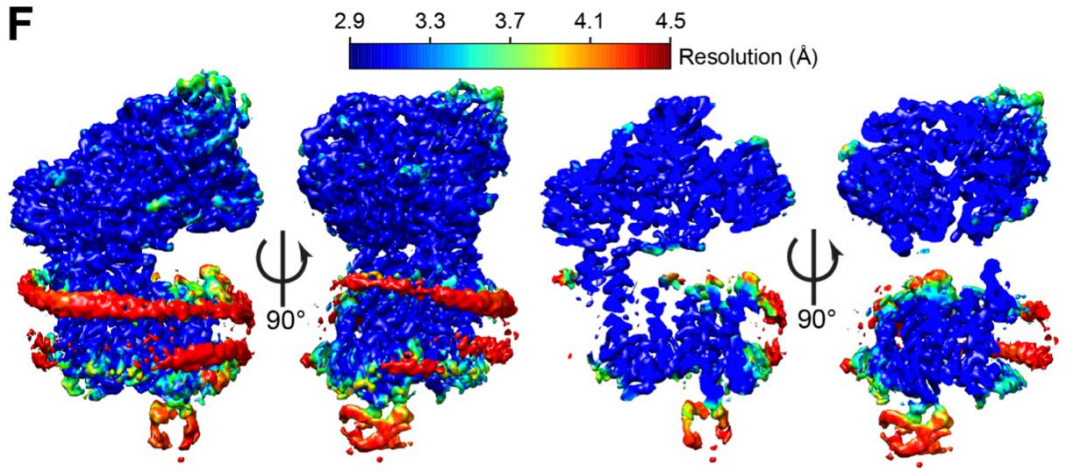


Figure S3.2 – Lengths of AftD Homologs and Conditional Knock-out Scheme. ***(A)*** *Total number of residues of each AftD homolog against the genus. Corynebacterium genus (pink dots) has unusually short AftD sequence compared to the rest of the genus in the Actinobacteria class (blue dots), including the Mycobacterium genus (green dots).* ***(B)*** *Conditional knock-out scheme for aftD gene in M. smegmatis. The gene locus for WT M. smegmatis around the aftD gene is shown.*


A
Absorbance at 280 nm (mAU)
Elution volume (ml)
B
Ladder
Concentrated Sample
Molecular weight (kDa)
AftD
MSP 1E3D1
D
180 Å
C
5 µm
2 µm
100nm
E
Elevation
Azimuth
# of images


F
2.9 3.3 3.7 4.1 4.5
Resolution (Å)
90°
90°

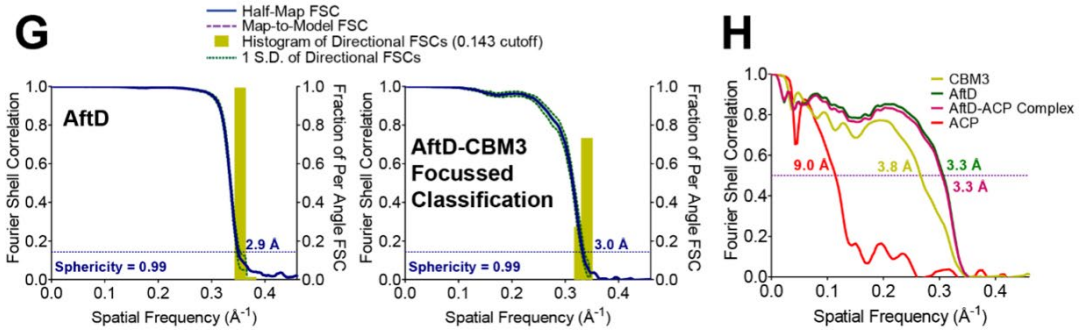

G
Half-Map FSC
Map-to-Model FSC
Histogram of Directional FSCs (0.143 cutoff)
1 S.D. of Directional FSCs
AftD
2.9 Å
Sphericity = 0.99
AftD-CBM3 Focussed Classification
3.0 Å
Sphericity = 0.99
Fourier Shell Correlation
Fraction of Per Angle FSC
Spatial Frequency (Å-1)
H
CBM3
AftD
AftD-ACP Complex
ACP
9.0 Å
3.8 Å
3.3 Å
3.3 Å
Fourier Shell Correlation
Spatial Frequency (Å-1)

Figure S3.3 – Purification and single-particle cryo-EM structural determination of wild-type M. abscessus AftD. (A) *Representative size-exclusion chromatography (SEC) trace of AftD incorporated into MSP-1E3D1 nanodiscs. Fractions corresponding to the peak were used for cryo-EM analysis.* ***(B)*** *SDS-PAGE gel of the concentrated sample from the peak fractions stained with Coomassie blue. The predicted molecular weight of AftD is 150 kDa, while that for MSP-1E3D1 is 32 kDa.* ***(C)*** *Representative square, hole and exposure magnification views of the Spotiton nanowire carbon patterned grids that were imaged. The exposure image had a defocus of -1.7 µm (estimated by CTFFind4 and pixel size of 1.0605 Å/pixel.* ***(D)*** *Representative 2D class averages.* ***(E)*** *Euler angle distribution plot of the final 3D reconstruction from CryoSPARC 2.* ***(F)*** *ResMap display of unsharpen AftD reconstructions in orthogonal views, showing the nanodisc belt as well as variation in local resolution.* ***(G)*** *Fourier shell correlation (FSC) curves for the full AftD map and AftD-CBM3 focused classified map describing the half-map (blue) resolutions (at 0.143 cut-offs), as well as histogram of directional resolutions sampled evenly over the 3DFSC (yellow). The corresponding sphericity value is also indicated.* ***(H)*** *Fourier shell correlation (FSC) curves for the atomic models of AftD, ACP, CBM3 of AftD and AftD with ACP against the whole composite map generated by combining the full AftD map with the AftD-CBM3 focused classified map using phenix.combine_focused_maps at 0.5 cut-off.*

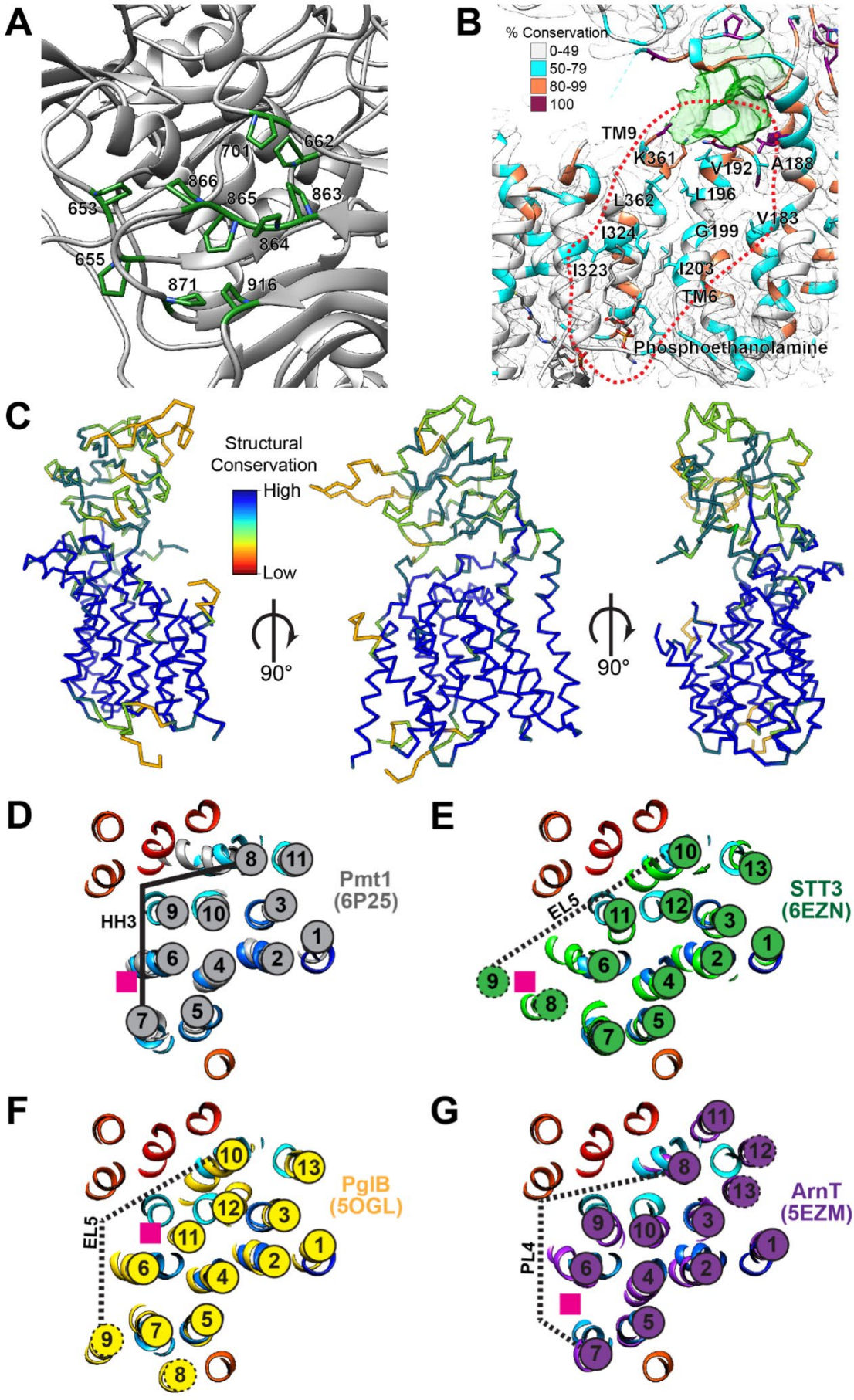

A
662
701
866
865
863
653
864
655
871
916
B
% Conservation
0-49
50-79
80-99
100
TM9
K361
V192
A188
L362
L196
V183
I324
G199
I323
I203
TM6
Phosphoethanolamine
C
Structural
Conservation
High
Low
90°
90°
D
Pmt1
(6P25)
HH3
E
STT3
(6EZN)
EL5
F
PglB
(5OGL)
EL5
G
ArnT
(5EZM)
PL4

Figure S3.4 – Structural Features of AftD. (A) *Stretch of consecutive prolines in CBM2 and neighboring prolines shown as green sticks. The rest of the structure is rendered as cartoon.* ***(B)*** *Location of the potential binding site (dotted red line) for the lipidic tail of the donor DPA. AftD is rendered as cartoon, with residues colored by conservation. The density map is shown as transparent grey.* ***(C)*** *Structural conservation of the GT-C fold of AftD (GT53 family, residues 1 to 600) against representative GT-C glycosyltransferases Pmt1-Pmt2 (GT39 family, PDB ID: 6P2R), STT3 (GT66 family, PDB ID: 6C26) and ArnT (GT83 family, PDB ID: 5EZM). Structural alignment was performed by the Dali server. Eukaryotic Pmt1-Pmt2 (GT39)* ***(D)****, eukaryotic STT3 (GT66)* ***(E)****, bacterial PglB (GT66)* ***(F)*** *and archaeal ArnT (GT83)* ***(G)*** *were superimposed onto full-length AftD and a slice through the transmembrane helices are shown as cartoon. The other glycosyltransferase structures were colored in grey, green, yellow and purple respectively, while full-length AftD was colored in rainbow. Helices of the other glycosyltransferases were labelled, and those that did not have a corresponding homologous helix in AftD were indicated with a dotted circle. Locations of the flexible loop that becomes ordered upon substrate binding (EL5, PL4) are indicated by the dotted lines. Pmt1-Pmt2's HH3 is already ordered as a helix without any substrate binding, and hence is shown as a solid line. The locations of the lipidic sugar donor substrate are indicated by pink boxes.*

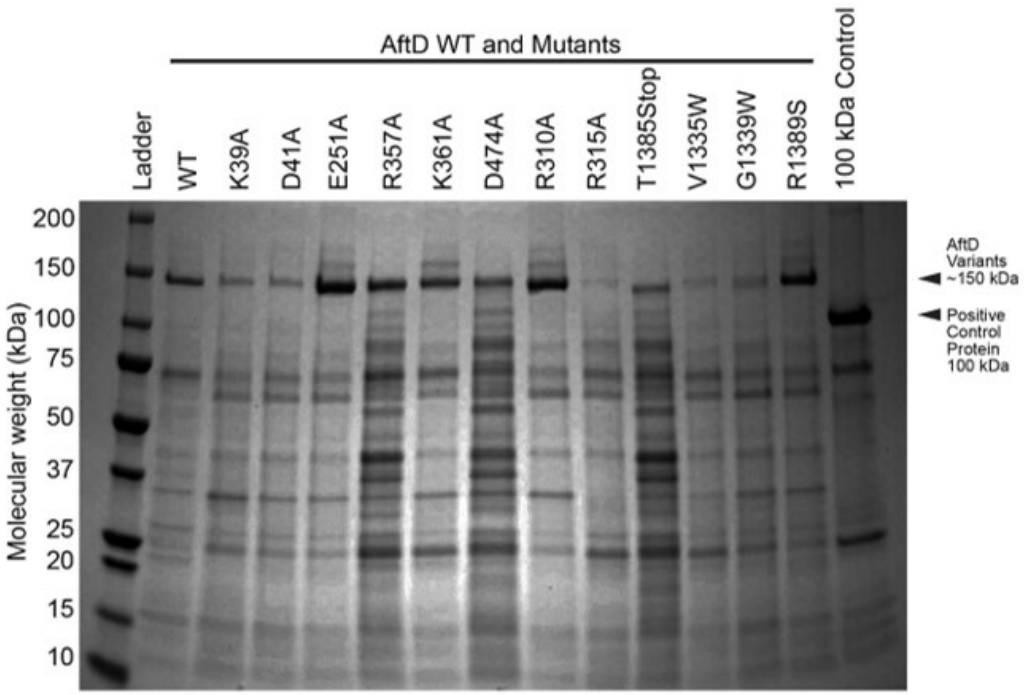


Figure S3.5 – Expression of AftD mutants in E. coli. *SDS-PAGE gel of samples from small scale purifications of WT AftD and AftD mutants, stained with Coomassie blue. The AftD WT and mutants were heterologously expressed in E. coli and purified for this experiment, without TEV cleavage.*

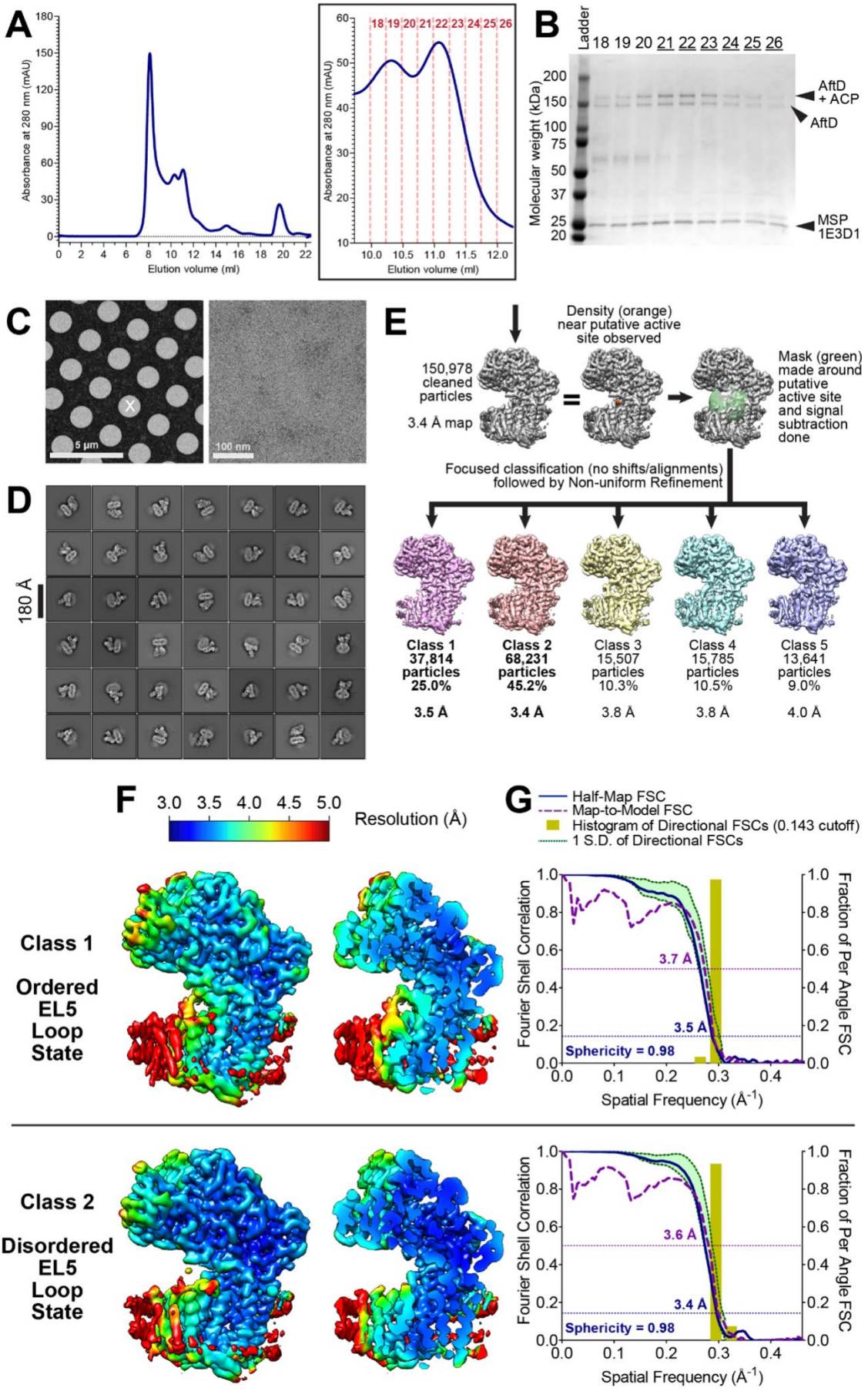

A
Absorbance at 280 nm (mAU)
Elution volume (ml)
18 19 20 21 22 23 24 25 26
B
Ladder
18 19 20 21 22 23 24 25 26
Molecular weight (kDa)
200
150
100
75
50
37
25
20
AftD + ACP
AftD
MSP 1E3D1
C
5 µm
100 nm
D
180 Å
E
Density (orange) near putative active site observed
150,978 cleaned particles
3.4 Å map
Mask (green) made around putative active site and signal subtraction done
Focused classification (no shifts/alignments) followed by Non-uniform Refinement
Class 1 37,814 particles 25.0% 3.5 Å
Class 2 68,231 particles 45.2% 3.4 Å
Class 3 15,507 particles 10.3% 3.8 Å
Class 4 15,785 particles 10.5% 3.8 Å
Class 5 13,641 particles 9.0% 4.0 Å
F
3.0 3.5 4.0 4.5 5.0
Resolution (Å)
Class 1
Ordered EL5 Loop State
Class 2
Disordered EL5 Loop State
G
Half-Map FSC
Map-to-Model FSC
Histogram of Directional FSCs (0.143 cutoff)
1 S.D. of Directional FSCs
Fourier Shell Correlation
Fraction of Per Angle FSC
3.7 Å
3.5 Å
Sphericity = 0.98
3.6 Å
3.4 Å
Sphericity = 0.98
Spatial Frequency (Å^{-1})

Figure S3.6 – Purification and single-particle cryo-EM structural determination of AftD-R1389S. (A) *Representative size-exclusion chromatography (SEC) trace of AftD-R1389S incorporated into MSP-1E3D1 nanodiscs. The insert shows a zoom in of the fractions collected for SDS-PAGE analysis.* ***(B)*** *SDS-PAGE gel of the concentrated sample from the peak fractions stained with Coomassie blue. The predicted molecular weight of AftD-R1389S is 150 kDa, while that for MSP-1E3D1 is 32 kDa. Mass spectrometry was used to determine that the higher band at around 150 kDa corresponds to AftD-R1389S with ACP still associated.* ***(C)*** *Representative square, hole and exposure magnification views of the gold grids that were imaged. The exposure image had a defocus of -1.6 µm (estimated by CTFFind4) and pixel size of 1.0605 Å/pixel.* ***(D)*** *Representative 2D class averages.* ***(E)*** *Signal subtraction and 3D focused classification scheme.* ***(F)*** *ResMap display of two classes from focused classification of unsharpen AftD R1389S.* ***(G)*** *Fourier shell correlation (FSC) curves for the two classes describing the half-map (blue) resolutions (at 0.143 cut-offs), map-to-model (purple) resolutions (at 0.5 cut-offs), as well as histogram of directional resolutions sampled evenly over the 3DFSC (yellow). The corresponding sphericity value is also indicated.*

Table S3.1 – Cryo-EM data collection and modeling statistics WT-AftD.

	AftD-WT	**AftD-WT-CBM3 Focus Classification**
EM data collection / processing		
Microscope	FEI Titan Krios	
Voltage (kV)	300	
Camera	Gatan K2 Summit with Energy Filter	
Energy Filter Slit Width (eV)	15	
Set defocus range (μm)	1.0 – 2.0	
Defocus mean ± std (μm)	1.6 ± 0.39	
Exposure time (s)	13.5	
Number of frames	90	
Dose rate (e-/pixel/s)	8.2	
Total dose (e-/ Å^2)	95.72	
Pixel size (Å)	1.0605	
Number of micrographs	7,274	
Number of particles (after initial cleanup)	889,985	
Number of particles (in final map)	490,616	–
Number of particles (after classification)	–	50,410
Symmetry	C1	C1
Resolution (global) (Å)	2.9	3.0
Local Resolution Range	2.3 – 4.1	2.3 – 4.1
Directional Resolution Range	2.7 – 2.9	2.9 – 3.1
Sphericity of 3DFSC	0.99	0.99
SCF Value*	0.836	0.836
Map sharpening b-factor (Å^2)	-56.04	-29.62
Model statistics	**AftD-WT**	
Residue range	2-1391	
Map CC	0.828	
RMSD [bonds] (Å)	0.010	
RMSD [angles] (Å)	1.224	
All-atom clashscore	2.27	
Ramachandran plot		
Favored (%)	95.45	
Allowed (%)	4.47	
Outliers (%)	0.08	
Rotamer outliers	0.00	
C-β deviations	0	
MolProbity Score	1.32	
EM-Ringer Score	3.8	

* The SCF value assumes that all orientations have been properly assigned and does not take into account false positive assignment.

Table S3.2 – Proteins identified by Liquid MS/MS from M. smegmatis purified AftD-WT M. abscessus.

Name	Accession UniProt	Unused ProtScore	Total ProtScore	Confidence (95%)	Species	Number of Peptides (95%)
ATP synthase subunit beta	sp|A0R200|ATPB_MYCS2	62.02	62.02	66.1	*M. smegmatis*	31
UPF0182 protein MSMEG_1959/MSMEI_	sp|A0QTT7|Y1959_MYCS2	38.08	38.08	25	*M. smegmatis*	21
ATP synthase subunit alpha	sp|A0R202|ATPA_MYCS2	37.14	37.19	50.4	*M. smegmatis*	19
AftD *Mycobacterium abscessus* 1948 F5/8 type C domain protein		**32.81**	**32.81**	**11.4**	***M. abscessus***	**18**
60 kDa chaperonin 1	sp|A0QQU5|CH601_MYCS2	28.05	28.05	39.2	*M. smegmatis*	14
Adenosylhomocysteinase	sp|A4ZHR8|SAHH_MYCS2	24.52	24.52	34.2	*M. smegmatis*	14
Elongation factor Tu	sp|A0QS98|EFTU_MYCS2	24.42	24.45	48	*M. smegmatis*	13
Probable malate:quinone oxidoreductase	sp|A0QVL2|MQO_MYCS2	18.94	18.94	24.7	*M. smegmatis*	10
Cell wall synthesis protein Wag31	sp|A0R006|WAG31_MYCS2	16.6	16.6	34.9	*M. smegmatis*	8
ATP synthase gamma chain	sp|A0R201|ATPG_MYCS2	12	12	30	*M. smegmatis*	6
Ketol-acid reductoisomerase (NADP(+))	sp|A0QUX8|ILVC_MYCS2	11.91	11.95	22.9	*M. smegmatis*	6
Glutamine synthetase	sp|A0R079|GLN1B_MYCS2	9.83	9.83	12.3	*M. smegmatis*	5
Galactofuranosyltransferase GlfT2	sp|A0R628|GLFT2_MYCS2	8.02	8.02	9.4	*M. smegmatis*	4
30S ribosomal protein S3	sp|A0QSD7|RS3_MYCS2	8.01	8.01	15.3	*M. smegmatis*	4
Uncharacterized protein MSMEG_4692/MSMEI_4575	sp|A0R1B5|Y4692_MYCS2	8	8	44	*M. smegmatis*	4
ATP synthase epsilon chain	sp|A0R1Z9|ATPE_MYCS2	8	8	41.3	*M. smegmatis*	4
30S ribosomal protein S5	sp|A0QSG6|RS5_MYCS2	6	6	25.7	*M. smegmatis*	3
Acyl carrier protein	tr|A0A0D6IL69|A0A0D6IL69_MYCSM	**6**	**6**	**53.5**	***M. smegmatis***	**3**
Pyruvate dehydrogenase E1 component	sp|A0R0B0|ODP1_MYCS2	5.81	5.87	3.8	*M. smegmatis*	3
Decaprenylphosphoryl-2-keto-beta-D-erythro-pentose reductase	sp|A0R610|DPRE2_MYCS2	5.07	5.13	12.6	*M. smegmatis*	3
Acetamidase	sp|Q07838|AMDA_MYCSM	4.01	4.02	5.4	*M. smegmatis*	2
50S ribosomal protein L5	sp|A0QSG1|RL5_MYCS2	4	4	14.4	*M. smegmatis*	2
ATP synthase subunit b	sp|A0R204|ATPF_MYCS2	2.29	2.29	7.6	*M. smegmatis*	1

Table S3.3 – Cryo-EM data collection and modeling statistics AftD R1389S.

	AftD R1389S Class 1	AftD R1389S Class 2
EM data collection / processing		
Microscope	FEI Titan Krios	
Voltage (kV)	300	
Camera	Gatan K2 Summit with Energy Filter	
Energy Filter Slit Width (eV)	15	
Set defocus range (μm)	0.8 – 1.6	
Defocus mean ± std (μm)	1.42 ± 0.35	
Exposure time (s)	13.5	
Number of frames	90	
Dose rate (e-/pixel/s)	8.34	
Total dose (e-/ Å^2)	96.78	
Pixel size (Å)	1.0605	
Number of micrographs	4,886	
Number of particles (after initial cleanup)	226,478	
Number of particles (in consensus map)	150,978	
Number of particles (after classification)	37,814	68,231
Symmetry	C1	C1
Resolution (global) (Å)	3.5	3.4
Local Resolution Range	3.0 – 8.7	3.0 – 7.2
Directional Resolution Range	3.3 – 3.6	3.2 – 3.6
Sphericity of 3DFSC	0.98	0.98
SCF Value*	0.833	0.844
Map sharpening B-factor (Å^2)	-125.09	-129.19
Model statistics		
Residue range	5-1,388	5-1,388
Map CC	0.8403	0.8045
RMSD [bonds] (Å)	0.012	0.010
RMSD [angles] (Å)	1.364	1.310
All-atom clashscore	4.12	3.23
Ramachandran plot		
Favored (%)	92.12	93.62
Allowed (%)	7.80	6.30
Outliers (%)	0.08	0.08
Rotamer outliers	0.11	0.32
C-β deviations	0	0
MolProbity Score	1.68	1.54
EM-Ringer Score	3.70	3.21

* The SCF value assumes that all orientations have been properly assigned and does not take into account false positive assignment.

CHAPTER FOUR

Cryo-EM structure of arabinosyltransferase EmbB from *Mycobacterium smegmatis*

This chapter contains published data:

Tan Y. Z., **Rodrigues, J.**, Keener J. E., Zheng R. B., Brunton R., Kloss B., Giacometti S. I., Rosário A. L., Zhang L., Niederweis M., Clarke O. B., Lowary T. L., Marty M. T., Archer M., Potter C. S., Carragher B. & Mancia F., "Cryo-EM structure of arabinosyltransferase EmbB from *Mycobacterium smegmatis*" *Nature Communications*, vol. 11, no. 1, p. 3396, 2020, doi: 10.1038/s41467-020-17202-8.

Author Contributions – J. Rodrigues

Genomics expansion and small-scale screening: A. L. Rosário, B. Kloss and *J. Rodrigues*

Manuscript preparation: Y. Z. Tan and F. Mancia with input from all authors.

Review and editing: all authors

1. Summary

Arabinofuranosyltransferase B (EmbB) belongs to a family of membrane-bound glycosyltransferases that build the lipidated polysaccharides of the mycobacterial cell envelope, and are targets of anti-tuberculosis drug ethambutol. We present the 3.3 Å resolution structure of *Mycobacterium smegmatis* EmbB solved by single-particle cryo-electron microscopy, providing insights on substrate binding and reaction mechanism. Mutations that confer ethambutol resistance map mostly around the putative active site, suggesting this pocket to be the location of drug binding.

2. Introduction

The cell envelope is crucial for growth and virulence of pathogenic mycobacteria like *M. tuberculosis* [1] and is a major contributor to resistance against common antibiotics [2]. Its main component is the mycolyl-arabinogalactan-peptidoglycan (mAGP) complex, which consists of peptidoglycan, a branched heteropolysaccharide arabinogalactan (AG) and long chain mycolic acids (Figure 4.1A). Another major component of the mycobacterial cell envelope is the lipidated heteropolysaccharide lipoarabinomannan (LAM) [3]. Of the enzymes involved in mycobacterial cell wall biosynthesis, arabinofuranosyltransferases (AraTs) are responsible for the addition of D-arabinofuranose sugar moieties to AG and LAM [2]. These transmembrane (TM) enzymes utilize the lipidic donor decaprenylphosphoryl-D-arabinofuranose (DPA) to transfer an arabinofuranose unit (Ara*f*) to the growing lipidated polysaccharides of the cell envelope [4].

Arabinosyltransferase B (EmbB), a 117 kDa integral membrane enzyme involved in the α-(1→5)-linked extension of the AG arabinan chain (Figure 4.1B), is one of the best characterized members of the aforementioned AraT family [5][6]. Its gene belongs to an operon coding for two other homologous – arabinofuranosyltransferase A (EmbA), which acts also on the extension of AG (EmbB and EmbA share 40% amino acid sequence identity in *M. smegmatis* and 42% in *M. tuberculosis*) and arabinofuranosyltransferase C (EmbC), which acts on LAM (EmbB and EmbC share 46% identity in *M. smegmatis*, and 44% in *M. tuberculosis*) (Supplementary Figures S4.1 and S4.2). The operon was named due to the sensitivity of the respective gene products to ethambutol, a first-line antibiotic against tuberculosis [9] and nontuberculous mycobacterial (NTM) disease [10].

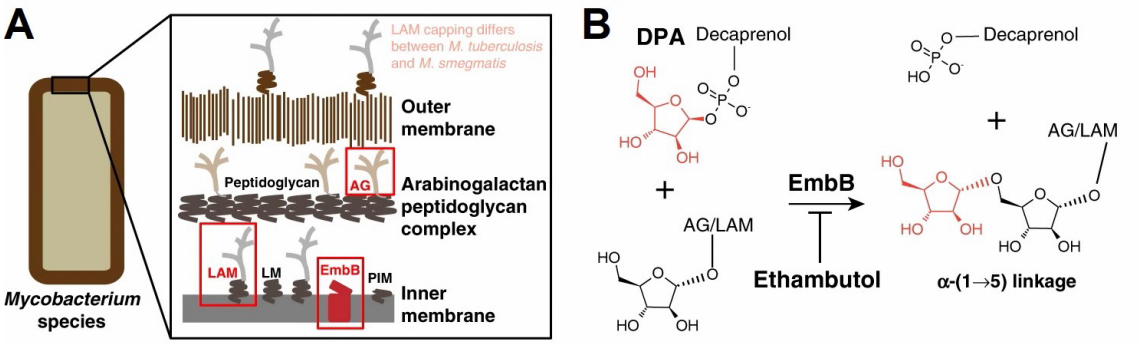


Figure 4.1 – Mycobacterial cell envelope and Ethambutol inhibition. (A) *Model of the cell envelope of mycobacterial cell envelope based on Dulberger et al.* [5]. *Red boxes highlight arabinogalactan (AG) and lipoarabinomannan (LAM) components, synthesized by EmbA/EmbB (red) and EmbC, respectively. LAM for M. tuberculosis is capped by α-mannose glycans while in M. smegmatis it is capped by phosphoinositol instead* [6]. ***(B)*** *Reaction catalyzed by EmbB, which is inhibited by ethambutol.*

EmbB and EmbC are reported to have mutations known to lead to ethambutol resistance [11], while a clear effect of the drug on EmbA activity has not yet been established [12]. The lack of atomic models of these proteins (only a high resolution structure of the C-terminal soluble domain of EmbC is available [13]) has thus far hindered our understanding of the catalytic action and drug resistance mechanisms of these proteins, although this situation has been recently remedied to a certain extent by the determination of the 3D structures of the EmbA-EmbB and EmbC-EmbC dimer complexes [14].

In this work, we report the 3D structure of *M. smegmatis* EmbB to 3.3 Å resolution, providing insights on substrate binding and catalytic mechanism. Mutations conferring resistance to ethambutol are mostly located around the putative active site, suggesting this drug to be a competitive inhibitor.

3. Materials and Methods

3.1. Sequence alignment

Protein sequences of EmbA, EmbB, and EmbC from *M. tuberculosis* and *M. smegmatis* were obtained from the Mycobrowser [15], with the following KEGG identifiers: EmbA *Mtb*-Rv3794, EmbB *Mtb*-Rv3795, EmbC *Mtb*-Rv3793, EmbA *Msm*-MSMEG_6388, EmbB *Msm*-MSMEG_6389, and EmbC *Msm*-MSMEG_6387. The sequences were then aligned using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) [16] and displayed using ESPript (http://espript.ibcp.fr) [17].

3.2. Genomic expansion and small-scale screening

EmbB genes were identified from a collection of 14 *Mycobacterium* genomes using a bioinformatics approach [18]. Ligation independent cloning (LIC) was used to clone these targets from the genomes into five LIC-adapted expression vectors (pNYCOMPS-Nterm, pNYCOMPS-Cterm, pNYCOMPS-N23, pNYCOMPS-C23, and pMCSG7-10x) that contained a tobacco etch virus (TEV) protease cleavage site (ENLYFQSYV) and decahistidine affinity tag. Small and medium scale expression was performed in a high throughput manner as described in detail in a previous protocol by Bruni and Kloss [19]. A number of orthologs could be cloned and expressed well (work described in Chapter 2: *High Throughput Expression Screening of Arabinofuranosyltransferases from Mycobacteria*), but *M. smegmatis embB* was chosen over the others because it represents a model organism used to study pathogenic *M. tuberculosis*.

M. smegmatis embB was ultimately cloned using LIC into a pMCSG21 expression vector [20] that contained a TEV protease cleavage site and Strep-tag on the 3′ end of the insert. This expression construct was used for all subsequent experiments.

3.3. EmbB expression, purification, and nanodisc reconstitution

M. smegmatis embB in the pMCSG21 plasmid was transformed into BL21 (DE3) pLysS *E. coli* competent cells and plated onto Luria-Bertani (LB) agar (Fisher) plates supplemented with 100 µg/mL ampicillin (Sigma) and 100 µg/mL spectinomycin (Sigma), and grown overnight at 37 °C. In the next day, a colony was picked and used to inoculate a starter culture containing 150 mL of 2xYT medium (Fisher) supplemented with 100 µg/mL ampicillin and 100 µg/mL spectinomycin. The starter culture was grown overnight at 37 °C in an incubator (New Brunswick Scientific) shaking at 240 rpm. The following day, six 2-L baffled flasks each with 800 mL of 2xYT medium (Fisher) supplemented with 100 µg/mL ampicillin and 100 µg/mL spectinomycin were inoculated with 10 mL of starter culture. The cultures were then grown at 37 °C shaking at 240 rpm until cells reached an OD_{600} of ~1.0 (~3 h). Temperature was then reduced to 22 °C and protein expression was induced by addition of 0.2 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) (Fisher). The culture was then incubated overnight shaking at 240 rpm. The next day, the cells were harvested by centrifugation at 4000 × *g* utilizing a H6000A/HBB6 rotor (Sorvall) for 30 min at 4 °C. The supernatant was discarded and the pellet was resuspended in chilled 1x phosphate buffered saline (PBS) and centrifuged again at 4000 × *g* for 30 min at 4 °C. The supernatant was again discarded and the pellet was resuspended in lysis buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 20 mM $MgSO_4$, 10 µg/mL DNase I (Roche), 8 µg/mL RNase A (Roche), 1 mM tris(2-carboxyethyl)phosphine hydrochloride (TCEP), 1 mM phenylmethylsulfonyl fluoride (PMSF), 1 tablet in 1.5 L buffer EDTA-free cOmplete protease inhibitor cocktail (Roche). For a 4.8 L of culture, the yield corresponded to ~10–20 g of wet cell pellet mass, which was resuspended with ~250 mL of lysis buffer. Cells were lysed by passing the suspension through a chilled Emulsiflex C3 homogenizer (Avestin) three times. The crude membrane fraction was isolated by ultracentrifugation at 37,000 × *g* in a Type 45 Ti Rotor (Beckman Coulter) at 4 °C for 30 min. The supernatant was discarded and the pellet was resuspended in the lysis buffer up to a volume of 240 mL and homogenized using a hand-held glass homogenizer (Konte). The membrane fraction was then stored at −80 °C until later use to purify protein.

The thawed membrane fraction was solubilized by adding n-dodecyl-β-D-maltopyranoside (DDM) to a final concentration of 1% (w/v) detergent for 2 h at 4 °C with gentle rotation. Insoluble material was removed by ultracentrifugation at 40,000 × *g* in Type 45 Ti Rotor at 4 °C for 30 min. 1.5 mg of avidin (IBA Lifesciences) was added to the

supernatant to block any endogenous biotin, and the mixture was left on ice for 5 min. Thereafter, the supernatant was added to six Falcon tubes containing pre-equilibrated Strep-Tactin® Superflow resin (IBA Lifesciences) and incubated with gentle rotation at 4 °C for 2 h. The resin was washed with 10 column volumes of buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 0.1% DDM and eluted with buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 50 mM D-biotin (Alfa Aesar), 0.05% DDM. The eluted protein was exchanged into a buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 0.05% DDM using a PD-10 desalting column (GE), and concentrated down using a 100-kDa concentrator (Pierce) to ~1 mg/mL.

The protein was then incorporated into lipid nanodisc [21] with a molar ratio 1:300:6 between EmbB:1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-(1′-rac-glycerol) (POPG) (Avanti):membrane scaffold protein 1E3D1 (MSP-1E3D1) and incubated for 2 h with gentle agitation at 4 °C. The POPG was prepared by adding the solid extract to deionized water to a final concentration of 20 mM. The mix was placed on ice and then gently sonicated with a tip sonicator (Fisher Scientific) to dissolve the lipids. The lowest power setting was used and sonication was stopped when the mixture turned from cloudy to semi-transparent, after approximately five cycles. No detergent was added to the lipid extract. Reconstitution was initiated by removing detergent with the addition of 150 mg Bio-beads (Bio-Rad) per mL of protein solution for overnight with constant rotation at 4 °C. Bio-beads were removed by passing the protein solution through an Ultrafree centrifugal filter unit (Fisher) at 4000 × *g* in a Centrifuge 5424R (Eppendorf) at 4 °C for 1 min and the nanodisc reconstitution mixture was re-bound to fresh Strep-Tactin® Superflow resin for 2 h at 4 °C in order to remove empty nanodisc. The resin was washed with 10 column volumes of wash buffer consisting of 20 mM HEPES pH 7.5, 200 mM NaCl, followed by three column volumes of elution buffer consisting of 20 mM HEPES pH 7.5, 200 mM NaCl, and 50 mM D-biotin.

The eluent was concentrated using a 100-kDa concentrator to under 500 µL and loaded onto a Superdex 200 Increase 10/300 GL size-exclusion column (GE Healthcare Life Sciences) in gel filtration buffer (20 mM HEPES pH 7.5 and 200 mM NaCl). Throughout the entire process of purification, 15 µL of samples were taken and added to 5 µL of 6X reducing Laemmli SDS sample buffer (Bioland Scientific). The samples were then loaded on a 4–20% Mini-PROTEAN TGX precast protein gel (Bio-Rad) for protein gel electrophoresis in a Tris/Glycine/SDS buffer. The gel was developed using InstantBlue (Sigma) protein stain.

3.4. Negative stain electron microscopy

Purified EmbB in nanodiscs was diluted to 0.005 mg/mL and applied onto copper grids (Ted Pella). These grids were overlaid by a thin (~1.5 nm) layer of continuous carbon that had been plasma-cleaned (Gatan Solarus) for 30 s using a mixture of H_2 and O_2. Thereafter, filter paper (Whatman 4) was used to remove the protein solution. Three microliters of 2% uranyl formate were then added and immediately removed by absorbing with filter paper (this was repeated seven times). The grid was imaged on a Tecnai TF20 microscope (FEI) equipped with a Tietz F416 CCD camera (Tietz) at 1.10 Å per pixel, respectively, using the Leginon software package [22]. Seventy seven images were collected and processed using the Appion software package [23] to obtain 2D classes with Relion 2.1 [47][48]. The micrographs showed good particle dispersion and homogeneity.

3.5. Single-particle Cryo-EM sample vitrification

Purified EmbB was concentrated using a 100-kDa concentrator (Pierce) between 5 and 20 µL of sample at ~8 mg/mL. 1 mM of ethambutol (Sigma) was added before vitrification. 2.5 µL of sample was added to a plasma-cleaned (Gatan Solarus) 0.6/1.0 µm holey gold grid (Quantifoil UltrAuFoil) and blotted using filter paper on one side for 2 s using the Leica GP plunger system before plunging immediately into liquid ethane for vitrification. The plunger was operating at 5 °C with >80% humidity to minimize evaporation and sample degradation.

3.6. Data acquisition

Images were recorded in two sessions. The first session was on a Titan Krios electron microscope (FEI) equipped with a Falcon III direct detector operating at 0.665 Å per pixel in electron counting mode using the Leginon software package [22]. Pixel size was calibrated after obtaining a preliminary map by docking the crystal structure of C-terminal domain of *M. tuberculosis* EmbC (PDB ID: 3PTY) [13]. Data collection was performed using a dose of ~78.02 e^- $Å^{-2}$ across 80 frames (1080 ms per frame) at a dose rate of

~0.40 e^{-} pix^{-1} s^{-1}, using a set defocus range of −0.5 to −2.5 µm. In all, 100-µm objective aperture was used. A total of 2,158 micrographs were recorded over 3 days using an image beam shift data collection strategy [26].

The second session was on a Titan Krios electron microscope (FEI) equipped with a K2 summit direct detector operating at 0.667 Å per pixel in counting mode using the Leginon software package [22]. Pixel size was calibrated in-house using a proteasome test sample. Data collection was performed using a dose of ~77.53 e^{-} $Å^{-2}$ across 80 frames (100 ms per frame) at a dose rate of ~4.3 e^{-} pix^{-1} s^{-1}, using a set defocus range of −0.3 to −2.9 µm. In all, 100-µm objective aperture was used. A total of 7,833 micrographs were recorded over three days using an image beam shift data collection strategy [26].

During data collection, movie frames were aligned using MotionCor2 [27] with 5 by 5 patches and B-factor of 100 through the Appion software package [23]. Micrograph CTF estimation was performed using both CTFFind4 [28] and GCTF [29], and best estimate based on confidence was selected within the Appion software package. The aligned frames and corresponding CTF allowed for monitoring of the collection process in real time.

3.7. Data processing

Data from the two sessions were processed separately and combined toward the end of the processing pipeline. For the first Falcon III dataset, movie frames were aligned using MotionCor2 [27] with 5 by 5 patches and B-factor of 500 for global alignment and 100 for local alignment through the Relion package [47][48]. Micrograph CTF estimates were imported from Appion. Ice thickness measurements were used to filter out micrographs containing ice thicker than 100 nm [30]. Template-free particle picking with Gautomatch (Kai Zhang, unpublished) using an extremely lenient threshold (to avoid missing any particles) was used to pick particles (extracted 384 box size binned to 256) that were transferred into Relion 2.1 for 2D classification. 2D class averages that were ice or showed no features were discarded, resulting in 162,271 particles. The particle stack was then brought into CryoSPARC [31] where repeated rounds of two class *ab initio* and 2D classification were used to clean up the particle stack down to 19,755 particles. The repeated rounds of two class *ab initio* classification were necessary because of the slight preferred orientation of EmbB – this allowed for the trimming of the dominant views while

retaining the less populated ones to give a more directional isotropic reconstruction. GCTF was then used to estimate per-particle CTF and the resulting particle stack was refined to 4 Å in resolution using CryoSPARC non-uniform refinement [32]. Particle polishing was then performed on the particle stack through Relion. The polished particle stack was then put through *cis*TEM [33] for CTF refinement to obtain better defocus values. The particles with the refined defocus values were then put through another round of CryoSPARC *ab initio* to further clean up the particle stack and a final non-uniform refinement produced a 3.4 Å map.

For the second K2 summit dataset, movie frames were aligned using MotionCor2 with 5 by 5 patches and B-factor of 500 for global alignment and 100 for local alignment through the Relion package. Micrograph CTF estimates were imported from Appion. Ice thickness measurements were used to filter out micrographs containing ice thicker than 100 nm. Template-free particle picking with Gautomatch using an extremely lenient threshold (to avoid missing any particles) was used to pick particles (extracted 384 box size binned to 256) that were transferred into Relion 2.1 for 2D classification. 2D class averages that were ice or showed no features were discarded, resulting in 700,201 particles. The particle stack was then brought into CryoSPARC where repeated rounds of two class *ab initio* and 2D classification were used to clean up the particle stack down to 39,702 particles, for the same rationale as stated for the first dataset. GCTF was then used to estimate per-particle CTF [29]. Particle polishing was not done for this dataset. The particle stack was then put through *cis*TEM [33] for CTF refinement to obtain better defocus values. The particles with the refined defocus values were then put through another round of CryoSPARC *ab initio* to further clean up the particle stack and a final non-uniform refinement produced a 3.3 Å map.

At this point, both datasets were combined, which was possible because of almost identical pixel size between them (0.9975 Å versus 1.0005 Å). A common pixel size value of 1.00 Å was used for this combined dataset of 57,970 particles. Non-uniform refinement in CryoSPARC produced a 3.3 Å final map, which was locally sharpened with a B-factor of −72.5 $Å^2$. Although resolution did not improve after combining, the map features look slightly better in the combined map versus the individual maps from either camera, hence the final map combined both stacks was used.

All conversions between Relion, CryoSPARC, and *cis*TEM were performed using Daniel Asarnow's pyem script (https://doi.org/10.5281/zenodo.3576630).

One millimolar of the drug ethambutol was added before vitrification for this dataset. However, when the data was collected without the drug, a 3.7 Å reconstruction was obtained and when compared to the 3.4 Å reconstruction where the drug was added, no differences were observed. Hence the higher resolution map was used for analysis and model building.

3.8. Model building and refinement

Density modification was applied to the map using phenix.resolve_cryo_em [34]. The crystallized structure of the C-terminal soluble domain of EmbC [13] was docked into EmbB with Chimera [35] and used as a starting point for model building. Coot [36] was used for manual model building. After the model was built, it was refined against the cryo-EM map utilizing real space refinement in the Phenix program [60][61]. Restraints for the lipids were generated using phenix.eLBOW and for the metal ions using phenix.ready_set. Thereafter, model adjustment and refinement were performed iteratively in Coot and Phenix, with the statistics being examined using Molprobity [39] until no further improvements were observed. Residues 1 – 20, 501 – 525, and the C-terminal purification tag had poor density and were not built in the model. The final map and model were then validated using (1) EMRinger [40] to compare a map with a model, (2) CryoSPARC's blocres implementation [41] to calculate map local resolution, (3) 3DFSC program suite [42] to calculate degree of directional resolution anisotropy through the 3DFSC, and (4) SCF program [43] to calculate the sampling compensation factor (SCF), which quantifies how inhomogeneity in Euler angle distributions contributes to attenuation of the FSC. Map-to-model FSCs were also calculated by first converting the model to a map using Chimera molmap function at Nyquist resolution (2 Å). A mask was made from this map using Relion (after low-pass filtering to 8 Å, extending by 1 pixel and applying a cosine-edge of 3 pixels), and was then applied to the density map. Map-to-model FSC was calculated using EMAN [44] proc3d between these maps.

3.9. Model analysis

A cavity search using the Solvent Extractor from Voss Volume Voxelator server [45] was performed using an outer probe radius of 5 Å and inner probe radius of 2 Å. In order to

search for other PDB structures with similar fold, a Dali server [46] search was performed, first globally and then against the different domains of the model. The Dali server was used to generate the structural conservation figures. Coot SSM superpose was used to align structures of other glycosyltransferases against EmbB. ConSurf [47] was used for generating sequence conservation data for the structure.

3.10. Mass spectrometry

EmbB was buffer exchanged into 0.2 M ammonium acetate at pH 6.8 (Sigma-Aldrich) with either 0.01% C12E8 (Anatrace) or 0.02% DDM (Anatrace) detergent using gel filtration. Native mass spectrometry (MS) analysis of EmbB in C12E8 detergent was performed using a Q-Exactive HF Orbitrap with Ultra High Mass Range modifications (Thermo Fisher Scientific) using previously described methods [48]. The HCD voltage was set to 150 V, and the capillary temperature was increased to 300 °C. Denatured intact liquid chromatography–mass spectrometry (LC-MS) was performed on EmbB in DDM using a SolariX FTICR mass spectrometer (Bruker). The online liquid chromatography (LC) separation was performed using a BioResolve RP mAB polyphenyl, 450 Å, 2.7 µM, 2.1 × 100 mm column (Waters) with the column temperature at 65 °C. The gradient was adjusted over 38 min from water to acetonitrile, each with 0.1% formic acid. The protein eluted at around 30/70 water/acetonitrile. For both native and denatured mass spectra, data were deconvolved and analyzed using UniDec [49]. Uncertainties were derived from the weighted standard deviation of masses measured at different charge states.

3.11. Screening, imaging and data analysis of glycan array

Glycan array analysis was done with *M. smegmatis* EmbB protein solubilized in DDM. Slides were prewetted in buffer A (25 mM Tris-HCl pH 7.8, 0.15 mM NaCl, 2 mM $CaCl_2$, and 0.05% Tween 20) for 5 min, rinsed with buffer B (25 mM Tris-HCl pH 7.8, 0.15 mM NaCl, and 2 mM $CaCl_2$) three times, and blocked overnight with buffer C (1% bovine serum albumin (BSA) in 25 mM Tris-HCl pH 7.8, 0.15 mM NaCl, and 2 mM $CaCl_2$) at 4 °C. Aliquots (500 µL) of serial dilutions of protein samples in buffer C were transferred to wells of the slide module immediately after aspiration of the blocking buffer. Wells

were sealed with an adhesive seal and incubated for 60 min at 37 °C. Protein was removed by aspiration, and slides were washed 10 times with buffer A and three times with buffer B. Fluorescence was measured directly or after addition of a secondary antibody in buffer C (1:1000 dilution). Slides were incubated with a secondary antibody at room temperature for 40 min before being washed repeatedly with buffer A and deionized water.

Before being scanned, slides were dried by centrifugation. Microarrays were scanned at 5-µm resolution with a GenePix 4000B scanner (Molecular Devices, Sunnyvale, CA). The fluorescent signal was detected at 532 nm for Cy3 or Alexa Fluor 555 and 488 nm for Alexa Fluor 488. The laser power was 100%, and the photomultiplier tube gain was 400. The fluorescent signals were analyzed by quantifying the pixel density (intensity) of each spot using GenePix ProMicroarray Image Analysis Software version 6.1. Fluorescence intensity values for each spot and its background were calculated. The local background signal was automatically subtracted from the signal of each separate spot, and the mean signal intensity of each spot was used for data analysis. Averages of triplicate experiments and standard deviations were calculated using Microsoft Excel.

3.12. Statistics

For calculations of Fourier shell correlations (FSC), the FSC cut-off criterion of 0.143 [50] was used. No statistical methods were used to predetermine sample size. The experiments were not randomized. The researchers were not blinded to allocation during experiments and outcome assessment.

4. Results

4.1. Cryo-EM structure of *M. smegmatis* EmbB

To better understand the function of EmbB at a molecular level, we adopted a structural genomics approach that identified the *M. smegmatis* ortholog (68% identical in *M. tuberculosis*), out of 14 screened, as a suitable candidate for structural studies based on expression levels in *E. coli* and stability in detergents compatible with structure determination [18]. After the heterologous expression and purification of EmbB from *M. smegmatis*, the protein was then reconstituted in lipid-filled nanodiscs. Once the purity was assessed by SDS-PAGE, we proceeded to observe the particles using negative stain electron microscopy (Figure 4.2).

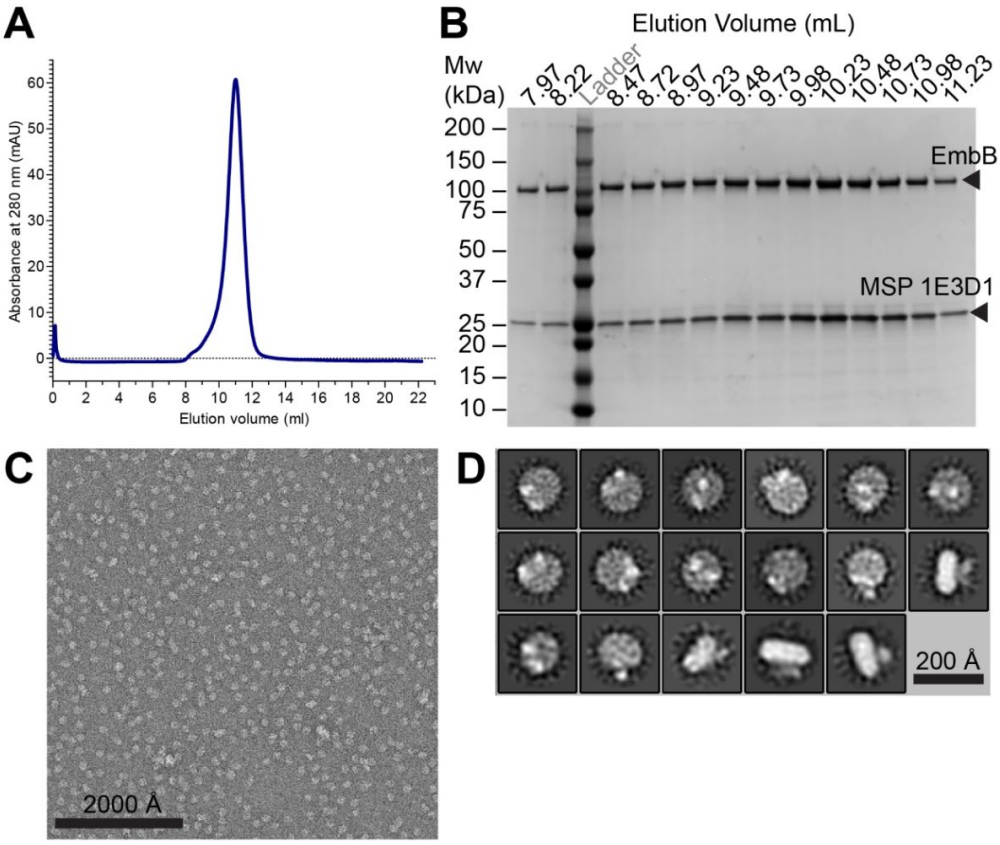


Figure 4.2 – Purification and Negative Stain Electron Microscopy of EmbB. (A) *Representative size-exclusion chromatography trace of EmbB incorporated into MSP-1E3D1 nanodiscs. Fractions corresponding to the peak were used for cryo-EM analysis.* ***(B)*** *SDS-PAGE of concentrated sample from the peak fractions stained with Coomassie blue. EmbB has a predicted molecular weight of 117 kDa and MSP-1E3D1 of 32 kDa.* ***(C)*** *Representative negative stain micrograph of EmbB.* ***(D)*** *2D negative stain class averages of EmbB.*

Using single-particle cryogenic electron microscopy (cryo-EM), we determined the structure of EmbB from *M. smegmatis* to 3.3 Å resolution (Figure 4.2, Supplementary Figures S4.3 – S4.5A, and Table S4.1).

Here, EmbB appears as a monomer, consisting of 15 TM helices and two distinct periplasmic carbohydrate binding modules (CBMs) (Figure 4.3). The first two TM helices are not found in other glycosyltransferase structures solved to date, and seem to serve to anchor the N-terminal CBM (N-CBM) to the membrane. The next 11 TM helices adopt a typical GT-C glycosyltransferase fold [51], structurally similar to enzymes from various glycosyltransferase families: mycobacterial AftD from GT53 [52], archaeal ArnT from GT83 [53], yeast Pmt1-Pmt2 from GT39 [54], and bacterial PglB from GT66 [55] (Supplementary Figure S4.6). The last two TM helices are shared only with ArnT. Thereafter, the polypeptide chain exits the membrane in the periplasm to form the second C-terminal CBM (C-CBM). The C-CBM then loops back around to complete a β-sheet with the N-CBM, likely to secure the N-terminal domain in place.

4.2. Mapping of putative active site

EmbB has a single large cavity (volume of ~1,120 $Å^3$) at the membrane–periplasm interface that encompasses juxtamembrane (JM) helix 4 and a disordered stretch of around 20 residues between JM4 and TM10 (Figure 4.4B and Supplementary Figure S4.5B). Conservation analyses reveal that this cavity contains highly conserved charged amino acid residues (D285, D286, R389, E391), where the residues corresponding to D285 and D286 are required for catalytic activity of corynebacterial EmbB (D297 and D298) [56] and mycobacterial EmbC (D293 and D294) [57] (Figure 4.4A). The putative active site has a number of negatively charged amino acids (D285, D286, and E313) which would help stabilize the carbon with a partial positive charge in the anticipated exploded S_N2-like transition state [58]. Structural alignments with the other GT-C structures show superimposition of their active site with this cavity.

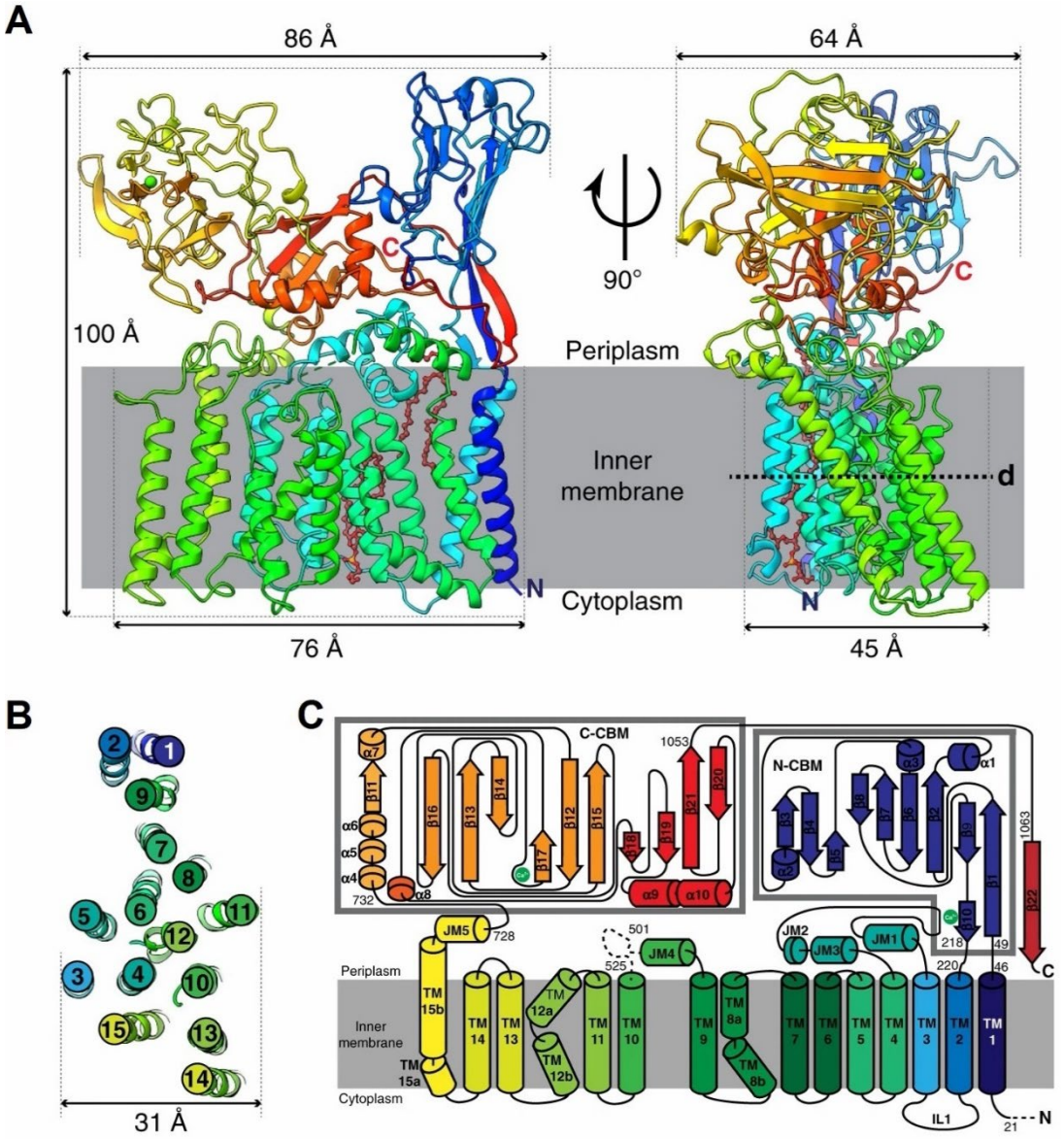


Figure 4.3 – Cryo-EM structure of EmbB. ***(A)*** *Single-particle cryo-EM structure of EmbB, rendered in cartoon form and colored in rainbow from N-terminus (blue) to C-terminus (red). A Ca^{2+} ion is shown as a green sphere. The bound lipids are represented as ball-and-sticks, colored in brown. Two orthogonal views that are perpendicular to the plane of the membrane are shown. Membrane boundaries were derived from the interface between the nanodisc lipids and solvent.* ***(B)*** *Arrangement of the TM helices of EmbB, viewed as a slice in the plane of the membrane, as indicated in (b) and magnified.* ***(C)*** *Two-dimensional topological diagram of EmbB. The two CBMs are enclosed in separate gray boxes. The topology diagram is rainbow colored from N-terminus (blue) to C-terminus (red). Unbuilt parts of the model, due to poor map density, are indicated by dotted lines. Bound Ca^{2+} atoms are shown as green circles.*

For instance, in PglB, both the donor, the acceptor and a catalytic Mn^{2+} localize here (Figure 4.4A, 4.4C). Based on this structural comparison, the lipidic donor DPA is likely to bind in the pocket formed by TM helices 7–9, on the right side of the cavity, with the soluble acceptor substrate binding to the left. As this structure was determined in the absence of any bound ligands, we expect the disordered residues between JM4 and TM10 to become ordered upon substrate binding, akin to what was shown for the PglB EL5 [55] and ArnT PL4 loops [53] (Supplementary Figure S4.6E, S4.6F).

While active site mutants in EmbB result in suppressed bacterial growth [56], a series of mutations (N630, W633, P641, N644, K648) have been shown to retain enzymatic activity yet reduce incorporation of arabinose, resulting in the formation of a truncated AG [56]. Similar findings have been reported for LAM in EmbC [20][22]. All these residues map to the loop between TM helices 13 and 14, situated at the entrance of the cavity for the putative sugar acceptor (Figure 4.4D), suggesting a role for it in regulating the oligosaccharides that can act as acceptors for EmbB [19][20].

4.3. Presence of tightly bound lipids in EmbB

We observe two bound lipids in our structure, in a pocket formed by TM helices 2, 5, 6, 7, and 9 on opposite leaflets (Figure 4.3C). The lipid on the outer leaflet appears to have a cation bound to its head group, mediating extensive interactions with the backbone and side chains of residues from TM2, JM1, and β10 (Figure 4.4E, Supplementary Figure S4.6G). To identify these lipids, we performed native mass spectrometry (MS) of EmbB, which showed that EmbB appeared as a dimer when solubilized in detergent C12E8 with a series of bound molecules with masses around 300-350 Da (Supplementary Figure S4.7A). In addition, there were two peaks that showed larger abundances, and we hypothesize that these corresponded to two or three bound lipid molecules with masses of ~750 Da (Supplementary Figure S4.7A). To remove bound adducts, we also performed denaturated LC-MS analysis. Under denaturing conditions, EmbB retained a tightly bound calcium ion (Supplementary Figure S4.7B). An additional peak was observed with 749 ± 21 Da mass, which also partially retained a calcium ion. This mass is consistent with a bound phosphatidylglycerol (PGly), which is also present in mycobacterial inner membranes (average molecular weight 761.073 Da) [23][24]. These lipids (or equivalent ones endogenously) may be important in stabilizing the protein structure (Supplementary Figure S4.6G).

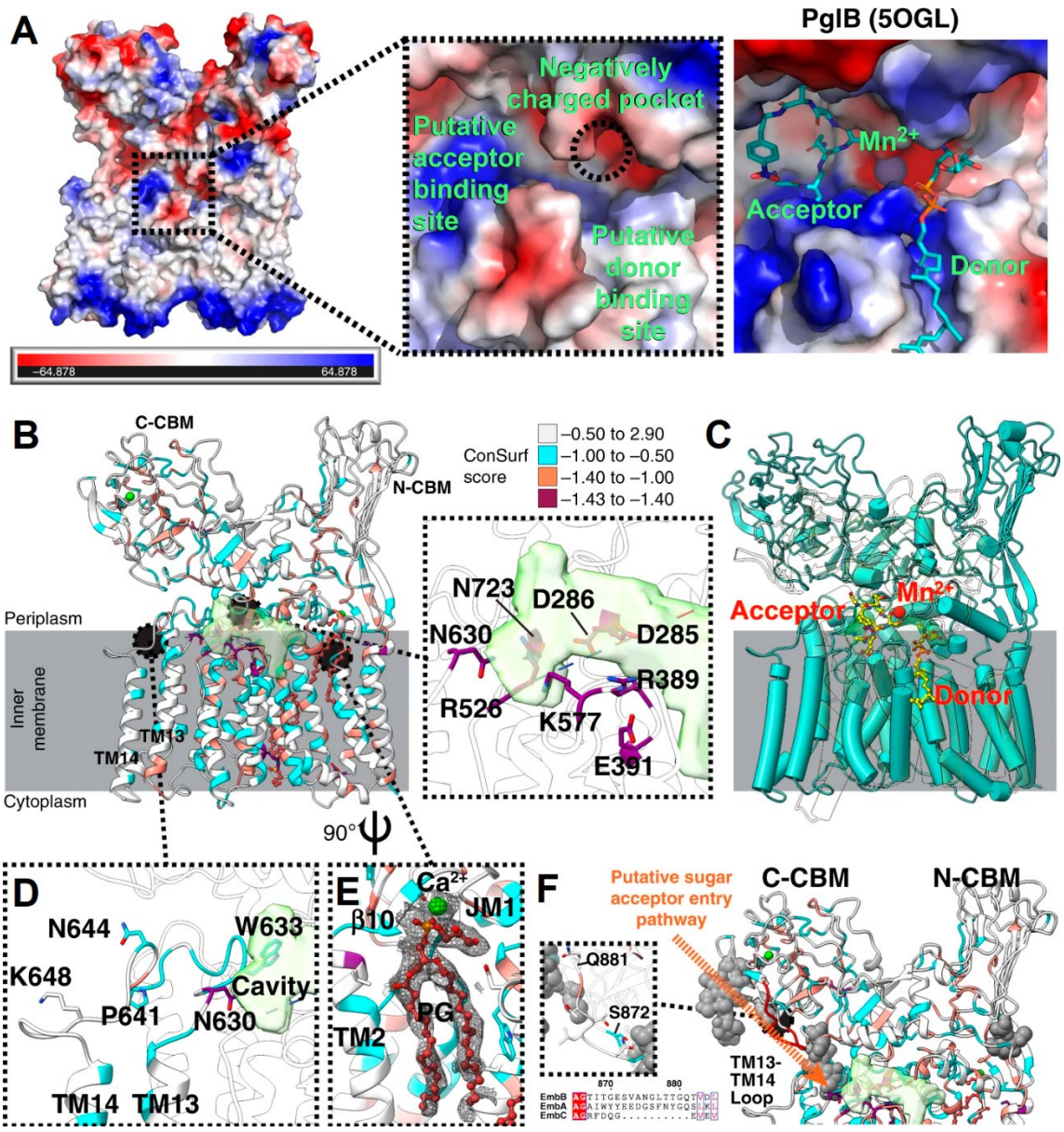


Figure 4.4 – Structural features of EmbB. ***(A)*** *Electrostatic representation of EmbB, with a zoom-in around the putative active site, labeled according to where the substrates are likely to bind, where red is more negatively charged and blue more positively charged. A comparison with the PglB active site with bound substrates is also shown, after PglB was aligned against EmbB.* ***(B)*** *Structure of EmbB, rendered in cartoon and colored based on ConSurf* [47] *score for sequence conservation (the more negative the score, the more conserved the residue). The putative active site cavity, generated by the Voss Volume Voxelator server* [45] *is colored in semi-transparent green. The insert shows the putative active site cavity with the strictly conserved residues labeled.* ***(C)*** *EmbB (pale blue) is superimposed on PglB (semi-transparent yellow), with the ligands and Mn^{2+} ion of PglB shown in yellow as sticks and a ball, respectively.* ***(D)*** *Residues that are known to maintain catalytic activity while altering substrate specificity on a loop between TM13 and TM14 are labeled and side chains are shown.* ***(E)*** *A tightly bound phosphatidylglycerol (PGly) shown as ball-and-stick) and calcium ion (shown as a ball) in a pocket between TM2, JM1, and β10 is shown. The density map of the lipid and the ion is displayed as mesh.* ***(F)*** *Glycan ligands for the top ten Dali server hits for both N-CBM and C-CBM were mapped onto the structure as a gray ball representation. 2WJS [10.2210/pdb2WJS/pdb] (PDB ID) was used for N-CBM, while 3PTY [10.2210/pdb3PTY/pdb] and 4GWM [10.2210/pdb4GWM/pdb] were used for the C-CBM. The missing loop in EmbC is colored in red, and the insert is a zoomed-in view. Sequence alignment of the region around the loop is appended below the insert. The putative sugar acceptor entry pathway is shown as an orange dotted line.*

4.4. EmbB has two carbohydrate binding modules

In the periplasmic region, the two CBMs exhibit β-sandwich folds. Using the top ten hits from the Dali server [46], structurally similar CBMs were aligned with the two CBMs of EmbB. This allowed us to interrogate the possible glycan-binding locations; potential substrates that were sterically hindered in the EmbB structure were discarded. As expected, the EmbB C-CBM structure is highly homologous to the corresponding CBM in EmbC (PDB ID: 3PTY) [13], with an RMSD of 1.2 Å. N-CBM's structural homologs bind to at most one monosaccharide unit, whereas structural homologs of the C-CBM reveal binding to more complex glycans, corroborating what was previously observed for the EmbC C-CBM [13]. The co-crystallized Ara-(1 → 5)-Ara-O-C8 ligand for the EmbC C-CBM maps to the loop of EmbB between TM helices 13 and 14, which we earlier proposed to control access of the acceptor to the active site (Figure 4.4F). This suggests a pathway for the binding of the acceptor in the active site exclusively via the C-CBM (Figure 4.4F). By screening purified EmbB against a synthetic array of mycobacterial glycan fragments [62], we found that the protein preferentially binds highly-branched arabinose-containing oligosaccharides (Supplementary Figure S4.8), supporting the hypothesis that the C-CBM binds to multiple monosaccharide residues.

A comparison of EmbA, EmbB, and EmbC sequences revealed that a loop region spanning from S872 to Q881 is missing in EmbC (Figure 4.4F and Supplementary Figure S4.2). In EmbB, this loop is located directly above the TM13-14 loop and is part of the C-CBM; it is in the path we propose the acceptor might bind. This could explain the different substrate specificities reported of the Emb family members: EmbA and EmbB act specifically on AG, while LAM is a substrate of EmbC, even though all enzymes catalyze the same reaction: addition of an arabinose residue α-(1→5) to an existing arabinan chain [1].

5. Discussion

5.1. Ethambutol resistance mutations map to putative active site

The structure of EmbB presents the unique opportunity to spatially map out data resulting from decades of known ethambutol resistance mutations in both EmbB and EmbC [9][27 – 30]. Focusing only on residues conserved between *M. tuberculosis* and *M. smegmatis* (Supplementary Tables S4.2 and S4.3), we found that mutations causing resistance all cluster around the putative active site (Figure 4.5A). Notably, these mutations are closer to the putative DPA binding site, suggesting that ethambutol might interfere with recruitment of the arabinose donor, thereby inhibiting enzyme function.

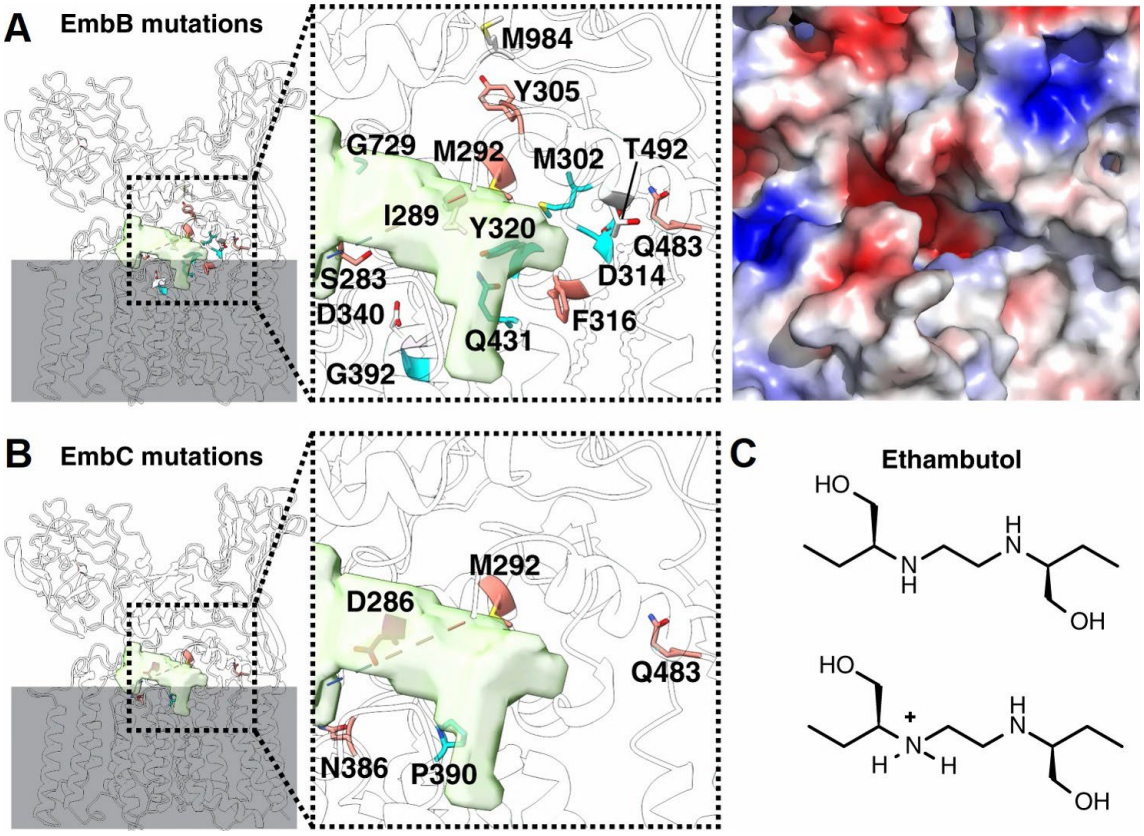


***Figure 4.5** – **Mutations known to confer resistance to ethambutol in EmbB (A)** and EmbC **(B)** are mapped on the overall structure and as a zoom-in, with their side chains displayed and colored based on the ConSurf score. The other regions of the EmbB model are rendered semi-transparent. The electrostatic potential of the zoom-in region of EmbB is shown on the right. **(C)** Chemical structure of (S, S)-ethambutol, showing both its neutral and one of two single positively charged forms.*

Most mutated residues are not highly conserved, following an evolutionary logic in terms of maintaining structural integrity and function, and also providing a template to predict residues that might be susceptible to drug-induced mutations in the future. Based on the reported values of pK_a for ethambutol (pK_{a1} = 6.35, pK_{a2} = 9.3) [67], the drug is expected to be positively charged at physiological pH, suggesting that ionic interactions are involved in drug binding (Figure 4.5C). Indeed, many of the mutations are conversions of negatively charged residues into uncharged ones (D314G/Y, D340A), or of uncharged residues into positively charged ones (Q431R, Q483K, T492R, M984R) (Figure 4.5A). The only two mutations decreasing the protein overall net charge (G392D, G729D) are from residues located at the periphery of the cavity. Moreover, the homology between EmbB and EmbC allow us to map onto the EmbB structure several EmbC mutations that contributed to drug resistance (Figure 4.5B and Supplementary Tables S4.2 and S4.3). Similarly, these mutations cluster around the putative DPA binding site. Surprisingly though, the highly conserved D286 residue is also involved in drug resistance in EmbC, which could be explained by the fact that D286G mutation reduces but does not abolish the catalytic activity of EmbC [65].

5.2. Comparison with heterodimeric Emb structures

Recently, the structures of heterodimeric *M. smegmatis* and *M. tuberculosis* EmbA, EmbB, and EmbC were determined [14]. *M. smegmatis* EmbB was solved as a dimer with EmbA, bound to either ethambutol or di-arabinofuranose (Supplementary Figure S4.9). Compared with our monomeric apo-structure, the disordered residues between JM4 and TM10 become ordered and the rest of the structure is very similar (Supplementary Figure S4.9C). Whether the ordering of these residues is caused by the addition of substrates or hetero-dimerization is yet to be determined. Notably, the dimeric *M. smegmatis* EmbA-EmbB was overexpressed endogenously and surprisingly had meromycolate extension acyl carrier protein (AcpM) bound, the same AcpM that is also associated to mycobacterial arabinofuranosyltransferase AftD [52]. AcpM in the EmbA-EmbB dimer structure extends its 4′-phosphopantetheine into the same pocket as that of inner leaflet PG present in our monomeric EmbB structure (Figure 4.3A, Supplementary Figure S4.9B). The lack of AcpM did not cause any significant conformational changes in the structure of EmbB around the AcpM binding site, suggesting that AcpM might not have a critical functional role, contrasting to AftD (Supplementary Figure S4.9C).

6. Conclusions

In conclusion, we report the full-length monomeric structure of a mycobacterial arabinosyltransferase from the Emb family. The structure was obtained by cryo-EM close to the native environment by its incorporation into a lipid-filled nanodisc, and the data show that EmbB has a conserved GT-C fold. Analysis of the structure allowed us to map the putative active site as well as substrate binding sites. We localized mutations that maintain catalytic activity while altering substrate specificity to a loop between TM13 and TM14, juxtaposed to the putative active site, which we propose controls the access of the acceptor to the active site. We also found in the density map a tightly bound phosphatidylglycerol lipid and calcium cation, which are likely to serve a structural purpose, and their presence and identity were confirmed using native and denaturing mass spectrometry. The mapping of known ethambutol mutations on the structure suggests that this drug binds in close proximity of the putative active site, providing a framework to better understand, if not predict, resistance-causing loci. Finally, our work provides a template for future structure-based drug design efforts aimed at enhancing the efficacy of this front-line drug used in the treatment against tuberculosis (*M. tuberculosis* [68], *M. bovis* [69], *M. microti* [70]) and NTM disease [68] (*M. avium*, *M. kansaii* [71]). This is of particular importance in the context of increasingly frequent infections with drug-resistant strains of *M. tuberculosis* [9] and other disease-causing mycobacteria [32][35 – 37]. Our solved structure of EmbB is from *M. smegmatis*, a model species for the entire mycobacteria family [73] that is non-pathogenic, therefore not all observations might be directly extrapolated to the aforementioned pathogenic mycobacterial species, but should instead serve as a guide for future studies of this family of enzymes.

7. References

[1] M. Jankute, J. A. Cox, J. Harrison, and G. S. Besra, “Assembly of the Mycobacterial Cell Wall” *Annu Rev Microbiol*, vol. 69, pp. 405–423, 2015

[2] K. A. Abrahams and G. S. Besra, “Mycobacterial cell wall biosynthesis: a multifaceted antibiotic target” *Parasitology*, vol. 145, no. 2, pp. 116–133, Feb. 2018

[3] A. E. Grzegorzewicz *et al.*, “Assembling of the *Mycobacterium tuberculosis* cell wall core” *J. Biol. Chem.*, vol. 291, no. 36, pp. 18867–18879, Sep. 2016

[4] B. A. Wolucka, M. R. McNeil, E. De Hoffmann, T. Chojnacki, and P. J. Brennan, “Recognition of the lipid intermediate for arabinogalactan/arabinomannan biosynthesis and its relation to the mode of action of ethambutol on mycobacteria” *J. Biol. Chem.*, vol. 269, no. 37, pp. 23328–23335, Sep. 1994

[5] C. L. Dulberger, E. J. Rubin, and C. C. Boutte, “The mycobacterial cell envelope — a moving target” *Nat. Rev. Microbiol.*, vol. 18, no. 1, pp. 47–59, Jan. 2020

[6] A. K. Mishra, N. N. Driessen, B. J. Appelmelk, and G. S. Besra, “Lipoarabinomannan and related glycoconjugates: structure, biogenesis and role in *Mycobacterium tuberculosis* physiology and host-pathogen interaction” *FEMS Microbiol. Rev.*, vol. 35, no. 6, pp. 1126–1157, Nov. 2011

[7] A. Telenti *et al.*, “The emb operon, a gene cluster of *Mycobacterium tuberculosis* involved in resistance to ethambutol” *Nat. Med.*, vol. 3, no. 5, pp. 567–570, 1997

[8] V. E. Escuyer *et al.*, “The role of the embA and embB gene products in the biosynthesis of the terminal hexaarabinofuranosyl motif of *Mycobacterium smegmatis* arabinogalactan” *J Biol Chem*, vol. 276, no. 52, pp. 48854–48862, 2001

[9] J. A. Caminero, G. Sotgiu, A. Zumla, and G. B. Migliori, “Best drug treatment for multidrug-resistant and extensively drug-resistant tuberculosis” *Lancet. Infect. Dis.*, vol. 10, no. 9, pp. 621–629, 2010

[10] S. K. Field and R. L. Cowie, “Treatment of *Mycobacterium avium-intracellulare* complex lung disease with a macrolide, ethambutol, and clofazimine” *Chest*, vol. 124, no. 4, pp. 1482–1486, 2003

[11] H. Safi *et al.*, “Evolution of high-level ethambutol-resistant tuberculosis through

interacting mutations in decaprenylphosphoryl-beta-D-arabinose biosynthetic and utilization pathway genes" *Nat Genet*, vol. 45, no. 10, pp. 1190–1197, 2013

[12] A. G. Amin, R. Goude, L. Shi, J. Zhang, D. Chatterjee, and T. Parish, "EmbA is an essential arabinosyltransferase in *Mycobacterium tuberculosis*" *Microbiology*, vol. 154, no. Pt 1, pp. 240–248, 2008

[13] L. J. Alderwick *et al.*, "The C-terminal domain of the Arabinosyltransferase *Mycobacterium tuberculosis* EmbC is a lectin-like carbohydrate binding module" *PLoS Pathog*, vol. 7, no. 2, p. e1001299, 2011

[14] L. Zhang and Y. Zhao, "Structures of cell wall arabinosyltransferases with the anti-tuberculosis drug ethambutol" *Science (80-.).*, vol. 368, no. 6496, pp. 1211–1219, 2020

[15] A. Kapopoulou, J. M. Lew, and S. T. Cole, "The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes" *Tuberculosis (Edinb).*, vol. 91, no. 1, pp. 8–13, Jan. 2011

[16] F. Sievers *et al.*, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega" *Mol. Syst. Biol.*, vol. 7, 2011

[17] X. Robert and P. Gouet, "Deciphering key features in protein structures with the new ENDscript server" *Nucleic Acids Res.*, vol. 42, no. Web Server issue, Jul. 2014

[18] J. Love *et al.*, "The New York Consortium on Membrane Protein Structure (NYCOMPS): A high-throughput platform for structural genomics of integral membrane proteins" *J. Struct. Funct. Genomics*, vol. 11, no. 3, pp. 191–199, Sep. 2010

[19] R. Bruni and B. Kloss, "High-Throughput Cloning and Expression of Integral Membrane Proteins in *Escherichia coli*" *Curr. Protoc. Protein Sci.*, vol. 74, no. 1, pp. 29.6.1-29.6.34, Nov. 2013

[20] L. Stols *et al.*, "New vectors for co-expression of proteins: structure of Bacillus subtilis ScoAB obtained by high-throughput protocols" *Protein Expr. Purif.*, vol. 53, no. 2, pp. 396–403, Jun. 2007

[21] T. H. Bayburt and S. G. Sligar, "Membrane protein assembly into Nanodiscs" *FEBS Lett.*, vol. 584, no. 9, pp. 1721–1727, May 2010

[22] C. Suloway *et al.*, "Automated molecular microscopy: the new Leginon system" *J.*

Struct. Biol., vol. 151, no. 1, pp. 41–60, Jul. 2005

[23] G. C. Lander *et al.*, “Appion: An integrated, database-driven pipeline to facilitate EM image processing” *J. Struct. Biol.*, vol. 166, no. 1, pp. 95–102, Apr. 2009

[24] S. H. W. Scheres, “RELION: Implementation of a Bayesian approach to cryo-EM structure determination” *J. Struct. Biol.*, vol. 180, no. 3, pp. 519–530, Dec. 2012

[25] D. Kimanius, B. O. Forsberg, S. H. W. Scheres, and E. Lindahl, “Accelerated cryo-EM structure determination with parallelisation using GPUS in RELION-2” *Elife*, vol. 5, Nov. 2016

[26] A. Cheng *et al.*, “High resolution single particle cryo-electron microscopy using beam-image shift” *J. Struct. Biol.*, vol. 204, no. 2, pp. 270–275, Nov. 2018

[27] S. Q. Zheng, E. Palovcak, J. P. Armache, K. A. Verba, Y. Cheng, and D. A. Agard, “MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy” *Nat. Methods 2017 144*, vol. 14, no. 4, pp. 331–332, Feb. 2017

[28] A. Rohou and N. Grigorieff, “CTFFIND4: Fast and accurate defocus estimation from electron micrographs” *J. Struct. Biol.*, vol. 192, no. 2, pp. 216–221, Nov. 2015

[29] K. Zhang, “Gctf: Real-time CTF determination and correction” *J. Struct. Biol.*, vol. 193, no. 1, pp. 1–12, Jan. 2016

[30] W. J. Rice *et al.*, “Routine determination of ice thickness for cryo-EM grids” *J. Struct. Biol.*, vol. 204, no. 1, pp. 38–44, Oct. 2018

[31] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker, “cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination” *Nat. Methods 2017 143*, vol. 14, no. 3, pp. 290–296, Feb. 2017

[32] A. Punjani, H. Zhang, and D. J. Fleet, “Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction” *Nat. Methods 2020 1712*, vol. 17, no. 12, pp. 1214–1221, Nov. 2020

[33] T. Grant, A. Rohou, and N. Grigorieff, “CisTEM, user-friendly software for single-particle image processing” *Elife*, vol. 7, Mar. 2018

[34] T. C. Terwilliger, S. J. Ludtke, R. J. Read, P. D. Adams, and P. V. Afonine, “Improvement of cryo-EM maps by density modification” *Nat. Methods 2020 179*, vol. 17, no. 9, pp. 923–927, Aug. 2020

[35] E. F. Pettersen *et al.*, “UCSF Chimera—A visualization system for exploratory research and analysis” *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004

[36] P. Emsley and K. Cowtan, “Coot: model-building tools for molecular graphics”, vol. 60, no. 12, pp. 2126–2132, Nov. 2004

[37] P. D. Adams *et al.*, “PHENIX: a comprehensive Python-based system for macromolecular structure solution”, vol. 66, no. 2, pp. 213–221, Jan. 2010

[38] P. V. Afonine *et al.*, “Real-space refinement in PHENIX for cryo-EM and crystallography”, vol. 74, no. 6, pp. 531–544, May 2018

[39] V. B. Chen *et al.*, “MolProbity: all-atom structure validation for macromolecular crystallography”, vol. 66, no. 1, pp. 12–21, Dec. 2009

[40] B. A. Barad *et al.*, “EMRinger: side chain–directed model and map validation for 3D cryo-electron microscopy” *Nat. Methods 2015 1210*, vol. 12, no. 10, pp. 943–946, Aug. 2015

[41] G. Cardone, J. B. Heymann, and A. C. Steven, “One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions” *J. Struct. Biol.*, vol. 184, no. 2, pp. 226–236, Nov. 2013

[42] Y. Zi Tan *et al.*, “Addressing preferred specimen orientation in single-particle cryo-EM through tilting” *Nat. Methods 2017 148*, vol. 14, no. 8, pp. 793–796, Jul. 2017

[43] P. R. Baldwin and D. Lyumkis, “Non-uniformity of projection distributions attenuates resolution in Cryo-EM” *Prog. Biophys. Mol. Biol.*, vol. 150, pp. 160–183, Jan. 2020

[44] S. J. Ludtke, P. R. Baldwin, and W. Chiu, “EMAN: Semiautomated Software for High-Resolution Single-Particle Reconstructions” *J. Struct. Biol.*, vol. 128, no. 1, pp. 82–97, Dec. 1999

[45] N. R. Voss and M. Gerstein, “3V: cavity, channel and cleft volume calculator and extractor” *Nucleic Acids Res.*, vol. 38, no. suppl_2, pp. W555–W562, Jul. 2010

[46] L. Holm and L. M. Laakso, “Dali server update” *Nucleic Acids Res.*, vol. 44, no. W1, pp. W351–W355, Jul. 2016

[47] H. Ashkenazy *et al.*, “ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules” *Nucleic Acids Res.*, vol. 44, no. W1, pp. W344–W350, Aug. 2016

[48] J. A. Townsend, J. E. Keener, Z. M. Miller, J. S. Prell, and M. T. Marty, “Imidazole Derivatives Improve Charge Reduction and Stabilization for Native Mass Spectrometry” *Anal. Chem.*, vol. 91, no. 22, pp. 14765–14772, Nov. 2019

[49] M. T. Marty, A. J. Baldwin, E. G. Marklund, G. K. A. Hochberg, J. L. P. Benesch, and C. V. Robinson, “Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles” *Anal. Chem.*, vol. 87, no. 8, pp. 4370–4376, Apr. 2015

[50] P. B. Rosenthal and R. Henderson, “Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy” *J. Mol. Biol.*, vol. 333, no. 4, pp. 721–745, Oct. 2003

[51] J. Liu and A. Mushegian, “Three monophyletic superfamilies account for the majority of the known glycosyltransferases” *Protein Sci.*, vol. 12, no. 7, pp. 1418–1431, Jul. 2003

[52] Y. Z. Tan *et al.*, “Cryo-EM Structures and Regulation of Arabinofuranosyltransferase AftD from Mycobacteria” *Mol. Cell*, vol. 78, no. 4, pp. 683-699.e11, 2020

[53] V. I. Petrou *et al.*, “Structures of aminoarabinose transferase ArnT suggest a molecular basis for lipid A glycosylation” *Science (80-.).*, vol. 351, no. 6273, pp. 608–612, 2016

[54] L. Bai, A. Kovach, Q. You, A. Kenny, and H. Li, “Structure of the eukaryotic protein O-mannosyltransferase Pmt1−Pmt2 complex” *Nat. Struct. Mol. Biol. 2019 268*, vol. 26, no. 8, pp. 704–711, Jul. 2019

[55] M. Napiórkowska *et al.*, “Molecular basis of lipid-linked oligosaccharide recognition and processing by bacterial oligosaccharyltransferase” *Nat. Struct. Mol. Biol. 2017 2412*, vol. 24, no. 12, pp. 1100–1106, Oct. 2017

[56] M. Seidel, L. J. Alderwick, H. Sahm, G. S. Besra, and L. Eggeling, “Topology and mutational analysis of the single Emb arabinofuranosyltransferase of *Corynebacterium glutamicum* as a model of Emb proteins of *Mycobacterium tuberculosis*” *Glycobiology*, vol. 17, no. 2, pp. 210–219, Feb. 2007

[57] A. Korkegian, D. M. Roberts, R. Blair, and T. Parish, “Mutations in the essential arabinosyltransferase EmbC lead to alterations in *Mycobacterium tuberculosis* lipoarabinomannan” *J Biol Chem*, vol. 289, no. 51, pp. 35172–35181, 2014

[58] L. L. Lairson, B. Henrissat, G. J. Davies, and S. G. Withers, “Glycosyltransferases:

Structures, Functions, and Mechanisms", vol. 77, pp. 521–555, Jun. 2008

[59] S. Berg *et al.*, "Roles of conserved proline and glycosyltransferase motifs of EmbC in biosynthesis of lipoarabinomannan" *J. Biol. Chem.*, vol. 280, no. 7, pp. 5651–5663, Feb. 2005

[60] L. Chiaradia *et al.*, "Dissecting the mycobacterial cell envelope and defining the composition of the native mycomembrane" *Sci. Rep.*, vol. 7, no. 1, p. 12807, 2017

[61] P. J. Brennan and H. Nikaido, "The envelope of mycobacteria" *Annu. Rev. Biochem.*, vol. 64, pp. 29–63, 1995

[62] R. B. Zheng *et al.*, "Insights into Interactions of Mycobacteria with the Host Innate Immune System from a Novel Array of Synthetic Mycobacterial Glycans" *ACS Chem. Biol.*, vol. 12, no. 12, pp. 2990–3002, Dec. 2017

[63] S. Sreevatsan *et al.*, "Ethambutol resistance in *Mycobacterium tuberculosis*: critical role of embB mutations" *Antimicrob. Agents Chemother.*, vol. 41, no. 8, pp. 1677–1681, 1997

[64] S. V. Ramaswamy *et al.*, "Molecular genetic analysis of nucleotide polymorphisms associated with ethambutol resistance in human isolates of *Mycobacterium tuberculosis*" *Antimicrob. Agents Chemother.*, vol. 44, no. 2, pp. 326–336, Feb. 2000

[65] R. Goude, A. G. Amin, D. Chatterjee, and T. Parish, "The arabinosyltransferase EmbC is inhibited by ethambutol in *Mycobacterium tuberculosis*" *Antimicrob Agents Chemother*, vol. 53, no. 10, pp. 4138–4146, 2009

[66] M. A. Lety, S. Nair, P. Berche, and V. Escuyer, "A single point mutation in the embB gene is responsible for resistance to ethambutol in *Mycobacterium smegmatis*" *Antimicrob. Agents Chemother.*, vol. 41, no. 12, pp. 2629–2633, 1997

[67] W. H. Beggs and F. A. Andrews, "Chemical characterization of ethambutol binding to *Mycobacterium smegmatis*" *Antimicrob. Agents Chemother.*, vol. 5, no. 3, pp. 234–239, 1974

[68] J. Richard J. Wallace, R. O'Brien, J. Glassroth, J. Raleigh, and A. Dutt, "Diagnosis and Treatment of Disease Caused by Nontuberculous Mycobacteria", vol. 142, no. 4, pp. 940–953, Dec. 2012

[69] Z. Lan, M. Bastos, and D. Menzies, "Treatment of human disease due to *Mycobacterium bovis*: a systematic review" *Eur. Respir. J.*, vol. 48, no. 5, pp.

1500–1503, Nov. 2016

[70] G. Panteix *et al.*, “Pulmonary tuberculosis due to *Mycobacterium microti*: a study of six recent cases in France” *J. Med. Microbiol.*, vol. 59, no. Pt 8, pp. 984–989, Aug. 2010

[71] N. Rastogi, K. S. Goh, A. Bryskier, and A. Devallois, “Spectrum of activity of levofloxacin against nontuberculous mycobacteria and its activity against the *Mycobacterium avium* complex in combination with ethambutol, rifampin, roxithromycin, amikacin, and clofazimine” *Antimicrob. Agents Chemother.*, vol. 40, no. 11, pp. 2483–2487, 1996

[72] D. Van Soolingen *et al.*, “A novel pathogenic taxon of the *Mycobacterium tuberculosis* complex, Canetti: Characterization of an exceptional isolate from Africa” *Int. J. Syst. Bacteriol.*, vol. 47, no. 4, pp. 1236–1245, Oct. 1997

[73] M. U. Shiloh and P. A. DiGiuseppe Champion, “To catch a killer. What can mycobacterial models teach us about *Mycobacterium tuberculosis* pathogenesis?” *Curr. Opin. Microbiol.*, vol. 13, no. 1, pp. 86–92, Feb. 2010

[74] Q. Sun *et al.*, “Mutations within embCAB Are Associated with Variable Level of Ethambutol Resistance in *Mycobacterium tuberculosis* Isolates from China” *Antimicrob. Agents Chemother.*, vol. 62, no. 1, Jan. 2018

8. Acknowledgments

Author Contributions

Conceptualization: F. Mancia and O. B. Clarke

Genomics expansion and small-scale screening: A. L. Rosário, B. Kloss and *J. Rodrigues*

Large-scale expression and purification: Y. Z. Tan and S. I. Giacometti

Negative-stain EM and Cryo-EM: Y. Z. Tan

Model building: Y. Z. Tan and O. B. Clarke

Mass spectroscopy: J. E. Keener and M. T. Marty

Glycan array experiments: R. B. Zheng (**Supervision**: T. L. Lowary)

Acceptor substrate synthesis: R. Brunton (**Supervision**: T. L. Lowary)

Project advisors: L. Zhang, M. Niederweis, M. Archer

EM analysis supervision: B. Carragher and C. S. Potter

Project supervision: F. Mancia

Manuscript preparation: Y. Z. Tan and F. Mancia with input from all authors.

Review and editing: all authors

We thank Christina Chen for help with nanodisc incorporation, Khuram Ashraf for help with expression optimization, Jonathan Kim for help with purifications, Vasileios Petrou for helpful discussions and Leora Hamberger for her assistance managing the Mancia laboratory (Columbia University). We thank Ed Eng, Bill Rice, Laura Kim, Mikhail Kopylov, and Kelsey Jordan (New York Structural Biology Center, Simons Electron Microscopy Center) for help with microscope setup. We thank Sargis Dallakyan, Carl Negro, Shaker Krit and Swapnil Bhatkar (New York Structural Biology Center, Simons Electron Microscopy Center) for computation support. We thank Dr. Krishna Parsawar and the University of Arizona Analytical & Biological Mass Spectrometry Facility for help with the FTICR mass spectrometer.

Funding

This work was supported by grants from NIH (R35 GM128624 to M.T.M., R01 GM111980, R35 GM132120, and R21 AI119672 to F.M., P41 GM103310 and OD019994 to C.S.P. and B.C.), Agency for Science, Technology and Research Singapore (to Y.Z.T.), University of Alabama at Birmingham (to M.N.), Fundação para a Ciência e Tecnologia, Portugal (PD/BD/128261/2016 to J.R.; PTDC/BIA-BQM/30421/2017 and IF/00656/2014 to M.A.), EU H2020 research and innovation program under the Marie Skłodowska-Curie grant No 823780 and Instruct-ULTRA (No 731005), an EUH2020 project to further develop the services of Instruct-ERIC (to J.R. and M.A.), Simons Foundation (SF349247 to C.S.P., B.C.), NYSTAR (to C.S.P., B.C.), Agouron Institute (F00316 to C.S.P., B.C.) and the Canadian Glycomics Network (to T.L.L.). Some of the work was performed at the Center for Membrane Protein Production and Analysis (COMPPÅ; P41 GM116799 to Wayne Hendrickson) and at the National Resource for Automated Molecular Microscopy at the Simons Electron Microscopy Center (P41 GM103310), both located at the New York Structural Biology Center. M.A. acknowledges MostMicro Research Unit (financially supported by LISBOA-01-0145-FEDER-007660 funded by FEDER funds through COMPETE2020 and by national funds through FCT), and iNOVA4Health (LISBOA-01-0145-FEDER-007344, co-funded by FEDER under PT2020).

9. Supplemental Information

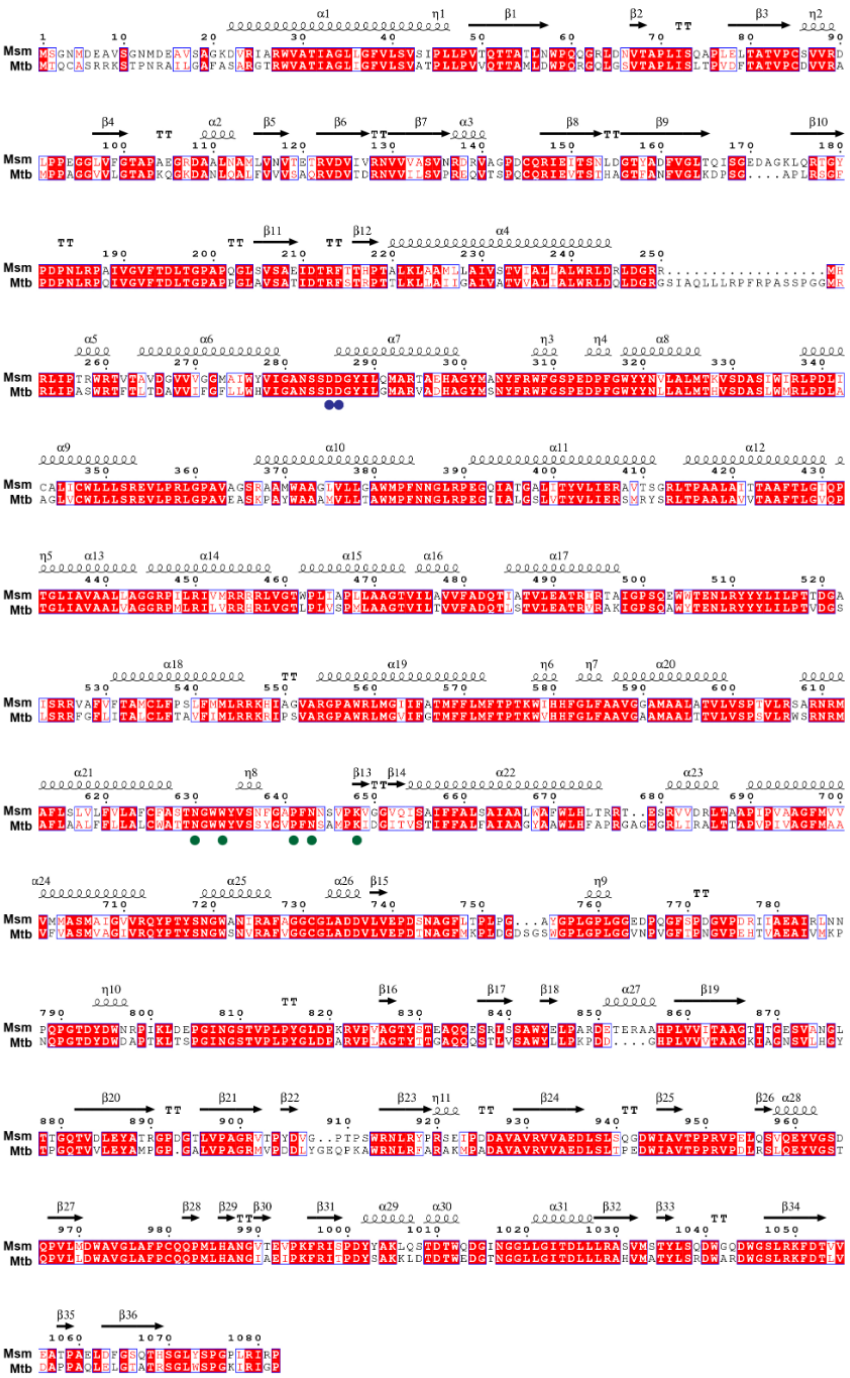


Figure S4.1 –Sequence alignment of EmbB from M. smegmatis versus M. tuberculosis.
The sequence alignment between M. smegmatis EmbB (MSMEG_6389) and M. tuberculosis EmbB (Rv3795) was generated using ClustalO [16] *and displayed using ESPript* [17] *. Secondary structure annotation based on the EmbB structure is displayed above the sequences. Identical residues are depicted in a red-filled box with white font, whereas residues with more than 70% similarity are marked in a blue bordered box with red font. Blue dots indicate the two catalytic aspartic acid residues, and green dots indicate the residues that when mutated retain enzymatic activity in the enzyme yet reduce incorporation of arabinose.*

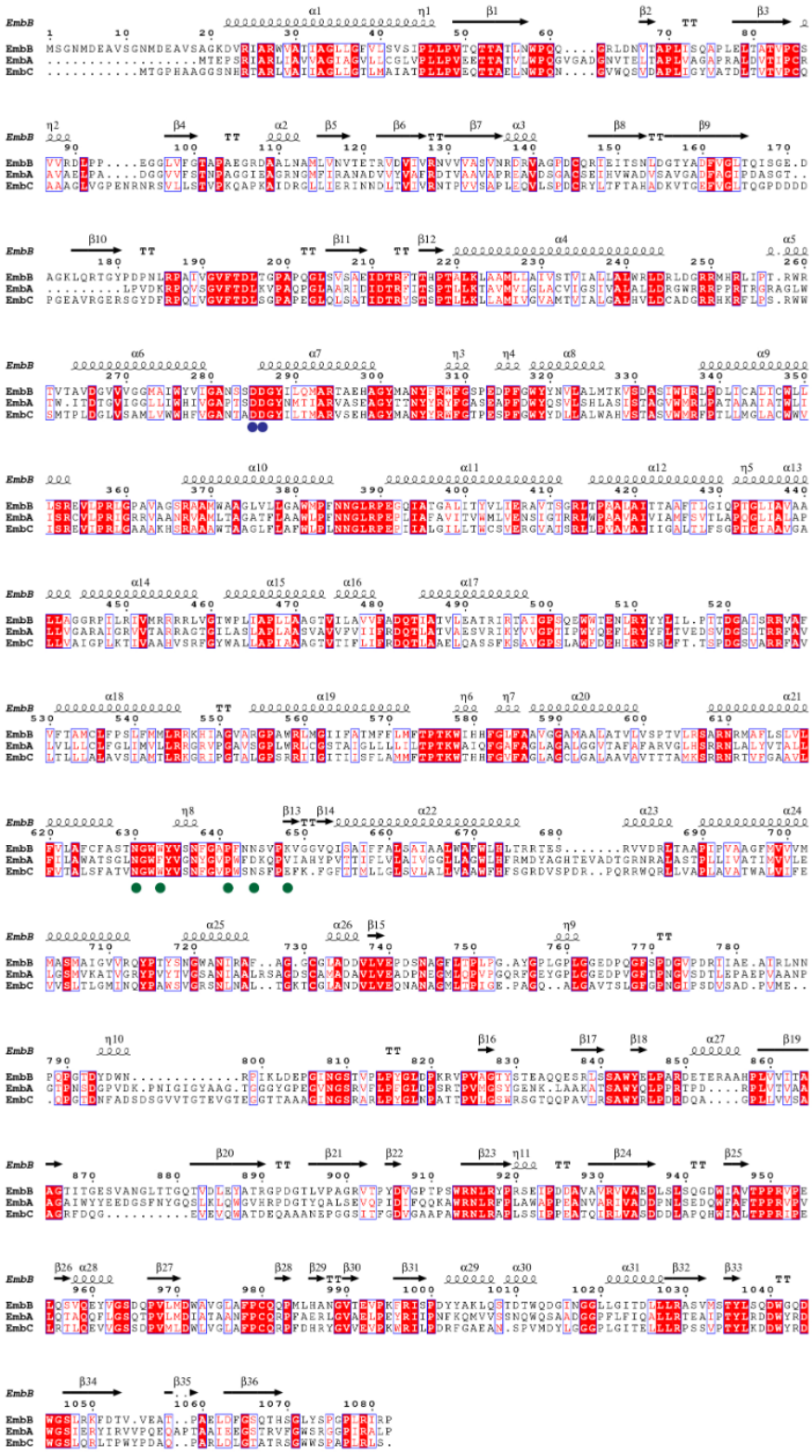


Figure S4.2 –Sequence alignment of M. smegmatis EmbA, EmbB and EmbC. *The sequence alignment between M. smegmatis EmbB (MSMEG_6389), EmbA (MSMEG_6388) and EmbC (MSMEG_6387), was generated using ClustalO* [16] *and displayed using ESPript* [17]*. Secondary structure annotation from the EmbB structure is displayed above the sequences. Identical residues are marked in a red filled box with white font, while residues above 70% similarity are depicted in a blue bordered box with red font. Blue dots indicate the two catalytic aspartic acid residues, and green dots indicate the residues that when mutated retain enzymatic activity in the enzyme yet reduce incorporation of arabinose.*

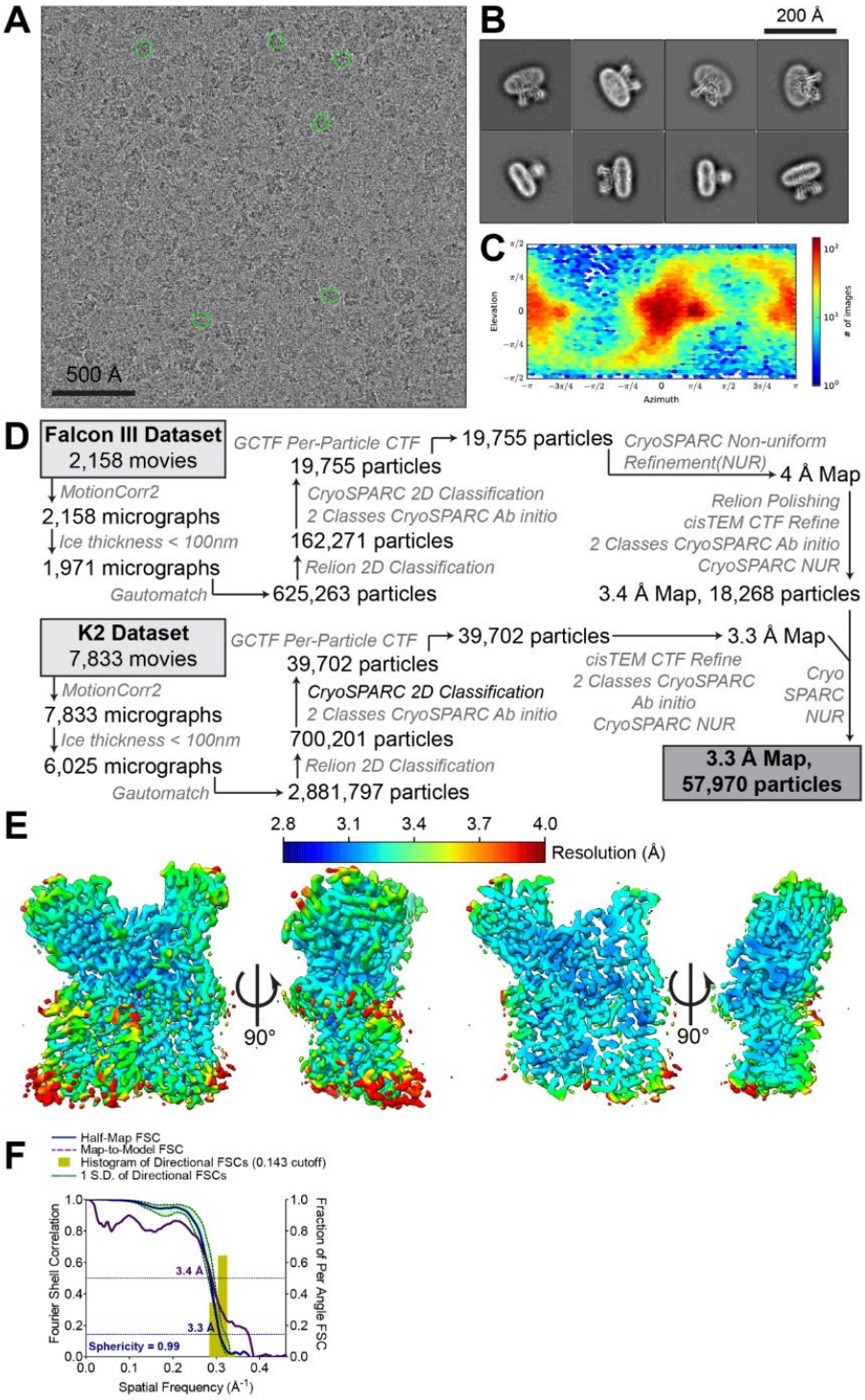


Figure S4.3 – Single-particle cryo-EM structural determination of EmbB. (A) *Representative micrograph from the first session of imaging using a Falcon III camera. The exposure image had a defocus of –1.9 µm (estimated by CTFFind4) and pixel size of 0.665 Å/pixel. Particles that went into the final reconstructions are circled in green.* ***(B)*** *Representative 2D class averages.* ***(C)*** *Euler angle distribution plot of the final 3D reconstruction.* ***(D)*** *Processing workflow for both cryo-EM datasets.* ***(E)*** *Local resolution display of EmbB reconstructions in orthogonal views.* ***(F)*** *Fourier shell correlation (FSC) curves for EmbB describing the half-map (blue) and map-to-model (purple) resolutions (at 0.143 and 0.5 cut-offs respectively), as well as histogram of directional resolutions sampled evenly over the 3DFSC (yellow). The corresponding sphericity value is also indicated.*

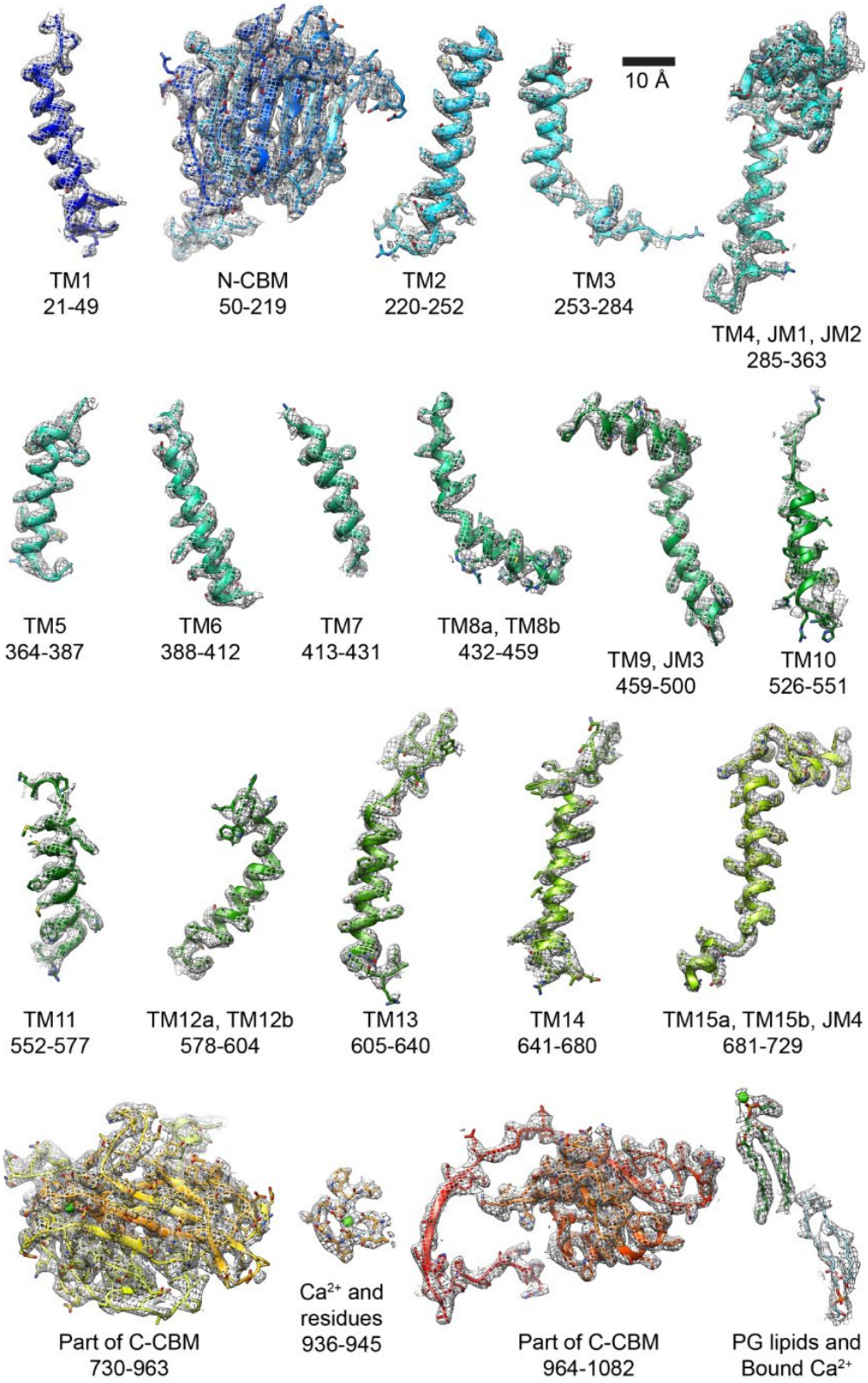


Figure S4.4 – EM Density of EmbB. *The atomic model for the structure of EmbB is colored in rainbow and rendered as a cartoon, with the side chains rendered as sticks. The map density is displayed as a mesh. The residues for each segment of the atomic model are indicated below the name of the segment. For all the segments except the PG lipids, the sigma value for display was 0.05, while for the PG lipids, it was 0.02.*

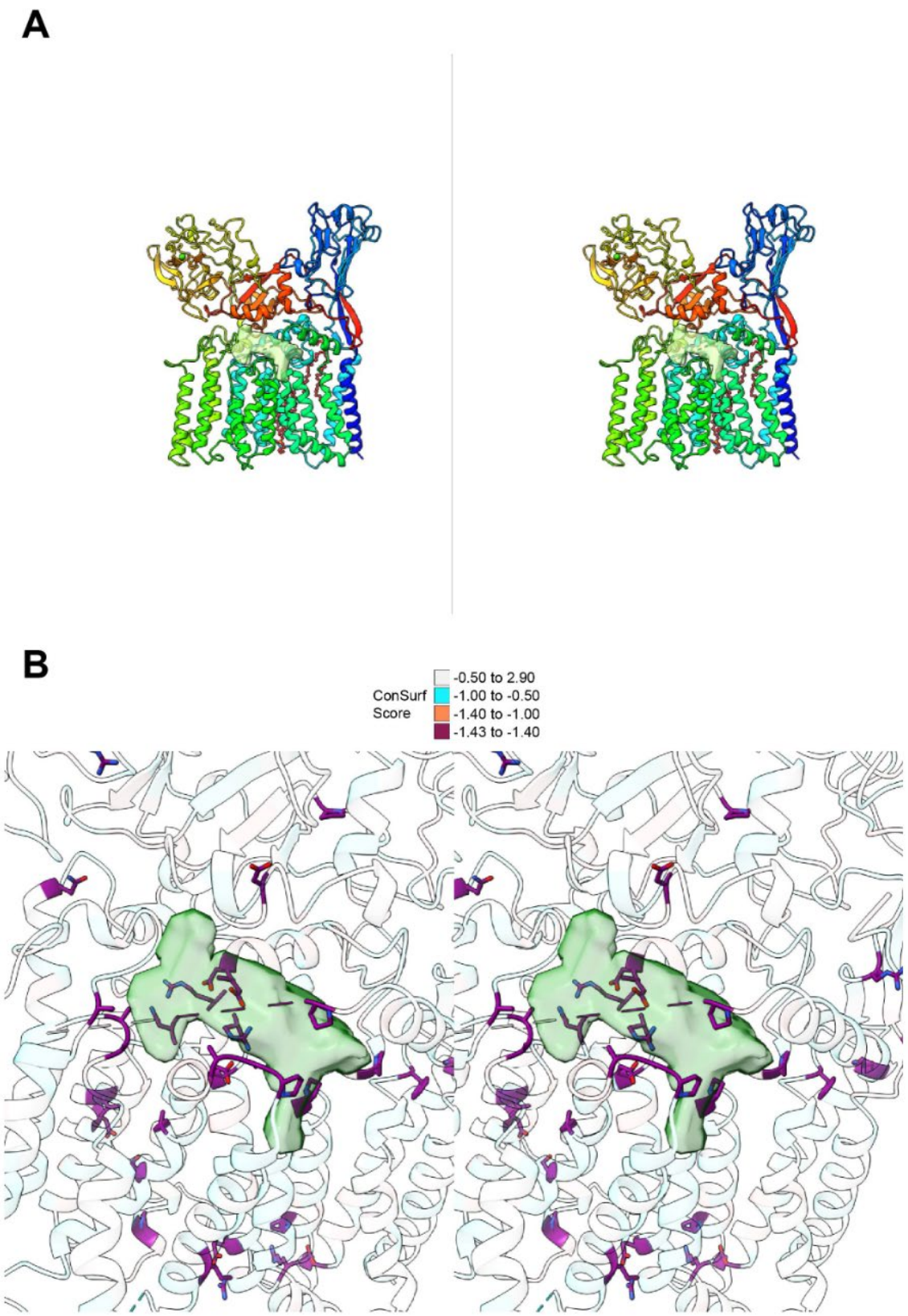


Figure S4.5 – Stereoscopic View of EmbB. (A) *Walleye representation of EmbB architecture as depicted in Fig. 4.3A.* ***(B)*** *Walleye representation of EmbB's active site as depicted in Fig. 4.4B.*

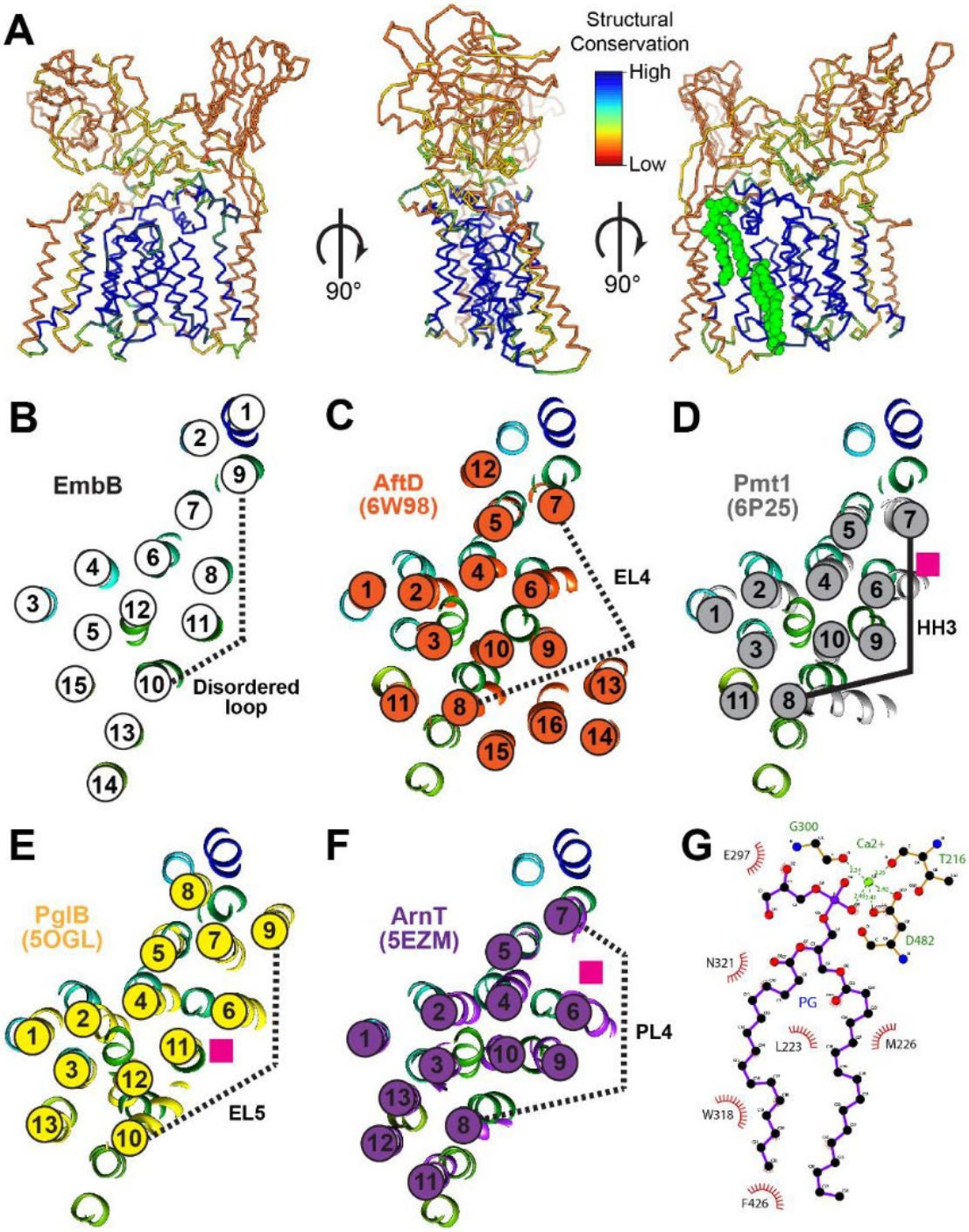


Figure S4.6 – Comparison of TM helices of EmbB with other GT-C Structures. ***(A)*** *Structural conservation of EmbB (GT53 family) against representative GT-C glycosyltransferases AftD (GT53 family, PDB ID: 6W98), Pmt1-Pmt2 (GT39 family, PDB ID: 6P2R), PglB (GT66 family, PDB ID: 5OGL) and ArnT (GT83 family, PDB ID: 5EZM). The protein is shown in cartoon, and the two PG lipids are shown in the last diagram as green spheres. Structural alignment was performed by the Dali server. Topology of the TM helices of EmbB with reference to the disordered loop is shown in* ***(B)****. The other glycosyltransferase structures, mycobacterial AftD* ***(C)****, eukaryotic Pmt1-Pmt2* ***(D)****, bacterial PglB* ***(E)*** *and archaeal ArnT* ***(F)****, were superimposed onto full-length EmbB and a slice through the transmembrane helices are shown as cartoon and colored in orange, grey, yellow and purple, respectively. Locations of the flexible loop that becomes ordered upon substrate binding are indicated by the dotted lines. Pmt1-Pmt2's HH3 is already ordered as a helix without any substrate binding, and hence is shown as a solid line. Locations of lipidic sugar donor substrate are indicated by pink boxes.* ***(G)*** *LigPlot+[3] rendering of the tightly bound PG with* Ca^{2+}.

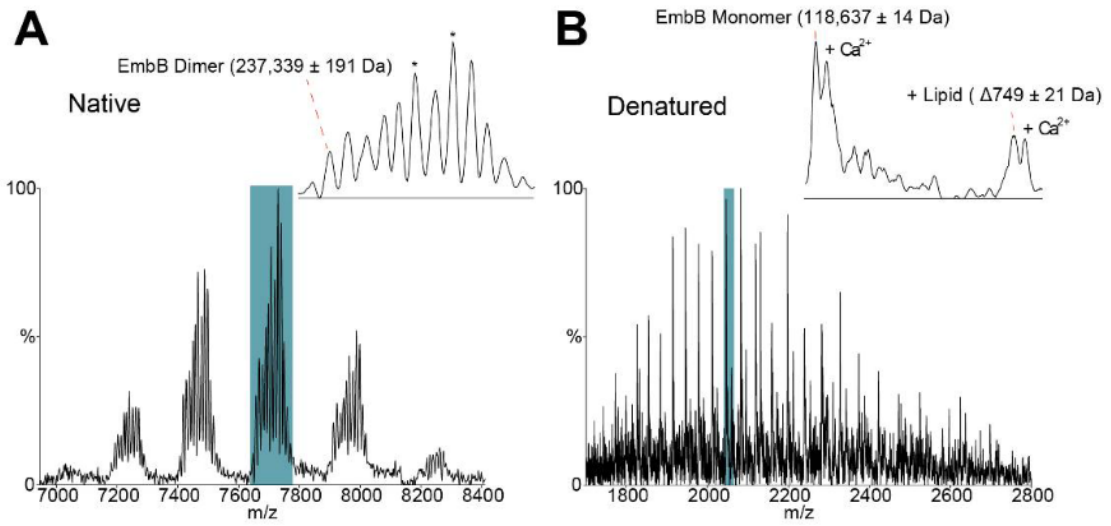


Figure S4.7 – Mass Spectrometry of EmbB. *Native* ***(A)*** *and denatured* ***(B)*** *mass spectra of EmbB in C12E8 and DDM detergent, respectively. Insets show the regions shaded in blue, and key peaks for EmbB monomer, dimer, and adducts are annotated. Peaks in* ***(A)*** *marked with an asterisk likely have two or three bound lipids.*

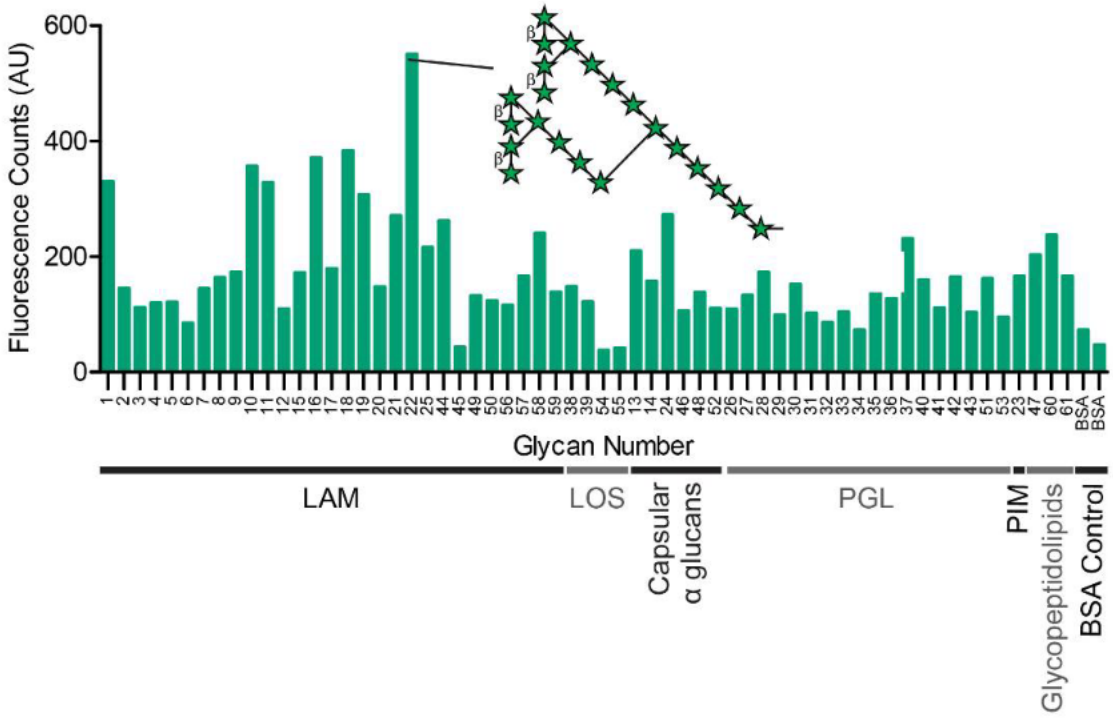


Figure S4.8 – Glycan Array Analysis of EmbB. *Glycan array analysis of EmbB, with top glycan motif hit indicated, where green star represents arabinofuranose. The classes of glycans used are: lipoarabinomannan (LAM), trehalose mycolates and lipooligosaccharides (LOS), capsular α glucans, phenolic glycolipids (PGL), phosphatidyl-myo-inositol mannoside (PIM), glycopeptidolipid and bovine serum albumin (BSA) control. For structures of all of the glycans on the array see* [62].

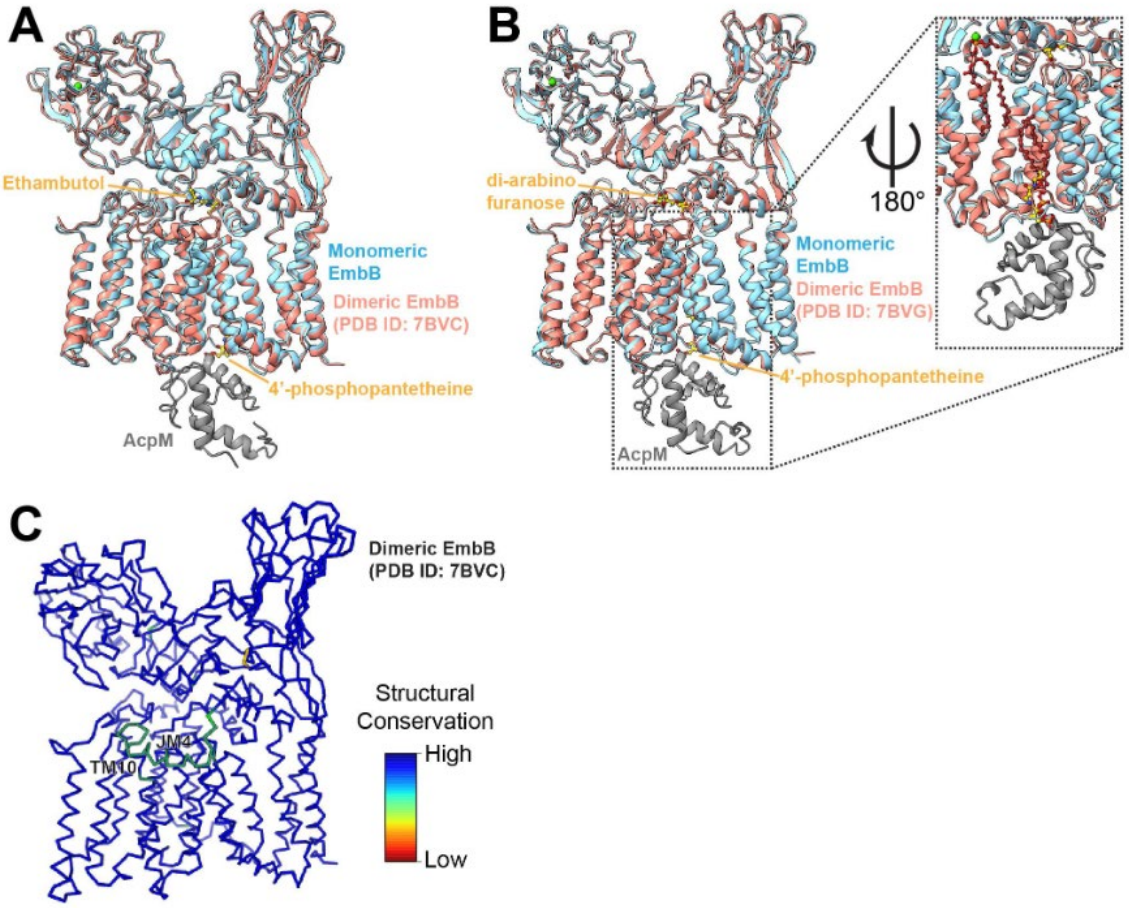


Figure S4.9 – Comparison with Structures of Dimeric EmbB. *Structures of M. smegmatis EmbB solved as a hetero-dimer with EmbA with either ethambutol inhibitor (PDB ID: 7BVC)* ***(A)*** *or di-arabinofuranose ligand (PDB ID: 7BVG)* ***(B)*** *are colored in salmon and superimposed with the monomeric apo-EmbB solved in this paper, colored in pale blue. The bound AcpM is colored in grey, and the ligands colored in gold. Zoom in of the two PG lipids in the monomeric apo-EmbB in brown is shown as an insert in* ***(B)****.* ***(C)*** *Structural conservation of monomeric EmbB against the two aforementioned dimeric EmbB, mapped onto the structure of the dimeric EmbB (PDB ID: 7BVC). Structural alignment was performed by the Dali server.*

Table S4.1 – Cryo-EM data collection and modeling statistics EmbB.

	Session 1	Session 2
EM data collection / processing		
Microscope	FEI Titan Krios	
Magnification	120,000	37,000
Voltage (kV)	300	
Camera	Falcon III	Gatan K2 Summit
Mode	Counting	Counting
Set defocus range (μm)	0.5 – 2.5	0.3 – 2.9
Defocus mean ± std (μm)	1.8 ± 0.29	1.9 ± 0.27
Exposure time (s)	86.4	8
Number of frames	80	80
Dose rate (e-/pixel/s)	0.4	4.3
Total dose (e-/ Å^2)	78.02	77.53
Pixel size (Å)	0.665	0.667
Number of micrographs	2,158	7,833
Number of particles (after initial cleanup)	162,271	700,201
Number of particles (in final map)	57,970	
Symmetry	C1	
Resolution (global) (Å)	3.3	
Local Resolution Range	2.8 – 16.0	
Directional Resolution Range	3.0 – 3.4	
Sphericity of 3DFSC	0.99	
SCF Value[a]	0.98	
Map sharpening B-factor (Å^2)	-72.5	
Model statistics		
Initial model used (PDB code)	3PTY	
Map-to-model resolution (Å)	3.4	
Model composition		
Non-hydrogen atoms	15,844	
Residue range	21–500, 526–1082	
Ligands	4	
Map CC	0.743	
RMSD [bonds] (Å)	0.0065	
RMSD [angles] (Å)	1.21	
All-atom clashscore	2.67	
B factors (Å^2)		
Protein	44.96	
Ligands	44.63	
Ramachandran plot		
Favored (%)	94.87	
Allowed (%)	5.05	
Outliers (%)	0.08	
Rotamer outliers	0.00	
C-β deviations	0	
MolProbity Score	1.41	
EM-Ringer Score	3.14	

[a] The SCF value is calculated as described [43], but currently assumes that all orientations have been properly assigned and does not take into account false positive assignment.

Table S4.2 – Ethambutol Resistance Mutations Found in EmbB. *Residues that are not conserved between EmbBs from M. tuberculosis (Mtb) and M. smegmatis (Msm) are colored in red.* [11][63][64][66][74]

Organism found in	Residue Number in *Mtb*	Residue in *Mtb*	Residue Number in *Msm*	Residue in *Msm*	Mutated to
***M. tuberculosis* (*Mtb*)**	297	S	283	S	A
	306	M	292	M	I L V
	319	Y	305	Y	C
	328	D	314	D	G Y
	330	F	316	F	V
	334	Y	320	Y	H
	354	D	340	D	A
	406	G	392	G	A C D S
	445	Q	431	Q	R
	497	Q	483	Q	K
	506	T	492	T	R
	745	G	729	G	D
	959	D	942	G	A
	1000	M	984	M	R
	1024	D	1008	S	N
	1082	T	1066	S	A
***M. smegmatis* (*Msm*)**	303	I	289	I	F M
	306	M	292	M	T

Table S4.3 – Ethambutol Resistance Mutations Found in the M. tuberculosis EmbC. *Residues that are not conserved among M. tuberculosis (Mtb) EmbC and M. smegmatis (Msm) EmbB and EmbC are colored in red.* [11][64][65]

Residue Number in *Mtb* EmbC	Residue in *Mtb* EmbC	Residue Number in *Msm* EmbC	Residue in *Msm* EmbC	Residue Number in *Msm* EmbB	Residue in *Msm* EmbB	Mutated to
294	D	280	D	286	D	G
300	M	286	M	292	M	L V
394	N	380	N	386	N	D
398	P	384	P	390	P	S
491	Q	477	Q	483	Q	R
502	L	488	F	494	I	P
738	R	719	R	721	W	Q

CHAPTER FIVE

AftA from *Mycobacterium vanbaalenii*

1. Summary

Arabinogalactan (AG) is one of the main components of the cell wall of *M. tuberculosis*, the causative organism of tuberculosis (TB). AG is a complex heteropolysaccharide, mostly comprised by a linear galactan core and a highly branched arabinan. Arabinofuranosyltransferase AftA is an essential enzyme responsible for the priming of the galactan core of AG with the first arabinofuranosyl (Ara*f*) residue, thus allowing further elongation and branching of arabinan performed by other arabinofuranosyltransferases (AraT). We present a preliminary low-resolution 3D model of a AftA-megabody protein complex. The use of nanobodies and megabodies to increase the size of small proteins facilitates the single particle Cryo-EM workflow for the structure determination of small proteins, however, sample heterogeneity and protein complex denaturation and/or dissociation constitute difficult challenges that need to be addressed in future studies.

Keywords: glycosyltransferase, *Mycobacterium tuberculosis*, arabinofuranose, membrane protein, nanodisc, nanobody, megabody, single-particle cryo-electron microscopy

2. Introduction

Mycobacterium tuberculosis (*Mtb*) possesses a unique cell envelope that protects against common antibiotics and contributes to the growth and virulence of *Mtb* [1][2]. The multilayered envelope comprises a cell wall core constituted mainly by the mycolyl-arabinogalactan-peptidoglycan (mAGP) complex, which is an essential feature targeted by anti-mycobacterial drugs [3]–[6].

Arabinogalactan (AG) is an heteropolysaccharide that constitutes approximately 35% of the mycobacterial cell wall [7][8] and is the main component of the mAGP complex, being composed by a linear galactan core comprising galactofuranosyl (Gal*f*) residues, and a branched arabinan, comprising arabinofuranosyl (Ara*f*) residues [8][9].

The synthesis of such intricate structure of AG involves several enzymes, wherein arabinofuranosyltransferases (AraT) are responsible for the polymerization of arabinan. These enzymes transfer one Ara*f* residue from the lipidic donor decaprenylphosphoryl-D-arabinofuranose (DPA) to the growing AG polymer chain [10][11].

AftA is an essential AraT responsible for the priming of galactan by transferring the first Ara*f* residue from the lipidic donor DPA to the galactan at the 8^{th}, 10^{th} and 12^{th} positions (Figure 5.1), and is one of the few AraT which are not inhibited by ethambutol, contrary to EmbA, EmbB and EmbC [12]–[14]. The essential role in the synthesis of AG makes AftA a prime candidate for drug development against TB, therefore the 3D structure of AftA coupled with a functional characterization of the protein will provide the basis for the development of novel therapeutics against TB.

AftA is a small membrane protein, with a theoretical mass of 69.5 kDa [15], which constitutes a challenge for the conventional single particle Cryo-EM workflow, as it is difficult to reconstitute a 3D volume of small particles, especially membrane proteins with subtle features outside the membrane [16].

Nanobodies (Nb) are a specific kind of antibodies produced in llamas that can be used as fiducial markers for single particle Cryo-EM, as they are stable proteins (10-15 kDa) that can form a protein complex with a target protein, providing an additional feature and a slight increase in size to the target protein [17].

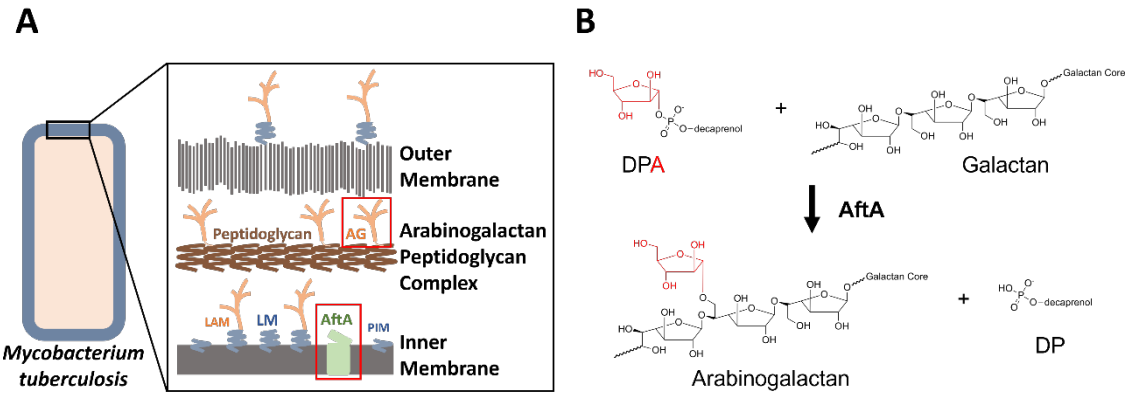


Figure 5.1 – Functional Analysis of the Role of the Arabinofuranosyltransferase AftA in the Biosynthesis of the Mycobacterial Cell. (A) *Model of the cell envelope of M. tuberculosis. Red boxes highlight the arabinogalactan (AG) and AftA.* ***(B)*** *Reaction catalyzed by AftA enzyme.*

Megabodies (Mb) are derived from nanobodies, in which a scaffold protein is fused to a nanobody with high affinity towards the target protein. The molecular weight of the megabodies can be tailored by choosing different scaffold proteins to fuse with a selected nanobody [18].

In this work, we screened several nanobodies against AftA from *Mycobacterium vanbaalenii* to select a fitting partner for the assembly of a protein complex with AftA, thus facilitating the single particle Cryo-EM workflow to determine the structure of this membrane protein. We analyzed by Cryo-EM a protein complex constituted by AftA and a 100 kDa megabody, derived from the previously selected anti-AftA nanobody, with the goal of determining the 3D structure of AftA. The preliminary data suggests a successful protein complex assembly of AftA with the selected 100 kDa megabody, however the number of particles of the complex collected by Cryo-EM was insufficient to determine a high-resolution structure of the AftA-megabody complex.

3. Materials and Methods

3.1. Genomic Expansion and Small-Scale Screening

AftA genes were identified from a collection of 14 *Mycobacterium* genomes using a bioinformatics approach [19]. Ligation independent cloning (LIC) was used to clone these targets from the genomes into five LIC-adapted expression vectors (pNYCOMPS-Nterm, pNYCOMPS-Cterm, pNYCOMPS-N23, pNYCOMPS-C23 and pMCSG7-10x) that contained a Tobacco Etch Virus (TEV) protease cleave site (ENLYFQSYV) and decahistidine affinity tag. Small and medium scale expression was done in a high throughput manner as described in detail in a previous protocol by Bruni and Kloss [20]. The *M. vanbaalenii* PYR-1 *aftA* construct in the pNYCOMPS-N23 plasmid was successfully cloned and expressed in small and medium scale, therefore it selected for further structural studies.

3.2. Sequence alignment

Protein sequence of AftA from *M. tuberculosis* was obtained from the Mycobrowser [15], with the KEGG identifier AftA *Mtb*-Rv3792. Protein sequence of AftA from *M. vanbaalenii* PYR-1 was obtained from GenBank, with the identifier ABM16394.1. The sequences were then aligned using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) [21] and displayed using ESPript (http://espript.ibcp.fr) [22].

3.3. *M. vanbaalenii* PYR-1 AftA Expression, Purification and Nanodisc Reconstitution

Mycobacterium vanbaalenii PYR-1 AftA WT in the pNYCOMPS-N23 plasmid was transformed into BL21 (DE3) pLysS *E. coli* competent cells and plated onto Luria broth (LB) agar plates supplemented with 100 µg/mL ampicillin (Sigma) and 35 µg/mL chloramphenicol (Sigma) and grown overnight at 37 °C. In the next day, a colony was picked and used to inoculate 100 mL of 2xYT medium supplemented with 100 µg/mL ampicillin and 35 µg/mL chloramphenicol in a 250 mL flask. The starter culture was

grown overnight at 37 °C with shaking (240 rpm) in an incubator shaker (Barnstead MaxQ 4000 Orbital Incubator Shaker, Thermo Scientific). In next day, twelve Thomson's Ultra Yield™ Flasks of 2.5 L (Oceanside, CA, USA), each with 800 mL of 2xYT medium supplemented with 100 mg/mL ampicillin and 35 mg/mL chloramphenicol, were inoculated with 8 mL of starter culture. The cultures were then grown at 37 °C with shaking (240 rpm) until OD_{600} reached 1.0 (approximately 3 hours). Temperature was then reduced to 22 °C and AftA expression was induced with 0.2 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) (Fisher). The culture was then incubated overnight with shaking (240 rpm). The cells were then harvested by centrifugation at 4,000 *g* in JA-10 fixed angle rotor (Beckman Coulter), for 30 min at 4 °C. The supernatant was discarded, and the pellet was resuspended in chilled 1x phosphate buffered saline (PBS) and centrifuged again at 4,000x *g* for 30 min at 4 °C. The supernatant was again discarded and the pellet was resuspended in lysis buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 20 mM $MgSO_4$, 10 µg/mL DNase I (Roche), 8 µg/mL RNase A (Roche), 1 mM tris(2-carboxyethyl)phosphine hydrochloride (TCEP) and 1 tablet/1.5 L buffer EDTA-free cOmplete protease inhibitor cocktail (Roche). For 9.6 L of culture, there was ~ 20 g of wet cell pellet mass, which was resuspended with ~200 mL of lysis buffer. Cell suspension was lysed after two passages at 15,000 psi on a cell disruptor (Constant Systems Ltd., Daventry, UK). Cell debris was discarded from the cell lysate after a centrifugation at 8,000x *g* in a JA-10 fixed angle rotor (Beckman Coulter), and the clear lysate was then ultracentrifuged at 197,215x *g* in a 45 Ti rotor (Beckman Coulter), for 1h30 min at 4 °C. The supernatant was discarded and the wet membrane pellet mass was ~ 8 g, which was resuspended in the lysis buffer up to a volume of 50 mL and manually homogenized using a Wheaton® glass homogenizer (DWK Life Sciences Limited, Stoke-on-Trent, UK).

The membrane fraction was solubilized by adding n-dodecyl-β-D-maltopyranoside (DDM) to a final concentration of 1% (w/v) detergent for 2 hours at 4 °C with gentle rotation. Insoluble material was removed by ultracentrifugation at 197,215x *g* in Type 45 Ti Rotor at 4 °C for 30 min. The supernatant was added to two Falcon tubes containing pre-equilibrated Ni^{2+}-NTA resin (QIAGEN) in the presence of 40 mM imidazole and incubated with gentle rotation at 4 °C for 2 hours. The resin was washed with 10 column volumes of wash buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 60 mM Imidazole, 0.1% DDM and eluted with elution buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 300 mM Imidazole, 0.05% DDM. The eluted protein buffer was exchanged into a buffer containing 20 mM HEPES pH 7.5, 200 mM NaCl, 0.05% DDM using a PD-

10 desalting column (GE) and concentrated using a 100-kDa concentrator (Pierce) to ~1.75 mg/mL.

The protein was then incorporated into lipid nanodisc [23] with the molar ratio 1:134:2.4 of 1-palmitoyl-2-oleoyl-glycero-3-phosphocholine (POPC) (Sigma) with membrane scaffold protein 1E3D1 (MSP-1E3D1) and incubated for 2 hours with gentle agitation at 4 °C. The POPC stock solution was prepared by solubilizing the solid lipid in 20 mM HEPES pH 7.5, 200 mM NaCl, 50 mM sodium cholate to a final concentration of 20 mM. Reconstitution was initiated by removing detergent with the addition of Bio-beads (Bio-Rad,120 mg per mL of protein solution), overnight with constant rotation at 4 °C. Nanodisc reconstitution mixture was removed from the bio-beads with a syringe and gently concentrated with a 100 kDa concentrator to decrease sample volume to perform a sucrose density gradient ultracentrifugation step. The sucrose density gradient ultracentrifugation step allows to separate empty and AftA loaded nanodiscs. To prepare the sucrose gradient tubes, 5% and 15% sucrose buffers were prepared (20 mM HEPES pH 7.5, 200 mM NaCl, 5% or 15% m/v sucrose) and the continuous gradient was set up using a Gradient Station™ (BioComp Instruments). The nanodisc reconstitution mixture was then carefully loaded on top of the 5 – 15% sucrose continuous gradient and centrifuged at 197,000x *g* for 11.5 h at 4 °C. The tubes were then top-down fractionated using the Piston Gradient Fractionator™ of the Gradient Station™ (BioComp Instruments). The continuous UV absorption of the fractionated sample was monitored by optical density at 280 nm (OD_{280}). Similar sucrose density gradient ultracentrifugation and fractionation steps were performed with nanodisc reconstitution mixture sample not containing AftA, to assess the purification profile of “empty” MSP-1E3D1 nanodiscs so that a better separation of empty and AftA-loaded nanodiscs is achieved. The collected fractions were run on 12% SDS-PAGE. The best fractions were pooled, exchanged to 20 mM HEPES pH 7.5, 200 mM NaCl and concentrated using a 100-kDa concentrator (Pierce).

3.4. Nanobody/Megabody selection, production and purification

Nanobodies against *M. vanbaalenii* PYR-1 AftA were generated in llama, as described in detail by Pardon E. *et al* [17]. The selection of nanobodies against *M. vanbaalenii* PYR-1 AftA was done by phage-display using two different approaches: (1) AftA was solid phase coated in coating buffer ($NaHCO_3$ pH 8.2); (2) AftA was trapped via anti-FLAG M2 monoclonal anti-bodies (mAb) in coating buffer ($NaHCO_3$ pH 8.2). The selection and

screening of nanobodies were done in 20 mM HEPES pH 7.5, 150 mM NaCl, 0.05% DDM. Specific nanobodies were detected using EPEA-tag at the C-terminus of each nanobody and the CaptureSelect™ Technology, wherein a CaptureSelect™ Biotin anti-C-tag Conjugate (1/4000 dilution) is mixed with streptavidin alkaline phosphatase (Promega) (1/1000 dilution) for 10 min before use. Signal was detected using 4-nitrophenyl phosphate disodium salt hexahydrate (2 mg/mL).

A set of 35 different nanobodies were discovered, which were further classified into 11 groups according to the sequence similarity of the complementarity-determining region (CDR3) of the nanobody (Supplementary Table S5.1). One nanobody from each group was then produced and purified to assess the binding affinity to AftA by Surface Plasmon Resonance (SPR).

The plasmid DNA corresponding to each selected nanobody was transformed into WK6 (Su^-) *E. coli* competent cells, which were then plated onto Luria-Bertani (LB) agar plates supplemented with 2% glucose and grown overnight at 37 °C. A colony was picked and used to inoculate a starter culture containing 50 mL of LB medium supplemented with 2% glucose, 100 µg/mL ampicillin and 1 mM $MgCl_2$. The starter culture was shaken overnight at 190 rpm, at 37 °C. The starter culture was used to inoculate 500 mL cultures of Terrific broth medium supplemented with 0.1% glucose, 100 µg/mL ampicillin and 1 mM $MgCl_2$. Cultures were grown at 37 °C with continuous shaking at 190 rpm until the OD_{600} reached 0.7 – 1.2, at which point the nanobody expression was induced with 1 mM IPTG. The cultures were then grown overnight at 28 °C. In the next day, the cells were harvested by centrifugation at 4,000 *g* in JA-10 fixed angle rotor (Beckman Coulter), for 30 min at 4 °C. After centrifugation, the supernatant is discarded.

Since the nanobodies are cloned behind a presignal (pelB signal sequence) which directs the nanobodies to the periplasmic space, the cells are not disrupted to retrieve the produced nanobodies. Instead, the cells are submitted to osmotic shock, which releases the nanobodies into the buffer without lysing the cells. Therefore, the cell pellet was resuspended in chilled 1x TES buffer (0,2 M Tris pH 8.0, 0.5 mM EDTA, 0.5 M sucrose), using 15 mL of 1x TES buffer per 1 L of cell culture. The resuspended cells were incubated on ice for 1 h, with constant stirring using a magnetic stirrer. After one hour, it is added 30 mL of 0.25x TES buffer per 1 L of cell culture, and the suspension is incubated for 45 min on ice with constant stirring. After the second incubation, the suspension is centrifuged at 4,000 *g* in JA-10 fixed angle rotor (Beckman Coulter), for 30 min at 4 °C. The supernatant corresponds to the periplasmic fraction, containing the produced nanobodies. A second osmotic shock may be performed to the cell pellet, by

resuspending the pellet overnight in 1x TES buffer, adding 0.25x TES buffer the next day. This second osmotic shock may yield as much as the first osmotic shock but will contain more contaminants.

The C-terminally His_6-tagged nanobody was purified using nickel affinity chromatography. The periplasmic fraction was added to Falcon tubes containing Ni^{2+}-NTA resin (QIAGEN), pre-equilibrated in 50 mM sodium phosphate pH 7.0, 1 M NaCl. The tubes were incubated with gentle rotation at 4 °C for 1 hour. The solution was then poured over an empty column and the flow-through was collected. The resin was washed with 4 column volumes of 50 mM sodium phosphate pH 7.0, 1 M NaCl, followed by washing with 10 column volumes of 50 mM sodium phosphate pH 6.0, 1 M NaCl. The protein was eluted with 50 mM sodium phosphate pH 7.0, 150 mM NaCl, 300 mM Imidazole. Eluted samples were concentrated using a 10-kDa concentrator (Pierce) and injected into a Superdex 200 column (Cytiva Europe GmbH, Freiburg, Germany) in 20 mM HEPES pH 7.5, 200 mM NaCl. The collected fractions were run on 15% SDS-PAGE. The best nanobody fractions were snap-frozen in liquid nitrogen and stored at -80 °C.

Megabodies (Mbs) consist in chimera molecules comprised by a selected nanobody grafted onto a protein scaffold to increase their molecular weight while maintaining full antigen binding specificity and affinity. We selected nanobody 48 (Nb48) for megabody development. Two megabody constructs were developed as described by Uchański, T. *et al* [18], wherein two protein scaffolds were fused to Nb48: β- turn S3-S4 of the adhesin domain of *H. pylori* [24] (HopQ, 45 kDa, PDB ID: 5LP2) and β-turn A'S1-A'S2 of the *E. coli* K12 Glucosidase YgjK [25] (YgjK, 86 kDa, PDB ID: 3W7T). The resulting megabodies were Mb36 and Mb41, having a molecular weight of 56 kDa and 101 kDa, respectively.

Production and purification of Mb36 and Mb41 was similar to the nanobodies. In summary, Mb36 and Mb41 were produced in WK6 (Su^-) *E. coli*, gene expression was induced by IPTG, and Mbs were extracted from the periplasmic fraction. Mbs were purified by nickel affinity chromatography, followed by size exclusion chromatography.

3.5. AftA nanodisc, nanobodies and megabodies protein-protein interaction

The interaction between AftA nanodiscs (AftA ND) and anti-AftA nanobodies (Nb) or anti-AftA megabodies (Mb) was assessed by SPR.

AftA ND were immobilized onto L1 (Series S) sensor chip, occupying a first pair of flow cells. Empty nanodiscs (empty-ND) were immobilized onto a different pair of flow cells of the sensor chip, for signal subtraction. For the immobilization of the nanodiscs (AftA-ND or empty-ND), a first injection of 20 mM CHAPS for 30 seconds (twice) was followed by a buffer wash for 120 seconds. The background buffer during immobilization was 100 mM sodium phosphate buffer pH 7.4, 50 mM NaCl. Empty-ND and AftA-ND were diluted to 0.1 µM and injected at a low flow rate (10 µL/min), having contact time with the carboxymethylated dextran matrix for 30 – 60 s. Negative controls were performed by injecting Bovine Serum Albumin (BSA) (Thermo Fisher Scientific) for 300 s at 0.2 mg/mL.

The anti-AftA nanobodies (Nb09, Nb42, Nb43, Nb44, Nb45, Nb46, Nb47, Nb48, Nb49, Nb50, Nb51) were directly dissolved in running buffer (100 mM sodium phosphate buffer pH 7.4, 50 mM NaCl) supplemented with 0.5 mg/mL of BSA, to avoid unspecific interactions. The nanobodies were tested at 10 different concentrations using a 2-fold dilution series, with the highest concentration tested being 0.2 µM. Interaction analysis cycles consisted of 240 s sample injection (30 µL/min; association phase) followed by 600 s of buffer flow (dissociation phase).

The anti-AftA megabodies (Mb36, Mb41) were directly dissolved in running buffer (100 mM sodium phosphate buffer pH 7.4, 50 mM NaCl) supplemented with 0.5 mg/mL of BSA, to avoid unspecific interactions. The nanobodies were tested at 10 different concentrations using a 2-fold dilution series, with the highest concentration tested being 0.2 µM. Interaction analysis cycles consisted of 220 s sample injection (30 µl/min; association phase) followed by 600 s of buffer flow (dissociation phase).

All sensorgrams were processed by first subtracting the binding response recorded from the control surface (reference Flow channel), followed by subtraction of the buffer blank injection from the reaction channel. All datasets were fitted to a simple 1:1 Langmuir interaction model (considering one molecule of AftA binding one nanobody/megabody), to determine the kinetic rate constants. Experiments were performed on a Biacore T200 (Biacore AB, GE Healthcare Life Sciences, Uppsala, Sweden) at 25 °C and the interactions were evaluated using the provided Biacore T200 evaluation software.

3.6. AftA – Megabody protein complex assembly

Protein complexes of AftA nanodiscs and megabody Mb41 were assembled using a 1:1.5 molar ratio. In summary, purified AftA nanodisc fractions at 20 µM were incubated with purified fractions of Mb41 (~15 µM), adjusting the volume of both proteins according to the molar ratio. The buffer for protein complex assembly was 20 mM HEPES pH 7.5, 200 mM NaCl. The mixture was incubated for 2 h, at 4 °C, with gentle rotation. The sample was then loaded on top of 5 – 15% sucrose continuous gradient and centrifuged at 197,000x *g* for 11.5 h at 4 °C. The tubes were then top-down fractionated using the Piston Gradient Fractionator™ of the Gradient Station™ (BioComp Instruments). The continuous UV_{280nm} absorption of the fractionated sample was monitored. The collected fractions were run on 12% SDS-PAGE. Densitometry was used to select the fraction with the best ratio AftA:Mb41:MSP for further characterization by Electron Microscopy. Sample aliquots were snap-frozen in liquid nitrogen and stored at -80 °C.

3.7. Negative staining imaging

A 10 µL aliquot of AftA-ND was thawed and diluted to 0.1 mg/mL. The sample was further diluted 10x and 2 µL of AftA-ND was loaded onto a glow-discharged carbon coated copper grids (Quimigen), secured in tweezers. The sample was absorbed for 30 seconds, and the excess sample was blotted with filter paper. The grid briefly touched droplets of distilled water (twice) and Nano-W™ stain (Quimigen), all blotted with filter paper, and afterwards touched a second droplet of Nano-W™ stain for 20 seconds, then blotted with filter paper and let dry completely. The grids were imaged in a Hitachi H-8100 Transmission Electron Microscope (TEM), at 200 kV.

3.8. Single-Particle Cryo-EM Sample Vitrification

Purified AftA-Mb41 was concentrated using a 100-kDa concentrator (Pierce) to 5 – 20 µL of sample at ~1 mg/mL. The sample was loaded on Quantifoil 1.2/1.3 µm holey grids, which were previously plasma cleaned for 45 seconds. The grids were blotted on the sample side for 1.5 s and vitrified using a Leica EM GP plunge freezer, at room temperature and 80% relative humidity.

3.9. Data Acquisition

Images were recorded on a 200 kV Talos Arctica electron microscope (FEI) equipped with a Falcon 3 camera, phase plate and EPU, at a nominal x150,000 magnification corresponding to a calibrated pixel size of 0.97 Å.

Data collection was performed using a dose of ~40 $e^-/Å^2$ across 40 frames at a dose rate of 0.65 $e^-/Å^2/s$ using a set defocus range of -1.2 to -4.0 µm. For AftA-Mb41, a total of 2,146 micrographs were recorded over two sessions.

3.10. Data Processing

The recorded micrographs of two separate data collections were aligned using MotionCor2 [26] and the contrast transfer function (CTF) parameters were first estimated using CTFfind4 [27] and Gctf [28], and lastly using a consensus CTF estimation done using Xmipp3 [29][30]. Images containing drift or astigmatism were discarded. Particles from each micrograph were picked using template free particle picking with Sphire – cryolo [31] using a box size of 200 pixels and a low confidence threshold to avoid missing any particles. A total of 315,733 particles were extracted and transferred into CryoSPARC [32] for 2D classification. 2D class averages that showed no features were discarded. After three cycles of 2D classification, the particle set was split into two distinct subsets corresponding to two different populations: a first subset of 48,020 particles and a second subset of 13,719 particles. The subsets were further processed separately. 2D classification was performed using CryoSPARC [32] and an *ab initio* 3D model and 3D classification were performed using RELION [33]–[36]. The overall resolution of the 3D volumes was determined using Xmipp3 [29][30] and the volume maps of the best class of each subset were viewed in Chimera [37].

4. Results

4.1. Genomic Expansion

To study the AftA function *in vitro* and determine its structure, we adopted a structural genomics approach [19][20][38] to identify homologs that yielded high expression levels in *E. coli* and were stable in detergents compatible with structure determination. High-throughput small-scale expression and purification screens of AftA from 14 mycobacterial species revealed that, under our experimental conditions, the *M. vanbaalenii* PYR-1 and *M. neoaurum* homologues were good targets for expression and purification [39]. Although the AftA homolog from *M. neoaurum* showed good potential for scale up production (work described in Chapter 2: *High Throughput Expression Screening of Arabinofuranosyltransferases from Mycobacteria*), we selected the AftA homolog from *M. vanbaalenii* PYR-1 for large scale production and purification, as we had done previous scale-up experiments with this target. This homolog has 67% of sequence identity with *M. tuberculosis* AftA (Supplementary Figure 5.1).

4.2. AftA purification and nanodisc reconstitution

AftA from *M. vanbaalenii* PYR-1 was successfully produced in *E. coli* and purified by nickel affinity chromatography in DDM detergent. Although the overall yield was greater than the previous AraTs studied (~0.5 mg per liter of culture), the protein became unstable and precipitated during concentration, once it reached a concentration higher than 0.8 mg/mL, which prevented any attempt for crystallization trials.

On one hand, AftA could not be concentrated to allow crystallization, on the other hand, the molecular weight of AftA around 70 kDa is considered too small to determine its structure by single particle Cryo-EM.

Therefore, we devised a strategy to increase the size of the particle by firstly reconstituting AftA into lipid filled nanodiscs, secondly by assembling protein complex of AftA-nanodiscs with an anti-AftA nanobody (~15 kDa), and lastly assembling a protein complex of AftA-nanodiscs with an anti-AftA megabody (50 – 100 kDa).

AftA was successfully reconstituted into POPC filed nanodiscs, wherein the AftA-ND were further purified using a 5 – 15% sucrose density gradient ultracentrifugation (Figure 5.2).

We used a sucrose density gradient ultracentrifugation instead of size exclusion chromatography (SEC) to improve the separation between empty nanodiscs and AftA-ND (Figure 5.2C), which we found to be difficult in previous SEC runs, but also because a significant amount of protein was lost during SEC. The purified AftA-ND sample was analyzed by negative staining under an electron microscope. The micrographs showed disc-like particles, wherein most particles are viewed on their side and often form stacks of particles (Figure 5.3).

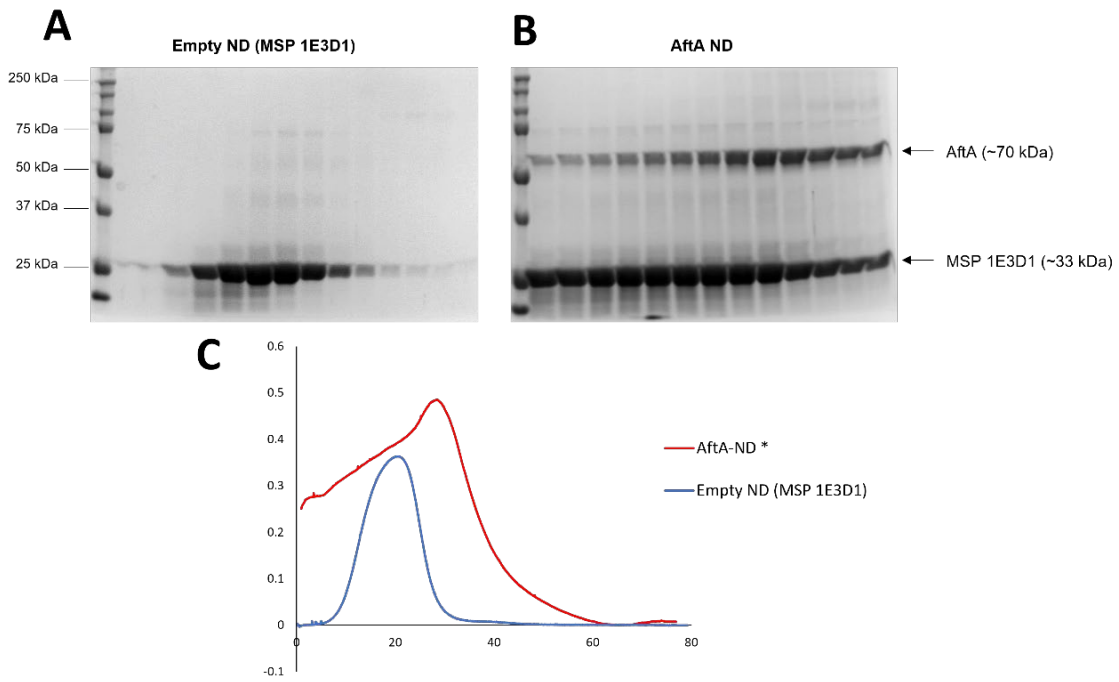


Figure 5.2 – Purification of AftA-ND by 5-15% sucrose density gradient ultracentrifugation. (A) *SDS-PAGE of the fractions of empty MSP-1E3D1 nanodiscs;* ***(B)*** *SDS-PAGE of the fractions of AftA nanodiscs;* ***(C)*** *Fractionation profile of empty ND vs AftA-ND: all empty ND are fractionated earlier than the peak fraction of AftA-ND in a 5-15% sucrose density gradient.*
** AftA-ND series was normalized to match the baseline of the empty ND series.*

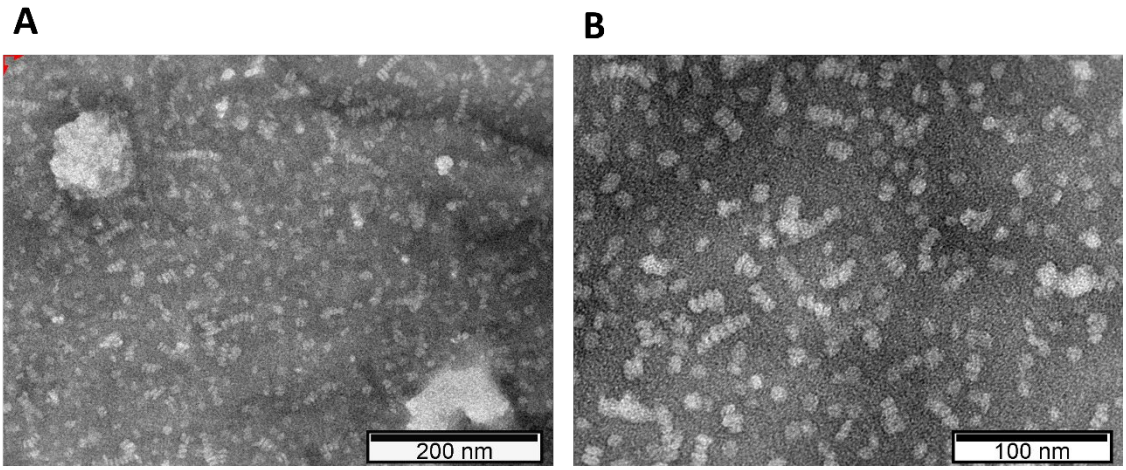


Figure 5.3 – Negative stain micrographs of purified AftA nanodiscs. (A) *AftA ND at 40,000 x magnification and at* ***(B)*** *70,000 x magnification.*

4.3. Anti-AftA nanobody and megabody purification

The anti-AftA nanobodies were purified by nickel-affinity chromatography followed by SEC and the protein-protein interaction of each nanobody with AftA-ND was assessed by SPR. The binding affinity of each nanobody are presented in Table 5.1.

Nb43 showed the highest binding affinity, having a measured K_D of 0.68 nM, which means that it has a strong interaction with AftA-ND. Nb09, Nb46, Nb49 and Nb51 showed no interaction up to 2 µM, suggesting that they do not have a strong interaction with AftA-ND. Although Nb43 would be the ideal candidate for megabody design, we selected Nb48 for megabody design, being another nanobody with strong interaction with AftA-ND (K_D = 12.6 nM).

The selection of Nb48 was done according to previous protein complex assembly experiments between AftA-ND and Nb48, in which the nanobody (~15 kDa) was co-eluted with AftA-ND (~70 kDa + ~30 kDa) in SEC, despite having a smaller molecular weight (Supplementary Figure S5.2).

Table 5.1 – Binding interaction by SPR of AftA-ND with anti-AftA nanobodies (Nb) or megabodies (Mb)

Ligand type	ID	Kinetic constant (K_D), M
Nanobody	Nb09	>2.00 E-07
	Nb42	4.86 E-08
	Nb43	6.81 E-10
	Nb44	3.76 E-09
	Nb45	3.91 E-09
	Nb46	>2.00 E-07
	Nb47	1.60 E-08
	Nb48	1.26 E-08
	Nb49	>2.00 E-07
	Nb50	1.35 E-08
	Nb51	>2.00 E-07
Megabody	Mb36	2.99 E-09
	Mb41	9.24 E-09

Two Nb48-derived megabodies were designed, having different molecular weights: Mb36 (~56 kDa) and Mb41 (~101 kDa). Both megabodies comprised Nb48 fused to a scaffold protein: *H. pylori* HopQ and *E. coli* K12 Glucosidase YgjK, respectively.

The anti-AftA megabodies were purified by nickel-affinity chromatography followed by SEC (Supplementary Figure S5.3) and the protein-protein interaction of each megabody with AftA-ND was assessed by SPR. The binding affinity of each megabody is presented in Table 5.1.

4.1. AftA – Megabody protein complex

Although Mb36 (~50 kDa) showed higher binding affinity to AftA than Mb41, we chose Mb41 (~101 kDa) as first candidate to form a protein complex with AftA, to obtain larger AftA-Megabody complex to characterize by Single Particle Cryo-EM. To assemble the complex, samples of purified AftA-ND and Mb41 were mixed with a molar excess of Mb41 to promote protein complex formation.

The AftA-Mb41 complex was purified using sucrose density gradient ultracentrifugation and fractions were analyzed by SDS-PAGE and densitometry to determine the best fractions (Figure 5.4A). Figure 5.4B shows the SDS-PAGE profile of the selected fraction,

being detected only protein bands attributable to AftA, Mb41 and MSP 1E3D1 (nanodisc scaffold protein).

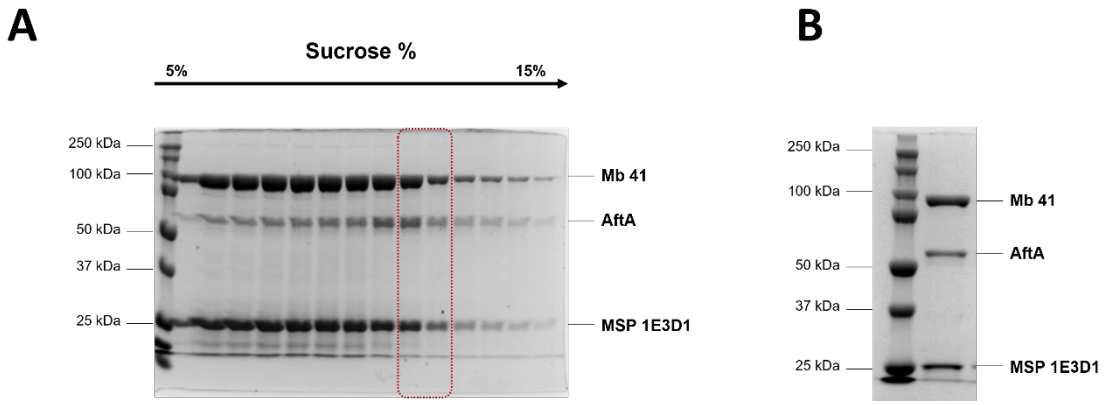


Figure 5.4 – AftA-Mb41 protein complex purification. (A) *SDS-PAGE of fractions from sucrose density gradient ultracentrifugation of AftA-Mb41 complex. The red box shows the selected fractions according to densitometry analysis.* ***(B)*** *SDS-PAGE of pooled fractions of AftA-Mb41 complex.*

4.1. AftA-Mb41 Single Particle Cryo-EM analysis

We prepared grids for Cryo-EM data collection of purified AftA-Mb41 complex. The overall scheme for the Cryo-EM data processing is presented in Figure 5.5. Micrograph image stacks from two sessions of data collection were processed and a total of 315,733 particles were extracted and their 2D projections classified to obtain a set of 2D classes.

After several iterations of 2D classification, it became clear that the sample was heterogeneous, having at least two different particle populations that differ in overall size. Therefore, the set of particles was split into two subsets, wherein the subset corresponding to the smaller particles comprised 48,020 particles and the other subset corresponding to the larger particles comprised 13,719 particles. Each subset was then further processed to refine the 2D classes prior to the 3D reconstruction step.

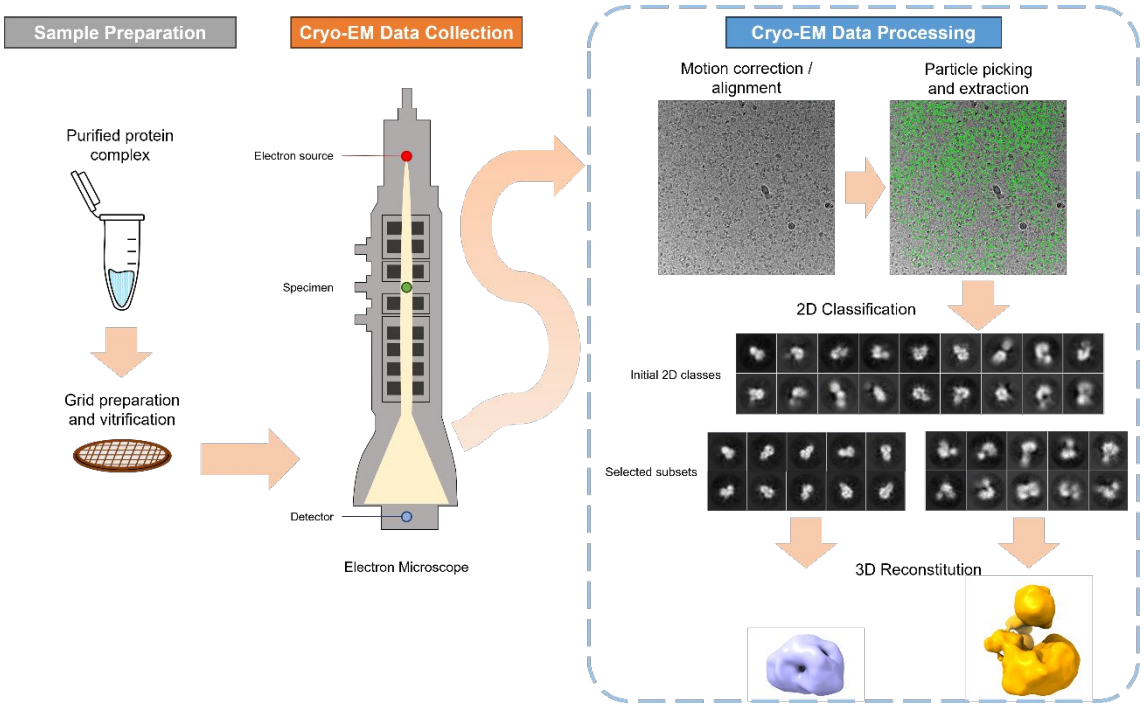


Figure 5.5 – Single Particle Cryo-EM workflow to analyze the protein complex AftA-Mb41.

The reconstructed 3D initial model of each particle subset is depicted in Figures 5.6A and 6B. As expected from the 2D classes of each subset, the corresponding initial 3D models are distinct, wherein from the largest subset of particles – subset 1 (Figure 5.6A) – we obtained a smaller 3D volume than the 3D volume obtained from the smaller subset of particles – subset 2 (Figure 5.6B).

Although each subset was further processed by performing 3D classification, the low number of particles (22,628 particles for subset 1; 2,527 particles for subset 2) made it impossible to improve 3D model of each subset and to obtain 3D models with identifiable features, either from AftA or from Mb41.

Considering that the protein sample should be a complex of two different proteins, AftA reconstituted into nanodiscs and the megabody, most likely the subset 1 3D models should correspond to unbound Mb41, and the larger 3D models of subset 2 should correspond to the assembled protein complex of AftA-ND with Mb41.

3D models of the best 3D class from each subset are presented in Figures 5.6C and 5.6D, having an overall resolution of 21.87 Å (subset 1 3D class, Fig. 5.6C) and 39.45 Å (subset 2 3D class, Fig. 5.6D).

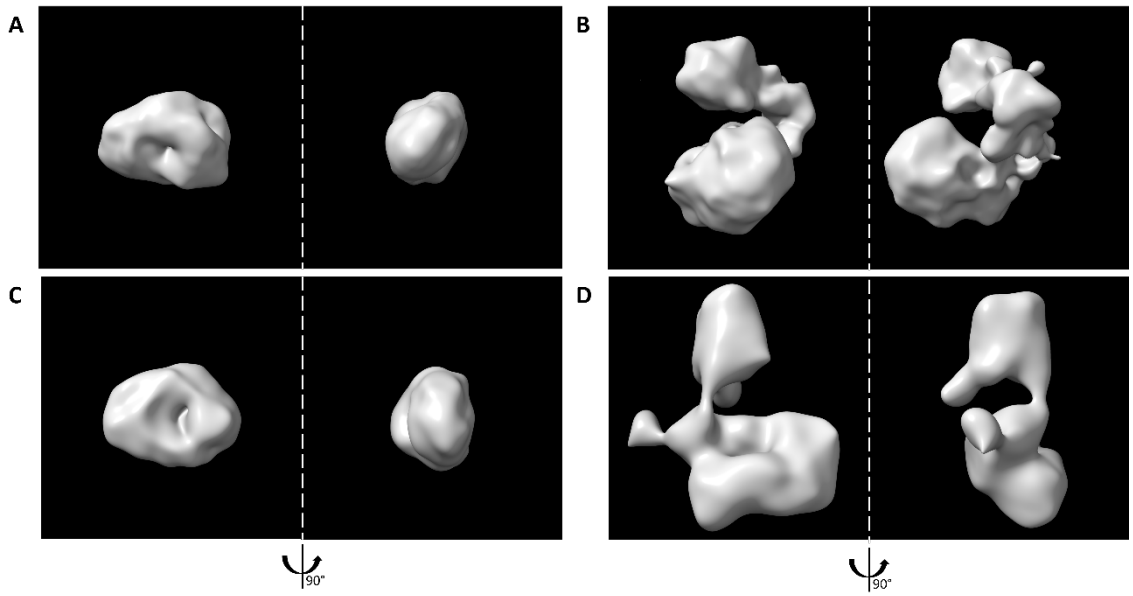


Figure 5.6 – 3D initial models and 3D classes generated from each subset of particles. ***(A)*** *3D initial model of subset 1 – 48,020 particles;* ***(B)*** *3D initial model of subset 2 – 13,719 particles;* ***(C)*** *3D class of subset 1 – 22,628 particles; and* ***(D)*** *3D class of subset 2 – 2,527 particles.*

5. Discussion

The collected dataset of AftA-Megabody complex was insufficient to determine the structure of AftA. Considering the low number of particles used throughout data processing, additional data collections are, therefore, necessary to increase the number of “good” particles to produce a reliable and detailed 3D model of AftA.

Nevertheless, the different 3D models produced in this preliminary work strongly suggest that the biological sample is heterogeneous, wherein the majority of particles are small (approximately 90% of the particles), and the number of larger particles that would correspond to the actual AftA-Mb41 complex is very low (~10% of the particles). In addition, there is no guarantee that the set of larger particles corresponding to the AftA-Mb41 complex would retain high resolution features, since at least part of those particles may be partially denatured by the air-water interface [40].

Considering that AftA-Mb41 protein complex is assembled *in vitro*, the sample was expected to have a heterogeneous population constituted by AftA-Mb41 protein complex and mainly Mb41, due to the molar excess of Mb41 in the sample mixture. However, the overwhelming amount of small particles suggests one of two scenarios: either AftA-ND and Mb41 are not suitable partners to form a stable complex, thus resulting in a large number of small particles; or the AftA-Mb41 is a stable protein complex in solution which dissociates during the grid preparation and vitrification.

Regarding the first scenario, the high affinity (K_D = 9.3 nM) of Mb41 indicates that there is a strong association to AftA-ND, which is a clear indication that the complex is stable in solution. Therefore, the second scenario is most likely to occur, wherein the stable complex present in the bulk solution must be dissociating or denaturing during grid preparation and vitrification, most likely due to diffusion of the complex to the air-water interface [40][41].

We addressed the issue of protein complex dissociation or denaturation with the addition of glutaraldehyde as a crosslinking agent to covalently bind AftA to Mb41, using the GraFix method [42][43]. However, the addition of a crosslinking agent caused a massive aggregation of the sample across the gradient, resulting in a “smear” of different molecular weight protein aggregates, which made it impossible to purify. Therefore, we were unsuccessful to produce a crosslinked AftA-Mb41 protein complex which would have the expected molecular weight, as most of the sample was permanently aggregated and highly heterogeneous (data not shown).

Other ways to solve the issue of complex dissociation or denaturation during grid preparation and vitrification can be highly dependent on instrumentation [44]. There are already several devices that are automated and programmed to either “spray” or “write” a very thin layer of sample onto the grid, thus controlling the ice thickness and reducing the exposure of the sample to the atmosphere by performing a fast plunge without requiring any blotting of the grid [45]–[51].

Developments in terms of the type of grids used in Cryo-EM constitute another option to optimize the sample preparation [52][53]. For instance, the use of polyethylene glycol (PEG) gold grids may help to reduce protein aggregation and improve particle orientation within the vitreous ice layer [54].

The use of additives prior to grid vitrification, such as mild detergents or amphipols, can reduce denaturation and dissociation of the protein complex [55], however the array of additives must be screened and extended, so frequent access to Cryo-EM facilities is necessary to identify the best conditions for grid preparation and vitrification.

Nevertheless, it was possible to partially fit the 3D models of AftA (predicted by Alphafold [56], P9WN02) and Mb41 scaffold protein (*E. coli* K12 Glucosidase YgjK, PDB ID: 3W7T) onto the 3D models generated (Figure 5.7A), wherein the 3D model of the megabody scaffold protein YgjK appears to match the size of the lower half of the generated volume, as presented in Figure 5.7A, while the fit of AftA into the volume is not as good, being the AftA model slightly larger than the generated volume. Considering that AftA is embedded in a lipid filled nanodisc, the expected volume would be larger and would account for the electron density of both AftA and the nanodisc, which includes the nanodisc scaffold protein and ordered lipids.

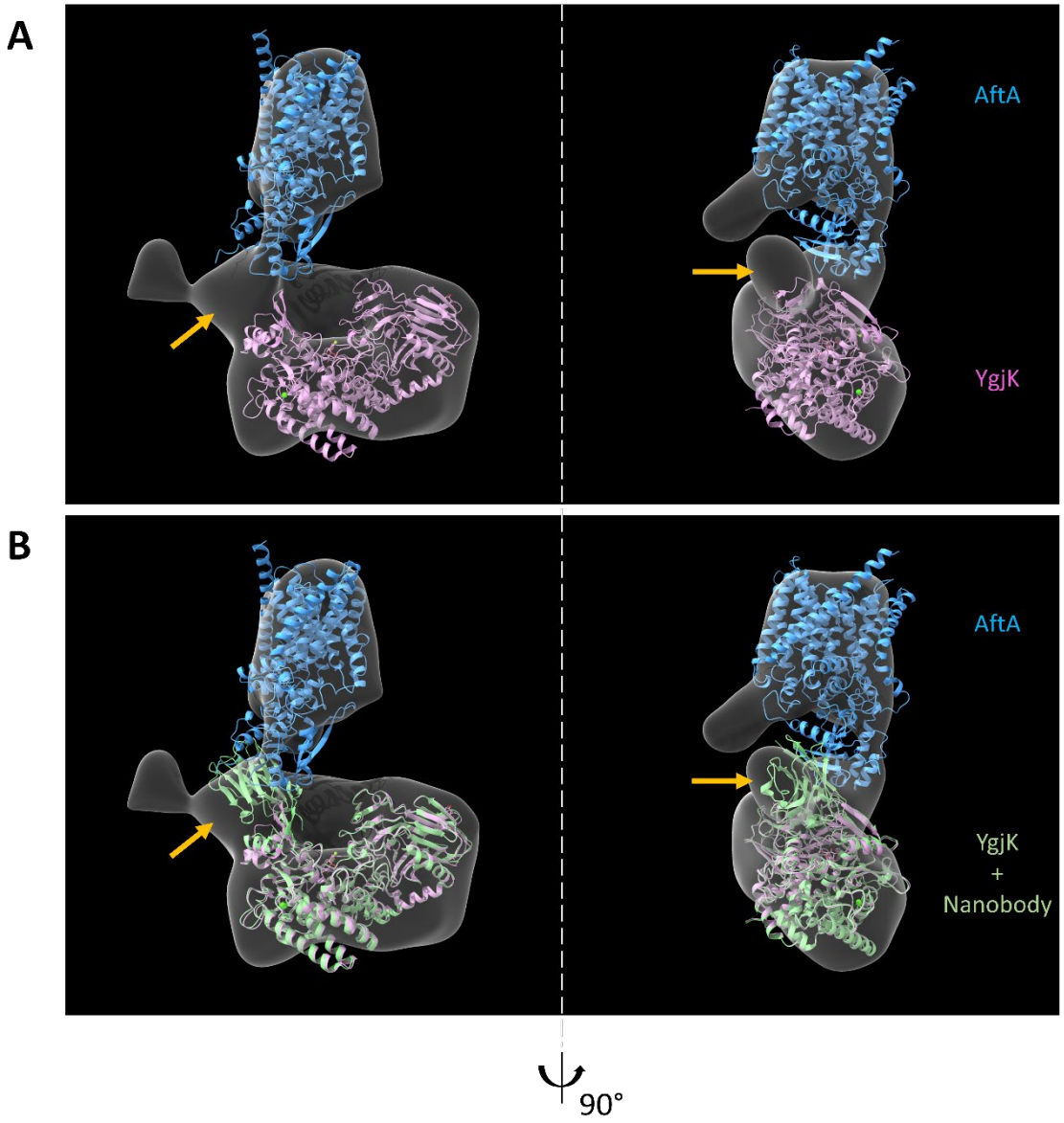


Figure 5.7 – Model fitting of AftA (Alphafold P9WN02 – blue), Mb41 scaffold protein (E. coli K12 Glucosidase YgjK, PDB ID: 3W7T – pink) and YgjK-based megabody (PDB ID: 6XUX – green) to the AftA-Mb41 3D class of subset 2. (A)(B) *AftA is fitted to the upper half of the 3D volume, although the model appears to be larger than the generated volume, while Mb41 protein scaffold fits the lower half of the 3D volume.* ***(C)(D)*** *The model of YgjK-based megabody model (green) overlaps perfectly with Mb41 protein scaffold (pink) wherein the nanobody portion of the model is positioned at the interface with AftA, indicated by a yellow arrow.*

It is possible that this small volume would be due to the small number of particles in the dataset, as well as their orientation. Moreover, Figure 5.7B shows the fit of AftA and the megabody scaffold protein YgjK fused to a nanobody (PDB ID: 6XUX) which corresponds to the crystal structure of a YgjK-based megabody used in a study for the development of Cryo-EM toolkits [18]. As expected, the model overlaps perfectly with the model of YgjK, wherein the YgjK-fused nanobody fills a volume which is close to the fitted AftA model. Although the 3D model is significantly incomplete and with poor resolution,

affecting model-fit algorithms, the overall shape and size of the generated volume suggests that it may indeed correspond to the protein complex of AftA with an anti-AftA megabody: Mb41.

Considering the limitations of the presented data, it is relevant to ask the question: what should be the next steps to determine the structure of the protein complex of AftA-Mb41?

In order to determine the structure of the protein complex, we have to address two issues: the biological sample and the limited dataset.

Regarding the sample, there is always room to optimize the sample purification, *e.g.*, reduce freeze-thaw cycles to the minimum and/or use crosslinking agents, although such optimization takes a lot of time and ultimately needs to be validated at the electron microscope, for which we do not have yet a regular and frequent access in Portugal.

Regarding the limited dataset, to determine any structure by single particle Cryo-EM, we need a dataset having a sufficient number of particles in all possible orientations, so that we can reconstruct a 3D model out of the 2D projections of all particles. Therefore, to determine the structure of AftA-Mb41 complex, we will have to prepare more grids and collect more data to massively increase the number of particles to have enough particles to solve the structure. This is a situation where there is a real “strength in numbers”, because even if the sample would still be heterogeneous and most particles would not correspond to the complex, if the number of particles in the dataset is high enough, we can confidently discard most of the particles because the remaining particles would be sufficient to solve the structure.

6. Conclusions

Here, we present a preliminary low-resolution 3D model of a AftA-megabody protein complex. The use of specific anti-AftA megabodies facilitate the single particle Cryo-EM workflow for the structure determination of small particles, by increasing the overall size of the particles. However, the structural characterization of protein complexes by Cryo-EM comprise additional challenges, such as sample heterogeneity and protein complex dissociation and/or denaturation. The collected dataset of AftA-Mb41 was heterogeneous and contained particles of different sizes, a small size population and a large size population, in which the percentage of small particles (~90%) far exceeded the percentage of large particles (~10%), wherein the large particles corresponded to the assembled AftA-Mb41 protein complex. Further studies are necessary to improve the resolution of the 3D model to fully determine the 3D structure of AftA and pave the way for structure-based drug design targeting AftA function of priming arabinogalactan synthesis in mycobacteria.

7. References

[1] C. E. Barry, “Interpreting cell wall ‘virulence factors’ of *Mycobacterium tuberculosis*,” *Trends Microbiol.*, vol. 9, no. 5, pp. 237–241, May 2001

[2] M. Jankute, J. A. Cox, J. Harrison, and G. S. Besra, “Assembly of the Mycobacterial Cell Wall,” *Annu Rev Microbiol*, vol. 69, pp. 405–423, 2015

[3] B. Zuber, M. Chami, C. Houssin, J. Dubochet, G. Griffiths, and M. Daffé, “Direct visualization of the outer membrane of mycobacteria and corynebacteria in their native state,” *J. Bacteriol.*, vol. 190, no. 16, pp. 5672–5680, Aug. 2008

[4] M. Sani, E. N. G. Houben, J. Geurtsen, J. Pierson, and K. De Punder, “Direct Visualization by Cryo-EM of the Mycobacterial Capsular Layer: A Labile Structure Containing ESX-1-Secreted Proteins,” *PLoS Pathog*, vol. 6, no. 3, p. 1000794, 2010

[5] M. Daffé, “The cell envelope of tubercle bacilli,” *Tuberculosis*, vol. 95, no. S1, pp. S155–S158, Jun. 2015

[6] M. Daffé and H. Marrakchi, “Unraveling the Structure of the Mycobacterial Envelope,” *Microbiol. Spectr.*, vol. 7, no. 4, Jul. 2019

[7] M. McNeil, S. J. Wallner, S. W. Hunter, and P. J. Brennan, “Demonstration that the galactosyl and arabinosyl residues in the cell-wall arabinogalactan of *Mycobacterium leprae* and *Mycobacterium tuberculosis* are furanoid,” *Carbohydr. Res.*, vol. 166, no. 2, pp. 299–308, Sep. 1987

[8] M. Jankute, S. Grover, A. K. Rana, and G. S. Besra, “Arabinogalactan and lipoarabinomannan biosynthesis: structure, biogenesis and their potential as drug targets,” *Futur. Microbiol*, vol. 7, no. 1, pp. 129–147, 2011

[9] M. McNeil, M. Daffe, and P. J. Brennan, “Evidence for the nature of the link between the arabinogalactan and peptidoglycan of mycobacterial cell walls.,” *J. Biol. Chem.*, vol. 265, no. 30, pp. 18200–18206, Oct. 1990

[10] B. A. Wolucka, M. R. McNeil, E. De Hoffmann, T. Chojnacki, and P. J. Brennan, “Recognition of the lipid intermediate for arabinogalactan/arabinomannan biosynthesis and its relation to the mode of action of ethambutol on mycobacteria.,” *J. Biol. Chem.*, vol. 269, no. 37, pp. 23328–23335, Sep. 1994

[11] K. Mikusova *et al.*, “Decaprenylphosphoryl arabinofuranose, the donor of the D-

arabinofuranosyl residues of mycobacterial arabinan, is formed via a two-step epimerization of decaprenylphosphoryl ribose," *J Bacteriol*, vol. 187, no. 23, pp. 8020–8025, 2005

[12] L. J. Alderwick, M. Seidel, H. Sahm, G. S. Besra, and L. Eggeling, "Identification of a novel arabinofuranosyltransferase (AftA) involved in cell wall arabinan biosynthesis in *Mycobacterium tuberculosis*," *J Biol Chem*, vol. 281, no. 23, pp. 15653–15661, 2006

[13] L. J. Alderwick *et al.*, "Deletion of Cg-emb in corynebacterianeae leads to a novel truncated cell wall arabinogalactan, whereas inactivation of Cg-ubiA results in an arabinan-deficient mutant with a cell wall galactan core," *J. Biol. Chem.*, vol. 280, no. 37, pp. 32362–32371, Sep. 2005

[14] L. Shi *et al.*, "Transfer of the first arabinofuranose residue to galactan is essential for *Mycobacterium smegmatis* viability," *J Bacteriol*, vol. 190, no. 15, pp. 5248–5255, 2008

[15] A. Kapopoulou, J. M. Lew, and S. T. Cole, "The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes," *Tuberculosis (Edinb).*, vol. 91, no. 1, pp. 8–13, Jan. 2011

[16] R. Nygaard, J. Kim, and F. Mancia, "Cryo-electron microscopy analysis of small membrane proteins," *Current Opinion in Structural Biology*, vol. 64. Elsevier Ltd, pp. 26–33, Oct. 01, 2020

[17] E. Pardon *et al.*, "A general protocol for the generation of Nanobodies for structural biology," *Nat. Protoc. 2014 93*, vol. 9, no. 3, pp. 674–693, Feb. 2014

[18] T. Uchański *et al.*, "Megabodies expand the nanobody toolkit for protein structure determination by single-particle cryo-EM," *Nat. Methods 2021 181*, vol. 18, no. 1, pp. 60–68, Jan. 2021

[19] J. Love *et al.*, "The New York Consortium on Membrane Protein Structure (NYCOMPS): A high-throughput platform for structural genomics of integral membrane proteins," *J. Struct. Funct. Genomics*, vol. 11, no. 3, pp. 191–199, Sep. 2010

[20] R. Bruni and B. Kloss, "High-Throughput Cloning and Expression of Integral Membrane Proteins in *Escherichia coli*," *Curr. Protoc. Protein Sci.*, vol. 74, no. 1, pp. 29.6.1-29.6.34, Nov. 2013

[21] F. Sievers *et al.*, "Fast, scalable generation of high-quality protein multiple

sequence alignments using Clustal Omega," *Mol. Syst. Biol.*, vol. 7, 2011

[22] X. Robert and P. Gouet, "Deciphering key features in protein structures with the new ENDscript server," *Nucleic Acids Res.*, vol. 42, no. Web Server issue, Jul. 2014

[23] T. H. Bayburt and S. G. Sligar, "Membrane protein assembly into Nanodiscs," *FEBS Lett.*, vol. 584, no. 9, pp. 1721–1727, May 2010

[24] A. Javaheri *et al.*, "*Helicobacter pylori* adhesin HopQ engages in a virulence-enhancing interaction with human CEACAMs," *Nat. Microbiol.*, vol. 2, Oct. 2016

[25] Y. Kurakata *et al.*, "Structural Insights into the Substrate Specificity and Function of *Escherichia coli* K12 YgjK, a Glucosidase Belonging to the Glycoside Hydrolase Family 63," *J. Mol. Biol.*, vol. 381, no. 1, pp. 116–128, Aug. 2008

[26] S. Q. Zheng, E. Palovcak, J. P. Armache, K. A. Verba, Y. Cheng, and D. A. Agard, "MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy," *Nat. Methods 2017 144*, vol. 14, no. 4, pp. 331–332, Feb. 2017

[27] A. Rohou and N. Grigorieff, "CTFFIND4: Fast and accurate defocus estimation from electron micrographs," *J. Struct. Biol.*, vol. 192, no. 2, pp. 216–221, Nov. 2015

[28] K. Zhang, "Gctf: Real-time CTF determination and correction," *J. Struct. Biol.*, vol. 193, no. 1, p. 1, Jan. 2016

[29] J. M. De la Rosa-Trevín *et al.*, "Xmipp 3.0: An improved software suite for image processing in electron microscopy," *J. Struct. Biol.*, vol. 184, no. 2, pp. 321–328, Nov. 2013

[30] C. O. Sorzano *et al.*, "Semiautomatic, high-throughput, high-resolution protocol for three-dimensional reconstruction of single particles in electron microscopy," *Methods Mol. Biol.*, vol. 950, pp. 171–193, 2013

[31] T. Wagner *et al.*, "SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM," *Commun. Biol.*, vol. 2, no. 1, Dec. 2019

[32] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker, "cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination," *Nat. Methods 2017 143*, vol. 14, no. 3, pp. 290–296, Feb. 2017

[33] D. Kimanius, L. Dong, G. Sharov, T. Nakane, and S. H. W. Scheres, "New tools

for automated cryo-EM single-particle analysis in RELION-4.0," *Biochem. J.*, vol. 478, no. 24, pp. 4169–4185, Dec. 2021

[34] D. Kimanius, B. O. Forsberg, S. H. W. Scheres, and E. Lindahl, "Accelerated cryo-EM structure determination with parallelisation using GPUS in RELION-2," *Elife*, vol. 5, Nov. 2016

[35] J. Zivanov *et al.*, "New tools for automated high-resolution cryo-EM structure determination in RELION-3," *Elife*, vol. 7, Nov. 2018

[36] S. H. W. Scheres, "RELION: Implementation of a Bayesian approach to cryo-EM structure determination," *J. Struct. Biol.*, vol. 180, no. 3, pp. 519–530, Dec. 2012

[37] E. F. Pettersen *et al.*, "UCSF Chimera—A visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004

[38] F. Mancia and J. Love, "High-throughput expression and purification of membrane proteins," *J. Struct. Biol.*, vol. 172, no. 1, pp. 85–93, Oct. 2010

[39] J. Rodrigues *et al.*, "High throughput expression screening of arabinofuranosyltransferases from mycobacteria," *Processes*, vol. 9, no. 4, p. 629, Apr. 2021

[40] E. D'Imprima, D. Floris, M. Joppe, R. Sánchez, M. Grininger, and W. Kühlbrandt, "Protein denaturation at the air-water interface and how to prevent it," *Elife*, vol. 8, Apr. 2019

[41] K. A. Taylor and R. M. Glaeser, "Retrospective on the early development of cryoelectron microscopy of macromolecules and a prospective on opportunities for the future," *J. Struct. Biol.*, vol. 163, no. 3, pp. 214–223, Sep. 2008

[42] B. Kastner *et al.*, "GraFix: Sample preparation for single-particle electron cryomicroscopy," *Nat. Methods*, vol. 5, no. 1, pp. 53–55, Jan. 2008

[43] H. Stark, "GraFix: Stabilization of fragile macromolecular complexes for single particle Cryo-EM," in *Methods in Enzymology*, vol. 481, no. C, Academic Press Inc., 2010, pp. 109–126.

[44] E. Y. D. Chua *et al.*, "Better, Faster, Cheaper: Recent Advances in Cryo-Electron Microscopy," *Annu. Rev. Biochem.*, vol. 91, pp. 1–32, 2022

[45] X. Feng *et al.*, "A Fast and Effective Microfluidic Spraying-Plunging Method for High-Resolution Single-Particle Cryo-EM," *Structure*, vol. 25, no. 4, pp. 663-

670.e3, Apr. 2017

[46] T. Jain, P. Sheehan, J. Crum, B. Carragher, and C. S. Potter, “Spotiton: A prototype for an integrated inkjet dispense and vitrification system for cryo-TEM,” *J. Struct. Biol.*, vol. 179, no. 1, pp. 68–75, Jul. 2012

[47] V. P. Dandey *et al.*, “Spotiton: New features and applications,” *J. Struct. Biol.*, vol. 202, no. 2, pp. 161–169, May 2018

[48] D. Ashtiani *et al.*, “Delivery of femtolitre droplets using surface acoustic wave based atomisation for cryo-EM grid preparation,” *J. Struct. Biol.*, vol. 203, no. 2, pp. 94–101, Aug. 2018

[49] S. A. Arnold *et al.*, “Blotting-free and lossless cryo-electron microscopy grid preparation from nanoliter-sized protein samples and single-cell extracts,” *J. Struct. Biol.*, vol. 197, no. 3, pp. 220–226, Mar. 2017

[50] R. B. G. Ravelli *et al.*, “Cryo-EM structures from sub-nl volumes using pin-printing and jet vitrification,” *Nat. Commun. 2020 111*, vol. 11, no. 1, pp. 1–9, May 2020

[51] J. L. Rubinstein *et al.*, “Shake-it-off: A simple ultrasonic cryo-EM specimen-preparation device Rubinstein John L.,” *Acta Crystallogr. Sect. D Struct. Biol.*, vol. 75, no. 12, pp. 1063–1070, Dec. 2019

[52] H. Wei *et al.*, “Optimizing ‘self-wicking’ nanowire grids,” *J. Struct. Biol.*, vol. 202, no. 2, pp. 170–174, May 2018

[53] Y. Z. Tan and J. L. Rubinstein, “Through-grid wicking enables high-speed cryoEM specimen preparation,” *Acta Crystallogr. Sect. D Struct. Biol.*, vol. 76, no. 11, pp. 1092–1103, Nov. 2020

[54] J. R. Meyerson *et al.*, “Self-assembled monolayers improve protein distribution on holey carbon cryo-EM supports,” *Sci. Reports 2014 41*, vol. 4, no. 1, pp. 1–5, Nov. 2014

[55] B. C. Choy, R. J. Cater, F. Mancia, and E. E. Pryor, “A 10-year meta-analysis of membrane protein structural biology: Detergents, membrane mimetics, and structure determination techniques,” *Biochim. Biophys. Acta - Biomembr.*, vol. 1863, no. 3, p. 183533, Mar. 2021

[56] J. Jumper *et al.*, “Highly accurate protein structure prediction with AlphaFold,” *Nat. 2021 5967873*, vol. 596, no. 7873, pp. 583–589, Jul. 2021

8. Acknowledgments

Author Contributions

Conceptualization: M. Archer, F. Mancia

Genomics expansion and small-scale screening: A. L. Rosário, B. Kloss and *J. Rodrigues*

Nanobody discovery: A. Lundqvist, E. Pardon, J. Steyaert

Megabody development: A. Lundqvist, E. Pardon, J. Steyaert

AftA Large-scale expression, purification and negative-stain EM: *J. Rodrigues*

Nanobody/Megabody Large-scale expression and purification: V. T. Almeida, *J. Rodrigues*

Surface Plasmon Resonance assays: Ana R. Lemos, Pedro M. F. Sousa

Grid vitrification: A. Domanska, *J. Rodrigues*

Cryo-EM data collection, processing and analysis: P. Laurinmäki, A. Domanska, *J. Rodrigues*, S. Butcher

Project supervision: M. Archer

Funding

This research was funded by Fundação para a Ciência e Tecnologia (FCT), Lisbon, Portugal; grants (PTDC/BIA-BQM/30421/2017 and PTDC/BIA-BQM/4056/2020 to M.A. and PD/BD/128261/2016 to J.R.), European Union's Horizon 2020 research and innovation programme under grant agreements: No. 857203 (Twinning), No. 823780 (MSCA-RISE) and No. 731005, Instruct-ULTRA, a project to further develop the services of Instruct-ERIC (M.A, J.R. and V.T.A.).

9. Supplemental Information

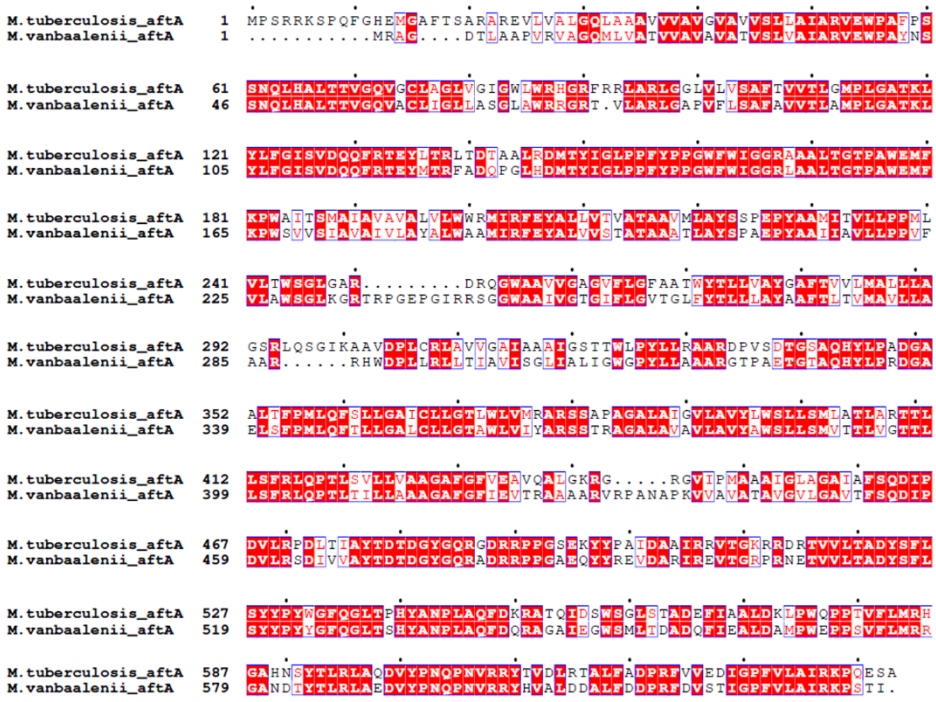


Figure S5.1 – Sequence alignment of AftA from M. vanbaalenii versus M. tuberculosis.
The sequence alignment between M. vanbaalenii AftA (GenBank: ABM16394.1) and M. tuberculosis AftA (Rv3792) was generated using ClustalO [21] *and displayed using ESPript* [22] *. Identical residues are depicted in a red-filled box with white font, whereas residues with more than 70% similarity are marked in a blue bordered box with red font.*

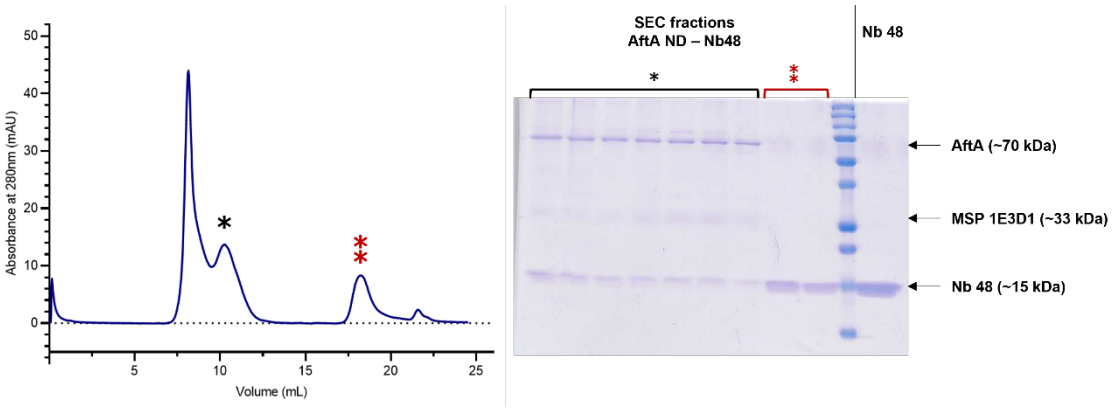


Figure S5.2 – Purification of protein complex of AftA nanodiscs (AftA ND) with nanobody 48 (Nb48). ***(A)*** *Size exclusion chromatography (SEC) profile of the purification of AftA ND-Nb48 complex.* ***(B)*** *SDS-PAGE of collected fractions. Asterisks indicate eluted peaks, wherein (⁎) corresponds to a high molecular weight peak, and (⁑) corresponds to a low molecular weight peak.*

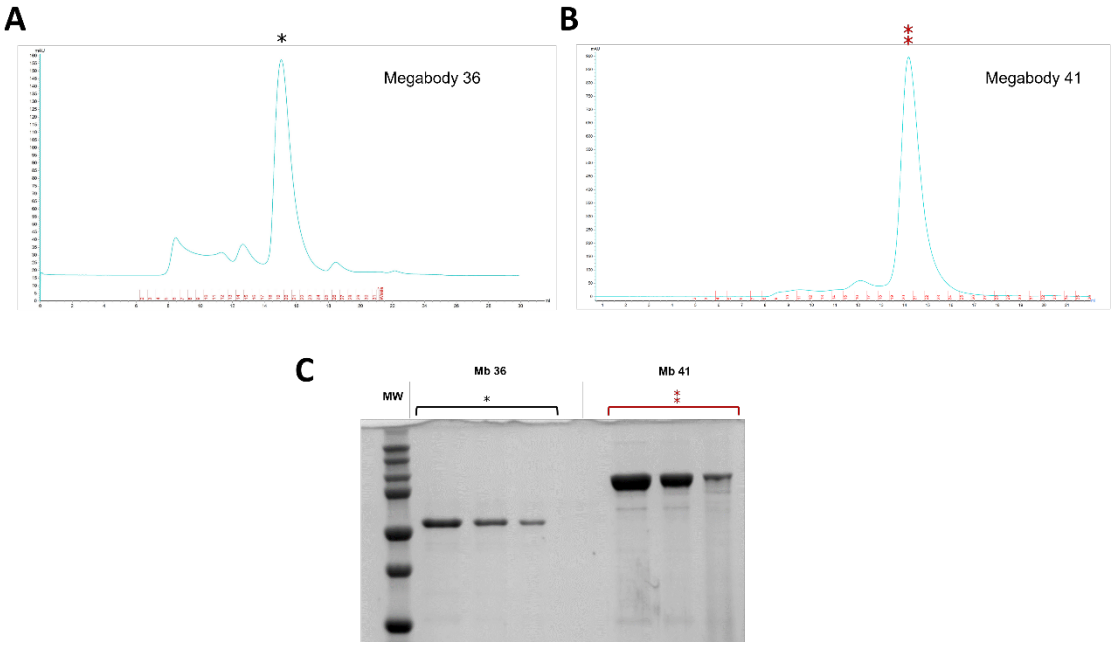


Figure S5.3 – Purification of anti-AftA megabodies: Mb36 (50 kDa) and Mb41 (100 kDa). (A) *Size exclusion chromatography (SEC) profile of the purification of Mb36.* ***(B)*** *Size exclusion chromatography (SEC) profile of the purification of Mb41.* ***(C)*** *SDS-PAGE of collected fractions of Mb36 and Mb41, respectively.*

Table S5.1 – Anti-AftA nanobody discovery and selection.

Immunization with AftA (*M. vanbaalenii* PYR-1)				
Nanobody ID	**Family[1]**	**Occurrence**	**Selection conditions in which the Nanobody is found**	
			Solid Phase Coated	**Flag Captured**
Nb09	1	53	✓	✓
Nb14				
Nb15				
Nb16				
Nb19				
Nb20				
Nb21				
Nb22				
Nb23				
Nb24				
Nb25				
Nb26				
Nb29				
Nb30				
Nb31				
Nb32				
Nb33				
Nb34				
Nb35				
Nb36				
Nb37				
Nb38				
Nb39				
Nb40				
Nb41				
Nb45	2	2	✓	
Nb43	3	2	✓	
Nb44	4	1	✓	
Nb46	5	1	✓	
Nb48	6	1	✓	
Nb42	7	1		✓
Nb47	8	1		✓
Nb49	9	2		✓
Nb50	10	1		✓
Nb51	11	1		✓

[1] A Nanobody family is defined as group of nanobody amino acid sequences with high similarity in the CDR3 sequence, thus binding to the same epitope in the target protein.

CHAPTER SIX

General Discussion and Future Work

General Discussion

This work aimed to advance knowledge in the fight against a disease that still afflicts millions of people worldwide every year. The goal was the structural characterization of mycobacterial arabinofuranosyltransferases (AraTs), which are responsible for the synthesis of AG, a major constituent of the cell wall, a main defensive structure of mycobacteria.

The proposed approach is not unique, in fact, once a protein is identified as a candidate therapeutic target, the corner stone of designing new drugs against this target is the determination of its 3D structure. Mycobacterial AraTs have long been identified as promising therapeutic targets, in which Embs are effectively targeted by the widely used first-line drug ethambutol to treat patients with active TB. Even though ethambutol is already a known inhibitor of EmbA, EmbB and EmbC, its specific inhibition mechanism was unknown as there was not a reported structure for these proteins, with or without ethambutol.

Therefore, the structure determination of any mycobacterial AraT would be a step forward to know more about how do these proteins function in the context of the mycobacterial cell wall synthesis; and how can their activity be inhibited by ethambutol or by novel compounds.

To elucidate the structure of any protein, it is essential to consistently produce the protein of interest, as most techniques for structural determination require a fair amount of protein. However, every so often the target protein is not easily expressed and/or produced, therefore a homolog structure may be determined instead. In this scenario, any feature resulting from the structural characterization of the homolog protein must be carefully inferred depending on the conservation degree to the actual target protein.

In Chapter II, we adopted a high throughput approach to screen the expression and production of a large set of AraTs, which included AraTs from *M. tuberculosis* orthologs from other *Mycobacterium* species, showing high degree of sequence identity to the corresponding AraT from *M. tuberculosis*. From the seven AraTs present in the set (AftA, AftB, AftC, AftD, EmbA, EmbB and EmbC), we successfully selected at least five AraT representatives: AftA and AftC from *M. neoaurum*, AftD from *M. abscessus* 1948 F5/8, EmbA from *M. marinum* M. and EmbB from *M. vanbaalenii* PYR-1.

Although the majority of targets in the initial set were either not expressed nor purified and would require further optimization, we find that the adopted strategy served its

purpose of selecting several mycobacterial AraT as candidates for structural studies. Nevertheless, it was important to scale-up the production of selected proteins to proceed to structural characterization, which require a substantial amount of protein. The scale-up step at the end of the high throughput approach allowed the identification of two distinct oligomeric states of EmbB from *M. smegmatis*, being this result confirmed by Cryo-EM structures of EmbB as monomer [1] (work described in Chapter IV: *Cryo-EM structure of arabinosyltransferase EmbB from Mycobacterium smegmatis*) and EmbB as homodimer [2].

This brute-force approach also set the base to elucidate the structures of AftD from *M. abscessus* [3] and EmbB from *M. smegmatis* [1], described in Chapters III and IV, respectively.

The two structures solved in this work, AftD and EmbB, were novel and belonged to the GT-C superfamily of glycosyltransferases [4]. Although the specific function and role of each protein within AG biosynthesis does not originate solely from the characteristic GT-C fold, the high structural conservation with other GT-C like proteins helped to fill some experimental gaps and gain insights into their reaction mechanisms and substrate binding, which would be impossible to infer exclusively from the *apo* structures of AftD and EmbB. The structural conservation between different GT-C glycosyltransferases helped to identify the catalytic pocket in both AftD and EmbB and explained the existence of the disordered external loop located across the catalytic pocket in both *apo* structures.

Ideally, such mechanistic insights would be provided by the characterization of acceptor or donor-bound AraT structures. Yet, such structures were not obtained in this work. Zhang and colleagues were able to solve the structure of a DPA-bound $EmbB_2$-$AcpM_2$ complex, wherein the lipid donor DPA is present at the interface between the EmbB protomers and drives conformational changes of several amino acid residues leading a periplasmic loop to form a helix, which contains residues that interact with the bound DPA [2]. These findings are aligned with the proposed mechanism for AftD, in which we suggested that N-EL4 loop should become ordered to interact with bound DPA.

Nevertheless, although the structural conservation of the GT-C core may help to unveil the overall mechanism of these two proteins, there are still significant structural differences that are worth emphasizing. AftD possesses three distinct CBMs, while EmbB possesses two CBM, wherein the spatial arrangement of the CBMs in each protein as well as the type of residues at the surface of each CBM may have influence in the selectivity of AraTs, in the sense that some AraTs may have preference towards branched arabinose chains, and other AraTs may prefer linear arabinose chains.

The other significant difference was the interaction of the C-terminal TM domain of AftD with an acyl carrier protein (ACP) from the host organism *E. coli*, in the cytoplasmic side of the membrane. This interaction with an endogenous partner from the host organism is not observed in the EmbB structure herein characterized.

However, as previously referred in Chapter IV, the recently solved structure of heterodimer EmbA-EmbB from *M. tuberculosis* showed both EmbA and EmbB C-terminal TM domains interacting with a meromycolate extension ACP (AcpM) of *M. smegmatis*, the host organism for protein production [5]. Moreover, another homodimeric structure of EmbB from *M. smegmatis* also showed two AcpM bound to the C-terminal TM domain of each EmbB protomer [2]. The same publication also described a deficient branching of AG when the interaction of ACP is abolished, which clearly suggests that ACP has a regulatory role on the transferase activity of EmbB [2].

These findings contradict the suggested hypothesis, in Chapter IV, that ACP would not have a regulatory role in EmbB, as there was not a significant conformational change of the monomeric EmbB in comparison with the published heterodimeric EmbA-EmbB structure. Then again, these findings are aligned with the proposed regulatory role of ACP for the transferase activity of AftD, in Chapter III.

Although the specific role of ACP as regulator of AraT activity is still not fully understood, the results obtained in this work regarding the interaction between *E. coli* ACP and AftD-WT or AftD-R1389S preceded the findings regarding the interaction between *M. tuberculosis* Emb proteins (EmbA, EmbB and EmbC) and the AcpM of the host organism *M. smegmatis* [5]. In addition, Zhang and colleagues also show that if the interaction between AcpM and EmbC is affected, the resulting LAM species is smaller [5].

Therefore, these findings hint that ACP does have a regulatory role to modulate the activity of AraTs, and further studies are required to fully understand how it works.

Regarding AftA, studied in Chapter V, we were not able to determine the structure of AftA from *M. vanbaalenii*. Although the experimental approach of using megabodies was successful to determine other protein structures [6]–[8], the collected Cryo-EM dataset did not produce enough particles to allow the characterization of AftA-megabody complex structure.

Recently, the structure of AftA from *M. tuberculosis* was solved by Cryo-EM, as a homodimer [9]. Comparing the solved structure of AftA with the preliminary 3D model obtained in this work, it appears that the model is representative of a protein complex between a single monomer of AftA from *M. vanbaalenii* PYR-1 and a megabody (Mb41),

which suggests that AftA from *M. vanbaalenii* PYR-1 is produced in *E. coli* as a monomer, similarly to the results obtained for AftD from *M. abscessus* [3] and EmbB from *M. smegmatis* [1], which were also produced in *E. coli*.

Although there is still no 3D structure of AftD produced in *M. tuberculosis* or *M. smegmatis*, the fact that EmbB and AftA produced in *M. smegmatis* are oligomers (homodimer AftA-AftA, EmbB-EmbB; or heterodimers EmbA-EmbB) suggest that the host organism chosen for protein production have an impact on the oligomeric state of these AraT. So, we are currently making efforts to solve the structure of AftD produced in *M. smegmatis*.

Interestingly, despite belonging to GT-C superfamily, sharing structural and functional conservation, AftA does not interact with any endogenous partner, such as AcpM from *M. smegmatis*. Although AftA does present a positively charged surface in the cytoplasmic side, similarly to AftD and all Emb proteins interacting with AcpM (with negatively charged surface), it does not have any hydrophobic pocket in its TM domain in which a 4'-phosphopantetheine could be inserted (Figure 1). This may suggest that the hydrophobic pocket in the TM domain for the 4'-phosphopantetheine of ACP might be a prerequisite for the interaction between AraTs and AcpM. This finding is relevant in the context of the present work, since it restricts the role of ACP as putative regulator to only *some* AraTs and not all of them.

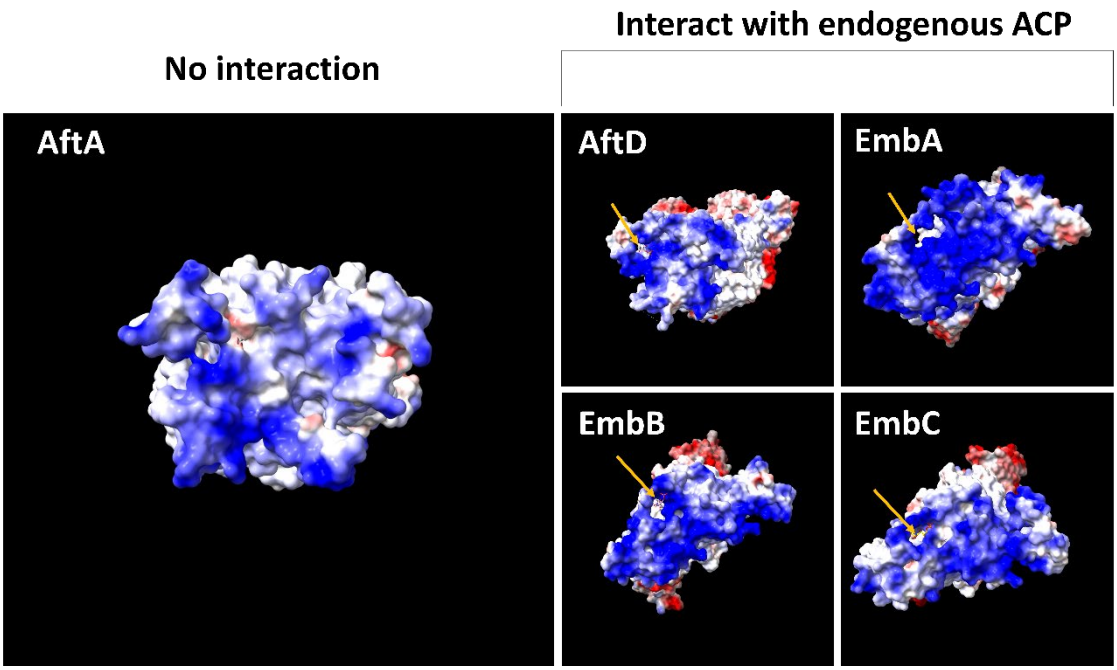


Figure 6.1 – Comparison of the cytoplasmic side of the solved structures of mycobacterial AraTs: AftA, AftD, EmbA, EmbB and EmbC. *The surfaces of the structures are colored according to their electrostatic potential, where red is more negatively charged and blue more positively charged. The arrows point to the hydrophobic pocket present in AftD, EmbA, EmbB and EmbC.*

Future Perspectives

Based on this work, there is still much more work to be done aiming for new therapeutics against TB. I would like to highlight some of the future research paths, both fundamental and applied research, some of which are already ongoing:

1. Study of the regulatory role of ACP in the cell wall biosynthesis.
2. Analysis of AraT mutations related to drug-resistance TB: EmbB.
3. Drug repurposing and drug design against AraTs: AftD and EmbB.
4. Structural characterization of the remaining mycobacterial AraTs: AftB and AftC.
5. Structural characterization of AftD, wild-type and mutants, produced in *M. smegmatis*.

The study of the regulatory role of ACP in the cell wall biosynthesis will be very challenging, consider that it is now known that ACP interacts with multiple AraTs, thereby having an impact in several stages of the mycobacterial cell wall biosynthesis. It will be important to have experimental snapshots of each step of the interaction/regulation of ACP with AraTs. We are currently developing activity assays for AraTs to study several AftD mutants in which the ACP interaction is hindered or abolished. At the same time, we hope to structurally characterize these mutants to have a better understanding of the conformational changes of AftD without ACP bound.

Following the drug design pathway, which starts with the determination of 3D structures of the target proteins, we now have the solved structures of AftA, AftD, EmbA, EmbB and EmbC, as well as the predicted structures of AftB and AftC by Alphafold. These structures are crucial to support *in silico* approaches to study the dynamic behavior of these proteins and to virtually test putative substrates or inhibitors. This approach is particularly useful to study EmbB, which is the AraT with most identified mutations conferring drug resistance. We are currently studying the dynamic behavior of ethambutol bound to the catalytic pocket of wild-type EmbB compared to the M306V mutant EmbB. Using molecular dynamics, we have found that the binding of ethambutol is severely weakened in the M306V mutant EmbB, in which we observe the outwards displacement of the ethambutol molecule from the EmbB catalytic pocket, which does not occur in the wild-type EmbB (see Appendix I). Other mutations can be studied using the same approach for EmbB and for other solved AraTs.

References

[1] Y. Z. Tan *et al.*, “Cryo-EM structure of arabinosyltransferase EmbB from *Mycobacterium smegmatis*,” *Nat. Commun.*, vol. 11, no. 1, p. 3396, 2020

[2] L. Zhang *et al.*, “Cryo-EM snapshots of mycobacterial arabinosyltransferase complex EmbB(2)-AcpM(2),” *Protein Cell*, vol. 11, no. 7, pp. 505–517, 2020

[3] Y. Z. Tan *et al.*, “Cryo-EM Structures and Regulation of Arabinofuranosyltransferase AftD from Mycobacteria,” *Mol. Cell*, vol. 78, no. 4, pp. 683-699.e11, May 2020

[4] J. Liu and A. Mushegian, “Three monophyletic superfamilies account for the majority of the known glycosyltransferases,” *Protein Sci.*, vol. 12, no. 7, pp. 1418–1431, Jul. 2003

[5] L. Zhang and Y. Zhao, “Structures of cell wall arabinosyltransferases with the anti-tuberculosis drug ethambutol,” *Science*, vol. 368, no. 6496, pp. 1211–1219, 2020

[6] C. E. Coupland *et al.*, “ll Structure, mechanism, and inhibition of Hedgehog acyltransferase,” *Mol. Cell*, vol. 81, pp. 5025-5038.e10, 2021

[7] Z. Chen *et al.*, “Cryo-EM structures of human SPCA1a reveal the mechanism of Ca^{2+}/Mn^{2+} transport into the Golgi apparatus,” *Sci. Adv.*, vol. 9, no. 9, Mar. 2023

[8] K. Wentinck, C. Gogou, and D. H. Meijer, “Putting on molecular weight: Enabling cryo-EM structure determination of sub-100-kDa proteins,” *Curr. Res. Struct. Biol.*, vol. 4, pp. 332–337, Jan. 2022

[9] Y. Gong *et al.*, “Structure of the priming arabinosyltransferase AftA required for AG biosynthesis of *Mycobacterium tuberculosis*,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 120, no. 23, p. e2302858120, Jun. 2023

Conclusion

In conclusion, the work disclosed in this PhD thesis contributed to better understand the function of mycobacterial AraTs through the structural characterization of AftD from *M. abscessus* and EmbB from *M. smegmatis*. It also revealed the interaction of mycobacterial AraTs with ACP, suggesting that ACP has a regulatory role on the transferase activity of mycobacterial AraTs affecting the cell wall biosynthesis in mycobacteria. This work provides a basis for a more in-depth study on their involvement in the biosynthesis of mycobacteria cell wall and lays the foundation for drug design studies for novel therapeutics against Tuberculosis.

Appendix I

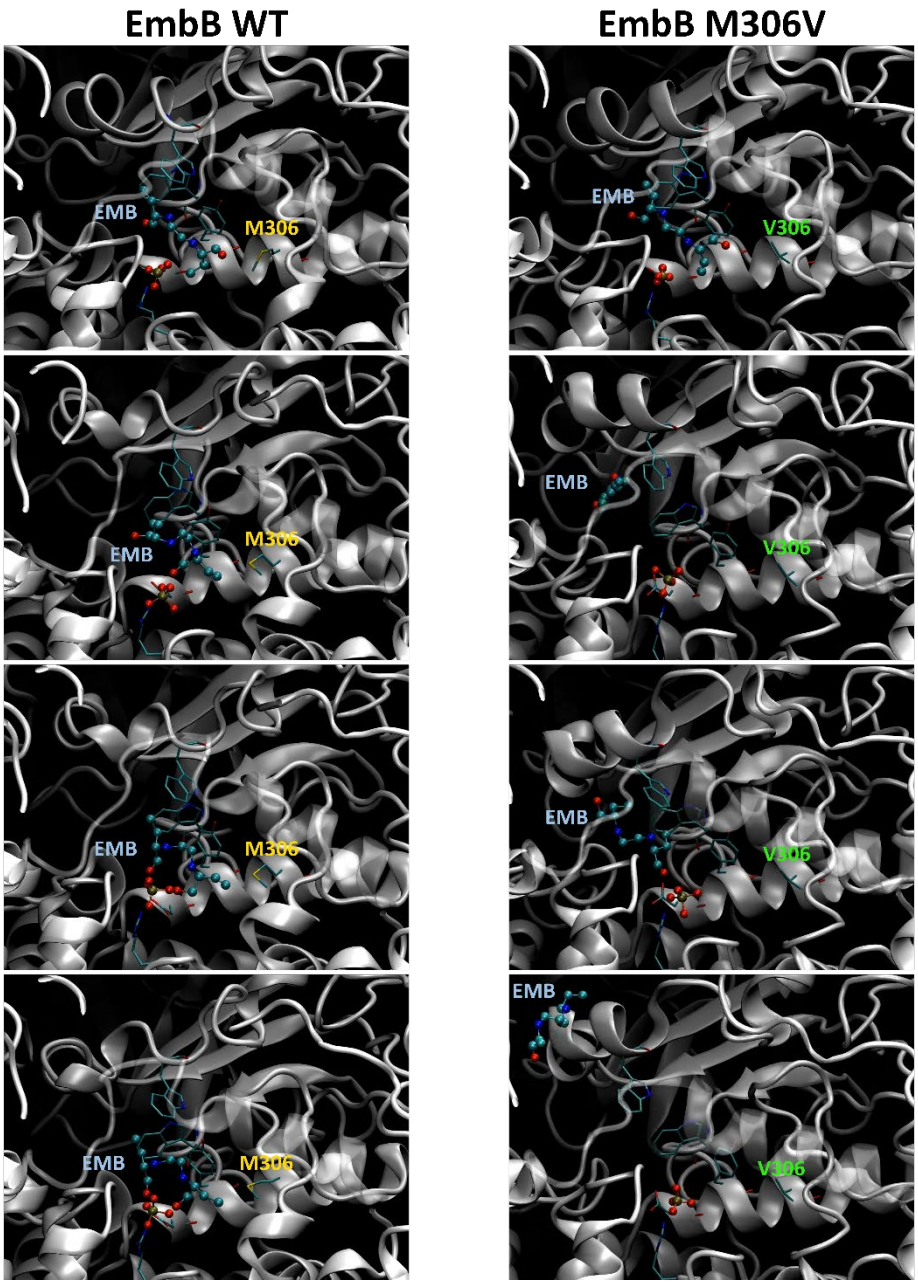


Figure A1 – Comparison of the Molecular Dynamics (MD) simulations of ethambutol (EMB) bound to the catalytic pocket of wild-type (WT) EmbB or mutant EmbB M306V. *In EmbB WT, the drug remains inside the catalytic pocket, while in the mutant EmbB M306V, the drug is displaced towards the outside of the catalytic pocket.*