eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Structural basis of bacterial glycans biosynthesis and processing: impact in human health and disease

Doctoral thesys by Itxaso Anso

Faculty of Science and Technology

Department of Biochemistry and Molecular Biology

University of Basque Country, June 2023

Thesis supervisors:

Marcelo E. Guerin (Director)

Xabier F. Contreras (Co-director)

PUBLICATIONS

This thesis is based on the following articles.

1- Trastoy, B.*, Naegeli, A.*, **Anso, I.***, Sjögren, J. & Guerin, M. E. Structural basis of mammalian mucin processing by the human gut *O*-glycopeptidase OgpA from *Akkermansia muciniphila*. *Nat Commun* **11**, 4844 (2020).

2- **Anso, I.***, Naegeli, A.*, Cifuente, J. O.*, Orrantia, A., Andersson, E., Zenarruzabeitia, O., Moraleda-Montoya, A., García-Alija, M., Corzana, F., Del Orbe, R. A., Borrego, F., Trastoy, B., Sjögren, J., & Guerin, M. E. Turning universal O into rare Bombay type blood. *Nat Commun* **14**, 1765 (2023).

3- **Anso, I***, Basso, L. G. M.*, Wang, L., Marina, A., Páez-Pérez, E. D., Jäger, C., Gavotto, F., Tersa, M., Perrone, S., Contreras, F. X., Prandi, J., Gilleron, M., Linster, C. L., Corzana, F., Lowary, T. L., Trastoy, B., & Guerin, M. E. Molecular ruler mechanism and interfacial catalysis of the integral membrane acyltransferase PatA. *Sci Adv* **7**, eabj4565 (2021).

Other publications:

4- Terrones, O.*, Olazar-Intxausti, J., **Anso, I.**, Lorizate, M., Nieto-Garai, J. A., & Contreras, F. X.. Raman Spectroscopy as a Tool to Study the Pathophysiology of Brain Diseases. *International Journal of Molecular Sciences* vol. 24 (2023).

5- Nieto-Garai, J. A.*, Olazar-Intxausti, J., **Anso, I.**, Lorizate, M., Terrones, O., & Contreras, F. X.. Super-Resolution Microscopy to Study Interorganelle Contact Sites. *International Journal of Molecular Sciences* vol. 23 (2022).

6- Boldrin, F.*, **Anso, I.**, Alebouyeh, S., Sevilla, I. A., Geijo, M., Garrido, J. M., Marina, A., Cioetto Mazzabò, L., Segafreddo, G., Guerin, M. E., Manganelli, R., & Prados-Rosales, R. The Phosphatidyl-*myo*-Inositol Dimannoside Acyltransferase PatA Is Essential for *Mycobacterium tuberculosis* Growth *In Vitro* and *In Vivo*. *J Bacteriol* **203**, e00439-20 (2021).

7- Rodrigo-Unzueta, A.*, Ghirardello, M., Urresti, S., Delso, I., Giganti, D., **Anso, I.**, Trastoy, B., Comino, N., Tersa, M., D'Angelo, C., Cifuente, J. O., Marina, A., Liebau, J., Mäler, L., Chenal, A., Albesa-Jové, D., Merino, P., & Guerin, M. E.. Dissecting the Structural and Chemical Determinants of the 'open-To-Closed' Motion in the Mannosyltransferase PimA from Mycobacteria. *Biochemistry* **59**, (2020).

8- Kalscheuer, R.*, Palacios, A., **Anso, I.**, Cifuente, J., Anguita, J., Jacobs, W. R., Jr, Guerin, M. E., & Prados-Rosales, R. The *Mycobacterium tuberculosis* capsule: A cell structure with key implications in pathogenesis. *Biochemical Journal* **476**, 1995–2016 (2019).

* These authors contributed equally.

COMMON ABBREVIATIONS AND SYMBOLS

ACP: Acyl carrier protein

Ac1PIM2: Monoacylated phosphatidyl-*myo*-inositol dimannoside

Ac2PIM2: Diacylated phosphatidyl-*myo*-inositol dimannoside

ADC: Albumin, dextrose, catalase

AG: Arabinogalactan

CAZYmes: Carbohydrate Active enzymes

CoA: Coenzyme A

Cryo-EM: Cryo-Electron microscopy

C16-CoA: Palmitoyl-CoA

C-terminal: Carboxyl terminal domain

DMPC: 1,2-dimyristoyl-*sn*-glycero-phosphatidylcholine

DMPG: 1,2-dimyristoyl-*sn*-glycero-3-phospho-(1′-rac-glycerol)

DOPC: 1,2-dioleoyl-*sn*-glycero- 3-phosphocholine

DOPG: 1,2-dioleoyl-*sn*-glycero-3-phospho- (1′-rac-glycerol)

DTNB: 5,5′-Dithiobis(2-nitrobenzoic acid)

FACS: Fluorescence-activated cell sorting

Fuc: Fucose

Gal: Galactose

GalNAc: *N*-acetylgalactosamine

GDP-Man: Guanosine diphosphate mannose

GHs: Glycoside hydrolases

GlcNAc: *N*-acetylglucosamine

GT: Glycosyltransferase

GTA: α-1,3-acetylgalactosaminyltransferase encoded by *A* allele in ABO blood system

GTB: *B* allele encoded α-1,3-galactosaminyltransferase in ABO blood system

G6PD: Glucose-6-phosphate dehydrogenase

HMO: Human milk oligosaccharide

IMD: Intracellular membrane domains

ITC: Isothermal titration calorimetry

LAM: Lipoarabinomannan

LC: Liquid chromatography

LC-MS: Liquid chromatography–mass spectrometry

LM: Lipomannan

Man: Mannose

MD: Molecular dynamics

MS: Mass spectrometry

NMR: Nuclear magnetic resonance

N-terminal: Amine terminal domain

OD: Optical density

PG: Peptidoglycan

PI: Phosphatidylinositol

PIMs: Phosphatidyl-*myo*-inositol mannosides

PIM2: Phosphatidyl-*myo*-inositol dimannoside

PIM6: Phosphatidyl-*myo*-inositol hexamannoside

PLs: Polysaccharide lyases

PM: Plasma Membrane

PTS: Proline, threonine and serine rich tandem domain

PUL: Polysaccharide utilization loci

RBC: Red blood cell

r.m.s.d.: root mean square deviation

RT: Room Temperature

Ser: Serine

SPR: Surface plasmon resonance

Sus: Starch utilization system

SUV: Small unilamellar vesicles

TB: tuberculosis

Thr: Threonine

TLC: Thin layer chromatography

VWC: Von Willebrand C domain

VWD: Von Willebrand D domain

6-*O*-MAN: 6-*O*-4-pentynoylmannopyranose

2-*N*-MAN: 1,3,4,6-tetra-*O*-acetyl-2-*N*-4-pentynoylmannopyranose

INDEX

1. INTRODUCTION

1.1. DISEASE AND HEALTH RELATIONSHIP BETWEEN HUMANS AND BACTERIA

In nature, living and non-living systems are all connected and organized in a community known as an ecosystem. The connections in the ecosystems can be analysed at many different levels, such as microorganisms and high animal living interactions. Microorganisms have co-evolved with the host based on different relationships. Some live in symbiotic relationships in which they have reached a mutualistic agreement. Microorganisms contribute to hosting metabolic functions, and the host, in turn, provides nutrients and comfortable living space for the guest. Nevertheless, a beneficial relationship is not always defined due to pathogenic or opportunistic microbial communities willing to unbalance the previous symbiotic community.[1–5]

The relationship between the host and the beneficial or harmful microorganism communities is in constant change, meaning that the mechanisms that maintain the equilibrium need to be very dynamic and constantly checked and updated. In that sense, different events must occur at the same time; on one hand, the host should be in constant recognition of the dynamic mutualistic and harmful organisms, which means that its immune system is able to differentiate between these two microbial populations. At the same time, the microorganism community advantaged from the beneficial relationship will also be recognized as host and detrimental community, and will try to reduce the second group to maintain the equilibrium. Lastly, pathogenic organisms will try to avoid all the mentioned above to to reach their aim.[1–5]

In conclusion, ecosystems are in constant communication with living organisms through many different recognition events which defines the host's health status. Dysbiosis, understood as the imbalance of this ecosystem, will conclude with the detriment of the host's health. Therefore,

further understanding at the molecular level of the recognition, communication and different mechanisms that govern the ecosystem balance is crucial.

In this thesis, I have studied the structure and activity of different carbohydrate recognizing or modifying enzymes that are essential in the beneficial or pathogenic interaction of bacteria and human hosts. In the first part of the thesis, I will focus on a set of enzymes that the mutualistic bacteria present in human gut microbiota, *Akkermansia muciniphila*, encodes to degrade mucins to use them as carbon sources. In the second part of the thesis, I will introduce the biosynthesis of unique glycolipids, phosphatidyl-*myo*-inositol mannosides (PIMs), and their complex derivatives, lipomannan (LM) and lipoarabinomannan (LAM), key structural elements but more importantly, in host interaction during infection of *Mycobacterium tuberculosis*, the causative agent of tuberculosis disease.

1.2. SYMBIOTIC RELATIONSHIP BETWEEN HUMANS AND GUT MICROBIOME

In nature, living organisms co-habitat and share space, which makes them establish connections. In the case of high animals, the body surfaces that are exposed to the environment are colonized by microbiota, mainly mouth, skin, vagina and intestinal tract. This relationship with surrounding microorganisms (bacteria mainly but also by protozoon, fungi, viruses and archaea) is so tight that human can be considered an holobiont.[5,6]

The microbial community that lives associated to us is called the microbiota and the genes they encode are known as our microbiome.[7] High animals and microorganisms have learned that co-living benefits both of them due to the various benefits that both obtain from this mutualistic relationship. Colonization of different microorganisms starts at birth, and it is modulated by different internal and environmental aspects. There is a considerable number and variability of different microorganisms in our body in constant change.[5] This highlights humans' flexibility to establish a new balanced relationship with the changing microorganism population due to internal

and external factors, but also reveals the huge adaptative ability that humans gain. Comensal microbiota encodes a vast set of enzymes that the guest lacks in its genome, providing a new and colossal range of degrading possibilities, very important for successfully digesting the food intake. The host genome is invariable, but the microbiota genome, the microbiome, is dynamic. This means that the host has different capabilities depending on the microorganism equilibrium reached at a certain moment.[5]

The equilibrium with microorganisms is beneficial and crucial in our lives; therefore, the microbiota is equally important in humans' health enhancement and deterioration. Microbiota is an external living organism whose habitat is close to the epithelial cells, which is also a key immunity problem.[5] The imbalance is known as dysbiosis, which occurs when microbiota is somehow affected and alters different systems such as the immune system or metabolic system.[6]

1.2.1. Protection against the external environment

The body surfaces exposed to the environment are protected from organic, inorganic and microbial intruders. Our skin protects the most external surfaces but the inner ones present different layers: (i) one or more layers of active cells (epithelial and Globet cells), (ii) mucus and (iii) glycocalyx (Figure 1).[8–11]

Epithelial cells covers the inside and outside of the body's surfaces and Globet cells are specialized secretory cells that synthesize mucins that are part of the mucus and glycocalyx.[8–12] Mucus covers the epithelial and glycocalyx surface. It provides protection and lubrication for epithelium apart from being the space where the microbiota lives. Therefore, it is crucial to maintein the homeostasis between gut microbiota and the host. Its thickness changes to accommodate the dynamic microbiota, but more importantly, to fulfil the requirements of the epithelium surface. The gastrointestinal tract and, therefore, its epithelium differs from the initial to the final tract depending on its function. The environment and function in the small or large intestine are very

different, and the mucus composition and structure changes accordingly to facilitate the tract's activity.[8–11,13] However, in general terms, mucus is composed mainly of water (95%)[14] and very low frequency but crucial importance components; proteins such as mucins, chloride channel accessory-1 (CLCA1), Fc fragment of IgG binding protein (FCGBP) and zymogen granule protein 16 (ZG16), as well as lipids and ions.[15]

Glycocalyx is the glycan cover that cell membranes, including intestinal enterocytes, show, composed of a highly variable and constantly renewed glycoproteins and glycolipids cover.[16]

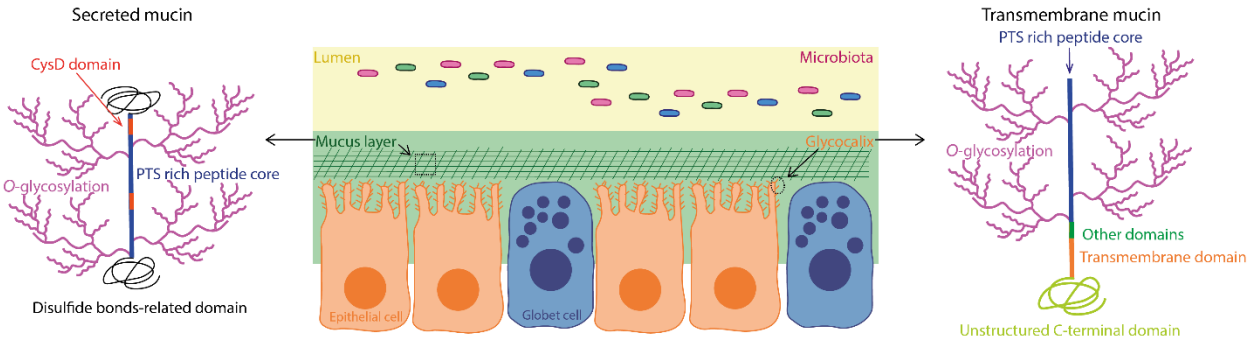

Figure 1: Cartoon representation of the gut surfaces protective layers. Active cells (Globet cells represented in blue and epithelial cells in orange), the mucus layer (in green), and glycocalyx covering the epithelial cells (described as little orange sticks). The gut lumen is represented in light yellow, and the microbiota present in the gut are in different colours spherical figures. Transmembrane mucins in the glycocalyx (left side of the figure) and secreted mucins in the mucus layer (right side) are also shown. PTS (proline, threonine and serine) rich tandem domains are represented in dark blue, CysD domains in dark orange, *O*-glycosylation in purple, disulfide bonds-related domains represent Willebrand D (VWD) domains, Willebrand C (VWC) domain and Cys-knot domain in black, transmembrane domain in light orange, other transmembrane domains such as SEA or NIDO–AMOP–VWD in dark green and unstructured C-terminal domain in light green.

1.2.2. Mucins

The glycoproteins that are the main component of mucus and glycocalyx are called mucins. They are composed of a peptide core highly decorated with a variable amount of glycans (Figure 1). They can be classified as transmembrane mucins when attached to the epithelial cells' surface as part of the glycocalyx or as water-soluble mucins, also called secreted mucins. These glycoproteins were first described in the 80s as the main component of mucus, which is why they were called MUCINs. The first described mucin and its encoding gene were called MUC1/*MUC1*, and since then, the mucins have been named following the description/discovery order.[9]

Nowadays, the Human Genome Organization gene nomenclature committee (HUGO; https://www.genenames.org/) has accepted 21 mucin-encoding genes enclosed in the *MUC* genes family.

Mucins show a peptide backbone. Its most characteristic feature is the high frequency of threonine, serine and proline residues arranged in a variable number of tandem repeats known as PTS domains and some cysteine-rich regions (Figure 1). Threonine and serine residues serve as *O*-glycosylation anchors, as will be discussed below, and proline residues ensure less structured sites where the glycosyltransferases can reach the peptide core at threonine or serine sites to further glycosylate them at the Golgi apparatus. These PTS domains are key to understanding mucins' structural and glycosylation heterogenicity. These repeating tandem domain variability in different mucin types creates different sizes and heterogeneous glycosylation patterned glycoproteins.[9]

Depending on the presence or absence of a transmembrane domain in the peptide core, mucins can be classified as transmembrane or secreted, respectively. At the same time, secreted mucins can be subclassified into monomeric and gel-forming mucins (later or better called polymer-forming mucins, which refers to their structural features to arrange in gel-textured networks) (Figure 1).[9,17,18]

Secreted monomeric mucins comprise a single PTS domain, and only two have been described until now; MUC7 and MUC8.[9,17,18] Polymer-forming or gel-forming mucin's structure, named like that due to the gel-textured mucus they create thanks to its polymerizing ability via disulfide bonds, is more complex. Apart from the typical mucin features (multiple PTS and CysD repeating central domains), these mucins show the following domains where their ability to create the disulfide bonds resides; 3 to 4 von Willebrand D (VWD) domains at the N-terminal domain, and a final Willebrand C (VWC) domain, another VWD domain and a Cys-knot domain at C-terminal.

We can find 5 different mucins expressed in humans that fulfil these features: MUC2, MUC5AC, MUC5B, MUC6 and MUC19 (Figure 1).[9,17,18]

Transmembrane mucins are located at the epithelium surface and are the main component of glycocalyx (Figure 1). They act as sensors to manage microbe-host interactions. They show an N-terminal extracellular domain where PTS repeating tandems are located, a transmembrane domain, and a relatively short unstructured C-terminal cytoplasmic tail. Depending on their complexity, a subclassification can be also performed in this group.[19] The simplest transmembrane mucins are MUC15, MUC21 and MUC22. More structurally complex mucins contain SEA domain (Sea urchin–Enterokinase–Agrin) between the transmembrane and PTS repeating domains (MUC1, MUC16, MUC3A, MUC3B, MUC12, MUC13, and MUC17) (Figure 4). The last subclassification is NIDO–AMOP–VWD mucins, which contain these three domains apart from the transmembrane and PTS repeating domain. The only one known in humans is MUC4.

Finally, there are three other mucins do not fit in this classification; MUC14 also known as EMCN (an endomucin or or mucin-like sialoglycoprotein), MUC9 also known as OVGP1 (oviductal glycoprotein 1) and MUC18, which is no longer considered mucin and it is related to the gene MCAM (melanoma cell adhesion molecule).[9,17,18]

Mucins are highly variable glycosylated proteins. This glycosylation density and variability is specie, tissue and individual dependent. Due to the high frequency of hydroxyl groups containing residues, serine and threonine, the most common glycosylation in mucins is *O*-glycosylation. This glycosylation starts at the Golgi apparatus thanks to the action of peptidyl-GalNAc transferases, which transfer a *N*-acetylgalactosamine (GalNAc) residue to the hydroxyl group of serine or threonine residues. Once the GalNac is anchored in the peptide bond, also known as Tn antigen, different glycosyltransferases highly decorate in two different levels; first branching and then capping with a variety of sugar monomers. Branching of the GalNac anchor is performed by the

addition of galactose (Gal), *N*-acetylglucosamine (GlcNAc) and GalNAc, following eight major distributions known as 1-8 Cores, but Core 1-4 (Core 1; Galβ1-3GalNAcα1-Ser/Thr, Core 2; Galβ1,3(GlcNAcβ1,6)GalNAcα1-Ser/Thr, Core 3; GlcNAcβ1,3GalNAcα1-Ser/Thr and Core 4; GlcNAcβ1,6(GlcNAcβ1,3)GalNAc α1-Ser/Thr) are the most common ones in intestinal mucins (Figure 2).[9,13,20] In humans, Core 1 and Core 2 glycosylation patterns are found in gastric and duodenal mucins, core 3 in the small intestine and core 3 and 4 in colonic mucin glycans.[20]

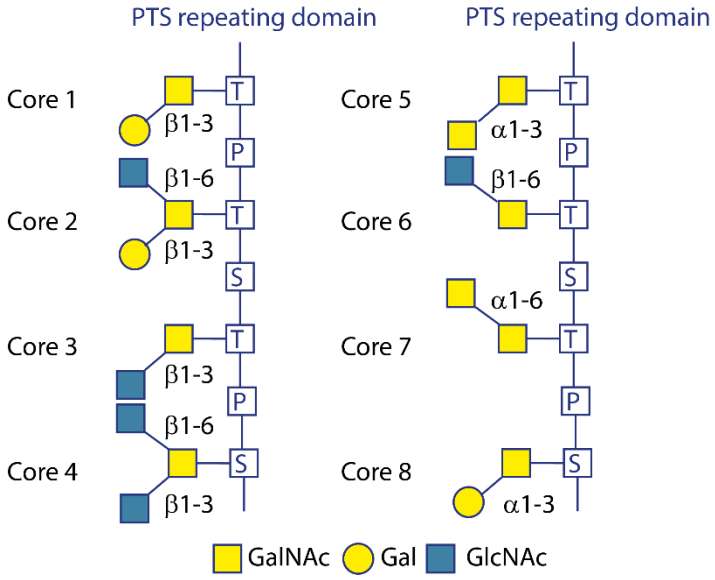

Figure 2: *O*-glycosylation epitopes present in mucins. Cores 1-8 present in mucins *O*-glycosylation schematic representation. Central T, P, and S squares presents threonine, proline and serine peptide core of mucin, the considered *O*-glycosylation starting points. Monosaccharides are represented following the Symbol Nomenclature For Glycans.[21]

These cores can be further brunched with GalNAc, Gal and GlcNAc residues and ended by capping motifs such as fucose (Fucα1-2, α1-3, α1-4), sialic acid [NeuAc(Gc)α2-3α2-6], residues and sulfate groups, which are known as glycan epitopes. The most common ones in gastrointestinal mucins are the histoblood type I and II ABH (blood group antigen A, B, and H) and Lewis antigens (Le^a, Le^b, Le^x and Le^y) (Figure 3).[9,12,20,22,23]

Therefore, we can conclude that mucins are a central peptide core, highly *O*-glycosylated and with acidic residues that cap the branching motifs, which charges the biomolecule negatively (Figure 4).[12,24]

Due to their structural variability, mucins can fulfil many different activity requirements. They are involved in many different functions and therefore, they are ubiquitous. The main functions where these biomolecules are involved are barrier features, dynamicity, hydration, lubrication, and bioactivity.[24]

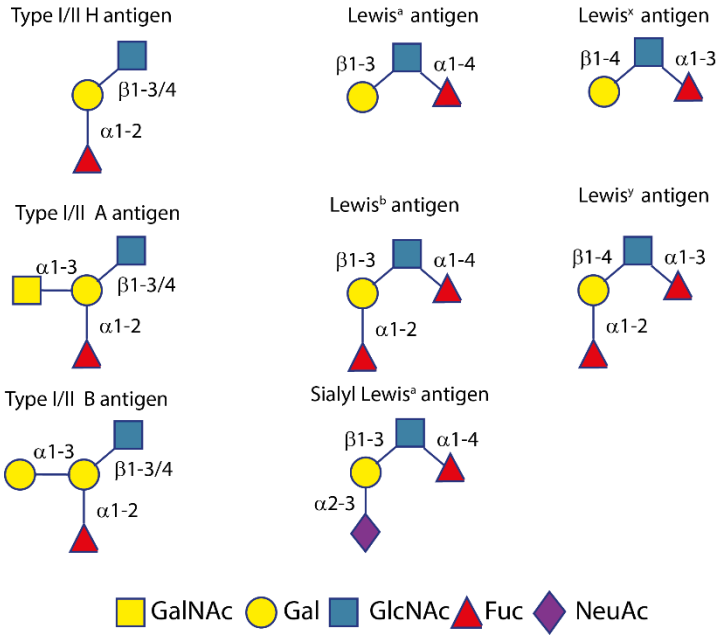

Figure 3: Most common gastrointestinal mucin glycan epitopes. Some external glycan motifs found in mucins belonging to ABH and Lewis histogroups represented following the Symbol Nomenclature For Glycans.[21] Type I and II A, B, H antigens depends on the starting Gal–GlcNAc linkage, Galβ1-3GlcNAc in type I and Galβ1-4GlcNAc in type II.

As discussed, mucins are organized in structured networks (secreted mucins are arranged in gel textured mucus via disulfide bonds, and transmembrane mucins are part of the glycocalyx covering). Therefore, the most obvious function of mucins is being a physical barrier against the different agents the epithelium is in contact with (food, microorganisms, mechaniscal stress from digested food).[24]

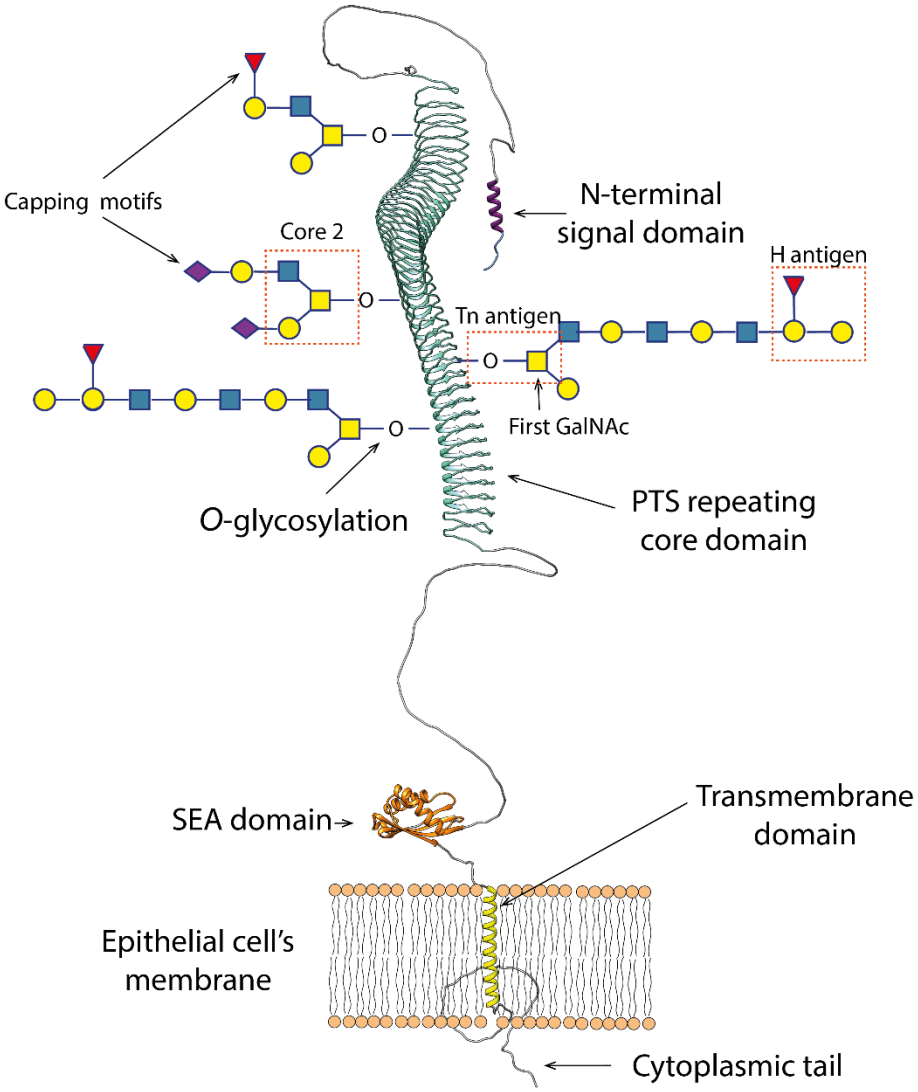

Figure 4: Cartoon representation of transmembrane mucin. Schematic representation of human MUC1 *O*-glycosylated transmembrane mucin. The protein model was modeled from Alphafold database prediction (https://alphafold.ebi.ac.uk/) and *O*-glycosylation patterns were selected as schematic examples from Glygen database (https://www.glygen.org). Glycans are represented following the Symbol Nomenclature For Glycans.[21]

However, the mucus barrier cannot be a compact wall. Gut epithelial's most important function is the digested metabolites' absorption; therefore, the mucins´protective layer should allow an effective but controlled molecules movement. This network shows different size pores thanks to mucin core's different glycosidic composition and density.[24]

Different sized and negatively charged pores prevent even the smallest undesired molecules filtration, acting as an affinity sorting filter and allowing the beneficial digested molecules.[24] The

coating of this physical barrier's architecture that protects our cell's surface is responsible for its huge dynamicity (the barrier can adapt to different functions that the gut shows along the intestinal tract) and avoids the water lost when cells are in contact with the air (hydration and surface's lubrication) due to its huge hygroscopic ability.[24] This barrier also overcomes physical protection and shows a bioactive function. It actively manages many host-immune interactions and distinguishes lipids, proteins, or carbohydrates from food, host microbiota, and external-pathogenic microorganisms, among many others.[12,24] Finally, a very studied function of mucins in the last decade is to host microbiota. Mucins are the substrate for the growth, adhesion and protection of microorganisms living in our gastro-intestinal tract.

1.2.3. Microbiota and immune system

The host develops a highly sophisticated immune system, capable of maintaining in a correct amount the mutualistic species that maintain homeostasis and avoid infections of opportunistic microorganisms. The host immune system needs to be adaptative enough to understand the dynamicity of the gut's environment and recognize the beneficial microorganisms' constant changes, avoid their overgrowing, and trigger immune system activation. This reflects the complexity of immune system building, exceeding the classical external vs. internal detection ability. The host immune system stabilishes two levels of external molecules detection, beneficial vs. harmful microorganisms. The microbiota needs to be involved in its development to build such a specific immune system. Microbiota, in addition to other factors such as host genetics or environmental factors, is involved in the development and maturation of lymphoid structures, modulates immune cell differentiation, tweaks the production of immune mediators such as cytokines and chemokines and influences many host physiological processes.[5]

The loss of the homeostatic relationship, known as dysbiosis, due to a perturbation of the microbiome has many different consequences in the host detrimental to its health.[6] The imbalance

could be understood as a loss of normal and healthy microbiota diversity or an overgrowing of pathogenic bacteria. The reasons why the balance is lost are highly diverse; it can be caused by host-specific factors (genetic background or lifestyle habits) or environmental factors (diet, xenobiotics, or hygiene).[25]

There are many known diseases related to alterations in the gut microbiota. It is believed that this alteration contributes to the development of inflammatory bowel disease[26], rheumatoid arthritis[27], type 1 diabetes mellitus[28], multiple sclerosis[29], cardiovascular events[30], systemic lupus erythematosus[31] and different types of cancer[32–35] among other diseases.[36]

Another relatively well-studied relationship is dysbiosis and obesity,[37,38] which increases the risk of hypercholesterolemia, hypertension, cardiovascular disease, stroke, and many others. Individuals with obesity show a higher *Firmicutes:Bacteroidetes* ratio, resulting in a smaller population of high-capacity carbohydrates processing bacteria and a higher population of bacteria related to alterations in gene promoter methylations,[39] which is linked to this condition.[36]

In conclusion, growing experimental evidence indicates a clear beneficial relationship between health and a balanced gut microbiota, and disease and perturbated gut microbiota.

1.2.4. Gut symbiotic microbiota

The genetics and species diversity in the human gut microbiota is far from being be fully understood but in the last decade two big projects, MetaHit[40,41] and the Human Microbiome Project[42–44], have identified and classified the microbiota species cohabitating in our gut. The improvement in molecular biology methods such as 16S rRNA sequencing and MALDI-TOL-MS technique utilization has facilitated the identification of 2172 species in tested human samples, classified into 12 different phyla, of which 93.5% belonged to Firmicutes (31.1%), Proteobacteria (29.5%), Actinobacteria (25.9%) and Bacteroidetes (7.1%).[42,44] The rest 5.9% of species belongs

to the rare taxa Spirochaetes, Tenericutes, Fusobacteria, Chlamydiaer and Synergistetes. The abundance of some of the microorganisms classified in this last percentage is so low that three phyla are represented by only one species isolated from humans: *Akkermansia muciniphila* from the phylum Verrucomicrobiota in gut microbiota, *Victivallis vadensis* from Lentisphaerae phylum and *Deinococcus aquaticus*, from Deinococcus-Thermus phylum. Scientists found it interesting that gut microbiota seems limited regarding phyla diversity but widely distributed in different species.[45]

Akkermansia muciniphila is a Gram-negative bacteria that belongs to the Verrucomicrobia phylum and was first isolated from the human gut (faeces and colon biopsy samples) in 2004.[46–48] *A. muciniphila* colonizes the mucosal layer of distal regions of the small and large intestines. According to recent studies, *Akkermansia muciniphila*'s abundance is between 3-5% of bacteria present in healthy populations in the colon and more than 1% of the microbiota in faeces. Although its presence in the human gut is a quite recent discovery, it is widely studied due to its relationship with a healthy intestine and its potential use as probiotic to enhance gut homeostasis.[49,50]

The close interaction of the diverse microorganism phyla co-living in the human gastrointestinal tract influences the host's health and disease. DAS, B. & Nair, G. B.[51] already listed that the human microbiota contributes to the synthesis of essential amino acids and vitamins,[52] modulates the efficacy and toxicity of xenobiotics,[53] participates in the maturation of the immune system[54] and lymphocyte homeostasis,[55] helps in tissue and organ development,[56] assists in the pathogens recognition[54] and digests the complex carbohydrates that the host is enabled to process.[57]

Akkermansia muciniphila is part of that 1% of human microbiota able to hydrolase mucins as a carbon source. It has complex genetic machinery able to produce specific enzymes to release from host mucins organic acids, such as acetate and propionate, and simple sugars and peptides.[46,58] Those enzymes capable of processing the complex oligosaccharides in mucins and diet

carbohydrates and convert them into simple monosaccharides are known as Glycoside Hydrolases (GHs)[59] and Polysaccharide Lyases (PLs).[60] GHs, which cleave two monosaccharides' linkages by a water molecule insertion and PLs, which cut the linkage by a β-elimination mechanism, are classified according to their structural and sequence homology and therefore, by their activity and substrate specificity, in different Carbohydrate Active enzymes families (CAZYmes) in CAZY database (http://www.cazy.org/).

The human genome encodes 97 GHs, some of which are directly (8 of them) or potentially (9 of them) related to the digestion of dietary carbohydrates. These dietary carbohydrates come from fruits, vegetables, and cereals, mainly, structurally diverse polysaccharides which require a huge set of different GHs to breakdown and obtain the simple sugars transformed into ATP, the energy source of our cells. It is easy to conclude that the predicted 17 GHs in the human genome are not enough to digest such a diverse amount of sugar linkages, and microbial fermentation in the large intestine is more than needed. For example, to obtain fermentable monosaccharides from type I rhamnogalacturonan, a pectin component takes part of the plant cell wall structure, at least 12 different GHs are activated.[61]

Kaoutari, A *et al.*[57] built a reference human microbiome based on the most representative human gut species (177 in total) belonging to 13 different phyla, and they found that Bacteroidetes phylum genome encodes the highest number of GHs and PLs, even though it is not the most abundant phylum in gut microbiota. It represents only 17% (29 over 177 total selected genomes in the study) of the genomes present in that mini-microbiome (Figure 5a) but encodes almost 4% of the total amount of GHs and shows an average of 137.1 GHs and PLs encoding genes in its genome (Figure 5b).[57]

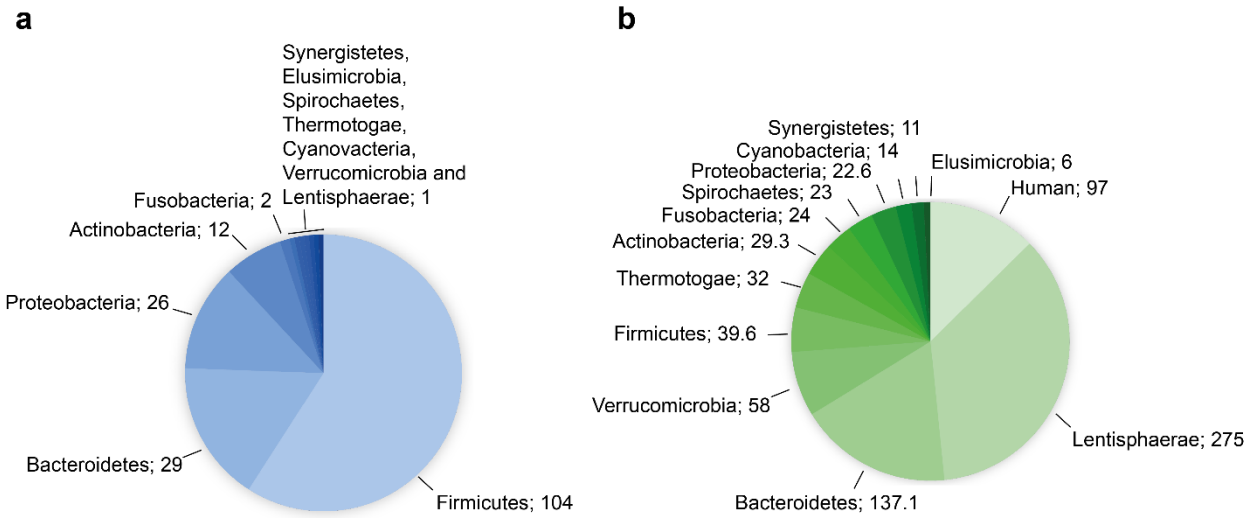

Figure 5: Comparison of abundance and mean GHs and PLs in humans and gut microbiota. **a** Pie chart represents the genomes per phylum present in the designed mini-microbiome. **b** Pie representation of the mean GHs and PLs genes and the total number of genomes per phylum present in the representative mini-microbiome ratio. The total number of GHs and PLs in the human genome is also included in the representation.[57]

One of the most studied and well-known effects of gut microbiota in humans is, in general, carbohydrate digestion, in particular, mucin degradation. 1% of the microorganisms can degrade the oligosaccharides decorating the epithelial cells in different parts of the human gastrointestinal tract. Mucin's most important function is the physical and chemical protection of epithelial cells; therefore, removing such protection could be considered a negative effect on human homeostasis.

That 1% of microbiota, thanks to a suit of different glycoside hydrolases and sulfatases, can use carbohydrates from the diet and present in epithelial walls as carbon sources for them and for the surrounding microorganisms. Therefore, the first symbiotic relationship is established between the mucin-degrading and non-mucin-degrading microbiota.

Complete sugar degrading of complex mucins is not an individual work but a cooperative job where other species are specialized in different sugar residues or/and linkages. It was observed that mixed bacterial colonies were able to degrade mucins complex mucin sugars rather than pure cultures of *Bacteroides fragilis, Bifidobacterium longum* and *Clostridium perfringens.*[62]

Degrading complex mucins to obtain simpler sugars requires different specificity GHs and sulfatases first to remove the capping decorative sugars and sulfate groups. Therefore, the microbiome should encode other families of GHs such as GH33 (sialidases or neuraminidases) and sulfatases to remove external and decorative sialic acid and sulfate groups; GH29 and GH95 fucosidases; GH84, GH85 and GH89 to remove exo- and endo- acting *N*-acetylglucosaminidases; GH101 and GH129 *N*-acetylgalactosaminidases and GH2, GH20, GH35 GH42 and GH98 galactosidases and GH16 endo-acting-*O*-glycanases to cleave large oligosaccharides. Based on that, it's been shown that no microbe encodes all the different types of families needed to completely process the mucins, meaning that the cooperation between different phyla and their symbiosis is vital.[58,63–65]

The most well-known mucin-degrading bacteria species are *Akkermansia* and *Bacteroides. Akkermansia muciniphila* can process 85% of the linkages present in mucins, due to the presence of 9 different encoding GH families in its genome. Other species that encode at least 9 mucin-degrading hydrolases are *B. bifidum and B. tethaiotaomicron*. Many other species in gut microbiota show mucin-degradation-related GHs in their genome, specifically 62% of all microbes and 83% of human gut microbes, meaning that they can partially release some residues from mucins if *Akkermansia* spp. or *Bacteroides spp*. starts the breakdown.[63,64]

The major carbohydrate-degrader gut commensal genome, *Bacteroidetes tethaiotaiomicron*, it's been intensely studied.[66] Bacteroidetes phylum has developed a highly orchestrated polysaccharides degrading encoding machinery called PULs, Polysaccharide Utilization Loci.[67] PULs are colocalized and coregulated clustered genes working together to degrade and integrate the simple sugars from complex dietary or host oligosaccharides. The main goal of the PUL is to orchestrate the detection, attachment, degradation and product transport of specific carbohydrates. The first studies of gene clusters for efficient dietary starch degradation in the 1980s[68,69] created a new paradigm on how this commensal could utilize complex carbohydrates, calling to that first

single gene cluster Starch Utilization System (Sus). Nowadays, as it is known that there are many clusters in *B. tethaiotaomicron* genome specialized in the degradation of structurally diverse oligosaccharides, the name of them has changed to PULs, but "Sus" is maintained to call to the cluster's participants and its architecture serves as a template for new PULs identification.[70–76]

The genes that usually define a PUL are; a sequential pair of SusC (encodes an integral outer membrane maltooligosaccharide transporter homologue also known as TonB-dependent transporters;TBDT) and SusD (encoding a maltooligosaccharide-binding protein homologue, a cell surface glycan-binding proteins; SGBP) homologues which transport the depolymerized sugars into the periplasm; SusE or SusF (substrate binding SGBPs); Carbohydrate modifying enzymes such as GH or PL; and a PUL sensor and transcriptional regulation mechanisms (SusR sensor/relator, extracytoplasmic function sigma ECF-sigma factor -anti-sigma-factor pair or hybrid two-component systems called HTCs) (Figure 6).[70–76]

Bacteroidetes phylum predicted and experimentally characterized PULs can be found at the PULDB Database (Polysaccharide Utilization Loci DB (cazy.org)).[77,78] PULs are defined in Bacteroides phylum as the co-localized and co-regulated SusC/SusD-like transport and binding encoding proteins (TBDT- SGBP) complemented with CAZYmes (e.g., GH, PL, and CE) able to hydrolase complex polysaccharides. Therefore, it seems logical to ask whether this system could also be found in other commensal microorganisms. Such a complex and rhw orchestrated system has not been found yet in other microorganisms. However, Grondin, J. M. *et al.*[76] authors discussed that TBDTs and CAZYmes encoding genes are also found in other Gram-negative bacteria close by each other in the genome, meaning that, even though the complete structure of PULs is not followed, simpler PUL-like systems could be found in other complex carbohydrates-degrading bacteria.

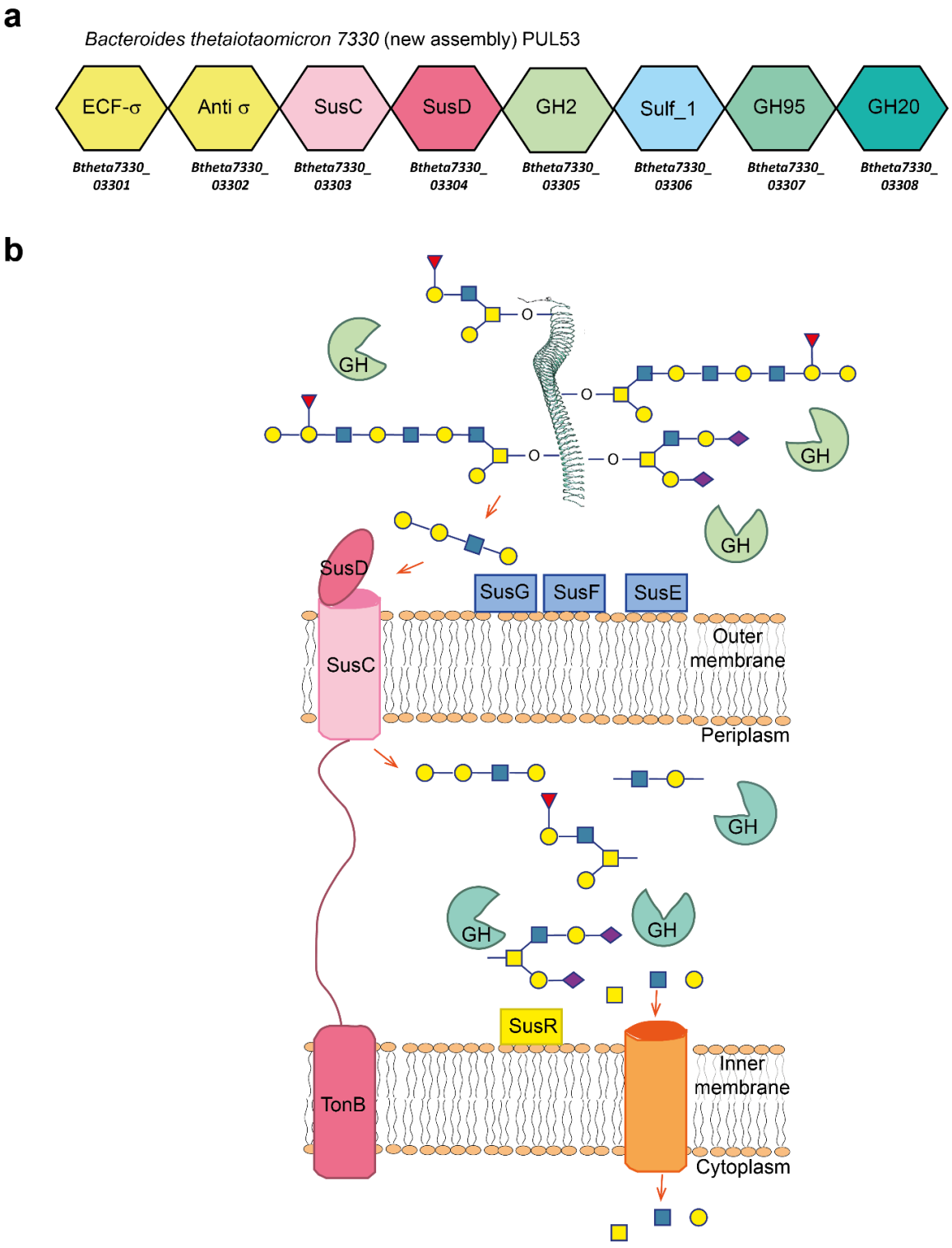

Figure 6: PUL description. **a** Representation of the genes and their encoding proteins' function that participates in the PUL53 in *B. thetaiotaomicron 7330* according to PULDB Database to show PULs organization. **b** Scheme of a generic PUL encoding proteins and mucin complex glycosylation degradation by extracellular and periplasmic GHs (in green), SusD and SusC and TonB transporter which transports the oligosaccharides to the periplasm (pink), SusG, SusF, and SusE (blue) oligosaccharides binding enzymes, SusR (yellow) sensor and transcriptional regulator and a transmembrane transporter (orange) to internalize the degraded monosaccharides to the cytoplasm.

1.2.5. Human red blood cells antigens and blood groups

The glycan epitopes, the specific glycan residues arrangement present in mucins' *O*-glycosylation pattern that we already discussed, are not mucin specific. They can be found in other situations in the body such as human milk, known as Human Milk Oligosaccharides (HMO), or in Red Blood Cells (RBCs) surface since they are usually related to recognition and signalling functions. The RBCs' surface glycan epitopes, known as blood group epitopes, are part of the structural base which defines the ABO Blood Group antigens, the most well-known blood group antigens due to their clinical relevance.

Blood group antigens are assigned to blood group systems based on their relationship to each other as determined by serological or genetic criteria.[79,80] According to the International Society of Blood Transfusion (ISTB; isbtweb.org), blood group systems are one or more antigens governed by a single gene or complex of two or more closely linked homologous genes. There are many blood group antigens based either on oligosaccharide epitopes, including the ABO, P, and Lewis antigens, or specific amino acid sequences, such as Rh, Kell, and Duffy antigens.[81] Currently 44 recognized blood group systems containing 354 red cell antigens. The 44 systems are genetically determined by 49 genes (December 2022) (https://www.isbtweb.org/). The deeply studied ABO antigens, due to their ability to evoke immune antibodies are capable of causing haemolysis in blood transfusions and organ transplantation, were discovered in 1900 by Karl Landsteiner by agglutination tests. The antigens present in this group are composed of specific oligosaccharides linked to lipids (nearly 10%) but primarely to proteins (almost 90%) and they can be classified as A, B or H and subclassified depending on the sugar composition and the variety of linkages (Figure 7a).[82–85]

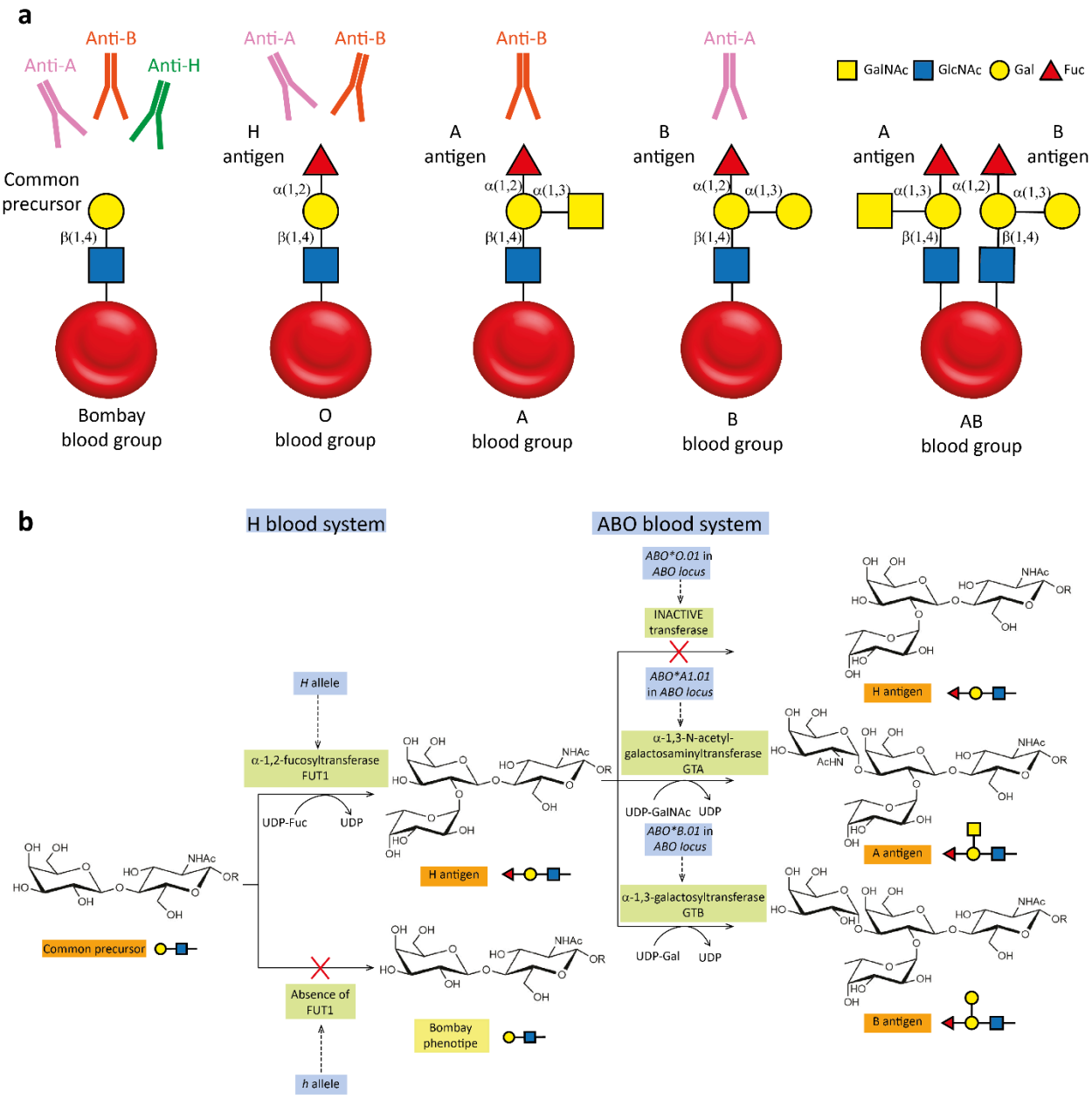

Figure 7: A, B, H and Bombay antigens in RBCs. a Cartoon representation of Bombay, O, A, B and AB blood groups antigens' carbohydrate epitopes present in RBC surface (RBC are depicted as red circles) and antibodies. **b** Enzymatic pathway for the biosynthesis of A, B and H antigens. Enzymes involved in the biosynthesis of A, B, and H antigensare highlighted in light green, encoding genes in blue and antigens' carbohydrate epitopes in orange and yellow. Figure edited from Anso *et al*. (2023).[88]

The A, B, and H antigens are formed by the sequential action of glycosyltransferases encoded by two blood group systems the ABO blood group system and H blood gorup system (Figure 7b).[86]

H blood system is governed by two genes, *FUT1* (composed by "*H" allele*) and *FUT2* (also called Secretor and composed by *"Se"* allele), located at *19q13.33* chromosome that encode two α-1,2-fucosyltransferases, FUT1 and FUT2. These two fucosyltransferases transfer a fucose residue

residue to a common precursor carbohydrate chain obtaining one antigen, H antigen, that can be a terminal antigen defining the O blood group, or the precursor structure for A and B histo-blood antigens in *ABO* blood group system, as it will be explained below. The fucose residue can be transferred to different carbohydrate arrangements on glycolipids and glycoproteins which defines the subclassification of H antigen; type I chains are Galβ(1,3)GlcNAcβ(1,3)-R and defines type I H antigen whereas type II are Galβ(1,4)GlcNAcβ (1,3)-R and defines type II H antigen. Type II H antigens' synthesis is catalysed during erythropoiesis by FUT1 and therefore, are anchored at RBC's (Red Blood Cell) surface. FUT2 transfers the fucose residue to type I chains, soluble molecules in secretions such as saliva (as part of mucins' structure), human milk on free oligosaccharides, or plasma (in glycosphingolipids or as part of lipoproteins). Sometimes, type I H antigen present in plasma may be adsorbed onto the surface of RBCs.[86,87]

A and B antigens are synthesized from H antigen by addition of different residues by specific glycosyltransferases encoded by the *ABO* locus (chromosome 9q34.2) by ABO blood group system.[22,89–91] . Any polymorphism or mutation on that gene defines the activity of the encoded enzyme and therefore alters the ABO phenotype, obtaining the different A, B or the unmodified H antigen. *ABO*A1.01* allele encodes an α-1,3-acetylgalactosaminyltransferase (GTA) adding a GalNAc to a galactose present in H antigen to obtain the A antigen. *ABO*B.01* allele encodes α-1,3-galactosaminyltransferase (GTB), which transfers a Gal to the third position of the galactose present in H antigen to obtain B antigen.[22,91,92] *ABO*O.01* allele encodes an inactive glycosyltransferase unable to modify the precursor. Therefore, depending on our genetic heritage, 4 blood groups can be determined: A blood group if A antigen is present, B blood group if B antigen is present, AB group if both antigens are present (*ABO*A1.01* and *ABO*B.01* alleles could be both co-expressed) or O blood group if H antigen is present (Figure 7b).[93]

These antigens' expression defines the presence or absence of naturally occurring antibodies, mostly IgG. The presence of a specific antigen (or antigens) lacks antibodies directed against it (them) while having antibodies for the antigens they do not express. In other words, the A blood group possesses B antibodies (anti-B), the B group possesses A antibodies (anti-A), the AB blood group possesses no antibodies and the H blood group possesses both of them.[85] These antibodies are responsible for acute intravascular transfusion reactions or acute transplant rejections due to incompatible blood or organ components (Figure 7a).[85]

In the 1950s Bhende *et al.*[94] described for the first time, 3 individuals with a blood group that could not be classified according to ABO blood group criteria in Mumbai, India.[95] Their serum samples showed the presence of Anti-A, Anti-B and Anti-H antibodies capable of agglutinating A, B, AB and O blood groups blood samples. Thanks to that finding, a new and rare blood group was described based on lacking of H, A or B antigens. Whether both, FUT1 and FUT2 enzymes expression is lost, due to mutations in the two homozygous recessive alleles, "*h/h*", of *FUT1* and "*se/se*" of *FUT2* genes, the consequence is a phenotype called Bombay (old Mumbai city name) or O_h phenotype. These individuals show no functional FUT1 and FUT2 enzymes and therefore, lack both types, type I and type II H antigens, together with the possibility of continuing the biosynthesis to A or B antigens. Para-Bombay or $H+^W$ phenotype can be also distinguished when a low amount of H antigen is expressed in two different situations; when the lack of H antigen is caused by inactive *FUT1* gene but *FUT2* gene is active (therefore, type I H antigen is found in secretions) or when a mutation in *FUT1* gene leads to a less active fucosyltransferase in combination with active or inactive *FUT2* gene. Consequently, in Para-Bombay individuals type I H antigen can be found in secretions but also passively adsorb onto the RBCs.[93,96]

Anti-H presence in Bombay and para-Bombay phenotypes, a mostly IgM nature antibody, can cause severe haemolytic transfusion reactions with intravascular haemolysis if the blood is

combined with any other ABO blood group samples. Thus, individuals with this characteristic phenotype should always be transfused with the same blood group samples.[93]

It is considered a rare blood group because it affects less than 1 person per 2,000 of the general population and shows a geographically asymmetric distribution. The lower frequency is found in the European population (1:1,000,000)[93] compared to the prevalence in Iran (1:125,000)[97] or India (1:10,000).[98] Moreover, the Southern and Western regions of India show the most significant number of Bombay phenotype populations, with the highest frequency in Bhuyan tribal population in Orissa, revealing the average prevalence of the Bombay phenotype to be 1 in 27821.[98] Endogamy and consanguinity might be the leading causes of the high prevalence of the rare Bombay blood group in India because they facilitate the homozygous expression of its rare recessive genetic character.[97,99]

1.3. PATHOGENIC MYCOBACTERIAL CELL ENVELOPE

1.3.1. Mycobacterial cell envelope

According to the World Health Organization (WHO; https://www.who.int/), until 2019 when COVID-19 worldwide pandemic occurred and this bacterial agent surpassed this ranking, tuberculosis (TB) was the leading infectious killing disease in the world. A total of 10.6 million people fell ill with this disease and 1.6 million people died from TB in 2021.

Mycobacterium tuberculosis, the causative agent of tuberculosis, is a Gram-negative bacteria belonging to Mycobacteriaceae family of Actinomycetes class. Mycobacteria species have a dynamic and immunomodulatory cell envelope. This bacteria infects and replicates alveolar macrophages. After that, a dynamic nodule of pulmonary tissue, immune cells, and bacteria is created, being a key step for *M. tuberculosis* growth, infection, and transmission. At this granuloma stage the infection may or not develop into necrotic tissue damage, irritation, coughing and aerosol transmission, resulting in two different patients, contagious or noncontagious.[100,101]

Understanding of the mycobacterial cell envelope's structure, composition and function has been a major challenge during the last decades due to its implication in the bacteria infection process, pathogenesis and drug resistance.[100] Mycobacterial cell envelope shows a more complex and unique chemical structure composition compared to other Gram-negative and positive bacterias' envelope, which structurally and functionally varies across the length of the cell and during the infection process.[100] The general Gram-negative and positive bacterias' envelope consists of an inner phospholipid membrane, a thinner or thicker peptidoglycan layer, respectively, and a glycolipid membrane called outer membrane in Gram-negatives.

The mycobacterial cell envelope is more complex and key in pathogenic processes such as host-pathogen interaction.[102] It comprises four main layers; (i) the innermost layer, called plasma membrane (PM). It is the typical phospholipid membrane except for the presence of unique mycobacterial glycolipids, lipoglycans and lipoproteins; (ii) the cell wall core composed of a peptidoglycan (PG) layer covalently attached to arabinogalactan (AG); (iii) the outer membrane, a mycolic acids layer covalently linked to AG decorated with different mycobacterial glycolipids, lipoglycans and lipoproteins; and (iv) the mycobacterial capsule.[100,103–106] This unusual cell envelope classifies mycobacteria as a Gram-positive bacteria, even though it possesse an outer membrane and a periplasmic space (Figure 8).

The PM is a glycerophospholipid bilayer that contains the phosphatidyl-*myo*-inositol mannosides (PIMs) and, in lower proportion, their derivatives lipomannan (LM) and lipoarabinomannan (LAM), cardiolipin (CL), phosphatidylglycerol, phosphatidylethanolamine (PE) and phosphatidylinositol (PI).[107]

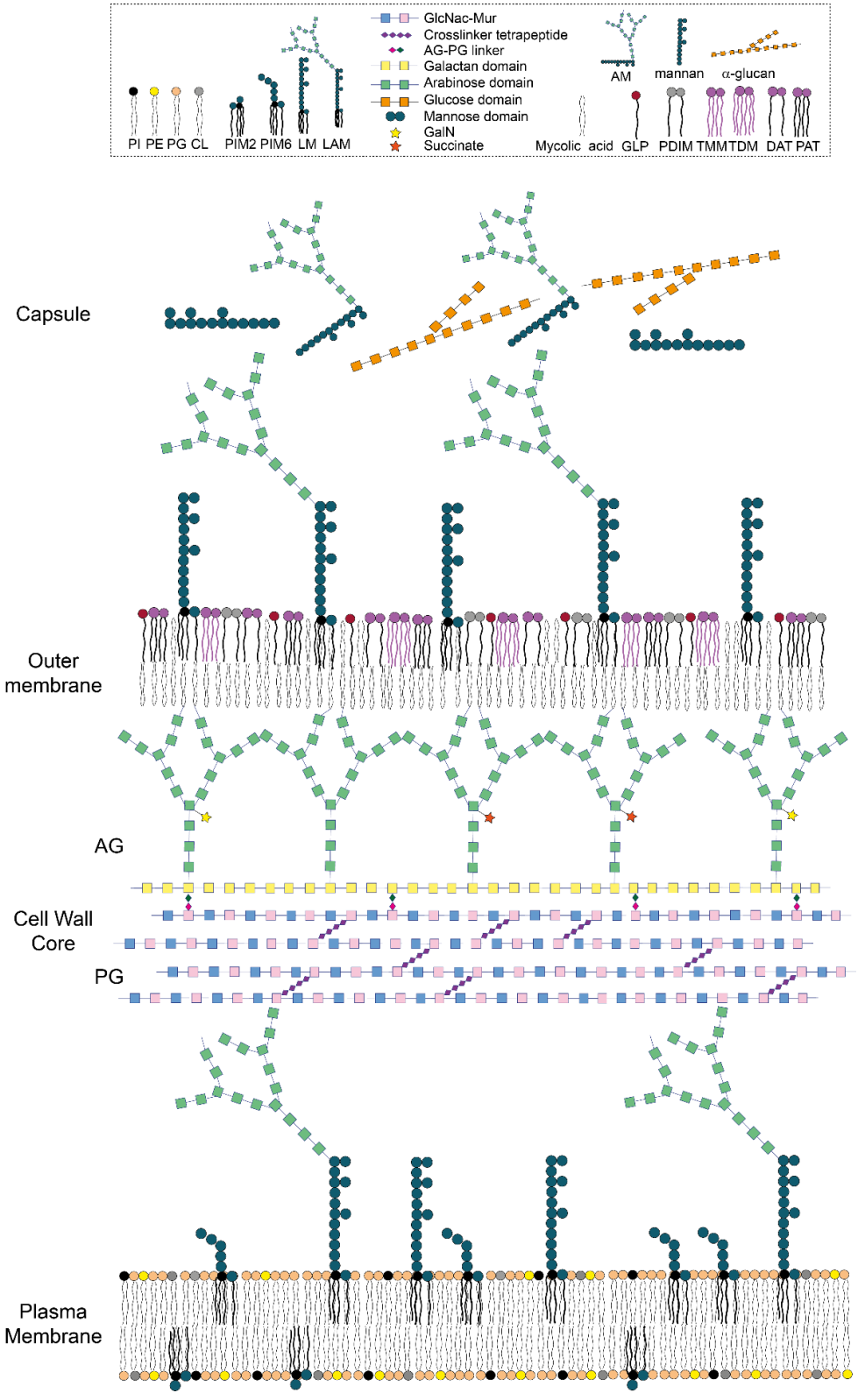

Figure 8: Mycobacterial cell envelope. Cartoon representation of the mycobacterial cell envelope. PM, cell wall core composed of PG and AG domains, outer membrane and capsule are shown. Di- and Hexa-Phosphatidyl-*myo*-inositol mannosides (PIM2 and PIM6), LM and LAM, CL, phosphatidylglycerol (PG), PE and PI, trehalose derivates (DAT, 2,3-diacyltrehaloses; PAT, pentaacyltrehaloses; TDM, trehalose dimycolates; TMM, trehalose monomycolates), phthiocerol dimycocerosates (PDIM), glycopeptidolipids (GPLs) and neutral polysaccharides (α-glucan, arabinomannan (AM) and mannan) are represented.

PM seems to have a dynamic behaviour due to the recently found lipid domains in *M. smegmatis* called Intracellular Membrane Domains (IMD). IMD are areas in the polar regions of the lipid bilayer where the membrane composition varies in response to environmental stresses and growth. They are located where the elongation of the cell takes place and contain a higher amount of cell envelope biosynthesis-involved enzymes.[108–111]

The cell wall core is based on two covalently attached layers. The innermost layer is the complex and rigid PG that maintains the shape of the cell envelope. PG is a glycan backbone that alternates GlcNAc and modified muramic acid (*N*-glycolylmuramic; MurNGlyc or *N*-acetylmuramic; MurNAc) through β-(1,4) linkages. The modified muramic acid units are decorated with the tetrapeptide (L-alanyl-D-glutamine-*meso*-diaminopimelyl-D-alanyl-D-alanine) that can be crosslinked in different positions and linkages with adjacent tetrapeptides.[103,104]

The 10-12% of modified muramic acid residues are attached to a rhamnosyl-GlcNAc disaccharide linker of the outermost layer of the cell wall core, AG, by a phosphodiester link (AG-PG linker).[103] AG is a heteropolysaccharide composed of 1 galactan domain, which contains 23 galactose Gal residues, and two arabinan domains, which contains 26 arabinoses (Ara) residues each of them. These arabinan domains contain an external 6 arabinose motif. Its terminal non-reducing ends serve as the anchoring points for the next mycobacterial envelope layer, the outer membrane mycolic acids, in addition to an internal core composed of arabinose residues with galactosamine and succinyl substituents.[103,104]

The outer membrane or mycomembrane, is a double leaflet composed of an inner α-alkyl-β-hydroxy long-chain (C60–C90) mycolic acids leaflet esterified at the nonreducing ends of AG and an outer leaflet composed of a diverse species such as LM and LAM lipoglycans, PIMs and different derivatives of trehalolipids (trehalose monomycolate;TMM, trehalose dimycolate;TDM, diacyltrehalose;DAT, and pentaacyltrehalose;PAT), phthiocerol dimycocerosates (PDIMs) lipids

and glycopeptidolipids (GPLs).[103,104,107] This asymmetric and highly hydrophobic layer gives unique features to mycobacteria and an intrinsic resistance to antibiotics.[106,112,113]

Lastly, the mycobacterial capsule is a loose, non-covalently attached layer mainly based on three types of neutral polysaccharides (α-glucan, arabinomannan and mannan) structurally identical to polysaccharides present in mycomembrane and proteins and lipids in less abundancy.[107,112–114]

1.3.2. Mannolipids; PIMs, LM and LAM

Phosphatidyl-*myo*-inositol mannosides (PIMs) are unique mycobacterial glycolipids located mostly in the plasma membrane but also found in the innermost layer of the mycomembrane.[103,106,115] These structures are only found in actinomycetes.[116] PIMs and their derivatives LM and LAM thanks to their chemical and structural features are critical for mycobacterial pathogenicity and resistance to antibiotics.

PIMs are based on a PI anchor formed by up to 6 residues of amnnose (Man) which can be acylated in 4 different sites: 6-OH of the Man residue attached to the 6-OH of the inositol ring, 3-OH of inositol ring and two acylations in the PI anchor. Palmitic (C16) and tuberculostearic (10-methyl-octadecanoic, C19) acids are the most common in PIMs structure. Less abundant are myristic (C14) and octadecenoic (C18:1) acids, and almost found in traces stearic (C18), hexadecenoic (C16:1) and heptadecanoic (C17) acids.[103,117–121] Even though the combination of these variables could give many structurally different PIMs, only tri- and tetra-acylated dimmanosides (PIM2) and hexamannosides (PIM6) species are found as the predominant in the plasma and mycomembranes.[115,122]

The biosynthesis of PIMs is under research due to many still unknowns such as non-identified enzymes involved and non-understand translocation and transportation strategies happening (Figure 9-10).

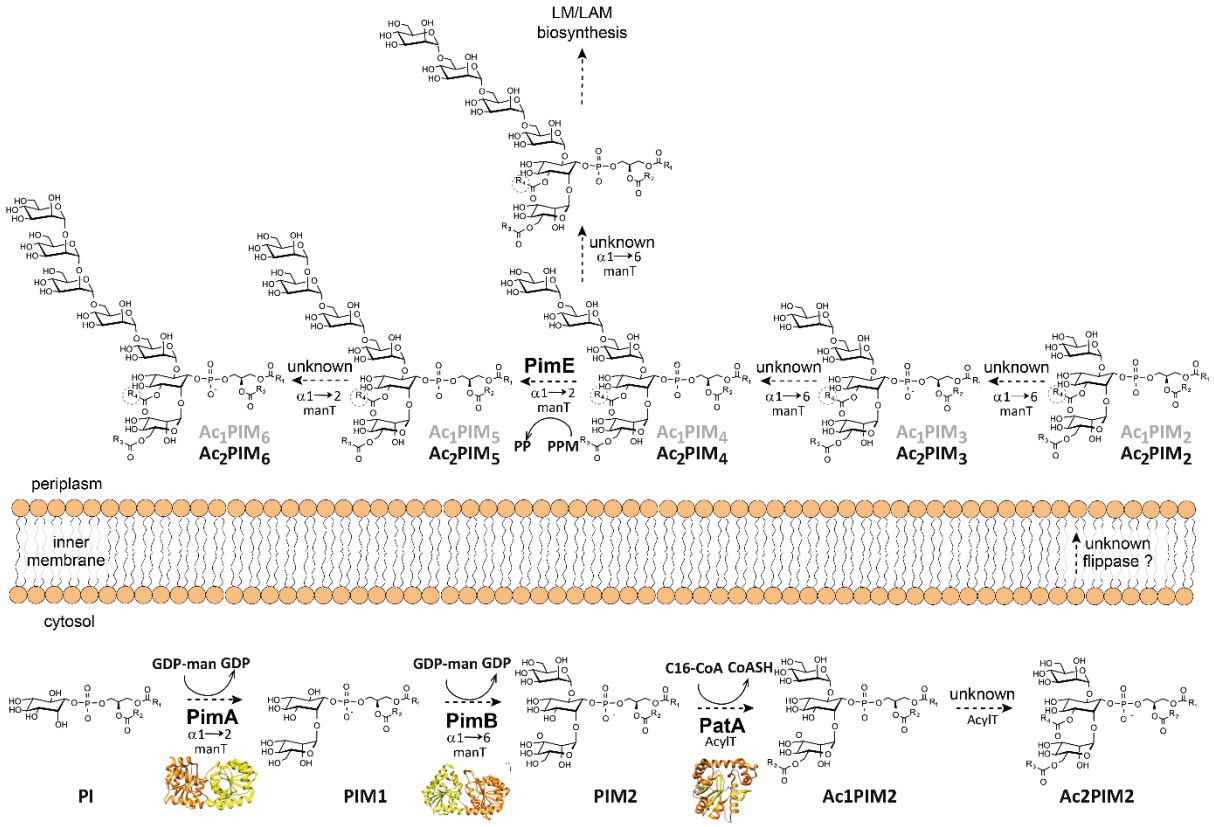

Figure 9: Biosynthetic pathway of PIMs. Schematic representation of mycobacterial phosphatidil-*myo*-inositol mannosides (PIMs) biosynthesis. Early stages occur in the inner leaflet of the inner membrane by the consequential mannose addition to PI by PimA and PimB mannosyltransferases from the water-soluble GDP-mannose. Then, PatA transfers a palmitate group from palmitoyl-CoA to PIM2. After adding two mannose residues, an acylation, and the transport to the outer leaflet of the inner membrane, Ac1PIM4/Ac2PIM4 could be converted into PIM5 by a linkage change to α-1,2 by the mannosyltransferase PimE, further to obtain the end product, Ac1PIM6/Ac2PIM6. PimE transfers a mannose residue from the lipophilic polyprenol-phosphate-Man (PPM). If the linkage is maintained, the addition of mannose residues to PIM4 changes the biosynthetic pathway to obtain the complex derivatives; LM and LAM.

According to the current model, the biosynthesis of PIMs starts in the inner leaflet of the innermost membrane, the plasma membrane, with the mannose transfer from GDP-Man (Guanosine diphosphate mannose) to the second position of PI to obtain monomannosylated PIM1 by the mannosyltransferase PimA (Rv2610c).[122–125] PimA is a retaining glycosyltransferase [126] from GT-4 family.[127] PimA displays a GT-B-fold usually found in this family, which comprises two Rossmann-fold domains with a deep groove at the interface where the catalytic residues are located.[115,127]

Then, a second mannosylation step is catalyzed by a mannosyltrasferase, PimB (Rv2188c), which transfers the second mannose to the 6th position in the inositol ring of PIM1 from GDP-Man to obtain the dimmanosylated PIM, PIM2.[128,129]

PIM1 and PIM2 are acylated by the acyltransferase PatA (Rv2611c) to obtain Ac1PIM1 or Ac1PIM2. PatA transfers the acyl group from palmitoyl-CoA to the 6th position of PimA added mannose residue. [130–132]

Interestingly, the enzymes involved in the biosynthesis of the end product Ac1PIM2, PgsA (involved in the biosynthesis of PI, the first substrate of the PIMs biosynthetic pathway),[133] PimA, PimB, and PatA, are essential for the growth of *M. smegmatis* and/or *M. tuberculosis*.[124,128,133–136] This fact supports the importance of deeply understanding the unknowns of this biosynthesis, due to the valuable therapeutic target they could be for further development of antibiotics or vaccines against this pathogenic mycobacteria. A second acylation step occurs in position 3 of the *myo*-inositol ring, but the identity of the enzyme involved, and the mechanism are still unknown.

By a still non-understand strategy, the mono- and di-acylated PIMs (Ac1PIM2 and Ac2PIM2) are transported to the outermost leaflet of the inner membrane to continue with tri- and tetra-processive mannosylations (PIM3 and PIM4) from the second added mannose residue by a non-identified α-1,6-mannosyltransferases.

A putative mannosyltransferas, PimC, was proposed as the enzyme responsible for PIM2 to PIM3 mannosylation step. However, the lack of phenotype repercussion after silencing *pimC* in the homologue *M. bovis*, and the poor number of orthologues of *pimC* found in clinical isolates suggest redundant gene(s) or an alternative mannosyltransferase pathway.[103,137]

At this step, the biosynthesis can differ in two different biosynthesis pathways to obtain, on one hand, the PIMs end product, mono- or di- acylated PIM6, and on the other hand, the complex lipoglycans LM and LAM (Figure 9-10).

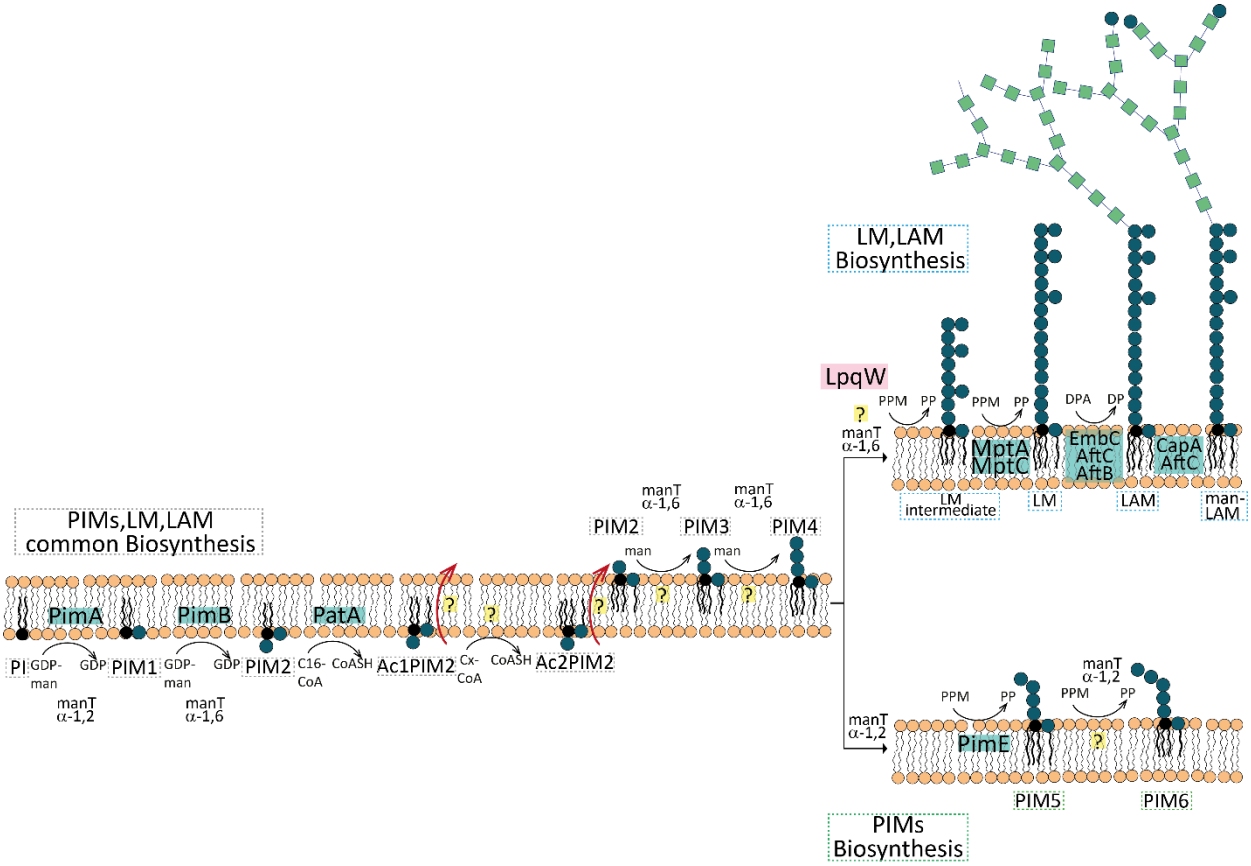

Figure 10: Biosynthetic pathway of PIMs, LM and LAM. Schematic representation of mycobacterial phosphatidil-*myo*-inositol mannosides (PIMs), LM and LAM biosynthesis. Early stages of biosynthesis are shared for PIMs, LM and LAM. Still non-identified enzymes are represented with a yellow question mark, reactions are highlighted with a black arrow and transport events throw the membrane by a red narrow. Enzymes participating in biosynthesis are highlighted in green and the regulatory enzymes are in pink. The transport to the outer leaflet of the inner membrane mechanisms is still unknown. It could be performed before or after the second acylation occurs, because mono- and di- acylated PIMs are found in mycomembranes as final products. Intending to simplify the scheme, after the translocation occurs from the inner to the outer membrane, the acylation degree is not specified in further PIMs. The fifth mannose addition to PIM4 defines the biosynthetic pathway to LM and LAM biosynthesis if the linkage is maintained (α-1,6-man addition) or PIMs biosynthesis if the mannose is added in a new α-1,2 linkage.

The donor substrate is changed to obtain further mannosylated PIMs; Ac2PIM6, LM and LAM. Instead of GDP-man which has a cytosolic localization and serves as Man donnor for GT-A/B superfamily of glycosyltransferases, the next mannosylation steps are dependent on a lipidic Man donor, polyprenol-phosphate-Man (PPM), found in outer and inner membranes This suggests a

different localization occurrence, from the cytosol to the inner or outer membrane, and a change into glycosyltransferases superfamily, to GT-,C responsible for the further reaction.[103,107]

If added mannose linkage changes from α-1,6 to α-1,2, no more than 2 mannose residues are added and the biosynthesis finishes with the end product PIM6. There are still many unknowns in these linkage change mechanisms, but according to the actual model, PimE (Rv1159)[138] seems to be the polyprenol-phosphate-mannose-dependent α-1,2-mannosyltransferase which transfers the fifth mannose to PIM4 (Figure 4).

Mishra, A. K. *et al.*[139] discussed that most of the GT-C from *M. tuberculosis* are already functionally characterized, leading the PIM5 to PIM6 mannosyltransferase candidates to three different options; or PimE is able to perform a second transferase. Rv0051 or Rv0051c,[123,140] still non-characterized GT-Cs could also be performing the reaction. The lack of LM/LAM or PIMs phenotypic consequences of *Rv0051c* silencing in *M. tuberculosis* points to Rv0051 as the best candidate.

Instead of adding the mannose residues in a different linkage, if the mannose residues are added in a processive way following the α-1,6- linkage, the biosynthesis derivates to the lipomannan and lipoarabinomannan lipoglycans. This linkage change seems to be regulated by PimE[138] and LpqW[141] in *M. smegmatis*. PimE drives the biosynthesis towards PIM6 generation changing the linkage to α−1,2, while LpqW, a regulatory lipoprotein, drives the biosynthesis towards the maintenance of the linkage, α-1,6.

To follow with the LM and LAM biosynthetic pathway, after 5-20 mannose residues addition by an unknown mannosyltransferase, MptA (Rv2174)[142], an α-1,6-mannosyltransferase, elongates the chain until 21-34 residues. MptC (Rv2181)[143] α-1,2-mannosyltransferase then decors the α-1,6-mannan backbone with 5-10 residues to obtain LM.

To further synthetise LAM from LM, around 70 arabinose residues domain is attached to the LM backbone, composed of a linear α-1,5 arabinan backbone with α-1,3 arabinan branches and decorated with 4 or 6 arabinose motifs. The arabinosyltransferase which primes the arabinan domain has not been identified yet, but EmbC (Rv3793)[144] is an α-1,5-arabinosyltransferase that elongates the backbone, AftC (Rv2673),[145] the α-1,3-arabinosyltransferase doing the branching and AftB (Rv3508c)[146] the β-1,2-arabinosyltransferase decorating the terminal motifs.[103,104,107] Depending on the species, these terminal motifs are capped with different structures; a variable number of mannose residues in *M. tuberculosis* different strains (Man-LAM) or phosphatidylinositol groups in *M. smegmatis* (PI-LAM), among others.[103,147,148]Two enzymes have been identified as able to mannose cap LAM structures, CapA (Rv1635c) together with MptC (Figure 10).[103,139,149]

As we have seen, PIMs, LM and LAM biosynthesis occurs in many different locations. The early stages happen in the cytosolic face of the inner membrane, then the PIM derivatives are translocated across the plasma membrane to continue the mannosylation and finally, some PIMs end products and their complex derivatives, LM and LAM, are transported and translocated to the outer membrane of the mycomembrane. A transport mechanism should assist such energetically unfavourable movements across the cell envelope.[150,151] Nevertheless, little is known about this transport mechanism.[152] Some putative transporter encoding genes have been proposed, such as LprG, Rv2190c,[139] the ABC-exporter encoded by Rv1747in *M. tuberculosis*,[100,153,154] a proposed TAG transporter P55[155] or other ABC putative Rv1457c-Rv1458c[156] and RND putative mmL13a and mmL13b[157] transporter pairs. Another proposed mechanism by some authors is the export of these glycolipids and lipoglycans by membrane vesicles. Still, further research is needed to clarify these transport events and movements across the mycobacterial cell envelope.

PIMs are critical for the cell wall integrity and fthe host-immune interaction. The presence of PIMs, LM and LAM, together with other mycobacterial envelope components, has been shown to be critical for the mycobacteria pathogenesis. According to the recent model, the mycobacterial infection in macrophages could be due to the several immunomodulatory properties of PIMs, LM and LAM show when they interact with the many immune system receptors. LM and LAM interact with the specific receptor DC-SIGN (dendritic cells specific intercellular adhesion molecule) carbohydrate recognition domain, which recognizes explicitly mannose-rich glycoconjugates, present in host dendritic cells (DC), critical cells in the host immune system that protects against infection, CD1 (including CD1b and CD1d), Toll-like receptor 2 (TLR2), dendritic cell immune activating receptor (DCAR), mannose-binding protein (MBP), lactosylceramide enriched lipid rafts among others.[158–161]

Some authors also suggest that the increase of PIMs is related to a rapid growth stage during macrophage infection and replication. This fact, together with the cytokines induction by PIMs and LMs to help in the granuloma establishment, may promote the progression of the infection.[100]

1.3.3. The mycobacterial palmitoyl-CoA acyltransferase; PatA

PatA is an acyltransferase from *M. tuberculosis*[120,121,135] involved in the biosynthesis of PIMs and their complex derivatives LM and LAM, unique glycolipids and lipoglycans populating the innermost and outermost layers of the complex mycobacterial cell envelope.[115,122,124,128,162] This enzyme is an integral membrane protein essential for the growth of *M. tuberculosis in vitro* and *in vivo.*[136] It transfers a palmitate group from palmitoyl-CoA (C16-CoA) to the 6-position of the mannose ring linked to the 2-position of inositol in mono- or di- mannosylated PIMs to obtain acylated PIM1 or PIM2 (Figure 11). PatA participates in the pathway's early stages, which occurs in the cytosolic face of the innermost membrane thanks to the consecutive action of three enzymes;

two mannosyltransferases, PimA and PimB and the acyltransferase, PatA (Figure 9-10).[124,134,163,164]

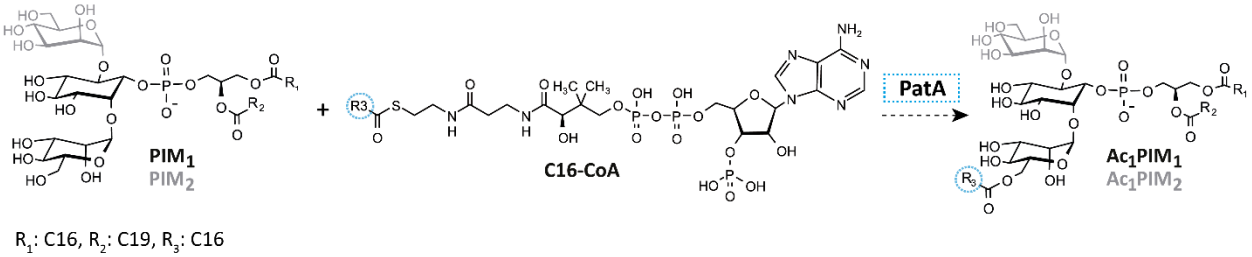

Figure 11: PatA catalysis. Schematic representation of PatA catalyzed reaction. This enzyme transfers a palmitate group from the water-soluble C16-CoA to the 6th position of the OH group in PimA previously transferred mannose residue preferably in PIM2, to obtain Ac1PIM2.

PatA from *M. smegmatis* mc^{2}155 (MSMEG_2934; 304 residues, 74% of sequence identity with PatA from *M. tuberculosis*) crystal structures in the presence of free palmitate, its naturally occurring acyl donor (PatA-C16),[131]a non-hydrolyzable analogue of palmitoyl–CoA (PatA-S-C16CoA),[131] and the 6-*O*-palmitoyl-α-D-mannopyranoside product (PatA-Man-C16)[132] were previously described (Figure 12). This experimental data together with a combination of site-directed mutagenesis activity measurements and docking calculation described at the molecular level the substrate recognition mode and mechanism of action of the enzyme.[131,132]

The central core of this enzyme consists of six stranded β-sheets surrounded by α helices and two grooves can be localized in its surface; (i) an open and long groove that runs parallel to the protein surface and contains the active site, defined as the acceptor binding site and (ii) a second narrow and mostly hydrophobic deep tunnel that goes perpendicular to the first one, the donor´s binding site.[131,132] The 4-phosphopantetheinate moiety of S-C16CoA is located at the entrance of the main groove, in close contact with a highly conserved region flanked by the β2-α8 (residues 149-153), β3-α9 (residues 174-180) and β4-β5 (residues 199-207) loops, and two α-helices, α9 (residues 181-190) and α10 (residues 221-230). The adenosine 3',5'-diphosphate (3',5'-ADP) moiety of the ligand sticks out from the core domain of the enzyme and is exposed to the bulk solvent. In contrast, the Man*p* ring of ManC16 is located within a cavity located at the end of the main groove

and comprising helices α4 and α8 and the connecting loops β1-α7 (residues 83-90), β2-α8 (residues 148-154), and α11-α12 (residues 282-291; Figure 12c). Specifically, the Man*p* ring is mainly stabilized by van der Waals interactions with the side chains of four aromatic residues Tyr83, Trp84, Trp130, and Phe160. In addition, the O6 atom of the Man*p* ring makes a hydrogen bond with the NE2 atom of the imidazole moiety in His126, whereas the O2 atom makes a weak hydrogen bond with the NH2 atom of Arg164 (Figure 12d).[131,132]

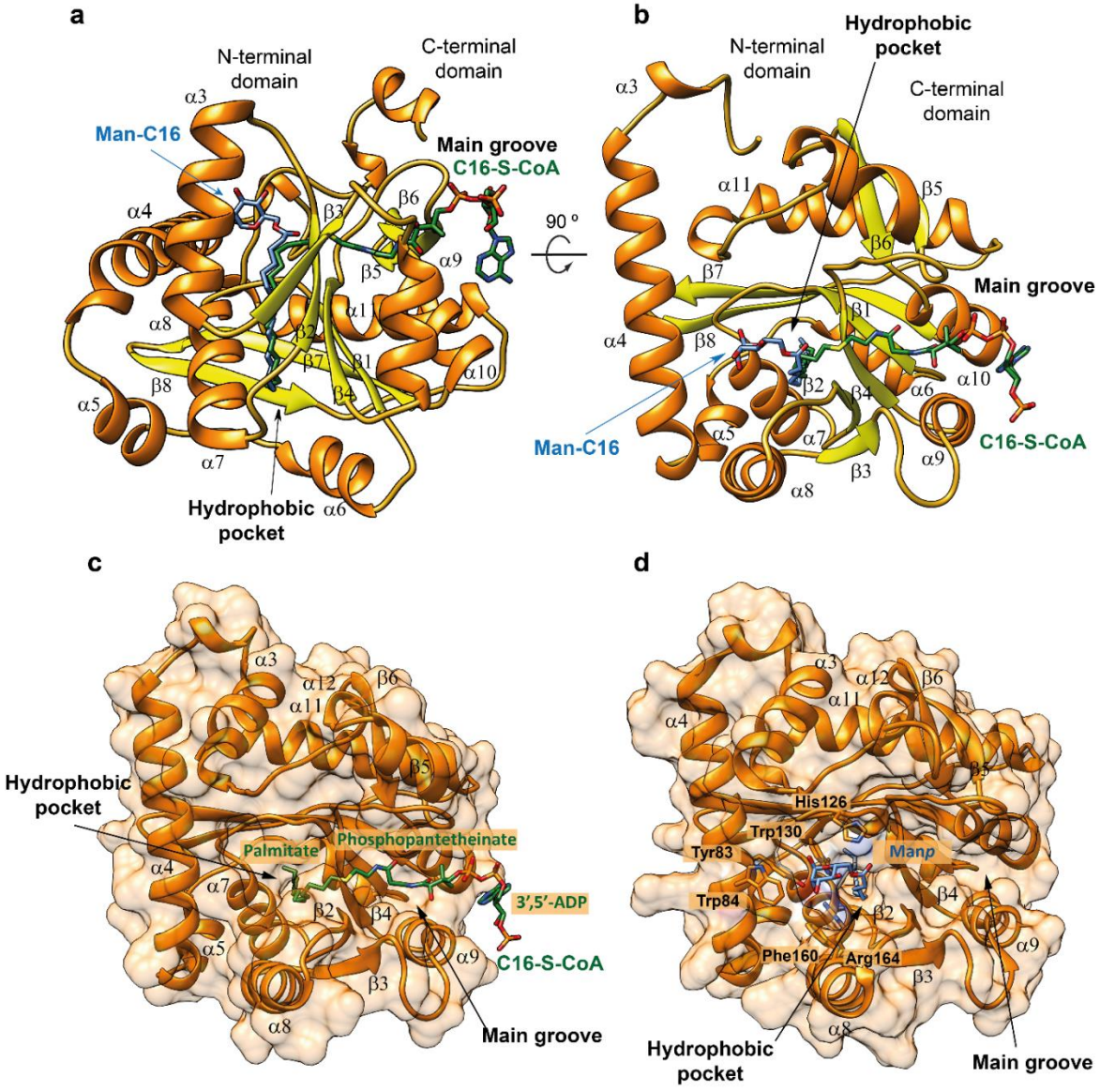

Figure 12: Overall crystal structure of PatA. a and **b** Two different views of the superimposed PatA-S-C16CoA (PDB code: 5F34) and PatA-Man-C16 (PDB code: 5OCE) X-ray crystal structures. Secondary structure β-sheets are highlighted in yellow and α-helices in orange. Manp-C16 end product in blue and S-C16CoA in green. **c** Upper view of PatA-S-C16CoA(PDB code: 5F34) cartoon representation and surface. S-C16CoA is shown in green. **d** Upper view of PatA-Man-C16 (PDB code: 5OCE) cartoon representation and surface. PatA-Man-C16 is shown in blue.

The first 40 residues of PatA do not crystallize, and the analysis of the amino-acid sequence of the enzyme revealed the lack of a signal peptide or hydrophobic transmembrane segments meaning that the membrane interaction remains unknown.

1.3.4. Interfacial catalysis and acyl chain recognition strategies of acyltransferases

Enzymes are proteins acting as catalysts of chemical reactions. Their structure is designed to bind and recognize, with different levels of specificity, the reactive substrates. In general, two distinct regions can be defined in these proteins' topology: The binding site, where the residues that are important to recognize the ligands are, and the active site, where the side chains that are accelerating the chemical reaction are.[165]

The binding site usually consists of a cavity in the protein's surface formed by some specific residues arrangement, where these amino acids could be far away in the protein sequence. However, thanks to the enzyme's conformation, they are brought together when the enzyme folds. The presence of different properties of amino acids in this cavity's surface has the capacity to modulate the environment, enabling and enhancing the substrate's binding. Enzymes that participate in similar chemical reactions and recognize resembling substrates, usually share their binding and/or catalytic mode. This is the criteria that CAZY database (http://www.cazy.org) bases its carbohydrate modifying enzymes' classification, for example.

Acyltransferases, such as the mycobacterial PatA, are enzymes that catalyse the transfer of fatty acid chains from activated fatty acid donors to a chemically diverse acceptor substrates. The fatty acid activation occurs when it is esterified with the thiol group of coenzyme A (CoA) obtaining acyl-CoA derivatives, or with an acyl carrier protein (ACP), knowing them as acyl-ACP derivatives. Several substrate recognition mechanisms depend on the substrate's properties and chemical reaction's requirements.[166] ***The ruler mechanism***, where the length of the binding site acts as a ruler, measuring and determining the length of the molecule to be transferred, is one of

the substrate recognition strategies found in nature. It is shared between many different enzymes to catalyze a determined volume or size of substrates. In the literature we can find several examples describing families of enzymes sharing this binding mode such as glycosyltransferases (GTs),[167] GHs,[168] RNases,[169,170] peptidases,[171–175] integrases,[176] dehydrogenases,[177] decarboxylases[178] or catalases,[179] among others.

Interestingly, this substrate recognition mode is highly shared between the acyltransferases that must distinguish between different length fatty acyl chain substrates. Some examples are the acyltransferase from *Acinetobacter baumannii* (*Ab*LpxM),[180] the acyltransferase (LpxA) from *Escherichia coli* involved in the biosynthesis of Lipid A[181], bacterial pimeloyl-CoA synthetase from *Bacillus subtilis* (BioW),[182,183] PagP, an *E. Coli* pathogenic bacterial outer membrane acyltransferase,[184–189] *Tm*PlsC, 1-acyl-glycerol-3-phosphate acyltransferase from *Thermotoga maritime* involved in membrane phospholipids biosynthesis[190] or the eukaryotic DHHC family of *S*-acyltransferases.[191,192] Actually, in some of the mentioned works,[185,190,191] the authors can modify the enzymes' specificity by modulating the length acyl chain's recognition groove by point mutations, supporting the idea of a hydrocarbon ruler mechanism as the substrates' recognition mode.

Acyltransferases are usually involved in the biosynthesis of phospholipids and glycolipids, which usually deal with different polarity substrates and opposite chemical environments catalysis.[193–195] These enzymes perform the catalysis between hydrophobic lipid donor substrates, which preferentially are located in lipid bilayers, and hydrophilic glycosidic acceptor substrates that reside into organelle lumens and this requires peculiar mechanistic solutions.

3 general mechanistic strategies have been identified[194] to solve the mentioned framework (Figure 13):

(i) Lateral diffusion and hydrophilic funnelling (Figure13a). This occurs when an integral membrane acyltransferase works inside the lipid bilayer and allows the entrance of hydrophilic substrates (water) by lateral diffusion to an internal cavity where the active site is located.

(ii) Desorb-and-modify. The acyltransferase is not inserted in the membrane, but it is associated with different structural motifs such as a N-terminal transmembrane helix. The lipophilic substrate, by desorption, is transferred from the membrane to the active site (Figure 13b).

(iii) Working at the interface. The enzyme's active site is located in the bilayer interface leaving its entrance accessible from both chemical environments (Figure 13c).

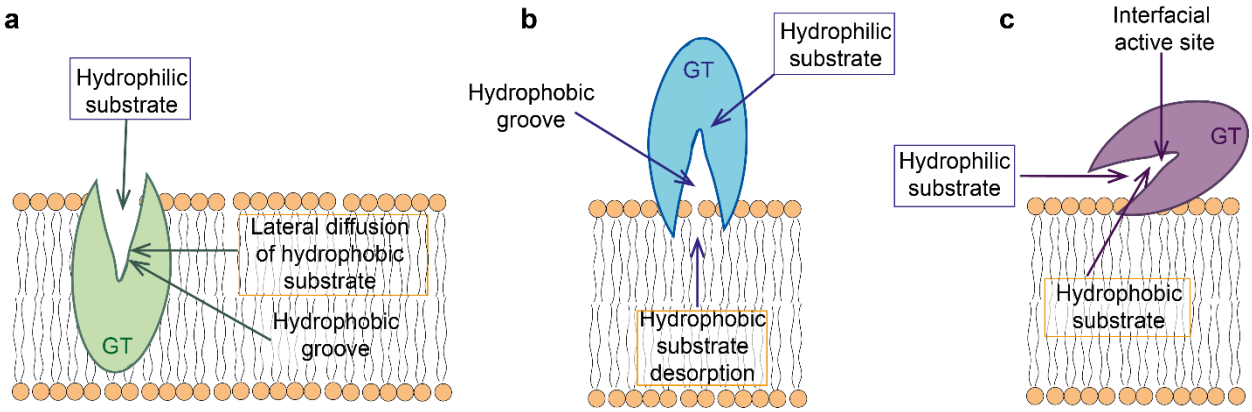

Figure 13: The strategies of acyltransferase catalysis. Schematic representation of the 3 general strategies of acyltransferases (GT) catalysis to solve the difficulties of different chemical nature substrates: **a** Lateral diffusion and hydrophilic funnelling strategy. **b** Desorb-and-modify strategy. **c** Working at the interface strategy.[194]

2. HYPOTHESIS AND OBJECTIVES

The long-term goal of our group is to study, at the molecular level, the structure, substrate recognition event, and mechanism of action of different bacterial carbohydrate recognizing or modifying enzymes that have an important role in the beneficial or pathogenic interaction with the human host.

In that context, during my thesis, I have been working on both beneficial, *Akkermansia muciniphila*, and pathogenic, *M. tuberculosis*, bacteria's enzymes to further understand their catalysis and their substrate specificity at the molecular level of detail. Therefore, this thesis will be defined in two sections: Mucin degradation enzymes machinery of human gut commensal *A. muciniphila* and the biosynthesis of mannolipids present in the complex cell envelope of pathogenic *M. tuberculosis.* Regarding to the first section, our hypothesis is that *A. muciniphila*, as a mucin-degrader human gut commensal, encodes carbohydrates recognizing enzymes in its genome such as *amuc_1119* and *amuc_1120*, nearly located to each other, that could be involved in an orchestrated mucins degradation. The second section's hypothesis is that PatA, an acyltransferase that participates in the biosynthesis of PIMs present in mycobacterial envelope, is an integral membrane protein which is able to differenciate CoA derivates based on a ruler mechanism.

The objectives of this thesis therefore are the followings:

1-Mucin degradation enzymes machinery of human gut commensal *Akkermansia muciniphila:*

- To determine the substrate specificity and mechanism of action of OgpA, *O*-glycopeptidase involved in mucin degradation encoded by *Amuc_1119,* by elucidating its crystal structure in complex with the donor and acceptor substrate, site-directed mutagenesis assays, and hydrolysis experiments.

-Understand the substrate specificity of FucOB, a GH95 family α-1,2-fucosidase potentially involved in mucin degradation encoded by *Amuc_1120*, by X-ray crystallography, site-directed mutagenesis, enzymatic activity and computational mehods

-Show FucOB's ability to convert universal O-type blood group into rare Bombay type by agglutination and flow cytometry-based techniques.

2-Biosynthesis of mannolipids present in the complex cell envelope of pathogenic *M. tuberculosis.*

-To unveil the membrane association mechanism of PatA, an acyltransferase involved in the early stages of PIMs biosynthesis, and the substrate substrate recognition ruler mechanism.

- To proposose an interfacial catalysis model, where PatA can allow the reaction between water-soluble acyl derivatives and membrane-associated PIMs.

- Design and synthesis bifunctional reporters that could be employed in the mycobacterial wall's glycolipids visualization and research.

- Optimization of *M smegmatis*, the non-pathogenic model of *M. tuberculosis*, growing conditions in presence of synthesized bifunctional PIMs analogues.

3. EXPERIMENTAL TECHNIQUES

In this chapter a short overview about the theorical and fundamental aspects of the two most important experimental techniques performed during my PhD development is given. Those techniques are X-ray crystallography for structural determination of bacterial carbohydrate modifying enzymes and bioorthogonal-click chemistry to monitor different living systems' events *in vitro.* Experimental details regarding the performed measurements with these two techniques are given in matherials and methods section of each project.

3.1. PROTEIN X-RAY CRYSTALLOGRAPHY AND STRUCTURE DETERMINATION

Proteins have a defined folding based on their amino acid sequence and the environment during the folding process, and this structure determinates the function of the protein. That is the classical paradigm that structural biologists follow to understand the function and mechanism of action of a particular protein in its biological environment.[196]

The final shape a protein adopts is usually the most energetically favourable. However, proteins do not freeze their structure and can make small movements and show different conformations to perform their function. To understand the structure and dynamicity of proteins, researchers mainly rely on three main structural techniques: X-ray crystallography, Nuclear Magnetic Resonance (NMR) and Cryo-Electron microscopy (Cryo-EM). Each of them shows different strengths and weaknesses, and depending on the scientific question and the specific situation we are trying to solve, single or an integrative approach[197] together with complementary techniques, such as computational techniques (eg., molecular docking, molecular dynamics (MD) or protein prediction software such as Alphafold), molecular biology, biophysical techniques, will improve our comprehension of a certain biological situation.[198]

Historically, X-ray crystallography and NMR have been the most powerful techniques to obtain structural information about a protein conformation. Thanks to the sample flash-freezing and preservation in vitrified ice utilization in EM, together with technological advances in microscopes, detectors, and image processing in the last years, Cryo-EM application to protein structural biology has revealed an increasing amount of near atomic-resolution protein models.[199]

Moreover, in recent years, the technique that has completely changed the structural biologist workflow is DeepMind's machine-learning protein structure prediction program, AlphaFold. Its developers explain that AlphaFold's success in highly accurate protein model prediction is based on incorporating novel neural network architectures and training procedures based on evolutionary, physical, and geometric constraints.[200–202]

In this case, a short overview of Protein X-ray Crystallography and structure determination is given, due to its importance during my PhD development. X-ray Crystallography has been the main structural biology technique in which I have been trained during my doctorate, and the primary technique that I have used to obtain and validate the models I have been studying.

3.1.1. Protein X-ray crystallography

Macromolecular X-ray Crystallography intends to obtain a 3D model of a macromolecule (proteins, nucleic acids, and large biological complexes) from a diffraction pattern obtained when the crystallized macromolecule is exposed to an X-ray beam.[203]

According to the PDB database (https://www.rcsb.org/) 172,859 structures have been solved by X-ray Crystallography to date (February 2023) since the first X-ray crystal structure was solved in the 1950s.[204] As can be seen by the massive amount of solved structures, it is a historically powerful technique that provides information at the atomic scale as well as important physical and

chemical information such as distances and bonding angles between different lateral chains, the conformation of trivial residues such as catalytic or structural relevant ones, among many others.

The workflow of this technique starts with obtaining good quality protein crystals, and exposing them to an X-ray beam later to obtain a diffraction pattern. That diffraction pattern owns the structural information to calculate the electron density map that allows us to build the structural model of the protein (Figure 14).[203]

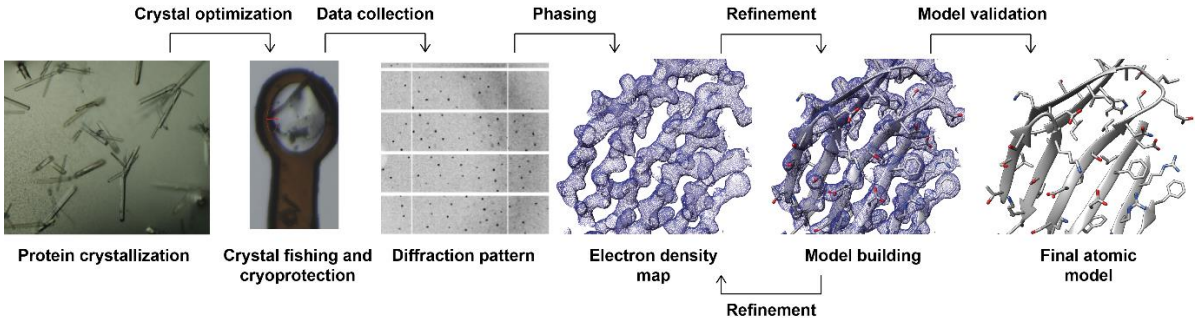

Figure 14: Protein X-ray Crystallization workflow scheme.

3.1.2. Protein crystallization

Protein crystal formation is the first and imperative step of the X-ray Crystallography workflow (Figure 14). Crystals consist of highly ordered microscopic structures due to the arrangement of the atoms in the space following a periodically repeating pattern. The smallest repeating unit with the crystal structure's full symetry is called the unit cell and it is defined by three axes (a,b,c) and three angles (α,β,γ). That ordered structure gives us later in the workflow the structural information as a diffraction pattern to build the model. Crystal obtention is essential for successful protein structure determination and, therefore, is the most significant and usual bottleneck in this technique and the most common challenging aspect.[203]

The physical principle governing this process is split into two steps: the nucleation and the crystal's growth.[205] The first step, nucleation, occurs when the protein shifts from a disordered state creating a partially ordered intermediate, known as the nucleus. Next, small and completely

ordered protein crystals grow and are are obtained from that nucleus. In this second step, the small nucleus extends into the space, maintaining the highly ordered organization.[205]

Three different states can be defined in a solution: two equilibrated (unsaturated and saturated) and a non-equilibrated supersaturated states. The third, non-equilibrated state is the one where the crystallization occurs.[206,207]

The experimental conditions that drive the protein solution to the unsaturated state are very difficult to predict. The correct combination of mild precipitating agents, temperature, ionic strength, pH and environment each protein needs to reach that nucleation step is specific and unique.[205,206,208]

A variety of approaches have been developes to increase the success of finding the appropriate combination of factors to achieve the crystals. Commercial screenings are usually chosen to try, as an initial screening and in a systematic procedure, the largest number of possible experimental combinations with different buffers, pHs, salt concentrations, and adjuvants, based on a trial and error procedure.[203,205] Together with the commercial screenings, there are different methodologies, such as the vapor diffusion method. This method is the most widely used nowadays and the one I used during my PhD. It is based on reaching the equilibrium between two different containers; the mother liquor, where the precipitant mixture is, and the protein, a smaller volume drop in the same precipitant mixture where the highly concentrated protein is dissolved.[205]

3.1.3. X-ray diffraction

X-rays are the electromagnetic waves produced by the emission of the internal electrons in the atom. This radiation is adequate for atomic resolution due to its wavelength, between 0.02-100 Å (around 1 Å polarized X-rays are usually employed in crystallography experiments), which is

similar to the size of atoms and therefore is useful to explore that resolution. Due to their short wavelength, X-rays are highly energetic and have a huge penetration ability.

When X-rays interacts with the electrons of an atom placed in a crystal, that interaction produces an electron-vibration coupling with the variation of the electric field, generating different consequences. After passing through the crystal and interacting with the electrons, part of the X-rays is annulated due to interference between the scattered waves. However, in other incident X-rays, if they interact in a phase of their waves with the electrons of the crystal atoms, a reinforced scattered photons wave is generated. When that in phase scattered waves hit against a detector create the diffraction pattern that we collect to obtain the structural information. This phenomenon known as Bragg's Law was first described mathematically by William Lawrence and William Henry Bragg and was awarded by Nobel Prize in Physics in 1915 (Figure 15).

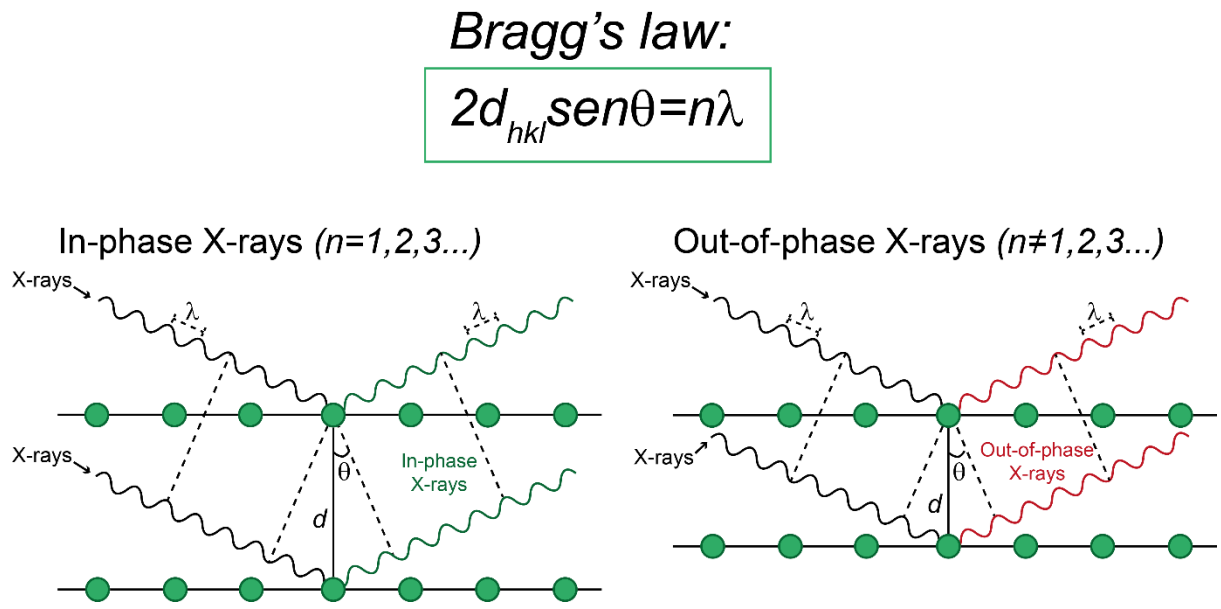

Figure 15: Braggs law scheme. Bragg's law equation is at the top of the figure. The bottom part pannels show incident x-ray wavelength (λ), different atoms' planes distance in the crystal (d_{hkl}), wavelength number of delays *(n)* and atoms arranged in planes represented as green circles. In-phase X-rays interaction representation in the bottom left panel and out-of-phase X-rays interaction in the right panel.

Due to of the three-dimensional periodicity of a crystal, the proteins are ordered perpendicular to certain directions arranged in planes (defined as Miller Indexes; *hkl*). According to Bragg's expression, if an incident X-ray beam with a certain wavelength (λ) hits and scatters in θ angle to

a set of planes and delays a whole number (*n*) regarding a certain distance (*d*) plane beam, the resulting waves will be in-phase and the signal will be reinforced.

To perform the X-ray experiment, a cryopreserved crystal is exposed to a monochromatic X-ray beam produced by an X-ray powder diffractometer at 100 K temperature, and a detector collects the scattering. This action is repeated the needed number of times after a small crystal rotation performed by the goniometer, depending on the protein's symmetry arrangement in the crystal to obtain all atoms' planes scattering, and therefore, the complete structural information, called the data set.

This powerful radiation damages the crystal. To avoid that effect, on one hand, radiation exposure time has been reduced thanks to technical advances. On the other hand, cryogenic temperatures, 100 K, are applied to the X-ray experiment. Due to the environmental humidity, at 100 K, ice crystals which also diffract and can interrupt and hide our crystal's diffraction, are formed. To avoid that ice formation, crystals are cryopreserved. After crystals are fished under the microscope with cryo-loops, they are soaked into a cryo-preservative solution before to swiftly plunging into liquid nitrogen at 77 K. That cryo-preservative solution is usually based on the crystallization buffer with different organic solvents such as glycerol, ethylene glycol or light polyethylene glycols.[209]

3.1.4. Model determination

Once the data is collected, its processing starts to calculate the electron density map, the experimental result where the model of the protein's structure will be built. This process consists of 4 steps: indexing, integration, scaling and structure factor calculation.

After these steps are calculated, to transform the diffraction spectrum (also known as Reciprocal Space) into electron density function (ρ; also known as Real Space), the Fourier transform function must be solved (Equation 1).

$$\rho(x,y,z) = \frac{1}{V}\sum_{h=-\infty}^{+\infty}\sum_{k=-\infty}^{+\infty}\sum_{l=-\infty}^{+\infty}|F_{hkl}|e^{-2\pi i(h_{x+}k_y+l_z-\phi_{hkl})}$$

Equation 1: Fourier transform function. Electron density calculation of each atom of the molecule (ρ(x,y,z)) also known as Real Scpace, is calculated by the Fourier Transform from the diffraction pattern, the Reciprocal Space. V is the crystal's unit cell volume. The two unknowns are the Structural Factor of each Miller's plane ($|F_{hkl}|$) calculated in the previous step, and the phase (ϕ_{hkl}), missing information in the experiment.

In that equation, two unknowns must be solved; the amplitudes associated with the diffracted beam extracted from the diffraction pattern thanks to the structure factor calculation ($|F_{hkl}|$) calculated in the previous step and its phase Φ_{hkl}. The phase is lost due to the physics that governs the experiment. This problem is known as the "phase problem" in crystallography and can be solved by different methods.

The most commonly used technique during this PhD was molecular replacement, where a known structure from a homologous protein is used as the initial model. The high structural identity percentage shared between two proteins is due to their similar folding, so the structural information (its orientation and atom positions) of the known homologue is used to estimate the phase and solve the Fourier transformation of the unknown protein.

Until now, in absence of a homologous crystal structure, more time-consuming and less successful methods were employed, such as Single/Multiple Isomorphous Replacement and Single/Multi-wavelength Anomalous Diffraction. Nowadays, thanks to AlphaFold accuracy, the computationally predicted model can be used as the initial model.

Once the initial electron density map is obtained, a circular refinement process will start, where the protein model will be manually built on the electron density map, and the result is checked

upon different refinement cycles against different quality parameters such as B-factor (atoms' thermal vibration), atoms occupancy in the electron density map, geometrical restrains (atoms angles, distances, dihedrals). Fitting of the model to the experimental electron density map is also evaluated by R-factor measurements, where the difference between the observed and calculated amplitudes is calculated.[210]

When all quality parameter values are within the acceptable range, the model is validated and shared with the scientific community after its submission to the protein data bank (PDB; https://www.rcsb.org/).

3.2. BIOORTHOGONAL-CLICK CHEMISTRY IN CELL BIOLOGY

The curiosity of understanding cellular events in nature has developed a great interest in chemists synthesizing biomolecules that can monitor different living systems' events. Researchers have devised many innovative strategies to follow those physiological processes *in vivo*.[211] However, the design of probes and reactions happening in living systems is not trivial. Apart from their complexity due to the presence of many other biocompounds, the reaction must be performed in physiological conditions (water-based reactions, with normal pressure and temperature) with non-toxic reagents apart from being a highly specific chemical reaction, to avoid interferences with the surrounding biological functions and structures, among many other requirements.

The most recent strategy is the combination of click chemistry and bioorthogonal chemical reporters, unique biomolecules modified with an orthogonal reactivity functional group.[211,212]

Click chemistry is an atom-economy and simple reaction used for joining two molecules. It must occur in one pot, generate minimal and inoffensive side products, preferably occurring in water (therefore non-disturbed by oxygen, water, or other solvents) and be characterized by a high thermodynamic driving force that pushes the reaction quickly and irreversibly.[213]

It was coined in 1998 by K. Barry Sharpless and fully described in 2001 by K. Barry Sharpless, Hartmuth C. Kolb, and M. G. Finn.[213,214]

The classic click chemistry reaction is the Copper(I)-catalysed azide-alkyne cycloaddition (CuAAC), the reaction catalysed by copper between an azide and an alkyne present in a 5-member ring. Two groups independently (Meldal[215] et al. and Sharpless et at.[216]) discovered the Copper(I)-catalysis of the Huisgen 1,3-dipolar cycloaddition.[217] The application of copper catalysis to this cycloaddition gives only the 1,4-isomer and the original 100 ºC is no longer needed during the reaction among other improvements (Figure 16a).[211,212]

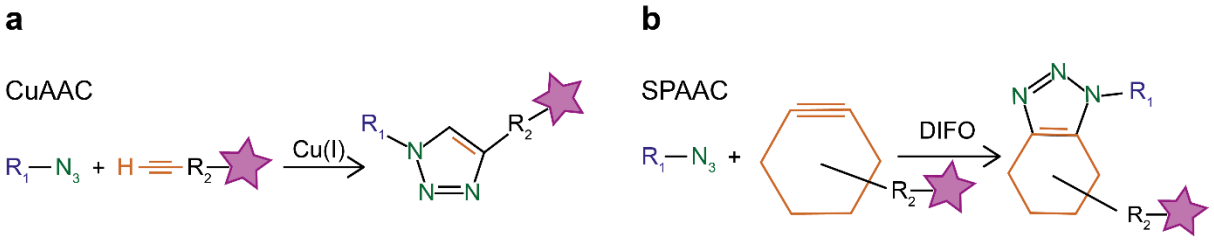

Figure 16: Click chemistry azide-alkyne cycloaddition reactions. In both panels different reaction partners are coloured to help with the reaction understanding and a fluorophore is represented with a purple star. **a** Cooper (I)-catalysed azide-alkyne cycloaddition (CuAAC) and **b** cooper-free bioorthogonal click chemistry known as Strain-promoted azide-alkyne cycloaddition (SPAAC) reaction schemes.

The adaption of this click chemistry to living cells is what is known as Bioorthogonal reactions. It is strictly a chemical reaction occurring inside a living system without interfering with the native environment.[211,212,218–220] In 2003 Bertozzi et al. coined the term. This synthetic strategy for i*n vitro* and *in vivo* application revolutionized the field.

It usually occurs in two steps: First, the bioorthogonal group containing substrate is given to the cells or living organism to be metabolized as the physiological substrate. Secondly, a complementary functional group is introduced to perform the bioorthogonal reaction and introduce the label of interest, such as a fishing target or a fluorophore, to visualize the clicked compound (Figure 17).

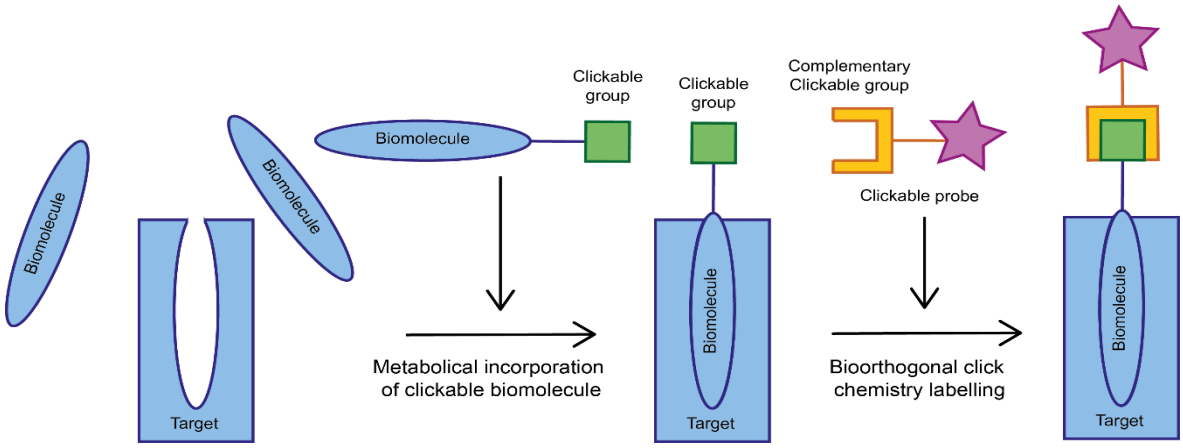

Figure 17: The bioorthogonal reaction strategy scheme. In the first step a biomolecule (in blue) modified with a clickable group (in green) is given to the organism to be metabolized as the natural substrate. Then, the bioorthogonal click reaction is performed in order to covalently tag the metabolized clickable biomolecule with an exogenous clickable probe (in orange and purple).

The characteristics this kind of reaction must fulfil are the same as click chemistry's but focused on biological systems: selectivity (the reaction must occur exclusively with the reaction partners and avoid side reactions with other biological molecules), biological and chemical inertness and reaction biocompatibility and accessible engineering (the bioorthogonal group containing substrate must behave as the natural substrate and avoid disturbing the natural biosynthetic pathways and physiology).[221]

One example of bioorthogonal chemistry is Bertozzi's group improvement into Huisgen's cycloaddition, avoiding the copper catalysis and therefore, the cytotoxicity of the reaction. The reaction is catalysed by a strained difluorooctyne (DIFO) and called strain-promoted azide-alkyne cycloaddition (SPAAC) (Figure 16b).[211,222,223] Since the first description of this reaction, many optimizations and improvements have been discovered, such as new cyclooctyne reagents like DIFO analogues or more water-soluble azacyclooctynes.[224]

This has been a revolution in live events monitorization strategies, and that's why Carolyn R. Bertozzi, Morten P. Meldal and K. Barry Sharpless were awarded Nobel Prize in chemistry in 2022.

4. THE STRUCTURAL BASES OF MUCIN PROCESSING *O*-PEPTIDASE OgpA FROM *A. MUCINIPHILA*

Akkermansia muciniphila is a Gram-negative bacteria present in gut microbiota. It colonizes the mucosal layer of the distal regions of small and large intestine. It was first isolated in 2004 but it has been widely studied due to its relationship with healthy intestines and its potential use as probiotic to enhance gut homeostasis.

More importantly, it is part of 1% of human microbiota able to use mucins as carbon source. It has a huge number of specific glycoside hydrolases, sulfatases, and proteases encoding genome, able to hydrolase 85% of the mucin and obtain simple sugars from highly glycosylated complex mucins.

However, compared to its mucin-degrader colleague Bacteroides phylum complex carbohydrates degradation system, PULs, little is known about the carbohydrates degradation genetic organization and regulation system of *A. muciniphila.*[46] How *A. muciniphila* processes the mucins in gut generating a positive impact on health still remains a challenge.

Careful inspection of *A. muciniphila ATCC BAA-835* strain genome shows some enzymes encoding genes, nearly located to each other, that could be involved into mucins degradation; (i) an encoding predicted sulphatase gene (*Amuc_1118*), (ii) an *O*-glycopeptidase encoding gene, *ogpA* gene, (*Amuc_1119*) (iii) an encoding putative glycoside hydrolase of the GH95 family gene, *fucOB,* (*Amuc_1120*). The GH95 family is composed of enzymes with reported α-L-fucosidase, α-1,2-L-fucosidase and α-L-galactosidase activities. Moreover, these genes are also conserved in other *A. muciniphila sp*., including *A. muciniphila CAG:154, A. muciniphila CAG:344 WG and A. muciniphila KLE1798*, suggesting that these enzymes could collaborate in the specific degradation of mucins (Figure 18).

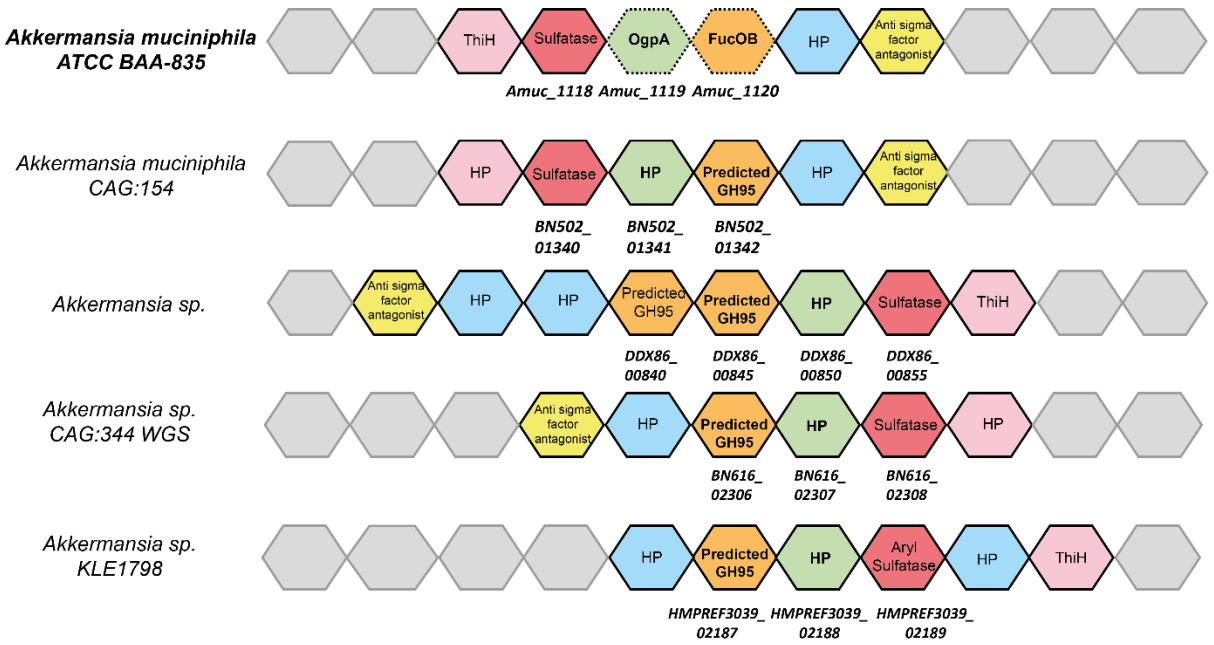

Figure 18: Genomic organization among different *A. muciniphila* strains. Linear distribution of similar functionality genes represented with hexagons. *A. muciniphila ATCC BAA-835* strain is on the top row and same genomic representation is shown for *CAG:154, sp. (DDX86_00850), CAG:344 WGS (BN616_02307)* and *KLE1798 (HMPREF3039_02188)* strains respectively in the following rows according to The National Center for Biotechnology Information (NCBI) Gene and Genome databases (https://www.ncbi.nlm.nih.gov). Similar functionality genes are highlighted in the same colours. The *FucOB(Amuc_1120)* gene and its homologues are highlighted in orange in the middle of the linear genomic representation. HP, hypothetical protein, refers to genes predicted to encode an unknown function protein, represented in blue. In light green the *ogpA* gene (*Amuc_1119*) and its homologues are indicated, sulfatases in red, anti-sigma factor encoding genes in yellow and a dihydroxy-acid dehydratase related to thiamin and biotin synthesis encoding genes in light pink. Grey hexagons represent those non-conserved genes along the different *Akkermansia* species. Figure edited from Anso *et al*. (2023).[88]

In this first part of the thesis, we will focus on the identification, structural analysis and substrate specificity determination of two different enzymes, OgpA (Amuc_1119; chapter 4) and FucOB (Amuc_1120; chapter 5), encoded by two closely related genes in *A. muciniphila's* genome, potentially involved into mucins degradation.

OgpA (encoded by *Amuc_1119*) from *A. muciniphila* is an *O*-glycopeptidase that exclusively hydrolyses the peptide bond N-terminal to serine or threonine residues substituted with an *O*-glycan.[225] This glycosylation pattern is found at the early steps of *O*-glycosylated mucins. Therefore, OgpA could be involved in mucin degradation.

The molecular mechanism by which OgpA specifically recognizes *O*-glycosylation sites remains to be defined. In order to unveil and further understand its substrate specificity and mechanism of

action, we applied *O*-glycopeptide chemistry, enzyme kinetics, computational methods, and X-ray crystallography to obtain high-resolution structures of the unliganded OgpA, and moreover, the complex with the substrate, glycodrosocin and its product.

4.1. MATERIAL AND METHODS

4.1.1. OgpA$_{WT}$ and inactive mutant OgpA$_{H205A/D206A}$ purification

OgpA (OgpA$_{WT}$) from *A. muciniphila*, the inactive mutant OgpA$_{H205A/E206A,}$ and the *O*-glycosylated peptide substrate are distributed as OpeRATOR®, GlycOCATCH® and glycodrosocin ($_{1}$GKPRPYSPRPT(Gal-GalNAc)SHPRPIRV$_{19}$; GD hereafter) from Genovis AB. company, Sweden. OgpA$_{WT}$ and OgpA$_{H205A/D206A}$ were recombinantly expressed in *E. coli*. Lyophilized OgpA$_{WT}$ and OgpA$_{H205A/D206A}$ were dissolved in 20 mM Tris-HCl pH 7.6 and loaded into a Superdex 200 10/300 column (24 mL; GE Healthcare), equilibrated in 20 mM Tris-HCl pH 7.6. The eluted proteins were concentrated at 10 mg mL^{-1} using an Amicon Ultra-4 centrifugal filter unit (Millipore) with a molecular cutoff of 10 kDa at 6,500 × *g*, and aliquots were stored at -80 °C.

4.1.2. Crystallization and data collection of unliganded OgpA$_{WT}$

OgpA$_{WT}$ was crystallized in two crystal forms, referred thereafter as OgpA$_{WT1}$ and OgpA$_{WT2}$. OgpA$_{WT1}$ was crystallized by mixing 0.25 µL of a protein solution at 10 mg mL^{-1} in 20 mM Tris-HCl pH 7.6, with 0.25 µL of 100 mM sodium HEPES/MOPS pH 7.5, 100 mM carboxylic acids mixture (sodium formate, ammonium acetate, sodium citrate tribasic dihydrate, sodium potassium tartrate tetrahydrate and sodium oxamate) and 50% (w/v) of precipitant mix based on 40% (w/v) PEG 500 MME and 20% (w/v) PEG 20,000 (Morpheus® protein crystallization screen). The crystals grew in 1 day. The crystals were transferred to the crystallization solution and frozen under liquid nitrogen.

The second crystal form $OgpA_{WT2}$ was obtained by mixing 0.25 µL of a protein solution at 10 mg mL^{-1} in 20 mM Tris-HCl pH 7.6 with 0.25 µL of 100 mM sodium imidazole/MES monohydrate pH 6.5, 100 mM carboxylic acids mixture (sodium formate, ammonium acetate, sodium citrate tribasic dihydrate, sodium potassium tartrate tetrahydrate and sodium oxamate) and 50% (w/v) of precipitant mix based on 25% (w/v) MPD, 25% (w/v) PEG 1000 and 25% (w/v) PEG 3350 (Morpheus® protein crystallization screen). The crystals grew in 1 day. The crystals were transferred to a cryo-protectant solution containing 10% ethylene glycol and frozen under liquid nitrogen.

Complete X-ray diffraction datasets for both crystal forms were collected at beamline I24 (Diamond Light source, Oxfordshire, UK). $OgpA_{WT1}$ crystal crystallized in the tetragonal space group $P\ 4_1\ 2_1\ 2$ with one molecule in the asymmetric unit and diffracted to a maximum resolution of 1.9 Å (Table 1). $OgpA_{WT2}$, crystallized in the orthorhombic space group $P\ 2_1\ 2_1\ 2_1$ with one molecule in the asymmetric unit and diffracted to a maximum resolution of 1.6 Å (Table 1). Both datasets were integrated and scaled with XDS following standard procedures.[226]

Table 1. $OgpA_{WT1}$ and $OgpA_{WT2}$ data collection and refinement statistics.

	OgpA WT SAD (Zn)	$OgpA_{WT1}$	$OgpA_{WT2}$
PDB code		6Z2D	6Z2O
Beamline	I24 (DLS)	I24 (DLS)	I24 (DLS)
Wavelength (Å)	1.28228	0.968820	1.170200
Resolution range (Å)	29.31-2.2	29.01 - 1.89	29.63 - 1.65
Space group	P 41 21 2	P 41 21 2	P 21 21 21
Unit cell	65.3, 65.3, 189.7, 90, 90, 90	64.9, 64.9, 187.4, 90, 90, 90	65.4, 70.2, 94.7, 90, 90, 90
Total reflections	372650 (7721)	805123 (62853)	670405 (53240)
Unique reflections	15013 (394)	32350 (2963)	52999 (5089)
Multiplicity	24.8 (19.6)	24.9 (21.2)	12.6 (10.5)
Completeness (%)	100 (97.8)	99.28 (93.14)	99.66 (97.03)
Mean I/sigma(I)	20.3 (28.1)	26.95 (3.78)	16.95 (3.29)
Wilson B-factor	20.1	26.11	23.23
R-merge	0.144 (0.094)	0.092 (0.84)	0.087 (0.57)
R-meas	0.15 (0.097)	0.094 (0.86)	0.091 (0.60)
CC1/2	0.99 (0.99)	1 (0.68)	0.99 (0.90)
CC*	0.99 (0.99)	1 (0.90)	0.99 (0.97)
Reflections used in refinement		32346 (2960)	52994 (5089)
Reflections used for R-free		1638 (131)	2729 (242)
R-work		0.18 (0.25)	0.19 (0.22)
R-free		0.22 (0.30)	0.20 (0.30)
CC(work)		0.96 (0.77)	0.96 (0.89)
CC(free)		0.92 (0.74)	0.95 (0.75)
Number of non-hydrogen atoms		2971	3089
macromolecules		2784	2803
ligands		34	29
Protein residues		354	354
RMS(bonds)		0.007	0.006
RMS(angles)		0.84	0.83
Ramachandran favored (%)		98.30	97.71
Ramachandran allowed (%)		1.70	2.29
Ramachandran outliers (%)		0.00	0.00
Rotamer outliers (%)		0.00	0.34
Clashscore		0.90	2.33
Average B-factor		29.46	27.99
macromolecules		29.20	27.30
ligands		36.88	34.73
solvent		32.52	34.75

4.1.3. Crystallization and data collection of inactive $OgpA_{H205A/D206A}$ mutant in complex with its substrate ($OgpA_{H205A/D206A}$-GD-SUB)

The inactive mutant, $OgpA_{H205A/D206A}$, in complex with its substrate, GD, hereafter; $OgpA_{H205A/D206A}$-GD-SUB complex, was crystallized by mixing 0.25 µL of a protein solution of $OgpA_{H205A/D206A}$ at 10 mg mL^{-1} in 20 mM Tris-HCl pH 7.6 and 2.5 mM of GD, with 0.25 µL of 100 mM MIB pH 5.0 and 25% (w/v) PEG 1,500. The crystals grew in 1 day. The crystals were transferred to a cryo-protectant solution containing 25% ethylene glycol and 2.5 mM GD, and frozen under liquid nitrogen.

A complete X-ray diffraction dataset was collected at beamline BL13-XALOC (ALBA, Cerdanyola del Valles, Spain). $OgpA_{H205A/D206A}$-GD-SUB complex crystallized in the tetragonal

space group *I 4* with one molecule in the asymmetric unit and diffracted to a maximum resolution of 2.16 Å (Table 2). The dataset was integrated and scaled with XDS following standard procedures.[226]

Table 2. Inactive mutant $OgpA_{H205A/D206A}$ in complex with the substrate, GD, ($OgpA_{H205A/D206A}$-GD-SUB) data collection and refinement statistics.

	$OgpA_{H205A/H206A}$-GD-SUB*
PDB code	6Z2P
Beamline	BL13-XALOC
Wavelength (Å)	0.979181
Resolution range (Å)	44.87 - 2.16
Space group	I 4
Unit cell	126.9, 126.9, 65.5, 90, 90, 90
Total reflections	111994 (9335)
Unique reflections	27920 (2619)
Multiplicity	4.0 (3.6)
Completeness (%)	99.02 (93.63)
Mean I/sigma(I)	15.40 (1.42)
Wilson B-factor	51.51
R-merge	0.076 (0.77)
R-meas	0.087 (0.90)
CC1/2	0.99 (0.53)
CC*	0.99 (0.83)
Reflections used in refinement	27837 (2617)
Reflections used for R-free	1393 (131)
R-work	0.19 (0.31)
R-free	0.23 (0.35)
CC(work)	0.95 (0.73)
CC(free)	0.95 (0.74)
Number of non-hydrogen atoms	2913
macromolecules	2849
ligands	26
Protein residues	366
RMS(bonds)	0.009
RMS(angles)	0.98
Ramachandran favored (%)	97.24
Ramachandran allowed (%)	2.76
Ramachandran outliers (%)	0.00
Rotamer outliers (%)	0.34
Clashscore	1.97
Average B-factor	54.11
macromolecules	54.18
ligands	53.55
solvent	49.77

4.1.4. Crystallization and data collection of OgpA in complex with the reaction product (OgpA -GD-PRO)

$OgpA_{WT}$ in complex with the reaction product, hereafter; OgpA-GD-PRO, was crystallized by mixing 0.25 µL of a protein solution of $OgpA_{WT}$ at 10 mg mL^{-1} in 20 mM Tris-HCl pH 7.6 and 2.5 mM GD, with 0.25 µL of 100 mM sodium HEPES/MOPS pH 7.5, 100 mM carboxylic acids

mixture (sodium formate, ammonium acetate, sodium citrate tribasic dihydrate, sodium potassium tartrate tetrahydrate and sodium oxamate) and 50% (w/v) of precipitant mix based on 40% (w/v) PEG 500 MME and 20% (w/v) PEG 20,000 (Morpheus® protein crystallization screen). The crystals grew in 1 day. The crystals were transferred to the crystallization solution and frozen under liquid nitrogen.

A complete X-ray diffraction dataset was collected at beamline I03 (Diamond Light source, Oxfordshire, UK). The $OgpA_{WT}$-GD-PRO complex crystallized in the tetragonal space group $P\,4_1\,2_1\,2$ with one molecule in the asymmetric unit and diffracted to a maximum resolution of 2.34 Å (Table 3). The dataset was integrated and scaled with XDS following standard procedures.[226]

Table 3. $OgpA_{WT}$ in complex with the reaction product (OgpA-GD-PRO) data collection and refinement statistics.

	$OgpA_{WT}$-GD-PRO
PDB code	6Z2Q
Beamline	I03 (DLS)
Wavelength (Å)	0.976220
Resolution range (Å)	29.23 - 2.34
Space group	P 41 21 2
Unit cell	65.4, 65.4, 187.7, 90, 90, 90
Total reflections	224733 (16495)
Unique reflections	17627 (1555)
Multiplicity	12.7 (10.6)
Completeness (%)	98.62 (87.13)
Mean I/sigma(I)	19.29 (2.91)
Wilson B-factor	35.98
R-merge	0.133 (0.96)
R-meas	0.139 (1.00)
CC1/2	0.999 (0.8)
CC*	1 (0.94)
Reflections used in refinement	17595 (1530)
Reflections used for R-free	935 (82)
R-work	0.20 (0.26)
R-free	0.25 (0.29)
CC(work)	0.95 (0.81)
CC(free)	0.93 (0.71)
Number of non-hydrogen atoms	2939
macromolecules	2806
ligands	39
Protein residues	357
RMS(bonds)	0.013
RMS(angles)	1.34
Ramachandran favored (%)	97.73
Ramachandran allowed (%)	2.27
Ramachandran outliers (%)	0.00
Rotamer outliers (%)	0.70
Clashscore	4.71
Average B-factor	37.02
macromolecules	37.02
ligands	43.04
solvent	34.32

4.1.5. OgpA$_{WT}$-SAD structure determination

The crystal structure of full-length OgpA in its unliganded form was solved using zinc single-wavelength anomalous dispersion (Zn-SAD).[227] Anomalous data of an OgpA$_{WT2}$ crystal were collected at the theoretical 9.6586 keV absorption edge of Zn (9.669 keV eV–1.28228 Å; OgpA$_{WT}$-SAD). The data were integrated and scaled with XDS.[226] The structure determination located two Zn atoms in the asymmetric unit (CC= 12.6, CC(weak)= 7.6 and CFOM= 20.2). Experimental phases were determined using the Big EP pipeline.[228] The partial model produced with CRANK2 was subsequently used for initial cycles of model building, density modifications and refinement by Buccaneer and the CCP4 suite.[229–231]

4.1.6. OgpA$_{WT1}$, OgpA$_{WT2}$, OgpA$_{H205A/D206A}$-GD-SUB and OgpA$_{WT}$-GD-PRO structure determination and refinement

The structure determination of OgpA$_{WT1}$, OgpA$_{WT2}$, OgpA$_{H205A/D206A}$-GD-SUB and OgpA$_{WT}$-GD-PRO were carried out by molecular replacement using Phaser and the PHENIX suite and the OgpA$_{WT}$-SAD structure as initial model.[232,233] Followed by cycles of manual rebuilding and refinement using Coot and phenix.refine, respectively.[234,235] The structures were validated by MolProbity.[236] Data collection and refinement statistics are presented in Table 1 (OgpA$_{WT1}$, OgpA$_{WT2}$ models), Table 2 (OgpA$_{H205A/D206A}$-GD-SUB model) and Table 3 (OgpA$_{WT}$-GD-PRO model). The atomic coordinates and structure factors of OgpA$_{WT1}$, OgpA$_{WT2}$, OgpA$_{H205A/D206A}$-GD-SUB and OgpA$_{WT}$-GD-PRO have been deposited with the Protein Data Bank, accession codes 6Z2D, 6Z2O, 6Z2P, and 6Z2Q, respectively. Molecular graphics and structural analyses were performed with the UCSF Chimera package.[237]

4.1.7. Structural analysis and sequence alignment

Structure-based sequence alignment analysis were performed using Chimera. Protein pocket volume was calculated using HOLLOW.[238] Z-score values were produced by using DALI.[239] Domain interface analysis was performed using PISA.[240] Conserved and similar residues were labelled using BoxShade server (http://embnet.vital-it.ch/software/BOX_form.html).

4.1.8. Synthesis of *O*-glycosylated peptides

The glycopeptides later used to measure the hydrolytic activity of OgpA by reverse-phase HPLC analysis (Tn, 3SC1, 6SC1, 3S6SC1, C2 and C3) were obtained as mentioned below. The 3SC1 glycopeptide was purchased from Sussex Research. To generate peptides C1, Tn and the unmodified control peptide, respectively, the 3SC1 glycopeptide at 400 µM in 25 µL of 50mM Tris-HCl pH 7.0 was treated with (i) 20 units of α-2,3-sialidase (SialEXO23; Genovis AB, Lund), (ii) 20 units of α-2,3-sialidase (SialEXO23; Genovis AB, Lund) and 20 units of β-galactosidase (GalactEXO; Genovis AB, Lund) or (iii) 20 units of α–2,3-sialidase (SialEXO23) and 20 units of *O*-glycosidase (OglyZOR; Genovis AB, Lund) at 37°C overnight. The C2 glycopeptide was synthesized by incubating the glycopeptide C1 at 37°C for 2 hs, under the following conditions: 400 µM glycopeptide, 10 mM UDP-GlcNAc and 2 µg of recombinant human β-1,3-galactosyl-*O*-glycosyl-glycoprotein β-1,6-*N*-acetylglucosaminyltransferase (GCNT1; R&D Systems) in 50 µL 100 mM MES, pH 6.0, supplemented with 5 mM $CaCl_2$ and 10 mM DTT. For the synthesis of glycopeptide C3, glycopeptide Tn (400 µM) was incubated at 37°C for 2 h, with 10 mM UDP-GlcNAc and 2 µg recombinant human acetylgalactosaminyl-*O*-glycosyl-glycoprotein β-1,3-*N*-acetylglucosaminyltransferase (B3GNT6; R&D Systems) in 50 µL 25 mM Tris-HCl, pH 7.5, 150 mM NaCl, 10 mM $MnCl_2$, 5 mM $CaCl_2$ and 20% DMSO. Glycopeptides 6SC1 and the di-sialylated 3S6SC1 were synthesized by incubating 400 µM C1 and 3SC1, respectively, with 2 µg

α-*N*-acetylgalactosaminyl α-2,6-sialyltransferase 2 (ST6GALNAC2; R&D Systems) and 10 mM CMP-Neu5Ac in 50 µL 25 mM Tris-HCl, pH 7.5, 10 mM $MnCl_2$.

The enzymes were removed from the reactions by passing the reaction mixture through a 10 kDa MWCO filter unit (PALL NanoSep). All glycopeptides were cleaned-up using Pierce C18 tips (Thermo Scientific), evaporated to dryness in a vacuum centrifuge and dissolved in 20 mM Tris-HCl, pH 7.0.

4.1.9. OgpA activity assays

The glycopeptides were incubated with the indicated amounts of OgpA and the reaction was stopped at different time points by diluting the reaction mixture in 0.1 % FA. Turnover was analysed by reverse phase HPLC on a Vanquish Duo UHPLC system equipped with an MSPac DS-10 desalting cartridge (both Thermo Fisher). The glycopeptides were separated by a 5 min. gradient from 1.8% to 12% ACN in 5 mM sodium phosphate pH 8.0 at 50°C and detected by fluorescence of the 5-FAM fluorophore (excitation :494 nm, emission: 521 nm). Quantification of product and substrate peaks allows for determination of product formation according to the following formula (equation 2):

$$n_{product} = \frac{peak\ area_{product}}{peak\ area_{educt} + peak\ area_{product}} \times n_{total}$$

Equation 2: Expression for quantification of product formation

$n_{product}$ refers to the amount of product formed (i.e., digested peptide), $peak\ area_{product/substrate}$ to the measured area under the curve for the product and substrate respectively and n_{total} is the total amount of peptide in the reaction. From this data, the initial turnover rate of each reaction could be determined by linear regression.

4.1.10. Molecular docking calculations

The Tn, 3SC1, 6SC1, 3S6SC1, C2 and C3 *O*-glycans were modelled using GLYCAM-Web website (Complex Carbohydrate Research Center, University of Georgia, Athens, GA; http://www.glycam.com). Ligand docking was performed using AutoDock Vina employing standard parameters and visualized using USCF Chimera.[237,241]

4.2. RESULTS AND DISCUSSION

4.2.1. The overall structure of full-length OgpA

OgpA from *A. muchiniphila* comprises 385 residues ($OgpA_{WT}$ hereafter; B2UR60, UniProt code) with a predicted signal peptide (residues 1–24) that was removed from the construct for crystallization purposes. The crystal structure of full-length OgpA in its unliganded form was solved using zinc single-wavelength anomalous dispersion (Zn-SAD) (see materials and methods section). A first crystal form of OgpA was obtained in the tetragonal *P* 4_1 2_1 *2* space group at 1.9 Å resolution ($OgpA_{WT1}$; PDB code 6Z2D; Table 1 in material and methods). A second crystal form of OgpA was obtained in the orthorhombic space *group P* 2_1 2_1 2_1 space group at 1.6 Å resolution ($OgpA_{WT2}$; PDB code 6Z2O; Table 1). The high quality of the electron density maps allowed the trace of residues 27 to 380 ($OgpA_{WT1}$) and 25 to 382 ($OgpA_{WT2}$).

The overall protein scaffold and the conformation of the $OgpA_{WT1}$ and $OgpA_{WT2}$ were essentially preserved (root mean square deviation, r.m.s.d., of 0.37 Å for 350 residues). However, connecting loop α4-α5 (loop 9; residues 227–246) is unstructured in $OgpA_{WT2}$, whereas in $OgpA_{WT1}$ displays a short α-helix. Furthermore, we were unable to fully model the α7–α8 connecting loop (loop 12; residues 309–320) on $OgpA_{WT2}$ crystal structure because it is disordered. We have decided to use the $OgpA_{WT2}$ crystal form for our description as it displays the highest resolution (Figure 19a,b).

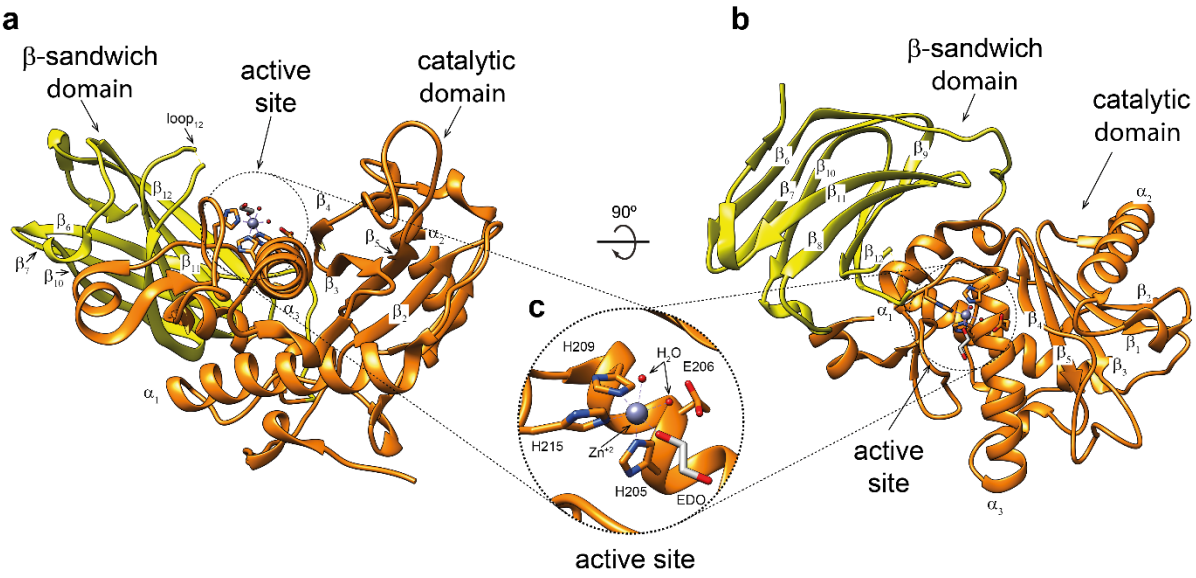

Figure 19: The overall structure and active site of OgpA. **a** Cartoon representation of the overall structure of $OgpA_{WT2}$. The two structural domains are highlighted: β-sandwich domain in yellow and catalytic domain in orange. **b** Another view of the overall structure of OgpA. **c** Closer view of the active site of OgpA. A zinc atom is shown (in purple), coordinated by three histidine residues (H205, H209 and H215), two water molecules (in red) and an ethylenglycol (EDO; in grey) molecule. The catalytic glutamic acid (E206) is also highlighted.

A close inspection of the two crystal structures revealed that the protein crystallized as a monomer. $OgpA_{WT}$ comprises two domains from the N- to the C-terminus: (i) a catalytic domain (residues 25-262) followed by (ii) a β-sandwich domain (residues 263-384). The central core of the N-terminal domain consists of a five-stranded β-sheet with topology β2–β1–β3–β5–β4 (β4 is antiparallel), surrounded by three long α-helices, α1, α2 and α3, with an overall size of 43 × 42 × 42 Å resolution (Figure 19a,b). The first β-sheet of the antiparallel β-sandwich domain consists of four β-strands with topology β6–β7–β8–β11, whereas the second β-sheet comprises four β-strands with topology β13–β12–β9–β10, with an overall size of 39 × 28 × 19 Å (Figure 19a,b). Both domains are separated by a long, open groove that runs parallel to the protein surface where the active site is located (Figure 19a,b).

Interestingly, a Zn^{2+} atom coordinates with the side chains of three histidine residues, H205, H209 and H215, a highly conserved triad among metalloproteases (Figure 19c). A glutamate residue, E206, is well-positioned to act as a general base/acid during catalysis. In addition, M231, lies within a Met-turn providing the hydrophobic base for the Zn^{2+} binding site (Figure 19c).[242–245] The

OgpA catalytic domain is annotated as an M11 peptidase, also called the gametolysin peptidase M11 family, according to the NCBI database. However, there are no *Akkermansia sp*. entries for the M11 family in MEROPS, a database that uses a hierarchical, structure-based classification of peptidases.[246] Interestingly, amino acid sequence homology analysis using the MEROPS scan data set (MEROPS-MPRO) indicates that the OgpA metal-binding motif shares high sequence identity with peptidases classified in the M12 family. Both families, M11 and M12, share the common motif HEXXHXXXXXH, in which the three His residues are zinc ligands and the Glu residue has a catalytic function. Structurally, the closest homologues also belong to the M12 family. However, it is worth noting that the M11 family is not structurally characterized. The M11 and M12 families belong to the MA(M) subclan, which comprises a metzincin fold with a methionine C-terminal to the Zn^{2+} atom.[245,247] Members of M11 and M12 families are proenzymes that require activation by limited proteolysis.[245] There is no experimental evidence to support that OgpA is a proenzyme. Taken together, all the data suggests that OgpA could be considered the founding member of a new family of peptidases.

A search for structural homologues using the DALI server (see Methods) revealed Zn-dependent metalloproteases with significantly low structural similarity to OgpA catalytic domain: (i) TNF-alpha converting enzyme (TACE; PDB code 3G42; Z-score of 11.3; r.m.s.d. value of 3.5 Å for 171 aligned residues; 12% identity)[248] (ii) atrolysin C from Crotalus atrox (PDB code 1ATL; Z-score of 10.9; r.m.s.d. value of 3.5 Å for 164 aligned residues, 12% identity),[249] and (iii) BaP1 from *Bothrops asper* (PDB code 1ND1; Z-score of 10.9; r.m.s.d. value of 3.4 Å for 163 aligned residues, 14% identity).[250] The three proteins are classified in the MEROPS M12 family (Figure 20a,b,c) and do not show specificity against *O*-glycopeptides. In addition, we found the following structural homologues for the β-sandwich domain using the same server: (i) the fibronectin type III (FnIII) domain of SleM from *Clostrium perfringens* (PDB code 5JIP; Z-score of 8.8; r.m.s.d. value of 2.8 Å for 90 aligned residues; 17% identity),[251] and (ii) FnIII domain of vacuolar protein

sorting-associated protein 26A (VPS26A) (PDB code 6H7W; Z-score of 9.1; r.m.s.d. value of 2.1 Å for 95 aligned residues; 14% identity),[252] and (iii) β-sandwich domain of glycoside hydrolase XacMan2A from *Xanthomonas axonopodis pv. citri* (PDB code 5DMY; Z-score of 9.0; r.m.s.d. value of 2.8 Å for 99 aligned residues; 11% identity; Figure 20d,e,f).[253] SleM is a peptidoglycan lysin, composed of an N-terminal catalytic domain similar to the GH25 family lysozymes and a C-terminal FnIII domain. The latter is involved in the formation of the SleM dimer.[251]

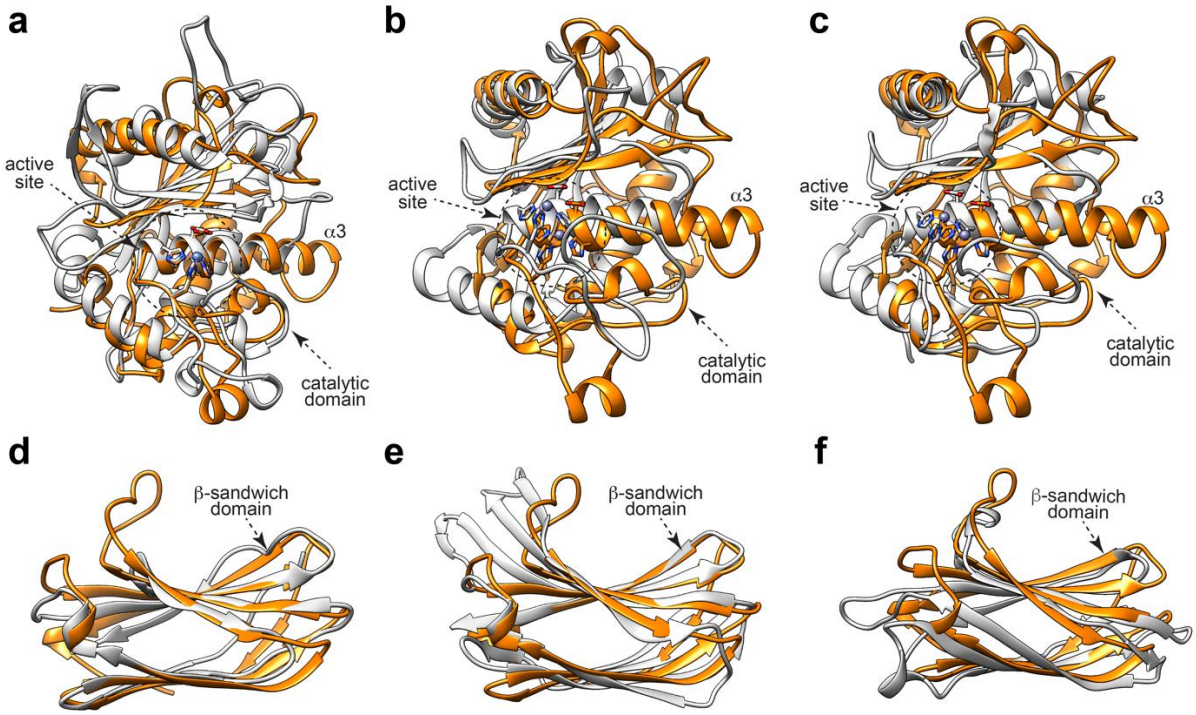

Figure 20: Structural homologues of OgpA. Structural superposition of the X-ray crystal structure of $OgpA_{WT}$ (orange) and structural homologues (grey): **a** TNF-alpha converting enzyme (TACE; PDB code 3G42), **b** Atrolysin C from *C. atrox* (PDB code 1ATL), **c** *Ba*P1 from *B. asper* (PDB code 1ND1), **d** FnIII domain of SleM from *C. perfringens* (PDB code 5JIP), **e** FnIII domain of vacuolar protein sorting-associated protein 26A (VPS26A; PDB code 6H7W), **f** β-sandwich domain of glycoside hydrolase XacMan2A from *X. axonopodis* pv. citri (5DMY; PDB code). Figure edited from Trastoy B. *et al*.(2020).[254]

4.2.2. The structure of OgpA in complex with *O*-glycan substrate

To obtain the crystal structure of the enzyme-substrate complex, we used a catalytically inactive version of OgpA, in which the residues H205, which coordinates the Zn^{2+} atom, and E206, the base/acid residue of the reaction, are mutated to alanine ($OgpA_{H205A/D206A}$).[242,247] The crystal

structure of $OgpA_{H205A/D206A}$ in complex with the *O*-glycopeptide GD was solved by molecular replacement methods ($OgpA_{H205A/D206A}$-GD-SUB; PDB code 6Z2P; Figure 21a,b; Table 2 and Methods section).

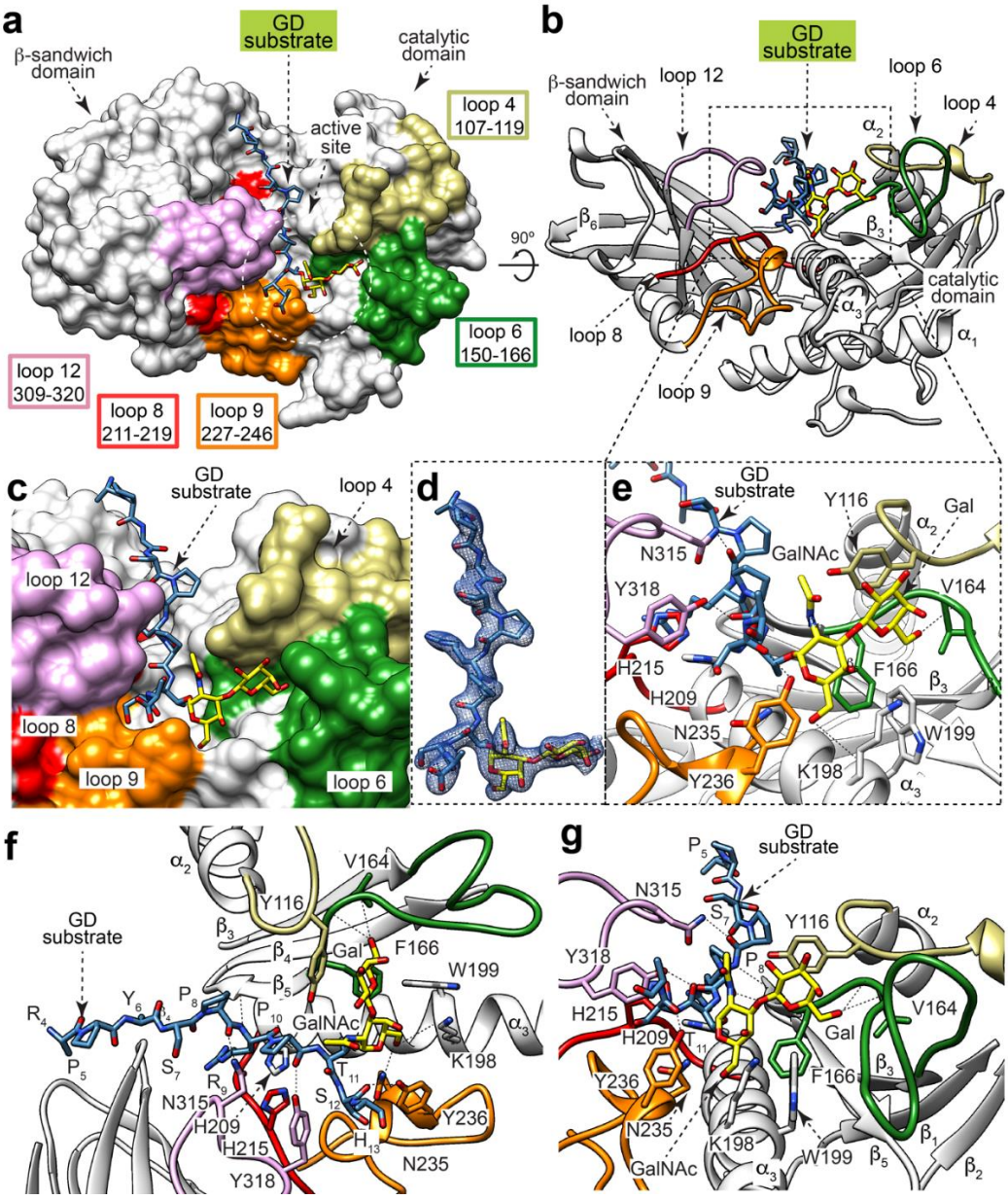

Figure 21: The *O*-glycopeptide substrate GD binding site. a Surface representation of the $OgpA_{H205A/D206A}$-GD-SUB crystal structure, with annotated domains and loops. **b** Cartoon representation of the $OgpA_{H205A/D206A}$-GD-SUB crystal structure. **c** Surface representation of OgpA showing the location of the GD *O*-glycopeptide substrate into the active site. **d** Electron density map of the GD substrate shown at 1.0 σ r.m.s.d. **e,f,g** Three different cartoon representations of OgpA showing the location of the GD *O*-glycopeptide into the active site, the main residues and secondary structure elements. The key hydrogen bond interactions between OgpA and the GD-SUB are shown in dotted lines. Figure edited from Trastoy B. *et al.*(2020).[254]

$OgpA_{H205A/D206A}$-GD-SUB crystallized in the tetragonal *I 4* space group and diffracted to a maximum resolution of 2.16 Å. The overall protein scaffold and the conformation of $OgpA_{H205A/D206A}$-GD-SUB were essentially preserved with respect to the $OgpA_{WT2}$ crystal structure (r.m.s.d. of 0.47 Å for 353 residues). The *O*-glycopeptide $_4$RPYSPRPT(Gal-GalNAc)SH_{13} substrate was unambiguously identified in the electron density map (Figure 21d). We assume that the full-length GD is in the crystal, but no density was observed for the residues 1-3 and 14-19, probably due to conformational flexibility.

The active site of metalloproteases has been defined in subsites that interact with the substrate side chains of the N-terminal region with respect to the scissile bond, non-primed side (P1, P2, P3, etc.), and the C-terminal region of the peptide to be cleaved, primed side (P1', P2', P3', etc.).[255,256] The subsites of the metalloproteases that interact with the N-terminal region are named S1, S2, S3, and those that interact with the C-terminal region S1', S2' and S3'. This peptidase-substrate recognition paradigm has been updated for *O*-glycopeptidases,[257] including subsites of the enzyme that specifically recognize the *O*-glycan moiety that has been termed G-sites (Figure 22).

Strikingly, the GD substrate is inserted as an extra parallel β-strand, relative to the other four strands of the β-sheet comprised in the catalytic domain (Figure 23), in a binding pocket located in the cleft between the N- and C-terminal domains of OgpA (Figure 21a,b,c,d,e; Figure 22). The *O*-glycopeptide is flanked by α3 and α4, as well as the connecting loops β2–α2 (loop 4; residues 107–119), β3-β4 (loop 6; residues 150–166), α3-α4 (loop 8; residues 211–219), α4-α5 (loop 9; residues 227-246) and β8-β9 (loop 12; residues 309-320; Figure 21b,c). Amino acid residues located at both ends of the *O*-glycopeptide, including $_4$RPYS_7 and H13, are solvent-oriented and do not interact with OgpA (Figure 21e,f,g).

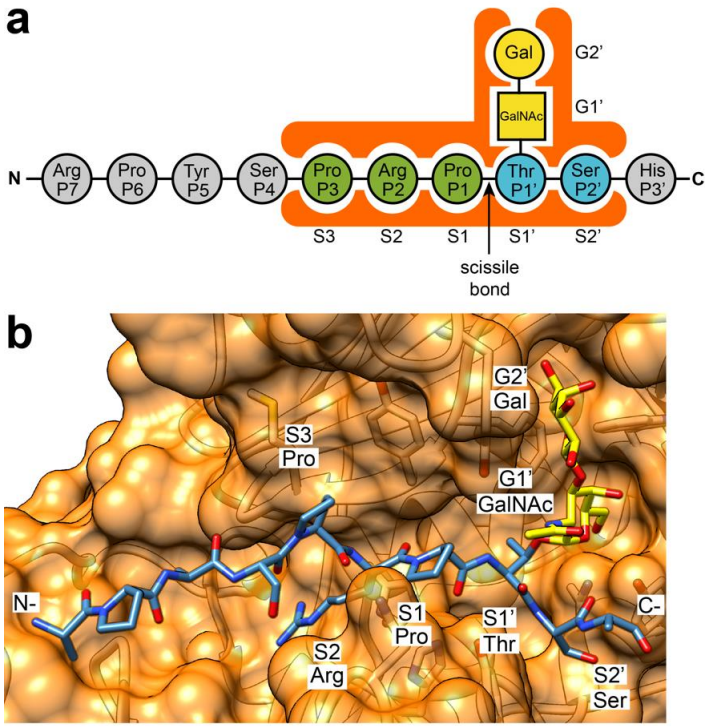

Figure 22: Subsite nomenclature for OgpA. a The proposed subsite nomenclature for OgpA substrate recognition. Amino acids that are solvent exposed and not interacting with OgpA are colored in grey; amino acids of the N-terminal with respect to the scissile bond that interact with OgpA subsites are colored in green; amino acids of the C-terminal with respect to the scissile bond that interact with OgpA subsites are colored in blue. Gal and GalNAc residues are colored in yellow, **b** Surface representation of the active site of the $OgpA_{H205A/D206A}$-GD-SUB crystal structure with annotated subsites involved in substrate recognition. Figure edited from Trastoy B. *et al*.(2020).[254]

Interestingly, we observe an extensive network of interactions, primarily mediated by hydrogen bonds, between GD's $_{8}$PRPTS$_{12}$ main chain residues and OgpA's N315, H215, Y318, and Y236 side chain residues (Figure 21e,f,g). The GD backbone forms hydrogen bonds with the M169 and Y167 main chains (Figure 21e,f,g). In addition, P6 side chain interacts with the N315 side chain, while the P8 side chain makes hydrophobic interactions with the M169 and Y168 side chains. Finally, the R9 side chain makes electrostatic interactions with the H209 and L213 main chains, and the S12 side chain with N235, and hydrophobic interactions with the Y236 and Y318 side chains. OgpA mainly interacts with neighbouring residues of the glycosylated threonine, suggesting that the enzyme shows a broad range specificity for the peptide backbone.

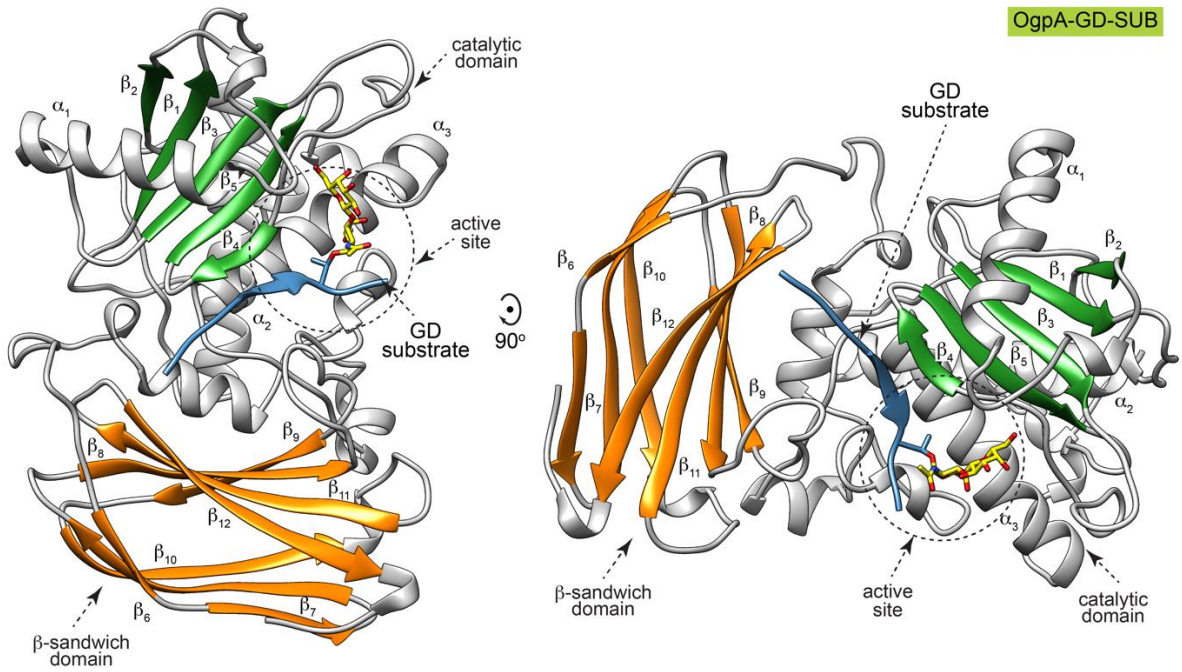

Figure 23: Cartoon representation showing the general fold and secondary structure organization of OgpA. Secondary structure elements are labelled. The central β-sheet of the catalytic domain and the β-sandwich are coloured in green and orange, respectively. The *O*-glycopeptide substrate is coloured in blue. It forms a parallel β-strand with respect to β2, β1, β3 and β5 comprised in the β-sheet of the catalytic domain. The disaccharide GalNAcGal is coloured in yellow. Figure edited from Trastoy B. *et al*.(2020).[254]

The crystal structure of $OgpA_{H205A/D206A}$-GD-SUB shows that protein-peptide interaction is primarily mediated by hydrogen bonds between the Pro-P3, Arg-P2, Pro-P1 and Ser-P1' backbone and the residues of OgpA that form the corresponding subsites (Figure 21, Figure 22). The side chains of these amino acid residues also interact with OgpA by hydrophobic interactions. Mucins, the potential target for OgpA *in vivo*, consist of a long, densely *O*-glycosylated domain with sequences rich in Pro, Thr, and Ser, often characterized by tandem repeats. His-P3', Ser-P4, Tyr-P5, pro-P6 and Arg-P7 that form the *O*-glycopeptide are exposed to the solvent and do not interact with the protein. This indicates that the side chain nature of these residues is not a key element for OgpA activity. In that sense, the structural comparison of the four structures obtained for OgpA indicates that the substrate-binding site is essentially preformed, which supports a conformational selection mechanism for the recognition of the *O*-glycopeptide.[258]

OgpA appears to show no preference for a specific amino acid sequence on the *O*-glycopeptide substrate. Indeed, a recent publication using OgpA mediated-digestion to map more than 3,000 *O*-glycosylation sites from different tissue and cell samples specifically addressed this issue and found no evidence of sequence biases or digestion at sites lacking *O*-glycans.[259] Previously known *O*-glycosylation sites were reliably detected regardless off sequence context, further validating the method.

In contrast, the disaccharide Gal-GalNAc is placed in a pocket decorated by four aromatic residues, Y116, F166, W199, and Y236 (Figure 21e,f,g). Specifically, the GalNAc moiety is surrounded by F166, W199, and Y236. The O6 of GalNAc of GD makes a hydrogen bond with the K198 main chain and also interacts with the N235 side chain. The Gal residue is surrounded by Y116 (CH-π interaction), F166, and W199. Furthermore, the O6 of Gal makes hydrogen bonds with the V164 main chain and also interacts with the G163 backbone. Finally, the carbonyl oxygen of the 2-acetamido group of GalNAc forms a hydrogen bond with the O2 of the Gal residue. This suggests that the interaction between the OgpA G-sites and the *O*-glycan are key structural determinants for the recognition mechanism and the reaction.

4.2.3. The structure of OgpA in complex with *O*-glycan product

The strategy for capturing a native binary enzyme-product complex was to perform co-crystallization experiments with full-length wild-type OgpA in the presence of the GD *O*-glycopeptide substrate. It is worth noting that the enzyme was active against GD (Methods section). Thus, we obtained a snapshot of OgpA in complex with the proteolysis product of GD (OgpA$_{WT}$-GD-PRO; PDB code 6Z2Q; Figure 24a,b; Table 3 and Methods section). OgpA$_{WT}$-GD-PRO crystallized in the tetragonal space group *P* 4_1 2_1 *2* and diffracted to a maximum resolution of 2.34 Å (Table 3).

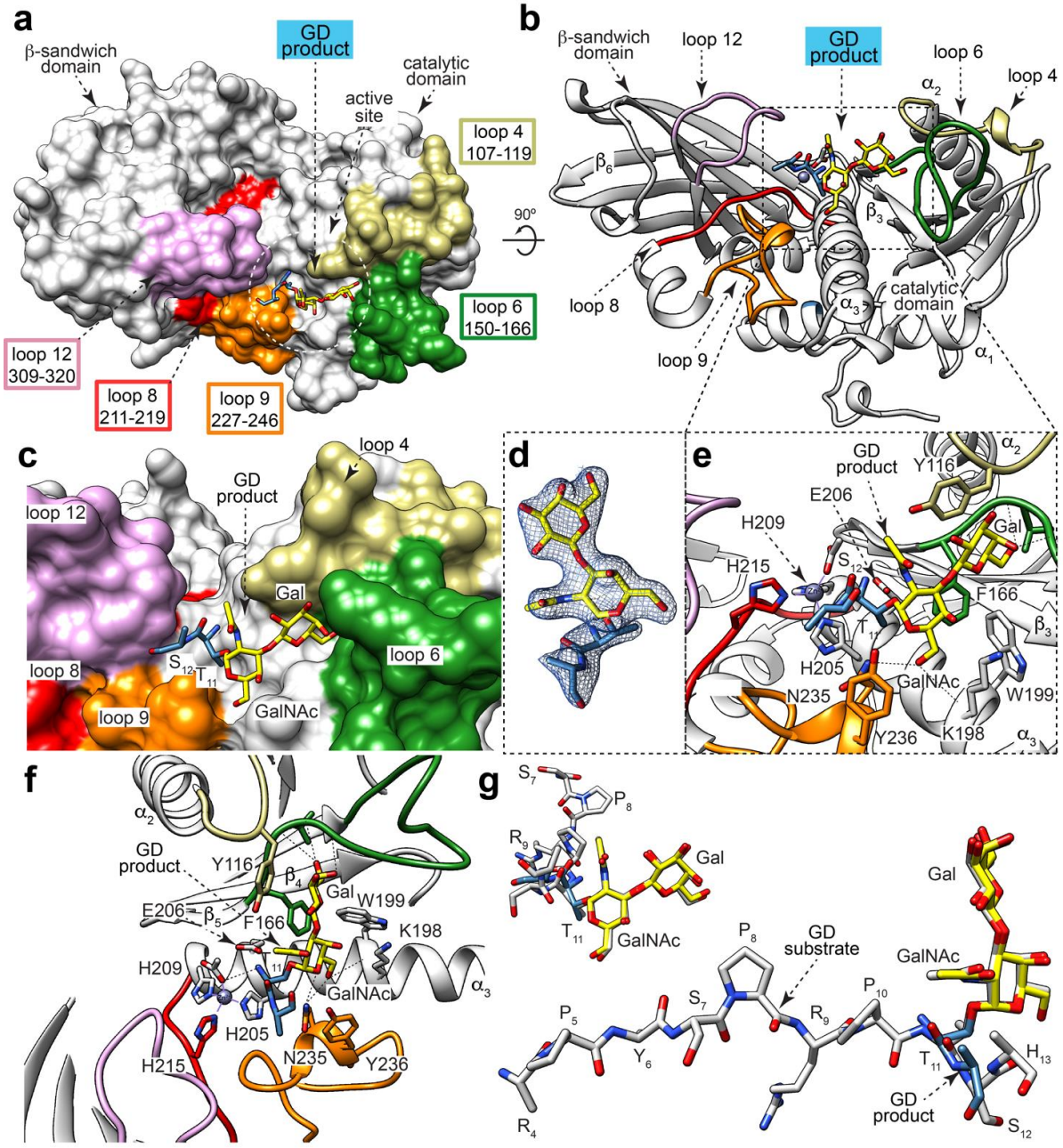

Figure 24: The *O*-glycopeptide GD product binding site. Surface representation of the $OgpA_{WT}$-GD-PRO crystal structure, with annotated domains and loops. **b** Cartoon representation of the $OgpA_{WT}$-GD-PRO crystal structure. **c** Surface representation of OgpA showing the location of the GD *O*-glycopeptide product into the active site. **d** Electron density map of the GD product shown at 1.0 σ r.m.s deviation. **e,f** Two different cartoon representations of OgpA showing the location of the GD *O*-glycopeptide product into the active site, the main residues and secondary structure elements. **g** Two views of the structural superposition of GD *O*-glycopeptide substrate (grey) and product (blue and yellow) in the $OgpA_{WT}$-GD-SUB and $OgpA_{WT}$-GD-PRO crystal structures, respectively. The key hydrogen bond interactions between OgpA and the GD-PRO are shown in dotted lines. Figure edited from Trastoy B. *et al*.(2020).[254]

The overall architecture and the conformation of $OgpA_{WT}$-GD-PRO were essentially preserved with respect to the crystal structure of $OgpA_{WT2}$ (r.m.s.d. of 0.37 Å for 344 residues) and that of the $OgpA_{H205A/D206A}$-glycodrosocin complex (r.m.s.d. of 0.42 Å for 353 residues). The $_{11}$T(Gal-

GalNAc)S_{12} *O*-glycopeptide product was unambiguously identified in the electron density map (Figure 24d). In contrast, there is no electron density for the $_{1}$GKPRPYSPRP$_{10}$ peptide, supporting the notion that it leaves the active site immediately after the proteolytic reaction takes place. As depicted in Figure 24g,f, the conformation of the disaccharide Gal-GalNAc group observed in the $_{4}$RPYSPRPT(Gal-GalNAc)SH_{13} substrate complex and in the $_{11}$T(Gal-GalNAc)S_{12} product complex, superimpose very well, as well as all residues located at the substrate/product binding site.

4.2.4. Structural and sequence comparison with other *O*-glycopeptides recognizing peptidases

To date, X-ray crystal structures of four enzymes with endopeptidase activity against *O*-glycopeptides have been reported: (i)–BT4244 from *B. thetaiotaomicron*,[257] (ii) IMPa from *Pseudomonas aeruginosa*,[257] (iii) ZmpB from *Clostridium perfringens* (strain ATCC 13124),[257] and (iv) StcE from enterohemorrhagic *Escherichia coli*.[260] No one of them appears as a structural homologue of OgpA, according to our structural homology search performed by DALI (see methods).

The BT4244 and ZmpB enzymes belong to the M60 family, while StcE and IMPa belong to the M66 and M88 families, respectively. Members of the M60, M66 and M88 families belong to the MA clan. They share the common catalytic mechanism and the conserved metal-binding motif HEXXH described for OgpA. Members of the M60 and M88 families coordinate the Zn^{2+} atom with two histidine residues and one aspartic acid, while members of the M66 family do not show the conserved methionine residue.

The overall structures of BT4244, ZmpB, IMPa and StcE are quite different compared to OgpA, and show additional domains. However, they all share some common structural elements that build

the catalytic and the substrate binding sites (Figure 25a,b,c,d). All five enzymes show a central α-helix (α3 in OgpA) and at least five-stranded β-sheet with different topology than OgpA.

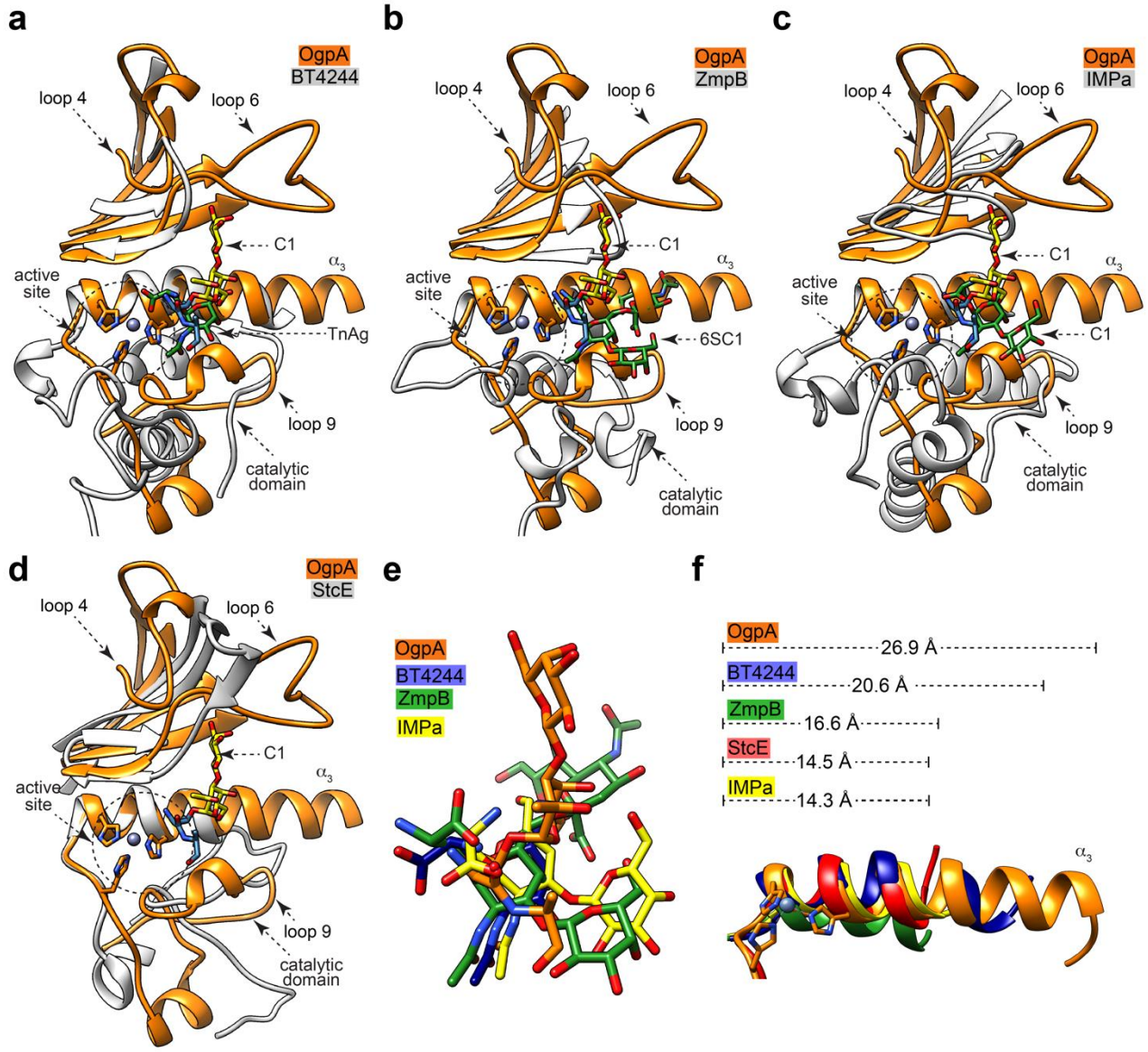

Figure 25: Structural basis of *O*-glycopeptide processing peptidases. Structural comparison of OgpA and **a** BT4244 from *B. thetaiotaomicron* (PDB code 5KD8), **b** ZmpB from *C. perfringens* (strain ATCC 13124) (PDB code 5KDU) **c** IMPa from *P. aeruginosa* (PDB code 5KDX), and **d** StcE from the enterohemorrhagic *E. coli* (PDB code 3UJZ). The α-helices, β-strands and loops that interact with the *O*-glycans are represented in orange for OgpA, and in grey for the other enzymes. The *O*-glycopeptide product found in the crystal structure of OgpA is colored in yellow and blue. The *O*-glycopeptide product found in the crystal structures of BT4244, ZmpB, ImpA and StcE, is colored in green. **e** Superposition of the *O*-glycan found in the X-ray crystal structure of OgpA (orange), BT4244 (blue), ZmpB (green) and IMPa (yellow) **f** The α-helix length in each OgpA (orange), BT4244 (blue), ZmpB (green) and IMPa (yellow)and StcE (red) *O*-glycan endopeptidases complexes. Figure edited from Trastoy B. *et al.*(2020).[254]

The nature and structural arrangement of the loops surrounding the *O*-glycan are also unique to each enzyme (Figure 25a,b,c,d). X-ray crystal structures of BT4244, ZmpB and IMPa in complex with an *O*-glycopeptide product revealed that the orientation of the glycan at the binding site also differs from that observed in OgpA. Specifically, in the OgpA-GD-PRO crystal structure, the *O*-glycan is found above α3 helix making interactions with residues of loops 4, 6, and 9, while in BT4244, ZmpB and IMPa structures the equivalent loop 6 is shorter than in OgpA; also loops 4 and 6 are far from the *O*-glycan binding site. This suggests that the loops that form the recognition subsites are different between the metalloprotease families. Moreover, the length of the α3 helix also differs markedly between these five enzymes (Figure 25e). OgpA and BT4244 show a longer α3 helix than ZmpB, IMPa and StcE.

Amino-acid sequence alignment revealed that OgpA shows a high degree of sequence identity with other putative metalloproteases from two phyla: Verrucomicrobia and Bacteroidetes. In addition to the proteases of other species of the genus *Akkermansia* of the phylum Verrucomicrobia, we found other homologues belonging to three classes of the phylum Bacteroidetes: *Bacteroidia* (*Marinifilum* and *Odoribacter*), *Chitinophagia* (*Chitonophaga*) and *Flavobacteriia* (*Elizabethkingia*, *Chryseobacterium*, *Zhouia* and *Imtechella*; Figure 26). The catalytic residue E206 and the residues that coordinate the Zn^{2+} atom, H205, H209 and H215, are conserved among these species. Moreover, residues that interact with the disaccharide Gal-GalNAc, Y116, F166, K198, W199, N235 and Y236 in OgpA, are also conserved among the species, supporting a similar mechanism of *O*-glycan recognition (Figure 26). In the case of BN783_01361 from *Odoribacter sp*. CAG:788, the α3 helix shows 5 residues less than the other homologues, suggesting that this putative metalloprotease could accept different *O*-glycopeptides than OgpA.

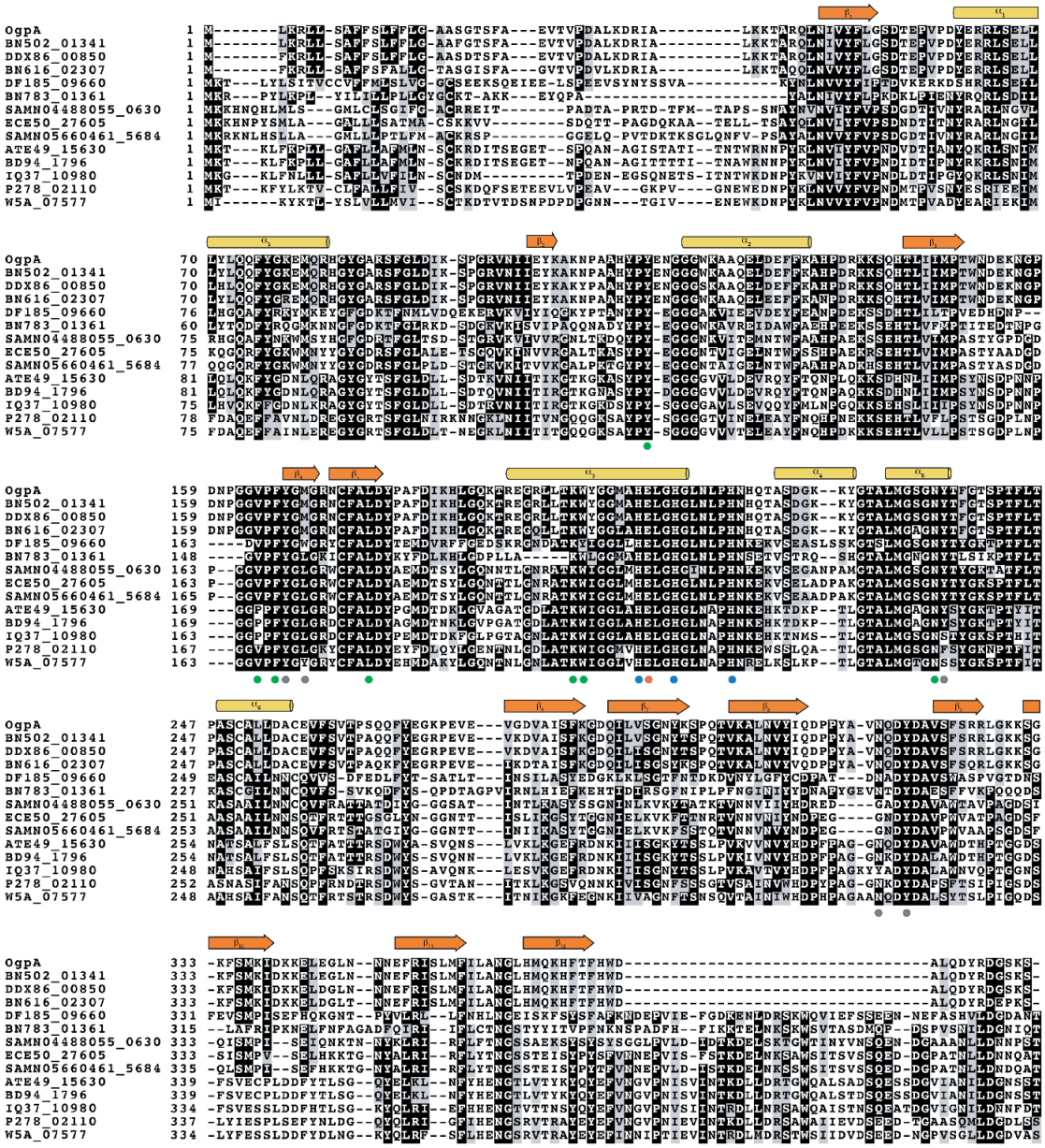

Figure 26: Sequence alignment of OgpA with homologues found in the Verrucomicrobia and Bacteroidetes phyla. Comparison of OgpA from *A. muciniphila* (B2UR60, Uniprot code), with BN502_01341 from *A. muciniphila* CAG:154 (R6J0R4, Uniprot code, Identity: 99 %), DDX86_00850 from *Akkermansia sp.* (A0A354E6N0, Uniprot code, Identity: 97 %), BN616_02307 from *Akkermansia sp. CAG:344* (A0A139TPP8, uniprot code, Identity: 91%), DF185_09660 from *Marinifilum breve* (A0A2V3ZZK2, uniprot code, Identity: 38 %), BN783_01361 from *Odoribacter sp. CAG:788* (R5PJX3, Uniprot code, Identity: 47%), SAMN04488055_0630 from *Chitinophaga niabensis* (A0A1N6DD36, Uniprot code, Identity: 41%), ECE50_27605 from *Chitinophaga sp. Mgbs1* (A0A3S1CWX3**,** Uniprot code, Identity: 43%), SAMN05660461_5684 from *Chitinophaga ginsengisegetis* (A0A1T5PAV1, Uniprot code, Identity: 44%), ATE49_15630 from *Elizabethkingia miricola* (A0A1A6C9I9, Uniprot code, Identity: 44%), BD94_1796 from *Elizabethkingia anophelis* (A0A077EDG2, Uniprot code, Identity: 39%), IQ37_10980 from *Chryseobacterium piperi* (A0A086BCN5, Uniprot code, Identity: 41%), P278_02110 from *Zhouia amylolytica* (W2URL6, Uniprot code, Identity: 38%), W5A_07577 from *Imtechella halotolerans* (I0WF82, Uniprot code, Identity: 41%). The catalytic residue and the residues that coordinate the Zn atom are marked with red and blue dots, respectively. The residues that interact with the Gal-GalNAc glycan and the aminoacids of the GD substrate in the crystal structure of $OgpA_{H205A/D206A}$-GD-SUB are marked with green and grey dots, respectively. Figure edited from Trastoy B. *et al.*(2020).[254]

4.2.5. Catalytic mechanism of OgpA

Solving the crystal structure of unliganded form of OgpA ($OgpA_{WT1}$ and $OgpA_{WT2}$) together with the complexes with the glycosylated peptide substrate (GD) and the reaction product, $OgpA_{H205A/D206A}$-GD-SUB and $OgpA_{WT}$-GD-PRO, respectively, show us the most relevant snapshots of the catalytic mechanism of OgpA. $OgpA_{H205A/D206A}$-GD-SUB crystal structure shows the binding of the *O*-glycopeptide substrate and $OgpA_{WT}$-GD-PRO, the binding of the reaction product ready to be released from the active site (Figure 27a)

OgpA displays a conserved metal binding site HEXXHXXXXXH.[246] The proposed catalytic mechanism is similar to that described for thermolysin from *Bacillus thermoproteolyticus* and bovine carboxypeptidase A [244,261,262] decades ago, and more recently for matrix metalloproteases (MMPs).[243,263] It corresponds to a single-displacement that comprises the nucleophilic attack of a catalytic solvent molecule polarized by the general base/acid glutamate and the catalytic Zn^{2+} atom (Figure 27b).[243,244,261–264]

The water molecule is part of the functional enzyme. It coordinates first with the metal atom, and is polarized by the base/acid glutamate, enhancing its nueclophlity. Once the peptide substrate is bound in the active cleft, the scissile carbonyl oxygen is polarized by the catalytic Zn^{2+} atom. Thus, the water molecule can attack the peptide's carbonyl oxygen of and simultaneusly transfer a proton to the general base glutamate. The nucleophilic attack produces a *gem*-diolate tetrahedral reaction intermediate that the Zn^{2+} atom and neighboring residues stabilize. This intermediate leads to the formation of a double-product complex mediated by scissile bond cleavage and double proton transfer to the new α-amino terminus. Finally, the two products are released from the enzyme, most likely the nonprimed product first.

The Zn^{2+} atom coordinates with three histidines residues, H205, H209 and H215, and three water molecules in the $OgpA_{WT1}$ crystal form, and two water molecules and one ethylenglycol molecule

in the $OgpA_{WT2}$ crystal form, adopting in both cases an octahedral geometry almost regular (Figure 27a). Although the lowest-energy ground-state coordination number of zinc bound to one acidic or two or more neutral protein ligands is 4,[263] the octahedral geometry has been observed in other unliganded metalloproteases.[263] The base/acid residue E206 is in close contact with the catalytic water molecule in both crystal forms, $OgpA_{WT1}$ and OgA_{WT2} (3 Å and 2.6 Å, respectively). In addition, Y236 makes hydrogen bonds with the base/acid residue E206 and partially occupies the substrate binding site in the $OgpA_{WT2}$ crystal structure, while in OgA_{WT1} loop 9 adopts a different conformation and Y236 is positioned distant to E206 (Figure 27a). Due to the mutation of H205 by alanine, the Zn^{2+} cation is not present in the active site of $OgpA_{H205A/D206A}$-GD-SUB structure. The entrance of the substrate does not cause conformational changes in the loops that surround the binding pocket. However, the position of Y236 changes to make hydrogen bonds with the *O*-glycopeptide backbone. In the absence of the Zn^{2+} atom, the scissile carbonyl oxygen is making hydrogen bonds with Y318 and H215, at a distance that could be coordinated by the Zn^{2+} atom showed in the unliganded crystal structures (2.4 Å in OgA_{WT1} and $OgA_{WT2,}$ respectively). In the $OgpA_{WT}$-GD-PRO crystal structure, the nonprimed peptide product is not present, supporting the notion that it is the first to leave the active site. In addition, the Zn^{2+} is tetracoordinated with an acetate molecule from the crystallization condition. The amino terminus of the GD-PRO interacts with E206 by hydrogen bonds and the rest of the interactions that maintain the primed peptide in the active site are mediated by the *O*-glycan and the protein.

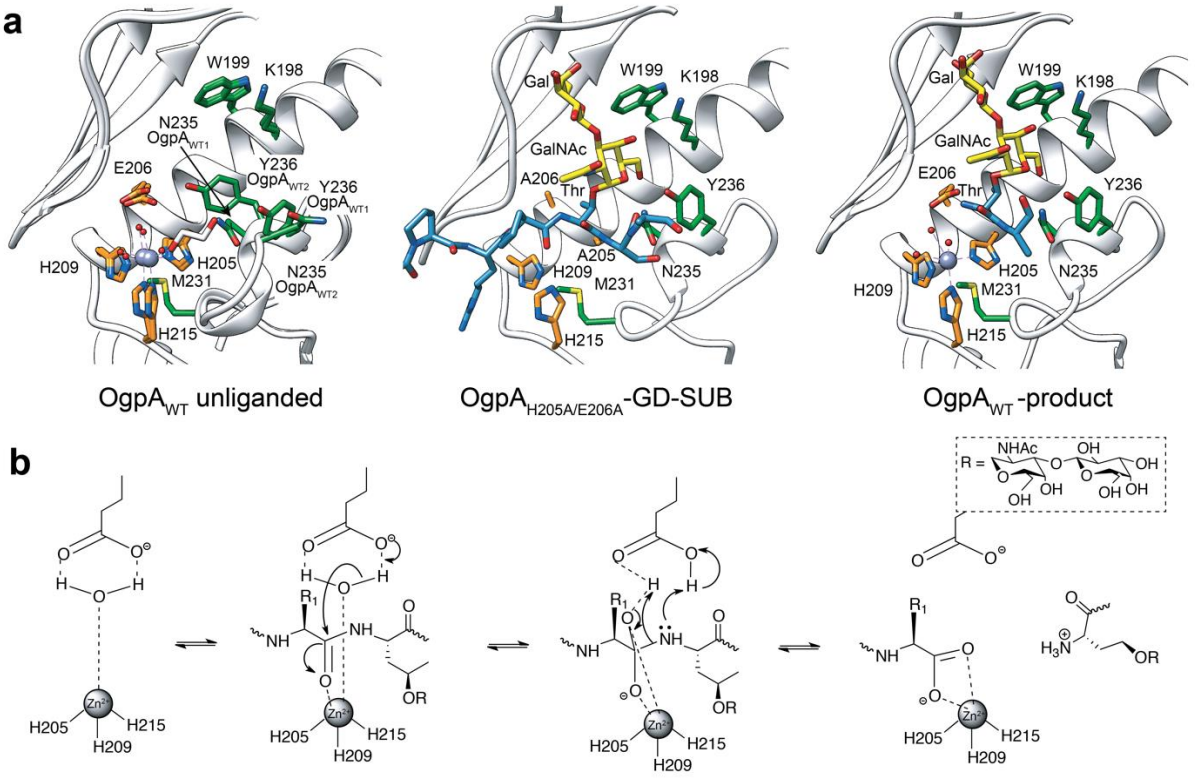

Figure 27: The catalytic mechanism of OgpA. a Catalytic site of OgpA in the OgpA$_{WT2}$ unliganded (left), OgpA$_{H205A/D206A}$-GD-SUB (center) and OgpA$_{WT}$-GD-PRO (right) crystal structures. **b** Proposed catalytic mechanism for OgpA. Figure edited from Trastoy B. *et al*.(2020).[254]

4.2.6. Structural basis of OgpA specificity for *O*-glycopeptides

O-glycans are structurally diverse, with up to 8 different core arrangements identified in mammals. We studied the substrate specificity of OgpA for the most common *O*-glycan core structures in vitro using synthetic *O*-glycopeptides as substrates. To this end, we synthesized short *O*-glycopeptides based on the MUC1 sequence region, with a fluorescent label at the C-terminus. OgpA activity was measured using reverse-phase HPLC (Figure 28). The peptide carrying a core 1 glycan (C1, Gal-β1,3-GalNAc) was by far the best substrate, while core 3 (C3, GlcNAc-β1,3-GalNAc) and α2,3-sialylated core 1 (3SC1, Neu5Ac-α2,3-Gal-β1,3-GalNAc), could be turned over although at very low rates. In contrast, there was no digestion of α-2,6-sialylated core 1 (6SC1, Gal-β1,3-(Neu5Ac-α2,6)-GalNAc) or core 2 (C2, Gal-β1,3-(GlcNAc-β1,6)-GalNAc) *O*-glycopeptides. Peptides lacking the β1,3-linked galactose (Tn antigen, TnAg) and non-glycosylated peptides were not digested by OgpA (Figure 28 and Figure 29a).

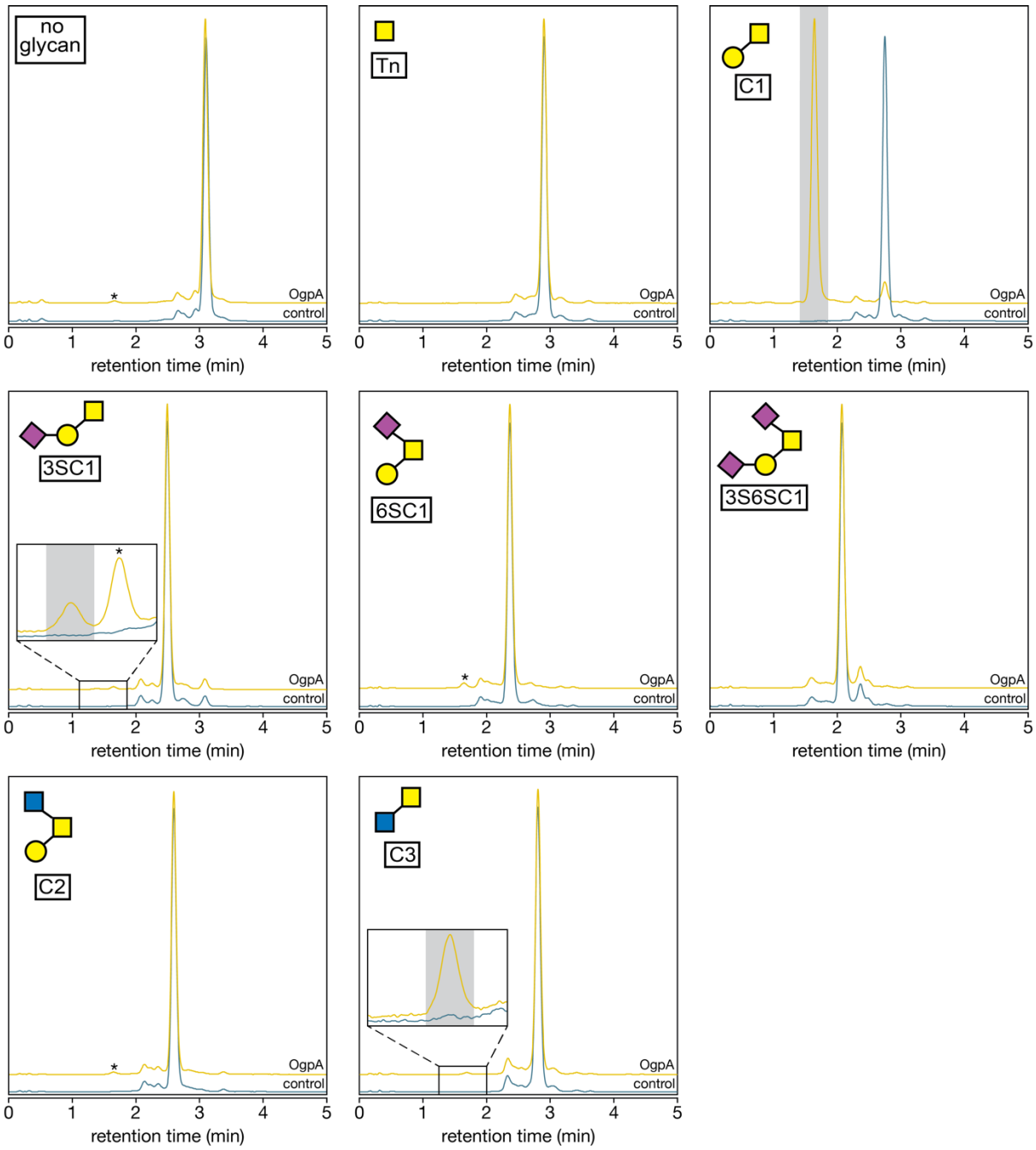

Figure 28: Substrate specificity of OgpA for *O*-glycopeptides. The different glycopeptides were incubated with OgpA overnight (18 hs) and the reaction mixture was analyzed by high pH reverse phase HPLC with fluorescence detection. Each panel shows an overlay of an OgpA digestion (yellow) with a control reaction without the enzyme (teal). Reaction products are highlighted in grey. The asterisk marks peaks corresponding to digested C1 peptide. These originated from incomplete turnover during synthesis of the other core structures using C1 as starting material. Figure edited from Trastoy B. *et al.*(2020).[254]

In vitro experiments with previously mentioned endopeptidases (BT4244, IMPa, ZmpB and StcE) show different substrate specificities. These experiments on defined synthetic *O*-glycopeptides (TnAg, C1, 6SC1 and 3SC1) revealed that IMPa can hydrolyze TnAg, C1, 6SC1 and 3SC1, while

BT4244 hydrolyzes TnAg, and ZmpB exclusively hydrolyzes 6SC1.[257] StcE cleaves densely *O*-glycosylated mucin proteins in the following residue to an *O*-glycosylated Ser or Thr and shows activity on elaborated C1 and C2 structures.[265]

To further understand the OgpA substrate specificity at the molecular level, we performed docking calculations with the natural C1 *O*-glycan substrate and the different *O*-glycan cores evaluated in the activity assays (Figure 29b,c,d,e,f,g,h.). We first studied TnAg, which only displays the GalNAc residue observed in C1 (Figure 29b,c). Although the TnAg can be placed in the binding pocket, it does not fulfil all of the requested interactions with the enzyme to be specifically recognized (figure 29c). When 3SC1 is placed in the binding pocket of OgpA, the terminal Neu5Ac makes clashes with residues in the connecting loops between β2-α2 and β3- β4 (loops 4 and 6; Figure 29d). Addition of a Neu5Ac residue to the GalNAc of C1, in the forms of Gal-β1,3-(Neu5Ac-α2,6)-GalNAc (6SC1) and Neu5Ac-α2,3-Gal-β1,3-(Neu5Ac-α2,6)-GalNAc (3S6SC1), induce major steric repulsions with the long α3 helix of OgpA, as observed in C2 (Figure 29e-g). As a consequence, the enzymatic activity is markedly reduced (3SC1) (Figure 29d) or completely abolished (6SC1, 3S6SC1 and C2; Figure 29e,f,g). Finally, slight shape changes in the second residue, as occurs in the C3 ligand, slow down activity 153-fold due to the proximity with the two mentioned connecting loops, more specifically, with residues Y116 an F166 (Figure 29h). In this case, the only differences compared to C1 are the arrangement of the 4-OH group that converts galactose to glucose and the acetylation of the 2-position of the GlcNAc residue (Figure 29b,h). Taken together, the combination of enzymatic activity measurements and molecular docking calculations, strongly support that OgpA is highly specific with respect to its *O*-glycan substrate, with the second Gal residue of C1 playing a prominent role.

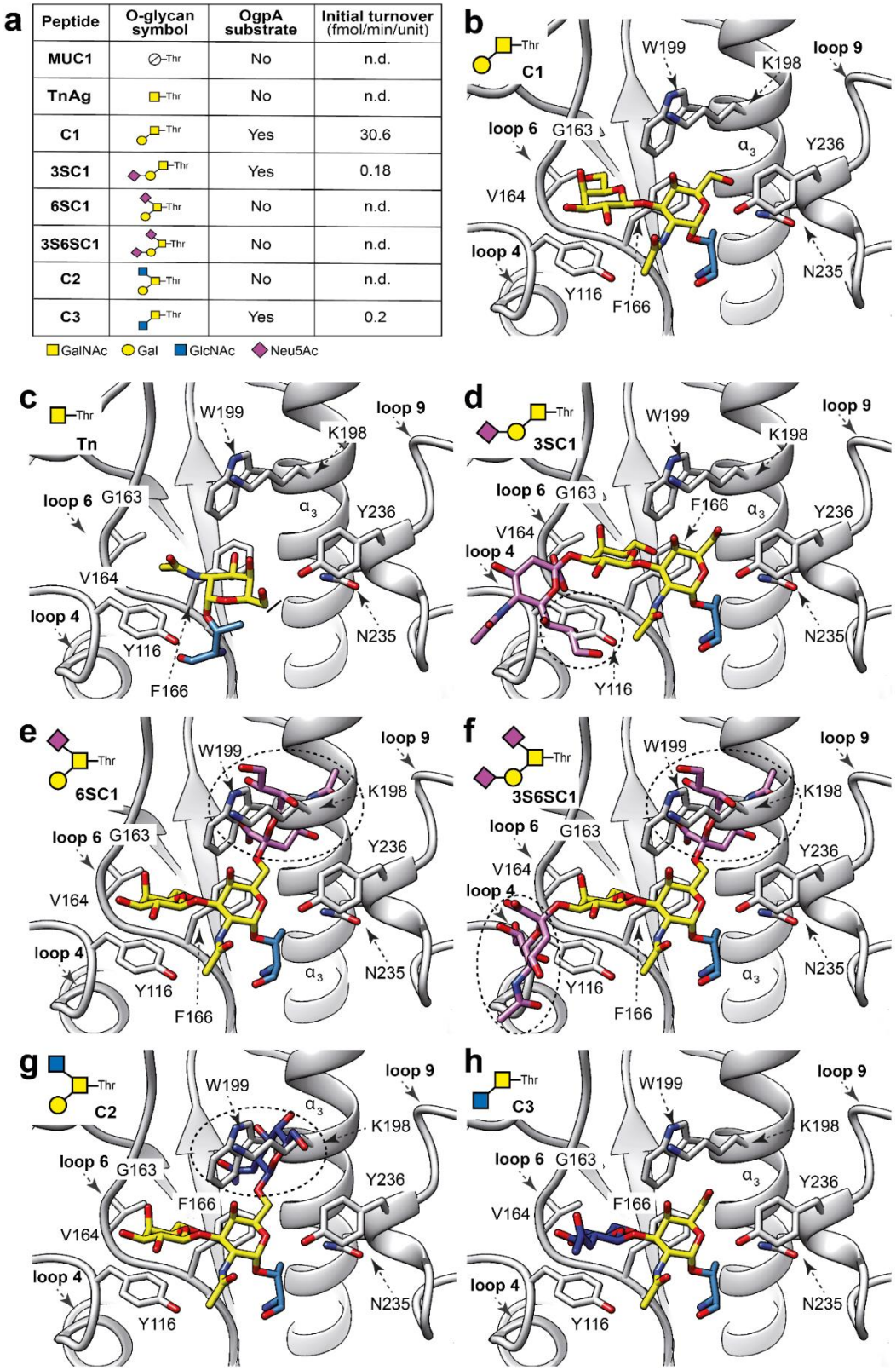

Figure 29: Structural basis of OgpA specificity for *O*-glycopeptides. a Hydrolytic activity of OgpA is shown, as determined by reverse-phase HPLC analysis. **b** X-ray crystal structure of $OgpA_{WT}$-GD-PRO with C1 glycan. Docking calculations of OgpA with different *O*-glycans, including **c** Tn, **d** 3SC1, **e** 6SC1, **f** 3S6SC1, **g** C2 and **h** C3. The predicted clashes are shown as dotted circles. Figure edited from Trastoy B. *et al*.(2020).[254]

Molecular docking calculations clearly indicate that the OgpA α3 helix blocks the entrance of *O*-glycopeptides with substitutions in the O6 group of the first GalNAc residue (Figure 29). Interestingly, BT4244 is unable to hydrolyze 2,6 sialylated *O*-glycopeptides, most likely due to the long α3 helix that blocks the access of the extra carbohydrate moiety at this position (Figure 25). Supporting this notion, ZmpB, IMPa and StcE which shows a short α3, can hydrolyze 2,6 sialylated *O*-glycopeptides (Figure 25). In summary, the α3 helix is a common structural element among the metalloproteases of the gluzincins and metzincins families involved in the generation of the Zn2+ binding motif, but also plays a key structural role in the selection of the *O*-glycan substrate in *O*-glycopeptidases.[244]

5. UNVEILING THE STRUCTURAL BASIS AND SUBSTRATE RECOGNITION EVENT OF FucOB, THE MUCIN-DEGRADER *A. MUCINIPHILA'S* α-1,2-FUCOSIDASE

FucOB (encoded by *Amuc_1120*) from *A. muciniphila* is a novel α-1,2-fucosidase able to hydrolyse Type I, Type II, Type III and Type V H antigens, glycosyl antigens found in many *O*-glycosylated proteins, such as mucins or red blood cells (RBCs) surface. We have first described FucOB's enzymatic activity, substrate specificity, and structural description thanks to the high-resolution X-ray crystallographic unliganded structure.

Moreover, we have described the ability of FucOB to convert O type universal blood into rare Bombay type blood thanks to agglutination and flow cytometry-based techniques, providing a potential biotechnological tool to facilitate blood transfusion to Bombay phenotype people.

5.1. MATERIAL AND METHODS

5.1.1. Materials

Blood group H antigen triaose type I, blood group H antigen triaose type V (2'-fucosyllactose), blood group H antigen triaose type II, blood group A antigen triaose type V, blood group B antigen

triaose type V were purchased from ELICITYL. 3-fucosyllactose was purchased from Carbosynth, GDP-Fucose and Lewis a trisaccharide from Sigma Aldrich, and 6-fucosylchitobiose (GlcNAcβ(1-4)[Fucα(1-6)]GlcNAc) from TCI. Anti-blood group H ab antigen antibody [97-I] (ab24213; the epitope is a carbohydrate moiety) and mouse IgM [B11/7]-Isotype control (ab91545) were purchased from Abcam, anti-blood Group H n/ab antigen antibody [86-M] (AGM-022YJ) was purchased from Creative Biolabs, goat anti-mouse IgM (Heavy chain) cross-adsorbed secondary antibody, PE and CaptureSelect IgG-Fc were purchased from ThermoFisher Scientific, and BV421 Mouse Anti-Human CD235a were purchased from BD Biosciences. Anti-H lectin reagent (extract of Ulex europaeus seeds) was purchased from BioRad and FITC-conjugated anti-H lectin (also extract of Ulex Europeaus) (L32476) was purchased from ThermoFisher Scientific. SialEXO, PNGase F and FabRICTOR were from Genovis AB and Recombinant Fut2 was purchased from R&D systems. Fucosidase 95A from *Bifidobacterium longum* (CZ0511) was purchased from NZYTech.

5.1.2. Cloning of wild-type and single point mutants of FucOB from *A. muciniphila* strain ATCC BAA-835 and *Bb*AfcA

The pET29a-*Amuc_1120*, pET29a-*Amuc*_1120^{E541A} (pET29a-*fucOB* and pET29a-*fucOB*^{E541A} genes, hereafter) and the catalytic domain of pET29a-*BbAfcA*^{266} were synthesized/sequenced by ATG:biosynthetics. The *fucOB*, *fucOB*^{E541A} and *BbAfcA* were introduced into the pET29a plasmid using the NdeI and HindIII sites. The recombinant FucOB and FucOB$_{E541A}$ (796 residues) have a deletion of the first 23 residues that were predicted as a signal peptide and an additional peptide of 17 amino acids at the N-terminus that includes a histidine tag. pET29a-*fucOB*^{W378A}, pET29a-*fucOB*^{H383A}, pET29a-*fucOB*^{N385A}, pET29a-*fucOB*^{N387A}, pET29a-*fucOB*^{T443A}, pET29a-*fucOB*^{S444A}, pET29a-*fucOB*^{W453A}, pET29a-*fucOB*^{H613A}, pET29a-*fucOB*^{W655A}, pET29a-*fucOB*^{H693A}, pET29a-*fucOB*^{D699A} were generated by QuikChange site-directed mutagenesis.[267] *Bb*AfcA also includes a histidine tag at N-terminus.

5.1.3. Expression and purification of wild-type and single point mutants of FucOB from *Akkermansia muciniphila* strain ATCC BAA-835 and *Bb*AfcA

Escherichia coli BL21 (DE3) cells transformed pET29a-*fucOB,* pET29a-*fucOB*^{E541A}, pET29a-*fucOB*^{W378A}, pET29a-*fucOB*^{H383A}, pET29a-*fucOB*^{N385A}, pET29a-*fucOB*^{N387A}, pET29a-*fucOB*^{T443A}, pET29a-*fucOB*^{S444A}, pET29a-*fucOB*^{W453A}, pET29a-*fucOB*^{H613A}, pET29a-*fucOB*^{W655A}, pET29a-*fucOB*^{H693A}, pET29a-*fucOB*^{D699A} or pET29a-*BbAfcA* were grown in Luria Broth (LB) medium supplemented with 50 μg mL^{-1} of kanamycin at 37 ºC. When the culture reached OD_{600}=0.6, protein expression was induced by adding 1.0 mM isopropyl β-thiogalactopyranoside (IPTG). After 20 h at 18 ºC, cells were harvested at 5,000 × *g* for 20 min at 4 °C and resuspended in 50 mL of 50 mM Tris-HCl pH 7.5, 500 mM NaCl, containing protease inhibitors (Complete EDTA-free, Roche) and 0.5 μL L^{-1} of culture of benzonase (Sigma Aldrich). Cells were then disrupted by sonication in 12 cycles of 10 s pulses, with 60 s cooling intervals between the pulses, and 60% of amplitude and the suspension was centrifuged for 30 min at 59,000 × *g* at 18 ºC. The supernatant was filtered by 0.22 μm pore size Merck Millipore Durapore™ PVDF Membrane Filters and subjected to Ni^{2+} affinity chromatography using a HisTrap Chelating column (5 mL, GE HealthCare) equilibrated in 50 mM Tris–HCl pH 7.5, 500 mM NaCl. Elution was performed with a linear gradient of 0–500 mM imidazole in 300 mL of 50 mM Tris–HCl pH 7.5, 500 mM NaCl at 4 mL min^{-1}. Fractions of interest were pooled, buffer exchanged to 50 mM HEPES pH 7.0 in a 30 kDa cut off centrifugal filter and loaded into a HiTrap SP HP column (1 mL; GE HealthCare), equilibrated in buffer 50 mM HEPES pH 7.0. Elution was performed with a linear gradient of 0–1000 mM NaCl in 30 mL of 50 mM HEPES pH 7.0, at 1 mL min^{-1}. Fractions of interest were pooled and loaded onto a Superdex 75 16/600GL (GE Healthcare) equilibrated in the corresponding buffer according to the experiment to be performed (20 mM Tris-HCl pH 7.5 for crystallization assays and 50 mM Tris-HCl pH 7.5 and 150 mM NaCl for blood group conversion assays). Fractions of interest were pooled and concentrated to 14 mg mL^{-1} and 15 mg mL^{-1},

respectively, in 20 mM Tris-HCl pH 7.5, using a 30 kDa cut off centrifugal filter (Millipore) for crystallization purposes. The resulting preparations displayed a single protein band by SDS-PAGE. Purified FucOB and $FucOB_{E541A}$ were produced at 5.0 and 6.8 mg L^{-1} of growth culture. Purified proteins were stored at -80 °C.

5.1.4. FucOB, *Bb*AfcA and *Bi*Fuc95A substrate specificity assays

120 nmol each of 3-fucosyllactose, Lewis-A trisaccharide, 6-fucosylchitobiose, blood group H antigen triaose Type I, Type II and Type V (2'-fucosyllactose), blood group A antigen tetraose Type V or blood group B antigen tetraose Type V were incubated with FucOB, *Bb*AfcA, and *Bl*Fuc95A, at a molar enzyme to substrate ratio of 1:100 000 in 20 mM Tris-HCl pH 6.8. After 30 min incubation at 37°C, the reactions were stopped by heating to 90°C for 10 min and the amount of released fucose in each reaction was determined using the L-fucose assay kit (Megazyme) according to the manufacturer's instructions. In short, L-fucose dehydrogenase and $NADP^+$ were added to the FucOB, *Bb*AfcA, and *Bi*Fuc95A digested samples and the formation of NADPH during oxidation of L-fucose, stoichiometric with the amount of free L-fucose in the sample, was monitored spectrophotometrically.

5.1.5. FucOB activity assay using *p*-nitrophenyl-α-L-fucose (*p*NP-Fuc)

500 nmol pNP-Fuc was incubated with 40 µg FucOB in 50 µL 20 mM Tris-HCl pH 6.8 at 37°C for 4 h. The reaction was stopped by addition of 0.1% formic acid and product was separated from educt by reverse-phase HPLC and quantified using UV detection at 300 nm.

5.1.6. Glycoengineering of TNFR

1 mg etanercept (TNFR/IgG1 Fc fusion protein) was incubated with 1000u SialEXO and 1000u PNGase F for 4 h at 37 ºC to desialylate the *O*-glycans and remove the *N*-glycans (to simplify the analysis). 8 mM GDP-fucose, 10 mM $MnCl_2$, 10 mM CaCl2, and 3 µg recombinant human Fut2

fucosyltransferase were added and the resulting mixture was incubated at 37 ºC overnight. The resulting fucosylated glycoprotein was purified using CaptureSelect IgG-Fc (multispecies) according to manufacturer's recommendations.

5.1.7. FucOB activity assays by LC-MS

20 µg of fucosylated etanercept was incubated with 1 µg FucOB for 1 h at 37 ºC in a total volume of 40 µL 20 mM Tris-HCl pH 6.8. To simplify the analysis, the TNFR domain carrying the fucosylated *O*-glycans was separated from the Fc region by digestion with FabRICATOR (IdeS) protease that cleaves the IgG Fc fusion protein at one specific site below the hinge. The resulting protein subunits were reduced and denatured by incubation for 1 h at 37 ºC in 4 M guanidine-HCl, 100 mM DTT and analysed by reverse phase LC-MS using a Bruker Impact II ESI-QTOF mass spectrometer.

5.1.8. Activity assay of FucOB mutants

120 nmol each of blood group H antigen triaose type II and V were incubated with FucOB at a molar enzyme to substrate ratio of 1:50 000 in 20 mM Tris-HCl pH 6.8. After 30 min incubation at 37°C, the reactions were stopped by addition of 1:6 (v:v) 1 M Tris-HCl pH 10 and the amount of released fucose in each reaction was determined using the L-fucose assay kit (Megazyme) according to the manufacturer's instructions.

5.1.9. FucOB and FucOB$_{E541A}$ crystallization and data collection

FucOB was crystallized by mixing 0.25 µL of a protein solution at 14 mg mL^{-1} in 20 mM Tris-HCl pH 7.5 with 0.05 µL of seed stock solution and 0.2 µL of 200 mM potassium fluoride, 20% (w/v) PEG 3500 (PEG ION HR2-126 protein crystallization screen, Hampton Research). Seeds stock solution was prepared following Hampton Research protocols and vortexed bead seed stock technique procedures.[268] Crystals grew in 120 days and were cryo-cooled in liquid nitrogen using

200 mM potassium fluoride, 20% (w/v) PEG 3500 and 20% glycerol, as cryo-protectant solution. The $FucOB_{E541A}$ was crystallized by mixing 0.25 µL of a protein solution at 15 mg mL^{-1} and 2.5 mM of A type V blood antigen in 20 mM Tris-HCl pH 7.5 with 0.05 µL of seed stock solution and 0.2 µL of 200 mM sodium chloride, 20% (w/v) PEG 3350 (PEG ION suite protein crystallization screening, Hampton Research). Crystals grew in 15 days and were cryo-cooled in liquid nitrogen using 200 mM sodium chloride 20% (w/v) PEG 3350 and 20% glycerol, as cryo-protectant solution.

Complete X-ray diffraction datasets for FucOB and $FucOB_{E541A}$ were collected at X06DA-PXIII beamline, at the Swiss Light Source, the Paul Scherrer Institute, Switzerland, and BL13-XALOC beamline at ALBA, Cerdanyola del Valles, Spain, respectively. FucOB crystallized in the orthorhombic space group $P\ 2_1\ 2_1\ 2_1$ with one molecule in the asymmetric unit and diffracted to a maximum resolution of 1.8 Å (Table 4). $FucOB_{E541A}$ crystallized in the monoclinic space group $P\ 2_1$ with one molecule in the asymmetric unit and diffracted to a maximum resolution of 1.95 Å (Table 4). All datasets were integrated and scaled with XDS following standard procedures. [226]

Table 4. FucOB and FucOB$_{E541A}$ data collection and refinement statistics.

	FucOB	FucOB$_{E541A}$
PDB code	7ZNZ	7ZO0
Beamline	X06DA-PXIII (SLS)	BL13-XALOC
Wavelength	0.9792	0.97926
Resolution range	47.33 - 1.8 (1.864 - 1.8)	52.24 - 1.95 (2.02 - 1.95)
Space group	P 21 21 21	P 1 21 1
Unit cell	53.67 100.40 156.98 90 90 90	53.58 87.50 95.58 90 102 90
Total reflections	975882 (65783)	423928 (43184)
Unique reflections	79452 (7833)	62656 (6188)
Multiplicity	12.3 (8.4)	6.8 (7.0)
Completeness (%)	99.93 (99.80)	99.84 (99.77)
Mean I/sigma(I)	26.40 (1.84)	12.84 (1.91)
Wilson B-factor	31.84	27.38
R-merge	0.05474 (1.058)	0.1299 (1.223)
R-meas	0.05709 (1.126)	0.1409 (1.323)
R-pim	0.01601 (0.3804)	0.05409 (0.4985)
CC1/2	1 (0.813)	0.998 (0.648)
CC*	1 (0.947)	0.999 (0.887)
Reflections used in refinement	79417 (7830)	62627 (6188)
Reflections used for R-free	3972 (392)	2951 (295)
R-work	0.1899 (0.3396)	0.2033 (0.2901)
R-free	0.2177 (0.3568)	0.2142 (0.3208)
CC(work)	0.954 (0.875)	0.957 (0.744)
CC(free)	0.899 (0.861)	0.948 (0.708)
Number of non-hydrogen atoms	6317	6012
macromolecules	5871	5730
ligands	56	30
Protein residues	422	761
RMS(bonds)	0.012	0.004
RMS(angles)	1.18	0.62
Ramachandran favored (%)	96.71	96.97
Ramachandran allowed (%)	3.16	2.90
Ramachandran outliers (%)	0.13	0.13
Rotamer outliers (%)	0.84	0.89
Clashscore	4.06	3.94
Average B-factor	43.18	29.86
macromolecules	42.99	29.55
ligands	50.81	40.02
solvent	45.41	35.69

5.1.10. FucOB and FucOB$_{E541A}$ structure determination and refinement

The structure determination of FucOB and FucOB$_{E541A}$ were carried out by molecular replacement methods implemented in Phaser and the PHENIX suite using the PDB code 2EAB as a search template.[232,233] Initial cycles of model building, density modifications, and refinement by Buccaneer[234] and the CCP4 suite.[231] The final manual building was performed with Coot[234] and refinement with phenix refine.[235] The structures were validated by MolProbity. [236] Data collection and refinement statistics are presented in Table 4. The atomic coordinates and structure factors

were deposited with the PDB, accession codes are 7ZNZ and 7ZO0. Molecular graphics and structural analyses were performed with the UCSF Chimera package.[237]

5.1.11. Structural analysis and sequence alignment

Structure-based sequence alignment analysis was performed using Chimera.[237] Protein pocket volume was calculated using HOLLOW.[238] Z-score values were produced by using DALI.[239] Domain interface analysis was performed using PISA.[240] Conserved and similar residues were labelled using Multiple Align Show server (https://www.bioinformatics.org/SMS/multi_align.html).

5.1.12. Molecular docking calculations

H, A and B antigens were modelled using GLYCAM-Web website (Complex Carbohydrate Research Center, University of Georgia, Athens, GA; http://www.glycam.com). Ligand docking was performed using AutoDock Vina employing standard parameters and visualized using USCF Chimera.[237,241] The active site of FucOB was defined taking into account the crystal structure of the homologous α-1,2-fucosidase *Bb*AfcA in complex with the substrate 2'-fucosyllactose (2'FL; Fucα1–2Galβ1–4Glc; PDB code 2EAD).

5.1.13. Molecular dynamics (MD) simulations

We used the *in silico* molecular docking structures shown in Figure 29d,e,f as the initial structures for the three complexed studied by MD simulations. The simulations were carried out with AMBER 20 package[269] implemented with ff14SB[270] and GLYCAM06j[271] force fields. The system was neutralized by adding explicit counter ions. Each complex was immersed in a water box with a 10 Å buffer of TIP3P water molecules.[272] A two-stage geometry optimization approach was performed. The first stage minimizes only the positions of solvent molecules, and the second stage is an unrestrained minimization of all the atoms in the simulation cell. The systems were then

gently heated by incrementing the temperature from 0 to 300 K under the constant pressure of 1 atm and periodic boundary conditions. Harmonic restraints of 30 kcal mol^{-1} were applied to the solute, and the Andersen temperature coupling scheme was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Long-range electrostatic effects were modelled using the particle-mesh-Ewald method.[273] An 8 Å cut-off was applied to Lennard-Jones interactions. Each system was equilibrated for 2 ns with a 2-fs time step at a constant volume and temperature of 300 K. Production trajectories were then run for additional 0.5 μs under the same simulation conditions.

5.1.14. Blood samples extraction, collection, and storage

Samples of RBCs concentrates from anonymous healthy donors were obtained from blood bag tubing segments at the Blood Bank of Cruces University Hospital. The samples containing standard preserving CPD-SAGM media[274] were stored at 4 ºC until use up to 35 days. RBCs from non-treated blood were collected the same day of the experiment from two healthy consenting donors into a citrate Vacutainer using a protocol approved by the Ethics Committee for Clinical Research of Cruces University Hospital (CEI E21/65) in accordance with the Spanish Law and the Declaration of Helsinki.

5.1.15. Enzymatic conversion of universal O into rare Bombay type blood group assay

To analyse the enzymatic conversion of universal O into rare Bombay type 2 mL of healthy donors' blood samples were first diluted with 18 mL of PBS to reach a 4% red cells suspension and washed twice with PBS (centrifuged at 5000 × *g* for 5 min at 4 ºC). After removing the supernatant, 18 mL of PBS were added and the suspension separated in 18 aliquots of 1 mL final volume. The samples were centrifuged at 200 × *g* for 5 min at RT. After removing the supernatant, the needed amount of enzyme was added over 200 μl final sample volume, to reach a final concentration of 200, 50, 5.0, 0.5, 0.05, and 0.005 μl mL^{-1} of FucOB or $FucOB_{E541A}$. The mixtures

were maintained in an incubator OPAQ I10-E and the orbital MaXI shaker OL30-ME (OVAN) at a constant stirring of 110 rpm for 30 min at 37 ºC. The cells were then washed twice with 1 mL of PBS. Finally, each sample was diluted in 200 µL of PBS per sample.

5.1.16. DG Gel column agglutination assay

DG Gel Neutral and DG Gel Coombs cards from Grifols (Diagnostic Grifols, S.A.) were used for blood group typing and agglutination assays based on the gel technique described in 1985 by Ives Lapierre.[275] We followed the manufacturer's recommendations. 10 µL of RBC sample was diluted in 1 mL of Grifols Diluent. 25 µL of Bombay serum was added to the selected DG Gel wells and 50 µL of previously diluted RBC sample was then added. The DG Gel Neutral cards were incubated for 15 min at RT, whereas DG Gel Coombs cards were incubated for 15 min at 37 ºC, and later centrifuged in DG SPIN centrifuge for 9 min at RT. RBCs that underwent agglutination with anti-H antibody present in Bombay serum were evaluated following the manufacturer's instructions. Briefly, presence of RBCs the pellet in the bottom of the gel-column indicates no agglutination (negative result) neither haemolysis in the sample. On the other hand, clumps of RBCs throughout the gel column indicate cells agglutinated in the sample (positive result). Each agglutination card has 8 buffered tubes to perform the experiments. Therefore, the results displayed in the corresponding Figures are shown in patches (Figure 38c and Figure 39).

5.1.17. Anti-H lectin agglutination assay

1 mL of two healthy O negative blood group donor's samples and 1 mL of one healthy B positive blood group donor sample were first washed twice with 9 mL of PBS and centrifuge at 5000 × g for 5 min at 4 ºC. The washed samples were diluted to 500 µL of 4% of red cells suspension. 5 µL of the needed amount of FucOB, $FucOB_{E541A}$ or PBS (as a control) was added over the red cells suspension to reach a final concentration of 5.0 µg mL^{-1} of enzyme. The mixtures were incubated for 10 min at 37 ºC. 25 µL of each sample was added over 50 µL of anti-H lectin reagent or PBS

(as a control) and results were read macroscopically. The presence of RBCs pellet in the bottom indicates no agglutination (negative result) neither haemolysis in the sample. In contrast, a reddish solution of RBCs indicates cells agglutinated in the sample (positive result).

5.1.18. Blood Smears

To visualize the RBC physical preservation after conversion, treated and control samples were observed by May-Grünwald-Giemsa stained smears.[276] Specifically, 200 μL of O negative blood samples were directly incubated with 50 μg mL^{-1} of active enzyme in an incubator OPAQ I10-E and the orbital MaXI shaker OL30-ME (OVAN) at constant stirring of 110 rpm for 30 min at 37 ºC. One drop of enzymatically converted blood sample was later spread in a glass slide and fixed dipping it in absolute methanol for three minutes. Then, equal volume of stain solution 1 (0.3 g May-Grünwald powder in 100 mL absolute methanol) was freshly mixed with buffer solution at pH 6.8 of 6.63 g KH_2PO_4, 2.56 g $Na_2HPO_4{\cdot}2H_2O$ and distilled water up to 1000 mL. The mixture was applied over the fixed sample for 5 minutes horizontally positioned. A dilution of stain solution 2 (1 g Giemsa stain powder dissolved in 66 mL glycerol, and heated to 56 ºC for 120 min to later add 66 mL of absolute methanol) in the same buffer solution (1:9; v/v) was added over the sample for 15 min. Finally, the stained sample was washed with water and visualized with an optical microscope.

5.1.19. Glucose-6-phosphate dehydrogenase assay

To perform glucose-6-phosphate dehydrogenase (G6PD) activity assay 200 μL of O negative blood sample was directly incubated with 50 μg mL^{-1} of active enzyme in an incubator OPAQ I10-E and the orbital MaXI shaker OL30-ME (OVAN) at constant stirring of 110 rpm incubator for 30 min at 37 ºC. Trinity Biotech Glucose-6-Phosphate Dehydrogenase reagents were used to perform G6PD deficiency assay in converted blood samples following manufacturer's instructions.[277] 0.2 mL of G6PD substrate solution (Glucose-6-Phosphate (4 μmol), NADP (1.6

μmol), Glutathione, oxidized (1.6 μmol), and lytic agent in 2 mL volume) were incubated at 37 ºC with 0.01 mL blood sample.

5.1.20. Flow cytometry studies

For FACS analysis, enzymatically treated O type RBCs were diluted (1:10) in PBS and then 1 μL of diluted blood was added to 100 μL of staining buffer (PBS + 1% fetal bovine serum). Next, cells were incubated with (1:100) mouse anti-blood group H ab antigen antibody (97-I) (from Abcam), (1:10) mouse anti-blood group H n/ab antigen antibody (86-M) (from Creative Biolabs) or with (1:10) mouse IgM Isotype control (B11/7) (from Abcam) for 30 minutes on ice. Then, cells were washed twice with staining buffer and incubated with (1:100) PE goat anti-mouse IgM antibody from Invitrogen for 30 minutes at RT. For the staining with lectins, cells were incubated with (1:1000) FITC-conjugated anti-H lectin (L32476) for 15 minutes at RT. Then, cells were washed twice with staining buffer and incubated with BV421 mouse anti-CD235a (GA-R2) from BD Bioscience for 30 minutes on ice. Lastly, samples were washed once and resuspended in staining buffer and acquired in a MACSQuant Analyzer 10 flow cytometer (Miltenyi Biotec). Flow cytometry data were analysed using FlowJoTM v10.4. The determination of the population positive for antigen H was based in the isotype control.

5.1.21. Samples and ethics statement

Tubular segments attached to RBCs concentrate units from anonymous healthy donors were obtained from the Blood Bank of Cruces University Hospital. The study was approved by the Ethics Committee for Clinical Research of Cruces University Hospital (CEI E21/65) in accordance with Spanish Law and the Declaration of Helsinki.

5.2. RESULTS AND DISCUSSION

5.2.1. The architecture of full-length FucOB

To obtain the crystal structure of the unliganded form of full-length FucOB, we purified the enzyme to apparent homogeneity (see Methods section for details). FucOB from *A. muciniphila* comprises 796 residues (UniProt code B2UR61; GeneBank code ACD04946.1) with a predicted signal peptide (residues 1-23) that was removed from the construct. The structure was determined at 1.8 Å resolution using molecular replacement methods (Table 4; see Methods section). The high quality of the electron density maps allowed the trace of residues 25-785 (Figure 30). A close inspection of the crystal structure revealed that FucOB crystallized as a monomer comprising three domains from the N- to the C-terminus: (i) a β-sandwich domain (residues 25-260), (ii) an $(\alpha/\alpha)_6$ helical barrel catalytic domain (residues 357-712) and (iii) a second β-sandwich domain (residues 713-785). It is worth noting that the N-terminal β-sandwich domain is connected to the catalytic helical barrel domain through a linker comprised of five α-helices (α3, α4, α6, α7 and a short α5; residues 261-356). The first β-sheet of the N-terminal β-sandwich domain consists of nine β-strands with topology β1–β6–β7–β8–β9–β16–β13–β12–β11 (β6, β8, β16, β12 are antiparallel), whereas the second β-sheet comprises seven β-strands with topology β2–β3–β4–β5–β10–β15–β14 (β2–β4–β10–β14 are antiparallel), with an overall size of 41 Å × 40 Å × 11 Å. The central catalytic core $(\alpha/\alpha)_6$ helical barrel domain consists of 12 α-helices (α8 to α19; Figure 30a,b,c). The C-terminal β-sandwich domain consists of a β-sheet of five β-strands with topology β17–β20–β21–β22–β23 (β20–β22–β23 are antiparallel), whereas the second β-sheet comprises eight β-strands with topology β24–β25 (β25 is antiparallel), with an overall size of 27 Å × 17 Å × 10 Å (Figure 30).

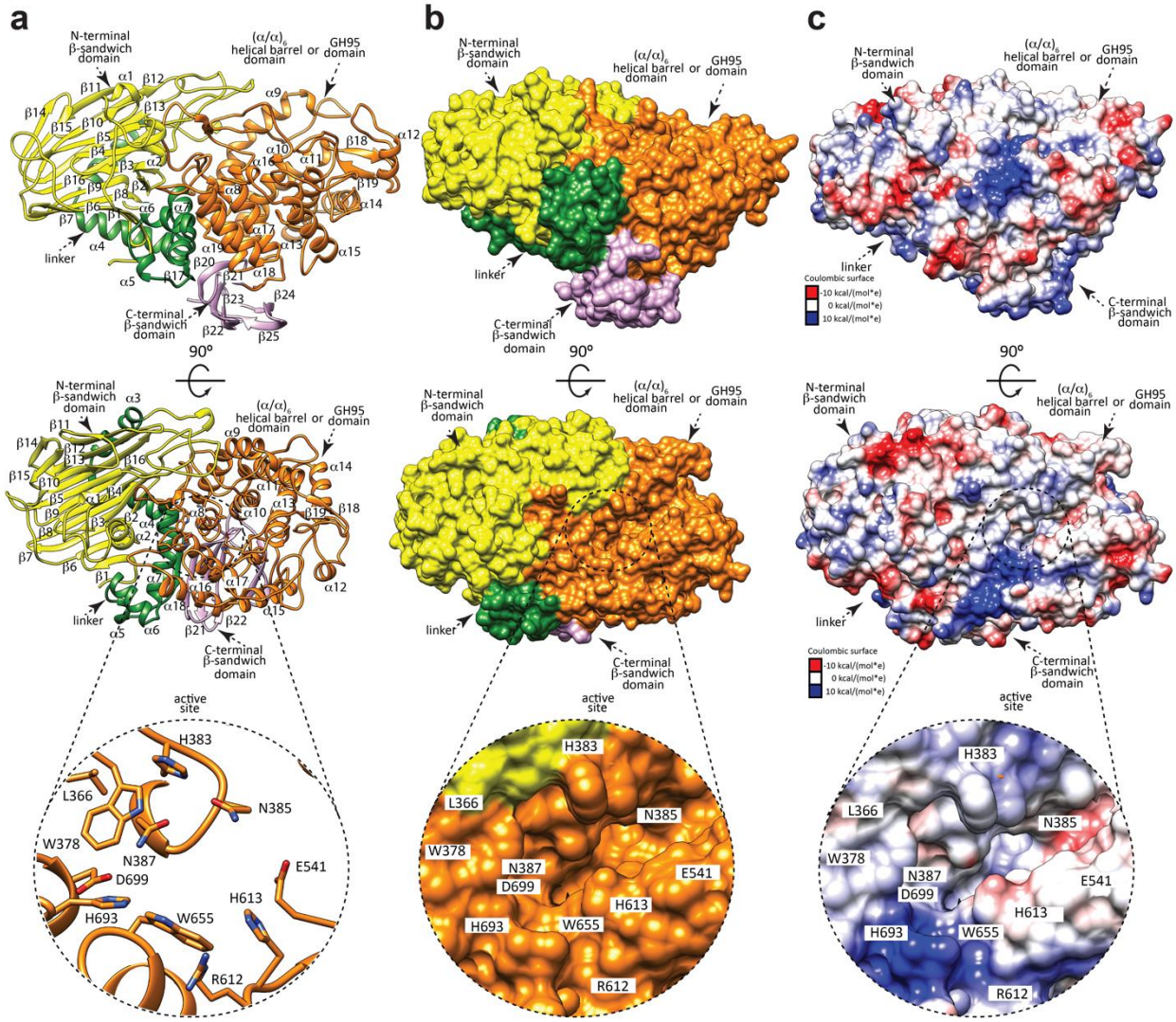

Figure 30: The overall structure of FucOB. Two views of the cartoon representation showing the general fold and secondary structure organization of FucOB, including the catalytic $(\alpha/\alpha)_6$ helical barrel domain or GH95 catalytic domain (orange) and the N- and C-terminal β-sandwich domains (yellow and pink, respectively) and α-helices linker (in green). The bottom panel corresponds to a close-up view of the active site, shown as a cartoon/stick representation. **b** Two views of the surface representation of FucOB. The bottom panel represents a close-up view of the active site of FucOB showing the substrate binding pocket. **c** Two views of the electrostatic surface representation of FucOB showing the location of the putative substrate binding site and the catalytic site. The bottom panel shows a close-up view of the active site of FucOB showing the substrate binding pocket. Figure edited from Anso *et al*. (2023).[88]

The $(\alpha/\alpha)_6$ helical barrel of FucOB displays a deep groove where the active site is located, mainly flanked by solvent-exposed and flexible loops, including loop 25 (α7−α8; residues 357-386), loop 28 (α9−α10; residues 429-458), loop 33 (β19-α13; residues 547-550), loop 35 (α14−α15; residues 587-614); loop 37 (α16−α17; residues 651-655), loop 39 (α18−α19; residues 683-699) and α-helix 8 (residues 387-392. A first glycerol molecule is deeply buried into a pocket defined by W378, H383, W655, and H693. The O1 and O2 atoms make hydrogen bonds with the side chains of N387

and Q697, respectively. A second glycerol molecule accommodates into an adjacent solvent exposed pocket mainly defined by Y106 and W453. The O3 and O2 atoms make hydrogen bonds with the side chains of S444 and H383, respectively. The two glycerol molecules are involved in additional hydrogen-bonding interactions with a few water molecules. A third glycerol molecule is located near the first glycerol molecule. The O1, O2 and O3 atoms interact with the side chains of H693, R612 and H613, respectively (third row panels of Figure 30)

According to DALI server, FucOB shows similar structural features to GH95 family, characterized as α-L-fucosidase (EC 3.2.1.51); α-1,2-L-fucosidase (EC 3.2.1.63); α-L-galactosidase (EC 3.2.1.-). Enzymes that belong to this family of GHs follow a single displacement inverting catalytic mechanism.[278] A conserved glutamic acid (E541 in FucOB) acts as a general acid catalyst. Mutation of E541 by alanine completely abolished the hydrolytic activity of the enzyme (see activity measurements in methods). There is no carboxylic acid residue at the appropriate position for a general base catalyst. Therefore, it is proposed that a water molecule acts as a general base in the reaction, activated by two asparagine residues (N385 and N387 in FucOB) and an aspartic acid residue (D699 in FucOB) to perform the nucleophilic attack to the fucose (Figure 31).[266,278]

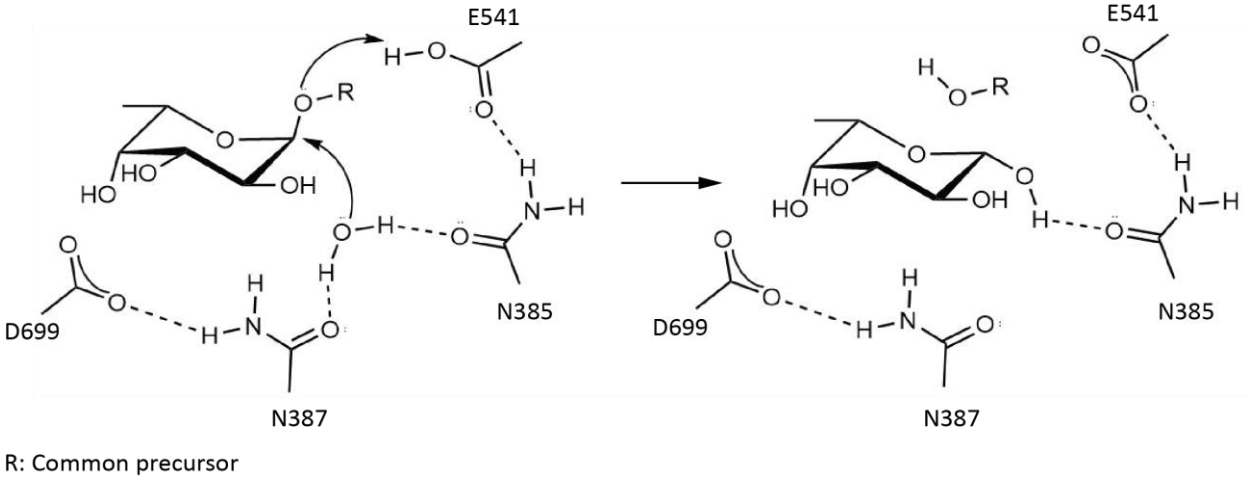

Figure 31: The catalytic mechanism of FucOB. FucOB follows a single displacement inverting catalytic mechanism, as proposed for *Bb*AfcA and other GH95 family members. A conserved glutamic acid (E541) acts as a general acid catalyst. A water molecule acts as a general base in the reaction, activated by two asparagine residues (N385 and N387) and an aspartic acid residue (D699) to perform the nucleophilic attack to the fucose. Figure edited from Anso *et al*. (2023).[88]

The structural homologs that DALI server found to date are, as mentioned above, the following GH95 family of enzymes: (i) α-1,2-fucosidase *Xac*Afc95 from *Xanthomonas citri* (PDB code 7KMQ; Z-score of 42; r.m.s.d. value of 1.8 Å for 679 aligned residues; 35% identity)[279], (ii) a putative GH95 member from *Bacillus halodurans* (PDB code 2RDY; Z-score of 41.3; r.m.s.d. value of 1.9 Å for 675 aligned residues, 36% identity), (iii) α-1,2-fucosidase *Bb*AfcA from *B. bifidum* (PDB code 2EAB; Z-score of 41.3; r.m.s.d. value of 2.2 Å for 698 aligned residues; 33% identity)[266] and (iv) α-L-galactosidase BACOVA_03438 from *Bacteroides ovatus* (PDB code 4UFC; Z-score of 40.5; r.m.s.d. value of 1.8 Å for 677 aligned residues; 35% identity)[280] (Figure 32). Interestingly, the genome of *A. muciniphila* strain ATCC BAA-835 encodes solely one additional enzyme that belongs to the GH95 family, Amuc_0186,[281] which shares 29% sequence identity with FucOB.

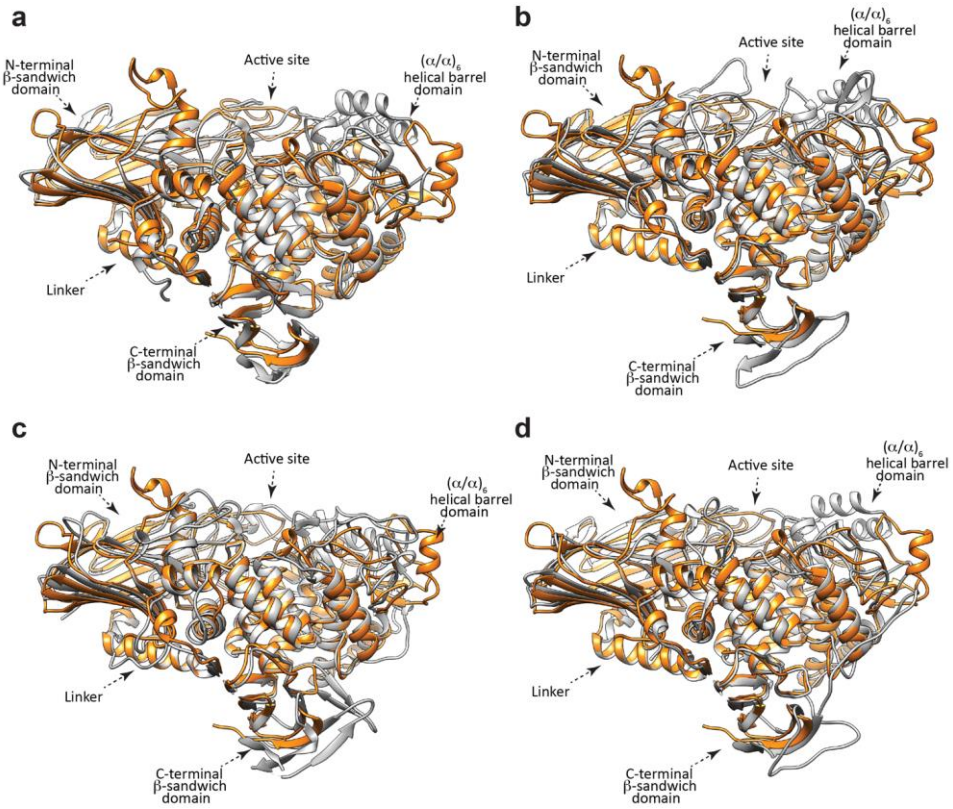

Figure 32: Structural homologues of FucOB. Structural superposition of the X-ray crystal structure of FucOB and structural homologues (grey): **a** α-1,2-fucosidase *Xac*Afc95 from *Xanthomonas citri* (PDB code 7KMQ), **b** a putative GH95 member from *Bacillus halodurans* (PDB code 2RDY), **c** α-1,2-fucosidase *Bb*AfcA from *Bifidobacterium bifidum* (PDB code 2EAB), **d** α-L-galactosidase BACOVA_03438 from *Bacteroides ovatus* (PDB code 4UFC). Figure edited from Anso *et al*. (2023).[88]

5.2.2. FucOB is an α-1,2-fucosidase that specifically cleaves H antigen

To study the enzymatic activity and substrate specificity of FucOB we incubated the enzyme with a series of fucosylated oligosaccharides and quantified the amount of released fucose *in vitro*. Although more than one GH activity has been reported for this family, FucOB was very specific to L-fucose, which is equivalent to 6-deoxy-L-galactose (Figure 33a). FucOB displays α-1,2-L-fucosidase activity and shows no activity against α-1,3, α-1,4 and α-1,6 fucosylated oligosaccharides (Figure 33a).

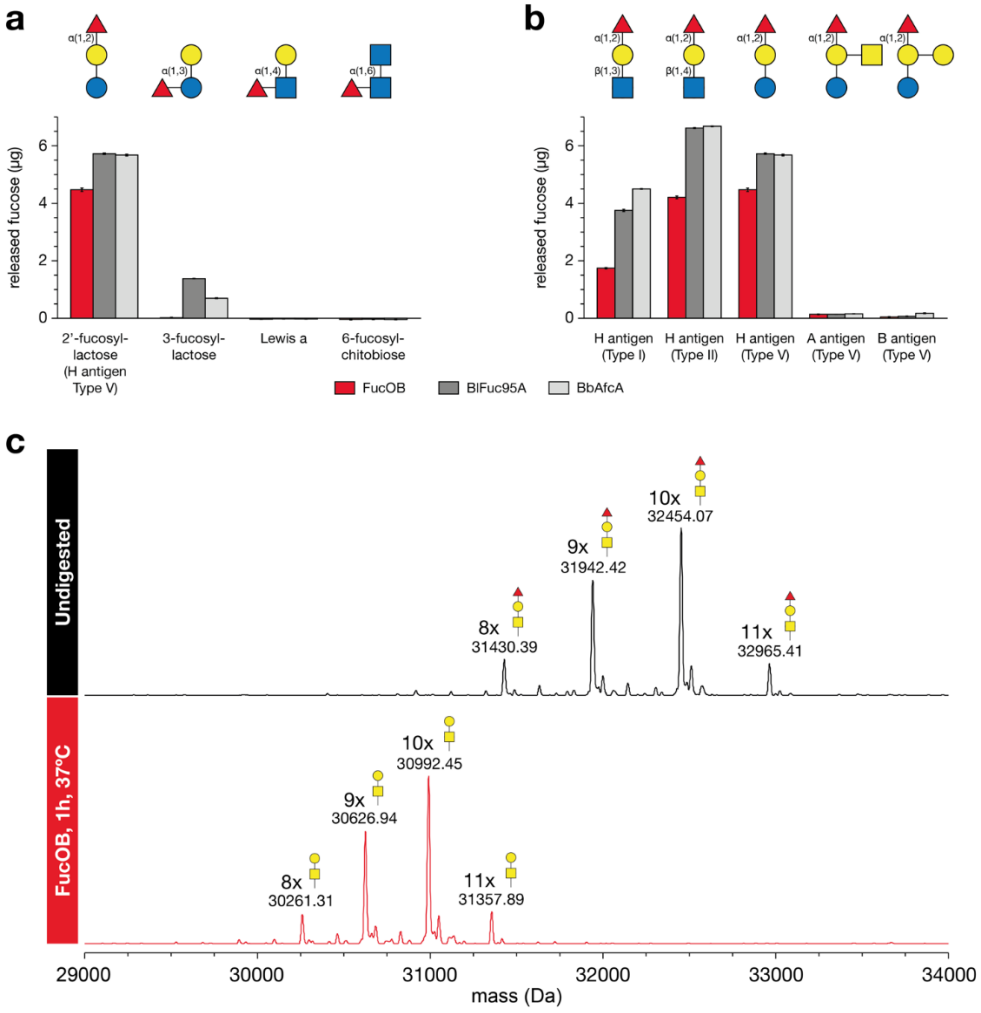

Figure 33: Substrate Specificity of FucOB. a,b Quantification of released fucose after incubation of synthetic oligosaccharides with FucOB, *Bi*Fuc95A and *Bb*AfcA for 30 min at 37 ºC. Free fucose levels were quantified spectrophotometrically and compared to a standard curve of known concentrations. Bars and error bars represent the mean and standard deviation of triplicate measurements. **c** Deconvoluted mass spectra of the TNFR fragment of glycoengineered etanercept before (black) and after (red) incubation with FucOB for 1 hour at 37 ºC. The major peaks are annotated with average mass and glycan composition. Figure edited from Anso *et al*. (2023).[88]

It is worth noting that the Fucα1-2Galβ epitope is part of the oligosaccharide anchor of blood group A, B and H antigens, present in many human tissue surfaces such as RBC or gastrointestinal epithelium. We therefore evaluated the ability of FucOB to process the α-1,2 linked fucose residue on synthetic oligosaccharides representing Type I, Type II, Type V H antigens, Type V A antigen and Type V B antigen (Figure 33b). Interestingly, we found that FucOB hydrolyses all three types of H antigen structures to obtain the afucosylated Bombay phenotype and very poorly hydrolyse the α-1,2 linked fucose in the branched structures of the A or B antigens (Figure 33b; see Methods section). We further compared the enzymatic activity and substrate specificity of two FucOB homologues from family GH95, *Bb*AfcA from *Bifidobacterium bifidum* (29% identity with FucOB) and Fucosidase 95A from *Bifidobacterium longum* CZ0511 (*Bi*Fuc95A; 29% identity with FucOB; NZYTech). As depicted in Figure 4a,b, both *Bb*AfcA and *Bi*Fuc95A are a bit faster than FucOB in cleaving α-1,2 fucose but clearly less specific with activity also towards α-1,3 fucose as well. These results highlight the value of FucOB, a very specific enzyme for α-1,2-L-fucosylated substrates and showing no activity against α-1,3, α-1,4 and α-1,6 fucosylated substrates. Lastly, we assessed the ability of FucOB to hydrolyse fucose not just from synthetic oligosaccharides but also in the context of a larger glycoconjugate such as a glycoprotein. To this end, we glycoengineered the TNFα receptor to carry 8-11 α-1,2-fucosylated core 1 *O*-glycans, corresponding to Type III H antigen structures and used this as a model substrate to assay FucOB activity. After 1 hour of incubation, the substrate was completely defucosylated by FucOB as demonstrated by reverse phase LC-MS (Figure 33c).

5.2.3. Structural basis of H antigen recognition and specificity by FucOB

To further understand the FucOB substrate specificity at the molecular level, we performed co-crystallization experiments with (i) wild-type FucOB and (ii) the catalytically inactive variant $FucOB_{E541A}$, both in the presence of V Type H blood group antigen as the substrate. Despite much effort, we could not crystallize FucOB or $FucOB_{E541A}$ in the presence of the substrate or the

corresponding product. The structural comparison of FucOB and the $FucOB_{E541A}$ revealed that the protein structure is mostly preserved and that there are no substantial conformational changes (r.m.s.d. of 0.477 Å for 761 residues). The crystal packing analysis of FucOB and $FucOB_{E541A}$ structures reveals a strong π–π stacking interaction between W453 with two prolines residues, P765 and P783, of the neighbor protomer. Consequently, this protomer restricts the entrance of the substrate into the active site, supporting the inability to obtain a complex crystal form.

To describe the architecture of the H type blood group epitope binding site, we thus generated a three-dimensional model of FucOB in complex with H, A and B antigens by *in silico* molecular docking calculations. To this end, we defined the putative FucOB substrate binding site considering the crystal structure of *Bb*AfcA in complex with the substrate 2'-fucosyllactose (2'FL; Fucα1–2Galβ1–4Glc; PDB code 2EAD). The H antigen's O4 and O5 atoms of the galactose ring make hydrogen bonds with W378, whereas O3 and O4 atoms make hydrogen bonds with H383, both residues located in loop 25. In contrast, the fucose residue in H antigen is more deeply buried into the active site of FucOB. The O4 atom makes hydrogen bond with H693 in loop 39; O5 makes hydrogen bond with N387 located in α-helix 8; whereas O4 and O3 atoms make hydrogen bonds with W655 in loop 37 (Figure 34a,b,c,d). Conversely, molecular docking calculations for the A type blood group epitope containing the GalNAcα1-3(Fucα1-2)Gal oligosaccharide show that the GalNAc residue exhibits significant clashes with the protein (Figure 34e). Specifically, GalNAc residue's ring shows clashes, including C1 and C2 with T443 located in loop 28, and C2, C3, C4, O3 and O4 with N385 located in loop 25. Moreover, the acetyl group clashes with T443 and S444 in loop 28, and H383 in loop 25. Furthermore, molecular docking calculations for the B type blood group epitope, containing the Galα1-3(Fucα1-2)Gal oligosaccharide, show similar significant clashes of the Gal residue with the protein as observed with the A antigen (Figure 34f).

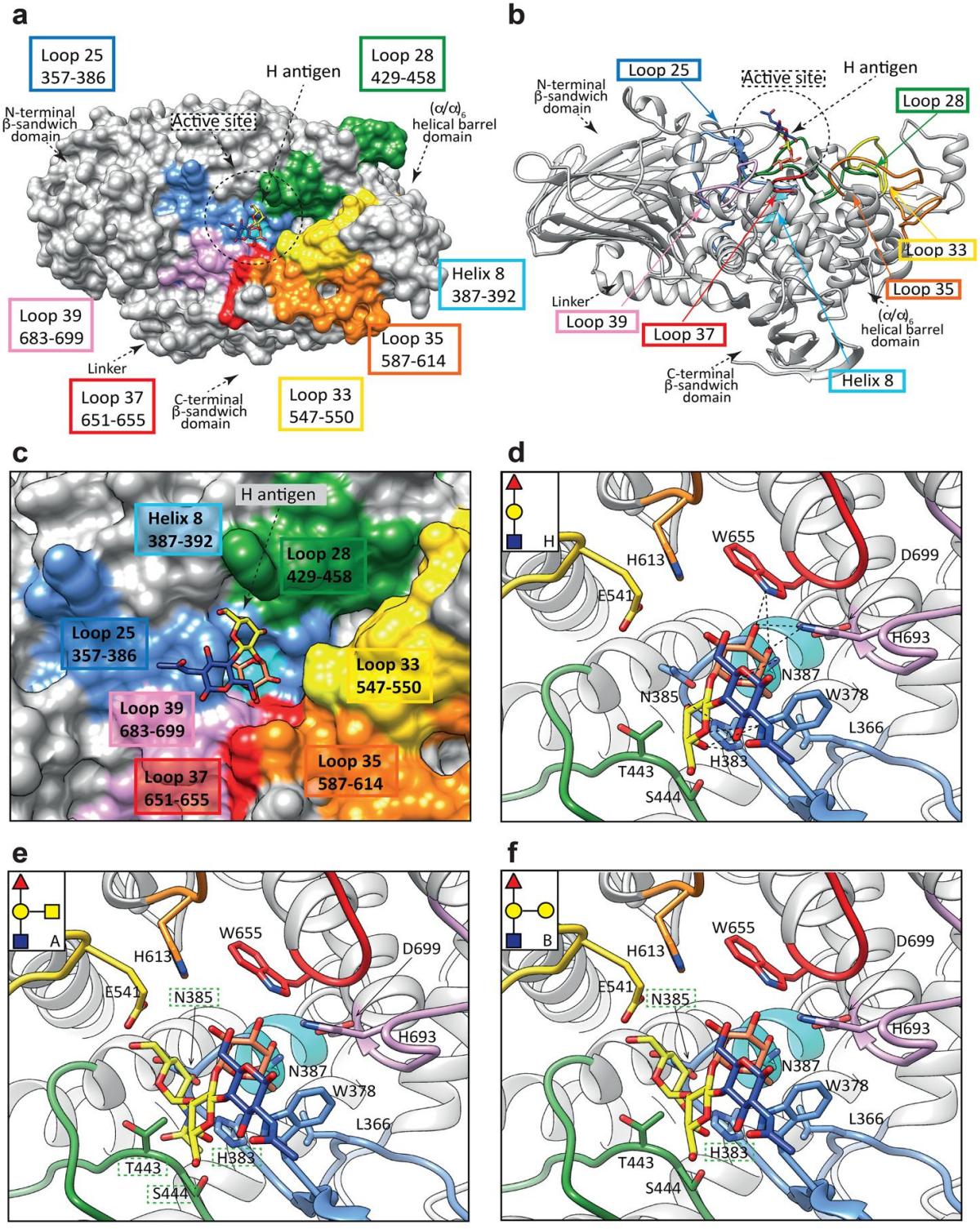

Figure 34: Structural basis of FucOB specificity for H type blood antigen. a Surface representation of the FucOB structure, with annotated domains and loops, showing the location of the Type II H antigen substrate. **b** Cartoon representation of the FucOB structure, showing the location of the Type II H antigen substrate. **c** Surface representation of FucOB showing the location of the Type II H antigen into the active site. Docking calculations of FucOB with different blood group antigens including **d** Type II H, **e** Type II A and **f** Type II B. The predicted clashes are shown as green dotted squares and H bonds are shown as black dotted lines. Figure edited from Anso *et al*. (2023).[88]

To further support these findings at the molecular level, we first performed 0.5 μs molecular dynamics (MD) simulations of the FucOB in complex with H, A, and B type antigens obtained by *in silico* molecular docking calculations. The relative movement between the ligands and the enzyme was analyzed by monitoring the distance between the center of the aromatic ring of W378 and the methyl group of the fucose residue. The MD simulations clearly show that the type H antigen is stable and maintains this interaction throughout the trajectory. Several hydrogen bonds between the antigen and FucOB are also formed. In contrast, the type A and B antigens explore other areas of FucOB or almost detach from the enzyme.

We then performed single-point alanine mutations of residues in the loops that decorate the β-barrel core of the enzyme and contact the H antigen substrate (Figure 35a,b). We studied the ability of 11 single-point mutants to process the fucose residue on two substrates, Type II and Type V H antigens (Figure 35c. Specifically, we individually mutated key residues in loop 25 (W378A, H383A, N385A), loop 28 (T443A, S444A, W453A), loop 35 (H613A); loop 37 (W655A), loop 39 (H693A and D699A) and α-helix 8 (N387A), in addition to the catalytically inactive E541A.

Collectively, the mutational analysis of the FucOB loops that contact the fucose residue (loops 25, 35, 37, and 39) and N387A located in α-helix 8, indicated they were critical for Type II and Type V H antigens recognition (Figure 35c). Interestingly, the mutations T443A and S444A were nearly dispensable for the enzymatic activity, while W453A resulted in a very significant reduction of the FucOB activity (Figure 35c). It is worth mentioning that residues T443, S444 and W453 of loop 28 contact the galactose ring. The side chains of T443 and S444 make hydrogen bonds with O3 and O4 atoms of galactose, while W453 stabilizes the sugar ring by a stacking interaction which, in light of the experimental data, seems to play an important role in FucOB substrate binding (Figure 35).

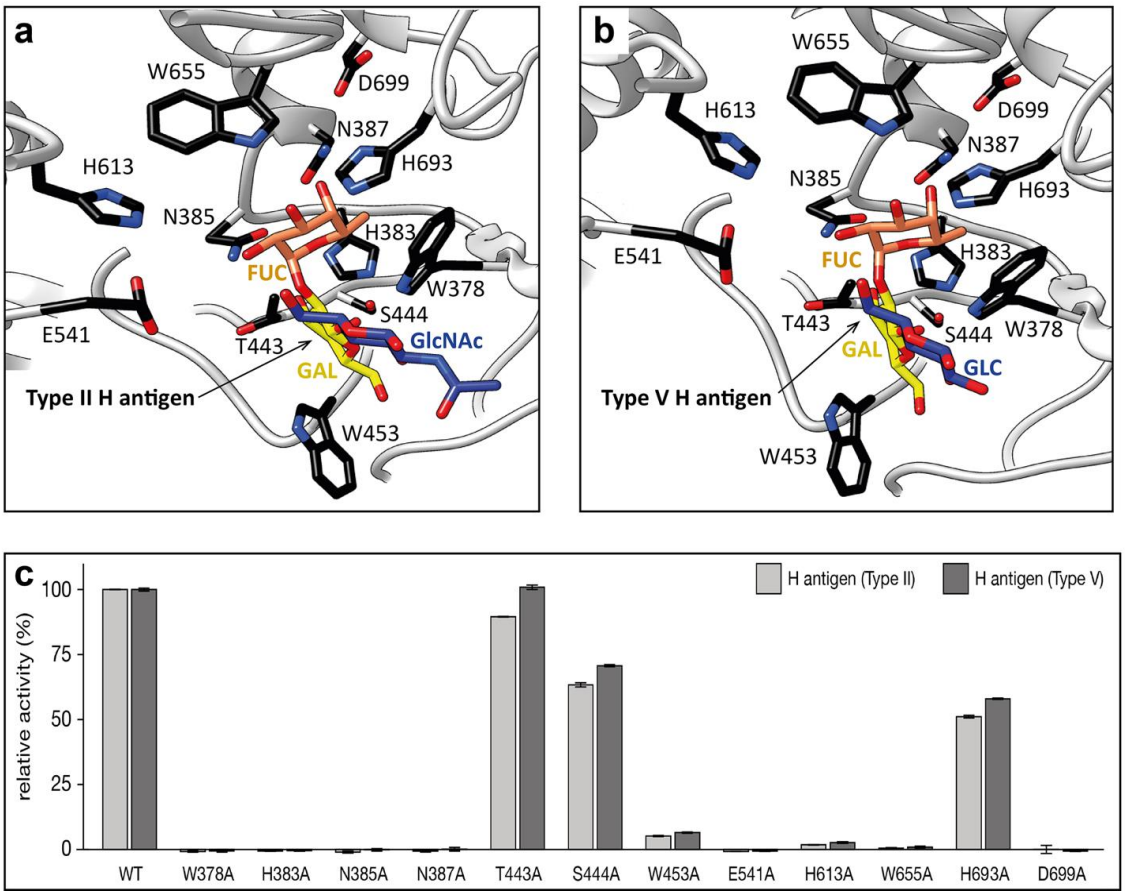

Figure 35: Alanine scanning mutagenesis of the FucOB-H antigen interface. a,b Structure of the FucOB (grey) complex with Type II H-antigen (Fuc in orange; Gal in yellow; GlcNAc in blue; panel a) and Type V H-antigen (Fuc in orange; Gal in yellow; Glc in blue; panel b) in which the amino acids mutated by alanine are in black. **c** Amount of fucose released by FucOB and FucOB mutants, from Type II H-antigen and Type V H-antigens. All values are shown relative to the wild-type enzyme. All enzymatic activity measurements were determined in triplicates. Figure edited from Anso *et al*. (2023).[88]

A detailed comparison of FucOB structures with that of the only four GH95 members for which experimental structural information is currently known provides additional deep insights into the molecular mechanism of H type blood group epitope recognition and specificity. The three-domain architecture is essentially preserved between the five GH95 members, with FucOB displaying the largest substrate binding groove (Figure 32). However, the structural variability of the loops that decorate the active site might account for the differences in enzymatic activity and substrate specificity. In that sense, the available structural information of enzyme-ligands complexes is also limited to just four: α-L-galactosidase BACOVA_03438 in the presence of β-L-Gal (PDB code 4UFC), and α-1,2-fucosidase *Bb*AfcA in the presence of the substrate 2'FL (PDB

code 2EAD), the products α-L-Fuc and Galβ1–4Glc (PDB code 2EAE) and the deoxyfuconojirimycin inhibitor (DFJ; PDB code 2EAC) (Figure 36).

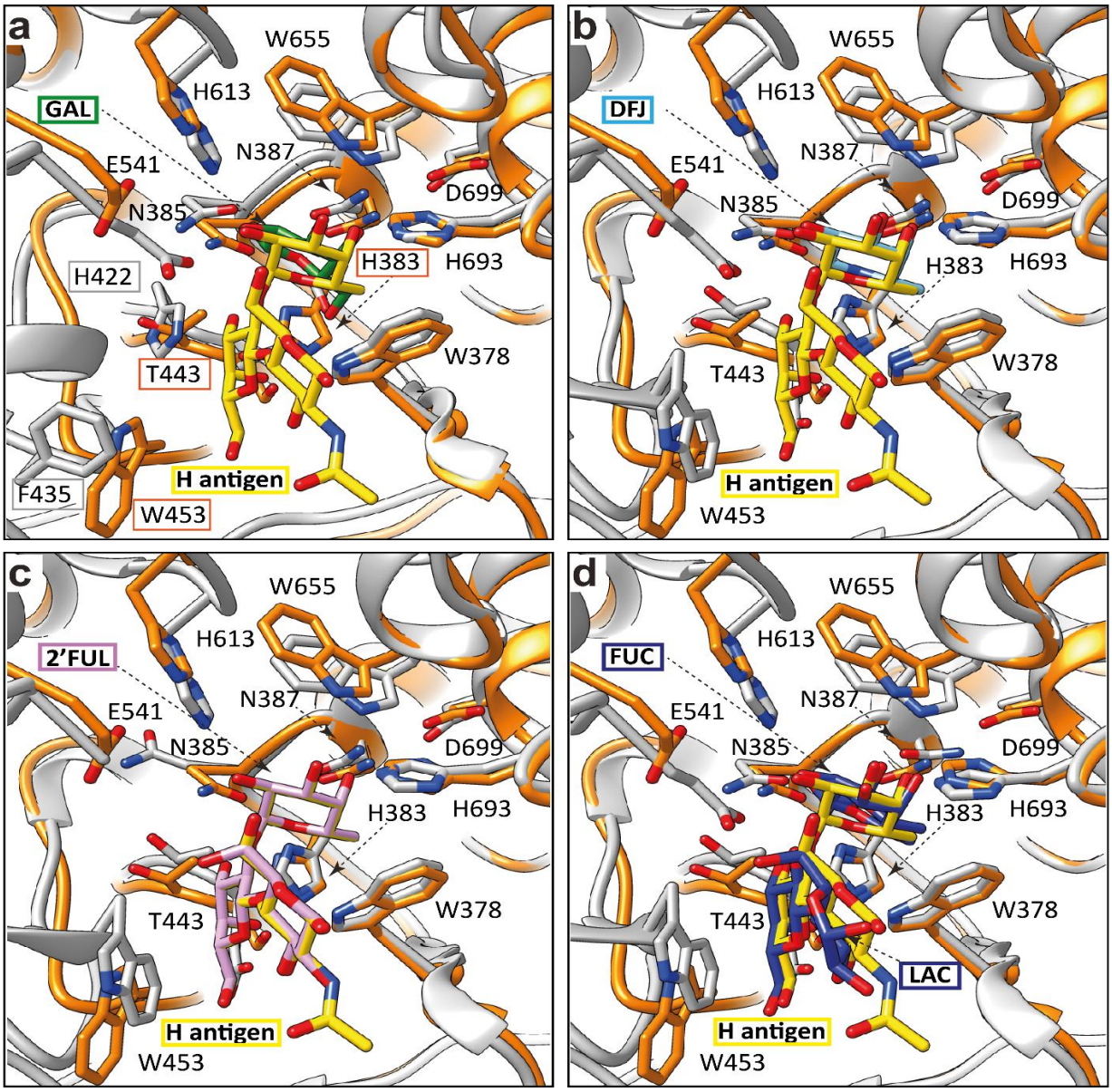

Figure 36: Structural comparison of FucOB with GH95 homologues in complex with the ligands. Binding site comparison between the FucOB (in orange) in complex with the H Type II antigen obtained by *in silico* molecular docking calculations (in yellow) with that of crystal structures of GH95 homologues in complex with ligands (in grey). FucOB conserved residues that participate in binding are highlighted. **a** α-L-galactosidase BACOVA_03438 from *Bacteroides ovatus* in complex with galactose (PDB code 4UFC; Gal residue in green), **b** α-1,2-fucosidase *Bb*AfcA from *Bifidobacterium bifidum* in complex with DFJ inhibitor (PDB code 2EAC; DFJ in light blue), **c** *Bb*AfcA from *Bifidobacterium bifidum* in complex with the substrate 2'FL (PDB code 2EAD; 2'FL in pink) **d** α-1,2-fucosidase *Bb*AfcA from *Bifidobacterium bifidum* in complex with the products α-L-Fuc and Galβ1-4Glc (PDB code 2EAC; Fuc and Lac in blue). Figure edited from Anso *et al*. (2023).[88]

The active site along the GH95 family shows two different conformations of the loop where the catalytic E541 is located. In the unliganded structures of FucOB and *Bb*AfcA (PDB code 2EAB; Figure 37b), this loop moves away from the binding site creating an open groove. However, this loop adopts a different conformation in GH95 family crystal complexes of *Bb*AfcA (PDB code 2EAC and 2EAE) and BACOVA_03438 (PDB code 4UFC) with glycan substrates, moving around 2.8 Å closer to the substrate and closing the active site (Figure 36).

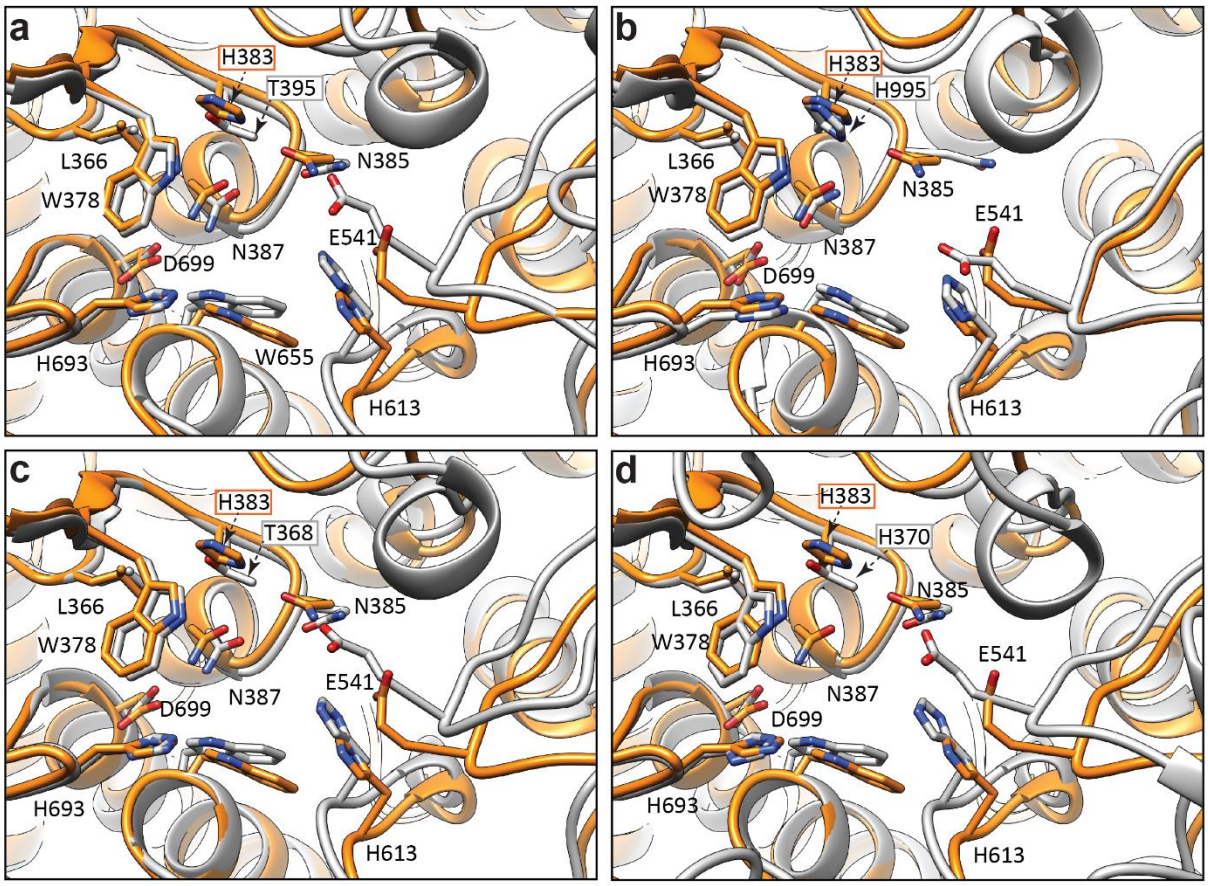

Figure 37: Structural comparison of FucOB with GH95 homologues. Binding site comparison between FucOB (in orange) and crystal structures of GH95 homologues (in grey). It is worth mentioning that the active site's architecture along the GH95 family is highly conserved. FucOB conserved residues that participate in binding are highlighted. The non-conserved residue is squared in orange (FucOB) or grey (homologues). **a** α-1,2-fucosidase *Xac*Afc95 from *Xanthomonas citri* (PDB code 7KMQ), **b** α-1,2-fucosidase *Bb*AfcA from *Bifidobacterium bifidum* (PDB code 2EAB), **c** a putative GH95 member from *Bacillus halodurans* (PDB code 2RDY) **d** α-L-galactosidase BACOVA_03438 from *Bacteroides ovatus* (PDB code 4UFC). Figure edited from Anso *et al*. (2023).[88]

Structural comparison of the FucOB complex with the H type antigen obtained by *in silico* molecular docking calculations with that of *Bb*AfcA in complex with 2'FL reveals that both

ligands superimpose well (Figure 36b,c,d). Single-point mutational analyses in the FucOB and *Bb*AfcA enzymes support that residues N385, N387, E541 and D699 (N421, N423, E566, and D766 in *Bb*AfcA, respectively) play critical roles in the catalytic mechanism of GH95 members. FucOB residues W378, H383, H613, W655 and H693, which also comprise the fucose binding pocket, were found essential for the enzymatic activity and are largely conserved not only in *Bb*AfcA (W414, H419, H678, W722 and H760, respectively) but also other α-1,2-fucosidase members of the GH95 family. Our mutagenesis data identified W453 as an important residue for the interaction with the galactose residue at the +1 subsite. Interestingly, the structure-based alignment shows variability in the region comprising loop 28. However, the enzymes preserve an aromatic residue at this position (W500 in *Bb*AfcA, F435 in BACOVA_03438; Figure 36, Figure 37). Altogether our structural and enzymatic data strongly support a common substrate binding and catalytic mechanisms for the α-1,2-fucosidase members of the GH95 family.

Finally, a structural comparison between BACOVA_03438 (α-L-galactosidase) and *Bb*AfcA (α-1,2-L-fucosidase) suggested that a single residue located at the -1 subsite could determine the substrate specificity for L-galactose or L-fucose. Specifically, T370 in BACOVA_03438 interacts with the L-galactose O6 atom by hydrogen bonds, whereas H419 in *Bb*AfcA interacts with the L-fucose C6 methyl group by aliphatic interactions.[280] Nevertheless, very limited structural data are available for this family to conclude that this polymorphic position at the -1 subsite is the sole structural feature that determines the substrate specificity of this family.

In that sense, subsequent biochemical characterization and the structural determination of other GH95 family members demonstrated that this polymorphic position was not essential for substrate recognition in this family of enzymes. This is the case of *Xac*Afc95 from *Xanthomonas axonopodis pv. citri*,[257,282] and Blon_2355 from *Bifidobacterium longum* subsp. *Infantis*,[283] both of which contain a threonine residue at -1 subsite and showed α-1,2-L-fucosidase activity. In addition, the mutation T395H in *Xac*Afc95 (Figure 37a) maintains the substrate specificity for L-

fucose.[279] Interestingly, the catalytic domain of *Bb*AfcA did not liberate fucose from any of the artificial substrates examined, *p*-nitrophenyl (pNP)-α-L-fucoside, pNP-β-L-fucoside, and 4-methylumbelliferyl-α-L-fucoside.[278] Similarly, BACOVA_03438 (Figure 38d) lacks enzymatic activity against 4-nitrophenyl-α-L-galactopyranoside.[280] Therefore, it was suggested that the subsite +1 could also play a role in defining the enzymatic activity and specificity of GH95 enzymes.[280] However, according to our activity assays, FucOB and *Xac*Afc95,[279] hydrolyze fucose from *p*NP-α-L-fucoside artificial substrate. These findings support a more elaborate mechanism of substrate selectivity in the GH95 family members that is not restricted to direct interactions with the residues comprising the -1 subsite.

This experimental data explains at molecular level the substrate specificity and structural basis of FucOB, an α-1,2-fucosidase from *A. muciniphila*, a bacteria present in gut microbiota with a key role in human health and disease.[46] It is known that *Akkermansia* is a mucin-degrader thanks to the many GHs, sulfatases and peptidases that its genome encodes able to recognize and degrade glycosidic epitopes present in gut mucins, such as FucOB, which recognizes H antigens from ABH blood group epitopes. The mucin glycans that make up the mucus layer provide binding sites and a sustainable source of nutrients for bacteria that inhabit the mucus niche.[284] The peripheral terminal epitopes show considerable variation with a decreasing gradient of fucose and ABH blood group antigens expression from the ileum to the colon.[285] For example, H and A blood group antigens were shown to be present exclusively in the ileum and cecum.[286] It is worth noting that ABH blood group antigens can play a direct role in infection by serving as receptors and/or co-receptors for bacteria, parasites, and viruses.[287]

Moreover, the finding and characterization of the *fucOB* gene close to *ogpA*, an encoding paradigmatic *O*-glycopeptidase that exclusively hydrolyses the peptide bond N-terminal to serine or threonine residues substituted with an *O*-glycan, and a predicted sulfatase (Amuc_1118) (Figure 18), support that the two enzymes seem to be part of a conserved cluster dedicated to the mucin

degradation system. This experimental data helps us to further understand the still unknown mucin degradation system and regulation of *Akkermansia sp.*

5.2.4. FucOB converts universal O to rare Bombay type blood group

Our enzymatic activity assays clearly showed that FucOB can hydrolyse the α-1,2 linked fucose residue in Type I, Type II, Type III and Type V H antigens. In contrast, FucOB cannot cleave the α-1,2 linked fucose residue from Type V A antigen or the Type V B antigen.

The only structural difference between the ABO blood group antigens is the presence or absence of an external sugar residue linked to the common chain precursor in A/B and H antigens, respectively (Figure 18).[93] Therefore, the enzymatic modification of one sugar in the oligosaccharide could change the blood group of a sample into another one, being the conversion of A, B, AB blood groups into O, the called universal blood, highly important for the universal blood supply of blood banks in emergency situations.[288]

This concept was first proposed/demonstrated using an α-galactosidase from green coffee beans, to convert B type RBCs into universal O, following a subsequent successful transfusion.[289,290] Two families of α-*N*-acetylgalactosaminidases (GH109) and α-galactosidases (GH110) successfully converted their corresponding B and A RBCs.[291] However, large quantities of enzymes were needed, rendering the approaches impracticable. The complete removal of both the A and B antigens was achieved with the endo-galactosidase E-ABase from *Clostridium perfringens*, which cleaves the terminal trisaccharides.[292] More recently, a significant advance was made by functional metagenomic screening of the human gut microbiome for enzymes that can convert the A or B type antigens into universal O antigen. An enzymatic pathway from the obligate anaerobe *Flavonifractor plautii* comprising (i) an *N*-acetylgalactosamine deacetylase and (ii) a galactosaminidase (GH36), converted A^+ RBCs to O type universal donor RBCs via a unique mechanism.[293] Their ability to complete the conversion at very low enzyme concentrations in

whole blood will simplify their incorporation into blood transfusion practice, broadening blood supply.[293] In addition, both enzymes were also used to convert blood group A lungs to blood group O lungs using *ex vivo* lung perfusion.[294] The authors showed minimized antibody binding, complement deposition, and antibody-mediated injury after enzymatic treatment suggesting that this strategy has the potential to improve equity in the allocation of organs for transplantation.[294]

With this idea in mind, we wonder whether FucOB could convert universal O into rare Bombay type blood group. In order to determine this ability, type O blood group RBCs were incubated with 200, 50, 5, 0.5, 0.05, and 0.005 μg mL^{-1} of either FucOB or the catalytically inactive version of the enzyme $FucOB_{E541A}$,and analysed by agglutination assays against naturally containing anti-H Bombay serum (Figure 38).

Strikingly, as depicted in Figure 38c, RBCs pre-incubated with FucOB at 200, 50, and 5 μg mL^{-1} showed no agglutination in the presence of Bombay serum that contains anti-H antibodies. This result clearly indicates the cleavage of L-fucose present in the H antigen and therefore, the conversion of universal O into rare Bombay type blood group. It is worth noting that all O RBCs previously incubated (i) without enzyme or (ii) with the inactive $FucOB_{E541A}$, agglutinated due to the presence of anti-H antibodies in the Bombay serum, showing the classic haemolytic reaction described once Bombay blood is mixed with any kind of ABO blood group sample.[94] To check whether this phenomenon is reproducible and representative in larger number of samples, we repeated the experiment with RBCs from 20 donors, 10 O Rh negative and 10 O Rh positive. All universal O RBCs were converted to rare Bombay type blood group (Figure 39).

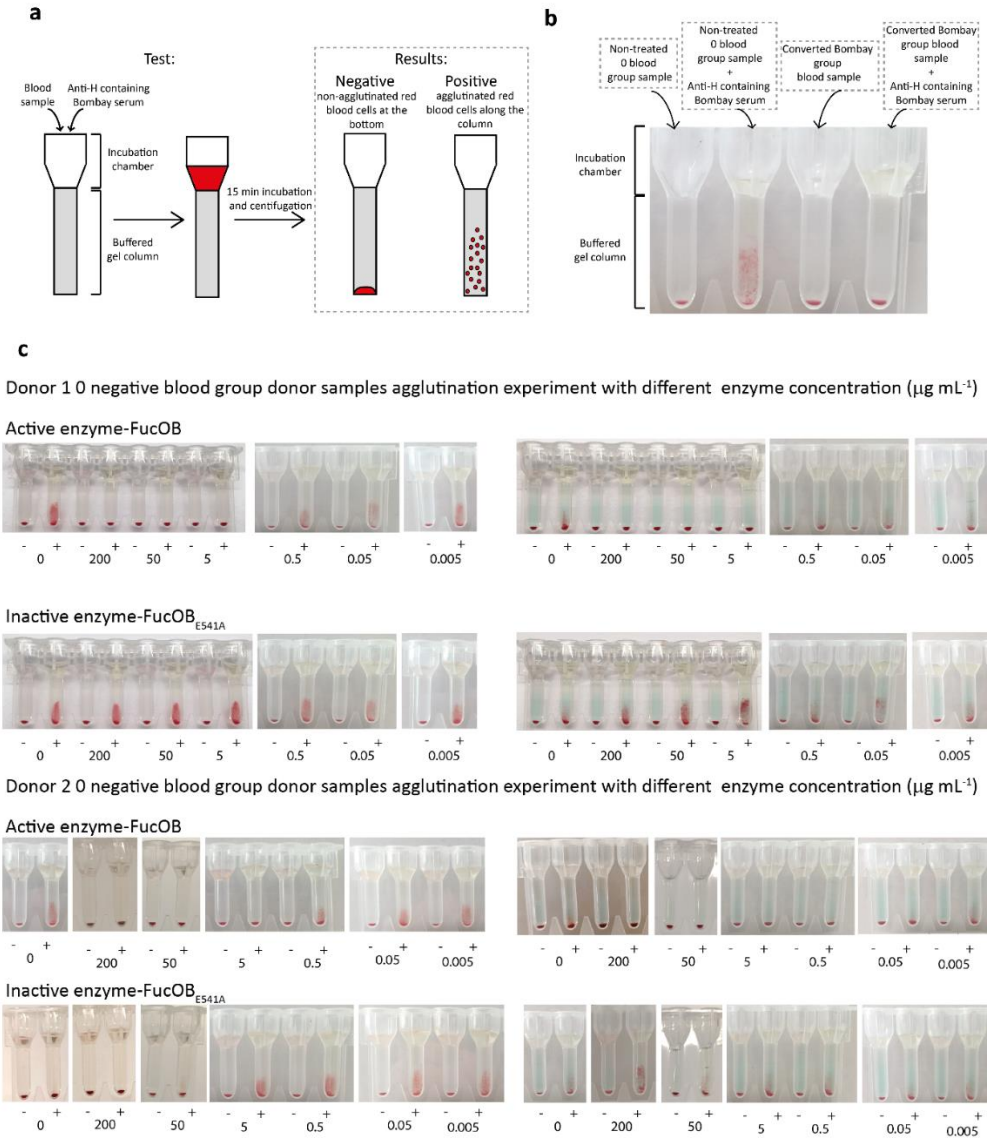

Figure 38: Turning universal O into rare Bombay type blood. **a** Schematic representation of agglutination assay with DG Gel cards from Grifols (Diagnostic Grifols, S.A.) and positive and negative results representation. **b** Agglutination experiment results picture; negative/non-agglutination in non-treated O blood group sample in the first column and positive/agglutination in the second column in non-treated O blood group sample in contact with H antibodies naturally containing Bombay serum as a negative and positive controls, respectively; negative/non-agglutination in the third column in converted Bombay blood group sample; negative/non-agglutinated result in converted Bombay blood group sample in contact with H antibodies naturally containing Bombay serum. **c** Agglutination assay results from 2 different donor 0 blood group samples (donor 1 in first and second row, donor 2 in third and fourth row). In the left panels appear the pictures of the gel-column of DG Gel Neutral cards and in the right panels appear the pictures from DG Gel Coombs cards. 1 and 3 rows show samples enzymatically treated with different concentrations of active FucOB (0, 200, 50, 5, 0.5 0.05 and 0.005 µg mL^{-1}). 2 and 4 rows show samples enzymatically treated with different concentrations of inactive $FucOB_{E541A}$ (0, 200, 50, 5, 0.5 0.05 and 0.005 µg mL^{-1}). Each sample was incubated in the absence (-) and presence (+) of naturally containing H antibodies Bombay serum. The pellet in the bottom of the gel column is a negative result, meaning that there is no agglutination or haemolysis in the sample. When clumps of cells appear throughout the gel column shows a positive result, meaning that cells are agglutinated in the sample. It is worth noting that each agglutination card has 8 buffered tubes to perform the experiments. Therefore, the results are shown in patches. Figure edited from Anso *et al*. (2023).[88]

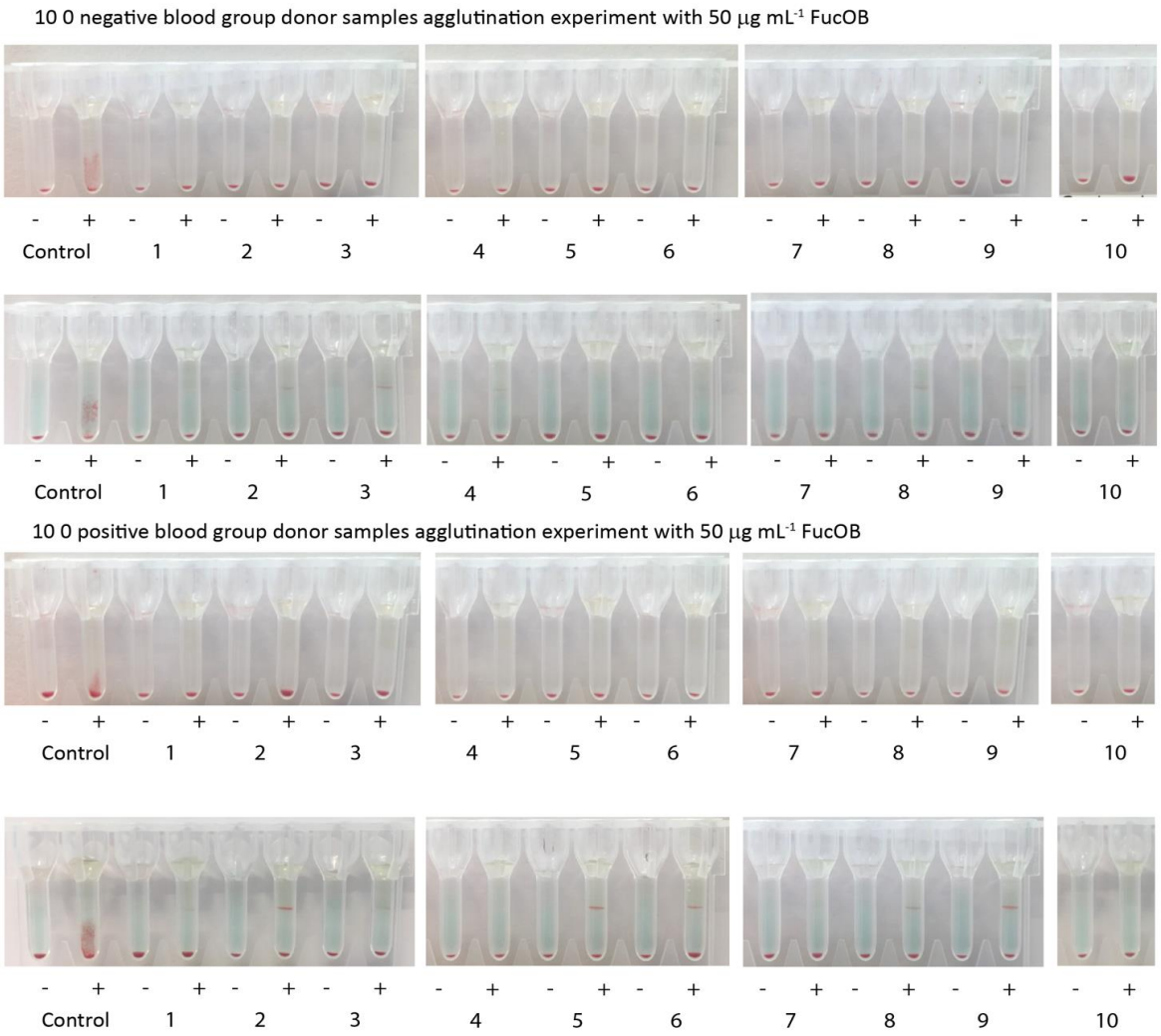

Figure 39: FucOB converts both O^+ and O^- into rare Bombay type blood. Agglutination assay results from 10 different (represented from 1 to 10) O Rh negative blood group donor samples (in first and second rows) and 10 different (described from 1 to 10) O Rh positive blood group donor samples (in third and fourth rows). The agglutination assay was performed with enzymatically treated blood samples with FucOB at 50 µg mL^{-1} concentration and non-treated blood sample as a positive control. In the first and third rows the pictures of the assays performed in the gel column of DG Gel Neutral cards are shown. In the second and fourth rows the pictures of the assays performed in DG Gel Coombs cards are shown. Each sample was incubated in absence (-) and presence (+) of naturally containing H antibodies Bombay serum. The pellet in the bottom of the gel column is a negative result, meaning that there is no agglutination or haemolysis in the sample. When clumps of cells appear throughout the gel column shows a positive result, meaning that cells are agglutinated in the sample. It is worth noting that each agglutination card has 8 buffered tubes to perform the experiments. Therefore, the results are shown in patches. Figure edited from Anso *et al*. (2023).[88]

To further support and validate our findings, we performed a secondary agglutination test using an anti-H lectin, a blood grouping reagent prepared from an extract of *Ulex europaeus* seeds (Figure 40). RBCs from O donors pre-incubated with FucOB 5 µg mL^{-1} showed no agglutination in the presence of the anti-H lectin. In contrast, the catalytically inactive $FucOB_{E541A}$ showed a

clear agglutination reaction after treatment with the anti-H lectin. These results also support the cleavage of L-fucose present in the H antigen by FucOB, and the conversion of universal O into rare Bombay type blood group.

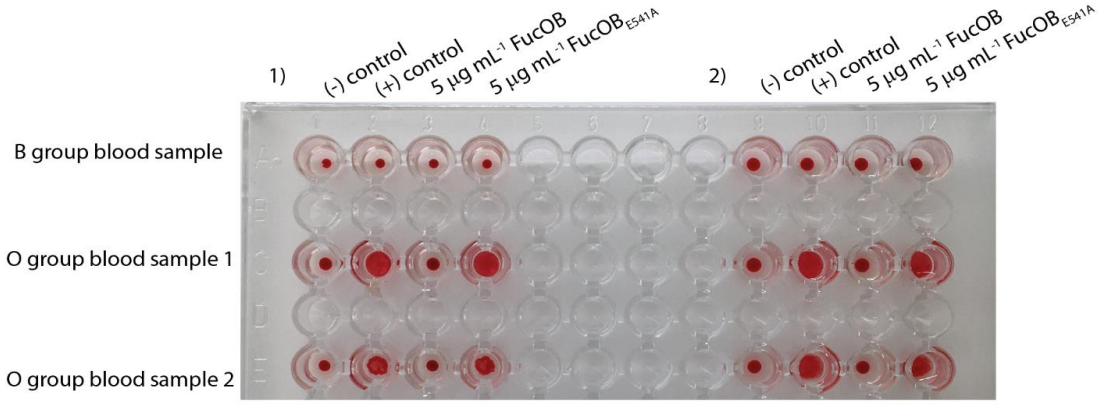

Figure 40: FucOB conversion of universal O into rare Bombay type blood typing by anti-H lectin agglutination assay. Duplicate (experiment 1 on the left side and experiment 2 on the right side of the plate) agglutination assay experiments of three blood samples, (i) one B group blood sample as a negative control in the first row and (ii) two different O group blood samples (sample 1 in row 2 and sample 2 in row 3). In the first column, blood samples were incubated with PBS as a negative (-) control. In the second column, as a positive (+) control, blood samples were incubated with the anti-H lectin. In the third column, blood samples treated with 5 µg mL^{-1} of active wild-type FucOB were incubated with the anti-H lectin. In the fourth column, blood samples treated with 5 µg mL^{-1} of the catalytically inactive $FucOB_{E541A}$ mutant were incubated with the anti-H lectin. The presence of RBCs pellet in the bottom indicates no agglutination (negative result) or haemolysis in the sample. A reddish RBC solution means cells agglutinated in the sample (positive result). Figure edited from Anso *et al.* (2023).[88]

The viability and integrity of converted Bombay RBCs by the action of FucOB were essentially preserved (Figure 41 and Figure 42). Supporting this notion, RBCs belonging to O negative and O positive blood groups were subjected to different tests that are carried out routinely in the clinic. On the one hand, RBCs from non-treated blood, as well as RBCs incubated with and without 50 µg mL^{-1} of FucOB at 37 ºC, were visualized in blood smears to check whereas the morphology of the erythrocytes is affected by (i) the incubation temperature or (ii) the action of FucOB on the H surface antigen of the erythrocytes. As shown in Figure 41 all RBCs display a normal biconcave morphology.[295]

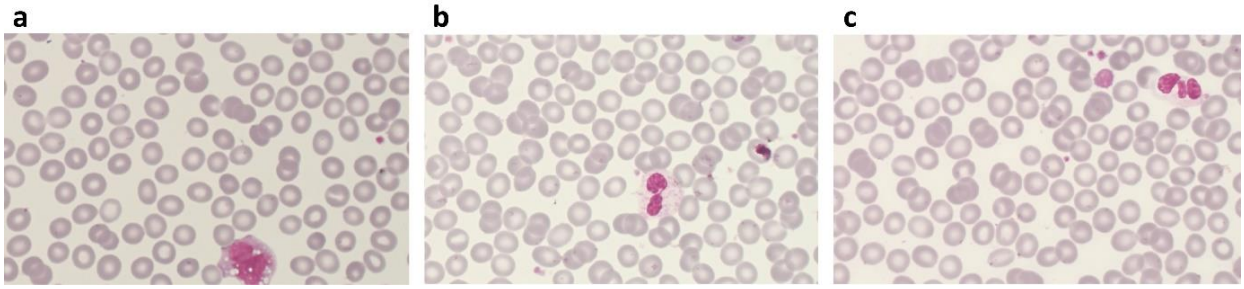

Figure 41: Integrity of RBCs. a Non-treated O blood group smear picture. **b** Blood sample smear picture after an incubation at 37ºC without enzyme. **c** Converted O blood group to Bombay sample's smear picture. RCBs are shown maintaining the normal biconcave morphology in the three panels.

On the other hand, we performed the G6PD assay in the same RBCs. G6PD is a ubiquitous enzyme present in the membrane of erythrocytes that plays a critical role in the redox metabolism of all aerobic cells. The enzyme catalyses the first and rate-limiting step of the pentose phosphate pathway, generating nicotinamide adenine-dinucleotide-phosphate hydrogen (NADPH) and ribose-5-phosphate, which is essential for the production of nucleotide coenzymes and nucleic acids and therefore, cell division.[277] The positive results in all of our samples indicate that the viability and integrity of the cell membrane is preserved (Figure 42).

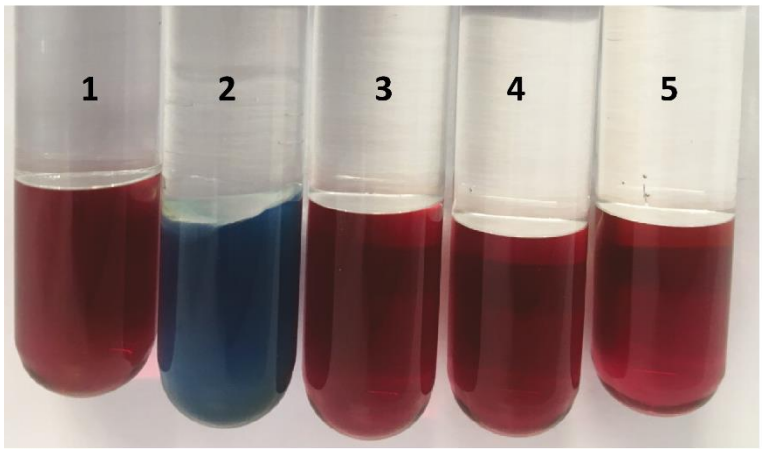

Figure 42: G6DP activity colorimetric assay results. The red color in the tube solution means that G6DP present in RBC is active, and the blue color indicates that it has lost its activity. 1 and 2 tubes are positive and negative controls, respectively. Tube 3 shows a blood sample after incubation at 37ºC without FucOB. Tube 4 and 5 show two different blood samples after incubation at 37ºC in presence of 50 µg mL^{-1} of active FucOB.

To evaluate whether the conversion of O type RBCs by FucOB was complete, we studied by fluorescence-activated cell sorting (FACS) analysis after treatment of O type RBCs with the active enzyme. For that, we first identified RBCs based on the expression of CD235a (glycophorin A), a transmembrane glycoprotein expressed by erythrocytes, and then determined the frequency of antigen H+ cells (Figure 43) with 2 different anti-H blood group monoclonal antibodies (clone 97-

I and clone 86-M anti-H monoclonal antibodies) and an H antigen recognizing conjugated lectin (Methods section).

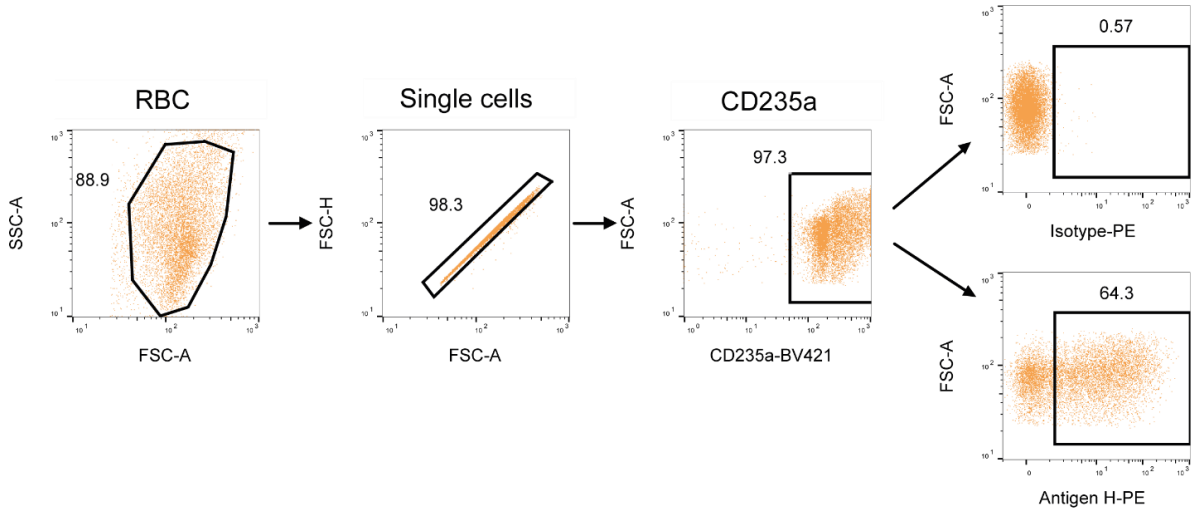

Figure 43: Gating strategy for the analysis of antigen H expression in RBCs. Dot plot graphs showing the gating strategy used for the analysis of antigen H expression on O type RBCs. Data from a representative example is shown. RBCs were electronically gated based on their forward (FSC) and side scatter (SSC) parameters and then single cells were selected. Next, the population positive for the CD235a (Glycophorin A), a transmembrane glycoprotein expressed by erythrocytes, was selected. Finally, the frequency of cells positive for antigen H was analyzed. The determination of the population positive for antigen H was based on the isotype control. Figure edited from Anso *et al*. (2023).[88]

The percentages of positive cells resulted in (i) 64.2%, when we used clone 97-I, and (ii) 56.7%, when we used clone 86-M. (Figure 44) Importantly, the experiments performed with the anti-H lectin showed that practically all group O red cells, 95.1%, were positive. It is worth noting that the inability of anti-H antibodies to label all O type RBCs may be related to the fact that the expression of blood antigens is variable depending on the donors[296,297] and/or the quality of the antibodies. Interestingly, in the literature, the expression of the H antigen by flow cytometry on O type RBCs is generally performed by staining with anti-H lectin.[298] Importantly, the binding of the anti-H monoclonal antibodies and the anti-H lectin completely disappeared when O type RBCs were pre-treated with FucOB (Figure 44).

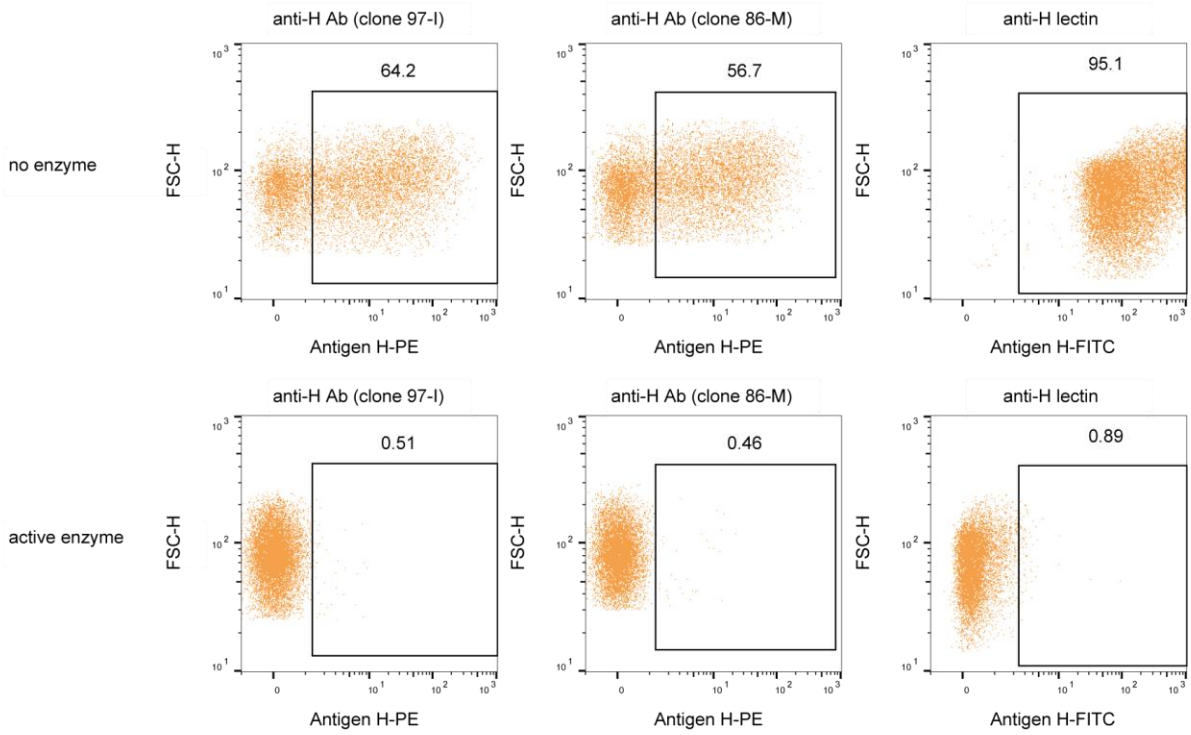

Figure 44: Enzymatic blood type conversion as determined by flow cytometry. Different labelling strategies for the analysis of antigen H expression in RBCs. Dot plot graphs showing the frequency of antigen H+ RBCs from an O blood group sample donor non-incubated and incubated with the active FucOB (5 µg mL^{-1} final concentration). Two different antibodies (clone 97-I left column panels and clone 86-M middle column panels) and an antigen H-specific lectin (right column panels) were used to detect antigen H+ RBCs. Figure edited from Anso *et al*. (2023).[88]

Our experimental data suggest that FucOB to converts universal O type into rare Bombay type blood, providing exciting possibilities to facilitate transfusion in recipients/patients with Bombay phenotype.

A rare blood donor phenotype occurs at least 1/1000 and includes high-frequency-antigen-negative and multiple-common-antigen-negative blood groups, including the Bombay phenotype. Due to the low frequency of each rare blood group, it is crucial to have an accurate international database of rare blood donors to ensure that patients who require lifesaving rare blood units can receive them. There is also a rare inherited primary immunodeficiency disorder called leukocyte adhesion deficiency (LAD) that causes Bombay phenotype. Type II (LAD II) disorder results from a defect in fucose metabolism which enables the correct fucosylation of glycoproteins, including H antigen biosynthesis.[299,300] Due to the low prevalence of each rare disease, medical expertise is

rare, knowledge is scarce, care offerings inadequate and research limited. Thus, despite their large overall number, rare disease patients are the orphans of health systems.[301,302]

Efforts were concentrated on the identification of an enzyme capable of performing the conversion of universal type H into rare Bombay-type blood. Up to date, only 14 GH95 family members were reported to have α-1,2-L-fucosidase activity (EC 3.2.1.63; Fucα1-2Galβ as substrate; 11 from bacterial species and three eukaryotic enzymes). The removal of the fucose residue from the H-antigen oligosaccharide in RBCs was initially explored using an α-1,2-fucosidase from *Aspergillus niger*.[303] However, the enzyme displayed an optimum pH of 4.5, limiting its practical application.[304] An α-1,2-fucosidase successfully modified the Type II chain H antigen on RBCs. However, the Type III chain H antigen was unaffected. In addition, the enzyme displayed maximum activity at acidic pH values, reducing its application.[305] Interestingly, synthetic metallopeptides of 16 to 20 amino acids were designed as artificial fucosidases to remove fucose from Type II H antigen on RBCs.[306] More recently, a membrane α-1,2-fucosidase from *Elizabethkingia meningoseptica* showed activity on Type I, Type II, and Type IV H antigens, producing H-deficient RBCs.[307] Its amino acid sequence displayed ca. 23.9% identity with the soluble FucOB from *A. muciniphila*.[308] FucOB shows encouraging characteristics as an α-1,2-fucosidase. First, FucOB is easily produced in high yields and purity, 5.0 mg L^{-1}. Second, according to our Bombay conversion enzymatic assays, the amount of active enzyme able to hydrolyse the fucose from H antigen in O blood samples is very low, 5 μg mL^{-1}. Finally, FucOB displayed the ability to convert O group into Bombay in complete blood and non-washed RBCs samples. We directly incubated the enzyme at 37 ºC with blood samples collected from the Blood Bank to perform G6PD activity assay and blood smears and we measured that the enzyme can hydrolyse the fucose from H antigens in that condition. Altogether, we propose FucOB as a promising biotechnological and therapeutic tool to cleave the fucose present in O type blood RBCs and therefore, convert universal O type blood into rare Bombay type blood facilitating the

transfusion to Bombay phenotype individuals. Taking into account our extensive/thorough structural, biochemical and substrate specificity analysis of FucOB, it is tempting to speculate that other members of the GH95 family (e.g. *Bb*AfcA) will likely serve the same purpose as FucOB regarding the conversion of O blood group to the Bombay blood group.

6. MOLECULAR RULER AND INTERFACIAL CATALYSIS STUDY OF THE MYCOBACTERIAL ACYLTRANSFERASE, PatA.

Understanding of the mycobacterial cell envelope's structure, composition, and function has been a major challenge during the last decades due to its implication in the bacterial infection process, pathogenesis, and drug resistance.[100] PIMs and their derivatives, LM and LAM, are some of the unique glycolipids involved in those key functions, mainly located in the plasma membrane but also found in the innermost layer (Figure 9). Although research on PIMs, LM and LAM has been going on for some decades now, there are many unknowns regarding to their biosynthesis, translocation, and transport strategies.

That's why, in this second section of the thesis, I have been focused on better understanding the enzymes involved in the biosynthesis of PIMs, LM and LAM. In that sense, I studied the membrane association mechanism of PatA, an acyltransferase involved in the early stages of PIMs biosynthesis, by SPR and Langmuir Balance techniques and its substrate recognition by an acyl chain selectivity "ruler" mechanism by enzymatic activity assays, structural biology, binding measurements on chemically synthesized non-hydrolyzable CoA derivatives (chapter 6).

I was also involved into the design and synthesis of bifunctional reporters that could be employed in the mycobacterial wall's glycolipids visualization and research. (chapter 7). This chemical tool, in optimized growing conditions for mycobacteria, could be helpful into the identification of the set of proteins that participates in the biosynthesis of these glycolipids and remain unknown.

6.1. MATERIALS AND METHODS

6.1.1. Materials

The phospholipids 1,2-dioleoyl-*sn*-glycero- 3-phosphocholine (DOPC), 1,2-dioleoyl-*sn*-glycero-3-phospho- (1′-rac-glycerol) (DOPG), 1,2-dimyristoyl-*sn*-glycero-phosphatidylcholine (DMPC), and 1,2-dimyristoyl-*sn*-glycero-3-phospho-(1′-rac-glycerol) (DMPG); the CoA derivatives, 12:0 CoA (C12-CoA), 14:0 CoA (C14-CoA), 16:0 CoA (C16-CoA), 18:0 CoA (C18-CoA), and 20:0 CoA (C20-CoA) ammonium salts; and the nonhydrolyzable CoA derivatives 16:0 ether CoA and 14:0 ether CoA ammonium salts, were purchased from Avanti Polar Lipids Inc. (Alabaster, AL). The 02:0 acetyl-CoA lithium salt (C2-CoA) and 08:0 octanoyl-CoA ammonium salt (C8-CoA) were purchased from Sigma- Aldrich (Missouri, USA). All reagents were used without further purification. PIM_2 was chemically synthesized as previously described.[132]

6.1.2. Expression and purification of PatA from *M. smegmatis* mc^{2}155

The full-length *patA* gene from *M. smegmatis* mc^{2}155 (*MSMEG_2934*; UniProt code A0QWG5) was inserted into the pJAM3 plasmid using the *Nde I* and *Xho I* sites (pJAM3-*patA*). The pJAM3 plasmid is a derivative of pJAM2 plasmid, where a *Nde I* site was generated in the multiple cloning site. *M. smegmatis* mc^{2}155 cells, transformed with pJAM3-*patA*, were grown in MM63 medium [15 mM $(NH_4)_2SO_4$ and 100 mM KH_2PO_4] supplemented with 0.5 mM $MgSO_4$, 0.025% (v/v) tyloxapol, and 17 mM succinic acid, adjusted to pH 7.0 with KOH, supplemented with kanamycin (20 μg mL^{-1}).[132] When the culture reached optical density at 600 nm (OD_{600}) = 0.6, the expression of PatA was induced by adding 0.2% acetamide. After 30 h at 37°C, cells were harvested at 5000 × *g* for 20 min at 4°C and resuspended in 20 mM Tris-HCl pH 7.5 containing protease inhibitors (Complete EDTA-free, Roche) and culture of benzonase (0.5 μL $liter^{-1}$; Sigma-Aldrich). Cells were disrupted by sonication in 36 cycles of 10-s pulses, with 60-s cooling intervals between the pulses, and 60% of amplitude. To solubilize the protein from the mycobacterial membrane, we

added a final concentration of 2 mM CHAPS, 300 mM NaCl, and 10% glycerol, to the previous disrupted cells mixture. The suspension was gently stirred for 1 h at 4°C and centrifuged at 59,000 × *g* for 30 min at 4°C. The supernatant was filtered by 0.22-μm pore-size Merck Millipore Durapore polyvinylidene difluoride (PVDF) membrane filters and applied to a Co^{2+} Talon Resin (GE Healthcare; ratio 1 mL of resin:1000 mL of culture volume) equilibrated in 20 mM Tris-HCl pH 7.5, 300 mM NaCl, and 10% glycerol and incubated with gently stirring on rotator for 1 hour at 4°C. The resin was then transferred to an empty column and washed with 25 column volumes (CV) of 20 mM Tris-HCl pH 7.5, 300 mM NaCl, 10 mM imidazole, and 10% glycerol. The proteins were then eluted by gravity flow with a step gradient of 20, 50, 75, 100, 200, 300, and 500 mM imidazole, in 20 mM Tris-HCl pH 7.5, 300 mM NaCl, and 10% glycerol, three CV each. The fractions of interest were pooled and dialyzed against 5 L of 50 mM Tris-HCl pH 7.5, 150 mM NaCl, and 10% glycerol overnight at 4°C. The protein was loaded onto a Superdex 200 10/300 GL (GE Healthcare) equilibrated in the 20 mM Tris-HCl pH 7.5 and 150 mM NaCl. The resulting preparation displayed a single-protein band when analysed by SDS–polyacrylamide gel electrophoresis (PAGE). The purified full-length PatA was stored at −80°C. We obtained 1.6 mg of purified PatA per liter of culture.

6.1.3. Expression and purification of PimB from *M. smegmatis* mc^2155

E. coli BL21(DE3) cells transformed with pET29a-*pimB* were grown in 2xYT medium supplemented with kanamycin (50 μg mL^{-1}) at 37°C (*29*). When the culture reached OD_{600} = 0.6, PimB expression was induced by adding 1.0 mM isopropyl-β-d-thiogalactopyranoside. After 20 h at 18°C, cells were harvested at 5000 × *g* for 20 min at 4°C and resuspended in 50 mM Tris-HCl pH 8., 500 mM NaCl, 10 mM imidazole, and 25% glycerol containing protease inhibitors (complete EDTA-free, Roche) and culture of benzonase (0.5 μL $liter^{-1}$; Sigma-Aldrich). Cells were then disrupted by sonication in 12 cycles of 10-s pulses, with 60-s cooling intervals between

the pulses, and 60% of amplitude, and the suspension was centrifuged for 30 min at 59,000 × *g* at 18°C. The supernatant was filtered by 0.22-μm pore-size Merck Millipore Durapore PVDF membrane filters and subjected to Ni^{2+}-affinity chromatography using a HisTrap Chelating column (1 mL; GE Healthcare) equilibrated in 50 mM Tris-HCl pH 8.0, 500 mM NaCl, 10 mM imidazole, and 25% glycerol. The elution was performed with a linear gradient of 20 to 250 mM imidazole at 1 mL min^{-1}. The fractions of interest were pooled and dialyzed overnight at 4°C against 50 mM tris-HCl (pH 7.5), 350 mM NaCl, and 25% glycerol. The protein was loaded onto a Superdex 200 10/300 GL (GE Healthcare) equilibrated in 50 mM Tris-HCl pH 7.5, 350 mM NaCl, and 25% glycerol. The resulting preparation displayed a single protein band when analysed by SDS-PAGE. The purified PimB was aliquoted and stored at −80°C. We obtained 6.25 mg of purified PimB per liter of growth culture.

6.1.4. Isolation of native lipids from *M. smegmatis* mc^2155

M. smegmatis mc^2155 strain was cultured in 200 mL of Middlebrook 7H9 supplemented with 10% (v/v) OADC (oleic acid, albumin, dextrose and catalase) enrichment medium [NaCl (8.8 g L^{-1}), albumin bovine fraction V (Amresco), dextrose (20 g L^{-1}), 0.05% (v/v) of 1% stock solution of oleic acid, catalase (40 mg L^{-1}), and 0.5% (v/v) glycerol] for 3 days at 37°C in absence of agitation. Bacterial cultures were centrifuged at 5000 × *g* for 10 min at 4°C. The resulting pellet was washed with 137 mM NaCl, 2.7 mM KCl, 8 mM Na_2HPO_4, and 2 mM KH_2PO_4 [phosphate-buffered saline (PBS)] at 4°C. The cells were resuspended in PBS, 250 mM sucrose, and 1 mM EDTA, supplemented with protease inhibitors, and disrupted in a FastPrep-24 5G bead beater using zirconia beads (0.1 mm zirconia/ silica; BioSpec Products), nine beating cycles of 30 s, with 60-s cooling intervals between the pulses. The lysed cells were centrifuged at 10,000 × *g* for 10 min at 4°C to remove the beads and unbroken cells. The supernatant was centrifuged at 100,000 × *g* for

1 hour at 4°C. The isolated *M. smegmatis* membranes were resuspended in PBS, 250 mM sucrose, 20 mM Tris-HCl pH 7.5, 300 mM NaCl, and 10% glycerol and quickly frozen at −80°C.

Liquid-liquid extraction technique was used to purify the membrane's lipids. To this end, 2.4 mL of $CHCl_3$:CH_3OH (2:1) were added to 0.3 mg of the *M. smegmatis* mc^2155 membranes. The mixture was vigorously mixed, and the organic and aqueous phases were separated because of the solvents immiscibility at 4°C. The organic phase was washed with water and 2 M NaCl, and the polar phase was washed with $CHCl_3$:CH_3OH (2:1) twice. All the saved organic phases were collected in a tube, dried under a N_2 flow, and centrifuged for 1 h under vacuum in a RVC 2-25 CD plus concentrator at 37°C to remove traces of organic solvents. Two milligrams of dry lipid film were obtained.

6.1.5. Preparation of lipid vesicles

For SPR (Surface Plasmon Resonance) measurements, SUV (Small Unilamellar vesicles) were prepared starting from a dry lipidic film [DOPC, DOPG:DOPC (60:40, w/w), DMPC DMPG:DMPC (60:40, w/w), or *M. smegmatis* mc^2155 membrane lipids]. The dry film was prepared from a $CHCl_3$:CH_3OH (2:1) solution by removing the solvent under a stream of N_2 and ultracentrifuged under vacuum for 1 h to remove traces of organic solvent. SUVs of each lipid (DOPC films to create DOPC-SUVs, DOPG:DOPC mixture film to create DOPG-SUVs, DMPC films to create DMPC-SUVs, DMPG:DMPC mixture film to create DMPG-SUVs, and *M. smegmatis* mc^2155 membrane lipids films to create *M. smegmatis* mc^2155-SUVs) were prepared by dispersion of the lipids 20 mM Tris-HCl pH 7.5 and 150 mM NaCl by vigorous vortexing and following sonication at 4°C with a probe microtip sonicator [MSE Soniprep 150 (MSE, UK)] of two cycles of 5 min (5 s on and off cycles) at 100-μm amplitude and a break of 5 min in between cycles.

6.1.6. Surface plasmon resonance (SPR) experiments

Surface plasmon resonance (SPR) experiments were performed in a Biacore 3000 system (GE Healthcare). The running buffer for PatA and PimB experiments was 50 mM Tris-HCl pH 7.5, 150 mM NaCl and 50 mM Tris-HCl pH 7.5, 350 mM NaCl 25% glycerol, respectively. The chip surfaces were first washed with 50 mM NaOH:isopropanol (2:3) and 20 mM CHAPS, and then the lipid vesicles were immobilized on the chip's surface. DOPC-SUVs and DMPC-SUVs were immobilized at 5 μL min^{-1} to reach a density of 2000 RU (resonance unit) in flow cell 1 as negative control because we did not observe binding of PatA either PimB to DOPC-SUVs and DMPC-SUVs in preliminary experiments. DOPG-SUVs, DMPG-SUVs, and *M. smegmatis* mc^2155 lipids SUVs were immobilized at 2 μL min^{-1} to reach the same density in flow cell 2. After lipid vesicle immobilization, BSA (1 mg mL^{-1}) was injected at 30 μL min^{-1} to fill any possible noncovered chip surface. After baseline stabilization, twofold concentration series (from 12 μM up to 0.05 μM) of 60 μL of the corresponding protein was injected at 30 μL min^{-1} for 1 min. After 5 min of dissociation time, the surface was regenerated with 50 mM NaOH:isopropanol (2:3) and 20 mM CHAPS between binding cycles. Experimental data were corrected for instrumental and bulk artifacts by referencing to flow cell 1, using the BIAevaluation 4.1 software. The affinity constants (K_D) were calculated using a general steady-state equilibrium model and the BIAevaluation 4.1 software.

6.1.7. Monolayer studies by Langmuir films

Protein-lipid interactions were studied with the Langmuir film balance technique using a Kibron microtensiometer (μTROUGH SX, Kibron Inc. Helsinki, Finland). We have prepared a solution of 1 mM of the DOPC- and DOPG-lipids [DOPG:DOPC (60:40, w/w)]. Monomolecular films of the indicated lipid were spread on 20 mM Tris-HCl pH 7.5, 150 mM NaCl for PatA and 50 mM Tris-HCl pH 7.5, 350 mM NaCl, 25% glycerol for PimB subphases (volume of 1250 μL). The

desired lipidic monolayer was created by spreading over the aqueous subphase of the DOPC or DOPG with a 10 μL Hamilton syringe inducing different superficial tension values (π_0) between 15 and 40 mN m^{-1}. After baseline equilibration, a final concentration of 100 μM protein (PatA, PimB, and BSA) was injected. PimB and BSA were used as negative controls. Superficial pressure changes were continuously recorded as a function of time, and the achieved pressure was saved when a plateau was reached ($\pi_{protein}$). The data were analysed with the FilmWareX 3.57 program (Kibron Inc.). The experiment results were plotted following a linear regression between the monolayer pressure difference before and after the protein addition ($\Delta\pi_{max} = \pi_{protein} - \pi_0$) versus the initial monolayer superficial tension values (π_0). The critical pressure of the protein insertion (π_c) is the value where, in the plot, the abscissa axis is crossed.

6.1.8. Molecular modelling

First, helices α1, α2, and partially α3 were modelled for PatA using the crystal structure of lipid A secondary acyltransferase LpxM as a template (PDB code 5KN7)[180] and with PyMOL 2.4 software. Subsequently, the PIM_2 glycolipid was docked to the PatA-ManC16 molecule, whose binding pose with the protein was previously revealed using x-ray crystallography by our group (PDB code 5OCE).[132]

Bilayers of 1,2-distearoyl-*sn*-glycero-3-phosphocholine lipid, together with the modelled complex, H_2O molecules, and Cl^- to neutralize the system, were constructed using the CHARMM Membrane Builder GUI (www.charmm-gui.org)[309] and converted to Lipid17 PDB format with the *charmmlipid2amber.x* script included in Amber20.[269] MD simulations were performed using the ff14SB force field, which is an evolution of the Stony Brook modification of the Amber 99 force field force field (ff99SB).[270] Lipid17 force field[310] was used to properly simulate the conformational behaviour of the lipid membrane. Parameters for the PIM_2 glycolipid were generated with the antechamber module of Amber20 using the general Amber force field[311] with

partial charges set to fit the electrostatic potential generated with HF/6-31G(d) by RESP.[312] The charges were calculated according to the Merz-Singh-Kollman scheme using Gaussian 09.[313] The TIP3P water model was used for H_2O.[272] The full system was minimized for 10,000 steps, of which the first 5000 steps used the steepest descent method and the remaining steps used the conjugate gradient method.[314] The system was then heated from 0 to 100 K using Langevin dynamics [315] for 5 ps at constant volume, with weak restraints on the lipid (force constant, 10 kcal mol^{-1} $Å^{-2}$). Following this, the volume was allowed to change freely and the temperature increased to 300 K with a Langevin collision frequency of $\gamma = 1.0$ ps^{-1} and anisotropic Berendsen regulation[316] (1 atm) with a time constant of 2 ps for 100 ps. The same weak restraint of 10 kcal mol^{-1} $Å^{-2}$ was maintained on the lipid molecules. Constant pressure and constant temperature (NPT) runs were performed using the AMBER 20. Bonds involving hydrogen were constrained using the SHAKE algorithm, allowing a 2-fs time step. Structural data were recorded every 10 ps. PME (particle mesh Ewald) was used to treat all electrostatic interactions with a real space cut-off of 10 Å. A long-range analytical dispersion correction was applied to the energy and pressure. All simulations were performed at constant pressure of 1 atm and constant target temperature. Temperature was controlled by the Langevin thermostat, with a collision frequency of $\gamma = 1.0$ ps^{-1}. Pressure was regulated by the anisotropic Berendsen method (1 atm) with a pressure relaxation time of 1.0 ps. The system was simulated for 0.5 μs.

6.1.9. PatA palmitoyl-CoA hydrolytic activity assay

The specific activity of full-length PatA using different acyl-CoA derivatives as acyl donors (C2-CoA, C8-CoA, C12-CoA, C14-CoA, C16-CoA, C18-CoA, and C20-CoA) was measured by the spectrophotometric disulfide DTNB (5,5′-Dithiobis(2-nitrobenzoic acid)) assay, a method described for other acyltransferases.[131,132,317,318] Specifically, PatA transfers the acyl chain from the acyl-CoA compound to PIM_2, producing CoA from the hydrolysis of the acyl donor. Thus, CoA-SH liberated in the reaction reacts with DTNB, releasing TNB^{-2} as a product. TNB^{-2}

absorbance is measured at 412-nm wavelength in a NanoDrop One spectrophotometer (Thermo Fisher Scientific). The reaction contained 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM DOPC-SUV, 200 μM PIM_2 mixture, 200 μM of the acyl-CoA derivative, 600 μM DTNB, and 1 μM full-length PatA. DOPC-SUV and PIM_2 were vortexed together in 20 mM Tris-HCl pH 8.3, in two cycles of 2 min with a 2-min time lapse in between. The protein was added after the blank was measured. After the reaction started, TNB^{-2} absorbance was measured every 3 s for the first 180 s and then every 7 s until 1600 s was reached. The absorbance values during the first 60 s were converted into the amount of hydrolysed acyl-CoA derivative per milligram of protein in 1 min and compared with the most active one activity (C14-CoA in our scenario). All enzymatic activity measurements were determined in duplicates.

6.1.10. PatA acyltransferase activity assay

PatA transferase activity was measured by high-resolution LC (Liquid chromatography) coupled to MS (Mass Spectrometry). A typical reaction contained 20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM DOPC-SUV, 200 μM PIM_2 mixture, 200 μM of the acyl-CoA derivative, 600 μM DTNB, and 3.5 μM full-length PatA. DOPC-SUV and PIM_2 are vortexed together in 20 mM Tris-HCl pH 8.3, in two cycles of 2 min with a 2-min time lapse in between. The reaction was incubated for 3 h at room temperature. The reaction was quenched by adding 60% (v/v) final concentration of acetonitrile. In negative control reactions, full-length PatA was added after 1 h and immediately quenched. The quenched samples were sonicated for 10 min. To 350 μL of sonicated mixture, 350 μL of chloroform was added, and the solution was incubated for 30 min at room temperature under shaking, followed by centrifugation for 5 min at 4°C and 21,000 × *g* for phase separation. 100 μL of the aqueous phase were dried at −4°C in a rotary vacuum concentrator (Labconco). Before LC-MS analysis, samples were resuspended and vortexed in 150 μL of a CH3OH:H_2O (2:1) mixture.

6.1.11. Liquid chromatography–mass spectrometry (LC-MS) experiments

Relative quantification of target compounds was performed using an Agilent 1290 LC coupled to an Agilent 6560 quadrupole time-off light (Q-TOF) MS system equipped with a Dual Agilent Jet Stream ESI (electrospray ionization) source. The column (Waters ACQUITY UPLC BEH C18, 100 × 1 mm, 1.7-μm particle size) used in this study was maintained at 40°C. The autosampler was kept at 4°C, and the injection volume was 10 μL. The flow rate was set to 0.12 mL min^{-1}, and the mobile phases were consisted of 10 mM ammonium formate in 60:40 CH3CN:H_2O (Eluent A, pH unadjusted) and 10 mM ammonium formate in 90:10 IPN:CH3CN (Eluent B, pH unadjusted). The run consisted of a linear gradient from 40 to 99% Eluent B over 6 min, followed by an isocratic delivery of 99% Eluent B over 7 min and a reequilibration phase on starting conditions with 40% Eluent B for 7 min. Total run time was 20 min per sample. MS experiments were performed using ESI in negative mode with a capillary voltage of 2000 V and a nozzle voltage of 500 V. The deprotonated target molecules were monitored in high-resolution mode (slicer position: 5) and Extended Dynamic Range (2 GHz) with the following Q-TOF MS conditions: drying gas temperature, 325°C; drying gas flow, 10 L min^{-1} (nitrogen); nebulizer, 35 psig; sheath gas temperature, 350°C; sheath gas flow, 12 L min^{-1}; fragmentor, 380 V; and Oct RF Vpp, 750 V. Full-scan spectra were acquired from mass/charge ratio (*m*/*z*) 50 to 1600 (1 spectrum s^{-1}). External mass calibration was performed before measurement of each set of samples. A reference solution (*m*/*z* 112.9855 and *m*/*z* 966.0007) was used for online mass correction during the acquisition. All data were acquired with Agilent Mass Hunter LC/MS Data Acquisition (ver B.09.00, Build 9.0.9044.0) and analysed with Agilent Mass Hunter Profinder (ver 10.0 SP1, Build 10.0.10142.1). Target compounds were identified by exact mass (mass error ± 5 parts per minute), isotopic pattern, and retention time (± 0.15 min) matching (Batch Targeted Feature Extraction). Semiquantification was based on the integrated peak area of the deprotonated target compound.

6.1.12. Chemical synthesis of nonhydrolyzable C12-CoA and C18-CoA derivatives

To a stirred solution of CoA-SH (1.0 equivalent) in a 1.0 M aqueous buffer solution of triethylammonium bicarbonate (1 mL, pH 8.1 to 8.5) was added the corresponding alkyl iodide (20 equivalent) at room temperature. The reaction mixture was heated at 50° to 60°C for 16 h before being cooled to room temperature, and then water was added. The mixture was extracted with CH_2Cl_2 to remove the unreacted alkyl iodide. The aqueous phase was concentrated, and the residue was purified by C18 chromatography (CH_3OH:H_2O from 1:1 to 1:3). The collected fractions were concentrated in vacuum and lyophilized to give the final products as fluffy white solids. See the Supplementary Materials for details.

6.1.12. PatA crystallization and data collection

PatA was crystallized by mixing 0.2 μL of a protein solution at 5.22 mg mL^{-1} in 20 mM Tris-HCl pH 7.5 and 150 mM NaCl in presence of 1 mM S-C12-CoA derivative, with 0.2 μL of 100 mM imidazole/ MES monohydrate pH 8.5, 60 mM divalent cations mixture (magnesium chloride hexahydrate and calcium chloride dihydrate), and 50% (w/v) of precipitant mix based on 20% (w/v) PEG 4000 and 40% (w/v) glycerol (Morpheus protein crystallization screen). The crystals grew in 7 days. The crystals were transferred to a cryo-protectant solution containing 30% glycerol and frozen under liquid nitrogen. Complete x-ray diffraction datasets for both crystal forms were collected at beamline i03 (Diamond Light Source, Oxfordshire, UK). PatA crystallized in the monoclinic space group $P2_1$ with four molecules in the asymmetric unit and diffracted to a maximum resolution of 3.67 Å (Table 5). All datasets were integrated and scaled with XDS following standard procedures.[226]

Table 5. Unliganded PatA crystal structure data collection and refinement statistics.

PatA	
PDB code	5F2Z
Wavelength	0.97625
Resolution range	40.65 - 3.67 (3.80 - 3.67)
Space group	P 1 21 1
Unit cell	81.31 92.97 81.09 90 90.33 90
Total reflections	90742 (7983)
Unique reflections	13263 (1233)
Multiplicity	6.8 (6.5)
Completeness (%)	99.07 (94.90)
Mean I/sigma(I)	9.96 (1.41)
Wilson B-factor	125.8
R-merge	0.1792 (1.459)
R-meas	0.194 (1.584)
R-pim	0.07376 (0.6112)
CC1/2	0.998 (0.7)
CC*	0.999 (0.908)
Reflections used in refinement	13216 (122
Reflections used for R-free	660 (62)
R-work	0.2676 (0.3608)
R-free	0.2837 (0.4449)
CC(work)	0.936 (0.737)
CC(free)	0.906 (0.580)
Number of non-hydrogen atoms	7144
macromolecules	7120
ligands	24
Protein residues	996
RMS(bonds)	0.004
RMS(angles)	0.83
Ramachandran favored (%)	98.48
Ramachandran allowed (%)	1.52
Ramachandran outliers (%)	0
Rotamer outliers (%)	0.16
Clashscore	3.86
Average B-factor	122.99
macromolecules	1222.99
ligands	122.83

6.1.13. PatA structure determination and refinement

Structure determination of PatA was carried out by molecular replacement methods implemented in Phaser[232] and the PHENIX suite[233] and using the PDB code 5F2Z as a search template. The final manual building was performed with Coot[234] and refinement with phenix.refine.[235] The structure was validated by MolProbity.[236] Data collection and refinement statistics are presented in Table 5. The atomic coordinates and structure factors have been deposited with the PDB code 7OJT. Molecular graphics and structural analyses were performed with the UCSF Chimera package.[237]

6.1.14. Isothermal titration calorimetry (ITC) measurements

ITC essays were performed on a MicroCal PEAQ-ITC (Malvern) at 25°C. 200 μL of 50 μM full-length PatA in 20 mM Tris-HCl pH 7.5 and 150 mM NaCl was degasified for 30 min and placed into the cell. 40 μL of 0.716 mM nonhydrolyzable acyl-CoA derivatives in 20 mM Tris-HCl pH 7.5 and 150 mM NaCl (S-C12-CoA, S-C14-CoA, S-C16-CoA, and S-C18-CoA) were added in 13 injections (0.4 μL of first injection and 3 μL of the rest of them with 150-s time lapse between each injection) at 750 rpm. Before data analysis was carried out, dilution experiments were performed and subtracted (buffer versus buffer, buffer versus full-length PatA, and CoA NH derivatives versus buffer) to the raw data. Subtracted experimental data were fitted, and K_D parameters were calculated with MicroCal PEAQ-ITC Analysis software (Malvern Instruments Ltd.) following One Set of Sites-Single Injection fitting model

6.2. RESULTS AND DISCUSSION

6.2.1. The ruler mechanism for acyl chains recognition

According to the experimental data PatA prefers palmitoyl-CoA as a donor substrate rather than any other acid chain length CoA derivative,[319] therefore, to understand how this enzyme differs between different length acyl chains, a structural homologue search was performed in previous studies using DALI server.[131,132,239] The results revealed two structural homologues; an acyltransferase from *Acinetobacter baumannii* (*Ab*LpxM)[180] and the glycerol-3-phosphate acyltransferase from *Cucurbita moschata* (*Cm*GPAT).[320] *Ab*LpxM highly conserves the hydrophobic tunnel and superimposes very well with PatA's donor substrate's pocket but displays a different acyl chain preference; this enzyme transfers 12 carbon length acyl chains over the other acyl chains length tested (C10, C14 and C16), instead of 16 carbon length acyl chains, as PatA does. In order to see whether *Ab*LpxM could accommodate a larger acyl chain CoA derivatives, a palmitoyl CoA molecule was fitted into the equivalent cavity and subjected to energy

minimization (Figure 45). The result showed that the hydrophobic groove is shorter than that observed in PatA crystal structure (Figure 45), suggesting that the specificity of the acyltransferase could be determined by the length of the groove, pointing to the occurrence of a hydrocarbon ruler mechanism.

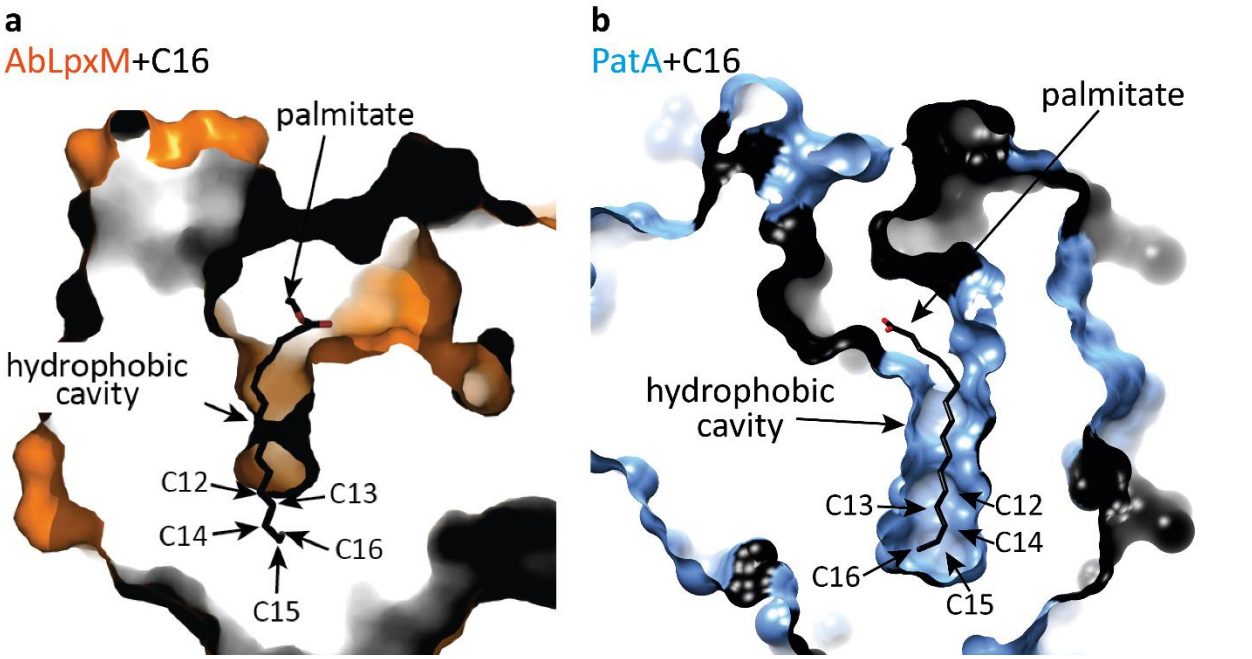

Figure 45: PatA and *Ab*LpxM common acyl donor binding site. a *Ab*LpxM hydrophobic groove's surface shown in orange superposed a palmitate group (C16) inside (in black). **b** PatA hydrophobic groove's surface shown in blue and black in presence of palmitate group (C16) shown in black.

To this end, we measured the specific activity of PatA against acyl-CoA derivatives with different lengths of the acyl chain, including C2-CoA, C8-CoA, C12-CoA, C14-CoA, C16-CoA, C18-CoA, and C20-CoA, by a spectrophotometric assay using 5,5′-dithio-bis- (2-nitrobenzoic acid) (DTNB; Figure 46a). The assay measures the absorption of the released TNB^{+2} produced by reaction between DTNB and the CoA released when PatA catalyses the transfer of the acyl chain from an acyl-CoA derivative to PIM_2. All reactions were performed in the presence of an artificial membrane model DOPC- SUVs, which is required for the PatA reaction.[132] The assays showed that the highest activity corresponds to C14-CoA, followed by the C12-CoA and C16-CoA derivatives; the last was previously described as the preferred acyl-CoA donor substrate of the enzyme (Figure 46a).[120,121] The formation of the reaction products was detected and confirmed by

mass spectrometry (MS) in all the reactions (Figure 46b). The increase or decrease in the acyl chain length in the CoA substrate derivatives, compared to C14-CoA, resulted in a gradual reduction of enzymatic activity, with almost no activity with C2-CoA and C20-CoA. The activity measurements showed a clear Gaussian-like curve profile, strongly supporting the notion that PatA can recognize subtle differences in acyl chain lengths in CoA derivatives. The C14-CoA derivative unexpectedly showed the highest specific activity compared to C16-CoA. This could be due to a different binding mode between the enzyme and the C14-CoA, with respect to the natural substrate C16-CoA.

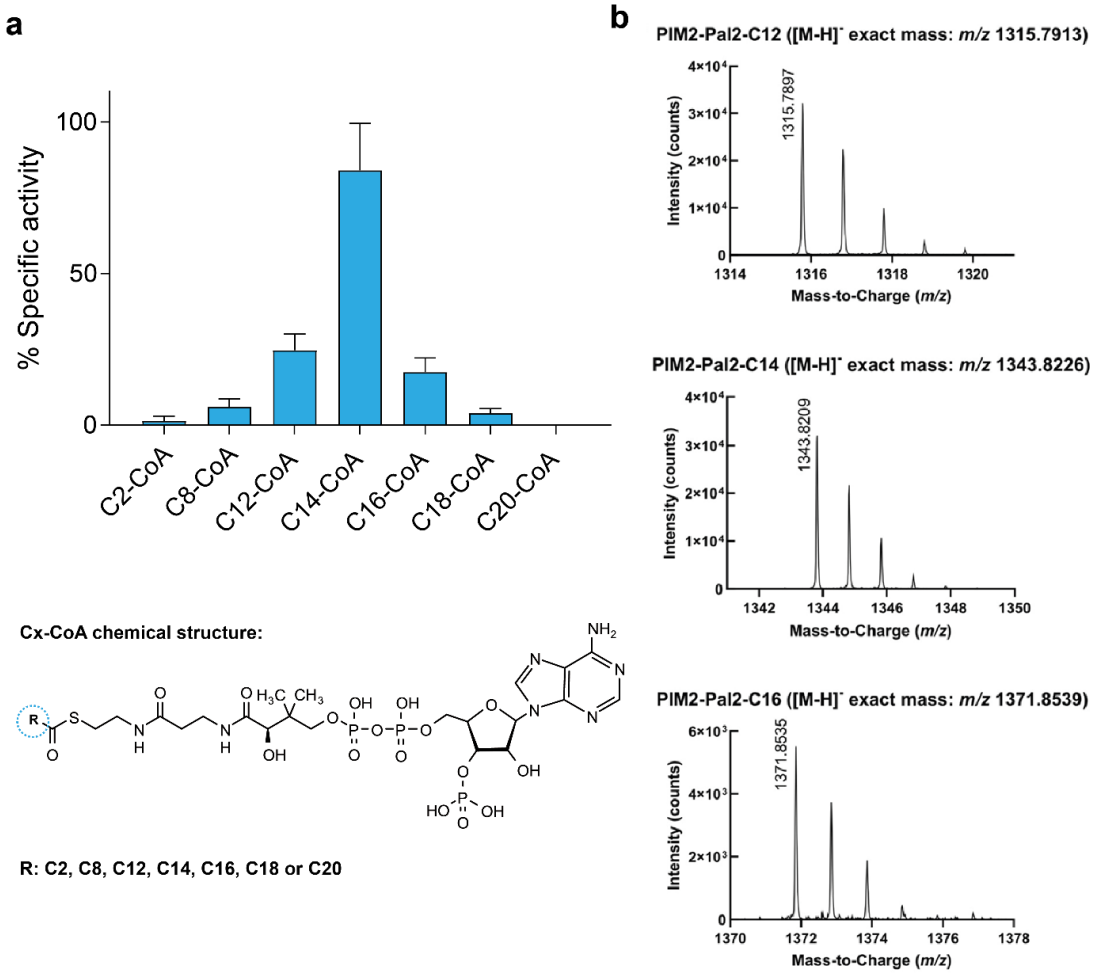

Figure 46: PatA substrate specificity. a PatA-specific hydrolytic activity measured by spectrophotometric DTNB assay with different acyl-CoA derivatives (C2-CoA, C8-CoA, C12-CoA, C14-CoA, C16-CoA, C18-CoA, and C20-CoA) in the presence of DOPC-SUVs as an artificial membrane model. The activity results of PatA for each compound are plotted in reference to the most active one, the C14-CoA derivative. **b** Representative three mass spectra of the deprotonated reaction product; PIM2-Pal2-C12 (after C12 acyl chain transfer to PIM2) in the top panel; PIM2-Pal2-C14 (after C14 acyl chain transfer to PIM2) in the middle panel and ; PIM2-Pal2-C16 (after C16 acyl chain transfer to PIM2) in the top panel.

With this in mind, we studied the impact of acyl chain length on the binding of different acyl-CoA derivatives to PatA, using ITC. To this end, we chemically synthesized the nonhydrolyzable ether substrates *S*-lauroyl-CoA (S-C12-CoA), *S*-stearoyl-CoA (S-C18-CoA), and *S*-arachidoyl-CoA (S-C20-CoA). Specifically, the nonhydrolyzable ether substrates were obtained by conjugation of CoA-SH with the corresponding alkyl iodide following described protocols.[321] In addition, *S*-tetradecyl-CoA (S-C14CoA) and *S*-hexadecyl-CoA (S-C16-CoA) were purchased in Avanti Polar Lipids. We detected binding exclusively with S-C16-CoA (K_D = 7.18 ± 1.17 μM), the derivative that was previously described as the preferred donor substrate of the enzyme (Figure 47).[119] The high activity of PatA in the presence of C14-CoA, a non-natural substrate of the enzyme, and the low binding affinity of the enzyme to the nonhydrolyzable homolog derivative with the same chain length might be explained by an easier release of the C14 product facilitating reaction turnover.

Together, the experimental data support a fatty acid ruler as the mechanism by which PatA specifically recognizes the donor acyl chain of CoA.

To better understand the recognition mechanism of the donor substrate, we attempted to obtain the cocrystal structure of PatA in complex with the nonhydrolyzable substrates S-C12-CoA, S-C14CoA, S-C18-CoA, and S-C20-CoA. We could obtain good-diffracting crystals for PatA in the presence of the S-C12-CoA derivative and solved the structure at 3.67-Å resolution in the space group P21 (Figure 48b; Table 5; see material and methods). Structural comparison of this new crystal structure with the previously reported PatA-S-C16-CoA complex (PDB code 5F34)[131] revealed that the overall protein structure is essentially preserved (r.m.s.d. of 0.36 Å for 245 residues; Figure 48). We could not find clear electron density to model the S-C12-CoA substrate in the hydrophobic cavity. It is worth noting that C16 was present in the hydrophobic cavity in all reported crystal structures of PatA (PDB codes 5OCE, 5F2Z, 5F2T, and 5F34). Unexpectedly, the hydrophobic cavity maintains its conformation although it is empty (Figure 48).

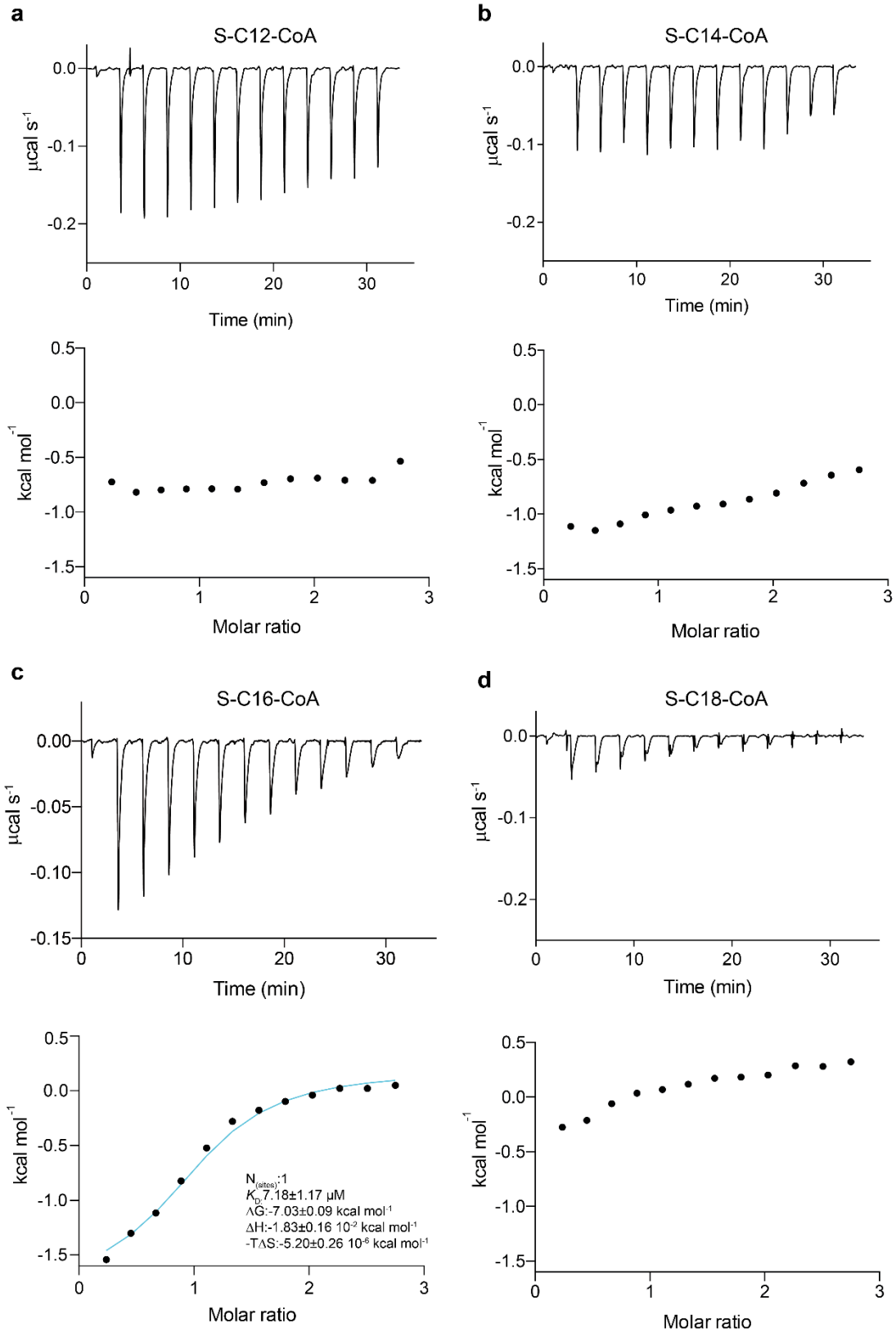

Figure 47: Acyl chain length binding experiments by ITC. ITC raw and integrated data for PatA and non-hydrolysable acyl-CoA derivatives **a** S-C12-CoA **b** S-C14-CoA **c** S-C16-CoA and **d** S-C18-CoA. The experimental points are represented as filled squares and the best fit of these points to a one-site binding model is represented as a solid curve for S-C16-CoA experiment. Figure edited from Anso *et al.*(2021).[322]

These data suggest that the S-C12-CoA derivative can access both the main grove and the hydrophobic cavity and displaces the palmitic acid coming from the purification.[131] However, since it displays a short acyl chain, the S-C12-CoA is released from the active site, as it is not sufficiently stabilized.

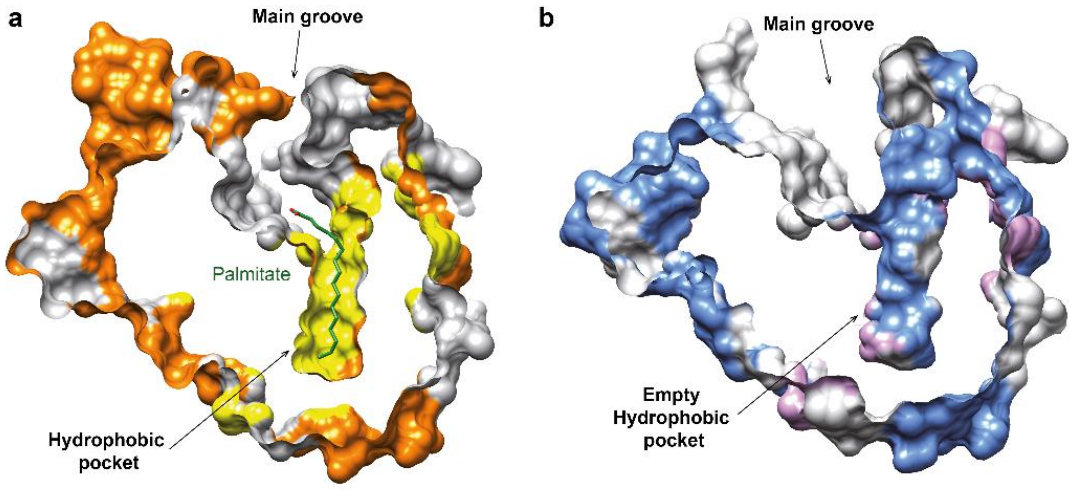

Figure 48: Acyl chain binding groove structure of PatA. a The hydrophobic pocket surface shape of PatA crystal structure in complex with palmitate (in green) (PDB code 5F2Z). **b** The empty hydrophobic pocket surface shape of the crystal structure PDB code 7OJT.

6.2.2. PatA's membrane association model

PatA participates in the early stages of PIMs biosynthetic pathway, which occurs in the cytosolic face of the innermost membrane. Therefore, a close interaction of the enzyme with the cytosolic face of the mycobacterial plasma membrane might be a strict requirement. Supporting this notion, PatA was found to colocalize in mycobacterial membrane fractions.[128,130,135]

In order to further understand the membrane's importance for PatA catalysis, we first performed binding experiments by SPR.[323,324] We first chose some synthetic anionic DOPG or zwitterionic DOPC phospholipids, to be our membrane models in the binding experiments. We immobilized SUVs composed of DOPG:DOPC (60:40, w/w) (DOPG-SUVs) or DOPC (DOPC-SUVs) into a L1 chip and injected different concentrations (micromolar range) of purified full-length PatA to

determine the apparent dissociation constants (K_D) (see Materials and Methods for details). It is worth noting that the SPR experiments with DOPG-SUVs and DOPC-SUVs were performed at 25°C, temperature at which these lipids are in the fluid phase (phase transition temperature for DOPG and DOPC are -18° and -17°C, respectively). PatA did not bind to DOPC-SUVs; however, the enzyme showed binding to the anionic DOPG- SUVs. Thus, we decided to use the zwitterionic DOPC-SUVs as the negative control. As depicted in Figure 49a, PatA binding to anionic DOPG-SUVs is consistent with the strict requirement that a negative charge needs to be present in the phospholipid for interaction with PatA to occur. It is worth noting that it was not possible to measure the K_D value for PatA, because the equilibrium was not reached during the injection phase. However, as PatA is injected, the SPR signal increases and when the protein injection is stopped, the SPR signal remains constant, indicating that the protein is firmly associated with the DOPG-SUVs (Figure 49a). We determined that PimB, the enzyme that carries out the second step of the PIM biosynthetic pathway,[128] also preferentially binds anionic DOPG-SUVs. However, in contrast to PatA, once the injection of PimB is stopped, the SPR signal decreased suggesting that the interaction with the lipid bilayer is temporarily (Figure 49b), typical of peripheral membrane proteins.[129] It was possible to measure the binding affinity of PimB for DOPG-SUVs (K_D), which is in the low micromolar range (0.309 ± 0.03 μM). We also determined the binding of PatA and PimB to (DMPG): (DMPC) (60:40, w/w) (DMPG-SUVs) or DMPC-SUVs at 30°C, temperature at which these lipids are also in the fluid phase (the phase transition temperature for DMPG and DMPC are 23° and 24°C, respectively; Figure 49c,d). DMPG, an anionic saturated phospholipid, is one of the most widely used lipid system mimicking bacterial membranes. PatA and PimB showed a similar binding interaction with DMPG-SUVs than DOPG-SUVs (Figure 49c,d). Specifically, PatA remains bound to the DMPG-SUVs after stopping the injection while PimB dissociates rapidly after the injection ends (Figure 49c,d). To further explore the differences between PatA and PimB in an environment more similar to its natural one, we analysed the binding

of both enzymes using SUVs generated with membrane lipids isolated from *M. smegmatis* mc^{2}155 cells (Figure 49e,f; see Materials and Methods). The behaviour of the two enzymes was reminiscent of that observed with the DOPG-SUVs; however, the association rates were faster for both enzymes, and the dissociation rate was faster for PimB, while that for PatA did not change. Together, the SPR data indicate that (i) PatA and PimB preferentially bind anionic phospholipids and (ii) PatA firmly associates with the lipid bilayer, while (iii) PimB associates temporarily with the lipid bilayer.

The insertion of PatA into phospholipid membranes was evaluated using Langmuir films at various initial pressure (π_0) values (Figure 49g,h). PatA was injected underneath DOPG:DOPC (60:40, w/w) (DOPG) and DOPC monolayers prepared at the air-water interface, and the interaction was quantified by surface pressure measurements. The critical pressure of insertion (π_c), determined by the abscissa intercept of the surface pressure increase versus π0 plot, indicates the protein propensity for membrane insertion. Hence, a larger πc indicates a higher propensity for protein insertion into monolayers. The lateral pressure of biological membranes corresponds to a surface pressure of 30 to 35 mN m^{-1} in the Langmuir monolayers experiments.[325] Therefore, it is generally accepted that πc values between 30 and 35 mN m^{-1} strongly suggest the insertion of the protein into the lipid monolayer.[326–328] PatA displayed a πc of 31.54 mN m^{-1} in DOPG (R^2 = 0.92) phospholipid monolayer and a πc of 30.96 mN m^{-1} (R^2 = 0.89) in DOPC monolayer experiments, supporting the notion that the enzyme inserts into both types of phospholipid monolayers but preferably into the DOPG monolayer. PimB showed a π_c value of 28.38 mN m^{-1} (R^2 = 0.95) in DOPG experiments and 21.5 mN m^{-1} (R^2 = 0.85) in DOPC experiments, consistent with a peripheral localization, while the soluble bovine serum albumin (BSA) has a πc value of 20.18 mN m^{-1} (R^2 = 0.99) and 20.91 mN m^{-1} (R^2 = 0.99) in DOPG and DOPC experiments, respectively (Figure 49g,h).

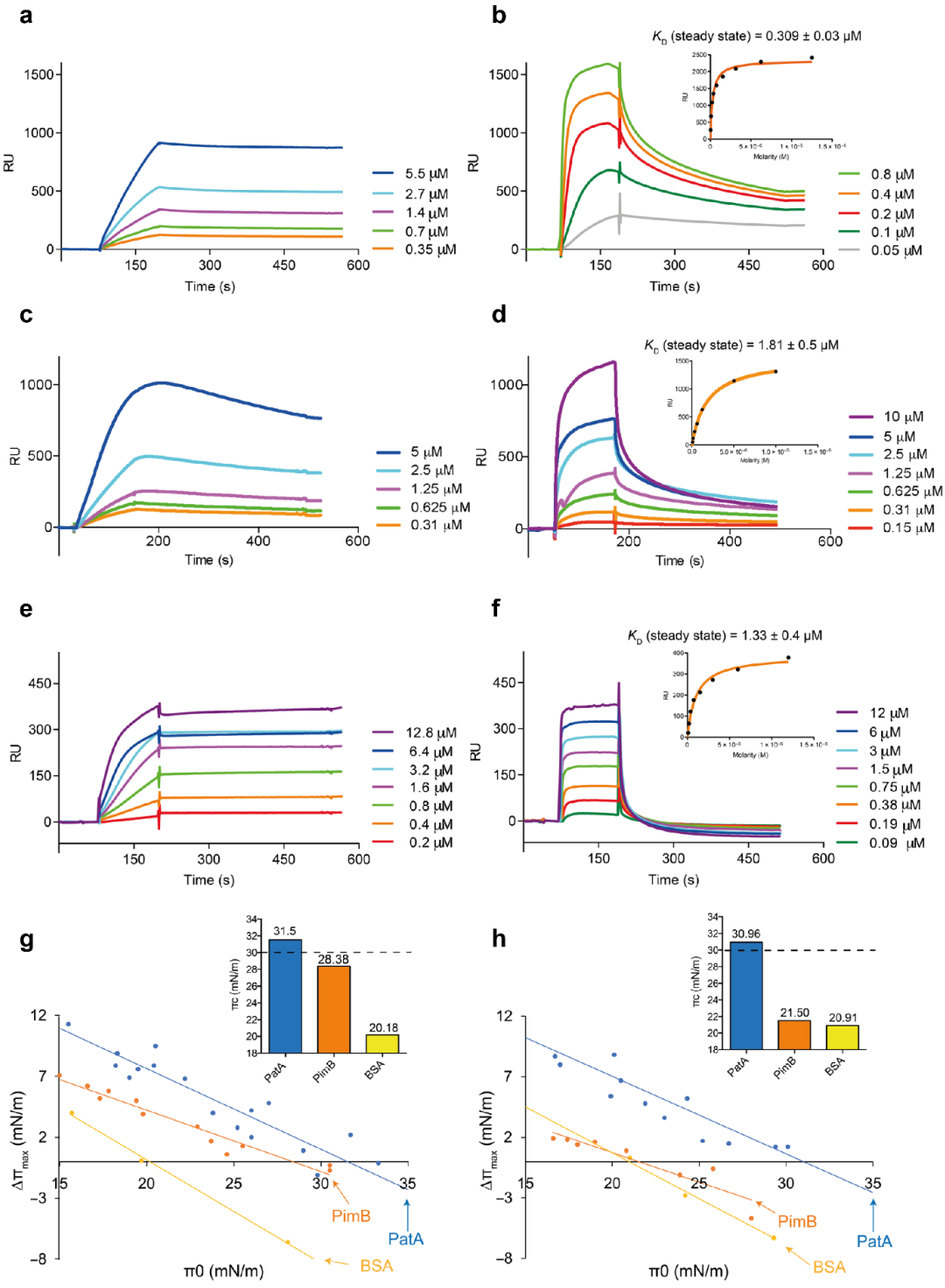

Figure 49: PatA interaction and insertion with different lipid membrane models. SPR sensograms of increasing concentrations of PatA with captured DOPG-SUVs at 25 ºC **a**, DMPG-SUVs at 30 ºC **c**, and *M. smegmatis* membrane lipid SUVs at 25 ºC **e**; and PimB with captured DOPG-SUVs at 25 ºC **b**, DMPG-SUVs at 30 ºC **d** and *M. smegmatis* membrane lipid SUVs **f**. **g, h** Surface pressure increase (Δπ) plotted as a function of the initial pressure (Δπ0) to determine the critical pressure of PatA and PimB in the presence of DOPG and DOPC, respectively. Critical pressure of protein insertion (πc) of the three proteins is represented on top-right corner of **g** and **h** pannels. The lateral biological membrane pressure is marked with a dashed black line (30 mN/m). BSA (in yellow) surface pressure data is also shown as a negative control. Figure edited from Anso *et al*.(2021).[322]

In conclusion, all these experimental data suggest that PatA is an integral membrane protein, permanently attached to the one side of the membrane.

6.2.3. Interfacial catalysis of the membrane-associated PatA

This enzyme is a paradigm of a large family of acyltransferases that work at the membrane-water interface. PatA catalyses a reaction between substrates of completely different nature: (i) the PIM2 glycolipid, which is deeply anchored in the mycobacterial IM, and (ii) water-soluble palmitoyl-CoA. It is worth noting that both PIM2 acyl chains are essential for PatA enzymatic activity, supporting the notion that they play a prominent structural role in the arrangement of the sugar moiety/moieties into the catalytic centre.[132]

Nevertheless, the understanding of the molecular mechanism of membrane association remains a challenge. Our experimental data suggest that PatA is an integral membrane protein because is permanently attached to the anionic lipid bilayers and is able to insert into anionic lipid monolayers.

Analysis of the amino acid sequence of PatA revealed the lack of a signal peptide or hydrophobic transmembrane segments, suggesting that PatA associates with only one side of the lipid bilayer, a feature typical of peripheral or monotopic integral membrane proteins.[329] The first 40 residues of PatA are not visualized in the crystal structures but are predicted to fold as two amphipathic helices, α1 and α2 (Figure 50 a,b,c,d). However, the positioning and involvement of these helices in the binding of the PIM_1/PIM_2 lipid acceptors and membrane interaction remain unknown.

The electrostatic potential of the PatA surface revealed a clear solvent-exposed area bordering the main groove containing several hydrophobic and aromatic residues interspersed with positively charged residues (Figure 50e,f). Specifically, this region comprises the amphipathic helices α3,

α4, and α8, and the connecting loop β2-α8 (residues 148 to 154; Figure 50c). In contrast, the opposite side of PatA exhibits a negatively charged surface, which would generate significant electrostatic repulsion with the anionic phospholipid bilayer. Therefore, the polar character of PatA seems to determine the correct orientation of the enzyme into the membrane (Figure 50e,f). To establish the preferred binding orientation. of full-length PatA within the phospholipid bilayer and the involvement of the helices α1 and α2, in PIM_1/PIM_2 acceptor and membrane association, we modelled helices α1, α2, and partially α3 and carried out 0.5-μs MD simulations of PatA embedded in one leaflet of the bilayer (see Materials and Methods). To this end, we used the coordinates of PatA containing (i) the C16-CoA donor molecule based on the PatA-S-C16-CoA crystal structure and (ii) the PIM_2 glycolipid molecule docked into the acceptor binding site based on the PatA-ManC16 structure (Figure 50c,d). The calculations show that the complex is stable in the membrane throughout the simulation time and that helices α1 and α2 interact with the polar fragments of the inner layer via several hydrogen bonds. In addition, the inositol moiety, one of the mannose, and the phosphate group of the PIM_2 glycolipid are involved in hydrogen bonds and hydrophobic contacts with the protein

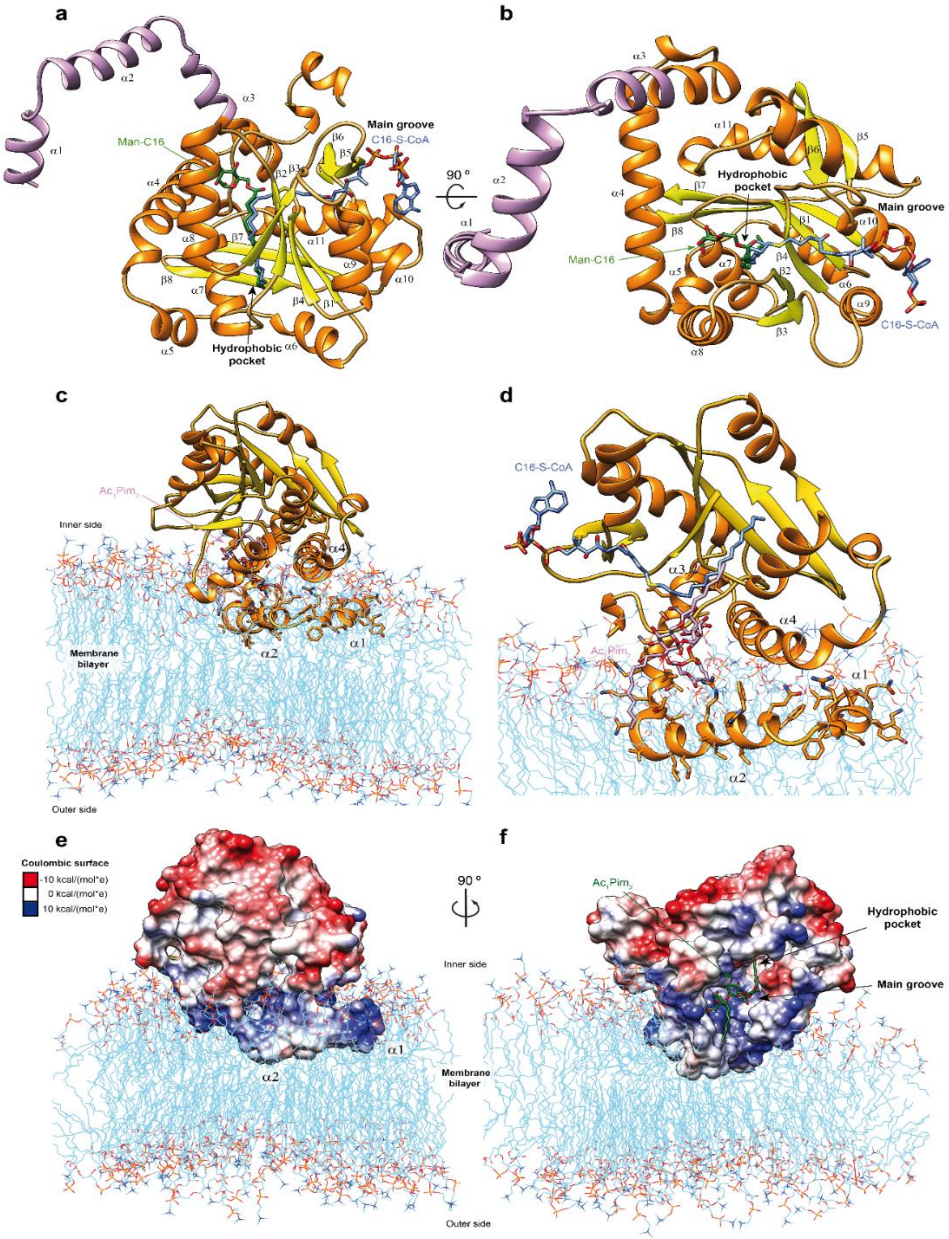

Figure 50: PatA overall structure and membrane association. a and **b** Two views of the x-ray crystal structures of PatA in complex with nonhydrolyzable S-C16-CoA [Protein Data Bank (PDB) code 5F34]. The Man-C16 product was placed by superposition of the crystal structure of PatA in complex with the Man-C16 product (PDB code 5OCE). S-C16-CoA is shown in blue, Man-C16 product in green, PatA's β-strands in yellow, and α helices in orange. The modelled α1, α2, and partial α3 are highlighted in purple. **c** Model of PatA and membrane interaction result from MD simulations. The membrane bilayer apolar face is shown in light blue, phospholipids external heads are shown in red and orange, the PatA β strands are shown in yellow, and α helices are shown in orange. The reaction product, Ac_1PIM_2, is modeled in purple. **d** A closer snapshot derived from 0.5-μs MD simulations superimposed with the nonhydrolyzable S-C16 CoA (in blue). **e** and **f** Two views of Coulombic surface representation of PatA and the membrane interaction model obtained by MD calculations. The polar character of the enzyme is shown in three colors; negatively charged patches in red (−10 kcal mol^{-1}), hydrophobic zones in white (0 kcal mol^{-1}), and positively charged residues in blue (10 kcal mol^{-1}). The modeled $A_{c1}PIM_2$ reaction product is colored in green. The palmitate chain is deeply buried in the hydrophobic pocket, the inositol and mannose rings are located in the main groove, and the acyl chains are located close to the membrane. Figure edited from Anso *et al*.(2021).[322]

Our experimental data support a model in which the reaction mediated by PatA occurs at the lipid-water (mycobacterial IM–cytosol) interface (Figure 51). PatA is permanently attached to the lipid bilayer. This interaction is mainly mediated by the insertion of N-terminal helices α1 and α2 into anionic lipid bilayers (Figure 50c,d). In addition, PatA displays several hydrophobic patches interspersed with clusters of positively charged residues adjacent to the main groove that contain the catalytic site Figure 50e,f). These hydrophobic and negatively charge residues certainly maintain and stabilize PatA into the bilayer through electrostatics and London dispersion forces with other bilayer components. In addition, the enzyme displays a negatively charged region at the opposite side of the enzyme, oriented toward the cytosol.[131,132] Therefore, the polar character of PatA also contributes to determine the correct orientation of the enzyme into the membrane.

In that context, PatA recognizes the polar head group of the PIM_1/PIM_2 glycolipid acceptor molecule in a region of the major groove mainly flanked by helix α4. The fatty acid chains of PIM_1/PIM_2 are essential for the activity of the enzyme, contributing to the mannose/s and inositol residues into the binding site.[132] Notably, PatA can sequester the acyl chain of the acyl-CoA donor substrate from the cytosol and incorporate it into a hydrophobic tunnel that extends entirely through the core of the enzyme. The adenosine 3',5'-ADP moiety of the acyl-CoA protrudes away from the groove and is exposed to bulk solvent, as observed in other acyl-CoA modifying enzymes. Thus, PatA transfers the C16 acyl group to PIM_1/PIM_2 in a two-step reaction mechanism in which His^{126} acts as a general base to deprotonate the hydroxyl group of a mannose ring in PIM_1/PIM_2, facilitating nucleophilic attack on the thioester of the acyl-CoA. The reaction is assisted by Glu^{200}, which modulates the pK_a of the central histidine as a general base or acid during the catalytic cycle, similar to the charge-relay system of serine proteases (Figure 51).

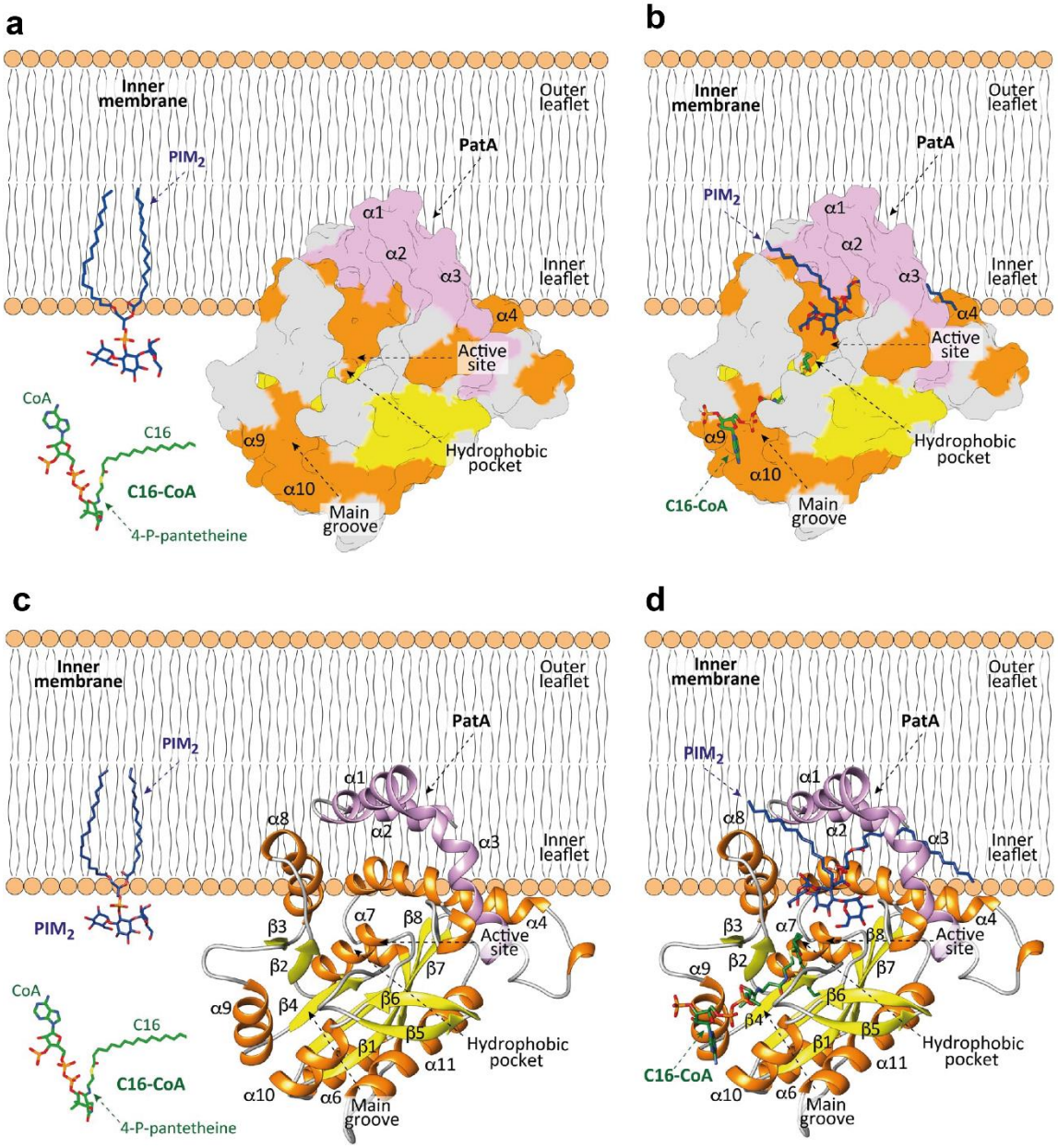

Figure 51: Model of membrane association mechanism of PatA. a and **c** The N-terminal helices α1 and α2 and the polarity of PatA determines the right orientation of the enzyme on the membrane. PIM_2 is anchored into the inner leaflet of the mycobacterial IM. C16-CoA is localized in the cytosol. **b** and **d** PatA sequesters the acyl chain of the acyl-CoA donor substrate from the cytosol and incorporates it into a hydrophobic tunnel that extends entirely through the core of the enzyme. In this orientation, the polar head of PIM_2 enters into a cavity located in the major groove to react with the acyl-CoA, at the protein-membrane interphase. The Ac_1PIM_2 product is formed and CoA-SH is released to the cytosol. Figure edited from Anso *et al*.(2021).[322]

7. SYNTHESIS OF BIOORTHOGONAL BIFUNCTIONAL PIMS ANALOGUES

There is considerable interest in developing chemical tools that allow researchers to further understand bacterial envelope dynamics and composition in its natural environment without disturbing its physiological interactions.[330] Some of the latest described strategies are synthetizing functionalized analogues containing clickable or photoactivable groups.

The biomolecule which contains a clickable group can be visualized (if the click reaction is performed with a clickable fluorophore) or fished (if the click reaction is performed with a clickable affinity tag molecule) when the click chemistry reaction is performed with a complementary click tool.

A photoactivable group, such as a diazirine group, crosslinks with a biomolecule in close vicinity due to the formation of a highly reactive carbene group after UV irradiation.[331] This strategy allows us to trap the interactions between a biomolecule and surrounding proteins in their native environment at a distance below 3 Å. To fulfill its goal, the photoactivable group must be a small, non-activated group that does not interfere with biomolecule`s activity and the physiological environment. At the same time, it is given as a modified substrate to the cells or bacterial culture. Under UV exposition, it rapidly reacts with the surrounding proteins to avoid long UV exposition to the cell and minimize its potential damage. The quick reaction time also enhances the specificity of crosslinking. Many different photoactivable groups are nicely reviewed in [332–336].

This strategy evolved into more complex reporters. Bifunctionalised analogues are biomolecules that contain both functionalized groups, a photoactivable and a clickable group within the same biomolecule. On one hand, these analogues can target the interactions between the biomolecules of interest by a photoactivable group that covalently links protein nearby the tagged molecule. On the other hand, it enables the detection or visualization of the crosslinked complex.[336,337] It was first reported by Hosoya and co-workers in 2004, who applied this techonology to the catalytic

domain of HMG-CoA reductase, the rate-limiting enzyme in cholesterol biosynthesis.[338] Since then, growing examples of bifunctionalised analogues have been synthesized.[337,339,340]

These bifunctional reporters could also be employed in the mycobacterial wall and lipids research. Thanks to the visualization by a clicked fluorophore in a clickable glycolipids we could better understood PIMs distribution and study their subcellular localization in mycobacterial membrane. Moreover, this technology could help in the identification of the set of proteins that specifically interact with PIMs, LM and LAM *in situ*, solving the identity of the unknown enzymes that participate in the biosynthesis of PIMs.

7.1. MATERIAL AND METHODS

7.1.1. Synthesis of the clickable 6-*O*-4-pentynoylmannopyranose (6-*O*-MAN)

A suspension of mannose (1) in dichloromethane (CH_2Cl_2) was treated with triethylamine (Et_3N) and trimethylsilyl chloride (TMSCl) to obtain the fully protected 1,2,3,4,6-penta-*O*-(trimethylsilyl)-mannopyranose (2). Then, to achieve the partial deprotection of the hydroxyl group in position 6 (1,2,3,4-tetra-*O*-(trimethylsilyl)-mannopyranose; (3)), the reaction product was subjected to controlled alkaline hydrolysis using a catalytic amount of potassium carbonate (K_2CO_3) in a methanol/dichloromethane (3:1) solvent mixture. To acylate the free 6 hydroxyl group of the mannose residue to obtain 6-*O*-4-pentynoyl-1,2,3,4-tetra-*O*-(trimethylsilyl)-mannopyranose (4), (3) was subjected to a DCC-mediated coupling with 4-pentynoic acid. The final product, 6-*O*-4-pentynoylmannopyranose (5) or 6-*O*-MAN was obtained upon removal of TMS protecting groups using a brief treatment with a Dowex 50WX8-200 ion exchange resin in methanol (Figure 53).

The reaction steps were followed by Thin Layer Chromatography (TLC) using 60 F_{254} silica gel plates (Figure 54) and ^{1}H and ^{13}C NMR confirmed the first and the last reaction steps products.

Depending on the reaction products' polarity, different mobile phasese were used for TLC plates development. α-naphthol staining was the chosen staining to visualize TLC plates, which contains 10.5 mL 15% ethanolic solution of 1-naphthol, 6.5 mL 97% sulfuric acid, 40.5 mL ethanol and 4 mL water.

7.1.2. Optimization of *M. smegmatis* growing conditions

A pre-inoculum of *M. smegmatis mc*^{155} transformed with a kanamycin resistance cassette plasmid and growth until the late-logarithmic phase in standard mycobacteria medium Middlebrook 7H9 supplemented with 10% ADC (albumin, dextrose, and catalase), 0.5% glycerol, and 0.05% Tween-80 and 20 mg mL^{-1} kanamycin.

Then, the growth was diluted to an OD_{600} of 0.2 in new fresh media and mixed with clickable mannose. Cells were incubated 4 h at 37 °C until the late-log phase was reached, OD_{600} 0.5-0.8. Two different media were tested; 7H9 complemented with 10% ADC, 0.5% glycerol, and 0.05% Tween-80[341,342] and 7H9 complemented with 0.5% glycerol and 0.05% Tween-80.[343] Two different clickable mannoses were also tested; the synthesized 6-*O*-MAN, (5) at 50 μM final concentration and the previously synthetized in the lab 1,3,4,6-tetra-*O*-acetyl-2-*N*-4-pentynoylmannopyranose (2-*N*-MAN) at 80 μM final concentration. After that, cells were harvested, and membrane-free lipids were extracted following standard procedures.[344]

7.1.3. Membrane lipid extraction

Membrane free lipids were extracted following standard procedures[344] and previously described in 6.1.4 section (Isolation of native lipids from *M. smegmatis* mc^{2}155) in material and methods. The lipid profile of *M. smegmatis* grown in different conditions was revealed by TLC, visualized first under UV radiation, then stained with α-naphthol staining. This is a specific staining that reveals glycolipids (Figure 55).

7.1.4. Click Chemistry

Click chemistry was performed as described before.[341,342] Dry lipids dissolved in 7 μL of chloroform were mixed in 30 μL of click chemistry reaction mixture to perform the click chemistry reaction at 42ºC in a heating block without shaking until all solvent is condenser under the lid of the tube. The reaction mixture contains 5 μL of 44 mM 3-azido-7-hydroxycroumarin in ethanol, 500 μL of 10 mM $[acetonitrile]_4CuBF_4$ in acetonitrile and 2 mL of ethanol. After removing the condensed solvent, dry lipids are dissolved in 30 μL chloroform to be visualized by TLC.

7.2. RESULTS AND DISCUSSION

7.2.1. Design and synthesis of bioorthogonal analogue

Bioorthogonal or bifunctional reporters could also be employed in the mycobacterial wall and lipids visualization and research.[345] We can find examples of this strategy in the research of mycobacterial cell envelopes such as bioorthogonal arabinose[346] and trehalose-based reporters[342,347,348] and bifunctionalized trehalose-base reporters.[349]

In this context, we aim to synthesize bifunctional analogues that are specifically incorporated into PIMs. With this approach, we aim to (i) monitor the protein-PIMs analogue interactions to further understand the biosynthetic pathway and transportation of PIMs across the cell envelope. This type of biological chemistry will enable us to identify the enzymes that participate in biosynthesis of PIMs that are still unknown. (ii) To visualize the different compounds in the mycobacterial membrane to study whether the mycobacteria membrane is heterogeneously divided into different liquid-liquid phases at the nanoscopic level and whether the glycolipids present in these membranes play an important role in liquid-liquid phase separation.

For those goals, we propose to synthesize two small and different analogues, one for each functional group: a clickable mannose residue 6-*O*-4-pentynoylmannopyranose and a 16 carbon long photoactivable fatty acid which will be photoactivable thanks to the presence of a diazirine

ring in the acyl chain (Figure 52). We hypothetized that the mycobacteria would incorporate those two analogues into its PIMs biosynthesis as natural substrates easier rather than a bifunctionalized analogue of PIMs due to the biggest size and complexity of the single bifunctionalized analogue (Figure 52).

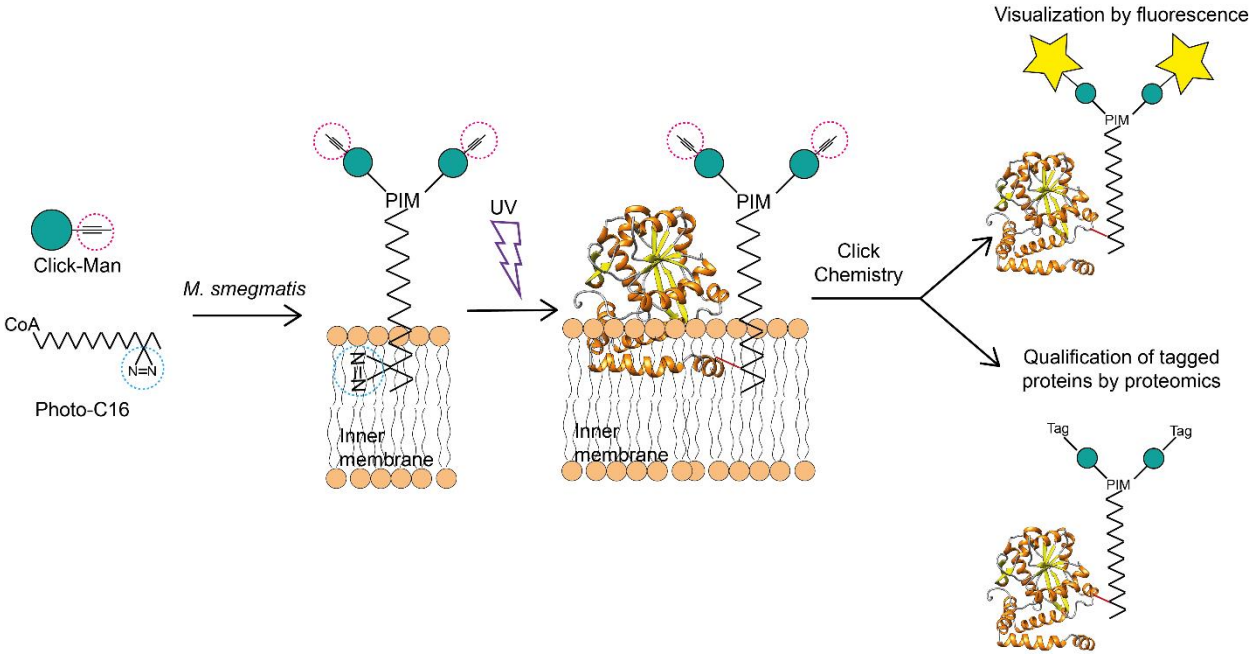

Figure 52: Bifunctional PIM biosynthesis experiment scheme. Cartoon representation of the bifunctionalized PIMs biosynthesis. *M. smegmatis* metabolizes clickable mannose (click-man) and photoactivable C16-CoA derivative (photo-C16) to obtain the bifunctionalized PIM that will be crosslinked after UV irradiation with the surrounding proteins. After that, *M. smegmatis* can be treated with different clickable fluorophores to be visualized by fluorescence microscopy or tagged by different to identify the proteins that directly intect with PIM and build PIMs' interactome.

Following a previously described synthesis of Maradolipid[350] as shown in Figure 53 and Figure 54, we synthesized 18.35 mg (0.07 mmol) of 6-*O*-MAN (see section 7.1.1. in material and methods; (5)).

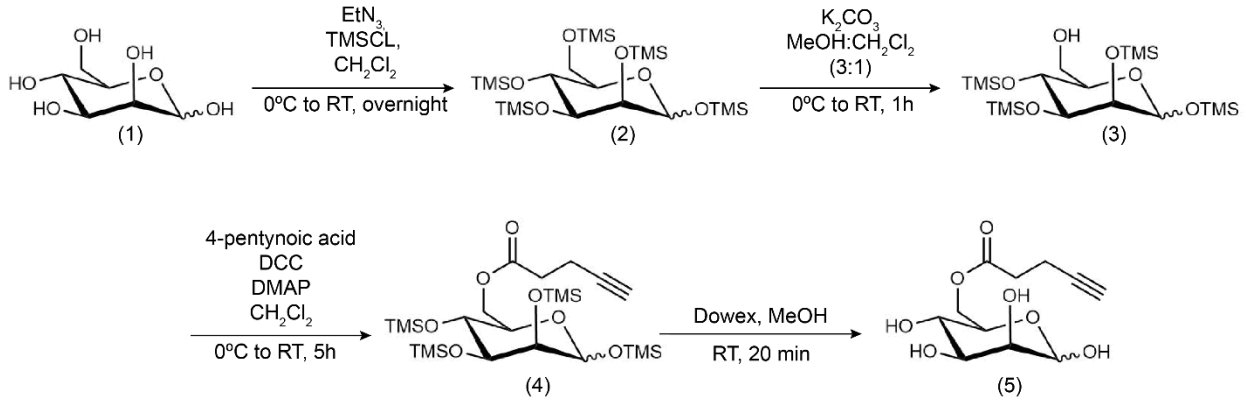

Figure 53: Scheme of 6-*O*-4-pentynoylmannopyranose (5) synthesis.

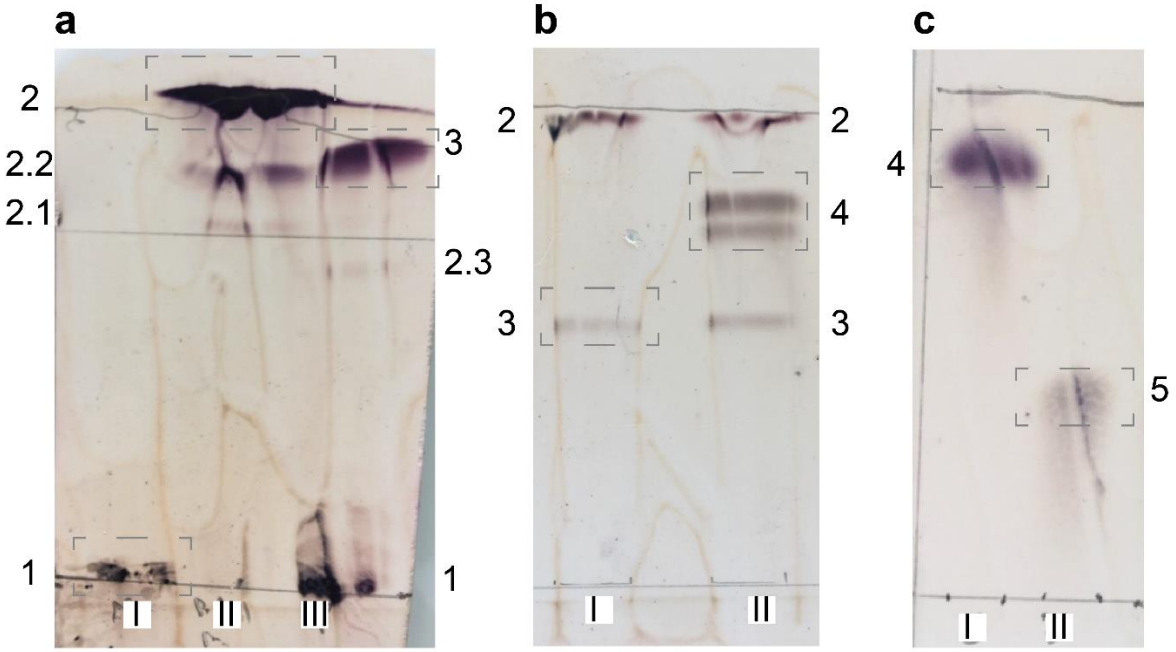

Figure 54: TLC plates of the 4 reaction steps visualized with α-naphthol. The most relevant bands are highlighted in grey dashed boxes. The bands are identified as 1, 2, 3, 4, 5 the already defined in 7.1.1 methods section reaction intermediates and final product. 2.1, 2.2, 2.3 are partially protected by TMS group mannose **a** Row I shows the first reaction step substrate, mannose. Row II shows the first reaction step product mixture and row II shows the product mixture of reaction step two. Mixtures were separated by EtAcO:petroleum ether (1:9) **b** Row I shows the third step's reactants and row II shows the reaction products mixture. EtAcO:petroleum ether (1:8) was the chosen mobile phase**. c** Row I shows the fourth reaction step starting reagent row II shows the final product of the synthesis. This TLC was separated by the mobile phase methanol:EtAcO (1:9).

7.2.2. Optimization of *M. smegmatis* growing conditions

The next step was the optimization of *M. smegmatis*, as a non-pathogenic model of *M. tuberculosis*, growing conditions in the presence of the clickable mannose residue and

experimental confirmation by TLC that is incorporated into the mycobacteria cell envelope as PIMs (Figure 52).

Following the previous work in clickable trehalose in *M. smegmatis* growth[341,342], the pre-inoculum of *M. smegmatis mc*^{155} was diluted to an OD_{600} of 0.2 in new fresh media and mixed with clickable mannose and cells were incubated until the late-log phase was reached. The tested experimental conditions were the following; two different media (7H9 growing media, 0.5% glycerol, and 0.05% Tween-80 with and without 10% ADC)[341–343] in the presence of two different clickable mannoses analogues; the synthesized 6-*O*-MAN (50 μM final concentration) and 2-*N*-MAN (80 μM final concentration). After that, cells were harvested, membrane free lipids were extracted following standard procedures,[344] and click chemistry was performed as described above (see methods section).[341,342] The lipid profile of *M. smegmatis* grown in different conditions was revealed by TLC, visualized first under UV radiation to visualize the potentially click-tagged mannose, and then stained with α-naphthol staining, a specific staining that reveals glycolipids (Figure 55).

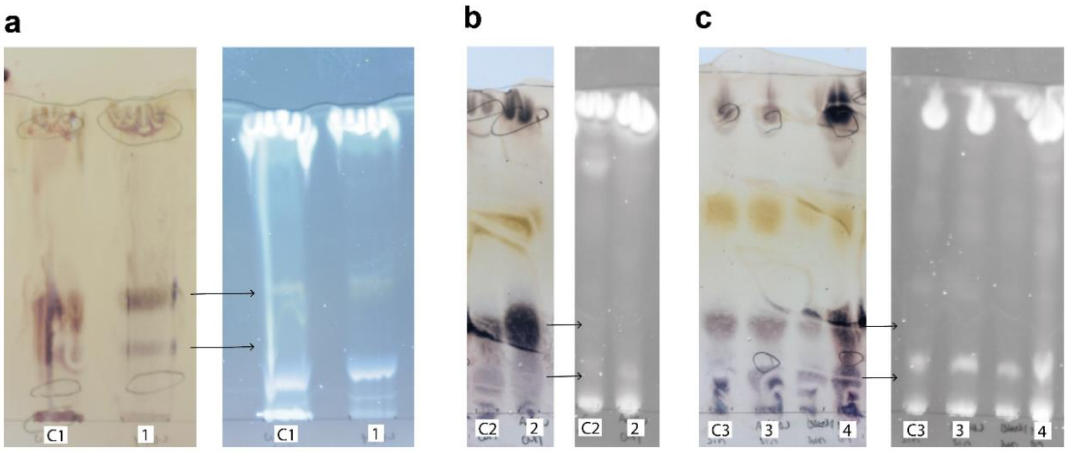

Figure 55: Lipid profile of *M. smegmatis* separated by TLC. Panels in right show the TLC stained with α-naphthol staining and panels in left show the profile revealed with UV irradiation. **a** *M. smegmatis* growth in 7H9 medium complemented with 1% ADC, 0.5% glycerol, and 0.05% Tween-80 and 20 mg mL^{-1} kanamycin and 0.1% methanol in C1 and 50 μM 6-*O*-MAN 1. **b** *M. smegmatis* growth in 7H9 medium complemented with 1% ADC, 0.5% glycerol, and 0.05% Tween-80 and 20 mg mL^{-1} kanamycin and 0.1% methanol in C2 and 80 μM 2-*N*-MAN in 2. **c** *M. smegmatis* growth in 7H9 medium complemented with 1% ADC without dextrose, 0.5% glycerol, and 0.05% Tween-80 and 20 mg mL^{-1} kanamycin and 0.1% methanol in C3, 50 μM 6-*O*-MAN in 3 and 80 μM 2-*N*-MAN in 4.

As shown in Figure 55, there is no difference between the lipid profile of non-treated (negative control) and clickable mannose-treated mycobacteria samples, neither on the α-naphthol stained profile nor in the UV-revealed profile. Moreover, there is no coincidence in bands between the sample's UV and α-naphthol stained profiles, meaning that the mycobacteria have not metabolised and introduced the modified mannose into its glycolipids. The bands in the UV profile could be due to free fluorophore or its aggregates.

There are still many unknowns regarding to this project due to its youth. I've been involved in the design and proposal of this long-term project and I started working in the very first steps. We achieved the first goal, the synthesis of the clickable mannose component. However, deeper research and experimental work is needed to further understand the mannose metabolism of *M. smegmatis* to find the correct conditions in which the mycobacteria obtains the clickable mannose from external sources.

8. GENERAL CONCLUSIONS

The conclusions of this thesis, therefore, are the following:

1-Mucin degradation enzymes machinery of human gut commensal *Akkermansia muciniphila:*

-We have identified a potential genes cluster in the human commensal *A. muciniphila*, conserved along the different *Akkermansia* species, which encloses genes with different encoding enzymes that could be involved in mucin degradation.

-That potential gene cluster that neighbores *ogpA* encodes proteins potentially involved in mucin degradation such as glycoside hydrolases and sulfatases, suggesting that these enzymes could work together with others in a common mechanism.

-We elucidated molecular mechanism by which OgpA from *A. muciniphila* specifically recognizes and processes *O*-glycopeptides at the molecular level detail.

-The crystal structures of unliganded OgpA and in complex with the substrate and product, together with activity assays and docking calculations, revealed that OgpA can process specifically core 1 *O*-glycosilated peptides and decreases its ability to recognize the *O*-glycosilated core in core 2, core 3, Tn antigen and sialylated core 1. Structural analysis confirm that only core 1 *O*-glycan substrate fulfils the requested interactions with the enzyme in the binding pocket. However, the crystal structure of $OgpA_{H205A/D206A\text{-}GD\text{-}SUB}$ shows that protein-peptide interaction is primarily mediated by hydrogen bonds between the backbone and the residues of OgpA that form the corresponding subsites, suggesting that the side chain nature of these residues is not a key element for OgpA activity.

- OgpA could be considered as the founding member of a new family of peptidases due to the inexistence of a family of Zn dependent proteases where this enzyme could belong to. The structural homologues according to DALI server are Zn-dependent metalloproteases

with significant low structural similarity, and sequence homology search run by MEROPS data base annotates as M11 or M12 families of proenzymes that require activation by limited proteolysis.

-We have discovered a highly specific α-1,2-fucosidase, FucOB, able to remove fucose from type I, type II, and type V H antigens but not into type II A or B blood antigens.

- FucOB crystallized as a monomer comprising three domains a β-sandwich domain, an $(\alpha/\alpha)_6$ helical barrel catalytic domain a second β-sandwich domain connected to the catalytic helical barrel domain through a linker comprised of five α-helices. This folding shows similar structural features to GH95 family, a single displacement inverting catalytic mechanism α-1,2-L-fucosidase and α-L-galactosidase family of GHs with a highly conserved active site.

- We explained the mechanism of substrate specificity of FucOB by determination of the X-ray crystal structures of unliganded FucOB and the inactive mutant $FucOB_{E541A}$, computational methods and single point directed mutagenesis.

-We showed that FucOB is able to remove the fucose from the H antigen present in RCBs surface, converting the O blood group into the rare Bombay group by agglutination and flow cytometry experiments.

-FucOB conversion of the O blood group into Bombay group was consistent in a large number O blood group individuals. Moreover, it does not alter the RBCs's viability neither the morphology (shown in blood smears and G6DP experiments).

-We propose FucOB as a potential biotechnological tool to treat O group blood and facilitate its transfusion to Bombay individuals because of enzyme's high substrate specificity by H antigen, its production in high yields and purity, and its reaction efficiency.

2- Biosynthesis of mannolipids present in the complex cell envelope of pathogenic *M. tuberculosis.*

- PatA is an acyltransferase able to recognize the acyl chain length of the donor substrate based on a ruler mechanism. LC-MS detection, hydrolytic DNTB, and ITC assays show that PatA is able to transfer different acyl chain lengths CoA derivates, showing a Gaussian-type curve profile with a greater transfer ability against C14 and C16 chain length derivates and minor transfer activity with longer or shorter acyl chains. The length of the hydrophobic groove in PatA measures where the donor binding site is located, act as a ruler, and determines the acyl chain specificity.

- SPR and Langmuir balance assays all support the notion that PatA is permanently attached to anionic mono- or bi- lipid layers.

- Our experimental data support an interfacial catalysis model, where PatA is able to allow the reaction between water-soluble acyl derivatives and membrane-associated PIMs. PatA is permanently attached to the membrane thanks to its polar character and three structural alpha helices located in the N-terminal domain. In this context, PatA's active site is correctly oriented against the PIM1/PIM2 polar head group buried into the membrane and the cytosolic C16-CoA.

- The synthesis of bifunctionalised PIMs analogues is an on-going project in which we have designed a two-functionalised metabolites strategy to identify the unknown enzymes that participate in the biosynthesis of PIMs on the one hand, and to better understand the glycolipids distribution in the membrane on the other hand. In that context, we have synthetised the first functionalised metabolite, a bioorthogonal mannose analogue 6-*O*-MAN and we have started with the growing conditions of *M. smegmatis* in presence of 6-*O*-MAN.

9. BIBLIOGRAPHY

1. McFall-Ngai, M. Adaptive immunity: Care for the community. *Nature* vol. 445 (2007).

2. Curtis, M. M. & Sperandio, V. A complex relationship: The interaction among symbiotic microbes, invading pathogens, and their mammalian host. *Mucosal Immunology* vol. 4 (2011).

3. Martens, E. C., Neumann, M. & Desai, M. S. Interactions of commensal and pathogenic microorganisms with the intestinal mucosal barrier. *Nature Reviews Microbiology* vol. 16 (2018).

4. Gonçalves, S. M. *et al.* The microbiome-metabolome crosstalk in the pathogenesis of respiratory fungal diseases. *Virulence* vol. 8 (2017).

5. Sommer, F. & Bäckhed, F. The gut microbiota-masters of host development and physiology. *Nature Reviews Microbiology* vol. 11 227–238 (2013).

6. Postler, T. S. & Ghosh, S. Understanding the Holobiont: How Microbial Metabolites Affect Human Health and Shape the Immune System. *Cell Metabolism* vol. 26 (2017).

7. Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The impact of the gut microbiota on human health: An integrative view. *Cell* vol. 148 (2012).

8. Günther, J. & Seyfert, H. M. The first line of defence: insights into mechanisms and relevance of phagocytosis in epithelial cells. *Seminars in Immunopathology* vol. 40 (2018).

9. Hansson, G. C. Mucins and the Microbiome. *Annual Review of Biochemistry* vol. 89 (2020).

10. Corfield, A. P. Mucins: A biologically relevant glycan barrier in mucosal protection. *Biochimica et Biophysica Acta - General Subjects* vol. 1850 236–252 (2015).

11. Johansson, M. E. V., Sjövall, H. & Hansson, G. C. The gastrointestinal mucus system in health and disease. *Nature Reviews Gastroenterology and Hepatology* vol. 10 352–361 (2013).

12. Martens, E. C., Neumann, M. & Desai, M. S. Interactions of commensal and pathogenic microorganisms with the intestinal mucosal barrier. *Nature Reviews Microbiology* vol. 16 (2018).

13. Derrien, M. *et al.* Mucin-bacterial interactions in the human oral cavity and digestive tract. *Gut Microbes* **1**, (2010).

14. Bansil, R. & Turner, B. S. Mucin structure, aggregation, physiological functions and biomedical applications. *Current Opinion in Colloid and Interface Science* vol. 11 (2006).

15. Suriano, F., Nyström, E. E. L., Sergi, D. & Gustafsson, J. K. Diet, microbiota, and the mucus layer: The guardians of our health. *Frontiers in Immunology* vol. 13 (2022).

16. Ouwerkerk, J. P., de Vos, W. M. & Belzer, C. Glycobiome: Bacteria and mucus at the epithelial interface. *Best Practice and Research: Clinical Gastroenterology* vol. 27 25–38 (2013).

17. Wagner, C. E., Wheeler, K. M. & Ribbeck, K. Mucins and Their Role in Shaping the Functions of Mucus Barriers. *Annu Rev Cell Dev Biol* **34**, (2018).

18. Thornton, D. J., Rousseau, K. & McGuckin, M. A. Structure and function of the polymeric mucins in airways mucus. *Annual Review of Physiology* vol. 70 (2008).

19. Paone, P. & Cani, P. D. Mucus barrier, mucins and gut microbiota: The expected slimy partners? *Gut* vol. 69 (2020).

20. Tailford, L. E., Crost, E. H., Kavanaugh, D. & Juge, N. Mucin glycan foraging in the human gut microbiome. *Front Genet* **6**, 81 (2015).

21. Varki, A. *et al.* Symbol nomenclature for graphical representations of glycans. *Glycobiology* **25**, (2015).

22. Patenaude, S. I. *et al.* The structural basis for specificity in human abo(h) blood group biosynthesis. *Nat Struct Biol* **9**, 658–690 (2002).

23. Henry, S., Oriol, R. & Samuelsson, B. Lewis Histo-Blood Group System and Associated Secretory Phenotypes. *Vox Sanguinis* vol. 69 (1995).

24. Petrou, G. & Crouzier, T. Mucins as multifunctional building blocks of biomaterials. *Biomaterials Science* vol. 6 (2018).

25. Hrncir, T. Gut Microbiota Dysbiosis: Triggers, Consequences, Diagnostic and Therapeutic Options. *Microorganisms* vol. 10 (2022).

26. Yu, J. & Cheon, J. H. Microbial Modulation in Inflammatory Bowel Diseases. *Immune Network* vol. 22 (2022).

27. Xu, X. *et al.* The bridge of the gut–joint axis: Gut microbial metabolites in rheumatoid arthritis. *Frontiers in Immunology* vol. 13 (2022).

28. del Chierico, F. *et al.* Pathophysiology of Type 1 Diabetes and Gut Microbiota Role. *International Journal of Molecular Sciences* vol. 23 (2022).

29. Yousefi, B. *et al.* Gastrointestinal Tract, Microbiota and Multiple Sclerosis (MS) and the Link Between Gut Microbiota and CNS. *Current Microbiology* vol. 80 (2023).

30. Mansuri, N. M. *et al.* Role of Gut Microbiome in Cardiovascular Events: A Systematic Review. *Cureus* (2022).

31. Zhan, Y., Liu, Q., Zhang, B., Huang, X. & Lu, Q. Recent advances in systemic lupus erythematosus and microbiota: from bench to bedside. *Frontiers of Medicine* (2022).

32. Yang, H. J. & Kim, J. H. Role of microbiome and its metabolite, short chain fatty acid in prostate cancer. *Investigative and clinical urology* vol. 64 3–12 (2023).

33. Niekamp, P. & Kim, C. H. Microbial Metabolite Dysbiosis and Colorectal Cancer. *Gut Liver* (2023).

34. Lin, B. *et al.* Gut microbiota in brain tumors: An emerging crucial player. *CNS Neurosci Ther* (2023).

35. Asili, P. *et al.* The Association of Oral Microbiome Dysbiosis with Gastrointestinal Cancers and Its Diagnostic Efficacy. *Journal of Gastrointestinal Cancer* (2023).

36. Martinez, J. E. *et al.* Unhealthy Lifestyle and Gut Dysbiosis: A Better Understanding of the Effects of Poor Diet and Nicotine on the Intestinal Microbiome. *Frontiers in Endocrinology* vol. 12 (2021).

37. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: Human gut microbes associated with obesity. *Nature* **444**, (2006).

38. Karlsson, F., Tremaroli, V., Nielsen, J. & Bäckhed, F. Assessing the human gut microbiota in metabolic diseases. *Diabetes* vol. 62 (2013).

39. Kumar, H. *et al.* Gut microbiota as an epigenetic regulator: Pilot study based on whole-genome methylation analysis. *mBio* **5**, (2014).

40. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* **32**, (2014).

41. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, (2010).

42. Hugon, P. *et al.* A comprehensive repertoire of prokaryotic species identified in human beings. *The Lancet Infectious Diseases* vol. 15 (2015).

43. Methé, B. A. *et al.* A framework for human microbiome research. *Nature* **486**, (2012).

44. Proctor, L. M. *et al.* The Integrative Human Microbiome Project. *Nature* **569**, (2019).

45. Thursby, E. & Juge, N. Introduction to the human gut microbiota. *Biochemical Journal* **474**, 1823–1836 (2017).

46. Derrien, M., Vaughan, E. E., Plugge, C. M. & de Vos, W. M. Akkermansia municiphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *Int J Syst Evol Microbiol* **54**, (2004).

47. Derrien, M. Mucin utilisation and Host Interactions of the Novel Intestinal Microbe Akkermansia muciniphila. *Narcisinfo* (Wageningen University, Wageningen, The Netherlands, 2007).

48. Derrien, M., Collado, M. C., Ben-Amor, K., Salminen, S. & de Vos, W. M. The mucin degrader Akkermansia muciniphila is an abundant resident of the human intestinal tract. *Appl Environ Microbiol* **74**, 1646–1648 (2008).

49. Effendi, R. M. R. A. *et al.* Akkermansia muciniphila and Faecalibacterium prausnitzii in Immune-Related Diseases. *Microorganisms* vol. 10 (2022).

50. Derrien, M., Belzer, C. & de Vos, W. M. Akkermansia muciniphila and its role in regulating host functions. *Microbial Pathogenesis* vol. 106 (2017).

51. DAS, B. & Nair, G. B. Homeostasis and dysbiosis of the gut microbiome in health and disease. *Journal of biosciences* vol. 44 (2019).

52. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* vol. 474 (2011).

53. Ursell, L. K. & Knight, R. Xenobiotics and the human gut microbiome: Metatranscriptomics reveal the active players. *Cell Metabolism* vol. 17 (2013).

54. Chung, H. *et al.* Gut immune maturation depends on colonization with a host-specific microbiota. *Cell* **149**, (2012).

55. Hooper, L. v. & MacPherson, A. J. Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nature Reviews Immunology* vol. 10 (2010).

56. Bouskra, D. *et al.* Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis. *Nature* **456**, (2008).

57. Kaoutari, A. el, Armougom, F., Gordon, J. I., Raoult, D. & Henrissat, B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat Rev Microbiol* **11**, (2013).

58. Wardman, J. F., Bains, R. K., Rahfeld, P. & Withers, S. G. Carbohydrate-active enzymes (CAZymes) in the gut microbiome. *Nature Reviews Microbiology* vol. 20 (2022).

59. Henrissat, B. & Bairoch, A. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochemical Journal* **293**, (1993).

60. Lombard, V. *et al.* A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochemical Journal* **432**, (2010).

61. Martens, E. C. *et al.* Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol* **9**, (2011).

62. Willis, C. L., Cummings, J. H., Neale, G. & Gibson, G. R. In vitro effects of mucin fermentation on the growth of human colonic sulphate-reducing bacteria. *Anaerobe* **2**, (1996).

63. Raimondi, S., Musmeci, E., Candeliere, F., Amaretti, A. & Rossi, M. Identification of mucin degraders of the human gut microbiota. *Sci Rep* **11**, (2021).

64. Glover, J. S., Ticer, T. D. & Engevik, M. A. Characterizing the mucin-degrading capacity of the human gut microbiota. *Sci Rep* **12**, (2022).

65. González-Morelo, K. J., Vega-Sagardía, M. & Garrido, D. Molecular Insights Into O-Linked Glycan Utilization by Gut Microbes. *Frontiers in Microbiology* vol. 11 (2020).

66. Xu, J. *et al.* A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science (1979)* **299**, (2003).

67. Bjursell, M. K., Martens, E. C. & Gordon, J. I. Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period. *Journal of Biological Chemistry* **281**, (2006).

68. Anderson, K. L. & Salyers, A. A. Biochemical evidence that starch breakdown by Bacteroides thetaiotaomicron involves outer membrane starch-binding sites and periplasmic starch-degrading enzymes. *J Bacteriol* **171**, (1989).

69. Anderson, K. L. & Salyers, A. A. Genetic evidence that outer membrane binding of starch is required for starch utilization by Bacteroides thetaiotaomicron. *J Bacteriol* **171**, (1989).

70. Kmezik, C., Krska, D., Mazurkewich, S. & Larsbrink, J. Characterization of a novel multidomain CE15-GH8 enzyme encoded by a polysaccharide utilization locus in the human gut bacterium Bacteroides eggerthii. *Sci Rep* **11**, (2021).

71. Lapébie, P., Lombard, V., Drula, E., Terrapon, N. & Henrissat, B. Bacteroidetes use thousands of enzyme combinations to break down glycans. *Nat Commun* **10**, (2019).

72. Ndeh, D. *et al.* Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature* **544**, (2017).

73. Briliūtė, J. *et al.* Complex N-glycan breakdown by gut Bacteroides involves an extensive enzymatic apparatus encoded by multiple co-regulated genetic loci. *Nat Microbiol* **4**, (2019).

74. Xu, J. *et al.* Evolution of Symbiotic Bacteria in the Distal Human Intestine. *PLoS Biol* **5**, (2007).

75. Flint, H. J., Bayer, E. A., Rincon, M. T., Lamed, R. & White, B. A. Polysaccharide utilization by gut bacteria: Potential for new insights from genomic analysis. *Nature Reviews Microbiology* vol. 6 (2008).

76. Grondin, J. M., Tamura, K., Déjean, G., Abbott, D. W. & Brumer, H. Polysaccharide utilization loci: Fueling microbial communities. *Journal of Bacteriology* vol. 199 (2017).

77. Terrapon, N., Lombard, V., Gilbert, H. J. & Henrissat, B. Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics* **31**, (2015).

78. Terrapon, N. *et al.* PULDB: The expanded database of Polysaccharide Utilization Loci. *Nucleic Acids Res* **46**, (2018).

79. Daniels, G. The molecular definition of red cell antigens. *International Society of Blood Transfusion, ISBT Science Series* **5**, 300–302 (2010).

80. Reid, M. E., Lomas-Francis, C. & Olsson, M. L. *The Blood Group Antigen*. *The Blood Group Antigen* (Elsevier Ltd, 2012).

81. Daniels, G. *Human Blood Groups*. *Human Blood Groups: 3rd edition* (Wiley-Blackwell, 2013).

82. Landsteiner, K. On Agglutination of Normal Human Blood. *Transfusion (Paris)* **1**, (1961).

83. Kabat, E. A. *Blood group substances—their chemistry and immunochemistry*. *Journal of the American Pharmaceutical Association (Scientific ed.)* vol. 45 (1956).

84. Watkins, W. M. & Morgan, W. T. J. Specific inhibition studies relating to the Lewis blood-group system. *Nature* **180**, 1038–1040 (1957).

85. Quraishy, N. & Sapatnekar, S. Advances in Blood Typing. in *Advances in Clinical Chemistry* vol. 77 221–269 (2016).

86. Stanley, P. & Cummings, R. D. *Structures Common to Different Glycans*. *Essentials of Glycobiology* (2015).

87. Jajosky, R. P. *et al.* iScience ABO blood group antigens and differential glycan expression: Perspective on the evolution of common human enzyme deficiencies. (2023).

88. Anso, I. *et al.* Turning universal O into rare Bombay type blood. *Nat Commun* **14**, 1765 (2023).

89. Hosoi, E. Biological and clinicel aspects of ABO blood group system. *Journal of Medical Investigation* **55**, 174–182 (2008).

90. Oriol, R. Genetic control of the fucosylation of ABH precursor chains. Evidence for new epistatic interactions in different cells and tissues. *Int J Immunogenet* **17**, 235–245 (1990).

91. Alfaro, J. A. *et al.* ABO(H) blood group A and B glycosyltransferases recognize substrate via specific conformational changes. *Journal of Biological Chemistry* **283**, 10097–10108 (2008).

92. Albesa-Jové, D., Sainz-Polo, M. Á., Marina, A. & Guerin, M. E. Structural Snapshots of α-1,3-Galactosyltransferase with Native Substrates: Insight into the Catalytic Mechanism of Retaining Glycosyltransferases. *Angewandte Chemie* **129**, (2017).

93. Scharberg, E. A., Olsen, C. & Bugert, P. The H blood group system. *Immunohematology* **32**, 112–118 (2016).

94. BHENDE, Y. M. *et al.* A 'new' blood group character related to the ABO system. *Lancet* **1**, 903–904 (1952).

95. Bhatia, H. M. The 'Bombay' (Oh) Blood Group. *Vox Sang* **52**, 1–2 (1987).

96. Kelly, R. J. *et al.* Molecular basis for H blood group deficiency in Bombay (O(h)) and para- Bombay individuals. *Proc Natl Acad Sci U S A* **91**, 5843–5847 (1994).

97. Shahverdi, E. *et al.* The First Comprehensive Study of H-Deficient Phenotypes in Iran. *Transfusion Medicine and Hemotherapy* **46**, 376–380 (2019).

98. Balgir, R. Identification of a rare blood group, 'Bombay (Oh) phenotype,' in Bhuyan tribe of Northwestern Orissa, India. *Indian J Hum Genet* **13**, 109–113 (2007).

99. Mallick, S., Kotasthane, D. S., Chowdhury, P. S. & Sarkar, S. Bombay blood group: Is prevalence decreasing with urbanization and the decreasing rate of consanguineous marriage. *Asian J Transfus Sci* **9**, 129–132 (2015).

100. Dulberger, C. L., Rubin, E. J. & Boutte, C. C. The mycobacterial cell envelope — a moving target. *Nat Rev Microbiol* **18**, (2020).

101. Cambier, C. J., Falkow, S. & Ramakrishnan, L. Host evasion and exploitation schemes of Mycobacterium tuberculosis. *Cell* vol. 159 (2014).

102. Silhavy, T. J., Kahne, D. & Walker, S. The bacterial cell envelope. *Cold Spring Harb Perspect Biol* **2**, a000414 (2010).

103. Angala, S. K., Belardinelli, J. M., Huc-Claustre, E., Wheat, W. H. & Jackson, M. The cell envelope glycoconjugates of Mycobacterium tuberculosis. *Crit Rev Biochem Mol Biol* **49**, 361–399 (2014).

104. Jankute, M., Cox, J. A. G., Harrison, J. & Besra, G. S. Assembly of the Mycobacterial Cell Wall. *Annu Rev Microbiol* **69**, 405-423. (2015).

105. Bendre, A. D., Peters, P. J. & Kumar, J. Recent Insights into the Structure and Function of Mycobacterial Membrane Proteins Facilitated by Cryo-EM. *Journal of Membrane Biology* vol. 254 (2021).

106. Jackson, M. The mycobacterial cell envelope-lipids. *Cold Spring Harb Perspect Med* **4**, a021105 (2014).

107. Rahlwes, K. C., Sparks, I. L. & Morita, Y. S. Cell Walls and Membranes of Actinobacteria. in *Subcellular Biochemistry* (ed. Kuhn, A.) vol. 92 417–469 (Springer US, 2019).

108. Hayashi, J. M. *et al.* Spatially distinct and metabolically active membrane domain in mycobacteria. *Proc Natl Acad Sci U S A* **113**, 5400–5405 (2016).

109. Hayashi, J. M. *et al.* Stress-induced reorganization of the mycobacterial membrane domain. *mBio* **9**, e01823-17 (2018).

110. Hayashi, J. M. & Morita, Y. S. Mycobacterial membrane domain, or a primordial organelle? *Yale Journal of Biology and Medicine* **92**, 549–556 (2019).

111. Puffal, J., Mayfield, J. A., Moody, D. B. & Morita, Y. S. Demethylmenaquinone Methyl Transferase Is a Membrane Domain-Associated Protein Essential for Menaquinone Homeostasis in Mycobacterium smegmatis. *Front Microbiol* **9**, 3145 (2018).

112. Ortalo-Magne, A. *et al.* Molecular composition of the outermost capsular material of the tubercle bacillus. *Microbiology (N Y)* **141**, 1609–1620 (1995).

113. Lemassu, A. *et al.* Extracellular and surface-exposed polysaccharides of non-tuberculous mycobacteria. *Microbiology (N Y)* **142**, 1513–1520 (1996).

114. Kalscheuer, R. *et al.* The Mycobacterium tuberculosis capsule: A cell structure with key implications in pathogenesis. *Biochemical Journal* **476**, 1995–2016 (2019).

115. Guerin, M. E., Korduláková, J., Alzari, P. M., Brennan, P. J. & Jackson, M. Molecular Basis of Phosphatidyl-myo-inositol Mannoside Biosynthesis and Regulation in Mycobacteria. *Journal of Biological Chemistry* **285**, 33577–33583 (2010).

116. Brennan, P. J. & Nikaido, H. The envelope of mycobacteria. *Annual Review of Biochemistry* vol. 64 (1995).

117. BALLOU, C. E., VILKAS, E. & LEDERER, E. Structural studies on the myo-inositol phospholipids of Mycobacterium tuberculosis (var. bovis, strain BCG). *J Biol Chem* **238**, (1963).

118. Ballou, C. E. & Lee, Y. C. The Structure of a Myoinositol Mannoside from Mycobacterium Tuberculosis Glycolipid. *Biochemistry* **3**, (1964).

119. Gilleron, M., Lindner, B. & Puzo, G. MS/MS approach for characterization of the fatty acid distribution on mycobacterial phosphatidyl-myo-inositol mannosides. *Anal Chem* **78**, 8543–8548 (2006).

120. Gilleron, M. *et al.* Acylation State of the Phosphatidylinositol Mannosides from Mycobacterium bovis Bacillus Calmette Guérin and Ability to Induce Granuloma and Recruit Natural Killer T Cells. *Journal of Biological Chemistry* **276**, 34896–34904 (2001).

121. Gilleron, M., Quesniaux, V. F. J. & Puzo, G. Acylation state of the phosphatidylinositol hexamannosides from Mycobacterium bovis bacillus Calmette Guérin and Mycobacterium tuberculosis H37Rv and its implication in toll-like receptor response. *Journal of Biological Chemistry* **278**, 29880–29889 (2003).

122. Morita, Y. S. *et al.* Inositol lipid metabolism in mycobacteria: Biosynthesis and regulatory mechanisms. *Biochim Biophys Acta Gen Subj* **1810**, 630–641 (2011).

123. Berg, S., Kaur, D., Jackson, M. & Brennan, P. J. The glycosyltransferases of Mycobacterium tuberculosis - Roles in the synthesis of arabinogalactan, lipoarabinomannan, and other glycoconjugates. *Glycobiology* **17**, 35-56R (2007).

124. Korduláková, J. *et al.* Definition of the first mannosylation step in phosphatidylinositol mannoside synthesis: PimA is essential for growth of mycobacteria. *Journal of Biological Chemistry* **277**, 31335–31344 (2002).

125. Giganti, D. *et al.* Secondary structure reshuffling modulates glycosyltransferase function at the membrane. *Nat Chem Biol* **11**, 16–18 (2015).

126. Guerin, M. E., Jackson, M., Buschiazzo, A., Alzari, P. M. & Korduláková, J. Crystallization and preliminary crystallographic analysis of PimA, an essential mannosyltransferase from Mycobacterium smegmatis. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **61**, 518–520 (2005).

127. Guerin, M. E. *et al.* Molecular recognition and interfacial catalysis by the essential phosphatidylinositol mannosyltransferase PimA from mycobacteria. *Journal of Biological Chemistry* **282**, 20705–20714 (2007).

128. Guerin, M. E. *et al.* New insights into the early steps of phosphatidylinositol mannoside biosynthesis in mycobacteria: PimB′ is an essential enzyme of Mycobacterium smegmatis. *Journal of Biological Chemistry* **284**, 25687–25696 (2009).

129. Batt, S. M. *et al.* Acceptor substrate discrimination in phosphatidyl-myo-inositol mannoside synthesis: Structural and mutational analysis of mannosyltransferase corynebacterium glutamicum PimB′. *Journal of Biological Chemistry* **285**, 37741–37752 (2010).

130. Svetlíková, Z., Baráth, P., Jackson, M., Korduláková, J. & Mikušová, K. Purification and characterization of the acyltransferase involved in biosynthesis of the major mycobacterial cell envelope glycolipid - Monoacylated phosphatidylinositol dimannoside. *Protein Expr Purif* **100**, 33–39 (2014).

131. Albesa-Jové, D. *et al.* Structural basis for selective recognition of acyl chains by the membrane-associated acyltransferase PatA. *Nat Commun* **7**, 10906 (2016).

132. Tersa, M. *et al.* The Molecular Mechanism of Substrate Recognition and Catalysis of the Membrane Acyltransferase PatA from Mycobacteria. *ACS Chem Biol* **13**, 131–140 (2018).

133. Jackson, M., Crick, D. C. & Brennan, P. J. Phosphatidylinositol is an essential phospholipid of mycobacteria. *Journal of Biological Chemistry* **275**, 30092–30099 (2000).

134. Boldrin, F. *et al.* The phosphatidyl-myo-inositol mannosyltransferase PimA is essential for Mycobacterium tuberculosis growth in vitro and in vivo. *J Bacteriol* **196**, 3441–3451 (2014).

135. Kordulakova, J. *et al.* Identification of the required acyltransferase step in the biosynthesis of the phosphatidylinositol mannosides of mycobacterium species. *J Biol Chem* **278**, 36285–36295 (2003).

136. Boldrin, F. *et al.* The Phosphatidyl-myo-Inositol Dimannoside Acyltransferase PatA Is Essential for Mycobacterium tuberculosis Growth In Vitro and In Vivo. *J Bacteriol* **203**, e00439-20 (2021).

137. Kremer, L. *et al.* Characterization of a putative α-mannosyltransferase involved in phosphatidylinositol trimannoside biosynthesis in Mycobacterium tuberculosis. *Biochemical Journal* **363**, (2002).

138. Morita, Y. S. *et al.* PimE is a polyprenol-phosphate-mannose-dependent mannosyltransferase that transfers the fifth mannose of phosphatidylinositol mannoside in mycobacteria. *Journal of Biological Chemistry* **281**, 25143–25155 (2006).

139. Mishra, A. K., Driessen, N. N., Appelmelk, B. J. & Besra, G. S. Lipoarabinomannan and related glycoconjugates: Structure, biogenesis and role in Mycobacterium tuberculosis physiology and host-pathogen interaction. *FEMS Microbiology Reviews* vol. 35 1126–1157 (2011).

140. Liu, J. & Mushegian, A. Three monophyletic superfamilies account for the majority of the known glycosyltransferases. *Protein Science* **12**, (2003).

141. Crellin, P. K. *et al.* Mutations in pimE restore lipoarabinomannan synthesis and growth in a Mycobacterium smegmatis lpqW mutant. *J Bacteriol* **190**, (2008).

142. Mishra, A. K. *et al.* Identification of an α(1→6) mannopyranosyltransferase (MptA), involved in Corynebacterium glutamicum lipomanann biosynthesis, and identification of its orthologue in Mycobacterium tuberculosis. *Mol Microbiol* **65**, (2007).

143. Mishra, A. K. *et al.* Lipoarabinomannan biosynthesis in Corynebacterineae: The interplay of two α(1→2)-mannopyranosyltransferases MptC and MptD in mannan branching. *Mol Microbiol* **80**, (2011).

144. Shi, L. *et al.* The carboxy terminus of EmbC from Mycobacterium smegmatis mediates chain length extension of the arabinan in lipoarabinomannan. *Journal of Biological Chemistry* **281**, (2006).

145. Birch, H. L. *et al.* Biosynthesis of mycobacterial arabinogalactan: Identification of a novel α(1→3) arabinofuranosyltransferase. *Mol Microbiol* **69**, (2008).

146. Seidel, M. *et al.* 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. *Journal of Biological Chemistry* **282**, (2007).

147. Nigou, J., Gilleron, M. & Puzo, G. Lipoarabinomannans: From structure to biosynthesis. *Biochimie* (2003).

148. Khoo, K. H., Dell, A., Morris, H. R., Brennan, P. J. & Chatterjee, D. Inositol phosphate capping of the nonreducing termini of lipoarabinomannan from rapidly growing strains of Mycobacterium. *Journal of Biological Chemistry* **270**, (1995).

149. Kaur, D. *et al.* Lipoarabinomannan of Mycobacterium: Mannose capping by a multifunctional terminal mannosyltransferase. *Proc Natl Acad Sci U S A* **105**, (2008).

150. Daleke, D. L. Phospholipid flippases. *Journal of Biological Chemistry* vol. 282 (2007).

151. Sanyal, S., Frank, C. G. & Menon, A. K. Distinct flippases translocate glycerophospholipids and oligosaccharide diphosphate dolichols across the endoplasmic reticulum. *Biochemistry* **47**, (2008).

152. Jackson, M., Stevens, C. M., Zhang, L., Zgurskaya, H. I. & Niederweis, M. Transporters Involved in the Biogenesis and Functionalization of the Mycobacterial Cell Envelope. *Chemical Reviews* vol. 121 (2021).

153. Glass, L. N. *et al.* Mycobacterium tuberculosis universal stress protein Rv2623 interacts with the putative ATP binding cassette (APC) transporter Rv1747 to regulate mycobacterial growth. *PLoS Pathog* **13**, e1006515 (2017).

154. Cabarca, S. *et al.* Structure of the Mycobacterium tuberculosis cPknF and conformational changes induced in forkhead-associated regulatory domains. *Curr Res Struct Biol* **3**, 165–178 (2021).

155. Schneewind, O. & Missiakas, D. Lipoteichoic acids, phosphate-containing polymers in the envelope of gram-positive bacteria. *Journal of Bacteriology* vol. 196 (2014).

156. Mishra, A. K. *et al.* Identification of a novel α(1→6) mannopyranosyltransferase MptB from Corynebacterium glutamicum by deletion of a conserved gene, NCgl1505, affords a lipomannan- and lipoarabinomannan-deficient mutant. *Mol Microbiol* **68**, (2008).

157. Viljoen, A. *et al.* The diverse family of MmpL transporters in mycobacteria: from regulation to antimicrobial developments. *Molecular Microbiology* vol. 104 (2017).

158. Maeda, N. *et al.* The cell surface receptor DC-SIGN discriminates between Mycobacterium species through selective recognition of the mannose caps on lipoarabinomannan. *Journal of Biological Chemistry* **278**, 5513–5516 (2003).

159. Pitarque, S. *et al.* Deciphering the molecular bases of Mycobacterium tuberculosis binding to the lectin DC-SIGN reveals an underestimated complexity. *Biochemical Journal* **392**, (2005).

160. Driessen, N. N. *et al.* Role of phosphatidylinositol mannosides in the interaction between mycobacteria and DC-SIGN. *Infect Immun* **77**, (2009).

161. Zhou, K. L., Li, X., Zhang, X. L. & Pan, Q. Mycobacterial mannose-capped lipoarabinomannan: a modulator bridging innate and adaptive immunity. *Emerging Microbes and Infections* vol. 8 (2019).

162. Kaur, D., Guerin, M. E., Škovierová, H., Brennan, P. J. & Jackson, M. Chapter 2 Biogenesis of the Cell Wall and Other Glycoconjugates of Mycobacterium tuberculosis. *Advances in Applied Microbiology* vol. 69 (2009).

163. Brennan, P. & Ballou, C. E. Biosynthesis of mannophosphoinositides by Mycobacterium phlei. The family of dimannophosphoinositides. *Journal of Biological Chemistry* **242**, 3046–56 (1967).

164. Lea-Smith, D. J. *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* **283**, 6773–6782 (2008).

165. Uzman, A. Molecular Cell Biology (4th edition) Harvey Lodish, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore and James Darnell; Freeman & Co., New York, NY, 2000, 1084 pp., list price $102.25, ISBN 0-7167-3136-3. *Biochemistry and Molecular Biology Education* **29**, (2001).

166. Vesely, P. Molecular biology of the cell. By Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts and Peter Walter. ISBN 0-8153-3218-1; hardback; 1,616 pages; $110.00 Garland Science Inc., New York, 2002. *Scanning* **26**, (2006).

167. Ramírez, A. S. *et al.* Structural basis of the molecular ruler mechanism of a bacterial glycosyltransferase. *Nat Commun* **9**, 445 (2018).

168. Roth, C. *et al.* Amylose recognition and ring-size determination of amylomaltase. *Sci Adv* **3**, 1–10 (2017).

169. Liang, Y. H., Lavoie, M., Comeau, M. A., Elela, S. A. & Ji, X. Structure of a Eukaryotic RNase III Postcleavage Complex Reveals a Double-Ruler Mechanism for Substrate Selection. *Mol Cell* **54**, 431–44 (2014).

170. Song, H. *et al.* The Functional Cycle of Rnt1p: Five Consecutive Steps of Double-Stranded RNA Processing by a Eukaryotic RNase III. *Structure* **25**, (2017).

171. Gandhi, A., Lakshminarasimhan, D., Sun, Y. & Guo, H. C. Structural insights into the molecular ruler mechanism of the endoplasmic reticulum aminopeptidase ERAP1. *Sci Rep* **1**, 1–6 (2011).

172. Sui, L. & Guo, H. C. ERAP1 binds peptide C-termini of different sequences and/or lengths by a common recognition mechanism. *Immunobiology* **226**, (2021).

173. Sui, L., Gandhi, A. & Guo, H. C. Crystal structure of a polypeptide's C-terminus in complex with the regulatory domain of ER aminopeptidase 1. *Mol Immunol* **80**, (2016).

174. Peters, J., Schönegge, A. M., Rockel, B. & Baumeister, W. Molecular ruler of tripeptidylpeptidase II: Mechanistic principle of exopeptidase selectivity. *Biochem Biophys Res Commun* **414**, (2011).

175. Subramanian, S. R., Singam, E. R. A., Berinski, M., Subramanian, V. & Wade, R. C. Identification of an Electrostatic Ruler Motif for Sequence-Specific Binding of Collagenase to Collagen. *Journal of Physical Chemistry B* **120**, (2016).

176. Wright, A. v. *et al.* A Functional Mini-Integrase in a Two-Protein-type V-C CRISPR System. *Mol Cell* **73**, (2019).

177. Smits, S. H. J., Mueller, A., Schmitt, L. & Grieshaber, M. K. A Structural Basis for Substrate Selectivity and Stereoselectivity in Octopine Dehydrogenase from Pecten maximus. *J Mol Biol* **381**, (2008).

178. Ray, S. S. *et al.* Cocrystal structures of diaminopimelate decarboxylase: Mechanism, evolution, and inhibition of an antibiotic resistance accessory factor. *Structure* **10**, (2002).

179. Putnam, C. D., Arvai, A. S., Bourne, Y. & Tainer, J. A. Active and inhibited human catalase structures: Ligand and NADPH binding and catalytic mechanism. *J Mol Biol* **296**, (2000).

180. Dovala, D. *et al.* Structure-guided enzymology of the lipid a acyltransferase LpxM reveals a dual activity mechanism. *Proc Natl Acad Sci U S A* **113**, E6064–E6071 (2016).

181. Wyckoff, T. J. O., Lin, S., Cotter, R. J., Dotson, G. D. & Raetz, C. R. H. Hydrocarbon rulers in UDP-N-acetylglucosamine acyltransferases. *Journal of Biological Chemistry* **273**, 32569–32372 (1998).

182. Wang, M. *et al.* Using the pimeloyl-CoA synthetase adenylation fold to synthesize fatty acid thioesters. *Nat Chem Biol* **13**, (2017).

183. Estrada, P. *et al.* The pimeloyl-CoA synthetase BioW defines a new fold for adenylate-forming enzymes. *Nat Chem Biol* **13**, (2017).

184. Cuesta-Seijo, J. A. *et al.* PagP crystallized from SDS/Cosolvent reveals the route for phospholipid access to the hydrocarbon ruler. *Structure* **18**, (2010).

185. Ahn, V. E. *et al.* A hydrocarbon ruler measures palmitate in the enzymatic acylation of endotoxin. *EMBO Journal* **23**, 2931–2941 (2004).

186. Bishop, R. E. The lipid A palmitoyltransferase PagP: Molecular mechanisms and role in bacterial pathogenesis. *Mol Microbiol* **57**, 900–12 (2005).

187. Khan, M. A., Moktar, J., Mott, P. J. & Bishop, R. E. A thiolate anion buried within the hydrocarbon ruler perturbs pagP lipid acyl chain selection. *Biochemistry* **49**, 2368–79 (2010).

188. Khan, M. A. *et al.* Inscribing the perimeter of the PagP hydrocarbon ruler by site-specific chemical alkylation. *Biochemistry* **49**, (2010).

189. Khan, M. A. *et al.* Gauging a hydrocarbon ruler by an intrinsic exciton probe. *Biochemistry* **46**, 4565–79 (2007).

190. Robertson, R. M. *et al.* A two-helix motif positions the lysophosphatidic acid acyltransferase active site for catalysis within the membrane bilayer. *Nat Struct Mol Biol* **24**, 666–671 (2017).

191. Rana, M. S. *et al.* Fatty acyl recognition and transfer by an integral membrane S-acyltransferase. *Science (1979)* **359**, eaao6326 (2018).

192. Panina, I. S., Krylov, N. A., Chugunov, A. O., Efremov, R. G. & Kordyukova, L. v. The Mechanism of Selective Recognition of Lipid Substrate by hDHHC20 Enzyme. *Int J Mol Sci* **23**, (2022).

193. Rottig, A. & Steinbuchel, A. Acyltransferases in Bacteria. *Microbiology and Molecular Biology Reviews* **77**, 277–321 (2013).

194. Forneris, F. & Mattevi, A. Enzymes without borders: Mobilizing substrates, delivering products. *Science (1979)* **321**, 213–216 (2008).

195. Dufrisne, M. B., Petrou, V. I., Clarke, O. B. & Mancia, F. Structural basis for catalysis at the membrane-water interface. *Biochim Biophys Acta Mol Cell Biol Lipids* **1862**, 1368–1385 (2017).

196. Su, X.-D. Textbook of structural biology – Second edition. *Crystallogr Rev* **24**, (2018).

197. Rout, M. P. & Sali, A. Principles for Integrative Structural Biology Studies. *Cell* vol. 177 (2019).

198. *Advances in Protein Molecular and Structural Biology Methods*. *Advances in Protein Molecular and Structural Biology Methods* (2022).

199. Kühlbrandt, W. The resolution revolution. *Science* vol. 343 (2014).

200. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, (2021).

201. Varadi, M. *et al.* AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**, (2022).

202. Subramaniam, S. & Kleywegt, G. J. A paradigm shift in structural biology. *Nature Methods* vol. 19 (2022).

203. Smyth, M. S. & Martin, J. H. J. x Ray crystallography. *Journal of Clinical Pathology - Molecular Pathology* **53**, 8–14 (2000).

204. Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, (1958).

205. McPherson, A. & Gavira, J. A. Introduction to protein crystallization. *Acta Crystallographica Section F:Structural Biology Communications* **70**, 2–20 (2014).

206. McPherson, A. Introduction to protein crystallization. *Methods* **34**, (2004).

207. McPherson, A. An Overview of Macromolecular Crystallography. in *Introduction to Macromolecular Crystallography* (2009).

208. McPherson, A. *Introduction to Macromolecular Crystallography: Second Edition*. *Introduction to Macromolecular Crystallography: Second Edition* (2008).

209. Garman, E. F. Summary of lecture and practical session at Biophysics and Structural Biology at Synchrotrons workshop: cryo-cooling in macromolecular crystallography—why and how? in *Biophysical Reviews* vol. 11 (2019).

210. Shabalin, I. G., Porebski, P. J. & Minor, W. Refining the macromolecular model–achieving the best agreement with the data from X-ray diffraction experiment. *Crystallography Reviews* vol. 24 (2018).

211. Sletten, E. M. & Bertozzi, C. R. Bioorthogonal chemistry: Fishing for selectivity in a sea of functionality. *Angewandte Chemie - International Edition* vol. 48 (2009).

212. Prescher, J. A. & Bertozzi, C. R. Chemistry in Living Systems. *Nature Chemical Biology* vol. 1 (2005).

213. Devaraj, N. K. & Finn, M. G. Introduction: Click Chemistry. *Chemical Reviews* vol. 121 (2021).

214. Kolb, H. C., Finn, M. G. & Sharpless, K. B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angewandte Chemie - International Edition* vol. 40 (2001).

215. Tornøe, C. W., Christensen, C. & Meldal, M. Peptidotriazoles on solid phase: [1,2,3]-Triazoles by regiospecific copper(I)-catalyzed 1,3-dipolar cycloadditions of terminal alkynes to azides. *Journal of Organic Chemistry* **67**, (2002).

216. Rostovtsev, V. v., Green, L. G., Fokin, V. v. & Sharpless, K. B. A stepwise huisgen cycloaddition process: Copper(I)-catalyzed regioselective 'ligation' of azides and terminal alkynes. *Angewandte Chemie - International Edition* **41**, (2002).

217. Huisgen, R. 1,3-Dipolar Cycloadditions. Past and Future. *Angewandte Chemie International Edition in English* **2**, (1963).

218. Battigelli, A., Almeida, B. & Shukla, A. Recent Advances in Bioorthogonal Click Chemistry for Biomedical Applications. *Bioconjugate Chemistry* vol. 33 (2022).

219. Bertozzi, C. R. A decade of bioorthogonal chemistry. *Accounts of Chemical Research* vol. 44 (2011).

220. Hang, H. C., Yu, C., Kato, D. L. & Bertozzi, C. R. A metabolic labeling approach toward proteomic analysis of mucin-type O-linked glycosylation. *Proc Natl Acad Sci U S A* **100**, (2003).

221. Patterson, D. M., Nazarova, L. A. & Prescher, J. A. Finding the right (bioorthogonal) chemistry. *ACS Chemical Biology* vol. 9 (2014).

222. Agard, N. J., Prescher, J. A. & Bertozzi, C. R. A strain-promoted [3 + 2] azide-alkyne cycloaddition for covalent modification of biomolecules in living systems. *J Am Chem Soc* **126**, (2004).

223. Baskin, J. M. *et al.* Copper-free click chemistry for dynamic in vivo imaging. *Proc Natl Acad Sci U S A* **104**, (2007).

224. Sletten, E. M. & Bertozzi, C. R. A hydrophilic azacyclooctyne for cu-free click chemistry. *Org Lett* **10**, (2008).

225. Yang, S. *et al.* Deciphering Protein O-Glycosylation: Solid-Phase Chemoenzymatic Cleavage and Enrichment. *Anal Chem* **90**, 8261–8269 (2018).

226. Kabsch, W. XDS. *Acta Crystallogr D Biol Crystallogr* **66**, 125–132 (2010).

227. Cha, S. S. *et al.* Experimental phasing using zinc anomalous scattering. *Acta Crystallogr D Biol Crystallogr* **68**, (2012).

228. Sikharulidze, I., Winter, G. & Hall, D. R. Big EP: automated structure solution pipeline deployment at Diamond Light Source. *Acta Crystallogr A Found Adv* **72**, (2016).

229. Skubák, P. & Pannu, N. S. Automatic protein structure solution from weak X-ray data. *Nat Commun* **4**, (2013).

230. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr* **62**, (2006).

231. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography* vol. 67 (2011).

232. McCoy, A. J. *et al.* Phaser crystallographic software. *J Appl Crystallogr* **40**, 658–674 (2007).

233. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213–221 (2010).

234. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501 (2010).

235. Afonine, P. v. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* **68**, 352–367 (2012).

236. Chen, V. B. *et al.* MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12–21 (2010).

237. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612 (2004).

238. Ho, B. K. & Gruswitz, F. HOLLOW: Generating accurate representations of channel and interior surfaces in molecular structures. *BMC Struct Biol* **8**, 49 (2008).

239. Holm, L. & Rosenström, P. Dali server: Conservation mapping in 3D. *Nucleic Acids Res* **38**, W545-549 (2010).

240. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J Mol Biol* **372**, 774–797 (2007).

241. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **31**, 455–461 (2009).

242. Tallant, C., Marrero, A. & Gomis-Rüth, F. X. Matrix metalloproteinases: Fold and function of their catalytic domains. *Biochimica et Biophysica Acta - Molecular Cell Research* vol. 1803 (2010).

243. Bertini, I. *et al.* Snapshots of the reaction mechanism of matrix metalloproteinases. *Angewandte Chemie - International Edition* **45**, (2006).

244. Auld, D. S. Chapter 78 - Catalytic Mechanisms for Metallopeptidases A2 - Rawlings, Neil D. in *Handbook of Proteolytic Enzymes* (2013).

245. Cerdà-Costa, N. & Gomis-Rüth, F. X. Architecture and function of metallopeptidase catalytic domains. *Protein Science* vol. 23 (2014).

246. Rawlings, N. D. *et al.* The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res* **46**, (2018).

247. Bode, W., Gomis-Rüth, F. X. & Stöckler, W. Astacins, serralysins, snake venom and matrix metalloproteinases exhibit identical zinc-binding environments (HEXXHXXGXXH and Met-turn) and topologies and should be grouped into a common family, the 'metzincins'. *FEBS Lett* **331**, (1993).

248. Park, K. *et al.* Synthesis and activity of tryptophan sulfonamide derivatives as novel non-hydroxamate TNF-α converting enzyme (TACE) inhibitors. *Bioorg Med Chem* **17**, (2009).

249. Zhang, D. *et al.* Structural interaction of natural and synthetic inhibitors with the venom metalloproteinase, atrolysin C (form d). *Proc Natl Acad Sci U S A* **91**, (1994).

250. Watanabe, L. *et al.* Amino acid sequence and crystal structure of BaP1, a metalloproteinase from Bothrops asper snake venom that exerts multiple tissue-damaging activities. *Protein Science* **12**, (2009).

251. Al-Riyami, B., Üstok, F. I., Stott, K., Chirgadze, D. Y. & Christie, G. The crystal structure of Clostridium perfringens SleM, a muramidase involved in cortical hydrolysis during spore germination. *Proteins: Structure, Function and Bioinformatics* **84**, (2016).

252. Kovtun, O. *et al.* Structure of the membrane-assembled retromer coat determined by cryo-electron tomography. *Nature* **561**, (2018).

253. Domingues, M. N. *et al.* Structural basis of exo--mannanase activity in the GH2 family. *Journal of Biological Chemistry* **293**, (2018).

254. Trastoy, B., Naegeli, A., Anso, I., Sjögren, J. & Guerin, M. E. Structural basis of mammalian mucin processing by the human gut O-glycopeptidase OgpA from Akkermansia muciniphila. *Nat Commun* **11**, 4844 (2020).

255. Abramowitz, N., Schechter, I. & Berger, A. On the size of the active site in proteases II. Carboxypeptidase-A. *Biochem Biophys Res Commun* **29**, (1967).

256. Schechter, I. & Berger, A. On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun* **27**, (1967).

257. Noach, I. *et al.* Recognition of protein-linked glycans as a determinant of peptidase activity. *Proc Natl Acad Sci U S A* **114**, E679–E688 (2017).

258. Guerin, M. E., Stirnemann, G. & Giganti, D. Conformational entropy of a single peptide controlled under force governs protease recognition and catalysis. *Proc Natl Acad Sci U S A* **115**, (2018).

259. Yang, W., Ao, M., Hu, Y., Li, Q. K. & Zhang, H. Mapping the O-glycoproteome using site-specific extraction of O-linked glycopeptides (EXoO). *Mol Syst Biol* **14**, (2018).

260. Yu, A. C. Y., Worrall, L. J. & Strynadka, N. C. J. Structural insight into the bacterial mucinase StcE essential to adhesion and immune evasion during enterohemorrhagic E. coli infection. *Structure* **20**, (2012).

261. Matthews, B. W. Structural Basis of the Action of Thermolysin and Related Zinc Peptidases. *Acc Chem Res* **21**, (1988).

262. Vallee, B. L. & Auld, D. S. Active-site zinc ligands and activated H2O of zinc enzymes. *Proc Natl Acad Sci U S A* **87**, (1990).

263. Pelmenschikov, V. & Siegbahn, P. E. M. Catalytic mechanism of matrix metalloproteinases: Two-layered ONIOM study. *Inorg Chem* **41**, (2002).

264. Dudev, T. & Lim, C. Tetrahedral vs octahedral zinc complexes with ligands of biological interest: A DFT/CDM study. *J Am Chem Soc* **122**, (2000).

265. Malaker, S. A. *et al.* The mucin-selective protease StcE enables molecular and functional analysis of human cancer-associated mucins. *Proc Natl Acad Sci U S A* **116**, (2019).

266. Nagae, M. *et al.* Structural basis of the catalytic reaction mechanism of novel 1,2-α-L-fucosidase from Bifidobacterium bifidum. *Journal of Biological Chemistry* **282**, 18497–18509 (2007).

267. Liu, H. & Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol* **8**, (2008).

268. Hampton Research. Crystal Growth 101: Seeding. *Hampton Research* 1–8 (2020).

269. Case, D. A. *et al.* AMBER 2020. (2020).

270. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713 (2015).

271. Kirschner, K. N. *et al.* GLYCAM06: A generalizable biomolecular force field. carbohydrates. *J Comput Chem* **29**, (2008).

272. K. Kiyohara, K. E. Gubbins & A. Z. Panagiotopoulos. Phase coexistence properties of polarizable water models. *Mol Phys* **94**, 803–808 (1998).

273. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* **98**, (1993).

274. Antonelou, M. H. *et al.* Red blood cell aging markers during storage in citrate-phosphate-dextrose-saline-adenine-glucose-mannitol. *Transfusion (Paris)* **50**, 376–389 (2010).

275. Y. Lapierre MD, MSc, D. Rigal, J. Adam, D. Josef, F. Meyer, S. Greber, C. D. The gel test: a new way to detect red cell antigen-antibody reactions. *Transfusion (Paris)* **30**, 109–113 (1990).

276. Houwen, B. Blood film preparation and staining procedures. *Clin Lab Med* **22**, 1–14 (2002).

277. Minucci, A., Giardina, B., Zuppi, C. & Capoluongo, E. Glucose-6-phosphate dehydrogenase laboratory assay: How, when, and why? *IUBMB Life* **61**, 27–34 (2009).

278. Katayama, T. *et al.* Molecular cloning and characterization of Bifidobacterium bifidum 1,2-α-L-fucosidase (AfcA), a novel inverting glycosidase (glycoside hydrolase family 95). *J Bacteriol* **186**, 4885–4893 (2004).

279. Vieira, P. S. *et al.* Xyloglucan processing machinery in Xanthomonas pathogens and its role in the transcriptional activation of virulence factors. *Nat Commun* **12**, 4049 (2021).

280. Rogowski, A. *et al.* Glycan complexity dictates microbial resource allocation in the large intestine. *Nat Commun* **6**, (2015).

281. Kostopoulos, I. *et al.* Akkermansia muciniphila uses human milk oligosaccharides to thrive in the early life conditions in vitro. *Sci Rep* **10**, 14330 (2020).

282. Déjean, G., Tauzin, A. S., Bennett, S. W., Creagh, A. L. & Brumer, H. Adaptation of syntenic xyloglucan utilization loci of human gut Bacteroidetes to polysaccharide side chain diversity. *Appl Environ Microbiol* **85**, e01491-19 (2019).

283. Sela, D. A. *et al.* Bifidobacterium longum subsp. infantis ATCC 15697 α-fucosidases are active on fucosylated human milk oligosaccharides. *Appl Environ Microbiol* **78**, 795–803 (2012).

284. Etienne-Mesmin, L. *et al.* Experimental models to study intestinal microbes–mucus interactions in health and disease. *FEMS Microbiol Rev* **43**, 457–489 (2019).

285. Robbe, C., Capon, C., Coddeville, B. & Michalski, J. C. Structural diversity and specific distribution of O-glycans in normal human mucins along the intestinal tract. *Biochemical Journal* **384**, 307–316 (2004).

286. Jensen, P. H., Kolarich, D. & Packer, N. H. Mucin-type O-glycosylation - Putting the pieces together. *FEBS Journal* **277**, 81–94 (2010).

287. Cooling, L. Blood groups in infection and host susceptibility. *Clin Microbiol Rev* **28**, 801–870 (2015).

288. Garratty, G. Modulating the red cell membrane to produce universal/stealth donor red cells suitable for transfusion. *Vox Sang* **94**, 87–95 (2008).

289. Goldstein, J., Siviglia, G., Hurst, R., Lenny, L. & Reich, L. Group B erythrocytes enzymatically converted to group O survive normally in A, B, and O individuals. *Science (1979)* **215**, 168–170 (1982).

290. Kruskall, M. S. *et al.* Transfusion to blood group A and O patients of group B RBCs that have been enzymatically converted to group O. *Transfusion (Paris)* **40**, 1290–1298 (2000).

291. Liu, Q. P. *et al.* Bacterial glycosidases for the production of universal red blood cells. *Nat Biotechnol* **25**, 454–464 (2007).

292. Anderson, K. M. *et al.* A clostridial endo-β-galactosidase that cleaves both blood group A and B glycotopes: The first member of a new glycoside hydrolase family, GH98. *Journal of Biological Chemistry* **280**, 7720–7728 (2005).

293. Rahfeld, P. *et al.* An enzymatic pathway in the human gut microbiome that converts A to universal O type blood. *Nat Microbiol* **4**, 1475–1485 (2019).

294. Wang, A. *et al.* Ex vivo enzymatic treatment converts blood type A donor lungs into universal blood type lungs. *Sci Transl Med* **14**, eabm7190 (2022).

295. Ford, J. Red blood cell morphology. *Int J Lab Hematol* **35**, 351–357 (2013).

296. Bianco, T., Farmer, B. J., Sage, R. E. & Dobrovic, A. Loss of red cell A, B, and H antigens is frequent in myeloid malignancies. *Blood* **97**, (2001).

297. Hult, A. K. & Olsson, M. L. Many genetically defined ABO subgroups exhibit characteristic flow cytometric patterns. *Transfusion (Paris)* **50**, (2010).

298. Sharon, R. & Fibach, E. Quantitative flow cytometric analysis of ABO red cell antigens. *Cytometry* **12**, (1991).

299. Becker, D. J. & Lowe, J. B. Leukocyte adhesion deficiency type II. *Biochim Biophys Acta Mol Basis Dis* **1455**, 193–204 (1999).

300. Das, J., Sharma, A., Jindal, A., Aggarwal, V. & Rawat, A. Leukocyte adhesion defect: Where do we stand circa 2019? *Genes Dis* **7**, 107–114 (2020).

301. Hartley, T. *et al.* New Diagnostic Approaches for Undiagnosed Rare Genetic Diseases. *Annual Review of Genomics and Human Genetics* vol. 21 (2020).

302. Boycott, K. M., Dyment, D. A., Sawyer, S. L., Vanstone, M. R. & Beaulieu, C. L. Identification of genes for childhood heritable diseases. *Annu Rev Med* **65**, (2014).

303. Bahl, O. P. Glycosidases of aspergillus niger. II. Purification and general properties of 1,2-alpha-L-fucosidase. *Journal of Biological Chemistry* **245**, 299–304 (1970).

304. Doinel, C., Ropars, C. & Rufin, J. M. I and H activities of human red blood cells treated with an 1, 2-α-L-Fucosidase from aspergillus niger. *Rev Fr Transfus Immunohematol* **23**, 259–269 (1980).

305. Zhang, W. & Zhu, Z. yan. Structural modification of H histo-blood group antigen. *Blood Transfus* **13**, 143–149 (2015).

306. Yu, Z. & Cowan, J. A. Design of Artificial Glycosidases: Metallopeptides that Remove H Antigen from Human Erythrocytes. *Angewandte Chemie - International Edition* **56**, 2763–2766 (2017).

307. Tiansheng Li, Juan Ye, Lei Wang, Lin Zou, Yameng Guo, Linlin Hou, Danfeng Shen, Xiaohong Cai, Haobo Huang, Guiqin Sun, L. C. Bacterial fucosidase enables the production of Bombay red blood cells. *bioRxiv* (2019).

308. Sun, G. *et al.* Complete genome sequence of Elizabethkingia meningoseptica, isolated from a T-cell non-Hodgkin's lymphoma patient. *Genome Announc* **3**, e00673-15 (2015).

309. Lee, J. *et al.* CHARMM-GUI Membrane Builder for Complex Biological Membrane Simulations with Glycolipids and Lipoglycans. *J Chem Theory Comput* **15**, 775–786 (2019).

310. Dickson, C. J. *et al.* Lipid14: The amber lipid force field. *J Chem Theory Comput* **10**, 865–879 (2014).

311. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general Amber force field. *J Comput Chem* **25**, 1157–1174 (2004).

312. Bayly, C. I., Cieplak, P., Cornell, W. D. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *Journal of Physical Chemistry* **97**, 10269–10280 (1993).

313. Frisch, M. J. *et al.* Gaussian 09, Revision A.02. **92**, 549–556 (2016).

314. Press, W. H., Teukolsky, S. A., Vettering, W. T. & Flannery, B. P. *NUMERICAL RECIPES The Art of Scientific Computing Third Edition*. *CAMBRIDGE UNIVERSITY PRESS* (2007).

315. Pastor, R. W., Brooks, B. R. & Szabo, A. An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Mol Phys* **65**, 1409–1419 (1988).

316. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J Chem Phys* **81**, 3684–3690 (1984).

317. Rodgers, J. B. Assay of acyl-CoA:monoglyceride acyltransferase from rat small intestine using continuous recording spectrophotometry. *J Lipid Res* **10**, 427–32 (1969).

318. Ellman, G. L. Tissue sulfhydryl groups. *Arch Biochem Biophys* **82**, 70–7 (1959).

319. Brennan, P. & Ballou, C. E. Biosynthesis of mannophosphoinositides by Mycobacterium phlei. Enzymatic acylation of the dimannophosphoinositides. *Journal of Biological Chemistry* **243**, 2975–84 (1968).

320. Tamada, T. *et al.* Substrate recognition and selectivity of plant glycerol-3-phosphate acyltransferases (GPATs) from Cucurbita moscata and Spinacea oleracea. *Acta Crystallogr D Biol Crystallogr* **60**, (2004).

321. Zheng, W. & Cole, P. A. Novel bisubstrate analog inhibitors of serotonin N-acetyltransferase: The importance of being neutral. *Bioorg Chem* **31**, 398–411 (2003).

322. Anso, I. *et al.* Molecular ruler mechanism and interfacial catalysis of the integral membrane acyltransferase PatA. *Sci Adv* **7**, eabj4565 (2021).

323. Sastre, D. E. *et al.* Membrane fluidity adjusts the insertion of the transacylase PlsX to regulate phospholipid biosynthesis in Gram-positive bacteria. *Journal of Biological Chemistry* **295**, 2136–2147 (2020).

324. Šakanovič, A., Hodnik, V. & Anderluh, G. Surface Plasmon Resonance for Measuring Interactions of Proteins with Lipids and Lipid Membranes. *Methods in Molecular Biology* **2003**, 53–70 (2019).

325. Bernardino De La Serna, J. *et al.* Compositional and structural characterization of monolayers and bilayers composed of native pulmonary surfactant from wild type mice. *Biochim Biophys Acta Biomembr* **1828**, 2450–2459 (2013).

326. Marsh, D. Components of the lateral pressure in lipid bilayers deduced from HII phase dimensions. *Biochim Biophys Acta Biomembr* **1279**, 119–123 (1996).

327. Söderlund, T., Alakoskela, J. M. I., Pakkanen, A. L. & Kinnunen, P. K. J. Comparison of the effects of surface tension and osmotic pressure on the interfacial hydration of a fluid phospholipid bilayer. *Biophys J* **85**, 2333–2341 (2003).

328. Zhao, H. *et al.* Interaction of the antimicrobial peptide pheromone Plantaricin A with model membranes: Implications for a novel mechanism of action. *Biochim Biophys Acta Biomembr* **1758**, 1461–1474 (2006).

329. Blobel, G., Walter, P. & Gilmore, R. Intracellular protein topogenesis. *Prog Clin Biol Res* **168**, 3–10 (1984).

330. Siegrist, M. S., Swarts, B. M., Fox, D. M., Lim, S. A. & Bertozzi, C. R. Illumination of growth, division and secretion by metabolic labeling of the bacterial cell surface. *FEMS Microbiology Reviews* vol. 39 (2015).

331. Ford, F., Yuzawa, T., Platz, M. S., Matzinger, S. & Fülscher, M. Rearrangement of dimethylcarbene to propene: Study by laser flash photolysis and ab initio molecular orbital theory. *J Am Chem Soc* **120**, (1998).

332. Dormán, G. & Prestwich, G. D. Benzophenone Photophores in Biochemistry. *Biochemistry* **33**, (1994).

333. Brunner, J. New photolabeling and crosslinking methods. *Annual Review of Biochemistry* vol. 62 (1993).

334. Das, J. Aliphatic diazirines as photoaffinity probes for proteins: Recent developments. *Chemical Reviews* vol. 111 (2011).

335. Korneev, S. M. Valence isomerization between diazo compounds and diazirines. *European Journal of Organic Chemistry* (2011).

336. Haberkant, P. & Holthuis, J. C. M. Fat & fabulous: Bifunctional lipids in the spotlight. *Biochim Biophys Acta Mol Cell Biol Lipids* **1841**, 1022–1030 (2014).

337. Haberkant, P. *et al.* In vivo profiling and visualization of cellular protein-lipid interactions using bifunctional fatty acids. *Angewandte Chemie - International Edition* **52**, 4033–4038 (2013).

338. Hosoya, T. *et al.* Novel bifunctional probe for radioisotope-free photoaffinity labeling: Compact structure comprised of photospecific ligand ligation and detectable tag anchoring units. *Org Biomol Chem* **2**, (2004).

339. Wang, D. *et al.* Global Mapping of Protein–Lipid Interactions by Using Modified Choline-Containing Phospholipids Metabolically Synthesized in Live Cells. *Angewandte Chemie - International Edition* **56**, 5829–5833 (2017).

340. Liu, X., Dong, T., Zhou, Y., Huang, N. & Lei, X. Exploring the Binding Proteins of Glycolipids with Bifunctional Chemical Probes. *Angewandte Chemie - International Edition* **55**, 14330–14334 (2016).

341. Foley, H. N., Stewart, J. A., Kavunja, H. W., Rundell, S. R. & Swarts, B. M. Bioorthogonal chemical reporters for selective in situ probing of mycomembrane components in mycobacteria. *Angewandte Chemie - International Edition* **55**, 2053–2057 (2016).

342. Swarts, B. M. *et al.* Probing the mycobacterial trehalome with bioorthogonal chemistry. *J Am Chem Soc* **134**, (2012).

343. Patterson, J. H., Waller, R. F., Jeevarajah, D., Billman-Jacobe, H. & McConville, M. J. Mannose metabolism is required for mycobacterial growth. *Biochemical Journal* **372**, (2003).

344. Parish, T. & Kumar, A. *Mycobacteria Protocols Methods and Protocols Fourth Edition Methods in Molecular Biology 2314*. http://www.springer.com/series/7651.

345. Biegas, K. J. & Swarts, B. M. Chemical probes for tagging mycobacterial lipids. *Current Opinion in Chemical Biology* vol. 65 (2021).

346. Kolbe, K. *et al.* Azido Pentoses: A New Tool To Efficiently Label Mycobacterium tuberculosis Clinical Isolates. *ChemBioChem* **18**, (2017).

347. Backus, K. M. *et al.* Uptake of unnatural trehalose analogs as a reporter for Mycobacterium tuberculosis. *Nat Chem Biol* **7**, (2011).

348. Urbanek, B. L. *et al.* Chemoenzymatic synthesis of trehalose analogues: Rapid access to chemical probes for investigating mycobacteria. *ChemBioChem* **16**, (2015).

349. Kavunja, H. W. *et al.* Photoactivatable Glycolipid Probes for Identifying Mycolate-Protein Interactions in Live Mycobacteria. *J Am Chem Soc* **142**, 7725–7731 (2020).

350. Sarpe, V. A. & Kulkarni, S. S. Synthesis of maradolipid. *Journal of Organic Chemistry* **76**, 6866–6870 (2011).