Investigating the Functional Roles of the Sialidase NanH and Glycopeptidase Amuc_1438 in the Enzymatic Degradation of Mucin

by

Brendon J. Medley
B.Sc. (Hon), University of Victoria, 2020

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Biochemistry and Microbiology

Investigating the Functional Roles of the Sialidase NanH and Glycopeptidase Amuc_1438 in the Enzymatic Degradation of Mucin

By

Brendon J. Medley

B.Sc. (Hon), University of Victoria, 2020

Supervisory Committee

Dr. Alisdair Boraston
Department of Biochemistry and Microbiology
Supervisor

Dr. John Burke
Department of Biochemistry and Microbiology
Departmental Member

Dr. Fraser Hof
Department of Chemistry
Outside Member

Abstract

The mucosal layer within the gastrointestinal tracts of animals and humans serves a critical function in safeguarding the epithelial layer against pathogens, including *Clostridium perfringens*. This layer consists of mucin-based glycoproteins, forming a dual-layered structure: an inner layer firmly attached to the epithelial cells, acting as a protective barrier, and an outer layer that fosters a favourable environment for commensal organisms. Both commensal and pathogenic bacterial species possess an array of enzymes designed for mucin degradation, fulfilling two main purposes: utilizing the abundant carbohydrate network as an energy and carbon source, and breaking down the mucus layer during pathogenic invasion. However, our understanding of these enzymatic tools employed by the gut microbiota remains incomplete, leaving gaps in our knowledge concerning pathogenic invasion and host-microbe interactions. This thesis presents a comprehensive analysis of two enzymes involved in mucin degradation: the intracellular sialidase NanH from *Clostridium perfringens* and the previously uncharacterized Amuc_1438 from *Akkermansia muciniphila*. The thesis investigates the structure of NanH, comparing it to its extracellular counterpart NanI, and provides structural insight into the specificity for sialic acids linked to glycans with α(2,3) over α(2,6) linkages. Additionally, through structural and functional investigations, Amuc_1438 is revealed to possess glycopeptidase activity, targeting specifically glycopeptides containing a Tn-antigen on a serine or threonine residue. Intriguingly, this work outlines that Amuc_1438 likely belongs to an uncharacterized family of glycopeptidases.

The primary aim of this thesis is to demonstrate how these two distinct enzymes fit into the enzymatic pathway of mucin degradation observed in both commensal and pathogenic gut bacteria. By shedding light on the structural characteristics and functional roles of NanH and Amuc_1438, this research contributes to a deeper understanding of the intricate enzymatic processes involved in mucin degradation, thus enhancing our knowledge of pathogenic mucosal-invasion strategies and host-microbe interactions.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my sincere gratitude to Dr. Boraston, my supervisor, for accepting me as a student during my research experience course in my third year of schooling. I am truly thankful for your unwavering faith in me and your constant support in pursuing entry into the honours program and inevitably graduate school within your research group. Without your invaluable mentorship, none of these achievements would have been possible. Thank you to my committee, Dr. John Burke, and Dr. Fraser Hof for providing interest in a successful master's degree, while giving guidance along the way.

I am deeply thankful to the numerous members, both past and present Boraston lab. I would like to extend my heartfelt thanks to Dr. Andrew Hettle Dr. Benjamin Pluvinage, and Dr. Jo Hobbs for imparting their extensive knowledge, and crafting my understanding of the scientific world. I would also like to acknowledge Berna Alvarez for living in the lab and always being able to turn off my amicon, and Ashley Deventer for our daily 2 pm quad laps, reminding me that there is a world beyond our lab. Cheers to the Coffee Crew, Liam Mihalynuk**,** Emily Knudson Goerner, Olivia Canil, Claire Stevens, Rory Hills, Eric Scott-Iversen, Matt Shortill, and the many innocent honours students that stuck around to patiently listen to my rants.

I extend my warmest thank you to my mom and dad, brother Jake, sister Rae and partner Ally for being my unwavering pillars of support through my entire educational journey. I am forever grateful for all your efforts in making me the person I am today.

List of abbreviations

GalNAc - *N*-acetylgalactosamine

GlcNAc - *N*-acetylglucosamine

VWD - von Willebrand Domains

MUC – Mucin glycoprotein

CysD – Cystein rich domain

Fuc - Fucose

GalNAc Ts - UDP-*N*-acetylgalactosaminyltransferase

CAZymes – Carbohydrate Active Enzymes

CBM – Carbohydrate-Binding Module

GHXXX – Glycosyl hydrolase from family XXX

Neu5Ac – *N*-acetylneuraminic acid

Neu5Gc – *N*-Glycolylneuraminic acid

CPA – *Clostrium perfringens alpha toxin*

STag Sialylated T-Antigen

SLN - Neu5,7,9Ac-α(2,6)LAcNac

ALT – Amuc_1438 construct with the alternative start site

ALTL – Amuc_1438 construct with the alternative start site and linker

CAT – Amuc_1438 catalytic construct

SUS – Starch utilization system

Chapter 1.0 Introduction

1.1 The Mucosal layer of the gastrointestinal tract.

Many organs are protected from their extracorporeal environment with the use of a barrier that is known as a mucosal layer. Organs that are exposed to particulate matter and potentially harmful agents use the mucosal layer as a primary form of protection, commonly associated with the gastrointestinal tract, the lungs, and the reproductive tracts[1–3]. In addition to the protection the mucosal layer provides, this layer provides lubrication, particularly when discussing the cellular barriers of the lower gastrointestinal tract[4]. The main composition of this barrier consists of three layers: the deepest layer made up of smooth muscle called the muscularis mucosae, the layer of epithelial cells on the surface, which are separated by the layer of connective tissue known as the lamina propria.

Given the presence of mucosal layers in different environments within the body, it is reasonable to expect that these layers would adapt to their respective circumstances. For instance, the gastrointestinal mucosal layers facilitate the secretion of gel-forming mucin glycoproteins on the exterior of the epithelial cells to form a protective layer of mucus[5]. As the gastrointestinal tract contains trillions of microorganisms including opportunistic pathogens, the mucus layer is key for regulating resident organism homeostasis while preventing the establishment of pathogenic infections[6]. The gastrointestinal tract mucus can be divided into two layers: a tight and dense inner mucus layer that's bound to the epithelial cells that are replenished by goblet cells in the epithelial layer, and a loose outer mucus layer that is colonized by commensal species of the gastrointestinal tract[6,7]. With the main goal of lubrication and protection, the outer layer is the initial barrier against host invasion from bacterial, viral, and chemical attempts to reach the epithelial cells which would inevitably cause an inflammatory response [4,7]. The dense and constrictive inner layer that eventually feeds into the outer layer keeps the epithelial cells free from any bacteria by providing an impenetrable

barrier[8]. On the other hand, the outer layer provides a wealth of nutrients for organisms that target this outer layer as their habitat. The thickness and composition of the mucus layer are modified depending on the location in the gastrointestinal tract.

Generally in the mammalian gut, the stomach and the colon are both lined with thick inner and outer mucus layers, whereas a single mucus layer is used to protect the small intestine[9,10]. Dependent on the animal, there is significant variation in mucus layer thickness throughout the gastrointestinal tract[11]. Due to the highly acidic and deconstructive environment of the stomach, the thick mucus barrier provides a preventative measure for the stomach acid from degrading the epithelial cell walls[12]. The mucus layer in the duodenum measures a considerable decrease in thickness comparably to the stomach, which allows for an increase in nutrient absorption through the intestinal lining[13]. In mice and rats, it is shown that the dense inner layer ranges in thickness from 15–30 μM in the small intestine to ~100 μM in the colon where the outer layer can be 100-400 μM in the small intestine and up to 700 μM in the colon[6]. The gastrointestinal tract is a fascinating system consisting of various organs involved in digestion. What makes it particularly remarkable is its ability to protect itself from self-digestion, thanks to the presence of the mucus layer.

An accurate human model is hard to study due to the difficulty in extracting a sample from the source, however, the general composition of the mucus layer is still understood[11]. So far, a total of 21 distinct genes have been identified that belong to the MUC family, responsible for producing mucin glycoproteins that are distributed throughout the body[14,15]. The mucin glycoproteins that these genes encode are categorized into gel-forming mucins; the mucins that get secreted and assemble as polymeric structures, and the non-gel-forming transmembrane mucins that are characterized by COOH-terminal domains that are useful for attaching to surface membranes and found covering intestinal epithelial cells[16,17]. Goblet cells are a specialized form of epithelial cell that originate from crypts within the epithelial layer found in the mucosa with the purpose of secreting the mucin glycoproteins which are responsible for the formation and preservation of the mucus layer (figure 1)[17,18]. Within

the colon, the predominate transmembrane mucins are MUC1, MUC3A/B, MUC4, MUC13, and MUC17[19]. Somewhat less varied, there are three polymeric gel-forming mucins that are secreted by the mucosal barriers found throughout the gastrointestinal tract MUC2, MUC5B and MUC6[11]. MUC6 is secreted from Brunner's glands in the duodenum, and MUC5B can be found secreted in the colon, however, MUC2 is the dominant mucin found throughout the entire distal gastrointestinal environment, and well represented in the colon[11]. MUC2 is constructed of a peptide backbone and *O*-linked to the serine and threonine residues are densely packed glycan oligosaccharides that are responsible for around ~80% of the mass of MUC2. The oligosaccharide chains are extensively variable in length and structure but are all comprised of monosaccharide building blocks including galactose, fucose, hexosamines (*N*-acetylgalactosamine [GalNAc], *N*-acetylglucosamine [GlcNac]), and sialic acid[20]. MUC2, the large gel-forming mucin-based protein provides much of the structure in the mucous layer by forming a largely impenetrable inner layer adhered to the epithelial cells, and a loose outer layer that provides a nutrient-rich environment that serves as a habitat for commensal organisms[21]. The MUC2 mucin dimerizes through intramolecular disulphide bonds in the cysteine knot domains found at the C-terminal end of the glycoprotein followed by polymerizing through intramolecular forces between localized MUC2 molecules via N-terminal von Willebrand Domains (VWD) creating a gel-like network (figure 1) [5,22].

Many factors can affect the composition of the mucus layer including diet, antibiotic treatments, pre/probiotics and pathogens[11]. As well, many of these conditions can compromise the mucus layer, leading to variations in the viscosity and thickness of the protective mucus. During the event of mucin thinning, the epithelial cells are more exposed to the microbiota, triggering inflammation of the gut mucus layer (figure 1)[23].

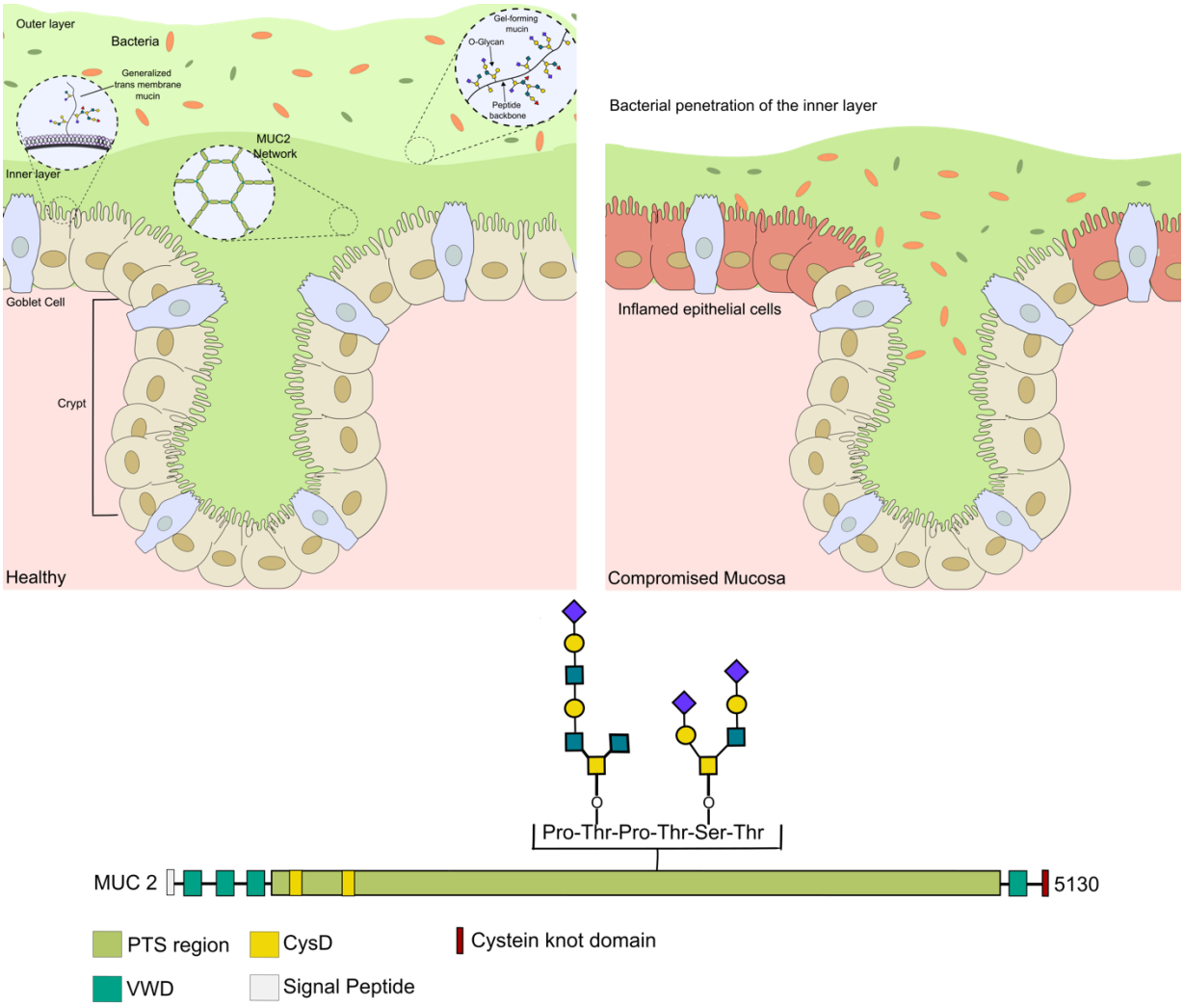


Figure 1. Generalized mucosal layer within the colon. Representation of the Inner and outer mucus layers that provide protection and nutrients for the microbiota within the colon. MUC2 is secreted from the goblet cells, forming a tight and sterile environment. *On the right,* is a brief diagram of the same mucosal layer in a compromised state with bacterial invasion of the inner layer. A brief MUC2 diagram is seen below[24].

One specification of mucins is the peptide backbone of mucins contains repeating domains that are rich in proline, threonine as well as serine residues that are heavily glycosylated by *O*-linked glycosylation on the serine and threonine residues, known as PTS domains [25,26]. The five secreted polymerization mucins found throughout the body (MUC2, MUC5AC, MUC6 and MUC5B) have all evolved from one single originating gene which explains their homologous structures, however, the amino acid composition of the PTS domain contained in each mucin is where they become distinct from one

another, allowing for unique glycosylation initiation sites and inevitably glycosylation patterns[27]. MUC2, found in the gastrointestinal system, typically consists of more than 5000 amino acids. It has two distinct regions with a high concentration of glycosylation sites. The first region is a sizable segment with approximately 100 repetitions of a 23-amino acid sequence (PTTTPITTTTTVTPTPTPTGTQT). The second region is a 347-amino acid domain that contains abundant glycosylation, both regions are separated by a ~150 amino acid spacer that is rich in cysteine (CysD)[13,28].

The significance of these mucin glycoproteins in preserving the host's health has been demonstrated through numerous experiments. These findings reveal that genetic abnormalities in the *MUC* genes frequently contribute to various diseases such as cystic fibrosis and inflammatory bowel disease[16,19,29].

1.2 Carbohydrate structure.

Within the gastrointestinal system, glycosylation is a common post-translational modification that accounts for around 80% of the total weight of the mucin glycoprotein protein[30]. Typically, carbohydrate chains are adhered to the peptide backbone of glycoproteins and protrude outward, giving these glycoproteins a 'bottle brush' shape[31]. For mucin glycoproteins to form the 'gel' network, hydrophobic interactions between surrounding negatively charged polysaccharide side chains in conjunction with the disulphide bonds between the carboxyl and NH_2 domains of the peptides discussed earlier[32,33]. The complex glycosylated oligosaccharide chains of the mucin glycoproteins play a key role in the biological and physical protection seen within the mucosal layer, lubricating the epithelia of the host, as well as providing protection against foreign adjuvants[4]. The incredibly extensive networks of glycan chains that initiate from the mucin cores are largely responsible for a fully glycosylated monomeric MUC2 weighing ~2.5 MDa[34].

1.2.1 O-linked mammalian glycans.

O-linked glycosylation is a common post-translational modification that occurs on the hydroxyl groups of serine and threonine residues in the mucin protein[5]. A Serine/threonine *O*-glycosylated with a single *N*-acetylgalactosamine (GalNAc) provides what's known as a Tn-antigen, one of the simplest glycan additions[35]. The *O*-linked Tn-antigen builds the base structure for eight glycosylation events found on mucins known as the "core groups". Among these groups, cores 1-4 are the predominant cores found in mammalian mucins that can initiate the formation of longer glycan chains. Modifications to the mucin cores can be done by linking GalNAc, galactose (gal), *N*-acetylglucosamine (GlcNAc), fucose (fuc), and sialic acid to create what is called the "extended Muc cores" (figure 2, core 2 and core 4 are given as an example of being extended)[36]. In eukaryotes, the *O*-linked Glycosylation of mucin glycoproteins begins in the Golgi apparatus as well as the endoplasmic reticulum, whilst in prokaryotes this event occurs within the cytoplasm[37,38]. Within these environments, the addition of the GalNAc from UDP-GalNAc to a serine/threonine is spearheaded by the enzyme family classified as polypeptide UDP-*N*-acetylgalactosaminyltransferase (GalNAc Ts)[39].

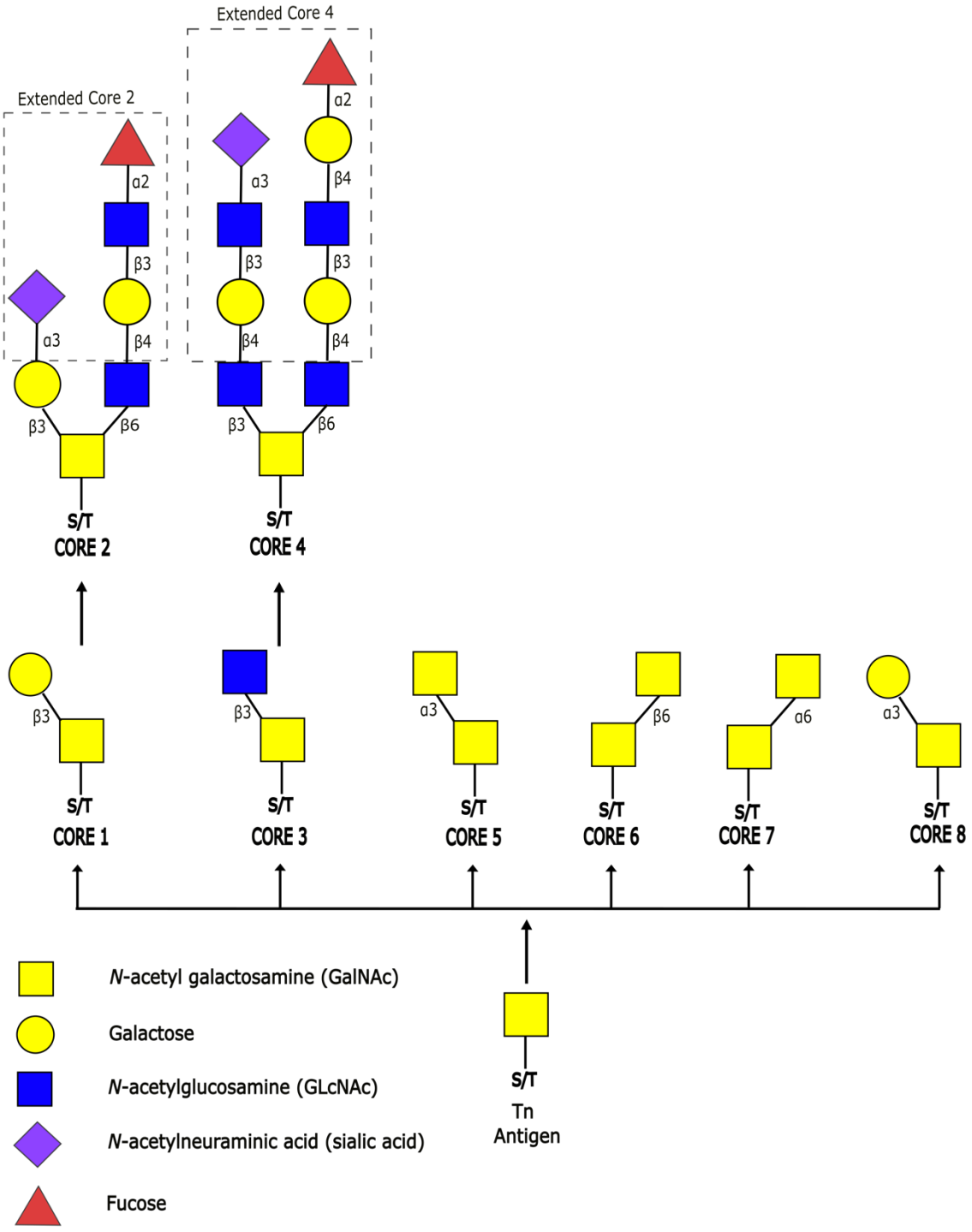


Figure 2. Mucin cores. All cores are generated from the single Tn-antigen using a serine or threonine residue. Glycan branching of GalNAc, galactose and GcNAc sugars generates the core structures. The extended versions of core 2 and core 4 are given as an example, where the extensions are highlighted in the dashed rectangle[40].

There are twenty known isoforms of GalNAcT that possess the ability to initiate glycosylation. Each of the isoforms is differentially expressed and each having peptide specificities outlines the complexity and specificity of mucin construction[41]. Further, the glycosyl transferase family responsible for extending the mucin cores is extensive due to the combinational variety of linked sugars that each addition can be attached to[42]. It is intriguing that the families of transferase enzymes used to extend the *O*-glycans contain a drastically smaller number of transferase enzymes that carry out the same function, in comparison to the 20 GalNAcT isoforms. For instance, one of the most common extensions of the Tn-antigen is through the Core 1 structure that is carried out by one of three 'C2GnTs' that add a GlcNAc through a β1–3 linkage to the GalNAc[42]. Core 2 is a branched version of core 1 containing a β1–3 linked GlcNac, that can be further extended by GlcNAc-galactose (*N*-acetyllactosamine repeats), commonly seen in mucin structures[43]. The structure of core 3 resembles core 1. In this instance, β3Gn-T6 transferase is used to create a β1-3 linked GlcNAc to the GalNAc, which can also be extended with a β1-6 GlcNac in a branched manner as we see in core 4[42,44]. Even though there are eight cores in total, cores one through four are the versions that are the most represented in humans, whereas cores 5-8 are rarer in nature[41]. Commonly incorporated into the extended version of these glycan structures, these core groups are also terminally modified through sialylation, fucosylation or sulfation which play a large role in effecting target binding[45].

O-linked glycosylation is a posttranslational modification that has a significant role in various important biological processes. These include blood group typing, antibody recognition, leukocyte receptor recognition, and particularly relevant to this thesis, the production of mucin[46–49]. Mucin production leads to the creation of attachment sites for commensal bacterial colonization and serves as an energy source for bacterial proliferation.

1.3 The microbiome.

It is said that the key to strong gut health is diet, the mucosa, and the commensal microbiota that lives within the host[50]. The microbiota is referred to as the conglomeration of commensal and pathogenic bacteria, fungi, archaea, viruses, and protozoa that is present within an environment[51], including inside and on multicellular plants and animals[52].

The definition of the microbiome has varied over time, where the most referenced definition by Dr. Joshua Lederberg describes it as an ecological system of commensal, symbiotic and perhaps pathogenic microorganisms that reside in the human body[53]. As recent as 2020, Berg *et al* have proposed that the microbiome is far more complex than this rudimentary definition and should now address: the members of the microbiome, interactions amongst members and the formed microbial networks, the spatial and temporal characteristics of microbiomes in their environment, the core microbiota, the phenotypes of the resident species, and the interactions between the microbiota and the host or environment it resides within[54]. The microbiome is an important player in maintaining the health and well-being of the organisms it shares a tightly connected relationship with. Concurrent with the immune system, humans are born essentially free of any residential bacterial presence, developing a microbiota when introduced to the world of microorganisms[51].

The importance of the microbiome is specifically noteworthy during times of dysbiosis, where there an imbalance of microbial community occurs, creating an opening for an opportunistic pathogen to create a foothold and proliferate[55]. These variations of the microbiota are correlated to several diseases that could be detrimental to the mucosal tissue, including type 2 diabetes, obesity, intestinal inflammatory bowel syndrome and even cardiovascular disease[56–59]. Not only does the microbiome affect health and disease, but the microbiome is also known for strengthening immunity and increasing nutrient availability through the release of beneficial short-chain fatty acids (e.g., acetic, propionic, and butyric acid) from dietary complex carbohydrate fermentation[60–62]. Given

that the microbiome of a mammalian host comprises trillions of microbes, it is impressive that most of this bacterial load resides within the gastrointestinal system[63]. The density and population notably increase closer to the distal regions of the gastrointestinal tract. For instance, the stomach contains roughly 10^{3-4} bacteria per mL of content (based on the volume of the organ) which is consistent with the bacterial load seen in the duodenum and jejunum, before reaching the denser regions of the ileum (10^8/mL). However, the colon stands out as having the highest bacterial concentration (10^{11}/L), translating to roughly 70% of microbes found in the human body[64]. Interestingly, the increasing bacterial concentration closer to the distal regions of the gastrointestinal tract correlates well with the increasing thickness of the mucin layer further into the gastrointestinal tract, as discussed in section 1.1. As previously mentioned, the gastrointestinal tract is a dynamic environment that affects the diversity of residential bacteria, which is estimated to be made up of over 400 bacterial species that typically fall under the phyla *Bacteroidetes, Firmicutes, Actinobacteria, Proteobacteria and Verrucomicrobia.* The phyla *Firmicutes* and *Bacteroidetes* represent 90% of the common microbiota[65]. Studies have indicated that the ratio of gastrointestinal residents varied dependent on diet and environmental conditions. For instance, obese mice have an increase in species belonging to the *Firmicutes* phylum, where a decrease in *Bacteroides* is typically evident in comparison to lean animals[66]. Following, one study indicates in a comparison between 123 non-obese to 169 obese individuals that those with low bacterial diversity exhibit increased adiposity and a more pronounced inflammatory profile[66]. With so much possibility for variation within one environment, it is challenging to provide a clear picture of what is considered a healthy microbiota. However, as studies continue and more exploration continues in the microbiome, research helps us understand the key relationship the microbiome plays in maintaining host health.

The composition and structure of the gastrointestinal system provide an environment that is rich with nutrients for bacterial species to forage from[57]. Even though the mucosal layer is commonly respected as the first line of defence in the gastrointestinal tract[67], there is a homeostasis between commensal species and this protective layer. The

bacterial community in this environment relies on the undigested complex polysaccharides from the host's diet, as well as the glycoproteins harvested within the mucus layer[68]. However, to break down the gigantic mucin glycoproteins efficiently and effectively, the residents of the gastrointestinal tract must generate strategies to deconstruct the biomass.

1.4 *Clostridium perfringens:* tailored for mucin degradation.

Clostridium perfringens is a Gram-positive, opportunistic pathogen that is commonly found within the gastrointestinal tract of humans and animals[69,70]. It is worth noting that the *Clostridium* genus accounts for 95% of the *Firmicutes* phylum, which is among the most prevalent bacterial phyla found in the gastrointestinal tract[65]. This endo-spore-forming bacterium is common in the eye of the public due to its widespread ability to cause numerous diseases that include food poisoning and diarrhea, necrotic enteritis (commonly associated with poultry, with an annual estimated cost of six billion dollars US globally), clostridial myonecrosis (gas gangrene), and human enterotoxaemia[71–74]. In the United States, there is an estimation that *Clostridium perfringens* is responsible for one million cases of foodborne illness each year, making it the second most common cause of food-based infections[75].

The onset of the diseases *C. perfringens* causes is due to the extensive tool belt of toxins that it can produce. The repertoire of roughly 20 toxins that *C. perfringens* can produce are classified as either an α -toxin, β-toxin, ε-toxin or ɪ-toxin[71,76]. Each strain of *C. perfringens* can be categorized by the assortment of toxins it can produce[73]. Each strain of *C. perfringens* is classified into two toxinotypes: isolates that produce *C. perfringens* enterotoxin (CPE) but not β-toxin, ε-toxin or ɪ-toxin are classified as type F strains that are commonly associated with food poisoning[71]. Strains that produce *C. perfringens* α-toxin (CPA) and the necrotic enteritis toxin B-like (NetB) are known as Type G strains and are responsible for necrotic enteritis in poultry[71,72].

Even though CPA and NetB are considered the essential toxins in *Clostridium perfringens* pathogenic strategy, this bacterium is also equipped with supporting virulence factors to assist in pathogenicity that includes extracellular enzymes[77,78]. The battery of secreted *C. perfringens* enzymes contains a diverse array of CAZymes, encoding 89 in total, such as *α*-L-fucosidases, glycoproteases, and glycosyl hydrolases which play a vital role in mucin degradation and pathogenicity (discussed further in greater detail) [69,74,79]. To show the mucin-degrading specialty of *Clostridium perfringens,* studies have shown that *C. perfringens* could not only grow on porcine and chicken mucus substrates, but it was also able to survive off chicken mucin as well as the monosaccharides sialic acid, mannose, galactose and GlcNAc as a sole carbon source[74]. This observation highlights the mucolytic capabilities of *C. perfringens* and its adaptation to mucosal invasion.

1.5 *Akkermansia muciniphila*: From discovery to probiotic.

Discovered only in 2004, *Akkermansia muciniphila* is a Gram-negative commensal organism residing in the gastrointestinal tract and accounting for approximately 3% of the microbiota composition[80,81]. *Akkermansia muciniphila* had been the only species from the *Akkermansia* genus until the discovery of *Akkermansia glycaniphila* in 2016[82]. Despite its late discovery, *A. muciniphila* has gained widespread attention due to its inverse relationship with the severity of disease including obesity hepatic inflammation and type 2 diabetes[81,83]. In addition to *Akkermansia muciniphila's* potential health benefits, this anaerobic organism is known for its proliferation within the mucus outer layer of the gastrointestinal tract[80,84].

Akkermansia muciniphila is considered beneficial in maintaining healthy microbiota and has even been introduced into the human diet as a probiotic supplement[84]. This organism appears to serve a purpose in maintaining the equilibrium of the gut mucus layer. Evidence of this was observed when an obese mouse with a depleted mucus layer was inoculated with *Akkermansia muciniphila,* resulting in the restoration of the mucus layer thickness, a result not seen when inoculated with heat-killed *Akkermansia muciniphila*[85]. This restoration is due to the organism's ability to break down mucin and

produces short-chain fatty acids, which can be absorbed through the epithelium to help stimulate the host epithelium to synthesize and secrete more mucin[86]. Therefore, *Akkermansia muciniphila* is likely responsible for promoting the homeostasis of the mucus layer and preventing the colonization of pathogens within the environment it resides in.

The naming of the organism, which translates to “mucin-loving”, reflects its proficiency in breaking down the mucus layer in which it resides. *Akkermansia muciniphila’s* role as a mucin specialist is demonstrated by its ability to utilize mucin as a sole carbon source, as well as to subsist solely on monosaccharides derived from mucin, including GalNAc, GlcNAc, galactose and fucose[87]. The upregulation of mucin-degrading enzymes by *Akkermansia muciniphila* plays a critical role in bacterial energy acquisition in the absence of dietary fibres[88]. Remarkably, 26% of the species proteome (567 proteins) contains a predicted signal peptide, where 11% of these secreted proteins (61 proteins) have been annotated as enzymes that are active on carbohydrates[89].

1.6 Carbohydrate Active Enzymes.

The prolific bacterial species that utilize host-associated glycans are equipped with an extensive tool belt of carbohydrate-active enzymes (CAZymes). Generally, CAZymes provide an essential tool for commensal species to graze on host carbohydrate chains, harvesting the abundant nutrient source found in the mucus outer layer[67]. However, gastrointestinal pathogenic species are also equipped with mucolytic CAZymes that are key for penetrating the mucus layer during invasion[90].

The enzymes classified as CAZymes can be split into five different groups with differentiating functions: Glycoside hydrolases, polysaccharide lyases, carbohydrate esterases glycosyltransferases as well as accessory domains[67]. All five enzyme classifications of CAZymes are responsible for the interaction and modification of carbohydrate chains, a niche ability in the world of enzymes. In brief, the function of CAZymes usually involves either the assembly of monosaccharides into

oligosaccharides or polysaccharide chains or inversely, the disassembly[91]. An enzyme that falls under the classification of glycoside hydrolase, polysaccharide lyase or esterase is responsible for the dissociation of the linkage to another constituent binding partner, which could be another carbohydrate, lipid, or protein posttranslational modification[92–94]. To date, glycoside hydrolases form the largest classification of CAZyme, containing 183 families according to the CAZyme database. The way that CAZymes are classified in the database is based on the structure of the domains that make up the protein. This often results in numerous hydrolase families that have the same substrate: an example being the families 5, 6, 7, 8, 9 12, 44, 45 and 48 all containing cellulase activity on cellulose, yet are still distinct due to their structural variations[95]. However, some families are more diverse in their substrate selectivity, an example represented in the group GH16 that contains members with κ -carrageenase activity towards marine red algae polysaccharides, while other GH16's can be found in gastrointestinal microbiota that have *O*-glycanase activity that is key in the initiating steps of breaking down mucin carbohydrates[95–97].

The most prevalent CAZymes within the gastrointestinal tract are from the enzyme class glycoside hydrolase (GH)[98]. Gut-residing bacteria are equipped with glycosyl hydrolase tools to trim off the GlcNAc, GalNAc, galactose, fucose, sialic acid and sometimes mannose mono and oligosaccharides from diet and the glycan chains found on mucin glycoproteins[69,99]. Subsequently, these processed substances are either imported into the periplasm to serve as an energy source or shared among the microbial community[99]. As the sugar building blocks that build the carbon chains in mucin glycoproteins are just a selection of the complex sugars found on Earth, this group of enzymes is found within all domains of life and takes on many different approaches to breaking down carbohydrate-derived biomass[100]. The function of a hydrolase is to hydrolyze a glycosidic bond of a carbohydrate linked to a separate molecular component. This idea is similar to the polysaccharide lyase that also breaks down carbohydrate bonds, differentiating in its mechanism by cleaving through hydrolysis instead of elimination that would be used in the mechanism of a polysaccharide lyase[101]. The third group of CAZymes that are useful in carbohydrate degradation are

the carbohydrate esterases that are responsible for cleaving the ester bonds between carbohydrates and acyl or acetyl groups[102]. The significance of an esterase in bacterial degradation of polysaccharides becomes evident when the example of plants evolving to contain a defence mechanism involving the acetylation of polysaccharides in their cell walls effectively inhibiting hydrolytic activity[103]. Bacterial utilization of an esterase deacetylates the polysaccharide, allowing for hydrolysis of the modified glycan structure. Commonly, a CAZyme with the capability to disassemble carbohydrates is associated with an auxiliary module known as a carbohydrate-binding module (CBM)[104]. These modules lack catalytic function, however, aid in their ability to bind carbohydrates and bring the catalytic machinery it is affixed to within proximity of the substrate. CBM's have been found to be associated with glycoside hydrolases, lyases as well as esterases in numerous bacterial and fungal species[105]. The organization of complex hydrocarbons into glycans can vary greatly in nature, resulting in dynamic structural arrangements, resulting in the evolutionary race to create these CAZymes that are able to build and degrade them.

The gut microbiota largely contains a greater arsenal of degradative enzymes than the host that it resides in, making it very proficient in targeting the various carbohydrates found in both the mucus layer and in the host diet[106]. Carbohydrates are an important source of energy for the human body, but not all carbohydrates can be absorbed through the small intestine. The undigested carbohydrates travel through the large intestine where they can be fermented by the resident gut microbiota, creating a mutualistic relationship between the host and bacterial residents[106]. However, during times of low dietary fibre, some species including *Bacteroides thetaiotaomicron* selectively induce their mucin-degrading glycosyl hydrolases to target host-derived carbohydrates[107]. What is even more unusual and quite fascinating, is bacterial evolution to adopt CAZymes into their genomes for carbohydrates that have been adapted into the diet of the host. According to Hehemann *et al.,* in Japanese culture, there is a large consumption of the species of red marine seaweed called *Porphyra,* in various culinary dishes. Usually, marine bacterial species that are common degraders of this type of biomass contain the CAZymes typical in depolymerizing these

carbohydrates. What this study discovered was the gut bacterium species *Bacteroides plebius*, through gene transfer from a marine species, took up these marine CAZyme genes including porphyranases and agarases enabling the degradation of this novel carbohydrate source[108].

As unique oligosaccharides evolved in all facets of life to create complex carbohydrate structures, it's natural that the gastrointestinal microbiota has adapted to process the carbohydrates provided by the host as well as through diet. Many organisms are known for having a high concentration of mucin-degrading enzymes including species of *Bacteroides, Clostridium* and *Akkermansia muciniphila,* that encode the genetic tools to disassemble up to 85% of structures found in mucin chains[109].

1.6.1 Sialidases.

The term sialidase (also known as neuraminidase) refers to the type of glycosyl hydrolase that is present in all kingdoms of life that is responsible for removing sialic acid from carbohydrates[110]. There are currently four different family classifications of sialidases. Sialidases belonging to families GH33, GH34 and GH83 are known for their exo activity, i.e., removing a terminal sialic acid off a glycan chain, where the endo-acting, GH53 can cleave in between glycan chains to remove a sialic acid[110,111]. One of the most well-known sialidase is the GH34 viral neuraminidase which is utilized in the invasion strategy of an influenza virus. To invade a host, the influenza virus locates on the exterior of the cell by binding the sialic acids on the host glycoproteins. However, as the virus requires mobility across the host cell, the GH34 is used to cleave the linkage, allowing for the influenza virus to relocate to another sialic acid to find a point of entry[112]. Both GH33 and GH34 sialidases cleave the glycosidic bond that retains the anomeric orientation at the cleavage point, following the mechanism seen in figure 4[113]. These retaining sialidases use an atypical tyrosine as a nucleophile that is activated by a local glutamic acid, compared to the classical retaining mechanism seen in most glycosyl hydrolases involving two residues with carboxylate groups[114]. After the nucleophilic attack at the anomeric carbon, a transition state occurs in this mechanism

as a proton is donated from a residue acting as a local general acid/base. To complete the catalysis, the general acid/base accepts the proton, allowing for the hydrolysis of the glycosidic linkage, and the release of the covalently attached catalytic tyrosine.

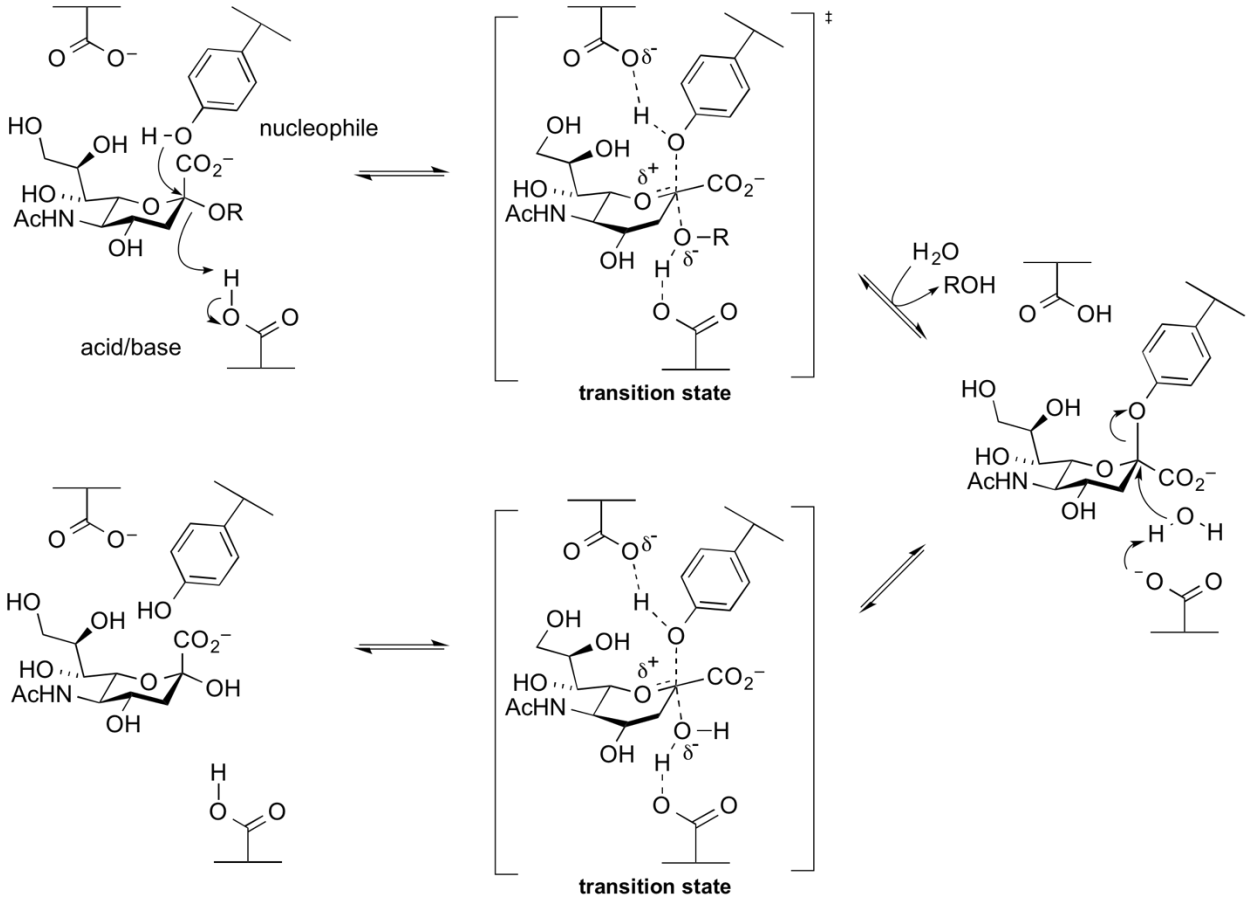


Figure 4. Retaining mechanism used by GH33 and GH34 hydrolytic sialidases. This mechanism figure was sourced from Sialidases from Juge, *et al.*; Gut Bacteria: a mini-review[21].

Sialic acids are unique compared to the sugars found in vertebrates that typically are built with five to six carbon atoms, whereas sialic acids are 9-carbon sugars[115]. These α-keto sugars are spread widely throughout animal tissues, and found in viruses, bacteria, fungi and protozoa but are not present in plants[116]. In glycoconjugates found within human tissue, *N*-acetylneuraminic acid (Neu5Ac) is the most prevalent form of sialic acid, yet through post-translational modifications, there are more than 50 naturally

occurring sialic acids[117]. Neu5Ac can be modified through acetylation, lactylation, sulfation, phosphorylation, methylation, lactonization and hydroxylation[118]. What is interesting is that human evolution has removed the production of one of the other most common sialic acids, *N*-acetylglyconeuraminic acid (Neu5Gc) that is often found within many mammals including the closest evolutionary-related primates[119]. It is suggested that ancestral hominins escaped malarial parasites that preferentially bind Neu5Gc by eliminating Neu5Gc production, causing the over-representation of Neu5Ac in humans[120].

The sialic acid monosaccharide can be linked to other glycans through three main configurations, either as an α2,3, α2,6 or α2,8, where α2,8 is considered slightly more niche within human cells[121]. Within the gastrointestinal tract where sialic acid is a 'terminal cap' on the carbohydrate chains of the mucin, most of the sialic acid is linked via an α2,3 or α2,6 linkage[21]. Since this terminal sugar impedes other glycosyl hydrolase enzymes from accessing the underlying glycan chain, sialidase recognition and removal of the sialic acid is considered key for the initiation of some mucin carbohydrate breakdown strategies[122]. One strategy to prevent the removal of the sialic acid is the *O*-acetyl ester modifications to resist the recognition by sialidase enzymes, often adding acetyl groups onto the carbon 7 or carbon 9 of the sialic acid[123]. Therefore, this results in an evolutionary arms race for bacterial species to develop sensitivity to these acetylated sialic acids to harvest[124]. Bacterial strains that reside within the gastrointestinal tract are commonly adapted to the removal of sialic acids due to the high abundance within their environment. This includes multiple species of *Clostridia, Bifidobacterium, Bacteroides* and *Akkermansia muciniphila* that encode sialidases[21,69,125]. Now what happens to the sialic acid differs depending on the bacterial strain that's doing the removal. *Streptococcus pneumoniae* D39 possesses the GH33 sialidase NanA and harbours the *nan* gene cluster which encodes the genes for importing and metabolizing sialic acid[126]. Consequently, the bacterium can utilize the sialic acid cleaved by NanA as a source of carbon, energy, and nitrogen. On the other hand, species like *Bacteroides thetatiotaomicron* encode the sialidase, yet lack the *nan* operon to utilize the sialic acid as a nutrient source[127]. In the instance of mucin

degradation, this strategy is likely used for the access of the underlying glycan chains of the sialoglycans, and the sialic acid will often be taken up by other species of the gastrointestinal tract that possess the catabolic machinery.

1.6.2 The sialidases of *Clostridium perfringens*.

Included in the *C. perfringens* mucin degrading toolbelt are three *exo-a*-sialidase genes named *nanJ, nanI and nanH* that are classified in the glycosyl hydrolase family 33 (GH33) (figure 5)[69]. Typically, sialidases in the gastrointestinal tract are known to initiate mucin degradation by cleaving the protective terminal sialic acid from the underlying glycan chain of the mucin[69]. It is uncommon for an organism to possess three distinct enzymes performing the same function, as redundant systems are not typical in nature.

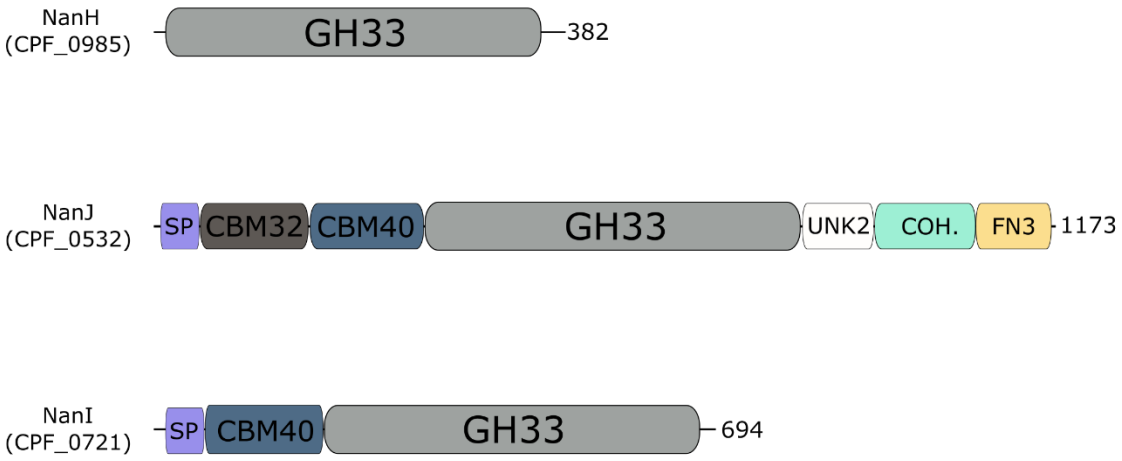


Figure 5. Modularity of the *nan* genes from *Clostridium perfringens* strain ATCC13124. *nanH* is constructed solely of the GH33 catalytic domain. *nanJ* contains 6 different modules and a signal peptide, from the N-terminus: a CBM32, CBM40, GH33 (catalytic module), a module of unknown function, an "X-module" of unknown function belonging to family 82, and a fibronectin type III-like domain. *nanI* also contains a signal peptide that is N-terminus to a CBM40 domain, followed by the catalytic GH33 module.

The three sialidases are constructed from the GH33 catalytic module but differentiate from one another in their extra modular composition. The *nanJ* and *nanI* genes encode N-terminal CBMs from different families. Similarly, the GH33 catalytic module in both genes is C-terminal to a CBM40, however, the *nanJ* gene also encodes an N-terminal

CBM32 not present in *nanI*[69]*.* On the contrary, the *nanH* gene completely lacks any secondary modules, existing solely as a catalytic GH33. What is most fascinating is that the *nanJ* and *nanI* genes both have a signal peptide that indicates these enzymes are likely secreted extracellularly. As targeting and degrading mucin in the environment is a priority for *Clostridium perfringens*, the secretion tag is logical. However, the *nanH* gene does not encode a signal peptide. This suggests that the secreted NanJ and NanI sialidases act as virulence factors for the pathogen by interacting with sialylated host glycans directly, where NanH likely cleaves the sialic acid group off glycans imported into the bacterium. *Clostridium perfringens* encodes several different sugar transporters, as well as putative phosphoenolpyruvate-dependent phosphotransferase systems (PTS) that are known for trehalose and sucrose uptake, suggesting oligosaccharide import is plausible[128–130].

Previous literature has described in detail the hydrolase activity between sialidases and sialic acids, including studies on NanI[131]. Crystal structures of the CBM32 of NanJ have been solved in complex with either galactose or GalNAc, and the CBM40 has been complexed with sialic acid[132]. These modules likely function to aid the sialidase enzyme in the targeting of glycans that terminate with either galactose/GalNAc or sialic acid[132,133]. The catalytic domains of NanI and NanJ from *Clostridium perfringens* strain ATCC13124 exhibit a significant similarity with a sequence identity of 60%. However, when compared to NanH, NanJ and NanI show much lower identities at 27% and 29%, respectively. Not only can strains of *Clostridium perfringens* that contain the Nan sialidases support growth on mucin media, but evidence has also shown its ability to grow on free sialic acid - further suggesting the use of these sialidases in bacterial survival[134,135].

To conclude, the utility of a sialidase enzyme has been harnessed by many organisms across multiple kingdoms to remove the sialic acid from glycan chains. It is generally accepted that this is an initiating step in bacterial utilization of the glycans found within the gastrointestinal mucins, whether as a method of pathogenic invasion, or potential nutrient harvesting.

1.7 Peptidases.

A peptidase, synonymous with a protease, is a type of enzyme that specializes in the degradation of peptides into smaller peptides or amino acids through hydrolysis. The importance of breaking down protein material is signified by the conservation of peptidase enzymes within all kingdoms of life[136]. Peptidases are incredibly useful tools, highlighted by the fact that ~ 2% of mammalian genomes encode peptidase genes that will be used in the development, coagulation, cell death, inflammation, and T-cell activation[137,138]. With the importance of proteolysis being essential for many natural biological processes, evolution has driven the peptidase family to unique specializations. The classification of peptidases can be broken down into several groups based on the residues used in its catalytic machinery. Conveniently, the nomenclature of each group is signified by its catalytic motif. To date, peptidases are broken down as either an aspartic acid, cysteine, serine, threonine, glutamic, asparagine or metallo-peptidase, with the metallopeptidase being the second largest family next to cysteine peptidases according to the MEROPS database[139].

A metallopeptidase is unique in the sense that the catalytic mechanism of the enzyme requires a metal ion which is often zinc, but sometimes cobalt, manganese or nickel[138]. Most metallopeptidases contain the consensus zinc-binding motif “HEXXH”, where the histidines are responsible for the chelation of the metal ion, and the glutamate is the residue used in catalysis (figure 6)[140]. Most of the members of the metallopeptidase family function by standard Michaelis-Menton kinetics: binding a peptide at the active site of the enzyme that is in proximity with the metal ion, followed by hydrolysis of the peptide using a single-displacement reaction[141]. Once the peptide is bound, the metal ion helps polarize the carbonyl oxygen from the peptide, therefore promoting nucleophilic attack by a solvent water molecule, helping transfer the proton to the catalytic glutamate residue[141]. Following, the zinc ion is further used to stabilize the tetrahedral intermediate, before a double proton transfer (onto the catalytic glutamate and new amino end), resulting in the cleavage of the scissile bond[141].

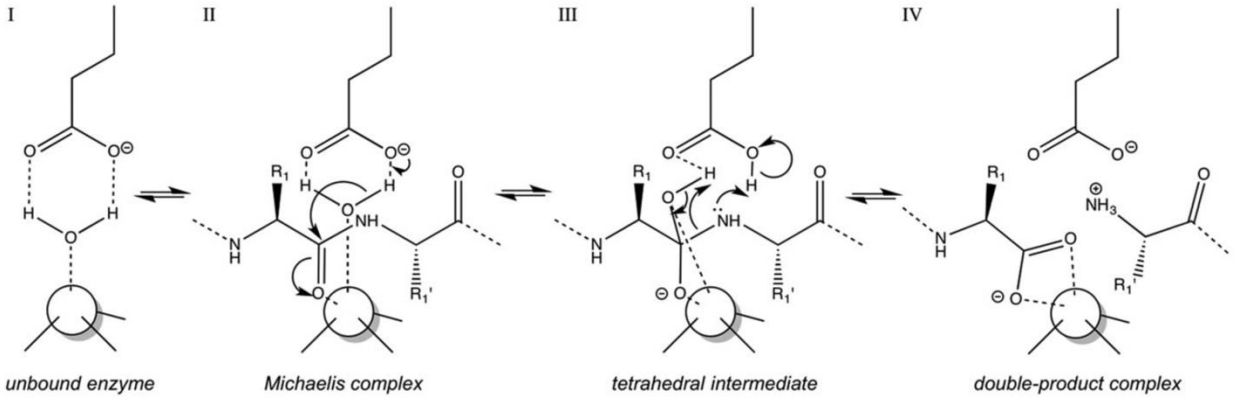


Figure 6. Metallopeptidase cleavage mechanism. Carboxyl group represents the catalytic glutamate, assisted by the zinc ion in proximity[142]. (Figure used from: Architecture and function of metallopeptidase catalytic domains, Cerda-Costa *et al.* 2014)[141].

One subdivision of the metallopeptidase clan is named the metzincin family name the loop containing a conserved methionine in the active site, known as the “Met-turn” and the requirement of a metal ion in catalysis[136,141]. Metzincins are calcium-dependent peptidases that are reported to play an important role in many biological processes including cancer-progressing related processes, degradation of the extracellular matrix and protein-protein complexes[143]. Peptidase enzymes, like CAZymes, may also possess supplementary domains used for multiple purposes including binding onto specific motifs. One type of domain that can be found coded next to the catalytic domain of a metallopeptidase is carbohydrate-binding modules.

In enzymatic mucin degradation pathways, the CAZymes found within the gastrointestinal environment can only degrade the carbohydrate chains of the mucin glycoproteins, leaving trimmed glycopeptides which require alternative strategies to complete the degradation. Often these glycoproteins will be decorated with varieties of simple carbohydrate decorations including the mucin core groups, T and Tn antigen moieties that are treated as defence mechanisms against peptidases, clashing with the enzyme’s active site and impeding peptide breakdown[144]. Interestingly, there are certain peptidase enzymes that can account for this by incorporating a unique hydrophobic pocket that can accommodate the glycan decoration on the peptide. These peptidases, classified as glycopeptidases, are a great tool utilized by many species to break down

the trimmed glycoprotein into smaller peptides and single amino acids. These glycopeptidase enzymes require both peptide and glycan to function, i.e., this enzyme class will not cleave a peptide without glycan decorations or oligosaccharides unbound to a peptide[145].

With upwards of 50% of proteins thought to be glycosylated through post-translational modification, the existence of regulatory glycopeptidases is evident in nature[145]. The metallopeptidases that have the dual specificity for both peptide and glycan often contain accessory domains for aiding in glycopeptide binding. For instance, the glycopeptidases from *Clostridium perfringens* ZmpA, ZmpB and ZmpC classified in the M60 MEROPS family all contain multiple *N*-terminal CBM 32 domains as well as multiple CBM51 domains on the *C*-terminal side of the catalytic domain[146]. Even a more specific example would be the family M88 IMPa from *Pseudomonas aeruginosa which* contains an *N*-terminal module that is not only used for *O*-linked glycan recognition on a serine/threonine in the glycopeptide but also contains a proline recognition domain that specifies cleavage sites only containing a proximal proline amino acid[144]. The glycopeptidase substrate recognition follows the peptidase nomenclature that contains active site subsites known as 'S-sites', and 'P-residues' that refer to the amino acids and peptides binding to the
active site (figure 7)[146].

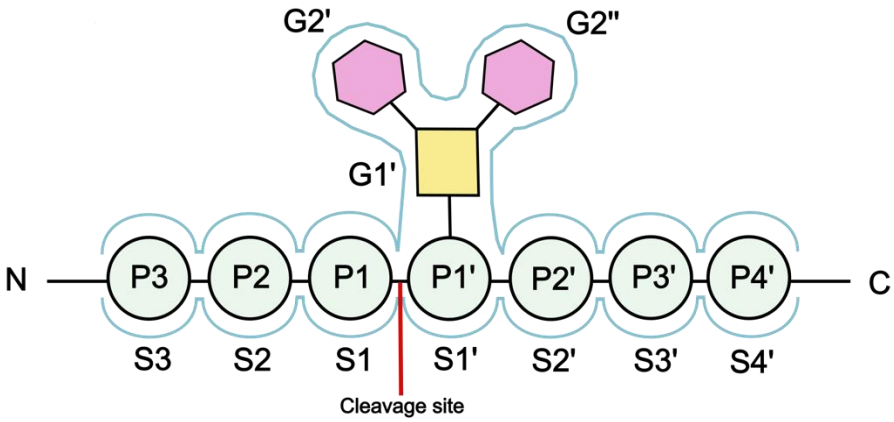


Figure 7. Generalized binding sites of a glycopeptidase. The peptide binding sites are labeled as S1-S3 on the N-terminus of the cleavage site, and S1'- S4' on the C-

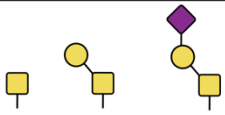
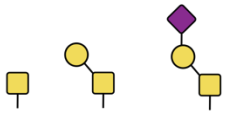
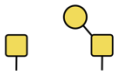
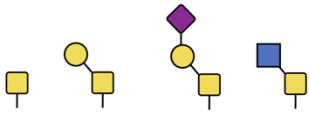
terminus, accommodating the glycopeptide. The glycan binding pockets are seen attached to the amino acid (serine or threonine) in the P1' position, starting with the G1' pocket. The Yellow square in the G1 site represents the GlcNAc sugar, and the pink hexagons are representative of generalized sugars.

The amino acid that is directly C-terminal to the peptide bond being hydrolyzed (serine/threonine in *O*-linked glycans) is labelled as S1', followed by S2' and so forth. Correspondingly, S1 is directly N-terminal, followed by S2. The glycan binding site that coordinates with the *O*-linked GalNAc on the serine/threonine in the P1' binding site is labelled G1' followed by the incremental binding site G2'. If a glycopeptidase can recognize branched glycan chains, the first branched glycan binding site is known as G2''.

1.7.1 Glycopeptidases from *Akkermansia muciniphila*.

Glycopeptidases are crucial tools for the enzymatic mucin degradation strategies of both pathogenic and commensal species. Assisting in the microbiota's ability to fully deconstruct the glycoproteins constructing the gastrointestinal mucosal layer, *Akkermansia muciniphila* encodes and deploys *O*-glycopeptidases that target the mucin backbone have been depolymerized by the secreted CAZymes. To date, there have been four characterized glycopeptidases that have been reported on from *Akkermansia muciniphila* (other than the glycopeptidase discussed in this document). Three of them with the locus tags Amuc_0627, Amuc_0908 and Amuc_1514 are documented as M60-like and cleave peptide bonds N-terminal to a serine or threonine that bears *O*-glycosylation[147]., OgpA (Amuc_1119) is the fourth enzyme with glycopeptidase activity from *Akkermansia muciniphila*. , OgpA is classified as an M11 peptidase according to the NCBI database, however, there are no M11 entries from *Akkermansia sp.* In the MEROPS database based on the structure of the enzyme[148]. All four of these glycopeptidases have similar, yet slightly distinctive glycan specificity (table 1)[148].

Table 1. Known glycopeptidases from *A. muciniphila.* MEROPS classification is referring to the catalytic domain of the enzyme.

Peptidase	MEROPS	Glycan Specificity
Amuc_0627	M60	
Amuc_0908	M60	
Amuc_1514	M98	
Amuc_1119 (OgpA)	M11	

As seen in table one, all four enzymes are sensitive to T and Tn antigen (T antigen being a Tn-antigen with an β1-3 branched galactose)[149]. Amuc_1514, belonging to family M98 is the only one out of the four that is not active on sialylated glycan-bearing peptides. Interestingly, OgpA is also sensitive to peptides bearing the core 3 moiety (GalNAc β1-3 GlcNAc), not seen by the other glycopeptidases[148,149]. The activity of OgpA, AM0627 and AM1514 significantly decreased when a terminal sialic acid was attached to a T-antigen, indicating that the desialylation of the mucin cores is crucial for these glycopeptidases to breakdown the mucin peptide core[147,148].

In brief, *A. muciniphila* is well equipped with enzymes that targets peptides that express the common MUC core-group glycans found in the mucus glycoproteins. However, within the multiple studies of *Akkermansia muciniphila* mucinase enzymes, one putative glycoprotease enzyme had yet to be explored. A bioinformatic analysis has outlined that there is an unexplored gene Amuc_1438, that could share similar structural and functional properties to the previously discovered glycopeptidases[87].

1.8 Objective and Hypothesis.

The enzymatic pathways used by commensal and pathogenic species incorporate a consortium of enzymes that play a part in the degradation and consumption of mucin glycoprotein. This thesis aims to elucidate the roles of two enzymes, namely the glycopeptidase Amuc_1438 from *Akkermansia muciniphila* and the intracellular sialidase NanH from *Clostridium perfringens,* within the enzymatic pathways involved in mucin processing used by gastrointestinal species. Amuc_1438 was selected for characterization as it had yet been flagged as a mucin-degrading enzyme, and no enzymatic or structural studies had yet commenced. NanH was also chosen as it had also very little previous characterization and little insight into its role in depolymerizing mucin oligosaccharides. This thesis hypothesizes that: ***Amuc_1438 encodes a glycopeptidase that exhibits a specific attraction to mucin-based glycopeptides that have undergone trimming by CAZymes. Furthermore, it is proposed that NanH selectively focuses on imported short sialylgylcans displaying a distinctive specificity when compared to the extracellular sialidase NanI.*** Chapter two focuses on investigating the potential involvement of NanH in degrading imported sialylated oligosaccharides from the mucosal region. This investigation includes the analysis of complexes formed between NanH and various sialylated glycans present in the environment. In chapter three, the thesis explores the structural characteristic and enzymatic functionality of the previously undiscovered fifth glycoprotease, Amuc_1438, originating from *Akkermansia muciniphila,* through a comprehensive examination.

Chapter 2: Exploration of Substrate Specificity of the Intracellular Sialidase NanH from *Clostridium perfringens*

Notice of collaboration:

This chapter was part of a collaboration: the Neu5,7,9Ac-α(2,6)LAcNAc substrate was created and sent by the Geert-Jan Boons group from Utrecht University.

2.1 Introduction.

The recognition of sialylated glycan chains in the gastrointestinal tract is a crucial process in regulating the thickness of the mucus layer and promoting the growth of commensal species through nutrient harvesting. Alas, pathogenic species of the gastrointestinal system have adapted to recognize the same sialylated glycans; often used in host invasion strategies. These pathogens are responsible for causing incredible tissue damage that promotes numerous diseases[150]. *Clostridium perfringens is* one of the more commonly known gastrointestinal pathogenic species known, largely in part by its correlation with food-borne illness in humans, and necrotic enteritis in poultry[151]. Necrotic enteritis in poultry is characterized by necrotic lesions in the intestinal mucosa, associated with the pore-forming cytotoxin NetB from *C. perfringens* type A strains[152]. Following, *Clostridium perfringens* uses its plethora of CAZymes to dismantle the complex network of glycoproteins found in the mucus barrier to access the epithelial cells while harvesting the released glycans as a nutrient source[69].

Known for its mucin desialylation behaviour, the genome of *C. perfringens* encodes three sialidase enzymes NanJ, NanI, and NanH[69]. Both NanI and NanJ are multi-modular enzymes containing a GH33 catalytic domain that shares roughly 60% sequence identity between catalytic domains. Both sialidases contain domains used in carbohydrate binding, and a signal peptide for extracellular protein secretion (figure 4)[69]. However, NanH is a curious case as it exists solely as a GH33 domain that lacks signal peptide, indicating that it likely functions intracellularly and never encounters the large mucin glycoproteins. Previous studies have solved the structure of NanI using X-ray crystallography, revealing the common sialidase six-bladed β-propeller catalytic fold both apo versions and complexed with sialo-derived ligands[131]. This structural insight of

ligand-bound NanI can suggest the catalytic activity and specificity of NanI in the presence of sialo-mucins[131,153].

As previously mentioned in section 1.4, CAZymes (including sialidases[154]) can contain multiple subsites used in more complex interactions with glycan chains. Sialic acids can be linked to many different combinations of sugars and linkages within the gastrointestinal tract, reasoning why NanH might contain subsite binding pockets that differentiate from NanI and NanJ to accommodate unique sialoglycans. To date, there has yet to be any structural evaluation completed on NanH from *Clostridium perfringens*, therefore, there has been no reasoning for why *Clostridium perfringens* would require a third, intracellular sialidase when it already secretes two. The presence of NanH leads to the **hypothesis that the NanH catalytic domain has a sialo-glycan specificity used in nutrient harvesting that is not shared with its extracellular counterparts NanI and NanJ.** To elucidate the binding capability of NanH, this thesis reports on unique X-ray crystallography structures of NanH in complex with multiple glycan chains. Also, to provide a comparison with extracellular counterparts, I am using the previously published NanI catalytic structures as a reference (as the catalytic domain of NanJ has yet to be published)[131].

2.2 Results.

2.2.1 Overall structure of NanH and its catalytic machinery.

Using X-ray crystallography techniques and molecular replacement, the apo structure of NanH was solved to 1.85Å using the apo structure of NanI as a template (figure 7a). Like NanI, NanH has the classical 6-bladed β propellor folding pattern seen in *exo*-active sialidase GH33 domains with a significant structural similarity (figure 7b). An overlap of the NanH sialidase and the catalytic domain of NanI reveals the structural similarity between enzymes. One noticeable difference is the existence of the loop in NanI catalytic domain between residues 361- 427 (the orange loop), which is not present in NanH. This loop is not within proximity of the catalytic site of the sialidases, therefore likely not directly impacting the specificity of the sialidase.

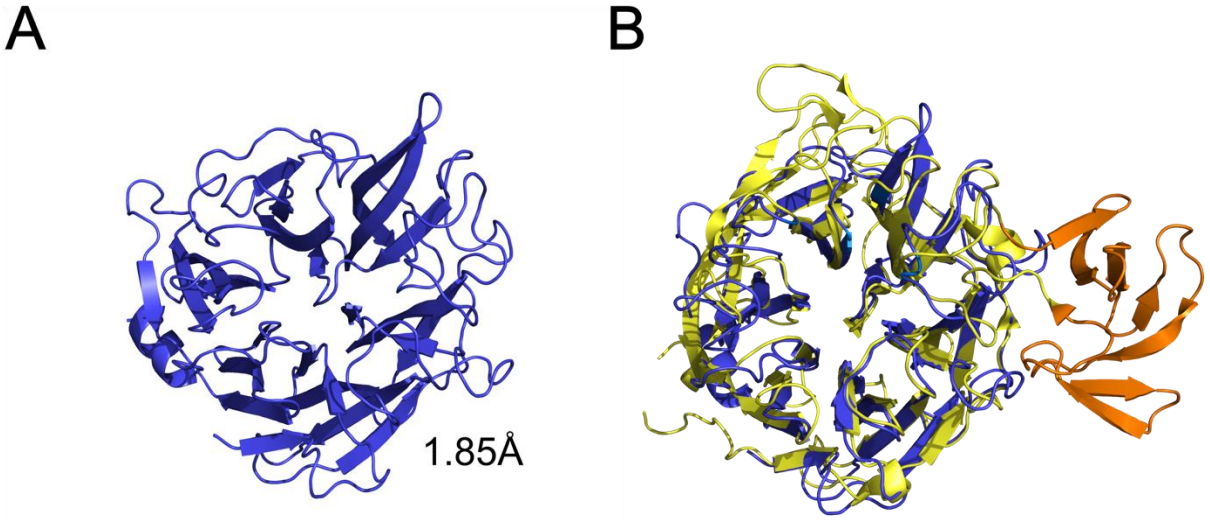


Figure 7. Structure of NanH sialidase. A. Structure of NanH, solved to 1.85Å. *B.* Structure of NanH overlapped with the catalytic domain of NanI (seen in yellow). The 361-427 loop in NanI is highlighted in orange (PDB code for NanI: 2VK7).

In the active site of the NanH sialidase in complex with the sialic acid Neu5Ac, the carboxylate group of Neu5Ac is accommodated by the argininal triad consisting of Arg37 Arg312 and Arg245 as depicted in figure 8a. The nitrogen species of the guanidino groups of the asparagine residues are situated between 2.84-2.98Å of the carboxylate group of Neu5Ac, a sialic acid stabilization mechanism similarly used in NanI[131]. The bottom of the binding pocket that accommodates the Neu5Ac carboxylate group is supported by Tyr347 and Glu230. In NanH's hydrolysis mechanism, Tyr347 is utilized as the catalytic nucleophile and faces its hydroxyl group 3.2Å away from the anomeric carbon of a Neu5Ac[155]. The Asp62 is required in the catalytic mechanism by standing in as a general acid/base, donating a proton during the nucleophilic attack of the tyrosine and generating the transition state. A partnering glutamate group to the Tyr347, Glu230, finds itself within distance for hydrogen bonding with the hydroxyl group (~2.53-3.11Å).

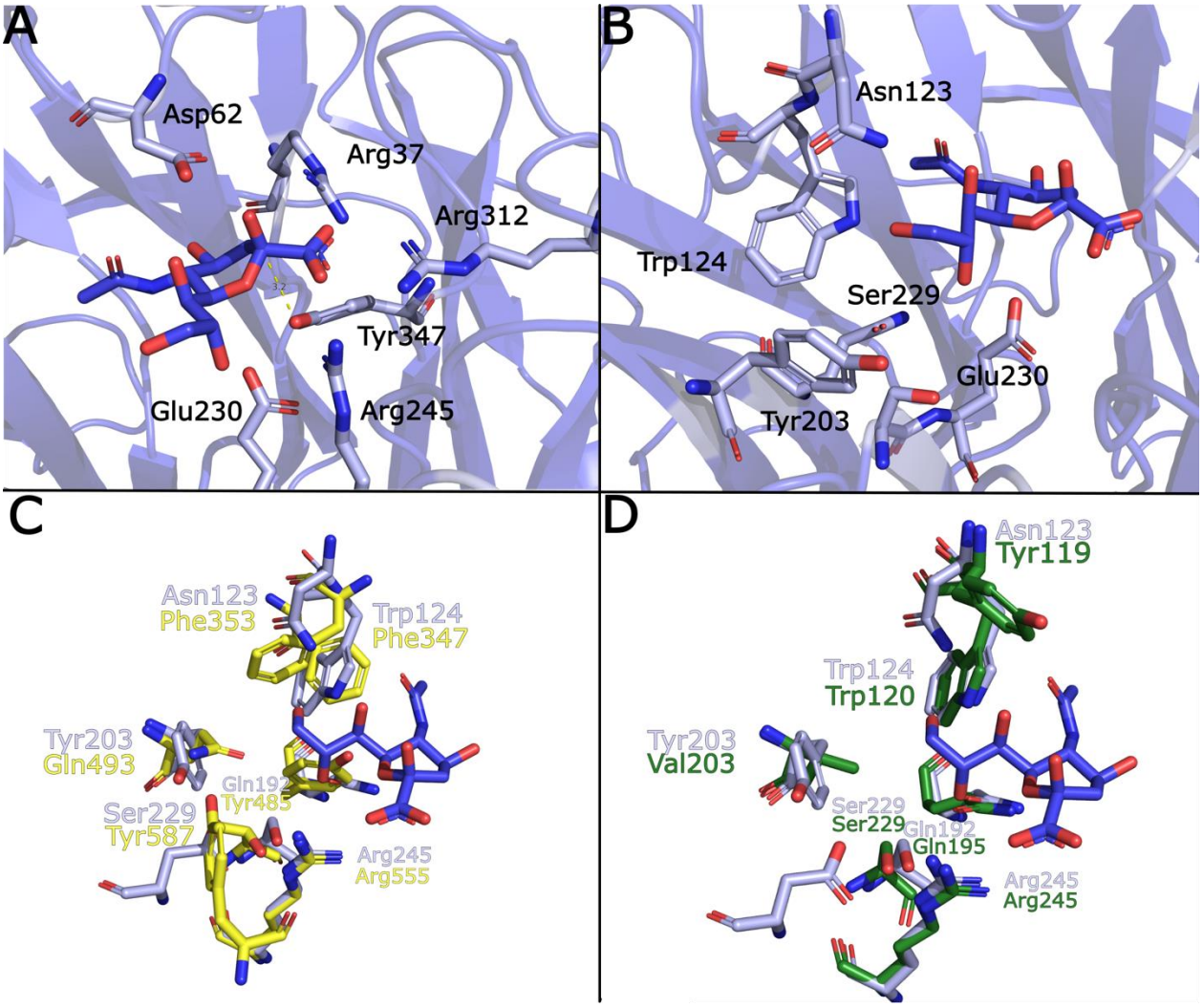


Figure 8. Interacting residues between the -1 active site in NanH and Neu5Ac. *A.* The arginines in position 37, 312 and 245 form the argininal triad. General acid/base Asp62 is seen above of the anomeric carbon of Neu5ac, while the rest of the catalytic machinery (Tyr347 and Glu230) are positioned underneath for hydrolysis. *B.* stabilizing residues that build the binding pocket for the C7-C9 tail of Neu5Ac. *C.* NanH (light blue) active site residues overlayed with NanI (yellow) interacting with Neu5Ac (NanI PDB code: 2BF6) *D.* NanH (light blue) active site residues overlayed with GH33 trans-sialidase from *Trypanosoma cruzi* (green) interacting with Neu5Ac (PDB code: 1s0i).

As previously mentioned, there are over fifty known sialic acid derivatives, where one common modification region of the sialic acid is the tail consisting of carbons C7-C9[117]. This highly modifiable sugar requires sialidases to adapt to the sialic acids they interact with, often modifying the binding pockets within its active site to adapt to the sialic acids it's targeting. Therefore, not all sialidases conform to the same active site structure, even within members of GH33. The active site of NanH constructs a pocket that

accommodates this C7-C9 tail, built out of Asn123, Trp124, Tyr203, Arg245 and Glu230 (figure 8b). A structural alignment of the active sites of NanH and NanI shows the variation in residues that construct this pocket. Out of the six main residues that build this pocket, only Arg245 from NanH and Arg555 from NanI share identity and location (figure 8c). Interestingly, the other residues in NanH and NanI with identical locations vary greatly in functional groups and aromaticity. To compare to a non-clostridial sialidase, NanH was aligned with one of the most well-known sialidases belonging to the GH33 family: the trans-sialidase from *Trypanosoma cruzi* that shares ~32% sequence identity with NanH[131,155]. Contrary to the variation of residues seen when aligning with NanI, NanH utilizes surprisingly similar residues to interact with the C7-C9 tail of sialic acid (figure 8d). Only Tyr203 and Tyr123 of NanH didn't correspond to the eukaryotic sialidase.

2.3.2 Structural preference of α(2,3) over α(2,6) sialylglycans.

The sialic acids that terminally decorate mucin glycoproteins are predominately linked through an α2,3 or α2,6 glycosidic linkage[156]. One study tested the sialic acid glycosidic linkage sensitivity of the *Clostridium perfringens* sialidases by using a double knockout of the two sialidases that were not being tested and then subjecting the *C. perfringens* culture supernatant (extracellular) or sonicated cell culture (intracellular) to sialylated glycans and measuring for free sialic acid, indicating cleavage had occured[157]. This rudimentary experiment suggested that NanH shows a preference for *α*2,8 > *α*2,3 > *α*2,6 linkages. However, this test only concluded that the sonicated cell culture shows some preference for this linkage order and did not test isolated enzymes to assess the sensitivity of the sialidases in vitro. To get a better understanding of the glycosidic linkage preference of NanH, multiple sialoglycans were soaked into inactivated NanH protein crystals to gain crystal complexes.

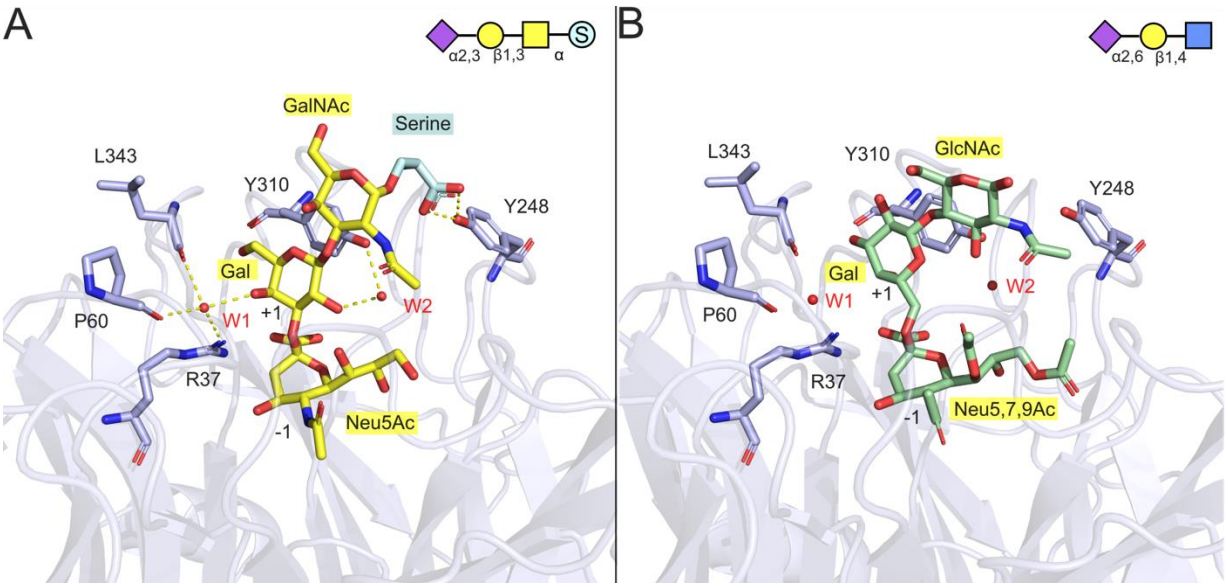


Figure 9. Crystal complex structures of NanH with variable sialoglycans. The subsites are labeled as -1 and +1. *A,* NanH in complex with sialylated α(2,3)T-antigen seen in yellow. The serine is highlighted in cyan, connected to the GalNAc on the glycan chain. Sub-site residues are represented in stick conformation, and conserved water molecules are labelled as W1 and W2 *B,* NanH complexed with Neu5,7,9Ac-α(2,6)LAcNac seen in mint. Sugar representations are in the top right corner of each panel.

In this study, Neu5,7,9Ac-α(2,6)LAcNac (SLN) and sialylated α(2,3)T-antigen (STag) were soaked into catalytically inactivated NanH crystals (D62N) to trap ligand in the active site. A galactose sugar in both the STag and SLN structures sits within the -1 subsite of the NanH active site (figure 9a). Within this complex, the oxygen species on carbon four of the galactose (O4) is within 2.8Å of the water molecule labelled W1, which is conserved throughout the NanH structures reported in this thesis. This water molecule is siatuated in proximity to the anomeric carbon of the sialylated species, perhaps contributing to the hydrolysis of the glycan linkage incoming sialyl glycan. Residues P60, R37 and L343 are within hydrogen bonding proximity (2.4Å, 3.1Å and 2.9Å, respectively) with W1, likely impacting the stability of this ligand. Likewise, the O2 oxygen group of the galactose is sharing a stabilizing interaction with the functional group of Y310 by hydrogen bonding with the W2 molecule (3.0Å and 2.6Å respectively). Interestingly, even though NanH successfully captured two trisaccharide substrates, there are no direct residues that interact with the GalNAc sugar bound to the serine suggesting, that NanH does not contain a +2 sugar-binding site. On top of that, the

carboxyl functional group of the serine attached to the GalNAc is directed toward Y248 and within 3.0Å (figure 9a). Almost all the interactions mentioned in the STag complex do not apply to the NanH complex of SLN. This ligand introduces a α(2,6) linkage between the Neu5Ac and the extended glycan. This linkage repositions the +1 galactose where it is no longer within interacting distance with W1 and W2, therefore providing a less stable interaction with the active site of NanH.

Perhaps the most important interaction between these complexes is through the CH – π stabilization provided by Y310. The galactose residues of STag and SLN align within the +1 binding site creating an interface between the face of the 6-membered carbohydrate ring and the aromatic Tyrosine 310 (figure 10a). The STag complex is orientated so that hydroxyl species on the galactose (O4) faces away from this stacking interface, presenting axial protons to interact with the electron-rich aromatic tyrosine through a CH – π interface (<5Å distance)[158]. However, the introduction of an α(2,6) linkage in the SLN ligand reorientates the O4 hydroxyl species and directs it towards the aromatic face[159]. This reorientation could have the potential to cause steric hindrance which could affect stability of the CH – π interface.

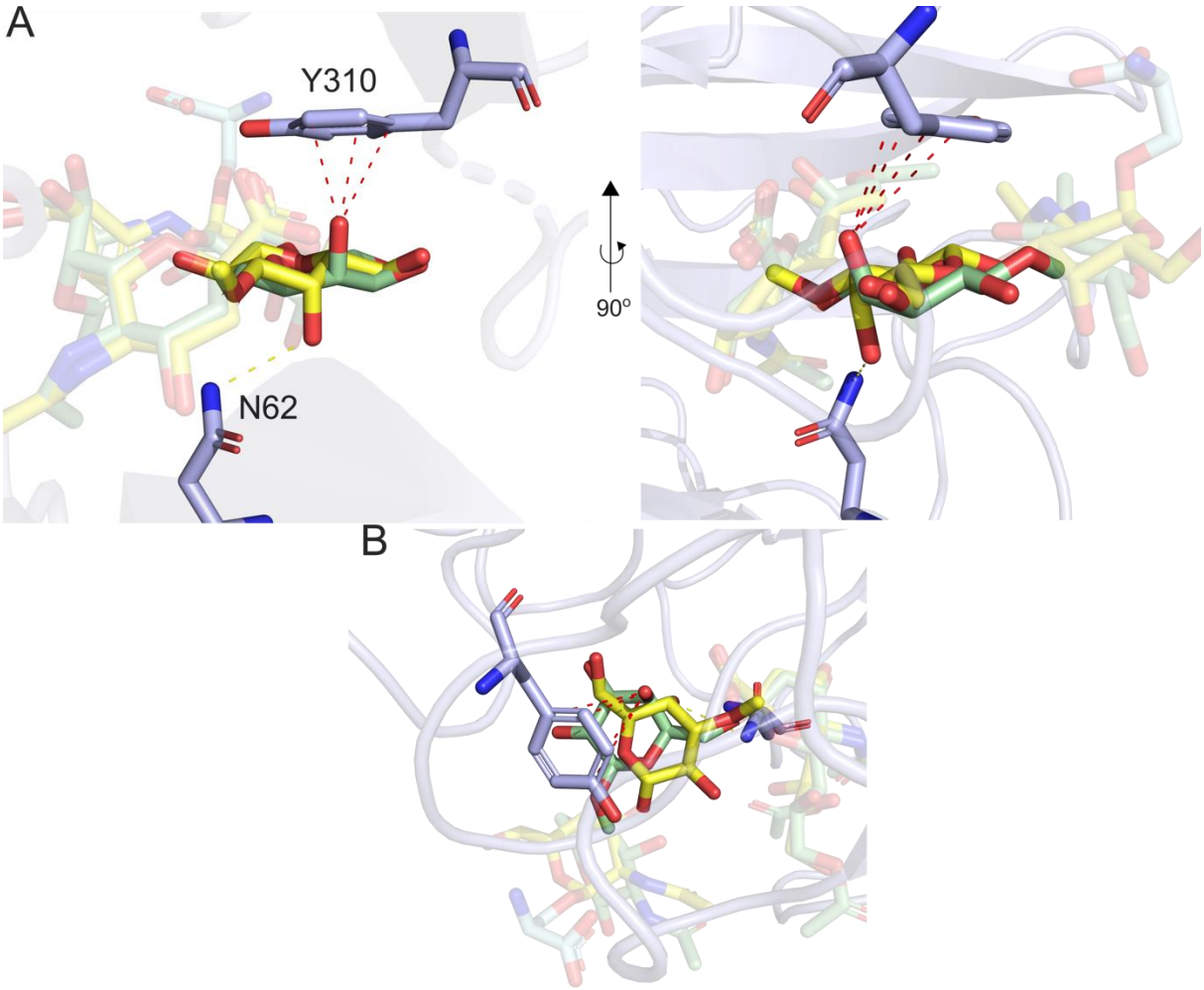


Figure 10. CH – π interactions in NanH active site. *A,* Y310 of NanH interacting with the galactose of both Neu5,7,9Ac-α(2,6)LAcNac (mint) and sialylated α(2,3)T-antigen (yellow). Red lines indicate the proximity of the O4 group within the Ch–π interface(<4Å). *B,* Over top view of each substrate in complex in the active site of NanH.

When the O4 of the STag galactose is positioned away from the Y310, it provides a hydrogen bond contact with the amide of Asn62 (3.2Å distance). As expected, this interaction is not noticed within the SLN structure. It is worth noting that asparagine is only present in the catalytically inactivated versions of NanH, as the aspartic acid (D62) is utilized as the general acid/base in hydrolysis. Therefore, the amide group of asparagine would be replaced by a carboxylic acid, and theoretically still provide ligand stabilization through non-covalent interactions. As a whole, the non-covalent hydrogen and aromatic interactions could provide reasoning for the α(2,3) > α(2,6) specificity of NanH.

2.3.3 Glycopeptide groove.

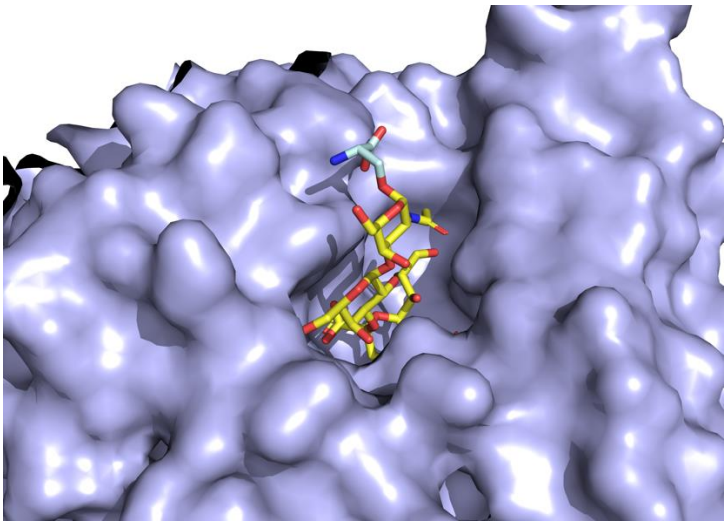


Figure 11. NanH Glycopeptide substrate accommodation. Surface rendition of NanH (D62N) in complex with STag ligand.

There have been reports on sialidases (including family GH33 sialidases) bound to di- and trisaccharides, indicating the existence of +1 and +2 subsites in sialidases. Using a crystal structure, Shuoker *et al.* demonstrate the catalytic site of the sialidase *Am*GH181 from *A. muciniphila* bound to a disaccharide representing T-antigen disaccharide (Galactose-GlcNAc), where the inhibitor DANA (*N*-acetyl-2,3-dehydro-2-deoxyneuraminic acid) was in the -1 active site of the sialidase[125]. Unfortunately, the published structure of this *Am*GH181 complex did not indicate the presence of a peptide that would be associated with a T-antigen within the active site. Here, is the first structural representation of a GH33 sialidase binding to a glycopeptide. The surface rendition shows that the GalNAc is not directly positioned within the active site of NanH, where it is partially exposed to the outside surface. The STAg complex is positioned in such a way that the serine residue is sitting within a surface groove that looks as though it could accommodate a longer peptide chain where the serine from the Tn-antigen is located.

2.3 Discussion.

Mucin glycoprotein that protects mucosal epithelia in many mammalian organs often contains a terminal negatively charged sialic acid, utilized in host defence. The connection between the underlying carbohydrate chains and the terminal sialic acids in mucin glycoproteins is typically either an α(2,3) or an α(2,6) linkage[156]. Some species, including pathogens, have evolved to not only remove the sialic acids serving as a glycan cap to access the underlying glycan layer but also utilize the released sialic acid as a nutrient source. One species known for its glycan degradation strategies is *Clostridium perfringens which* encodes three sialidases known as NanJ, NanI and NanH. Intriguingly, both NanJ and NanI share roughly 60% sequence identity and contain signal peptides to get secreted out of the bacterium. However, NanH lacks the signal peptide, indicating it is likely a sialidase that works intracellularly. It is interesting to imagine why one species might contain three enzymes that seemingly complete an identical function; however, *Clostridium perfringens* is not the only bacteria to display this behaviour. The non-pathogenic bacterium *Steptomyces avermitilis* also encodes three sialidases: NeuA1, NeuA2, and NeuA3[160]. Likewise, to *C. perfringens*, NeuA1 and NeuA2 are likely extracellular functioning sialidases, where NeuA3 exists intracellularly, with a 22.1% sequence identity with NanH[160].

2.3.1 Specificity for α(2,3) >α(2,6) glycans.

An assumption could be made that NanH is responsible for deconstructing sialyl glycans that have been imported intracellularly that have not been broken down by NanJ or NanI. As it was previously mentioned that the glycosidic linkage that attaches a sialic acid to a glycan chain affects the sensitivity of *C. perfringens* sialidase, its preference for α(2,8) > α(2,3) > α(2,6) could also provide reason for its conservation through evolution. Unfortunately, due to a lack of resources, a complex with a sialoglycan that contains an α(2,8) linkage was not obtained, nor is there a published structure on a GH33 sialidase that is complexed with an α(2,8) that NanH could be compared to. Therefore, the enzymatic substrate specificities of the NanH sialidase will

have to be drawn from the structural interactions with α(2,6) and α(2,3) linked sialo-glycans.

Comparing the NanH complexes of two mucin-based glycosylated substrates, STag containing an α(2,3) linkage to the SLN substrate shows a staggering effect on galactose in the -1 subsite due to carbon 2 of Neu5Ac linking to carbon 6 of galactose, comparatively to carbon 3 that is within the cyclic sugar ring (figure 10b)[161,162]. This distances the CH–π interaction between the α(2,3) ligand and Y310, theoretically decreasing the affinity between the binding site and ligand and binding pocket. However, the literature suggests that NanH preferentially interacts with ligands with α(2,3) linkages between Neu5Ac and the underlying glycan chain. This is likely due to the α(2,3) linkage orientating the O4 on the galactose away from the Y310 CH – π interface and directing the C-H protons from the carbon ring towards the aromatic electron pi cloud. Further, it introduces an interaction between the carbon four hydroxyl group and the functional group with asparagine 62. In the wild type NanH, aspartic acid is in the place of asparagine 62. This switch of functional groups from an amide to a carboxylic acid could possibly have a stronger interaction due to the replacement of the nitrogen atom for more electronegative oxygen.

2.3.2 Structural stabilization in the -1 and +1 subsites.

Within the -1 subsite, residues R37, P60 and L343 are seen to coordinate a conserved water molecule that provides stability by interacting with the O4 in the galactose present in α(2,3) linked ligands (figure 9). Similarly, the Y310 residue that provides CH – π stability also is in proximity to a conserved water group interacting with the galactose O2 molecule pointing in the opposite direction. Since the O4 in the α(2,6) linked glycan is directly opposite of these residues, this water interaction is not seen and strengthens the structural explanation of the α(2,3) binding preference of NanH.

All three ligands that are complexed with NanH (D62N) in this work contain a GalNAc sugar on the non-reducing end of the oligosaccharide. Interestingly, there are no residues within the binding pocket that provide interactions within 3.2Å for hydrogen

bonding, therefore, stability for the GalNAc is likely due to its proximity just within the surface of the binding pocket. The failed efforts to obtain a complex with larger polysaccharides suggest that without interaction with the active site of NanH, anything longer than a trisaccharide is too unstable and disassociates from the glycan chain. However, a NanH complex with α(2,3) STAg introduces a serine residue linked to the GlcNAc sugar that provides a carboxyl group to interact with Y248 not seen in the SLN complex (figure 9). It is puzzling to see this interaction since NanH is an intracellular sialidase that would likely not encounter glycopeptides, therefore, the proximity between Y248 and the STag serine could be purely coincidental and not key for binding glycopeptides. Perhaps this potential peptide groove could be a conserved trait of the bacterial GH33 sialidase that is usually secreted into the environment. *Clostridium perfringens* can utilize the extracellular NanJ and NanI to free sialic acid from the mucosal environment and use a permease (NanT) to bring sialic acid across the membrane. Further,as previously suggested, importing short sialo-oligosaccharides across the membrane using a putative ABC transporter is plausible[128,163]. Therefore, NanH is presumed to release the sialic acid from the sialo-oligosaccharides that have been imported and have not been cleaved by NanJ or NanI. This hypothesis is supported by the shallow pocket active site design of NanH, able to envelop short oligosaccharides but not the long glycan chains found in the mucosal chains. Due to the sialic acids more commonly linked through α(2,3) over α(2,8) glycosidic linkages, it is probable that NanH is more useful in degrading the abundant α(2,3) linked sialo glycans[164]. Although as the literature states that NanH had a α(2,8) > α(2,3) > α(2,6) linkage preference that is not seen in NanJ or NanI, it could be suggested that NanH exists to cleave the sialylglycans containing the α(2,8) linkage that were not cleaved by NanJ or NanI and were imported intracellularly[157].

Perhaps the reason that NanH was shown to successfully show binding with a glycopeptide is due to the evolution of the GH33 catalytic structure NanH possesses, displaying a strong similarity to the extracellular sialidases that are present in the environment of *Clostridium perfringens.* What it does emphasize is that these sialidases could target glycopeptides, highlighting their utilization in mucin degradation.

Chapter 3: The glycopeptidase AMUC_1438 from *Akkermansia muciniphila*

Notice of collaboration:

This chapter was a part of a collaboration: the glycopeptides used in the study was created and sent from the Wakarchuk group at the University of Alberta, and the mass spectrometry work was done and modelled by the Malaker group from Yale University.

Adapted from the publication:

A previously uncharacterized *O*-glycopeptidase from *Akkermansia muciniphila* requires the Tn-antigen for cleavage of the peptide bond. *J. Biol. Chem.* 2023 Aug 30;298(10).

3.1 Introduction.

The mucins that build the protective mucosal layer are rich in carbohydrates, providing a unique bacterial proliferation environment. As most of the glycoproteins that make up the mucosal layer are extended chains of *O*-linked glycans, CAZymes are responsible for a large part of the deconstruction of mucosal glycoproteins. However, in terms of the gastrointestinal mucosal layer, there is still a >5000 amino acid backbone that requires to be degraded to complete the break down of the mucin outer layer. Host-adapted bacteria have evolved to utilize *O*-glycopeptidases that recognize and cleave the peptide bond of protein substrates that bear glycan decorations, establishing their essential niche in mucin degradation. To gain a competitive edge within the gastrointestinal environment, both commensal and pathogenic species have adapted. *Clostridium perfringens* is well equipped with the *O*-glycopeptidases that belong to peptidase family M60 named ZmpA, ZmpB and ZmpC, which compete for environmental glycopeptide with other extracellular glycopeptidases, such as the commensal M60 glycopeptidase BT4244 from *Bacteroides thetaiotaomicron*[165].

Akkermansia muciniphila is an interesting case where four enzymes have glycopeptidase activity reported (three belonging to family M60: Amuc_0627, Amuc_0908 and Amuc_1514 and OgpA (Amuc_1119) belonging to an unreported

family)[149,166]. Ottman *et al. have* described a transcriptome analysis of *Akkermansia muciniphila* growing in cultures using a mucin carbon source in comparison to a glucose carbon source and reported 1074 genes were differentially expressed when growing on the mucin based media (657 of them being more than two-fold[87]). Amongst these differentially regulated genes are directly correlated to mucin degradation, i.e., several glycosyl hydrolases. *amuc_1438* is a gene flagged as a "putative glycosyl hydrolase family 98 protein", with a 3.585-fold increase in expression when utilizing a mucin-based media. However, a bioinformatic analysis using InterPro for domain prediction indicates that the *amuc_1438* gene encodes a multi-modular enzyme that contains a metallopeptidase catalytic domain, signified by the HEXXH zinc binding motif[141,167]. The bioinformatics also indicates the C-terminal domain of this gene encodes a carbohydrate-binding module from family 51 (figure 12). A sequence alignment indicates little identity with the four known glycopeptidases from *Akkermansia muciniphila* (~27% supplemental figure 1), however, Amuc_1438 and OgpA shared significant sequence similarities around the catalytic HEXXH binding domain (Figure 13c).

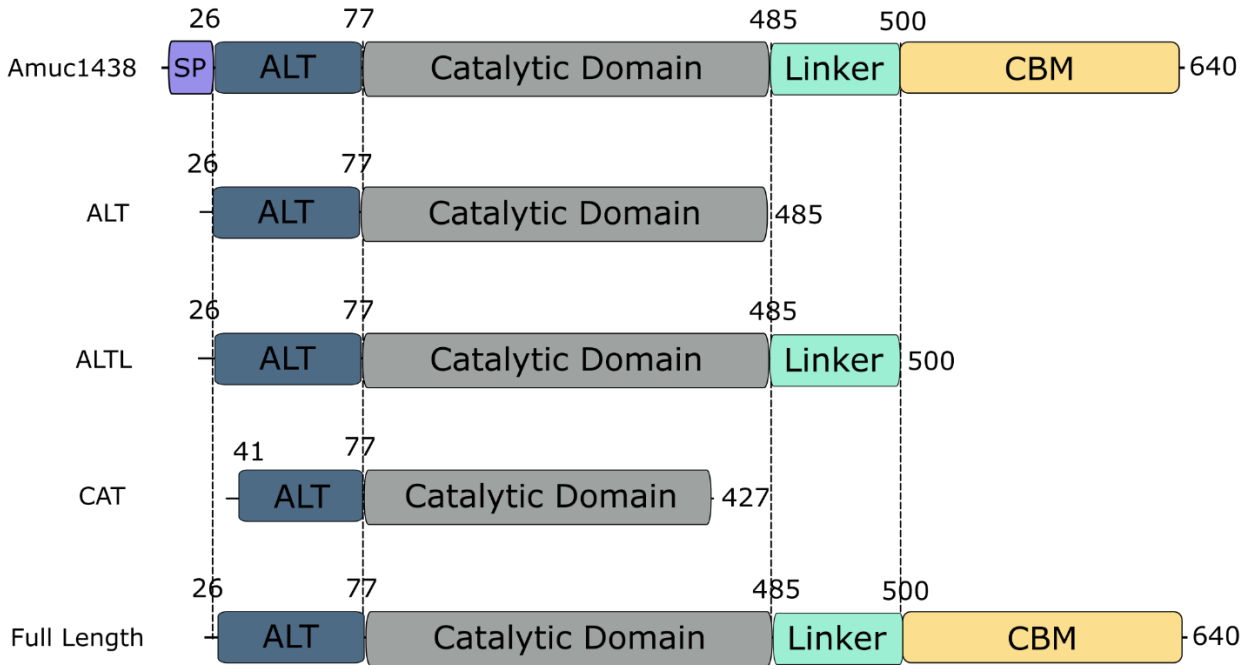


Figure 12. Bioinformatic analysis of the multi-modular amuc_1438 from *A. muciniphila*. Depiction of the genetic constructs used in this study. Numbers above the segments indicate the translated amino acid borders of the predicted domains.

Together, this outlines the hypothesis of this chapter: ***The amuc_1438* gene encodes a peptidase that has specificity for glycosylated peptides.** This chapter takes a functional and structural approach to characterize Amuc_1438 as a glycopeptidase utilized in the enzymatic mucin degradation pathway.

3.2 Results.

3.2.1 Structural analysis of Amuc_1438.

The bioinformatic analysis predicted five separate domains, yet which of these domains was required for the proper assembly of a functioning enzyme to be deciphered. Multiple constructs of *amuc_1438* were designed with varying boundaries of the gene to elucidate the full catalytic machinery of the enzyme (figure 12). The ALT construct was successfully crystallized and solved to a resolution of 2.35 Å and the ALTL (containing the linker region) was solved at 2.50 Å. Within the ALT crystal structure, the N and C terminus of the Amuc_1438 enzyme were disassociated (by 15 and ~60 residues respectively) leading to a proposed idea that the actual boundary for the properly folded catalytic machinery is between residues 41 and 427. The CAT construct was created and utilized based on the boundaries found in the crystallography experiments.

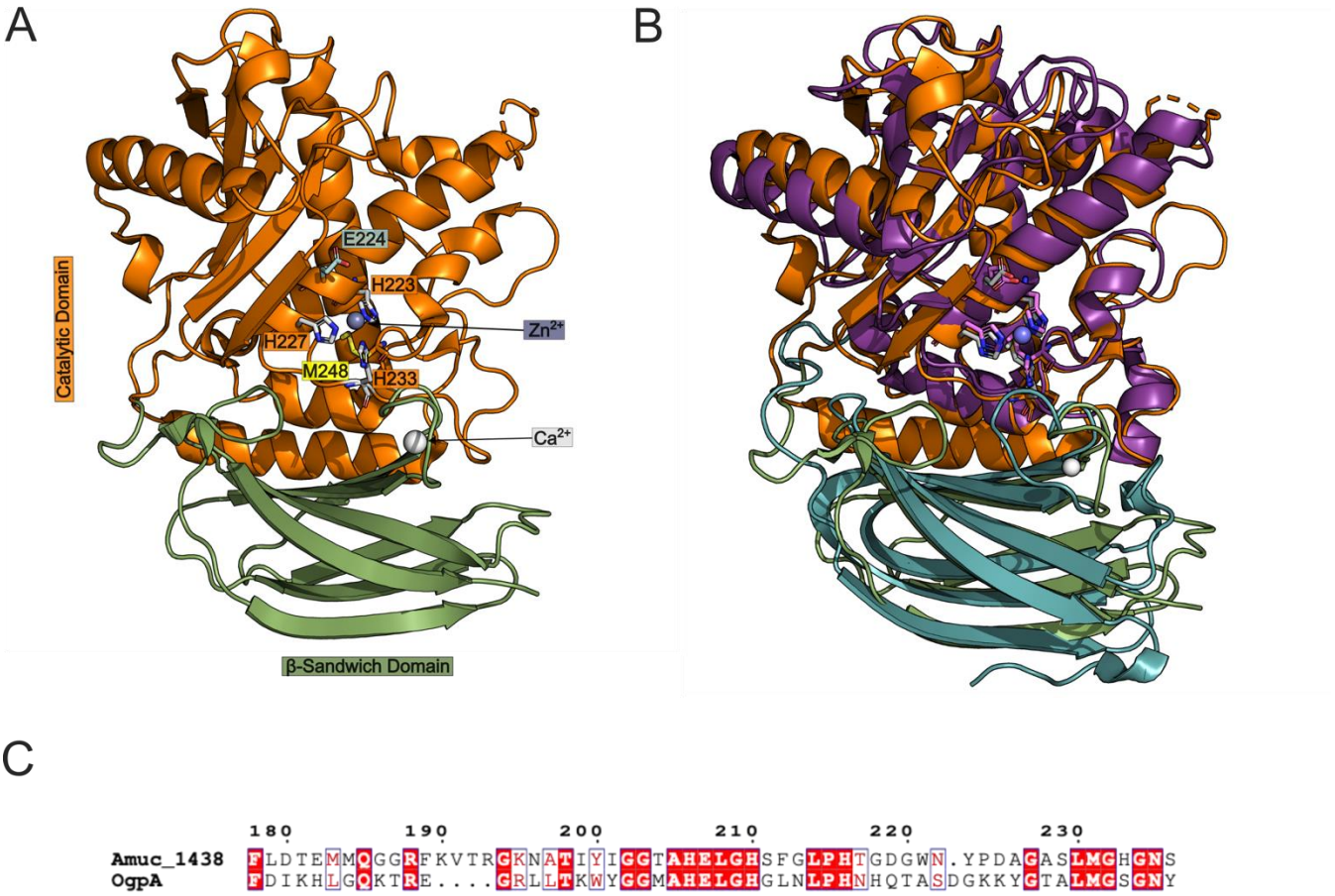


Figure 13. Structure of the catalytic domain of Amuc_1438. *A.* crystal structure of the 'ALT' construct of Amuc_1438. Catalytic machinery is outlined in the center of the enzyme, along with two spheres representing a zinc ion and a calcium ion. *B.* Overlay of Amuc_1438 with the glycopeptidase OgpA from *A. muciniphila* (catalytic domain is in purple, β-sandwich domain in teal). The catalytic machinery of OgpA is highlighted in light purple and overlayed with Amuc_1438 machinery in light grey. *C.* Sequence alignment of the catalytic region of Amuc_1438 and OgpA. Alignment was created using ClustalW and visualized using ESPript 3.0. The numbering above refers to the position in the alignment.

Solving the structure of the ALT construct revealed both a catalytic domain and a β-sandwich domain that is adjacent to the catalytic machinery. As expected for an enzyme containing a metallopeptidase catalytic domain, this crystal structure was complexed with a Zn^{2+} ion that was coordinated by H223, H227 and H233, a part of the HELGH motif (figure 13a). The catalytic glutamate E224 and the methionine M248 are represented in the active site as well, M248 attached to the Met-turn that gives structure to the active site. Even though the sequence alignment with the OgpA glycopeptidase

was limited, the structural overlay of both enzymes indicates that these are very similarly built enzymes on a structural level (figure 13b). An alignment of the sequences containing both enzyme's catalytic machinery displays how conserved the active site is between both genes (figure 13c). As the structure of an enzyme often leads to function, this structural analysis is very suggestive that Amuc_1438 perhaps functions in the same fashion as the glycopeptidase OgpA.

3.2.2 Activity Assays on glycosylated peptides.

Section 1.7.1 indicated that the reported glycopeptidases from *Akkermansia muciniphila* have specificity for the types of glycosylation that are decorated on the glycopeptides they cleave. The specificity of Amuc_1438 was tested against the peptide sequence ‘GPAPGSTAPPAE’, a repeating sequence in the gastrointestinal glycoprotein MUC2, bearing various glycan decorations on the threonine C-terminal to the serine (Figure 14). The CAT construct of Amuc_1438 did not recognize a ‘naked’ peptide (bearing no glycans), indicating that this enzyme was not active on bare peptide chains. Following this, the enzyme was then tested against a peptide containing a single Tn-antigen, a core 1, core 2, and a sialylated Sia-2,3 Core 1. The known glycoprotease IMPa from *Pseudomonas aeruginosa* was used simultaneously as a positive control[145]. A shift in the BODIPY labelled glycopeptide band was seen when CAT was incubated with Tn-antigen bearing peptide, in comparison to the negative control and assured by the IMPa positive control, Indicating cleavage. However, for the Core 1, Core 2 and Sia-2,3 Core 1 experiments, there was no evidence of cleavage to be seen, indicating that the CAT enzyme had a specificity to this peptide bearing only the Tn-antigen. An interesting, and correlated result to what we have seen with the other *A. muciniphila* glycoproteases.

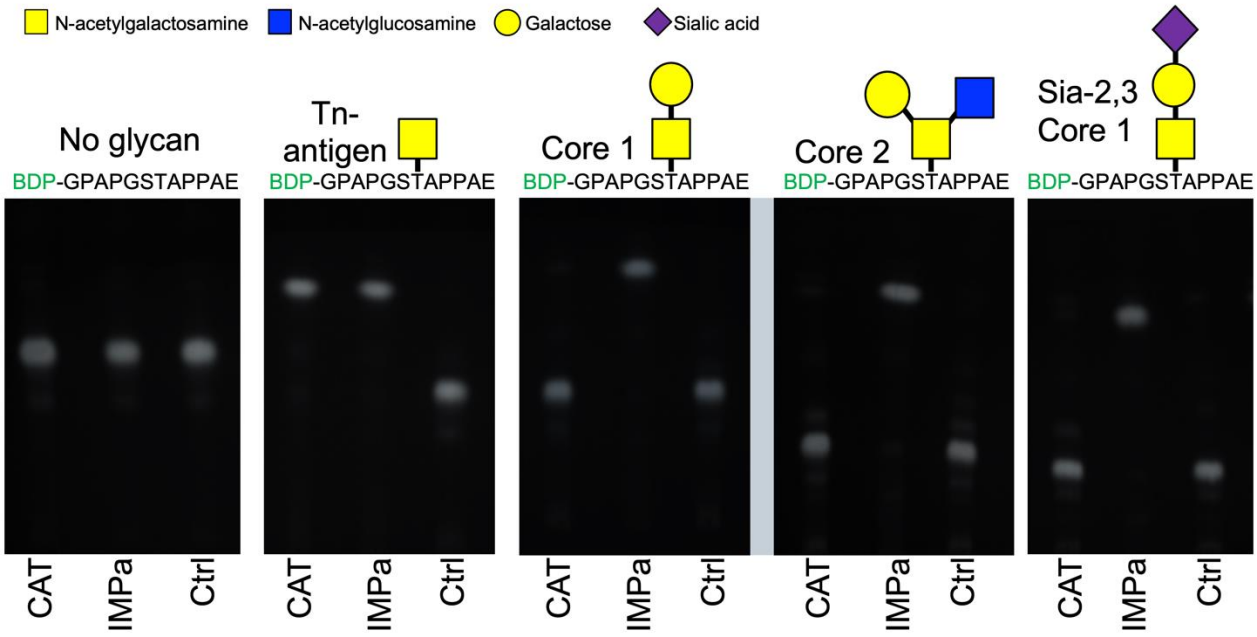


Figure 14. Thin layer chromatography panels of Amuc_1438 glycopeptide digests. Each panel represents a separate experiment done using the indicated enzyme (written below panel) with the substrate above. Positive control was the known glycopeptidase IMPa[145], negative control was substrate without the presence of enzyme.

Glycopeptidases can be quite selective of what type of substrate they are sensitive to, as observed with Amuc_1438. The glycosylation patterns of the substrate are key for targeting and cleavage of the peptide, however, in some cases the peptide sequence also affects targeting. For instance, the secreted protease of C1 esterase inhibitor (StcE) is a secreted glycoprotease from *Escherichia coli* that requires the peptide sequence S/T*(X)S/T, where X can be any amino acid, the S/T* requires glycosylation, and StcE cleaves before the S/T after the (X) residue[168]. Therefore, to fully characterize Amuc_1438, it is deemed necessary to explore amino acid sensitivity.

3.2.3 Sequence specificity.

To understand whether Amuc_1438 targets and cleaves based on peptide sequence, the CAT construct was tested against several glycopeptides from proteins that are commonly found throughout the body including the gastrointestinal, respiratory, and urogenital tracts[169–172]. The peptides were decorated with a single Tn-antigen except for

MUC1(diTn) was the exception where it contained an identical peptide sequence to MUC1 and a second glycosylation site. The CAT construct was able to degrade all the tested glycopeptides except for the substrate from the fetuin source (figure 15). Since the only variation in this experiment is the amino acid sequence, this indicates that there are some specifications required for Amuc_1438 to target a glycopeptide based on residue sequence. The question is, what peptide sequence does Amuc_1438 deem acceptable to cleave?

Peptide source	Peptide sequence	Activity
MUC1	PAPGSTAPPAHGVTSAPDTRPAPG	✓
MUC1(diTn)	PAPGSTAPPAHGVTSAPDTRPAPG	✓
IgA1	VPSTPPTPSPSTPPTPSPSC	✓
Fetuin	GAEAEAPSAVPDAAG	✗
MUC5AC	GTTPSPVPTTSTTSAP	✓

Figure 15. Representation of Tn-decorated glycopeptide degradation. The glycopeptides were sourced from the identity in the first column, and the sequence isolated is represented in the center with denoted with glycan positioning. If cleavage was observed by a shift in the banding pattern on a TLC plate (supplemental figure 2), this would indicate activity.

To further investigate the glycopeptide specificity, reactions containing CAT construct and mucin-like glycoproteins containing mapped glycosylation events were incubated and analyzed using mass spectrometry (figure 16)[147,168]. The observed cleavage sites were directly N-terminal to a glycosylated serine/threonine residue containing a Tn-antigen, supporting the previous data presented.

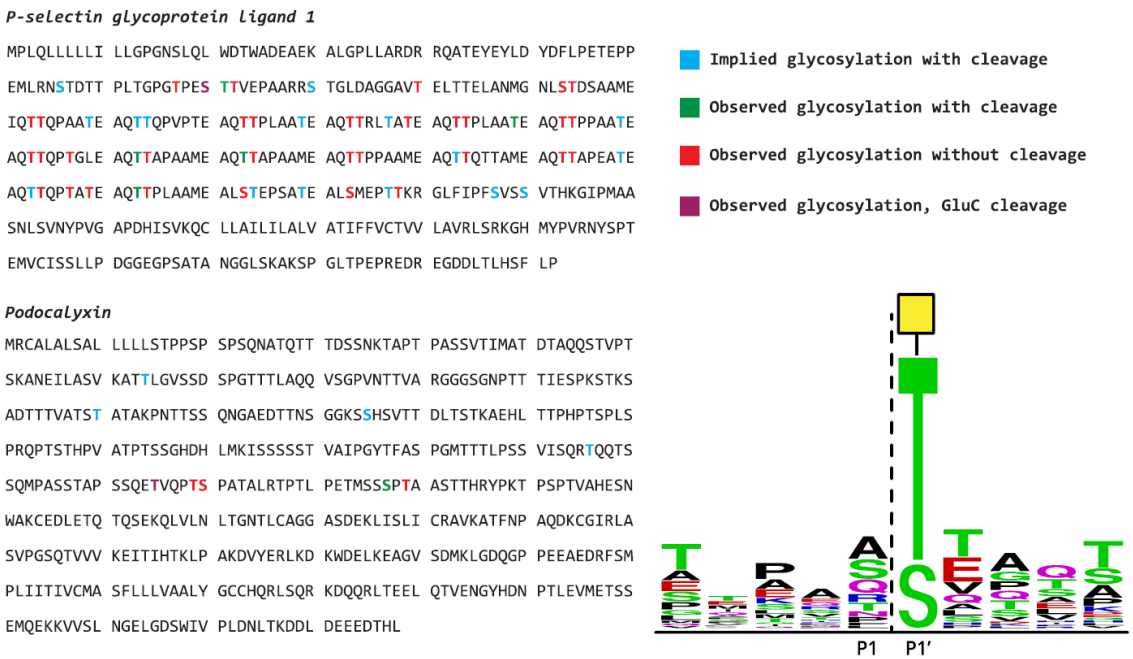


Figure 16. Glycoprotein breakdown by Amuc_1438 and mapped through mass spectrometry. The protein sequence is outlined with specific residues highlighted based on glycosylation states, or cleavage information. The bottom right is a logo plot representing the experiment's cumulative data. The dashed line N-terminal indicates the cleavage site to the glycosylation site.

Many of the known glycosylation sites with observed glycosylation of the P-selectin glycoprotein ligand 1 and podocalyxin were not cleaved by CAT. Likely, this could be the result of many of the glycosylation sites containing more decorated glycans than just the Tn-antigen. There were also twenty unexpected S/T cleavage sites between both glycoproteins, therefore implying that these are sites of glycosylation. Overall, observing the CAT construct only cleaving at Tn-antigen-bearing peptides in this experiment solidified the requirement for this decoration, yet no reasonable pattern was determined for a required peptide sequence.

3.2.4 Amuc_1438 contains a catalytic G1 glycan binding site.

As the structure of an enzyme is commonly associated with the function of the enzyme, the previous data suggests the catalytic machinery of Amuc_1438 has a G1 binding site to interact with the glycopeptide. To map this hypothesized site, the most beneficial way would be through gaining a crystal complex with a Tn-containing glycopeptide. Unfortunately, attempts to obtain a crystal complex with a glycopeptide were

unsuccessful. However, a structural alignment of Amuc_1438 with OgpA complexed with glycodrosocin allows for an accurate representation of where the G1 glycan binding pocket likely exists[148].

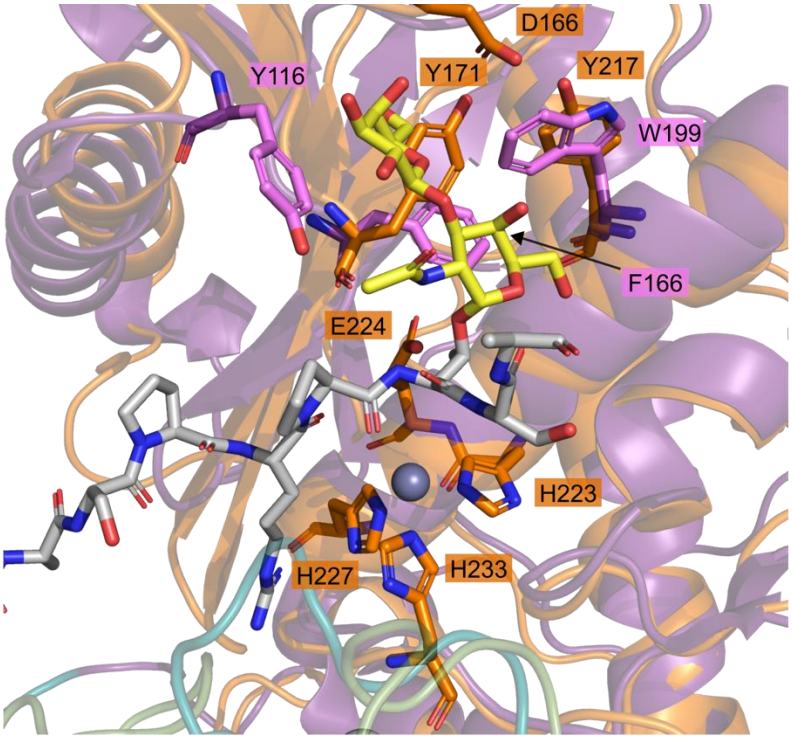


Figure 17. Amuc_1438 binding sites. Cartoon representation of the Amuc_1438 structure overlayed with OgpA that is in complex with glycodrosocin (grey peptide chain with yellow glycan disaccharide). Amuc_1438 is represented in orange (both cartoon and key residues) while OgpA is in purple. Protein Data Bank ID for OgpA complex: 6Z2P).

These two enzymes differ in their active sites around the construction of their respective glycan binding sites. OgpA utilizes Y116, W199 and F166 to interact with glycans, creating an ‘open’ conformation and allowing larger glycans to fit within the glycan binding sites. This alignment outlined three aromatic residues that correspond with the G1 pocket of Amuc_1438. The positions of the Amuc_1438 residues Y171, Y217 and D166 are within proximity with the GAlNac sugar attached to the peptide chain, suggesting their responsibility in creating a sugar-binding pocket (figure 17). It also shows the steric clash that would occur between the galactose and Y171, limiting the glycan binding to a single Tn-antigen. Further, D166 is seen to be at the ‘top’ for the G1

binding site, possibly contributing to glycan stabilization through hydrogen bonding. This data corresponds well with the enzymatic analysis of the Tn-antigen selectivity.

3.2.5 Point mutations in the active site.

To confirm the hypothesis of what catalytic and sugar-binding residues Amuc_1438 uses for substrate recognition, single point mutations were used on each residue to build a small library of Amuc_1438 knockouts. Alanine was used as the mutant amino acid to remove all potential chances of a conserved function of the mutants. A fluorescent resonant energy transfer (FRET) based mechanism was designed using the glycosylated IgA1 peptide sequence and attaching a fluorophore N-terminally as well as a C-terminal dark quencher QXL520 to test for enzymatic cleavage[173]. This FRET IgA1 glycopeptide was used to test the point mutations, as well as a kinetic analysis of both the CAT and full-length constructs.

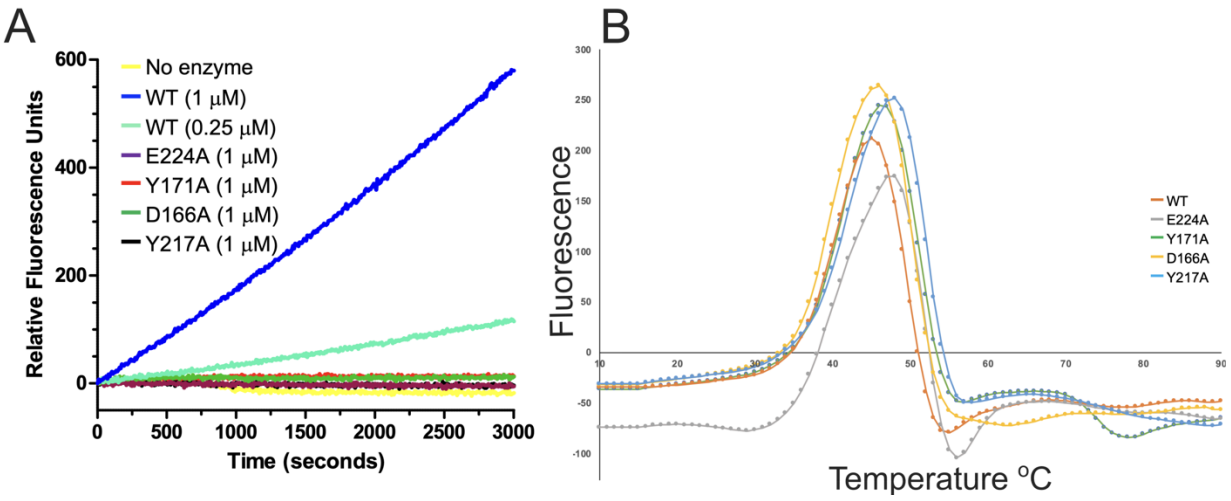


Figure 18. Kinetic analysis of Amuc_1438 mutants using a FRET-based glycopeptide. *A,* Rates of FRET cleavage by wildtype CAT enzyme (in both 0.25 and 1.0μM quantities, compared to catalytic and sugar pocket mutants of CAT. *B,* Differential scanning fluorimetry graph depicting mutant protein stability compared to wildtype based on temperature.

The wildtype CAT construct was able to degrade the FRET IgA1 glycopeptide in both 0.25μM and 1.00μM concentrations, as indicated by the increase in relative

fluorescence units over time (figure 18a). The knockout of the catalytic glutamate E224A completely eliminates any sign of activity, with no seen increase in fluorescence in comparison to the no enzyme negative control. It is intriguing that all three residue knockout strains selected based on the structural identity between OgpA seen in Figure 17 also did not cleave the IgA1 glycopeptide. This result suggests that each of the selected amino acids is utilized in forming a sugar-binding pocket, and the removal of even one is enough to destabilize the interaction between enzyme and substrate.

The abolition of released fluorescence by the mutant experiments creates the worry that perhaps the loss of activity is due to the mutation causing the protein to misfold. Using differential scanning fluorimetry, the folding of the mutant protein could be compared to the wild-type enzyme as temperature increases. As the temperature increased, the wildtype Amuc_1438 started to degrade around ~32^{o}C as indicated by the increase of fluorescence and had a maximum fluorescence reading at ~44^{o}C. To support the results in Figure 18a, the mutant enzymes Y171A, D166A, and Y217A all shared very similar melting plots to the wildtype enzyme, suggesting that the mutants are properly folded regarding the wildtype CAT.

3.2.6 Michaelis-Menten kinetics.

The FRET IgA1 glycopeptide that was utilized in a qualitative manner to test mutant knockouts can also be used quantitatively in kinetics. Both the CAT construct as well as the full-length enzyme were tested (figure 19 a/b). CAT had indicated a K_M of 300 (±70) µM and a k_{cat} of 1.7 (±0.2) min^{-1}. The full-length enzyme had a K_M of 122 (±30) µM and a k_{cat} of 1.4 (±0.1) min^{-1}.

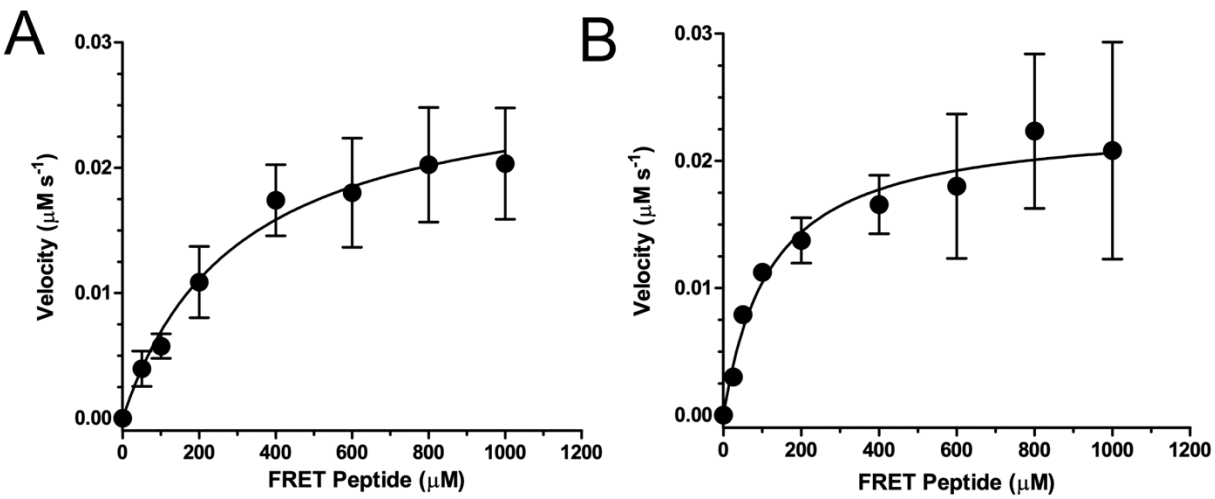


Figure 19. Michaelis-Menten kinetics. *A,* Michaelis-Menten plot of the CAT construct using the FRET peptide. *B,* Michaelis-Menten plots of the full-length construct. Both plots are based on six replicates using the FRET IgA1 substrate.

It is worth noting that much like other M60-like glycopeptidases, Amuc_1438 contains a C-terminal carbohydrate-binding module[174]. Unfortunately, this domain was not crystalized nor tested for sugar binding affinity. The Amuc_1438 CBM51 structure that was predicted using Alphafold2 displays a strong correlation with the CBM51 domain from a GH95 in *Clostridium perfringens*[175]. This CBM51 from *C. perfringens* (*cp*CBM51) is in complex with a galactose monosaccharide, which shares a significant structural comparison with Amuc_1438 CBM51 (*am*CBM51), suggesting that both modules might share similar glycan binding functions (figure 20a).

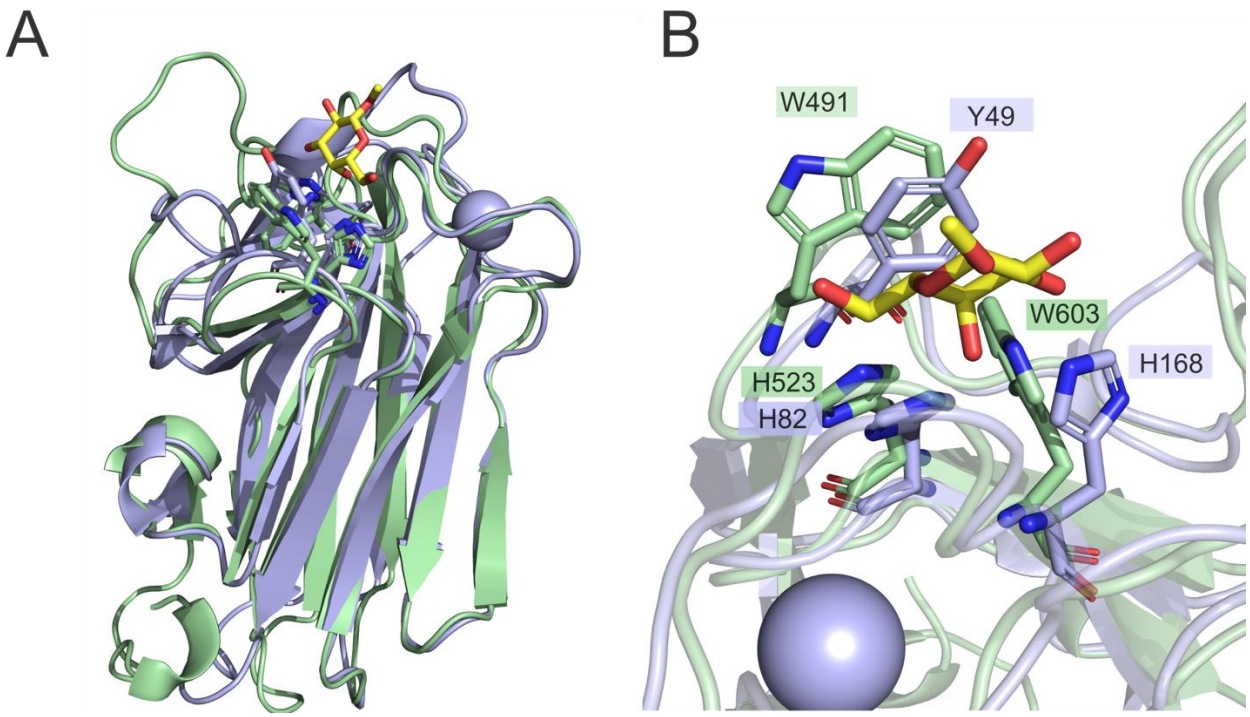


Figure 20. Amuc_1438 CBM51. *A.* structural alignment of *am*CBM51 (green) and *cp*CBM51 (blue) in complex with methyl galactose (yellow) and a Ca^{2+} ion (PDB: 2VMG). *B.* Close-up of the galactose binding site in the *cp*GH95 CBM51, with binding residues indicated as sticks. Presumptive sugar-binding residues from the *am*CBM51 protein are outlined as sticks as well.

Taking a closer look at the binding site overlay indicates that much of the glycan binding structure is conserved in *am*CBM51 (figure 20b). One of the histidines of *cp*CBM51 has been replaced with tryptophan in position 603, although its orientation allows its nitrogen to be in proximity to hydrogen bond with an incoming glycan. Alternatively, the other histidine in the binding site is directly conserved in *am*CBM51. The main interacting tyrosine is also modified to a tryptophan, yet the orientation conserves the planar aromatic binding interface that would be used in glycan interaction.

Without structural studies or functional characterization, whether this moiety from Amuc_1438 would bind glycans is unknown. Yet, this alignment suggests that glycan binding is plausible and would likely have a synergistic effect with the functionality of the peptidase catalytic domain. Like many other enzymes that have an affinity for glycans, the amCBM51 is very likely used to adhere to a polysaccharide and bring the catalytic

domain in more intimate proximity with the glycopeptide target and allowing for faster hydrolysis[176].

3.4 Discussion

Analyzing the gene sequence of Amuc_1438 allowed for the hypothesis that it encodes a multi-modular enzyme that carries the metallopeptidase catalytic motif HEXXH and carbohydrate-binding motifs, with specificity for glycopeptides. Interestingly, there was little sequence similarity between the other characterized glycopeptidases from *Akkermansia muciniphila,* all showing around 30% identity. Solving the structure of CAT indicated the catalytic machinery seen in metallopeptidase enzymes. The structural alignment with OgpA indicated a strong resemblance between the two enzymes (RMSD of 2.1Å). One key feature seen in the structure of Amuc_1438 is the ~30Å long helix that contains H227, H223 and E224 which belong to the HEXXH zinc binding and catalytic domain, sharing a very similar length and orientation to the helix seen in OgpA. Even though the overall sequence identity isn't very high between the enzymes, the percentage increases to around 40% over 47 residues in the active site, including the catalytic helix. The physical similarity in the 'catalytic helix' is not seen throughout all the M60-like enzymes, however. According to Trastoy *et. al,* the peptidases BT4244, ZmpB, StcE and IMPa are between ~30% - 50% shorter in length[148]. Not all glycopeptidases from *A. Muciniphila* exhibit a long catalytic helix; in. fact, the helix of Amuc_0627 is approximately 29% short in comparison to Amuc_1438 and OgpA[177]. Possibly, the variation in helix length may impact glycopeptide specificity since this helix houses the catalytic machinery for these enzymes. Overall, the lack of sequence similarities between any known MEROPS family and the significant structural alignments with OgpA allows for the assumption that Amuc_1438 and OgpA are founding members of a novel family of glycopeptidase.

The CAT enzyme requires a Tn-antigen to cleave the peptide bond of a glycosylated peptide. Similarly, to the enzymes BT4244, ZmpB, StcE, ImpA that belong to the M60-like superfamily, CAT cleaves *N*-terminally to the site of glycosylation. What is unique to

Amuc_1438 in comparison to the four other glycopeptidases from *A. muciniphila* is its specificity for nothing more decorated than a Tn-antigen. In comparison Amuc_0627, Amuc_0908, Amuc_1514, and OgpA are capable of cleaving glycopeptides bearing both Tn and core 1 decoration, and more so in the case of all but Amuc_1514, somewhat sensitive to sialylated glycans[147,148]. Amuc_1438 exhibits a narrower specificity compared to these four enzymes. It is possible that Amuc_1438 would be employed specifically in situations where there is a higher abundance of Tn-antigen during mucin degradation. However, it is important to highlight that one study indicated when grown on mucin as opposed to glucose Amuc_1438 displayed a 3.6-fold increase in expression, whereas Amuc_0908, Amuc_1514 and OgpA all indicate lower expression levels (2.63, 1.36 and 1.05, respectively)[87]. This suggests that Amuc_1438 may be employed more frequently in mucin breakdown than the other enzymes, but significantly less frequently than Amuc_0627, demonstrating an 18.8-fold increase. The limitation of binding Tn-decorated glycopeptides is seen in the G1 binding site. The superimposition of glycodrosocin from the OgpA shows the glycan specificity of Amuc_1438 is limited due to the amino acid construction of the G1 binding site. Whereas Y171 is clashing with the galactose, providing to be a limiting factor to bind the disaccharide from glycodrosocin. However, as this is just a superimposition, this is only theoretical until a crystal structure of Amuc_1438 is in complex with a glycopeptide is present to better understand the glycopeptide binding nature of Amuc_1438.

Except for fetuin, all the *O*-linked glycopeptides that were examined were successfully cleaved by the CAT construct. This discrepancy might be because the fetuin peptide originates from an *O*-glycosylated protein that is not derived from a mucin source[169]. However, the glycosylated peptide from the hinge region of the IgA1 antibody was cleaved. Hence, it can be concluded that specificity is not solely dependent on peptides generated from glycosylated mucin regions. To quantify the hydrolysis potential of the catalytic and full length, a custom *O*-glycopeptide was developed utilizing a FRET-based system. Both CAT and full-length enzymes were tested against the FRET system containing the sequence -TPSP**S**TPPTK- based on the IgA1 peptide sequence. The CAT enzyme had a k_{cat} of 1.7 (±0.2) min^{-1} and the full-length enzyme had a similar k_{cat}

of 1.4 (±0.1) min^{-1}, yet both are quite slow in comparison to the “average enzyme” (exhibiting a k_{cat} of ~$10s^{-1}$) [178]. Unfortunately, there could be several factors that could have affected the speed of catalysis. One idea is that the addition of the fluorescent group (Hilyte488) and the dark quencher (QXL520) interferes with the interaction between the peptide and the active site, therefore reducing the natural rate of cleavage. In contrast, this peptide was derived from the IgA1 hinge peptide segment found in the antibody, rather than a mucin glycoprotein. Since no activity was observed against fetuin, a non-mucin protein, it implies that glycopeptides not originating from mucins may have limited potential for interaction and suggests perhaps the IgA1 hinge peptide might not be the best substrate for Amuc_1438. To date, kinetic studies on M60-like glycopeptidases are quite limited. The only other glycopeptidase that has had kinetic evaluation is the M60 ZmpB from *Clostridium perfringens,* yielding a K_M of 532.5 (±81.8)μM and a k_{cat} of 0.026 (±0.003) s^{-1} which is still considerably less than the average enzyme[165]. As FRET-based kinetics were used for Amuc_1438 and ZmpB, it is difficult to say whether the low-level kinetics are based on the assay itself, or perhaps a feature of the enzyme class.

In conclusion, the comprehensive analysis of Amuc_1438 provides insights into its role as a tool used in host mucin degradation with specificity for peptides carrying a Tn-antigen on a serine or threonine. The bioinformatic and structural analysis suggests this enzyme contains glycopeptidase activity that is like the known glycopeptidase OgpA, likely belonging to an unknown family of M60-like glycopeptidases.

Chapter 4: Amuc_1438 and NanH fit into the enzymatic breakdown pathway of mucin.

4.1 Final Discussion and conclusion.

The mammalian gut is a complex system that requires microbial-host interactions to maintain host health. The intestinal epithelium is protected by the mucin-rich mucus layer that helps shape the gut microbiota by providing a habitat for residents of the gut[166]. Many species have been characterized as mucin degraders, with a spotlight on both *Akkermansia muciniphila* and species within *Bacteroides*[179]*.* As most mucin is made of complex combinations of glycan chains, the bacterial colonizers utilize the exposed *O*-glycans as adhesion points to initiate colonization[125,179]. Complexity is required within the glycan organization of the mucin glycoproteins to promote microbiota selectivity, as simplicity in the glycan chains would lead to over-harvesting and thinning of this protective layer. The complexity of glycan chains in the gastrointestinal tract is not consistent throughout. For example, mucin glycoproteins often contain sialic acids, sulphation, and fucose in various terminal positions[180]. In humans, fucosylation is most abundant in the small intestines but gradually decreases as it moves toward the distal colon. On the other hand, sialylation and sulphation are more prevalent in the distal colon and less common in the small intestine[125,181]. As a result, bacteria residing in the gastrointestinal tract need to adjust to utilize these glycans, requiring their repertoire of enzymes to adapt accordingly to their specific environment.

To flourish within the gastrointestinal tract, it is important for the microbial residents to contain proper enzymatic pathways to break down the glycoprotein mucin layer within their environment, particularly in times of low ingested dietary nutrients[88]. For a bacterial species to undertake mucin glycoprotein breakdown for energy harvesting, their enzymatic repertoire must consist of CAZymes able to initiate the breakdown of the carbohydrate chains, and glycopeptidases that target and break down the trimmed glycopeptides. In the big picture, mucin breakdown initiates with the removal of the

terminal capping structures including sialic acid (sialidases), fucose (fucosidases), sulphates (sulphatases) and blood-antigen groups on the glycan chains, before additional CAZymes can act on the underlying glycan chains (figure 21)[180]. In conjunction with the protected underlying glycan chains being targeted and degraded by a variety of CAZymes, bacterial enzymes will have access to the trimmed glycopeptide backbone, presenting a perfect substrate for the extracellular glycopeptidases. Following the liberation of monosaccharides, oligosaccharides and glycopeptides, the degradation products are found to be either imported into the bacteria or else it is likely that the products are shared within the vast microbial community of the gastrointestinal system[127]. Once imported intracellularly, a multitude of intracellular enzymes degrade the oligosaccharides and glycopeptides to utilize for nutrients[182].

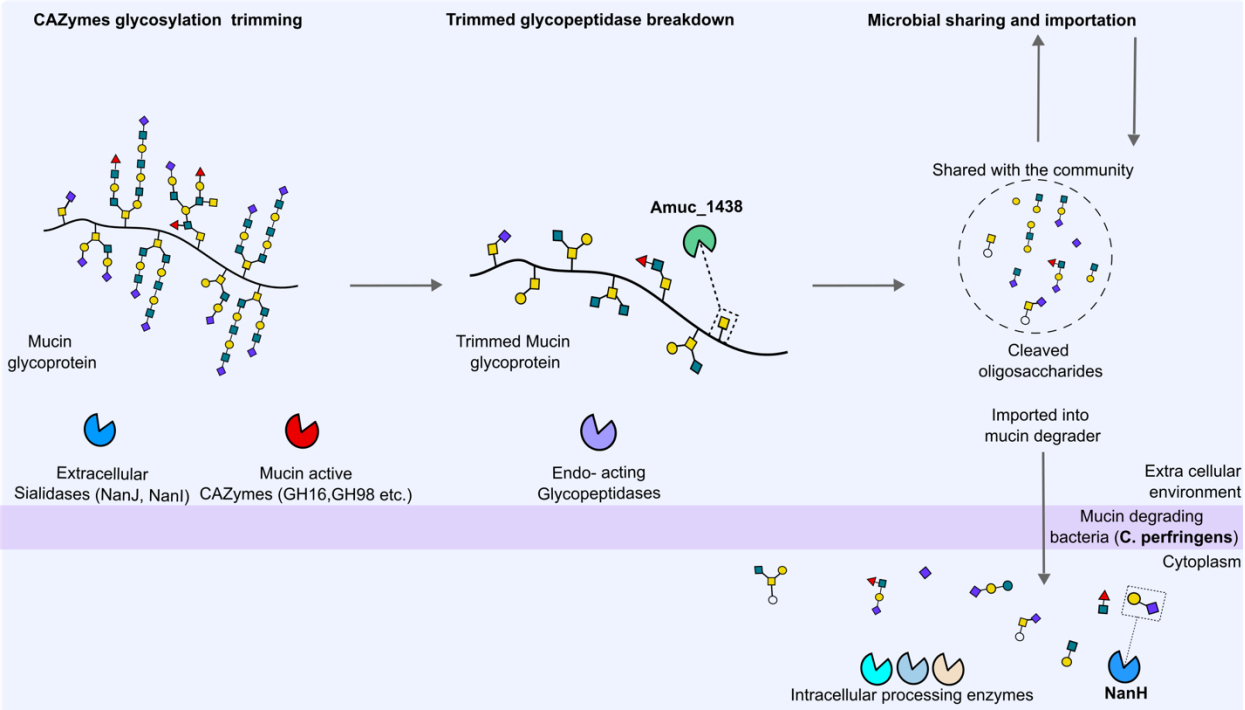


Figure 21. Mucin degrading pathways by enzymatic systems. The overall concept revolves around the breakdown of mucin glycoproteins using bacterial enzymatic systems utilized by both commensal and pathogenic inhabitants. Included is the idea that Amuc_1438 (green) assists in breaking down the trimmed glycopeptides, as well as NanH targeting the sialyl glycans into *Clostridium perfringens*. The clear circles without any colouration represent amino acids attached to the glycans.

Two of the many species that target the mucosal layer of the gastrointestinal system are the commensal species *Akkermansia muciniphila* and the pathogen *Clostridium*

perfringens[87,183]. To date, there have been 96 individuals characterized CAZymes reported to the CAZy database associated with *Akkermansia muciniphila* (ATCC BAA-835), and 110 from *Clostridium perfringens.* This thesis showcased two enzymes from either mucin-degrading tool belt that assists in the enzymatic pathway: the glycopeptidase Amuc_1438 as well as the intracellular NanH. Both species contain an extensive toolbelt that could be utilized in mucin degradation[69,87].

In both species, the importance of sialic acid cleavage (whether for initiating mucin degradation and or nutrient harvesting) is evident. *Akkermansia muciniphila* encodes two sialidases (Amuc_0625 and Amuc_1835) that are both capable of cleaving α(2,3), α(2,6) and α(2,8)[125]. Interestingly, studies showed that *A. muciniphila* sialidase inhibition dramatically decreased growth initiation on mucin, providing evidence that desialylation is crucial for the enzymatic mucin degradation pathway[125]. Within the genome of *Clostridium perfringens,* the presence of the three sialidases *nanJ*, *nanH* and *nanI* enhance pathogenic invasion of the mucosal layer. While most strains possess all three sialidases, certain strains may have one or two of them[134]. One study suggested that in strains of *C. perfringens* that encode all three sialidases, NanI is seen to account for up to 70% of the extracellular sialidase activity[77]. Therefore, in the enzymatic mucin degradation pathway, NanI and NanJ are utilized by *C. perfringens* for exposing the underlying glycan layers by removing the extracellular sialic acid to be imported into the cell[77].

Similarly, to the mucin degrading bacteria of the gut, the CAZymes found within *C. perfringens* encodes many *O*-glycanases including galactosidases, fucosidases, and an assortment of carbohydrate-binding modules that are essential for the following step of mucin glycan disassembly[69]. Included in this list is an endo acting GH16 that is valuable in its capability of releasing the disaccharide GlcNAc-α(1,3)-Gal, as well as the endo-β-galactosidase from the GH98 family that releases trisaccharides (known as blood group antigen A and B) that are found terminally on mucin chains[180,184,185]. The use of GH16 enzymes is commonly considered part of the initiating step of mucin breakdown, as they target the poly LacNAc that comprises the structure of lots of mucin

glycan chains[186]. However, research has shown that GH16 enzymes cannot cleave this glycan moiety if a sialic acid is present, necessitating the use of a sialidase before cleavage[182]. While glycosyl hydrolases play a crucial role in mucin degradation, they are only limited to trimming down the glycan chains that are attached to the peptide backbone. Even during glycan degradation, having the presence of signal glycan moieties such as Tn-antigen and the muc-cores on the peptide backbone can interfere with pancreatic digestive enzymes like trypsin: working as another stage of microbial selection[187]. In such cases, glycoproteases with an affinity of glycopeptides can work in conjunction with glycosyl hydrolases to complete the stages of mucin breakdown.

Glycopeptidases have been characterized for both commensal and pathogenic species, including *Akkermansia muciniphila* and *Clostridium perfringens*, fitting this type of enzymes into an impressive set of glycoconjugate-acting enzymes used in the mucin degradation pathway[148,165]. Amuc_1438 is the most recently discovered glycopeptidase from Akkermansia *muciniphila* that targets and cleaves peptides that are decorated with a single GalNAc sugar on a serine or threonine. This suggests that in the enzymatic mucin-degrading pathway, Amuc_1438 is utilized by *Akkermansia muciniphila* when the extracellular glycosyl hydrolases have trimmed the glycan chains down to a single GalNAc before Amuc_1438 targets cleave the glycosylated peptide bond for degradation. As Amuc_1438 is the fifth enzyme from *A. muciniphila* to be characterized with glycopeptidase activity, it is clear how important this class of enzyme is for mucin breakdown. Perhaps Amuc_1438 could stand out in the mucin degradation pathway when Tn-antigen is highly represented on mucin glycoproteins in comparison to the slightly broader glycan specificity seen with the other glycopeptidases. Further, the enzymatic trials indicated Amuc_1438 is sensitive to a certain peptide combination. It is reasonable to imagine that the biological significance of this could be Amuc_1438 targets a certain peptide combination that is less preferable than the alternative glycopeptidases of *A. muciniphila,* carving out its own niche in the pathway.

The glycopeptidase and endo-acting glycanase processing of the mucus glycoproteins can generate large oligosaccharides that are available for bacterial uptake for further

degradation[99]. There have been several different strategies taken by bacteria to import oligosaccharides, perhaps the best-known being the starch utilization system (SUS) of *Bacteroides*[188]*.* Unfortunately, a sialylated glycoconjugate uptake strategy in *Clostridium perfringens* has yet to be discovered. Yet, there are several encoded CAZymes belonging to *C. perfringens* that target mucosal glycosidic linkages that lack a signal peptide, indicating that they reside intercellularly to degrade imported oligosaccharides[69]. Many of which are multi-modular with the addition of CBMs. Within this group of intracellular CAZymes is NanH. Likely, the functionality of NanH is to remove the sialic acid group off imported sialoglycans, followed by a breakdown by a selection of the enzymes found within the nan operon of free sialic acid into fructose 6-phosphate for use in glycolysis[163].

One interesting concept is the idea that nutrients get shared with the microbial community resulting from the deconstruction of mucin[186]. Some evidence indicates that cross-feeding of nutrients occurs during mucin degradation. One example shows that some bacterial strains contain sialidases to release terminal sialic acids, yet lack the machinery to catabolize the sugar, whereas another resident bacteria cant cleave sialic acid but can degrade it[189]. Further, another example shows that co-culturing *A. muciniphila* with Clostridia demonstrated a dramatic increase in butyrate production, indicating that the sialic acid released by *Akkermansia muciniphila* likely is being utilized in the microbial community[125]. One benefit to this strategy would be removing the terminal sugars that pathogenic species use to adhere to the mucin layer, maintaining a balance of the commensal microbiome[189].

This thesis primarily examined the roles of two enzymes that play a part in the large map of mucin-degrading enzymes that function within the breakdown of mucin glycoproteins within the gastrointestinal system. Despite their distinct functions, both enzymes contribute to the overall process of glycoprotein breakdown. While CAZymes are responsible for dismantling the glycan chains of mucin, Amuc_1438 specifically targets and cleaves the Tn-Antigen-bearing glycopeptides present in the mucin backbone. This action completes the disassembly of the complex bottlebrush-shaped

mucin structure. Both enzymes operate within a dynamic environment alongside an array of unique and fascinating enzymes, all with the common objective of extracting nutrients from the abundant and intricately structured world of the gastrointestinal system.

Chapter 5: Materials and Methods

5.1 Cloning and transformation.

The construct AMUC1438WT was designed from the *Amuc_1438* gene and amplified from *Akkermansia muciniphila* (ATCC ® BAA-835™) genomic DNA using the Takara-Bio In–Fusion cloning kit. This construct was amplified using the Amuc1438WT_Fwd (NheI) and Amuc1438WT_Rev (XhoI) primer set (Table 2). The amplified genomic DNA was purified using the NucleoSpin plasmid miniprep kit (Takarabio) and ligated into a linearized pET28a plasmid vector. The pET28a plasmids for NanI, NanJ and NanH were sent from the Abbott laboratory at the University of Lethbridge, therefore no genomic amplification was required. Using the primer combinations outlined in table 2, QuickChange site directed mutagenesis took place to mutate the catalytic and key sugar binding pocket residues (of *Amuc_1438*). The collection of wildtype and mutated plasmids were transformed into highly competent Stellar Cells (Takarabio). For transformation, the competent cells were thawed from -80ºC on ice for 10 minutes prior to the addition of 1.5uL of ~ 80ng/µL plasmid DNA. This mixture was stored on ice for approximately 20 minutes followed by a 45-second heat shock in a 42ºC water bath. These cells were suspended on ice for five minutes before an addition of 1mL lysogeny broth (LB) and incubated for one hour at 37ºC. These grown cells were pelleted and then spread plated onto LB + 500µg/mL kanamycin agar plates. These plates were stored overnight at 37ºC to allow for colony growth for DNA sequencing.

Table 2. Primers for gene amplification

Amplicon	Primer Name	Sequence	TmºC
	Amuc1438WT_Fwd (NheI)	5’ – CAG CCA TAT GGCTAGC GACAGGGAGGGAGCGG -3’	69.9
	Amuc1438WT_Rev	5’ – GGTGGTGGTGCTCGAGTTATCCCCGGGCATCCATCC-3’	71.0

Amuc1438	(XhoI)		
	Amuc1438Y202A_Fwd	5’-GAAAGAACGCCACCATTGCCATAGGCGGTACGGCCC-3’	70.5
	Amuc1438Y202A_Rev	5’-GGGCCGTACCGCCTATGGCAATGGTGGCGTTCTTTC-3’	70.5
	Amuc1438E208A_Fwd	5’-CGGTACGGCCCATGCATTGGGGCATTCCT-3’	69.2
	Amuc1438E208A_Rev	5’-AGGAATGCCCCAATGCATGGCCCGTACCG-3’	69.2
	Amuc1438E208Q_Fwd	5’-GCCCATCAATTGGGGCATTCCTTCG-3’	62.6
	Amuc1438E208Q_Rev	5’- CCCCAATTGATGGGCCGTACCG-3’	62.8
	Amuc1438D151A_Fwd	5’-CCAGCTTCCGGCCGGCGTGGGTC-3’	71.0
	Amuc1438D151A_Rev	5’-GACCCACGCCGGCCGGAAGCTGG-3’	71.0
	Amuc1438Y151A_Fwd	5’-CGGCGTGGGTCCTGCTTACGGCGGCGGC-3’	74.9
	Amuc1438Y151A_Rev	5’-GGCGCCGCCGTAAGCAGGACCCACGCCG-3’	74.9
	AmucY156A_2_Fwd	5’-GGGTCCTGCTTACGGCGGCGGCTTCAGC-3’	71.5
	AmucY156A_2_Fwd	5’-CCGTAAGCAGGACCCACGCCGTCCGG-3’	70.2
	Amuc1438cbm51_Fwd	5’-CAGCCATATGGCTAGCGCCAGTATCAGCCTGAATGACTGCAAGCCTTCCG-3'	71.3
	Amuc1438cbm51_Rev	5’-GGTGGTGGTGCTCGAGTTACCGGGTCAGCATGCCGTTGGCTATAATGCCCC-3'	73.4
NanH	NanH_MutD62N_Fwd	5’-CCGGACAACCACGCGTACATCGAC-3’	63.8
	NanH_MutD62N_Rev	5’-CGCGTGGTTGTCCGGACCGTTG-3’	65.2
	NanH_MutY347F_Fwd	5’GGTGGCTTCTCTTGCCTGAGC-3’	60.7
	NanH_MutY347F_Rev	5’- GCAAGAGAAGCCACCACCCAG-3'	60.8
	NanHMutD62N_Fwd	5’-CCGGACAACCACGCGTACATCGAC-3'	63.8
	NanH_MUTD62N_Rev	5’-CGCGTGGTTGTCCGGACCGTTG-3'	65.2
NanI	NanIMutY655F_Fwd	5’-TACGCGTTCTCCTGCCTGACCGAAC-3'	64.4
	NanIMutY655F_Rev	5’-GCAGGAGAACGCGTAGTAACCCGG-3'	63.4
	NanIMutD291N_Fwd	5’-GGTGCTAACGCGCCGAACAACGAC-3'	64.7
NanJ	NanIMutD291N_Rev	5’-CGGCGCGTTAGCACCACCGTG-3'	66
	NanJMutD424N_Fwd	5’-GACCACAACGCGCCGAACAACAAC-3'	63.3
	NanJMutD424N_Rev	5’-CCGCGCGTTGTGGTCGCC-3'	65.7
	NanJMutY787F_Fwd	5’-TTCGCTTTCAGCTGCCTGACC-3'	60.3
	NanJMutY787F_Rev	5’-GCAGCTGAAAGCGAAGGAACC-3'	59.1

5.2 Protein expression and purification.

The pET28a plasmids containing the constructs were transformed (using the methods described in section 2.1) into the competent *Escherichia coli* strain BL21 DE3*. Individually six liters of 2x YT media was inoculated with each construct and grown in a 37ºC shaking incubator for approximately 5-7 hours. Once an optical density of around 0.9 (600nm) was measured, the cells were chemically induced with [0.5mM] final isopropyl-β-D-1-thiogalacto-pyranoside (IPTG) at 16ºC shaking overnight. The cell cultures were pelleted by centrifugation at 6000 rpm for 15 minutes at 4ºC.

The cell pellets underwent chemical lyses following methods described in Noach *et al.* 2017[145]. The pellet was resuspended in sucrose solution (25% sucrose, 20mM tris-HCl pH 8.0) on a stir plate, with the addition of 10mg of lysozyme to stir for 20 minutes. Double the sucrose volume of deoxycholate solution (1% deoxycholate, 1% triton X100, 50 mM tris-HCl pH 8.0, 100mM NaCl) was added to the cellular lysate. Once the lysate became very viscous, a final concentration of 200µM MgCl2 and 90µl of DNase I (2mg/mL) was added and stirred until the cellular lysate returned to its liquidated form. The lysate was span down at 4^{o}C at 16,500 x g for 45 minutes to separate the supernatant from the cellular content. The expressed protein was isolated by loading the supernant onto a Ni^{2+} immobilized metal affinity chromatography (IMAC) resin (GE Healthcare Streamline Chelating beads) bed. The IMAC column is equilibrated and washed with binding buffer (500mM NaCl, 20mM Tris pH 8.0, 10% glycerol) before an imidazole gradient ranging from 20mM – 500mM is used to separate the bound, his-tagged protein of interest. Samples were taken from each fraction and run on a 12% SDS/PAGE gel at 220V for 45 minutes before being stained with Coommassie dye to assess the purity of the fractions. Selecting the purest fractions for further experimentation, the eluant was concentrated on an ultrafiltration unit (Amicon) using a 10kDa membrane (EMD milipore). The concentrated protein was separated further from contamination by size exclusion chromatography, using a Sephacryl S-200 HR column (GE healthcare). Finally, concentration of the purified protein was determined through measuring the absorbance at 280nm while using the respective molecular weight and extinction coefficient of each protein (Table 3).

Table 3. Protein characteristics.

Construct	Molecular weight (Da)	Extinction coefficient $M^{-1}cm^{-1}$
Amuc1438	42209	53080
NanI	50511	74510
NanJ	50002	71865
NanH	42812	82530

5.3 Enzymatic activity assays.

5.3.1 Glycan specificity assay.

Thin layer chromatography (TLC) was used to detect AMUC1438 enzyme activity on a selection of uniquely decorated *O*-glycopeptides. AMUC1438 protein was concentrated at 10µM and incubated with unlabeled *O*-linked glycopeptide (5 µg/µL) in tris-HCl pH 7.5 for three hours at 37^{o}C. The reactions (3µL final) were spotted on a silica gel TLC plate. Similarly, the peptides labeled with BODIPY were incubated at 1 µg/ µL with Amuc1438 (1 µM) in 20mM tris-HCl pH 7.5 for three hours at 37^{o}C. The reaction products for both plates were separated in a butanol:acetic acid:H_2O (45:35:30 v:v:v) solvent. Once the silica plates were dry, the unlabelled *O*-linked glycopeptide plate was developed using a ninhydrin solution (1g in 95mL pyridine and 5 mL glacial acetic acid) at 110^{o}C for 15 minutes, where the BODIPY substrate plate was imaged under UV at a wavelength of 365 nm.

5.3.2 Fluorescence based kinetics on IGA1 derived glycopeptide.

The IGA1 based FRET glycopeptide was designed and ordered through Anaspec. The glycopeptide was based off the peptide hinge region of the IGA1 antibody, containing the sequence TPSP**S**TPPT with the bolded serine containing an *O*-linked GalNac. This glycopeptide contained an N-terminal HilLyte488 fluorophore, and a C-terminal dark quencher QXL520. All the kinetic experiments were completed at 25^{o}C in 384 well microtiter plates while read on a SpectraMax M5 plate reader using SoftMax pro 6.2.1 software. 1µM of Amuc1438 and a 0 – 1000µM range of FRET IGA1 was resuspended in 20mM Tris pH 7.0 and 100µM Zinc chloride. Fluorescence was measured using 492 and 530 nm wavelengths for respective excitation and emission including a 515 nm cut off filter. A standard curve was generated using a peptide based upon the fluorescent product of the experimental hydrolysis (HiLyteFluor 488-TPSPS). The inner filter effects of the cleavage of the FRET peptide were corrected as previously described in (9 pluvinage). The Michaelis-Menton equation was used to fit the raw data to determine the kinetic data for both CAT and FL of the Amuc1438 enzyme.

5.5 Protein melting curves of Amuc1438.

Protein melting curves were carried out in 200µL PCR tubes and measured in a Biorad CFX 96 Deepwell/C1000 Touch Thermocycler. The CAT construct, the catalytic mutant (E208Q), and the sugar-binding pocket mutants (E224A, Y171A, D166A, Y217A) were diluted to 1mg/mL in a 20mM tris pH range from pH 4 – pH9 and a 10X final SPYRO Orange Protein Gel Stain (Sigma-Aldrich). Each experiment was exposed to an increase of temperature from 10^{o}C – 95^{o}C and the amount of fluorescence was measured stepwise using an excitation wavelength of 470 nm and emission of 570 nm. The first derivative of each measurement was taken to construct the melting curves.

5.6 Protein crystallization and optimization.

NanH was crystallized by the sitting drop vapour diffusion method at 18^{o}C. The crystal conditions for NanH were determined by mixing a 1:1 ratio of purified protein with the range of conditions found in the MCSG 1-4 (Anatrace) and Index (HT) crystallization screens. NanH crystallized in the condition 0.1M HEPES:NaOH pH7.5, 20% PEG 8000. To obtain the NanH complexes with ligand, NanH crystals were soaked with 15mM of either Neu5Ac, STAg or SLN to obtain a crystal complex, where catalytically inactive D62N NanH mutants were used for the STAg and SLN complexes. The substrate complexes were obtained by soaking the crystals in the respective ligand for 30 minutes before flash cooling in liquid N_2. The crystals were cryoprotected using a 20% v/v ethylene glycol addition before exposure to X-ray detection. ALT crystals were grown in a 1:1 ratio of purified protein at a concentration of 63mg/mL with the crystallization condition 0.2 M $(NH_4)_2SO_4$, 20% (w/v) PEG3350, and 0.1 M Hepes pH 7.5.

5.7 X-ray diffraction data collection and Structural refinement.

Diffraction data was collected on a machine built out of a Pilatus 200K 2D detector coupled with a MicroMax-007HF X-ray generator, a VariMaxTM-HF ArcSec Confocal Optical System, and an Oxford Cryostream 800. Collected diffraction data sets were processed using HKL2000, shown in Table 4**.** Molecular replacement using PHASER[190]

determined the structures of NanH by utilizing the coordinates of apo NanI (PDB code 2VK6) as a search model. NanH was determined by the single isomorphous replacement and anomalous scattering (SIRAS) method that used a native data set and an iodide derivative. Using REFMAC[191] as an initial refinement software, the electron density maps were assessed for evidence of ligand binding and manual structure correction was completed using COOT[192]. The additions of water were performed in COOT using FINDWATERS, with manual revisions. Polishing of each structure's refinement was carried out using the software PHENIX Refine. All the data sets were monitored by flagging 5% of all observations as "free"[193]. Validation of the models was performed using MOLPROBITY[194]. The ALT structure of Amuc_1438 was also determined using SIRAS, using both a native data set and an iodide derivative. The SHARP/autoSHARP pipeline was used initially to determine the phases[195]. Following, phases were improved using PARROT to perform density modification and noncrystallographic averaging[196]. An initial model was constructed by autobuilding using BUCANNEER[197]. Similarly, to the NanH structures, the ALT structure was finalized by successive rounds of model building with Coot and refinement with REFMAC.

Table 4. X-ray data collection and structure statistics.

	NanH Native	NanH Neu5Ac	NanH Neu59Ac	NanH (Diacyl sialo(2,6)-LAcNAc)	Amuc_1438 ALT iodide	Amuc_1438 ALT native	Amuc_1438 ALTL
Data Collection							
Beamline	Home Beam	Home Beam	Home Beam	Home Beam	Home Beam	Home Beam	Home Beam
Wavelength (Å)	1.54178	1.54178	1.54178	1.54178	1.54178	1.54178	1.54178
Space Group	C 1 2 1	C 1 2 1	C 1 2 1	C 1 2 1	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Cell Dimensions							
a, b, c (Å)	93.44, 64.49, 65.75	93.86, 64.38, 66.81	94.04, 64.26, 66.27	93.18, 64.69, 65.58	88.7, 145.8, 147.5	88.6, 146.1, 147.3	72.2, 91.5, 161.7
α, β, γ (º)	90.00,102.71, 90.00	90.0, 102.45, 90.0	90.00 102.7, 90.00	90.00,102.75, 90.00	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Resolution (Å)	30.00-1.80	23.00-1.85	22.9 – 1.89	24.00-2.46	25.00 – 2.40 (2.44-2.40)	25.00-2.35 (2.39-2.35)	2500-2.500 (2.54-2.50)
R_{meas}	0.094 (0.661)	0.100 (0.730)	0.142 (0.443)	0.083 (0.438)	0.153 (1.131)	0.159 (0.393)	0.092 (0.346)
R_{pim}	0.046 (0.366)	0.040 (0.446)	0.057 (0.263)	0.037 (0.272)	0.030 (0.339)	(0.072 (0.228)	0.043 (0.195)
CC1/2	0.991* (0.704)	0.997* (0.762)	0.981* (0.881)	0.993* (0.815)	0.998 (0.715)	0.986 (0.902)	0.992 (0.860)
<I/σI>	14.9 (2.0)	15.9 (2.0)	11.6 (2.3)	16.1 (2.5)	24.7 (2.0)	8.2 (1.9)	15.2

Completeness (%)	99.9 (98.6)	99.8 (99.8)	100.00 (99.9)	99.3 (96.4)	99.8 (97.6)	98.2 (96.9)	98.7 (96.5)
Redundancy	4.0 (2.9)	4.3 (2.4)	4.6 (2.6)	3.9 (2.0)	24.8 (10.0)	3.8 (2.5)	4.1 (2.9)
No. of reflections	140406	155440	141496	101754	1,880,433	289,777	147,473
No. Unique	35412 (1757)	36096 (1823)	30625 (1493)	25987 (1267)	75,769 (3634)	78,394 (3818)	36,634 (1773)
Refinement							
Resolution (Å)	1.80	1.85	1.89	2.46		25.00-2.35	25.00-2.50
R_{work}/R_{free}	0.16/0.20	0.18/0.22	0.18/0.21	0.18/0.24		0.23/0.27	0.23/0.28
No. of atoms							
Protein	2849 (A)	2838 (A)	2839 (A)	2816 (A)		2800 (A), 2820 (B)<,2809 (C), 2830 (D)	3265 (A), 3292 (B)
Ligand	42 ACE, 56 EDO	21 SIA, 60 EDO	24 5N6, 56 EDO	52 Diacyl sialo(2,6)-LAcNAc, 40 EDO		4 Zn 4 Ca	2 Zn
Water	274	246	228	129		320	60
B-factors							
Protein	19.7 (A)	20.8 (A)	21.0 (A)	34.3 (A)		33.1 (A)/33.7 (B)/36.8 (C)/34.2 (D)	34.4 (A), 40.2 (B)
Ligand	28.2 (ACE), 27.4	23.3 (SIA), 31.1	26.72 (5N6), 37.0 (EDO)	44.8 (Diacyl sialo(2,6)-LAcNAc), 39.9 EDO		37 (Zn), 34.4 (Ca)	41.2
Water	25.8	24.9	26.5	34.5		30.7	27.8
r.m.s.d							
Bond lengths (Å)	0.011	0.010	0.013	0.002		0.002	0.002
Bond angles (°)	1.078	1.071	1.338	0.562		0.552	0.449
Ramachandran (%)							
Preferred	97.2	96.1	96.7	95.8		96.8	97.4
Allowed	2.5	3.6	3.0	3.9		2.9	2.6
Disallowed	0.3	0.3	0.3	0.3		0.3	0.0

Values for highest resolution shells are shown in parenthesis.
*Value refers to low resolution shell.

5.8 Mass spectrometry.

***Note;* This work was completely done by the Malaker laboratory, as well as the methodology description.**

Recombinantly expressed podocalyxin, CD43, and PSGL-1 were purchased from R&D

Systems (1658-PD, 9680-CD, and 3345-PS, respectively). C1 esterase inhibitor from human plasma (catalog no.: E0518) and sialidase (catalog no.: 11080725001) were purchased from Sigma. Each protein was reconstituted in 100 ng/µl of 50 mM ammonium bicarbonate. For each protein, four 1 µg samples were prepared. CAT was added to two of the samples at a 1:10 enzyme:protein ratio. Sialidase (100 µU) was added to two samples: one without CAT and one including CAT. The digestion was incubated at 37 followed by one time of 1 ml rinse of 0.1% formic acid in water ("buffer A"). The samples were then added to the column and rinsed with 150 µl of 0.1% formic acid. Finally, the samples were eluted twice with 150 µl of 0.1% formic acid in 30% CAN and dried by vacuum centrifugation. The samples were reconstituted in 10 µl of buffer A for MS analysis. Samples were analyzed by online nanoflow liquid chromatography–tandem MS using an Orbitrap Eclipse Tribrid mass spectrometer (Thermo Fisher Scientific) coupled to a Dionex Ultimate 3000 HPLC (Thermo Fisher Scientific). A portion of the sample (400 ng) was loaded via autosampler isocratically onto a C18 nano precolumn using buffer A. For preconcentration and desalting, the column was washed with 2% ACN and 0.1% formic acid in water ("loading pump solvent"). Subsequently, the C18 nano precolumn was switched in line with the C18 nano separation column (75 µm × 250 mm EASYSpray containing 2 µm C18 beads) for gradient elution. The column was held at 35 C using a column heater in the EASY-Spray ionization source (Thermo Fisher Scientific). The samples were eluted at a constant flow rate of 0.3 µl/min using a 60 min gradient. The gradient profile was as follows: 0-0-35-95-95-2%B in 0-5-65-70-75 to 77 min, respectively. The instrument method used an MS1 resolution of 60,000 full width at half maximum at 400 m/z, an automatic gain control (AGC) target of 3e5, and a mass range from 300 to 1500 m/z. Dynamic exclusion was enabled with a repeat count of 3, repeat duration of 10 s, and exclusion duration of 10 s. Only charge states 2 to 6 were selected for fragmentation. MS2s were generated at top speed for 3 s. Higher energy collisional dissociation (HCD) was performed on all selected precursor masses with the following parameters: isolation window of 2 m/z, 28% collision energy, orbitrap detection (resolution of 7500), maximum injection time of 75 ms, and an AGC target of 1e4 ions. Electron-transfer/higher energy collision dissociation with supplemental activation was performed if (1) the precursor

mass was between 300 and 1500 m/z and (2) three of nine HexNAc or NeuAc fingerprint ions (126.055, 138.055, 144.07, 168.065, 186.076, 204.086, 274.092, and 292.103) were present at ±0.1 m/z and greater than 5% relative intensity. Electron-transfer/higher energy collision dissociation parameters were as follows: Orbitrap detection (resolution of 7500) calibrated charge-dependent electron transfer dissociation times, 15% normalized collision energy for HCD, maximum injection time of 250 ms, reagent AGC target of 5e5, and precursor AGC target of 1e4. Raw files were searched using O-Pair search with Meta-Morpheus against directed databases containing the recombinant protein of interest. Files were searched using nonspecific cleavage specificity. Mass tolerance was set at 10 ppm for MS1s and 20 ppm for MS2s. Cysteine carbami-domethylation was set as a fixed modification, and methionine oxidation was allowed as a variable modification. The default O-glycan database was included, and a maximum number of glycosites per peptide was set to 4. Peptide hits were filtered using a 1% false discovery rate. All peptides were manually validated and/or sequenced using Xcalibur software (Thermo Fisher Scientific). After all peptides unique to the mucinase-C overnight. Samples were then reduced in 2 mM DTT at 65 digested samples were sequenced, peptides ±5 amino acids from the cleavage site were input into **weblogo.berkeley.edu** to generate the consensus motif.

Chapter 6: References.

1. Brotman, R. M., Ravel, J., Bavoil, P. M., Gravitt, P. E. & Ghanem, K. G. Microbiome, sex hormones, and immune responses in the reproductive tract: Challenges for vaccine development against sexually transmitted infections. *Vaccine* **32**, 1543–1552 (2014).
2. Ahluwalia, B., Magnusson, M. K. & Öhman, L. Mucosal immune system of the gastrointestinal tract: maintaining balance between the good and the bad. *Scandinavian Journal of Gastroenterology* vol. 52 1185–1193 Preprint at https://doi.org/10.1080/00365521.2017.1349173 (2017).
3. Pilette, C., Ouadrhiri, Y., Godding, V., Vaerman, J.-P. & Sibille, Y. Lung mucosal immunity: immunoglobulin-A revisited. *European Respiratory Journal* **18**, 571–588 (2001).
4. Rose, M. C. Mucins: structure, function, and role in pulmonary diseases. *Lung Cell. Mol. Physiol* **263**, 413–429 (1992).
5. Johansson, M. E. V. *et al.* The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proc Natl Acad Sci U S A* **105**, 15064–15069 (2008).
6. Rogier, E. W., Frantz, A. L., Bruno, M. E. C. & Kaetzel, C. S. Secretory IgA is concentrated in the outer layer of colonic mucus along with gut bacteria. *Pathogens* **3**, 390–403 (2014).
7. 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 Preprint at https://doi.org/10.1038/nrgastro.2013.35 (2013).
8. Johansson, M. E. V., Holmén Larsson, J. M. & Hansson, G. C. The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. *Proc Natl Acad Sci U S A* **108**, 4659–4665 (2011).
9. Varum, F. J. O., Veiga, F., Sousa, J. S. & Basit, A. W. Mucus thickness in the gastrointestinal tract of laboratory animals. *Journal of Pharmacy and Pharmacology* **64**, 218–227 (2012).

10. Strugala, V., Dettmar, P. W. & Pearson, J. P. Thickness and continuity of the adherent colonic mucus barrier in active and quiescent ulcerative colitis and Crohn's disease. *Int J Clin Pract* **62**, 762–769 (2008).
11. Paone, P. & Cani, P. D. Mucus barrier, mucins and gut microbiota: The expected slimy partners? *Gut* vol. 69 2232–2243 Preprint at https://doi.org/10.1136/gutjnl-2020-322260 (2020).
12. McGuckin, M. A., Lindén, S. K., Sutton, P. & Florin, T. H. Mucin dynamics and enteric pathogens. *Nat Rev Microbiol* **9**, 265–278 (2011).
13. Allen, A., Hutton, D. A. & Pearson, J. P. The MUC2 gene product: A human intestinal mucin. *International Journal of Biochemistry and Cell Biology* **30**, 797–801 (1998).
14. Callaghan, M. & Voynow, J. A. Respiratory Tract Mucin Genes and Mucin Glycoproteins in Health and Disease. *the American Physiological Society* **86**, 245–278 (2006).
15. Engevik, M. A. *et al.* Mucin-Degrading Microbes Release Monosaccharides That Chemoattract Clostridioides difficile and Facilitate Colonization of the Human Intestinal Mucus Layer. *ACS Infect Dis* **7**, 1126–1142 (2021).
16. Kesimer, M., Makhov, A. M., Griffith, J. D., Verdugo, P. & Sheehan, J. K. Unpacking a gel-forming mucin: A view of MUC5B organization after granular release. *Am J Physiol Lung Cell Mol Physiol* **298**, 15–22 (2010).
17. Grondin, J. A., Kwon, Y. H., Far, P. M., Haq, S. & Khan, W. I. Mucins in Intestinal Mucosal Defense and Inflammation: Learning From Clinical and Experimental Studies. *Frontiers in Immunology* Preprint at https://doi.org/10.3389/fimmu.2020.02054 (2020).
18. Knoop, K. A. & Newberry, R. D. Goblet cells: multifaceted players in immunity at mucosal surfaces. *Mucosal Immunology* vol. 11 1551–1557 Preprint at https://doi.org/10.1038/s41385-018-0039-y (2018).
19. Boltin, D., Perets, T. T., Vilkin, A. & Niv, Y. *Mucin Function in Inflammatory Bowel Disease An Update*. *J Clin Gastroenterol* vol. 47 www.jcge.com (2013).
20. Sekirov, I., Russell, S. L., Caetano M Antunes, L. & Finlay, B. B. Gut microbiota in health and disease. *Physiological Reviews* vol. 90 859–904 Preprint at https://doi.org/10.1152/physrev.00045.2009 (2010).
21. Juge, N., Tailford, L. & Owen, C. D. Sialidases from gut bacteria: A mini-review. *Biochem Soc Trans* **44**, 166–175 (2016).
22. Ambort, D. *et al.* Calcium and pH-dependent packing and release of the gel-forming MUC2 mucin. *Proc Natl Acad Sci U S A* **109**, 5645–5650 (2012).
23. Cornick, S., Tawiah, A. & Chadee, K. Roles and regulation of the mucus barrier in the gut. *Tissue Barriers* vol. 3 Preprint at https://doi.org/10.4161/21688370.2014.982426 (2015).
24. Johansson, M. E. V. & Hansson, G. C. Immunological aspects of intestinal mucus and mucins. *Nature Reviews Immunology* vol. 16 639–649 Preprint at https://doi.org/10.1038/nri.2016.88 (2016).
25. Lang, T., Hansson, G. C. & Samuelsson, T. Gel-forming mucins appeared early in metazoan evolution. *Proc Natl Acad Sci U S A* **104**, 16209–16214 (2007).
26. Strous, G. J. & Dekker, J. Mucin-type glycoproteins. *Crit Rev Biochem Mol Biol* **27**, 57–92 (1992).

27. Lang, T. *et al.* Searching the Evolutionary Origin of Epithelial Mucus Protein Components - Mucins and FCGBP. *Mol Biol Evol* **33**, 1921–1936 (2016).
28. Yamashita, M. S. de A. & Melo, E. O. Mucin 2 (MUC2) promoter characterization: an overview. *Cell Tissue Res* **374**, 455–463 (2018).
29. Morrison, C. B., Markovetz, M. R. & Ehre, C. Mucus, mucins, and cystic fibrosis. *Pediatric Pulmonology* vol. 54 S84–S96 Preprint at https://doi.org/10.1002/ppul.24530 (2019).
30. Low, K. E., Smith, S. P., Abbott, D. W. & Boraston, A. B. The glycoconjugate-degrading enzymes of Clostridium perfringens: Tailored catalysts for breaching the intestinal mucus barrier. *Glycobiology* vol. 31 681–690 Preprint at https://doi.org/10.1093/glycob/cwaa050 (2021).
31. Perez, B. H. & Gipson, I. K. Focus on Molecules: Human mucin MUC16. *Experimental Eye Research* vol. 87 400–401 Preprint at https://doi.org/10.1016/j.exer.2007.12.008 (2008).
32. Kesimer, M., Makhov, A. M., Griffith, J. D., Verdugo, P. & Sheehan, J. K. Unpacking a gel-forming mucin: A view of MUC5B organization after granular release. *Am J Physiol Lung Cell Mol Physiol* **298**, 15–22 (2010).
33. 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**, 189–215 (2018).
34. Bergstrom, K. S. B. & Xia, L. Mucin-type O-glycans and their roles in intestinal homeostasis. *Glycobiology* vol. 23 1026–1037 Preprint at https://doi.org/10.1093/glycob/cwt045 (2013).
35. Ju, T., Otto, V. I. & Cummings, R. D. The Tn antigena-structural simplicity and biological complexity. *Angewandte Chemie - International Edition* vol. 50 1770–1791 Preprint at https://doi.org/10.1002/anie.201002313 (2011).
36. Guzman-Aranguez, A. & Argüeso, P. Structure and Biological Roles of Mucin-type O-glycans at the Ocular Surface. *Ocul. Surf.* **8**, 8–17 (2010).
37. Dell, A., Galadari, A., Sastre, F. & Hitchen, P. Similarities and differences in the glycosylation mechanisms in prokaryotes and eukaryotes. *Int J Microbiol* **2010**, (2010).
38. Röttger, S. *et al.* Localization of three human polypeptide GalNAc-transferases in HeLa cells suggests initiation of O-linked glycosylation throughout the Golgi apparatus. *J Cell Sci* **111**, 45–60 (1998).
39. Becker, J. L., Tran, D. T. & Tabak, L. A. Members of the GalNAc-T family of enzymes utilize distinct Golgi localization mechanisms. *Glycobiology* **28**, 841–848 (2018).
40. González-Morelo, K. J., Vega-Sagardía, M. & Garrido, D. Molecular Insights Into O-Linked Glycan Utilization by Gut Microbes. *Frontiers in Microbiology* vol. 11 Preprint at https://doi.org/10.3389/fmicb.2020.591568 (2020).
41. Jensen, P. H., Kolarich, D. & Packer, N. H. Mucin-type O-glycosylation - Putting the pieces together. *FEBS Journal* vol. 277 81–94 Preprint at https://doi.org/10.1111/j.1742-4658.2009.07429.x (2010).
42. Tarp, M. A. & Clausen, H. Mucin-type O-glycosylation and its potential use in drug and vaccine development. *Biochim Biophys Acta Gen Subj* **1780**, 546–563 (2008).

43. Ujita, M. *et al.* Synthesis of Poly-N-acetyllactosamine in Core 2 Branched O-Glycans the Requirement of Novel-1,4-Galactosyltransferase IV and-1,3-N-Acetylglucosaminyltransferase. *J Biol Chem* **273**, 34843–34849 (1998).
44. Yang, J.-M. *et al. Alterations of O-glycan biosynthesis in human colon cancer tissues*. *Glycobiology* vol. 4 https://academic.oup.com/glycob/article/4/6/873/768264 (1994).
45. Guzman-Aranguez, A. & Argueso, P. Structure and Biological Roles of Mucin-type O-glycans at the Ocular Surface. *Laboratory Science* **1**, 8–17 (2010).
46. Arike, L. & Hansson, G. C. The Densely O-Glycosylated MUC2 Mucin Protects the Intestine and Provides Food for the Commensal Bacteria. *Journal of Molecular Biology* vol. 428 3221–3229 Preprint at https://doi.org/10.1016/j.jmb.2016.02.010 (2016).
47. Van Den Steen, P., Rudd, P. M., Dwek, R. A. & Opdenakker, G. Concepts and principles of O-linked glycosylation. *Critical Reviews in Biochemistry and Molecular Biology* vol. 33 151–208 Preprint at https://doi.org/10.1080/10409239891204198 (1998).
48. Hounsell, E. F. O-linked protein glycosylation structure and function. . *Glycoconj J* **13**, 19–26 (1996).
49. Novak, J., Julian, B. A., Tomana, M. & Mestecky, J. IgA Glycosylation and IgA Immune Complexes in the Pathogenesis of IgA Nephropathy. *Semin Nephrol* **28**, 78–87 (2008).
50. Patil, Y., Gooneratne, R. & Ju, X. H. Interactions between host and gut microbiota in domestic pigs: a review. *Gut Microbes* **11**, 310–334 (2020).
51. Malard, F., Dore, J., Gaugler, B. & Mohty, M. Introduction to host microbiome symbiosis in health and disease. *Mucosal Immunology* vol. 14 547–554 Preprint at https://doi.org/10.1038/s41385-020-00365-4 (2021).
52. Christian, N., Whitaker, B. K. & Clay, K. Microbiomes: Unifying animal and plant systems through the lens of community ecology theory. *Frontiers in Microbiology* vol. 6 Preprint at https://doi.org/10.3389/fmicb.2015.00869 (2015).
53. Liu, X. Microbiome. *Yale Journal of Biology and Medicine* **89**, 275–276 (2016).
54. Berg, G. *et al.* Microbiome definition re-visited: old concepts and new challenges. *Microbiome* vol. 8 Preprint at https://doi.org/10.1186/s40168-020-00875-0 (2020).
55. Degruttola, A. K., Low, D., Mizoguchi, A. & Mizoguchi, E. Current understanding of dysbiosis in disease in human and animal models. *Inflamm Bowel Dis* **22**, 1137–1150 (2016).
56. Devkota, S. & Chang, E. B. Nutrition, microbiomes, and intestinal inflammation. *Current Opinion in Gastroenterology* vol. 29 603–607 Preprint at https://doi.org/10.1097/MOG.0b013e328365d38f (2013).
57. Tremaroli, V. & Bäckhed, F. Functional interactions between the gut microbiota and host metabolism. *Nature* vol. 489 242–249 Preprint at https://doi.org/10.1038/nature11552 (2012).
58. Carding, S., Verbeke, K., Vipond, D. T., Corfe, B. M. & Owen, L. J. Dysbiosis of the gut microbiota in disease. *Microb Ecol Health Dis* **26**, (2015).
59. Bien, J., Palagani, V. & Bozko, P. The intestinal microbiota dysbiosis and Clostridium difficile infection: Is there a relationship with inflammatory bowel

disease? *Therapeutic Advances in Gastroenterology* vol. 6 53–68 Preprint at https://doi.org/10.1177/1756283X12454590 (2013).
60. Kim, C. H. Immune regulation by microbiome metabolites. *Immunology* vol. 154 220–229 Preprint at https://doi.org/10.1111/imm.12930 (2018).
61. Den Besten, G. *et al.* The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Journal of Lipid Research* vol. 54 2325–2340 Preprint at https://doi.org/10.1194/jlr.R036012 (2013).
62. Kim, S. *et al.* Mucin degrader Akkermansia muciniphila accelerates intestinal stem cell-mediated epithelial development. *Gut Microbes* **13**, 1–20 (2021).
63. Dieterich, W., Schink, M. & Zopf, Y. Microbiota in the Gastrointestinal Tract. *Medical sciences (Basel, Switzerland)* vol. 6 Preprint at https://doi.org/10.3390/medsci6040116 (2018).
64. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol* **14**, (2016).
65. Rinninella, E. *et al.* What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. *Microorganisms* **7**, (2019).
66. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
67. Cai, R. *et al.* Interactions of commensal and pathogenic microorganisms with the mucus layer in the colon. *Gut Microbes* vol. 11 680–690 Preprint at https://doi.org/10.1080/19490976.2020.1735606 (2020).
68. Zhang, P. Influence of Foods and Nutrition on the Gut Microbiome and Implications for Intestinal Health. *International Journal of Molecular Sciences* vol. 23 Preprint at https://doi.org/10.3390/ijms23179588 (2022).
69. Low, K. E., Smith, S. P., Abbott, D. W. & Boraston, A. B. The glycoconjugate-degrading enzymes of Clostridium perfringens: Tailored catalysts for breaching the intestinal mucus barrier. *Glycobiology* vol. 31 681–690 Preprint at https://doi.org/10.1093/glycob/cwaa050 (2021).
70. Petit, L., Gibert, M. & Popoff, M. R. Clostridium perfringens: toxinotype and genotype. *Trends in microbiology* 104–110 (1999).
71. Rood, J. I. *et al.* Expansion of the Clostridium perfringens toxin-based typing scheme. *Anaerobe* **53**, 5–10 (2018).
72. Forti, K. *et al.* Molecular characterization of clostridium perfringens strains isolated in Italy. *Toxins (Basel)* **12**, (2020).
73. Li, J. & McClane, B. A. NanH Is Produced by Sporulating Cultures of Clostridium perfringens Type F Food Poisoning Strains and Enhances the Cytotoxicity of C. perfringens Enterotoxin. *mSphere* **6**, (2021).
74. MacMillan, J. L. *et al.* Structural analysis of broiler chicken small intestinal mucin O-glycan modification by Clostridium perfringens. *Poult Sci* **98**, 5074–5088 (2019).
75. Grass, J. E., Gould, L. H. & Mahon, B. E. Epidemiology of foodborne disease outbreaks caused by clostridium perfringens, United States, 1998-2010. *Foodborne Pathog Dis* **10**, 131–136 (2013).
76. Freedman, J. C., Shrestha, A. & McClane, B. A. Clostridium perfringens enterotoxin: Action, genetics, and translational applications. *Toxins* vol. 8 Preprint at https://doi.org/10.3390/toxins8030073 (2016).

77. Chiarezza, M. *et al.* The NanI and NanJ sialidases of Clostridium perfringens are not essential for virulence. *Infect Immun* **77**, 4421–4428 (2009).
78. Yan, X. X. *et al.* Structural and functional analysis of the pore-forming toxin NetB from clostridium perfringens. *mBio* **4**, (2013).
79. Adams, J. J. *et al.* Structural basis of Clostridium perfringens toxin complex formation. *PNAS* **105**, 12194–12199 (2008).
80. Derrien, M., Vaughan, E. E., Plugge, C. M. & de Vos, W. M. Akkermansia muniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *Int J Syst Evol Microbiol* **54**, 1469–1476 (2004).
81. Geerlings, S. Y., Kostopoulos, I., de Vos, W. M. & Belzer, C. Akkermansia muciniphila in the human gastrointestinal tract: When, where, and how? *Microorganisms* vol. 6 Preprint at https://doi.org/10.3390/microorganisms6030075 (2018).
82. Ouwerkerk, J. P., Aalvink, S., Belzer, C. & de Vos, W. M. Akkermansia glycaniphila sp. nov., an anaerobic mucin-degrading bacterium isolated from reticulated python faeces. *Int J Syst Evol Microbiol* **66**, 4614–4620 (2016).
83. Cani, P. D. & de Vos, W. M. Next-generation beneficial microbes: The case of Akkermansia muciniphila. *Frontiers in Microbiology* vol. 8 Preprint at https://doi.org/10.3389/fmicb.2017.01765 (2017).
84. Zhang, T., Li, Q., Cheng, L., Buch, H. & Zhang, F. Akkermansia muciniphila is a promising probiotic. *Microbial Biotechnology* vol. 12 1109–1125 Preprint at https://doi.org/10.1111/1751-7915.13410 (2019).
85. Everard, A. *et al.* Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proc Natl Acad Sci U S A* **110**, 9066–9071 (2013).
86. Zhou, K. Strategies to promote abundance of Akkermansia muciniphila, an emerging probiotics in the gut, evidence from dietary intervention studies. *J Funct Foods* **33**, 194–201 (2017).
87. Ottman, N. *et al.* Genomescale model and omics analysis of metabolic capacities of Akkermansia muciniphila reveal a preferential mucin-degrading lifestyle. *Appl Environ Microbiol* **83**, (2017).
88. Desai, M. S. *et al.* A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility. *Cell* **167**, 1339-1353.e21 (2016).
89. van Passel, M. W. J. *et al.* The genome of Akkermansia muciniphila, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. *PLoS One* **6**, (2011).
90. Labourel, A. *et al.* O -Mucin-degrading carbohydrate-active enzymes and their possible implication in inflammatory bowel diseases . *Essays Biochem* (2023) doi:10.1042/ebc20220153.
91. Chao, L. & Jongkees, S. High-Throughput Approaches in Carbohydrate-Active Enzymology: Glycosidase and Glycosyl Transferase Inhibitors, Evolution, and Discovery. *Angewandte Chemie* **131**, 12880–12890 (2019).
92. Samuel, G. & Reeves, P. Biosynthesis of O-antigens: Genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. *Carbohydrate*

Research vol. 338 2503–2519 Preprint at https://doi.org/10.1016/j.carres.2003.07.009 (2003).
93. Goth, C. K., Petäjä-Repo, U. E. & Rosenkilde, M. M. G Protein-Coupled Receptors in the Sweet Spot: Glycosylation and other Post-translational Modifications. *ACS Pharmacology and Translational Science* vol. 3 237–245 Preprint at https://doi.org/10.1021/acsptsci.0c00016 (2020).
94. Prasad, Y. S. *et al.* Enzymatic synthesis and self-assembly of glycolipids: robust self-healing and wound closure performance of assembled soft materials. *RSC Adv* **8**, 37136–37145 (2018).
95. Berlemont, R. & Martiny, A. C. Glycoside Hydrolases across Environmental Microbial Communities. *PLoS Comput Biol* **12**, 1–16 (2016).
96. Michel, G. *et al.* The κ-carrageenase of P. carrageenovora Features a Tunnel-Shaped Active Site. *Structure* **9**, 513–525 (2001).
97. Crouch, L. I. *et al.* Prominent members of the human gut microbiota express endo-acting O-glycanases to initiate mucin breakdown. *Nat Commun* **11**, (2020).
98. Onyango, S. O., Juma, J., De Paepe, K. & Van de Wiele, T. Oral and Gut Microbial Carbohydrate-Active Enzymes Landscape in Health and Disease. *Front Microbiol* **12**, (2021).
99. Wardman, J. F., Bains, R. K., Rahfeld, P. & Withers, S. G. Carbohydrate-active enzymes (CAZymes) in the gut microbiome. *Nature Reviews Microbiology* vol. 20 542–556 Preprint at https://doi.org/10.1038/s41579-022-00712-1 (2022).
100. Sathya, T. A. & Khan, M. Diversity of glycosyl hydrolase enzymes from metagenome and their application in food industry. *J Food Sci* **79**, R2149–R2156 (2014).
101. Lombard, V. *et al.* A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochemical Journal* **432**, 437–444 (2010).
102. Nakamura, A. M., Nascimento, A. S. & Polikarpov, I. Structural diversity of carbohydrate esterases. *Biotechnology Research and Innovation* **1**, 35–51 (2017).
103. Sista Kameshwar, A. K. & Qin, W. Understanding the structural and functional properties of carbohydrate esterases with a special focus on hemicellulose deacetylating acetyl xylan esterases. *Mycology* **9**, 273–295 (2018).
104. Sidar, A. *et al.* Carbohydrate Binding Modules: Diversity of Domain Architecture in Amylases and Cellulases From Filamentous Microorganisms. *Frontiers in Bioengineering and Biotechnology* vol. 8 Preprint at https://doi.org/10.3389/fbioe.2020.00871 (2020).
105. Shoseyov, O., Shani, Z. & Levy, I. Carbohydrate Binding Modules: Biochemical Properties and Novel Applications. *Microbiology and Molecular Biology Reviews* **70**, 283–295 (2006).
106. Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P. & Forano, E. Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* vol. 3 Preprint at https://doi.org/10.4161/gmic.19897 (2012).
107. Sonnenburg, J. L. *et al.* Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science (1979)* **307**, 1955–1959 (2005).
108. Hehemann, J. H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010).

109. Glover, J. S., Ticer, T. D. & Engevik, M. A. Characterizing the mucin-degrading capacity of the human gut microbiota. *Sci Rep* **12**, 1–14 (2022).
110. Giacopuzzi, E., Bresciani, R., Schauer, R., Monti, E. & Borsani, G. New Insights on the Sialidase Protein Family Revealed by a Phylogenetic Analysis in Metazoa. *PLoS One* **7**, (2012).
111. Kim, J. Y. *et al.* Selective and slow-binding inhibition of shikonin derivatives isolated from Lithospermum erythrorhizon on glycosyl hydrolase 33 and 34 sialidases. *Bioorg Med Chem* **20**, 1740–1748 (2012).
112. Dou, D., Revol, R., Östbye, H., Wang, H. & Daniels, R. Influenza A virus cell entry, replication, virion assembly and movement. *Frontiers in Immunology* vol. 9 Preprint at https://doi.org/10.3389/fimmu.2018.01581 (2018).
113. Pons, T., Naumoff, D. G., Martínez-Fleites, C. & Hernández, L. Three Acidic Residues are at the Active Site of a β-Propeller Architecture in Glycoside Hydrolase Families 32, 43, 62, and 68. *Proteins: Structure, Function and Genetics* **54**, 424–432 (2004).
114. Davies, G. & Henrissat, B. Structures and mechanisms of glycosyl hydrolases. *Structure* 853–859 (1995).
115. Angata, T. & Varki, A. Chemical diversity in the sialic acids and related α-keto acids: An evolutionary perspective. *Chem Rev* **102**, 439–469 (2002).
116. Schauer, R. & Kamerling, J. P. Exploration of the Sialic Acid World. *Adv Carbohydr Chem Biochem* **75**, 1–213 (2018).
117. Wang, B. Sialic acid is an essential nutrient for brain development and cognition. *Annu Rev Nutr* **29**, 177–222 (2009).
118. Sokolovskaya, O. M., Tan, M. W. & Wolan, D. W. Sialic acid diversity in the human gut: Molecular impacts and tools for future discovery. *Current Opinion in Structural Biology* vol. 75 Preprint at https://doi.org/10.1016/j.sbi.2022.102397 (2022).
119. Tangvoranuntakul, P. *et al.* Human uptake and incorporation of an immunogenic nonhuman dietary sialic acid. *Proc Natl Acad Sci U S A* **100**, 12045–12050 (2003).
120. Varki, A. Uniquely human evolution of sialic acid genetics and biology. *Proc Natl Acad Sci U S A* **107**, 8939–8946 (2010).
121. Ulloa, F. & Real, F. X. *Differential Distribution of Sialic Acid in 2,3 and 2,6 Linkages in the Apical Membrane of Cultured Epithelial Cells and Tissues*. *The Journal of Histochemistry & Cytochemistry* vol. 49 http://www.jhc.org (2001).
122. Marcobal, A., Southwick, A. M., Earle, K. A. & Sonnenburg, J. L. A refined palate: Bacterial consumption of host glycans in the gut. *Glycobiology* vol. 23 1038–1046 Preprint at https://doi.org/10.1093/glycob/cwt040 (2013).
123. Robinson, L. S., Lewis, W. G. & Lewis, A. L. The sialate O-acetylesterase EstA from gut Bacteroidetes species enables sialidase-mediated cross-species foraging of 9-O-acetylated sialoglycans. *Journal of Biological Chemistry* **292**, 11861–11872 (2017).
124. Robinson, L. S., Lewis, W. G. & Lewis, A. L. The sialate O-acetylesterase EstA from gut Bacteroidetes species enables sialidase-mediated cross-species foraging of 9-O-acetylated sialoglycans. *Journal of Biological Chemistry* **292**, 11861–11872 (2017).

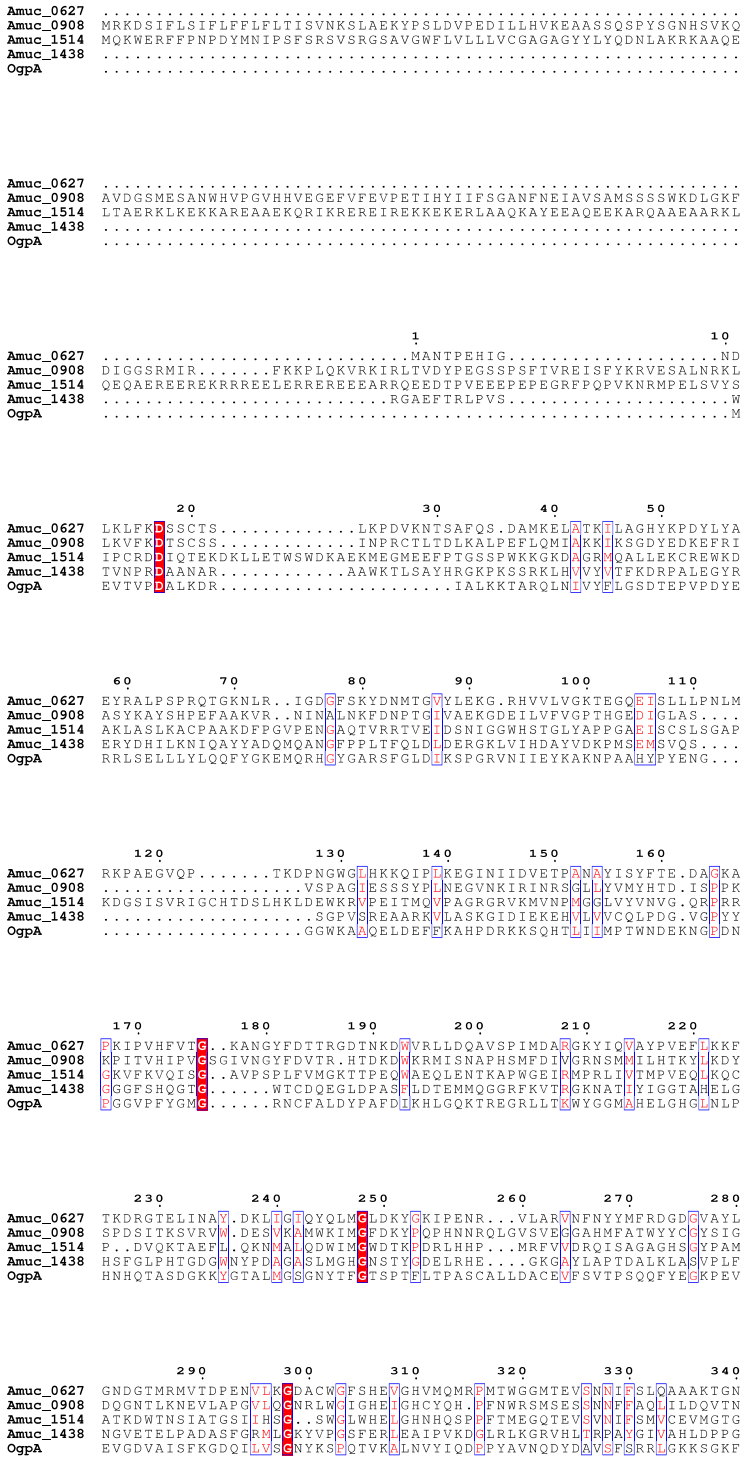
125. Shuoker, B. *et al.* Sialidases and fucosidases of Akkermansia muciniphila are crucial for growth on mucin and nutrient sharing with mucus-associated gut bacteria. *Nat Commun* **14**, 1833 (2023).
126. Afzal, M., Shafeeq, S., Ahmed, H. & Kuipers, O. P. Sialic acid-mediated gene expression in Streptococcus pneumoniae and role of nanr as a transcriptional activator of the nan gene cluster. *Appl Environ Microbiol* **81**, 3121–3131 (2015).
127. Ng, K. M. *et al.* Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. *Nature* **502**, 96–99 (2013).
128. Walters, D. M., Stirewalt, V. L. & Melville, S. B. Cloning, sequence, and transcriptional regulation of the operon encoding a putative N-acetylmannosamine-6-phosphate epimerase (nanE) and sialic acid lyase (nanA) in Clostridium perfringens. *J Bacteriol* **181**, 4526–4532 (1999).
129. Shimizu, T. *et al.* Complete genome sequence of Clostridium perfringens, an anaerobic flesh-eater. *PNAS* **99**, 996–1001 (2002).
130. Park, M., Mitchell, W. J. & Rafii, F. Effect of trehalose and trehalose transport on the tolerance of clostridium perfringens to environmental stress in a wild type strain and its fluoroquinolone-resistant mutant. *Int J Microbiol* **2016**, (2016).
131. Newstead, S. L. *et al.* The structure of Clostridium perfringens NanI sialidase and its catalytic intermediates. *Journal of Biological Chemistry* **283**, 9080–9088 (2008).
132. Boraston, A. B., Ficko-Blean, E. & Healey, M. Carbohydrate recognition by a large sialidase toxin from Clostridium perfringens. *Biochemistry* **46**, 11352–11360 (2007).
133. Boraston, A. B., Ficko-Blean, E. & Healey, M. Carbohydrate recognition by a large sialidase toxin from Clostridium perfringens. *Biochemistry* **46**, 11352–11360 (2007).
134. Wang, Y. H. Sialidases From Clostridium perfringens and Their Inhibitors. *Front Cell Infect Microbiol* **9**, 1–11 (2020).
135. Li, J. & Mcclane, B. A. NanI Sialidase Can Support the Growth and Survival of Clostridium perfringens Strain F4969 in the Presence of Sialyated Host Macromolecules (Mucin) or Caco-2 Cells. *Infect Immun* **86**, (2018).
136. Rivera, S., Khrestchatisky, M., Kaczmarek, L., Rosenberg, G. A. & Jaworski, D. M. Metzincin proteases and their inhibitors: Foes or friends in nervous system physiology? *Journal of Neuroscience* **30**, 15337–15357 (2010).
137. Antalis, T. M., Shea-Donohue, T., Vogel, S. N., Sears, C. & Fasano, A. Mechanisms of disease: Protease functions in intestinal mucosal pathobiology. *Nature Clinical Practice Gastroenterology and Hepatology* vol. 4 393–402 Preprint at https://doi.org/10.1038/ncpgasthep0846 (2007).
138. Van Der Velden, V. H. J. & Hulsmann, A. R. Peptidases: structure, function and modulation of peptide-mediated effects in the human lung. *Clinical and Experimental Allergy* **29**, 445–456 (1999).
139. Oda, K. New families of carboxyl peptidases: Serine-carboxyl peptidases and glutamic peptidases. *Journal of Biochemistry* vol. 151 13–25 Preprint at https://doi.org/10.1093/jb/mvr129 (2012).
140. Menach, E., Hashida, Y., Yasukawa, K. & Inouye, K. Effects of conversion of the zinc-binding motif sequence of thermolysin, HEXXH, to that of dipeptidyl

peptidase III, HEXXXH, on the activity and stability of thermolysin. *Biosci Biotechnol Biochem* **77**, 1901–1906 (2013).
141. Cerdà-Costa, N. & Gomis-Rüth, F. X. Architecture and function of metallopeptidase catalytic domains. *Protein Science* vol. 23 123–144 Preprint at https://doi.org/10.1002/pro.2400 (2014).
142. Laronha, H. & Caldeira, J. Structure and Function of Human matrix Metalloproteinases. *Cells* **9**, 1076 (2020).
143. Xu, Y. *et al.* Genetic variants in the metzincin metallopeptidase family genes predict melanoma survival. *Mol Carcinog* **57**, 22–31 (2018).
144. Noach, I. & Boraston, A. B. Structural evidence for a proline-specific glycopeptide recognition domain in an O-glycopeptidase. *Glycobiology* **31**, 385–390 (2021).
145. 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).
146. Pluvinage, B. *et al.* Architecturally complex O-glycopeptidases are customized for mucin recognition and hydrolysis. doi:10.1073/pnas.2019220118/-/DCSupplemental.
147. Shon, D. J. *et al.* An enzymatic toolkit for selective proteolysis, detection, and visualization of mucin-domain glycoproteins. *PNAS* **117**, 21299–21307 (2020).
148. 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**, 1–14 (2020).
149. Shon, D. J., Kuo, A., Ferracane, M. J. & Malaker, S. A. Classification, structural biology, and applications of mucin domain-targeting proteases. *Biochemical Journal* vol. 478 1585–1603 Preprint at https://doi.org/10.1042/BCJ20200607 (2021).
150. Ribet, D. & Cossart, P. How bacterial pathogens colonize their hosts and invade deeper tissues. *Microbes and Infection* vol. 17 173–183 Preprint at https://doi.org/10.1016/j.micinf.2015.01.004 (2015).
151. Niilo, L. Clostridium perfringens in Animal Disease: A Review of Current Knowledge. *The Canadian Veterinary Journal* **21**, 141–148 (1980).
152. Keyburn, A. L., Bannam, T. L., Moore, R. J. & Rood, J. I. NetB, a Pore-Forming Toxin from Necrotic Enteritis Strains of Clostridium Perfringens. *Toxins* vol. 2 1913–1927 Preprint at https://doi.org/10.3390/toxins2071913 (2010).
153. Lee, Y. *et al.* Crystal structure of the catalytic domain of Clostridium perfringens neuraminidase in complex with a non-carbohydrate-based inhibitor, 2-(cyclohexylamino)ethanesulfonic acid. *Biochem Biophys Res Commun* **486**, 470–475 (2017).
154. Bule, P. *et al.* Inverting family GH156 sialidases define an unusual catalytic motif for glycosidase action. *Nat Commun* **10**, (2019).
155. Nardy, A. F. F. R., Freire-de-Lima, C. G., Pérez, A. R. & Morrot, A. Role of Trypanosoma cruzi Trans-sialidase on the escape from host immune surveillance. *Front Microbiol* **7**, (2016).
156. Baos, S. C., Phillips, D. B., Wildling, L., McMaster, T. J. & Berry, M. Distribution of sialic acids on mucins and gels: A defense mechanism. *Biophys J* **102**, 176–184 (2012).

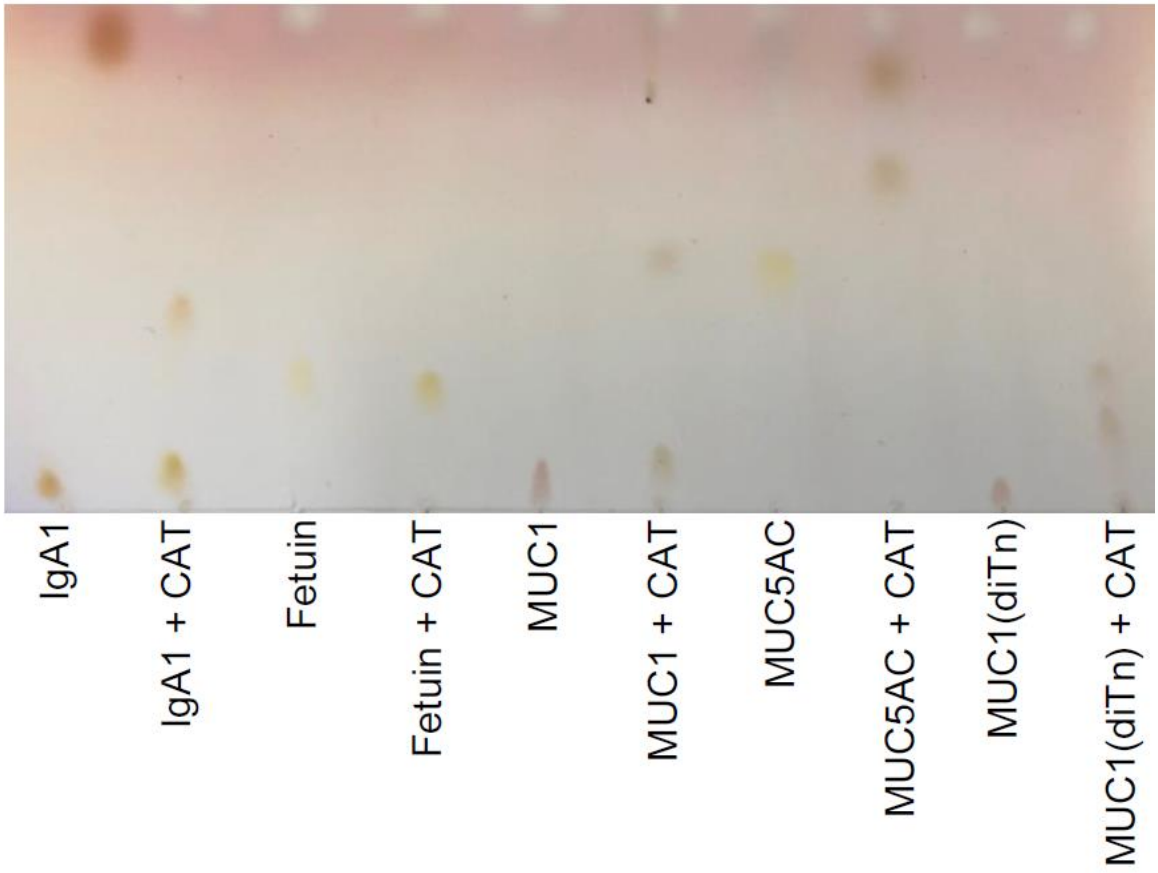
157. Li, J. & McClane, B. A. The sialidases of Clostridium perfringens type D strain CN3718 differ in their properties and sensitivities to inhibitors. *Appl Environ Microbiol* **80**, 1701–1709 (2014).
158. Hudson, K. L. *et al.* Carbohydrate-Aromatic Interactions in Proteins. *J Am Chem Soc* **137**, 15152–15160 (2015).
159. Spiwok, V. CH/π interactions in carbohydrate recognition. *Molecules* vol. 22 Preprint at https://doi.org/10.3390/molecules22071038 (2017).
160. Guo, J. *et al.* Identification and functional characterization of intracellular sialidase NeuA3 from Streptomyces avermitilis. *Process Biochemistry* **50**, 752–758 (2015).
161. Yao, Y. *et al.* Mucus sialylation determines intestinal host-commensal homeostasis. *Cell* **185**, 1172-1188.e28 (2022).
162. Owen, C. D. *et al.* Unravelling the specificity and mechanism of sialic acid recognition by the gut symbiont Ruminococcus gnavus. *Nat Commun* **8**, (2017).
163. Therit, B., Cheung, J. K., Rood, J. I. & Melville, S. B. NanR, a transcriptional regulator that binds to the promoters of genes involved in sialic acid metabolism in the anaerobic pathogen clostridium perfringens. *PLoS One* **10**, 1DIMMY (2015).
164. Cohen, M. & Varki, A. The sialome-far more than the sum of its parts. *OMICS A Journal of Integrative Biology* vol. 14 455–464 Preprint at https://doi.org/10.1089/omi.2009.0148 (2010).
165. Pluvinage, B. *et al.* Architecturally complex O-glycopeptidases are customized for mucin recognition and hydrolysis. *PNAS* (2021) doi:10.1073/pnas.2019220118/-/DCSupplemental.
166. Medley, B. J. *et al.* A previously uncharacterized O-glycopeptidase from Akkermansia muciniphila requires the Tn-antigen for cleavage of the peptide bond. *Journal of Biological Chemistry* **298**, (2022).
167. Paysan-Lafosse T *et al.* InterPro in 2022. *Nucleic Acids Res* (2022).
168. 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**, 7278–7287 (2019).
169. Ramírez-Vélez, R., García-Hermoso, A., Hackney, A. C. & Izquierdo, M. Effects of exercise training on Fetuin-a in obese, type 2 diabetes and cardiovascular disease in adults and elderly: A systematic review and Meta-analysis. *Lipids in Health and Disease* vol. 18 Preprint at https://doi.org/10.1186/s12944-019-0962-2 (2019).
170. von Mensdorff-Pouilly, S., Snijdewint, F. G. M., Verheijen, R. H. M. & Kenemans, P. Human MUC1 mucin and cancer immunotherapy. *Int J Biol Markers* **15**, 343–356 (2000).
171. Ding, L., Chen, X., Cheng, H., Zhang, T. & Li, Z. Advances in IgA glycosylation and its correlation with diseases. *Frontiers in Chemistry* vol. 10 Preprint at https://doi.org/10.3389/fchem.2022.974854 (2022).
172. Magalhães, A. *et al.* Muc5ac gastric mucin glycosylation is shaped by FUT2 activity and functionally impacts Helicobacter pylori binding. *Sci Rep* **6**, (2016).
173. Cummings, R. T. *et al.* A peptide-based fluorescence resonance energy transfer assay for Bacillus anthracis lethal factor protease. *Proc Natl Acad Sci U S A* **99**, 6603–6606 (2002).

174. Nakjang, S., Ndeh, D. A., Wipat, A., Bolam, D. N. & Hirt, R. P. A novel extracellular metallopeptidase domain shared by animal Host-Associated mutualistic and pathogenic microbes. *PLoS One* **7**, (2012).
175. Gregg, K. J., Finn, R., Abbott, D. W. & Boraston, A. B. Divergent modes of glycan recognition by a new family of carbohydrate-binding modules. *Journal of Biological Chemistry* **283**, 12604–12613 (2008).
176. Shoseyov, O., Shani, Z. & Levy, I. Carbohydrate Binding Modules: Biochemical Properties and Novel Applications. *Microbiology and Molecular Biology Reviews* **70**, 283–295 (2006).
177. Shon, D. J., Fernandez, D., Riley, N. M., Ferracane, M. J. & Bertozzi, C. R. Structure-guided mutagenesis of a mucin-selective metalloprotease from Akkermansia muciniphila alters substrate preferences. *Journal of Biological Chemistry* **298**, (2022).
178. Bar-Even, A. *et al.* The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402–4410 (2011).
179. Raimondi, S., Musmeci, E., Candeliere, F., Amaretti, A. & Rossi, M. Identification of mucin degraders of the human gut microbiota. *Sci Rep* **11**, (2021).
180. Raba, G. & Luis, A. S. Mucin utilization by gut microbiota: recent advances on characterization of key enzymes. *Essays Biochem* (2023) doi:10.1042/ebc20220121.
181. Robbe, C. *et al.* Evidence of regio-specific glycosylation in human intestinal mucins: Presence of an acidic gradient along the intestinal tract. *Journal of Biological Chemistry* **278**, 46337–46348 (2003).
182. Crouch, L. I. *et al.* Prominent members of the human gut microbiota express endo-acting O-glycanases to initiate mucin breakdown. *Nat Commun* **11**, (2020).
183. Deplancke, B. *et al. Selective growth of mucolytic bacteria including Clostridium perfringens in a neonatal piglet model of total parenteral nutrition*. *Am J Clin Nutr* vol. 76 www.ncbi.nlm.nih.gov/BLAST/ (2002).
184. Ashida, H. *et al.* A Novel Endo-β-galactosidase from Clostridium perfringens that Liberates the Disaccharide GlcNAcα1→4Gal from Glycans Specifically Expressed in the Gastric Gland Mucous Cell-type Mucin. *Journal of Biological Chemistry* **276**, 28226–28232 (2001).
185. 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).
186. Crouch, L. I. *et al.* Prominent members of the human gut microbiota express endo-acting O-glycanases to initiate mucin breakdown. *Nat Commun* **11**, (2020).
187. Berkhout, M. D., Plugge, C. M. & Belzer, C. How microbial glycosyl hydrolase activity in the gut mucosa initiates microbial cross-feeding. *Glycobiology* **32**, 182–200 (2022).
188. Foley, M. H., Cockburn, D. W. & Koropatkin, N. M. The Sus operon: a model system for starch uptake by the human gut Bacteroidetes. *Cellular and Molecular Life Sciences* vol. 73 2603–2617 Preprint at https://doi.org/10.1007/s00018-016-2242-x (2016).

189. Almagro-Moreno, S. & Boyd, E. F. Insights into the evolution of sialic acid catabolism among bacteria. *BMC Evol Biol* **9**, (2009).
190. McCoy, A. J. *et al.* Phaser crystallographic software. *J Appl Crystallogr* **40**, 658–674 (2007).
191. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* **67**, 355–367 (2011).
192. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486–501 (2010).
193. Brunger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **335**, (1992).
194. Davis, I. W. *et al.* MolProbity: All-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* **35**, (2007).
195. Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. *Automated Structure Solution With autoSHARP*. *From: Methods in Molecular Biology* vol. 364 http://www.globalphasing.com/sharp/. (2007).
196. Cowtan, K. Recent developments in classical density modification. *Acta Crystallogr D Biol Crystallogr* **66**, 470–478 (2010).
197. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr* **62**, 1002–1011 (2006).



Supplemental Figure 1. Sequence alignment of the known glycopeptidases from Akkermansia muciniphila.



Supplemental Figure 2. TLC image of glycopeptide digest. The representation seen in figure 15 is based on the TLC of CAT Amuc_1438 digesting Tn-antigen-containing glycopeptides.