Chemical Biology of Protein *O*-Glycosylation

by

Patrick Chaffey

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Chemistry and Biochemistry

2016

This thesis entitled:

Chemical Biology of Protein O-Glycosylation

written by Patrick Chaffey

has been approved for the Department of Chemistry and Biochemistry

____

Zhongping Tan, Ph.D., Committee Chair

____

Wei Zhang, Ph. D., Committee Member

Date________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Chaffey, Patrick (Ph.D., Chemistry and Biochemistry)

Chemical Biology of Protein O-Glycosylation

Thesis directed by Assistant Professor Zhongping Tan

Protein glycosylation, the covalent attachment of carbohydrates to amino acid side chains of proteins, is a ubiquitous post-translational modification across all branches of life. Due to many factors including the vast structural complexity of glycans and the convoluted processes regulating their construction, protein glycosylation is a significantly understudied phenomenon. In particular, the study of protein *O*-glycosylation, where the carbohydrate moieties are attached via the oxygen of a serine or threonine residue, is lacking because there exists no well-defined consensus sequence for its occurrence and the enzymatic construction of *O*-glycosylated proteins in a controlled manner is very difficult. We employed chemical synthesis for the construction of homogeneous and well-defined *O*-glycoproteins with a large variety of structures and used these synthetic biomolecules to systematically and quantitatively investigate the effects of glycosylation on the biophysical and biological properties of proteins. Our ultimate goal is to develop a set of principles that can be widely applied to the rational engineering of enzymes and therapeutic proteins through glycosylation and other post-translational modifications. The initial focus was on examining the effects of *O*-glycosylation on the properties of a carbohydrate binding module (CBM) of an industrially important fungal cellulase. We used a wide range of biochemical assays to characterize a library of 51 differently-glycosylated CBM isoforms, and observed strong effects of glycosylation, in a pattern specific manner, on the folding, stability, solubility, chromatographic behavior, binding affinity and specificity of this small domain. In the long term, this project is expected to lead to fungal cellulases with optimal stability, specificity, activity required to achieve efficient saccharification of biomass for biofuels production. We then expanded our methodology to investigate the influence of *O*-glycosylation on two important therapeutic peptides: insulin and glucagon-like peptide-1 (GLP-1). As with the CBM system, we observed that glycosylation can significantly impact physical and/or functional properties of these molecules. We have identified specific isoforms of both insulin and GLP-1 that have

increased stability and unchanged biological functions. It is our hope that further development of the most promising lead candidates for insulin and GLP-1 could lead to better therapies for the treatment of metabolic disorders.

Contents

Chapter 1

Introduction: Chemical Biology of Protein *O*-Glycosylation

1.1 Introduction

Glycosylation is an extremely common and complex post-translational modification of many biomolecules. While this chapter will focus on the glycosylation of proteins, other types of biomolecules are known to carry glycans as well, most notably lipids. Protein glycosylation is commonly divided into three different categories based on the atom linking the glycan to the protein: *N*-type, *O*-type and the somewhat rare *C*-type. For protein *N*-glycosylation, the carbohydrate chain is attached to the side chain of an asparagine (Asn) residue through the nitrogen atom of the terminal amide. *O*-glycosylation, the subject of this thesis, is the term for glycans attached to the side chain of a serine (Ser) or threonine (Thr) residue via the oxygen of the hydroxyl group. In *C*-glycosylation, the anomeric carbon of a sugar residue is attached directly to an aromatic carbon on the side chain of a tryptophan (Trp) residue through a carbon-carbon bond. Naturally occurring *C*-glycosylation always involves the sugar mannose (Man) and has been observed almost exclusively within thrombospondin repeat domains.[1] *N*-glycosylation, while present on a much wider variety of proteins, has a similarly restrictive consensus sequence for its incorporation into a protein; and it occurs at the Asn in the consensus sequence Asn-X-Ser/Thr where X is any amino acid except proline.[2] The fact that these types of glycosylation follow such easily identifiable, and thus controllable, patterns has greatly facilitated their study in biological systems. Unlike *N*- and *C*-glycosylation, however, most *O*-glycosylation sites have no obvious consensus sequence for their occurrence. This fact has hindered many traditional approaches to studying the phenomenon based on molecular biology, such as mutating in or out glycosylation sites in proteins of interest.

Nevertheless, a major goal of glycoscience is to understand the structure-function and the composition-function relationships of protein glycosylation (Figure 1.1). A better understanding of these relationships will lead to the development of better industrial and therapeutic proteins, and better diagnostic tools for human disease. While the complex nature of protein glycosylation makes such studies difficult, recent years have seen many advances from an interdisciplinary approach to the problem. New technologies and the chemical biology mindset have inspired studies that combine chemical synthesis of complex biomolecules and biologically relevant assays to reveal robust and important discoveries. The field of proteomics, or when applied to the study of glycoproteins: glycoproteomics, has led to extremely useful tools for characterizing naturally occurring glycosylation and is also helping to advance our understanding in this area.

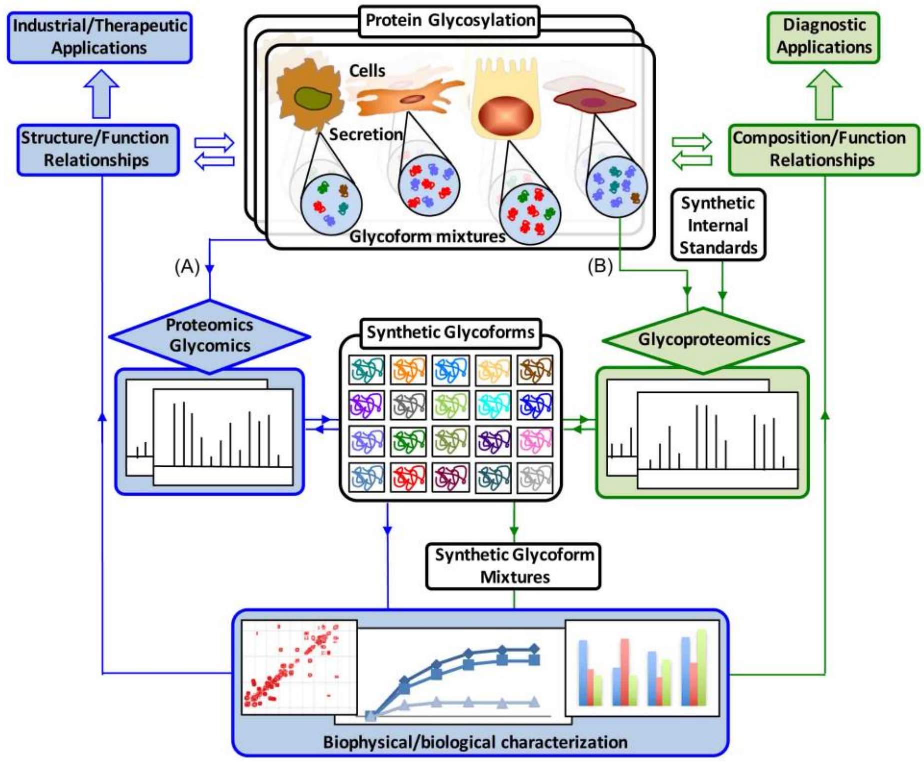


Figure 1.1 - The overall strategy of chemical glycobiology research. (A) Identification of the naturally occurring glycans and glycosylation sites by proteomics and glycomics will inform the construction of a library of synthetic glycoforms, which will be individually characterized to determine the correlations between glycosylation patterns and functional properties of glycoproteins. (B) Synthetic glycoforms and internal glycopeptide standards will be used to enable the quantitative analysis of the composition of naturally-secreted mixtures of glycoforms under different conditions. Assays of re-constructed mixtures will lead to correlations between glycoform composition and mixture function. This type of knowledge is expected to greatly advance the understanding of protein glycosylation and promote its applications in the areas of enzyme and therapeutic protein engineering and disease diagnosis.

Protein *O*-glycosylation is a very complex type of post-translational modification (PTM) that covers a wide variety of structures.[3,4] It is usually divided into several different categories based on the identity of initial carbohydrate residue attached to the protein. In mammals, there are six commonly acknowledged categories of *O*-glycosylation, each started by the following seven monosaccharides: *N*-acetylglucosamine (GalNAc), mannose (Man), fucose (Fuc), glucose (Glc), *N*-acetylglucosamine (GlcNAc), galactose (Gal) and xylose (Xyl) (Figure 1.2). By far the most common *O*-glycan is the *O*-GalNAc-, or mucin-, type glycan. This structure is extremely common on secreted proteins of many varieties and on most extracellular, membrane-bound proteins. *O*-Man was thought for many years to only occur on the protein α-dystroglycan, which is critical for cell adhesion, but has more recently become appreciated as the second most common type of *O*-glycan in the nervous system. *O*-Fuc glycans occur on many different extracellular proteins, but always in the context of specific peptide domains: epidermal growth factor-like repeats (EGF repeats) and thrombospondin-like repeats (TSRs). These domains are commonly involved in cellular communication. *O*-Glc glycans also appear on EGF repeats on a variety of membrane proteins. *O*-GlcNAc is most well known as a cytoplasmic modification of proteins somewhat akin to phosphorylation. Since its initial discovery, this type of glycosylation has become appreciated as an important aspect of signal transduction within the cell. Since this work focuses mostly on extracellular glycosylation, this unique form of glycosylation will not be discussed in detail here. Recently, however, the *O*-GlcNAc modification was identified on EGF repeats of several important proteins. This extracellular *O*-GlcNAc appears to be controlled independently of the intracellular variety and is relevant to this work. *O*-Gal glycans occur mostly on collagen and are attached to the special, non-proteogenic amino acids hydroxylysine and hydroxyproline. Finally, *O*-Xyl glycans include both heparan sulfate and chondroitin sulfate, which are important members of the proteoglycan polysaccharide family.[2]

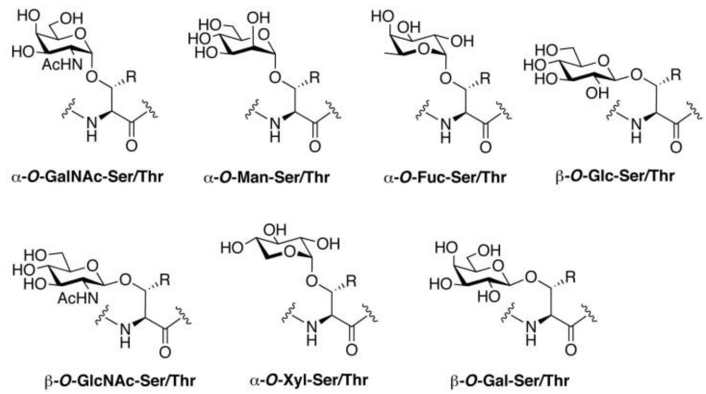


Figure 1.2 - The seven types of *O*-glycosylation in mammals. R = H/Me.

These long and unbranched, poly-anionic carbohydrate chains are critical for cell adhesion and cellular communication events. Proteoglycans are not the focus of our work and will not be covered.

This chapter will review the current literature discussing the 5 most important types of mammalian *O*-glycans, including α-*O*-GalNAc, α-*O*-Man, α-*O*-Fuc, β-*O*-Glc, and β-*O*-GlcNAc, which are naturally found outside the cell on secreted and membrane-bound proteins. The focus will be on their biosynthesis pathways, what is known of their *in vivo* functions, and how chemical biology has helped to advance the understanding of this important post-translational protein modification.

1.2 Biosynthesis of *O*-Glycoproteins

Protein *O*-glycosylation is an enzyme-catalyzed process that occurs in the endoplasmic reticulum (ER) and Golgi compartments (Figure 1.3). Each type of *O*-glycosylation has a unique biosynthetic pathway and uses a different range of enzymes and substrates.

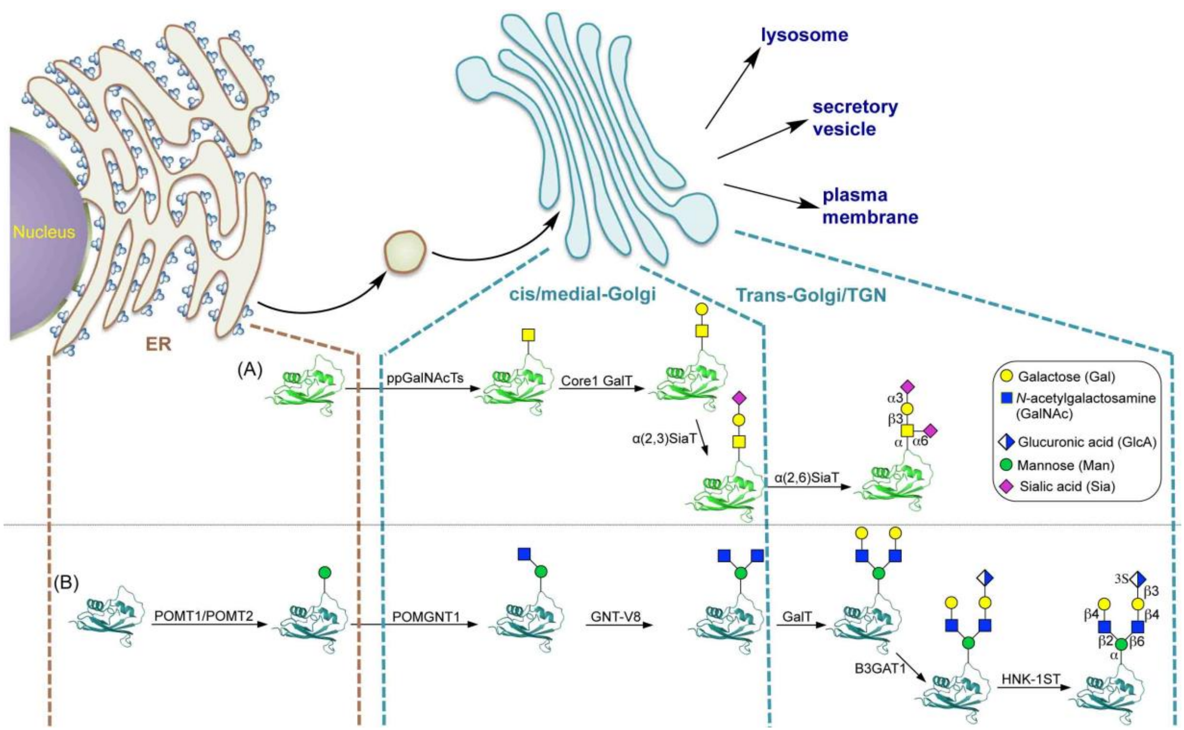


Figure 1.3 - Biosynthesis of representative (A) GalNAc-type and (B) Man-type *O*-glycoproteins in ER and Golgi apparatus.

1.2.1 α-*O*-GalNAc

GalNAc is added to Ser or Thr residues in an α-linkage by a large family of over 20 different polypeptide *N*-acetylgalactosamine transferases (ppGalNAcTs). Unlike most glycosylation, which happens in the ER, GalNAc *O*-glycosylation occurs in the Golgi after protein folding.[5] Each ppGalNAcT has two domains: a catalytic glycosyltransferase domain and ricin-homology lectin domain. This lectin domain has been implicated in governing the selectivity and activity of several members of the family.[6,7] Unlike *N*-glycosylation and several other types of *O*-glycosylation, to date no consensus sequence has been identified for the addition of *O*-GalNAc; it has even been suggested that no such sequence exists and ppGalNAcTs recognize and modify protein secondary structures, like the β-turn, instead of specific primary sequences.[8] What is known, as demonstrated by several studies discussed in greater detail later on, is that each ppGalNAcT homolog has a slightly different activity profile for peptide and/or glycopeptide substrates. Additionally, the expression of each homolog varies spatially across tissues and temporally during development.[5,9] Such a high level regulation during cell differentiation and organismal development suggests critical roles for these enzymes and the *O*-GalNAc modification they impart. Interestingly, the deletion of the ppGalNAcT-1, -T-4, -T-5, or -T-13 genes in mice fails to result in any observable phenotype.[10] This might suggest a high level of redundancy and overlap between the functions of each family member. However this situation does not hold for all

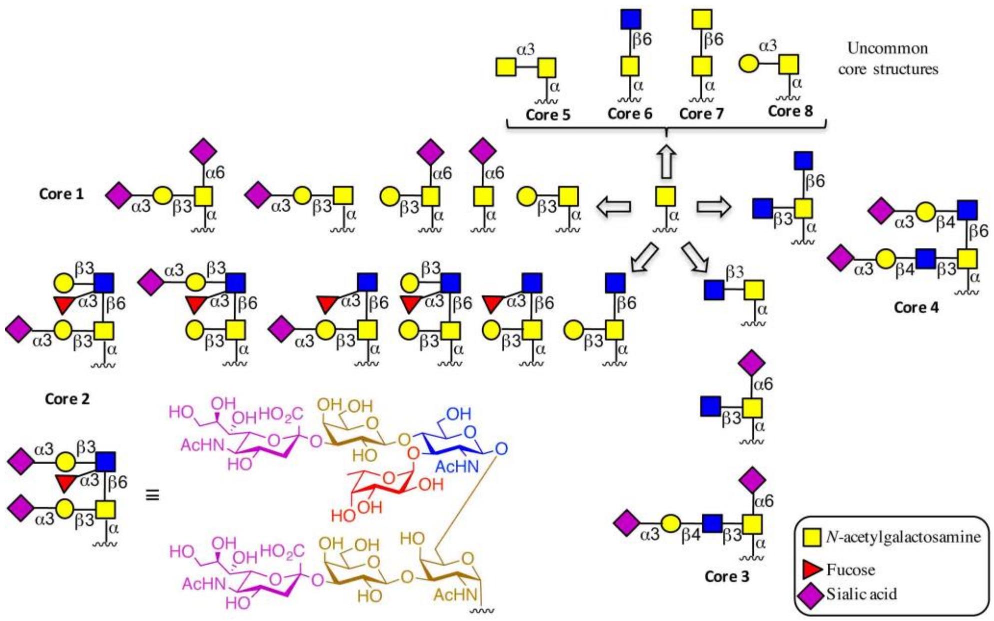


Figure 1.4 - Core structures and their elaboration as seen in mucin-type *O*-glycans.

systems studied as the deletion of any one of 5 different *Drosophila melanogaster* *O*-GalNAc transferases is lethal,[9] and inactivating mutations of ppGalNAcT-3 in humans cause the rare autosomal recessive metabolic disorder familial tumoral calcinosis.[11,12] Clearly, in certain cellular contexts, a single *O*-GalNAc transferase can have unique and critical functions. After the initial *O*-GalNAc transfer, further elaboration is accomplished in the Golgi apparatus by a suite of up to 30 glycosyltransferases to yield a huge variety of mature "mucin-type" *O*-GalNAc glycans (see Figure 1.4).

1.2.2 α-*O*-Man

It is now recognized that *O*-Man glycans, like *O*-GalNAc glycans, can have a variety of branching architectures and end groups. It has been proposed that the naturally occurring mammalian *O*-Man glycans can be divided into three groups, each having a different core structure, similar to how mucin type glycans are classified (see Figure 1.5). All *O*-Man glycans are initiated by addition of mannose by a pair of mammalian protein *O*-mannosyl-transferases: POMT1 and POMT2, which appear to act as a functional heterodimer *in vivo*.[13] These enzymes have different expression patterns across tissues in humans, and different tissue specific isoforms formed through alternative gene splicing.[14,15]

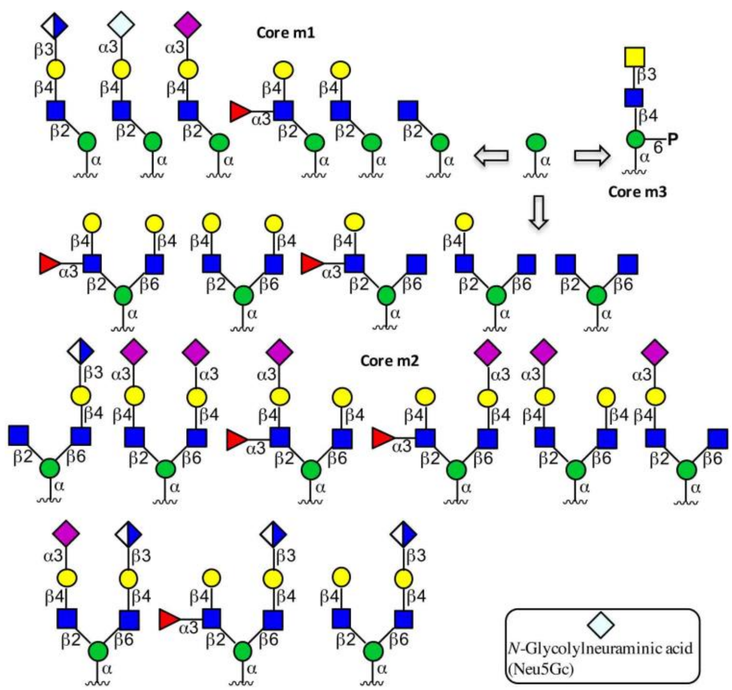


Figure 1.5 - *O*-Mannose glycans found in mammals.

Like most *O*-glycan-transferases, both POMT1 and POMT2 are found in the ER, but interestingly they utilize a unique sugar donor: dolichol phosphate mannose (Dol-P-Man), unlike other glycosyltransferases which use nucleotide sugars.[16] After the attachment of the first mannosyl residue, glycoproteins are transported to the Golgi apparatus. Both Core M1 and Core M2 structures are then extended by addition of GlcNAc through a β1-2 linkage by the UDP-GlcNAc protein *O*-mannose β1,2-N-acetylglucosaminyltransferase (POMGNT1).[17] POMGNT1 acts in the Golgi, and appears to be completely independent of the enzymes that catalyze formation of the GlcNAcβ1-2Man linkages in *N*-glycans, since those two enzymes (GnT-I and GnT-II) show no reactivity towards *O*-mannosyl peptides *in vitro*.[18] Core M1 structures are then further elongated by Golgi-resident glycosyltransferases including galactosyltransferases (GALTs), sialyltransferases (SIATs), glucuronyltransferases (GLCATs), sulfotransferase HNK-1ST and α1,3-fucosyltransferase 9 (FUT9) to give a variety of end structures, but detailed mechanisms for how these steps are regulated are not well-understood.[16] Core M2 glycans, thought to exist only in the brain, have a branched trisaccharide core: GlcNAcβ1–6(GlcNAcβ1–2)Man-Ser/Thr. After synthesis of the Core M1 dimer, a branching GlcNAcβ1-6 can be added to form Core M2 structures. This is done by the brain-specific *N*-acetylglucosaminyltransferase IX (GnT-IX) enzyme,[19] which is also involved in synthesis of highly branched complex-type *N*-glycans.[20]

Core M3 is a uniquely branched phosphodiester-containing trisaccharide: GalNAcβ1-3GlcNAcβ1-4(phosphate-6)Man-Ser/Thr. The phosphoglycan core structure was first identified and structurally characterized in 2010.[21] Unlike other complex *O*-glycans, this core structure is entirely assembled in the ER.[22] So far, this structure has only been found on a few specific sites on α-dystroglycan, a critical component of the dystroglycan complex that links the cytoskeleton of muscle cells to the extracellular matrix.[16] After initial mannosyltransfer to the peptide, glycosyltransferase-like domain-containing 2 (GTDC2/POMGNT2) adds GlcNAc, β1-3*N*-acetylgalactosaminyltransferase 2 (B3GALNT2) adds GalNAc and the final trisaccharide is phosphorylated by protein kinase-like protein SGK196/POMK to produce the core M3 structure.[22] Notably, POMK is missing key catalytic residues found in other known kinases, and it has been suggested that it could be the first of a new class of kinases with a potentially novel catalytic mechanism.[23] After phosphorylation, the glycan is extended by a polymer of repeating [-3-Xylα1-3-GlcAβ-1-] units synthesized by the enzyme LARGE.[24] This polymeric oligosaccharide has many similarities to glycosaminoglycans (GAGs), like heparin, and has even been proposed to be better thought of as a member of that glycan class.[25] However, LARGE cannot modify the phosphor-trisaccharide product of POMK. Recently, B4GAT1 was found to add a glucuronic acid to xylose (most likely a β-Xyl) and the resulting dimer is capable of serving as the "primer" for (-GlcA-β-1,3-Xyl-α-1,3-) polymer synthesis by LARGE.[26] It is not yet known what the xylose that B4GAT1 modifies is attached to and so the exact details of how the LARGE glycan is attached to the Core M3 structure are not known. It has been confirmed that most, if not all, post-phosphorylation modification of Core M3 *O*-Man glycans occurs in the Golgi, and that neither sialic acid, galactose or fucose are involved in those transformations.[27] LARGE2, a paralog of LARGE, appears to have identical enzymatic activity, but a different optimal pH range and different tissue-expression patterns.[28,29] The significance of LARGE2 is not yet known. Further complicating the picture, sulfotransferase HNK-1ST has been shown to add a sulfate group to the post-phosphoryl carbohydrate moiety and block the activity of LARGE,[30] a probable negative regulatory mechanism for controlling the length of LARGE glycan polymer.

1.2.3 α-*O*-Fuc

O-linked fucose appears on two distinct kinds of protein domain, both cysteine knots present in many different proteins: EGF repeats and thrombospondin-like repeats (TSRs). Each one has its own protein *O*-fucosyl transferase: POFUT1 for EGF repeats and POFUT2 for TSRs.[31] Like most other protein *O*-glycosyltransferases, POFUT1 and POFUT2 are localized to the ER.[32,33] After initial protein *O*-fucosylation, each domain then follows a unique extension pattern. Interestingly, there appears to be no cross-talk between the two pathways, enzymes that build the *O*-Fuc glycans on TSRs have no activity towards EGF repeats and *vice versa*.[31] POFUT1 adds fucose to EGF repeats with the consensus sequence $C^2X_{4\text{-}5}(S/T)C^3$, where C^2 and C^3 are the second and third cysteines in the conserved EGF repeat sequence.[34] Under certain conditions, the initial fucose can be elongated with a β-GlcNAc at the 3-OH of fucose by the Fringe family of enzymes.[35] In mammals, Gal can be added to GlcNAcβ1-3Fuc by β4galactosyltransferase 1(β4GalT1),[36] but in flies no extension past the disaccharide has been observed.

Further elongation by sialic acid to a tetrasaccharide (Siaα2-3Galβ1-4GlcNAcβ1-3Fuc-α-*O*-Ser) is also seen in mammals.[37] The consensus sequence for fucosylation of TSRs is $C^1X_{2-3}(S/T)C^2X_2G$, where C^1 and C^2 are the first and second Cys of a TSR sequence.[38] TSR *O*-Fuc glycans can be extended by the β1,3-glucosyltransferase B3GLCT to a disaccharide in the ER.[39,40]

1.2.4 β-*O*-Glc

Like *O*-Fuc, *O*-linked glucose is most commonly found on EGF repeats. In *Drosophila*, the protein Rumi was identified as the protein *O*-glucosyltransferase responsible for the modification[41] and the mammalian ortholog, protein *O*-glucosyltransferase 1 (POGLUT1), has also been identified and confirmed to serve the same function.[42,43] These enzymes add glucose to the serine within the C^1XSXP/AC^2 consensus sequence.[37,44] On several blood factors in both cows and humans as well as on the Notch receptor, the glucose-initiated glycan structure was found to be extended to a trisaccharide Xyl-α1-3-Xyl-α1-3-Glc-β1-O-Ser.[45,46] In humans the first xylose is added by either glucoside xylosyltransferase 1 (GXYLT1) or GXYLT2 and the second is added through action of xyloside xylosyltransferase 1 (XXYLT1). All three of these glycosyltransferases are active in the ER.[47,48] In flies, there is only one GXYLT, also known as Shams,[49] and extension to the trisaccharide has not been identified. It is not currently known why humans have two functional glucoside xylosyltransferases.[50]

1.2.5 β-*O*-GlcNAc

Recent years have established internal β-*O*-GlcNAc as an important and wide-spread modification catalyzed by the ubiquitously expressed protein *O*-glucosaminyltransferase (OGT). Advances in understanding the biology of that unique form of glycosylation have been reviewed previously. Relevant to this work, however, is the discovery that β-*O*-GlcNAc also exists extracellularly on the EGF repeats of several proteins. The ER-resident glycosyltransferase for this glycosylation has been identified in flies,[51] mice,[52] and humans,[53] and is known as EGF-specific protein *O*-glucosaminyltransferase (EOGT). Unexpectedly, this enzyme was found to share almost no sequence similarity to OGT, and instead is most closely related to a xylosyltransferase from plants.[52] Characterization of the known carriers of this glycan has revealed the consensus sequence for EOGT activity to be $C^5X_2GX(S/T)GX_2C^6$, where C^5 and C^6 are the fifth and sixth cysteines of the EGF sequence.[54] This consensus sequence shows up in many proteins in both flies and mammalian genomes, although most of these potential sites have not been experimentally validated.

1.3 Chemical Biology in Studying the Structural and Functional Consequences of Protein O-Glycosylation

While the unique features of *O*-glycosylation make its study difficult, recently the power of modern chemical biology has made it increasingly possible to investigate this important post-translational modification. This is due in large part because chemical synthesis has advanced to the point where many homogenous glycoforms are available and now commonly used in studies of the phenomenon. Here, we will review the progress in using these new techniques and methods to study the chemical biology of protein *O*-glycosylation.

1.3.1 α-O-GalNAc

O-GalNAc glycosylation is the most common form of *O*-glycosylation in mammals, occurring in an estimated 10% of all mammalian proteins and half of all proteins passing though the secretory system.[55] The most well-known examples of proteins carrying this type of glycosylation are the mucins (Figure 1.6), and in fact *O*-GalNAc glycosylation is referred to as “mucin-type” glycosylation in the literature due to the strength of this association.[4,8,56] Mucins can be divided into two categories: secreted mucins that form extensive oligomers and result in a viscous, mucosal layer around tissues; and membrane-bound mucins that are monomeric and form a significant amount of the glycocalyx that surrounds and protects many cells.[57] The most relevant feature of mucin proteins is the so-called “mucin-domain” which consists of a large number of repeats of a mucin-subtype specific sequence.[58,59] These repeated sequences, often referred to as “tandem repeats”, are heavily glycosylated with densely-clusted *O*-GalNAc glycans.[8]

Additionally, there are a number of non-mucin proteins contain domains that are structurally very similar to the heavily glycosylated, tandem-repeat regions of mucins, and these are also referred to as "mucin domains".[58,60]

O-GalNAc glycans in these mucin domains are frequently acknowledged for their ability to prevent proteolytic degradation.[61,62] Early light microscopy data showed that mucin domains adopt a rigid, extended conformation in solution. A multitude of studies since then have established that this structure is a general feature of mucin domains and that this rod-like structure is dependent on the presence of *O*-GalNAc glycan clusters throughout the sequence.[63,64] Therefore, the function of *O*-GalNAc glycosylation in these domains has traditionally been thought of as many structural: producing a stable, extended structure on the cell surface for display of other, more functionally important protein domains, with little importance placed on the actual glycosylation pattern(s) present.[6] For example, glycans on extracellular proteins are known to inhibit oligomerization through what is likely the nonspecific charge-charge repulsion of negatively charged end groups.[65,66] The fact that the sequence of mucins and mucin-domains has been only lightly conserved during evolution has also been used as evidence for this assumption.[67]

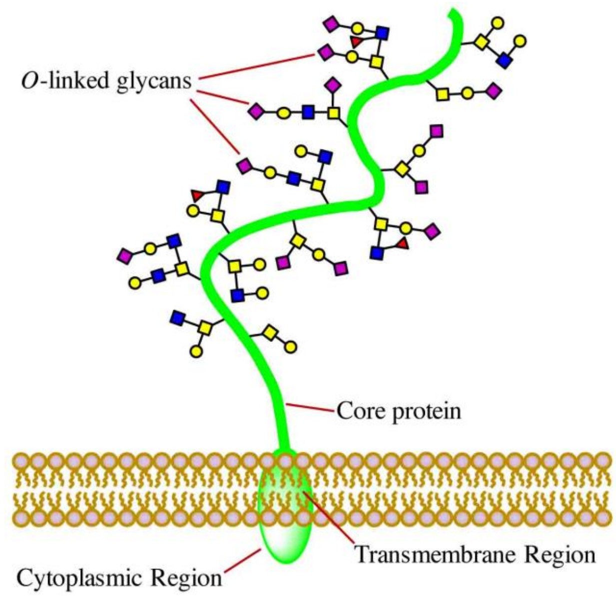


Figure 1.6 - Schematic representation of the mucin structure. Mucins are a diverse family of proteins that are heavily *O*-glycosylated. Their protein core domains are often rich in serine, theronine, proline, alanine and glycine residues and adopt an extended "bottle brush" conformation.

While there is undeniably an important structural aspect to mucin glycosylation, numerous examples have emerged of glycans acting in a much more specific manner, for example, it has been suggested that they are integral to molecular code used in everyday cellular communication and recognition.[68] As evidence of these more specific roles, the glycosylation patterns of many proteins appears to be tightly regulated during cellular maturation,[69] and both temporally and spatially during development.[70] Glycans patterns on mucins have also been shown to vary in a region-specific way in several different body systems.[57,71] The intestinal tract is a prime example, where different core structures are specific to each region and sialic acids appear in a well-defined gradient increasing from the ileum to the colon.[72,73] Mucin type glycosylation is also known to be important in governing selectivity in cell-cell communication and signaling in several different systems including T cell maturation,[74,75] Notch signaling,[76] leukocyte migration[77] and mucin signaling.[78] This type of glycosylation is critical to proper processing of many proteins from the proprotein form to maturity[79] and important in regulating cleavage by ADAM proteases and subsequent shedding of the extracellular domains of many diverse membrane proteins including Notch[76] and TNF-α.[80] Interestingly, this regulation has been observed as both positive and negative, so it is not a simple steric-interference mechanism and may involve specific binding interactions.[80] Cytokines, secreted soluble proteins that are involved in cell communication, are also known to be regulated through several types of glycosylation, including *O*-glycosylation with mucin type glycans.[81,82]

Chemical biology has clearly aided in the studies of *O*-GalNAc glycosylation. As discussed below, much of the early work on structural and functional effects of *O*-GalNAc glycosylation focused on small glycopeptides. The readily available and pure homogenous glycoforms made it possible to draw strong conclusions about the different substrate specificities of ppGalNAcTs and the ability of *O*-GalNAc glycans to affect the structural and different functional properties of proteins.

1.3.1.1 Substrate Preferences of ppGalNAcTs

One of the most difficult aspects of studying *O*-glycosylation as compared to *N*-glycosylation is the lack of a defined consensus motif for the former. This is despite the effort of many to characterize the sequence specificities of the large family of ppGalNAcTs over the years. Nevertheless, the ready availability of homogenous glycopeptide substrates that chemical synthesis allows for has made strides in this area possible in recent years. In a seminal work by the Bertozzi group, for example, the specificities of many individual ppGalNAcT enzymes were characterized using an impressively large synthetic glycopeptide library.[10] The glycopeptide sequence was taken from a fragment of rat submandibular mucin and contains six potential glycosylation sites, all threonines. The 56-member library based on this sequence was composed of every possible combination of mono-, di-, tri-, and tetraglycosylated peptide and all of the glycans were single GalNAc residues. Each member of the synthetic library was exposed to the eight pp-GalNAc-Ts studied and the activity of the transferase was subsequently quantified. The authors found distinct preferences for each of the transferases studied and were able to divide them into three groups: those that prefer unglycosylated or monoglycosylated substrates (ppGalNAcT-1, -T-2 and –T-5), those that prefer diglycosylated peptides (ppGalNAcT-3 and –T-4), and those that prefer substrates with three or more glycans (ppGalNAcT-10). Interestingly, ppGalNAcT-7 and –T-11 had unique and highly specific selectivities, suggesting more specialized roles for these two glycosyltransferases. From these results, the authors proposed a model of mucin glycosylation in the Golgi were the three groups of glycosyltransferases act in sequence to generate the highly clustered GalNAc glycans observed *in vivo*.[10]

It is well known that members of the ppGalNAcT family have two domains, a catalytic transferase domain and a carbohydrate-binding lectin domain, separated by a flexible linker.[5] In order to more explicitly investigate the influence of this lectin domain on the site-specificity of the ppGalNAcT family, the Gerken group used a series of randomized glycopeptides bearing a single GalNAc residue and a possible acceptor site between 6 and 16 amino acids away.[6] They then compared the ability of several ppGalNAcTs to act on these chemically synthesized substrates, and found distinct *N*- or *C*-terminal preferences for many of the enzymes. In particular, ppGalNAcT-1, -T-2, and –T-14 were found to prefer sites on the *N*-terminal side of pre-existing glycosylation and ppGalNAcT-3 and –T-6 preferred the opposite orientation. Other transferases studied displayed no obvious preference. They also found that the optimal distance for new glycosylation varied across the enzymes from 10 residues for ppGalNAcT-1 and –T-2 to 16 residues for ppGalNAcT-3, reflecting differences in the linker connecting the catalytic and lectin domains of the various family members.[6] Follow up work by the same group used a similar strategy to characterize the role of the catalytic domain in selectivity of the glycopeptide-preferring transferases. A set of random glycopeptides, each containing a single *O*-GalNAc-Thr residue and a possible glycosyltransfer acceptor site were tested as substrates for the ppGalNAcTs known to prefer glycopeptide substrates: ppGalNAcT-4, -T-7, -T-10 and –T-12 and the fly ppGalNAcT-7 ortholog PGANT7.[7] It was found that all transferases tested except ppGalNAcT-4 significantly prefer acceptor sites near previous sites of glycosylation, and in addition, each transferase was highly specific in which site was modified. The authors propose that this is indicative of GalNAc binding sites within the catalytic domain of these transferases, and the binding of these sites is much more important for catalytic activity than comparable sites found in the lectin domains of the same enzymes. With their previous work, the authors propose that three distinct kinds of binding are used to various degrees by each member of the ppGalNAcT family: distal glycan recognition by the lectin domain, neighboring glycan recognition by the catalytic domain, and peptide sequence recognition by the catalytic domain. Together these three binding modes allow for glycosyltransfer to a diverse range of substrates with varying degrees of glycosylation and could explain the unusually large number of isoforms for this family.[7]

1.3.1.2 Structural Effects of O-GalNAc Glycosylation

Due to the repetitive sequence in mucin domain with the presence of only a few different amino acids, it is often very difficult to obtain high quality information from NMR about the structure of the whole domain.[83] Therefore, most structural studies have been carried out on small glycopeptides. Since the extended structure of mucins ensures minimal interaction between regions of the glycopeptide more than a few amino acid residues apart, it is thought that even a very small model system could be relevant to much larger constructs.[64] This assumption has been repeated shown to be valid in heavily glycosylated systems.

Glycophorin A, which is an important glycoprotein found on erythrocytes that carries the epitope determining the MN blood type, is one of the earliest model systems used to study the structure of O-glycosylated peptides. It was recognized early on that the immunogenic nature of this glycoprotein was due to a small region of the *N*-terminus that is heavily glycosylated.[84] Since it was of interest how the glycosylation affected the structure, and thus immunogenicity of the peptide, it was commonly studied as a model system for heavily glycosylated peptide domains. Synthesis of mono-glycosylated pentapeptides matching the *N*-terminal sequence of Glycophorin A^M, each with an *O*-GalNAc at a different one of four Ser/Thr residues, was carried out by the Dill group and natural abundance ^{13}C-NMR was used to characterize them.[85] It was concluded that the placement of the glycan had noticeable effects on the properties and NMR signals of the pentapeptide, and in general the greatest changes where at the location of the glyco-amino acid in the sequence. Follow up work by the same lab with a higher field instrument, was able to show that glycosylated threonine residues in the sequence were more conformationally constrained than glycosylated serines.[86] More complicated constructs based on glycophorin were also noteworthy contributors. In an important 1999 study, a decapeptide containing six consecutive *O*-GalNAc residues was synthesized and studied.[87] The authors were able to obtain an impressive number of NOE-distance and backbone angle restrictions, which were then used in a hydrated molecular modeling simulation of the structure. This simulation resulted in an extended "wave-like" structure for the glycosylated peptide, showing that the extended structure of densely glycosylated peptides is most-strongly dependent on the first carbohydrate unit.[87] The authors were also able to confirm that glycosylated Thr is a more rigid unit than glycosylated Ser.

Around this time, structural studies on mucin-derived glycopeptides were also reported. One such study looked at NMR structures of glycopeptides corresponding to fragments of MUC7.[88] Much like the glycophorin constructs, they found an extended, rigid random-coil similar to a polyproline type II (PPII) coil for the peptide backbone. Also identified were NOE contacts between the backbone and the *O*-GalNAc of Thr-linked glycans but not Ser-linked glycans, suggesting that the GalNAc-Thr linkage is more rigid than that of serine, which is similar to the results found in glycophorin systems.[86,87] The stability differences between Thr and Ser glycosylation are a general theme of *O*-GalNAc type glycosylation. More detailed study into the structures of glycosylated Thr and Ser confirmed that the GalNAc-Thr is a much more rigid structure, and attributed this to a significant change in the way water is structured around each of the glyco-amino acids in solution.[89] The same group was able to show later that these structural differences are reflected in the binding preferences of several lectins, were a significant amount of selectivity was observed for GalNAc-Thr or GalNAc-Ser over the other in identical peptide sequences.[90] This shows that lectin recognition motifs can depend on the underlying peptide sequence as well as the glycan epitope.

There is also the very impressive work of the Danieshefsky and Live groups on structure of the mucin domain of CD43. The authors first prepared, through chemical synthesis, four glycosylated pentapeptides, each carrying three glycans on consecutive amino acid side chains.[83] The glycans were varied in size from single *O*-GalNAc to a sialylated trisaccharide. From the detailed NMR analysis of the synthetic glycopeptides, it was revealed that the glycans induce an extended and extremely stable structure, highly unusual for a peptide of the size.[83] Additionally, through synthesis of the β-linked stereoisomer of one of the glycopeptides, it was shown that the native α-linkage between Ser/Thr and GalNAc is critical to any stabilizing effect. By comparing the structures of glycopeptides bearing different size glycans, it was revealed that the distal carbohydrate units played a very minor role in stabilizing the observed structure, confirming the conclusions of others. Together, these observations led to a general model for glycosylation in mucin domains where the peptide backbone and initial α-O-GalNAc carbohydrate residues form a very stable scaffold upon which important end-group carbohydrate epitopes can be mounted and act as signals in intercellular communications.[83] The authors followed up this study with more detailed work on even higher field magnets a few years later.[91] In addition to the substrates previously examined, they also synthesized and characterized a slightly longer heptamer sequence containing three consecutive hexasaccharides. Much like the previous structures reported, they found an extremely rigid structure for the glycopeptides, and the length or size of the glycans did not have a significant effect on the stability of the molecules, even with hexasaccharides. With the more detailed

NMR results, the authors could identify a highly organized network of hydrogen bonds linking the first α-*O*-GalNAc to the surrounding peptide backbone, which helps to explain the remarkable stabilization observed in the systems.[91]

MUC1, which is implicated in tumor progression and is thus a potential therapeutic target in several types of cancers,[64] has also been widely studied in terms of its structure. The 20-residue tandem repeat sequence of MUC1 contains two interesting structural features: a PDTRP motif that is responsible for immunogenicity of MUC1 and a GVTSAP motif that is a target for ppGalNAcT-1 and -T-3. In cancers, MUC1 is often under-glycosylated and this leads to the revealing of the immunogenic PDTRP motif, which has important implications in cancer diagnosis and treatment.[92] To study possible influences on the immunogenic motif by glycosylation of the sequence, a synthetic 15-mer peptide and its glycosylated analog were synthesized and characterized with NMR.[93] Although the most populated NMR structures of the area immediately surrounding the glycosylation site were common to both glycosylated and unglycosylated constructs, there was also a unique conformation for each showing some difference in conformational space available to the glycosylated peptide compared to the unglycosylated one and *vice versa*. Additionally, the authors were able to identify hydrogen bonding between the amide proton of the GalNAc moiety and the peptide backbone, much like in studies of other systems as discussed above.[91,93] This hints at a general mechanism for stabilization. The calculated structure closely matched a crystal structure of an anti-MUC1 antibody in complex with a 13-mer sequence containing the PDTRP motif.[94] A study of glycosylation directly on the PDTRP motif showed that the region adopted a β-I-type turn in the absence of glycosylation, but a much more extended conformation upon glycosylation by α-*O*-GalNAc at the Thr.[95] The Kunz group was able to synthesize and study the complete 20-mer tandem repeat sequence containing a GalNAc within the GVTSAP motif.[96] They found that the entire sequence from the glycosylation site through the immunogenic motif (nine amino acids) was very rigid and stable. The glycosylation site adopted an extended, wave-like conformation very similar to those observed in the glycophorin and CD43 mucin domain systems discussed previously,[83,87] while the immunogenic domain was found in a β-turn structure.[96] Further work on the MUC1 repeat has also been carried out by the Nishimura group, who was able to chemically synthesize a series of glycopeptides of the full 20-mer repeat domain of MUC1 each containing one core-2 based tetrameric structure at one of the five potential glycosylation sites.[97] Enzymatic extension on both arms of the core oligosaccharide gave a series of glycopeptides displaying di-sialylhexasaccharides. NMR analysis and structural modelling of these glycopeptides showed that sialylation did indeed induce small changes to the backbone conformation, but only with in the PDTRP immunogenic motif. The authors conclude that the underlying sequence is a noticeable factor in how distal sialylic acids alter the conformation of a glycopeptide. Together, these studies on MUC1 indicate that the immunogenicity of certain sequences can be modulated by conformational switches that are in turn controlled by the glycosylation state of the peptide. Since, aberrant glycosylation is an extremely common feature of most cancers; this has the potential to inform new understanding of cancer biology and new therapies.[98]

1.3.1.3 Functional Effects of O-GalNAc Glycosylation

Chemical biology has also contributed to the study of functional aspects of mucin-type glycans in biological systems. As mentioned earlier, glycosylation patterns on extracellular membrane proteins are very important for cellular communication and adhesion. For example, during inflammation, leukocytes adhere to endothelial cells near the site of injury before migration towards the injury site.[99] This adherence is the result of P-selectin on the endothelial cells binding specific ligands on leukocytes, the most important of which is known as P-selectin glycoprotein ligand-1 (PSGL-1). P-selectin binds to the extreme terminus of PSGL-1, and both tyrosine sulfation and *O*-glycosylation of PSGL-1 in this region have been shown to be critical for high affinity binding by studying genetic knockout models.[99] To identify the particular modifications required for binding in this context, the Cummings group chemo-enzymatically synthesized a 23-mer glycosulfopeptide based on the *N*-terminal sequence of PSGL-1 that contained three sulfotyrosines and a sialyl Lewis x (sLe^x) hexasaccharide.[100,101] To avoid problems with site-specificity of ppGalNAcTs, the peptide was synthesized on solid phase with the first GalNAc at position Thr 57 in place. Enzymatic transformations by a series of five glycosyltransferases constructed the full hexasaccharide and lastly addition of the sulfate groups was also accomplished enzymatically. It was found that the sulfotransferase was able to add sulfates to all three tyrosines in the glycopeptide simultaneously.[100,101] In addition, the authors synthesized a glycosulfopeptide with a core-1 sLe^X-containing hexasaccharide and the same three sulfotyrosines through a very similar chemoenzymatic approach. The authors then tested the two final glycosulfopeptides for binding affinity to P-selectin and found that only the construct with the core-2 glycan structure was bound to a measurable degree.[100,101] In addition, when the terminal sialic acid was removed from the glycan, binding was abolished. Moreover, they found that the synthetic intermediate carrying the full length glycan, but not the sulfates, did not bind, nor did the isolated hexasaccharide bind. This study suggests that both modifications, sulfation and glycosylation, are critical to high affinity binding between PSGL-1 and P-selectin. However, because the authors were unable to chemo-enzymatically prepare any partially-sulfated glycopeptide products, they could not address the question of whether or not specific sulfo-tyrosines were more important to binding than others. This question was solved in a follow-up study by the same group that used chemical synthesis to introduce the sulfo-tyrosine residues in a completely controlled manner.[101] In this study, they synthesized a large library of glycosulfopeptides containing each possible combination of 0-3 sulfate groups and several different glycan structures. Characterization of each construct and comparison of the results revealed that the most important determinant of binding was the presence of a fucose in the sLe^X epitope on the peptide. Although the number and pattern of sulfate groups as well as the terminal sialic acid had strong effects on binding. Also noteworthy is the site-specific nature of this glycan dependence. Moving the full length glycan from the natural glycosylation site to a nearby site (both threonines) resulted in complete loss of binding.[101]

1.3.2 α-*O*-Man

O-mannosylation, originally thought to exist only in yeast and fungi, is now a well-known modification of mammalian systems. First discovered on proteoglycan core proteins in the brain and assumed to be relatively rare in mammals,[102] more recent estimates put mannose-initiated *O*-glycan structures as accounting for 30% of the total *O*-glycans in the brain.[103] Significantly, in contrast to yeast *O*-mannosyl glycans, which are mostly linear chains,[104] the *O*-mannosyl glycans observed in mammals have complicated structures with a wide variety of monosaccharides, branching, and charged terminal structures like Le^X and HNK-1 epitopes.[105-108] Over the last two decades, *O*-Man-linked glycans have been confirmed on a wide variety of proteins in both the brain and muscle cells, including the cadherin and plexin families of cell membrane receptors,[109,110] IgG2,[111] CD24,[112] neurofascin 186,[113] receptor tyrosine phosphatase β (RPTPβ),[114] and brain-derived lectican proteins: aggrecan, versican, neurocan and brevican.[115] Many of these proteins are important for their role in cell adhesion and the list includes almost every known member of the perineuronal net, which is responsible for stabilizing neuronal synapses in mature brain tissue.[115] Along with their wide spread expression throughout the nervous system suggests, the severe phenotypes that result from mutations in their biosynthetic machinery confirms that *O*-mannosyl glycans are functionally important in mammals. While many of the known carriers of this modification remain largely uncharacterized, there are a few examples of studies examining the detailed functional consequences of *O*-mannosylation. For example, recent evidence has pointed to the expression of HNK-1 epitopes on 2,6-branched *O*-mannosylglycans, particularly on RPTPβ in the brain, as critical to proper myelination and brain function.[114,116] *O*-Man-linked glycans have also been implicated in critical stages of early neuronal development by serving as the most important scaffolds for both HNK-1 and Le^X epitopes in the developing brain.[117] But by far the best characterized *O*-mannosylated protein in humans is α-dystroglycan (α-DG).[118]

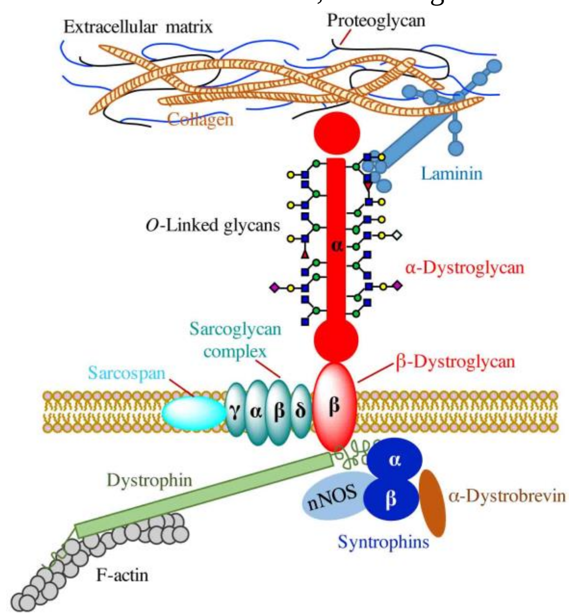


Figure 1.7 - Schematic representation of the interactions between dystroglycan, dystrophin and the extracellular matrix.

Dystroglycan is the product of a single gene, *DAG1*, which is cleaved into two mature proteins: α-and β-DG. β-DG is a transmembrane protein that binds intracellular dystrophin, which binds the cytoskeleton protein F-actin.[25] α-DG is secreted and binds the extracellular portion of β-DG at one end and several extracellular matrix components, most notably laminin, on the other end.[25] Together these proteins form the dystroglycan complex (DGC), which physically links the cytoskeleton of a cell to the surrounding basement membrane.[16] α-DG has a dumbbell shape with two globular domains separated by a conserved mucin-like core, which has over 40 potential *O*-glycosylation sites and is known to harbor both classic mucin-type (*O*-GalNAc) and *O*-Man glycans in close proximity to one another (Figure 1.7).[119] This glycosylation, and in particular the *O*-mannosylation, of α-DG is critical to the function of the complex.[118,120] Thanks to the many studies employing the powerful tools of chemical biology, over the past 20 years, our knowledge and understanding of this important system has been substantially enriched.

1.3.2.1 Biosynthetic Pathway of *O*-Man Glycans

As in the study of *O*-GalNac glycosylation, several groups have attempted to settle the question of whether or not *O*-mannosylation has a consensus motif in recent years through the use of synthetic peptides. The Endo group, in a particularly relevant study, synthesized a series of 20-mer peptides covering the entire mucin-like domain of human α-DG and assayed each for ability to accept mannose transfer from recombinant POMT1 and 2.[121] From this assay, they were able to identify two regions that were particularly prone to mannosylation corresponding to residues 336-355 and 401-420. Comparing these two regions of α-DG to one another and to regions of mucin proteins that are not able to be mannosylated in the same assay suggested that the consensus sequence for *O*-mannoyslation is: IXPT(P/X)TXPXXXXPTX(T/X)XX.[121] However, this sequence has since been only mildly predictive of the observed mannosylation patterns in several recent mapping studies of α-DG.[119,122-124] Later work has pointed to a cis-regulatory 41-residue sequence located adjacent to the α-DG mucin that, along with the peptide primary sequence around glycosylation sites, influences *O*-mannosylation *in vivo*.[125]

After initial mannose transfer, POMGNT1 is responsible for extending the glycan by attachment of a GlcNAc through a β1-2 linkage.[17] Study of this step in the biosynthetic pathway of both core M1 and M2 mannose-type glycans has also been advanced through the use of chemical biology. Initially, the presence of β1-2-GlcNAc transferase activity in brain extracts was confirmed using synthetic peptides and glycopeptides as probes for enzymatic activity.[18] That same study then made extensive use of synthetic carbohydrates as authentic chromatography standards to characterize the structure of the resulting disaccharide and classify the glycosyltransferase as a β-1,2-*N*-acetylglucosaminyltransferase. The stereochemistry and connectivity of the linkage was later confirmed by NMR studies on glycopeptides synthesized chemoenzymatically with recombinant POMGNT1.[126] Mutations in the *POMGNT1* gene cause a severe form of muscular dystrophy, muscle-eye-brain (MEB) disease, and in MEB patients, α-DG is significantly hypoglycosylated and lacks laminin binding affinity.[17,127] These facts have prompted a deeper look at the functioning of the POMGNT1 protein. One such study looked at the effect of known MEB-causing mutations on the enzymatic activity by comparing the ability of mutant enzymes to incorporate GlcNAc into synthetic mannosyl-glycoproteins that mimic the natural sequence of α-DG.[128] Although they had difficulty correlating the structural effects of mutations on the enzyme with severity of the resulting phenotype, they did conclude that a complete loss of enzyme function is not necessary for an observable phenotype since several mutants studied still had appreciable activity towards synthetic glycopeptides *in vitro*. Also significant, the authors of the study found that only small subset of the synthetic glycopeptides used as enzymatic probes were actually modified by POMGNT1, showing that, at least in this *in vitro* assay, only specific regions of α-DG are modified with extended mannose-type glycans.[128] The clear sequence dependence of POMGNT1 activity was confirmed in a very similar study that assayed for enzymatic activity using a series of glycopeptides of varying lengths.[129] The authors of that study concluded not only that the sequence surrounding the *O*-mannose affects glycan elongation, but also that a minimum of eight residues is required for efficient binding of POMGNT1.

Curiously, despite the fact that *POMGNT1* mutations result in MEB phenotypes,[17] the actual binding of α-DG to the basement membrane is mediated by a glycan structure that does not require POMGNT1 for assembly: the core M3 O-mannose glycan.[21] Using synthetic glycoconjugate substrates and recombinant enzymes, the entire biosynthetic pathway for the core M3 phosphotrisaccharide was recently laid out in more detail.[22] The authors started by examining the activity of GTDC2 (POMGNT2) towards a synthetic mannosyl-glycopeptide corresponding to residues 316-329 of human α-DG's mucin-like domain, one of the select few regions of the domain where this structure occurs.[120] They found that GTDC2 was able to incorporate UDP-GlcNAc but not UDP-GalNAc to the synthetic glycopeptide substrate.[22] To more rigorously characterize the linkage and stereochemistry of the disaccharide, the synthetic glycoconjugate 4-methylumbelliferyl-α-D-mannoside was used as the enzyme substrate and the product was analyzed with NMR. The authors were able to clearly identify the disaccharide glycosyl transfer product as GlcNAc-β-1,4-Man, proving that GTDC2 has protein *O*-mannose β-1,4-*N*-acetylglucosaminyltransferase activity.[22] Next, the authors used this chemoenzymatically produced GlcNAc-β-1,4-Man glycoconjugate to test if B3GALNT2 possessed its predicted β-1,3-*N*-acetylgalactosaminyltransferase activity and found that it did. The authors thus concluded that POMGNT2 (GTDC2) and B3GALNT2 act in series to construct the trisaccharide moiety of M3 core glycans on α-DG, and turned their attention to the phosphorylation of the glycan which had previously been shown as necessary for laminin binding.[21,120] This trisaccharide was then mixed with recombinant SGK196 (POMK) and ATP, which verified that SGK196 was, in fact, the kinase responsible for *O*-Man glycan phosphorylation.[22] The phosphorylation site was confirmed as being on the 6-position of the mannose by NMR. Interestingly, SGK196 was only able to phosphorylate the full-length trisaccharide, which helps to explain how mutations in *POMGNT2 (GTDC2)* and *B3GALNT2* cause dystroglycanopathy phenotypes.

The enzyme LARGE has been known has important in the functioning of α-DG for several years.[120,127] In recent years, the understanding of the function of this enzyme has also been vastly improved by using synthetic glycoconjugates and other tools common to chemical biology. The specific transferase activity of LARGE was confirmed by Inamori *et al*. in 2012 through a series of chemical biology-base experiments.[130] They took recombinant LARGE enzyme and mixed it with either an α-xyloside glycoconjugate in the presence of UDP-GlcA or a β-glucuronide glycoconjugate in the presence of UDP-Xyl. In both cases disaccharides were formed. Incubation of these dimer products with LARGE, UDP-Xyl and UDP-GlcA resulted in a polymer of repeating disaccharides [-Xyl-α-1,3-GlcA-β-1,3-]. Using NMR, the authors were able to confirm the anomeric stereochemistry and regioselectivity of each carbohydrate linkage, that LARGE alone is capable of forming carbohydrate polymers and that it possesses bifunctional glycosyltransferase activity.[130] They continued this study a year later to examine a LARGE paralog in mammals, LARGE2. In that follow up study, they found LARGE2 to have the same biufunctional glycosyltransferase activity, but a slightly different optimum pH range.[29] While this did answer the questions surrounding the nature of LARGE's enzymatic activity, it did not address several remaining questions concerning relationship between the structures of LARGE-synthesized glycans on α-DG and binding affinity towards laminin. Work by the same group began to answer these later types of questions with the study of synthetically produced LARGE glycans.[24] In this study, chemoenzyamtically synthesized LARGE-glycan repeats were immobilized on ELISA assay plates and laminin binding affinity was measured. They found that the length of the repeating glycan was directly proportional to the binding affinity towards ECM components like laminin.[24] They went on to show that the expression of LARGE, and as a result the length of the LARGE-glycan repeats on α-DG, was temporally regulated during muscle regeneration after injury. Together, these two pieces of evidence were used by the authors to propose a model where the binding affinity of α-DG towards the basement is regulated by the length of LARGE-glycan repeats through variations in the level of expression of LARGE.[24] This has potential functional implications for muscular dystrophies where mutations in LARGE inhibit its ability to synthesize glycans of proper length. However, LARGE cannot modify the phosphotrisaccharide core of M3 glycans on α-DG, which suggests further modifications on the glycan are necessary before LARGE polymerization can take place. A recent paper found that the misnamed β-1,3-*N*-acetylglucosaminyltransferase 1 (B3GNT1) is actually a β-1,4-glucuronyltransferase with a preference for xylose acceptors.[26] Through a combination of synthetic substrates and recombinant enzymes, the authors showed that B3GNT1 transfers a GlcA with a β-linkage to the 4-position of a xylose acceptor, and they propose a new name, β-1,4-glucuronyltransferase 1 (B4GAT1), for this enzyme. B4GAT1 was not found to be capable of forming polymeric carbohydrates, but it did have a significantly higher activity towards monosaccharide acceptors than LARGE. Because of these activity differences, it was proposed by the authors that this newly characterized enzyme (B4GAT1) acts as a "priming" enzyme by transferring the first GlcA to a Xyl acceptor to form a dimer structure that is easily recognized and elongated by LARGE.[26] They did not report, however, on the nature of the connection between the M3 core glycan structure and the potential Xyl residue that B4GAT1 requires for activity, and so mysteries remain in the biosynthesis of *O*-mannose type glycans.

While understanding the biosynthetic pathways and regulation of those pathways behind each individual glycan is important, there are also interactions and possible cooperative regulatory effects between different glycans to consider. For example, it is well known that the core M3 glycan structure responsible for laminin binding is only present on a few specific Thr residues of the α-DG sequence, and that those residues are in close proximity to sites modified by the more common core M1 glycans.[119,122-124] Furthermore, mutations present in the enzyme POMGNT1, which is only involved in the synthesis of M1 glycans, are known to have a strong effect on α-DG laminin binding in patients with MEB.[17,127] These pieces of evidence together point to significant cross talk between different types of *O*-mannosylation at different sites. Here again, the unique approaches and mindset of chemical biology can achieve significant advances. After the initial discovery and characterization of the unique M3 core structure,[21] the synthesis of the novel phosphoglycan component and its incorporation into glycopeptides was seen as a top priority in order to allow future *in vitro* studies to be carried out. The synthesis of the M3 core structure was successfully completed and the resulting phosphotrisaccharide as well as glycopeptides corresponding to the natural α-DG sequence known to bear the M3 structure were studied.[131] Using mono-mannosylated glycopeptides corresponding to the α-DG sequence surrounding the known M3 glycan site, they found that POMGNT1 was able to extend a mannose at the known M3 site (T379), but not at a site two residues away (T381).[131] This, along with the studies discussed previously,[128,129] illustrates the sequence dependence of POMGNT1. In addition, synthetic glycopeptides containing single mannose units at both the M3 site and the nearby site were found to be modified by POMGNT1 at both mannose residues, although not both simultaneously. This suggests that POMGNT1 activity and selectivity can be significantly altered by clustered presentations of glycans on a peptide backbone.[131] The fact that POMGNT1 can modify the site of M3 glycans *in vitro* and yet no M1-type glycans have been observed at this site *in vivo* was later explained when the biosynthetic pathway for the M3 core-glycan was found to occur entirely in ER, before possible exposure to POMGNT1 in the Golgi.[22]

The probable interplay between classical mucin-type *O*-GalNAc initiated glycans and *O*-Man initiated glycans in the mucin region of α-DG was more explicitly investigated in a later study.[132] In that study, the authors examined the efficiency and selectivity of several different ppGalNAcTs towards glycopeptides already containing one or more *O*-Man glycans. This mimics the natural biosynthetic pathways of the two glycans where *O*-Man is added in the ER[16] and *O*-GalNAc is added later in the Golgi.[5] They found that the presence of *O*-Man glycans on the sequence has a significant effect on the activity of ppGalNAcTs, and furthermore that such effects are site specific in nature.[132] These conclusions were drawn from studying synthetic glycopeptides derived from two regions of the α-DG sequence, each sequence containing four consecutive sites for potential glycosylation. The first sequence was from a region of the protein shown to only contain *O*-Man type glycans *in vivo*, and the second from a region thought to contain both *O*-Man and *O*-GalNAc type glycans. Both of the sequences were shown to be able to accept GalNAc transfer, unless specific *O*-mannosylation patterns were present before exposure to ppGalNAcTs.[132] This highlights the importance of specific glycosylation patterns in regulating subsequent post-translational modification and, eventually, protein function. Also noteworthy is that fact that the location of *O*-GalNAc incorporation depended on the *O*-mannosylation pattern, showing that *O*-

Man glycans have the ability to not only up- or down-regulate the activity of subsequent glycosyltransferase enzymes, but can also strongly affect the sites at which those enzymes act.[132]

1.3.2.2 Biophysical and Biological Effects of *O*-Mannosylation

Understanding the consequences of *O*-mannosylation on the biophysical and biological properties of a protein or peptide is also an important research area. Given the well documented structural effects of *O*-GalNAc type glycans on mucin domains in other systems, it was of significant interest to investigate how the presence of both *O*-Man and *O*-GalNAc glycans on the mucin domain of α-DG might affect the structure of the protein. As with *O*-GalNAc glycosylation, biophysical and structural studies of synthetic model systems allowed for quick and robust conclusions. For example, by synthesizing a series of glycopeptides derived from the α-DG mucin region and containing up to four consecutive *O*-Man residues, it was possible to conclude that the glycopeptides existed in a mostly unordered structure.[133] Interestingly, when the glycans were switched to *O*-GalNAc, the same sequence was considerably more ordered, as would be expected given the previous work on mucin-domain structures.[91,95] This comparison held across three different biophysical characterization methods: circular dichroism (CD), NMR and hydrogen-deuterium exchange (HDX).[133] Combined with the well-documented importance of *O*-Man glycans in the functioning of α-DG,[16,23] these observations led to a proposed model where the *O*-GalNAc glycans provide the mucin domain with a defined, rigid structure that supports optimal display of the functionally important end-groups of the *O*-Man glycans.[133,134] Structural studies of the glycopeptide bearing a negatively charged 6-*O*-phosphomannose residue also supported this assertion. As with previous work on the structural impact of *O*-mannosylation, the authors found little structural difference between the unglycosylated peptide and the phosphomannose-containing glycopeptide.[131] Replacing the mannose derivative with an *O*-GalNAc, however, imparted significant changes in the NMR spectra of the glycopeptide.

In addition to the studies of the effects of *O*-glycosylation on unstructured mucin domains, many groups, including our own, have undertaken research programs to get this knowledge with respect to of folded peptides and proteins. This is the subject of the remaining chapters of this thesis and will be discussed further in those chapters.

One example of this type of research is the work done by the Tan group on a Family 1 carbohydrate-binding module (CBM) of a fungal cellulase. The CBM is a small, 36-residue peptide domain and is responsible for recognition and binding of cellulose (Figure 1.8). It is naturally glycosylated by several *O*-glycans, which have been implicated as critical in the functions of this domain.[135] In order to characterize the specific effects of *O*-mannosylation on the CBM, a library of CBM glycoforms bearing mono- and oligo-mannose chains at each of the three possible glycosylation sites were chemically synthesized. By comparing the individual members of this glycoform library, the Tan group was able to show that *O*-glycosylated CBM was more stable and bound tighter to cellulose.[136] Particularly intriguing, the authors found that glycosylation at Ser3, more so than either of the other two sites, was the single most important factor in the thermostability of the glycoforms (Figure 8). This finding shows that in certain contexts, glycosylation can site-specifically control the physical properties of glycoproteins. It also hinted at as-yet-underappreciated interplay between glycan structures and the underlying peptide sequence or structure that can lead to a specific outcome in the context of one

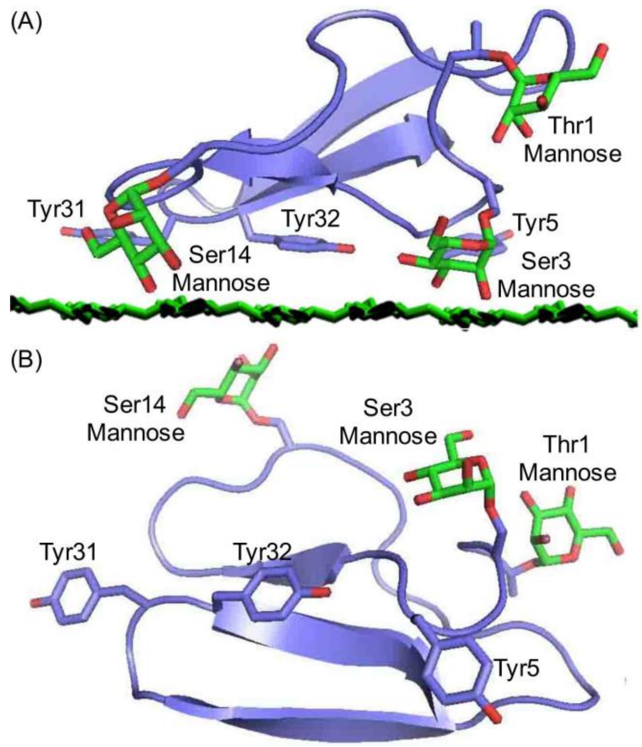


Figure 1.8 - The NMR structure of the *O*-mannosylated Family 1 CBM. (A) Side view. (B) top view. The top layer of the cellulose crystal is shown in green. The tyrosine residues that form the binding face to cellulose and the three mannose monosaccharides are show in sticks.

glycosylation site but not another. This hypothesis warranted further investigation, and a follow-up study by the same group looked at the possible cooperative effects of amino acid side chains and glycan structures on the physical and functional properties of the CBM.[137] As with the previous study, a library approach was chosen and the authors synthesized 31 new CBM isoforms with amino acid mutations and a wide range of glycan structures varying in size, linkage stereochemistry, branch structure and charge state and each library member was rigorous characterized through a panel of biophysical and functional assays. By systematically investigating each structural feature, the authors determined that planar polar (Gln) and aromatic (Tyr) side-chains near the Ser3 site were critical to the increased stabilization observed upon glycosylation. Additionally, the stereochemistry of the linkage between glycan and peptide had to be in the α configuration. Thus, by comparing biophysical characteristics across library members, this study provided new insights into the molecular basis for the effects of *O*-glycosylation at Ser3 of the CBM. The strong conclusions made by these studies were only possible because of the ability to controllably synthesize a large variety of glycoforms. This ensured that any structural feature, including linkage stereochemistry and branching structure, which are notoriously difficult to control in natural expression systems or enzymatic syntheses, could be thoroughly investigated.

1.3.3 α-*O*-Fuc

Many proteins have been validated as carrying α-*O*-fucosylation, including plasminogen activator proteins,[138] blood coagulation factors,[34,139,140] Cripto,[140] and Notch.[141] In addition, numerous proteins containing EGF repeats and TSRs have the consensus sequence for *O*-fucosylation, although many of these have not been experimentally verified as glycosylation sites *in vivo*.[142] The effects of *O*-fucosylation on different proteins have also not been systematically investigated. However, the facts that POFUT1 and POFUT2 only recognize and modify properly folded substrates[143] and that *O*-fucosylation is essential for the secretion of at least two ADAMTS superfamily proteases[144,145] suggested that *O*-fucosylation might be important for folding or quality control as well.[146]

Specific roles for *O*-fucosylation have also been identified, although they have not been extensively characterized. For example, defucosylated urokinase-type plasminogen activator (uPA) was shown to be able to bind its receptor as normal, but did not activate mitogenic activity in the bound cells, so it does not appear to stimulate the normal signal upon binding.[147] Recently, the presence of an *O*-Fuc on the EGF repeat of Cripto was identified and was initially thought to be essential for Cripto to mediate Nodal-dependent signaling since POFUT1 knock outs displayed a lethal phenotype related to Nodal signaling.[140,148] Later investigations showed that while POFUT1 does indeed fucosylate Cripto and the Thr that gets fucosylated is essential, the carbohydrate itself is not. Even substituting the Fuc-Thr for Fuc-Ser resulted in the same lethal phenotype as an alanine mutation.[149] So far, the roles of *O*-fucosylation and the POFUT1 enzyme in Cripto/Nodal signaling system remain cryptic.

The best characterized and most extensively investigated *O*-fucosylation substrate is Notch. Notch is an intercellular juxtacrine receptor critical for cell-fate decision making, development, and homeostatsis.[150] Notch structure and signaling is well conserved across metazoans, and all Notch receptors are large single-pass type I transmembrane proteins with a long extracellular domain containing between 29 and 36 EGF repeats in tandem.[151] These EGF repeats mediate interactions between Notch and its ligands and many of these repeats are post-translationally glycosylated with a variety of structures including Ser/Thr-linked *O*-Fuc and *O*-Glc glycans.[152] These glycans were first identified on mammalian Notch1 in 2000.[37] *O*-Fuc is added to Notch by a single glycosylfucosyl transferase: protein *O*-fucosyltransferase-1, POFUT1 in mammals and OFUT1 in *Drosophila*.[150] Mutations which destroy the transferase activity of POFUT1 or OFUT1 imply that *O*-fucosylation is necessary for correct Notch signaling.[150] In *Drosophila*, the glycosyltransferase Fringe extends the glycan to a disaccharide: GlcNAcβ1,3Fuc.[35,153] In mammals, there is some evidence that this disaccharide can be further extended to a tetrasaccharide (Siaα2/3,6Galβ1,4GlcNAcβ1,3Fuc) by the galactosyl transferase B4GalT1[34,36] and a sialic acid transferase.[34] In flies, it is well established that glycan extension by Fringe makes Notch more sensitive to the ligand Delta and simultaneously less sensitive to the competing ligand named Serrate.[154] However, in mammals, there are numerous isoforms of Notch, Fringe and each of the ligands, making the mammalian system significantly more complicated. Despite significant effort, a clear picture of how glycosylation affects Notch signaling in mammals has not emerged. It is known that EGF repeats 11 and 12 of Notch are responsible for interactions with the ligand Delta.[155] Since EGF repeat 12 is known to be both *O*-fucosylated and *O*-glucosylated, EGF repeat 12 has been the subject of several studies in this area. Recent work has revealed that Fringe-catalyzed extension of *O*-Fuc on EGF 12 results in an increased binding affinity of human Notch1 towards both Jagged-1 and Delta-like-1 (DLL1) ligands.[156] There is also evidence that the extension of *O*-Fuc effects the interaction of Notch1 with the different isoforms of Fringe found in mammals and the DLL1 ligand.[157] In another study, structural and computational analysis of a Notch1 fragment in complex with a DLL4 fragment showed that extension of the *O*-Fuc on EGF 12 adds a significant amount of contact area between the Fringe-added GlcNAc and both DLL4 and Notch1 surfaces.[158] Further extension of the glycan to a tri- or tetrasaccharide has also been investigated as a potential regulatory event. Elongation of the glycan to a trisaccharide by B4GalT1 is necessary for Lunatic fringe dependent inhibition of Jagged1 activation of Notch in mammals.[36]

Along with traditional biochemical approaches, the recently-reported chemical synthesis of homogenously glycosylated EGF repeat 12 has been instrumental in shedding light on the specific structural interactions of the glycopeptide domain. Both mouse and human EGF repeat 12 have been chemically synthesized by the Nishimura group in recent years, following a very similar synthetic strategy (Figure 1.9).[159] The mouse-derived sequence was first synthesized in 2010 by the group with a single *O*-Fuc-initiated disaccharide.[159,160] They were also able to enzymatically elongate the glycan after synthesis and folding to the full-length, naturally occurring tetrasaccharide. NMR studies of the synthetic glycoprotein domains revealed that the *O*-Fuc glycan is involved in several key contacts with amino acid residues opposite it on the anti-parallel β-sheet strand. The authors suggest that this is indicative of the glycan's stabilizing role as a sort of bridge across strands of the β-sheet and that it shows the importance of *O*-fucosylation for the structural stability of EGF repeat 12.[159] X-ray structures of glycosylated human

Notch a few years later confirmed many of these contacts.[156] Most recently, the Nishimura group has completed the synthesis of human EGF repeat 12 with both an *O*-fucose-initiated disacchride at Thr466 and an *O*-glucose-initiated Xyl-α1,3-Xyl-α1,3-Glc-β1-O-Ser trisaccharide at Ser458 (Figure 1.9).[160] They modified the previous folding protocol to include calcium ions in the folding buffer, which are known to be necessary for the final structure of EGF repeat 12.[156] As with the previous work, enzymatic elongation of the *O*-Fuc disaccharide yielded the full-length, sialylated tetrasaccharide at Thr466.[160] NMR analysis of the resulting glycopeptide domains confirmed many of the same contacts identified previously with glycosylated EGF repeat 12 from mouse Notch1. Additionally, they found that there were significant contacts between the terminal xylose of the *O*-Glc initiated glycan and amino acid side chains in the structurally critical β-sheet of the EGF repeat.[160] By including calcium ions in the folding buffer for this synthesis, the authors were also able to investigate the influence of divalent metal ions on the structure. Most notably, they were able to show for the first time that calcium ions not only helped to stabilize the final structure, but actually accelerated the folding reaction and decreased the presence of mis-folded glycopeptide structures. Finally, by comparing the NMR structures obtained of both mouse and human EGF repeat 12, as well as X-ray structures, the authors were able to identify structural contributions by both types of *O*-glycans as well as the coordinated calcium ion to EGF repeat 12.[156,159,160] This led to the authors' conclusion that all three factors are important for maintaining the correct fold of the EGF domain, and thus most likely all three act as modulators of Notch signaling in concert.

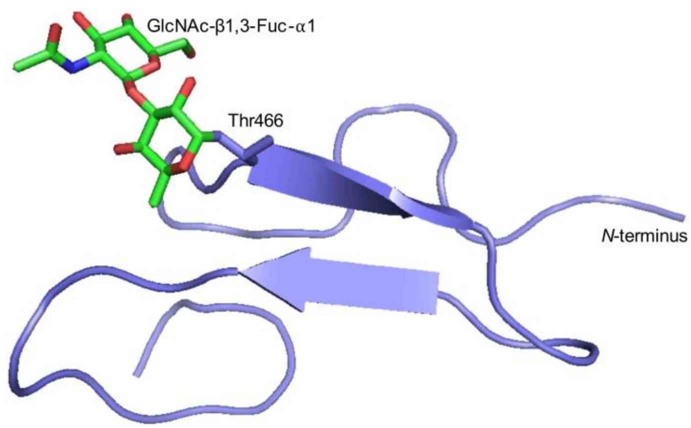


Figure 1.9 - Three-dimensional structure of synthetic human Notch 1 EGF repeat 12. The *O*-fucosyl-linked Thr466 is shown as sticks.

In addition to EGF repeat 12, *O*-fucosylation of EGF repeats 26 and 27 is known to significantly affect the binding of Notch1 towards its ligands, even though this region is not implicated in the binding event.[161] Recently it has also become appreciated that the Notch ligands themselves are also *O*-fucosylated by POFUT1[162] and that these modifications are likewise necessary for proper signaling,[163,164] but the molecular details of these effects are not known.

To date, almost all the functional work has been done on EGF domains and the role of fucosylation in the function of TSRs remains largely uncharacterized. TSRs and EGF repeats share many characteristics and both are involved in numerous physiologically-relevant protein-protein interactions.[165] Thus it is possible that fucosylation of TSRs has some functions analogous to those observed for fucosylation of EGF repeats. It is noteworthy that several important interactions between TSRs and ligands are thought to occur in sequences containing proposed *O*-fucosylation sites; but characterizations of how glycosylation affects these interactions have not been carried out.[146] Glucose can be added to Fuc by B3GlcT. The exact role of extension of *O*-Fuc on TSRs by Glc is not understood currently, but it is known that mutations in the glucosyltransferase responsible for the modification (B3GlcT) cause the disease Peter's Plus Syndrome.[166]

1.3.4 β-*O*-Glc

As with *O*-fucosylation, *O*-glucosylation is found on EGF repeats in several proteins, including blood coagulation factors,[167] but so far, only Notch *O*-Glc modification has been heavily studied.[37] The fact that POGLUT1 only recognizes and modifies properly folded substrates suggests that *O*-glucosylation may play some role in protein quality control.[168] In flies, half of the EGF repeats on Notch contain the consensus sequence for *O*-glucosylation, and these 18 potential sites seem to act redundantly to both promote correct folding of the Notch protein and to enable cleavage after ligand binding.[50] Despite the fact that the ligand binding site of Notch (EGF repeats 11-13) is *O*-glucosylated, so far *O*-Glc has not been shown to directly influence the binding of any ligands tested.[156,169] A recently solved crystal structure of Notch EGF repeats 11-13 in complex with DLL4 showed two *O*-glucose residues covering hydrophobic patches away from the sight of ligand binding, which raises the interesting possibility that these glycans function by preventing aggregation of Notch receptors on the cell surface and thus promote proteolytic cleavage and signaling.[158] Further indirect effects on Notch signaling are evident in mammals, where EGF repeat 28 *O*-glucosylation appears to regulate DLL1 but not Jagged1 binding even though it is far from the known binding site for either ligand.[44] It is also worth noting that most Notch ligands in mammals contain several *O*-glucosylation consensus sequences, although validation of these sites as being glycosylated *in vivo* or functional characterization of the consequences of any glycosylation on these proteins has not been investigated.[170] Finally, investigation of the extension of protein O-glucose by xylose has not yet been carried out in detail in mammalian cells. Studies of flies with mutated, non-functional Shams, however, seem to indicate that such an extension negatively regulates activity of Notch in flies.[49]

1.3.5 β-*O*-GlcNAc

The β-*O*-GlcNAc glycosylation of the extracellular domain of *Drosophila* Notch in EGF repeat 20 was first identified accidently while the authors were looking for the well-known *O*-Fuc and *O*-Glc modifications.[171] Subsequent investigations have confirmed that extracellular GlcNAc exists in humans and mice, as well as flies, on several EGF repeat-containing membrane proteins in the brain and even on a secreted cytokine, AIMP1.[54] In addition, the glycosyltransferase that catalyzes the glycosylation appears to be ubiquitously expressed in mice.[52]

The *Drosophila* protein Dumpy is perhaps the best characterized example of a protein with this type of glycosylation, although it has no obvious mammalian ortholog.[51] Dumpy, which is thought to be critical to interactions between epithelial cells and cuticle cells in the fly, appears to require EOGT-catalyzed *O*-GlcNAcylation to function properly and mutation of *EOGT* in flies is lethal.[51] Curiously, even though Notch in *Drosophila* is known to be *O*-GlcNAcylated, the glycan does not appear to be important for Notch signaling.[51] As with other EGF repeat specific glycosyltransferases, EOGT appears to recognize exclusively fully folded EGF repeats which suggests possible roles in protein quality control.[52] In humans, mutations in the *EOGT* gene have been found in individuals with Adams-Oliver Syndrome,[172] and the mutant EOGT proteins found in those individuals have been shown to impair the transferase activity of the enzyme.[173] So far no definitive role of β-O-GlcNAc has emerged in either *Drosophila* or mammals.

1.4 Chemical Biology in Studying the Composition of Mixtures of O-Glycoproteins

Natural glycoproteins are often secreted as a complex mixture of many glycoforms. To better understand protein glycosylation, it is also important to appreciate the biological consequences of forming such mixtures. To achieve this goal, it is essential to know the composition of glycoforms that are secreted by different cells and under different physiological conditions. The consequences can be determined by analyzing the function of mixtures with known compositions.

Many different methods, including non-mass spectrometric (MS) and MS methods, have been developed for the analysis of protein glycosylation. MS methods are generally more sensitive and rapid, and have thus become more commonly used. The generally accepted workflow for MS methods typically involves four steps, all of which should be optimized for each particular analyte for best results: initial purification, degradation, on-line analytical separation and mass analysis (Figure 1.10).

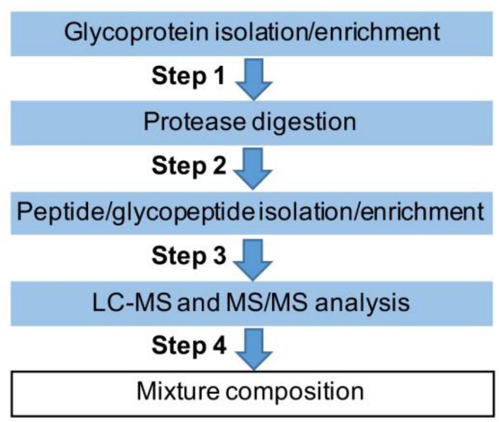


Figure 1.10 - Experimental workflow for glycoprotein mixture analysis.

1.4.1 Glycoprotein Purification and Enrichment

The first step is necessary to separate the target glycoprotein(s) from other molecules in the sample; for example, isolating a single protein and its various glycoforms from a blood-serum sample. This can involve target-specific affinity purifications or less specific separations.[174] Gel electrophoresis, either 1D-

SDS-PAGE or 2D-PAGE is a very common initial purification step in glycoform analysis.[175] Reverse-phase liquid chromatography is also very common for purification early on in the process.[174]

Intact protein molecules with relatively simple glycosylation, for example, most monoclonal antibodies, can be analyzed directly by ESI-MS, and the ratios of unglycosylated to glycosylated protein can be calculated based on the intensities of deconvoluted MS peaks corresponding to each glycoform.[176] However, it is much more common for glycoproteins to carry a variety of glycans at multiple sites, and the mass spectra usually become too complicated to assign in such circumstances. It is thus common practice in many laboratories to degrade the glycoproteins using proteases and to analyze the glycoforms at the peptide level.[177]

The specific recognition of antibodies can be exploited to enrich glycoproteins. For example, *O*-GlcNAcylated proteins have been enriched and subsequently analyzed from rat brain samples using immunoaffinity chromatography that relied on an *O*-GlcNAc-specific antibody immobilized on a solid support.[178] Since antibody recognition is also reliant on the peptide backbone, particularly for small glycans, using multiple antibodies simultaneously significantly increases the coverage of enrichment.[179]

Lectins are another class of specific-epitope binding molecules that recognize carbohydrates and there are a large variety of lectins available for use in enriching glycoproteins.[180] Often these lectins are immobilized on a solid support and packed in columns.[181] As with antibodies, combined use of multiple lectins significantly increases the coverage towards the full spectrum of glycan structures.[182] This is particularly relevant in discovery applications where the target glycan structure is perhaps unknown. While highly effective for enriching full-length glycoproteins, lectins tend to recognize clusters of glycans across multiple binding sites; which can make their use in purifying or enriching small, singly glycosylated peptides challenging. To solve this problem, long columns packed with immobilized lectin beads and isocratic elution conditions have been used in what is known as lectin weak affinity chromatography (LWAC) to enrich *O*-GlcNAcylated glycopeptides[183] or *O*-mannosylated peptides.[109] For

O-GalNAcylated glycopeptides, jacalin, peanut agglutinin (PNA) and Vicia villosa lectin (VVA) are widely used, as jacalin and PNA prefer the T-antigen ligand motif and VVA prefers the Tn-antigen ligand motif. For example, the truncated GalNAc-type *O*-Glycoproteome of "SimpleCell" cultures was mapped by using LWAC packed with VVA to enrich the glycoforms at the both glycoprotein stage and glycopeptide stage.[184]

Due to the unique concentration of vicinal-diol moieties in carbohydrates, orthogonal chemical reactions can be used to selectively capture glycans as well. One of the most common methods here is non-specific oxidation of cis-diols in glycans by periodate to create aldehyde groups, and subsequent covalent capture on hydrazide-coated beads.[185] Unbound, unglycosylated peptides can be easily washed away after glycopeptide capture. Since this procedure is irreversible, the captured glycans are cleaved from the peptides and the formerly-glycosylated peptides are analyzed. In the case of *N*-glycans, peptides can be released by PNGase F digestion,[186] and isotopic labels can be introduced through the use of $H_2{}^{18}O$ in the cleavage step to label the former glycosylation sites,[187,188] although this might not be as selective an isotope label as desired.[189] For *O*-glycans, no such universal glycosidase exists, but several other approaches to peptide release have been explored including β-elimination and direct hydrazine bond cleavage by mildly acidic hydroxylamine treatment.[190] Milder oxidation by periodate can selectively target sialic acids on both *N*- and *O*-glycoproteins. This is advantageous because sialic acids are often the most vulnerable towards acidic hydrolysis, and so the de-sialyated glycopeptides can be easily released after capture and most of the glycan structure as well as site information can be characterized.[191] Obviously, though, all information on the sialic acid content is lost with this approach. Along similar lines, boronic acids can react selectively with the diols of glycans in the context of biological samples, and thus solid support-immobilized boronic acids have also been used to enrich glycoproteins.[192] Nicely, this reaction is reversible and so captured glycoproteins can be released intact under mild acidolysis.[193]

1.4.2. Glycoprotein Digestion and Glycopeptide Separation

Once purified and, if necessary enriched, glycoproteins are often degraded by proteases before analysis. Breaking full length glycoproteins into glycopeptides and peptides allows for much more complete analysis. The protease digestion step can be performed in-gel or in-solution before desalting and MS or LC-MS analysis of the resulting glycopeptide fragments.[175] The digestion can also be performed following hydrazide capture, which can be done while the glycoproteins are still bound to the beads for convenience.[191]

Most often, the protease trypsin is used for glycoprotein digestion.[194] In some cases, other enzymes are necessary, for example due to a lack of trypsin cleavage sites in a particular sample, and many other proteases have been used successfully.[181] In recent years, proteinase enzymes immobilized on solid support and packed into columns have become commercially available.[195] These columns have the advantage of quicker, milder reaction conditions and minimal purification after digestion as compared to the traditional wet chemistry methods. Non-specific proteases have also been used for peptide digestion.[194] In particular, they are effective in combination with more specific proteases (like trypsin) for the analysis of glycoproteins that are difficult to digest and analyze such as densely *O*-glycosylated mucin domains, which also consist of peptide sequences with few cleavage sites for specific proteases.[196]

Finally, often immediately before mass analysis, it is usually advantageous to separate the protease digestion product glycopeptides from one another. It is very helpful to temporally separate the analytes for clearest analysis and ideally, this would allow a single species into the mass analysis at a time. Many methods have been examined to separate and enrich small glycopeptides. Hydrophilic Interaction Chromatography (HILIC) is one such method. HILIC combines the highly polar stationary phases of normal-phase chromatography (sepharose, cellulose, silica) and the moderately polar mobile phases of reverse-phase chromatography (acetonitrile and water).[197] In this context, glycopeptides can be retained in the column while the comparatively more hydrophobic peptides are washed out; increasing the polarity of the mobile phase elutes the glycopeptides. Since its development, HILIC has become a standard part of glycoprotein/glycopeptide analysis.[198] Porous Graphitized Carbon is also commonly used in the analysis and separation of glycans and glycopeptides.[199] It has the advantage of separating not only based on polarity but also size and shape, so isomorphic glycans or glycopeptides containing isomorphic glycans can be separated from one another.[200] Ion-mobility can also separate based on size and shape, but in the gas-phase.[201] This has been used, for example, to separate two glycopeptides with identical sequence and glycan composition but different sites of glycosylation.[202]

Typically, this type of analytical separation is performed on-line, right before induction into the MS for mass analysis. Reverse-phase liquid chromatography (RPLC) is a very common method here. In RPLC the peptide backbone is the biggest determinant of the retention time, and so all the variously glycosylated isoforms of a given peptide sequence often elute very close to one another.[203] RPLC is commonly combined with other separation methods to achieve the best results, for example in combination with HILIC to isolate and characterize *N*-linked glycoproteins from the rat brain.[204] Multi-dimensional separations, done by combining three types of chromatography: electrostatic repulsion HILIC, jacalin affinity chromatography and RPLC, resulted in more comprehensive coverage of mucin-type core 1 glycopeptides from bovine serum.[205]

1.4.3. Glycopeptide Analysis

Almost all mass analysis of glycopeptides is done using tandem mass spectroscopy, also known as MS/MS or MS^2.[180] Within MS/MS techniques, several methods exist for fragmenting the precursor ions with various levels of success. Collision-Induced Dissociation (CID) is one of the earliest and most widely available fragmentation methods for tandem MS. In a CID experiment, high energy collisions with an inert gas in the fragmentation chamber cause ions to fragment, generating smaller ions that are then analyzed.[180] This almost exclusively results in the cleavage of glycosidic bonds, which makes CID very useful for characterization of glycans.[206] Unfortunately, information on the peptide sequence is almost absent from most CID experiments, and it is often impossible to located the glycosylation site. Since CID reliably generates easily-visible oxonium ions from glycosidic bond cleavage, it can be used as an easy way to identify chromatographic peaks that contain glycopeptides in an unknown sample. Other ionization methods can then be used to analyze the peptide sequence and get site information either in parallel or in subsequent runs. This technique was applied to great effect in a study of Haptoglobin, even when the authors employed a novel nano-LC set-up that only used femtomolar quantities of analyte.[207] In certain circumstances, peptide bond cleavage can be achieved, and CID has been successfully used to sequence glycopeptides, identify glycosylation sites and characterize glycans. For example, CID paired with nanospray ESI-MS/MS and TOF detection has been successfully used to characterize sites of *O*-fucosylation.[208] Collision-based fragmentation has also seen use in the context of relatively new linear ion-trap orbitrap (LTQ Oritrap) instruments, where high energy collisions (HCD) can take place before analysis in the orbitrap portion.[209] Because these collisions are higher in energy than standard CID, peptide backbone fragmentation is observed as well as glycosydic bond cleavage. HCD has even been combined with traditional CID for a more complete analysis of glycoproteins in the rat brain.[204]

In some respects, electron-based fragmentation methods have an advantage over collision-based methods in the analysis of glycopeptides since electron-based fragmentation tends to favor backbone cleavage, leaving glycans attached to the resulting fragments intact.[181] This can simplify the assignment of glycosylation sites and peptide sequences. Electron capture dissociation (ECD) is a common feature of Fourier transform ion-cyclotron resonance (FT-ICR) mass spectroscopy that involves bombardment of multiply-charged ions with low-energy electrons to generate odd-electron radical cations, which readily dissociate across the N-Cα bonds of the peptide backbone.[210] ECD is a powerful technique that can allow for complete glycoform analysis, including differentiating glycoforms carrying identical glycans on different sites and glycoforms carrying different glycans on identical sites, as shown by a recent analysis of IgA *O*-glycosylation analysis.[211] Electron transfer dissociation (ETD) uses a carrier gas to transfer the electron to the multiply charged precursor ions, and can be used in a much wider variety of MS instruments.[194] Much like ECD, this electron transfer results in the formation of radical cations that fragment preferentially across N-Cα bonds, leading to fragments of the peptide backbone with intact glycans.[181] A complication of both ETD and ECD is that fragmentation efficiency is heavily dependent on highly charged precursor ions. One solution to this problem is the use of "super-charging" reagents that favor highly charged ion formation, which was found to significantly improve analysis and site-specific glycan assignment of glycopeptides derived from both erythropoietin and a monoclonal antibody, trastuzumab.[212] ETD has also been shown to be highly efficient at identifying glycosylation sites and peptide sequences for densely-glycosylated O-glycopeptides, such as those from the mucin family.[213]

1.4.4. Importance of Synthetic Glycopeptides in Protein Glycosylation Analysis

Each of the steps outlined above has been the subject of numerous efforts at optimization and improvement to get to the current state of the art. Throughout this process of advancement, synthetic glycopeptides have served as one of the most useful tools for validation of newer methods, and analysis of their relative strengths and weaknesses.[194] For example, novel high-affinity, pan-specific antibodies were raised against an epitope derived from casein kinase II (CKII) in a study that made use of synthetic glycopeptides in several ways.[179] The authors first used synthetic glycopeptides and native chemical ligation to create a novel vaccine composed of three parts: an *O*-GlcNAc-ylated glycopeptide taken from CKII, a helper T-cell epitope and a Toll-like receptor 2 agonist as an adjuvant. This vaccine was then used for the successful production of many highly selective antibodies against the *O*-GlcNAcylated CKII peptide epitope. Synthetic glycopeptides and peptides were then used to validate each of the antibodies and show conclusively that many were selective for only the *O*-GlcNAc-containing sequence.[179] Such clear validation of a method is often easiest with well-characterized synthetic standards.

Synthetic glycopeptides' usefulness is also reflected in the development of negative-mode CID paired with ECD as a powerful method to sequence both glycan and peptide moieties of *O*-glycopeptides containing both fucose and sialic acids.[214] The authors were able to show that negative mode CID favors a clean break between the peptide and glycan, without shedding of the normally labile fucose or sialic acid residues. MS^3 spectra of the cleaved glycan precursor ion and comparison with known fragmentation patterns of commercial standards allowed easy characterization of the complete *O*-glycans. Additionally, the peptide portion of the synthetic model molecule was sequenced through positive-ion ECD MS/MS in parallel. This study shows that using synthetic standards of known structures for validation greatly expedites the process of developing novel analytical methods.[214] This study also highlights the value and convenience that synthetic standards can provide when examining detailed fragmentation patterns in MS^n spectra.

Synthetic glycopeptides were especially convenient for validating ion-mobility methods as a way to separate and analyze isomorphic glycopeptides, such as positional isomers. In one such study, the authors used two synthetic standards which shared common peptide sequences and *O*-glycans, but differed in the glycosylation site.[202] They were able to show that, although the pair of positional isomers co-eluted on RPLC, they had different drift times under the experimental conditions, and could thus be separated with ion-mobility in the gas phase. This experiment would have been much more difficult had the two standards not been so readily available and well characterized.

Quantifying specific glycopeptides in complex mixtures is an important goal for the determination of the composition of mixtures of glycoproteins. Validating methods for quantification is another area where synthetic standards have proven invaluable, such as one study that aimed to quantify *O*-GlcNAc glycopeptides.[215] With the ability to control the exact amount of a previously characterized synthetic standard, the authors were able to carry-out detailed proof-of-concept experiments that would have been impossible otherwise. Thanks to such rigorous validation of their method, the authors convincingly showed that extremely targeted ion measurements using multiple-reaction monitoring in tandem MS spectra is a great way to quantify low abundance *O*-glycosylated peptides amongst a complex background. Using this newly developed method, they went on to characterize a full length glycoprotein: glycogen synthase kinase 3 beta (GSK-3β). Thanks to the in-depth studies carried out with model molecules, the authors confidently identified three new *O*-GlcNAcylation sites on the protein and quantified the increase in *O*-glycosylation upon inhibition of a glycosylhydrolase.[215]

Isotopically labelled synthetic peptides are commonly used to quantify proteins in complex biological samples.[216] Although similarly labeled standards for glycopeptides are not as widely available, they have also proven to be extremely helpful in quantification.[217]

1.5 Conclusion

Recent advances in the fields of chemical biology and glycoproteomics have combined to reveal an unprecedented amount of information regarding protein *O*-glycosylation. New technologies for the chemical synthesis of complex glycopeptides have allowed scientists to study many new glycan structures in a systematic and meaningful way. This is leading to a new understanding of the role of *O*-glycans in many biological systems and a renewed appreciation for the complicated mechanisms governing their introduction by glycosyltransferases. Thanks to the hard work of many in the glycoproteomics field, we know more about the natural glycan structures and glycosylation patterns of human proteins than ever before and this knowledge is helping to make significant strides towards universal disease biomarkers and their use in diagnosis and treatment of human disease. Going forward, a concerted effort will be required to conquer the remaining challenges in the field. Two important goals for the future are quantifying individual glycoforms even in the presence of complex biological backgrounds and characterizing the composition of naturally produced glycoform mixtures under a variety of culture conditions and cellular contexts. Achieving this deep level of understanding will open new avenues for both the study of glycoprotein biology and their use in controlling human health.

1.6 References

1. Y. Ihara, Y. Inai, M. Ikezaki, I.-S. L. Matsui, S. Manabe and Y. Ito, in *Glycoscience: Biology and Medicine*, eds. N. Taniguchi, T. Endo, W. G. Hart, H. P. Seeberger and C.-H. Wong, Springer Japan, Tokyo, 2015, DOI: 10.1007/978-4-431-54841-6_67, pp. 1091-1099.
2. A. Varki, *Essentials of glycobiology*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 2009.
3. P. V. d. Steen, P. M. Rudd, R. A. Dwek and G. Opdenakker, *Crit Rev Biochem Mol Biol* 1998, **33**, 151-208.
4. R. G. Spiro, *Glycobiology*, 2002, **12**, 43R-56R.

5. E. P. Bennett, U. Mandel, H. Clausen, T. A. Gerken, T. A. Fritz and L. A. Tabak, *Glycobiology*, 2012, **22**, 736-756.
6. T. A. Gerken, L. Revoredo, J. J. C. Thome, L. A. Tabak, M. B. Vester-Christensen, H. Clausen, G. K. Gahlay, D. L. Jarvis, R. W. Johnson, H. A. Moniz and K. Moremen, *J Biol Chem*, 2013, **288**, 19900-19914.
7. L. Revoredo, S. Wang, E. P. Bennett, H. Clausen, K. W. Moremen, D. L. Jarvis, K. G. Ten Hagen, L. A. Tabak and T. A. Gerken, *Glycobiology*, 2015, DOI: 10.1093/glycob/cwv108.
8. H. C. Hang and C. R. Bertozzi, *Bioorg Med Chem*, 2005, **13**, 5021-5034.
9. D. T. Tran, L. Zhang, Y. Zhang, E. Tian, L. A. Earl and K. G. Ten Hagen, *J Biol Chem*, 2012, **287**, 5243-5252.
10. M. R. Pratt, H. C. Hang, K. G. Ten Hagen, J. Rarick, T. A. Gerken, L. A. Tabak and C. R. Bertozzi, *Chem Biol*, 2004, **11**, 1009-1016.
11. O. Topaz, D. L. Shurman, R. Bergman, M. Indelman, P. Ratajczak, M. Mizrachi, Z. Khamaysi, D. Behar, D. Petronius, V. Friedman, I. Zelikovic, S. Raimer, A. Metzker, G. Richard and E. Sprecher, *Nat Genet*, 2004, **36**, 579-581.
12. K. Kato, C. Jeanneau, M. A. Tarp, A. Benet-Pagès, B. Lorenz-Depiereux, E. P. Bennett, U. Mandel, T. M. Strom and H. Clausen, *J Biol Chem*, 2006, **281**, 18370-18377.
13. K. Akasaka-Manya, H. Manya, A. Nakajima, M. Kawakita and T. Endo, *J Biol Chem*, 2006, **281**, 19339-19345.
14. L. A. P. Jurado, A. Coloma and J. Cruces, *Genomics*, 1999, **58**, 171-180.
15. T. Willer, W. Amselgruber, R. Deutzmann and S. Strahl, *Glycobiology*, 2002, **12**, 771-783.
16. T. Endo, *J Biochem*, 2015, **157**, 1-12.
17. A. Yoshida, K. Kobayashi, H. Manya, K. Taniguchi, H. Kano, M. Mizuno, T. Inazu, H. Mitsuhashi, S. Takahashi, M. Takeuchi, R. Herrmann, V. Straub, B. Talim, T. Voit, H. Topaloglu, T. Toda and T. Endo, *Dev Cell*, 2001, **1**, 717-724.
18. S. Takahashi, T. Sasaki, H. Manya, Y. Chiba, A. Yoshida, M. Mizuno, H.-K. Ishida, F. Ito, T. Inazu, N. Kotani, S. Takasaki, M. Takeuchi and T. Endo, *Glycobiology*, 2001, **11**, 37-45.
19. K.-i. Inamori, T. Endo, Y. Ide, S. Fujii, J. Gu, K. Honke and N. Taniguchi, *J Biol Chem*, 2003, **278**, 43102-43109.
20. K.-i. Inamori, T. Endo, J. Gu, I. Matsuo, Y. Ito, S. Fujii, H. Iwasaki, H. Narimatsu, E. Miyoshi, K. Honke and N. Taniguchi, *J Biol Chem*, 2004, **279**, 2337-2340.
21. T. Yoshida-Moriguchi, L. P. Yu, S. H. Stalnaker, S. Davis, S. Kunz, M. Madson, M. B. A. Oldstone, H. Schachter, L. Wells and K. P. Campbell, *Science*, 2010, **327**, 88-92.
22. T. Yoshida-Moriguchi, T. Willer, M. E. Anderson, D. Venzke, T. Whyte, F. Muntoni, H. Lee, S. F. Nelson, L. Yu and K. P. Campbell, *Science*, 2013, **341**, 896-899.
23. J. L. Praissman and L. Wells, *Biochemistry*, 2014, **53**, 3066-3078.
24. M. M. Goddeeris, B. Wu, D. Venzke, T. Yoshida-Moriguchi, F. Saito, K. Matsumura, S. A. Moore and K. P. Campbell, *Nature*, 2013, **503**, 136-140.
25. T. Yoshida-Moriguchi and K. P. Campbell, *Glycobiology*, 2015, **25**, 702-713.
26. J. L. Praissman, D. H. Live, S. Wang, A. Ramiah, Z. S. Chinoy, G.-J. Boons, K. W. Moremen and L. Wells, *eLife*, 2014, **3**.
27. T. Willer, K. Inamori, D. Venzke, C. Harvey, G. Morgensen, Y. Hara, D. B. V. de Bernabe, L. P. Yu, K. M. Wright and K. P. Campbell, *elife*, 2014, **3**.

28. A. Ashikov, F. F. Buettner, B. Tiemann, R. Gerardy-Schahn and H. Bakker, *Glycobiology*, 2013, **23**, 303-309.
29. K.-i. Inamori, Y. Hara, T. Willer, M. E. Anderson, Z. Zhu, T. Yoshida-Moriguchi and K. P. Campbell, *Glycobiology*, 2013, **23**, 295-302.
30. N. Nakagawa, H. Takematsu and S. Oka, *Glycobiology*, 2013, **23**, 1066-1074.
31. Y. Luo, A. Nita-Lazar and R. S. Haltiwanger, *J Biol Chem*, 2006, **281**, 9385-9392.
32. Y. Luo and R. S. Haltiwanger, *J Biol Chem*, 2005, **280**, 11289-11294.
33. Y. Luo, K. Koles, W. Vorndam, R. S. Haltiwanger and V. M. Panin, *J Biol Chem*, 2006, **281**, 9393-9399.
34. R. J. Harris and M. W. Spellman, *Glycobiology*, 1993, **3**, 219-224.
35. D. J. Moloney, V. M. Panin, S. H. Johnston, J. Chen, L. Shao, R. Wilson, Y. Wang, P. Stanley, K. D. Irvine, R. S. Haltiwanger and T. F. Vogt, *Nature*, 2000, **406**, 369-375.
36. J. Chen, D. J. Moloney and P. Stanley, *Proc Natl Acad Sci U S A*, 2001, **98**, 13716-13721.
37. D. J. Moloney, L. H. Shair, F. M. Lu, J. Xia, R. Locke, K. L. Matta and R. S. Haltiwanger, *J Biol Chem*, 2000, **275**, 9604-9611.
38. A. Gonzalez de Peredo, D. Klein, B. Macek, D. Hess, J. Peter-Katalinic and J. Hofsteenge, *Mol Cell Proteomics*, 2002, **1**, 11-18.
39. K. Kozma, J. J. Keusch, B. Hegemann, K. B. Luther, D. Klein, D. Hess, R. S. Haltiwanger and J. Hofsteenge, *J Biol Chem*, 2006, **281**, 36742-36751.
40. T. Sato, M. Sato, K. Kiyohara, M. Sogabe, T. Shikanai, N. Kikuchi, A. Togayachi, H. Ishida, H. Ito, A. Kameyama, M. Gotoh and H. Narimatsu, *Glycobiology*, 2006, **16**, 1194-1206.
41. M. Acar, H. Jafar-Nejad, H. Takeuchi, A. Rajan, D. Ibrani, N. A. Rana, H. Pan, R. S. Haltiwanger and H. J. Bellen, *Cell*, 2008, **132**, 247-258.
42. R. Fernandez-Valdivia, H. Takeuchi, A. Samarghandi, M. Lopez, J. Leonardi, R. S. Haltiwanger and H. Jafar-Nejad, *Development*, 2011, **138**, 1925-1934.
43. H. Takeuchi, R. C. Fernández-Valdivia, D. S. Caswell, A. Nita-Lazar, N. A. Rana, T. P. Garner, T. K. Weldeghiorghis, M. A. Macnaughtan, H. Jafar-Nejad and R. S. Haltiwanger, *Proc Natl Acad Sci U S A*, 2011, **108**, 16600-16605.
44. N. A. Rana, A. Nita-Lazar, H. Takeuchi, S. Kakuda, K. B. Luther and R. S. Haltiwanger, *J Biol Chem*, 2011, **286**, 31623-31637.
45. H. Nishimura, S. Kawabata, W. Kisiel, S. Hase, T. Ikenaka, T. Takao, Y. Shimonishi and S. Iwanaga, *J Biol Chem*, 1989, **264**, 20320-20325.
46. S. Hase, H. Nishimura, S. Kawabata, S. Iwanaga and T. Ikenaka, *J Biol Chem*, 1990, **265**, 1858-1861.
47. M. K. Sethi, F. F. R. Buettner, V. B. Krylov, H. Takeuchi, N. E. Nifantiev, R. S. Haltiwanger, R. Gerardy-Schahn and H. Bakker, *J Biol Chem*, 2010, **285**, 1582-1586.
48. M. K. Sethi, F. F. R. Buettner, A. Ashikov, V. B. Krylov, H. Takeuchi, N. E. Nifantiev, R. S. Haltiwanger, R. Gerardy-Schahn and H. Bakker, *J Biol Chem*, 2012, **287**, 2739-2748.
49. T. V. Lee, M. K. Sethi, J. Leonardi, N. A. Rana, F. F. R. Buettner, R. S. Haltiwanger, H. Bakker and H. Jafar-Nejad, *PLoS Genet*, 2013, **9**, e1003547.
50. A. R. Haltom and H. Jafar-Nejad, *Glycobiology*, 2015, **25**, 1027-1042.
51. Y. Sakaidani, T. Nomura, A. Matsuura, M. Ito, E. Suzuki, K. Murakami, D. Nadano, T. Matsuda, K. Furukawa and T. Okajima, *Nat Commun*, 2011, **2**, 9.

52. Y. Sakaidani, N. Ichiyanagi, C. Saito, T. Nomura, M. Ito, Y. Nishio, D. Nadano, T. Matsuda, K. Furukawa and T. Okajima, *Biochem Biophys Res Commun*, 2012, **419**, 14-19.
53. R. Muller, A. Jenny and P. Stanley, *Plos One*, 2013, **8**, 17.
54. J. F. Alfaro, C. X. Gong, M. E. Monroe, J. T. Aldrich, T. R. W. Clauss, S. O. Purvine, Z. H. Wang, D. G. Camp, J. Shabanowitz, P. Stanley, G. W. Hart, D. F. Hunt, F. Yang and R. D. Smith, *Proc Natl Acad Sci U S A*, 2012, **109**, 7280-7285.
55. D. J. Gill, H. Clausen and F. Bard, *Trends Cell Biol*, 2011, **21**, 149-158.
56. K. W. Moremen, M. Tiemeyer and A. V. Nairn, *Nat Rev Mol Cell Biol*, 2012, **13**, 448-462.
57. A. P. Corfield, *Biochim Biophys Acta*, 2015, **1850**, 236-252.
58. J. Perez-Vilar and R. L. Hill, *Journal of Biological Chemistry*, 1999, **274**, 31751-31754.
59. J.-L. Desseyn, D. Tetaert and V. Gouyer, *Gene*, 2008, **410**, 215-222.
60. A. P. Corfield and M. Berry, *Trends Biochem Sci*, 2015, **40**, 351-359.
61. K. Kozarsky, D. Kingsley and M. Krieger, *Proc Natl Acad Sci U S A*, 1988, **85**, 4335-4339.
62. E. B. Maryon, S. A. Molloy and J. H. Kaplan, *J Biol Chem*, 2007, **282**, 20376-20387.
63. R. Shogren, T. A. Gerken and N. Jentoft, *Biochemistry*, 1989, **28**, 5525-5536.
64. J. J. Barchi, *Biopolymers*, 2013, **99**, 713-723.
65. Z. Xu and A. Weiss, *Nat Immunol*, 2002, **3**, 764-771.
66. K. W. Wagner, E. A. Punnoose, T. Januario, D. A. Lawrence, R. M. Pitti, K. Lancaster, D. Lee, M. von Goetz, S. F. Yee, K. Totpal, L. Huw, V. Katta, G. Cavet, S. G. Hymowitz, L. Amler and A. Ashkenazi, *Nat Med*, 2007, **13**, 1070-1077.
67. T. Lang, G. C. Hansson and T. Samuelsson, *Proc Natl Acad Sci U S A*, 2007, **104**, 16209-16214.
68. H. J. Gabius, *The Sugar Code: Fundamentals of Glycosciences*, Wiley, 2009.
69. M. Amado, Q. Yan, E. M. Comelli, B. E. Collins and J. C. Paulson, *J Biol Chem*, 2004, **279**, 36689-36697.
70. D. T. Tran and K. G. Ten Hagen, *J Biol Chem*, 2013, **288**, 6921-6929.
71. M. Kilcoyne, J. Q. Gerlach, R. Gough, M. E. Gallagher, M. Kane, S. D. Carrington and L. Joshi, *Anal Chem*, 2012, **84**, 3330-3338.
72. C. Robbe, C. Capon, E. Maes, M. Rousset, A. Zweibaum, J.-P. Zanetta and J.-C. Michalski, *J Biol Chem*, 2003, **278**, 46337-46348.
73. C. ROBBE, C. CAPON, B. CODDEVILLE and J.-C. MICHALSKI, *Biochem J*, 2004, **384**, 307-316.
74. L. A. Earl, S. G. Bi and L. G. Baum, *J Biol Chem*, 2010, **285**, 2242-2254.
75. M. C. Clark, L. G. Baum and N. Y. A. S. Annals, *Glycobiology of the Immune Response*, 2012, **1253**, 58-67.
76. M. T. Boskovski, S. Yuan, N. B. Pedersen, C. K. Goth, S. Makova, H. Clausen, M. Brueckner and M. K. Khokha, *Nature*, 2013, **504**, 10.1038/nature12723.
77. J. C. Brazil, R. P. Liu, R. Sumagin, K. N. Kolegraff, A. Nusrat, R. D. Cummings, C. A. Parkos and N. A. Louis, *J Immunol*, 2013, **191**, 4804-4817.
78. Y. Mori, K. Akita, M. Yashiro, T. Sawada, K. Hirakawa, T. Murata and H. Nakada, *J Biol Chem*, 2015, **290**, 26125-26140.
79. K. T. B. G. Schjoldager and H. Clausen, *Biochim Biophys Acta, Gen Subj*, 2012, **1820**, 2079-2094.
80. C. K. Goth, A. Halim, S. A. Khetarpal, D. J. Rader, H. Clausen and K. T.-B. G. Schjoldager, *Proc Natl Acad Sci U S A*, 2015, DOI: 10.1073/pnas.1511175112.

81. G. Opdenakker, P. M. Rudd, M. Wormald, R. A. Dwek and J. Van Damme, *FASEB J*, 1995, **9**, 453-457.
82. C. Dong, A. Chua, S. Ganguly, A. M. Krensky and C. Clayberger, *J Immunol Methods*, 2005, **302**, 136-144.
83. D. H. Live, L. J. Williams, S. D. Kuduk, J. B. Schwarz, P. W. Glunz, X.-T. Chen, D. Sames, R. A. Kumar and S. J. Danishefsky, *Proc Natl Acad Sci U S A*, 1999, **96**, 3489-3493.
84. W. L. Bigbee, R. G. Langlois, M. Vanderlaan and R. H. Jensen, *J Immunol*, 1984, **133**, 3149-3155.
85. K. Dill, R. D. Carter, J. M. Lacombe and A. A. Pavia, *Carbohydr Res*, 1986, **152**, 217-228.
86. K. Dill, S. H. Hu, E. Berman, A. A. Pavia and J. M. Lacombe, *J Protein Chem*, 1990, **9**, 129-136.
87. O. Schuster, G. Klich, V. Sinnwell, H. Kränz, H. Paulsen and B. Meyer, *J Biomol NMR*, 1999, **14**, 33-45.
88. G. A. Naganagowda, T. E. L. Gururaja, J. Satyanarayana and M. J. Levine, *J Pept Res*, 1999, **54**, 290-310.
89. F. Corzana, J. H. Busto, G. Jiménez-Osés, M. García de Luis, J. L. Asensio, J. Jiménez-Barbero, J. M. Peregrina and A. Avenoza, *J Am Chem Soc*, 2007, **129**, 9458-9467.
90. D. Madariaga, N. Martinez-Saez, V. J. Somovilla, L. Garcia-Garcia, M. A. Berbis, J. Valero-Gonzalez, S. Martin-Santamaria, R. Hurtado-Guerrero, J. L. Asensio, J. Jimenez-Barbero, A. Avenoza, J. H. Busto, F. Corzana and J. M. Peregrina, *Chem Eur J*, 2014, **20**, 12616-12627.
91. D. M. Coltart, A. K. Royyuru, L. J. Williams, P. W. Glunz, D. Sames, S. D. Kuduk, J. B. Schwarz, X.-T. Chen, S. J. Danishefsky and D. H. Live, *J Am Chem Soc*, 2002, **124**, 9833-9844.
92. V. Apostolopoulos and I. F. McKenzie, *Crit Rev Immunol*, 1994, **14**, 293-309.
93. L. Kirnarsky, O. Prakash, S. M. Vogen, M. Nomoto, M. A. Hollingsworth and S. Sherman, *Biochemistry*, 2000, **39**, 12076-12082.
94. P. Dokurno, P. A. Bates, H. A. Band, L. M. D. Stewart, J. M. Lally, J. M. Burchell, J. Taylor-Papadimitriou, D. Snary, M. J. E. Sternberg and P. S. Freemont, *J Mol Biol*, 1998, **284**, 713-728.
95. J. Schuman, A. P. Campbell, R. R. Koganty and B. M. Longenecker, *J Pept Res*, 2003, **61**, 91-108.
96. S. Dziadek, C. Griesinger, H. Kunz and U. M. Reinscheid, *Chem Eur J*, 2006, **12**, 4981-4993.
97. T. Matsushita, N. Ohyabu, N. Fujitani, K. Naruchi, H. Shimizu, H. Hinou and S. I. Nishimura, *Biochemistry*, 2013, **52**, 402-414.
98. S. Nath and P. Mukherjee, *Trends Mol Med*, 2014, **20**, 332-342.
99. R. D. Wright and D. Cooper, *Glycobiology*, 2014, **24**, 1242-1251.
100. A. Leppänen, P. Mehta, Y.-B. Ouyang, T. Ju, J. Helin, K. L. Moore, I. van Die, W. M. Canfield, R. P. McEver and R. D. Cummings, *J Biol Chem*, 1999, **274**, 24838-24848.
101. A. Leppänen, S. P. White, J. Helin, R. P. McEver and R. D. Cummings, *J Biol Chem*, 2000, **275**, 39569-39578.
102. J. Finne, T. Krusius, R. K. Margolis and R. U. Margolis, *J Biol Chem*, 1979, **254**, 10295-10300.
103. W. Chai, C.-T. Yuen, H. Kogelberg, R. A. Carruthers, R. U. Margolis, T. Feizi and A. M. Lawson, *Eur J Biochem*, 1999, **263**, 879-888.
104. M. Lommel and S. Strahl, *Glycobiology*, 2009, **19**, 816-828.
105. A. Chiba, K. Matsumura, H. Yamada, T. Inazu, T. Shimizu, S. Kusunoki, I. Kanazawa, A. Kobata and T. Endo, *J Biol Chem*, 1997, **272**, 2156-2162.

106. C.-T. Yuen, W. Chai, R. W. Loveless, A. M. Lawson, R. U. Margolis and T. Feizi, *J Biol Chem*, 1997, **272**, 8924-8931.
107. T. Sasaki, H. Yamada, K. Matsumura, T. Shimizu, A. Kobata and T. Endo, *Biochim Biophys Acta, Gen Subj*, 1998, **1425**, 599-606.
108. N. R. Smalheiser, S. M. Haslam, M. Sutton-Smith, H. R. Morris and A. Dell, *J Biol Chem*, 1998, **273**, 23698-23703.
109. M. B. Vester-Christensen, A. Halim, H. J. Joshi, C. Steentoft, E. P. Bennett, S. B. Levery, S. Y. Vakhrushev and H. Clausen, *Proc Natl Acad Sci U S A*, 2013, **110**, 21018-21023.
110. M. Lommel, P. R. Winterhalter, T. Willer, M. Dahlhoff, M. R. Schneider, M. F. Bartels, I. Renner-Müller, T. Ruppert, E. Wolf and S. Strahl, *Proc Natl Acad Sci U S A*, 2013, **110**, 21024-21029.
111. T. Martinez, D. Pace, L. Brady, M. Gerhart and A. Balland, *J Chromatogr A*, 2007, **1156**, 183-187.
112. C. Bleckmann, H. Geyer, A. Lieberoth, F. Splittstoesser, Y. Liu, T. Feizi, M. Schachner, R. Kleene, V. Reinhold and R. Geyer, *Journal*, 2009, **390**, 627.
113. S. Pacharra, F.-G. Hanisch and I. Breloy, *J Proteome Res*, 2012, **11**, 3955-3964.
114. K. L. Abbott, R. T. Matthews and M. Pierce, *J Biol Chem*, 2008, **283**, 33026-33035.
115. S. Pacharra, F.-G. Hanisch, M. Mühlenhoff, A. Faissner, U. Rauch and I. Breloy, *J Proteome Res*, 2013, **12**, 1764-1771.
116. K. Kanekiyo, K.-i. Inamori, S. Kitazume, K. Sato, J. Maeda, M. Higuchi, Y. Kizuka, H. Korekane, I. Matsuo, K. Honke and N. Taniguchi, *J Neurosci*, 2013, **33**, 10037-10047.
117. S. Yaji, H. Manya, N. Nakagawa, H. Takematsu, T. Endo, R. Kannagi, T. Yoshihara, M. Asano and S. Oka, *Glycobiology*, 2015, **25**, 376-385.
118. L. Wells, *J Biol Chem*, 2013, **288**, 6930-6935.
119. A. G. Toledo, M. Raducu, J. Cruces, J. Nilsson, A. Halim, G. Larson, U. Ruetschi and A. Grahn, *Glycobiology*, 2012, **22**, 1413-1423.
120. Y. Hara, M. Kanagawa, S. Kunz, T. Yoshida-Moriguchi, J. S. Satz, Y. M. Kobayashi, Z. Zhu, S. J. Burden, M. B. A. Oldstone and K. P. Campbell, *Proc Natl Acad Sci U S A*, 2011, **108**, 17426-17431.
121. H. Manya, T. Suzuki, K. Akasaka-Manya, H.-K. Ishida, M. Mizuno, Y. Suzuki, T. Inazu, N. Dohmae and T. Endo, *J Biol Chem*, 2007, **282**, 20200-20206.
122. S. H. Stalnaker, S. Hashmi, J.-M. Lim, K. Aoki, M. Porterfield, G. Gutierrez-Sanchez, J. Wheeler, J. M. Ervasti, C. Bergmann, M. Tiemeyer and L. Wells, *J Biol Chem*, 2010, **285**, 24882-24891.
123. J. Nilsson, J. Nilsson, G. Larson and A. Grahn, *Glycobiology*, 2010, **20**, 1160-1169.
124. R. Harrison, P. G. Hitchen, M. Panico, H. R. Morris, D. Mekhaiel, R. J. Pleass, A. Dell, J. E. Hewitt and S. M. Haslam, *Glycobiology*, 2012, **22**, 662-675.
125. I. Breloy, T. Schwientek, B. Gries, H. Razawi, M. Macht, C. Albers and F.-G. Hanisch, *J Biol Chem*, 2008, **283**, 18832-18840.
126. S. H. Stalnaker, K. Aoki, J.-M. Lim, M. Porterfield, M. Liu, J. S. Satz, S. Buskirk, Y. Xiong, P. Zhang, K. P. Campbell, H. Hu, D. Live, M. Tiemeyer and L. Wells, *J Biol Chem*, 2011, **286**, 21180-21190.

127. D. E. Michele, R. Barresi, M. Kanagawa, F. Saito, R. D. Cohn, J. S. Satz, J. Dollar, I. Nishino, R. I. Kelley, H. Somer, V. Straub, K. D. Mathews, S. A. Moore and K. P. Campbell, *Nature*, 2002, **418**, 417-421.
128. J. Voglmeir, S. Kaloo, N. Laurent, Marco M. Meloni, L. Bohlmann, Iain B. H. Wilson and Sabine L. Flitsch, *Biochem J*, 2011, **436**, 447-455.
129. K. Akasaka-Manya, H. Manya, M. Mizuno, T. Inazu and T. Endo, *Biochem Biophys Res Commun*, 2011, **410**, 632-636.
130. K. Inamori, T. Yoshida-Moriguchi, Y. Hara, M. E. Anderson, L. P. Yu and K. P. Campbell, *Science*, 2012, **335**, 93-96.
131. K. F. Mo, T. Fang, S. H. Stalnaker, P. S. Kirby, M. Liu, L. Wells, M. Pierce, D. H. Live and G. J. Boons, *J Am Chem Soc*, 2011, **133**, 14418-14430.
132. D. T. Tran, J. M. Lim, M. Liu, S. H. Stalnaker, L. Wells, K. G. Ten Hagen and D. Live, *J Biol Chem*, 2012, **287**, 20967-20974.
133. M. A. Liu, A. Borgert, G. Barany and D. Live, *Biopolymers*, 2008, **90**, 358-368.
134. A. W. Barb, A. J. Borgert, M. Liu, G. Barany and D. Live, *Methods Enzymol*, 2010, **478**, 365-388.
135. C. B. Taylor, M. F. Talib, C. McCabe, L. Bu, W. S. Adney, M. E. Himmel, M. F. Crowley and G. T. Beckham, *J Biol Chem*, 2012, **287**, 3147-3155.
136. L. Chen, M. R. Drake, M. G. Resch, E. R. Greene, M. E. Himmel, P. K. Chaffey, G. T. Beckham and Z. Tan, *Proc Natl Acad Sci U S A*, 2014, **111**, 7612-7617.
137. X. Guan, P. K. Chaffey, C. Zeng, E. R. Greene, L. Chen, M. R. Drake, C. Chen, A. Groobman, M. G. Resch, M. E. Himmel, G. T. Beckham and Z. Tan, *Chem Sci*, 2015, **6**, 7185-7189.
138. E. J. Kentzer, A. Buko, G. Menon and V. K. Sarin, *Biochem Biophys Res Commun*, 1990, **171**, 401-406.
139. H. Nishimura, T. Takao, S. Hase, Y. Shimonishi and S. Iwanaga, *J Biol Chem*, 1992, **267**, 17520-17525.
140. S. G. Schiffer, S. Foley, A. Kaffashan, X. Hronowski, A. E. Zichittella, C.-Y. Yeo, K. Miatkowski, H. B. Adkins, B. Damon, M. Whitman, D. Salomon, M. Sanicola and K. P. Williams, *J Biol Chem*, 2001, **276**, 37769-37778.
141. T. Okajima and K. D. Irvine, *Cell*, 2002, **111**, 893-904.
142. R. Rampal, K. B. Luther and R. S. Haltiwanger, *Curr Mol Med*, 2007, **7**, 427-445.
143. C. I. Chen, J. J. Keusch, D. Klein, D. Hess, J. Hofsteenge and H. Gut, *EMBO J*, 2012, **31**, 3183-3197.
144. L. W. Wang, M. Dlugosz, R. P. T. Somerville, M. Raed, R. S. Haltiwanger and S. S. Apte, *J Biol Chem*, 2007, **282**, 17024-17031.
145. L. M. Ricketts, M. Dlugosz, K. B. Luther, R. S. Haltiwanger and E. M. Majerus, *J Biol Chem*, 2007, **282**, 17014-17023.
146. K. B. Luther and R. S. Haltiwanger, *Int J Biochem Cell Biol*, 2009, **41**, 1011-1024.
147. S. A. Rabbani, A. P. Mazar, S. M. Bernier, M. Haq, I. Bolivar, J. Henkin and D. Goltzman, *J Biol Chem*, 1992, **267**, 14151-14156.
148. Y.-T. Yan, J.-J. Liu, Y. Luo, C. E, R. S. Haltiwanger, C. Abate-Shen and M. M. Shen, *Mol Cell Biol*, 2002, **22**, 4439-4449.
149. S. Shi, C. Ge, Y. Luo, X. Hou, R. S. Haltiwanger and P. Stanley, *J Biol Chem*, 2007, **282**, 20133-20141.

150. H. Takeuchi and R. S. Haltiwanger, *Biochem Biophys Res Commun*, 2014, **453**, 235-242.
151. R. Kopan and M. X. G. Ilagan, *Cell*, 2009, **137**, 216-233.
152. P. Stanley and T. Okajima, in *Notch Signaling*, ed. R. Kopan, Elsevier Academic Press Inc, San Diego, 2010, vol. 92, pp. 131-164.
153. K. Bruckner, L. Perez, H. Clausen and S. Cohen, *Nature*, 2000, **406**, 411-415.
154. A. Xu, N. Haines, M. Dlugosz, N. A. Rana, H. Takeuchi, R. S. Haltiwanger and K. D. Irvine, *J Biol Chem*, 2007, **282**, 35153-35162.
155. I. Rebay, R. J. Fleming, R. G. Fehon, L. Cherbas, P. Cherbas and S. Artavanis-Tsakonas, *Cell*, 1991, **67**, 687-699.
156. P. Taylor, H. Takeuchi, D. Sheppard, C. Chillakuri, S. M. Lea, R. S. Haltiwanger and P. A. Handford, *Proc Natl Acad Sci U S A*, 2014, **111**, 7290-7295.
157. X. Hou, Y. Tashima and P. Stanley, *J Biol Chem*, 2012, **287**, 474-483.
158. V. C. Luca, K. M. Jude, N. W. Pierce, M. V. Nachury, S. Fischer and K. C. Garcia, *Science*, 2015, **347**, 847-853.
159. K. Hiruma-Shimizu, K. Hosoguchi, Y. Liu, N. Fujitani, T. Ohta, H. Hinou, T. Matsushita, H. Shimizu, T. Feizi and S.-I. Nishimura, *J Am Chem Soc*, 2010, **132**, 14857-14865.
160. S. Hayakawa, R. Koide, H. Hinou and S.-I. Nishimura, *Biochemistry*, 2016, DOI: 10.1021/acs.biochem.5b01284.
161. R. Rampal, J. F. Arboleda-Velasquez, A. Nita-Lazar, K. S. Kosik and R. S. Haltiwanger, *J Biol Chem*, 2005, **280**, 32133-32140.
162. V. M. Panin, L. Shao, L. Lei, D. J. Moloney, K. D. Irvine and R. S. Haltiwanger, *J Biol Chem*, 2002, **277**, 29945-29952.
163. J. Muller, N. A. Rana, K. Serth, S. Kakuda, R. S. Haltiwanger and A. Gossler, *Plos One*, 2014, **9**, 9.
164. K. Serth, K. Schuster-Gossler, E. Kremmer, B. Hansen, B. Marohn-Kohn and A. Gossler, *Plos One*, 2015, **10**, 21.
165. J. C. Adams and R. P. Tucker, *Dev Dyn*, 2000, **218**, 280-299.
166. S. A. J. Lesnik Oberstein, M. Kriek, S. J. White, M. E. Kalf, K. Szuhai, J. T. den Dunnen, M. H. Breuning and R. C. M. Hennekam, *Am J Hum Genet*, 2006, **79**, 562-566.
167. S. Hase, S. Kawabata, H. Nishimura, H. Takeya, T. Sueyoshi, T. Miyata, S. Iwanaga, T. Takao, Y. Shimonishi and T. Ikenaka, *J Biochem*, 1988, **104**, 867-868.
168. H. Takeuchi, J. Kantharia, M. K. Sethi, H. Bakker and R. S. Haltiwanger, *J Biol Chem*, 2012, **287**, 33934-33944.
169. J. Leonardi, R. Fernandez-Valdivia, Y.-D. Li, A. A. Simcox and H. Jafar-Nejad, *Development*, 2011, **138**, 3569-3578.
170. H. Jafar-Nejad, J. Leonardi and R. Fernandez-Valdivia, *Glycobiology*, 2010, **20**, 931-949.
171. A. Matsuura, M. Ito, Y. Sakaidani, T. Kondo, K. Murakami, K. Furukawa, D. Nadano, T. Matsuda and T. Okajima, *J Biol Chem*, 2008, **283**, 35486-35495.
172. R. Shaheen, M. Aglan, K. Keppler-Noreuil, E. Faqeih, S. Ansari, K. Horton, A. Ashour, Maha S. Zaki, F. Al-Zahrani, Anna M. Cueto-González, G. Abdel-Salam, S. Temtamy and Fowzan S. Alkuraya, *Am J Hum Genet*, 2013, **92**, 598-604.
173. M. Ogawa, S. Sawaguchi, T. Kawai, D. Nadano, T. Matsuda, H. Yagi, K. Kato, K. Furukawa and T. Okajima, *J Biol Chem*, 2015, **290**, 2137-2149.

174. P. Pompach, Z. Brnakova, M. Sanda, J. Wu, N. Edwards and R. Goldman, *Mol Cell Proteomics*, 2013, **12**, 1281-1293.
175. D. Kolarich, P. H. Jensen, F. Altmann and N. H. Packer, *Nat Protocols*, 2012, **7**, 1285-1298.
176. J. P. Skinner, L. Chi, P. F. Ozeata, C. S. Ramsay, R. L. O'Hara, B. B. Calfin and S. Y. Tetin, *Anal Chem*, 2012, **84**, 1172-1177.
177. S. Ongay, A. Boichenko, N. Govorukhina and R. Bischoff, *J Sep Sci*, 2012, **35**, 2341-2372.
178. L. Wells, K. Vosseller, R. N. Cole, J. M. Cronshaw, M. J. Matunis and G. W. Hart, *Mol Cell Proteomics*, 2002, **1**, 791-804.
179. C. F. Teo, S. Ingale, M. A. Wolfert, G. A. Elsayed, L. G. Nöt, J. C. Chatham, L. Wells and G.-J. Boons, *Nat Chem Biol*, 2010, **6**, 338-343.
180. W. R. Alley, B. F. Mann and M. V. Novotny, *Chem Rev*, 2013, **113**, 2668-2732.
181. Z. Zhu and H. Desaire, *Annu Rev Anal Chem*, 2015, **8**, 463-483.
182. E. Ruiz-May, S. Hucko, K. J. Howe, S. Zhang, R. W. Sherwood, T. W. Thannhauser and J. K. C. Rose, *Mol Cell Proteomics*, 2014, **13**, 566-579.
183. K. Vosseller, J. C. Trinidad, R. J. Chalkley, C. G. Specht, A. Thalhammer, A. J. Lynn, J. O. Snedecor, S. Guan, K. F. Medzihradszky, D. A. Maltby, R. Schoepfer and A. L. Burlingame, *Mol Cell Proteomics*, 2006, **5**, 923-934.
184. S. Y. Vakhrushev, C. Steentoft, M. B. Vester-Christensen, E. P. Bennett, H. Clausen and S. B. Levery, *Mol Cell Proteomics*, 2013, **12**, 932-944.
185. H. Zhang, X.-j. Li, D. B. Martin and R. Aebersold, *Nat Biotech*, 2003, **21**, 660-666.
186. R. Chen, X. Jiang, D. Sun, G. Han, F. Wang, M. Ye, L. Wang and H. Zou, *J Proteome Res*, 2009, **8**, 651-661.
187. J. Gonzalez, T. Takao, H. Hori, V. Besada, R. Rodriguez, G. Padron and Y. Shimonishi, *Anal Chem*, 1992, **205**, 151-158.
188. D. F. Zielinska, F. Gnad, J. R. Wiśniewski and M. Mann, *Cell*, 2010, **141**, 897-907.
189. G. Palmisano, M. N. Melo-Braga, K. Engholm-Keller, B. L. Parker and M. R. Larsen, *J Proteome Res*, 2012, **11**, 1949-1957.
190. E. Klement, Z. Lipinszki, Z. Kupihár, A. Udvardy and K. F. Medzihradszky, *J Proteome Res*, 2010, **9**, 2200-2206.
191. J. Nilsson, U. Ruetschi, A. Halim, C. Hesse, E. Carlsohn, G. Brinkmalm and G. Larson, *Nat Meth*, 2009, **6**, 809-811.
192. Y. Wang, M. Liu, L. Xie, C. Fang, H. Xiong and H. Lu, *Anal Chem*, 2014, **86**, 2057-2064.
193. M. Chen, Y. Lu, Q. Ma, L. Guo and Y.-Q. Feng, *Analyst*, 2009, **134**, 2158-2164.
194. G. Zauner, R. P. Kozak, R. A. Gardner, D. L. Fernandes, A. M. Deelder and M. Wuhrer, *Biol Chem*, 2012, **393**, 687-708.
195. J. Ma, L. Zhang, Z. Liang, Y. Shan and Y. Zhang, *Trends Analyt Chem*, 2011, **30**, 691-702.
196. M. N. Christiansen, D. Kolarich, H. Nevalainen, N. H. Packer and P. H. Jensen, *Anal Chem*, 2010, **82**, 3500-3509.
197. B. Buszewski and S. Noga, *Anal Bioanal Chem*, 2011, **402**, 231-247.
198. G. Zauner, A. M. Deelder and M. Wuhrer, *Electrophoresis*, 2011, **32**, 3456-3466.
199. C. West, C. Elfakir and M. Lafosse, *J Chromatogr A*, 2010, **1217**, 3201-3216.
200. S. Hua, C. C. Nwosu, J. S. Strum, R. R. Seipert, H. J. An, A. M. Zivkovic, J. B. German and C. B. Lebrilla, *Anal Bioanal Chem*, 2011, **403**, 1291-1302.
201. R. Guevremont, *J Chromatogr A*, 2004, **1058**, 3-19.

202. A. J. Creese and H. J. Cooper, *Anal Chem*, 2012, **84**, 2597-2601.
203. H. Desaire, *Mol Cell Proteomics*, 2013, **12**, 893-901.
204. B. L. Parker, M. Thaysen-Andersen, N. Solis, N. E. Scott, M. R. Larsen, M. E. Graham, N. H. Packer and S. J. Cordwell, *J Proteome Res*, 2013, **12**, 5791-5800.
205. Z. Darula, J. Sherman and K. F. Medzihradszky, *Mol Cell Proteomics*, 2012, **11**, O111 016774.
206. R. Goldman and M. Sanda, *Proteomics Clin Appl*, 2015, **9**, 17-32.
207. D. Wang, M. Hincapie, T. Rejtar and B. L. Karger, *Anal Chem*, 2011, **83**, 2029-2037.
208. B. Maček, J. Hofsteenge and J. Peter-Katalinić, *Rapid Commun Mass Spectrom*, 2001, **15**, 771-777.
209. J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning and M. Mann, *Nat Meth*, 2007, **4**, 709-712.
210. M. Mormann and J. Peter-Katalinić, *Rapid Commun Mass Spectrom*, 2003, **17**, 2208-2214.
211. K. Takahashi, A. D. Smith, K. Poulsen, M. Kilian, B. A. Julian, J. Mestecky, J. Novak and M. B. Renfrow, *J Proteome Res*, 2012, **11**, 692-702.
212. J. P. Williams, S. Pringle, K. Richardson, L. Gethings, J. P. C. Vissers, M. De Cecco, S. Houel, A. B. Chakraborty, Y. Q. Yu, W. Chen and J. M. Brown, *Rapid Commun Mass Spectrom*, 2013, **27**, 2383-2390.
213. M. Thaysen-Andersen, B. L. Wilkinson, R. J. Payne and N. H. Packer, *Electrophoresis*, 2011, **32**, 3536-3545.
214. K. Deguchi, H. Ito, T. Baba, A. Hirabayashi, H. Nakagawa, M. Fumoto, H. Hinou and S.-I. Nishimura, *Rapid Commun Mass Spectrom*, 2007, **21**, 691-698.
215. J. J. P. Maury, D. Ng, X. Bi, M. Bardor and A. B.-H. Choo, *Anal Chem*, 2014, **86**, 395-402.
216. M. A. Kuzyk, D. Smith, J. Yang, T. J. Cross, A. M. Jackson, D. B. Hardie, N. L. Anderson and C. H. Borchers, *Mol Cell Proteomics*, 2009, **8**, 1860-1877.
217. N. Leymarie, P. J. Griffin, K. Jonscher, D. Kolarich, R. Orlando, M. McComb, J. Zaia, J. Aguilan, W. R. Alley, F. Altmann, L. E. Ball, L. Basumallick, C. R. Bazemore-Walker, H. Behnken, M. A. Blank, K. J. Brown, S.-C. Bunz, C. W. Cairo, J. F. Cipollo, R. Daneshfar, H. Desaire, R. R. Drake, E. P. Go, R. Goldman, C. Gruber, A. Halim, Y. Hathout, P. J. Hensbergen, D. M. Horn, D. Hurum, W. Jabs, G. Larson, M. Ly, B. F. Mann, K. Marx, Y. Mechref, B. Meyer, U. Möginger, C. Neusüβ, J. Nilsson, M. V. Novotny, J. O. Nyalwidhe, N. H. Packer, P. Pompach, B. Reiz, A. Resemann, J. S. Rohrer, A. Ruthenbeck, M. Sanda, J. M. Schulz, U. Schweiger-Hufnagel, C. Sihlbom, E. Song, G. O. Staples, D. Suckau, H. Tang, M. Thaysen-Andersen, R. I. Viner, Y. An, L. Valmu, Y. Wada, M. Watson, M. Windwarder, R. Whittal, M. Wuhrer, Y. Zhu and C. Zou, *Mol Cell Proteomics*, 2013, **12**, 2935-2951.

Chapter 2

Specificity of *O*-Glycosylation in Enhancing the Stability and Cellulose Binding Affinity of Family 1 Carbohydrate-Binding Modules

2.1 – Introduction

Terrestrial plant biomass is primarily degraded in nature by fungi and bacteria, which secrete synergistic cocktails of enzymes that work in concert to degrade polysaccharides and sometimes lignin.[1-4] In many cases, the enzymes used by these organisms are multi-modular consisting of one or more catalytic domains of various function[2-8] linked to a carbohydrate-binding module (CBM) that targets plant cell wall polysaccharides through specific recognition mechanisms.[9] To date, 67 families of CBMs have been discovered[10] and many of these families contain members important in biomass depolymerization. Nearly all known CBM-bearing lignocellulose-degrading enzymes from fungi are Family 1 CBMs,[10] which are small proteins that consist of less than 40 amino acids. Kraulis *et al*. solved the first Family 1 CBM structure from the well-characterized glycoside hydrolase (GH) Family 7 cellobiohydrolase from the fungus *Trichoderma reesei* (*Hypocrea jecorina*), or *Tr*Cel7A[11]. The structure of the *Tr*Cel7A CBM revealed a β-sheet rich structure with two disulfide bridges and a flat face decorated with aromatic and polar residues that forms the putative binding face for adsorption to the hydrophobic face of crystalline cellulose microfibrils (Figure 2.1).[11-15]

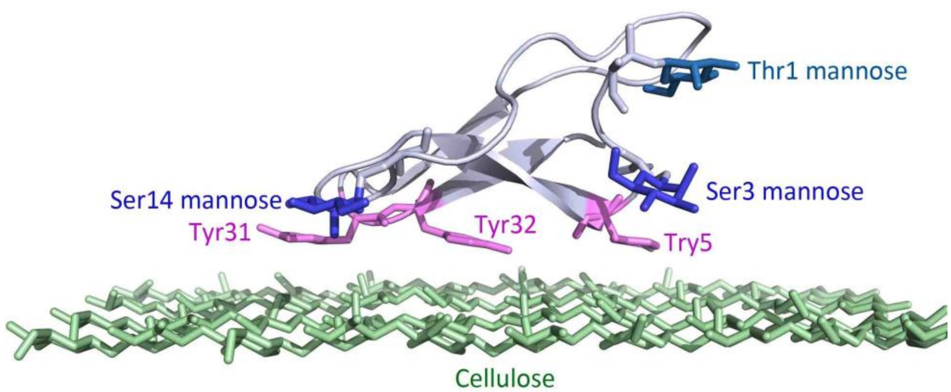


Figure 2.1 - The NMR structure of the Family 1 CBM and the top layer of cellulose. The tyrosine residues are shown in purple. The O-linked mannoses are shown in cyan and blue.

Glycosylation is an important heterogeneous post-translational modification in fungal enzymes that degrade biomass.[14,16] To date, few studies have been conducted to examine glycosylation in secreted fungal enzymes to determine the extent and factors that control it including growth conditions and extracellular glycan-trimming enzymes. Catalytic domains can exhibit both *N*- and *O*-linked glycans,[17,18]whereas the linkers connecting enzymatic domains to CBMs are decorated with *O*-linked glycosylation, which has long been attributed to protease protection[19] and more recently implicated in substrate binding.[20] For *Tr*Cel7A, Harrison *et al.* published the original characterization of the glycosylation pattern on the *Tr*Cel7A linker.[21] Notably, the last five residues analyzed in their study (TQSHY) form the *N*-terminus of the CBM, and the threonine and serine residues (Thr1, Ser3, respectively) were shown to both natively exhibit mannosylation.[21] Given that these residues are highly conserved, this is likely a common feature of all Family 1 CBMs.[22] It is also possible that mannosylation may be natively found on the highly conserved Ser14 residue, but this has not been experimentally characterized to our knowledge. We recently employed free energy calculations to predict that the mannosylation will improve the CBM binding affinity to crystalline cellulose.[23] Our results suggested that the glycan structure, their locations, and the number of occupied glycosylation sites will impact the affinity of CBMs for crystalline cellulose.[23]

To quantitatively assess the impact of glycosylation on Family 1 CBMs, here we present a systematic experimental study of proteolytic stability, thermostability, and cellulose binding affinity of a library of Family 1 CBM glycoforms. As glycosylation depends on host and culture conditions, multiple glycoforms of the same protein are often observed in biological production systems, which are often difficult to separate. Thus, a new method for the routine production of specific Family 1 CBM glycoforms was developed using emerging tools in chemical glycoprotein synthesis.[24-26] Recent advances in this field have made it possible to generate a variety of homogeneous glycoforms for structure-function studies.[27] Since chemical glycosylation is not dictated by the amino acid sequences of proteins, it allows the generation of homogeneous glycoforms, thus enabling us to assess if the effects on the stability and function of the *Tr*Cel7A CBM are general effects of glycosylation or are specific to certain sites and sugar moieties.

2.2 – Results

2.2.1 Chemical synthesis of CBM glycoforms

9-Fluorenylmethoxycarbonyl (Fmoc)-based solid-phase peptide synthesis (SPPS) was employed for the synthesis of glycosylated CBM variants because of its compatibility with acid-sensitive glycosidic bonds.[12,28] Our synthesis started with the optimization of the conditions of most of the steps involved in the SPPS. By using a preloaded trityl resin, Fmoc-Leu-NovaSyn® TGT, a pseudoproline dipeptide Fmoc-Ala-Ser(psiMe,Mepro)-OH during the SPPS process and prolonged coupling time, we could efficiently prepare the glycopeptide library (Figure 2.2).[29] To make the preparation of folded CBMs less labor intensive, we next examined the feasibility of obtaining the correctly folded CBM glycoforms via a one-pot deprotection/folding sequence. To this end, we found that all acetyl protecting groups in the crude glycopeptides can be completely removed in less than 30 min using 5% hydrazine. Importantly, we found that *O*-mannosylation at Thr1, Ser3, and Ser14 sites does not impair CBM folding. As

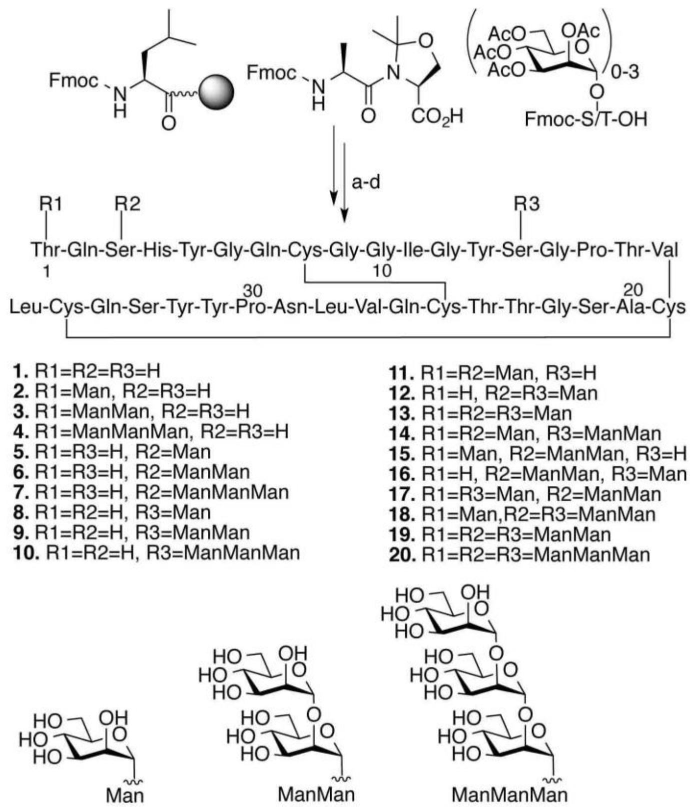


Figure 2.2 - One-pot synthesis of the *Tr*Cel7A CBM. Reagents and conditions: (a) HATU, *N*,*N*-diisopropylethylamine, DMSO; piperidine, 1,8-diazabicyclo[5.4.0]undec-7-ene, DMSO; (b) TFA/H_2O/triisopropylsilane (95:2.5:2.5); (c) NH_2NH_2, H_2O (for glycosylated CBMs); (d) Tris-acetate, L-glutathione reduced, L-glutathione oxidized, H_2O, pH 8.2.

anticipated, folding of the deprotected glycopeptides can be initiated by direct dilution of the deprotection mixture in a mixed glutathione-folding buffer without further extraction or purification. [30] Properly folded CBM glycoforms display much shorter retention times on HPLC (high performance liquid chromatography) than side products, allowing for facile purification. Additional details on the synthesis method can be found in the “Experiments” section.

Using our one-pot synthesis/deprotection/folding/purification method, we first synthesized a library of CBM glycoforms containing either a mono-, di-, or tri-saccharide at each of the three separate glycosylation sites for a total of 9 CBM glycoforms (**2-10**). Subsequently, we synthesized two additional series of CBMs that had specific glycans at more than one of the three glycosylation sites (**11-20**) (Figure 2.2). Depending on the glycosylation patterns, the yields for the synthesis of these glycovariants ranged from 20% for **12** to 3% for **20**. Glycoform identities and homogeneity were experimentally verified by liquid chromatography–mass spectrometry (LC-MS) (“Experiments” section). Their conformations were examined by circular dichroism (CD). The CD spectra of **2-20** were similar to that of **1**, indicating that the glycosylation at positions Thr1, Ser3, and Ser14 did not significantly alter the conformation and structure of the *Tr*Cel7A CBM (“Experiments” section).

2.2.2 Site-specific impacts of *O*-mannosylation on CBM properties

We first investigated how mannose at each of the three *O*-glycosylation sites affects proteolytic stability, thermostability, and cellulose binding affinity (Figure 2.3A and Table 2.2). To elucidate the potential site-specific effects of CBM glycans on the proteolytic stability, nine mono-glycosylated variants, **2-10**, were compared with the nonglycosylated CBM **1** by thermolysin digestion. Thermolysin is capable of digesting the CBM into several small fragments. One cleavage site is close to the *N*-terminus of CBM, and the truncation causes a detectable change in molecular mass. Therefore, by monitoring the first order exponential decay of the full-length CBM using quantitative MALDI-TOF mass spectrometry, the CBM glycoform half-life to thermolysin degradation was calculated.[31-33] As shown in Figure 2.3A (top panel), *O*-mannosylation can substantially impact and improve the proteolytic stability of the CBM in a site-specific and glycan size-dependent manner. The nonglycosylated CBM **1** has a half-life of thermolysin degradation of about 0.2 h, whereas the Ser3 glycosylated CBM variants, **6** and **7**, have half-lives of more

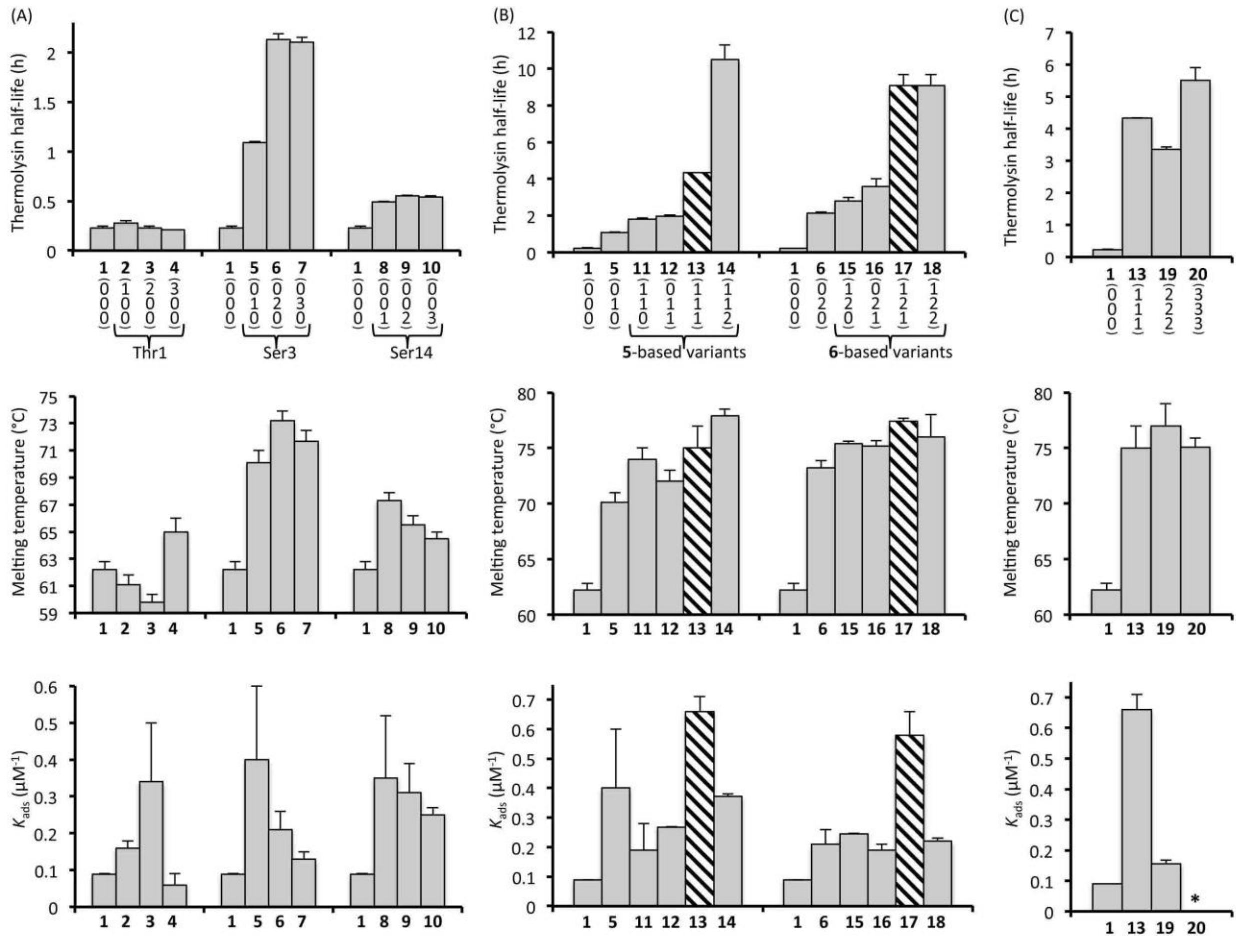


Figure 2.3 - The effects of *O*-mannose glycans on the proteolytic stabilty (half-life to thermolysin degradation), thermostability (melting temperatures measured by variable temperature circular dichroism), and binding affinity (K_{ads} values on bacterial microcrystalline cellulose) of the *Tr*Cel7A CBM. (A) The site-specific contribution of mono-, di- and tri-mannoses at each of the three O-glycosylation sites. (B) The combined effects of multiple O-linked glycans on CBMs. (C) The effects of glycosylation density on the properties of CBMs. Bolded numbers represent the identity of CBM glycoforms as per Figure 2.2. Top panel numbers in parentheses represent the glycoform pattern *i.e.* (100) representing a single mannose at Thr1, (010) representing a single mannose at Ser3, and (001) representing a single mannose at Ser14. All error bars reported are standard deviations of data achieved from three (thermolysin half-life and melting temperature) and two separate trials (K_{ads} values on BMCC). The hatched pattern indicates the glycoforms with multiple enhanced properties. ∗ Observable binding noted, nonlinear least squares curving fitting failed to converge.

than 2 h, an increase of over 10-fold. In contrast, the glycoforms bearing large *O*-linked mannoses at Thr1 and Ser14 sites show much less or no increase in half-lives compared to the unglycosylated CBM **1**.

O-mannosylation can also affect the CBM thermostability in a site-specific manner. The thermostability of each variant was assessed by its melting temperature, which can be directly measured by variable temperature CD.[27] As shown in Figure 2.3A (middle panel), *O*-mannosylation at Ser3 leads to the most substantial stabilization, with the increase in melting temperature of 11 °C as compared to nonglycosylated CBM **1**, which has a melting point of 62 °C. Mannosylation at Ser14 also leads to noticeable, but less pronounced stabilization than Ser3 mannosylation. Thr1 mannosylation displayed the least stabilizing ability. The glycan size does not appear to be directly related to the magnitude of the stabilizing effect, similar to observations for protein *N*-glycosylation.[34,35]

The binding affinity of the *Tr*Cel7A CBM glycoforms to crystalline cellulose was also measured, using a similar method to our previous studies.[20] The binding affinity of each CBM glycoform for bacterial microcrystalline cellulose (BMCC) was fitted to a Langmuir isotherm and is reported as K_{ads}, which correlates with the strength of CBM adsorption to BMCC. As shown in Fig. 3A (bottom panel), all three sites are potentially involved in CBM-substrate interactions. The addition of a single mono-mannose motif to the Ser3 or Ser14 positions provided a substantial increase in affinity and a di-mannose motif to Thr1 caused a similarly large increase in binding affinity. Conversely, increased affinity towards BMCC is diminished with attachment of larger glycans beyond a mono-mannose at Ser3 or Ser14 or a di-mannose at Thr1. Overall, by systematically comparing the properties of the mono-glycosylated variants, the importance and site-specific and size-dependent effects of *O*-mannosylation were demonstrated. Two glycoforms, **5** and **6**, were identified with multiple enhanced properties (Figure 2.3A).

2.2.3 Effects of *O*-mannosylation at multiple glycosylation sites

To understand the impact of *O*-glycosylation on the *Tr*Cel7A CBM in the physiological context [21], it is necessary to examine glycoforms with *O*-mannose glycans at multiple glycosylation sites. Thus, we conducted two additional series of comparative studies (Fig. 3B and Table 2.3). Because glycoforms with *O*-mannose residues at Ser3 (**5** and **6**) have multiple enhanced properties, we focused our studies on changes of the properties of these two variants. In each series of studies, the effects of the addition of *O*-linked glycans at Thr1 and/or Ser14 on the proteolytic and thermostability and binding affinity of the CBM were examined.

As shown in Figure 2.3B (top panel), the attachment of an additional mono-mannose to Thr1 or Ser14 can lead to further increase in the half-life of CBM to thermolysin degradation (compare **5**, **11**, **12** and **6**, **15**, **16**), although the half-life of mono-mannosylated **2** is essentially the same as that of **1**. The greatest half-life enhancement was achieved with higher glycan density, such as additional attachment of glycans at more positions and greater length of glycans (compare **5**, **13**, **14** and **6**, **17**, **18**). The correlation between thermostability and glycan density is much less obvious than that of the proteolytic stability (Fig. 3B, middle panel). The mannosylation of Ser3 with a di-mannose leads to the most significant increases in the melting temperature (compare **1** and **6**). Additional glycosylation site occupancy and the greater glycan length past the di-mannose structure have much less impact (compare **6** and **15**, **16**, **17**, **18**). An interesting affinity trend was identified from the BMCC adsorption studies of multi-mannosylated CBM glycoforms. As shown in Fig. 3B (bottom panel), it seems that the three *O*-mannosylation sites synergistically modulate the binding affinity enhancements. The CBM glycoforms with fully occupied glycosylation sites show better binding than those with partially occupied sites (compare partially occupied **11**, **12**, **15**, **16** and fully occupied **13**, **17**). Intriguingly, increases in glycan sizes lead to decreased binding affinities.

Lastly, to further confirm the influence of the size of *O*-linked mannoses on the properties of CBM, we compared the stability and binding affinity of **1**, **13**, **19**, and **20** (Fig. 3C and Table 2.4). As expected, glycoform **20**, which contains a tri-mannose at each glycosylation site, has a higher proteolytic stability, a similar thermostability, and a much lower binding affinity compared with **13**. The overall properties of **19** are better than those of **20**, but are less favored than those of **13**.

2.3 Discussion

Protein glycosylation is one of the most prevalent post-translational modifications with more than 50% of proteins in eukaryotes containing glycans.[36] Glycosylation can modulate both the physical and biological properties of proteins[37,38] and can aid in protein folding and secretion.[39] Indeed, *O*-linked glycans on cellulase linkers confer proteolytic resistance[40] and have been shown to impart affinity towards crystalline cellulose.[20] Furthermore, it has been shown that small *O*-linked glycans exist on the *Tr*Cel7A CBM near the binding face (Fig. 1) [21], which were predicted through computational studies to improve CBM affinity towards crystalline cellulose.[23] Alternatively, *O*-linked glycans distant from the binding face of a Family 2a CBM do not affect cellulose affinity[41] and large high-mannose type *N*-linked glycans near the Family 2a CBM binding face detrimentally affect cellulose affinity.[42] Family 1 CBM experimental studies to date have examined the functional role of many structural features, but no work has systematically considered the effects of the natural *O*-mannosylation.[9,12,13,15,41,43] To address the various potential effects of *O*-mannosylation on the *Tr*Cel7A CBM, we performed the comparative study using synthetic homogenous glycoforms.

Using synthetic glycoforms, we systematically demonstrated that *O*-glycosylation enhances the stability and cellulose binding affinity of a model Family 1 CBM. This study further demonstrates the feasibility and reliability of chemical synthesis in exploring the effects of glycosylation and allows for the identification of the *O*-glycosylation site that has the largest impact on the functional properties of CBM, Ser3, and the identification of the glycoforms with better overall properties, **13** and **17**, (Figure 2.3, highlighted by hatching). In addition, this study provides unique insights into the varied roles of different *O*-linked mannoses in modulating the properties directly related to the performance of the CBM, which would not be possible using heterogeneous natural mixtures of glycoforms.

During biomass depolymerization in Nature, organisms secrete proteolytic enzymes capable of cellulase degradation. The secretion of proteases aids in the competition for resources and also as a means for pathogen defense mechanisms.[44] It is clear from previous studies that glycans can protect against proteolysis.[40] Our results also demonstrate such protection, showing that glycans can protect the peptide backbone from proteolytic attack, likely through a steric hindrance mechanism. It is also clear that mannosylation at Ser3 leads to larger increases in proteolytic stability, possibly because glycosylation hinders thermolysin access to the *N*-terminal cleavage site at Tyr5. Although the protease protection conferred by mannosylation shows site-specific differences, the density of glycans on the surface of CBM causes far more dramatic increases in the thermolysin resistance. The more heavily glycosylated CBMs all have much longer half-lives to proteolytic degradation by thermolysin, indicating that the backbone of the CBM can be more effectively shielded by increased glycan length and density.[38]

Thermostability is a highly preferred trait of industrial enzymes. As such, many studies have engineered cellulases for improved thermostability through amino acid substitutions or through domain and sequence shuffling.[45,46] We observed that marked increases in CBM thermostability are conferred by glycosylation and that mannosylation at Ser3, specifically, plays a more substantial role in increasing the melting temperature (T_m) of CBM than glycans at Ser14 and Thr1. The results show that even a mono-mannose glycan at Ser3 substantially increases the CBM thermostability. Di-mannose at Ser3 leads to a T_m enhancement of 11 °C compared to the nonglycosylated molecule, which is the largest increase observed by the addition of a glycan to a single site. Further attachment of mannose to the CBM only induces minor changes in T_m. This finding is similar to the observations for studies of *N*-glycosylation, indicating that the large enhancements caused by *O*-mannosylation at Ser3 might be at least partially due to interactions between the first two mannose units attached to Ser3 and its local amino acid residues.[34,47]

An improvement of CBM-mediated adsorption onto insoluble crystalline cellulose has been shown experimentally to be beneficial for activity of the *Humicola grisea* GH Family 7 cellobiohydrolase on crystalline cellulose.[48] Though previous Type-A CBM glycoengineering efforts were met with limited success,[41,42] we have shown that small mannosyl residues at each of the three glycosylation sites are able to increase the binding affinity of CBM to BMCC (Figure 2.3A, bottom panel). Greater affinity enhancement could be achieved with a di-mannose moiety at Thr1, or with a mono-mannose at Ser3 or

Ser14. Intuitively, this may be due to the greater distance of Thr1 from the binding face than the other glycosylation sites (Figure 2.1). Therefore, to gain any increase in hydrogen bonding potential with the cellulose surface, a longer glycan would be required on Thr1. Moreover, the addition of extra mannosyl units actually decreased affinity for crystalline cellulose, similar to work from Boraston *et al*.,[42] who reported large *N*-linked glycans as detrimental to a Family 2 CBM adsorption to cellulose. This result could be explained by a steric hindrance effect, such that longer glycans can interfere with the interactions between the hydrophobic surface of BMCC and highly conserved Tyr5, Tyr31, and Tyr32 residues as has been hypothesized previously (Figure 2.1).[41,42]

O-linked mannoses at the three sites studied here act synergistically to enhance the binding affinity of CBM to BMCC. From the data presented in Figure 2.3B (bottom panel), it was also confirmed that the chief binding enhancement could be achieved through addition of monosaccharides on all three glycosylation sites. This observation further supports the theory that large glycans inhibit additional affinity of the Cel7A CBM towards BMCC. Taken together, these data show that binding affinity of Family 1 CBMs is intimately tied to *O*-mannosylation patterns and the affinity is best enhanced with smaller glycan structures. Notably, the trend from the experimental binding data is quite similar to that reported by Taylor *et al*.,[23] signifying that the combination of computational predictions and experimental verifications could be a useful tool in understanding cellulase glycosylation.

The affinity values achieved for the non-glycosylated CBM**1** are in agreement with other studies,[20,43,49] We report affinity enhancements of similar magnitude as those reported in Takashima *et al.*[48] and Linder and coworkers[12] with affinity of CBM **13** and **17** approaching or within range of K_{ads} values obtained for both the whole *Tr*Cel7A or the multi-glycosylated *Tr*Cel7A linker-CBM domain,[20,50-53] suggesting that glycoengineering of Family 1 CBMs as a viable strategy for activity enhancement of cellulases. Moreover, the *O*-linked glycosylation sites studied here are highly conserved across Family 1 CBMs,[22] suggesting that the glycosylation enhancing properties observed here likely occur throughout this ubiquitous CBM family. It has also been proposed that other cellulolytic enzymes contain similarly beneficial glycosylation for substrate binding, such as the glycans near the binding surfaces of lytic polysaccharide monooxygenases.[54,55] This suggests that glycans impart multiple beneficial properties to cellulases, and glycosylation may be strategy commonly used by biomass-degrading organisms for cellulase enhancement.

2.4 Conclusions

In summary, we utilized chemical synthesis to develop a practical, one-pot method for quickly and conveniently obtaining a collection of representative glycoforms of a Family 1 CBM. These homogeneous glycoforms are valuable tools for developing a quantitative understanding of protein glycosylation. Using these structurally well-defined glycoforms, we have shown that *O*-linked mannose residues increase the proteolytic stability of CBM in a glycan size-dependent manner, thermostability in a glycosylation site-specific manner, and binding affinity in a glycosylation pattern-dependent manner. Our data also support the theory that large glycans decrease the ability of CBMs to bind to crystalline cellulose. Taken together, our study demonstrates the importance of *O*-mannosylation in regulating the properties of the Family 1 CBM. This regulation may allow biological systems to fine-tune how the CBM binds to crystalline cellulose during degradation. We anticipate that the concepts put forth here will find broad applicability in the study of other protein post-translational modifications, and in the glycoengineering of industrially and therapeutically-important proteins.

2.5 Experiments

2.5.1 Materials

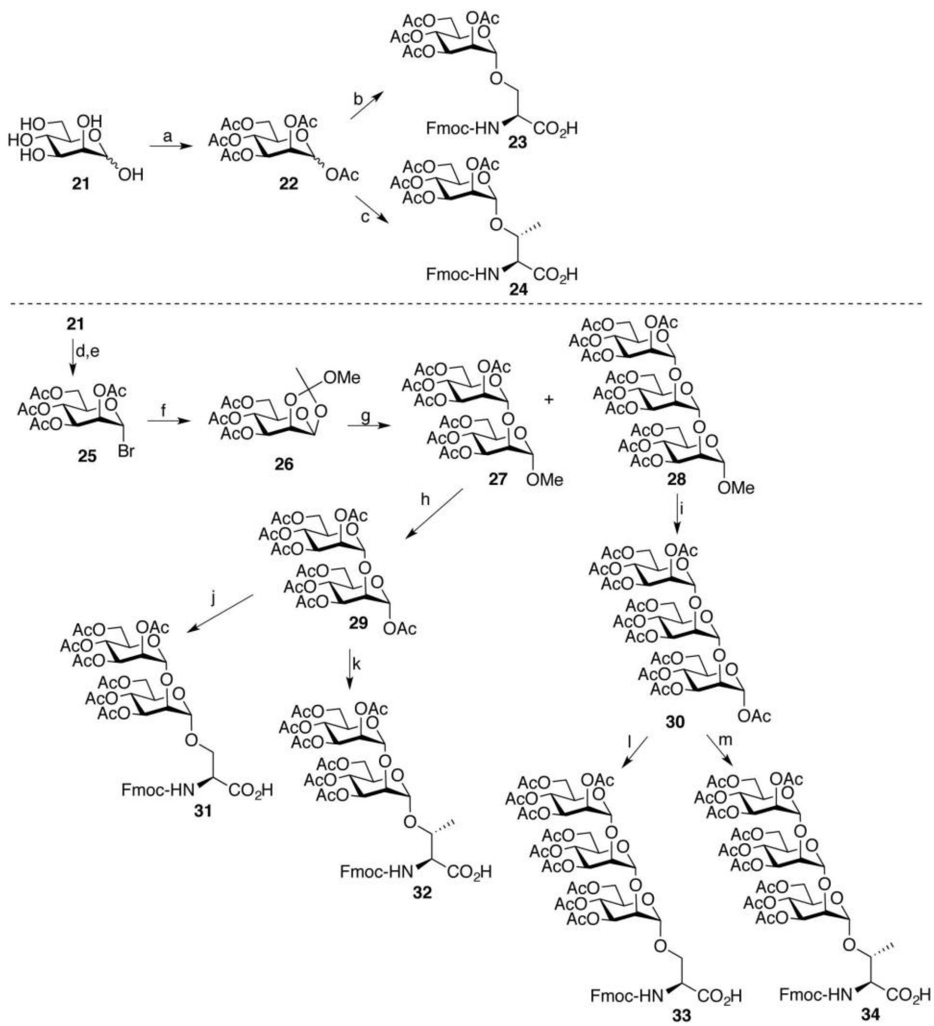
All commercial reagents and solvents were used as received. Unless otherwise noted, all reactions and purifications were performed under air atmosphere at room temperature. All LC-MS analyses were performed using a Waters Acquity™ Ultra Performance LC system equipped with Acquity UPLC® BEH 300 C4, 1.7μm, 2.1 x 100 mm column at flow rates of 0.3 and 0.5 mL/min. The mobile phase for LC-MS analysis was a mixture of H_2O (0.1% formic acid, v/v) and acetonitrile (0.1% formic acid, v/v). All preparative separations were performed using a LabAlliance HPLC solvent delivery system equipped with a Rainin UV-1 detector and a Varian Microsorb 100-5, C18 250x21.4mm column at a flow rate of 16.0 mL/min. The mobile phase for HPLC purification was a mixture of H_2O (0.05% TFA, v/v) and acetonitrile (0.04% TFA, v/v). A Waters SYNAPT G2-S system was used mass spectrometric analysis. All circular dichroism (CD) spectra were obtained using an Applied PhotophysicsChirascan^{TM}-plus CD spectrometer.

2.5.2 Chemical synthesis of the *O*-mannosylated amino acid building blocks

The glycoamino acid building blocks, Fmoc-Ser(Ac4Manα1)-OH, Fmoc-Thr(Ac4Manα1)-OH, Fmoc-Ser(Ac4Manα1-2Ac$_3$Manα1)-OH, Fmoc-Thr(Ac4Manα1-2Ac$_3$Manα1)-OH, Fmoc-Ser(Ac4Manα1-2Ac$_3$Manα1-2Ac$_3$Manα1)-OH, and Fmoc-Thr(Ac4Manα1-2Ac$_3$Manα1-2Ac$_3$Manα1)-OH were prepared according to the previously reported methods.[56-59]

Compound 25: The mixture of compound **21** (20 g, 111.05 mmol) and NaOAc (12 g, 146.30 mmol) in Ac_2O (200 ml) was stirred at 110 ^oC for 2 h. The solution was concentrated to remove Ac_2O, poured into ice-cold aq. $NaHCO_3$, extracted with DCM, washed with brine, and dried with Na_2SO_4. The resulting solution was filtered and concentrated to give 57.70 g syrup **22**. 200 ml 33% HBr-AcOH solution was added to the above syrup, and the resulting mixture was stirred at room temperature for 1h, diluted with DCM, and transferred to a separatory funnel containing ice. The organic phase was washed with ice-water until neutral pH was obtained. The solution was dried with Na_2SO_4, filtered and concentrated to give 40.11 g **25** as a syrup. Yield: 88%. The product was used directly in the next step without purification.

Compound 26: Compound **25** (22 g, 53.6 mmol) was dissolved in $CHCl_3$ (150 ml). 2,6-lutidine (15 ml, 128.77 mmol) in MeOH (150 ml) was then added. The resulting mixture was stirred at room temperature overnight. The mixture was then diluted with $CHCl_3$, washed with ice-cold 3% aq. $NaHCO_3$. The aqueous layer was extracted with $CHCl_3$. The combined organic portion was washed with ice-cold water, and dried over Na_2SO_4. The product solution was filtered, concentrated, and washed with hexanes to give a white solid (15.5 g). 5 g of the solid was purified by chromatography on a silica gel column (Hexane/Ethyl acetate =3:1 → 2:1) to give 4.6 g of pure **26**. Yield: 74%.



Scheme 2.1 - Synthesis of *O*-mannosylated Ser and Thr. Reagents and conditions: (a) Ac_2O, NaOAc, 110 °C, 91%; (b) Fmoc-Ser-OH, $BF_3{\cdot}OEt_2$, CH_3CN, 44%; (c) Fmoc-Thr-OH, $BF_3{\cdot}OEt_2$, CH_3CN, 41%; (d) Ac_2O, NaOAc, 110 °C; (e) HBr, AcOH, CH_2Cl_2, 88% over two steps; (f) 2,6-lutidine, MeOH, $CHCl_3$, 74%; (g) TMSOTf, CH_2Cl_2, -30 °C, **27**, 58%, **28**, 16%; (h) Ac_2O, H_2SO_4, 0 °C, 89%; (i) Ac_2O, H_2SO_4, 0 °C, 73%; (j) $BF_3{\cdot}OEt_2$, CH_3CN, 41%; (k) $BF_3{\cdot}OEt_2$, CH_3CN, 64%; (l) $BF_3{\cdot}OEt_2$, CH_3CN, 57%; (m) $BF_3{\cdot}OEt_2$, CH_3CN, 58%;

*Compound **27** and **28**:* To the solution of compound **26** (6 g, 16.56 mmol) in DCM (200 ml) at -30^{o}C, TMSOTf (9.6 ml, 49.71 mmol) was added. After 5 min, the reaction mixture was diluted with DCM and quenched with addition of aq. $NaHCO_3$. The aqueous layer was extracted with DCM. The combined organic solution was washed with brine, and dried over anhydrous Na_2SO_4. After drying, the product solution was filtered, concentrated, and purified by chromatography on a silica gel column (Hexane/Ethyl acetate =1:1 → 1:2) to give 3.12 g of **27** as a white foam. Yield: 58%. And 830 mg of **28** as a white foam. Yield: 16%.

*Compound **30**:* Compound **28** (550 mg, 0.59 mmol) was dissolved in Ac_2O (7 ml), and cooled to 0 ^oC. H_2SO_4 (0.02 mL) was then added dropwise. The mixture was stirred at 0 ^oC for 5 h. After that, it was diluted with DCM and neutralized with aq. $NaHCO_3$. The organic portion was washed with water and brine, and dried over Na_2SO_4. The product solution was filtered, concentrated, and purified by chromatography on a silica gel column (Hexane/Ethyl acetate =1:1 → 1:2) to give 410 mg of **30** as a white foam. Yield: 73%.

*General procedures for the synthesis of compound **23**, **24**, **31**, **32**, **33** and **34**:* To the solution of mono-, di- or tri-mannose (1.0 eq) and Fmoc-Ser/Thr-OH (1.2 eq) in acetonitrile (20 mL per mmol sugar), $BF_3{\cdot}OEt_2$ (3.0 eq) was added. The resulting mixture was stirred at room temperature for 24 h. The solvent was removed under reduced pressure. The residue was dissolved in ethyl acetate, washed with water and dried over Na_2SO_4. The resulting solution was filtered and concentrated under reduced pressure, purified by chromatography on a silica gel column (Hexane/Ethyl acetate/AcOH) to afford the title product as a white foam.

*Fmoc-Ser (Ac4Manα)-OH **23**:* ^{1}H-NMR (400 MHz, $CDCl_3$) δ 7.75 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.61 (t, *J* = 6.8 Hz, 2H, H-Fmoc), 7.38 (t, *J* = 7.4 Hz, 2H, H-Fmoc), 7.29 (t, *J* = 8.0 Hz, 2H, H-Fmoc), 7.46 (d, *J* = 8.8 Hz, 1H, NH), 5.42 (dd, *J* = 10.0, 3.6 Hz, 1H, H-3), 5.23-5.28 (m, 2H, H-2, H-4), 4.86 (d, *J* = 1.6 Hz, 1H, H-1), 4.65-4.68 (m, 1H, H-α), 4.29-4.45 (m, 2H, CH_2- Fmoc), 4.19-4.27 (m, 2H, CH- Fmoc, H-6), 4.05-4.11 (m, 4H, H-5, H-6, CH_2-β), 2.14 (s, 3H, CH_3-Ac), 2.06 (s, 3H, CH_3-Ac), 2.01 (s, 3H, CH_3-Ac), 1.97 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 172.4, 170.8, 170.7, 170.3, 169.8, 156.1, 143.8, 141.3, 127.7, 127.1, 125.2, 120.0, 98.1, 69.4, 69.3, 69.0, 67.3, 66.1, 54.2, 47.1, 20.9, 20.8, 20.7, 20.6. ESI-MS: Calc. for $C_{32}H_{35}NO_{14}$: 657.2058. Found: 680.3007 $[M+Na]^+$.

*Fmoc-Thr (Ac4Manα)-OH **24**:* ^{1}H-NMR (300 MHz, DMSO-d6) δ 7.90 (d, *J* = 7.5 Hz, 2H, H-Fmoc), 7.76 (d, *J* = 6.9 Hz, 2H, H-Fmoc), 7.40-7.46 (m, 2H, H-Fmoc), 7.31-7.37 (m, 2H, H-Fmoc), 5.29 (dd, *J* = 9.9, 3.6 Hz, 1H, H-3), 5.05 (dd, *J* = 3.9, 1.8 Hz, 1H, H-2), 5.04 (t, *J* = 9.9, 1H, H-4), 4.97 (d, *J* = 1.5 Hz, 1H, H-1), 4.22-4.32 (m, 4H, CH-β, CH- Fmoc, CH_2- Fmoc) 4.02-4.20 (m, 4H, H-5, H-6, H-6, H-α), 2.08 (s, 3H, CH_3-Ac), 2.03 (s, 3H, CH_3-Ac), 2.01 (s, 3H, CH_3-Ac), 1.92 (s, 3H, CH_3-Ac), 1.24 (d, *J* = 6.3 Hz, 3H, CH_3-γ). ^{13}C-NMR (75 MHz, DMSO-d6) δ 178.8, 170.5, 170.0, 169.9, 169.8, 157.1, 144.3, 141.2, 128.1, 127.6, 125.9, 120.6, 98.5, 76.4, 69.5, 69.0, 68.5, 66.5, 62.6, 59.3, 47.1, 21.5, 21.0, 20.93, 20.87, 18.4. ESI-MS: Calc. for $C_{33}H_{37}NO_{14}$: 671.22. Found: 694.33 $[M+Na]^+$.

Fmoc-Ser (Ac4Manα1-2Ac$_3$Manα)-OH ***31****:* ^{1}H-NMR (500 MHz, $CDCl_3$) δ 7.76 (d, *J* = 7.5 Hz, 2H, H-Fmoc), 7.63 (d, *J* = 6.5 Hz, 2H, H-Fmoc), 7.40 (t, *J* = 7.5 Hz, 2H, H-Fmoc), 7.32 (t, *J* = 7.5 Hz, 2H, H-Fmoc), 6.01 (d, *J* = 8.0 Hz, 1H, NH), 5.38 (dd, *J* = 10.0, 3.0 Hz, 1H, H-3'), 5.26-5.32 (m, 3H, H-3, H-4, H-4'), 5.23-5.25 (m, 1H, H-2'), 4.96 (d, *J* = 2.5 Hz, 1H, H-1), 4.91 (d, *J* = 2.0 Hz, 1H, H-1'), 4.58-4.60 (m, 1H, H-α), 4.38-4.42 (m, 2H, CH_2- Fmoc), 4.30 (dd, *J* = 12.5, 3.0 Hz, 1H, H-6), 4.20-4.25 (m, 2H, CH- Fmoc, H-6), 4.05-4.17 (m, 5H, H-5', H-6', H-6', CH_2-β), 4.02-4.03 (m, 1H, H-2), 3.96-3.99 (m, 1H, H-5), 2.14 (s, 3H, CH_3-Ac), 2.12 (s, 3H, CH_3-Ac), 2.11 (s, 3H, CH_3-Ac), 2.10 (s, 3H, CH_3-Ac), 2.05 (s, 3H, CH_3-Ac), 2.01 (s, 3H, CH_3-Ac), 2.00 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 172.4, 171.1, 171.0, 170.9, 169.9, 169.8, 169.6, 169.5, 143.7, 141.3, 127.8, 127.1, 125.1, 120.0, 99.0, 98.8, 77.3, 76.6, 70.2, 69.7, 69.1, 69.0, 68.4, 67.3, 66.5, 62.5, 66.3, 62.1, 54.3, 47.1, 20.82, 20.76, 20.67, 20.65, 20.63, 20.61. ESI-MS: Calc. for $C_{44}H_{51}NO_{22}$: 945.29. Found: 968.39 $[M+Na]^+$.

Fmoc-Thr (Ac4Manα1-2Ac$_3$Manα)-OH ***32****:* ^{1}H-NMR (400 MHz, Actone-d6) δ 7.89 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.74 (t, *J* = 6.4 Hz, 2H, H-Fmoc), 7.43 (t, *J* = 7.6 Hz, 2H, H-Fmoc), 7.33-7.38 (m, 2H, H-Fmoc), 5.38 (dd, *J* = 10.4, 3.6 Hz, 1H, H-3), 5.27-5.32 (m, 4H, H-2', H-3', H4, H4'), 5.25 (d, *J* = 1.6 Hz, 1H, H-1), 5.02 (d, *J* = 1.6 Hz, 1H, H-1'), 4.51-4.54 (m, 1H, CH-β), 4.33-4.44 (m, 3H, H-α, CH_2- Fmoc), 4.27 (t, *J* = 6.6 Hz, 1H, CH-Fmoc), 4.10-4.23 (m, 7H, H-2, H-5, H5', H6, H6'), 2.12 (s, 3H, CH_3-Ac), 2.09 (s, 3H, CH_3-Ac), 2.08 (s, 3H, CH_3-Ac), 2.05 (s, 3H, CH_3-Ac), 2.03 (s, 3H, CH_3-Ac), 2.03 (s, 3H, CH_3-Ac), 1.97 (s, 3H, CH_3-Ac), 1.42 (d, *J* = 6.4 Hz, 3H, CH_3-γ). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 171.8, 170.1, 170.0, 169.8, 169.4, 169.3, 169.2, 169.1, 142.2, 141.2, 127.6, 127.1, 125.3, 119.9, 99.9, 98.0, 77.4, 76.8, 70.0, 69.4, 69.31, 69.28, 69.0, 68.6, 66.40, 66.35, 65.9, 62.2, 62.1, 58.8, 47.2, 19.90, 19.88, 19.84, 19.79, 19.76, 19.74, 19.68, 17.9. ESI-MS: Calc. for $C_{45}H_{53}NO_{22}$: 959.31. Found: 982.30 $[M+Na]^+$.

Fmoc-Ser (Ac4Manα1-2Ac$_3$Manα1-2Ac$_3$Manα)-OH ***33****:* ^{1}H-NMR (400 MHz, Actone-d6) δ 7.88 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.74 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.44 (t, *J* = 7.4 Hz, 2H, H-Fmoc), 7.33-7.38 (m, 2H, H-Fmoc), 7.16 (d, *J* = 8.4 Hz, 1H, NH), 5.24-5.42 (m, 8H, H-1, H-2'', H-3, H-3', H-3'', H-4, H-4', H-4''), 5.20 (d, *J* = 2.0 Hz, 1H, H-1'), 5.17 (d, *J* = 1.6 Hz, 1H, H-1''), 4.58-4.62 (m, 1H, H-α), 4.33-4.42 (m, 2H, CH_2- Fmoc), 4.12-4.29 (m, 13H, H-2', H-5, H-5', H-5'', H-6, H-6', H-6'', CH_2-β, CH-Fmoc), 4.04-4.09 (m, 1H, H-2), 2.13 (s, 3H, CH_3-Ac), 2.11 (s, 3H, CH_3-Ac), 2.10 (s, 3H, CH_3-Ac), 2.08 (s, 3H, CH_3-Ac), 2.05 (s, 3H, CH_3-Ac), 2.04 (s, 3H, CH_3-Ac), 2.034 (s, 3H, CH_3-Ac) 2.028 (s, 3H, CH_3-Ac), 1.98 (s, 6H, CH_3-Ac). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 170.3, 170.1, 169.5, 169.4, 169.3, 169.14, 169.10, 144.4, 141.2, 127.6, 127.1, 125.3, 119.9, 99.8, 99.2, 98.7, 77.1, 70.4, 69.7, 69.4, 69.29, 69.25, 68.8, 68.6, 66.02, 65.98, 62.4, 62.0, 61.9, 47.2, 20.0, 19.94, 19.89, 19.82, 19.79, 19.76, 19.7. ESI-MS: Calc. for $C_{56}H_{67}NO_{30}$: 1233.37. Found: 1256.90 $[M+Na]^+$.

Fmoc-Thr (Ac4Manα1-2Ac$_3$Manα1-2Ac$_3$Manα)-OH ***34****:* ^{1}H-NMR (400 MHz, Actone-d6) δ 7.88 (d, *J* = 7.2 Hz, 2H, H-Fmoc), 7.73 (t, *J* = 6.8 Hz, 2H, H-Fmoc), 7.43 (t, *J* = 7.6 Hz, 2H, H-Fmoc), 7.34 (t, *J* = 7.4 Hz, 2H, H-Fmoc), 5.25-5.48 (m, 9H, H-1, H-1', H-2'', H-3, H-3', H-3'', H-4, H-4', H-4''), 5.19 (d, *J* = 1.2 Hz, 1H, H-1''), 4.45-4.50 (m, 1H, H-β), 4.08-4.43 (m, 15H, H-2, H-2', H-5, H-5', H-5'', H-6, H-6', H-6'', H-α, CH-Fmoc, CH_2-Fmoc), 2.12 (s, 3H, CH_3-Ac), 2.11 (s, 3H, CH_3-Ac), 2.08 (s, 3H, CH_3-Ac), 2.06 (s, 3H, CH_3-Ac), 2.052 (s, 3H, CH_3-Ac), 2.045 (s, 3H, CH_3-Ac), 2.03 (s, 3H, CH_3-Ac) 2.024 (s, 3H, CH_3-Ac), 2.015 (s, 3H, CH_3-Ac), 1.97 (s, 3H, CH_3-Ac), 1.36 (d, *J* = 6.4 Hz, 3H, CH_3-γ). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 170.5, 170.3, 170.0, 169.5, 169.4, 169.3, 169.2, 169.0, 144.3, 141.2, 127.6, 127.1, 125.3, 119.9, 99.6, 99.4, 99.2, 77.5, 77.3, 77.1, 76.4, 69.8, 69.5, 69.31, 69.26, 68.9, 68.6, 66.0, 65.8, 62.1, 61.9, 48.5, 47.2, 20.01, 19.97, 19.84, 19.82, 19.80, 19.75, 19.68, 18.1. ESI-MS: Calc. for $C_{57}H_{69}NO_{22}$: 1247.39. Found: 1270.63 $[M+Na]^+$.

2.5.2 General synthetic procedure for CBM variants

Solid-phase peptide synthesis was performed on a Pioneer™ Peptide Synthesis System. Peptides and glycopeptides were synthesized by Fmoc chemistry on a pre-loaded Fmoc-Leu-Novasyn® TGT resin. The following Fmoc amino acid building blocks and pseudoproline dipeptides from Chem-Impex, EMD Millipore, and AAPPTec were employed in the synthesis: Fmoc-Asn(Trt)-OH, Fmoc-Cys(Trt)-OH, Fmoc-Gln(Trt)-OH, Fmoc-Gly-OH, Fmoc-His(Trt)-OH, Fmoc-Ile-OH, Fmoc-Leu-OH, Fmoc-Pro-OH, Fmoc-Ser(tBu)-OH, Fmoc-Thr(tBu)-OH, Fmoc-Tyr(tBu)-OH, Fmoc-Val-OH, and Fmoc-Ala-Ser(psiMe,MePro)-OH. Synthetic cycles were completed with a standard coupling time of 15 min using Fmoc protected amino acids (4 eq.), 2-(1H-7-Azabenzotriazol-1-yl)-1,1,3,3-tetramethyl uroniumhexafluorophosphatemethanaminium (4 eq.) and N,N-diisopropylethylamine (8 eq), except for a prolonged coupling time of 2 h for glycoamino acids. The deblocking was performed by mixing with DMF/piperidine/1,8-Diazabicyclo[5.4.0]undec-7-ene (100/2/2, v/v/v) for 5 min. Upon completion, the resin was washed into a peptide cleavage vessel with dichloromethane. Cleavage and side-chain deprotection was performed by treatment with TFA/H_2O/triisopropylsilane (95/2.5/2.5, v/v/v) solution for 45 min. The filtered cleavage mixture was then concentrated using a gentle stream of air and precipitated at 0 °C by the addition of cold diethyl ether. After centrifugation, the resulting pellet was dissolved in H_2O/acetonitrile (1/1, v/v) and lyophilized to dryness for further use.

The acetyl groups of the sugar moiety were removed by stirring the unpurified synthetic glycopeptides in a hydrazine solution (hydrazine/H_2O, 5/100, v/v) at room temperature for 30 min under helium atmosphere. The final concentration of the glycopeptides was 4 mM. Upon completion, the reaction was quenched with a solution of AcOH (AcOH/H_2O, 5/100, v/v) and the pH was adjusted to ~8.

The folding was initiated by diluting the fully-unprotected peptides/glycopeptides to a final concentration of ~0.05 mM in a folding buffer (0.2 M Tris-acetate, 0.33 mM oxidized glutathione, 2.6 mM reduced glutathione, pH 8.2). The folding solution was stirred at room temperature for 12 h under helium atmosphere. The solution was then concentrated to a small volume using 3 kDa cut-off centrifugal filter units (Amicon) before RP-HPLC purification. The HPLC purification was performed on a Versagrad Preparation-HPLC system using a semi-preparative C-18 column. The products were detected by UV absorption at 275 nm.

Synthesis of CBM variant ***1:*** The unglycosylated peptide was synthesized on a 0.05 mmol scale. After SPPS, 168.2 mg of the crude peptide was obtained. 16 mg (4.28 μmol) of the crude peptide was dissolved in 80 ml of folding buffer and stirred at room temperature for 12 h. After concentration and HPLC purification with a linear gradient of 20→40% acetonitrile in H_2O over 30 min, 5.18 mg of **1** was obtained as a white solid (30% yield based on resin loading).

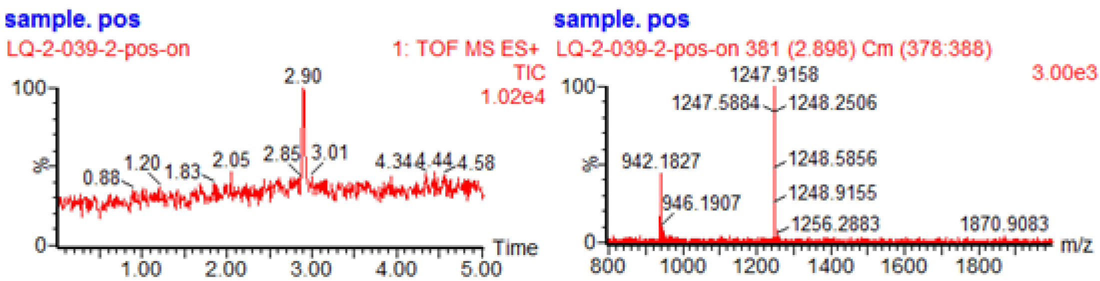


Figure 2.5 - LC-MS trace and ESI-MS of purified compound **1**. MS (ESI) Calcd for $C_{159}H_{235}N_{43}O_{54}S_4$: $[M+2H]^{2+}$ m/z = 1870.29, $[M+3H]^{3+}$ m/z = 1247.19, $[M+4H]^{4+}$ m/z = 935.65.

Synthesis of CBM variant ***2:*** The glycosylated peptide was synthesized on a 0.05 mmol scale. After cleavage and lyophilization, 188.7 mg of the crude glycopeptide was obtained. 7 mg (1.71 μmol) of it was dissolved in 450 μl of the solution of hydrazine and stirred at room temperature for 30 min. The reaction was quenched with the solution of AcOH and the pH was adjusted to ~8. The resulting solution was diluted by the addition of 40 mL of folding buffer. After folding and centrifugal concentration, HPLC purification with a linear gradient of 20→40% acetonitrile in H_2O over 30 min afforded the correctly folded **2** (1.04 mg, white solid, 15% yield based on resin loading).

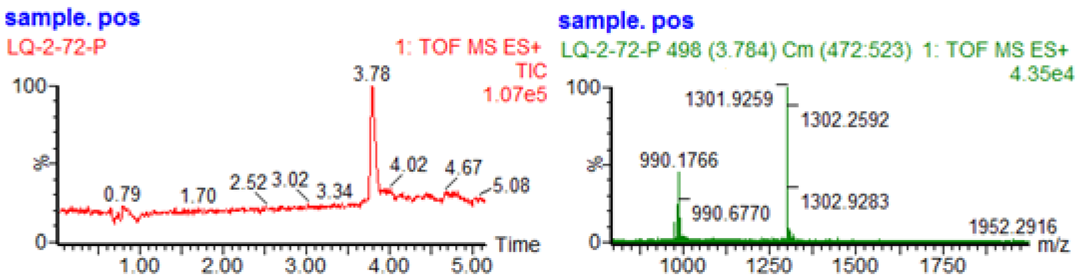
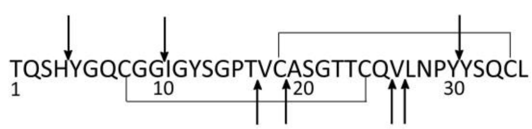


Figure 2.6 - LC-MS trace and ESI-MS of purified compound **2**. MS (ESI) Calcd for $C_{165}H_{245}N_{43}O_{59}S_4$: $[M+2H]^{2+}$ m/z = 1951.32, $[M+3H]^{3+}$ m/z = 1301.21, $[M+4H]^{4+}$ m/z = 976.16.

Confirming disulfide linkages. In order to confirm the disulfide linkages, CBM **1** was digested with thermolysin and the resulting fragments were analyzed. The lyophilized Thermolysin, which was obtained from *Bacillus thermoproteolyticusrokko*, was purchased from Promega Corporation. The digestion was performed in a 50 mM Tris-HCl buffer (pH = 8) with 0.5 mM $CaCl_2$ at a temperature of 37 °C. The reaction was performed in 100 µL of solution with an initial concentration of 270 µM. The solution was prepared so that **1** and thermolysin were initially present in a 20:1 molar ratio. The reaction was monitored over time by taking 10 µL aliquots and quenching them with an equal volume of 5% AcOH. The aliquots were analyzed using the Waters Acquity UPLC and a Waters SYNAPT G2-S mass spectrometer. As shown in Figure S23, the peptide fragments that were observed for the digest were consistent with the appropriate disulfide bond pattern (C8 to C25 and C19 to C35). The important CBM peptide peaks are identified.



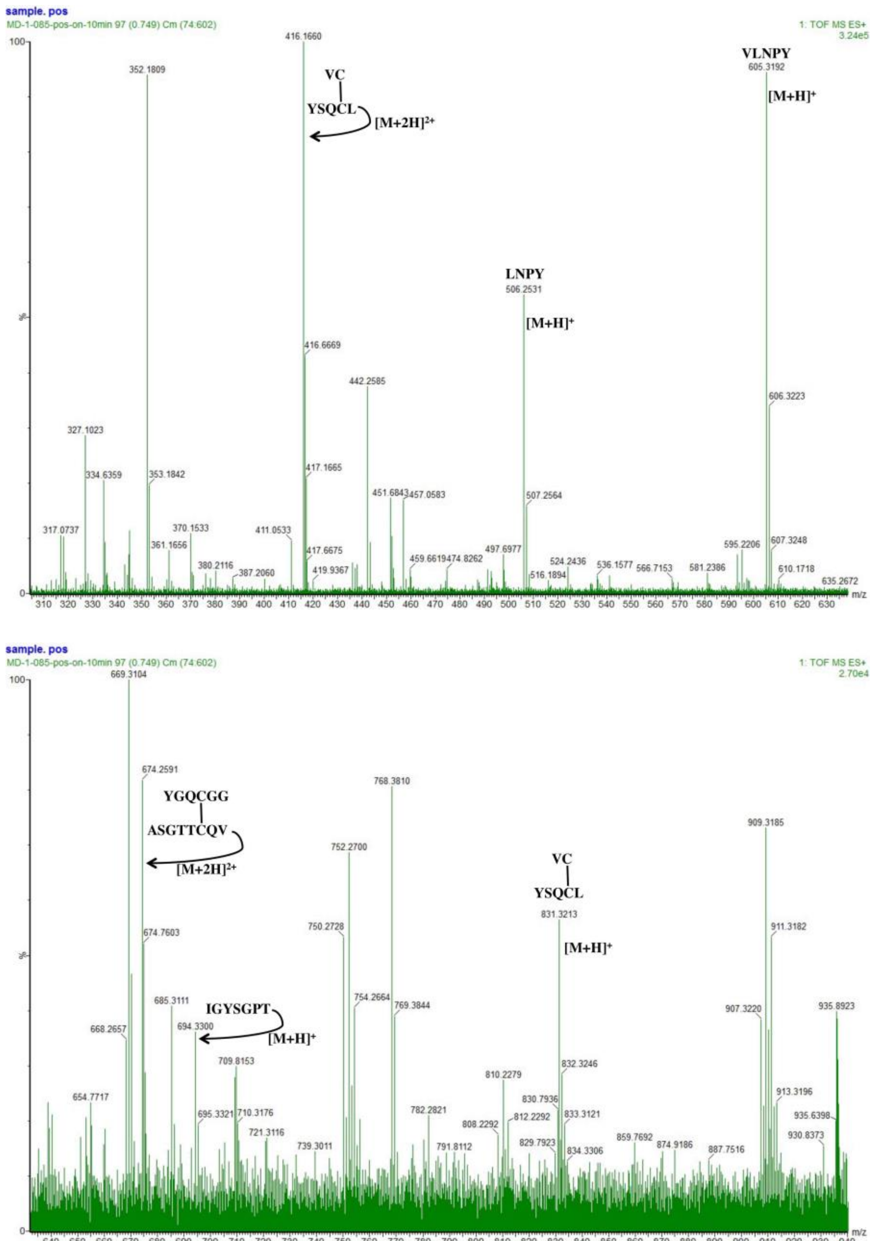


Figure 2.7 - Determination of the disulfide-bonding pattern of CBM **1** by thermolysin digestion

Circular dichroism (CD). All CD spectra were acquired using an Applied PhotophysicsChirascan^{TM}-plus CD spectrometer. In all cases, the spectra were acquired in a 0.5 mm quartz cuvette under nitrogen at a flow rate of 1 L/min. Each CBM analog was dissolved in 10 mM NaOAc buffer with a pH of 5.2. The peptide concentration was 0.2 mg/mL in all tests. CD spectra were obtained at 20 °C with a step of 0.5 nm, 0.5 s per point and a spectral width of 200-240 nm. The spectra are the average of 4 scans with an averaged 4 scan buffer baseline subtracted. The resulting CD spectra were used to calculate the secondary structure fractions of each CBM analog using the CDPro software provided by Colorado State University. The secondary structure fractions of each CBM analog are the average of the results of the three CDPro programs (SELCON3, CDSSTR and CONTIN) and are displayed in Table 2.1.

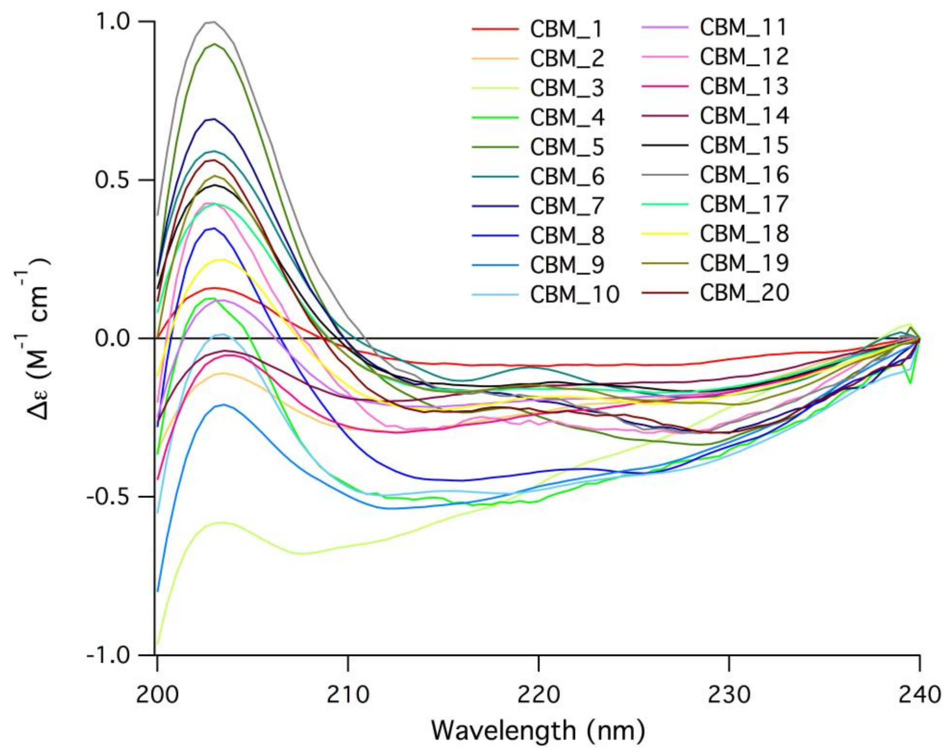


Figure 2.8 - The CD spectra of the 20 glyco-variants of the *Tr*Cel7A CBM.

Table 2.1 - The secondary structure percentages of each CBM analog based on the CD results.

	Secondary Structure Percent			
CBM Variant	B-Sheet	α-helix	Turn	Unordered
1	43.4	2.4	22.9	29.3
2	34.9	3.0	24.0	36.4
3	30.9	1.7	25.6	41.8
4	30.9	0.7	25.2	42.6
5	27.3	0.5	25.1	46.3
6	35.1	0.2	23.5	38.2
7	33.7	0.6	25.1	39.6
8	30.3	1.3	25.7	42.5
9	28.6	1.4	26.1	43.4
10	28.6	0.9	25.9	44.0
11	33.5	~0	23.3	37.0
12	30.6	1.3	24.9	42.5
13	34.2	2.9	23.0	37.6
14	32.3	0.1	25.8	36.2

15	35.3	0.5	23.3	37.7
16	32.3	~0	25.5	42.3
17	34.1	1.6	24.9	38.4
18	34.1	1.6	25.1	38.2
19	34.2	0.7	24.5	40.1
20	34.6	0.8	24.7	39.4

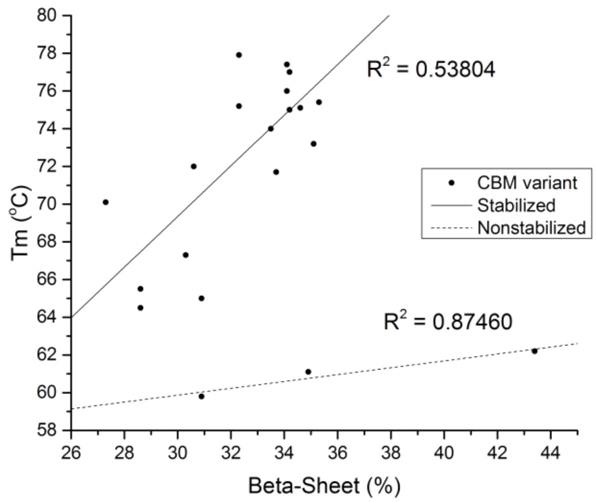


Figure 2.9 - Correlation of thermostability (melting temperatures measured by variable temperature circular dichroism) and % β-Sheet characteristic in the secondary structure of the *Tr*Cel7A CBM. Data points represent averaged T_m data and % β-sheet determined by deconvoluted CD spectra for each CBM glycoform. Lines represent the linear least squares fitting. For glycoforms **4-20** with an increased melting temperature over CBM **1** there exists a positive linear correlation between % β-sheet and T_m. Additionally, for **1-3** where glycosylation effectively lowered thermostability, this same correlation is apparent, but divergent from glycoforms **4-20.** The results for **1-3** are similar to those achieved elsewhere for *O*-glycosylation of interferon α2b resulting in a lower melting temperature.[2]

2.5.3 Biophysical and biological characterization

Thermolysin digests. The digestions were performed at 37°C in 100 μL of solution (50 mM Tris-HCl buffer, 0.5 mM $CaCl_2$, pH = 8.0) with an initial CBM variant concentration of 270 μM. The CBM variant and thermolysin were initially present in a 20:1 molar ratio. 10 μL aliquots were taken at specific time intervals and quenched with an equal volume of 5% AcOH. Each sample was analyzed by Quantitative MALDI-TOF Mass Spectrometry (described below) to calculate the change in CBM concentration with time. The digestion rate was determined by monitoring and fitting data to the first order exponential decay of the full length CBM glycoform over time [31-33].

Quantitative MALDI-TOF mass spectrometry. For absolute CBM quantitation, internal reference standard solutions of each CBM glycoform were prepared per experiment by serial dilution (10 μL per concentration) [60]. To all sample aliquots, 150 pmol of a CBM internal standard peptide ($\Delta m/z \geq 162$ Da) in H_2O:MeCN:AcOH (1:1:3.3% 3 μL) was added. 0.5 μL of each sample was spotted directly on a 100 well MALDI target plate with 1.126 μL of α-cyano-4-hydroxycinnamic acid (CHCA) matrix (6.2 mg/ml) in MeOH:MeCN:H_2O (36:56:8) and allowed to air dry (~5 min). Spectra were acquired on a Voyager-DE™ STR MALDI-TOF mass spectrometer (Applied Biosystems) in linear positive ion mode, with 50 shots per spectra. The laser intensity was set to 1950, the accelerating voltage was set to 20,000 V, the extraction delay time was 100 ns, and the grid voltage was set to 94%. The low mass gate was set to 500 Da and data were collected from 3200-5000 Da (5500 Da for glycoform **20**). An in-house MATLAB program was written to determine the ratio of analyte ion intensities between the CBM and the CBM internal standard. From these data, a standard linear calibration curve was generated for each experiment to calculate the absolute CBM concentration from CBM:CBM internal standard ion intensity ratios.

Thermostability Assay. All CD spectra were acquired using an Applied PhotophysicsChirascan™-plus CD spectrometer in a 0.5 mm quartz cuvette under nitrogen at a flow rate of 1 L/min. Lyophilized CBM glycoforms were suspended in 10 mM sodium acetate (pH = 5.2) at a concentration of 0.2 mg/ml. The melts were performed by ramping the temperature of the sample from 20 to 94°C at a rate of 1 °C/min while monitoring the CD signal at 217 nm. The melts resulted in roughly sigmoidal melting curves and the point of inflection of the curve was interpreted to be the melting point of the analog [61].

BMCC Adsorption Assay. Adsorption isotherms were performed as described elsewhere [12,43]. Briefly, lyophilized CBM glycoforms were suspended and serially diluted in 50 mM sodium acetate, 50 mM sodium chloride buffer (pH = 5.0). CBM suspensions were added 1:1 with 2.4 mg/ml bacterial microcrystalline cellulose from *Acetobacter xylinus sub sp. Sucrofermentans* (in 50 mM sodium acetate, 50 mM sodium chloride, pH = 5.0, total volume 100 µL) in microcentrifuge tubes containing magnetic stir bars. The samples were stirred to equilibrium at 1100 rpm, 4°C for two hours before centrifugation at 14,000 x g for 10 min. Two 10 µL aliquots were taken from the supernatant and analyzed by quantitative MALDI-TOF Mass Spectrometry (see above) to calculate unbound CBM concentration. Data were fitted to a single site Langmuir adsorption model (Eqn. 2.1) using OriginPro 9 software,

$$[B] = \frac{B_{max} \times K_{ads} \times [F]}{1 + K_{ads} \times [F]} \quad (2.1)$$

where B_{max} represents the total binding capacity of the CBM glycoform, K_{ads} represents the binding affinity and $[B]$ and $[F]$ represent bound and free concentrations respectively, as has been done elsewhere [49,52,62].

Table 2.2 - Summary of Site-specific Impacts of *O*-mannosylation on CBM Proteolytic Stability, Thermostability, and Adsorption to Crystalline Cellulose - Half-Life to Thermolysin Degradation and T_m results are presented as mean of three trials ± SD. Adsorption affinity consant, K_{ads} and B_{max} results are presented as the mean of two trials ± SD. *Denotes an averaged value of four trials ± SD.

CBM Variant	Half-Life to Thermolysin Degradation (hr)	T_m (°C)	K_{ads} (µM^{-1})	B_{max} (µmol/g)
1	0.23 ± 0.02	62.2 ± 0.6	0.0894 ± 0.0007*	24 ± 5*
2	0.28 ± 0.02	61.1 ± 0.7	0.16 ± 0.02	6.5 ± 0.7
3	0.23 ± 0.02	59.8 ± 0.6	0.34 ± 0.16	25 ± 7
4	0.208 ± 0.002	65 ± 1	0.06 ± 0.03	22 ± 11
5	1.09 ± 0.01	70.1 ± 0.9	0.4 ± 0.2	6 ± 1.3
6	2.13 ± 0.06	73.2 ± 0.7	0.21 ± 0.05	3.6 ± 0.8
7	2.10 ± 0.05	71.7 ± 0.8	0.13 ± 0.02	3 ± 1.0
8	0.49 ± 0.01	67.3 ± 0.6	0.35 ± 0.17	5 ± 2
9	0.55 ± 0.01	65.5 ± 0.7	0.31 ± 0.08	5.9 ± 0.7
10	0.54 ± 0.01	64.5 ± 0.5	0.25 ± 0.02	10.5 ± 0.6

Table 2.3 - Summary of the Impacts of Mixed *O*-mannosylation at Multiple Sites on CBM Proteolytic Stability, Thermostability, and Adsorption to Crystalline Cellulose - Half-Life to Thermolysin Degradation and T_m results are presented as mean of three trials ± SD. Adsorption affinity consant, K_{ads} and B_{max} results are presented as the mean of two trials ± SD. *Denotes an averaged value of four trials ± SD.

CBM Variant	Half-Life to Thermolysin Degradation (hr)	T_m (°C)	K_{ads} (μM^{-1})	B_{max} (μmol/g)
11	1.82 ± 0.04	74 ± 1	0.19 ± 0.09	13 ± 6
12	1.96 ± 0.07	72 ± 1	0.268 ± 0.002	16 ± 1.2
13	4.33 ± 0	75 ± 2	0.66 ± 0.05	6.6 ± 0.5
14	10.5 ± 0.8	77.9 ± 0.6	0.373 ± 0.008	9.6 ± 0.11
15	2.8 ± 0.2	75.4 ± 0.2	0.245 ± 0.003	5.6 ± 0.18
16	3.6 ± 0.4	75.2 ± 0.5	0.19 ± 0.02	7.1 ± 0.9
17	9.1 ± 0.6	77.4 ± 0.3	0.58 ± 0.08	4.9 ± 0.15
18	9.1 ± 0.6	76 ± 2	0.22 ± 0.01	13.4 ± 0.4

Table 2.4 - Summary of the Impacts of Uniform *O*-mannosylation at Multiple Sites on CBM Proteolytic Stability, Thermostability, and Adsorption to Crystalline Cellulose - Half-Life to Thermolysin Degradation and T_m results are presented as mean of three trials ± SD. Adsorption affinity consant, K_{ads} and B_{max} results are presented as the mean of two trials ± SD. *Denotes an averaged value of four trials ± SD. **Weak affinity to cellulose noted, no K_{ads} value could be obtained.

CBM Variant	Half-Life to Thermolysin Degradation (hr0.23)	T_m (°C)	K_{ads} (μM^{-1})	B_{max} (μmol/g)
19	3.36 ± 0.08	77 ± 2	0.155 ± 0.012	8.1 ± 0.5
20	5.5 ± 0.4	75.1 ± 0.8	~0**	~0**

2.6 References

1. T. K. Kirk and R. L. Farrell, *Annu Rev Microbiol*, 1987, **41**, 465-505.
2. L. R. Lynd, P. J. Weimer, W. H. van Zyl and I. S. Pretorius, *Microbiol Mol Biol Rev* 2002, **66**, 506-577.
3. M. E. Himmel, S. Y. Ding, D. K. Johnson, W. S. Adney, M. R. Nimlos, J. W. Brady and T. D. Foust, *Science*, 2007, **315**, 804-807.
4. S. P. Chundawat, G. T. Beckham, M. E. Himmel and B. E. Dale, *Annu Rev Chem Biomol Eng*, 2011, **2**, 121-145.
5. E. A. Bayer, J. P. Belaich, Y. Shoham and R. Lamed, *Annu Rev Microbiol*, 2004, **58**, 521-554.
6. A. L. Demain, M. Newcomb and J. H. Wu, *Microbiol Mol Biol Rev*, 2005, **69**, 124-154.
7. C. M. Fontes and H. J. Gilbert, *Annu Rev Biochem*, 2010, **79**, 655-681.
8. R. Brunecky, M. Alahuhta, Q. Xu, B. S. Donohoe, M. F. Crowley, I. A. Kataeva, S. J. Yang, M. G. Resch, M. W. Adams, V. V. Lunin, M. E. Himmel and Y. J. Bomble, *Science*, 2013, **342**, 1513-1516.
9. A. B. Boraston, D. N. Bolam, H. J. Gilbert and G. J. Davies, *Biochem J*, 2004, **382**, 769-781.

10. V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho and B. Henrissat, *Nucleic Acids Res*, 2013, **42**, D490-495.
11. J. Kraulis, G. M. Clore, M. Nilges, T. A. Jones, G. Pettersson, J. Knowles and A. M. Gronenborn, *Biochemistry*, 1989, **28**, 7241-7257.
12. M. Linder, G. Lindeberg, T. Reinikainen, T. T. Teeri and G. Pettersson, *FEBS Lett.*, 1995, **372**, 96-98.
13. M. Linder and T. T. Teeri, *Proc Natl Acad Sci U S A*, 1996, **93**, 12251-12255.
14. M. R. Nimlos, G. T. Beckham, J. F. Matthews, L. Bu, M. E. Himmel and M. F. Crowley, *J Biol Chem*, 2012, **287**, 20603-20612.
15. J. Lehtio, J. Sugiyama, M. Gustavsson, L. Fransson, M. Linder and T. T. Teeri, *Proc Natl Acad Sci U S A*, 2003, **100**, 484-489.
16. N. Deshpande, M. R. Wilkins, N. Packer and H. Nevalainen, *Glycobiology*, 2008, **18**, 626-637.
17. J. P. Hui, P. Lanthier, T. C. White, S. G. McHugh, M. Yaguchi, R. Roy and P. Thibault, *J Chromatogr B Biomed Sci Appl*, 2001, **752**, 349-368.
18. I. Stals, K. Sandra, S. Geysens, R. Contreras, J. Van Beeumen and M. Claeyssens, *Glycobiology*, 2004, **14**, 713-724.
19. R. Pinto, J. Carvalho, M. Mota and M. Gama, *Cellulose*, 2006, **13**, 557-569.
20. C. M. Payne, M. G. Resch, L. Chen, M. F. Crowley, M. E. Himmel, L. E. Taylor, 2nd, M. Sandgren, J. Stahlberg, I. Stals, Z. Tan and G. T. Beckham, *Proc Natl Acad Sci U S A*, 2013, **110**, 14646-14651.
21. M. J. Harrison, A. S. Nouwens, D. R. Jardine, N. E. Zachara, A. A. Gooley, H. Nevalainen and N. H. Packer, *Eur J Biochem*, 1998, **256**, 119-127.
22. G. T. Beckham, J. F. Matthews, Y. J. Bomble, L. Bu, W. S. Adney, M. E. Himmel, M. R. Nimlos and M. F. Crowley, *J Phys Chem B*, 2010, **114**, 1447-1453.
23. C. B. Taylor, M. F. Talib, C. McCabe, L. Bu, W. S. Adney, M. E. Himmel, M. F. Crowley and G. T. Beckham, *J Biol Chem*, 2012, **287**, 3147-3155.
24. C. Unverzagt and Y. Kajihara, *Chem Soc Rev*, 2013, **42**, 4408-4420.
25. R. M. Wilson, S. Dong, P. Wang and S. J. Danishefsky, *Angew Chem Int Ed*, 2013, **52**, 7646-7665.
26. P. Wang, S. Dong, J. H. Shieh, E. Peguero, R. Hendrickson, M. A. Moore and S. J. Danishefsky, *Science*, 2013, **342**, 1357-1360.
27. J. L. Price, D. Shental-Bechor, A. Dhar, M. J. Turner, E. T. Powers, M. Gruebele, Y. Levy and J. W. Kelly, *J Am Chem Soc*, 2010, **132**, 15359-15367.
28. G. B. Fields and R. L. Noble, *Int J Pept Protein Res*, 1990, **35**, 161-214.
29. Z. Tan, S. Shang, T. Halkina, Y. Yuan and S. J. Danishefsky, *J Am Chem Soc*, 2009, **131**, 5424-5431.
30. G. Johansson, J. Stahlberg, G. Lindeberg, A. Engstrom and G. Pettersson, *FEBS Lett*, 1989, **243**, 389-393.
31. U. Arnold, A. Schierhorn and R. Ulbrich-hofmann, *Eur J Biochem*, 1999, **259**, 470-475.
32. S. Ahmad, V. Kumar, K. B. Ramanand and N. M. Rao, *Protein Sci*, 2012, **21**, 433-446.
33. E. V. Hackl, *Biopolymers*, 2014, **101**, 591-602.
34. S. E. O'Connor, J. Pohlmann, B. Imperiali, I. Saskiawan and K. Yamamoto, *J Am Chem Soc*, 2001, **123**, 6187-6188.
35. S. R. Hanson, E. K. Culyba, T. L. Hsu, C. H. Wong, J. W. Kelly and E. T. Powers, *Proc Natl Acad Sci U S A*, 2009, **106**, 3131-3136.

36. A. Varki, *Essentials of glycobiology*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 2009.
37. A. Varki, *Glycobiology*, 1993, **3**, 97-130.
38. R. J. Sola, J. A. Rodriguez-Martinez and K. Griebenow, *Cell Mol Life Sci*, 2007, **64**, 2133-2152.
39. D. Shental-Bechor and Y. Levy, *Curr Opin Struc Biol*, 2009, **19**, 524-533.
40. M. L. Langsford, N. R. Gilkes, B. Singh, B. Moser, R. C. Miller, Jr., R. A. Warren and D. G. Kilburn, *FEBS Lett.*, 1987, **225**, 163-167.
41. A. B. Boraston, L. Sandercock, R. A. Warren and D. G. Kilburn, *J Mol Microbiol Biotechnol*, 2003, **5**, 29-36.
42. A. B. Boraston, R. A. Warren and D. G. Kilburn, *Biochem J*, 2001, **358**, 423-430.
43. M. Linder, M. L. Mattinen, M. Kontteli, G. Lindeberg, J. Stahlberg, T. Drakenberg, T. Reinikainen, G. Pettersson and A. Annila, *Protein Sci*, 1995, **4**, 1056-1064.
44. A. Schuster and M. Schmoll, *Appl Microbiol Biotechnol*, 2010, **87**, 787-799.
45. P. Heinzelman, C. D. Snow, I. Wu, C. Nguyen, A. Villalobos, S. Govindarajan, J. Minshull and F. H. Arnold, *Proc Natl Acad Sci U S A*, 2009, **106**, 5610-5615.
46. C. M. Dana, P. Saija, S. M. Kal, M. B. Bryan, H. W. Blanch and D. S. Clark, *Biotechnol Bioeng*, 2012, **109**, 2710-2719.
47. E. K. Culyba, J. L. Price, S. R. Hanson, A. Dhar, C. H. Wong, M. Gruebele, E. T. Powers and J. W. Kelly, *Science*, 2011, **331**, 571-575.
48. S. Takashima, M. Ohno, M. Hidaka, A. Nakamura and H. Masaki, *FEBS Lett*, 2007, **581**, 5891-5896.
49. J. Guo and J. M. Catchmark, *Biomacromolecules*, 2013, **14**, 1268-1277.
50. J. Medve, J. Stahlberg and F. Tjerneld, *Appl Biochem Biotechnol*, 1997, **66**, 39-56.
51. H. Palonen, M. Tenkanen and M. Linder, *Appl Environ Microbiol*, 1999, **65**, 5229-5233.
52. K. Igarashi, T. Uchihashi, A. Koivula, M. Wada, S. Kimura, T. Okamoto, M. Penttila, T. Ando and M. Samejima, *Science*, 2011, **333**, 1279-1282.
53. D. Gao, S. P. Chundawat, A. Sethi, V. Balan, S. Gnanakaran and B. E. Dale, *Proc Natl Acad Sci U S A*, 2013, **110**, 10922-10927.
54. X. Li, W. T. t. Beeson, C. M. Phillips, M. A. Marletta and J. H. Cate, *Structure*, 2012, **20**, 1051-1061.
55. M. Wu, G. T. Beckham, A. M. Larsson, T. Ishida, S. Kim, C. M. Payne, M. E. Himmel, M. F. Crowley, S. J. Horn, B. Westereng, K. Igarashi, M. Samejima, J. Stahlberg, V. G. Eijsink and M. Sandgren, *J Biol Chem*, 2013, **288**, 12828-12839.
56. C. Uriel, J. Ventura, A. M. Gomez, J. C. Lopez and B. Fraser-Reid, *J Org Chem*, 2012, **77**, 795-800.
57. D. Varon, C. Unverzagt, E. Lioy, M. E. Patarroyo and X. Schratt, *Aust J Chem*, 2002, **55**, 161-165.
58. L. Chen and Z. Tan, *Tetrahedron Lett*, 2013, **54**, 2190-2193.
59. T. Toyokuni, J. S. Dileep Kumar, P. Gunawan, E. S. Basarah, J. Liu, J. R. Barrio and N. Satyamurthy, *Mol Imaging Biol*, 2004, **6**, 324-330.
60. M. W. Duncan, H. Roder and S. W. Hunsucker, *Brief Funct Genomic Proteomic*, 2008, **7**, 355-370.
61. S. P. Voutilainen, S. Nurmi-Rantala, M. Penttila and A. Koivula, *Appl Microbiol Biotechnol*, 2014, **98**, 2991-3001.
62. N. Sugimoto, K. Igarashi, M. Wada and M. Samejima, *Langmuir*, 2012, **28**, 14323-14329.

Chapter 3

Molecular-Scale Features that Govern the Effects of *O*-Glycosylation on a Carbohydrate-Binding Module

3.1 – Introduction

The capability of glycans to affect protein properties opens the possibility of custom-designed glycan motifs that can be introduced to produce proteins with desirable properties.[2,3] Regrettably, due to the current lack of quantitative knowledge about the effects of protein glycosylation, such glycoengineering approaches are still largely empirical, which makes research in this area challenging, time-consuming, and costly.[5] A detailed, molecular-level understanding of the features and factors associated with the effects of natural glycosylation of proteins would facilitate the process. Recent studies of protein *N*-glycosylation have clearly demonstrated that such information is useful in guiding the glycoengineering of proteins.[6-9] Unfortunately, unlike *N*-glycosylation, no universal consensus sequence has been identified for *O*-glycosylation, which seriously limits access to glyco-variants and hampers the detailed study and application of *O*-glycosylation.[10-13]

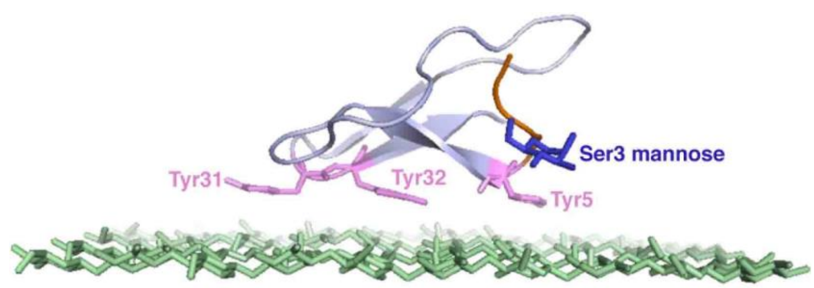


Figure 3.1 - The NMR structure of the Family 1 CBM and the top layer of cellulose.[1] The tyrosine residues are shown in purple. The *O*-linked mannose at Ser3 site is shown in blue.[4]

In the present study, we have chosen to investigate the molecular features that control the effects of *O*-glycosylation at a specific site, Ser3, in the Family 1 carbohydrate-binding module (CBM) of the Glycoside Hydrolase Family 7 cellobiohydrolase from the cellulolytic fungus, *Trichoderma reesei* (*Tr*Cel7A), a key enzyme in the cellulosic biofuels industry (Figure 3.1). Family 1 CBMs are small, natively glycosylated, synthetically tractable, and their glycosylation poses interesting stability and functional questions, making them excellent model systems to study *O*-glycosylation.[14,15] The amino acid Ser3 was chosen for in depth study after we established that, for the CBM, glycosylation at this position is responsible for the most significant enhancements in desirable enzyme properties: proteolytic stability against thermolysin degradation, thermostability, and binding affinity towards bacterial microcrystalline cellulose (BMCC).[16] The fact that glycosylation at this site caused the largest, and hence most detectable, changes makes glycosylation at Ser3 an ideal choice for identifying the molecular determinants of natural *O*-glycosylation's observed effects in this system.

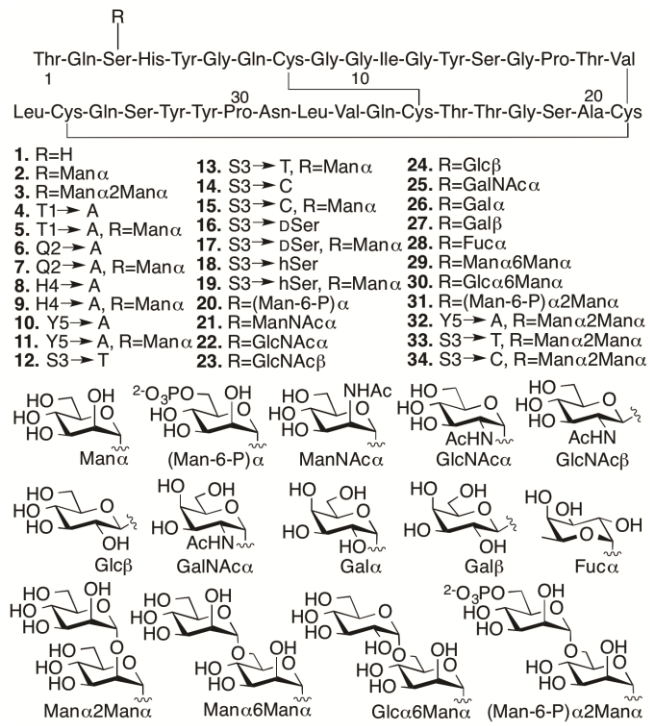


Figure 3.2 - Synthetic CBM isoforms and the structures of *O*-linked glycans.

We conducted several comparative studies to determine the contributions of multiple molecular features (Figure 3.2).[16,17] Like our previous studies, we first designed and prepared 31 new CBM isoforms with systematic variations in amino acid sequence, glycopeptide linkage, glycan structure, and anomeric configuration to assess the importance of each of these structural elements in mediating the effects of *O*-glycosylation (Figure 3.2, **4-36**).[18-20] Three previously characterized CBM isoforms, which all have the natural amino acid sequence and either no glycans (**1**), a single mannose (**2**), or a single di-mannose (**3**), were also included as controls.[16]

3.2 Results and discussion

Since chemical glycosylation is not controlled by the structural features of peptides, it is capable of generating almost any glyco-variant.[10,11,13] Synthesis of CBM isoforms was conducted with Fmoc-based solid-phase peptide synthesis (SPPS). During SPPS, all sugar hydroxyl and/or phosphate groups on the side chains of the glycoamino acid building blocks where protected as acetyl[16] or benzyl esters,[21] respectively, which are stable during peptide coupling procedures and easily removed under carbohydrate-compatible conditions. Since most of the glycoamino acid building blocks used in this study are not commercially available, we first identified efficient synthetic methods to quickly prepare glycosylated Fmoc-Ser, Fmoc-Thr, Fmoc-D-serine (DSer), and homoserine (hSer) in gram scales ("Experiments" Section). To ensure strict control over anomeric stereochemistry, reaction conditions were carefully chosen for high diastereomeric selectivity and every synthetic glycoamino acid building block was analyzed using 2D HSQC NMR to confirm absolute anomeric configuration. After synthesizing all the desired building blocks, our previously developed one-pot synthesis and folding method enabled us to quickly generate all 31 desired CBM isoforms in high purity and with good yields for glycopeptide synthesis (ranging from 30% for **6** to 6% for **20**) ("Experiments" Section).[16,22]

With the CBM isoform library completed, we began by investigating how amino acid side chains close to the glycosylation site alter the effects of *O*-glycosylation using Ala-scanning mutagenesis; four mutations were used for this. For each mutation, the unglycosylated CBM was compared to the corresponding mono-mannosylated glycopeptide in terms of proteolytic stability, thermostability, and binding affinity, following previously described protocols (Figure 3.3).[8,16,23] As shown in the left side of Figure 3.3 (top panel), Ala mutations at any residue adjacent to the Ser3 glycosylation site (Thr1, Gln2, His4, or Tyr5) did not significantly alter the thermolysin half-life of the unglycosylated CBMs (compare **1**, **4**, **6**, **8** and **10**). Our previous study established that glycosylation of Ser3 significantly stabilized the CBM towards protease degradation,[16] but this trend holds true in only two of the four Ala-mutant sequences (T1A, compare **4** and **5** and H4A,

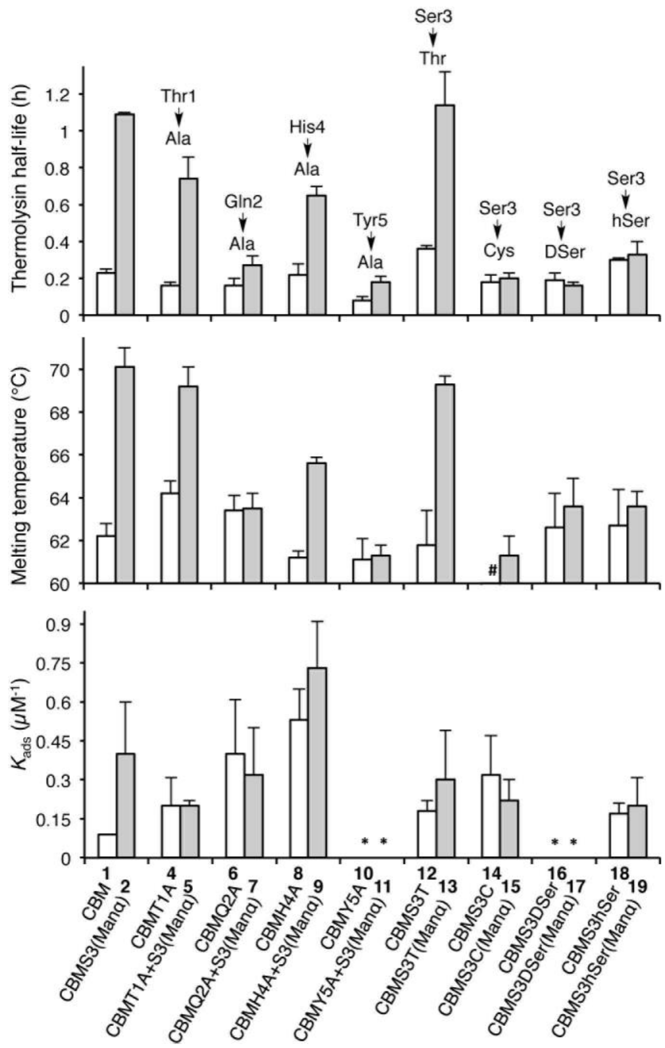


Figure 3.3 - The contributions of amino acids to the effects of the Ser3 glycosylation on the proteolytic stability (half-life to thermolysin degradation), thermostability (melting temperatures measured by variable temperature CD), and binding affinity (K_{ads} values on BMCC) of the *Tr*Cel7A CBM. All error bars reported are standard deviations of data achieved from three separate trials. The structural feature of each isoform is implied by its name, i.e. CBMS3(Manα) representing the isoform containing a single mannose α-linked to Ser3, CBMQ2A+S3(Manα) representing the isoform containing a Gln-to-Ala mutation at position 2 and a single mannose α-linked to Ser3, and CBMS3hSer(Manα) representing the isoform containing a Ser-to-hSer mutation at position 3 and a single mannose α-linked to hSer3. # 53°C. * No observable binding noted.

compare **8** and **9**). In contrast, the attachment of a single mannose to Ser3 in both the Q2A mutant (compare **6** and **7**) and the Y5A mutant (compare **10** and **11**) leads to almost no increase in their thermolysin half-life. Thermostability of these CBM sequences follows a similar trend (Figure 3.3, middle panel). The binding affinity exhibits a very different pattern (Figure 3.3, bottom panel). For unglycosylated isoforms, replacing Thr1, Gln2, His4, or Tyr5 with Ala induces pronounced and widely variable changes in BMCC binding, from large increases (Q2A, **6** and H4A, **8**) to totally eliminating binding (Y5A, **10**).[14] Mono-mannosylation of any of these mutants gives only small negative or positive deviations to the binding constant.

To quantify how side-chain properties like hydrophobicity, glycosidic bond character, side-chain orientation, and length alter the influence of mannosylation, Ser3 was replaced by four similar amino acids: Thr **12**/**13**, Cys **14**/**15**, DSer **16**/**17**, and hSer **18**/**19.** As shown in the right side of Figure 3.3 (top and middle panel), replacement of Ser3 by Thr has little influence on the stability of either unglycosylated or mannosylated CBM. Replacement by Cys, DSer, or hSer, however, significantly diminishes the stabilizing effect of mannose. Thermostability followed a comparable trend. Interestingly, CBM variant **14** has a 10°C lower melting temperature than that of CBM **1**. This may be a result of less stable disulfide bonds in the presence of a free Cys.[24] Capping the free sulfhydryl group with a mannose brings the melting temperature back up to 61 °C. Binding affinity of the unglycosylated CBM increased upon substitution of Ser3 by Thr, Cys, or hSer (compare **1** to **12**, **14,** and **18**), but mannosylation of these mutant CBMs shows a very different trend. Both Thr and hSer-containing isoforms showed insignificant increases in binding affinity upon glycosylation (compare **12** to **13** and **18** to **19**), while glycosylation of the Ser-to-Cys mutation results in a small decrease (compare **14** and **15**). Neither of the DSer mutants (**16** and **17**) shows any obvious binding to BMCC.

Understanding the impact of glycan composition and linkage stereochemistry on the effects of Ser3 glycosylation was our next goal. For this, we directly compared CBM glycoforms with systematically varied glycan structures in two final studies (Figure 3.4). To elucidate the potentially variable influence of different mono-saccharides nine CBM glycoforms, **20**-**28**, were compared to unglycosylated **1** and mannosylated **2**. As shown in Figure 3.4A, half-lives towards thermolysin degradation and melting temperatures vary in a remarkably similar pattern across these isoforms, with the mannosylated isoform **2** having the highest of both types of stability. Changes to binding affinity followed a distinctly different pattern, although the three CBM glyco-variants with the lowest stabilities (**23**, **24**, and **27**), also have low affinities to the BMCC substrate. Of particular note, we observe that the anomeric stereochemistry of the glycosidic linkage has a more significant influence than most other structural features on the effects of glycosylation (compare **1** to **2**, **22**, and **23**). While the α-linked mono-saccharides on **2** and **22** gave significant improvements over the unglycosylated **1**, the β-linked mono-saccharide on **23** had almost no effect on the proteolytic stability, thermostability, or binding affinity of the CBM. Similarly, the α-linked galactose on **26** significantly improved the melting temperature and modestly improved the proteolytic stability, but the same galactose attached through a β-linkage in **27** gave almost no increase in either property. To probe the influence of a second glycan unit, we also examined six new CBM glyco-variants containing either α1,2- (**3**, **31**, **32**, **33**, and **34**) or α1,6- (**29** and **30**) glycosidic linkages. Once agai n, as shown in Figure 3.4B, the proteolytic stability and thermostability exhibit similar trends after attachment of the additional sugar residues while the binding affinity varies independently. Only the attachment of α1,2-linked mono-mannose to Man-α-Ser (**3**) and Man-α-Thr (**33**) causes a further increase over mono-mannosylated CBM (**2**) in either stability measure. Similarly, only the attachment of α1,6-linked mono-glucose to Man-α-Ser (**30**) and α1,2-linked mono-mannose to Man-α-Thr (**33**) causes a further increase in the binding

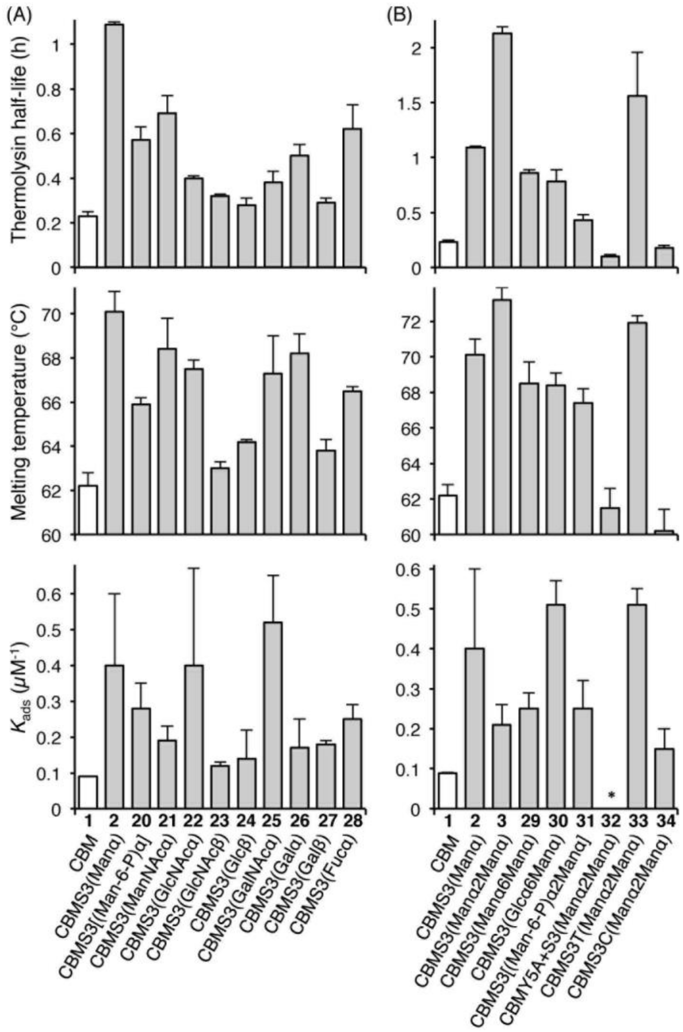


Figure 3.4 - The contributions of different glycans to the effects of Ser3 glycosylation on the proteolytic stability, thermostability, and binding affinity of the *Tr*Cel7A CBM. All error bars reported are standard deviations of data achieved from three separate trials. * No observable binding noted.

affinity. Mutating Tyr5 to Ala or Ser3 to Cys significantly diminishes or even abolishes the effects of glycosylation (compare **3** to **32** and **34**). Phosphorylation of the 6-hydroxyl of both mono- and di-mannose, which may naturally occur in Family 1 CBMs, adversely impacts the effects of mannosylation (compare **2** to **20** and **3** to **31**).[25]

The results obtained by comparing the properties of 34 CBM isoforms provide new insights into the molecular determinants of the effects of *O*-glycosylation on the stability and function of this protein. A well-established effect of protein glycosylation is an increase in proteolytic stability, either by increasing the rigidity of the protein, or by providing a steric barrier that hinders protease access to the peptide bonds.[19,26-28] Our results indicate that steric hindrance may be less important than peptide rigidity in the case of CBM *O*-glycosylation. Support for this conclusion comes from the CBM variants **4-19** (Figure 3.3). Since the sizes of their glycan moieties are identical, the differences observed in their susceptibilities to thermolysin hydrolysis can be attributed to altered conformational rigidity.[29] More specifically, the rigidity seems largely controlled by Gln2, Tyr5 and the glycosylated amino acid residue because glyco-variants with Gln2-to-Ala, Tyr5-to-Ala, or Ser3-to-Cys, DSer, or hSer mutations do not exhibit large changes to the proteolytic stability upon glycosylation. Further support for the limited role of steric hindrance in thermolysin resistance comes from the results of the analysis of CBM variants **20-34**. As shown in Figure 3.4, different extents of proteolytic stability are conferred by different mono- or di-saccharides of similar sizes at Ser3 and the stereochemistry at the anomeric carbon plays a large role in modulating the proteolytic stability.

Thermostability is another important property known to be affected by glycosylation.[3] Recent studies have suggested that local interactions, such as carbohydrate-aromatic interactions, strongly contribute to the large stabilizing impact of *N*-glycosylation.[3,8,30]. Other studies into O-glycosylation have also revealed the importance of local interactions between carbohydrate and peptide for O-glycopeptide conformation.[31] Our results here continue to support this conclusion for O-glycosylation. Mutating Tyr5 to Ala (compare **11**, **32** and **2)** led to a substantial decrease in the thermostability. In addition, we observe a loss of mannosylation-induced stability for the Q2A mutant. The specific role played by Gln2 is not clear, but previous findings from studies of protein-carbohydrate interactions suggest that its planar polar side chain may be involved in several hydrogen bonds linking the protein and glycan.[32,33] The importance of these local interactions in stabilizing the CBM is further underscored by the fact that the β-linked glycans have very limited effects on CBM thermostability. This can be explained by decreased contact between glycan and nearby amino acids since the β-glycosidic linkages directs the glycan away from the peptide.[34]

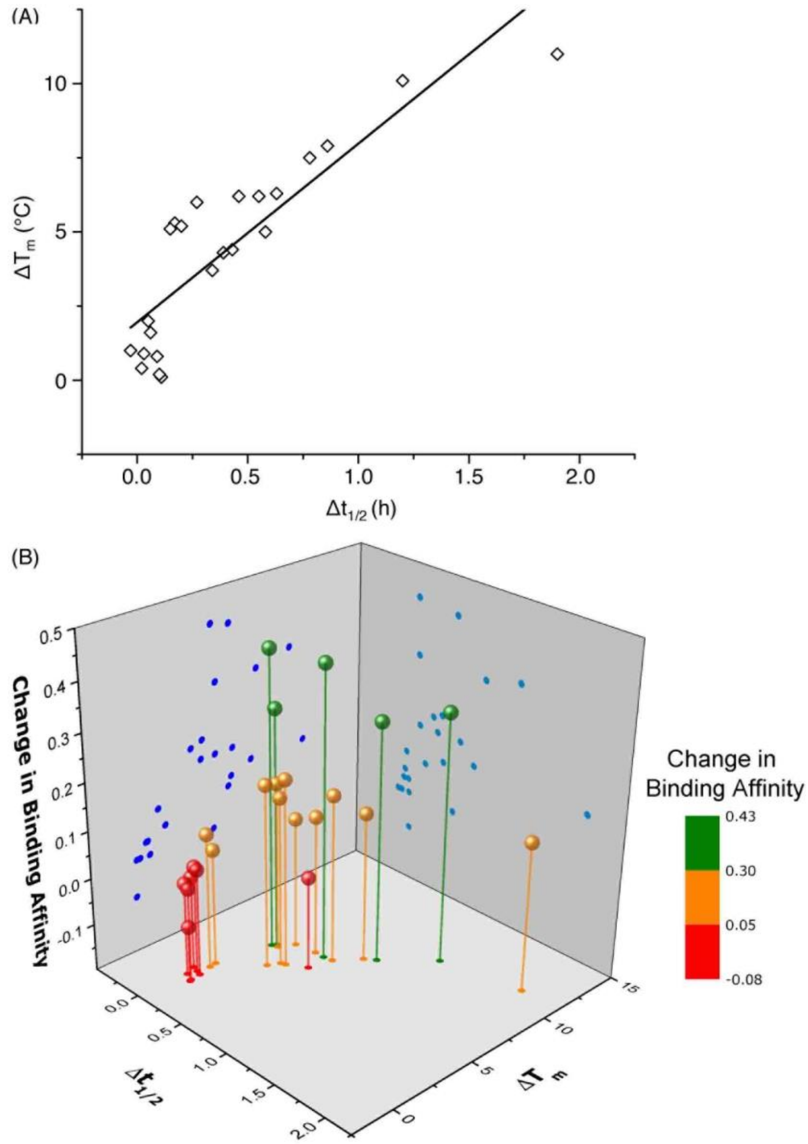


Figure 3.5 - Correlation of (A) the change in melting temperature (ΔT_m) and change in half-life during thermolysin degradation ($\Delta t_{1/2}$), and (B) the change in binding affinity upon glycosylation. Data points represent differences between CBM glyco-variants and their corresponding unglycosylated counterparts. The data for the CBM pairs **15**/**14** and **35**/**14** are not included in the plot because of their unique characteristics.

One important questions in glycobiology is whether altered biophysical properties and biological function of glycoproteins are related.[3,35] The answer to this question is critical to the practice of glycoengineering. A positive answer would imply that it is possible to simultaneously increase protein stability and function by glycosylation. As shown in Figure 3.5, our results reveal a striking correlation between variations in the CBM's proteolytic stability and thermal stability, suggesting common molecular forces are responsible for both the thermostabilizing effects of mannosylation and increasing the rigidity of the same site.[31] Most interestingly, our study reveals a strong link between glycoprotein stability and function: CBM glyco-variants with much lower affinities towards BMCC generally also have low stabilities, those with higher binding affinities often have intermediate stabilities, and the highest stabilities do not necessarily correlate with the highest binding affinities (Figure 3.5B). Existing theories shed some light on these observations: intermediate stability or flexibility would allow the CBM to maintain its native structure in solution while permitting the peptide to adopt optimal conformations for dynamically binding to cellulose.[36]

3.3 Conclusion

In summary, by using chemical synthesis, we were able to systematically vary the amino acid sequence at the *N*-terminal end of a model Family 1 CBM and the glycan structures at Ser3, a highly conserved and functionally important glycosylation site.[17] By comparing these variants' characteristics, this study provides new insights into the molecular basis for the effects of CBM Ser3 *O*-glycosylation. We have shown that planar polar (Gln) and aromatic amino acid (Tyr) residues as well as *O*-glycans α-linked to Ser or Thr are important for the effects of CBM *O*-glycosylation. More importantly, our data suggest that CBM proteolytic and thermostability are linearly related while the CBM function (i.e., binding affinity) peaks at moderate levels of stability. This type of knowledge is expected to facilitate future investigations into the glycosylation of other proteins, including those with therapeutic and industrial relevance.

Although there are many challenges remaining, this work is one small but significant contribution to the currently opaque process of rationally engineering proteins, and provides an illustrative example of simultaneously improving stability and function.

3.4 Experiments

3.4.1 Materials

All commercial reagents and solvents were used as received. Unless otherwise noted, all reactions and purifications were performed under air atmosphere at room temperature. All LC-MS analyses were performed using a Waters Acquity™ Ultra Performance LC system equipped with Acquity UPLC® BEH 300 C4, 1.7μm, 2.1 x 100 mm column at flow rates of 0.3 and 0.5 mL/min. The mobile phase for LC-MS analysis was a mixture of H_2O (0.1% formic acid, v/v) and acetonitrile (0.1% formic acid, v/v). All preparative separations were performed using a LabAlliance HPLC solvent delivery system equipped with a Rainin UV-1 detector and a Varian Microsorb 100-5, C18 250x21.4 mm column at a flow rate of 16.0 mL/min. The mobile phase for HPLC purification was a mixture of H_2O (0.05% TFA, v/v) and acetonitrile (0.04% TFA, v/v). A Waters SYNAPT G2-S system was used mass spectrometric analysis. All circular dichroism (CD) spectra were obtained using an Applied PhotophysicsChirascan™-plus CD spectrometer.

3.4.2 Synthesis of glycoamino acids

The glycoamino acid building blocks Fmoc-Ser(Ac4Manα1)-OH (**35**), Fmoc-Ser(Ac4Manα1-2Ac$_3$Manα1)-OH (**36**), Fmoc-Thr(Ac4Manα1)-OH (**37**), and Fmoc-Thr(Ac4Manα1-2Ac$_3$Manα1)-OH (**53**) were prepared according to the previously reported method.[37] Glycoamino acid building block **44** was purchased from AnaSpec. All the other building blocks, **38-43**, **45-52**, and **54** are prepared as described below. The spectroscopic characterizations (^{1}H NMR, ^{13}C NMR, IR, and high-resolution MS) of all new compounds are reported.

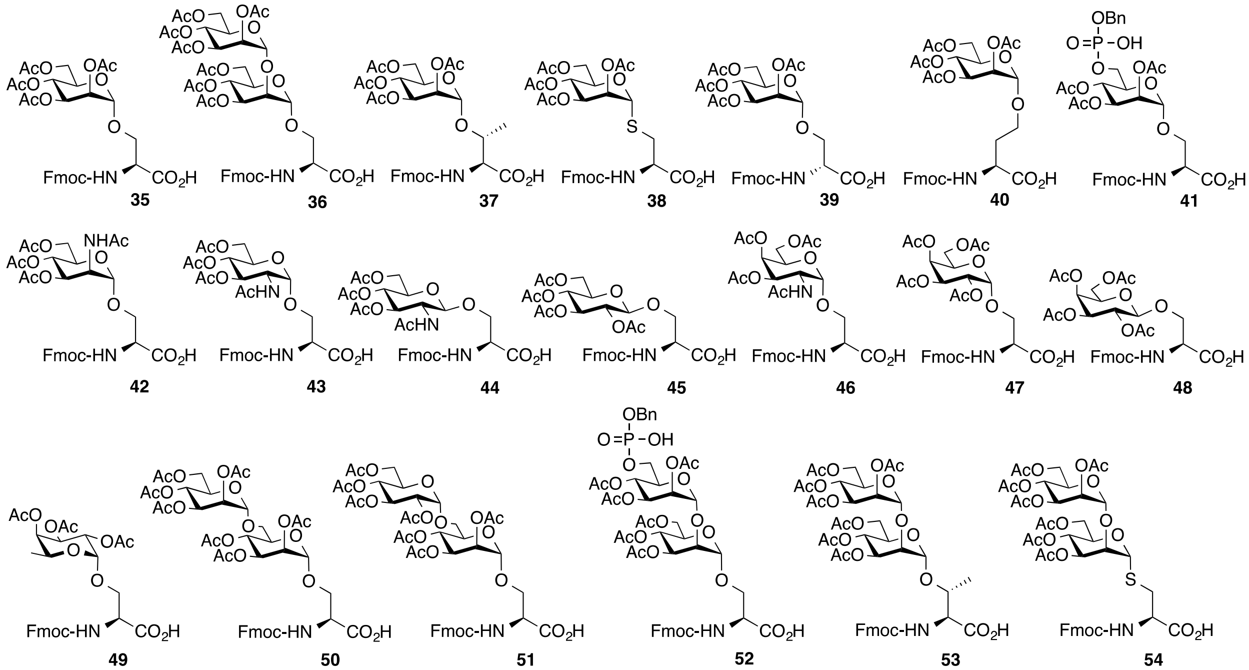
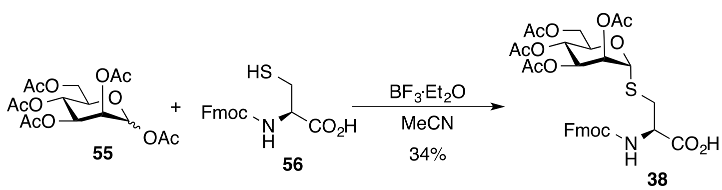
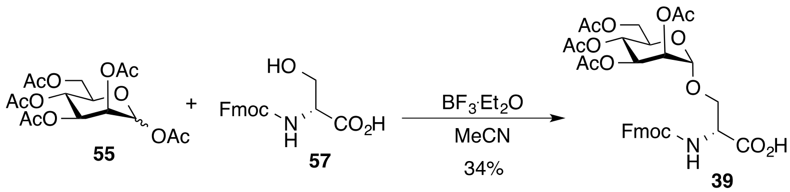


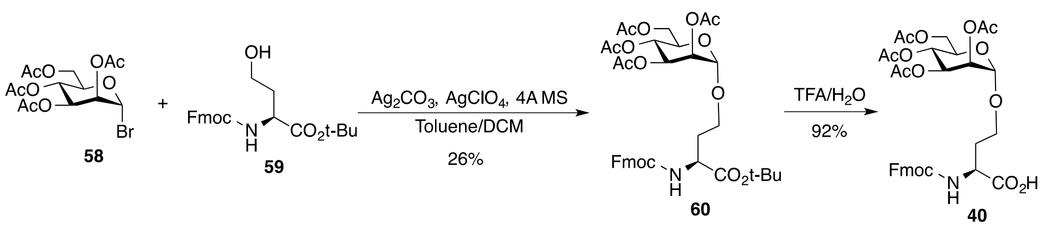
Figure 3.6 - *O*-linked glycoamino acid building blocks used for the synthesis of CBM glyco-variants.



Synthesis of glycoamino acid **38.** The 1,2,3,4,6-penta-O-acetyl-D-mannopyranose **55** was prepared as reported in the literature[37]. To the solution of **55** (682 mg, 1.75 mmol) and **56** (Fmoc-Cys-OH, 900 mg, 2.62 mmol) in MeCN (35 ml), $BF_3{\cdot}OEt_2$ (0.81 ml, 5.25 mmol) was added. The resulting mixture was stirred at room temperature for 28 h under argon. The solvent was removed under reduced pressure, and the resulting residue was diluted with EtOAc then washed with water. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH = 4:1:0.5→3:1:0.4→2:1:0.3) to give **38** (404 mg, 34%) as a white foam.

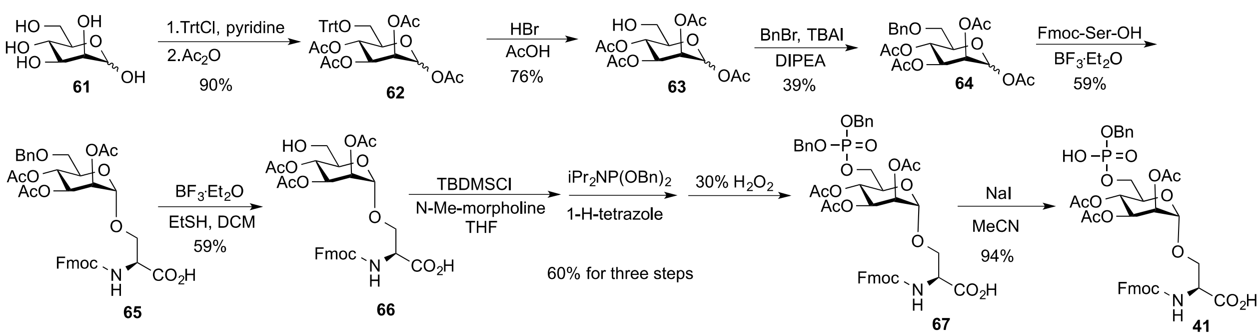


Synthesis of glycoamino acid **39.** To the solution of **55** (390 mg, 1.0 mmol) and **57** (Fmoc-DSer-OH, 490 mg, 1.5 mmol) in MeCN (12 ml), $BF_3{\cdot}OEt_2$ (0.38 ml, 3.0 mmol) was added. The resulting mixture was stirred at room temperature for 23 h under argon. The solvent was removed under reduced pressure, and the resulting residue was diluted with EtOAc then washed with water. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH = 4:1:0.5→3:1:0.4→2:1:0.3) to give **39** (223 mg, 34%) as a white foam.



Synthesis of glycoamino acid **40.** The 2,3,4,6 tetra-O-acetyl mannosyl bromide **58** was prepared as reported in the literature.[37] A solution of **59** (1.0 g, 2.51 mmol), Ag_2CO_3 (1.03 g, 3.77 mmol), 4A molecular sieves (4A MS) (2.5 g) in toluene (8.5 ml) and DCM (12.8 ml) was stirred at 0 ^{o}C for 30 minutes[38]. Then, a solution of $AgClO_4$ (130 mg, 0.63 mmol) in toluene (3.1 ml) was added dropwise at 0 ^{o}C. Subsequently, a solution of **58** (1.47 g, 3.77 mmol) in a mixture of DCM (6.4 ml) and toluene (6.4 ml) was very slowly added dropwise at 0 ^{o}C. The mixture was stirred, in the dark, at room temperature for 21 h. The reaction was diluted with DCM, filtered through Celite and washed with H_2O then $NaHCO_3$ (sat.). The organic layer was dried with Na_2SO_4 and concentrated under reduced pressure. The residue was purified by flash chromatography on a silica gel column (Hex/EtOAc = 2:1→1:1) to give **60** (474 mg, 26%) as a white foam.

60 (470 mg, 0.65 mmol) was dissolved in a solution of TFA/H_2O (95:1, 8.0 ml) and stirred at room temperature for 2.5 h. The solvent was removed under reduced pressure by co-evaporation with toluene and the remaining residue was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH = 1:1:0→1:1:0.2) to give **40** (400 mg, 92%) as a white foam. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.82 – 7.75 (m, 2H), 7.62 (d, *J* = 7.5 Hz, 2H), 7.46 – 7.37 (m, 2H), 7.33 (td, *J* = 7.4, 1.3 Hz, 2H), 5.81 (d, *J* = 6.6 Hz, 1H), 5.38 – 5.24 (m, 2H), 5.21 (s, 1H), 4.81 (s, 1H), 4.50 (d, *J* = 5.7 Hz, 1H), 4.43 (d, *J* = 7.1 Hz, 2H), 4.36 – 4.27 (m, 1H), 4.23 (t, *J* = 6.7 Hz, 1H), 4.12 (d, *J* = 12.5 Hz, 1H), 4.01 (d, *J* = 8.9 Hz, 1H), 3.90 (s, 1H), 3.53 (dd, *J* = 10.7, 5.3 Hz, 1H), 2.35 – 2.20 (m, 2H), 2.15 (s, 3H), 2.11 (s, 3H), 2.07 (d, *J* = 1.6 Hz, 3H), 2.01 (s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 171.37 , 170.80 , 170.37 , 169.80 , 155.82 , 143.68 , 141.31 , 127.77 , 127.08 , 125.08 , 125.03 , 120.02 , 120.00 , 97.55 , 69.63 , 69.47 , 68.92 , 66.98 , 65.65 , 63.84 , 62.51 , 51.35 , 47.18 , 31.33 , 20.87 , 20.79 , 20.75 , 20.71. IR (NaCl, film): 3341, 3066, 2952, 1748, 1522, 1451, 1371, 1228, 1138, 1117, 1049, 980 cm^{-1}. HRMS (ESI) Calcd. for $C_{33}H_{37}NNaO_{14}$ $[M + Na]^+$ requires 694.2107, Found: 694.2103.



Synthesis of glycoamino acid **41.** To a stirred solution of D-(+)-mannose **61** (5 g, 27.8 mmol) in pyridine (25 ml) trityl chloride (TrtCl) (8.5 g, 30.55 mmol) was added and the resulting mixture was stirred for 1.5 h at 40 ^{o}C. After cooling to 0 ^{o}C, Ac_2O (15 ml) was added to the mixture and the resulting solution was stirred overnight at room temperature. The reaction mixture was poured into ice water and extracted with DCM. The organic phase was dried over anhydrous Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column

(Hex/EtOAc = 5:1) to give **62** (14.71 g, 90%) as a white foam. Product matched the previously known spectra of **62**[39].

To a stirred solution of **62** (3 g, 5.08 mmol) in AcOH (10 ml) was added 33% HBr in AcOH (1.0 ml). The resulting mixture was stirred for 1 min. The Ph_3CBr formed was immediately removed by suction filtration. The filtrate was diluted with cold water and extracted with DCM. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc =2:1→1:1) to give **63** (1.34 g, 76%) as a white foam. Product matched the previously known spectra of **63**[39].

To a flask with **63** (23 g, 66 mmol) and tetra-n-butylammonium iodide (TBAI) (7.3 g, 19.8 mmol) was added DIPEA (45 ml, 264 mmol) and BnBr (31.6 ml, 264 mmol). The resulting mixture was stirred at 90 ^{o}C for 4 h. The reaction was diluted with DCM and washed with water. The organic phase was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc =4:1→2:1) to give **64** (11.16 g, 39%) as an oil (α/β=2:1). ^{1}H-NMR (400 MHz, $CDCl_3$) δ 7.27-7.36 (m, 7.5H, H-Ph), 6.10 (d, *J* = 2.0 Hz, 1H, H-1α), 5.85 (d, *J* = 1.2 Hz, 0.5H, H-1β), 5.47 (dd, *J* = 3.2 Hz, 1.2 Hz, 0.5H, H-2β), 5.31-5.41 (m, 2.5H, H-3α, H-4α, H-4β), 5.25 (dd, *J* = 3.2 Hz, 2.0 Hz, 1H, H-2α), 5.11 (dd, *J* = 10.0 Hz, 3.2 Hz, 0.5H, H-3β), 4.46-4.59 (m, 3H, CH_2-Bn), 3.98-4.02 (m, 1H, H-5α), 3.73-3.77 (m, 0.5H, H-5β), 3.57-3.60 (m, 3H, H-6α, H-6β), 2.21 (s, 1.5H, CH_3-Acβ), 2.162 (s, 3H, CH_3-Acα), 2.155 (s, 3H, CH_3-Acα), 2.09 (s, 1.5H, CH_3-Acβ), 2.002 (s, 3H, CH_3-Acα), 1.997 (s, 1.5H, CH_3-Acβ), 1.92 (s, 3H, CH_3-Acα), 1.90 (s, 1.5H, CH_3-Acβ). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 170.3, 170.1, 169.9, 169.8, 169.62, 169.58, 168.4, 168.2, 137.7, 137.6, 128.6, 128.4, 128.0, 127.9, 127.8, 127.7, 90.7, 90.5, 77.2, 74.5, 73.67, 73.61, 72.0, 70.9, 68.9, 68.8, 68.6, 68.4, 68.3, 66.4, 66.3, 20.9, 20.82, 20.79, 20.74, 20.68, 20.57. IR (NaCl, film): 3064, 3031, 2937, 2870, 1761, 1454, 1432, 1370, 1237, 1149, 1055, 972, 738, 701 cm^{-1}. HRMS (ESI) Calcd. for $C_{21}H_{26}NaO_{10}$ $[M + Na]^+$ requires 461.1419, Found: 461.1417.

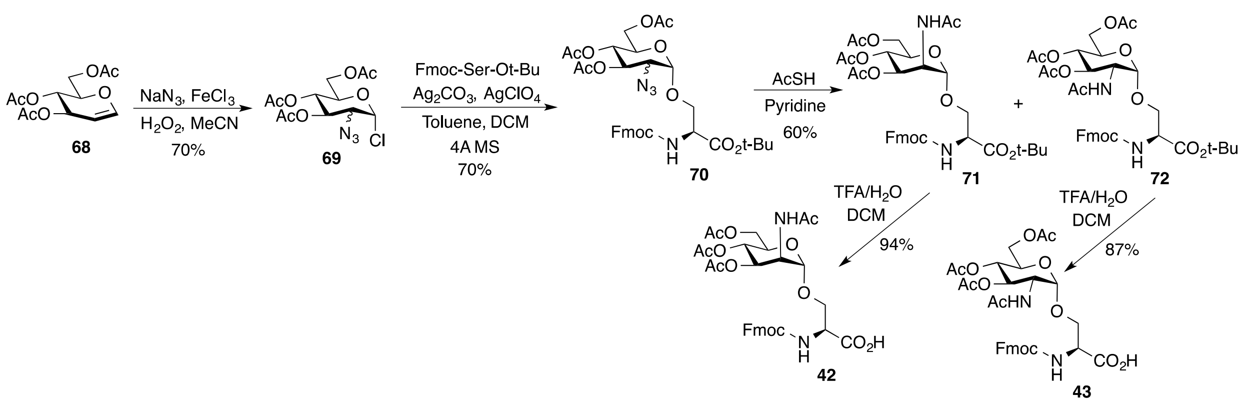
To the solution of **64** (400 mg, 0.91 mmol) and Fmoc-Ser-OH (448 mg, 1.37 mmol) in MeCN (22 mL), $BF_3 \cdot OEt_2$ (0.42 ml, 2.74 mmol) was added[37]. The resulting mixture was stirred at room temperature for 24 h under argon. The solvent was removed under reduced pressure and the residue was diluted with EtOAc and washed with water. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH =2:1:0.3) to give **65** (379 mg, 59%) as a white foam. ^{1}H-NMR (400 MHz, $CDCl_3$) δ 7.74 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.56-7.58 (m, 2H, H-Fmoc), 7.24-7.38 (m, 9H, H-Fmoc, H-Ph), 6.54 (d, *J* = 8.8 Hz, 1H, H-NH), 5.33-5.43 (m, 3H, H-2, H-3, H-4), 4.87 (s, 1H, H-1), 4.68 (brs, 1H, H-α), 4.49 (dd, *J* = 43.2 Hz, 12.0 Hz, 2H, CH_2-Bn), 4.03-4.44 (m, 6H, H-5, CH_2-β, CH_2-Fmoc, CH-Fmoc), 3.51 (s, 2H, H-6), 2.14 (s, 3H, CH_3-Ac), 2.00 (s, 3H, CH_3-Ac), 1.84 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 170.7, 170.3, 169.8, 156.1, 143.8, 141.27, 141.24, 137.5, 128.8, 128.3, 127.92, 127.86, 127.7, 127.1, 125.3, 119.9, 98.3, 73.5, 70.1, 69.5, 68.4, 67.3, 66.7, 47.1, 43.9, 20.9, 20.8, 20.7. IR (NaCl, film): 3339, 3065, 1754, 1707, 1370, 1224, 1050, 740, 700 cm^{-1}. HRMS (ESI) Calcd. for $C_{37}H_{39}NNaO_{13}$ $[M + Na]^+$ requires 728.2314, Found: 728.2304.

The mixture of **65** (300 mg, 0.42 mmol), EtSH (2.1 ml) and $BF_3 \cdot OEt_2$ (523 μl, 3.40 mmol) in DCM (4.2 ml) was stirred for 6 h at room temperature under argon. The reaction was quenched with water and extracted with EtOAc. The organic phase was washed with brine, dried (Na_2SO_4), filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH =3:2:0.5→1:1:0.2) to give **66** (152 mg, 59%) as an oil. ^{1}H-NMR (400 MHz, CD_3OD) δ 7.82 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.72-7.74 (m, 2H, H-Fmoc), 7.39-7.43 (m, 2H, H-Fmoc), 7.32-7.36 (m, 2H, H-Fmoc), 5.23-5.36 (m, 3H, H-2, H-3, H-4), 4.91 (s, 1H, H-1), 3.89-4.48 (m, 7H, H-5, H-α, CH_2-β, CH_2-Fmoc, CH-Fmoc), 3.66 (dd, *J* = 12.4 Hz, 2.4 Hz, 1H, H-6), 3.56 (dd, *J* = 12.4 Hz, 5.2 Hz, 1H, H-6), 2.12 (s, 3H, CH_3-Ac), 1.97 (s, 3H, CH_3-Ac), 1.95 (s, 3H, CH_3-Ac). ^{13}C-NMR (400 MHz, CD_3OD) δ 170.22, 170.17, 170.1, 157.0, 144.1, 143.9, 141.18, 141.12, 127.34, 127.32, 126.82, 126.79, 125.1, 124.9, 119.47, 119.45, 98.1, 71.0, 69.6, 69.5, 68.5, 66.8, 66.2, 60.5, 19.21, 19.19. IR

(NaCl, film): 1755, 1706, 1370, 1227, 1084, 1047, 760, 740 cm^{-1}. HRMS (ESI) Calcd. for $C_{30}H_{33}NNaO_{13}$ $[M + Na]^+$ requires 638.1845, Found: 638.1844.

To a solution of **66** (150 mg, 0.24 mmol) in THF (1.5 ml) were added N-methyl-morpholine (27 μl, 0.24 mmol, dissolved in 0.4 ml THF) and tert-Butyldimethylchlorosilane (TBDMSCl) (36 mg, 0.24 mmol, dissolved in 0.5 ml THF)[40]. After stirring for 30 minutes, 1H-tetrazole (2.5 ml, 1.13 mmol, 0.45M in CH_3CN) and dibenzyl N,N-diisopropylphosphoramidite [$iPr_2NP(OBn)_2$] (166 μl, 0.5 mmol) were added. The reaction mixture was stirred for 3 h at room temperature, cooled to 0 ^oC, and then 30% H_2O_2 (aq., 64 μl, 0.64 mmol) was added. The resulting mixture was slowly warmed to room temperature over 30 minutes, saturated Na_2SO_3 (1.5 ml) was then added. After stirring vigorously for 30 minutes, the mixture was diluted with saturated Na_2SO_3, extracted with EtOAc. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH =2:1:0.3→3:2:0.5) to give **67** (127 mg, 60%) as a white foam. ^{1}H-NMR (400 MHz, CD_3OD) δ 7.81 (d, *J* = 7.2 Hz, 2H, H-Fmoc), 7.69 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.30-7.42 (m, 14H, H-Fmoc, H-Ph), 5.33-5.38 (m, 2H, H-3, H-4), 5.26 (d, *J* = 2.0 Hz, 1H, H-2), 5.04-5.12 (m, 4H, CH_2-Bn), 4.90 (s, 1H, H-1), 4.40-4.48 (m, 2H, H-α, CH-Fmoc), 4.23-4.31 (m, 2H, CH_2-Fmoc), 4.05-4.17 (m, 4H, H-5, H-6, CH_2-β), 3.90 (dd, *J* = 10.4 Hz, 6.0 Hz, 1H, H-6), 2.01 (s, 3H, CH_3-Ac), 1.96 (s, 3H, CH_3-Ac), 1.95 (s, 3H, CH_3-Ac). ^{13}C-NMR (400 MHz, CD_3OD) δ 170.04, 170,00, 169.91, 144.0, 143.8, 141.2, 135.8, 128.32, 128.30, 128.28, 127.81, 127.75, 127.4, 126.83, 126.80, 125.0, 119.5, 98.3, 69.55, 69.52, 69.49, 69.45, 69.4, 69.2, 68.4, 66.8, 65.6, 65.3, 54.7, 19.21, 19.16, 19.15. ^{31}P-NMR (400 MHz, CD_3OD) δ -1.77. IR (NaCl, film): 3065, 3035, 2360, 2343, 1756, 1718, 1521, 1371, 1246, 1220, 1011, 882, 740, 698 cm^{-1}. HRMS (ESI) Calcd. for $C_{44}H_{46}NNaO_{16}P$ $[M + Na]^+$ requires 898.2447, Found: 898.2438.

67 (120 mg, 0.14 mmol) was dissolved in CH_3CN (1.5 ml). To this solution was added NaI (42 mg, 0.28 mmol)[41]. The reaction was stirred at 45 ^oC for 12 h under argon. The reaction mixture was concentrated and dissolved in small amount EtOAc. Hexanes was added until white solid formed. The suspension was centrifuged and the resulting solid was dissolved in H_2O/CH_3CN=1:1. The resulting solution was frozen and lyophilized to give **41** (103 mg, 94%) as a white solid. ^{1}H-NMR (400 MHz, CD_3OD) δ 7.80 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.69-7.72 (m, 2H, H-Fmoc), 7.22-7.41 (m, 9H, H-Fmoc, H-Ph), 5.26-5.36 (m, 3H, H-2, H-3, H-4), 4.91-4.95 (m, 2H, CH_2-Bn), 4.85 (d, *J* = 1.6 Hz, 1H, H-1), 4.41-4.46 (m, 1H, CH-Fmoc), 4.22-4.29 (m, 3H, H-α, CH_2-Fmoc), 3.83-4.13 (m, 5H, H-5, H-6, CH_2-β), 2.09 (s, 3H, CH_3-Ac), 1.95 (s, 3H, CH_3-Ac), 1.90 (s, 3H, CH_3-Ac). ^{13}C-NMR (400 MHz, CD_3OD) δ 170.3, 170.2, 170.1, 156.9, 144.1, 143.9, 141.17, 141.13, 138.5, 138.4, 128.2, 127.9, 127.3, 127.1, 126.9, 126.8, 125.1, 124.9, 119.5, 98.0, 69.8, 69.7, 69.5, 68.7, 66.79, 66.73, 66.1, 63.77, 63.72, 56.1, 19.30, 19.27, 19.24. ^{31}P-NMR (400 MHz, CD_3OD) δ 0.31. IR (NaCl, film): 3405, 3065, 2952, 1753, 1613, 1524, 1452, 1416, 1371, 1228, 1138, 1048, 868, 761, 740, 699 cm^{-1}. HRMS (ESI) Calcd. for $C_{37}H_{40}NNaO_{16}P$ $[M + Na]^+$ requires 808.1977, Found: 808.1980.



Synthesis of glycoamino acid **42** *and* **43**. To a solution of Tri-O-acetyl-D-glucal **68** (2.3 g, 8.43 mmol) in MeCN (38 mL) at -30 ^oC was added $FeCl_3 \cdot 6H_2O$ (2.50 g, 9.27 mmol), NaN_3 (602.5 mg, 9.27 mmol) and H_2O_2 (30%, aq., 1.26 ml, 12.65 mmol) and the reaction was stirred at -30 ^oC for 6 h[42]. The mixture was diluted with Et_2O and washed with H_2O, $NaHCO_3$ (sat.), and brine. The organic layer was dried with Na_2SO_4 and concentrated under reduced pressure to give **69** (2.06 g, 70%) as a viscous oil. The product was used directly without further purification.

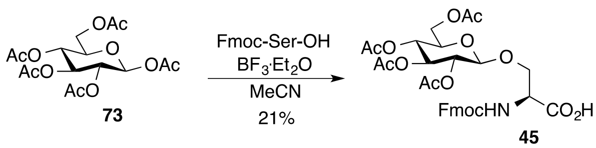
To a solution of Fmoc-Ser-OH (1.0 g, 3.05 mmol) in EtOAc (15.3 ml) was slowly added a solution of tert-Butyl 2,2,2-trichloroacetimidate (TBTA) (918 mg, 5.13 mmol) in cyclohexane (6.1 ml) over the course of 15 minutes with stirring at room temperature[43]. The mixture was allowed to stir at room temperature for 5 hours, then quenched with $NaHCO_3$ (sat., aq.) and extracted with EtOAc. The organic layers were combined, washed with H_2O, then brine, dried over Na_2SO_4, filtered and concentrated under reduced pressure. The residue was purified by flash chromatography on a silica gel column (Hex/EtOAc = 10:1→2:1) to yield Fmoc-Ser-Ot-Bu (643.6 mg, 55%) as a white solid.

A solution of Fmoc-Ser-Ot-Bu (2.03 g, 5.31 mmol), Ag_2CO_3 (2.44 g, 8.85 mmol), 4A MS (3.5 g) in toluene (12 ml) and DCM (18 ml) was stirred at 0 ^{o}C for 30 min[42]. Then, a solution of $AgClO_4$ (306.8 mg, 1.48 mmol) in toluene (6 mL) was added dropwise at 0 ^{o}C. Subsequently, a solution of **69** (2.06 g, 5.9 mmol) in a mixture of DCM (12 ml) and toluene (12 mL) was very slowly added dropwise at 0 ^{o}C. The mixture was stirred, in the dark, at room temperature for 24 h. The reaction was diluted with EtOAc, filtered through Celite and washed with H_2O then $NaHCO_3$ (sat.). The organic layer was dried with Na_2SO_4 and concentrated under reduced pressure. The residue was purified by flash chromatography on a silica gel column (PE/EtOAc = 5:1→2:1) to give **70** (2.86 g, 70%) as a white foam.

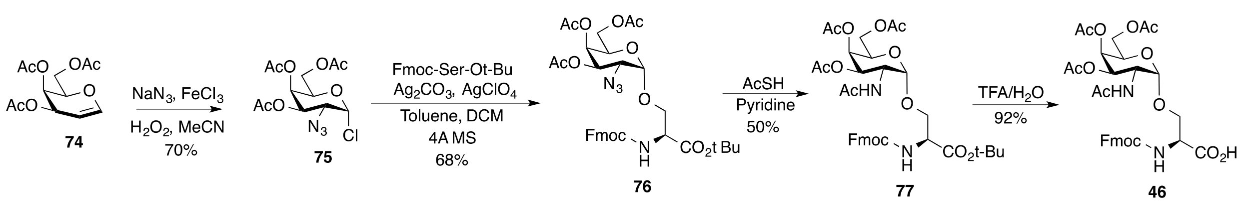
70 (2.6 g, 3.73 mmol) was dissolved in a solution of pyridine (4 ml) and AcSH (8 ml) and stirred at room temperature for 40 hours[44]. The mixture was diluted with EtOAc and washed with HCl (1M, aq.), and $NaHCO_3$ (sat., aq.). The organic layers were dried with Na_2SO_4 and concentrated under reduced pressure. The residue was purified by flash chromatography on a silica gel column (DCM/EtOAc = 3:2→1:1→0:1) to give **71** (780 mg, 29%) and **72** (830 mg, 31%) as white foams.

71 (780 mg, 1.09 mmol) was dissolved in a solution of TFA (4 ml), H_2O (0.2 mL) and DCM (4 ml) and stirred at room temperature for 3 h. The solvent was removed under reduced pressure by co-evaporation with toluene and the remaining residue was purified by flash chromatography on a silica gel column (DCM:MeOH = 15:1→10:1) to give **42** (680 mg, 94%) as a white foam.

72 (830 mg, 1.16 mmol) was dissolved in a solution of TFA (4 ml), H_2O (0.2 ml) and DCM (3 ml) and stirred at room temperature for 3 h. The solvent was removed under reduced pressure by co-evaporation with toluene and the remaining residue was purified by flash chromatography on a silica gel column (DCM:MeOH = 15:1→10:1) to give **43** (670 mg, 87%) as a white foam.



Synthesis of glycoamino acid **45.** To the solution of penta-O-acetyl-β-D-glucopyranose **73** (1.0 g, 2.56 mmol) and Fmoc-Ser-OH (1.0 g, 3.07 mmol) in MeCN (30 ml), $BF_3 \cdot OEt_2$ (1.0 mL, 7.68 mmol) was added. The resulting mixture was stirred at room temperature for 24 h under argon[37]. The solvent was removed under reduced pressure, and the resulting residue was diluted with EtOAc then washed with water. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH = 4:1:0.5→3:1:0.4→2:1:0.3) to give **45** (361 mg, 21%) as a white foam. Product matched the previously known spectra of **45**[45].

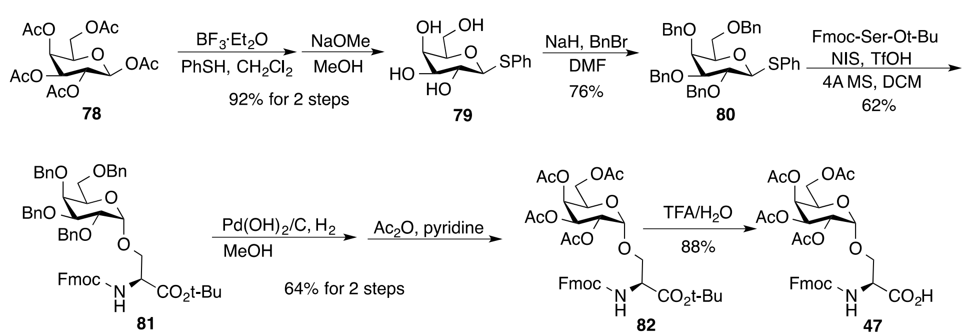


Synthesis of glycoamino acid **46.** To a solution of 3,4,6-Tri-O-acetyl-D-galactal **74** (1.0 g, 3.67 mmol) in MeCN (30 ml) at -30 °C was added $FeCl_3 \cdot 6H_2O$ (0.79 g, 2.94 mmol), NaN_3 (263 mg, 4.04 mmol) and H_2O_2 (30%, aq., 0.42 ml, 4.04 mmol) and the reaction was stirred at -30 °C for 6 h[42]. The mixture was diluted with Et_2O and washed with H_2O, $NaHCO_3$, and brine. The organic layer was dried with Na_2SO_4 and concentrated under reduced pressure to give **75** (900 mg, 70%) as a viscous oil. The product was used directly without further purification.

A solution of Fmoc-Ser-Ot-Bu (560 mg, 1.46 mmol), Ag_2CO_3 (600 mg, 2.19 mmol), 4A MS (1.7 g) in toluene (5 ml) and DCM (7.5 ml) was stirred at 0 ^oC for 30 minutes. Then, a solution of $AgClO_4$ (75 mg, 0.37 mmol) in toluene (2 ml) was added dropwise at 0^oC. Subsequently, a solution of **75** (900 mg, 2.58 mmol) in a mixture of DCM (3.75 ml) and toluene (3.75 ml) was very slowly added dropwise at 0 ^oC. The mixture was stirred, in the dark, at room temperature for 19 h. The reaction was diluted with DCM, filtered through Celite and washed with H_2O then $NaHCO_3$ (sat.). The organic layer was dried with Na_2SO_4 and concentrated under reduced pressure. The residue was purified by flash chromatography on a silica gel column (Tol/EtOAc = 10:1) to give **76** (692 mg, 68%) as a white foam.

76 (500 mg, 0.72 mmol) was dissolved in a solution of pyridine (0.8 ml) and AcSH (1.6 ml) and stirred at room temperature for 24 h. The mixture was diluted with DCM and washed with H_2O, HCl (1M, aq.), and $NaHCO_3$ (sat., aq.). The organic layers were dried with Na_2SO_4 and concentrated under reduced pressure. The residue was purified by flash chromatography on a silica gel column (Hex/EtOAc = 1:1→0:1) to give **77** (253 mg, 50%) as a white foam.

77 (240 mg, 0.34 mmol) was dissolved in a solution of TFA/H_2O (95:5, 4.0 ml) and stirred at room temperature for 2 h. The solvent was removed under reduced pressure by co-evaporation with toluene and the remaining residue was suspended in MeCN/H_2O (1:1), frozen and lyophilized to give a white solid. The solid was dissolved in DCM and purified by flash chromatography on a silica gel column (DCM:MeOH = 10:1→5:1) to give **46** (206 mg, 92%) as a white foam.



Synthesis of glycoamino acid **47.** To a solution of β-D-galactose pentaacetate **78** (20.0 g, 51.3 mmol) in CH_2Cl_2 (100 ml) was added thiophenol (7.3 mL, 72.0 mmol). The resulting mixture was cooled to 0 °C. $BF_3 \cdot Et_2O$ (7.7 ml, 61.5 mmol) was then added dropwise and the reaction mixture was allowed to warm to room temperature. After being stirred at room temperature for 2 h, the mixture was diluted with CH_2Cl_2 (100 ml), washed with 2 M NaOH solution, H_2O, dried over $MgSO_4$ and concentrated under reduced pressure. The residue was dissolved in MeOH (100 ml) and MeONa (138 mg, 2.56 mmol) was added to the solution. The reaction was stirred at room temperature overnight, then neutralized with Amberlite IR-120 resin, filtered and concentrated to give **79** (12.8 g, 92%) as a white foam[46].

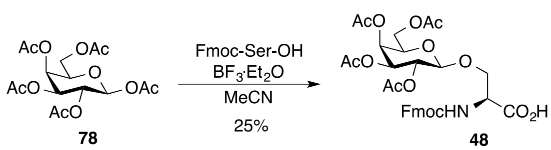
To a suspension of NaH (9.4 g, 235 mmol, 60% in mineral oil) in DMF (150 mL) at 0 °C was added dropwise a solution of **79** (12.8 g, 47.0 mmol) in DMF (70 ml), which was followed by the addition of a solution of BnBr (27.8 mL, 235 mmol) in DMF (80 mL). The resulting mixture was stirred at room temperature overnight, then diluted with EtOAc, washed with H_2O, and concentrated under reduced pressure. The oily residue was purified by flash chromatography on a silica gel column (Hex/EtOAc = 6:1) to afford **80** (22.6 g, 76%) as a white solid[46].

To a solution of **80** (632 mg, 1.00 mmol) and Fmoc-Ser-Ot-Bu (421 mg, 1.10 mmol) in DCM (15 ml) at -30 ^{o}C were added 4A MS (300 mg), N-iodosuccinimide (NIS) (470 mg, 2.00 mmol) and trifluoromethanesulfonic acid (TfOH) (9 µL, 0.10 mmol). The resulting mixture was stirred at -30 ^{o}C for 10 min under argon before it was quenched with Na_2SO_3 (sat., aq.). The mixture was diluted with water and extracted with EtOAc. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on silica gel (Hex/EtOAc = 5:1) to give **81** (561 mg, 62%) as a white foam. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.79 (d, *J* = 7.6 Hz, 2H), 7.63 (dd, *J* = 7.5, 4.0 Hz, 2H), 7.50 – 7.17 (m, 24H), 6.29 (d, *J* = 8.5 Hz, 1H), 4.99 (d, *J* = 11.3 Hz, 1H), 4.90 – 4.82 (m, $^1J_{CH}$ = 167.2 Hz, 3H), 4.79 (d, *J* = 11.7 Hz, 1H), 4.72 (d, *J* = 11.9 Hz, 1H), 4.62 (d, *J* = 11.4 Hz, 1H), 4.53 (d, *J* = 12.0 Hz, 1H), 4.49 – 4.40 (m, 2H), 4.40 – 4.33 (m, 2H), 4.29 – 4.19 (m, 2H), 4.11 (dd, *J* = 10.1, 3.7 Hz, 1H), 4.06 (t, *J* = 6.5 Hz, 1H), 4.01 (d, *J* = 2.8 Hz, 1H), 3.95 (dd, *J* =

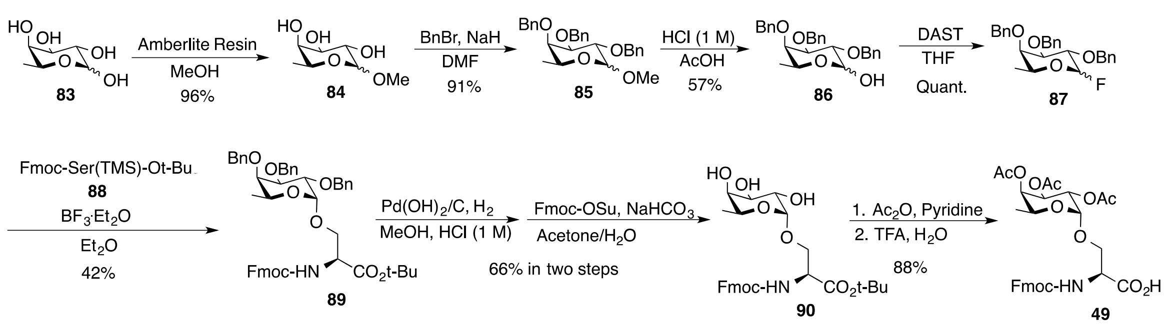
10.1, 2.7 Hz, 1H), 3.87 (dd, *J* = 11.2, 3.0 Hz, 1H), 3.64 (dd, *J* = 9.3, 6.1 Hz, 1H), 3.55 (dd, *J* = 9.2, 6.6 Hz, 1H), 1.51 (s, 9H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 169.15 , 156.14 , 143.97 , 143.95 , 141.28 , 138.71 , 138.59 , 138.57 , 137.98 , 128.42 , 128.39 , 128.37 , 128.29 , 127.96 , 127.76 , 127.71 , 127.67 , 127.64 , 127.58 , 127.51 , 127.10 , 125.27 , 119.95 , 99.52 , 82.36 , 78.74 , 74.96 , 74.80 , 73.48 , 73.41 , 73.09 , 70.73 , 70.02 , 68.93 , 67.06 , 55.24 , 47.16 , 28.04. IR (NaCl, film): 3340, 3064, 3031, 2978, 2929, 1725, 1497, 1453, 1369, 1347, 1248, 1155, 1100, 1057, 739, 697 cm^{-1}. HRMS (ESI) Calcd. for $C_{56}H_{59}NNaO_{10}$ $[M + Na]^+$ requires 928.4032, Found: 928.4026.

A solution of **81** (420 mg, 0.46 mmol) in MeOH (8 ml) was stirred with Pearlman's catalyst [$Pd(OH)_2$/C, 50 mg] under a hydrogen atmosphere at room temperature for 24 h. The reaction was filtered through Celite and the filtrate was concentrated under reduced pressure. The residue was dissolved in pyridine (2 ml) and Ac_2O (2 ml) was added dropwise. The resulting mixture was stirred at room temperature under argon overnight. The mixture was poured into ice-water and extracted with EtOAc. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on silica gel column (Hex/EtOAc = 2:1) to give **82** (212 mg, 64%) as a white foam. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.75 (d, *J* = 7.5 Hz, 2H), 7.62 (d, *J* = 6.8 Hz, 2H), 7.39 (t, *J* = 7.3 Hz, 2H), 7.35 – 7.28 (m, 2H), 5.83 (d, *J* = 8.0 Hz, 1H), 5.45 (d, *J* = 2.4 Hz, 1H), 5.30 (dd, *J* = 10.9, 3.2 Hz, 1H), 5.15 (dd, *J* = 10.9, 3.5 Hz, 1H), 5.04 (d, *J* = 3.5 Hz, 1H), 4.46 – 4.35 (m, 3H), 4.24 (t, *J* = 7.1 Hz, 1H), 4.19 (t, *J* = 6.6 Hz, 1H), 4.10 – 4.02 (m, 2H), 3.98 (dd, *J* = 10.4, 2.8 Hz, 1H), 3.92 (dd, *J* = 10.6, 2.9 Hz, 1H), 2.13 (s, 3H), 2.05 (s, 3H), 1.99 (s, 3H), 1.95 (s, 3H), 1.48 (s, 9H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 170.42 , 170.31 , 170.17 , 169.97 , 168.61 , 155.84 , 143.83 , 141.29 , 127.77 , 127.11 , 125.13 , 120.02 , 96.96 , 82.82 , 69.31 , 67.98 , 67.81 , 67.40 , 67.26 , 66.82 , 61.74 , 54.74 , 47.08 , 28.00 , 20.76 , 20.68 , 20.65 , 20.61. IR (NaCl, film): 3357, 3066, 2979, 1751, 1519, 1451, 1371, 1229, 1156, 1065, 761, 741 cm^{-1}. HRMS (ESI) Calcd. for $C_{36}H_{43}NNaO_{14}$ $[M + Na]^+$ requires 736.2576, Found: 736.2579.

Compound **82** (120 mg, 0.17 mmol) was dissolved in a TFA-water mixture (95:5, 1 ml) and stirred at room temperature for 2 h. The solvent was evaporated and the residue was co-evaporated with toluene to afford **47** (98 mg, 88%) as a white foam. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.76 (d, *J* = 7.5 Hz, 2H), 7.61 (t, *J* = 7.2 Hz, 2H), 7.40 (t, *J* = 7.5 Hz, 2H), 7.31 (t, *J* = 7.4 Hz, 2H), 6.50 (d, *J* = 8.7 Hz, 1H), 5.50 – 5.39 (m, 2H), 5.17 (d, $^1J_{CH}$ = 167.2 Hz, *J* = 3.5 Hz, 1H), 5.09 – 5.03 (m, 1H), 4.66 (d, *J* = 8.3 Hz, 1H), 4.39 (d, *J* = 7.1 Hz, 2H), 4.24 (d, *J* = 6.5 Hz, 2H), 4.12 – 4.01 (m, 2H), 4.01 – 3.91 (m, 2H), 2.14 (s, 3H), 2.03 (s, 3H), 2.01 (s, 3H), 1.91 (s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 172.46 , 170.79 , 170.64 , 170.35 , 156.10 , 143.82 , 143.69 , 141.29 , 127.81 , 127.08 , 125.12 , 125.05 , 120.05 , 96.73 , 77.26 , 69.28 , 68.25 , 67.94 , 67.89 , 67.38 , 66.75 , 61.89 , 54.16 , 47.03 , 20.80 , 20.64 , 20.54 , 20.52. IR (NaCl, film): 3336, 3067, 2954, 1750, 1527, 1451, 1372, 1229, 1153, 1063, 761, 741 cm^{-1}. HRMS (ESI) Calcd. for $C_{32}H_{36}NNaO_{14}$ $[M + Na]^+$ requires 658.2131, Found: 658.2144.



Synthesis of glycoamino acid **48.** To the solution of β-D-Galactose pentaacetate **78** (1.0 g, 2.56 mmol) and Fmoc-Ser-OH (1.0 g, 3.07 mmol) in MeCN (30 ml), $BF_3{\cdot}OEt_2$ (1.0 ml, 7.68 mmol) was added. The resulting mixture was stirred at room temperature for 24 h under argon. The solvent was removed under reduced pressure, and the resulting residue was diluted with EtOAc then washed with water. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH = 4:1:0.5→3:1:0.4→2:1:0.3) to give **48** (417.1 mg, 25%) as a white foam. Product matched the previously known spectra of **48**[45].



Synthesis of glycoamino acid **49.** To a solution of L-fucose **83** (3.3 g, 20.1 mmol) in MeOH (33 ml) was added Amberlite IR120 Resin (MeOH pre-treated H^+ form, 5.3 g). The mixture was heated to reflux and allowed to stir at reflux for 3 h. The reaction was then filtered and concentrated under reduced pressure. The residue was purified by recrystallization from EtOH to give **84** (3.41 g, 96%) as an off-white solid[47].

To a solution of **84** (3.41 g, 19.1 mmol) in DMF (85 ml) was slowly added NaH (60% in oil, 5.04 g, 126.06 mmol) over the course of 20 min with stirring. The mixture was allowed to stir for 1 h before BnBr (6.77 ml, 57.3 mmol) was added. The mixture was stirred at room temperature overnight. The reaction was quenched by the slow addition of H_2O and extracted with EtOAc. The organic layers were combined, washed with $NaHCO_3$ (sat.) and brine, dried over Na_2SO_4 and concentrated under reduced pressure. The residue was purified by flash chromatography on a silica gel column (Hex/EtOAc = 12:1→6:1) to give **85** (7.84 g, 91%) as an oil[47].

85 (7.81 g, 17.4 mmol) was dissolved in AcOH (112.75 mL) and heated to 95 ^{o}C with stirring. HCl (aq., 1 M, 31.3 mL) was added and the mixture was allowed to stir at 95 ^{o}C for 2 h. The reaction was cooled to room temperature and extracted with $CHCl_3$. The organic layers were combined, washed first with ice-cold $NaHCO_3$ (sat.) until neutral, then brine, dried over Na_2SO_4, filtered and concentrated under reduced pressure. The residue was purified by flash chromatography on a silica gel column (Hex/EtOAc = 4:1→3:1) to give **86** (4.3 g, 57%) as a thick syrup[47].

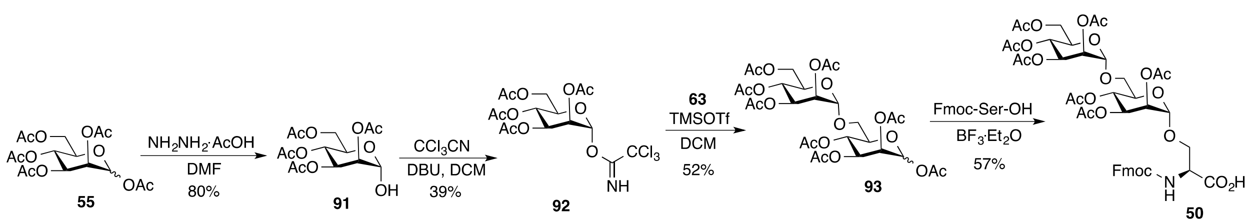
86 (1.17 g, 2.68 mmol) was dissolved in THF (30 ml) and cooled to -30 ^{o}C. (Diethylamino)sulfur trifluoride (DAST) (0.37 mL, 2.81 mmol) was added at -30 ^{o}C and stirred for 5 min. The reaction was quenched with H_2O (4 ml) at -30 ^{o}C and stirred another 5 min, after which the cooling bath was removed and the reaction was diluted with EtOAc. The organic layer was washed with brine, dried over Na_2SO_4, filtered and concentrated under reduced pressure to almost-dryness. The residue was quickly purified by flash chromatography on a short silica gel column (Hex/EtOAc = 6:1) to give **87** (1.2 g, quant.) as a white oil[48].

To a solution of Fmoc-Ser-Ot-Bu (643.6 mg, 1.68 mmol) in DMF (3.5 ml) was added imidazole (456.8 mg, 6.71 mmol) and trimethylchlorosilane (TMSCl) (0.426 ml, 3.36 mmol). The mixture was allowed to stir at room temperature for 5 h. The reaction was quenched with brine and extracted with EtOAc. The organic layers were combined, washed with brine, dried over Na_2SO_4, filtered and concentrated under reduced pressure to give **88** (712.3 mg, 93%) as a thick oil. The product was used directly without further purification.

This and subsequent steps of the synthesis of compound **49** are based on a previously published procedure[49]. **87** (670.5 mg, 1.53 mmol) and **88** (698.9 mg, 1.53 mmol) were dissolved in Et_2O (23 ml) and cooled to -20 ^{o}C. $BF_3{\cdot}OEt_2$ (57 µl, 0.46 mmol) was added to the reaction and the solution was allowed to stir at -20 ^{o}C for 1 h, after which the reaction was allowed to warm to 0 ^{o}C and stirred for another hour. The cooling bath was then removed and the reaction was stirred at room temperature for an additional 3 h. The reaction was quenched with $NaHCO_3$ (sat., aq.). The aqueous layer was extracted with Et_2O. The organic layers were then combined, washed with brine, dried over Na_2SO_4 and concentrated under reduced pressure. The residue was purified by flash chromatography on a silica gel column (Hex/EtOAc = 4:1) to give **89** (511.1 mg, 42%) as a hard white foam.

A solution of **89** (479.4 mg, 0.6 mmol) in EtOH (17 ml) and HCl (aq., 1M, 0.5 ml) was stirred with Pearlman's catalyst [$Pd(OH)_2$/C, 120 mg] under a hydrogen atmosphere at room temperature for 3 h. The reaction was filtered through Celite and the filtrate was neutralized with $NaHCO_3$ (sat., aq.) then concentrated under reduced pressure. The resulting residue was dissolved in H_2O (20 ml) and $NaHCO_3$ (200 mg), acetone (40 ml) and Fmoc-OSu (202 mg, 0.6 mmol) were added with vigorous stirring. The reaction was allowed to stir at room temperature for 1 h. The reaction was concentrated under reduced pressure, diluted with H_2O and extracted with $CHCl_3$. The organic layers were combined, dried with Na_2SO_4 and concentrated under reduced pressure. The residue was purified by flash chromatography on a silica gel column (DCM/MeOH = 50:1→25:2) to give **90** (210.6 mg, 66% in two steps from **89**) as a hard, white foam.

90 (201.4 mg, 0.38 mmol) was dissolved in a mixture of Ac_2O (1 ml) and pyridine (0.8 ml) and allowed to stir at room temperature for 17 h. The reaction was concentrated under reduced pressure and the residue was co-evaporated with toluene three times before being dissolved in a solution of TFA and H_2O (95:5, 1 ml) and stirred for an additional hour at room temperature. Solvent was removed by co-evaporating the reaction mixture with toluene under reduced pressure and the resulting residue was purified by flash chromatography on a silica gel column ($CHCl_3$/MeOH = 40:1→20:1) to give an oil, which was then dissolved in H_2O/CH_3CN=1:1. The solution was frozen and lyophilized to give **49** (184.2 mg, 88% over two steps from **90**) as a white solid. Product matched the previously known spectra of **49**[50].



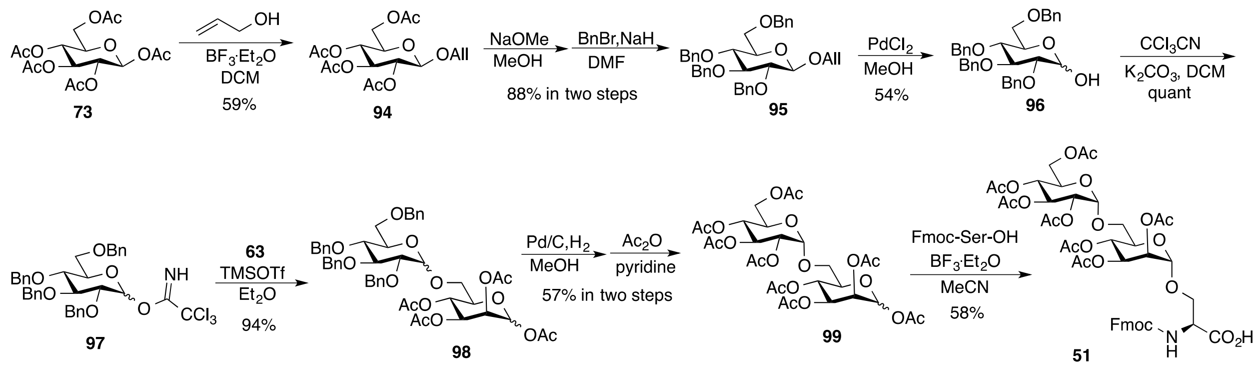
Synthesis of glycoamino acid **50.** To a stirred solution of **55** (6.6 g, 16.92 mmol) in DMF (66 ml) was added $N_2H_4 \cdot$AcOH (1.87 g, 20.30 mmol). The resulting mixture was stirred for 4 h at room temperature under Ar. The reaction mixture was diluted with EtOAc, washed by water and brine. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc =1:1) to give **91** (4.69 g, 80%) as a syrup[51].

To a stirred solution of **94** (4.6 g, 13.21 mmol) in DCM (60 ml), CCl_3CN (13.21 ml, 132.10 mmol) and DBU (3.95 ml, 24.43 mmol) were added. After stirring overnight, the solvent was evaporated under reduced pressure and the residue was purified by flash chromatography on a silica gel column (Hex/EtOAc =1:1) to give **92** (2.54 g, 39%) as a syrup[51].

A solution of **63** (500 mg, 1.43 mmol) and **92** (776 mg, 1.58 mmol) in DCM (20 ml) was stirred with 4A molecular sieves (450 mg) under argon for 15 min. Then, a solution of Trimethylsilyl triflate (TMSOTf)

(113 μl, 0.40 mmol) in DCM (2 ml) was added dropwise. The resulting mixture was stirred for 4 h at room temperature under argon. The reaction mixture was diluted with DCM and washed with sat. aq. $NaHCO_3$. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc =2:1→3:2→1:1) to give **93** (504 mg, 52%) as a white foam.

To the solution of **93** (500 mg, 0.73 mmol) and Fmoc-Ser-OH (363 mg, 1.11 mmol) in MeCN (18 ml), $BF_3{\cdot}OEt_2$ (0.34 ml, 2.22 mmol) was added. The resulting mixture was stirred at room temperature for 24 h under argon. The solvent was removed under reduced pressure, and the resulting residue was diluted with EtOAc then washed with water. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH = 2:1:0.3→3:2:0.5→1:1:0.2) to give **50** (390 mg, 57%) as a white foam. ^{1}H-NMR (400 MHz, Acetone-d6) δ 7.86 (d, J = 7.6 Hz, 2H, H-Fmoc), 7.74 (d, J = 7.6 Hz, 2H, H-Fmoc), 7.42 (t, J = 7.2 Hz, 2H, H-Fmoc), 7.34 (t, J = 7.4 Hz, 2H, H-Fmoc), 5.22-5.37 (m, 6H, H-2, H-3, H-4, H-2’, H-3’, H-4’), 4.96 (s, $^1J_{CH}$ = 176 Hz, H-1), 4.94 (s, $^1J_{CH}$ = 172 Hz, H-1’), 4.03-4.55 (m, 10H, H-α, CH_2-β, CH-Fmoc, CH_2-Fmoc, H-5, H-5’, H-6’), 3.69-3.83 (m, 2H, H-6), 2.13 (s, 3H, CH_3-Ac), 2.10 (s, 3H, CH_3-Ac), 2.02 (s, 6H, CH_3-Ac), 2.01 (s, 3H, CH_3-Ac), 1.953 (s, 3H, CH_3-Ac), 1.947 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, Acetone-d6) δ 169.9, 169.5, 169.4, 169.3, 144.3, 144.2, 141.18, 141.15, 127.6, 127.1, 125.4, 125.3, 119.9, 98.2, 97.5, 69.4, 69.2, 68.6, 66.6, 65.9, 62.3, 47.1, 19.86, 19.83, 19.80, 19.76. IR (NaCl, film): 3361, 2954, 1751, 1371, 1225, 1138, 1087, 1047, 761, 741 cm^{-1}. HRMS (ESI) Calcd. for $C_{44}H_{51}NNaO_{22}$ $[M + Na]^+$ requires 968.2795, Found: 968.2784.



Synthesis of glycoamino acid **51. 73** (5 g, 12.80 mmol) was dissolved in DCM (30 ml) under argon. The stirred solution was cooled to 0 °C and $BF_3 \cdot Et_2O$ (2.44 ml, 19.22 mmol) was added by syringe. After stirring for 10 min at 0 °C, allyl alcohol (1.31 ml, 19.22 mmol) was added. The ice bath was removed after completion of the addition and the reaction stirred at room temperature for overnight. The reaction was then cooled to 0 °C and quenched with $NaHCO_3$ (sat. aq.). After dilution with water, the organic layer was separated and aqueous layer was extracted with DCM. The combined organic layer was washed with brine, dried over Na_2SO_4, filtered and concentrated under reduced pressure. The product was purified by flash chromatography on a silica gel column (Hex/EtOAc =4:1→3:1→2:1) to give **94** (2.92 g, 59%) as a white solid[52].

To a stirred solution of **94** (2.9 g, 7.47 mmol) in MeOH (30 ml) was added NaOMe (20.5 mg, 0.38 mmol). After stirring for 1 h, the reaction was neutralized with Dowex H^+ and then filtered. The solvent was concentrated under reduced pressure to give a white foam (1.68g, 7.47 mmol). The product was dissolved in DMF (5 ml). The resulting solution was used directly in the next step without purification. To a suspension of NaH (60% in oil, 1.52 g, 38 mmol) in DMF (25 ml) was added dropwise the above solution at 0 °C under argon. The resulting mixture was stirred at 0 °C for 30 min and then BnBr (4.5 ml, 38 mmol) was added dropwise. The reaction was allowed to warm up to room temperature and stirred overnight. With caution, the reaction was quenched with water at 0°C. The resulting mixture was diluted with EtOAc and washed with water then brine. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc =20:1→15:1→10:1) to give **95** (3.8 g, 88% over two steps) as a white solid[53].

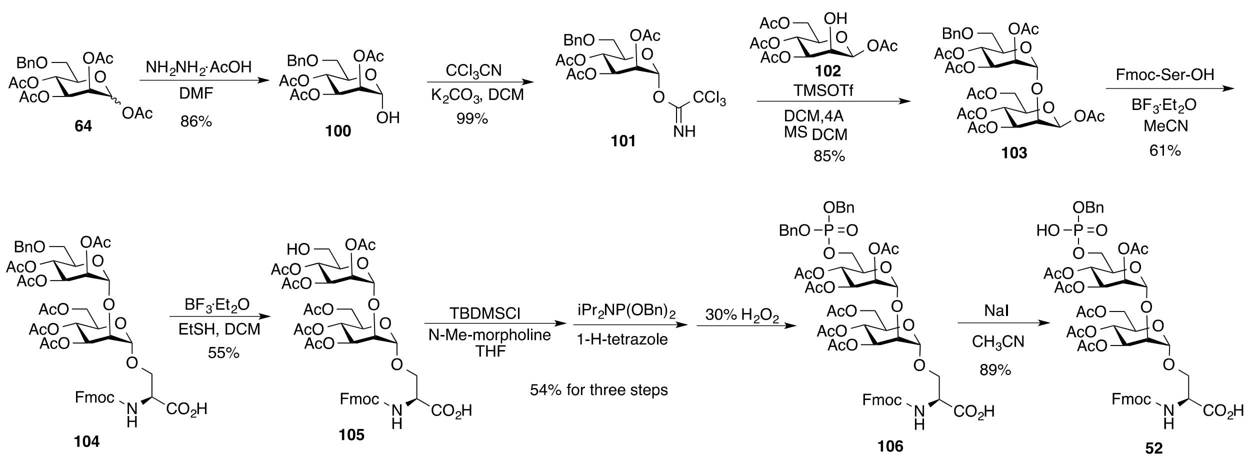
A solution of **95** (1.8 g, 3.10 mmol) and $PdCl_2$ (109 mg, 0.62 mmol) in MeOH (20 ml) was stirred vigorously at room temperature overnight. The reaction was diluted with diethyl ether and filtered through Celite. The solvent was concentrated under reduced pressure and the residue was purified by flash chromatography on a silica gel column (Hex/EtOAc =4:1→2:1) to give **96** (900 mg, 54%) as a white solid.[53]

A solution of **96** (900 mg, 1.66 mmol), CCl_3CN (1.72 ml, 21.6 mmol) and K_2CO_3 (1.15 g, 8.3 mmol) in DCM (22 ml) was stirred vigorously at room temperature under argon overnight. The reaction was filtered through Celite and the filtrate was concentrated under reduced pressure to give **97** (1.21 g, 100%) as a white foam. The product was used directly in the next step without purification.

A solution of **63** (293 mg, 0.84 mmol) and **97** (690 mg, 1.01 mmol) in diethyl ether (20 ml) was stirred with 4A MS (700 mg) under argon for 1 h. The reaction was cooled to -40 ^{o}C and TMSOTf (76 μl, 0.42 mmol) was added. The resulting mixture was allowed to warm up to room temperature slowly and stirred for 4 h. The reaction was quenched with Et_3N (500 μl) and stirred for 10 additional minutes, then filtered. The filtrate was concentrated under reduced pressure and the residue was purified by flash chromatography on a silica gel column (Hex/EtOAc =5:1→2:1→1:1) to give **98** (688 mg, 94%) as a white foam.

A solution of **98** (680 mg, 0.78 mmol) in MeOH (25 ml) was stirred at room temperature under a hydrogen atmosphere for 24 h in the presence of 10% Pd/C (200 mg). The reaction was filtered through Celite and the filtrate was concentrated under reduced pressure. The residue was dissolved in pyridine (3 ml) and Ac_2O (1 ml) was added dropwise. The resulting mixture was stirred at room temperature under argon overnight. The mixture was slowly poured into ice-water and extracted with DCM. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc =2:1→3:2) to give first **99** (304 mg, 57%). ^{1}H-NMR (400 MHz, $CDCl_3$) δ 6.06 (d, *J* = 2.0 Hz, 1H, H-1), 5.47 (dd, *J* = 10.0 Hz, 9.2 Hz, 1H, H-3'), 5.38 (dd, *J* = 10.0 Hz, 3.2 Hz, H-3), 5.31 (t, *J* = 10.0 Hz, 1H, H-4), 5.27 (dd, *J* = 3.6 Hz, 2.0 Hz, 1H, H-2), 5.07-5.12 (m, 2H, H-1', H4'), 4.86 (dd, *J* = 10.0 Hz, 3.6 Hz, 1H, H-2'), 4.26-4.30 (m, 1H, H-6'), 4.09-4.13 (m, 2H, H-5', H-6'), 4.03-4.07 (m, 1H, H-5), 3.79 (dd, *J* = 11.2 Hz, 6.0 Hz, H-6), 3.62 (dd, *J* = 10.8 Hz, 2.8 Hz, H-6), 2.21 (s, 3H, CH_3-Ac), 2.20 (s, 3H, CH_3-Ac), 2.11 (s, 3H, CH_3-Ac), 2.102, (s, 3H, CH_3-Ac), 2.099 (s, 3H, CH_3-Ac), 2.07 (s, 3H, CH_3-Ac), 2.04 (s, 3H, CH_3-Ac), 2.03 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 170.7, 170.2, 169.98, 169.97, 169.8, 169.7, 169.6, 168.1, 95.3, 90.3, 71.1, 70.9, 69.9, 68.8, 68.30, 68.27, 67.3, 67.1, 66.3, 61.8, 20.84, 20.75, 20.72, 20.69, 20.66, 20.62. IR (NaCl, film): 1751, 1370, 1223, 1149, 1040, 977 cm^{-1}. HRMS (ESI) Calcd. for $C_{28}H_{38}NaO_{19}$ $[M + Na]^+$ requires 701.1900, Found: 701.1902.

To the solution of **99** (400 mg, 0.59 mmol) and Fmoc-Ser-OH (290 mg, 0.88 mmol) in MeCN (15 mL), $BF3{\cdot}OEt_2$ (0.27 ml, 1.77 mmol) was added. The resulting mixture was stirred at room temperature for 24h under argon. The solvent was removed under reduced pressure. The resulting residue was diluted with EtOAc, and then washed with water. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH =2:1:0.3→3:2:0.5→1:1:0.2) to give **51** (321 mg, 58%) as a white foam. ^{1}H-NMR (400 MHz, $CDCl_3$) δ 7.76 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.63 (d, *J* = 6.4 Hz, 2H, H-Fmoc), 7.40 (t, *J* = 7.4 Hz, 2H, H-Fmoc), 7.31 (t, *J* = 7.6 Hz, 2H, H-Fmoc), 5.92 (d, *J* = 8.4 Hz, 1H, H-NH), 5.44 (t, *J* = 9.8 Hz, 1H, H-3'), 5.34 (dd, *J* = 10.0 Hz, 3.6 Hz, 1H, H-3), 5.27-5.29 (m, 1H, H-2), 5.18 (t, *J* = 10.2 Hz, 1H, H-4), 5.10 (d, *J* = 4.0 Hz, 1H, H-1'), 5.03 (t, *J* = 9.4 Hz, 1H, H-4'), 4.86 (dd, *J* = 10.0 Hz, 3.6 Hz, 1H, H-2'), 4.83 (s, 1H, H-1), 4.67-4.72 (m, 1H, H-α), 4.42 (d, *J* = 7.6 Hz, 2H, CH_2-Fmoc), 4.25 (t, *J* = 7.0 Hz, 1H, CH-Fmoc), 4.07-4.15 (m, 5H, H-β, H-5, H-5', H-6'), 3.90-3.94 (m, 1H, H-β), 3.79 (dd, J = 10.8 Hz, 6.8 Hz, 1H, H-6), 3.53 (dd, J = 10.8 Hz, 2.8 Hz), 2.14 (s, 3H, CH_3-Ac), 2.08 (s, 3H, CH_3-Ac), 2.06 (s, 3H, CH_3-Ac), 2.05 (s, 3H, CH_3-Ac), 2.01 (s, 3H, CH_3-Ac), 2.00 (s, 3H, CH_3-Ac), 1.99 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 170.77, 170.75, 170.6, 170.3, 170.0, 169.9, 167.7, 156.1, 143.81, 143.77, 141.3, 127.7, 127.10, 127.08, 125.2, 120.0, 98.0, 95.4, 77.2, 70.7, 70.4, 69.9, 69.2, 69.0, 68.5, 68.4, 67.4, 67.3, 67.1, 66.6, 61.9, 54.1, 47.1, 20.80, 20.79, 20.77, 20.74, 20.70, 20.64, 20.58. IR (NaCl, film): 3355, 3065, 2952, 1751, 1521, 1452, 1370, 1223, 1139, 1042, 763, 741 cm^{-1}. HRMS (ESI) Calcd. for $C_{44}H_{51}NNaO_{22}$ $[M + Na]^+$ requires 968.2795, Found: 968.2793.



Synthesis of glycoamino acid **52.** To a solution of **64** (4 g, 9.13 mmol) in DMF (37 ml) was added N_2H_4·AcOH (1.0 g, 10.96 mmol). The resulting mixture was stirred for 4 h at room temperature under argon. The reaction was diluted with EtOAc and washed with water and brine. The organic layer was dried over anhydrous Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc =2:1→3:2) to give **100** (3.11 g, 86%) as a white foam. ^{1}H-NMR (400 MHz, $CDCl_3$) δ 7.27-7.36 (m, 5H, H-Ph), 5.40 (dd, J = 10.0 Hz, 3.2 Hz, 1H, H-3), 5.19-5.26 (m, 3H, H-1, H-2, H-4), 4.54 (dd, J = 25.2 Hz, 12.0 Hz, 2H, CH_2-Bn), 4.17-4.22 (m, 1H, H-5), 3.57 (dd, J = 10.8 Hz, 6.4 Hz, 1H, H-6), 3.49 (dd, J = 10.4 Hz, 2.4 Hz, 1H, H-6), 3.27 (brs, 1H, H-OH), 2.14 (s, 3H, CH_3-Ac), 1.99 (s, 3H, CH_3-Ac), 1.93 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 170.2, 170.0, 169.9, 137.5, 128.4, 128.1, 127.8, 92.1, 73.6, 70.0, 69.6, 69.1, 68.8, 66.9, 20.9, 20.72, 20.69. IR (NaCl, film): 3423, 2937, 2871, 1751, 1454, 1433, 1372, 1227, 1078, 1051, 739, 701 cm^{-1}. HRMS (ESI) Calcd. for $C_{19}H_{24}NaO_9$ $[M + Na]^+$ requires 419.1313, Found: 419.1320.

100 (3.1 g, 7.83 mmol) in DCM (100 ml) was stirred vigorously with CCl_3CN (8.10 ml, 101.73 mmol) and K_2CO_3 (5.40 g, 39.13 mmol) at room temperature under argon overnight. The reaction mixture was filtered through celite and the filtrate was concentrated under reduced pressure to give **101** (4.39 g, 99%) as a syrup. The product was used directly to the next step without further purification. ^{1}H-NMR (400 MHz, $CDCl_3$) δ 8.75 (s, 1H, H-NH), 7.27-7.35 (m, 5H, H-Ph), 6.29 (d, *J* = 2.0 Hz, 1H, H-1), 5.38-5.50 (m, 3H, H-2, H-3, H-4), 4.54 (dd, *J* = 40.4 Hz, 12.0 Hz, 2H, CH_2-Bn), 4.14-4.18 (m, 1H, H-5), 3.60 (d, *J* = 4.0 Hz, 2H, H-6), 2.18 (s, 3H, CH_3-Ac), 2.00 (s, 3H, CH_3-Ac), 1.93 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 169.9, 169.8, 169.6, 137.7, 128.3, 127.9, 127.7, 94.7, 73.5, 72.5, 69.0, 68.5, 67.9, 66.1, 20.8, 20.69, 20.65. IR (NaCl, film): 3323, 2917, 2869, 1751, 1678, 1454, 1432, 1369, 1246, 1159, 1089, 1050, 976, 944, 836, 798, 737, 701 cm^{-1}. HRMS (ESI) Calcd. for $C_{21}H_{24}Cl_3NNaO_9$ $[M + Na]^+$ requires 562.0409, Found: 562.0406.

The 1,3,4,6-Tetra-O-acetyl-β-D-mannopyranose **102** was prepared as reported in the literature[37]. A solution of **102** (1.46 g, 4.20 mmol) and **101** (3 g, 5.58 mmol) in DCM (45 ml) was stirred with 4A MS (1.5 g) under argon for 30 min. The reaction was cooled to -30 ^oC and a solution of TMSOTf (234 μl, 1.26 mmol) in DCM (2 ml) was added dropwise. After stirring at -30 ^oC for 15 min, the reaction mixture was allowed to warm up to room temperature slowly and stirred for 5 h. The reaction was quenched with Et_3N (3 ml), stirred for an additional 10 minutes and then filtered. The filtrate was concentrated under reduced pressure and the residue was purified by flash chromatography on a silica gel column (Hex/EtOAc =3:2→1:1) to give **103** (2.6 g, 85%) as a white foam. ^{1}H-NMR (400 MHz, $CDCl_3$) δ 7.27-7.35 (m, 5H, H-Ph), 5.78 (d, *J* = 1.2 Hz, 1H, H-1), 4.99 (dd, *J* = 10.0 Hz, 3.2 Hz, 1H, H-3'), 5.29-5.41 (m, 3H, H-2', H-4, H-4'), 5.12 (dd, *J* = 9.6 Hz, 2.8 Hz, 1H, H-3), 5.00 (d, *J* = 2.0 Hz, 1H, H-1'), 4.50 (dd, *J* = 34.0 Hz, 8.0 Hz, 2H, CH_2-Bn), 4.36-4.41 (m, 1H, H-5'), 4.26 (dd, *J* = 12.4 Hz, 4.8 Hz, 1H, H-6), 4.16-4.20 (m, 2H, H-2, H-6), 3.76-3.80 (m, 1H, H-5), 3.55 (d, *J* = 4.0 Hz, 2H, H-6'), 2.14 (s, 3H, CH_3-Ac), 2.12 (s, 3H, CH_3-Ac), 2.10 (s, 3H, CH_3-Ac), 2.06 (s, 3H, CH_3-Ac), 2.03 (s, 3H, CH_3-Ac), 2.02 (s, 3H, CH_3-Ac), 1.92 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 171.0, 170.3, 169.92, 169.85, 169.6, 169.3, 168.7, 137.6, 128.4, 127.9, 127.8, 98.4, 90.9, 74.6, 73.7, 73.1, 72.2, 70.2, 70.0, 69.3, 68.4, 67.1, 65.8, 61.8, 21.0, 20.77, 20.74, 20.70, 20.6, 20.5. IR (NaCl, film): 3064, 2939, 2869, 1751, 1454, 1433, 1370, 1221, 1055, 933, 737, 702 cm^{-1}. HRMS (ESI) Calcd. for $C_{33}H_{42}NaO_{18}$ $[M + Na]^+$ requires 749.2264, Found: 749.2261.

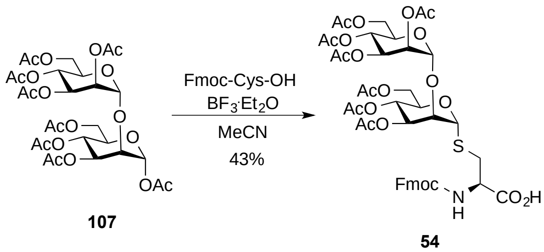
To the solution of **103** (2.5 g, 3.44 mmol) and Fmoc-Ser-OH (1.69 g, 5.16 mmol) in MeCN (80 mL), $BF_3 \cdot OEt_2$ (1.6 ml, 10.33 mmol) was added. The resulting mixture was stirred at room temperature for 24 h under argon. The solvent was removed under reduced pressure; the resulting residue was diluted with EtOAc and washed with water. The organic layer was dried (Na_2SO_4), filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH =2:1:0.3→3:2:0.5) to give **104** (2.10 g, 61%) as a white foam. ^{1}H-NMR (400 MHz, $CDCl_3$) δ 7.77 (d, J = 7.6 Hz, 2H, H-Fmoc), 7.64 (d, J = 7.2 Hz, 2H, H-Fmoc), 7.41 (t, J = 7.4 Hz, 2H, H-Fmoc), 7.30-7.36 (m, 7H, H-Fmoc, H-Ph), 5.81 (d, J = 8.0 Hz, 1H, H-NH), 5.22-5.40 (m, 4H, H-2', H-3, H-3', H-4'), 5.09 (t, J = 10.0 Hz, 1H, H-4), 5.02 (d, J = 0.8 Hz, 1H, H-1), 4.98 (d, J = 1.6 Hz, 1H, H-1'), 4.65 (dd, J = 56.4 Hz, 12.4 Hz, 2H, CH_2-Bn), 4.50-4.52 (m, 1H, H-α), 3.90 (d, J = 6.8 Hz, 2H, CH_2-Fmoc), 3.96-4.27 (m, 8H, H-2, H-5, H5', H-6, CH-Fmoc, CH_2-β), 3.66 (dd, J = 10.4 Hz, 7.2 Hz, 1H, H-6'), 3.56 (dd, J = 10.4 Hz, 2.8 Hz, 1H, H-6'), 2.12 (s, 6H, CH_3-Ac), 2.09 (s, 3H, CH_3-Ac), 2.02 (s, 3H, CH_3-Ac), 2.00 (s, 3H, CH_3-Ac), 1.93 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 171.0, 170.7, 170.4, 170.0, 169.8, 169.4, 155.9, 143.79, 143.76, 141.3, 136.0, 128.6, 128.5, 128.4, 127.8, 127.1, 125.20, 125.17, 120.0, 100.2, 98.2, 74.5, 73.8, 70.23, 70.17, 70.15, 69.7, 69.6, 69.0, 68.0, 67.4, 67.3, 66.7, 62.1, 54.7, 47.1, 20.9, 20.69, 20.67, 20.65. IR (NaCl, film): 3338, 3065, 2952, 1751, 1521, 1452, 1370, 1225, 1136, 1047, 760 cm^{-1}. HRMS (ESI) Calcd. for $C_{49}H_{55}NNaO_{21}$ $[M + Na]^+$ requires 1016.3159, Found: 1016.3148.[37]

The mixture of **104** (1.9 g, 1.91 mmol), EtSH (7.5 ml) and $BF_3 \cdot OEt_2$ (2.42 ml, 15.30 mmol) in DCM (15 ml) was stirred for 6 h at room temperature under argon. The reaction was quenched with water and extracted with EtOAc. The organic phase was washed with brine, dried (Na_2SO_4), filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH =1:1:0.2→2:3:0.5) to give **105** (945 mg, 55%) as a white foam. ^{1}H-NMR (400 MHz, CD_3OD) δ 7.82 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.71-7.74 (m, 2H, H-Fmoc), 7.39-7.43 (m, 2H, H-Fmoc), 7.32-7.36 (m, 2H, H-Fmoc), 5.42-5.41 (m, 5H, H-2', H-3, H-3', H-4, H-4'), 5.15 (d, *J* = 2.0 Hz, 1H, H-1), 5.00 (d, *J* = 1.6 Hz, 1H, H-1'), 4.47 (dd, *J* = 10.0 Hz, 6.4 Hz, 2H, CH_2-Fmoc), 3.93-4.34 (m, 9H, H-2, H-5, H-5', H-6, H-α, CH_2-β, CH-Fmoc), 3.66 (dd, *J* = 12.4 Hz, 2.4 Hz, 1H, H-6'), 3.59 (dd, *J* = 12.0 Hz, 5.6 Hz, 1H, H-6'), 2.14 (s, 3H, CH_3-Ac), 2.11 (s, 3H, CH_3-Ac), 2.052 (s, 3H, CH_3-Ac), 2.046 (s, 3H, CH_3-Ac), 2.00 (s, 3H, CH_3-Ac), 1.97 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, CD_3OD) δ 171.3, 170.5, 170.22, 170.20, 170.1, 170.0, 157.0, 144.0, 143.8, 141.19, 141.15, 127.41, 127.38, 126.84, 126.80, 125.0, 124.8, 119.53, 119.51, 99.2, 99.0, 76.9, 71.6, 70.3, 69.7, 69.0, 68.8, 68.4, 66.8, 66.4, 66.1, 54.8, 19.4, 19.29, 19.27, 19.23, 19.18. IR (NaCl, film): 3350, 3065, 1751, 1371, 1229, 1046, 740 cm^{-1}. HRMS (ESI) Calcd. for $C_{42}H_{49}NNaO_{21}$ $[M + Na]^+$ requires 926.2690, Found: 926.2693.

To a solution of **105** (850 mg, 0.94 mmol) in THF (5.875 ml) were added N-methyl-morpholine (106 μl, 0.94 mmol, dissolved in 1.57 ml THF) and TBDMSCl (141 mg, 0.94 mmol, dissolved in 1.96 ml THF). After stirring for 30 minutes, 1H-tetrazole (9.79 ml, 4.42 mmol, 0.45M in CH_3CN) and $iPr_2N(OBn)_2$ (650 μl, 1.98 mmol) were added. The reaction mixture was stirred for 3 h at room temperature, cooled to 0 ^oC, and then 30% H_2O_2 (aq., 250 μl, 2.49 mmol) was added. The resulting mixture was slowly warmed to room temperature over 30 min, saturated Na_2SO_3 (6 ml) was then added. After stirring vigorously for 30 min, the mixture was diluted with saturated Na_2SO_3, extracted with EtOAc. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH =1:1:0.2) to give **106** (586 mg, 54%) as a white foam. ^{1}H-NMR (400 MHz, CD_3OD) δ 7.81 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.69-7.72

(m, 2H, H-Fmoc), 7.31-7.42 (m, 14H, H-Fmoc, H-Ph), 5.27-5.41 (m, 5H, H-2’, H-3, H-3’, H-4, H-4’), 5.04-5.13 (m, 5H, H-1, CH_2-Bn), 4.94 (d, *J* = 1.2 Hz, 1H, H-1’), 4.04-4.45 (m, 12H, H-2, H-5, H-5’, H-6, H-6’, H-α, CH_2-β, CH_2-Fmoc, CH-Fmoc), 3.92 (dd, *J* = 10.4 Hz, 6.0 Hz, 1H, H-6), 2.09 (s, 3H, CH_3-Ac), 2.04 (s, 6H, CH_3-Ac), 2.00 (s, 6H, CH_3-Ac), 1.97 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, CD_3OD) δ 171.2, 170.4, 170.03, 170.99, 169.91, 169.86, 158.3, 157.0, 144.0, 143.8, 141.19, 141.14, 135.74, 135.67, 128.39, 128.36, 128.33, 128.31, 127.9, 127.8, 127.39, 127.37, 126.84, 126.80, 125.0, 124.9, 119.5, 99.1, 98.8, 77.1, 70.2, 69.57, 69.55, 69.51, 69.50, 69.45, 68.8, 68.4, 66.8, 65.93, 65.85, 65.80, 65.6, 61.7, 19.4, 19.28, 19.25, 19.20, 19.19, 19.13. ^{31}P-NMR (400 MHz, CD_3OD) δ -1.51. IR (NaCl, film): 2956, 1750, 1452, 1370, 1225, 1046, 740 cm^{-1}. HRMS (ESI) Calcd. for $C_{56}H_{62}NNaO_{24}P$ $[M + Na]^+$ requires 1186.3292, Found: 1186.3287[40].

106 (480 mg, 0.41 mmol) was dissolved in CH_3CN (4.4 ml). To this solution was added NaI (124 mg, 0.82 mmol). The reaction was stirred at 45 ^oC for 12 h under argon. The reaction mixture was concentrated and dissolved in small amount EtOAc. Hexane was added until white solid formed. The suspension was centrifuged and the resulting solid was dissolved in H_2O/CH_3CN=1:1. The resulting solution was frozen and lyophilized to give **52** (392 mg, 89%) as a white solid. ^{1}H-NMR (400 MHz, CD_3OD) δ 7.81 (d, *J* = 7.6 Hz, 2H, H-Fmoc), 7.71-7.74 (m, 2H, H-Fmoc), 7.22-7.42 (m, 9H, H-Fmoc, H-Ph), 5.24-5.40 (m, 5H, H-2’, H-3, H-3’, H-4, H-4’), 5.14 (d, *J* = 1.6 Hz, 1H, H-1), 4.93-5.00 (m, 2H, CH_2-Bn), 4.86 (d, *J* = 2.0 Hz, 1H, H-1’), 3.93-4.47 (m, 13H, H-2, H-5, H-5’, H-6, H-6’, H-α, CH_2-β, CH_2-Fmoc, CH-Fmoc), 2.13 (s, 3H, CH_3-Ac), 2.10 (s, 3H, CH_3-Ac), 2.02 (s, 3H, CH_3-Ac), 1.99 (s, 3H, CH_3-Ac), 1.98 (s, 3H, CH_3-Ac), 1.97 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, CD_3OD) δ 171.3, 170.3, 170.2, 170.1, 170.0, 143.9, 141.2, 127.9, 127.3, 127.0, 126.9, 126.8, 119.46, 119.43, 99.0, 98.9, 76.7, 70.4, 70.2, 69.6, 69.0, 68.6, 66.8, 66.7, 66.2, 66.1, 63.9, 61.7, 60.2, 55.4, 19.4, 19.28, 19.27, 19.25, 19.22, 19.15. ^{31}P-NMR (400 MHz, CD_3OD) δ 0.28. IR (NaCl, film): 3431, 2954, 1749, 1623, 1452, 1370, 1227, 1081, 1046, 740 cm^{-1}. HRMS (ESI) Calcd. for $C_{49}H_{56}NNaO_{24}P$ $[M + Na]^+$ requires 1096.2822, Found: 1096.2830[41].



Synthesis of glycoamino acid **54. 107** was prepared as reported in the literature. [37] To the solution of **107** (678 mg, 1.0 mmol) and Fmoc-Cys-OH (514 mg, 1.5 mmol) in MeCN (20 ml), $BF_3{\cdot}OEt_2$ (0.46 mL, 3.0 mmol) was added.[54] The resulting mixture was stirred at room temperature for 19 h under argon. The solvent was removed under reduced pressure, and the resulting residue was diluted with EtOAc then washed with water. The organic layer was dried (Na_2SO_4), filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on a silica gel column (Hex/EtOAc/AcOH = 2:1:0.3→3:2:0.5) to give **54** (414 mg, 43%) as a white foam. ^{1}H-NMR (400 MHz, $CDCl_3$) δ 7.76 (d, J = 7.6 Hz, 2H, H-Fmoc), 7.61 (d, J = 7.2 Hz, 2H, H-Fmoc), 7.40 (t, J = 7.4 Hz, 2H, H-Fmoc), 7.32 (t, J = 7.4 Hz, 2H, H-Fmoc), 6.05 (d, J = 7.2 Hz, 1H, NH), 5.48 (s, 1H, H-1), 5.40 (dd, 3H, J = 10.0, 2.7 Hz, H-3’), 5.31-5.36 (m, 2H, H-4, H-4’), 5.17-5.23 (m, 2H, H-2’, H-3), 4.92 (d, J = 1.2 Hz, 1H, H-1’), 4.64 (s, 1H, H-α), 4.27-4.45 (m, 5H, H-5’, H-6, CH_2- Fmoc), 4.22 (t, 1H, J = 7.0 Hz, CH-Fmoc), 4.13-4.17 (m, 4H, H-2, H-5, H-6’), 3.24 (dd, 2H, J = 66.0, 13.6 Hz, CH_2-β), 2.14 (s, 3H, CH_3-Ac), 2.13 (s, 3H, CH_3-Ac), 2.12 (s, 3H, CH_3-Ac), 2.08 (s, 3H, CH_3-Ac), 2.04 (s, 3H, CH_3-Ac), 2.02 (s, 3H, CH_3-Ac), 2.01 (s, 3H, CH_3-Ac). ^{13}C-NMR (100 MHz, $CDCl_3$) δ 171.9, 171.1, 170.3, 169.8, 169.7, 169.6, 169.4, 155.8, 143.74, 143.72, 141.3, 127.8, 127.1, 125.1, 120.0, 99.2, 83.6, 78.4, 77.2, 70.2, 69.62, 69.56, 69.1, 68.4, 67.3, 66.6, 66.2, 62.4, 62.1, 53.6, 47.1, 33.6, 20.9, 20.8, 20.67, 20.65, 20.63. IR (NaCl, film): 3350, 2956, 1749, 1370, 1228, 1045, 741 cm^{-1}. HRMS (ESI) Calcd. for $C_{44}H_{51}NNaO_{21}S$ $[M + Na]^+$ requires 984.2567, Found: 984.2556.

3.4.3 Synthesis and characterization of CBM glycoforms

General procedure for the synthesis of unglycosylated CBM variants. The crude peptide was prepared using the previously reported protocol[54]. 16 mg of the crude peptide was dissolved in 80 ml of folding buffer (0.2 M Tris-acetate, 0.33 mM oxidized glutathione, 2.6 mM reduced glutathione, pH 8.2) and stirred at room temperature for 12 h under a helium atmosphere. The solution was then concentrated to a small volume (around 6 ml) using 3 kDa cut-off centrifugal filter units (Amicon) before RP-HPLC purification. The RP-HPLC purification was performed on a Versagrad Preparation-HPLC system using a semi-preparative C18 column. The products were detected by UV absorption at 275 nm. After HPLC purification with a linear gradient of 20→40% MeCN in H_2O over 30 min, the fractions were collected and checked by ESI+ MS. The pure fractions were combined and lyophilized to give the desired product as a white solid.

General procedure for the synthesis of glycosylated CBM variants. The crude glycopeptide was prepared using the previously reported protocol[54]. 16 mg of the crude peptide was dissolved in 1 ml of hydrazine solution (hydrazine/H_2O, 5/100, v/v) and stirred at room temperature for 30 min under helium. The reaction was quenched with 2 ml of acetic acid solution (AcOH/H_2O, 5/100, v/v). The resulting mixture was diluted to 80 mL with folding buffer (0.2 M Tris-acetate, 0.33 mM oxidized glutathione, 2.6 mM reduced glutathione, pH 8.2, 80 ml) and stirred at room temperature for 12 h under a helium atmosphere. The solution was then concentrated to a small volume (around 6 ml) using 3 kDa cut-off centrifugal filter units (Amicon) before RP-HPLC purification. The RP-HPLC purification was performed on a Versagrad Preparation-HPLC system using a semi-preparative C18 column. The products were detected by UV absorption at 275 nm. After HPLC purification with a linear gradient of 20→40% MeCN in H_2O over 30 min, the fractions were collected and checked by ESI+ MS. The pure fractions were combined and lyophilized to afford the desired product as a white solid.

LC-MS analysis of purified CBM variants. LC-MS was performed under two flow rates with C4 column: (1) 0.5 ml/min with a linear gradient of 15% to 35% acetonitrile in water over 3 min and (2) 0.3 ml/min with a linear gradient of 15% to 35% acetonitrile in water over 5 min.

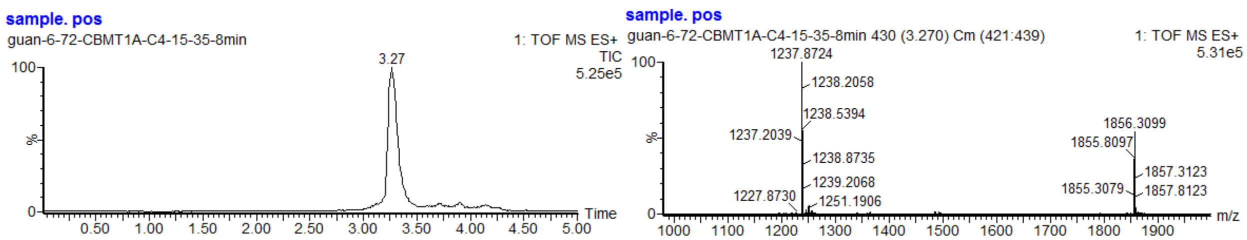


Figure 3.7 - LC-MS trace and ESI-MS of purified CBM **4** (yield: 28%). LC-MS condition: 0.3 mL/min, 15%-35% MeCN in H_2O over 5 min. MS (ESI) Calcd for **4** $C_{158}H_{233}N_{43}O_{53}S_4$ $[M+2H]^{2+}$ m/z = 1855.2949, $[M+3H]^{3+}$ m/z = 1237.1992.

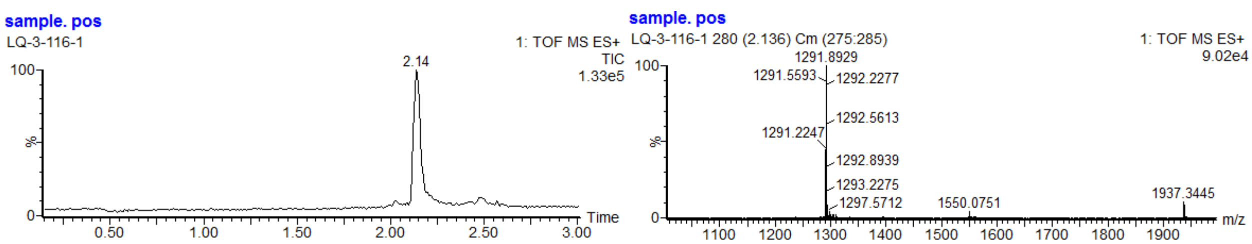


Figure 3.8 - LC-MS trace and ESI-MS of purified CBM **5** (yield: 18%). LC-MS condition: 0.5 ml/min, 15%-35% MeCN in H_2O over 3 min. MS (ESI) Calcd for **5** $C_{164}H_{243}N_{43}O_{58}S_4$ $[M+2H]^{2+}$ m/z = 1936.3135, $[M+3H]^{3+}$ m/z =1291.2090.

Confirming disulfide linkages. The folding of the CBMs was confirmed as described in our previous report.[18] After HPLC purification, the UPLC-MS trace of the folded CBMs showed a single peak, which indicated the homogeneity of the product. The observed mass loss of 4 Da is consistent with the formation of two disulfide bridges. The far-ultraviolet CD spectra of the CBMs were very similar to previously obtained spectrum of the unglycosylated CBM, which suggested that the synthetic peptide adopted the appropriate secondary structure upon folding. Moreover, the UPLC-MS analysis revealed that the thermolysin digestion of CBMs produces two fragments that contain two short peptide chains, VC/YSQCL and YGQCGG/ASGTTCQV (or AGQCGG/ASGTTCQV for CBM **11**). These short peptide chains are connected by disulfide linkages, which clearly confirmed the correct disulfide connectivity (Figure 3.9).

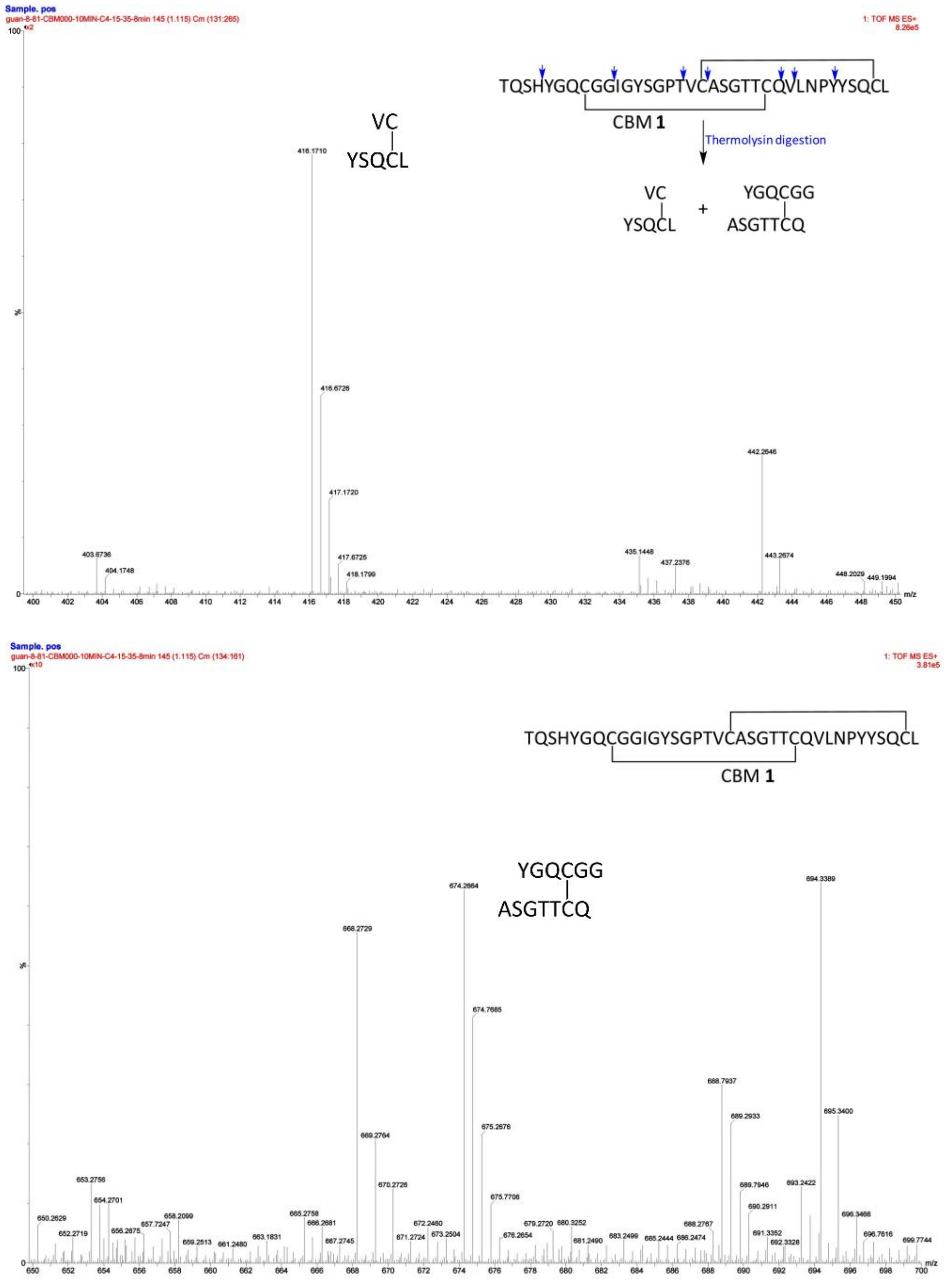


Figure 3.9 - Confirming disulfide linkages of representative CBM variants by thermolysin digestion. MS (ESI) Calcd for VC/YSQCL $C_{34}H_{54}N_8O_{12}S_2$ $[M+2H]^{2+}$ m/z = 416.1730; MS (ESI) Calcd for YGQCGG/ASGTTCQV $C_{52}H_{82}N_{16}O_{22}S_2$ $[M+2H]^{2+}$ m/z = 674.2693; MS (ESI) Calcd for AGQCGG/ASGTTCQV $C_{46}H_{78}N_{16}O_{21}S_2$ $[M+2H]^{2+}$ m/z = 628.2562.

3.5 References

1. J. Kraulis, G. M. Clore, M. Nilges, T. A. Jones, G. Pettersson, J. Knowles and A. M. Gronenborn, *Biochemistry*, 1989, **28**, 7241-7257.
2. R. M. Anthony, F. Wermeling and J. V. Ravetch, *Ann. N Y Acad. Sci.*, 2012, **1253**, 170-180.
3. J. L. Price, E. K. Culyba, W. Chen, A. N. Murray, S. R. Hanson, C. H. Wong, E. T. Powers and J. W. Kelly, *Biopolymers*, 2012, **98**, 195-211.
4. C. B. Taylor, M. F. Talib, C. McCabe, L. Bu, W. S. Adney, M. E. Himmel, M. F. Crowley and G. T. Beckham, *J. Biol. Chem.*, 2012, **287**, 3147-3155.
5. A. M. Sinclair and S. Elliott, *J. Pharm. Sci.*, 2005, **94**, 1626-1635.
6. S. E. O'Connor, J. Pohlmann, B. Imperiali, I. Saskiawan and K. Yamamoto, *J. Am. Chem. Soc.*, 2001, **123**, 6187-6188.
7. R. M. Anthony, F. Nimmerjahn, D. J. Ashline, V. N. Reinhold, J. C. Paulson and J. V. Ravetch, *Science*, 2008, **320**, 373-376.
8. E. K. Culyba, J. L. Price, S. R. Hanson, A. Dhar, C. H. Wong, M. Gruebele, E. T. Powers and J. W. Kelly, *Science*, 2011, **331**, 571-575.
9. P. Wang, S. Dong, J. H. Shieh, E. Peguero, R. Hendrickson, M. A. Moore and S. J. Danishefsky, *Science*, 2013, **342**, 1357-1360.
10. K. Kodier and C.-H. Wong, in *Glycoscience: Chemistry and Chemical Biology I–III*, eds. B. Fraser-Reid, K. Tatsuta and J. Thiem, Springer Berlin Heidelberg, 2001, DOI: 10.1007/978-3-642-56874-9_56, ch. 56, pp. 2305-2352.
11. L. A. Marcaurelle and C. R. Bertozzi, *Glycobiology*, 2002, **12**, 69R-77R.
12. D. P. Gamblin, E. M. Scanlan and B. G. Davis, *Chem. Rev.*, 2009, **109**, 131-163.
13. A. Fernandez-Tejada, J. Brailsford, Q. Zhang, J. H. Shieh, M. A. Moore and S. J. Danishefsky, *Top. Curr. Chem.*, 2015, DOI: 10.1007/128_2014_622.
14. M. Linder, M. L. Mattinen, M. Kontteli, G. Lindeberg, J. Stahlberg, T. Drakenberg, T. Reinikainen, G. Pettersson and A. Annila, *Protein. Sci.*, 1995, **4**, 1056-1064.
15. J. Lehtio, J. Sugiyama, M. Gustavsson, L. Fransson, M. Linder and T. T. Teeri, *Proc. Natl. Acad. Sci. U S A*, 2003, **100**, 484-489.
16. L. Chen, M. R. Drake, M. G. Resch, E. R. Greene, M. E. Himmel, P. K. Chaffey, G. T. Beckham and Z. Tan, *Proc. Natl. Acad. Sci. U S A*, 2014, **111**, 7612-7617.
17. X. Wu and P. G. Schultz, *J. Am. Chem. Soc.*, 2009, **131**, 12497-12515.
18. W.-g. Wu, L. Pasternack, D.-H. Huang, K. M. Koeller, C.-C. Lin, O. Seitz and C.-H. Wong, *J. Am. Chem. Soc.*, 1999, **121**, 2409-2417.
19. D. Russell, N. J. Oldham and B. G. Davis, *Carbohydr. Res.*, 2009, **344**, 1508-1514.
20. S. R. Hanson, E. K. Culyba, T. L. Hsu, C. H. Wong, J. W. Kelly and E. T. Powers, *Proc. Natl. Acad. Sci. U S A*, 2009, **106**, 3131-3136.
21. K. B. Hojlys-Larsen and K. J. Jensen, *Methods Mol. Biol.*, 2013, **1047**, 191-199.
22. G. Johansson, J. Stahlberg, G. Lindeberg, A. Engstrom and G. Pettersson, *FEBS Lett.*, 1989, **243**, 389-393.
23. C. M. Payne, M. G. Resch, L. Chen, M. Crowley, M. E. Himmel, L. E. Taylor II, M. Sandgren, J. Stahlberg, I. Stals, Z. Tan and G. T. Beckham, *Proc. Natl. Acad. Sci. U S A*, 2013, **110**, 14646–14651.
24. E. R. Lacy, M. Baker and M. Brigham-Burke, *Anal. Biochem.*, 2008, **382**, 66-68.

25. I. Stals, K. Sandra, B. Devreese, J. Van Beeumen and M. Claeyssens, *Glycobiology*, 2004, **14**, 725-737.
26. P. M. Rudd, H. C. Joao, E. Coghill, P. Fiten, M. R. Saunders, G. Opdenakker and R. A. Dwek, *Biochemistry*, 1994, **33**, 17-22.
27. E. S. Radisky and D. E. Koshland, Jr., *Proc. Natl. Acad. Sci. U S A*, 2002, **99**, 10316-10321.
28. A. Amore, A. Serpico, A. Amoresano, R. Vinciguerra and V. Faraco, *Biotechnol. Appl. Biochem.*, 2015, DOI: 10.1002/bab.1325.
29. A. Fontana, P. P. de Laureto, B. Spolaore, E. Frare, P. Picotti and M. Zambonin, *Acta. Biochim. Pol.*, 2004, **51**, 299-321.
30. J. L. Asensio, A. Arda, F. J. Canada and J. Jimenez-Barbero, *Acc. Chem. Res.*, 2013, **46**, 946-954.
31. E. E. Simanek, D.-H. Huang, L. Pasternack, T. D. Machajewski, O. Seitz, D. S. Millar, H. J. Dyson and C.-H. Wong, *J. Am. Chem. Soc.*, 1998, **120**, 11567-11575.
32. F. A. Qulocho, *Pure Appl. Chem.*, 1989, **61**, 1293-1306.
33. M. Nagae, K. Yamanaka, S. Hanashima, A. Ikeda, K. Morita-Matsumoto, T. Satoh, N. Matsumoto, K. Yamamoto and Y. Yamaguchi, *J. Biol. Chem.*, 2013, **288**, 33598-33610.
34. A. W. Barb, A. J. Borgert, M. Liu, G. Barany and D. Live, *Methods Enzymol.*, 2010, **478**, 365-388.
35. D. Shental-Bechor and Y. Levy, *Curr. Opin. Struct. Biol.*, 2009, **19**, 524-533.
36. K. Teilum, J. G. Olsen and B. B. Kragelund, *Biochim. Biophys. Acta.*, 2011, **1814**, 969-976.
37. L. Chen and Z. Tan, *Tetrahedron Lett*, 2013, **54**, 2190-2193.
38. H. K. Cui, Y. Guo, Y. He, F. L. Wang, H. N. Chang, Y. J. Wang, F. M. Wu, C. L. Tian and L. Liu, *Angew Chem Int Ed Engl*, 2013, **52**, 9558-9562.
39. H. Yu and X. Chen, *Org Lett*, 2006, **8**, 2393-2396.
40. Y. Liu, J. Marshall, Q. Li, N. Edwards and G. Chen, *Bioorg Med Chem Lett*, 2013, **23**, 2328-2331.
41. L. Zervas and I. Dilaris, *Journal of the American Chemical Society*, 1955, **77**, 5354-5357.
42. C. Plattner, M. Hofener and N. Sewald, *Org Lett*, 2011, **13**, 545-547.
43. D. M. Rothman, M. E. Vazquez, E. M. Vogel and B. Imperiali, *J Org Chem*, 2003, **68**, 6795-6798.
44. T. Rosen, I. M. Lico and D. T. W. Chu, *The Journal of Organic Chemistry*, 1988, **53**, 1580-1582.
45. L. A. Salvador, M. Elofsson and J. Kihlberg, *Tetrahedron*, 1995, **51**, 5643-5656.
46. J. Ohlsson and G. Magnusson, *Carbohydrate Research*, 2000, **329**, 49-55.
47. Y. Nishi and T. Tanimoto, *Biosci Biotechnol Biochem*, 2009, **73**, 562-569.
48. K. C. Nicolaou, C. W. Hummel and Y. Iwabuchi, *Journal of the American Chemical Society*, 1992, **114**, 3126-3128.
49. K. Hiruma-Shimizu, K. Hosoguchi, Y. Liu, N. Fujitani, T. Ohta, H. Hinou, T. Matsushita, H. Shimizu, T. Feizi and S. Nishimura, *J Am Chem Soc*, 2010, **132**, 14857-14865.
50. S. Peters, T. L. Lowary, O. Hindsgaul, M. Meldal and K. Bock, *Journal of the Chemical Society-Perkin Transactions 1*, 1995, DOI: Doi 10.1039/P19950003017, 3017-3022.
51. L. Wu and N. S. Sampson, *ACS Chem Biol*, 2014, **9**, 468-475.
52. Y. A. Lin, J. M. Chalker and B. G. Davis, *J Am Chem Soc*, 2010, **132**, 16805-16811.
53. J. Zeng, S. Vedachalam, S. Xiang and X. W. Liu, *Org Lett*, 2011, **13**, 42-45.
54. L. Chen, M. R. Drake, M. G. Resch, E. R. Greene, M. E. Himmel, P. K. Chaffey, G. T. Beckham and Z. Tan, *Proc Natl Acad Sci U S A*, 2014, **111**, 7612-7617.

Chapter 4

Quantitative Effects of *O*-Mannosylation on the Folding, Biophysical and Chromatographic Properties of a Family 1 Carbohydrate-Binding Module

4.1 – Introduction

Almost all secreted and integral membrane proteins are co-translationally passed, unfolded, through the secretory 61 (SEC61) translocon complex on the membrane of the endoplasmic reticulum (ER) into the ER lumen.[1] As the nascent peptide enters the ER lumen, numerous chaperones bind the growing chain to prevent misfolding events and aggregation.[2] In addition, numerous other proteins located in the ER lumen are involved in the covalent modification of the nascent peptide including proteases that cleave signal peptides, oxidoreductases that form disulfide bonds, glycosylphosphatidylinositol (GPI) transamidase that attaches GPI anchors, ADP-ribosyltransferases that attach ADP-ribose, and glycosyltransferases that add a variety of *N*- and *O*-glycans.[1] These modifications are critical for the proper maturation and exit from the ER of the final protein product, and most, if not all occur co-translationally as the protein is being extruded into the ER lumen.[2] Although there are many modifications that occur in the ER, glycosylation and disulfide bond formation have been shown to have particularly large effects on the folding process and maturation of proteins that pass through the secretory system. The details of how these two covalent

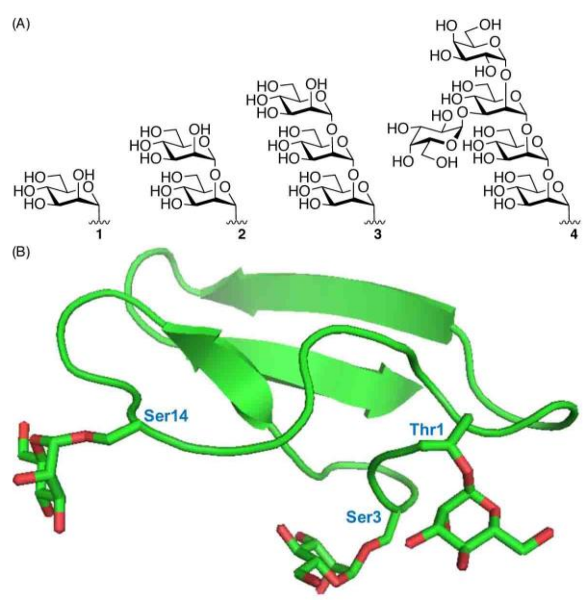


Figure 4.1 – (A) *O*-linked glycans cores on yeast and fungal proteins and (B) NMR structure of the Family 1 CBM with O-mannosyl residues at Thr1, Ser3, and Ser14.

modifications interact with one another during protein folding, particularly in the case of *O*-glycosylation, are not well understood.

Disulfide bonds are very important for proper folding, function and stability of many secreted proteins, and are thought to aid in thermodynamic stabilization of the final folded structure mainly by destabilizing the unfolded state.[3] A protein's final structure forms as a result of two closely intertwined processes: establishment of disulfide bonds and construction of a defined secondary/tertiary structure. Combined together these pathways are commonly known as oxidative folding.[3] In the ER both disulfide bond pairing and global folding events are highly regulated by a large suite of enzymes and chaperones to ensure most proteins achieve the correct fold and structure.[2] Protein disulfide isomerases (PDIs) are a large family of proteins located in the ER that catalyze thiol exchange reactions and are responsible for correct disulfide bond pairing and protein folding.[4]

Protein *O*-mannosylation is the second most common type of *O*-glycosylation found in yeast, fungal, and mammalian systems.[5-7] The transfer of the first *O*-linked mannose (Man) to serine (Ser) and threonine (Thr) by protein-*O*-D-mannosyltransferase occurs, similar to disulfide bond formation and chaperone-assisted folding, co-translationally in the endoplasmic reticulum (ER).[8] The initial mannose residue is then elongated in the Golgi apparatus by the sequential addition of more mannose or other monosaccharide residues.[5] In yeast and filamentous fungi, the core structures of *O*-linked glycans on proteins are predominantly Man1-Ser/Thr **1**, Manα1,2Manα1-Ser/Thr **2**, and Manα1,2Manα1,2Manα1-Ser/Thr **3** (Figure 4.1A). Further modifications of the structural cores with glucose (Glc), galactopyranose (Galp), or galactofuranose (Galf) residues also occur, leading to the formation of more complex and branched structures such as **4**.[9]

Other forms of glycosylation are strongly associated with the folding processes of glycoproteins in the ER. For example, *N*-glycans have long been known to act as signals for the lectin-chaperones Calnexin and Calreticulin, which are critical for proper maturation of *N*-glycoproteins.[2] Additionally, the intrinsic effects of *N*-glycosylation on protein structure and folding have been studied by several groups.[10-14] Such direct links between encouraging the correct folding of *O*-glycoproteins and *O*-mannosylation have not yet been found, however there is evidence that the *O*-mannosylation of proteins and oxidative folding pathways are connected. For example, when *O*-mannoyslation is blocked, accumulation and aggregation of unfolded proteins in the ER greatly increases, suggesting that *O*-mannosylation is involved in clearing abnormal proteins via the ERAD pathway, possibly by solubilizing the misfolded transcripts.[15] Also interesting is the observation that blocking *N*-glycosylation with tunicamicin causes a large up-regulation in *O*-mannosylation and even proteins that don't normally receive the modification become *O*-mannosylated under such conditions.[16] Possibly, this is a compensatory adjustment to help limit misfolding of proteins normally aided in folding by *N*-glycans. Additionally, it has been shown that certain members of the protein-*O*-mannosyltransferase family interact with chaperones in the ER of yeast.[17] Finally, recent studies propose that proteins that take too long to fold are "tagged" with *O*-mannose, which moves them out of the ER through the ERAD pathway.[18] *O*-mannosylation in this context is thought to stop the wasteful use of chaperones and energy on molecules that cannot fold properly. Together these studies hint at a strong role of *O*-mannosylation in the oxidative folding of proteins, but so far a detailed study on the intrinsic effects of this form of glycosylation on the folding of glycoproteins has not been carried out.

We have chosen to investigate the *O*-mannosylation and folding kinetics of a Family 1 carbohydrate-binding module (CBM) of the glycoside hydrolase Family 7 cellobiohydrolase from the cellulolytic fungus, *Trichoderma reesei* (*Tr*Cel7A), a key enzyme in the cellulosic biofuels industry (Figure 4.1B)[19-21] This CBM molecule is small (36 amino acids long), natively mannosylated at three sites (Thr1, Ser3, and Ser14), folded with two disulfide bonds and synthetically tractable, making it an excellent system to study the interplay between *O*-mannosylation and folding.[22]

Previously, we applied a chemical glycobiology approach to reveal that *O*-mannosylation can site-specifically affect numerous biophysical properties of the CBM. In this study we used a similar strategy to characterize the unfolding rate and rate of disulfide bond formation for a variety of differently glycosylated isoforms (glycoforms) of the *Tr*Cel7A CBM. Our results revealed a strong dependence on glycosylation pattern for both rates.

4.2 Results

4.2.1 Design and synthesis CBM variants

We have demonstrated that combining the data from a wide variety of glycoforms can provide a comprehensive view of the influence of protein *O*-glycosylation.[19-21] Therefore, we selected 22 CBM glycoforms for the present study, which can be divided into three series for a systematic study of three different features of CBM *O*-glycosylation (Figure 4.2). The first series (Figure 4.2, **6-14**) was meant to investigate both the site-specific and glycan-size dependent consequences of glycosylation. The second series (Figure 4.2, **15-27**) looked at the synergistic effects of small glycans at multiple sites on the peptide backbone. The final series (Figure 4.2, **19-27**) probed how the stereochemistry of the glycan units and linkage along with the amino acid side chains in close proximity to the glycan site alter the physical effects of *O*-glycosylation. The unglycosylated CBM peptide **5** was also included as a control.

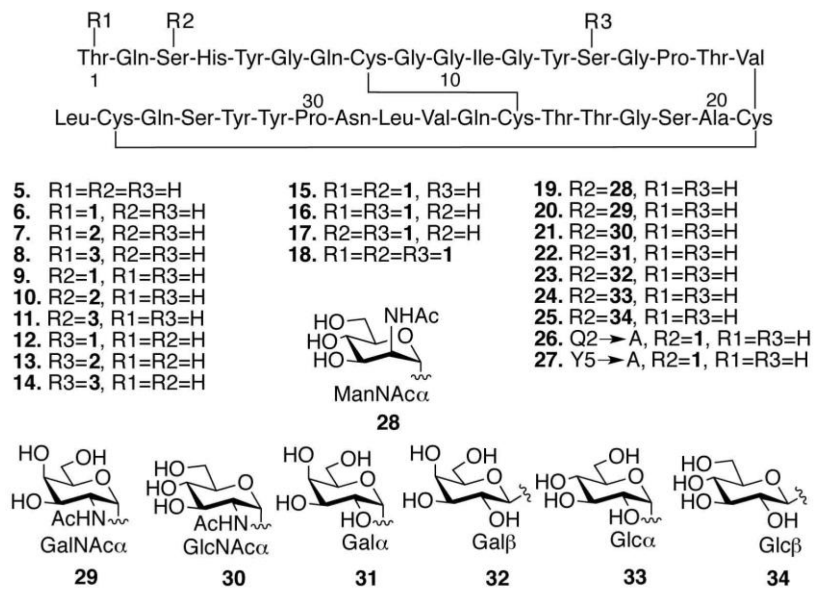


Figure 4.2 - Synthetic CBM varaints that are used in this study.

Although enzymatic synthesis is a valid method of producing modified peptides and has been successfully employed to great effect by others,[23-25] we chose chemical synthesis as our means to obtain the glycoform library. Chemical synthesis presents a unique level of control over all factors in the construction of modified peptides that enzymatic synthesis, although sometimes more convenient, cannot duplicate.[26-34] This allowed us the opportunity to thoroughly investigate the influence of any feature of protein *O*-glycosylation we desired. In particular, investigating the intricacies of stereochemistry in both the carbohydrate ring hydroxyl groups and the anomeric linkage between glycan and peptide was made significantly easier by the total chemical synthesis of the glycopeptides used in the study. This is because enzymatic glycosylation reactions are controlled by a great many factors including enzyme substrate specificities, amino acid sequences, and local peptide conformations, which makes the controlled synthesis of some glycosylated peptides differing by only minor structural alterations difficult. Chemical glycosylation reactions, on the other hand, are not nearly as beholden to such factors and, as a result, can easily produce closely related glycoforms of a given glycopeptide.

We were able to use our previously developed one-pot chemical synthesis and folding procedure for the production of all CBM glycoforms used in this study[20,21] Synthesis was carried out with 9-fluorenylmethoxycarbonyl (Fmoc)-based solid-phase peptide synthesis (SPPS), which easily accommodated the site-specific incorporation of protected amino acid and glycoamino acid building blocks. During SPPS, all sugar hydroxyl groups on the side chains of the glycoamino acid building blocks where protected as acetyl esters, which are stable during peptide coupling procedures and are easily removed under mild, carbohydrate-compatible conditions. Although most of the glycoamino acid building blocks used to synthesize the CBM glycoforms in this study are not commercially available, they have been previously synthesized and these procedures were followed. All anomeric stereochemical configurations of the synthetic building blocks were verified with 2D coupled-HSQC NMR spectra, as previously demonstrated. After synthesis of the building blocks, our previously-developed one-pot synthesis and folding method allowed us to produce all the desired glycoforms in high yield. Purity and homogeneity of each glycopeptide was experimentally verified by analytical liquid chromatography-mass spectroscopy (LC-MS).

4.2.2 Rate of disulfide bond formation

As mentioned above, disulfide bonds are critical to the structural integrity and function of the many proteins that contain them. Given the widely acknowledged role of *N*-glycan structures in oxidative folding, we expected to see some similar function for the *O*-glycan structures that are also added in the ER at the same time. In order to better understand any links between *O*-glycan structure and oxidative folding, we began by quantifying the rate of disulfide bond formation for each glycoform we had synthesized above. The CBM peptide domain contains two disulfide bonds when fully folded (Figure 4.2). This leads to a highly noticeable mass difference of 4 Da between the fully folded and unfolded peptides, which allowed us to conveniently use high resolution mass spectroscopy to monitor the formation of the disulfide bonds. Small scale folding reactions were set up with concentrations of peptide, buffer salts and redox reagents identical to those of the folding conditions used in the production of fully folded CBM glycoforms.[36,37] At different time intervals, aliquots were removed from the folding reaction and quenched by adjusting the pH to 3. Optimization experiments showed that disulfide bond formation was minimized around pH 3 and such mildly acidic conditions are commonly used to slow disulfide bond reaction in proteins.[3] Once quenched, the samples were stored at 4 ºC for no more than 24 hours before being analyzed by high resolution LC-MS. Control experiments verified that these storage conditions did not significantly alter the amount of oxidized peptide in the samples. Previous experiments with the synthesis of CBM glycoforms had resulted in an LC protocol that effectively separated the pure product from various side-products produced during SPPS and deprotection, which allowed us to easily compare the masses of pure CBM peaks at each time point. Unfortunately, we could not achieve reliable separation of all the partially oxidized intermediates in all cases, and so the average mass of the peptides in each sample was calculated. Since this average mass corresponds with the amount of oxidized peptide in the sample at the time of injection, it will also give us a measure of the overall rate of oxidative folding for each of the glycoforms. These curves displayed nearly linear behavior during the majority of the folding reaction, and so the slopes of linear fits for each glycoform were compared as a way to quantify the difference in rate of disulfide bond formation between glycoforms.

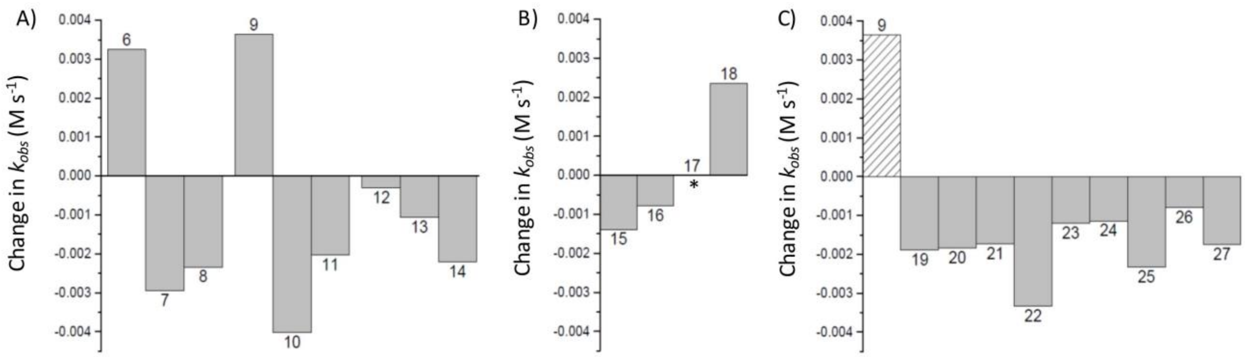


Figure 4.3 - Change in the observed rate of disulfide bond formation (k_{obs}) relative to the unglycosylated CBM peptide **5**. (A) Singly glycosylated CBM glycoforms. (B) Multiply glycosylated CBM glycoforms. *bi-phasic folding kinetics, could not be fit to simple zero order rate law equation (C) Derivatives of CBM **9** carrying a variety of different monosaccharides at the Ser3 glycosylation as well as glycoforms **26** and **27** which contain amino acid mutations. The change in observed rate for **9** is included for comparison.

Looking at the three possible glycosylation sites individually, our data shows that addition of a single mannose at either Thr1 or Ser3 significantly increased the rate of disulfide bond formation while Ser14 had very little effect on the rate as compared to unglycosylated CBM (Figure 4.3A, compare the folding of **6-14**). For each site, increasing the size of the linear mannose chains decreased the rate of bond formation with Ser3 giving the largest decrease in rate as a result of increasing glycan size (Figure 4.3A, compare the folding of **9-11**). Overall, only the presence of a single mannose at either Ser3 or Thr1 increased the rate of disulfide bond formation, while large glycans at any site had a detrimental effect on the rate.

Multiple glycosylation of the CBM with single mannose glycans can result in slower disulfide formation, even when the glycans are at the same sites that individually increase the rate. Of the multiply glycosylated analogs, the glycoform with a single mannose glycan at each of the three possible glycosylation sites displayed an increase in the rate of disulfide bond formation (Figure 4.3B, **15-18**).

Interestingly, CBM glycoform **17** seemed to display a biphasic folding reaction with a very fast initial phase that ended after loss of 2 hydrogens and second slower phase. This suggests that one of the disulfide bonds in this molecule is unusually stable, although the exact reason for this is not immediately obvious. Further work with **17** is currently being carried to verify these results and further investigate the possible causes.

Previously, glycosylation at Ser3 was shown to have the largest effect on many biophysical properties of the CBM compared to the other two possible sites. Our data also shows that this site can strongly influence the rate of oxidative folding for the CBM, and so we chose to further investigate the glycosylation at this site in particular. We started by attaching a variety of carbohydrates to the Ser3 site that included both alternate functional groups on the ring, such as the amide containing carbohydrates in **19-21**, and alternate orientations of the ring hydroxyl groups (**22** and **24**). We also included pairs of anomeric analogs to compare the α- and β-linked versions (**22/23** and **24/25**). Earlier work suggests that the anomeric linkage stereochemistry is critical to the interaction between glycan and peptide.[38] and we were curious how this could affect things. We found that any differences between the glycosylated isoforms carrying non-mannose sugars and the unglycosylated control were decreases in the rate of disulfide bond formation. All of the amide-containing carbohydrates seemed to behave very similar in this assay (Figure 4.3C, compare the folding of **19-21**). The anomeric linkage stereochemistry seems to have a definite effect on the magnitude of the rate change, and is additionally dependent on the identity of the carbohydrate residue (Figure 4.3C, compare the folding of **22-25**). Comparing the α- and β-linked galactose containing glycoforms shows that the α-linked carbohydrate had a much larger negative effect on the rate than the β-linked glycoform. For glucose, however, the opposite was observed: the β-linked glucose had the larger (still negative) effect. We also investigated the role played by specific amino acid side chains in the vicinity of the glycosylation site. These were chosen based on our previous work indicating that Gln2 and Tyr5 were the most important side chains for stabilizing the CBM peptide. Here we saw that mutating either Gln2 or Tyr5 to Ala resulted in a significant decrease in the rate of disulfide bond formation as compared to the unglycosylated control (Figure 4.3C, compare the folding of **26** and **27**). All of these rate changes are in stark contrast to the glycoform contain a single mannose at the same position (Figure 4.3C, compare the folding of **9** and **19-27**).

4.2.3 Secondary structure folding kinetics

Stabilizing the final structure of a glycoprotein seems to be a common feature of many types of naturally occurring glycosylation.[13,38] We have previously shown *O*-glycosylation to have a strong effect and thermodynamic stability of the CBM peptide domain, and we were interested to see if similar glycosylation patterns could also affect the kinetics of unfolding.[11,12] To better understand the role of *O*-glycosylation in stabilizing the CBM peptide, we chose to use circular dichroism (CD) spectroscopy to monitor the rate of secondary structure loss upon rapid heating.[39]

Fully folded and purified samples of each CBM glycoform were dissolved in a minimal amount of buffer and transferred to the CD cuvette. The temperature of the sample was very quickly raised to 80 ºC by adding hot buffer to the sample and rapidly placing the cuvette in the CD machine, which had been pre-equilibrated at 80 ºC. This procedure resulted in only about 12-15 seconds of dead time between initiating the thermal unfolding and starting the data collection, which was not a significant problem for any of the glycoforms in this study since the average time to complete the unfolding reaction was in excess of 200 seconds. A similar procedure was attempted in reverse to quickly cool a heated and unfolded sample in order to observe the formation of secondary structure. Unfortunately, the folding reaction was found to occur extremely rapidly and even the use of a stop-flow apparatus with only 10 milliseconds of dead time for the measurements did not allow us to observe any amount of folding.

Secondary structure loss was monitored by continually measuring the CD reading at 217 nm, which was determined to be the wavelength at which the largest change occurred during the transition between folded and unfolded states. For each sample, complete spectra in both folded and unfolded states were collected in addition to the kinetic measurements to verify that proper folding occurred in each trial. In order to verify that the folding and unfolding process was indeed fully reversible for each analog, we cycled the temperature and collected complete CD spectra for the two states several times in succession. Collected CD data was fit with a first order reaction kinetics model and the rate constants for each fit were compared to give a quantitative measure of the rate of unfolding for each glycoform. Comparing these parameters across members of the glycoform library allowed us to quantify the influence of different glycosylation patterns and different glycans on the kinetic stability of the CBM secondary structure (Figure 4.4).

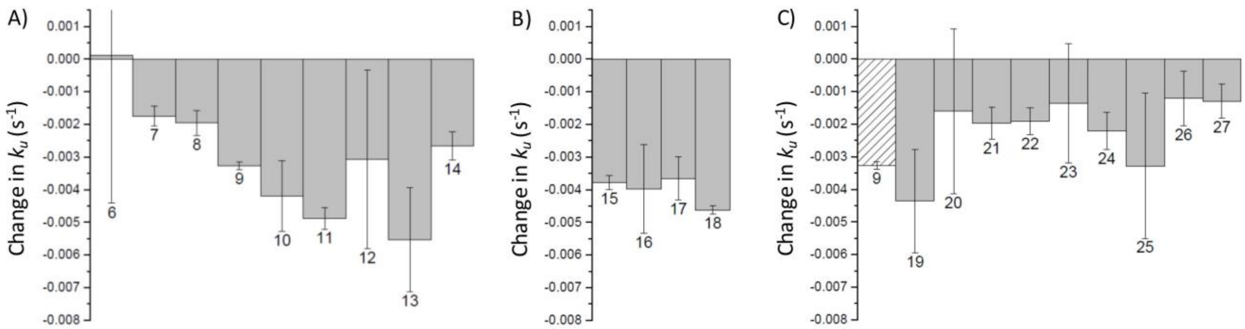


Figure 4.4 - Change in the rate of thermal unfolding relative to the unglycosylated CBM peptide **5**. Error bars reflect the standard deviation of three trials. (A) Singly glycosylated CBM glycoforms. (B) Multiply glycosylated CBM glycoforms. (C) Derivatives of CBM **9** carrying a variety of different monosaccharides at the Ser3 glycosylation as well as glycoforms **26** and **27** which contain amino acid mutations. The change in unfolding rate for **9** is included for comparison.

First, looking at each site individually, we can see that single mannose moieties at either Ser3 or Ser14 significantly stabilize the folded CBM structure, which is reflected by the lower rate constant and hence slower rate of unfolding. Interestingly, the same glycan (single mannose) at Thr1 did not have a very large effect one way or the other on the rate of unfolding under these conditions (Figure 4.4A, compare **6**, **9**, and **12**). Increasing the size of the glycan at any of the sites showed an increase in the kinetic stability of the secondary structure. For both the Thr1 and Ser3 sites, this increasing stability correlated with the size of the glycan, but at Ser14 site the stability peaked at di-mannose and fell upon addition of the third mannose to the linear glycan chain (Figure 4.4A, compare **7**, **8**, **10**, **11**, and **13**, **14**). Attaching multiple mannose residues to the CBM peptide appears to result in only modest improvements in the kinetic stability as compared to single glycosylations at Ser3 or Ser14, with the highest stability observed for a single mannose at each of the three sites (Figure 4.4B).

Further investigation of the structural details of carbohydrate residues attached at Ser3 showed interesting results. Changing the sugar moiety from mannose to *N*-acetylmannosamine resulted in a small but noticeable decrease in the rate of unfolding. This was in contrast to the other two amide-containing carbohydrates tested at the same position which stabilized the structure slightly less than mannose although still greater than no glycan at all (Figure 4.4C, compare **19**, **20** and **21**). Glucose gave an almost identical degree of stabilization as mannose when the glucose was linked via a β-linkage to the peptide while galactose was found to stabilize the peptide backbone structure to a small degree independent of the anomeric stereochemistry (Figure 4.4C, compare **22**, **23**, **24** and **25**). Mutating either Gln2 or Tyr5 to Ala while maintaining the single mannose glycan decreased the stabilization observed upon glycosylation, although those glycoforms were still shown to be slightly more stable than the unglycosylated peptide control (Figure 4.4C, compare **26** and **27**).

4.2.4 Solubility

We also quantified the solubility of each CBM glycoform using a simple and direct assay.[40] Solubility of peptides is important in many contexts,[41] and *O*-glycosylation is generally acknowledged to increase the solubility of many proteins. *O*-mannoyslation in particular has been recently implicated in the protein quality control mechanisms of yeast by solubilizing mis-folding protein transcripts in the ER lumen.[18] We were curious how the glycosylation pattern of the CBM could affect the solubility of the peptide domain, and how the various different types of *O*-glycans compared to *O*-mannose in this context.

To measure the solubility limits of each glycoform in the library, a small amount of buffer was added to samples of fully folded and purified lyophilized peptides. These samples were then allowed to equilibrate at 4 ºC for thirty minutes before being pelleted in a centrifuge. Aliquots of the supernatant from each sample were collected and the concentration of CBM dissolved in the supernatant was quantified using our previously established quantitative MALDI-TOF method.[20] In order to make sure that equilibrium was reached in the experimental time frame, select samples were allowed to equilibrate under identical conditions for an extended time. After 24 hours these samples did not show any appreciable difference in the amount of dissolved peptide present in the supernatant than those samples left for only 20 minutes.

We started by looking at the site-specific effects of glycosylation on the CBM peptide. Adding glycans at any site resulted in a higher solubility for the glycopeptide as compared to the unmodified peptide. At both Ser3 and Ser14 sites, the solubility further increased as the linear mannose chain was elongated. This was not the case, however, for the glycan at Thr1 site, which showed no large amount of change in solubility after the glycan chain was extended past the first carbohydrate (Figure 4.5A, compare **6-14**). Pairs of small glycans at any combination of sites gave only a modest improvement in solubility over singly glycosylated isoforms, and interestingly occupying all three possible sites with single mannose glycans did not make the glycopeptide significantly more soluble than having only two of the three occupied. Comparing the glycoforms carrying larger glycans at a single spot and several small glycans at multiple sites, it seems that the solubility is increased most with large glycans at a single spot even though the number of carbohydrate residues is the exact same (Figures 4.5A and 4.5B, compare **11, 14** and **18**).

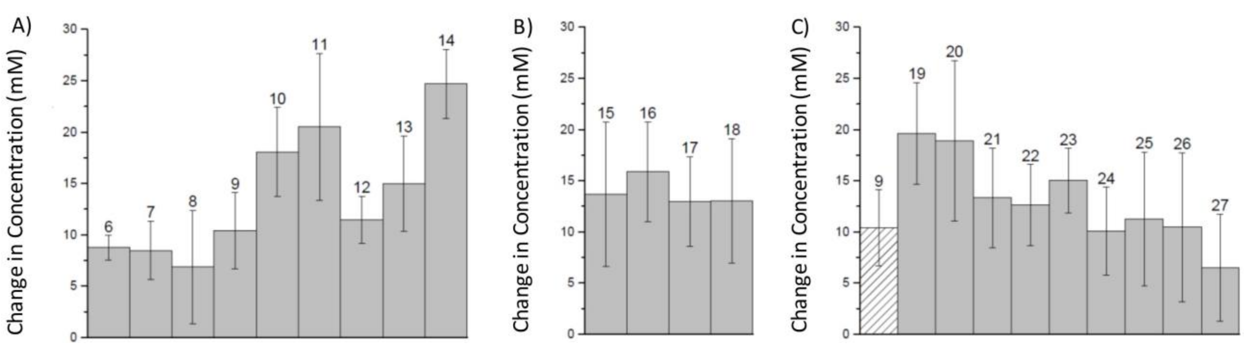


Figure 4.5 - Change in the concentration at the limit of solubility relative to the unglycosylated CBM peptide **5**. Error bars reflect the standard deviation of three trials. (A) Singly glycosylated CBM glycoforms. (B) Multiply glycosylated CBM glycoforms. (C) Derivatives of CBM **9** carrying a variety of different monosaccharides at the Ser3 glycosylation as well as glycoforms **26** and **27** which contain amino acid mutations. The change in solubility for **9** is included for comparison.

Changing the carbohydrate identity to a variety of sugars that are not mannose showed that most other carbohydrates gave a similar increase in solubility relative to the unglycosylated peptide control. Of the amide-containing carbohydrates both *N*-acetylgalactoseamine and *N*-acetylmannoseamine resulted in the largest increases in solubility (Figure 4.5C, compare the solubility limits for **19-21)**. The anomeric stereochemistry of the glycan-peptide linkage seems to have only a minor effect on the solubility of the glycopeptide, as shown by the slight difference between glycoforms containing the α- and β-linked galactose or glucose residues (Figure 4.5C, compare **22** to **23** and **24** to **25**). Of the two amino acid side chains tested here, only Tyr5 seems to be important for the solubility, as mutating this tyrosine to alanine resulted in a small decrease to the solubility relative to the mannose-containing wild-type sequence (Figure 4.5C, compare **27** to **9**). Removing the side chain of Gln2, on the other hand, had no significant effect on the solubility as shown by the almost identical solubility limits measured for **9** and **26** (Figure 4.5C).

4.2.5 Retention time

Retention time was also examined as both an important property on its own and as a measure of the overall hydrophobicity of the glycoforms.[42] This was done by individually running a uniform amount of each glycoform under identical chromatographic conditions as used for analytical LC-MS verification of glycoform purity. Retention time was taken at the highest point of the resulting total ion count chromatograms. The difference in observed retention times for any pair of glycoforms was found to be very constant and repeatable, and so the difference in retention time between the unglycosylated control **5** and each of the glycoforms was taken as an easily comparable measure.

All of the glycosylated analogs tested in this study displayed a lower retention time than the unglycosylated control, most likely reflecting the hydrophilicity of the carbohydrate residues. Of the three sites examined here, glycosylation at Ser3 had by far the greatest effect on the retention time. The difference was fairly pronounced with small glycans consisting of a single mannose and increasingly noticeable as the length of the linear glycan chain increased. Unlike the other two sites, which both showed larger retention time shifts with larger glycans attached, glycosylation at the Thr1 site showed no significant change in the retention time after the first mannose (Figure 4.6A, compare **6-14**). These site-specific trends were largely reflected in the glycoforms carrying single mannose moieties at multiple individual glycosylation sites. Comparing **15**, **16**, and **17**, it can be seen that both of the glycoforms with carbohydrates at the Ser3 position have significantly greater changes to their retention times (**15** and **17**) than the one that does not (**16**). This reflects the fact that Ser3 glycosylation on its own had a much higher

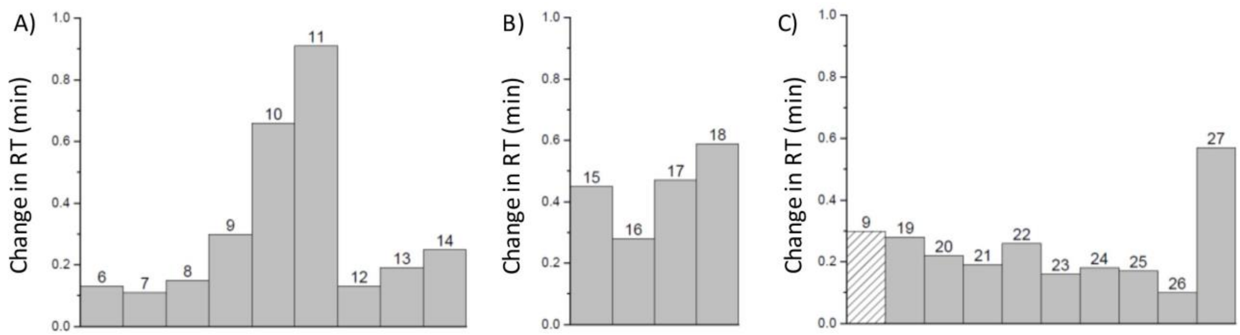


Figure 4.6 - Change in the retention time relative to the unglycosylated CBM peptide **5**. (A) Singly glycosylated CBM glycoforms. (B) Multiply glycosylated CBM glycoforms. (C) Derivatives of CBM **9** carrying a variety of different monosaccharides at the Ser3 glycosylation as well as glycoforms **26** and **27** which contain amino acid mutations. The change in solubility for **9** is included for comparison.

effect on the retention time than either Thr1 or Ser14, as mentioned previously. Since Thr1 and Ser14 are roughly equivalent to one another in their ability to lower the retention time of the CBM peptide when acting individually both **15** and **17** have similarly equivalent retention time shifts. Glycosylating all three sites on the CBM peptide results in small shift relative to the doubly glycosylated isoforms (Figure 4.6B, compare **15-18**).

Changing the glycan at Ser3 to carbohydrates other than mannose shows fairly small changes to the retention time shift (Figure 4.6C, compare **19-27**). That is to say, any carbohydrate at this particular position on the peptide sequence seems to have a similar effect on the retention time, independent of its structure (Figure 4.6C, compare **9** to **19-25**). Anomeric stereochemistry also appears to have only a minor effect on the retention time of the glycoforms. Removing the side-chains of the critical Gln2 or Tyr5 amino acids, however, did significantly alter the retention time of the mannose-containing CBM glycopeptide. Mutating Gln2 to Ala lowers the retention time shift, bringing it back much closer to that observed for the unglycosylated control peptide (Figure 4.6C, compare **9** and **26**). Tyr5 side chain removal results in a much larger change in the retention time than the presence of a mannose alone does (Figure 4.6C, compare **9** and **27**).

4.3 Discussion

Protein glycosylation is one of the most prevalent post-translational modifications and previous work by many groups has shown that glycosylation can have large effects on the properties of glycoproteins in both glycan site- and glycan size-specific manners.[43,44] In addition, there is considerable evidence that the interplay between the glycan and local amino acids is necessary for the changes glycosylation can induce to occur.[45,46] While the most well-known type of extracellular *O*-glycosylation, mucin-type glycosylation, occurs in the Golgi after folding, most other types of *O*-glycosylation in mammals occur in the endoplasmic reticulum including *O*-mannosylation, *O*-fucosylation, and *O*-glucosylation. Given the strong effects of co-translational *N*-glycosylation on the folding dynamics of glycoproteins in the ER, and previous observations that *O*-glycosylation can greatly alter the biophysical properties of glycoproteins, we reasoned that *O*-glycosylation might also have a role in the oxidative folding of glycopeptides. In addition, the solubility of proteins during oxidative folding in the ER is thought have a large impact on protein maturation and many mechanisms exist in the secretory pathway to control the solubility of proteins during the process. Indeed, *O*-mannosylation has been previously implicated as one such mechanism to solubilize misfolded proteins in the ER of yeast.[18]

Thus, we were also interested in quantifying the site-specific and glycan-specific effects of *O*-glycosylation on the solubility of small, glycosylated peptide domains.

Since it has the potential to affect such a wide range of important properties, there has been a recent interest to use protein glycosylation as a tool for engineering glycoproteins with desirable physical and functional properties.[46,47] Critical to the successful implementation of such a strategy is a deep understanding of the relationships relating the glycoprotein structure to the physical properties and functions of the glycoprotein. Currently, this type of understanding is lacking in most glycoprotein systems and uncovering the necessary information is a time-consuming and expensive process. Thus, a quick and generally reliable way to assess the potential certain modifications would be a boon to the field by allowing for the expedited analysis of many more glycoproteins. During the course of this study we found that the thermostability of each of the glycoforms correlates with their retention time on UPLC. We propose that this can be taken advantage of to quickly steer future investigations into the structure-function relationships of glycoproteins.

It seems possible that retention time can be of some use in identifying which sites have the most potential for glycoengineering (Figure 4.7). Looking at solubility, both the Ser3 and Ser14 had the most significant effects. For disulfide bond formation, it seems that both Thr1 and Ser3 had the largest effects. And for secondary structure unfolding, glycosylation at Ser3 consistently had strongest effects, although the CBM glycoform **13**, which contains a di-mannose at Ser14 seems to be an exception to this. Together, this points at Ser3 as the single most important site to physically modify the peptide for the largest shifts in observable physical properties. Conveniently, glycosylation at Ser3 also gave

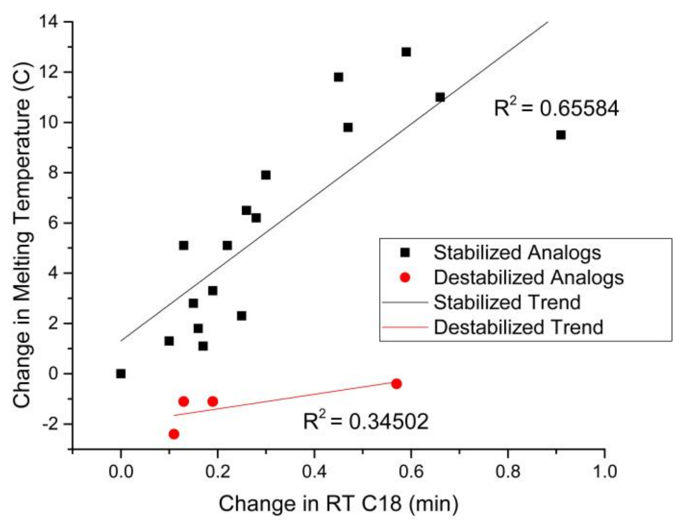


Figure 4.7 - Correlation of thermostability and retention time of the *Tr*Cel7A CBM. Data points represent averaged Tm data and retention time for each CBM glycoform. Lines represent the linear least squares fitting. For glycoforms with an increased melting temperature over **5**, there exists a positive linear correlation between retention time and Tm. Additionally, for glycoforms with a lowered thermostability, this same correlation is apparent, but divergent from the rest glycoforms.

by far the largest shifts to the retention time of the resulting glycopeptide. This suggests that screening a library of glycoforms by retention time first could save time down the road by narrowing the window of library members that need to be synthesized and characterized. Further work is currently being done in our lab to examine the scope of this observation and verify its usefulness in a wide variety of systems.

We also focused on quantifying the influence of glycosylation on the folding kinetics of the CBM peptide. While our attempts to monitor the formation of secondary structure after thermally unfolding the glycopeptides were unsuccessful, we were able to apply CD spectroscopy to the quantification of unfolding kinetics for each CBM glycoform in the library. From these data, it can be seen that glycosylation at either the Ser3 or Ser14 sites gave an increase in the kinetic stability of the CBM glycopeptide. When glycosylated at Ser3, the degree of stabilization increased with each increase in the size of the glycan, but at Ser14 the effect peaked upon addition of the second mannose and fell significantly upon extension of the glycan to the trimer. Interestingly, this kinetic stability followed a very similar site-specific pattern to the thermal stability of the CBM glycoforms studied previously. In that work, it was found that glycans at either Ser3 or Ser14 of the CBM peptide backbone significantly increased the melting temperature of the glycopeptide relative to the unglycosylated CBM peptide. This melting temperature is a quantitative measure of the thermostability of the secondary structure of the glycopeptide, much like the rate observed in the current study is a quantitative measure of the kinetic stability of the glycopeptides. The dependence of these two kinds of stability on the size of the glycans does appear to be different, however. We found that increasing the size of the glycan from one to two mannose residues increased the kinetic stability of CBM regardless of the site of that glycan. Earlier work with the thermostability of the CBM glycoforms examined here shows the opposite to be true at most of the possible sites. Only Ser3 glycosylation showed increasing thermostability when the glycan was extended to a dimer, however further extension to the trimer reversed that stability. It is thus noteworthy that large glycans can contribute positively to the kinetic stability while simultaneously decreasing the thermostability of glycopeptide secondary structure.

Also interesting is the comparison between disulfide bond formation kinetics and secondary structure kinetics. Under our experimental conditions, the formation of secondary structure for each of the glycoforms tested was too fast to observe. The stop-flow apparatus we tried to use for these measurements has a dead time of around 1.5 msec between solution mixing and starting the measurement. The fact that we could not observe folding means that it must have completed during that dead time. Tests with a lysozyme standard show that folding completed its initial phase in 70 msec, and given that lysozyme is a larger protein than the CBM glycopeptides tested here, this time scale does not seem unreasonable. This folding is very fast compared to the speed at which disulfide bond formation was observed to occur in this experiment where we saw formation of these bonds take place over the course of several hours.

Looking at the rate of disulfide bond formation and the rate of unfolding together, it is also fascinating to consider the biosynthesis of glycoproteins. The oxidative folding process, which is the combination of thiol oxidation to disulfide bonds and structure formation as a result of folding, takes place in the ER co-translationally. Initial *O*-mannosylation also takes place in the ER co-translationally. After maturation the glycoprotein is transferred to the Golgi where the short *O*-mannose glycans are extended and further *O*-glycosylation of the mucin type can take place. Given this biological context, it is particularly noteworthy that small glycans were found here to accelerate the rate of disulfide bond formation at specific sites while large glycans were found to kinetically stabilize fully folded protein structures. These *in vitro* studies hint at an as yet under explored role for *O*-mannosylation in the oxidative folding of disulfide-containing glycoproteins, and it seems entirely plausible that *O*-glycans are used to aid in the structural integrity of proteins in two distinct ways. Small glycans added in the ER during the oxidative folding process might act to accelerate correct folding of the peptide *in vivo*, as was observed here. Subsequent extension of these *O*-glycans to larger oligosaccharides in the Golgi would then help to stabilize the final folded structure. Separation of these two roles would be critical in such scenario since our data shows that large *O*-glycans impede the folding process.

The variations in solubility among the glycoforms tested are site-specific in nature. This means that increased solubility cannot be due simply to the fact that additional very hydrophilic structures (glycans) have been added to the peptide. Compare, for example **8** and **14**, they have identical glycan structures but very different solubility limits in our assay. It is possible that glycans at certain sites in the CBM sequence act to cover relatively hydrophobic patches on the surface of the peptide domain, shielding them from solvent exposure and contributing to the overall hydrophilicity of the glycopeptide and its solubility in aqueous solutions. This type of interaction has been observed in other systems, for example mammalian Notch1 carries several *O*-glucose glycans that are hypothesized to prevent cell-surface aggregation of the receptors by covering hydrophobic patches on the surface. Further work to uncover the details of how solubility is affected in such a site-specific manner is currently underway. Solubility is an important factor in the stability of any proteins and is intimately tied to the tendency to form aggregates under many conditions. In the ER this is particularly relevant as the unfolded nascent peptides often have a significant propensity to form aggregates and insoluble oligomers due to exposed hydrophobic amino acids, which are shielded in the protein's interior after folding. Numerous mechanisms are used by cells to prevent such aggregation, including many chaperones and glycosylation. Our work reveals that the solubility of glycoproteins is not controlled simply by the number of hydrophilic carbohydrate units attached to the peptide backbone, but also to a large extent by the specific sites occupied by those glycans.

4.5 Conclusion

By systematically investigating the effects of different glycans on the three glycosylation sites of a carbohydrate-binding module, our findings suggest a possible link between changes in different biophysical and biochemical properties and highlight the possible usefulness of glycoform retention times in reversed-phase high-performance liquid chromatography in the estimation of the range of glycosylation effects. Most interestingly however, we have shown that folding and kinetic stability of glycoproteins is closely correlated with the size and location of *O*-mannosylation sites in a manner consistent with the biosynthesis pathway or such molecules. We found that small *O*-mannose glycans, like those added in the

ER during the oxidative folding process, accelerate the rate of formation for disulfide bonds in the fungal-derived model system studied. We also found that large glycans, which are naturally synthesized in the Golgi after folding, hinder the oxidative folding reaction but kinetically stabilize the folded structure. Together these observations reveal the integrated nature of the co- and post-translational modification of proteins in living systems.

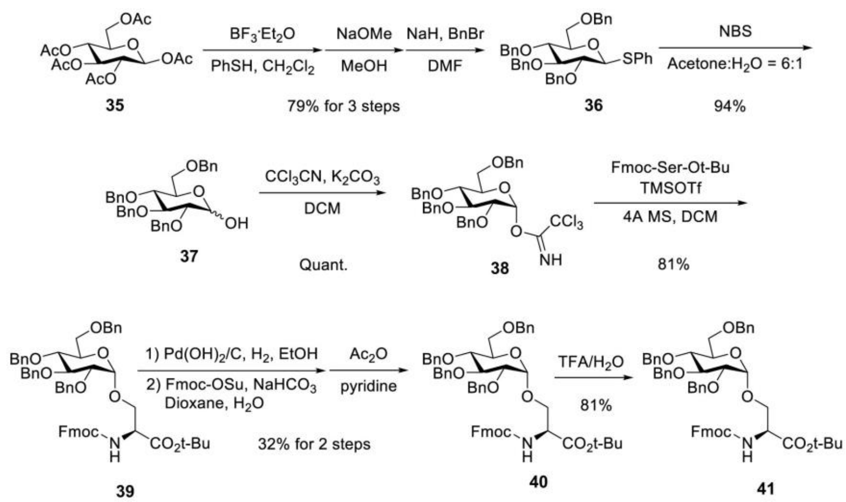
4.6 Experiments

4.6.1 Materials

All commercial reagents and solvents were used as received. Unless otherwise noted, all reactions and purifications were performed under air atmosphere at room temperature. All LC-MS analyses were performed using a Waters Acquity™ Ultra Performance LC system equipped with either Acquity UPLC® BEH 300 C4, 1.7μm, 2.1 x 100 mm or Acquity UPLC® BEH C18, 1.7μm, 2.1 x 100 mm columns at a flow rate of 0.3 mL/min. The mobile phase for LC-MS analysis was a mixture of H_2O (0.1% formic acid, v/v) and acetonitrile (0.1% formic acid, v/v). All preparative separations were performed using a LabAlliance HPLC solvent delivery system equipped with a Rainin UV-1 detector and a Varian Microsorb 100-5, C18 250x21.4 mm column at a flow rate of 16.0 mL/min. The mobile phase for HPLC purification was a mixture of H_2O (0.05% TFA, v/v) and acetonitrile (0.04% TFA, v/v). A Waters SYNAPT G2-S system was used mass spectrometric analysis. All circular dichroism (CD) spectra were obtained using an Applied PhotophysicsChirascan™-plus CD spectrometer.

4.6.2 Synthesis of glycoamino acids

Synthesis of glycoamino acid **41.** To a solution of β-D-glucose pentaacetate **35** (5.0 g, 12.8 mmol) in CH_2Cl_2 (50 ml) was added thiophenol (1.83 mL, 17.9 mmol). The resulting mixture was cooled to 0 °C. $BF_3 \cdot Et_2O$ (1.94 ml, 15.4 mmol) was then added dropwise and the reaction mixture was allowed to warm to room temperature. After being stirred at room temperature for 2 h, the mixture was diluted with CH_2Cl_2 (50 ml), washed with 1 M NaOH solution, H_2O, dried over $MgSO_4$ and concentrated under reduced pressure. The residue was dissolved in MeOH (50 ml) and MeONa (34.56 mg, 0.64 mmol) was added to the solution. The reaction was stirred at room temperature overnight, then neutralized with Amberlite IR-120 resin, filtered and concentrated. The resulting white foam was dissolved in



Scheme 4.1 - Chemical synthesis of glycoamino acid **41**.

DMF (48 mL), cooled to 0°C and NaH (2.56 g, 64 mmol, 60% in mineral oil) was added. The resulting mixture was allowed to warm slowly to room temperature while stirring over 20 minutes, then cooled to 0 °C and BnBr (27.8 mL, 235 mmol) was added slowly with stirring. The resulting mixture was stirred at room temperature overnight, then diluted with EtOAc, washed with H_2O, and concentrated under reduced pressure. The oily residue was purified by flash chromatography on a silica gel column (Hex/EtOAc = 20:1 to 10:1) to afford **36** (6.37 g, 79% over 3 steps) as a white solid[48].

To a solution of **36** (3.11 g, 4.91 mmol) in Acetone:H_2O = 6:1 (84 mL) was added N-bromosuccinimide (NBS) (2.62 g, 14.76 mmol) at room temperature. The resulting mixture was stirred at room temperature for 2 hours under argon before it was quenched with Na_2SO_3 (sat., aq.). The mixture was extracted with EtOAc and the organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on silica gel (Hex/EtOAc = 5:1 then CH_2Cl_2/MeOH = 20:1) to give **37** (2.49 g, 94%) as a white foam.

To a solution of **37** (620 mg, 1.15 mmol) in CH_2Cl_2 (4 mL) was added K_2CO_3 (39.45 mg, 0.29 mmol) at room temperature. To the resulting suspension was slowly added CCl_3CN (1.15 mL, 11.47 mmol) and the resulting mixture was stirred at room temperature for 12 hours. The mixture was filtered through Celite, washed with CH_2Cl_2, coevaporated with Hexanes 3 times, and concentrated under reduced pressure to give **38** (785.5 mg, quant.) as a clear oil.

To a solution of **38** (785.5 mg, 1.15 mmol) and Fmoc-Ser-Ot-Bu (483.6 mg, 1.26 mmol) in CH_2Cl_2 (15 ml) at room temperature was added 4A MS (800 mg) and the resulting mixture was stirred at room temperature for 20 minutes. The reaction was then cooled to -78 ^{o}C and TMSOTf (21 μL, 0.11 mmol) was added dropwise. The resulting mixture was stirred for 5 hours under argon while it slowly warmed to room temperature. The reaction was quenched with Et_3N (8 drops) and purified by flash chromatography on silica gel (Hex/EtOAc = 6:1 to 3:1) to give **39** (840 mg, 81%) as a white foam.

A solution of **39** (840 mg, 0.93 mmol) in EtOH (25.5 mL) was stirred with Pearlman's catalyst [$Pd(OH)_2$/C, 168 mg] and HCl (1 M, aq., 0.83 mL) under a hydrogen atmosphere at room temperature for 16 hours. The reaction was filtered through Celite and the filtrate was washed with EtOH 3 times and concentrated under reduced pressure. The residue was dissolved in pyridine (5 mL) and Ac_2O (1.75 mL) was added dropwise. The resulting mixture was stirred at room temperature under argon overnight. The mixture was poured into ice-water and extracted with EtOAc. The organic layer was dried over Na_2SO_4, filtered and the filtrate was concentrated under reduced pressure. The resulting oil was purified by flash chromatography on silica gel column (Hex/EtOAc = 2:1) to give **40** (211 mg, 32% for 2 steps) as a white foam. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.78 (d, *J* = 7.5 Hz, 2H), 7.66 (d, *J* = 7.5 Hz, 2H), 7.43 (t, *J* = 7.2 Hz, 2H), 7.33 (t, *J* = 7.2 Hz, 2H), 5.82 (d, *J* = 8.0 Hz, 1H), 5.46 (t, *J* = 9.8 Hz, 1H), 5.13 – 4.97 (m, 2H), 4.92 (dd, *J* = 10.2, 3.8 Hz, 1H), 4.44 (m, 2H), 4.34 – 4.17 (m, 2H), 4.12 (m, 1H), 4.08 – 3.90 (m, 3H), 2.15 – 2.00 (m, 12H), 1.52 (s, 9H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 170.6, 170.1, 170.0, 169.6, 168.6, 155.9, 143.8, 143.8, 141.3, 141.3, 127.8, 127.7, 127.1, 125.2, 125.1, 120.0, 120.0, 96.6, 82.9, 70.5, 69.9, 69.6, 68.4, 67.8, 67.2, 61.8, 54.8, 47.1, 28.0, 20.7, 20.7, 20.7, 20.6. IR ν/ cm^{-1}: 3359, 2978, 1750, 1368, 1224, 1039, 761, 741; HRMS (ESI^{+}) *m/z* Calc. for $C_{36}H_{42}NO_{14}Na$ [M + Na]$^{+}$: 736.2576, found 736.2573.

Compound **40** (184 mg, 0.26 mmol) was dissolved in a TFA-water mixture (95:5, 3 mL) and stirred at room temperature for 2 h. The solvent was evaporated and the residue was co-evaporated with toluene. The resulting white foam was purified by flash chromatography on silica gel column (CH_2Cl_2/MeOH = 30:1 to 10:1) to give **41** (138 mg, 81%) as a white foam. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.79 (d, *J* = 7.6 Hz, 2H), 7.65 (dd, *J* = 7.6, 3.6 Hz, 2H), 7.42 (t, *J* = 7.5 Hz, 2H), 7.34 (t, *J* = 7.5 Hz, 2H), 6.20 (d, *J* = 8.5 Hz, 1H), 5.54 (t, *J* = 9.8 Hz, 1H), 5.17 (d, *J* = 3.8 Hz, 1H), 5.05 (t, *J* = 9.8 Hz, 1H), 4.83 (dd, *J* = 10.2, 3.8 Hz, 1H), 4.65 (d, *J* = 8.2 Hz, 1H), 4.56 – 4.35 (m, 2H), 4.34 – 4.18 (m, 2H), 4.17 – 3.94 (m, 3H), 2.12 – 1.96 (m, 12H) *ppm*; ^{13}C NMR (101 MHz, $CDCl_3$) δ 170.7, 170.7, 170.6, 169.6, 156.1, 143.7, 141.3, 127.8, 127.1, 125.1, 125.1, 120.0, 96.6, 90.6, 70.9, 70.6, 70.2, 69.3, 68.7, 68.4, 67.7, 67.4, 65.5, 62.1, 61.8, 54.2, 47.1, 29.7, 20.8, 20.7, 20.6, 20.5 *ppm*. IR ν/ cm^{-1}: 3341, 3066, 2927, 1750, 1368, 1228, 1038, 740; HRMS (ESI^+) *m/z* Calc. for $C_{32}H_{34}NO_{14}Na$ $[M + Na]^+$: 680.1950, found 680.1948.

4.6.3 Assays

*Rate of Disulfide Bond Formation for Unglycosylated CBM Variant **5**:* 8 mg of the crude peptide was dissolved in 40 mL with folding buffer (0.2 M Tris-acetate, 4 mM reduced glutathione, pH 8.2) and stirred at room temperature for 24 h under a helium atmosphere. At certain time intervals after the addition of folding buffer (5 min, 15 min, 30 min, 60 min, 2 hr, 3 hr, 4 hr, 5 hr, 6 hr, 7 hr, 8 hr, 12 hr, and 24 hr), 40 uL aliquots of the folding reaction were removed and quenched by adding the aliquots directly to 136 uL of a MeCN/H_2O/TFA solution (3 mL MeCN, 17 mL H_2O, 100 uL TFA) in an LC-MS vial and briefly mixing. Quenched aliquots were stored at 4 ^oC for not more than 24 hours before LC-MS analysis. LC-MS analysis was done by injecting 2 uL of sample into a C4 UPLC column and eluting with a linear gradient of 15→35% MeCN in H_2O over 5 min at a flow rate of 0.3 mL/min. LC-MS traces were analyzed by combining ESI+ MS spectra where the CBM variant was detected. The $[M+3H]^{+3}$ peak was chosen for average mass calculation because it was the highest intensity peak. The weight average mass (WAM) was calculated for the CBM at each time point and then the difference between the WAM at each time point and the completely reduced WAM was calculated. Completely reduced WAM was calculated from LC-MS traces taken of samples of crude CBM peptide dissolved in TCEP solution (5 mg TCEP in 250 uL of quench solution). The change in WAM over time was plotted (Figure 4.8) and the slope of the line-of-best-fit for the linear portion of the plot was taken as the initial rate of disulfide bond formation.

*Rate of Disulfide Bond Formation for Glycosylated CBM Variants **6-27**:* 8 mg of the crude peptide was dissolved in 0.5 ml of hydrazine solution (hydrazine/H_2O, 5/100, v/v) and stirred at room temperature for 30 min under helium. The reaction was quenched with 1 ml of acetic acid solution (AcOH/H_2O, 5/100, v/v). The resulting mixture was diluted to 40 mL with folding buffer (0.2 M Tris-

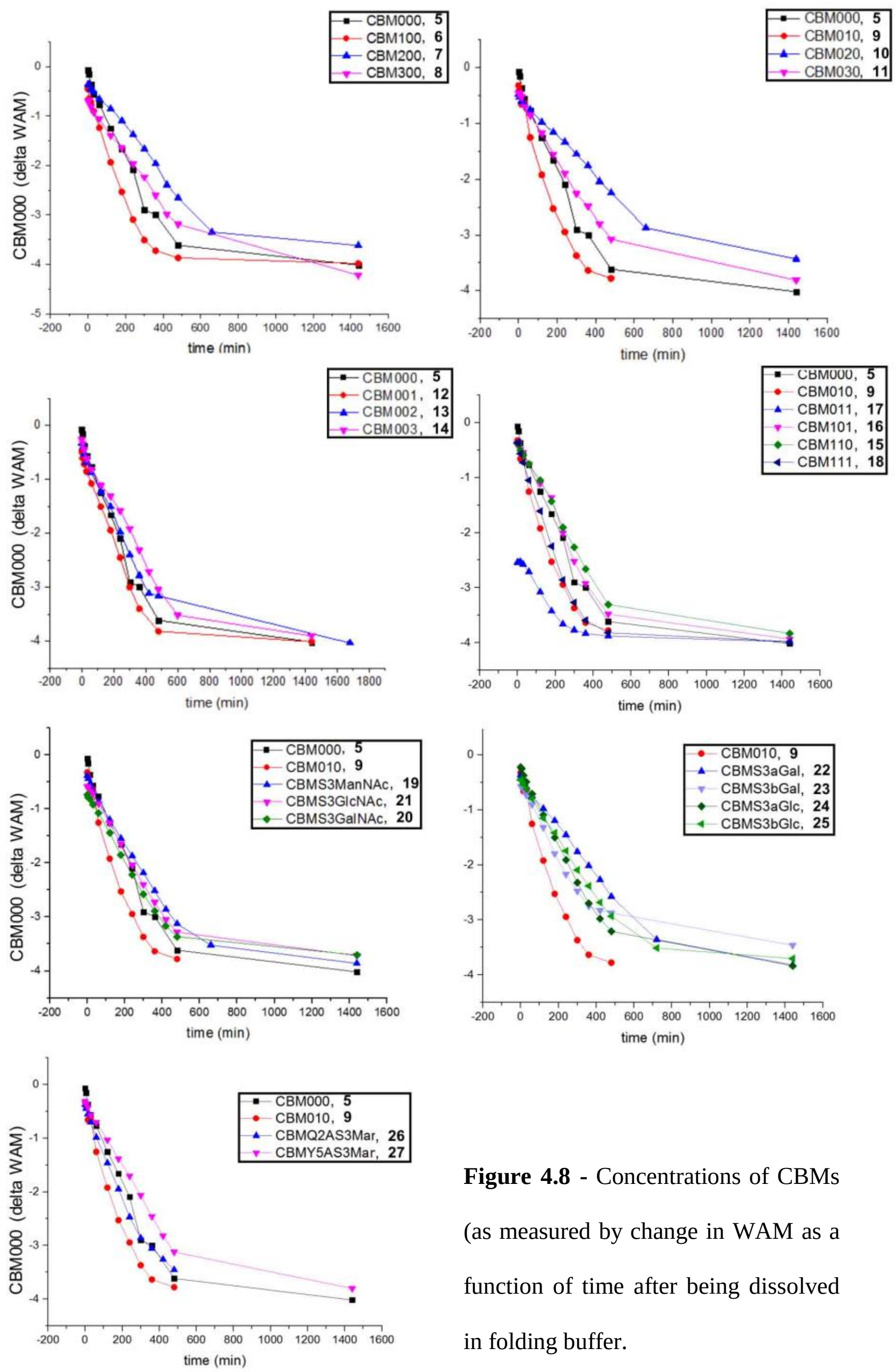


Figure 4.8 - Concentrations of CBMs (as measured by change in WAM as a function of time after being dissolved in folding buffer.

acetate, 4 mM oxidized glutathione, pH 8.2, 40 ml) and stirred at room temperature for 24 h under a helium atmosphere. At regular time intervals, 40 uL aliquots of the folding reaction were removed and quenched as described above for unglycosylated CBM variants. LC-MS conditions and data analysis was done as described above for unglycosylated CBM variants.

Secondary Structure Unfolding Kinetics: For each CBM variant, 20 uL of a 4 mg/mL stock solution of glycopeptide in buffer (10 mM NaOAc, pH = 5.0) was added to a 10mm by 10mm CD cuvette with a stirbar at room temperature. The cuvette holder temperature was raised to 80 ^oC, 2 mL of 80 ^oC buffer (10 mM NaOAc, pH = 5.0) was added to the sample and as quickly as possible the cuvette with the sample was placed into the cuvette holder and data collection was commenced. Circular diochroism (CD) of the sample was measured at 217 nm every 0.5 seconds for 600 seconds. The sample was then cooled to 20 ^oC and a CD spectrum was taken from 200 nm to 260 nm to confirm refolding had taken place. The concentration of folded CBM in the sample at each time point was calculated based on the CD measurements taken of completely unfolded and completely folded samples for each CBM variant, and the concentration of folded CBM as a function of time was plotted. The slope of the line-of-best-fit for the linear portion of this plot was taken as the initial rate of secondary structure unfolding. This was done in triplicate for each CBM variant and the resulting rates were averaged.

Limit of Solubility: For each CBM variant, to 1 mg of purified peptide was added 5 uL of buffer (50 mM NaOAc, 50 mM NaCl, pH = 5.0) at 4 ^oC. The resulting mixture was allowed to equilibrate at 4 ^oC for 20 minutes before being pelleted. Two aliquots of 0.5 uL of supernatant were removed from the sample, diluted with buffer and the concentration of glycopeptide in each was measured using our previously developed mass spectroscopy based method (see Chapter 2). This was done in triplicate for each CBM variant and the resulting rates were averaged.

4.7 References

1. C. Xu and D. T. Ng, *Nat Rev Mol Cell Biol*, 2015, **16**, 742-752.
2. I. Braakman and D. N. Hebert, *Cold Spring Harb Perspect Biol*, 2013, **5**, a013201.

3. J. L. Arolas, F. X. Aviles, J. Y. Chang and S. Ventura, *Trends Biochem Sci*, 2006, **31**, 292-301.
4. O. B. Oka, H. Y. Yeoh and N. J. Bulleid, *Biochem J*, 2015, **469**, 279-288.
5. M. Loibl and S. Strahl, *Biochim Biophys Acta-Mol Cell Res*, 2013, **1833**, 2438-2446.
6. S. H. Stalnaker, R. Stuart and L. Wells, *Curr Opin Struct Biol*, 2011, **21**, 603-609.
7. M. Lommel and S. Strahl, *Glycobiology*, 2009, **19**, 816-828.
8. L. Wells, *J Biol Chem*, 2013, **288**, 6930-6935.
9. M. Goto, *Biosci Biotechnol Biochem*, 2007, **71**, 1415-1427.
10. Y. Jitsuhara, T. Toyoda, T. Itai and H. Yamaguchi, *J Biochem*, 2002, **132**, 803-811.
11. J. L. Price, D. Shental-Bechor, A. Dhar, M. J. Turner, E. T. Powers, M. Gruebele, Y. Levy and J. W. Kelly, *J Am Chem Soc*, 2010, **132**, 15359-15367.
12. D. Shental-Bechor and Y. Levy, *Curr Opin Struct Biol*, 2009, **19**, 524-533.
13. D. Shental-Bechor and Y. Levy, *Proc Natl Acad Sci U S A*, 2008, **105**, 8256-8261.
14. S. R. Hanson, E. K. Culyba, T. L. Hsu, C. H. Wong, J. W. Kelly and E. T. Powers, *Proc Natl Acad Sci U S A*, 2009, **106**, 3131-3136.
15. K. Nakatsukasa, S. Okada, K. Umebayashi, R. Fukuda, S. Nishikawa and T. Endo, *J Biol Chem*, 2004, **279**, 49762-49772.
16. C. Harty, S. Strahl and K. Romisch, *Mol Biol Cell*, 2001, **12**, 1093-1101.
17. V. Goder and A. Melero, *J Cell Sci*, 2011, **124**, 144-153.
18. C. Xu, S. Wang, G. Thibault and D. T. Ng, *Science*, 2013, **340**, 978-981.
19. C. B. Taylor, M. F. Talib, C. McCabe, L. Bu, W. S. Adney, M. E. Himmel, M. F. Crowley and G. T. Beckham, *J Biol Chem*, 2012, **287**, 3147-3155.
20. L. Chen, M. R. Drake, M. G. Resch, E. R. Greene, M. E. Himmel, P. K. Chaffey, G. T. Beckham and Z. Tan, *Proc Natl Acad Sci U S A*, 2014, **111**, 7612-7617.
21. X. Guan, P. K. Chaffey, C. Zeng, E. R. Greene, L. Chen, M. R. Drake, C. Chen, A. Groobman, M. G. Resch, M. E. Himmel, G. T. Beckham and Z. Tan, *Chem Sci*, 2015, **6**, 7185-7189.
22. R. M. Happs, X. Guan, M. G. Resch, M. F. Davis, G. T. Beckham, Z. Tan and M. F. Crowley, *FEBS J*, 2015, **282**, 4341-4356.
23. N. Ohyabu, H. Hinou, T. Matsushita, R. Izumi, H. Shimizu, K. Kawamoto, Y. Numata, H. Togame, H. Takemoto, H. Kondo and S. Nishimura, *J Am Chem Soc*, 2009, **131**, 17102-17109.
24. M. Pudelko, J. Bull and H. Kunz, *Chembiochem*, 2010, **11**, 904-930.
25. H. Malekan, G. Fung, V. Thon, Z. Khedri, H. Yu, J. Qu, Y. Li, L. Ding, K. S. Lam and X. Chen, *Bioorg Med Chem*, 2013, **21**, 4778-4785.
26. K. Yoshida, B. Yang, W. Yang, Z. Zhang, J. Zhang and X. Huang, *Angew Chem Int Ed Engl*, 2014, **53**, 9051-9058.
27. J. P. Giddens and L. X. Wang, *Methods Mol Biol*, 2015, **1321**, 375-387.
28. H. Yu, K. Lau, Y. Li, G. Sugiarto and X. Chen, *Curr Protoc Chem Biol*, 2012, **4**, 233-247.
29. L. Li, Y. Liu, C. Ma, J. Qu, A. D. Calderon, B. Wu, N. Wei, X. Wang, Y. Guo, Z. Xiao, J. Song, G. Sugiarto, Y. Li, H. Yu, X. Chen and P. G. Wang, *Chem Sci*, 2015, **6**, 5652-5661.
30. R. Chen and T. J. Tolbert, *J Am Chem Soc*, 2010, **132**, 3211-3216.
31. A. Fernandez-Tejada, J. Brailsford, Q. Zhang, J. H. Shieh, M. A. Moore and S. J. Danishefsky, *Top Curr Chem*, 2015, **362**, 1-26.
32. H. C. Hang and C. R. Bertozzi, *Bioorg Med Chem*, 2005, **13**, 5021-5034.
33. T. Buskas, S. Ingale and G. J. Boons, *Glycobiology*, 2006, **16**, 113R-136R.
34. K. M. Koeller and C. H. Wong, *Nat Biotechnol*, 2000, **18**, 835-841.

35. L. Chen, M. R. Drake, M. G. Resch, E. R. Greene, M. E. Himmel, P. K. Chaffey, G. T. Beckham and Z. Tan, *Proc. Natl. Acad. Sci. U S A*, 2014, **111**, 7612-7617.
36. S. Chandrasekhar, B. S. Moorthy, R. Xie and E. M. Topp, *Pharm Res*, 2016, DOI: 10.1007/s11095-016-1879-3.
37. S. Chandrasekhar and E. M. Topp, *J Pharm Sci*, 2015, **104**, 1291-1302.
38. A. W. Barb, A. J. Borgert, M. Liu, G. Barany and D. Live, *Methods Enzymol*, 2010, **478**, 365-388.
39. N. J. Greenfield, *Nat Protoc*, 2006, **1**, 2527-2535.
40. F. Wang, L. Qin, C. J. Pace, P. Wong, R. Malonis and J. Gao, *Chembiochem*, 2012, **13**, 51-55.
41. R. M. Kramer, V. R. Shende, N. Motl, C. N. Pace and J. M. Scholtz, *Biophys J*, 2012, **102**, 1907-1915.
42. C. T. Mant, N. E. Zhou and R. S. Hodges, *J Chromatogr*, 1989, **476**, 363-375.
43. A. Varki, *Glycobiology*, 1993, **3**, 97-130.
44. D. N. Hebert, L. Lamriben, E. T. Powers and J. W. Kelly, *Nat Chem Biol*, 2014, **10**, 902-910.
45. J. L. Price, E. K. Culyba, W. Chen, A. N. Murray, S. R. Hanson, C. H. Wong, E. T. Powers and J. W. Kelly, *Biopolymers*, 2012, **98**, 195-211.
46. E. R. Greene, M. E. Himmel, G. T. Beckham and Z. Tan, *Adv Carbohydr Chem Biochem*, 2015, **72**, 63-112.
47. A. M. Sinclair and S. Elliott, *J Pharm Sci*, 2005, **94**, 1626-1635.
48. J. Ohlsson and G. Magnusson, *Carbohydr Res*, 2000, **329**, 49-55.

Chapter 5

Effects of *O*-Glycosylation on the Substrate Binding Specificity of a Cellulose Binding Module

5.1 - Introduction

The solar energy captured by plants through photosynthesis has the potential to provide a large portion of the world's transportation fuel requirements. Within plants, this energy is stored in the polymers of the highly stable cell-wall complex. Efficiently converting these natural polymers, often collectively termed lignocellulosic biomass, to more convenient energy-storage compounds, such as ethanol, for use in existing transportation infrastructure is the major goal of the biofuels industry.[1] Lignocellulosic biomass is mainly composed of three different polymers: cellulose, hemicellulose and lignin.[2] The most abundant component is cellulose, which is formed from long chains of glucose units.[3,4] Hemicellulose is the next most abundant component of biomass and is also a polysaccharide, but is composed of several different 5- and 6-carbon monosaccharide building blocks including glucose, xylose and mannose. In addition, hemicellulose has a branched structure and forms numerous covalent cross-links to other cell-wall components, which provide much of the physical rigidity to the complex as a whole.[5] Lignin is the final component and is constructed from a wide variety of aromatic phenylpropanoid monomers linked through chemically stable ether bonds. These three structural polymers are organized into macro-scale fibrils that bundle together and give strength and rigidity to the cell-wall complex (see Figure 5.1).[6]

Biochemical conversion of lignocellulosic biomass to fuels involves several steps. There is first a thermochemical decomposition step, often called pre-treatment, that is aimed at breaking apart the macroscopic structure of the cell wall and improves accessibility of the cellulose during subsequent steps. There are many methods currently being explored for this step including treatment of biomass with sulfuric acid, ammonia, lime or water at temperatures between 100 ºC and 200 ºC.[1,6] After pretreatment, the biomass is exposed to a cocktail of synergistic enzymes designed to rapidly hydrolyze the cellulose components. These cocktails contain several types of enzymes, including exoglucanases, endoglucanases and β-glucosidases, which act collectively to depolymerize cellulose.[7,8]

Lignin is a large problem for the enzymatic hydrolysis of biomass. Along with hemicellulose, it is a critical component of the cell-wall complex that physically obstructs the cellulose.[9] Breaking apart the physical barriers formed by lignin is an important function of any pretreatment step.[10] However, total separation of lignin and cellulose components of biomass is not currently feasible prior to enzymatic degradation, and cellulases tend to bind the residual lignin in the samples.[9] This is problematic because it can sequester the enzymes in a non-productive place away from the cellulose, and can lead to significant problems for recycling hydrolytic enzymes.[11,12]

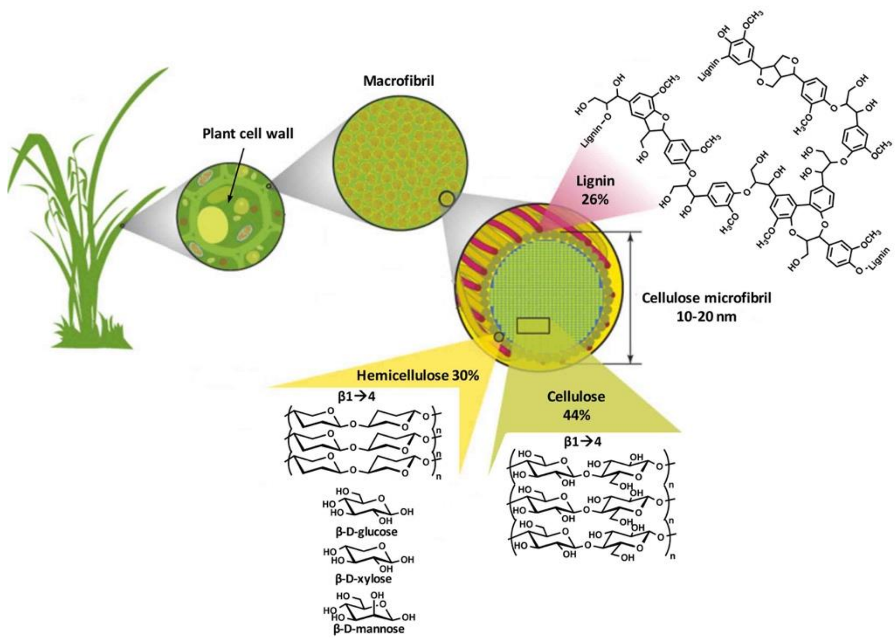


Figure 5.1 - Structure of lignocellulose. Cellulose, hemicellulose and lignin form structures called microfibrils, which are organized into macrofibrils that mediate structural stability in the plant cell wall.

Several solutions to this problem have been explored and include attempts change the CBM binding surface, lignin chemistry or processing conditions. Genetic approaches have also been explored, for example, it might possible to genetically engineer a plant to produce a minimal amount of lignin or predominantly lignin that won’t cause problems for the enzymatic hydrolysis.[1,6] Lignin chemistry is known to vary based on its source, and characterization of lignin isolated from new potential biomass sources might reveal crops with more amenable lignin.[11,12] This approach would also identify crops with the most potential for further genetic engineering efforts.[6,13] Various pretreatments are known to alter the lignin chemistry and this has also been explored as a way to decrease the interference of lignin in the process.[11,12,14] Even without a drastic effect on the structure of the biomass components, altering the hydrolysis conditions have been proposed to help. For example, a slight increase in pH was found to decrease lignin-cellulase interactions; which was attributed to additional charges introduced on both enzyme and lignin as the pH raised.[15] Cellulose in contrast, has almost no pH-dependent shift in charge and so binding of cellulase to cellulose substrates was near constant across the same range of pH.[11,12] The use of surfactants or sacrificial proteins has also been explored and shown to have some amount of benefit, most likely due to blocking the non-productive binding of lignin by cellulases.[16,17]

Similarly, the structure of the cellulose substrate is thought to play a role in the rate of enzymatic hydrolysis. Crystalline cellulose has several different structural allomorphs, all hydrolyzed at different rates by cellulases, and furthermore the crystal structure is known to change as a result of certain pretreatment conditions.[18] Although the exact details and limits of these changes are not yet known, this suggests that reaction conditions could be optimized to produce the most desirable cellulose crystal structure. Additionally, cellulose can exist across a wide range of organizational states from highly crystalline to mostly amorphous. This can be quantified through several different spectrographic methods including X-ray diffraction (XRD), solid-state ^{13}C-NMR, and infrared or Raman spectroscopy. Most often, the degree of crystal organization in a sample is expressed as a crystallinity index (CI) value.[19] Previous studies have shown that the measured crystallinity index strongly correlates with the initial hydrolysis rate of cellulose sample, which has been attributed to both increased binding affinity of cellulases towards less crystalline cellulose[20] and intrinsic differences in the reactivity of different cellulose structures towards hydrolysis.[19] Furthermore, CBMs from different cellulases are known to preferentially bind, and thus target their tethered catalytic domains towards, either crystalline or amorphous cellulose structures.[21] For example, Type A CBMs are known to favor crystalline cellulose substrates while Type B CBMs prefer amorphous ones.[21]

Cellulase-focused protein engineering approaches to the problem have also been explored. Since, it has been established that the cellulose binding module (CBM) domain is most responsible for the unproductive lignin binding,[11,14] it is a natural place to begin such protein engineering efforts. Amino acid mutations on the binding surface of the CBM have shown that increased hydrophobicity of the face correlates with increased binding to both cellulose[22] and lignin.[15] Additionally, the CBM-cellulose interaction is thought to occur mainly as a result of stacking between the aromatic tyrosine side-chains on the CBM face and the hydrophobic crystal face of cellulose, while very similar pi-pi stacking between the aromatic rings of lignin and those same tyrosine side-chains are the main factor driving the CBM-lignin interaction. Furthermore, lignin binding has been shown to have a direct correlation with the computationally calculated hydrophobic surface patch score for the entire cellulase enzyme.[14,23] Thus, both binding events seem to be driven by very similar hydrophobic interactions between particular aromatic residues on the CBM binding face and the substrate surface. These studies suggest that amino acid mutations alone are unlikely to lead to helpful changes.

On the other hand, binding affinity for lignin varies across different naturally occurring cellulases, which suggests that increased efficiency can be achieved through protein engineering.[9] Furthermore, the fact that many of these enzymes have similar hydrolytic activity towards cellulose in the absence of lignin shows that it is possible to decrease lignin binding without detrimental effects on the desired enzymatic activity.[9] A detailed understanding of how different sequences or post-translational modifications lead to altered lignin binding properties might open the door to engineered cellulases that are even more resistant to lignin binding.

We chose to investigate the effects of CBM glycosylation on CBM binding affinity for a variety of lignocellulosic-derived polymers. We chose to investigate several different kinds of cellulose that had been isolated and treated differently to yield varying degrees of crystallinity. This was to investigate how glycosylation of the CBM effects binding to cellulose substrates across a range of crystallinity values. We also investigated several lignin substrates in order to find out how different CBM glycoforms bind to lignin. Ultimately, we wished to compare each CBM glycoform's binding affinity to lignin and cellulose substrates in order to quantify the effects of glycan structure on the cellulose specificity of CBM binding.

To carry out this study, we synthesized a library of glycosylated CBMs carrying unique glycosylation patterns. For each member of the library, binding affinity towards each substrate was quantified by our previously developed and validated mass spectroscopy based binding assay.[24,25] We found that the glycosylation pattern of the CBM domain had a strong influence on the binding affinity towards both lignin and cellulose. Furthermore, specific glycoforms were identified that displayed simultaneous increases in cellulose binding affinity and decreases in lignin binding affinity. This work shows the glycoengineering of cellulose binding domains has the potential to both increase the hydrolytic efficiency of the enzymes towards cellulose and decrease the problematic, non-productive binding of lignin during the saccharification process.

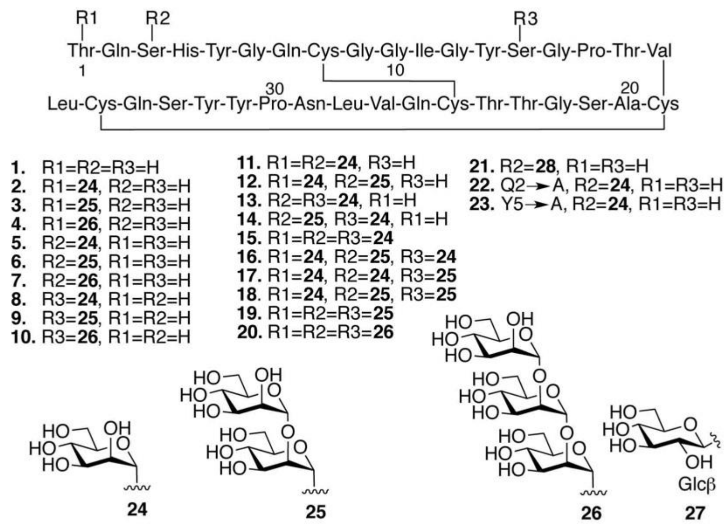


Figure 5.2 - Synthetic CBM varaints that are used in this study.

5.2 - Results

We began by synthesizing 23 CBM glycoforms using our previously developed one-pot synthesis and folding procedure (Figure 5.2). Chemical synthesis was chosen because it is a very efficient method for the synthesis of homogeneous glycoforms and allows for almost complete control over all aspects of the glycosylation pattern including stereochemistry and glycan site occupancy which are difficult to control properly in enzymatic synthesis methods. The carbohydrate-amino acid conjugate building blocks were synthesized according to literature procedures.[24,25]

We tested three cellulose substrates and two lignin substrates in this study. In order to make the results of the study as relevant to real-world conditions as possible, two of these substrates were generated from corn stover through the commonly used clean fractionation (CF) pretreatment process.[26] By means of a mixture of organic solvents, water and sulfuric acid catalysis, CF pretreatment separates biomass into three fractions: an organic fraction enriched in lignin, an aqueous fraction enriched in hemicellulose, and an insoluble fraction enriched in cellulose. Both the lignin-enriched and cellulose-enriched fractions were used in this study. Phosphoric acid-swollen cellulose (PASC) generated from cotton linen was also studied since it is known to have a significantly more amorphous structure than crystalline cellulose substrates.[19] Finally, two commercially available substrates were chosen: purified crystalline Avicel cellulose and purified Kraft lignin.

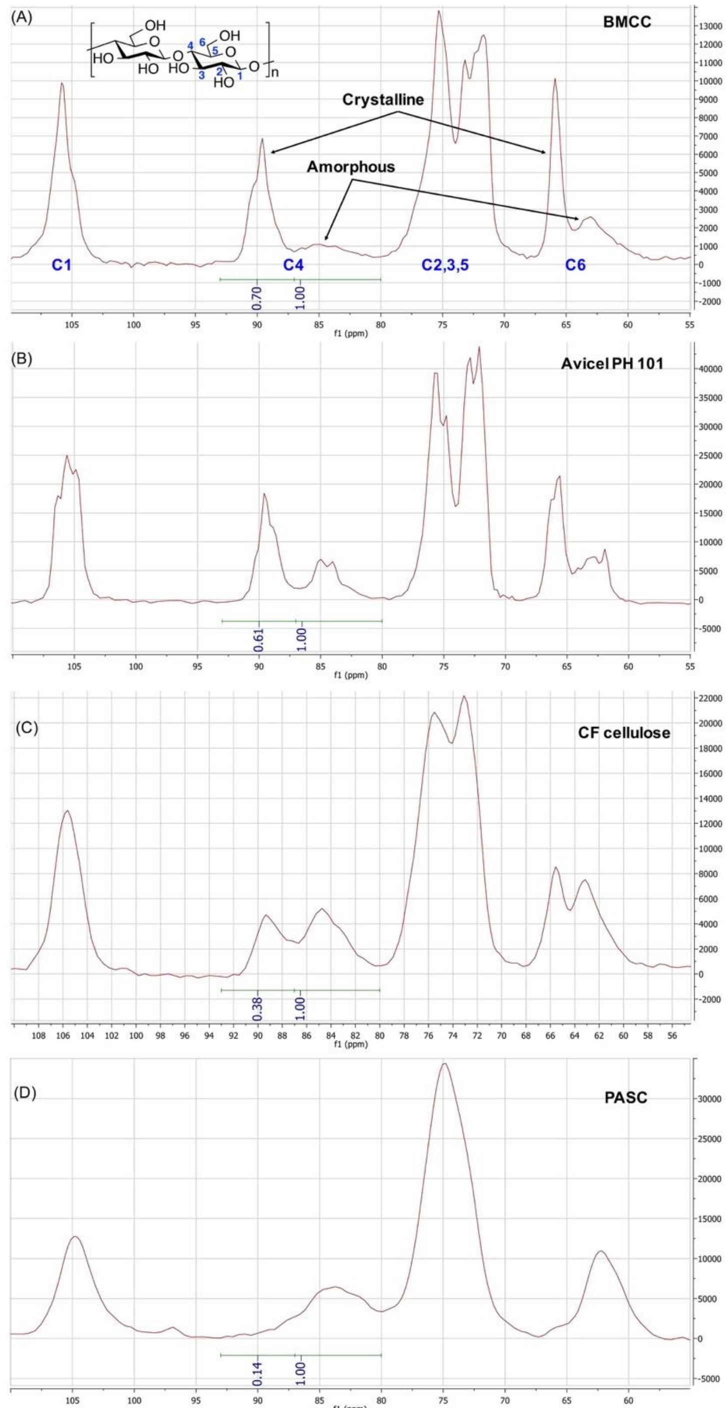


Figure 5.3 - Solid 13C NMR spectra of different types of cellulose. Based on the integration of the crystalline and amorphous C4 peaks, the CI was determined to be 70% (BMCC), 61% (Avicel PH 101), 38% (CF cellulose) and 14% (PASC).

Solid-state ^{13}C-NMR was used to

quantify the relative amounts of crystalline and amorphous structure in each of the four cellulose substrates used in this study. A crystallinity index (CI) was then calculated for each substrate by close examination of C4 signal, which is split into two slightly overlapping signals.[19] The peak at 89 ppm is taken as the crystalline cellulose and the peak at 84 ppm is the amorphous cellulose structure. CI is taken as ratio of the area under the peak corresponding to the crystalline C4 signal to the total C4 signal (Figure 5.3). As expected, the phosphoric-acid swollen cellulose (PASC) had the lowest CI (Figure 5.3D) and hence was the least crystalline or most amorphous of the samples while the bacterial microcrystalline cellulose (BMCC) was the most crystalline of the samples (Figure 5.3A).[19]

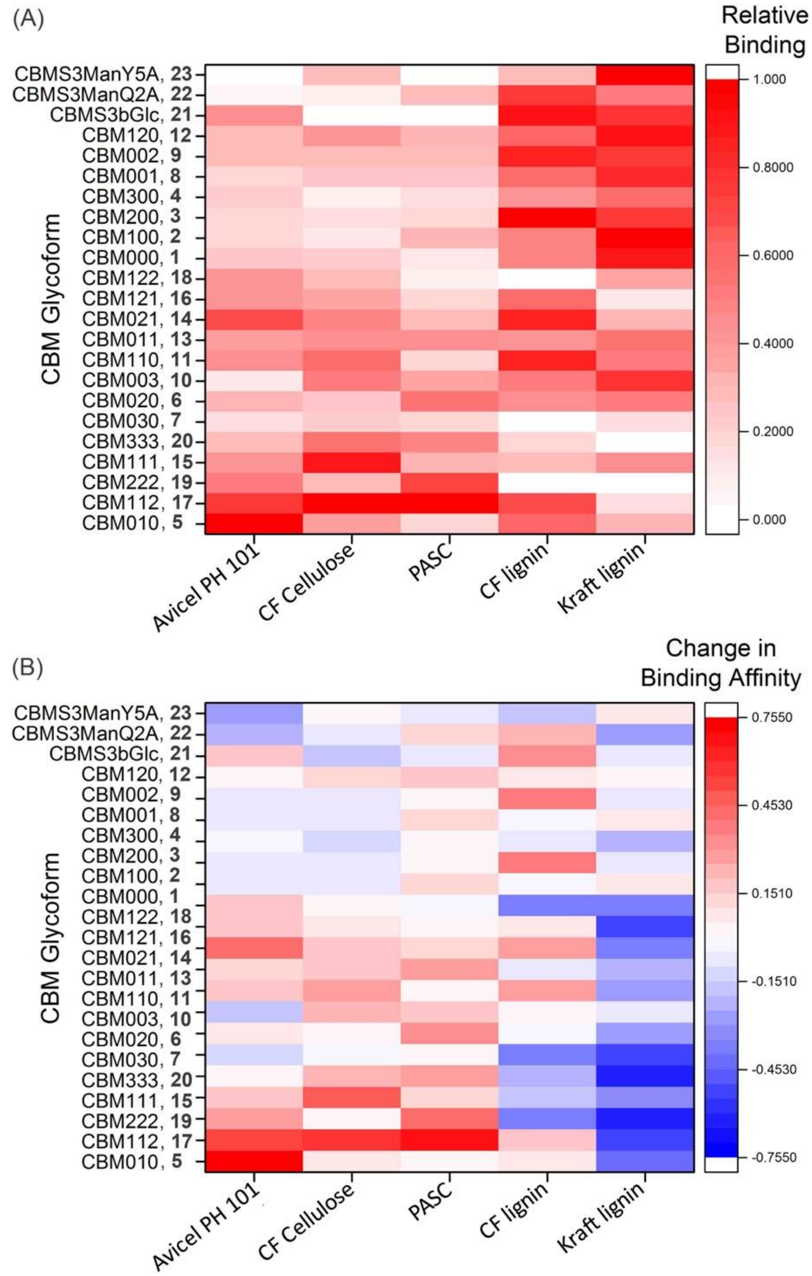


Figure 5.4 - Binding affinity of each glycoform towards different substrates (A) and changes in binding affinities caused by glycosylation (B).

We next tested the binding affinity of each glycoform towards each substrate. In total we quantified 115 individual binding affinities using our previously developed method based on mass spectroscopy.[24,25] The results from this study are summarized in Figure 5.4A. Since the absolute binding affinities are difficult to compare across substrates, each individual value was normalized to a substrate-specific scale where the lowest affinity measured for that particular substrate was assigned to 0, the highest assigned to 1, and all values in between expressed relative to those two end points. The glycoforms can be roughly divided into three groups: those that bind strongly to lignin and weakly to cellulose, those that bind with very similar affinities to both lignin and cellulose and those that bind strongly to cellulose and weakly to lignin.

To further explain the differences in binding caused by glycosylation, the change in binding affinity, relative to the unglycosylated CBM control peptide, was also calculated (Figure 5.4B). This data shows that differences in binding affinity were both positive and negative in direction depending on the exact glycoform and substrate being examined. For example, relative to CBM **1**, CBM **5** bound PASC, and both cellulose- and lignin-enriched corn stover fractions equally well, but it showed a large increase in binding affinity towards Avicel cellulose and large decrease in binding affinity towards Kraft lignin. Other glycoforms, like CBM **8**, displayed little change in binding affinity towards any of the substrates as compared to the unglycosylated control.

5.3 - Discussion

As shown in Figure 5.4A, from comparing the relative binding affinities towards each of the three cellulose substrates, we did not observe an ability to discriminate amongst the different types of cellulose for most of the glycoforms in this study. In other words, glycoforms that weakly bound one type of cellulose, tended to bind the other cellulose substrates poorly as well. Similarly, most of the CBM glycoforms we examined had similar binding affinities for both varieties of lignin. This finding supports previous work that attributes the well-known increase in hydrolysis rate of amorphous cellulose to an increase in the reactivity of the substrate rather than an increased ability for cellulases to bind the unstructured substrates.[19] We did, however, observe several exceptions to this, most notably CBM **5**, which bound everything poorly except the Avicel cellulose. Previous work has shown that glycosylation of Ser3, as is the case for CBM **5** significantly increased CBM binding affinity towards BMCC.[24,25]

Comparing the CI values calculated in this study for BMCC, Avicel, CF cellulose and PASC, BMCC and Avicel are far more crystalline than either of the other two cellulose forms (Figure 5.3). These data suggest that glycosylation of the CBM peptide sequence at Ser3 significantly increases its binding affinity toward crystalline cellulose substrates while having little effect on disordered cellulose binding.

From Figure 5.4B, it can be seen that many of the CBM glycoforms studied show only modest changes in cellulose binding affinity from the unglycosylated CBM peptide. Three notable exceptions to this are CBM **17**, **15** and **5**. CBM **17** displayed relatively large increases in binding affinity towards all three of the cellulose substrates tested here. CBM **5**, on the other hand, showed no increase in affinity towards CF cellulose or PASC but the largest increase measured in this study towards Avicel cellulose. CBM **15** was found to bind much better than unglycosylated CBM to CF cellulose, and bound only marginally better to the other two cellulose substrates. Binding to lignin, however, was decreased in almost half of the CBM glycoforms studied, particularly for the Kraft lignin substrate. This suggests that a decrease in lignin binding is a more general consequence of CBM *O*-glycosylation than an increase in binding affinity towards cellulose.

For the commercial use of cellulases for biomass hydrolysis, a strong binding affinity to cellulose combined with as little propensity for lignin binding as possible is highly desired.[9,12] While many glycoforms we characterized showed a decreased lignin binding capacity, only a select few glycosylation patterns resulted in increased cellulose binding. Thus, the data collected in this study point to small glycans distributed across all available sites as the most beneficial glycosylation pattern for commercial cellulase CBMs. In particular, CBM **17** and **15** coupled significant increases in binding affinity towards cellulose substrates with a decreased tendency to bind lignin. Larger glycans at each site are also helpful in that CBM glycoforms with such glycosylation patterns (CBM **19** and **20**) show much lower affinities towards lignin than the unglycosylated peptide, but the increases in cellulose binding affinity observed for these glycoforms were small. A single mannose at the Ser3 site could also be helpful as this resulted in a large increase in binding affinity, but only for highly crystalline cellulose substrates.

Our results highlight the effect glycosylation can have on CBM binding preferences across many different biomass-derived substrates. We have shown that specific patterns of *O*-glycosylation can lead to simultaneous increases in cellulose binding and deceases in affinity for lignin. In addition, most glycoforms studied here bound ordered and disordered cellulose equal well, meaning most of the time glycosylation will not cause an increased preference for a specific type of cellulose over another. The exception to this seems to be mono-mannosylation at Ser3, which resulted in a large increase in the binding affinity of only very crystalline cellulose substrates. Lignin is a ubiquitous problem during the conversion of biomass to biofuels, largely due to its ability to bind and sequester cellulases.[9,12] A widely applicable and generally reliable way to prevent or decrease lignin binding by cellulases would be a welcome advance. Our results suggest that *O*-glycosylation could be a way to significantly reduce enzyme-lignin interactions while simultaneously increasing the enzymes' binding affinity towards cellulose.

5.4 Experiments

5.4.1 Materials

All commercial reagents and solvents were used as received. Unless otherwise noted, all reactions and purifications were performed under air atmosphere at room temperature. All LC-MS analyses were performed using a Waters Acquity™ Ultra Performance LC system equipped with Acquity UPLC® BEH 300 C4, 1.7μm, 2.1 x 100 mm column at flow rates of 0.3 and 0.5 mL/min. The mobile phase for LC-MS analysis was a mixture of H_2O (0.1% formic acid, v/v) and acetonitrile (0.1% formic acid, v/v). All preparative separations were performed using a LabAlliance HPLC solvent delivery system equipped with a Rainin UV-1 detector and a Varian Microsorb 100-5, C18 250x21.4 mm column at a flow rate of 16.0 mL/min. The mobile phase for HPLC purification was a mixture of H_2O (0.05% TFA, v/v) and acetonitrile (0.04% TFA, v/v). A Waters SYNAPT G2-S system was used mass spectrometric analysis.

Solid-state ^{13}C-NMR was done on a Varian INOVA 400 MHz NMR instrument equipped with a 4 mm cross-polarization magic angle spinning (CP-MAS) probe.

5.4.2 Synthesis of the glycoamino acid building blocks

The glycoamino acid building blocks Fmoc-Ser(Ac$_4$Manα1)-OH, Fmoc-Ser(Ac$_4$Manα1-2Ac$_3$Manα1)-OH, Fmoc-Ser(Ac$_4$Manα1-2Ac$_3$Manα1-2Ac$_3$Manα1)-OH, Fmoc-Thr(Ac$_4$Manα1)-OH, Fmoc-Thr(Ac$_4$Manα1-2Ac$_3$Manα1)-OH, Fmoc-Thr(Ac$_4$Manα1-2Ac$_3$Manα1-2Ac$_3$Manα1)-OH and Fmoc-Ser(Ac$_4$Glcβ1)-OH were prepared according to the previously reported methods.

5.4.3 Synthesis of CBM glycoforms

General procedure for the synthesis of unglycosylated CBM variants. The crude peptide was prepared using the previously reported protocol[24]. 16 mg of the crude peptide was dissolved in 80 ml of folding buffer (0.2 M Tris-acetate, 0.33 mM oxidized glutathione, 2.6 mM reduced glutathione, pH 8.2) and stirred at room temperature for 12 h under a helium atmosphere. The solution was then concentrated to a small volume (around 6 ml) using 3 kDa cut-off centrifugal filter units (Amicon) before RP-HPLC purification. The RP-HPLC purification was performed on a Versagrad Preparation-HPLC system using a semi-preparative C18 column. The products were detected by UV absorption at 275 nm. After HPLC purification with a linear gradient of 20→40% MeCN in H_2O over 30 min, the fractions were collected and checked by ESI+ MS. The pure fractions were combined and lyophilized to give the desired product as a white solid.

General procedure for the synthesis of glycosylated CBM variants. The crude glycopeptide was prepared using the previously reported protocol[24]. 16 mg of the crude peptide was dissolved in 1 ml of hydrazine solution (hydrazine/H_2O, 5/100, v/v) and stirred at room temperature for 30 min under helium. The reaction was quenched with 2 ml of acetic acid solution (AcOH/H_2O, 5/100, v/v). The resulting mixture was diluted to 80 mL with folding buffer (0.2 M Tris-acetate, 0.33 mM oxidized glutathione, 2.6 mM reduced glutathione, pH 8.2, 80 ml) and stirred at room temperature for 12 h under a helium atmosphere.

The solution was then concentrated to a small volume (around 6 ml) using 3 kDa cut-off centrifugal filter units (Amicon) before RP-HPLC purification. The RP-HPLC purification was performed on a Versagrad Preparation-HPLC system using a semi-preparative C18 column. The products were detected by UV absorption at 275 nm. After HPLC purification with a linear gradient of 20→40% MeCN in H_2O over 30 min, the fractions were collected and checked by ESI+ MS. The pure fractions were combined and lyophilized to afford the desired product as a white solid.

LC-MS analysis of purified CBM variants. LC-MS was performed under two flow rates with C4 column: (1) 0.5 ml/min with a linear gradient of 15% to 35% acetonitrile in water over 3 min and (2) 0.3 ml/min with a linear gradient of 15% to 35% acetonitrile in water over 5 min.

5.4.4 Characterization of CBM glycoforms

Solid State NMR: Solid state ^{13}C-NMR spectra were collected for each sample using an acquisition time of 0.016 sec. Peak assignments were based on those of Park.[19] The amorphous peak was taken as 80 to 87 ppm and the crystalline peak was taken as 87 to 93 ppm.

5.5 References

1. C. Alvarez, F. M. Reyes-Sosa and B. Diez, *Microb Biotechnol*, 2016, **9**, 149-156.
2. M. F. Li, S. Yang and R. C. Sun, *Bioresour Technol*, 2016, **200**, 971-980.
3. T. Wang and M. Hong, *J Exp Bot*, 2016, **67**, 503-514.
4. M. Foston, *Curr Opin Biotechnol*, 2014, **27**, 176-184.
5. Q. Li, J. Song, S. Peng, J. P. Wang, G. Z. Qu, R. R. Sederoff and V. L. Chiang, *Plant Biotechnol J*, 2014, **12**, 1174-1192.
6. E. M. Rubin, *Nature*, 2008, **454**, 841-845.
7. L. R. Lynd, P. J. Weimer, W. H. van Zyl and I. S. Pretorius, *Microbiol Mol Biol Rev*, 2002, **66**, 506-577, table of contents.
8. C. M. Payne, B. C. Knott, H. B. Mayes, H. Hansson, M. E. Himmel, M. Sandgren, J. Stahlberg and G. T. Beckham, *Chem Rev*, 2015, **115**, 1308-1448.
9. A. Berlin, N. Gilkes, A. Kurabi, R. Bura, M. Tu, D. Kilburn and J. Saddler, *Appl Biochem Biotechnol*, 2005, **121-124**, 163-170.
10. R. H. Narron, H. Kim, H. M. Chang, H. Jameel and S. Park, *Curr Opin Biotechnol*, 2016, **38**, 39-46.
11. H. Palonen and L. Viikari, *Biotechnol Bioeng*, 2004, **86**, 550-557.
12. J. L. Rahikainen, R. Martin-Sampedro, H. Heikkinen, S. Rovio, K. Marjamaa, T. Tamminen, O. J. Rojas and K. Kruus, *Bioresour Technol*, 2013, **133**, 270-278.

13. D. Lee, A. Chen and R. Nair, *Biotechnol Genet Eng Rev*, 2008, **25**, 331-361.
14. D. Gao, C. Haarmeyer, V. Balan, T. A. Whitehead, B. E. Dale and S. P. Chundawat, *Biotechnol Biofuels*, 2014, **7**, 175.
15. J. L. Rahikainen, J. D. Evans, S. Mikander, A. Kalliola, T. Puranen, T. Tamminen, K. Marjamaa and K. Kruus, *Enzyme Microb Technol*, 2013, **53**, 315-321.
16. B. Yang and C. E. Wyman, *Biotechnol Bioeng*, 2006, **94**, 611-617.
17. J. B. Kristensen, J. Börjesson, M. H. Bruun, F. Tjerneld and H. Jørgensen, *Enzyme and Microbial Technology*, 2007, **40**, 888-895.
18. S. P. Chundawat, G. Bellesia, N. Uppugundla, L. da Costa Sousa, D. Gao, A. M. Cheh, U. P. Agarwal, C. M. Bianchetti, G. N. Phillips, Jr., P. Langan, V. Balan, S. Gnanakaran and B. E. Dale, *J Am Chem Soc*, 2011, **133**, 11163-11174.
19. S. Park, J. O. Baker, M. E. Himmel, P. A. Parilla and D. K. Johnson, *Biotechnol Biofuels*, 2010, **3**, 10.
20. M. Hall, P. Bansal, J. H. Lee, M. J. Realff and A. S. Bommarius, *FEBS J.*, 2010, **277**, 1571-1582.
21. J. Siroky, T. A. Benians, S. J. Russell, T. Bechtold, J. Paul Knox and R. S. Blackburn, *Carbohydr Polym*, 2012, **89**, 213-221.
22. M. Linder, G. Lindeberg, T. Reinikainen, T. T. Teeri and G. Pettersson, *FEBS Lett*, 1995, **372**, 96-98.
23. D. W. Sammond, J. M. Yarbrough, E. Mansfield, Y. J. Bomble, S. E. Hobdey, S. R. Decker, L. E. Taylor, M. G. Resch, J. J. Bozell, M. E. Himmel, T. B. Vinzant and M. F. Crowley, *J Biol Chem*, 2014, **289**, 20960-20969.
24. L. Chen, M. R. Drake, M. G. Resch, E. R. Greene, M. E. Himmel, P. K. Chaffey, G. T. Beckham and Z. Tan, *Proc Natl Acad Sci U S A*, 2014, **111**, 7612-7617.
25. X. Guan, P. K. Chaffey, C. Zeng, E. R. Greene, L. Chen, M. R. Drake, C. Chen, A. Groobman, M. G. Resch, M. E. Himmel, G. T. Beckham and Z. Tan, *Chem Sci*, 2015, **6**, 7185-7189.
26. R. Katahira, A. Mittal, K. McKinney, P. N. Ciesielski, B. S. Donohoe, S. K. Black, D. K. Johnson, M. J. Biddy and G. T. Beckham, *ACS Sustainable Chem Eng*, 2014, **2**, 1364–1376.

Chapter 6

Glycoengineering of Therapeutic Peptides for Improved Treatment of Human Diseases

6.1 – Introduction

As part of a fast-growing class of therapeutics in the biopharmaceutical market, short peptides are being widely used to treat human diseases.[1,2] In general, peptides can be highly specific and potent but are unfortunately susceptible to acid/base hydrolysis and proteolytic degradation.[1,3] Over the past three decades, research from many disciplines has established the importance of glycoengineering in overcoming the limitations of peptides, and mounting evidence is pointing to the likelihood that glycosylation of therapeutic peptides can lead to increased stability, biological activity, and reduced aggregation and immunogenicity.[4,5] Such changes could lead to less frequent injection for greater convenience and better patient compliance or even orally available peptide drugs.[6,7] But despite extensive research and effort in the area, many aspects of glycoengineering peptides for optimal performance remain unclear.[8] The deficiency in knowledge mainly stems from the lack of systematic studies of the impact of glycosylation on the physicochemical and biological properties of therapeutic peptides, which in turn, is due to the inaccessibility of peptides bearing structurally-defined glycans.

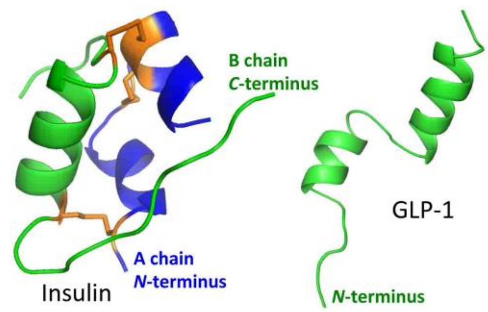


Figure 6.1 - Structure of human insulin and GLP-1. The disulfide bonds are highlighted in orange.

To better understand the impact of peptide glycosylation, homogeneous samples of individual glycoforms with well-defined glycan structures are indispensable.[9-17] Our studies and previous studies by others have clearly demonstrated that the characterization of such pure glycoforms can provide definitive information regarding the roles of glycosylation in modulating peptide stability, aggregation propensity, and biological activity.[18-22] We recently developed and optimized a convenient, efficient process for preparing large collections of glycopeptides that carry systematic variations in both glycan structure and amino acid sequence.[20] Easier access to libraries of homogeneous glyco-variants is expected to greatly facilitate the development of a more universal set of guidelines for peptide glycoengineering.

As an important step to achieve this goal, we choose to study the glycoengineering of two representative peptides, human insulin and glucagon-like peptide-1 (GLP-1), each representing one of the two groups of therapeutic peptides: those stabilized by disulfide bonds and those without disulfide bonds (Figure 6.1).[23,24] Human insulin is a small protein molecule made up of two separate polypeptide chains – A chain (1-21) and B chain (1-30) – which are intra- and inter-connected through three disulfide bridges.[25] GLP-1 is a short peptide that can form a stable α-helix in aqueous solution. Both of them are widely employed for the management of type 1 and late-stage type 2 diabetes. As an extremely important short peptide, insulin has long been used to develop new strategies for protein sequencing,[26] synthesis,[27] expression,[28] structure determination,[29] and engineering.[30] Similarly, previous studies of GLP-1 have also helped to establish new approaches for the engineering of therapeutic peptides.[31,32] Therefore, by systematically analyzing homogeneous glyco-variants of human insulin and GLP-1, it is highly possible to obtain a set of general glycoengineering guidelines for future development of peptides with improved therapeutic properties.

Chemical synthesis will be used to prepare site-specifically glycosylated peptides. Although other methods, such as biological expression, enzymatic synthesis, or “click-like” conjugation methods, are more practical means for large-scale production of glycopeptides, chemical synthesis offers greater flexibility for introducing variations into glycopeptides and for totally controlling every aspect of glycan structure and amino acid sequence.[22,33] This is a direct consequence of the fact that chemical glycosylation is not dictated by the chemical properties, under-lying amino acid sequences, or local conformations of peptides. It thus allows for more diversity in glyco-variant structures, which will enable us to define an as-comprehensive-as-possible set of glycoengineering guidelines.

6.2 Results and discussion

We applied a chemical glycobiology approach developed in our laboratory to achieve the proposed goals, beginning with the glycoengineering of human insulin.[19,20] The effects of glycosylation on insulin's physicochemical and biological properties were be established by comparing each glycoform to the unglycosylated insulin and any closely related insulin glyco-variants. At the same time, an identical chemical glycobiology approach was applied to the engineering of a structurally different peptide, GLP-1. We expected that the guidelines for glycoengineering of peptide therapeutics would be unveiled by gathering the rules that are applicable to both cases.

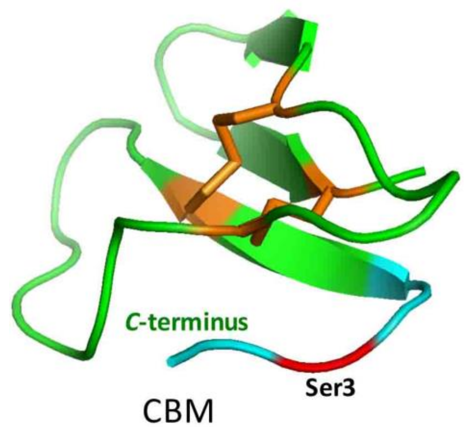


Figure 6.2 - Glycoengineering of CBM. The *N*-terminal region with systematically varied amino acids sequences is highlighted in cyan. The disulfide bonds are highlighted in orange. The glycosylation site is highlighted in red.

Through our work on a 36-mer peptide, a Family 1 carbohydrate-binding module (CBM), we have demonstrated the feasibility and effectiveness of the proposed strategy in developing new glycoengineering rules for improving the performance of peptides (Figure 6.2).[19,20] By systematically comparing 51 CBM glyco-variants, we were able to reveal that variations in the CBM's proteolytic stability and thermal stability followed a similar trend. Both of them are very likely controlled by conformation-stabilizing effects of local glycans. Larger glycans generally confer better proteolytic stability. We have also shown that planar polar (Gln) and aromatic amino acid (Tyr) residues as well as *O*-glycans α-linked to Ser or Thr are important for these effects. Furthermore, we found that the attachment of glycans to residues that lie at the termini, close to secondary structure elements and disulfide-bonds often had little or even a negative impact on both binding affinity and stability. Taken together, our study of the CBM indicated that glyco-variants with better overall properties could be generated by coordinately varying the structures of glycans and amino acids near the glycosylation site.

For the largest effects, it was also critical that the glycans be in an unstructured region important for substrate binding and susceptible to proteolytic cleavage.

6.2.1 Glycoengineering of human insulin

CBM is a naturally glycosylated single-chain peptide. The glycoengineering guidelines derived from studying this molecule may be limited to peptides with a similar size, structure, and/or sequence. To develop general glycoengineering guidelines for improving the performance of peptides, it is necessary to first verify if what we observed during the study of CBM glycosylation is still valid for other disulfide bridged peptides, especially those that are naturally unglycosylated.

In order to investigate how general the glycoengineering guidelines derived from the studies of CBM are, we started investigating the effects of glycosylation on human insulin, a 51-mer peptide that contains three disulfide bonds. This work may lead to the identification of insulin variants with better therapeutic properties, especially those with more suitable properties for oral administration.[34,35] As a peptide, orally administrated insulin can be quickly degraded in the stomach and small intestine before being absorbed into the bloodstream and reaching its intended targets. Happily, through the use of pH sensitive capsules, insulin can now be reliably protected from the harsh environment of the stomach and can be selectively delivered to the intestinal track.[36] However, insulin's high oligomerization propensity further complicates the issue. The intestinal epithelium forms a selective barrier which is generally impermeable to large molecules, including insulin oligomers.[37] This makes the absorption of insulin in the small intestine inefficient.[38] Also in the small intestine, a variety of proteases exist that easily chew up short peptides like insulin.[39] Therefore, insulin variants that can be used orally should have improved resistance to proteolytic degradation, lower oligomerization propensities, and better or unchanged biological activity.[35]

As an important step to examine the effects of insulin glycosylation, we first prepared 12 different insulin glycoforms, **2-13**, each containing an *O*-linked *N*-acetylgalactosamine (GalNAcα), monomannose (Manα), dimannose (Manα2Manα), or trimannose (Manα2Manα2Manα), at either SerA9, SerA12, SerB9, ThrB27 or ThrB30 site (Figure 6.3A). ThrA8 was not used in this study because it is adjacent to SerA9 and the effects of its glycosylation can be roughly represented by those of SerA9 glycosylation. Currently, biosynthesis and isolation of glycoforms with random changes in their glycosylation patterns has not been well optimized and so chemical synthesis was used to prepare wild-type insulin and its variants (**1-13**, Fig. 3). Due to the presence of difficult-to-synthesize sequences and the possibility of forming non-native disulfide bonds, chemical synthesis has yet to prove itself as a convenient technique for the preparation of insulin glyco-variants.[40,41] By systematically optimizing each of the synthetic steps, we have developed a robust, practical and efficient one-pot process to synthesize, fold, deprotect, and purify uniform insulin glycoforms. Using this method, we were able to quickly generate highly pure unglycosylated insulin **1** and insulin glycoforms **2-13** in sufficient amount for biophysical and biological characterizations. Notably, the average total time for preparing each variant is only three days.

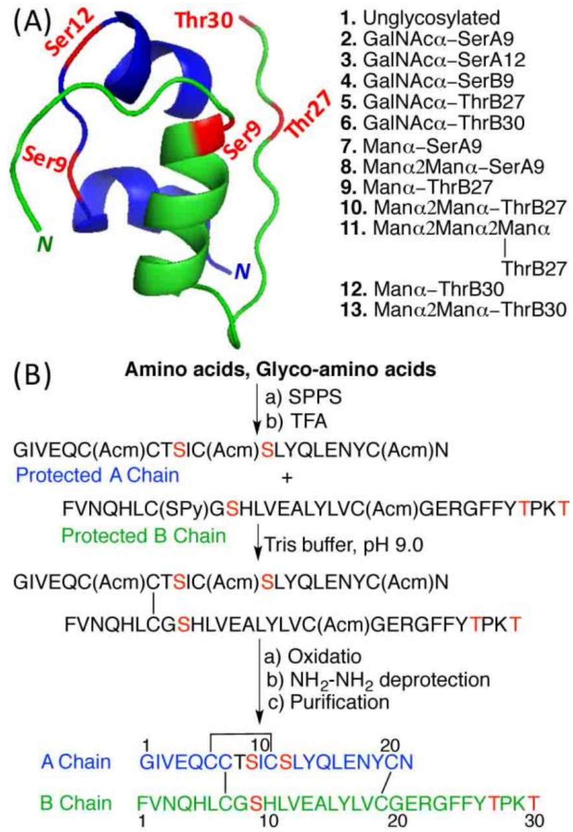


Figure 6.3 - Design and synthesis of insulin analogs. (A) The structure of human insulin and its glyco-variants. The structural feature of each glycoform is implied by its name, *i.e.* GalNAcα-SerA9 representing the glycoform containing a single GalNAc α-linked to the A chain Ser9, Manα2Manα2Manα-ThrB27 representing the glycoform containing an α1,2-linked trimannose at the B chain Thr27 site. (B) The optimized synthetic route to glycosylated insulin variants. The *O*-glycosylated Ser and Thr residues are highlighted in red.

With the synthetic insulin glycoforms in hand, we first investigated if *O*-linked glycans at any one of the five glycosylation sites in particular affects the stability of human insulin in the presence of α-chymotrypsin, a protease synthesized by the pancreas and secreted into the lumen of the small intestine. Chymotrypsin is capable of cleaving human insulin at the *C*-terminus of its B chain, an important region for receptor binding and activation, thus diminishing its biological activity.[42,43] This cleavage also causes an easily detectable change in molecular mass, and therefore each insulin glycoform's half-life towards α-chymotrypsin degradation can be calculated by monitoring the first-order exponential decay of the full-length glycoform using quantitative Matrix Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS).[19,20] The role of *O*-linked glycans in human insulin proteolytic stability can be established by comparing the half-lives of synthetic glycoforms with that of the unglycosylated insulin. As shown in Figure 6.4A, *O*-glycosylation with a GalNAcα (**2-6**) or Manα (**7**, **9**, **12**) moiety does not positively impact the proteolytic stability. However, as observed in our CBM studies, we found that dimannosylation (**10**) and trimannosylation (**11**) at ThrB27, which is adjacent to one of the cleavage sites, leads to noticeable improvement in proteolytic stability. The half-life of trimannosylated **11** is twice as long as that of unglycosylated insulin **1**.

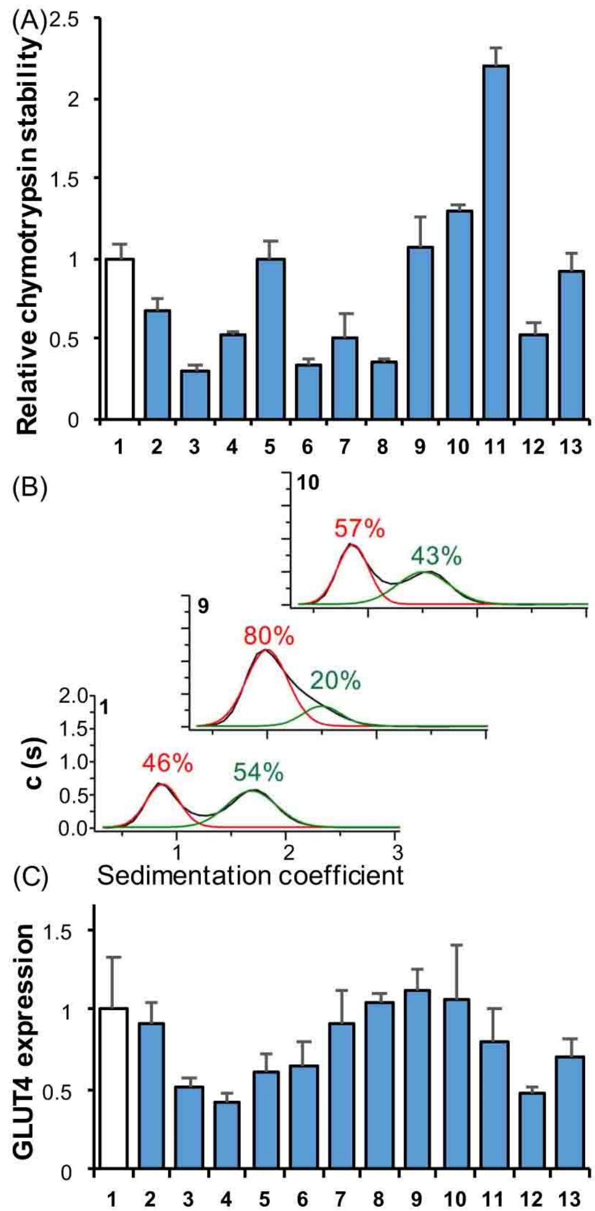


Figure 6.4 - Characterization of synthetic insulin glyco-variants. (A) The effects of *O*-glycosylation on the proteolytic stability (relative half-life to α-chymotrypsin degradation). (B) oligomerization propensity (sedimentation coefficient distribution. Monomer is highlighted in red. Dimer in green). (C) Insulin stimulated translocation of HA-GLUT4. All error bars reported are standard deviations of data achieved from three separate trials.

In addition to improving the proteolytic stability of insulin, we found that *O*-glycosylation could also decrease the oligomerization propensity of insulin. Oligomerization is a critical regulatory factor in insulin absorption and it is well known that aggregation of insulin can decrease its absorption.[40] By analyzing the sedimentation velocity though analytical ultracentrifugation (AUC), we derived a distribution of insulin molecular species that have different degrees of self-association (Fig. 4B).[44] The area under each peak gives the relative concentration of that species. As suggested by the data, the attachment of *O*-linked mannose to Thr27 can also significantly decrease insulin self-association (compare **9** and **10** to **1**).

The promise of glycoengineering design is to create insulin products of increased beneficial properties while not sacrificing any biological activity. In order to understand the relationship between the glycosylation and the biological activity of human insulin, we used a quantitative fluorescence assay to compare the trafficking of hemagglutinin (HA)-tagged glucose transporter type 4 (GLUT4) in 3T3-L1 adipocytes.[45,46] GLUT4 is an insulin-regulated glucose transporter that is responsible for insulin-regulated glucose uptake into liver, muscle, and fat cells. Insulin can increase the cell-surface level of GLUT4 by stimulating the translocation of GLUT4 to the plasma membrane. Therefore, the biological activity of each insulin analog can be evaluated by analyzing the level of cell-surface GLUT4. As shown in Figure 6.4C, the attachment of linear mannose chains to Thr27 did not lead to significantly decreased activity.

6.2.2 Glycoengineering of human GLP-1

To demonstrate the feasibility and usefulness of glycoengineering in addressing the challenges faced by peptide drugs, it is necessary to investigate the effects of glycans on the stability and actions of many therapeutic peptides across different sizes, amino acid compositions, and structural and functional features. Based on these criteria, human GLP-1 was chosen as another molecule to evaluate the effectiveness of peptide glycoengineering. Although both are short peptides, GLP-1 has one significant structural difference from insulin: it lacks disulfide bonds.[47] GLP-1 is naturally circulated as a pair of biologically active isoforms, GLP-1 (7-37) and GLP-1 (7-36) amide, which appear to be equipotent in their actions (Fig. 6A).[48] Because of its multiple beneficial functions in glucose disposal, various GLP-1 analogs are currently used or being tested for the treatment of diabetes. Sadly, as a short peptide, the proteolytic stability of GLP-1 is very low and it is rapid inactivated by proteases like dipeptidyl peptidase IV (DPP IV).[24] Development of GLP-1 analogs with better stability would greatly reduce the inconvenience and side effects associated with frequent injections during treatment. This objective can be pursued by attaching sugar units to the unstructured N-terminus of GLP-1, a region important for its biological activity and a sequence that is susceptible to proteolytic cleavage.

As shown in Figure 6.5A, Human GLP-1 contains one Ser and two Thr residues in its *N*-terminus, which can be used as *O*-glycosylation sites. Thr11 is closer to the protease cleavage site and is flanked by Glu and Phe residues; and this makes Thr11 a better choice than either Thr13 or Ser14 for the glycosylation site. As the initial step for evaluating the effects of GLP-1 glycosylation, we prepared and investigated a collection of homogeneously glycosylated GLP-1 analogs with systematically varied glycan structures. As shown in Figure 6.5B and 6.5C, our results confirmed that *O*-glycosylation at Thr11 confers protease protection.

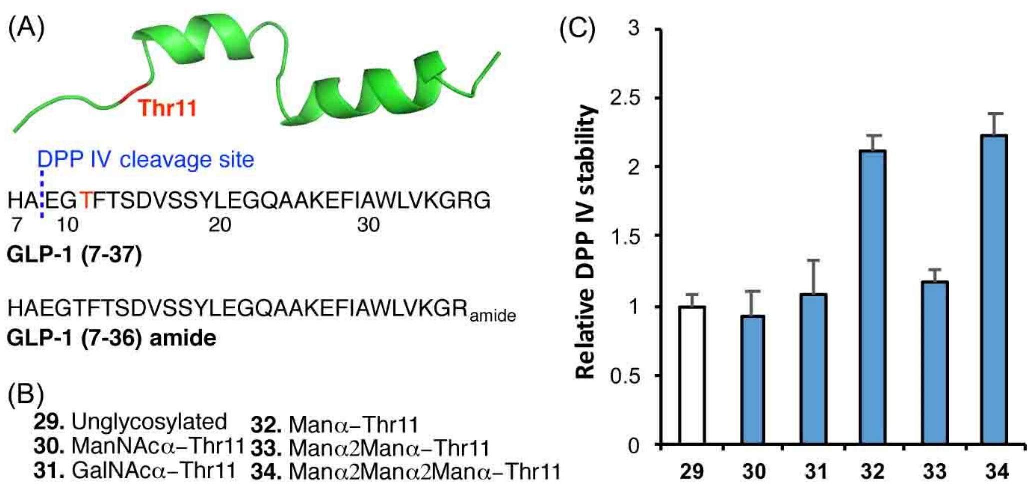


Figure 6.5 - Characterization of synthetic GLP-1 glyco-variants. (A) The NMR structure, amino acid sequence, and DPP IV cleavage site of human GLP-1. The glycosylated Thr residue is highlighted in red. (B) Synthesized GLP-1 variants. (C) The effects of *O*-glycosylation on the proteolytic stability (relative half-life to DPP IV degradation). All error bars reported are standard deviations of data achieved from three separate trials.

6.4 Conclusion

Together with previous studies, our work has confirmed the feasibility and effectiveness of the glycoengineering approach in increasing the beneficial properties of human insulin and GLP-1. Moreover, it has demonstrated the guidelines derived from the studies of an unrelated glycopeptide CBM can be applied to the glycoengineering of naturally unglycosylated protein molecules. These observations make us confident that our research represent a unique opportunity to develop more glycoengineering guidelines and to further improve the performance of therapeutic peptides. In the long term, we will further investigate the effects of glycosylation on the receptor binding specificity[49] and immunogenicity[50] of insulin and GLP-1 variants and develop better technologies for large-scale and cost effective production of promising candidates for clinical assessment.[51] Moreover, we will apply the guidelines derived from the proposed studies to the glycoengineering of other interesting therapeutic peptides, like enfuvirtide, calcitonin, and teduglutide, with the goal of addressing their possible therapeutic challenges.[2]

6.5 Experiments

6.5.1 Materials

All commercial reagents and solvents were used as received. Unless otherwise noted, all reactions and purifications were performed under air atmosphere at room temperature. All LC-MS analyses were performed using a Waters Acquity™ Ultra Performance LC system equipped with Acquity UPLC® BEH 300 C4, 1.7μm, 2.1 x 100 mm column at flow rates of 0.3 and 0.5 mL/min. The mobile phase for LC-MS analysis was a mixture of H_2O (0.1% formic acid, v/v) and acetonitrile (0.1% formic acid, v/v). All preparative separations were performed using a LabAlliance HPLC solvent delivery system equipped with a Rainin UV-1 detector and a Varian Microsorb 100-5, C18 250x21.4mm column at a flow rate of 16.0 mL/min. The mobile phase for HPLC purification was a mixture of H_2O (0.05% TFA, v/v) and acetonitrile (0.04% TFA, v/v). A Waters SYNAPT G2-S system was used mass spectrometric analysis.

6.5.2 General procedure for the synthesis of insulin variants.

Solid-Phase Peptide Synthesis, Cleavage and Activation - Automated peptide synthesis was performed on an Applied Biosystems Pioneer continuous flow peptide synthesizer. Peptides were synthesized under standard automated Fmoc conditions. The deblock solution was a mixture of 100/5/5 of DMF/piperidine/DBU. Fmoc protected amino acid (4.0 equiv), HATU (4.0 equiv) and DIEA (8.0 equiv) were used for the coupling steps.

A-Chain: GIVEQC(Acm)CTSIC(Acm)SLYQLENYC(Acm)N

B-Chain: FVNQHLC(SPy)GSHLVEALYLVC(Acm)GERGFFYTPKT

The synthesis of A-Chain was conducted on 0.05 mmol Fmoc-ASN-NovaSyn® TGT resin from EMD Millipore. After cleavaged from the resin and precipitated by cold ether, the crude peptide was dissolved in 20mL of MeCN/H2O (1:1) and lyophilized to dry for next folding step without further purification.

The synthesis of B-Chain was conducted on 0.05 mmol Fmoc-Thr-NovaSyn® TGT resin from EMD Millipore. Cleavage was conducted by treating the 0.05 mmol resin with 10 mL 0f TFA/TIS/H2O (95: 2.5: 2.5), which contains 20.0 equiv of 2,2′-Dithiodipyridine (DTDP, 0.22 g) at rt for 2 hour. The resin was filtered off, the filtrate was blown away by condensed air and then precipitated by cold ether (30 mL). The precipitate was collected by centrifugation, then washed with cold ether (30 mL x 3). The crude peptide was dissolved in 20mL of MeCN/H2O=1:1 and lyophilized to dry for next folding step without further purification.

Insulin folding **-** A-chain (0.02 mmol, 1.0 equiv) and B-chain (0.0204 mmol, 1.2 equiv) were mixed in 2 mL of 8 M GnHCl, 0.1 M Tris buffer (pH 8.8). The mixture was vortexed vigorously until fully dissolved, then pH was raised to 8 by adding 20uL of 2 M NaOH (monitored by pH strip). This solution was stirred for 5 minutes before diluted by 16 mL of AcOH/H2O (4:1), followed by treatment with I2 (0.282 g) in MeOH (3.2 mL) for 15 minutes at rt. Then this mixture was treated by 1 M aq ascorbic acid

(4.8 mL), diluted by H2O (20 mL), then loaded to preparative RP-HPLC for purification. Preparative separations were performed using a LabAlliance HPLC solvent delivery system equipped with a Rainin UV-1 detector and a Varian Microsorb 100-5, C18 250x21.4mm column (100 Å pore size) at a flow rate of 16.0 mL/min. All separations involved a mobile phase consisting of 0.05% TFA (v/v) in water (solvent A) and 0.04% TFA in acetonitrile (solvent B). The products were detected by UV absorption at 230 nm. After HPLC purification with a linear gradient of 20→50% MeCN in H_2O over 30 min, the fractions were collected and checked by LCMS. The pure fractions with the desired mass and shortest retention time were combined and lyophilized to give the desired product as a white powder.

6.5.3 General procedure for the synthesis of glycosylated insulin variants.

Solid-Phase Peptide Synthesis, Cleavage and Activation **-**

A-Chain: GIVEQC(Acm)CTSIC(Acm)SLYQLENYC(Acm)N (The amino acids shown in red are designed glycosylation sites.)

The synthesis of A-Chain was conducted on 0.05 mmol Fmoc-ASN-NovaSyn® TGT resin from EMD Millipore. Fmoc-Ser(Ac3GalNAcα1)-OH, Fmoc-Ser(Ac4Manα1)-OH, or Fmoc-Ser(Ac4Manα1-2Ac$_3$Manα1)-OH was used to introduce the sugars to the peptides as building blocks. After cleavaged from the resin and precipitated by cold ether, the crude peptide was dissolved in 20mL of MeCN/H2O (1:1) and lyophilized to dry for next folding step without further purification.

B-Chain: FVNQHLC(SPy)GSHLVEALYLVC(Acm)GERGFFYTPKT

The synthesis of B-Chain was conducted on 0.05 mmol Fmoc-Thr-NovaSyn® TGT resin from EMD Millipore. Fmoc-Ser(Ac3GalNAcα1)-OH, Fmoc-Thr(Ac3GalNAcα1)-OH, Fmoc-Thr(Ac4Manα1)-OH, Fmoc-Thr(Ac4Manα1-2Ac$_3$Manα1)-OH or Fmoc-Thr(Ac4Manα1-2Ac$_3$Manα1-2Ac$_3$Manα1)-OH was used to introduce the sugars to the peptides as building blocks. Cleavage was conducted by treating the 0.05 mmol resin with 10 mL 0f TFA/TIS/H2O (95: 2.5: 2.5), which contains 20.0 equiv of DTDP (0.22 g) at rt for 2 hour. The resin was filtered off, the filtrate was blown away by condensed air and then precipitated by cold ether (30 mL). The precipitate was collected by centrifugation, then washed with cold ether (30 mL x 3). The crude peptide was dissolved in 20mL of MeCN/H2O=1:1 and lyophilized to dry for next folding step without further purification.

B-Chain: FVNQHLC(SPy)GSHLVEALYLVC(Acm)GERGFFYTPKT

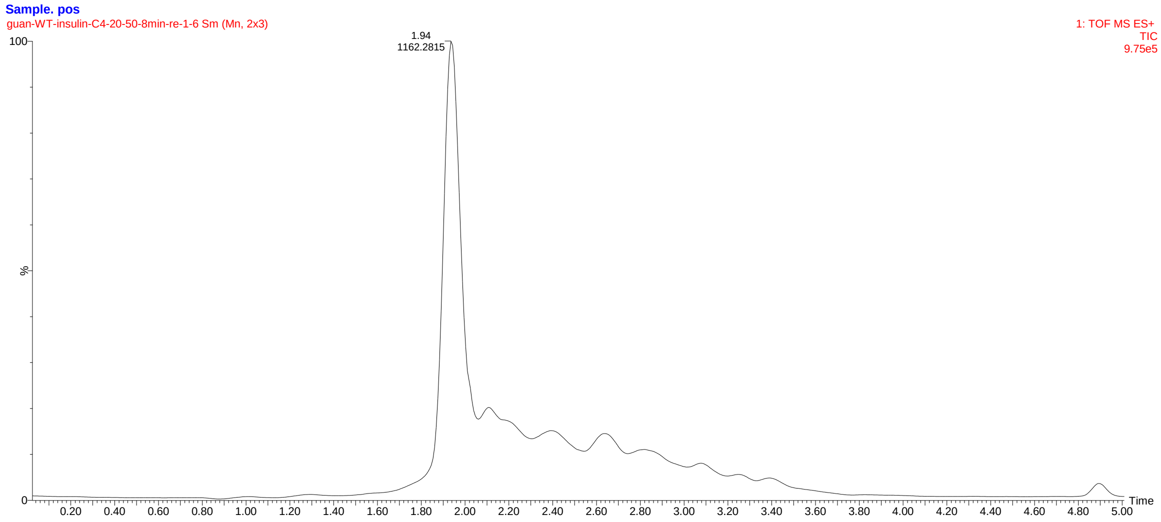
Fully protected peptide Boc-FVNQHLC(Trt)GSHLVEALYLVC(Acm)GERGFFYTPK-OH was synthesized on 0.05 mmol Fmoc-Lys-NovaSyn® TGT resin from EMD Millipore. Cleavage was conducted by treating the 0.05 mmol resin with 10 mL 0f DCM/TFE (7:3) at rt for 2 hour. The resin was filtered off, the filtrate was blown away by condensed air and then precipitated by cold ether (30 mL). The crude peptide was collected by centrifuge and dissolved in 20mL of MeCN/H2O=1:1 and lyophilized to dry for next coupling step without further purification. The fully protected peptide (0.0185 mmol, 1.1 equiv) and glycosylated Threonine building blocks (0.0168 mmol, 1.0 equiv) (H-Thr(Ac3GalNAcα1)-OtBu, H-Thr(Ac4Manα1)-OtBu or H-Thr(Ac4Manα1-2Ac$_3$Manα1)-OtBu) were dissolved in 710 uL of CHCl3/TFE (3:1). The mixture was cooled to -10^{o}C, then HOOBt (0.0185 mmol, 1.1 equiv) and EDCI (0.0185 mmol, 1.1 equiv) were added. The mixture was stirred at rt for 3h with centrifuge every 30 mins. After 3h, the solvent was blown away by condensed air and 1 mL of AcOH/H2O (1:20) was added. The supernate was discarded after centrifuge and the residue was dissolved in 4 mL of TFA/TIS/H2O (95: 2.5: 2.5), which contains 20.0 equiv of DTDP (0.088 g) at rt for 2 hour. The resin was filtered off, the filtrate was blown away by condensed air and then precipitated by cold ether (15 mL). The precipitate was collected by centrifugation, then washed with cold ether (15 mL x 3). The crude peptide was dissolved in 10mL of MeCN/H2O=1:1 and lyophilized to dry for next folding step without further purification.

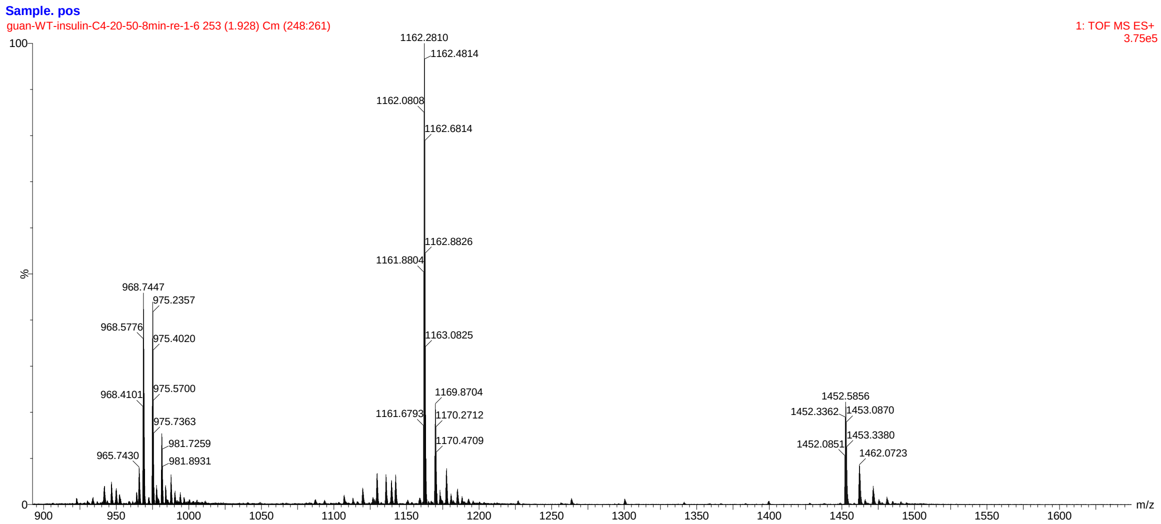
Insulin folding - A-chain (0.02 mmol, 1.0 equiv) and B-chain (0.0204 mmol, 1.2 equiv) were mixed in 2 mL of 8 M GnHCl, 0.1 M Tris buffer (pH 8.8). The mixture was vortexed vigorously until fully dissolved, then pH was raised to 8 by adding 20uL of 2 M NaOH (monitored by pH strip). This solution was stirred for 5 minutes before diluted by 16 mL of AcOH/H2O (4:1), followed by treatment with I2 (0.282 g) in MeOH (3.2 mL) for 15 minutes at rt. Then this mixture was treated by 1 M aq ascorbic acid (4.8 mL), diluted by H2O (20 mL), then loaded to preparative RP-HPLC for purification. Preparative separations were performed using a LabAlliance HPLC solvent delivery system equipped with a Rainin UV-1 detector and a Varian Microsorb 100-5, C18 250x21.4mm column (100 Å pore size) at a flow rate of 16.0 mL/min. All separations involved a mobile phase consisting of 0.05% TFA (v/v) in water (solvent A) and 0.04% TFA in acetonitrile (solvent B). The products were detected by UV absorption at 230 nm. After HPLC purification with a linear gradient of 25→60% MeCN in H_2O over 30 min, the fractions were collected and checked by ESI+ MS. The pure fractions with the desired mass were combined and lyophilized to give the Ac-protected product as a white powder.

Ac removal - The the Ac-protected product was dissloved in 1 mL hydrazine/H2O (1:20) and stirred at rt for 30 mins. Then the reaction was quenched with 1 mL AcOH/H2O (1:20). The mixture was loaded to preparative RP-HPLC for purification. After HPLC purification with a linear gradient of 25→60% MeCN in H_2O over 30 min, the fractions were collected and checked by LCMS. The pure fractions with the desired mass and shortest retention time were combined and lyophilized to give the desired product as a white powder.

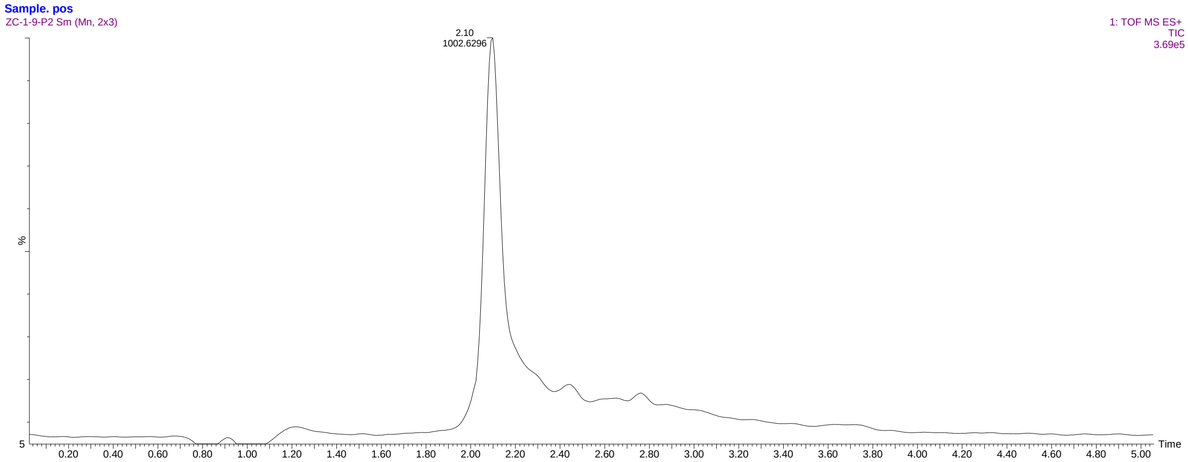
6.5.4 LC-MS Traces and Spectra for Synthetic Insulins 1-13

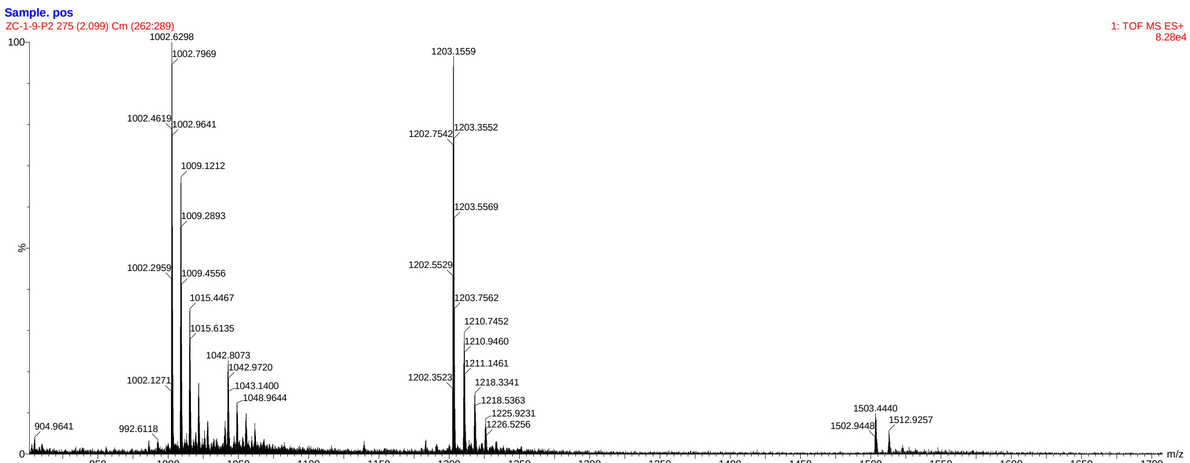
unglycosylated insulin **1**



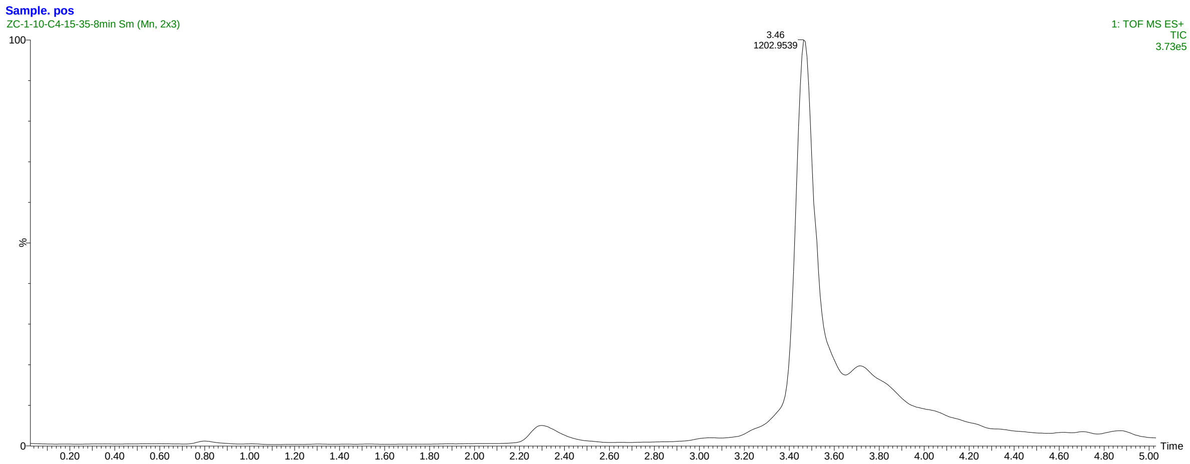


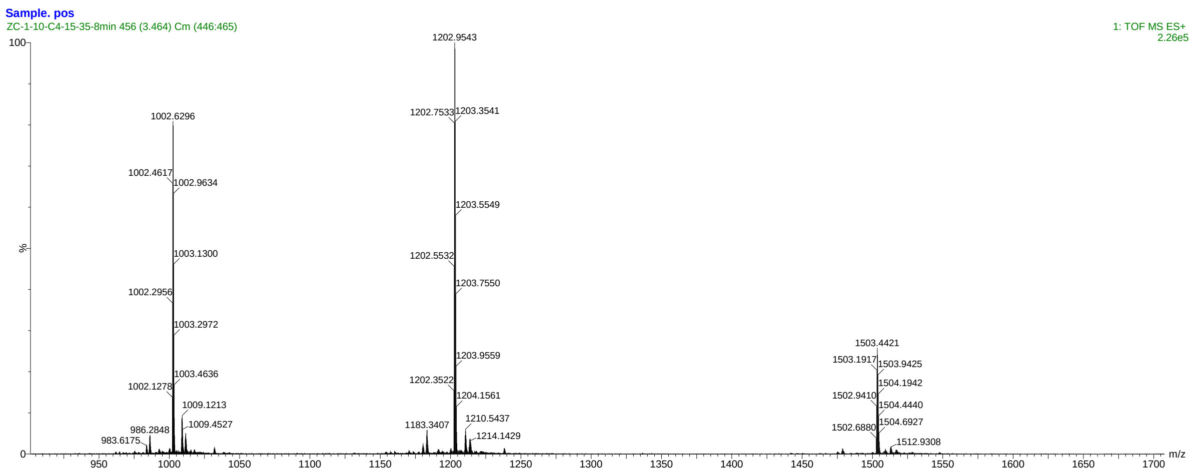
GalNAcα-SerA9 **2**



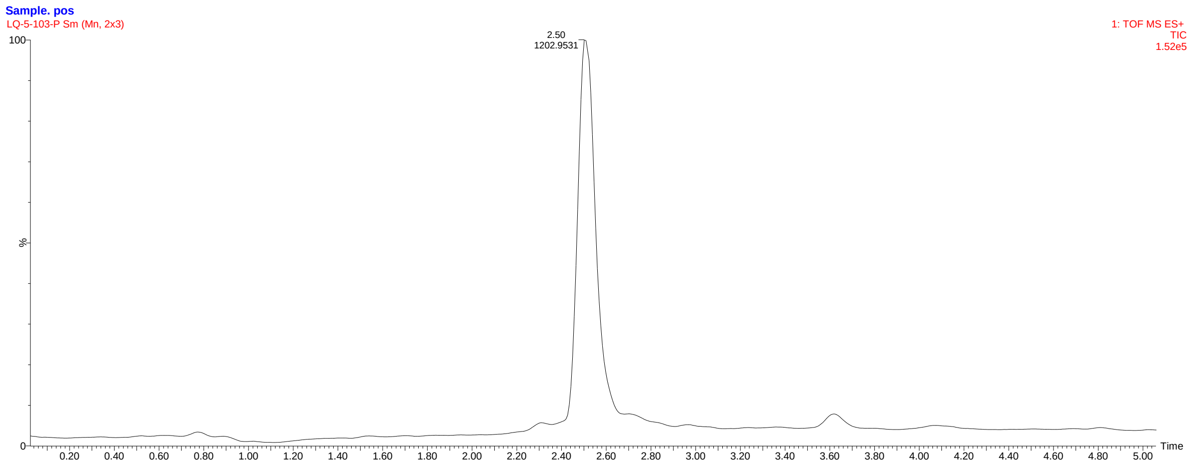


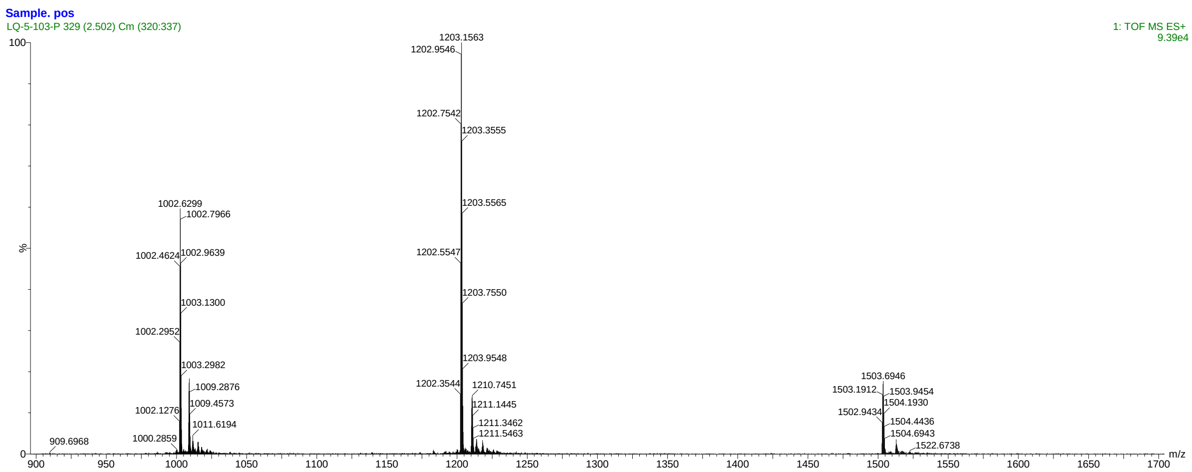
GalNAcα-SerA12 **3**



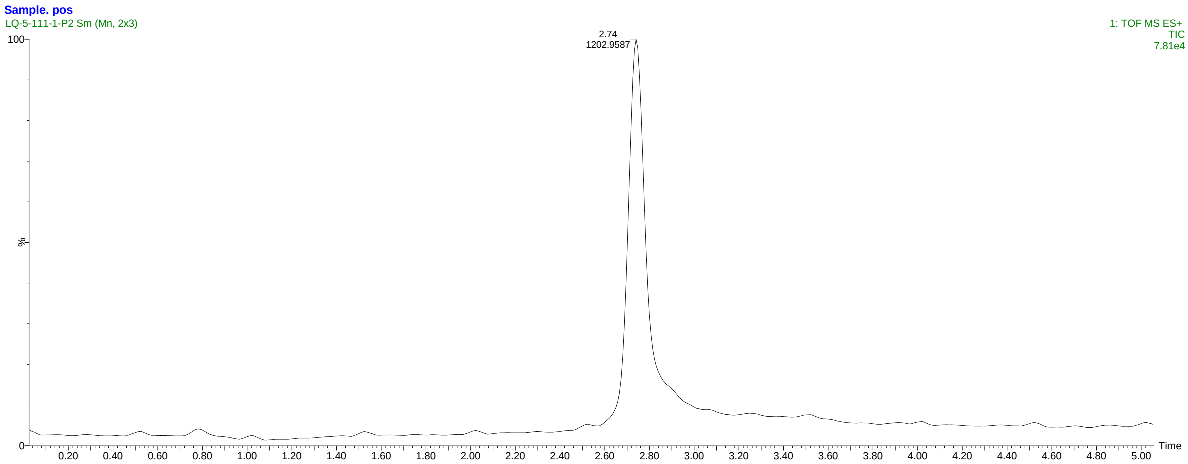


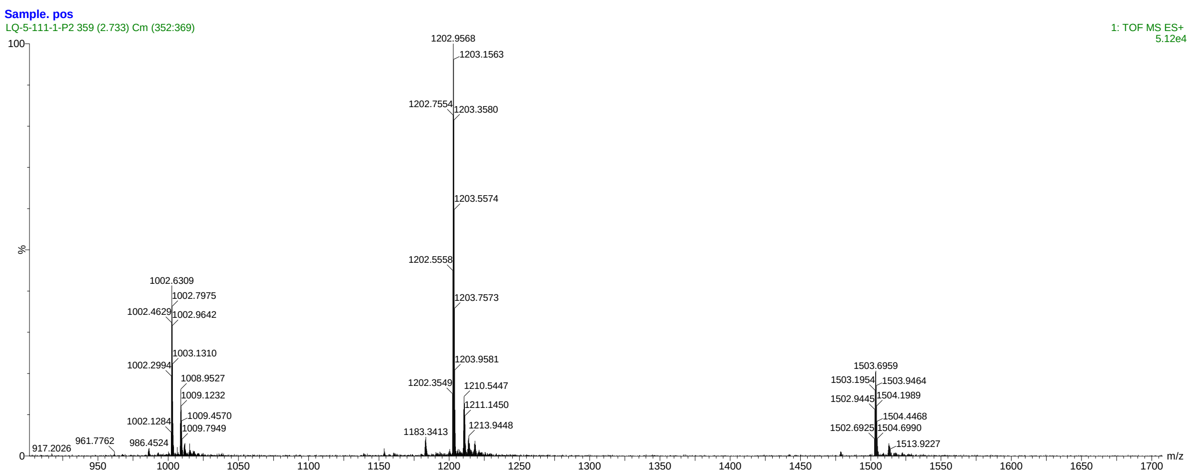
GalNAcα-SerB9 **4**



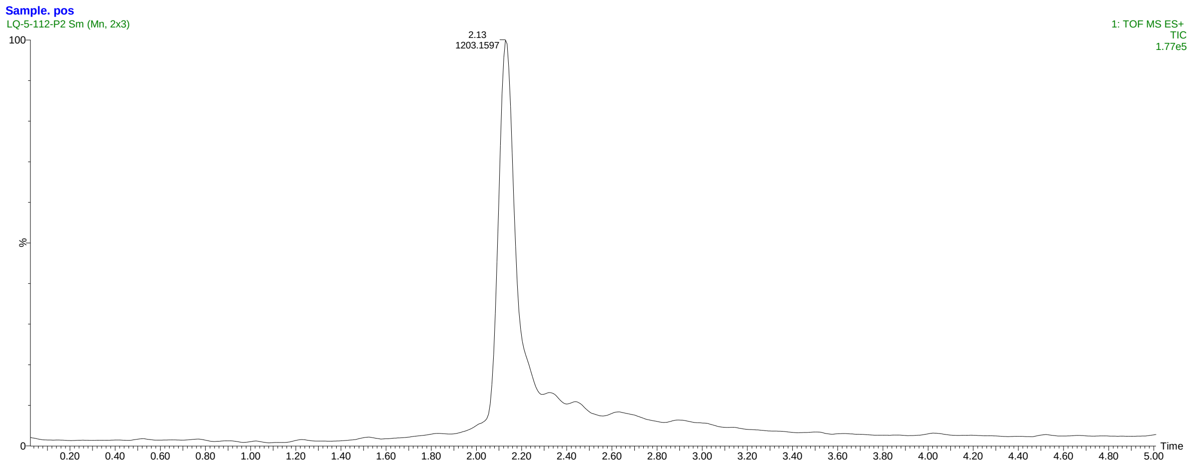


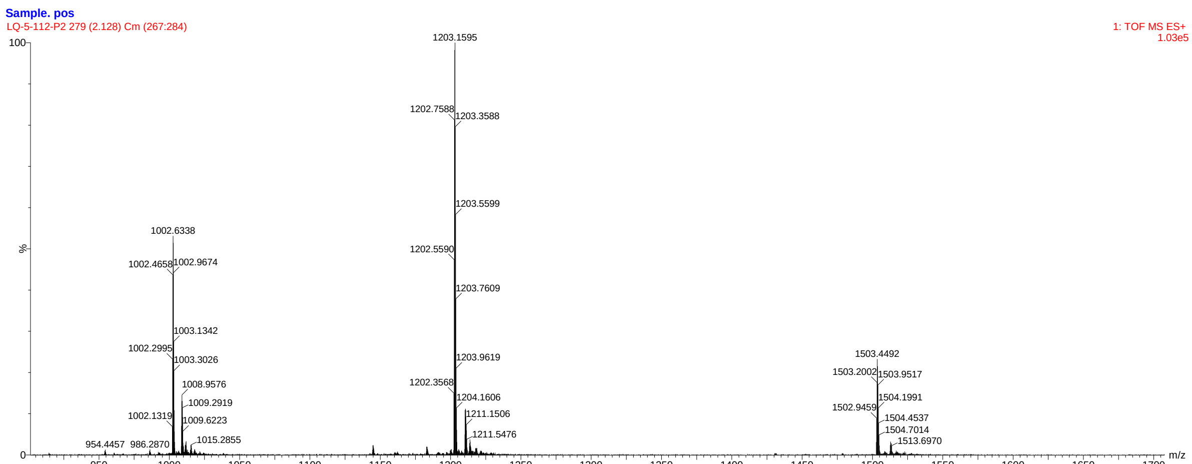
GalNAcα-ThrB27 **5**



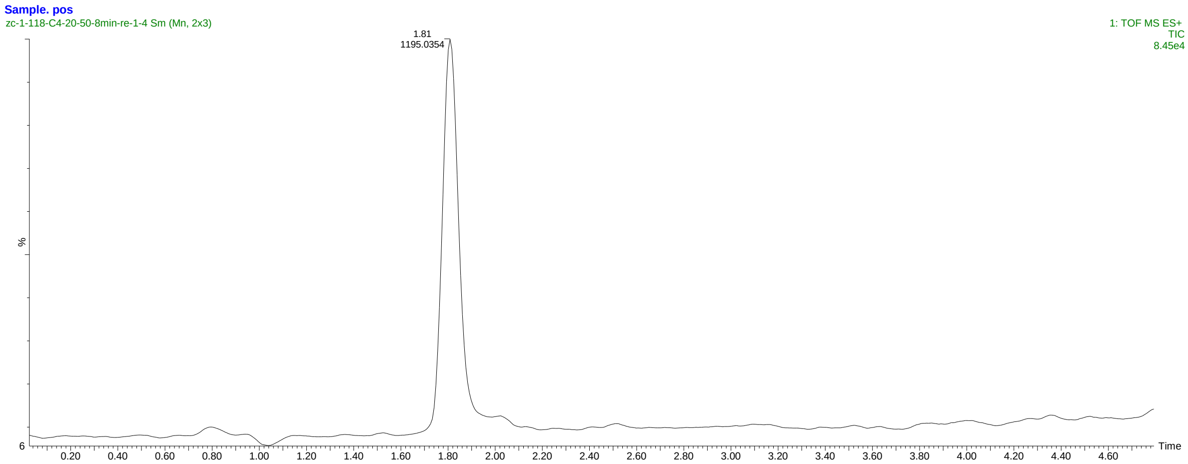


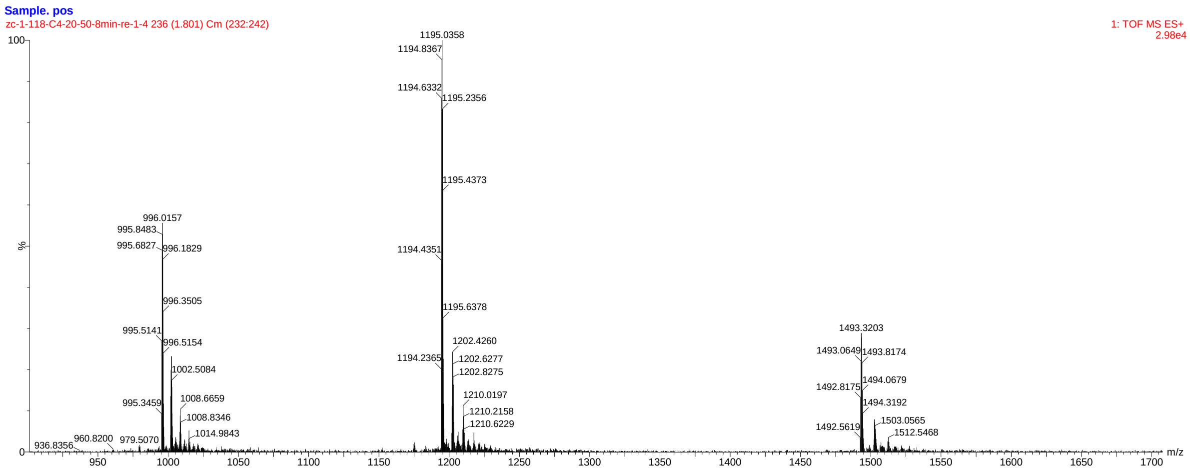
GalNAcα-ThrB30 **6**



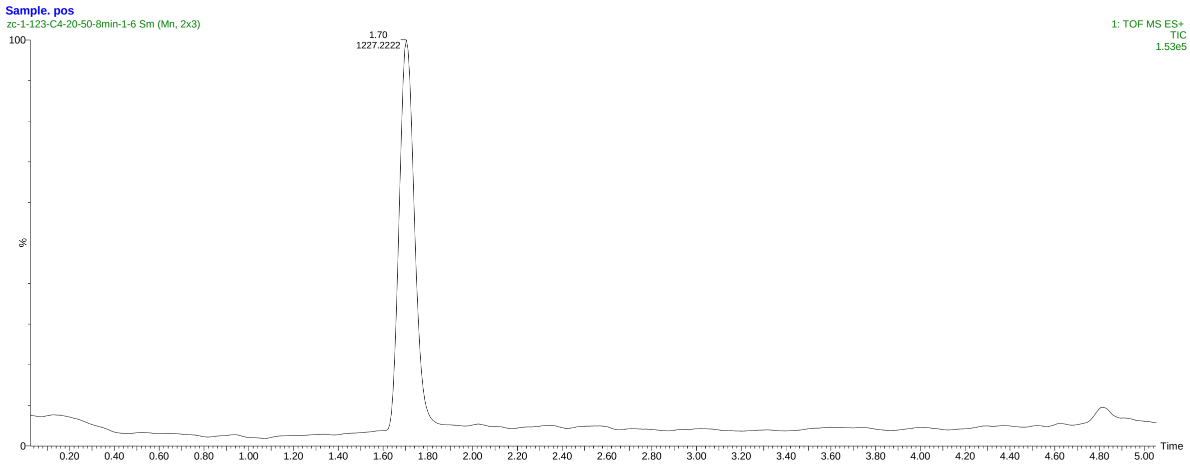


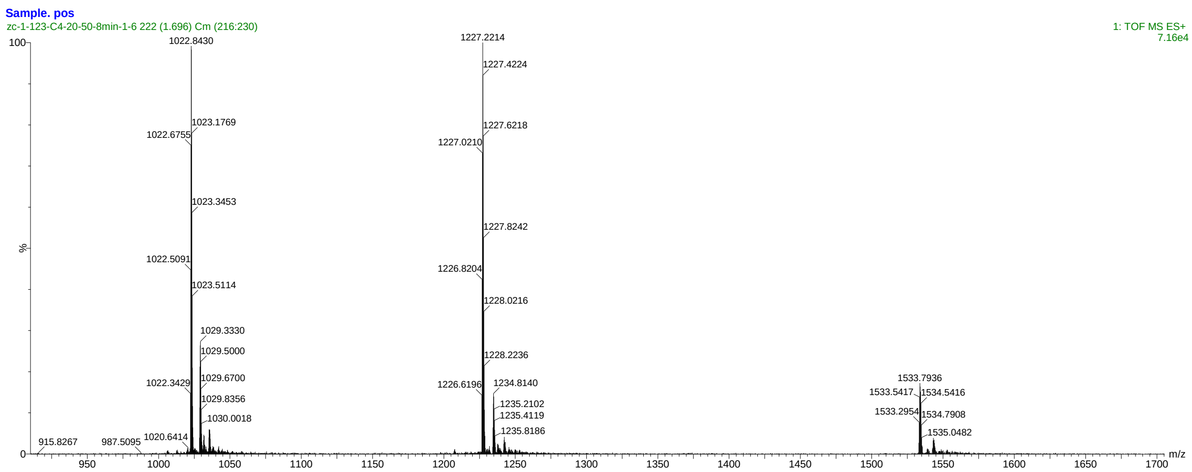
Manα-SerA9 **7**


Sample. pos
zc-1-118-C4-20-50-8min-re-1-4 Sm (Mn, 2x3)
1: TOF MS ES+
TIC
8.45e4
1.81
1195.0354
%
Time

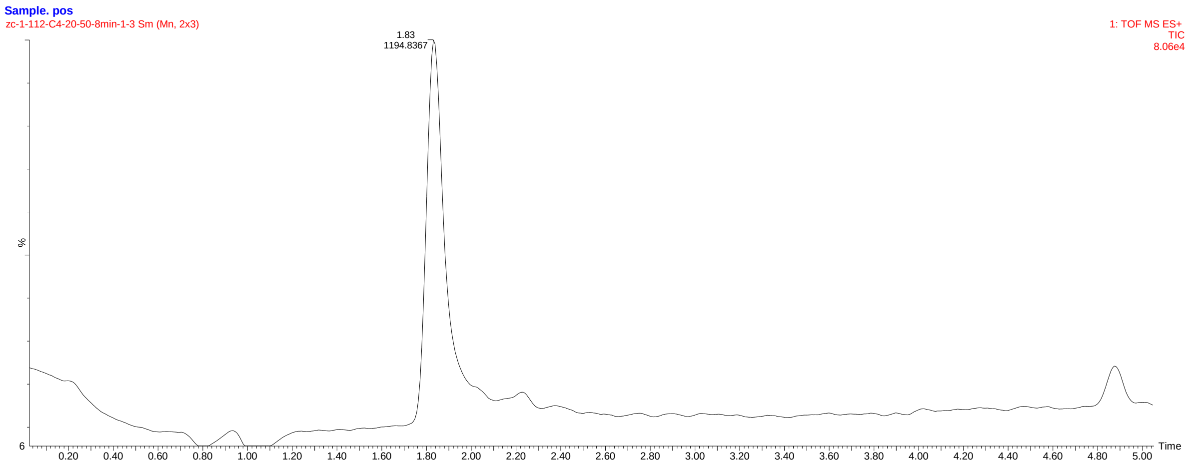

Sample. pos
zc-1-118-C4-20-50-8min-re-1-4 236 (1.801) Cm (232:242)
1: TOF MS ES+
2.98e4
1195.0358
1194.8367
1194.6332
1195.2356
1195.4373
1194.4351
1195.6378
1194.2365
996.0157
995.8483
995.6827
996.1829
996.3505
995.5141
996.5154
1002.5084
995.3459
1008.6659
1008.8346
1014.9843
936.8356
960.8200
979.5070
1202.4260
1202.6277
1202.8275
1210.0197
1210.2158
1210.6229
1493.3203
1493.0649
1493.8174
1492.8175
1494.0679
1494.3192
1492.5619
1503.0565
1512.5468
%
m/z

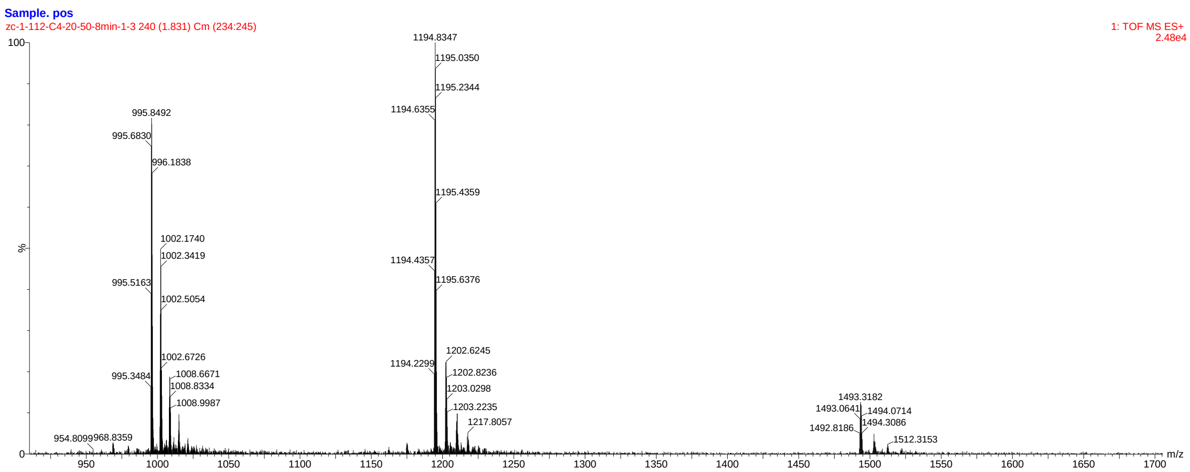
Manα2Manα-SerA9 **8**



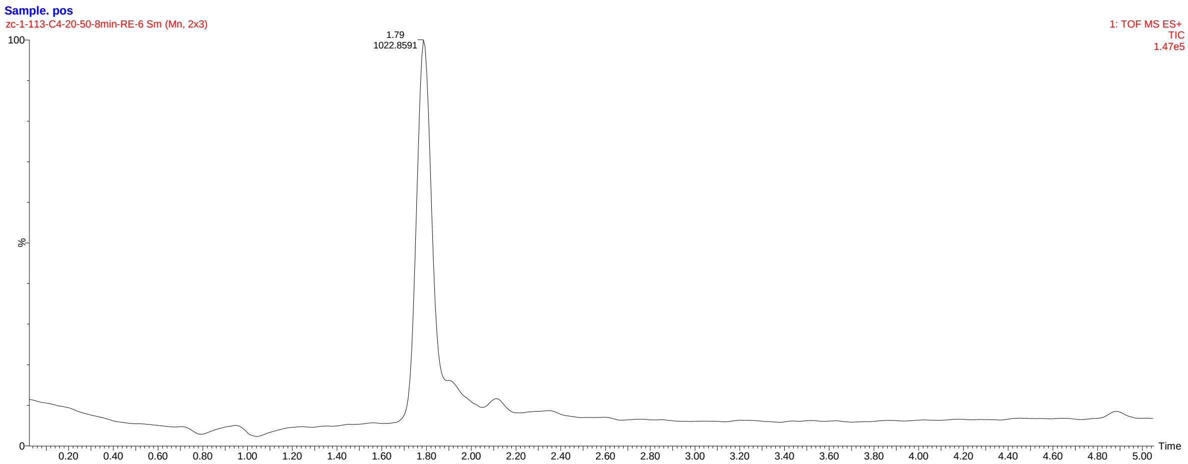


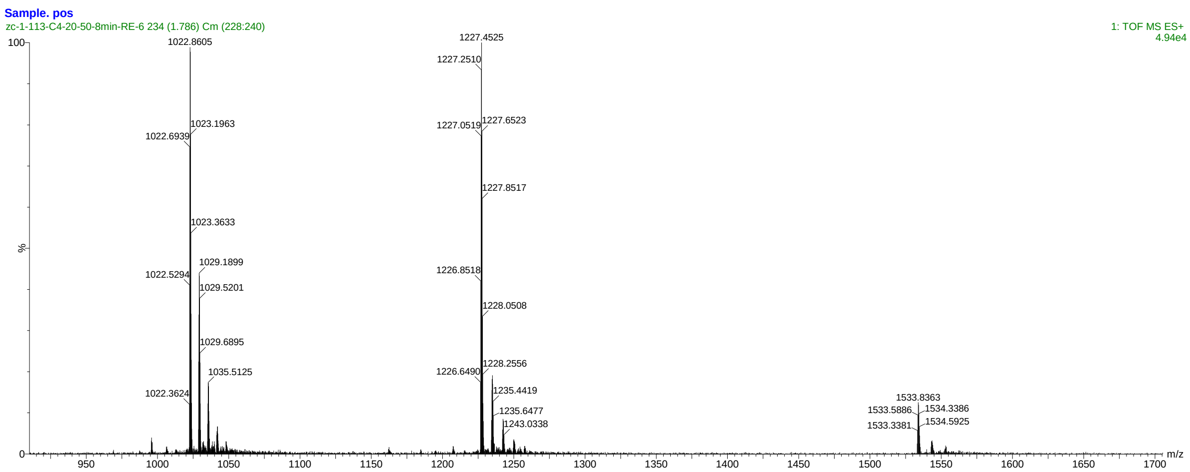
Manα-ThrB27 **9**



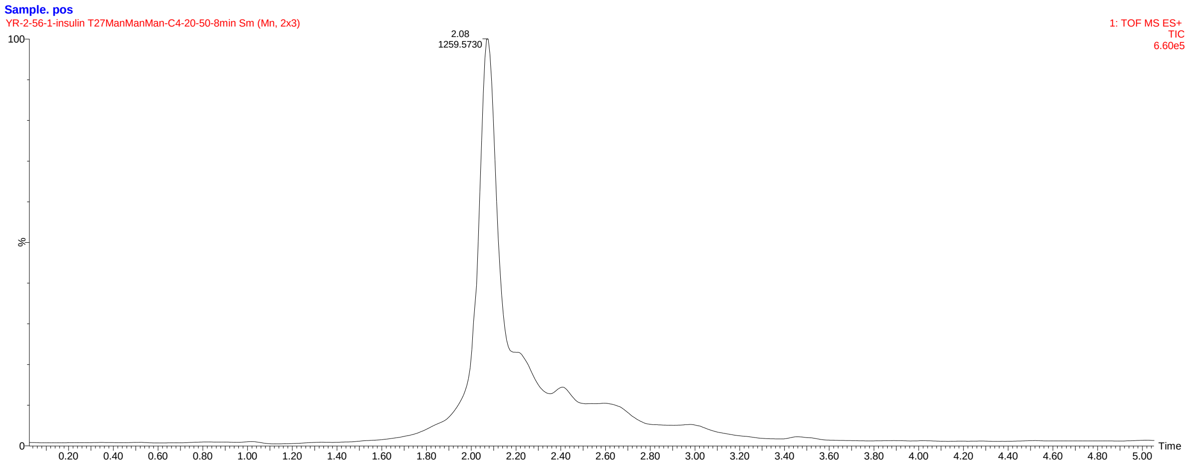


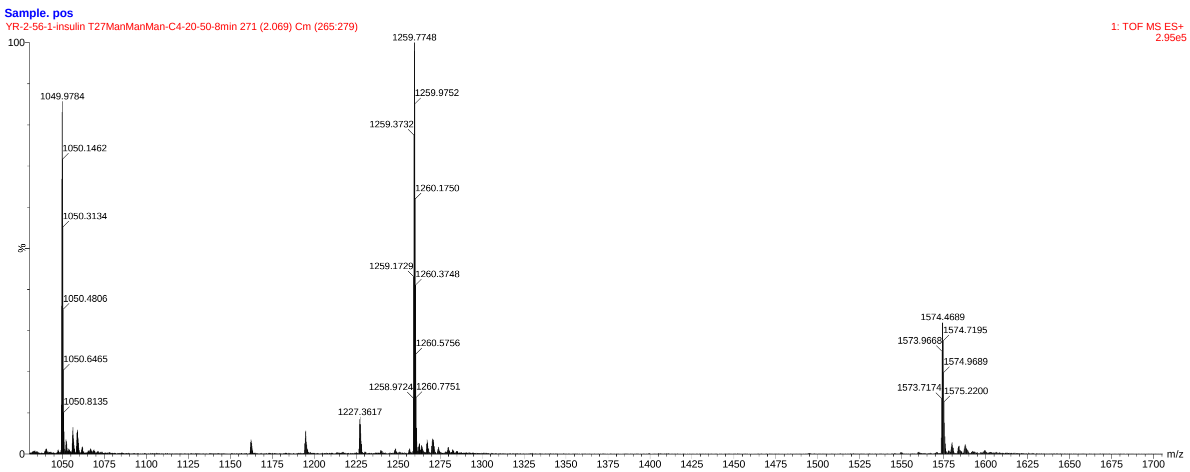
Manα2Manα-ThrB27 **10**



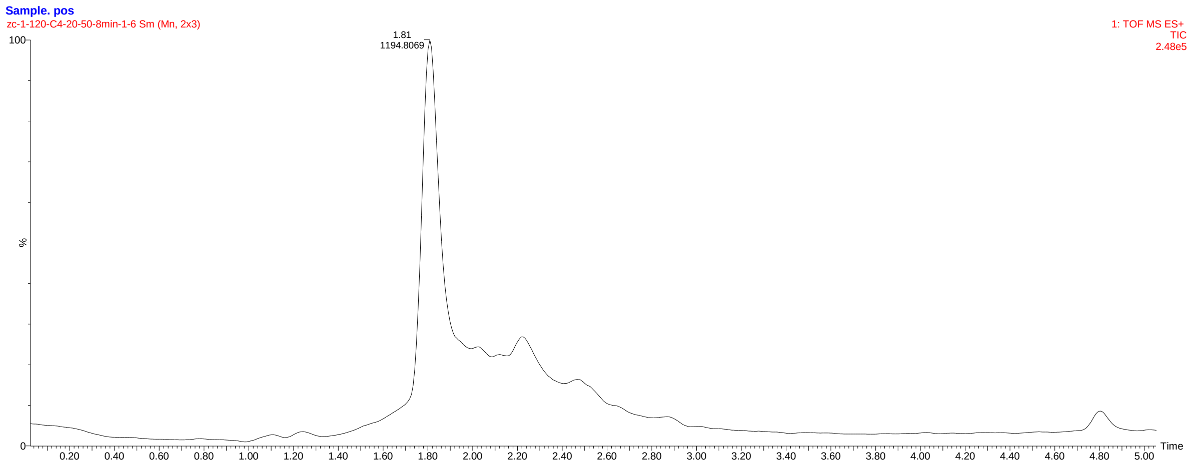


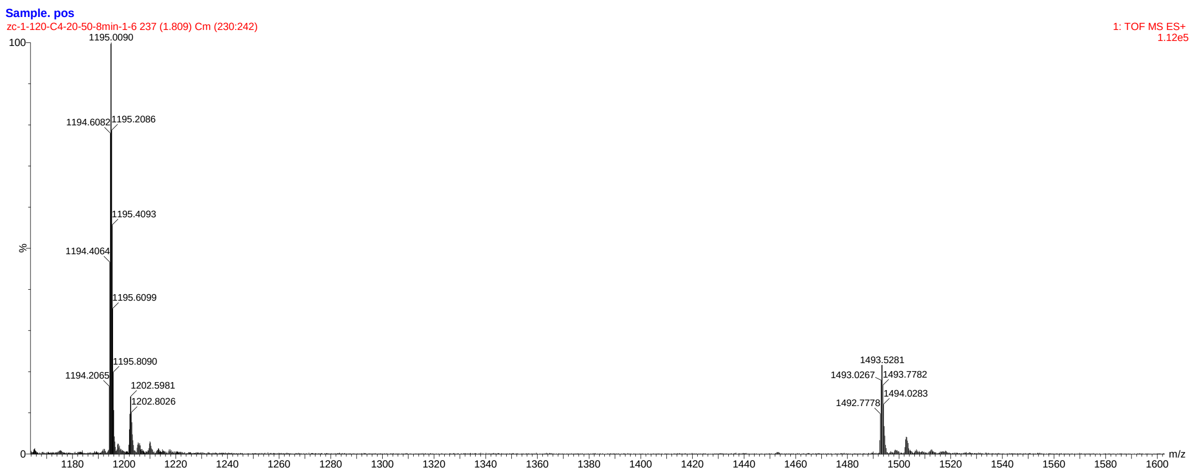
Manα2Manα2Manα-ThrB27 **11**



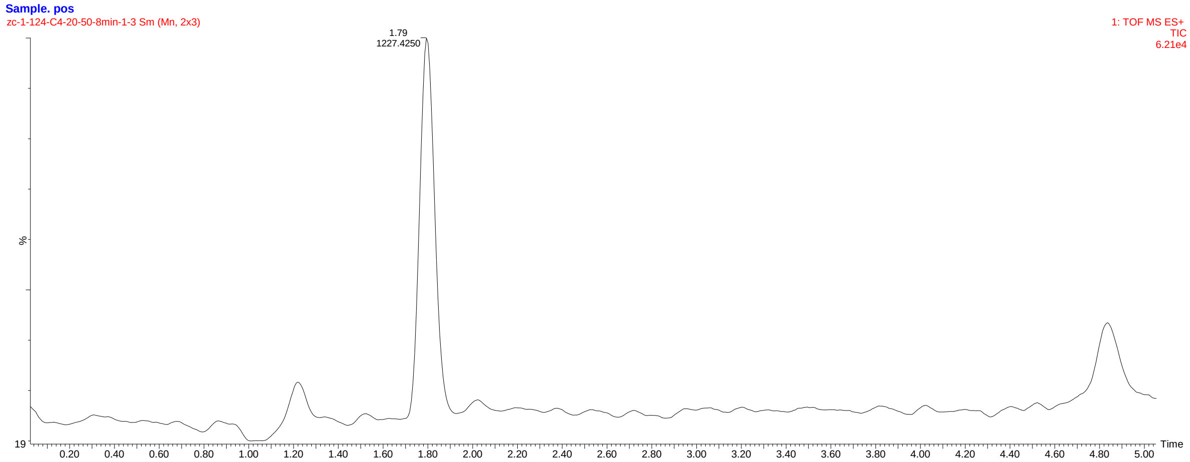


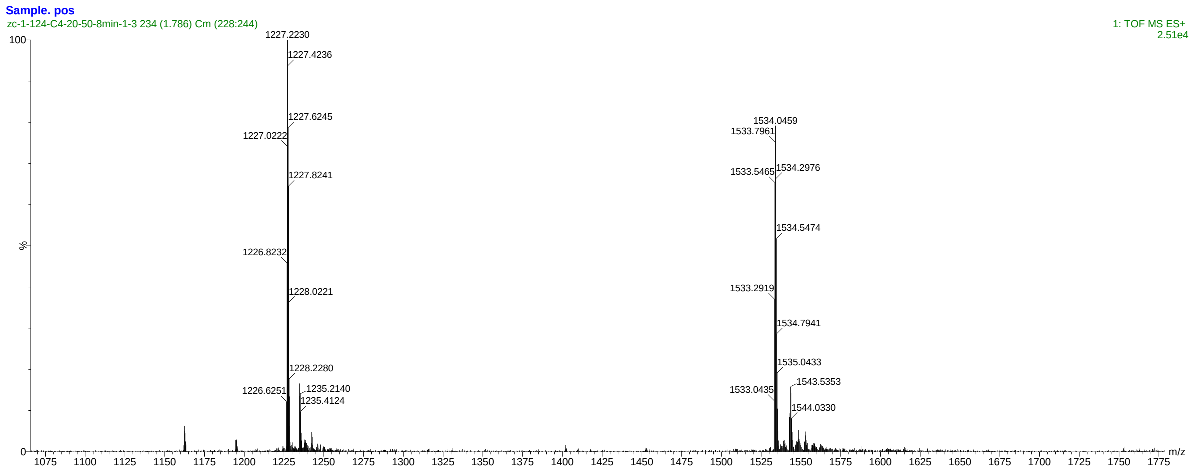
Manα-ThrB30 **12**





Manα2Manα-ThrB30 **13**

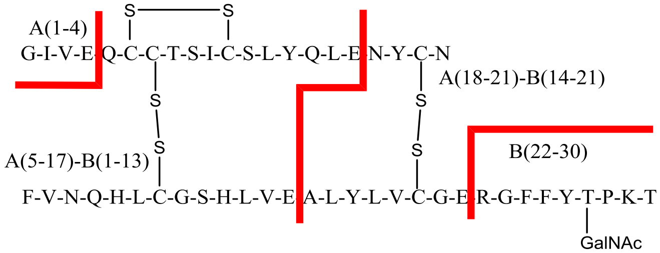


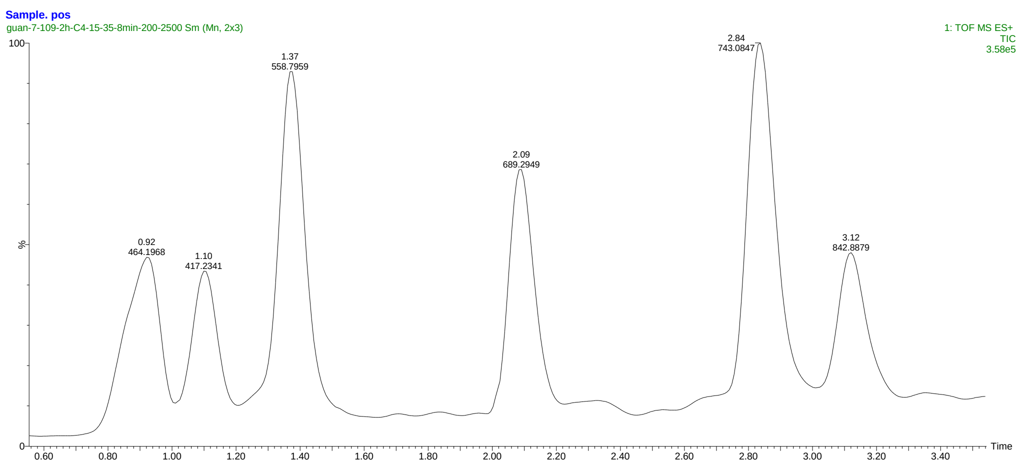


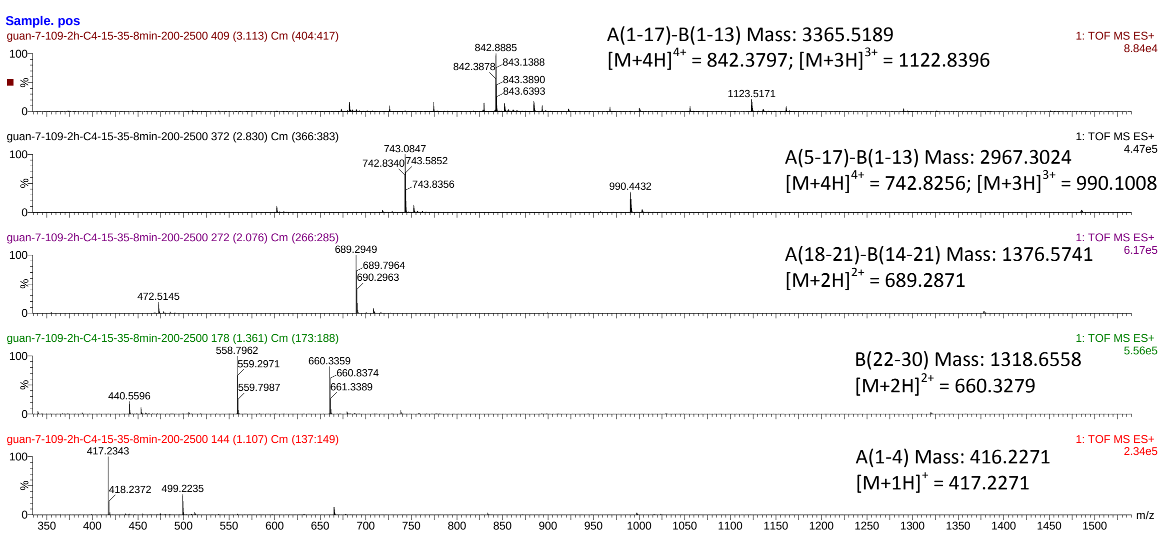
6.5.5 Confirmation of the Disulfide Bond Pattern of the Glycosylated Insulin

Glu-C Digestion –Glycosylated insulins (100 ug) were mixed with 10 ug Glu-C in 100 uL of 50 mM Tris buffer (pH 8.0) at room temperature and allowed to stand for 2 hours before being analyzed by LC-MS.

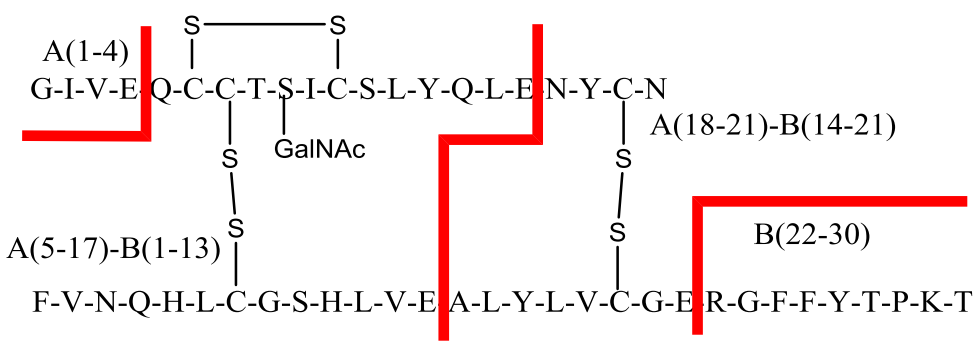
GalNAcα-ThrB27 **5** digested by Glu-C

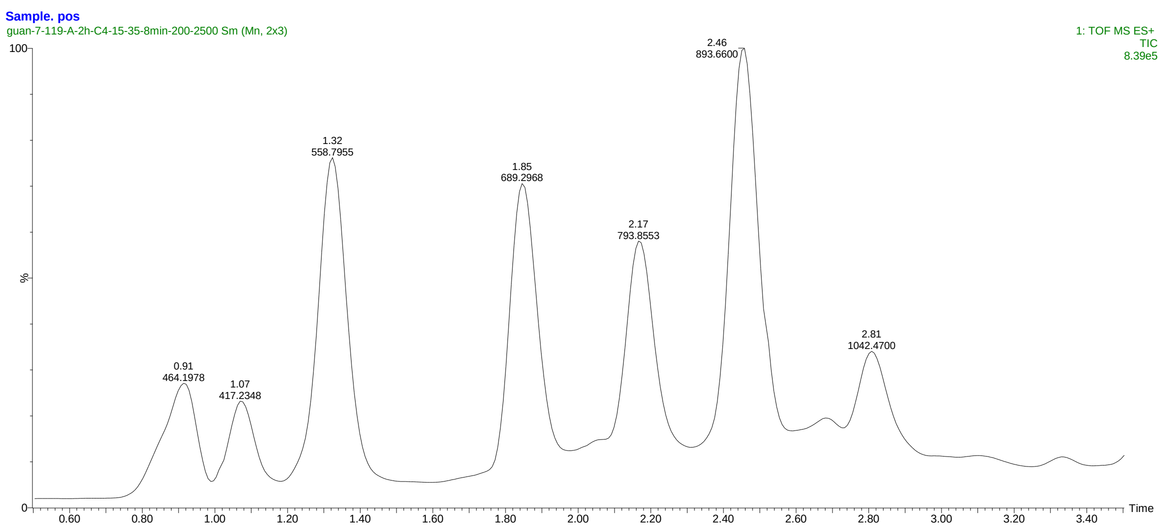


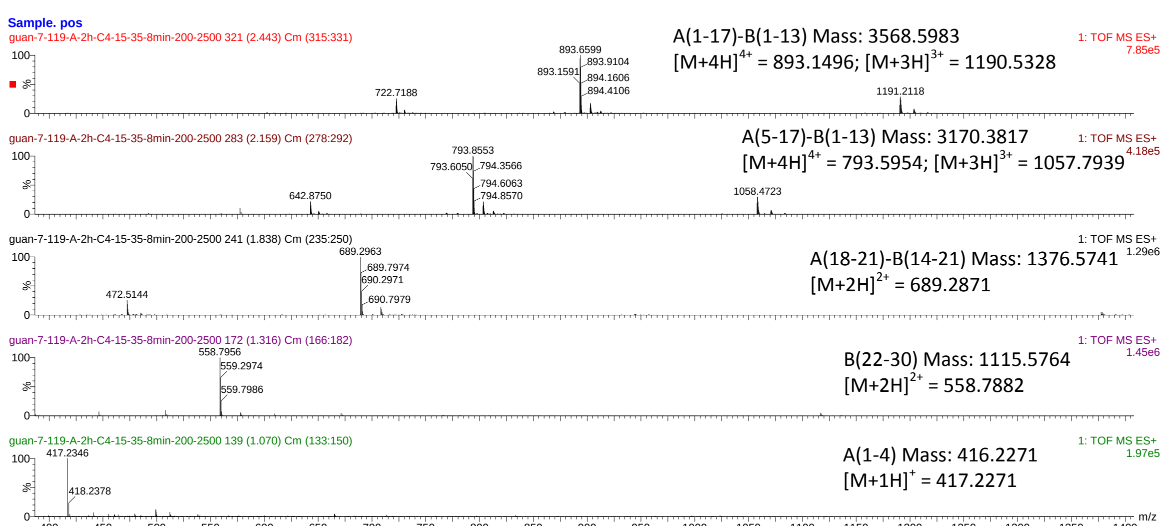




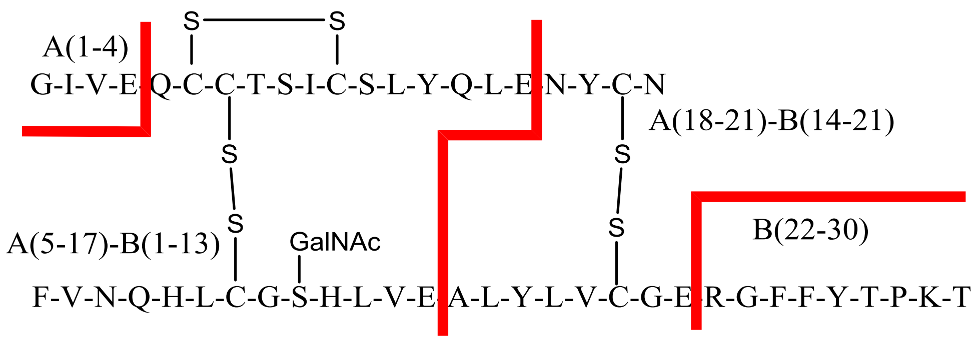
GalNAcα-SerA9 **2** digested by Glu-C

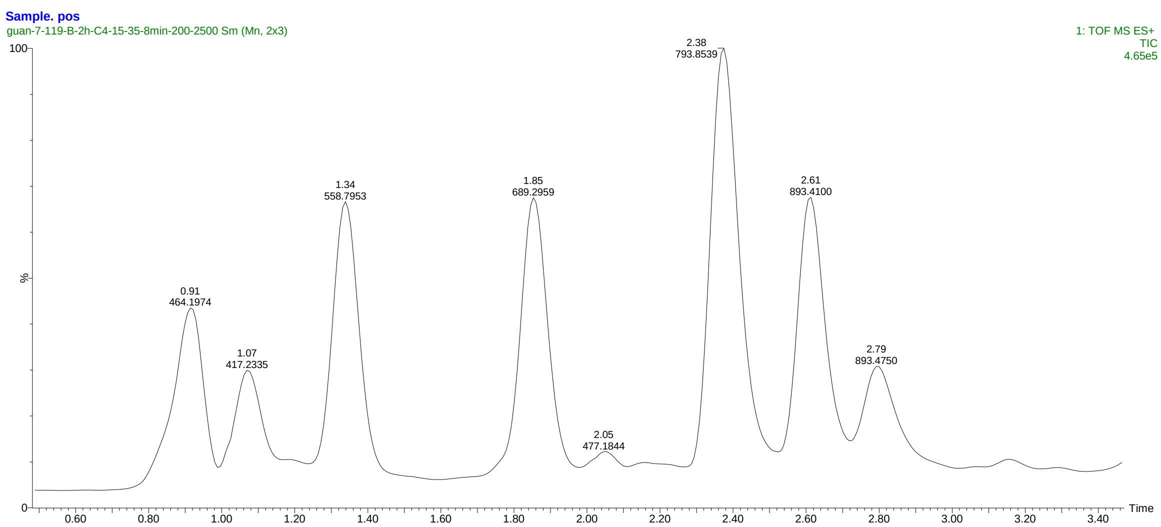


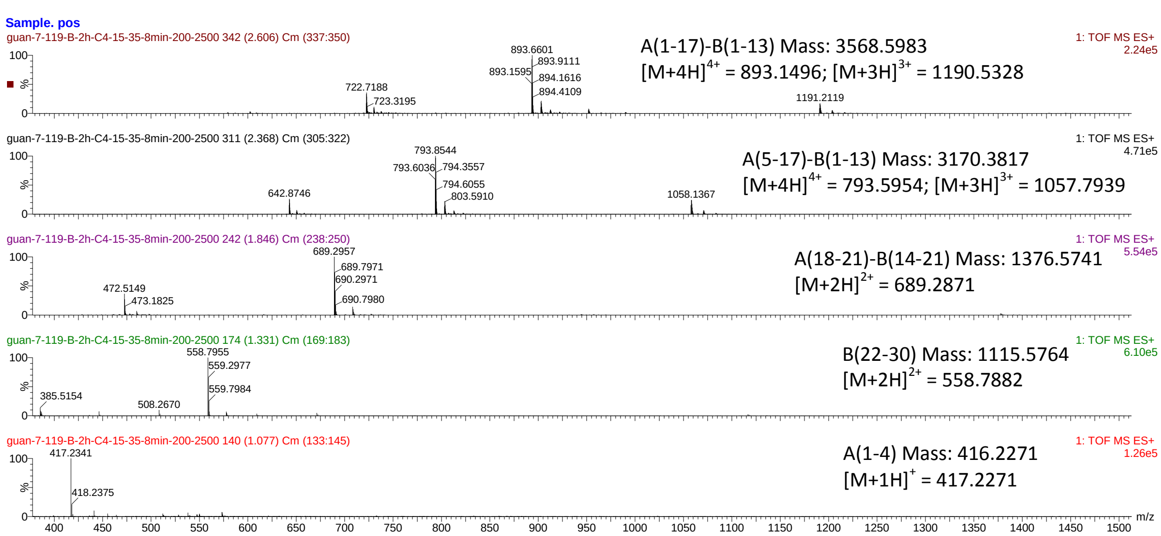




GalNAcα-SerB9 **4** digested by Glu-C







6.5.6 Chymotrypsin Digestion of Insulin

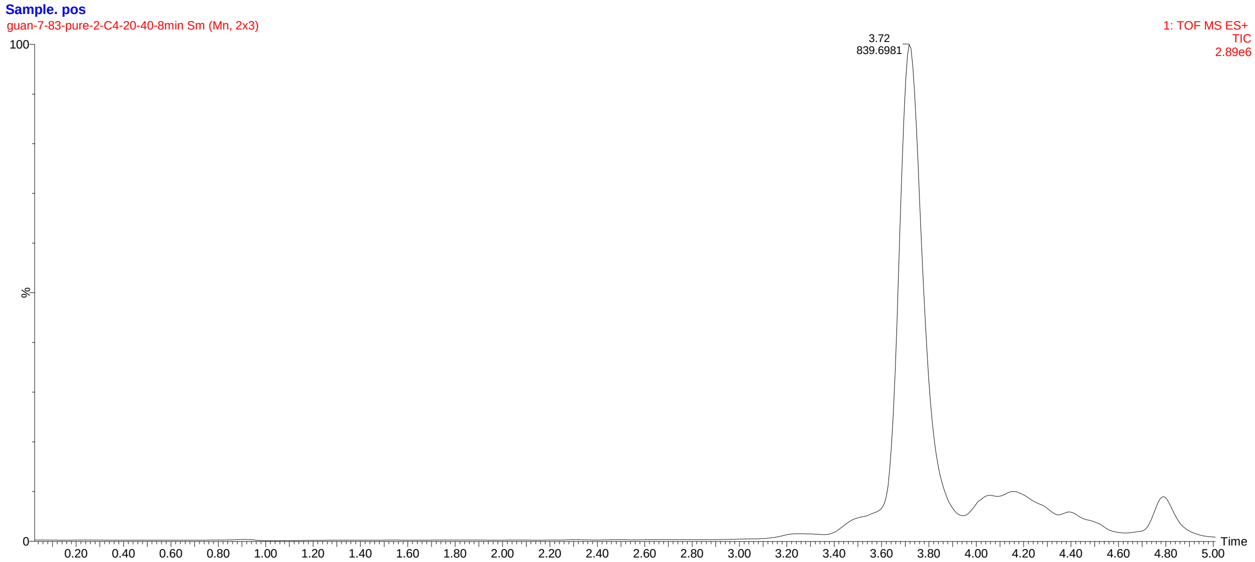
*Chymotrypsin Digestion of glycosylated synthetic insulin glycoforms **2-13***– 76 μL of a 0.5 μg/μL insulin solution was prepared in a buffer composed of 100 m*M* Tris and 1 m*M* $CaCl_2$ adjusted to pH 8.0. Prior to digestion this insulin solution was equilibrated to 37°C for 15 min. Immediately before addition of digestion enzyme, the insulin solution was vortexed for 2 sec and a 1 μL sample was taken as the zero-time sample and immediately added to 9.0 μL of a 0.2 % TFA solution containing 0.055 μg/μL unglycosylated synthetic insulin **1** as an internal standard. Digestion was begun by adding 4 μL of a chymotrypsin stock solution (0.25 μg/μL enzyme in a buffer composed of 100 m*M* Tris and 1 m*M* $CaCl_2$ adjusted to pH 8.0) to reach a final enzyme concentration of 0.0125 μg/μL. The resulting solution was vortexed for 2 sec and incubated at 37°C. After 1 min, 3 min, 5 min, 10 min, 20 min, 30 min, 40 min, 60 min, 90 min, 120 min, and 180 min 1 μL aliquots were removed from the digestion reaction and added to 9.0 μL of a 0.2 % TFA solution containing 0.055 μg/μL unglycosylated synthetic insulin **1** as an internal standard. These samples were vortexed for 2 sec and stored at -20°C until MALDI-TOF MS analysis could be carried out. MALDI-TOF analysis was done according to previous procedures (PNAS paper).

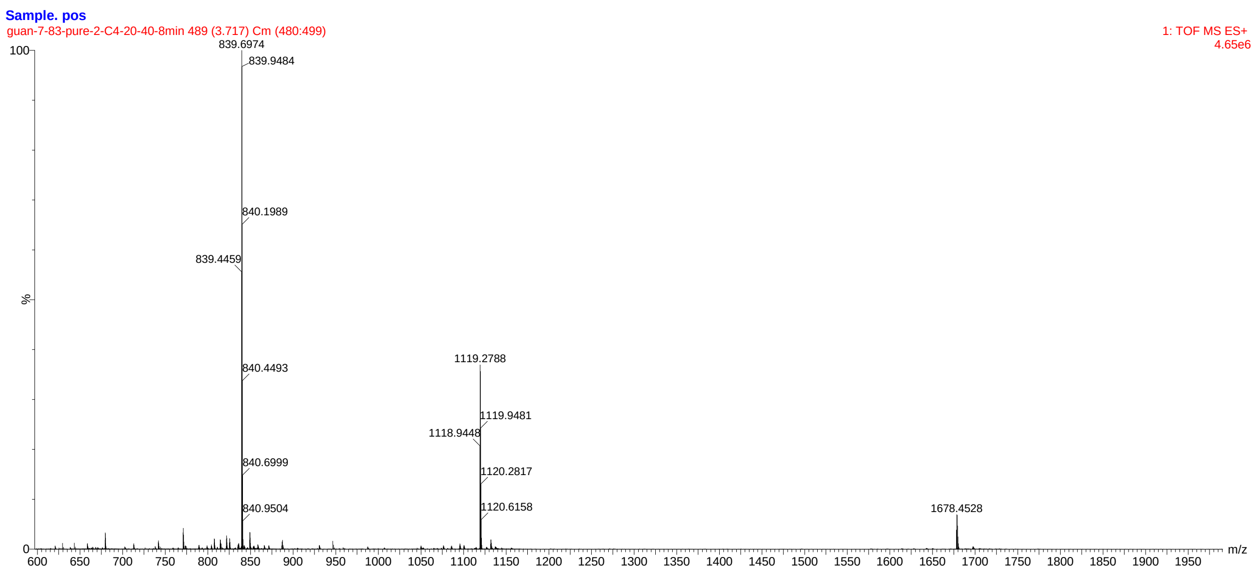
*Chymotrypsin Digestion of unglycosylated synthetic insulin glycoform **1***-Digestion and analysis of **1** was done exactly as described above for **2-13** except glycosylated synthetic insulin **5** was used as the internal standard during quantitative MALDI-TOF MS analysis.

6.5.7 LC-MS Traces and Spectra for GLP-1 analogs 29-34

Unglycosylated GLP-1 (**29**) -

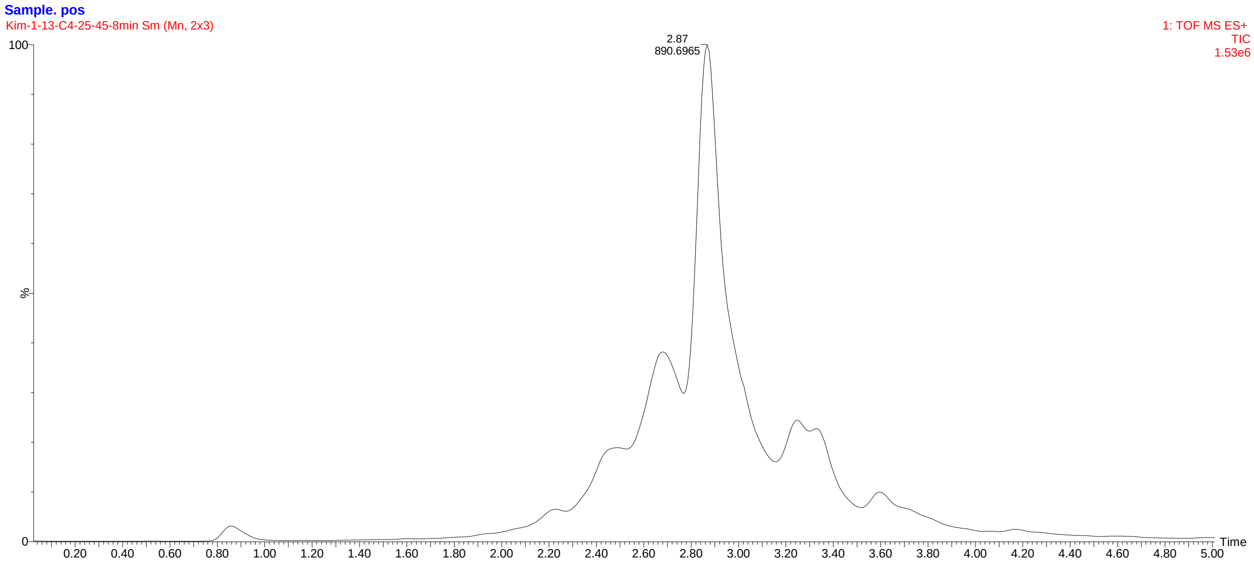
Mass: 3353.6681, $[M+4H]^{4+}$ = 839.4170, $[M+3H]^{3+}$ = 1118.8894, $[M+2H]^{2+}$ = 1677.8341

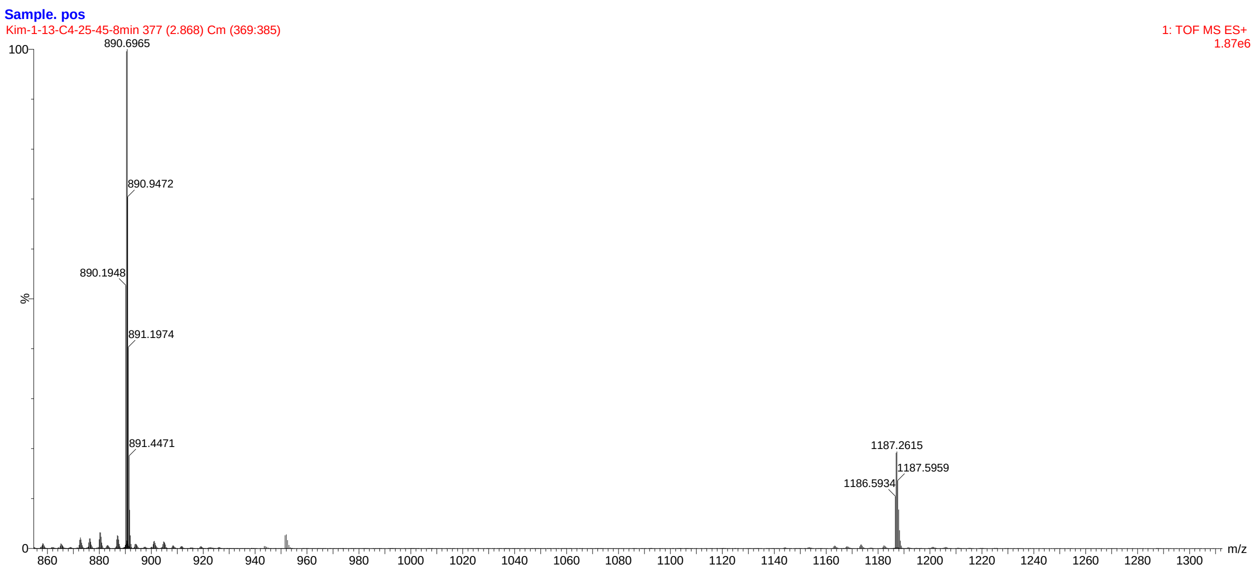




GLP-1 T11ManNAc (**30**) -

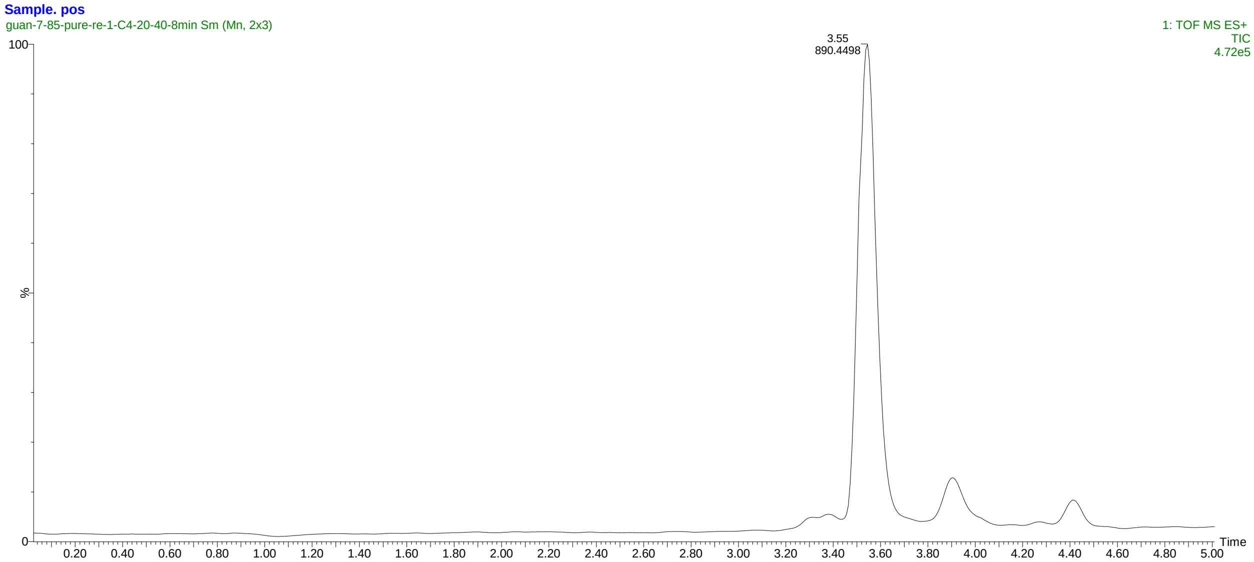
Mass: 3556.7474, $[M+4H]^{4+}$ = 890.1869, $[M+3H]^{3+}$ = 1186.5825

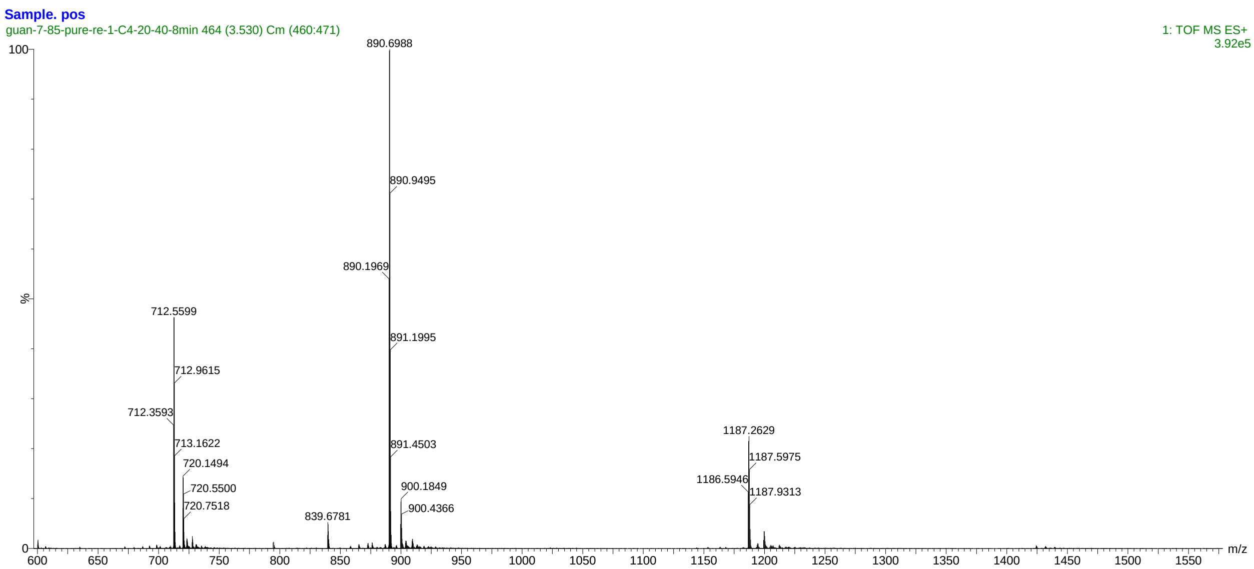




GLP-1 T11GalNAc (**31**) -

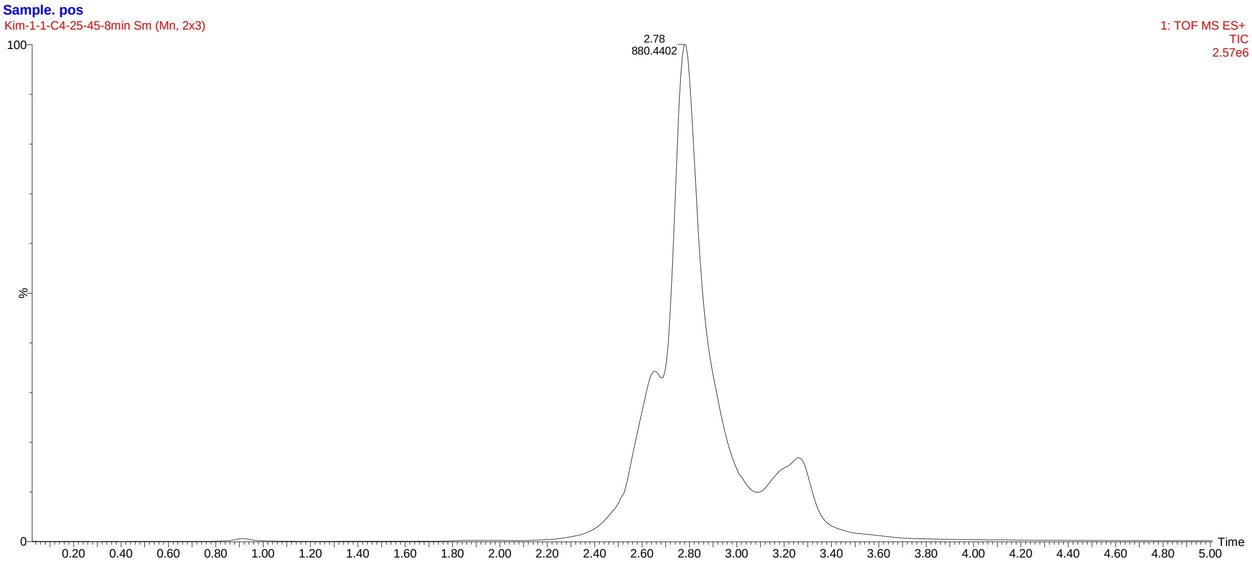
Mass: 3556.7474, $[M+5H]^{5+}$ = 712.3495, $[M+4H]^{4+}$ = 890.1869, $[M+3H]^{3+}$ = 1186.5825

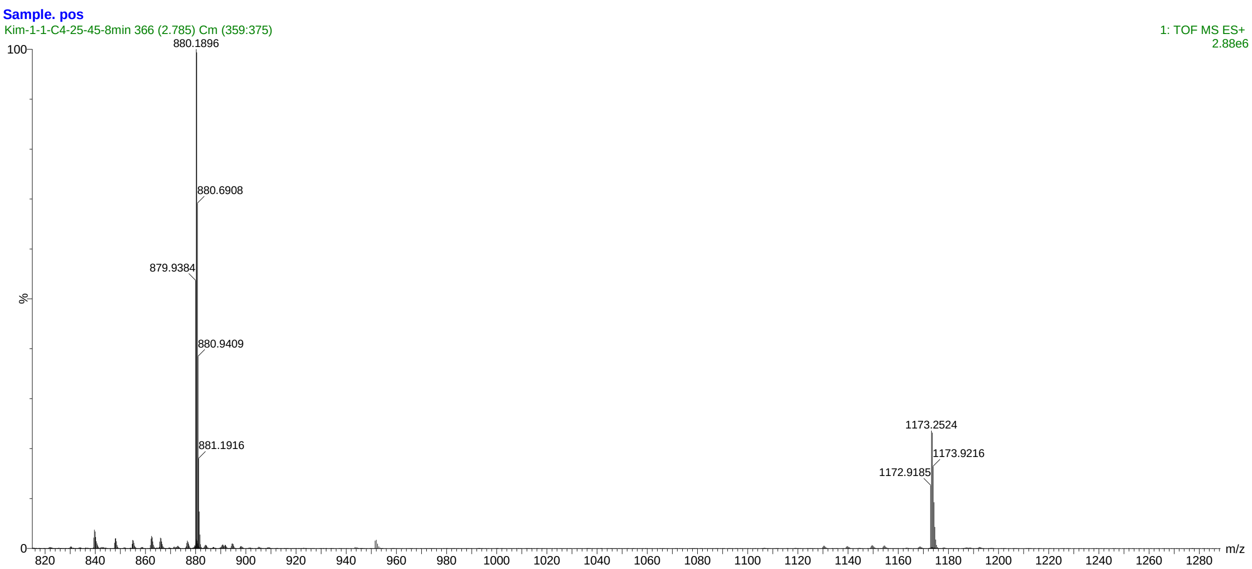




GLP-1 T11Man (**32**) -

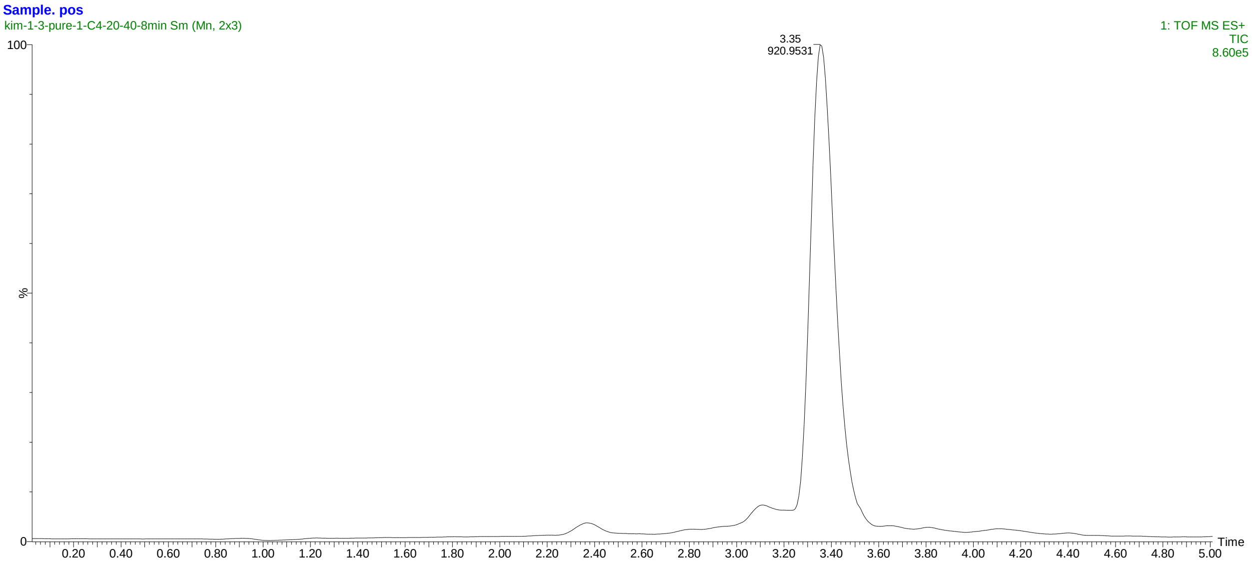
Mass: 3515.7209, $[M+4H]^{4+}$ = 879.9302, $[M+3H]^{3+}$ = 1172.9070

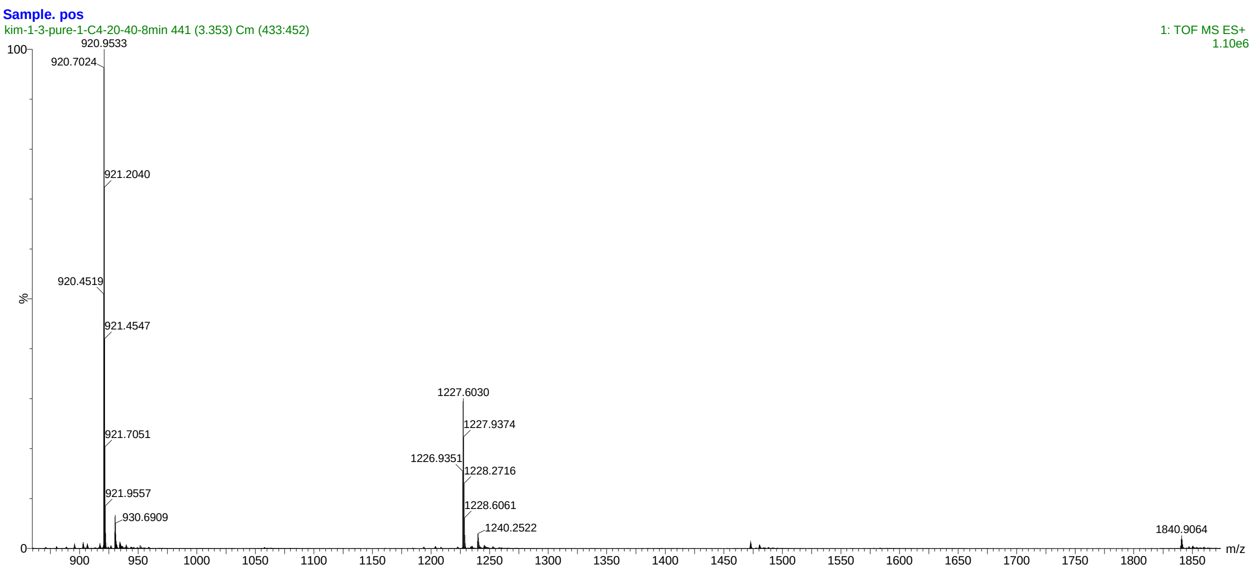




GLP-1 T11ManMan (**33**) -

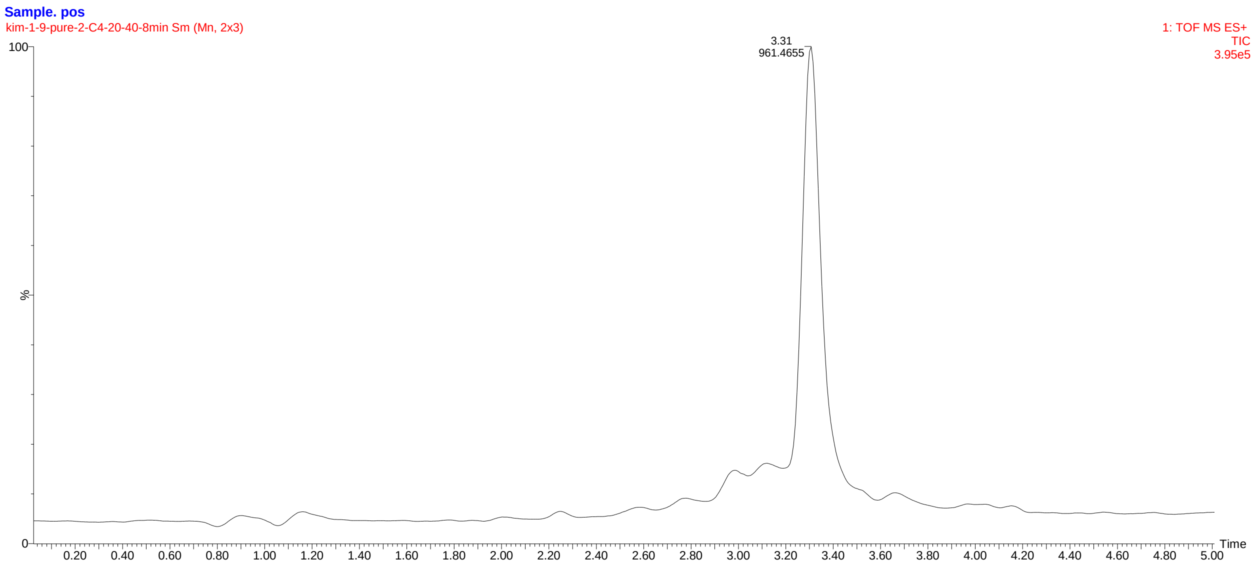
Mass: 3677.7737, $[M+4H]^{4+}$ = 920.4434, $[M+3H]^{3+}$ = 1226.9246, $[M+2H]^{2+}$ = 1839.8869

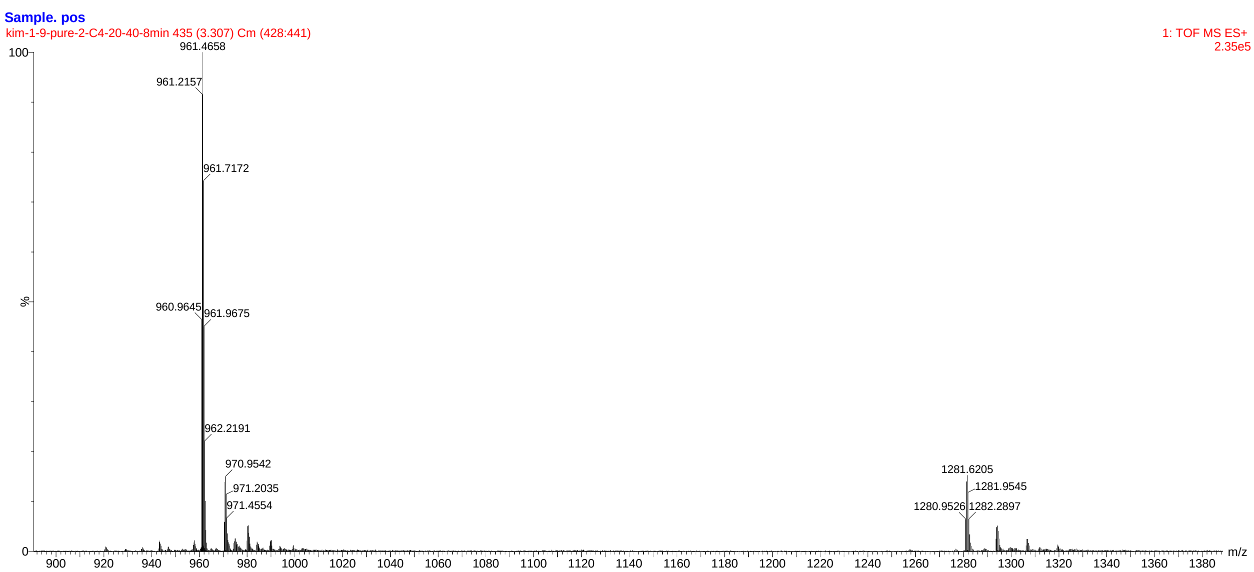




GLP-1 T11ManManMan (**34**) -

Mass: 3839.8265, $[M+4H]^{4+}$ = 960.9566, $[M+3H]^{3+}$ = 1226.9246, $[M+2H]^{2+}$ = 1280.9422





6.5.8 DPP IV digestion

50 µL of a 0.6 µg/µL GLP-1 solution was prepared in a buffer composed of 25 m*M* potassium phosphate, 25 m*M* potassium chloride, 5 m*M* magnesium chloride, adjusted to pH 7.0. Prior to digestion this GLP-1 solution was equilibrated to 37°C for 15 min. Immediately before addition of digestion enzyme, the GLP-

1 solution was vortexed for 2 sec and a 1 μL sample was taken as the zero-time sample and immediately added to 9.0 μL of a 0.2 % TFA solution containing 0.067 μg/μL internal standard GLP-1 molecule. Digestion was begun by adding 1.25 μL of a DPP-IV stock solution (0.04 μg/μL enzyme in a buffer composed of of 25 m*M* potassium phosphate, 25 m*M* potassium chloride, 5 m*M* magnesium chloride, adjusted to pH 7.0) to reach a final enzyme concentration of 0.976 μg/mL. The resulting solution was vortexed for 2 sec and incubated at 37^{o}C. After 1 min, 3 min, 5 min, 10 min, 20 min, 30 min, 40 min, 60 min, 90 min, 120 min, and 180 min 1 μL aliquots were removed from the digestion reaction and added to 9.0 μL of a 0.2 % TFA solution containing 0.067 μg/μL internal standard GLP-1 molecule. These samples were vortexed for 2 sec and stored at -20^{o}C until MALDI-TOF MS analysis could be carried out. MALDI-TOF analysis was done according to previous procedures (PNAS paper). For unglycosylated GLP-1 analog **29**, GLP-1 molecule **31** was used as the internal standard during MS analysis. For GLP-1 analogs **30-32**, GLP-1 molecule **34** was used as the internal standard during MS analysis. For GLP-1 analogs **33** and **34**, the unglycosylated GLP-1 molecule **29** was used as the internal standard during MS analysis.

6.6 References

1. K. Fosgerau and T. Hoffmann, *Drug Discov Today*, 2015, **20**, 122-128.
2. P. Vlieghe, V. Lisowski, J. Martinez and M. Khrestchatisky, *Drug Discov Today*, 2010, **15**, 40-56.
3. D. J. Craik, D. P. Fairlie, S. Liras and D. Price, *Chem Biol Drug Des*, 2013, **81**, 136-147.
4. A. M. Sinclair and S. Elliott, *J. Pharm. Sci.*, 2005, **94**, 1626-1635.
5. R. J. Sola and K. Griebenow, *J. Pharm. Sci.*, 2009, **98**, 1223-1245.
6. V. K. Pawar, J. G. Meher, Y. Singh, M. Chaurasia, B. Surendar Reddy and M. K. Chourasia, *J Control Release*, 2014, **196**, 168-183.
7. M. Peyrot, R. R. Rubin, D. F. Kruger and L. B. Travis, *Diabetes Care*, 2010, **33**, 240-245.
8. M. Dicker and R. Strasser, *Expert Opin Biol Ther*, 2015, **15**, 1501-1516.
9. K. Yoshida, B. Yang, W. Yang, Z. Zhang, J. Zhang and X. Huang, *Angew Chem Int Ed Engl*, 2014, **53**, 9051-9058.
10. J. P. Giddens and L. X. Wang, *Methods Mol Biol*, 2015, **1321**, 375-387.
11. H. Yu, K. Lau, Y. Li, G. Sugiarto and X. Chen, *Curr Protoc Chem Biol*, 2012, **4**, 233-247.
12. L. Li, Y. Liu, C. Ma, J. Qu, A. D. Calderon, B. Wu, N. Wei, X. Wang, Y. Guo, Z. Xiao, J. Song, G. Sugiarto, Y. Li, H. Yu, X. Chen and P. G. Wang, *Chem Sci*, 2015, **6**, 5652-5661.
13. R. Chen and T. J. Tolbert, *J Am Chem Soc*, 2010, **132**, 3211-3216.

14. A. Fernandez-Tejada, J. Brailsford, Q. Zhang, J. H. Shieh, M. A. Moore and S. J. Danishefsky, *Top Curr Chem*, 2015, **362**, 1-26.
15. H. C. Hang and C. R. Bertozzi, *Bioorg Med Chem*, 2005, **13**, 5021-5034.
16. T. Buskas, S. Ingale and G. J. Boons, *Glycobiology*, 2006, **16**, 113R-136R.
17. K. M. Koeller and C. H. Wong, *Nat Biotechnol*, 2000, **18**, 835-841.
18. R. J. Sola and K. Griebenow, *BioDrugs*, 2010, **24**, 9-21.
19. L. Chen, M. R. Drake, M. G. Resch, E. R. Greene, M. E. Himmel, P. K. Chaffey, G. T. Beckham and Z. Tan, *Proc Natl Acad Sci U S A*, 2014, **111**, 7612-7617.
20. X. Guan, P. K. Chaffey, C. Zeng, E. R. Greene, L. Chen, M. R. Drake, C. Chen, A. Groobman, M. G. Resch, M. E. Himmel, G. T. Beckham and Z. Tan, *Chem Sci*, 2015, **6**, 7185 - 7189.
21. J. L. Price, E. K. Culyba, W. Chen, A. N. Murray, S. R. Hanson, C. H. Wong, E. T. Powers and J. W. Kelly, *Biopolymers*, 2012, **98**, 195-211.
22. E. K. Culyba, J. L. Price, S. R. Hanson, A. Dhar, C. H. Wong, M. Gruebele, E. T. Powers and J. W. Kelly, *Science*, 2011, **331**, 571-575.
23. D. R. Owens, *Nat Rev Drug Discov*, 2002, **1**, 529-540.
24. A. J. Garber, *Diabetes Care*, 2011, **34 Suppl 2**, S279-284.
25. Q. Hua, *Protein Cell*, 2010, **1**, 537-551.
26. F. Sanger, *Annu Rev Biochem*, 1988, **57**, 1-28.
27. Y. C. Du, Y. S. Zhang, Z. X. Lu and C. L. Tsou, *Sci Sin*, 1961, **10**, 84-104.
28. D. V. Goeddel, D. G. Kleid, F. Bolivar, H. L. Heyneker, D. G. Yansura, R. Crea, T. Hirose, A. Kraszewski, K. Itakura and A. D. Riggs, *Proc Natl Acad Sci U S A*, 1979, **76**, 106-110.
29. T. L. Blundell, J. F. Cutfield, E. J. Dodson, G. G. Dodson, D. C. Hodgkin and D. A. Mercola, *Cold Spring Harb Symp Quant Biol*, 1972, **36**, 233-241.
30. D. F. Berenson, A. R. Weiss, Z. L. Wan and M. A. Weiss, *Ann N Y Acad Sci*, 2011, **1243**, E40-E54.
31. D. Ryan and A. Acosta, *Obesity (Silver Spring)*, 2015, **23**, 1119-1129.
32. D. Yabe and Y. Seino, *Expert Review of Endocrinology & Metabolism*, 2014, **9**, 659-670.
33. W. Chen, S. Enck, J. L. Price, D. L. Powers, E. T. Powers, C. H. Wong, H. J. Dyson and J. W. Kelly, *J Am Chem Soc*, 2013, **135**, 9877-9884.
34. B. F. Choonara, Y. E. Choonara, P. Kumar, D. Bijukumar, L. C. du Toit and V. Pillay, *Biotechnol Adv*, 2014, **32**, 1269-1282.
35. J. Renukuntla, A. D. Vadlapudi, A. Patel, S. H. Boddu and A. K. Mitra, *Int J Pharm*, 2013, **447**, 75-93.
36. V. Balamuralidhara, T. M. Pramodkumar, N. Srujana, M. P. Venkatesh, N. V. Gupta, K. K.L. and H. V. Gangadharappa, *Am J Drug Discovery Dev*, 2011, **1**.
37. T. J. Gibson and R. M. Murphy, *Protein Sci*, 2006, **15**, 1133-1141.
38. P. Fonte, F. Araujo, S. Reis and B. Sarmento, *J Diabetes Sci Technol*, 2013, **7**, 520-531.
39. B. J. Bruno, G. D. Miller and C. S. Lim, *Ther Deliv*, 2013, **4**, 1443-1467.
40. A. Belgi, M. A. Hossain, G. W. Tregear and J. D. Wade, *Immunol Endocr Metab Agents Med Chem*, 2011, **11**, 40-47.
41. F. Liu, E. Y. Luo, D. B. Flora and A. R. Mezo, *Angew Chem Int Ed Engl*, 2014, **53**, 3983-3987.
42. E. Ciszak, J. M. Beals, B. H. Frank, J. C. Baker, N. D. Carter and G. D. Smith, *Structure*, 1995, **3**, 615-622.
43. R. J. Schilling and A. K. Mitra, *Pharm Res*, 1991, **8**, 721-727.
44. S. A. Berkowitz, *AAPS J*, 2006, **8**, E590-605.

45. E. M. van Dam, R. Govers and D. E. James, *Molecular endocrinology*, 2005, **19**, 1067-1077.
46. R. Govers, A. C. Coster and D. E. James, *Mol Cell Biol*, 2004, **24**, 6456-6466.
47. X. Chang, D. Keller, S. I. O'Donoghue and J. J. Led, *FEBS Lett*, 2002, **515**, 165-170.
48. J. J. Holst, *Physiol Rev*, 2007, **87**, 1409-1439.
49. M. Zhao, Z. L. Wan, L. Whittaker, B. Xu, N. B. Phillips, P. G. Katsoyannis, F. Ismail-Beigi, J. Whittaker and M. A. Weiss, *J Biol Chem*, 2009, **284**, 32178-32187.
50. A. S. De Groot and D. W. Scott, *Trends Immunol*, 2007, **28**, 482-490.
51. B. L. Bray, *Nat Rev Drug Discov*, 2003, **2**, 587-593.