# Glycosidases as an analytical tool in glycomics assays

Rebello, O.D.

# Chapter 3

## Crystal structure of a GH20 *N*-acetylglucosaminidase with specificity for bisecting *N*-acetylglucosamines on *N*-glycans

.

# Abstract

Research into how the human gut microbiota degrades different glycans and polysaccharides has led to the characterisation of numerous Carbohydrate-Active enZymes (CAZymes) of huge diversity in terms of specificity. Recently, we described the set of CAZymes employed by a prominent member of the adult human gut (*Bacteroides thetaiotaomicron*) to degrade mammalian complex *N*-glycans [1]. This report follows up on one of these enzymes, BT 0456, with a molecular structure and discusses potential applications within glycoanalytics.

# Introduction

Post-translational glycan modifications are important for a variety of functions and processes in proteins. For instance, these glycans can provide protection of the protein from degradation and can be the ligands in many cellular processes [2, 3]. Changes to the composition of particular glycan decorations are characterised in a number of disease states, which means that they can be used as a biomarker for those diseases [4, 5]. To achieve this, a variety of glycoanalytic techniques can be applied to assess glycan composition [6, 7]. One way in which these approaches are being developed involves the application of enzymes that are highly specific towards particular sugars residues and their linkages, to characterise glycans [8-11]. These enzyme tools make current glycoanalytic techniques, such as liquid chromatography (LC) for fluorescently labelled glycans, much more powerful by adding an extra level of accurate characterisation [12]. Application of these enzymes can solve or side-step any issues with poor chromatography separation or mass spectrometry (MS) fragmentation identification, which is common with highly complex or low concentration samples [11, 13]. Therefore, the characterisation and quantification of glycan biomarkers that are associated with different diseases will be greatly helped by the discovery of glycoenzymes of different specificities.



*N*-glycans are common post-translational modifications and their composition varies between type of organism, tissue, age, and disease. One sugar decoration of interest is the bisecting *N*-acetylglucosamine (GlcNAc) on *N*-glycans, which is linked to the mannose core through a β1,4-linkage. Changes in the abundance of this sugar decoration is associated with several immunological diseases and cancers, for example, hepatitis B-induced liver cirrhosis [14], rheumatoid arthritis [15], type 2 diabetes [16], colorectal cancer [13], ovarian cancer [17], and prostate cancer [9]. Aside from particular diseases, the bisecting GlcNAc is also associated with an individual's metabolic age, which is indicative of overall health [19]. This sugar decoration is usually present in low abundance of approximately 6% in the human plasma *N*-glycome and currently is identified through MS/MS fragmentation. Having additional tools such as exoglycosidase that specifically cleaves these linkages will greatly improve the characterisation of this glycosylation as a potential disease biomarker [18].

Here we report the crystal structure of a glycoside hydrolase (GH) from family 20, BT 0456, originating from a human gut microbe, Bacteroides thetaiotaomicron. These are typically exo-acting β-GlcNAc'ases and BT 0456 has previously been characterised as removing the bisecting GlcNAc when all the galactose and sialic acid residues have been removed [1]. It is also capable of removing one of the antennary GlcNAc sugars from *N*-glycans, but this is not always to completion which makes this specificity inconclusive. Previous data also suggests that the α1,6-fucose commonly found on mammalian complex *N*-glycans does not prevent its activity [1].

# Materials and Methods

## Cloning, expression and purification of recombinant proteins.

DNA encoding the appropriate genes (excluding the signal sequences) was amplified from genomic DNA using appropriate primers and cloned into pET28b (Novagen) using NheI-XhoI restriction sites. These plasmids accommodated a HIS-tags which was encoded into the N-terminus of the protein. The recombinant plasmids were transformed into TUNER (Novagen) cells by heat shock treatment and then selected by culturing in 1mL of LB broth containing 10µg/mL kanamycin at 32 ℃ whilst shaking at 180 rpm. These cultures were then scaled up to 1 L cultures which were grown to mid-exponential phase in 2 L baffled flasks. Isopropyl B-D-thiogalactopyranoside (IPTG) was also added to these cultures at a final concentration of 0.2mM. These cells were then incubated for 16 hours at 16 ℃ on a rotary shaker table at 150 rpm. The cells were collected by ultracentrifugation and the cell pellets lysed by sonication on an ice bath. The recombinant His-tagged protein was purified from the cell free extracts using immobilized metal-affinity chromatography (Talon resin, Clontech) as previously described [1]. Finally, the purity and size of the proteins were analysed using SDS-PAGE and their concentration were determined by absorbance at 280nm (NanoDrop 2000c, Thermo Scientific) and their molar extinction coefficients.

## Purification of proteins for crystallization.

The metal-affinity purified proteins were purified by size-exclusion chromatography using a HiLoad Superdex 200 pg on an AKTA Pure FPLC system (Cytiva). The purity of the fractions was determined using SDS-PAGE and those fraction of satisfactory high purity were pooled and concentrated to 10mg/mL.

## Crystallization.

BT 0456 was initially screened using commercial kits (Molecular Dimensions and Hampton Research). The drops, composed of 0.1 µL or 0.2 µL of protein solution plus 0.1 µL of reservoir solution, were set up a Mosquito crystallization robot (SPT Labtech). The sitting drop method was used and the plates were incubated at 20 °C. The crystallisation condition was in 100 mM Hepes pH 7.5, 10 % PEG 8000 and 5 % ethylene glycol. Sample were cryo-protected with the addition of 20 % PEG 400 to the condition and flash-cooled in liquid nitrogen.

## Data collection, structure solution, model building, refinement and validation.

Diffraction data were collected at the synchrotron beamline I03 of Diamond light source (Didcot, UK) at a temperature of 100 K. The data set was integrated with XDS [20] and XIA2 [21]. The space group was confirmed with Pointless [22] and the data sere scaled with Aimless [23]. The phase problem was solved by experimental phasing selenium SAD using ShelX [24]. The initial model was generated with the CCP4cloud [25] task CCP4build [25, 26]. The model was refined with refmac [27] and manual model building with COOT [28]. The final model was

validated with MolProbity [29] and figures were made with PyMol [30]. Validation was done with Coot and Molprobity. Other program used were from the CCP4 suite [25].

## Bioinformatics.

The CAZy database (www.cazy.org) was used as the main reference for carbohydrate-active enzyme activity. Dali [31] and PDBefold [32] were used to carry out structural homology searches of protein modules.

# Results and discussion

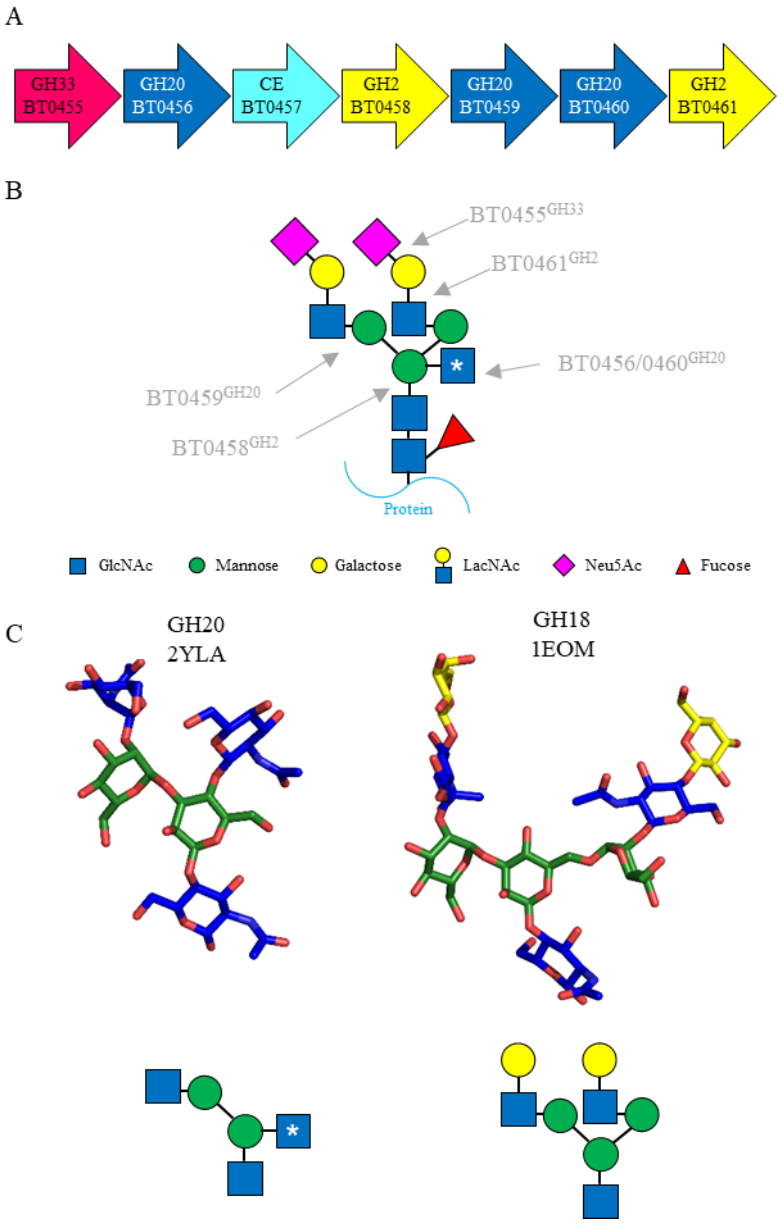


**Figure 1. Degradation of mammalian complex *N*-glycans by Bacteroides thetaiotaomicron.** **(A)** A schematic of the CAZyme loci containing BT 0456. CE: carbohydrate esterase. **(B)** The core *N*-glycan pentasaccharide consists of two GlcNAc and three mannose sugars. In the case of mammalian complex *N*-glycans, this pentasaccharide is built on with *N*-acetyl-D-lactosamine (LacNAc) disaccharides, which are themselves commonly decorated with sialic acids. A biantennary complex *N*-glycan is shown here, but up to four antennas are possible. Core α1,6-fucose decoration is also common for mammalian complex *N*-glycans. The bisecting GlcNAc is always linked through a β1,4-linkage (asterisk). The linkages where the CAZymes act are indicated. **(C)** The structures of two *N*-glycan fragments taken from two crystal structures to provide an indication of the steric hindrance different enzymes may face when degrading complex *N*-glycans.

β-GlcNAc'ases are present in glycoside hydrolase (GH) families of 3, 20, 73, 84 and 85 and BT 0456 belongs to GH20 (Figure 1) [33]. To explore the specificity shown by BT 0456, its structure was solved to 2.10 Å (Figure 2). Due to the lack of adequate homologous search models (highest identify of primary sequence was 26.3%), molecular replacement strategies were not used to solve for the phase problem. Therefore, selenomethionine derived BT 4056 protein samples were used for experimental phasing with the single anomalous diffraction (SAD) method using selenium atoms. Furthermore, attempts at co-crystallising BT 0456 with substrates was unsuccessful.

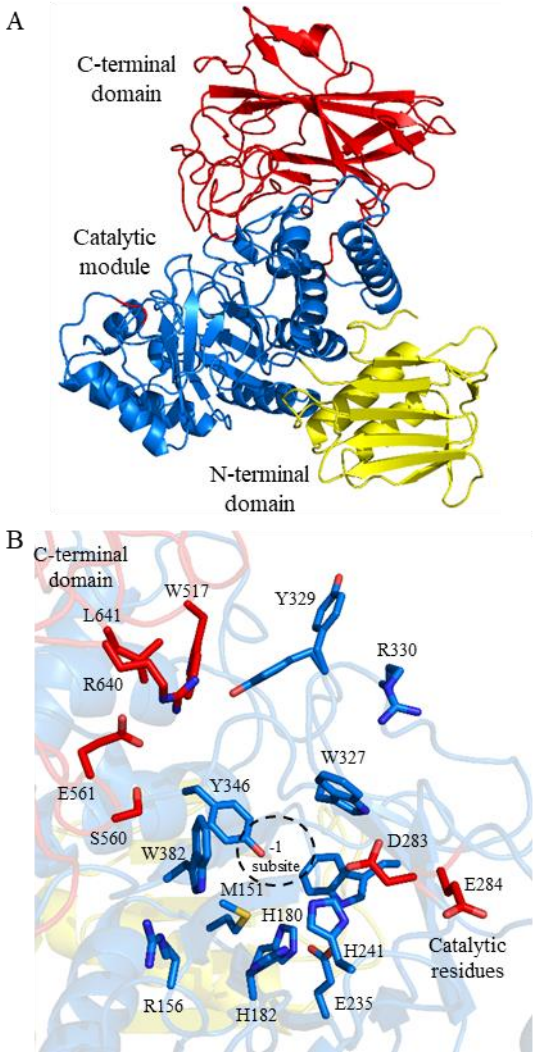


**Figure 2. Crystal structure of BT0456. (A)** The overall structure of BT0456 shown as a cartoon. The N-terminal domain, catalytic module, and the C-terminal domain are yellow, blue, and red, respectively. **(B)** The active site of BT 0456. The residues forming the glycan binding site are shown as sticks, the -1 subsite is encircled, and the catalytic residues are shown in red.

BT 0456 consist of three domains: a N-terminal domain consisting of a zincin-like fold, the catalytic domain is a (β/α)8 barrel fold typical of this family, and a C-terminal domain (Figure 2a). Crystal structures of other GH20 family members show that this N-terminal domain is a common feature, but additional accessory domains are less frequent amongst currently available crystal structures (Figure 3).

GH20 Family members typically have a HXGG(DE) catalytic motif, but in the case of BT 0456 this is HXGT(DE) [34, 35]. Interestingly, the general acid/base E284 is facing away from the active site (Figure 2B). This suggests flexibility in the active site and an element of induced fit upon substrate binding. Interestingly, the equivalent Glu from a Streptococcus gordonii GH20 crystal structure is also flipped away from the active site [34].

The residues surrounding the -1 subsite are largely the same when compared to other GH20 family members. The structure of BT 0459 from the same locus revealed a very open active site, which reflected its broad activity against GlcNAc substrates [1]. In comparison to this, BT 0456 has Y329 and R330 from the catalytic module contributing to the binding region in addition to residues provided by the C-terminal domain. These additional residues are likely driving the specificity of BT 0456 (Figure 2B).

There are two GH20 family members that have been previously characterised as having activity against complex *N*-glycans - GH20B [1, 34] and C [36] from Streptococcus pneumoniae. Both enzymes have specificity towards the antennary GlcNAcs and do not remove the bisecting GlcNAc. However GH20B can accommodate the bisecting GlcNAc to remove the antennary whereas this is not possible for GH20C. The specificity of these enzymes is conferred by residues from the catalytic module, whereas specificity drivers for BT 0456 activity likely come from both the catalytic domain and the C-terminal accessory module. The residues driving specificity in GH20B and C are not present in BT 0456. A substrate-enzyme complex would be required to outline the precise interactions used to drive specificity shown by BT 0456, but this was not successful under the conditions we explored.

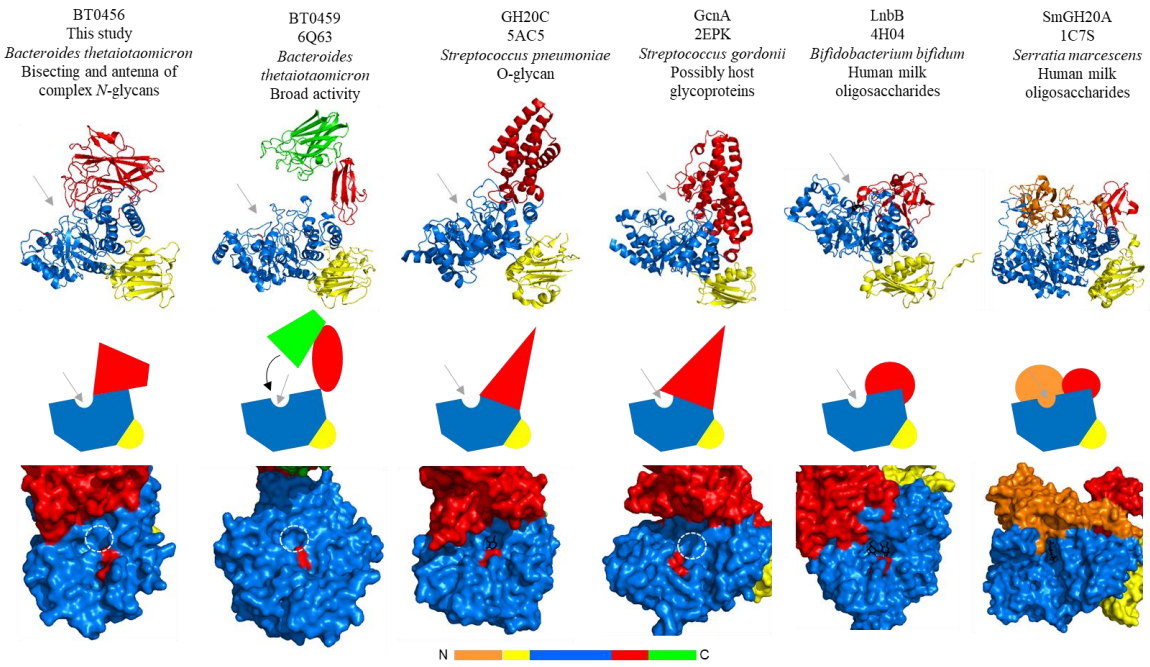


**Figure 3. Comparison of the crystal structure of BT 0456 with other GH20 family members that also have accessory modules.** The catalytic modules are in blue. The N-terminal domains are coloured yellow and a second N-terminal domain is coloured orange in the case of 1C7S. C-terminal domains are coloured red except for GH20 BT 0459, where the FN3 linker domain is coloured red and the C-terminal F5/F8 Type C domain is in green. Any bound glycans are shown as black sticks. A cartoon structure, a schematic, and a surface representation of each enzyme is provided. Grey arrows indicate where the active site is and in the case of the surface representations a white circle is the location of the -1 subsite where ligand is not present. The close association of the accessory domains with the active site make it likely that these contribute significantly to specificity. In the case of BT 0459, it is speculated that the green accessory module comes down over the active site to provide specificity or enhance binding (black arrow), which may be akin to what is seen for the SmGH20A (orange accessory domain).

One of the most striking aspects of this crystal structure is the positioning of the different domains relative to each other (Figure 2A). The C-terminal domain sits on the (β/α)8 barrel near the active site, so it is likely that the C-terminal domain contributes to the specificity of BT 0456. The residues in close proximity to this region includes S560, E561, R640, L641, and W517 (Figure 2B). The crystal structures of five other GH20 family members with accessory domains indicate that these accessory domains may also contribute to specificity based on their positioning (Figure 3).

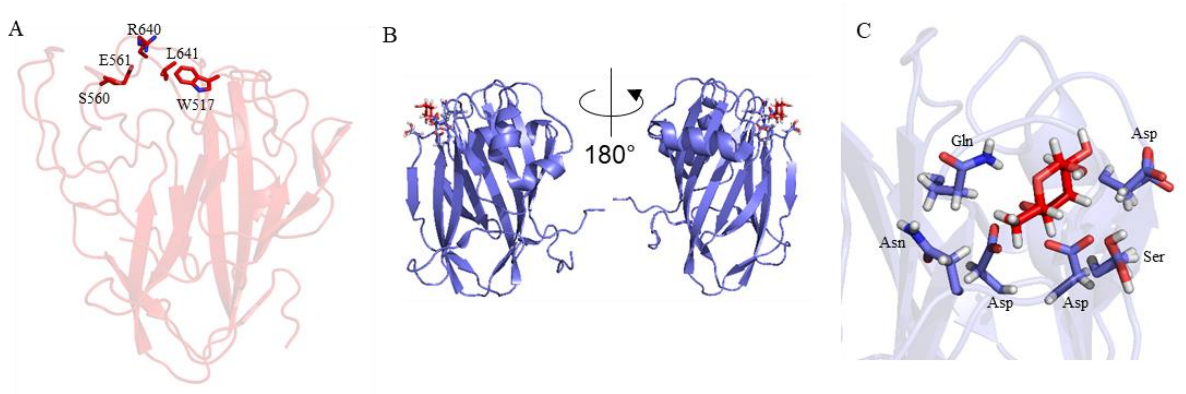


**Figure 4. Analysis of the BT 0456 C-terminal domain. (A)** The C-terminal domain of BT 0456 shown in more detail and the residues with likely association to the active site are shown as sticks. **(B)** The PA14 carbohydrate-binding domain from *Marinomonas primoryensis* (MpPA14) from both sides. **(C)** The details of the fucose binding site of MpPA14 and the residues comprising this active site.

The sequence of the BT0456 C-terminal domain did not produce any hits in either Pfam or SMART databases. When the structure of the C-terminal domain was used to search Dali database, a PA14 carbohydrate-binding domain from *Marinomonas primoryensis* with fucose bound was highlighted (MpPA14). A comparison of these two structures revealed that the binding site of fucose in MpPA14 was positioned in the same area as the residues that are highlighted being in close proximity to the active site for the C-terminal domain of BT 0456 (Figure 4).

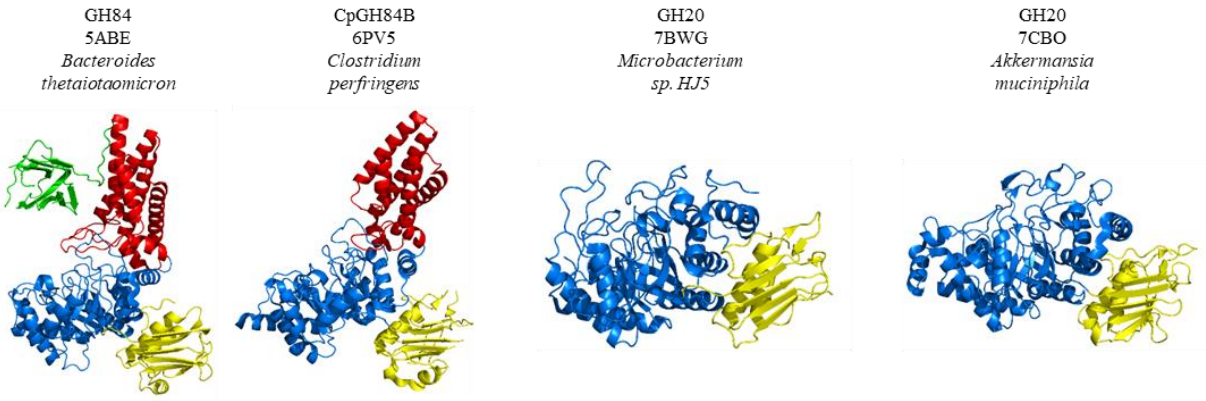


**Figure 5. N-terminal domain structural analysis.** The full crystal structures of the enzymes where the N-terminal domains (yellow) have close structural homology to the N-terminal domain from BT 0456. Catalytic domains are shown in blue and other accessory domains are in red and green where appropriate.

The N-terminal domain of BT0456 consist of a zincin-like fold. Structural matches of this domain were also found through Dali and include GlcNAcases from GH20 and GH84 families

(Figure 5). These N-terminal domains all show similar positioning relative to the catalytic module as that seen in BT 0456. The function of this domain is unknown but could be related to the structural stability of the overall protein.

## Acknowledgements

This research was supported by the European Union (GlySign, Grant No. 722095) and the NWO (Vernieuwingsimpuls Veni Project No. 722.016.008). Additionally, I also acknowledge Diamond Light Source (Oxfordshire, UK) for beamtime (proposal mx18598) and staff of beamline I03.

# References


1. Briliūtė, J., et al., Complex N-glycan breakdown by gut Bacteroides involves an extensive enzymatic apparatus encoded by multiple co-regulated genetic loci. Nature Microbiology, 2019. 4(9): p. 1571-1581.
2. Nakano, M., et al., Bisecting GlcNAc Is a General Suppressor of Terminal Modification of N-glycan*, [S]. Molecular & Cellular Proteomics, 2019. 18(10): p. 2044-2057.
3. Reily, C., et al., Glycosylation in health and disease. Nature Reviews Nephrology, 2019. 15(6): p. 346-366.
4. Saldova, R., et al., Association of Medication with the Human Plasma N-Glycome. Journal of Proteome Research, 2012. 11(3): p. 1821-1831.
5. Everest-Dass, A.V., et al., Human disease glycomics: technology advances enabling protein glycosylation analysis – part 2. Expert Review of Proteomics, 2018. 15(4): p. 341-352.
6. Pučić, M., et al., High Throughput Isolation and Glycosylation Analysis of IgG–Variability and Heritability of the IgG Glycome in Three Isolated Human Populations. Molecular & Cellular Proteomics, 2011. 10(10).
7. Vreeker, G.C.M., et al., Automated Plasma Glycomics with Linkage-Specific Sialic Acid Esterification and Ultrahigh Resolution MS. Analytical Chemistry, 2018. 90(20): p. 11955-11961.
8. Vanderschaeghe, D., et al., High-Throughput Profiling of the Serum N-Glycome on Capillary Electrophoresis Microfluidics Systems: Toward Clinical Implementation of GlycoHepatoTest. Analytical Chemistry, 2010. 82(17): p. 7408-7415.
9. Saldova, R., et al., Core fucosylation and α2-3 sialylation in serum N-glycome is significantly increased in prostate cancer comparing to benign prostate hyperplasia. Glycobiology, 2011. 21(2): p. 195-205.
10. Benicky, J., et al., Quantification of Fucosylated Hemopexin and Complement Factor H in Plasma of Patients with Liver Disease. Analytical Chemistry, 2014. 86(21): p. 10716-10723.
11. Rebello, O.D., et al., A Matrix-Assisted Laser Desorption/Ionization—Mass Spectrometry Assay for the Relative Quantitation of Antennary Fucosylated N-Glycans in Human Plasma. Frontiers in Chemistry, 2020. 8: p. 138.
12. Royle, L., et al., HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. Analytical Biochemistry, 2008. 376(1): p. 1-12.
13. Doherty, M., et al., Plasma N-glycans in colorectal cancer risk. Scientific Reports, 2018. 8(1): p. 8655.
14. Liu, X.-E., et al., N-glycomic changes in hepatocellular carcinoma patients with liver cirrhosis induced by hepatitis B virus. Hepatology, 2007. 46(5): p. 1426-1435.
15. Magorivska, I., et al., Glycosylation of random IgG distinguishes seropositive and seronegative rheumatoid arthritis. Autoimmunity, 2018. 51(3): p. 111-117.
16. Wu, Z., et al., IgG Glycosylation Profile and the Glycan Score Are Associated with Type 2 Diabetes in Independent Chinese Populations: A Case-Control Study. Journal of Diabetes Research, 2020. 2020: p. 5041346.

17. Hecht, E.S., et al., Relative Quantification and Higher-Order Modeling of the Plasma Glycan Cancer Burden Ratio in Ovarian Cancer Case-Control Samples. Journal of Proteome Research, 2015. 14(10): p. 4394-4401.
18. Reiding, K.R., et al., Human Plasma N-glycosylation as Analyzed by Matrix-Assisted Laser Desorption/Ionization-Fourier Transform Ion Cyclotron Resonance-MS Associates with Markers of Inflammation and Metabolic Health. Molecular & Cellular Proteomics, 2017. 16(2): p. 228-242.
19. Ruhaak, L.R., et al., Decreased Levels of Bisecting GlcNAc Glycoforms of IgG Are Associated with Human Longevity. PLOS ONE, 2010. 5(9): p. e12566.
20. Kabsch, W., XDS. Acta Crystallographica Section D, 2010. 66(2): p. 125-132.
21. Winter, G., xia2: an expert system for macromolecular crystallography data reduction. Journal of Applied Crystallography, 2010. 43(1): p. 186-190.
22. Evans, P., Scaling and assessment of data quality. Acta Crystallographica Section D, 2006. 62(1): p. 72-82.
23. Evans, P.R. and G.N. Murshudov, How good are my data and what is the resolution? Acta Crystallographica Section D, 2013. 69(7): p. 1204-1214.
24. Sheldrick, G., A short history of SHELX. Acta Crystallographica Section A, 2008. 64(1): p. 112-122.
25. Kovalevskiy, O., A. Lebedev, and E. Krissinel, Helping researchers to solve their structures: automation and user guidance in CCP4 Cloud. Acta Crystallographica Section A, 2021. 77(a2): p. C766.
26. The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr, 1994. 50(Pt 5): p. 760-3.
27. Murshudov, G.N., et al., REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallographica Section D, 2011. 67(4): p. 355-367.
28. Emsley, P. and K. Cowtan, Coot: model-building tools for molecular graphics. Acta Crystallographica Section D, 2004. 60(12 Part 1): p. 2126-2132.
29. Chen, V.B., et al., MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D, 2010. 66(1): p. 12-21.
30. Schrödinger, L., The PyMOL molecular graphics system, version 1.8. 2015.
31. Holm, L., et al., Using Dali for Structural Comparison of Proteins. Current Protocols in Bioinformatics, 2006. 14(1): p. 5.5.1-5.5.24.
32. Krissinel, E. and K. Henrick. Multiple Alignment of Protein Structures in Three Dimensions. in Computational Life Sciences. 2005. Berlin, Heidelberg: Springer Berlin Heidelberg.
33. Val-Cid, C., et al., Structural-Functional Analysis Reveals a Specific Domain Organization in Family GH20 Hexosaminidases. PLOS ONE, 2015. 10(5): p. e0128075.
34. Pluvinage, B., et al., Inhibition of the pneumococcal virulence factor StrH and molecular insights into N-glycan recognition and hydrolysis. Structure, 2011. 19(11): p. 1603-14.
35. Prag, G., et al., Structures of chitobiase mutants complexed with the substrate Di-N-acetyl-d-glucosamine: the catalytic role of the conserved acidic pair, aspartate 539 and glutamate 540. J Mol Biol, 2000. 300(3): p. 611-7.

36. Robb, M., et al., A Second β-Hexosaminidase Encoded in the Streptococcus pneumoniae Genome Provides an Expanded Biochemical Ability to Degrade Host Glycans*. Journal of Biological Chemistry, 2015. 290(52): p. 30888-30900.