Unveiling the Mystery of O-GlcNAc Processing Enzymes in Substrate Recognition

by

Hao Li

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Pharmaceutical Sciences)

at the

UNIVERSITY OF WISCONSIN-MADISON

2020

Date of final oral examination: July 1st, 2020

The dissertation is approved by the following members of the Final Oral Committee:

Jiaoyang Jiang, Associate Professor, School of Pharmacy
Tim S. Bugni, Professor, School of Pharmacy
Jennifer E. Golden, Assistant Professor, School of Pharmacy
Sandro Mecozzi, Professor, School of Pharmacy and Department of Chemistry
James L. Keck, Professor, School of Medicine and Public Health

ACKNOWLEDGEMENTS

This thesis becomes a reality with the kind support and help from many individuals. I would like to extend my sincerer thanks to all of them.

I would like to first express my deep gratitude to my dissertation advisor Professor Jiaoyang Jiang, for her constant support and encouragement during my graduate education and research. I enjoyed the process of learning in my Ph.D. research and I am really grateful that Professor Jiang encouraged me and provided the opportunity for me to explore brand new areas. In the lab, I not only obtained knowledge and skills, but also got trained to think in a deeper and broader perspective. My appreciation also extends to my committee members, Professor Jiaoyang Jiang, Professor Sandro Mecozzi, Professor Tim Bugni, Professor Jennifer Golden, Professor James Keck, and Professor Douglas Weibel. I benefited a lot from the discussions with them during my departmental seminars and annual meetings, and the suggestions helped me to keep on the right directions in my research.

At the same time, I would like to thank all Jiang Lab members. Dr. Baobin Li, Dr. Feng Wang, and Dr. Chia-Wei Hu taught me and guided me throughout my Ph.D. without reservation, providing valuable help and suggestions to me. I also collaborated with other enthusiastic and talented lab mates, especially Dr. Matthew Worth, Dacheng Fan, and all other members in the lab. I also gained lots of knowledge from our technician Dr. Kenneth Satyshur. It is a great pleasure working with them. I also appreciate the help from Analytical Instrumentation Center, Medicinal Chemical Center from the School of Pharmacy and LS-CAT, DND-CAT, Bio-CAT staff members from Argonne National Laboratory. My thanks go to my graduate coordinator, Kenneth Niemeyer, for all his help and support.

Last but not the least, this thesis would have not been possible without the love and support from my family and friends. I want to thank all my friends who made my life in Madison so joyful with their friendship. Also, I cannot emphasize my gratitude to my parents for their endless support too much. Most importantly, I want to thank my wife, Fang, whom I married during my Ph.D., and who accompanied me throughout my Ph.D.

ABSTRACT

O-GlcNAcylation is a post-translational modification where O-linked *N*-acetylglucosamine (GlcNAc) is attached on serine or threonine residues of protein substrates. Unlike other types of glycosylation that mostly present as polysaccharides on membrane proteins, O-GlcNAcylation is a monosaccharide, mainly found in the cytoplasm and nucleus. Compared to other PTMs, such as phosphorylation, which is catalyzed by hundreds of kinases and phosphatases, O-GlcNAcylation is only regulated by a single pair of enzymes, O-GlcNAc Transferase (OGT) and O-GlcNAcase (OGA). In the cell, there are more than 4,000 proteins modified by O-GlcNAc, and these proteins are involved in a large variety of important biological processes including transcription, translation, metabolism, signal transduction, and protein degradation. As O-GlcNAcylation plays significant roles in the cell, it is important to maintain a cellular homeostasis of O-GlcNAcylation level. As a consequence, for the disruption of homeostasis, dysregulation of O-GlcNAcylation has been linked to a number of diseases including cancer, diabetes, neurodegenerative diseases, and cardiovascular diseases, implicating O-GlcNAcylation as a potential therapeutic target. However, since O-GlcNAcylation is only regulated by a single pair of enzymes with thousands of substrates, adjusting O-GlcNAcylation on specific substrates remains elusive. We aim to utilize Chemical Biology, Structural Biology, and Cellular Biology tools to study how OGT and OGA recognize their substrates, which not only could facilitate understanding of O-GlcNAcylation functions and mechanisms, but could also establish O-GlcNAcylation as a powerful therapeutic target.

TABEL OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 O-GlcNAcylation

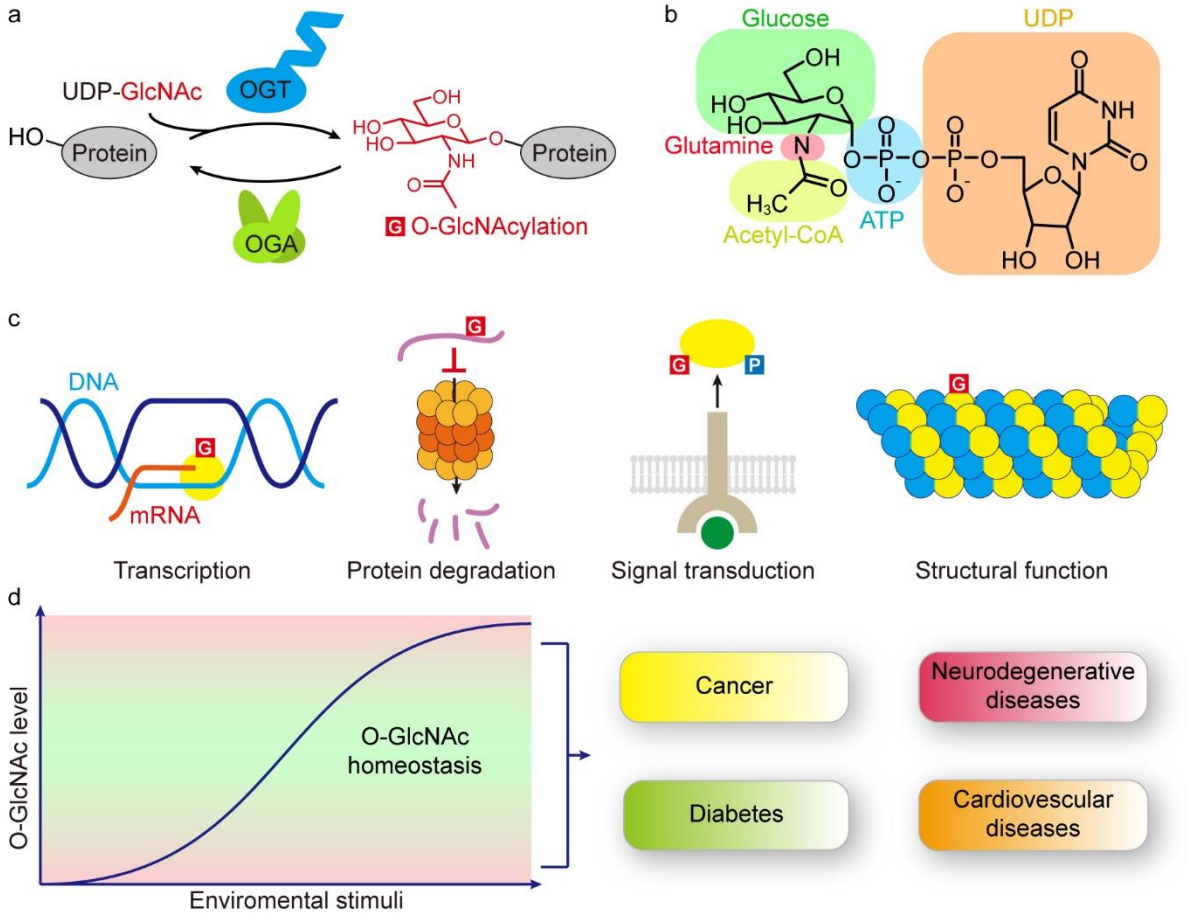


Figure 1.1 Overview of O-GlcNAcylation (a) The scheme of O-GlcNAcylation (b) The source of each part in UDP-GlcNAc synthesized through the HBP (c) Representative biological processes regulated by O-GlcNAcylation (d) The homeostasis of cellular O-GlcNAcylation and diseases caused by dysregulation of O-GlcNAcylation

O-GlcNAcylation is a nucleocytoplasmic post-translational modification (PTM),[1,2] where O-GlcNAc Transferase (OGT) transfers an *N*-acetylglucosamine (GlcNAc) moiety from the sugar donor uridine diphosphate *N*-acetylglucosamine (UDP-GlcNAc) onto serine or threonine residues of protein substrates and O-GlcNAcase (OGA) removes the modification. (Figure 1.1a) Unlike other types of surface glycans, which are mainly present on the cell surface as branched polysaccharides, O-GlcNAc is primarily found in the cytoplasm and nucleus as a monosaccharide.[3] As a result, we consider another type of PTM, phosphorylation, comparable to O-GlcNAc. While phosphorylation has similar labeling pattern as O-

GlcNAcylation, it is catalyzed by an array of kinases and phosphatases.[4] O-GlcNAcylation, on the other hand, is only regulated by a single pair of enzymes, OGT and OGA.

O-GlcNAc serves as a sensor of nutrients and stress through its sugar donor UDP-GlcNAc.[5] The biosynthesis of UDP-GlcNAc relies on the hexosamine biosynthetic pathway (HBP),[6] beginning with glucose or glucosamine that is sent through a series of metabolic pathways including glucose, nitrogen, nucleotide, and fatty acid metabolism. (Figure 1.1b) The O-GlcNAc level in the cell, as well as OGT activity, are highly responsive to the nutrient intake.[7] Hyperglycemia has been shown to alter the function and activity of O-GlcNAcylated proteins by increased O-GlcNAc level.

O-GlcNAc modification has been found on more than 4,000 protein substrates in the cell,[8] involved in a variety of cellular processes including transcription, translation, gene expression, signaling, protein degradation and structural functions.[9,10] (Figure 1.1c) Transcription is a complex process that is highly regulated by O-GlcNAcylation as numerous proteins are O-GlcNAc modified and their transcriptional activity is related to modification level. The C-terminal domain (CTD) of RNA polymerase II is highly O-GlcNAcylated at multiple sites following a distributive mechanism, which is essential for the formation of the transcriptional preinitiation complex. TET2 facilitates OGT glycosylation of histone 2B, positively regulating transcription.[11] O-GlcNAcylation of histone 2A is important in DNA repair mechanisms.[12] At the same time, O-GlcNAcylation on transcription factor Sp1 can inhibit its transcriptional ability.[13] O-GlcNAcylation is also happening on translational machinery, allowing the aggregation of untranslated messenger ribonuleoproteins into stress granules.[14]

O-GlcNAcylation also plays a significant role in signal transduction, mainly through its crosstalk with phosphorylation.[10] Protein phosphorylation is the attachment of a phosphate onto serine, threonine, and tyrosine residues, and serves as the most abundant PTM in eukaryotes, where 30% of all cellular proteins have been considered to be phosphorylated.[15] It has been found that in many cases, O-GlcNAcylation sites on proteins are in close proximity to phosphorylation sites, leading to competition or supplementation of the two PTMs.[16] O-GlcNAcylation and phosphorylation on the same site, T58 of c-

Myc inhibit each other, which affects the function of c-Myc in gene transcription.[17] On the other hand, both PTMs can happen simultaneously on nearby sites of insulin receptor substrate 1 (ISR-1).[18] O-GlcNAcylation can also affect phosphorylation indirectly by modifying kinases. Ca^{2+}/calmodulin-dependent protein kinase II (CaMKII) was reported to be activated by O-GlcNAc modification, regulating its chronic activation.[19]

Additionally, O-GlcNAc crosstalks with other types of PTMs.[20] Ubiquitination and O-GlcNAcylation were found to modify proteins simultaneously, and the increase of global O-GlcNAc level by OGA inhibitor PUGNAc treatment can increase the ubiquitination level as well.[21] As an opposing example, studies have indicated that during translation, O-GlcNAc will prevent the ubiquitination of nascent peptides, regulating proteolysis.[22]

With numerous important biological processes regulated by O-GlcNAcylation, it is essential to maintain a homeostasis of O-GlcNAc level in the cell.[23] (Figure 1.1d) Stimulations to the cells can lead to change of O-GlcNAc level and development of diseases. It has been reported that a variety of diseases are related to the dysregulation of O-GlcNAcylation, including cancer, diabetes, neurodegenerative disease and cardiovascular diseases.[9] Elevated O-GlcNAcylation or OGT expression level was found in most types of tumors, with a significant function in the proliferation and survival of tumor cells.[24] O-GlcNAcylation has long been related to neurodegenerative diseases. One example is tau protein, which is believed to be one of the important drivers of Alzheimer's progression, where impaired phosphorylation on multiple sites, after treatment with the OGA inhibitor PUGNAc, resulted in reduced Tau aggregation and cytotoxicity.[25,26] Further study revealed that O-GlcNAcylation at S400 of tau was reciprocal to phosphorylation at S396 and S404.[27] At the same time, another protein also involved in the progression of Parkinson's Disease, α-synuclein, has similar competition between O-GlcNAcylation and phosphorylation at nearby sites. O-GlcNAcylation of S72 of α-synuclein was reported to block the phosphorylation of S129, which has been proved to cause the aggregation of α-synuclein and play an important role in Parkinson's Disease.[28]

1.2 O-GlcNAc Transferase

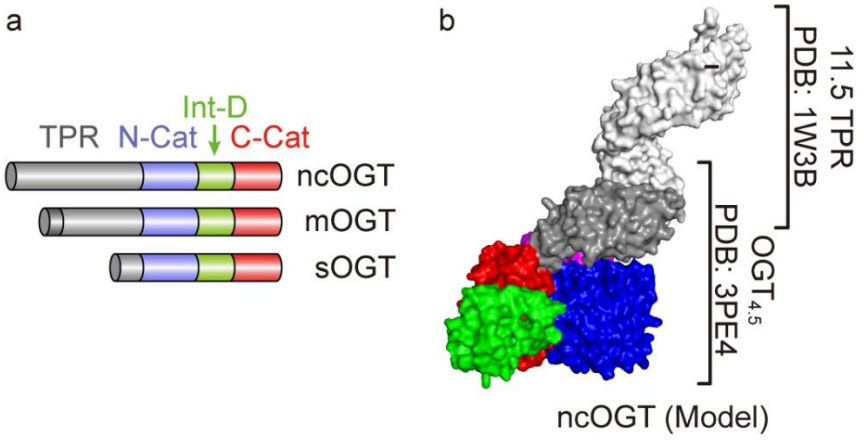


Figure 1.2 Introduction of OGT (a) The three isoforms of OGT and the domain representation (b) The full-length model of OGT built from two crystal structures

O-GlcNAc Transferase is responsible for the attachment of GlcNAc moieties onto serine or threonine residues of substrate proteins. OGT belongs to the GT41 glycosyltransferase family, with gene localized in the Xq13.1 region of chromosomal, and the full length protein is a 110 kDa multi-domain protein, consisting of an N-terminal tetratricopeptide repeat (TPR) domain, and two catalytic domains (N-cat and C-cat) split by an intervening domain.[29–31] Depending on the alternative splicing, OGT has three isoforms, nucleocytosolic OGT (ncOGT), mitochondrial OGT (mOGT), and short OGT (sOGT), that differ in length of TPR domain.[5] The mOGT splice variant is mainly localized in mitochondria through its mitochondrial targeting sequence. The sOGT isoform, on the other hand, has very similar localization in the cell as ncOGT, but its function is not fully understood.[32] (Figure 1.2a)

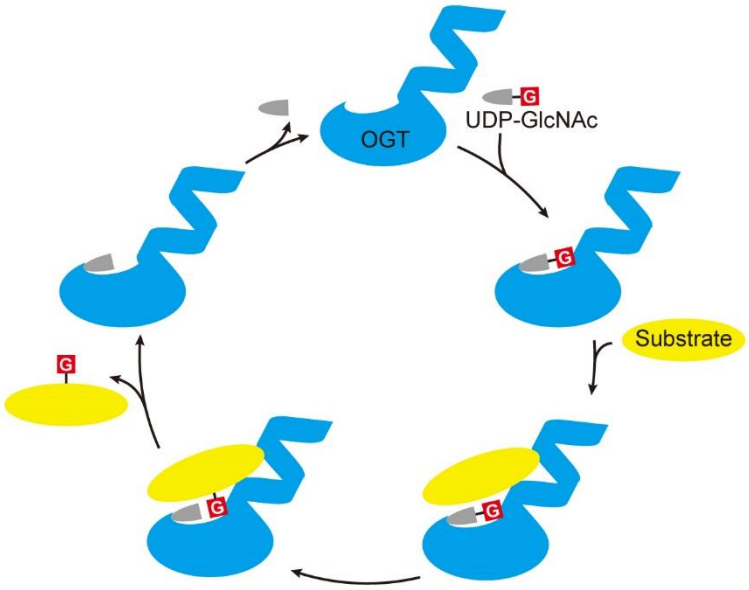


Figure 1.3 Ordered bi-bi mechanism of OGT

Efforts have been spent on determination of human OGT structure. However, with the TPR domain containing 13.5 TPR repeats, causing the overall conformation of OGT to be extremely flexible, the crystallization of full-length OGT has not been achieved. For now, a full-length model was built from two crystal structures, one from TPR domain, covering the first 11.5 TPR repeats,[33] and the other one with 4.5 TPR repeats together with the catalytic and intervening domains ($OGT_{4.5}$).[34] (Figure 1.2b) From the structure model of OGT, the TPR domain extends as a super helical structure, allowing the inner surface to provide binding pockets for protein substrates. The two catalytic lobes form the active site of OGT, providing binding cleft for UDP-GlcNAc. The intervening domain has a unique folding which is not found in other proteins. Although the function of the intervening domain remains elusive, it has been reported that it is responsible for the binding of phosphatidylinositol 3,4,5-triphosphate (PIP_3), suggesting the potential function of it in mediating interactions.[35] The TPR domain is connected to the N-catalytic domain through a transitional helix as a "hinge", yielding flexibility of the TPR domain. Due to the flexibility provided by the hinge, the TPR domain found in $OGT_{4.5}$ can adopt slightly different conformations in different crystal structures.

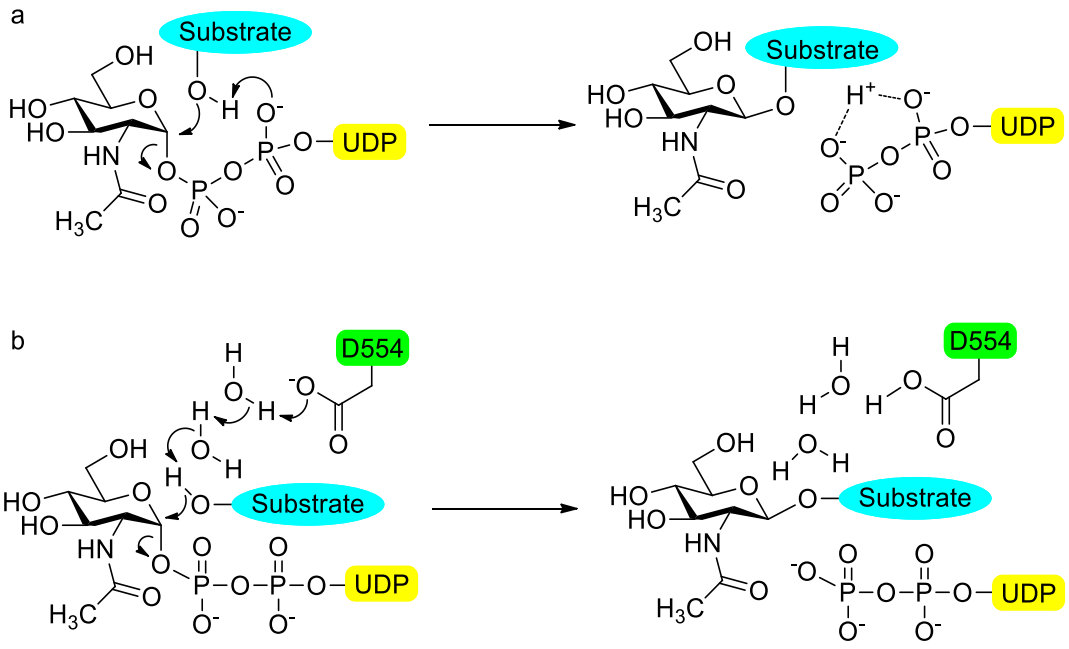


Figure 1.4 Proposed catalyzing mechanism of OGT (a) The mechanism of glycosylation catalyzed by α-phosphate (b) The mechanism of glycosylation catalyzed by D554 through the water chain

Based on the complex structure of OGA in complex with UDP and CKII peptide substrate, OGT adopts an ordered bi-bi mechanism where OGT first bind UDP-GlcNAc followed by its protein substrate for sugar transfer, the newly formed glycopeptide is released then UDP leaves.[36] (Figure 1.3) The detailed catalyzing mechanism of OGT is still under debate. One hypothesis involves the substrate facilitated mechanism where the α-phosphate acts as a catalytic base to activate the hydroxyl group of substrate serine or threonine, driving the nucleophilic attack on the anomeric carbon. The mechanism was proposed based on the discovery that replacing an oxygen on the α-phosphate to sulfur, which has weaker nucleophilicity, will largely affect the *in vitro* O-GlcNAcylation of substrates.[37] (Figure 1.4a) By solving a complex crystal structure of $OGT_{4.5}$ with UDP-GlcNAc and CKII peptide substrate, D554 was found to activate the substrate hydroxyl with two ordered water molecules.[36] (Figure 1.4b)

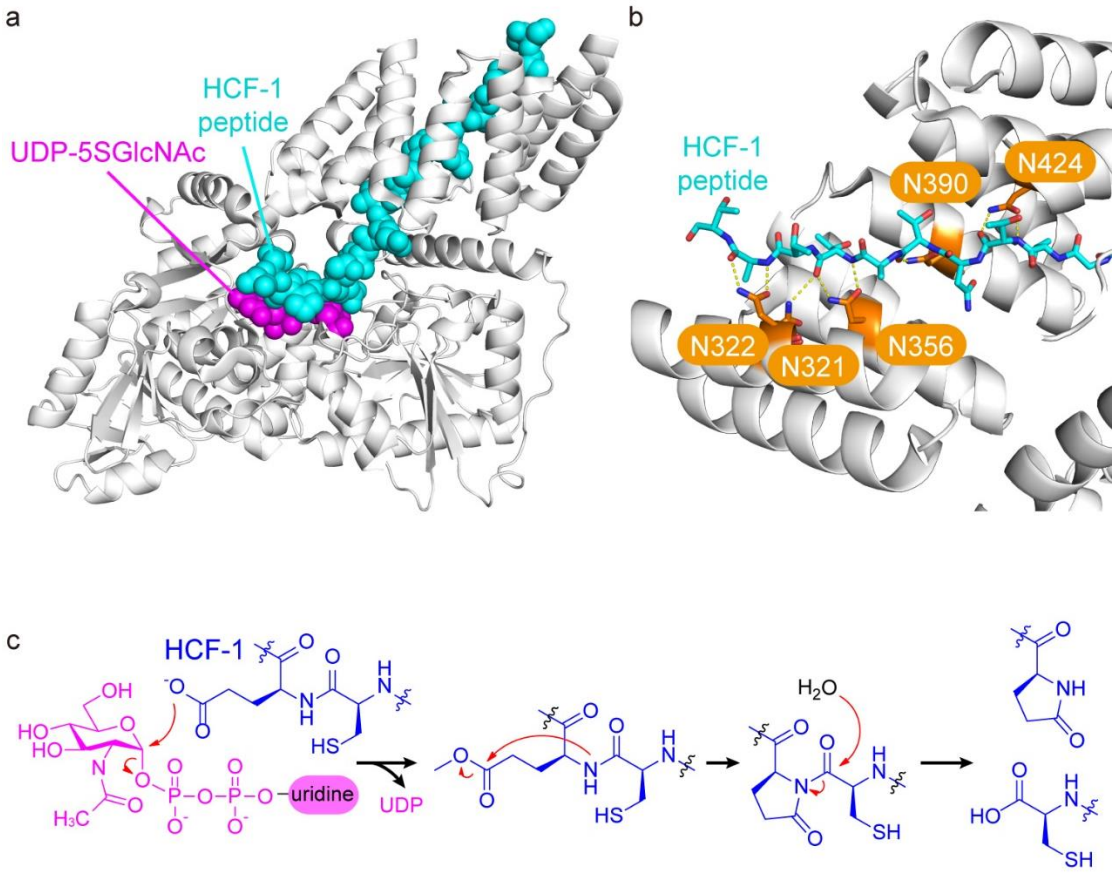


Figure 1.5 OGT cleaves HCF-1 in the active site (a) The tertiary structure of OGT4.5:UDP-5SGlcNAc:HCF-1 peptide, OGT is shown in white ribbon (b) HCF-1 peptide interacts with the asparagine ladder in OGT TPR domain, OGT is shown in white ribbon and interacting residues are highlighted with orange sticks (c) The mechanism of HCF-1 cleavage

A single protein OGT has thousands of substrates in cells, yet the substrate recognition of OGT remains unknown.[9] Efforts have been made to obtain complex structures of OGT with its substrates. As full length TPR is required for the binding of protein substrates, only the structures of $OGT_{4.5}$ in complex with peptides derived from casein kinase II (CKII),[34] TGF-Beta Activated Kinase 1 (TAB1),[37] Host Cell Factor 1 (HCF-1),[38] Retinoblastoma-like protein 2 (RBL2), Proto-oncogene tyrosine-protein kinase receptor Ret (Ret), Keratin-7 (KER7), and Lamin B1 (LAMIN), have been solved[39] In most of the structures, the peptides mainly formed interactions with UDP-GlcNAc, and the interaction with OGT is mainly focused on the peptide substrates' backbones. The only exception was the HCF-1 peptide which bound more extending towards the TPR domain and formed interactions with five spatially aligned asparagine

residues (N321, N322, N356, N390, N424, the asparagine ladder) in the inner surface of TPR. Further study fused TAB1 and HCF-1 peptides onto the N-terminal of $OGT_{4.5}$, and the same asparagine ladder formed interactions with both peptides, indicating that the TPR domain plays an important role in substrate binding and recognition.[40]

One special function of OGT is the cleavage of HCF-1, the transcriptional regulator of cell cycling.[41] In this scenario, OGT first transfers a GlcNAc onto a glutamate residue on HCF-1, facilitating the nucleophilic attack of the amide nitrogen to the formed ester of the glutamate to generate a 5-member ring intermediate. The amide bond is then cleaved to achieve HCF-1 proteolysis.[42] (Figure 1.5)

With OGT regulating numerous important substrates by O-GlcNAcylation, OGT has been considered as a potential drug target for a long time.[43] It has been reported that knocking down OGT will affect the growth rate of tumor cells.[44] However, as OGT is an essential protein, complete knock out will lead to cell death,[20] making the development of OGT inhibitors extremely valuable. I will review the development of OGT inhibitors in Chapter 3.

1.3 O-GlcNAcase

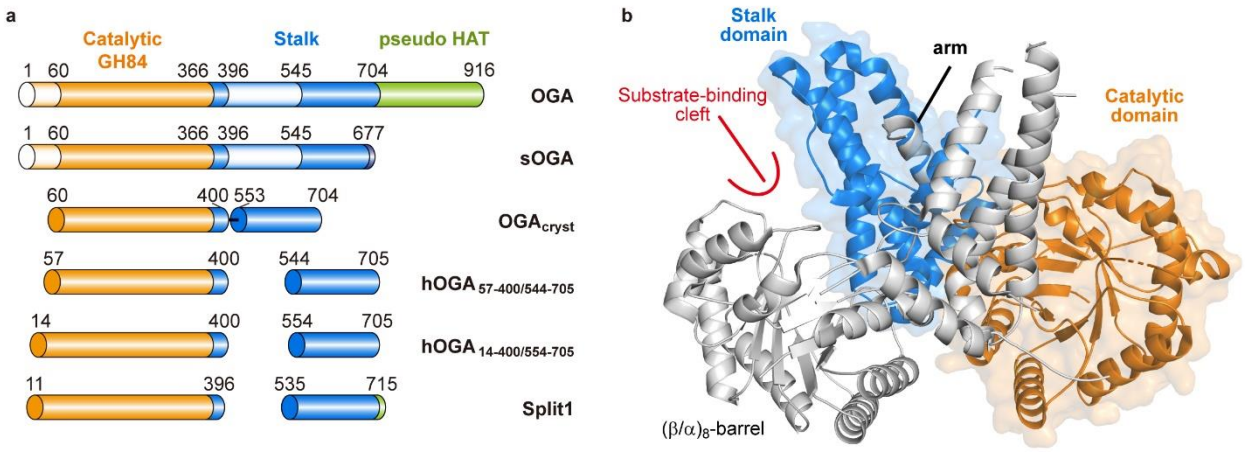


Figure 1.6 Introduction of OGA (a) The domain representation of two isoforms of OGA and the different constructs used for crystallization (b) The ribbon representation of OGA dimer structure. The catalytic domain and stalk domain of one monomer of OGA are shown in orange and blue ribbon, respectively. The other monomer is shown in white ribbon.

As the other enzyme regulating O-GlcNAcylation, OGA is responsible for removal of GlcNAc from O-GlcNAcylated substrates. OGA belongs to the GH84 glycosidase family and OGA gene is located in chromosome 10.[2,45] OGA has two isoforms, full-length OGT (lOGA) and short OGA (sOGA). The lOGA isoform is a 118 kDa protein mainly localized in the cytoplasm and nucleus. Full-length OGA contains an N-terminal catalytic domain, a C-terminal Histone Acetal Transferase (HAT) domain and a stalk domain in between. The other isoform of OGA, sOGA, lacks the HAT domain and was reported to accumulate on lipid droplets surface, relating sOGA to proteasome function.[45] (Figure 1.6a)

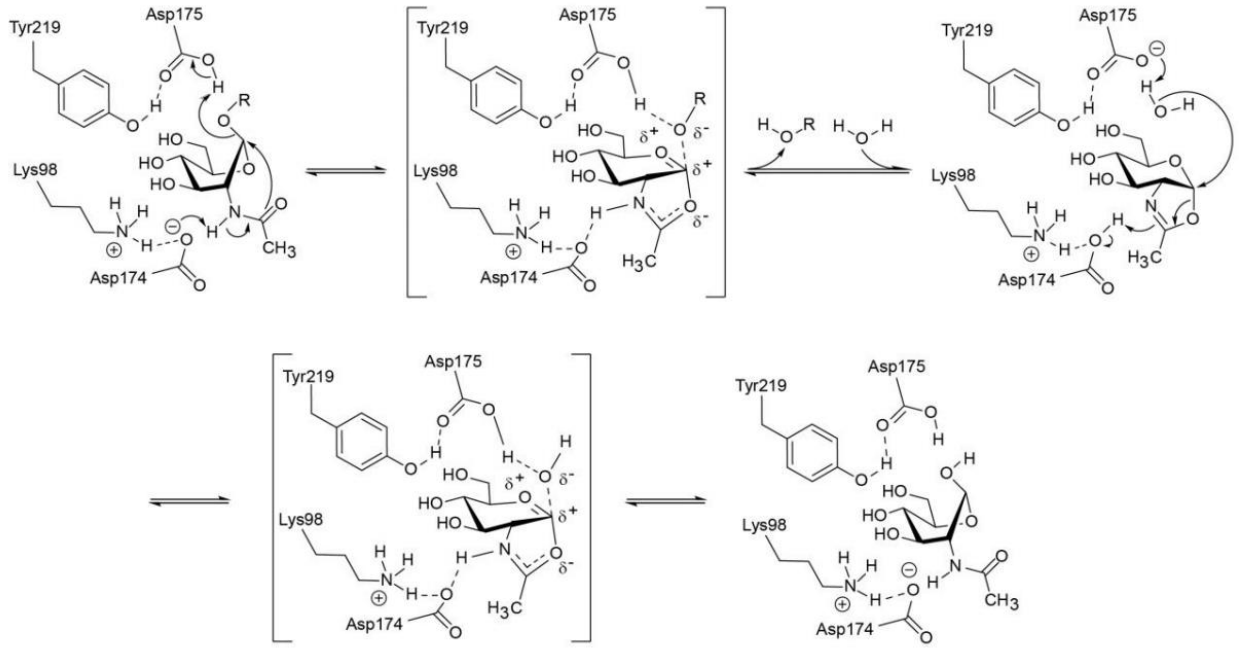


Figure 1.7 Substrate facilitated catalyzing mechanism of OGA

The structure of human OGA has been absent for more than 15 years after the discovery of this enzyme, largely due to the fact that OGA contains a large portion of structurally disordered regions, including the first 60 amino acids, a large part of the intervening domain, and most of the HAT domain, based on the protein secondary structure prediction.[46] In order to obtain structural information of human OGA, a number of bacterial homologues of OGA were used for the structural study, including *Bacteroides*

thetaiotaomicron hexosaminidase (BtGH84), *Clostridium perf ringens NagJ* (CpOGA), *Oceanicola granulosus* glycosidase (OgOGA), and *Thermobaculum terrenum* glycoside hydrolase (TtOGA).[9] The bacterial homologues share relatively high similarity in the catalytic domain compared to human OGA, and the conformation of sugar binding to the bacterial homologues could represent how OGA bind to the GlcNAc. From the structures of bacterial homologues in complex with either O-GlcNAc derived inhibitors or O-GlcNAcylated peptides, the catalytic mechanism could be derived. Two aspartate residues, D174 and D175, in the active site of OGA is responsible for the substrate assisted hydrolysis mechanism. First, D174 serving as a catalytic base, triggers the nucleophilic attack of the N-acetal group to the anomeric carbon, removing the substrate from the sugar and forming the oxazoline intermediate. D175 will catalyze the hydrolysis of the intermediate with a nearby water molecule, forming a free GlcNAc molecule. (Figure 1.7)

Despite the difficulty to crystallize human OGA, recently, our group and two others separately determined the first crystal structures of human OGA.[47–49] (Figure 1.6b) The constructs for crystallization all had truncation of the flexible region in the stalk domain and the HAT domain. In addition, we also determined the first crystal structures of human OGA in complex with a few glycopeptide substrates, revealing the substrate recognition pattern of OGA. I will discuss our discoveries in detail in chapter 4.

The flexible loop of the intervening domain (amino acids 396-554) and HAT domain of OGA are less studied. Although the 150-amino-acid loop cannot be crystallized, it was reported to serve as an important regulatory domain of OGA, with a number of PTMs including phosphorylation and O-GlcNAcylation, as well as a cleavage site by caspase 3.[32,50,51] OGA cleavage at D413 by caspase 3 was validated by in vitro experiments and it was proposed that OGA is an important regulator of apoptosis. Surprisingly, after cleavage, OGA still maintained its activity with two cleavage products remaining bound together. As OGA activity can regulate the O-GlcNAcylation level of itself, the loop is not only important in the regulatory role of OGA in the O-GlcNAc biology, but also apoptosis. The function of HAT domain has long been debated. It was reported that OGA purified from mammalian cells possessed the activity to

acetylate histone, while OGA purified from bacterial cells did not.[52] However, in another report, human OGA was proved to be a pseudo-acetyltransferase without essential residues for the activity when comparing with a bacterial putative acetyltransferase.[53] The reported activity of human OGA could potentially be derived from impurities of mammalian protein purification.

Same as OGT, OGA also serves as an important potential drug target.[54] As we discussed in the first section, O-GlcNAcylation dysregulation has been linked to the development of neurodegenerative diseases. Hyperphosphorylation could lead to the aggregation of tau protein, deconstructing the microtubule which is one hypothesis of the cause of Alzheimer's Disease progression.[55] O-GlcNAcylation will compete with phosphorylation, providing protection for the microtubule structure. One well developed inhibitor of OGA, thiamet-G, which was designed based on the oxazoline intermediate in the catalytic mechanism of OGA, showed suppression of tau protein phosphorylation and rescue of Alzheimer in mouse model.[56]

1.4 Reference

(1) Torres, C. R.; Hart, G. W. Topography and Polypeptide Distribution of Terminal N-Acetylglucosamine Residues on the Surfaces of Intact Lymphocytes. Evidence for O-Linked GlcNAc. *J. Biol. Chem.* **1984**, *259* (5), 3308–3317.

(2) Gao, Y.; Wells, L.; Comer, F. I.; Parker, G. J.; Hart, G. W. Dynamic O-Glycosylation of Nuclear and Cytosolic Proteins: Cloning and Characterization of a Neutral, Cytosolic Beta-N-Acetylglucosaminidase from Human Brain. *J. Biol. Chem.* **2001**, *276* (13), 9838–9845. https://doi.org/10.1074/jbc.M010420200.

(3) Holt, G. D.; Hart, G. W. The Subcellular Distribution of Terminal N-Acetylglucosamine Moieties. Localization of a Novel Protein-Saccharide Linkage, O-Linked GlcNAc. *J. Biol. Chem.* **1986**, *261* (17), 8049–8057.

(4) Kramer, I. M. *Signal Transduction*, Third edition.; AP, Academic Press, an imprint of Elsevier: Amsterdam Boston Heidelberg, 2016.

(5) Harwood, K. R.; Hanover, J. A. Nutrient-Driven O-GlcNAc Cycling - Think Globally but Act Locally. *J. Cell Sci.* **2014**, *127* (Pt 9), 1857–1867. https://doi.org/10.1242/jcs.113233.

(6) Peterson, S. B.; Hart, G. W. New Insights: A Role for O-GlcNAcylation in Diabetic Complications. *Crit. Rev. Biochem. Mol. Biol.* **2016**, *51* (3), 150–161. https://doi.org/10.3109/10409238.2015.1135102.

(7) Hardivillé, S.; Hart, G. W. Nutrient Regulation of Signaling, Transcription, and Cell Physiology by O-GlcNAcylation. *Cell Metab.* **2014**, *20* (2), 208–213. https://doi.org/10.1016/j.cmet.2014.07.014.

(8) Ma, J.; Hart, G. W. O-GlcNAc Profiling: From Proteins to Proteomes. *Clin. Proteomics* **2014**, *11* (1), 8. https://doi.org/10.1186/1559-0275-11-8.

(9) Worth, M.; Li, H.; Jiang, J. Deciphering the Functions of Protein *O* -GlcNAcylation with Chemistry. *ACS Chem. Biol.* **2017**, *12* (2), 326–335. https://doi.org/10.1021/acschembio.6b01065.

(10) Hart, G. W. Nutrient Regulation of Signaling and Transcription. *J. Biol. Chem.* **2019**, *294* (7), 2211–2231. https://doi.org/10.1074/jbc.AW119.003226.

(11) Chen, Q.; Chen, Y.; Bian, C.; Fujiki, R.; Yu, X. TET2 Promotes Histone O-GlcNAcylation during Gene Transcription. *Nature* **2013**, *493* (7433), 561–564. https://doi.org/10.1038/nature11742.

(12) Hayakawa, K.; Hirosawa, M.; Tani, R.; Yoneda, C.; Tanaka, S.; Shiota, K. H2A O-GlcNAcylation at Serine 40 Functions Genomic Protection in Association with Acetylated H2AZ or ΓH2AX. *Epigenetics Chromatin* **2017**, *10* (1), 51. https://doi.org/10.1186/s13072-017-0157-x.

(13) Yang, X.; Su, K.; Roos, M. D.; Chang, Q.; Paterson, A. J.; Kudlow, J. E. O-Linkage of N-Acetylglucosamine to Sp1 Activation Domain Inhibits Its Transcriptional Capability. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (12), 6611–6616. https://doi.org/10.1073/pnas.111099998.

(14) Ohn, T.; Kedersha, N.; Hickman, T.; Tisdale, S.; Anderson, P. A Functional RNAi Screen Links O-GlcNAc Modification of Ribosomal Proteins to Stress Granule and Processing Body Assembly. *Nat. Cell Biol.* **2008**, *10* (10), 1224–1231. https://doi.org/10.1038/ncb1783.

(15) Ubersax, J. A.; Ferrell Jr, J. E. Mechanisms of Specificity in Protein Phosphorylation. *Nat. Rev. Mol. Cell Biol.* **2007**, *8* (7), 530–541. https://doi.org/10.1038/nrm2203.

(16) Hart, G. W.; Slawson, C.; Ramirez-Correa, G.; Lagerlof, O. Cross Talk between O-GlcNAcylation and Phosphorylation: Roles in Signaling, Transcription, and Chronic Disease. *Annu. Rev. Biochem.* **2011**, *80*, 825–858. https://doi.org/10.1146/annurev-biochem-060608-102511.

(17) Chou, T. Y.; Hart, G. W.; Dang, C. V. C-Myc Is Glycosylated at Threonine 58, a Known Phosphorylation Site and a Mutational Hot Spot in Lymphomas. *J. Biol. Chem.* **1995**, *270* (32), 18961–18965. https://doi.org/10.1074/jbc.270.32.18961.

(18) Ball, L. E.; Berkaw, M. N.; Buse, M. G. Identification of the Major Site of O-Linked Beta-N-Acetylglucosamine Modification in the C Terminus of Insulin Receptor Substrate-1. *Mol. Cell. Proteomics MCP* **2006**, *5* (2), 313–323. https://doi.org/10.1074/mcp.M500314-MCP200.

(19) Erickson, J. R.; Pereira, L.; Wang, L.; Han, G.; Ferguson, A.; Dao, K.; Copeland, R. J.; Despa, F.; Hart, G. W.; Ripplinger, C. M.; Bers, D. M. Diabetic Hyperglycaemia Activates CaMKII and Arrhythmias by O-Linked Glycosylation. *Nature* **2013**, *502* (7471), 372–376. https://doi.org/10.1038/nature12537.

(20) Yang, X.; Qian, K. Protein O-GlcNAcylation: Emerging Mechanisms and Functions. *Nat. Rev. Mol. Cell Biol.* **2017**, *18* (7), 452–465. https://doi.org/10.1038/nrm.2017.22.

(21) Guinez, C.; Mir, A.-M.; Dehennaut, V.; Cacan, R.; Harduin-Lepers, A.; Michalski, J.-C.; Lefebvre, T. Protein Ubiquitination Is Modulated by O-GlcNAc Glycosylation. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **2008**, *22* (8), 2901–2911. https://doi.org/10.1096/fj.07-102509.

(22) Zhu, Y.; Liu, T.-W.; Cecioni, S.; Eskandari, R.; Zandberg, W. F.; Vocadlo, D. J. O-GlcNAc Occurs Cotranslationally to Stabilize Nascent Polypeptide Chains. *Nat. Chem. Biol.* **2015**, *11* (5), 319–325. https://doi.org/10.1038/nchembio.1774.

(23) Bond, M. R.; Hanover, J. A. A Little Sugar Goes a Long Way: The Cell Biology of O-GlcNAc. *J. Cell Biol.* **2015**, *208* (7), 869–880. https://doi.org/10.1083/jcb.201501101.

(24) de Queiroz, R. M.; Carvalho, E.; Dias, W. B. O-GlcNAcylation: The Sweet Side of the Cancer. *Front. Oncol.* **2014**, *4*, 132. https://doi.org/10.3389/fonc.2014.00132.

(25) Haltiwanger, R. S.; Grove, K.; Philipsberg, G. A. Modulation of O-Linked N-Acetylglucosamine Levels on Nuclear and Cytoplasmic Proteins in Vivo Using the Peptide O-GlcNAc-Beta-N-Acetylglucosaminidase Inhibitor O-(2-Acetamido-2-Deoxy-D-Glucopyranosylidene)Amino-N-Phenylcarbamate. *J. Biol. Chem.* **1998**, *273* (6), 3611–3617. https://doi.org/10.1074/jbc.273.6.3611.

(26) Liu, F.; Iqbal, K.; Grundke-Iqbal, I.; Hart, G. W.; Gong, C.-X. O-GlcNAcylation Regulates Phosphorylation of Tau: A Mechanism Involved in Alzheimer's Disease. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (29), 10804–10809. https://doi.org/10.1073/pnas.0400348101.

(27) Smet-Nocca, C.; Broncel, M.; Wieruszeski, J.-M.; Tokarski, C.; Hanoulle, X.; Leroy, A.; Landrieu, I.; Rolando, C.; Lippens, G.; Hackenberger, C. P. R. Identification of O-GlcNAc Sites within Peptides of the Tau Protein and Their Impact on Phosphorylation. *Mol. Biosyst.* **2011**, *7* (5), 1420–1429. https://doi.org/10.1039/c0mb00337a.

(28) Marotta, N. P.; Lin, Y. H.; Lewis, Y. E.; Ambroso, M. R.; Zaro, B. W.; Roth, M. T.; Arnold, D. B.; Langen, R.; Pratt, M. R. O-GlcNAc Modification Blocks the Aggregation and Toxicity of the Protein α-Synuclein Associated with Parkinson's Disease. *Nat. Chem.* **2015**, *7* (11), 913–920. https://doi.org/10.1038/nchem.2361.

(29) Hanover, J. A.; Yu, S.; Lubas, W. B.; Shin, S.-H.; Ragano-Caracciola, M.; Kochran, J.; Love, D. C. Mitochondrial and Nucleocytoplasmic Isoforms of O-Linked GlcNAc Transferase Encoded by a Single Mammalian Gene. *Arch. Biochem. Biophys.* **2003**, *409* (2), 287–297. https://doi.org/10.1016/S0003-9861(02)00578-7.

(30) Lubas, W. A.; Frank, D. W.; Krause, M.; Hanover, J. A. *O* -Linked GlcNAc Transferase Is a Conserved Nucleocytoplasmic Protein Containing Tetratricopeptide Repeats. *J. Biol. Chem.* **1997**, *272* (14), 9316–9324. https://doi.org/10.1074/jbc.272.14.9316.

(31) Shafi, R.; Iyer, S. P. N.; Ellies, L. G.; O'Donnell, N.; Marek, K. W.; Chui, D.; Hart, G. W.; Marth, J. D. The O-GlcNAc Transferase Gene Resides on the X Chromosome and Is Essential for Embryonic Stem Cell Viability and Mouse Ontogeny. *Proc. Natl. Acad. Sci.* **2000**, *97* (11), 5735–5739. https://doi.org/10.1073/pnas.100471497.

(32) Lazarus, B. D.; Love, D. C.; Hanover, J. A. Recombinant O-GlcNAc Transferase Isoforms: Identification of O-GlcNAcase, Yes Tyrosine Kinase, and Tau as Isoform-Specific Substrates. *Glycobiology* **2006**, *16* (5), 415–421. https://doi.org/10.1093/glycob/cwj078.

(33) Jínek, M.; Rehwinkel, J.; Lazarus, B. D.; Izaurralde, E.; Hanover, J. A.; Conti, E. The Superhelical TPR-Repeat Domain of O-Linked GlcNAc Transferase Exhibits Structural Similarities to Importin Alpha. *Nat. Struct. Mol. Biol.* **2004**, *11* (10), 1001–1007. https://doi.org/10.1038/nsmb833.

(34) Lazarus, M. B.; Nam, Y.; Jiang, J.; Sliz, P.; Walker, S. Structure of Human O-GlcNAc Transferase and Its Complex with a Peptide Substrate. *Nature* **2011**, *469* (7331), 564–567. https://doi.org/10.1038/nature09638.

(35) Janetzko, J.; Walker, S. The Making of a Sweet Modification: Structure and Function of O-GlcNAc Transferase. *J. Biol. Chem.* **2014**, *289* (50), 34424–34432. https://doi.org/10.1074/jbc.R114.604405.

(36) Lazarus, M. B.; Jiang, J.; Gloster, T. M.; Zandberg, W. F.; Whitworth, G. E.; Vocadlo, D. J.; Walker, S. Structural Snapshots of the Reaction Coordinate for O-GlcNAc Transferase. *Nat. Chem. Biol.* **2012**, *8* (12), 966–968. https://doi.org/10.1038/nchembio.1109.

(37) Schimpl, M.; Zheng, X.; Borodkin, V. S.; Blair, D. E.; Ferenbach, A. T.; Schüttelkopf, A. W.; Navratilova, I.; Aristotelous, T.; Albarbarawi, O.; Robinson, D. A.; Macnaughtan, M. A.; van Aalten, D. M. F. O-GlcNAc Transferase Invokes Nucleotide Sugar Pyrophosphate Participation in Catalysis. *Nat. Chem. Biol.* **2012**, *8* (12), 969–974. https://doi.org/10.1038/nchembio.1108.

(38) Lazarus, M. B.; Jiang, J.; Kapuria, V.; Bhuiyan, T.; Janetzko, J.; Zandberg, W. F.; Vocadlo, D. J.; Herr, W.; Walker, S. HCF-1 Is Cleaved in the Active Site of O-GlcNAc Transferase. *Science* **2013**, *342* (6163), 1235–1239. https://doi.org/10.1126/science.1243990.

(39) Pathak, S.; Alonso, J.; Schimpl, M.; Rafie, K.; Blair, D. E.; Borodkin, V. S.; Albarbarawi, O.; van Aalten, D. M. F. The Active Site of O-GlcNAc Transferase Imposes Constraints on Substrate Sequence. *Nat. Struct. Mol. Biol.* **2015**, *22* (9), 744–750. https://doi.org/10.1038/nsmb.3063.

(40) Rafie, K.; Raimi, O.; Ferenbach, A. T.; Borodkin, V. S.; Kapuria, V.; van Aalten, D. M. F. Recognition of a Glycosylation Substrate by the O-GlcNAc Transferase TPR Repeats. *Open Biol.* **2017**, *7* (6). https://doi.org/10.1098/rsob.170078.

(41) Zargar, Z.; Tyagi, S. Role of Host Cell Factor-1 in Cell Cycle Regulation. *Transcription* **2012**, *3* (4), 187–192. https://doi.org/10.4161/trns.20711.

(42) Janetzko, J.; Trauger, S. A.; Lazarus, M. B.; Walker, S. How the Glycosyltransferase OGT Catalyzes Amide Bond Cleavage. *Nat. Chem. Biol.* **2016**, *12* (11), 899–901. https://doi.org/10.1038/nchembio.2173.

(43) Lynch, T. P.; Reginato, M. J. *O* -GlcNAc Transferase: A Sweet New Cancer Target. *Cell Cycle* **2011**, *10* (11), 1712–1713. https://doi.org/10.4161/cc.10.11.15561.

(44) Wang, L.; Chen, S.; Zhang, Z.; Zhang, J.; Mao, S.; Zheng, J.; Xuan, Y.; Liu, M.; Cai, K.; Zhang, W.; Guo, Y.; Zhai, W.; Yao, X. Suppressed OGT Expression Inhibits Cell Proliferation While Inducing Cell Apoptosis in Bladder Cancer. *BMC Cancer* **2018**, *18* (1), 1141. https://doi.org/10.1186/s12885-018-5033-y.

(45) Keembiyehetty, C. N.; Krzeslak, A.; Love, D. C.; Hanover, J. A. A Lipid-Droplet-Targeted O-GlcNAcase Isoform Is a Key Regulator of the Proteasome. *J. Cell Sci.* **2011**, *124* (Pt 16), 2851–2860. https://doi.org/10.1242/jcs.083287.

(46) Ishida, T.; Kinoshita, K. PrDOS: Prediction of Disordered Protein Regions from Amino Acid Sequence. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W460-464. https://doi.org/10.1093/nar/gkm363.

(47) Li, B.; Li, H.; Lu, L.; Jiang, J. Structures of Human O-GlcNAcase and Its Complexes Reveal a New Substrate Recognition Mode. *Nat. Struct. Mol. Biol.* **2017**, *24* (4), 362–369. https://doi.org/10.1038/nsmb.3390.

(48) Roth, C.; Chan, S.; Offen, W. A.; Hemsworth, G. R.; Willems, L. I.; King, D. T.; Varghese, V.; Britton, R.; Vocadlo, D. J.; Davies, G. J. Structural and Functional Insight into Human O-GlcNAcase. *Nat. Chem. Biol.* **2017**, *13* (6), 610–612. https://doi.org/10.1038/nchembio.2358.

(49) Elsen, N. L.; Patel, S. B.; Ford, R. E.; Hall, D. L.; Hess, F.; Kandula, H.; Kornienko, M.; Reid, J.; Selnick, H.; Shipman, J. M.; Sharma, S.; Lumb, K. J.; Soisson, S. M.; Klein, D. J. Insights into Activity and Inhibition from the Crystal Structure of Human O-GlcNAcase. *Nat. Chem. Biol.* **2017**, *13* (6), 613–615. https://doi.org/10.1038/nchembio.2357.

(50) Shiromizu, T.; Adachi, J.; Watanabe, S.; Murakami, T.; Kuga, T.; Muraoka, S.; Tomonaga, T. Identification of Missing Proteins in the NeXtProt Database and Unregistered Phosphopeptides in the PhosphoSitePlus Database as Part of the Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2013**, *12* (6), 2414–2421. https://doi.org/10.1021/pr300825v.

(51) Butkinaree, C.; Cheung, W. D.; Park, S.; Park, K.; Barber, M.; Hart, G. W. Characterization of Beta-N-Acetylglucosaminidase Cleavage by Caspase-3 during Apoptosis. *J. Biol. Chem.* **2008**, *283* (35), 23557–23566. https://doi.org/10.1074/jbc.M804116200.

(52) Toleman, C.; Paterson, A. J.; Whisenhunt, T. R.; Kudlow, J. E. Characterization of the Histone Acetyltransferase (HAT) Domain of a Bifunctional Protein with Activable O-GlcNAcase and HAT Activities. *J. Biol. Chem.* **2004**, *279* (51), 53665–53673. https://doi.org/10.1074/jbc.M410406200.

(53) Rao, F. V.; Schüttelkopf, A. W.; Dorfmueller, H. C.; Ferenbach, A. T.; Navratilova, I.; van Aalten, D. M. F. Structure of a Bacterial Putative Acetyltransferase Defines the Fold of the Human O-GlcNAcase C-Terminal Domain. *Open Biol.* **2013**, *3* (10), 130021. https://doi.org/10.1098/rsob.130021.

(54) Lazarus, B. D.; Love, D. C.; Hanover, J. A. O-GlcNAc Cycling: Implications for Neurodegenerative Disorders. *Int. J. Biochem. Cell Biol.* **2009**, *41* (11), 2134–2146. https://doi.org/10.1016/j.biocel.2009.03.008.

(55) Gong, C.-X.; Iqbal, K. Hyperphosphorylation of Microtubule-Associated Protein Tau: A Promising Therapeutic Target for Alzheimer Disease. *Curr. Med. Chem.* **2008**, *15* (23), 2321–2328. https://doi.org/10.2174/092986708785909111.

(56) Yuzwa, S. A.; Macauley, M. S.; Heinonen, J. E.; Shan, X.; Dennis, R. J.; He, Y.; Whitworth, G. E.; Stubbs, K. A.; McEachern, E. J.; Davies, G. J.; Vocadlo, D. J. A Potent Mechanism-Inspired O-GlcNAcase Inhibitor That Blocks Phosphorylation of Tau in Vivo. *Nat. Chem. Biol.* **2008**, *4* (8), 483–490. https://doi.org/10.1038/nchembio.96.

Chapter 2 Development of GlcNAc Electrophilic Probes (GEPs) to decipher the substrate recognition of OGT

The results of this chapter were published on:

Hu, C.-W.*; Worth, M.*; Fan, D.*; Li, B.*; **Li, H.***; Lu, L.; Zhong, X.; Lin, Z.; Wei, L.; Ge, Y.; Li, L.; Jiang, J. Electrophilic Probes for Deciphering Substrate Recognition by O-GlcNAc Transferase. *Nat. Chem. Biol.* **2017**, *13* (12), 1267–1273. (*: equal contributions)

Contributions

This project is a highly collaborative project. I was involved in the organic synthesis part. Together with Dr. Matthew Worth, I synthesized the OGT probes **GEP1**, **GEP2**, **GEP1A**. At the same time, I also learned protein crystallography with Dr. Baobin Li.

2.1 Introduction

As we discussed before, OGT attaches O-GlcNAc onto serine or threonine residues of thousands of substrates, while the substrates recognition mechanism of OGT remains largely elusive. The understanding of how OGT recognizes its substrates is of great significance, not only for the study of O-GlcNAcylation function and mechanism, but also to be used as a potential therapeutic approach for a variety of diseases. Abnormal O-GlcNAcylation level could lead to cancer, diabetes, and neurodegenerative diseases,[1] and the effect of O-GlcNAcylated proteins have been found on proteins regulating these diseases. With the knowledge of specific structural features or residues for the recognition of substrates, we can achieve substrate specific tuning of O-GlcNAcylation level without affecting global modification.

Attempts have been made to investigate the structural features of OGT, or the residues that are involved in binding and recognition of substrates, but limited information could be concluded. In light of the structural study of OGT, only the crystal structure of $OGT_{4.5}$ was solved, providing information regarding binding of OGT with its substrates.[2] However, full-length TPR is required for the binding of protein substrates, so only the complex structures with short peptides are available. At the same time, due to the ordered bi-bi mechanism,[3] the binding of OGT with its substrates is largely facilitated by UDP-GlcNAc, and most of the binding was contributed by UDP-GlcNAc. Limited information of OGT-peptide substrate interactions could be derived near the active site. However, a few structures contain peptides extended into the inner groove of the 4.5 TPR repeats,[4,5] revealing that the asparagine ladder formed by five asparagine residues in the TPR domain (N321, N322, N356, N390, N424) was responsible for forming conserved interactions with two distinct substrates, HCF-1 and TAB1 peptides. The limitation of the construct for crystallization construct prohibited further insights in OGT substrate recognition to

be discovered. Traditional protein-protein interactions assays like binding assays (Isothermal titration calorimetry[6], Surface plasmon resonance[7], Microscale thermophoresis[8]) can also yield structural information of how an enzyme binds to its substrates. The difficulty in applying binding assays to OGT arises from its two substrates, where these traditional assays cannot distinguish between sugar or protein substrate binding effects. One accurate assay is the well-developed radio-labeled kinetic assay for OGT,[9] where UDP-GlcNAc was labeled with tritium, and the exact kinetic parameters, including K_M, k_{cat}, and catalytic efficiency can be determined. The drawbacks of radio-labeled kinetic assay are that it is time consuming and expensive. In this project, we developed a chemical probe of OGT, providing fast readout of OGT structural features responsible for the binding and recognition of specific substrates.

2.2 Material and methods

2.2.1 General Chemical Synthesis

All reagents were purchased from Sigma-Aldrich, MP Biomedicals, Alfa Aesar, TCI, or Thermo Fisher Scientific. Unless otherwise stated, all reactions were performed in flame-dried, roundbottomed flasks fitted with rubber septa under nitrogen atmosphere. Analytical thin layer chromatography (TLC) was conducted on silica gel plates (0.25 mm, 60 Å pore size) with fluorescent indicator on glass (254 nm) and visualized by ceric ammonium molybdate (CAM), basic $KMnO_4$, or UV light. Unless otherwise noted, all commercially available reagents were used as received. Compound purity was assessed with proton nuclear magnetic resonance (NMR) and mass spectrometry (MS). NMR spectra were obtained on Varian UI400 or UI500 spectrometers. Spectra were recorded at 400 or 500 MHz for ^{1}H NMR, 101 or 126 MHz for ^{13}C

NMR, and 162 MHz for ^{31}P NMR. Chemical shifts are expressed in parts per million (ppm, δ scale) and referenced to $CDCl_3$ or D_2O. Data for ^{1}H NMR spectra are reported as follows: chemical shift (ppm), multiplicity (s, singlet; d, doublet; t, triplet; q, quartet; m, multiplet; dd, doublet of doublets; ddd, doublet of doublet of doublets, dt, doublet of triplets; td, triplet of doublets; ddq, doublet of doublet of quartet), coupling constant (Hz), and integration. ^{13}C NMR spectra are expressed as chemical shifts. ^{31}P NMR spectra are represented as follows: chemical shifts, multiplicity, and coupling constant (Hz). High-resolution MS spectra were obtained from Q-TOF Maxis 4G (Bruker).

2.2.1.1 Synthesis of **GEP1** and **GEP2**

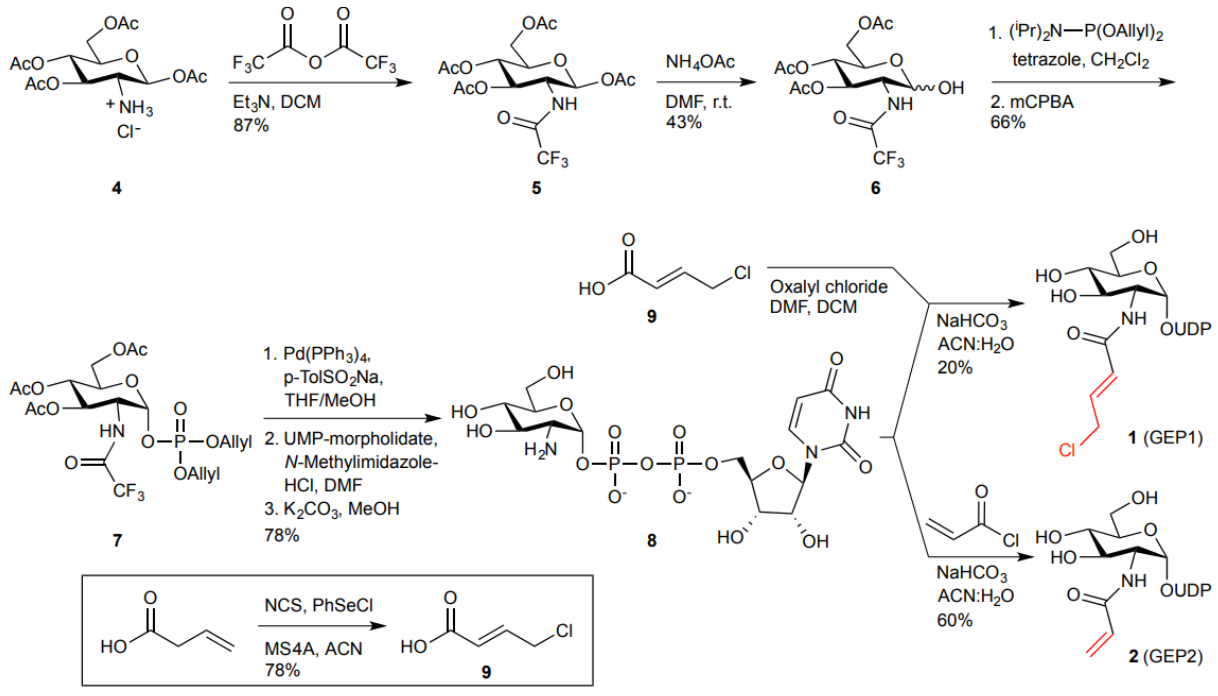


*Figure 2.1 Synthesis scheme for **GEP1** and **GEP2***

1,3,4,6-tetra-*O*-acetyl-2-trifluoroacetamido-2-deoxy-α,β-D-glucopyranose (5). Compound **4** (501 mg, 1.31 mmol), Et_3N (0.40 mL, 2.87 mmol), and CH_2Cl_2 (10 mL) were combined and cooled

to 0 ^{o}C in an ice-water bath. Trifluoroacetic anhydride was added dropwise. The solution was stirred at 0 ^{o}C for 20 min, then at room temperature (RT) for 4 h. The solution was then diluted with 30 mL EtOAc and washed with water (30 mL), saturated $NaHCO_3$ (2 x 30 mL), and brine (30 mL). The organic layer was dried over Na_2SO_4. The product was purified by silica gel chromatography using hexanes:EtOAc (7:2) and dried down to a white amorphous solid (507.8 mg, 87% yield). ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.32 (d, J = 9.7 Hz, 1H), 5.76 (d, J = 8.8 Hz, 1H), 5.36 (dd, J = 10.7, 9.4 Hz, 1H), 5.12 (t, J = 9.7 Hz, 1H), 4.38 (dd, J = 19.7, 9.4 Hz, 1H), 4.29 (dd, J = 12.5, 4.9 Hz, 1H), 4.17 (dd, J = 12.5, 2.2 Hz, 1H), 3.92 (m, 1H), 2.12 (s, 3H), 2.09 (s, 3H), 2.06 (s, 3H), 2.05 (s, 3H). [Lit.[10] ^{1}H NMR (300 MHz, $CDCl_3$) δ 7.24 (d, J = 9.0 Hz, 1H), 5.75 (d, J = 9.0 Hz, 1H), 5.31 (t, J = 10.0 Hz, 1H), 5.13 (t, J = 9.6 Hz, 1H), 4.35 (q, J = 9.9 Hz, 1H), 4.27 (dd, J = 12.6, 4.8 Hz, 1H), 4.15 (dd, J = 12.6, 2.1 Hz, 1H), 3.90 (ddd, J = 9.9, 4.8, 2.1 Hz, 1H), 2.12 (s, 3H), 2.10 (s, 3H), 2.06 (s, 3H), 2.05 (s, 3H)]. ^{13}C NMR (126 MHz, $CDCl_3$) δ 172.09, 170.93, 169.70, 169.68, 157.8 (q, J = 37.9 Hz), 115.8 (q, J = 287.9 Hz), 92.04, 73.21, 72.41, 68.29, 61.92, 53.36, 20.94, 20.93, 20.76, 20.64.

3,4,6-tri-*O*-acetyl-2-trifluoroacetamido-2-deoxy-α,β-D-glucopyranose (6). Compound **5** (445 mg, 1.00 mmol), NH_4OAc (155 mg, 2.00 mmol), and dry DMF (3 mL) were combined and stirred at RT for 28 h. The solution was then diluted with 30 mL EtOAc and washed with water (2 x 30 mL) and brine (30 mL). The organic layer was dried over Na_2SO_4. The product was purified on silica gel chromatography using hexanes:EtOAc (2:1) and dried down to a yellow-white solid (175 mg, 43% yield). ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.70 (d, J = 9.6 Hz, 1H), 5.36 (m, 2H), 5.16 (t, J = 9.4 Hz, 1H), 4.32 (td, J = 9.78, 3.47 Hz, 1H), 4.24 (m, 2H) 4.13 (m, 1H), 3.60 (s, 1H), 2.11 (s, 3H), 2.05 (s, 3H), 2.02 (s, 3H). [Lit.[11] ^{1}H NMR ($CDCl_3$) δ 6.64 (d, J = 9.1 Hz, 1H), 5.34 (dd, J = 9.9, 9.6 Hz, 1H), 5.33 (d, J = 3.3 Hz, 1H), 5.13 (dd, J = 9.6 Hz, 1H), 4.29 (ddd, 1H), 4.25-4.08 (m, 3H),

2.08 (s, 3H), 2.03 (s, 3H), 2.00 (s, 3H)]. ^{13}C NMR (126 MHz, $CDCl_3$) δ 171.33, 170.81, 169.33, 157.2 (q, J = 37.8 Hz), 115.5 (q, J = 287.8 Hz), 90.97, 70.38, 67.92, 67.73, 61.83, 52.74, 20.76, 20.59, 20.47.

Diallyl(3,4,6-tri-*O*-acetyl-2-trifluoroacetamido-2-deoxy-α-D-glucopyranosyl)phosphate (7). Compound **6** (76.7 mg, 0.191 mmol), tetrazole (0.45 M in ACN, 1.7 mL, 0.764 mmol), diallyl-N,N-diisopropylphosphoramidite (60.5 µL, 0.229 mmol), and CH_2Cl_2 (2 mL) were combined in a dry flask under N_2 and heated at reflux for 8.5 h. The solution was cooled to -45 ^{o}C in a dry ice-ACN bath, and mCPBA (215 mg, 0.955 mmol) was added. The mixture was stirred at -45 ^{o}C for 30 min, and was then allowed to warm to RT and stirred for 2 h. The solution was then diluted with 20 mL EtOAc and washed with 10% Na_2SO_3 (20 mL), saturated $NaHCO_3$ (20 mL) and water (20 mL). The organic layer was dried over Na_2SO_4 and separated on silica gel chromatography, eluting with hexanes:EtOAc (2:1) to give a white amorphous solid (71 mg, 66% yield). ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.04 (d, J = 8.7 Hz, 1H), 5.94 (m, 2H), 5.76 (dd, J = 6.3, 3.3 Hz, 1H), 5.39 (ddq, J = 16.8, 13.9, 1.4 Hz, 2H), 5.32 (m, 3H), 5.21 (t, J = 9.8 Hz, 1H), 4.60 (m, 4H), 4.39 (m, 1H), 4.28 (dd, J = 12.4, 4.1 Hz, 1H), 4.24 (m, 1H), 4.11 (dd, J = 12.4, 2.1 Hz, 1H), 2.09 (s, 3H), 2.05 (s, 3H), 2.03 (s, 3H). ^{31}P NMR (162 MHz, $CDCl_3$) δ -1.50 (s). Theoretical m/z calculated for $C_{20}H_{27}F_3NNaO_{12}P$ $[M+Na]^+$: 584.1115. HRMS found: 584.1102.

Uridine 5'-diphospho-2-amino-2-deoxy-α-D-glucopyranose (8). Compound **7** (100 mg, 0.178 mmol), sodium *p*-toluenesulfinate (64 mg, 0.356 mmol), $Pd(PPh_3)_4$ (10 mg, 0.0089 mmol), and MeOH:THF (1:1, 2.2 mL) were combined in a dry flask under N_2 and stirred at RT for 3.5 h. The solvent was evaporated under vacuum, and UMP-morpholidate (196 mg, 0.285 mmol), N-methylimidazolium choride (105 mg, 0.89 mmol), and dry DMF (3 mL) were added to the flask. The solution was stirred at RT for 21 h. The DMF was co-evaporated with toluene under vacuum.

K_2CO_3 (295 mg, 2.1 mmol) and MeOH:H_2O (1:1, 3 mL) were added, and the solution was stirred for 90 min at RT. The solvent was evaporated under vacuum. The residue was taken up in 20 mL water and washed with CH_2Cl_2 (20 mL) and the aqueous layer was concentrated under vacuum. The product was separated by Bio-Gel P-2 (Bio-Rad, CA) SEC column (60 cm) with 100 mM NH_4HCO_3 and lyophilized to white powder (79 mg, 78% yield). ^{1}H NMR (500 MHz, D_2O) δ 7.96 (d, J = 8.1 Hz, 1H), 5.99 (m, 2H), 5.86 (m, 1H), 4.38 (m, 2H), 4.20-4.32 (m, 3H), 3.96 (m, 2H), 3.89 (dd, J = 12.77, 1.70 Hz, 1H), 3.83 (dd, J = 12.58, 4.19 Hz, 1H), 3.58 (t, J = 9.6 Hz, 1H), 3.42 (m, 1H). Theoretical m/z calculated for $C_{15}H_{23}N_3O_{16}P_2$ $[M-2H]^{2-}$: 281.5282. HRMS found: 281.5252.

(*E*)-4-chloro-2-butenoic acid (9). 3-butenoic acid (0.50 mL, 5.9 mmol) was added to a solution of phenyl selenenyl chloride (113 mg, 0.59 mmol), activated 4Å molecular sieves (1.2 g) and acetonitrile (33 mL). A solution of *N*-chlorosuccinimide (0.86 g, 6.4 mmol) in acetonitrile (33 mL) was slowly added to the reaction mixture over 2 h. After 34 h stirring at RT, the solution was decanted into a separate flask and concentrated *in vacuo* to ~10 mL. Diethyl ether (80 mL) was added and the solution was washed twice with 35 mL water, and the organic layer was dried over Na_2SO_4. The product was purified on silica gel, eluting with hex:EtOAc (3:1). It was dried down to a light yellow solid (552 mg, 78% yield). ^{1}H NMR (400 MHz, $CDCl_3$) δ 11.80 (s, 1H), 7.10 (dt, *J* = 15.35, 5.9 Hz, 1H), 6.13 (d, *J* = 15.36 Hz, 1H), 4.20 (d, *J* = 5.87 Hz, 2H). [Lit.[12] ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.12 (dt, *J* = 15.0, 6.0 Hz, 1H), 6.16 (dt, *J* = 15.0, 1.5 Hz, 1H), 4.22 (dd, *J* = 6.0, 1.5 Hz, 2H)]. ^{13}C NMR (126 MHz, $CDCl_3$) δ 171.46, 144.63, 123.34, 42.38.

GEP1 (1). Compound **9** (21 mg, 0.18 mmol), dry CH_2Cl_2 (1 mL) and dry DMF (2 drops) were combined and cooled to 0 ^{o}C in an ice-water bath. Oxalyl chloride was added dropwise, and the solution was stirred at 0 ^{o}C for 30 min, then at RT for 3 h. The solvent was evaporated under

vacuum briefly, and the residue was cooled to 0 °C. The residue was taken up in dry ACN (1.7 mL) and added dropwise to a solution of **8** (10 mg, 0.018 mmol), $NaHCO_3$ (121 mg, 1.44 mmol) and ACN:H_2O (1:1, 10 mL) at 0 °C. The reaction was stirred at 0 °C for 4 h. The solvent was evaporated under vacuum, and the crude product was separated on a Bio-Gel P-2 SEC column (23 cm) with water, and further purified on silica gel chromatography using EtOAc:MeOH:H_2O (7:2:1). The product fraction was lyophilized to powder (2.3 mg, 20% yield). ^{1}H NMR (400 MHz, D_2O) δ 7.98 (t, J = 8.13 Hz, 1H), 6.88 (dt, J = 15.23, 6.15 Hz, 1H), 6.42 (dt, J = 15.29, 1.29 Hz, 1H), 5.99 (m, 2H), 5.58 (dd, J = 6.96, 3.2 H_Z, 1H), 4.37 (m, 2H), 4.24-4.31 (m, 3H), 4.18 (m, 1H), 4.13 (dt, J = 10.26, 2.84 Hz, 1H), 3.97 (m, 1H), 3.80-3.95 (m, 2H), 3.60 (t, J = 9.63 Hz, 1H). ^{31}P NMR (162 MHz, D_2O) δ -10.2 (d, J = 21 Hz), -11.5 (d, J = 21 Hz). Theoretical m/z calculated for $C_{19}H_{28}ClN_3O_{17}P_2$ $[M-2H]^{2-}$: 332.5291. HRMS found: 332.5176.

GEP2 (2). Compound **8** (5.1 mg, 9 μmol) and $NaHCO_3$ (100 mg, 0.25 mmol) were dissolved in ACN:water (1:1) and cooled to 0 °C in an ice bath. Acryloyl chloride (3.4 μL, 0.042 mmol) dissolved in ACN (0.4 mL) was added to the reaction mixture, which was stirred at 0 °C for 4 h. The solution was concentrated *in vacuo* and purified on a Bio-Gel P-2 SEC column (23 cm) with water. The product fractions were lyophilized to white powder (3.4 mg, 60% yield). ^{1}H NMR (400 MHz, D_2O) δ 7.96 (d, *J* = 8.14 Hz, 1H), 6.42 (dd, *J* = 17.27, 10.46 Hz, 1H), 6.23 (d, *J* = 17.14 Hz, 1H), 5.97 (m, 2H), 5.80 (d, *J* = 10.34 Hz, 1H), 5.56 (dd, *J* = 7.10, 3.13 Hz, 1H), 4.37 (m, 2H), 4.10-4.30 (m, 4H), 3.80-3.99 (m, 4H), 3.59 (t, *J* = 9.42 Hz, 1H). ^{31}P NMR (162 MHz, D_2O) δ -10.22 (d, *J* = 21.26, 1P), -12.01 (d, *J* = 21.09, 1P). Theoretical m/z calculated for $C_{18}H_{25}N_3O_{17}P_2$ $[M-2H]^{2-}$: 308.5335. HRMS found: 308.5300.

2.2.1.2 Synthesis of **GEP1A**

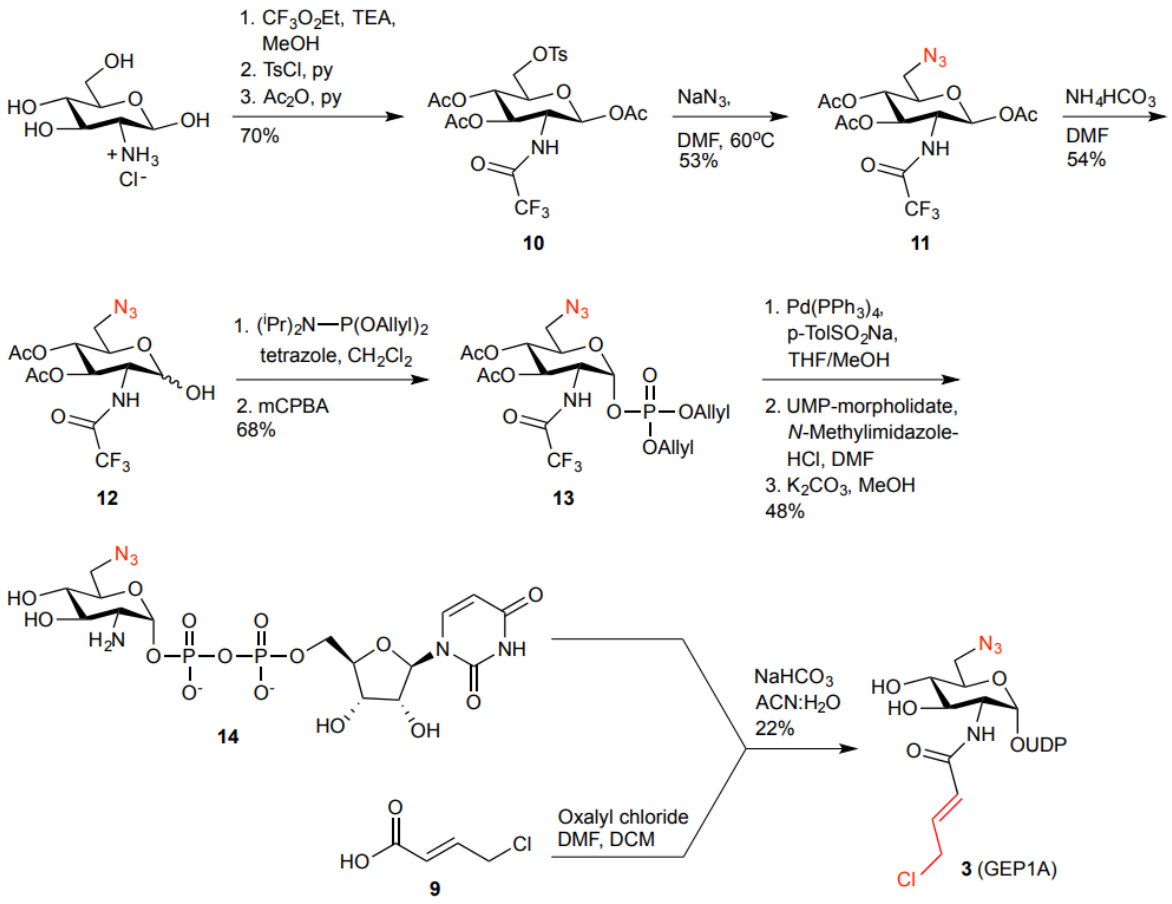


*Figure 2.2 Synthetic scheme for **GEP1A***

1,3,4-tri-*O*-acetyl-2-deoxy-2-trifluoroacetamido-6-O-tosyl-α,β-D-glucopyranose (10). Glucosamine hydrochloride (5.39 g, 25 mmol) was dissolved in MeOH (25 mL). Trimethylamine (3.6 mL, 26.4 mmol) was added to the reaction followed by ethyl trifluoroacetate (3.27 mL, 27.5 mmol), then stirred at RT overnight. The mixture was concentrated under vacuum. The residue was taken up in pyridine (15 mL) and the solution of tosyl chloride (7.6 g, 80 mmol) in pyridine was added dropwise via syringe at 0 °C. The mixture was stirred at 0 °C for 1 h and then allowed to warm to RT for 3 h when TLC showed complete conversion of substrates. Acetic anhydride was added to the mixture via syringe at 0 °C, then allowed to warm to RT for 2 h. Pyridine was

evaporated under vacuum and the residue was taken up in DCM and washed with 2 N HCl (100 mL) and saturated $NaHCO_3$ (100 mL) twice. The organic layer was dried over Na_2SO_4 and concentrated under vacuum to obtain the crude as an orange liquid. Silica gel chromatography was used to purify the product with hexanes:EtOAc (2:1) and dried down to white powder (70% yield). ^{1}H NMR (400 MHz, $CDCl_3$): δ 7.79 (m, 2H, α and β), 7.39 (m, 2H, α and β), 6.64 (d, J=8.8 Hz, 1H), 6.17 (d, 3.7 Hz, 1H, β), 5.76 (d, 8.8 Hz, 1H, α), 5.31 (m, 2H, β), 5.10 (m, 2H, α), 4.34 (m, 1H, α and β), 4.13 (m, 3H, α and β), 2.49 (s, 3H, β), 2.48 (s, 3H, α), 2.20 (s, 3H, β), 2.12 (s, 3H, α), 2.08 (s, 3H, β), 2.07 (s, 3H, β), 2.06 (s, 3H, α), 2.06 (s, 3H, α). ^{13}C NMR (101 MHz, $CDCl_3$) δ 171.83, 171.50, 169.36, 169.13, 169.11, 168.30, 157.43, 157.05, 145.35, 145.28, 129.92, 128.13, 116.76, 113.90, 91.62, 89.24, 72.62, 71.83, 69.99, 69.63, 68.03, 67.46, 67.03, 53.44, 53.07, 51.76, 21.70, 20.74, 20.59, 20.49, 20.45, 20.35. Theoretical m/z calculated for $C_{21}H_{24}F_3NNaO_{11}S$ $[M+Na]^+$: 578.0914. HRMS found: 578.0927.

1,3,4-tri-*O*-acetyl-2,6-dideoxy-2-trifluoroacetamido-6-azido-β-D-glucopyranose (11). Compound **10** (1.36 g, 2.45 mmol) was dissolved in DMF (30 mL), and sodium azide (477.5 mg, 7.35 mmol) was added. The mixture was then heated to 60 °C for 3 h when TLC showed complete conversion of substrates. DMF was evaporated under vacuum. The residue was dissolved in EtOAc (50 mL) and washed with brine (50 mL) for three times. The organic layer was dried over Na_2SO_4 and concentrated under vacuum. The crude product was purified by silica gel chromatography, eluting with hexanes:EtOAc (4:1) to give a white solid (53% yield). ^{1}H NMR (400 MHz, $CDCl_3$): δ 6.78 (d, J=9.8 Hz, 1H), 5.79 (d, J=8.8 Hz, 1H), 5.28 (dd, J=10.7, 9.4 Hz, 1H), 5.13 (dd, J=9.8, 9.4 Hz, 1H), 4.34 (dd, J=20.0, 9.3 Hz, 1H), 3.85 (dt, J=9.8, 4.5 Hz, 1H), 3.43 (d, J=4.5 Hz, 2H), 2.16 (s, 3H), 2.12 (s, 3H), 2.09 (s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 171.75, 170.88, 169.29,

146.96, 114.62, 91.75, 74.20, 71.75, 68.73, 51.04, 50.71, 20.78, 20.62, 20.39. Theoretical m/z calculated for $C_{14}H_{17}F_3N_4NaO_8$ $[M+Na]^+$: 449.0891. HRMS found: 449.0874.

3,4-di-*O*-acetyl-2,6-dideoxy-2-trifluoroacetamido-6-azido-β-D-glucopyranose (12). Compound **11** (180 mg, 0.42 mmol) was dissolved in DMF (2 mL), ammonium acetate (100 mg, 1.30 mmol) was added as a solid. The reaction mixture was stirred at RT for 24 h when TLC showed complete conversion of substrates. The reaction was poured into EtOAc (20 mL) and washed with brine (20 mL) for three times. The organic layer was dried over Na_2SO_4 and concentrated under vacuum. The product was purified by column chromatography, eluting with hexanes:EtOAc (2:1) to get a white solid in 54% yield. ^{1}H NMR (400 MHz, $CDCl_3$): δ 6.78 (d, J=9.0 Hz, 1H), 5.39 (dd, J=10.5, 9.5 Hz, 1H), 5.11 (dd, J=9.9 Hz, 9.5 Hz, 1H), 4.34 (dt, J=10.0, 3.8Hz, 1H), 4.24 (dd, J=10.3, 5.0 Hz, 1H), 4.02 (s, 1H), 3.38 (m, 2H), 2.09 (s, 3H), 2.06 (s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 171.38, 169.53, 157.09 (q), 116.96 (q), 90.81, 70.26, 69.02, 68.99, 52.77, 51.06, 20.63, 20.46. Theoretical m/z calculated for $C_{12}H_{15}F_3N_4NaO_7$ $[M+Na]^+$: 407.0785. HRMS found: 407.0769.

Diallyl(3,4-di-*O*-acetyl-2,6-dideoxy-2-trifluoroacetamido-6-azido-α-D-glucopanosyl) phosphate (13). Compound **12** (88.6 mg, 0.23 mmol) was dissolved in DCM, ^{1}H-tetrazol (0.45 M in ACN, 2.7 mL, 1.22 mmol) and diallyl N,N-diisopropylphosphoramidite (0.18 mL, 0.68 mmol) was added to the reaction. The mixture was stirred at RT for 3 h when TLC showed complete conversion of substrates. The reaction was cooled to -40 °C with dry ice/ACN. 3-chloroperbenzoic acid (77%, 268.4 mg, 1.20 mmol) was added as a solid. The reaction was stirred at -40 °C for 30 min, then allowed to warm to 0 °C for 30 min. The reaction mixture was diluted with 20 mL DCM followed by washed with 20 mL 10% Na_2SO_3 twice and 20 mL saturated $NaHCO_3$ twice. The organic layer was dried over Na_2SO_4 and concentrated under vacuum. The product was purified

by column chromatography using hexanes:EtOAc (1:1) to give a colorless liquid (68% yield). ^{1}H NMR (400 MHz, $CDCl_3$): δ 7.22 (d, J=8.8 Hz, 1H), 5.96 (m, 2H), 5.78 (dd, J=6.4, 3.3 Hz, 1H), 5.35 (m, 5H), 5.18 (dd, J=9.9, 9.7 Hz, 1H), 4.61 (m, 4H), 4.39 (dd, J=10.7, 8.7 Hz, 1H), 4.21 (ddd, J=8.2, 5.7, 2.8 Hz, 1H), 3.42 (dd, J=13.7, 2.8 Hz, 1H), 3.34 (dd, J=13.7, 5.7 Hz, 1H), 2.06 (s, 3H), 2.04 (s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 171.39, 169.38, 157.88 (q), 132.12 (q), 119.60, 119.52, 117.09, 114.23, 94.86, 71.12, 69.71, 69.28, 68.49, 52.89, 52.81, 50.95, 20.75, 20.62. Theoretical m/z calculated for $C_{18}H_{24}F_3N_4NaO_{10}P$ $[M+Na]^+$: 567.1074. HRMS found: 567.1068.

Uridine 5'-diphospho-2,6-dideoxy-2-amino-6-azido-α-D-glucopyranoside (14). Compound **13** (50 mg, 0.092 mmol), sodium *p*-toluenesulfinate (32.7 mg, 0.184 mmol) and tetrakis(triphenylphosphine)palladium (5.3 mg, 0.0045 mmol) were dissolved in THF:MeOH (1:1, 1.6 mL) and stirred at RT for 3 h. The solvent was evaporated under vacuum. Uridine 5′-monophosphomorpholidate 4-morpholine-N,N′-dicyclohexylcarboxamidine salt (100.9 mg, 0.147 mmol) and *N*-methylimidazolium chloride (37.7 mg, 0.459 mmol) was added to the crude product and dissolved in DMF (2 mL). The reaction was stirred at RT overnight and the solvent was evaporated under vacuum. The residue was dissolved in MeOH:H_2O (3:1, 2 mL) and K_2CO_3 (50 mg) was added. The reaction was stirred at RT for 2 h. The mixture was concentrated and the product was purified by silica gel column chromatography, eluting with EtOAc:MeOH:H_2O (5:2:1) to give a white solid in 48% yield. ^{1}H NMR (400 MHz, D_2O): δ 7.95 (d, J=8.1 Hz, 1H), 5.97 (m, 2H), 5.67 (m, 1H), 4.30-4.02 (m, 5H), 3.94-3.80 (m, 2H), 3.73-3.54 (m, 3H), 3.26 (m, 1H). ^{13}C NMR (101 MHz, D_2O) δ 162.99, 146.91, 144.75, 108.70, 95.92, 94.57, 80.66, 80.53, 79.74, 73.85, 71.08, 68.28, 61.17, 59.27, 53.67. Theoretical m/z calculated for $C_{15}H_{23}N_6O_{15}P_2$ $[M-H]^-$: 589.0702. HRMS found: 589.0763.

GEP1A (3). Compound **9** (20.9 mg, 0.17 mmol) was dissolved in DCM (1 mL) and two drops of DMF was added. The solution was cooled to 0 °C and oxalyl chloride (22.3 µL, 0.26 mmol) was added dropwise. The resulting solution was stirred at 0 °C for 20 min and allowed to warm to RT for 4 h. DCM and excess of oxalyl chloride were evaporated under vacuum. The residue was dissolved in ACN (1 mL) and added dropwise into the solution of **14** (5 mg, 0.0087 mmol) and $NaHCO_3$ (120 mg) in ACN:H_2O (1:1, 5 mL) at 0 °C. The reaction mixture was stirred at 0 °C for 4 h. The solution was concentrated under vacuum and taken up in H_2O (2 mL) and washed with DCM (2 mL). The organic layer was extracted with H_2O twice. The combined aqueous layer was concentrated under vacuum. The product was purified by Bio-Gel P-2 SEC column (23 cm) and lyophilized to white powder (1.3 mg, 22% yield). ^{1}H NMR (400 MHz, D_2O): δ 7.98 (d, J=8.0 Hz, 1H), 6.89 (td, J=15.3, 6.3 Hz, 1H), 6.42 (td, J=15.3, 1.3 Hz, 1H), 5.97 (m, 2H), 5.58 (dd, 7.0, 3.3Hz, 1H), 4.38 (m, 2H), 4.31 (dd, J=6.3, 1.2Hz, 2H), 4.29-4.19 (m, 2H), 4.16 (td, J=10.6, 3.0 Hz, 1H), 4.11 (td, J=10.0, 2.9 Hz, 1H), 3.87 (dd, J=10.3, 9.5 Hz, 1H), 3.77 (dd, J=13.7, 2.5 Hz, 1H), 3.70-3.62 (m, 2H), 3.38 (m, 1H). Theoretical m/z calculated for $C_{19}H_{27}ClN_6O_{16}P_2$ $[M-2H]^{2-}$: 345.0324. HRMS found: 345.0453.

2.2.2 Biochemical and mass spectrometric experiments

NUP62 protein expression and purification. Human NUP62 expression plasmid in pET21a vector (a kind gift from Dr. Suzanne Walker's lab) was transformed into *E. coli* BL21(DE3) competent cells and expressed as a fusion protein containing a C-terminal His_6-tag. The transformant was grown at 37 °C in LB medium. After OD_{600} reached 0.4, the culture was then induced with 0.2 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 3 h. Cells were pelleted and resuspended in lysis buffer (50 mM Tris pH 8.0, 10 mM EDTA, 0.5 M NaCl, and 1 mM PMSF). The cell suspension was lysed with ultra-high-pressure cell disrupter Emulsiflex-C5

(Avestin) followed by centrifugation at 20,500 g at 4 °C for 30 min. The collected pellets containing the inclusion body of NUP62 were then washed twice with 30 mL detergent solution (1.1 M urea, 2% Triton X-100) and resuspended in 30 mL denaturing buffer (8 M urea, 50 mM Tris pH 8.0, 1 mM EDTA, 2 mM DTT) and incubated at room temperature (RT) to dissolve the inclusion body. The lysate was then centrifuged at 16,100 g at 20 °C for 20 min to remove the unbroken pellets. The supernatant was centrifuged at 50,000 g at 20 °C for 15 min to further clear up the lysate. Proteins were dialyzed against 2 M urea in 20 mM Tris pH 7.5 for 4 h followed by 20 mM Tris pH 7.5 overnight at 4 °C. Following dialysis, samples were centrifuged at 16,100 g at 4 °C for 20 min, the supernatant containing NUP62 was collected, concentrated, and stored at –80 °C until use. The purity of NUP62 protein was examined by SDS-PAGE gel.

$OGT_{4.5}$ protein purification for crystallization. The $OGT_{4.5}$ expression plasmid was a kind gift from Dr. Suzanne Walker's lab. Briefly, the $OGT_{4.5}$ construct (spanning residues 313–1031 based on the numbering of the full-length human protein) was cloned into a pET24b vector. The plasmid was transformed into *E. coli* BL21(DE3) competent cells, and the transformant was grown at 37 °C in LB medium. After OD_{600} reached 1.0, the culture was induced by 0.2 mM IPTG at 16 °C for 16 h. The cells were pelleted, re-suspended in TBS (20 mM Tris pH 8.0, 150 mM NaCl) supplemented with 1 mM PMSF, and lysed with ultra-high-pressure cell disrupter. After centrifugation, the supernatant was subjected to Ni-NTA column for affinity purification. The recombinant protein was subsequently eluted with buffer containing 20 mM Tris (pH 8.0), 150 mM NaCl, 250 mM imidazole and 0.5 mM Tris(3-hydroxypropyl)phosphine (THP). The N-terminal His_6-tag was cleaved by HRV3C protease. Further purification was performed by size-exclusion chromatography (Superdex 200 increase 10/300, GE Healthcare) in the buffer containing 20 mM

Tris (pH 8.0), 150 mM NaCl, and 0.5 mM THP. The purified protein was concentrated to 8 mg/mL for crystallization.

Crystallization. All crystals were generated by mixing 1 μL of protein with 1 μL of reservoir solution and were equilibrated against 160 μL of reservoir solution using the hanging-drop vapor-diffusion method at 20 °C. For the glycosylated OGT:**GEP1**:CKII complex, $OGT_{4.5}$ was incubated with 1 mM **GEP1** and 2 mM CKII peptide for 1 h on ice. Crystals were obtained in the reservoir solution containing 0.08 M BIS-TRIS propane (pH 7.0), 0.02 M sodium cacodylate trihydrate (pH 6.5), 2.8 M sodium formate, 0.04 M ammonium sulfate, and 6% w/v polyethylene glycol 8,000. For the crosslinked OGT:**GEP1**:CKII complex, $OGT_{4.5}$ was preincubated with 1 mM **GEP1** for 30 min at RT before adding 2 mM CKII peptide and incubated for another 2 h on ice. Crystallization condition contains 0.08 M Tris (pH 8.5), 0.02 M sodium cacodylate trihydrate (pH 6.5), 1.24 M ammonium sulfate, and 6% w/v polyethylene glycol 8,000. All crystals were transferred into cryoprotectant solution containing their respective mother liquor plus 10% (v/v) glycerol, before being flash-frozen in liquid nitrogen for storage.

Data collection and X-ray structure determination. All X-ray data were collected on the Life Sciences Collaborative Access Team (LS-CAT) beamline 21-ID-D at Advanced Photon Source, Argonne National Laboratory, IL. The wavelength for data collection was 0.9785 Å. Datasets were processed using HKL2000 package.[13] The crystal of glycosylated OGT:**GEP1**:CKII belongs to the space group of *F*222 and contains one molecule per asymmetric unit, with cell parameters: a = 139.1 Å, b = 152.7 Å, c = 199.3 Å, α = β = γ = 90°. The crystal of crosslinked OGT:**GEP1**:CKII complex belongs to the space group of *C*2221 and has two molecules per asymmetric unit. The structures were solved by molecular replacement, using OGT as a search model (PDB 3PE3, ref. [2]). Iterative model building was performed in COOT[14] and refinement was completed in

PHENIX[15] and CCP4.[16] Final statistics were summarized in Supplementary Table 2. All structural figures were prepared using the program PyMOL (DeLano Scientific, http://www.pymol.org/). The coordinates and structure factors have been deposited with PDB codes 5VIF and 5VIE.

Mutagenesis and purification of OGA mutant. OGA-D175N and various OGT mutants were generated using the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent) according to the manufacturer's instructions. Human OGA and full-length OGT (plasmids were kind gifts from Dr. Suzanne Walker's lab) were used as the DNA templates along with the primers listed in Supplementary Table 5. The DNA sequences were verified by sequencing. Wild-type full-length OGT and mutant proteins were expressed and purified similarly as $OGT_{4.5}$ as mentioned above. For OGA-D175N, the mutant plasmid was transformed into *E. coli* Rosetta (DE3) competent cells, and the transformant was grown at 37 °C in LB medium. After OD_{600} reached 0.6, the culture was induced by 0.3 mM IPTG at 16 °C for 16 h. The cells were pelleted, lysed, and purified by Ni-NTA column followed by size-exclusion chromatography as mentioned above. The purified protein was concentrated to 5 mg/mL and stored at –80 °C before use.

Intact protein MS to detect the covalent modification of GEP1/GEP1A on OGT. A typical reaction was set up as below: purified OGT (20 μM wild-type $OGT_{4.5}$ or mutant) was incubated in a 25 μL reaction containing 200 μM of **GEP1**/**GEP1A** in the reaction buffer (10 mM Tris, pH 8.0, 75 mM NaCl, 0.5 mM THP, 200 U/mL CIP alkaline phosphatase, and 30 mM $MgCl_2$) at RT for 6 h followed by C8 StageTip[17] desalting. Samples were SpeedVac dried and redissolved in 20 μL 0.1% formic acid. Mass measurement of intact protein was performed on Q-TOF Maxis 4G (Bruker) or Impact II (Bruker) coupled with an ACQUITY UPLC (Waters). Samples were loaded onto a 2.1 × 100 mm BEH C4 column (1.7 μm, 300 Å, Waters). The injection volume was 9 μL with a flow rate of 300 μL/min. The mobile phases consisted of 0.1% formic acid (solvent A) and

0.1% formic acid in 95% acetonitrile (solvent B). LC program: 5% B for 10 min, 5–60% B for 15 min, 60–90% B for 1 min, 90% B for 5 min, and 90–5% B for 1 min. MS analysis was operated in positive mode with electrospray voltage of 3.8 kV. The end plate offset and nebulizer pressure were –500 V and 2.1 bar, respectively. The interface heater temperature was set at 220 °C with the dry gas flow rate 10 L/min. Data was acquired using one full MS scan (*m/z* 700–3,000) with the scan rate at 1 Hz. Funnel 1 RF and multiple RF were set to 400 eV for ion transfer. The ion energy of quadrupole was 3 eV, the collision energy was 6 eV, the transfer time was 120 μs, and the pre pulse storage time was 25 μs. Time-course experiments were performed similarly as above by incubating the reactions at 37 °C in two replicates for each indicated time points. Dose-dependent experiments were carried out with 40 or 200 μM of **GEP1** in a 10 μL reaction for 6 h at 37 °C. LC-MS data were processed and analyzed using Compass Data Analysis software (version 4.1, Bruker). The major LC peak corresponding to highest UV signal was selected for deconvolution performed by maximum entropy algorithm. The parameters of maximum entropy include mass range 50,000–100,000 Da, auto data point spacing, and the resolving power of 10,000. The increased mass resulting from labeling of **GEP1**/**GEP1A** was determined by subtracting the original protein mass from the deconvoluted protein mass.

Sample preparation for detection of GEP1/GEP1A-modified peptides. To detect **GEP1**/**GEP1A**-labeled OGT peptides, purified full-length OGT protein (70 μM) was incubated in a 10 μL reaction containing 350 or 700 μM of **GEP1**/**GEP1A** in the reaction buffer (10 mM Tris, pH 8.0, 75 mM NaCl, 0.5 mM THP, 200 U/mL CIP alkaline phosphatase, and 30 mM $MgCl_2$) at RT. The reaction was stopped at 6 h or other indicated time points by 10 mM DTT and 8 M urea. For trypsin digestion, the protein sample was diluted to 0.5 μg/μL in denature buffer (8 M urea and 10 mM Tris, pH 8.0). After reduction by 10 mM DTT for 30 min at RT and carbidomethylation

with 55 mM iodoacetamide in the dark for 30 min at RT, alkylated proteins were seven-fold diluted using 25 mM NH_4HCO_3 and then digested by Trypsin/Lys-C mix (protein:protease = 40:1) (Promega) at RT for 18 h. Formic acid was added to the sample to reach final 0.5% for enzyme inactivation and sample acidification. Peptides were SpeedVac dried after SDB-XC StageTip desalting. To measure the irreversible reaction rate of **GEP1** with C917, peptides (5 μg) from each reaction were redissolved in 50 μL of 0.1 M triethylammonium bicarbonate (TEAB) for dimethyl labeling.[18] The peptide samples prepared from the reactions of OGT with **GEP1** at concentration ratios of 1:5 and 1:10 were mixed with 2 μL of 4% formaldehyde-H_2 (Sigma-Aldrich) and 4% formaldehyde-D_2 (Sigma-Aldrich), respectively. Each sample was then mixed with freshly prepared 0.6 M sodium cyanoborohydride (2 μL) and then incubated for 1 h at RT. The reaction was quenched by addition of 8 μL of 1% NH_4OH. Formic acid (10%, 10 μL) was added to further stop the reaction and acidify the samples. Finally, the H- and D-labeled samples were combined at 1:1 ratio followed by SDB-XC StageTips desalting and SpeedVac dry. To detect the **GEP1**/**GEP1A**-glycosylated peptides, α-crystallin B chain peptide (200 μM, peptide sequence: 38–50, FPTSTSLSPFYLR, synthesized by Biomatik) or Lamin B1 peptide (200 μM, peptide sequence: 389–401, KLSPSPSSRVTVSK-biotin, synthesized by Biomatik) was incubated with 20 μM purified OGT in a 5 μL reaction containing 1 mM **GEP1**/**GEP1A** in the reaction buffer (10 mM Tris, pH 8.0, 75 mM NaCl, 0.5 mM THP, 200 U/mL CIP alkaline phosphatase, and 30 mM $MgCl_2$) at RT for 6 h. Peptide samples were then purified by SDB-XC StageTips and dried by SpeedVac. All peptide samples were stored at –80 °C until LC-MS/MS analysis.

nanoLC-MS/MS analysis. Peptide samples were dissolved in 0.1% formic acid for LC-MS/MS analysis on an Orbitrap Q-Exactive (Thermo Scientific) equipped with a nanoAcaquity UPLC system (Waters). Peptides were loaded onto a 75 μm × 15 cm 1.7 μm BEH C18 column at a flow

rate of 300 nL/min. Mobile phase A consisted of 0.1% formic acid, and solvent B was 0.1% formic acid in acetonitrile. A linear gradient of 0–4% B for 0.1 min, 4–35% B for 30 min, 35–75% B for 0.1 min, 75% B for 9.5 min, 75–95% B for 0.1 min, 95% B for 9.5 min, 95–0% B for 0.5 min, and 0% B for 9.5 min was employed throughout this study. Mass spectra from full scans were acquired in a data-dependent mode (*m/z* 200–2,000). The resolution of survey scan was set to 17,500 at *m/z* 400 with an automated gain control (AGC) value of 10^6. The top 15 most-intense precursor ions were selected from the MS scan for subsequent higher energy collisional dissociation (HCD, normalized collision energy 30 eV) MS/MS scan. Peptide identification was performed by MaxQuant (v1.5.6.5 or v1.5.8.3)[19,20] against a composite target-decoy protein sequence database containing Uniprot database (release 2015_04, subset human, 20,265 protein entries).[21] The search criteria used in this study include trypsin specificity allowing up to 2 missed cleavages, and variable modifications of **GEP1** ($C_{10}H_{15}NO_6$) on Cys, **GEP1A** ($C_{10}H_{14}N_4O_5$) on Cys, **GEP1** glycosylation ($C_{10}H_{14}ClNO_5$) on Ser/Thr, **GEP1A** glycosylation ($C_{10}H_{13}ClN_4O_4$) on Ser/Thr, carbamidomethyl on Cys, and oxidation on Met. The precursor mass tolerance and the fragment ion tolerance were set at ± 10 ppm and ± 0.6 Da, respectively. Peptide was considered identified based on the posterior error probability with a false discovery rate of 1%. The spectra of peptides were manually inspected.

To determine the reaction rate of **GEP1** with OGT, raw MS spectra from two biological samples were processed using MaxQuant and precursor intensities was calculated. The search criteria and database were set up as described above. Peptide abundance was normalized to the total protein intensity of OGT detected in the same sample. Based on the previous report,[22] the reaction rate (*k*) of **GEP1** with C917 of OGT in a second-order reaction can be calculated using eq. 1.

$$\ln([P]\,[X_0]/[X][P_0]) = -kt[X_0] + kt[P_0] \qquad \text{[eq. 1]}$$

Where $[P_0]$ is the initial abundance of unmodified C917 peptide, $[X_0]$ is the initial concentration of **GEP1**, $[P]$ is the abundance of unmodified C917 peptide at time *t*, $[X]$ is the concentration of **GEP1** at time *t*, which can be estimated using eq. 2.

$$X = X_0 \bullet \frac{(P_0 - P)}{P_0} \qquad \text{[eq. 2]}$$

GEP1A fluorescent assay. To detect the modification of **GEP1A** on OGT and NUP62 protein substrate, purified OGT (1 μM of $OGT_{4.5}$, full-length OGT, or mutants) was incubated with NUP62 (5 μM) and **GEP1A** (25 μM) in a 12 μL reaction containing buffer (20 mM Tris pH 8.0, 150 mM NaCl, and 0.5 mM THP) and incubated at 37 °C for 30 min. For competition assay, different doses of UDP-GlcNAc ranging from 0–50 μM were added into the reaction mixture together with **GEP1A**. The reactions were quenched by boiling at 95 °C for 5 min. SDS (1%) was added and incubated at RT for 5 min to help dissolve precipitated proteins. The samples were further diluted to final 0.1% SDS with TBS buffer. Click chemistry reagents at the final concentration of 1 mM $CuSO_4$, 0.1 mM Tris[(1-benzyl-1H-1,2,3-triazol-4-yl)methyl]amine (TBTA), 50 μM fluor 488-alkyne and 1 mM sodium ascorbate were sequentially mixed and immediately added to each sample. Click chemistry reaction was performed at RT for 1 h in the dark. Proteins were precipitated in ice-cold methanol for 2 h at –80 °C followed by 16,000 g for 10 min to pellet the proteins. The protein pellet was washed by MeOH and then re-dissolved in TBS buffer (20 mM Tris pH 8.0 and 150 mM NaCl) containing 4% SDS. The dissolved protein samples were separated on SDS-PAGE gel, followed by fluorescence detection and Coomassie Brilliant Blue staining for relative quantification. Both in-gel fluorescence scanning and Coomassie Blue stained gels were detected using ChemiDoc-It 2 imager equipped with BioLite MultiSpectral light source (UVP)

and quantified using ImageJ (v1.6.0_24).[57] The levels of **GEP1A**-modified proteins were normalized to the protein amount and relatively quantified. For the detection of **GEP1A** modification on OGT and OGA-D175N protein substrate, purified OGT (0.35 μM of $OGT_{4.5}$, full-length OGT or mutants) was incubated with OGA-D175N (16 μM) and **GEP1A** (25 μM) in a 12 μL reaction buffer (TBS with 0.5 mM THP) and incubated at 37 °C for 15 min. Subsequent click chemistry reactions were performed following the same protocol mentioned above. The reaction mixtures were then boiled with SDS loading buffer at 95 °C for 5 min without protein precipitation. In-gel fluorescence scanning and relative quantification were performed as mentioned above.

MCF7 nuclear extracts. The MCF7 cell line was purchased from American Type Culture Collection (ATCC). The cell line has been authenticated followed the guidelines from ATCC and was tested to be free of mycoplasma contamination. We periodically analyze the growth curve and check the morphology under microscopy to ensure the culture consistency and no adverse effect on cell behavior. MCF7 cells were maintained in Dulbecco's Modified Eagle Medium with 10% fetal bovine serum (FBS) at 37 °C in a 5% CO_2 incubator. Cells were plated at 70% confluence and starved in 2% FBS medium without glucose for 24 h before harvesting. Cell pellets were washed by PBS and lysed in buffer (10 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, and 10 mM KCl). The sample was centrifuged at 10,000 g for 20 min at 4 °C to pellet nucleus. The nuclear proteins were extracted by resuspending the nuclear pellet in buffer (20 mM HEPES, pH 7.9, 25% (v/v) glycerol, 0.42 M NaCl, 1.5 mM $MgCl_2$, and 0.2 mM EDTA) and incubated at 4 °C for 1 h. Cell debris was then removed by centrifugation at 16,000 g for 10 min at 4 °C. The nuclear extracts were stored at –80 °C in aliquots before use.

GEP1 preincubation-induced crosslinking of OGT with substrates. To detect the crosslinking of OGT with substrates, purified OGT (1 μM) was preincubated with 100 μM **GEP1** in the reaction

buffer (10 mM Tris, pH 8.0, 75 mM NaCl, 1 mM THP, 100 U/mL CIP alkaline phosphatase, and 30 mM $MgCl_2$) for 30 min at 37 °C, followed by adding 5 μM of NUP62 protein or 100 μM of biotinylated peptides to react at 37 °C for another 3 h. The reactions were analyzed by SDS-PAGE gel or western blot. Biotinylated peptide RBL2 (410–422, KENSPAVTPVSTAK-biotin), Lamin B1 (389–401, KLSPSPSSRVTVSK-biotin), and IRS1 (982–994, VPSSRGDYMTMQMK-biotin) were custom synthesized at Biomatik. For OGT crosslinking with proteins from the MCF7 nuclear extracts, full-length OGT (0.2 μM) and 50 μM **GEP1** were prepared similarly as described above and reacted with 40 μg of nuclear extracts at 37 °C for 5.5 h before western blot analysis. Nitrocellulose membrane (Life Technology) was blocked with 0.9% bovine serum albumin for 1 h and then probed with streptavidin-HRP (Sigma-Aldrich #GERPN1231) or anti-OGT antibody (Proteintech #11576-2-AP). All primary antibodies were used in 1: 5,000 dilution. Secondary anti-lgG-HRP antibody (Proteintech #SA00001-2) was used in 1: 50,000 dilution. Crosslinked peptides and proteins were detected using an enhanced chemiluminescence detection kit (Life Technology) on ChemiDoc XRS+ imager (Bio-Rad Hercules).

Data availability. The data that support the findings of this study are available in the Supplementary Information, Supplementary Note, and from the corresponding author upon reasonable request. X-ray crystallographic data that support the findings of this study have been deposited in the Protein Data Bank with the accession codes 5VIF and 5VIE for glycosylated OGT:**GEP1**:CKII and crosslinked OGT:**GEP1**:CKII complexes, respectively.

2.3 Results and discussions

2.3.1 Design and synthesis of the probes

The design of the probes was based on the crystal structure of OGT in complex with UDP-GlcNAc (PDB 4GZ5).[3] (Figure 2.3) We noticed that the N-acetyl group of UDP-GlcNAc was in close approximate to a cysteine, C917, in the active site of OGT. This cysteine does not have catalytic importance and it is not found in the same position of known glycotransferases with structural similarity to OGT. In light of this finding, we anticipated that we could extend the N-acetyl of UDP-GlcNAc with electrophilic groups, which may form covalent bond with C917 through nucleophilic attack, providing a covalent labeling of OGT with the probe. At the same time, the sugar could still be transferred to the substrate like UDP-GlcNAc, forming labeled substrates. According to the ordered bi-bi mechanism of OGT, a probe that can covalently label OGT will follow a revised kinetic scheme. (Figure 2.4a) Comparing the relative ratio of labeled OGT and modified substrate, we would be able to distinguish the sugar binding and protein binding ability of OGT. Thus, by testing a series of mutants of OGT with this probe, the information about which residue affecting sugar or protein binding could be readily read out by the assay, providing information on how OGT recognizes different substrates.

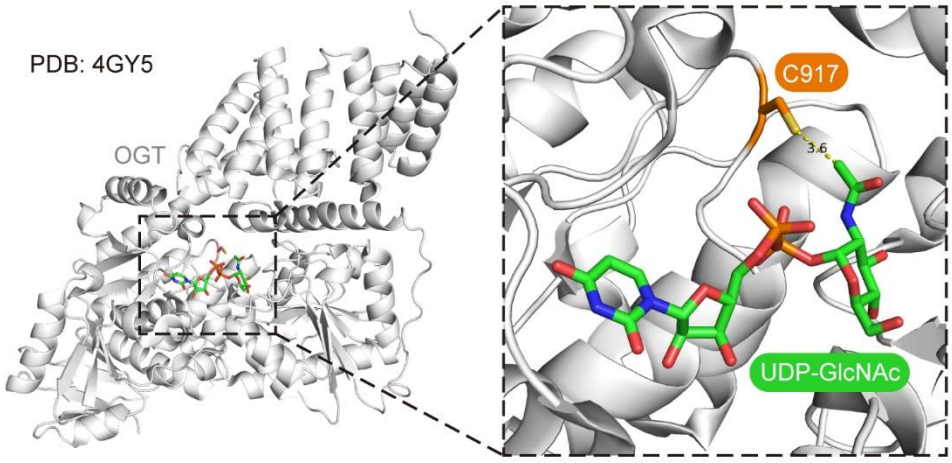


Figure 2.3 Design rational of GEPs. OGT is shown in white ribbon and C917 is highlighted with orange sticks

Following the design of GlcNAc Electrophilic Probe (GEP), we designed a panel of UDP-GlcNAc analogues carrying electrophilic groups with different length and reactivity. (Figure 2.4b) Two examples were shown here. **GEP1** had an extended allyl chloride side chain, while **GEP2** had a double bond connected to UDP-GlcNAc.

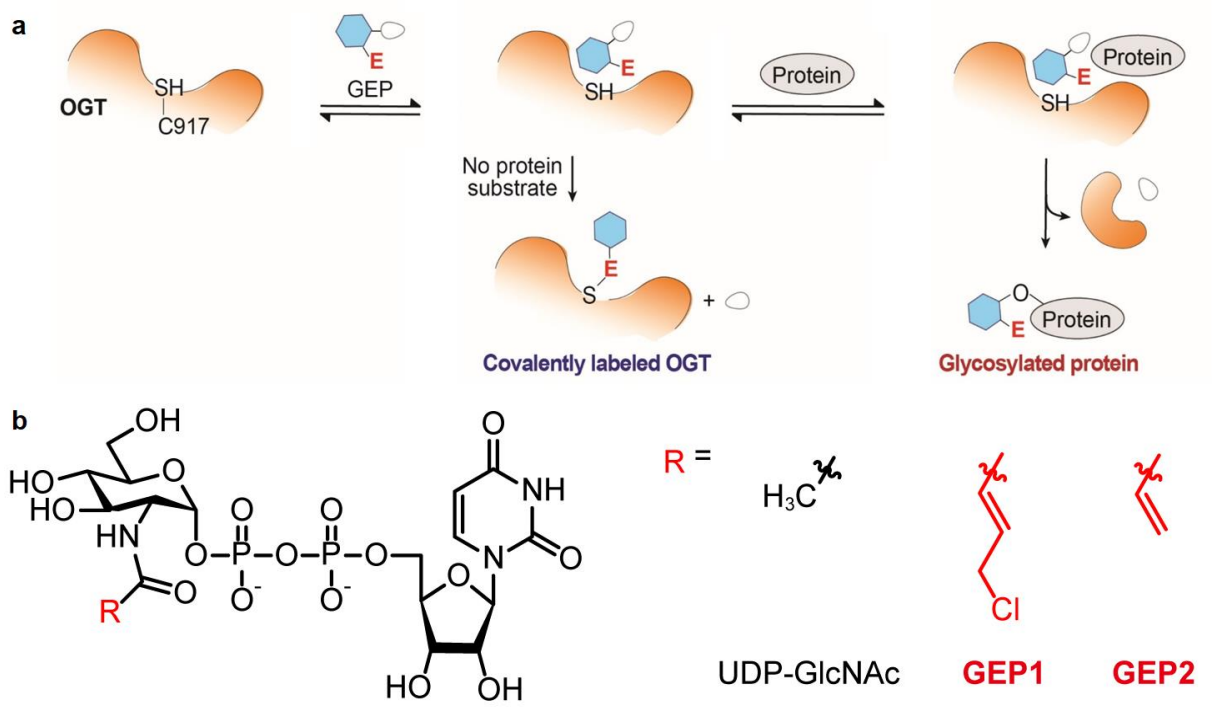


*Figure 2.4 Structure and reaction of the probes with OGT (a) **GEP1** can covalently label OGT and be transferred to the substrate (b) The structures of the designed **GEP**s*

We first carried out the organic synthesis of **GEP1** and **GEP2**. Started with 1,3,4,6-tetra-acetylated glucosamine, the free amine was first protected with a trifluoroacetyl (TFA) group. Originally, we tried to attach the side chain in the beginning of the synthesis. However, in the later part of the synthesis, the highly reactive electrophilic groups would undergo side reactions with reagents. In order to accommodate the reactive side chains, TFA was chosen as the protecting group. With the protected glucosamine, the 1-position was selectively deprotected for the attachment of UDP with ammonium acetate in DMF. Allyl protected phosphite was first attached onto the 1-position, followed by the oxidation by mCPBA to mono-phosphorylated GlcNAc. Next, the allyl protecting groups were hydrolyzed by tetrakis(triphenylphosphine)palladium(0), allowing the attachment of uridine monophosphate. All acetyl and trifluoroacetyl groups were deprotected before the

attachment of different electrophilic side chains. **GEP1** and **GEP2** were synthesized with 20% and 60% yield, respectively.

2.3.2 Validation of **GEP1** labeling OGT and protein substrate

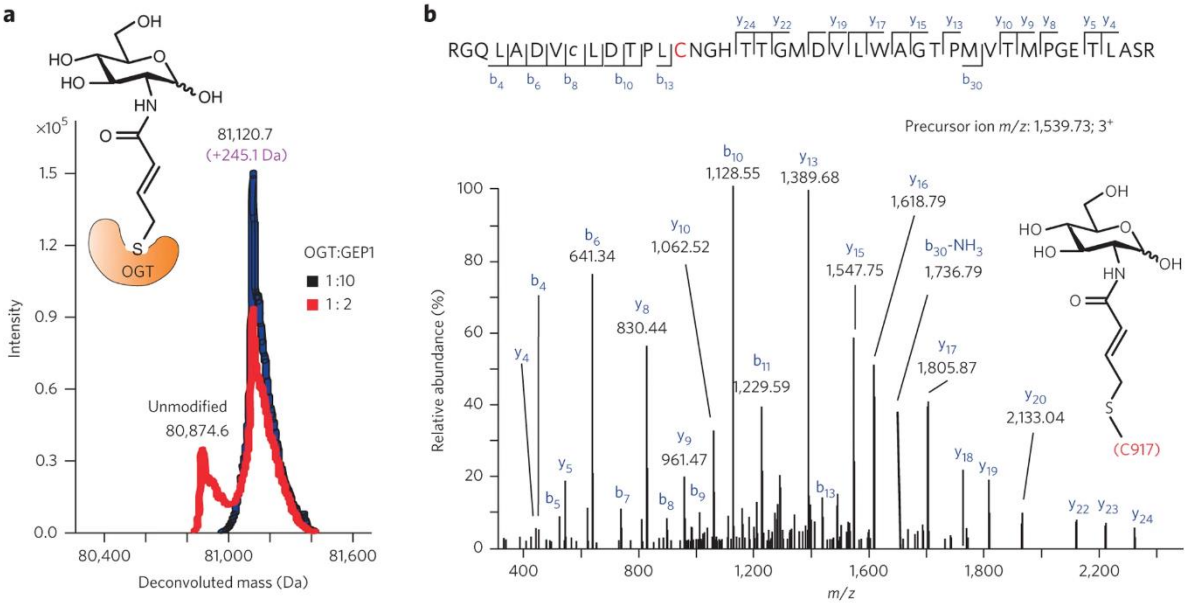


*Figure 2.5 MS data demonstrated that **GEP1** could label C917 of OGT (a) Intact protein MS indicated that **GEP1** labeled OGT with a 1:1 molar ratio (b) LC-MS/MS showed that **GEP1** specifically reacted with C917*

With the synthesized probes, we first tested whether the probes were able to modify OGT specifically on C917 as we proposed. Results showed that only **GEP1** was able to achieve the covalent labeling. After the incubation of **GEP1** and OGT at 2:1 molar ratio, the peak of one molecule **GEP1** labeled OGT was found by intact MS with an 80% labeling ratio. (Figure 2.5a) The ratio will increase to 100% with 10-fold excess of **GEP1**, while only mono-modified peak was found, indicating that only one cysteine could be modified despite there being a dozen surface cysteines on OGT. Further MS/MS experiments confirmed the modification site was C917, not other cysteines. (Figure 2.5b) With the same idea, the glycotransferase product was also confirmed by MS/MS, showing that **GEP1** sugar was transferred to the desired serine of substrate peptide

derived from α-crystallin B chain. (Figure 2.6a) Concurrently, we also solved the crystal structure of the $OGT_{4.5}$:**GEP1**:CKII peptide ternary complex. (Figure 2.6b and 2.6c) The crystal structure demonstrated that **GEP1** sugar was able to be transferred onto substrate, forming a similar conformation as native UDP-GlcNAc (Figure 2.6d). The results validated that **GEP1** was able to covalently modify OGT, and to be transferred to the substrates.

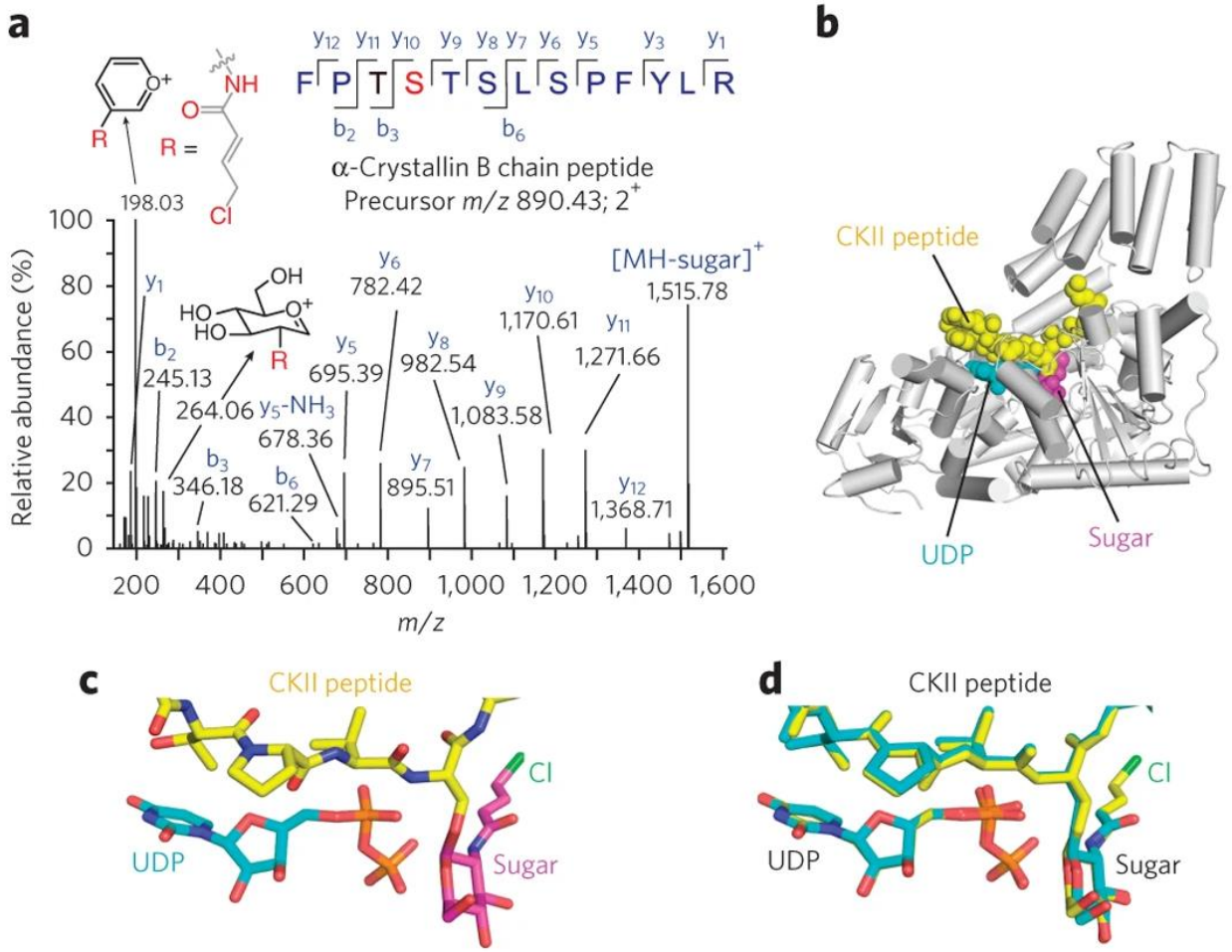


*Figure 2.6 MS and structure data demonstrated that **GEP1** could be transferred to substrates (a) LC-MS/MS mapped the modification of **GEP1** onto serine of substrate (b) Tertiary structure of OGT:**GEP1**:CKII peptide (c) The sugar part of **GEP1** was transferred to the peptide from the structure (d) Sugar transfer of **GEP1** would not affect the conformation of the glycosylated product. Glycosylated products of UDP-GlcNAc are shown in cyan sticks and products of **GEP1** are shown in yellow sticks.*

2.3.3 Design and synthesis of **GPE1A**

With the powerful probe **GEP1**, we sought to achieve rapid readout of two labeled species, labeled OGT and glycosylated substrate. One convenient method is the application of click chemistry to attach a fluorescent probe onto the sugar. From the structure of OGT in complex with UDP-GlcNAc, we anticipated that OGT was able to tolerate an azido group at the 6-position of GlcNAc, which represented the **GEP1A** probe. (Figure 2.7a) With **GEP1A**, the labeled OGT and modified protein can be easily quantified on a fluorescent gel. (Figure 2.7b and 2.7c)

Although **GEP1A** has only one functional group different as **GEP1**, the synthesis is much more complicated. Started with glucosamine, I first protected 6-position with tosyl group, 2-position with TFA group, and 1, 3, 4-position with acetyl groups in a one-pot reaction. Following that the tosyl groups could be easily replaced by azido. For the rest part, based on similar synthesis methods as **GEP1**, we obtained **GEP1A** at 22% yield.

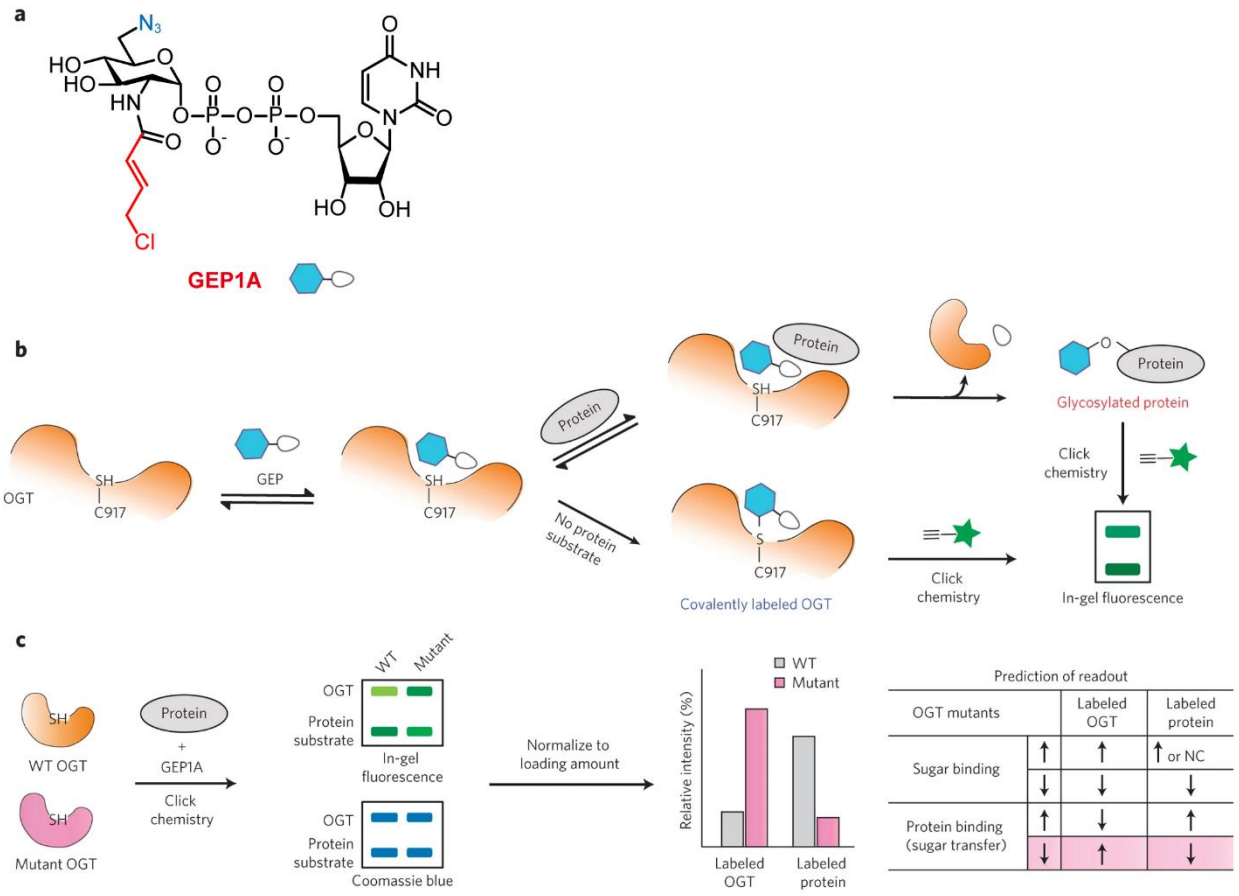


*Figure 2.7 Principal of **GEP1A** assay (a) Structure of **GEP1A** (b) Relative labeling of **GEP1A** could be detected by fluorescent gel (c) The predicted readout of **GEP1A** assay*

2.3.4 Validation of **GEP1A** assay

Similar MS test verified that **GEP1A** could also specifically modify C917 and be transferred to the substrate. Based on the revised kinetic scheme of GEP probes, we summarized a prediction of OGT mutant sugar and protein binding ability based on the readout of fluorescent gel for mutants compared with wild type (WT) OGT. (Figure 2.7c) For an example, if a mutant of OGT has impaired protein binding ability, the glycosylated protein amount will decrease, while the overall amount of modified proteins stays the same, resulting in the increase of modified OGT.

On the other hand, a sugar binding impaired mutant will suppress the total amount of modified protein. This also applies to the sugar or protein binding enhanced mutants of OGT.

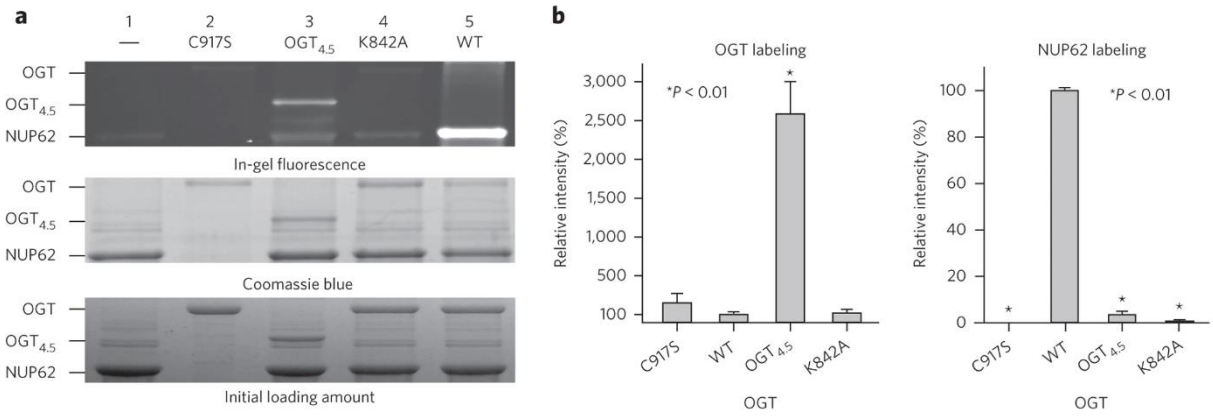


*Figure 2.8 Proof of concept for **GEP1A** assay (a) fluorescent gel result for WT OGT, OGT-C917, $OGA_{4.5}$, and OGT-K842A (b) Relative labeling level of each construct after normalization*

In order to validate our prediction and the concept of **GEP1A** assay, we tested a few known mutants of OGT with **GEP1A** against a well-studied protein substrate, NUP62. First of all, OGT-C917A was tested. With the covalent modification site mutated, C917 showed minimal fluorescent band, indicating that the modification of **GEP1A** is specific on C917. (Figure 2.8a, lanes 2 and 5) $OGT_{4.5}$, with truncated TPR domain, could not bind to protein substrates. From the **GEP1A** assay, $OGT_{4.5}$ showed decreased NUP62 labeling and increased OGT labeling, which agreed with our prediction. (Figure 2.8a, lanes 3 and 5) OGT-K842A represented the sugar binding defective OGT mutants, as K842 was important for the binding with diphosphate of UDP-GlcNAc. In the **GEP1A** assay, OGT-K842A showed declined labeling pattern of both OGT and NUP62 in the fluorescent gel. (Figure 2.8a, lanes 4 and 5) The initial results of **GEP1A** assay were perfectly aligned with our prediction, revealing the accuracy of the assay.

2.3.5 Application of **GEP1A** assay to identify OGT residues responsible for substrate binding

As a preliminary application of **GEP1A** assay, we screened a series of mutants in the TPR domain. One pair of mutations, N321A/N322A, showed a fluorescent pattern same as the prediction of impaired protein binding, indicating that the two asparagine residues might be important for the binding of OGT with NUP62. (Figure 2.9) Radio-labeled kinetic assay verified that the mutants could impair the binding between OGT and NUP62. Additionally, this finding was in agreement with the structural study showing that N321 and N322 was involved in the binding of HCF-1 peptide.

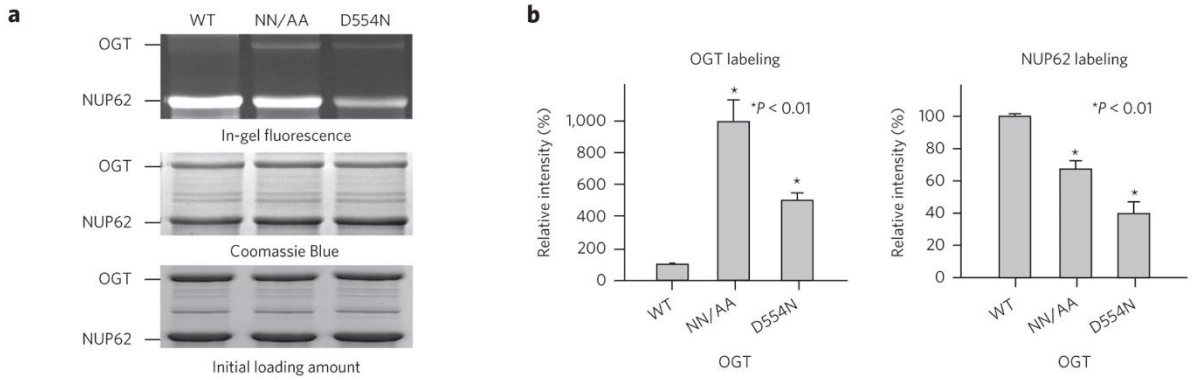


*Figure 2.9 Figure 2.9 **GEP1A** assay was utilized to identify a pair of residues for substrate recognition (a) fluorescent gel result for WT OGT, OGT-NN/AA, and OGT-D554N (b) Relative labeling level of each construct after normalization*

We also applied the **GEP1A** assay to test one of the hypotheses of OGT catalytic mechanism where D554 was involved. Indeed, OGT-D554 showed the labeling pattern same as sugar binding deficient mutant. Our results confirmed that D554 played an important role in the sugar attachment of OGT.

2.3.6 **GEP1** can crosslink OGT with its substrates

Aside from the two scenarios where the probe could modify OGT or be transferred onto substrate, **GEP1** could also serve as a crosslinker between OGT and its substrates under specific conditions. We found that pre-incubation of OGT and **GEP1** before the addition of substrates enabled crosslinking.

Identification of O-GlcNAcylation was extremely challenging, due to the transient cycling of O-GlcNAcylation, low modification level, and weak glycosidic bond. Traditionally, MS has been a powerful method for the detection of O-GlcNAcylated protein, with the problem of profound ion suppression effects on O-GlcNAcylation.

In vitro crosslinking of OGT with peptide or protein substrates by **GEP1** was detected by SDS-PAGE and western blot. Following that, we also solved the crystal structure of crosslinked OGT with CKII peptide. (Figure 2.10a) From the structure, the crosslinking did not perturb the overall conformation of OGT. The electron density in the active site showed unambiguous covalent linkage between the allyl chloride group and C917. (Figure 2.10b) At the same time, the overall conformation of the sugar and peptide didn’t change with the introduction of extended allyl chloride side chain, proving that the modification would not affect the binding of sugar. The results provided solid evidence of the crosslinking function of **GEP1**.

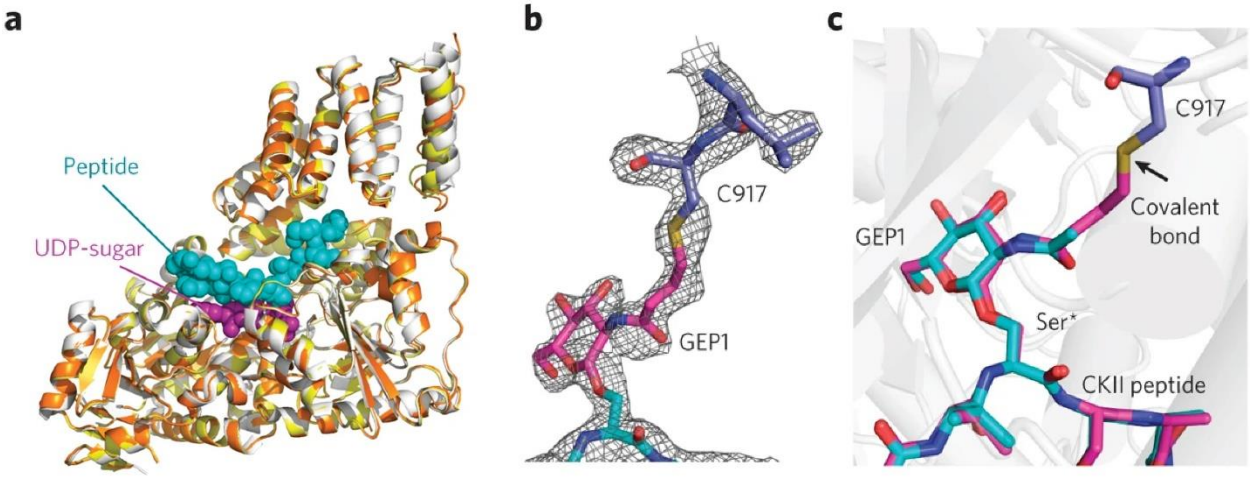


Figure 2.10 ***GEP1*** *could crosslink OGT with its substrate (a) Superposition of OGT:UDP:O-GlcNAcylated CKII peptide (PDB 4GYW, white) with crosslinked complex of OGT:****GEP1****:CKII peptide (orange) (b) Electron density showed the covalent linkage between* ***GEP1*** *and C917 (c) Superposition of ligands from the two complexes. OGT:UDP:O-GlcNAcylated CKII peptide in cyan sticks and crosslinked complex of OGT:****GEP1****:CKII peptide in magenta sticks.*

Taking advantage of the unique feature, we carried out an initial trial to detect OGT crosslinking partners in MCF7 cell lysate. Intriguingly, 102 proteins were enriched with **GEP1** for crosslinking with OGT, and 72 of them were previously identified substrates of OGT. Additionally, 30 more proteins were found to be potential novel substrates. This suggested that **GEP1** could serve as a valuable tool for the identification of OGT substrates.

2.4 Summary and future plan

In summary, we developed and synthesized the GlcNAc Electrophilic Probes (GEPs), which could covalently react with C917 in OGT active site. Applying the 6-azido form, **GEP1A**, we developed an efficient assay to generate fast readout of residues and structural features of OGT that contributed to the substrate binding and recognition. The assay results on a few already known mutants of OGT was in perfectly agreement with our prediction, and it was also applied to detect a few novel residues that are important for the binding of protein substrate.

At the same time, by tuning the reaction condition, the probe **GEP1** could also be applied as a crosslinker to detect OGT substrates. Our initial test indicated that **GEP1** could accurately and efficiently identify novel substrates.

For our future plan, we aim to apply **GEP1A** assay on a large scale of substrates and investigate the different residues in TPR that contributed to the binding of distinct substrates. This will provide valuable information of OGT substrate specificity, and also potential therapeutic solutions for diseases regulated by O-GlcNAcylation. We also plan to utilize **GEP1** to identify novel substrates or close binding partners of OGT.

2.5 References

1. Worth, M., Li, H. & Jiang, J. Deciphering the Functions of Protein *O* -GlcNAcylation with Chemistry. *ACS Chem. Biol.* **12**, 326–335 (2017).

2. Lazarus, M. B., Nam, Y. S., Jiang, J., Sliz, P. & Walker, Suzanne. Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature* **469**, 564-567. (2011).

3. Lazarus, M. B. *et al.* Structural snapshots of the reaction coordinate for O-GlcNAc transferase. *Nat. Chem. Biol.* **8**, 966–968 (2012).

4. Lazarus, M. B. *et al.* HCF-1 is cleaved in the active site of O-GlcNAc transferase. *Science* **342**, 1235–1239 (2013).

5. Rafie, K. *et al.* Recognition of a glycosylation substrate by the O-GlcNAc transferase TPR repeats. *Open Biol.* **7**, (2017).

6. Duff, M. R., Grubbs, J. & Howell, E. E. Isothermal titration calorimetry for measuring macromolecule-ligand affinity. *J. Vis. Exp. JoVE* (2011) doi:10.3791/2796.

7. Patching, S. G. Surface plasmon resonance spectroscopy for characterisation of membrane protein–ligand interactions and its potential for drug discovery. *Biochim. Biophys. Acta BBA - Biomembr.* **1838**, 43–55 (2014).

8. Jerabek-Willemsen, M. *et al.* MicroScale Thermophoresis: Interaction analysis and beyond. *J. Mol. Struct.* **1077**, 101–113 (2014).

9. Hu, C.-W. *et al.* Electrophilic probes for deciphering substrate recognition by O-GlcNAc transferase. *Nat. Chem. Biol.* **13**, 1267–1273 (2017).

10. Silva, D. J., Wang, H., Allanson, N. M., Jain, R. K. & Sofia, M. J. Stereospecific Solution- and Solid-Phase Glycosylations. Synthesis of β-Linked Saccharides and Construction of Disaccharide Libraries Using Phenylsulfenyl 2-Deoxy-2-Trifluoroacetamido Glycopyranosides as Glycosyl Donors [1]. *J. Org. Chem.* **64**, 5926–5929 (1999).

11. Sala, R. F., MacKinnon, S. L., Palcic, M. M. & Tanner, M. E. UDP-N-trifluoroacetylglucosamine as an alternative substrate in N-acetylglucosaminyltransferase reactions. *Carbohydr. Res.* **306**, 127–136 (1998).

12. Tunge, J. A. & Mellegaard, S. R. Selective Selenocatalytic Allylic Chlorination. *Org. Lett.* **6**, 1205–1207 (2004).

13. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).

14. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

15. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 1948–1954 (2002).

16. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **50**, 760–763 (1994).

17. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).

18. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* **4**, 484–494 (2009).

19. Cox, J. & Mann, Matthias. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

20. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).

21. UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42**, D191-198 (2014).

22. Kaur, P., Kiselar, J., Yang, S. & Chance, M. R. Quantitative protein topography analysis and high-resolution structure prediction using hydroxyl radical labeling and tandem-ion mass spectrometry (MS). *Mol. Cell. Proteomics* **14**, 1159–1168 (2015).

2.6 Supplementary information

Table 2.1Data collection and refinement statistics (molecular refinement)

	Glycosylated OGT:GEP1:CKII (PDB 5VIF)	**Crosslinked OGT:GEP1:CKII (PDB 5VIE)**
Data collection		
Space group	*F*222	*C2221*
Cell dimensions		
a, *b*, *c* (Å)	139.1, 152.7, 199.4	138.7, 200.0, 152.9
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90
Resolution (Å)	50.0–2.1 (2.17–2.13)*	50.0-2.60 (2.69–2.60)
R_{merge}	7.3 (36.0)	9.8 (62.9)
$I/\sigma(I)$	40.8 (4.0)	13.8 (1.9)
Completeness (%)	99.6 (98.3)	98.6 (99.9)
Redundancy	6.7 (6.0)	4.3 (4.6)
Refinement		

Resolution (Å)	50.0–2.2	50.0-2.6
No. reflections	58,556	68,209
R_{work} / R_{free}	18.2 / 22.8	24.7 / 30.8
No. atoms		
Protein	5,434	10,554
Ligand/peptide	121	191
Water	191	177
B factors	53.82	41.89
Protein	54.13	42.34
Ligand/peptide	48.14	31.23
Water	47.87	26.02
R.m.s. deviation		
Bond length (Å)	0.019	0.016
Bond angle (°)	2.05	1.97

*Each structure was determined from one crystal.

*Values in parentheses are for highest-resolution shell.

Chapter 3 Development of targeted covalent inhibitors of OGT

The results of this chapter were published on:

Worth, M.; Hu, C.-W.; **Li, H.**; Fan, D.; Estevez, A.; Zhu, D.; Wang, A.; Jiang, J. Targeted Covalent Inhibition of *O* -GlcNAc Transferase in Cells. *Chem. Commun.* **2019**, *55* (88), 13291–13294.

Contributions

This project is a highly collaborative project. I was involved in the organic synthesis part. Together with Dr. Matthew Worth and Arielis Estevez, I synthesized the OGT inhibitors **ES1**, **6AzES1**, and **UDP-ES1**. I also solved the crystal structure of OGT in complex with **UDP-ES1** and CKII peptide.

3.1 Introduction

The dysregulation of O-GlcNAcylation has been related to a variety of diseases, including cancer, diabetes, neurodegenerative diseases, and cardiovascular diseases.[1] The relation between O-GlcNAc, OGT and cancer has been widely studied.[2] In almost every type of tumor cell, OGT level or O-GlcNAc level was found to be elevated. Thus, the suppression of O-GlcNAc, or OGT activity could be a potential therapy for cancer. It has been reported that the knockdown of OGT would suppress the progression of tumor cells.[3] However, OGT as an essential protein, could not be completely knocked out. As a result, the development of OGT inhibitors could be a potential solution for cancer therapy.

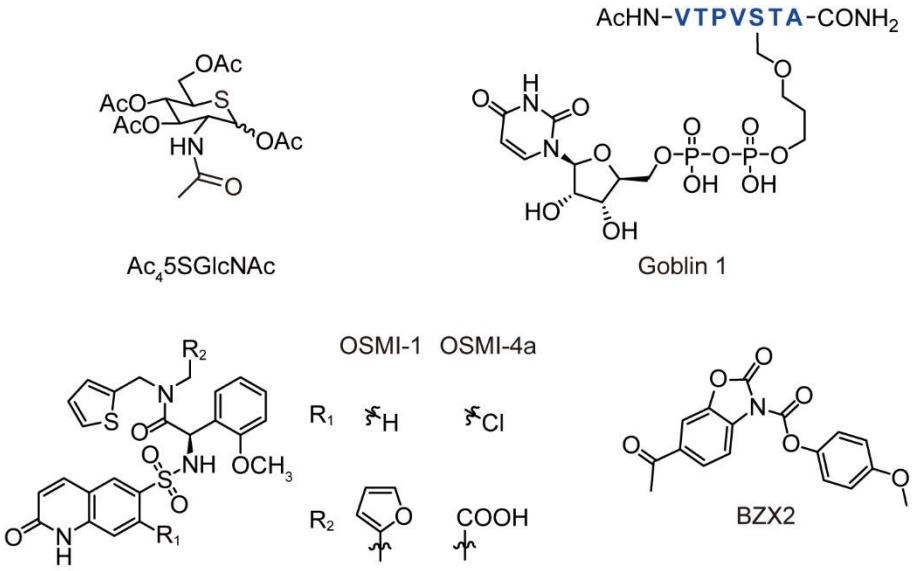


Figure 3.1 Current best inhibitors of OGT

By now, a number of OGT inhibitors have been developed, most of which are reversible inhibitors targeting the active site. Efforts have been made to develop inhibitors based on the substrate of OGT, UDP-GlcNAc. By replacing the oxygen in the pyranose ring of UDP-GlcNAc with sulfur, UDP-5SGlcNAc and its prodrug $Ac_4$5SGlcNAc were designed.[4] The thiol-sugar prevented sugar transfer onto the substrate, inhibiting OGT by competition with UDP-GlcNAc. UDP-5SGlcNAc is one of the best OGT inhibitors, with a Ki value of 8 μM. The drawback of UDP-5SGlcNAc is that it does not have high

specificity for OGT, as it has been reported to inhibit other glycotransferases as well. OGT has two substrate, UDP-GlcNAc and the protein or peptide substrate, which are mimicked by another type of inhibitors derived from the combination of UDP-GlcNAc and peptide. Goblin-1 represents the bi-substrate inhibitors with an IC_{50} value of 18 μM *in vitro*.[5] UDP-peptide inhibitors suffer from the problems that they are not suitable for cellular use. Aside from rationally designed inhibitors, high-throughput screening also yielded some useful probes. The OSMI series compounds were proved to be the most potent inhibitors of OGT.[6] The most up-to-date compound OSMI-4a binds OGT extremely tightly with K_d of 8 nM.[7] However, the selectivity of these compounds requires further test.

Covalent inhibitors are a class of probes that form reversible or irreversible covalent linkage with the protein target.[8] One advantage of covalent inhibitors compared to non-covalent inhibitors, is that they can avoid competition with endogenous ligands, such as UDP-GlcNAc which is the second highest concentration metabolite in cell. Other advantages include prolonged pharmacodynamics and increased efficiency and specificity. By now, the only covalent inhibitor for OGT is BZX2, which covalently reacts with cysteine and another nucleophilic residue.[9] BZX2 suffered from the problem of low specificity since it is not targeted to the active site of OGT.

In this project, we aim to adopt the previous design of GlcNAc Electrophilic Probes (GEPs) to design novel targeted covalent inhibitors of OGT.

3.2 Material and methods

3.2.1 Organic synthesis

3.2.1.1 General chemical synthesis

All reagents were purchased from Sigma-Aldrich, MP Biomedicals, Alfa Aesar, TCI, or Thermo Fisher Scientific. Unless otherwise stated, all reactions were performed in flame-dried, round-bottomed flasks fitted with rubber septa under nitrogen atmosphere. Analytical thin layer chromatography (TLC) was

conducted on silica gel plates (0.25 mm, 60 Å pore size) with fluorescent indicator on glass (254 nm) and visualized by ceric ammonium molybdate (CAM), basic $KMnO_4$, or UV light. Unless otherwise noted, all commercially available reagents were used as received. Compound purity was assessed with proton nuclear magnetic resonance (NMR) and mass spectrometry (MS). NMR spectra were obtained on Varian UI400 or UI500 spectrometers. Spectra were recorded at 400 or 500 MHz for ^{1}H NMR, 101 or 126 MHz for ^{13}C NMR, and 162 MHz for ^{31}P NMR. Chemical shifts are expressed in parts per million (ppm, δ scale) and referenced to $CDCl_3$ or D_2O. Data for ^{1}H NMR spectra are reported as follows: chemical shift (ppm), multiplicity (s, singlet; d, doublet; t, triplet; q, quartet; m, multiplet; dd, doublet of doublets; ddd, doublet of doublet of doublets, dt, doublet of triplets; td, triplet of doublets; ddq, doublet of doublet of quartet), coupling constant (Hz), and integration. ^{13}C NMR spectra are expressed as chemical shifts. ^{31}P NMR spectra are represented as follows: chemical shifts, multiplicity, and coupling constant (Hz). High-resolution MS spectra were obtained from Q-TOF Maxis 4G (Bruker). $Ac_4$5SGlcNAc was synthesized as previously described.[4]

3.2.1.2 Synthesis of **ES1-4**

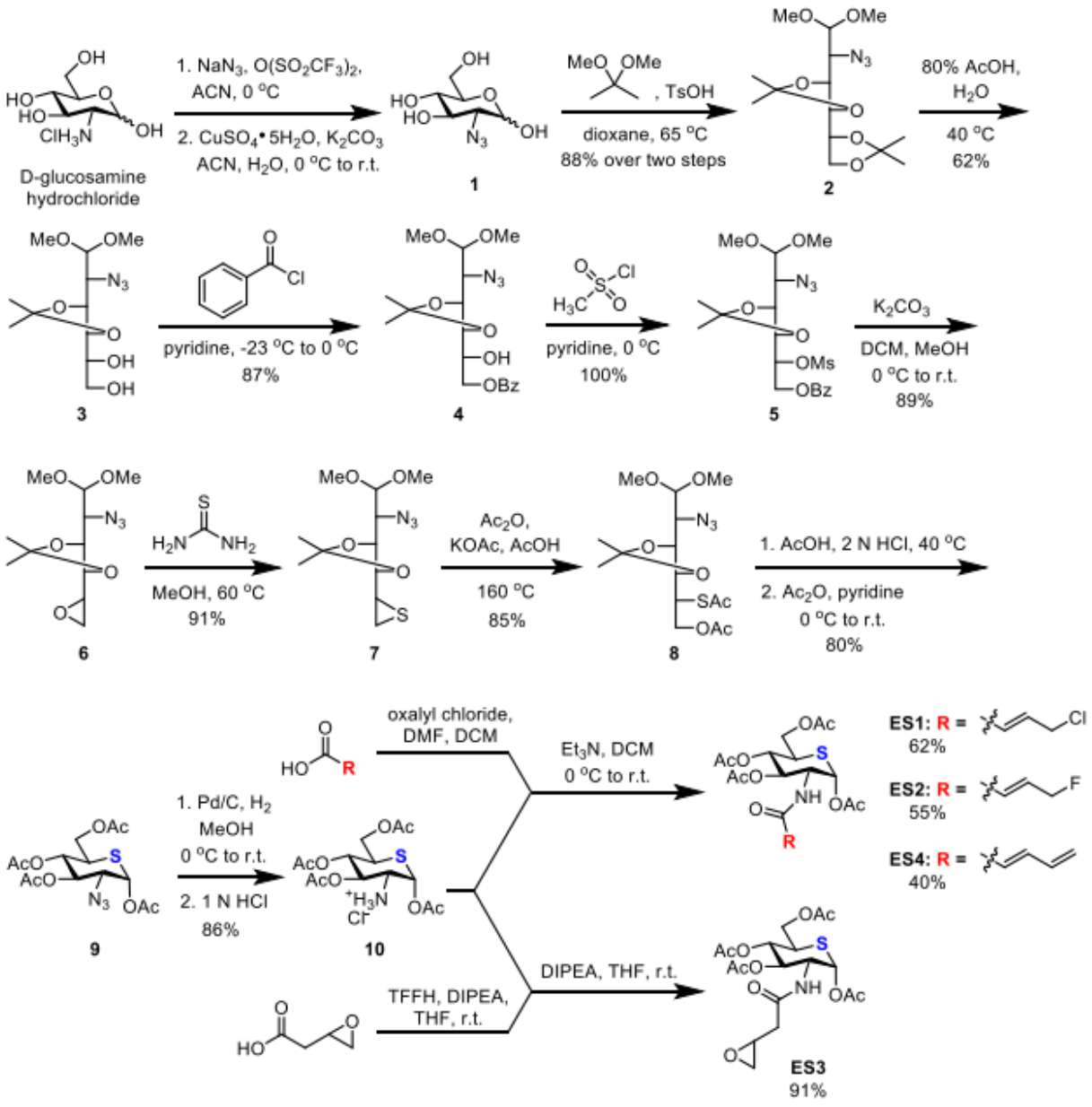


*Figure 3.2 Synthetic scheme for **ES1-ES4***

2-azido-2-deoxy-D-glucose (**1**):[10] NaN_3 (1.45 g, 22.2 mmol) was suspended in dry acetonitrile (26 mL) and cooled to 0 °C. Trifluoromethane sulfonic anhydride (2.85 mL, 16.9 mmol) was slowly added over 20 minutes and then stirred at 0 °C for 5 h. A solution of D-glucosamine-hydrochloride (3.01 g, 13.9 mmol),

$Cu(II)SO_4 \cdot 5H_2O$ (35 mg, 0.139 mmol), K_2CO_3 (3.84 g, 27.8 mmol) and water (15 mL) was prepared. The azide solution was added through a glass wool filter to the glucosamine solution at 0 °C. The mixture was stirred on ice for 1 h, then at room temperature (r.t.) overnight. TLC at 18 h showed product, and at 19 h the solution was concentrated in vacuo. It was separated on a silica gel with DCM:MeOH and eluted with DCM:MeOH 2:1. The product was dried to ~2.8 g white solid (yield determined after second step due to extra mass giving over-quantitative yield). ^{1}H NMR (400 MHz, $CDCl_3$) δ 5.35 (d, J = 3.6 Hz, 0.4H), δ 4.71 (d, J = 8.2 Hz, 0.6H), δ 3.92–3.72 (m, 3H), δ 3.54–3.44 (m, 2.6H), δ 3.28 (dd, J = 8.2, 1.4 Hz, 0.6H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 121.20, 118.68, 95.39, 91.49, 76.31, 74.67, 71.90, 71.78, 70.11, 69.79, 67.21, 63.87, 60.92, 60.77.

2-azido-2-deoxy-3,4,5,6-di-*O*-isopropylidene-aldehydo-D-glucose dimethyl acetal (**2**): Compound **1** (2.8 g, 13.9 mmol), dry 1,4-dioxane (31 mL), 2,2-dimethoxypropane (28 mL, 228 mmol), and tosic acid (600 mg, 3.5 mmol) were combined in a dry flask and stirred at 65 °C and monitored with ^{1}H NMR. At 6 h the flask was cooled to r.t. and the mixture was concentrated *in vacuo*. It was taken up in 60 mL ethyl acetate and washed with 60 mL each sat. $NaHCO_3$ and water and dried over Na_2SO_4. It was separated on silica gel eluting with hexanes:EtOAc 98:2 and dried down to give **2** as a thick amber oil (4.07 g, 88% yield over two steps). ^{1}H NMR (500 MHz, $CDCl_3$) δ 4.60 (d, J = 8.2 1H), 4.15–4.11 (m, 2H), 4.03 (m, 1H), 3.97 (dd, J = 8.5, 4.4, 1H), 3.94 (t, J = 8.4, 1H), 3.51 (m, 1H), 3.50 (s, 3H), 3.46 (s, 3H), 1.44 (s, 3H), 1.40 (s, 3H), 1.36 (s, 3H), 1.30 (s, 3H). ^{13}C NMR (126 MHz, $CDCl_3$) δ 110.07, 109.93, 103.93, 78.80, 77.46, 77.32, 68.06, 61.65, 55.01, 54.85, 27.24, 26.85, 26.75, 25.36.

2-azido-2-deoxy-3,4-*O*-isopropylidene-aldehydo-D-glucose dimethyl acetal (**3**):[4] Compound **2** (4.07 g, 12.3 mmol) and 80% acetic acid in water (43 mL) were combined in a flask and stirred at 40 °C for 7 h. The mixture was concentrated *in vacuo* and separated on silica gel, eluting with hexanes:EtOAc 1:1 to give **3** as a thick oil (2.20 g, 62% yield). ^{1}H NMR (500 MHz, $CDCl_3$) δ 4.62 (d, J = 8.0 Hz, 1H), 4.22 (dd, J = 7.9, 1.9 Hz, 1H), 4.02 (t, J = 7.7 Hz, 1H), 3.83 (dd, J = 11.0, 3.2 Hz, 1H), 3.74 (m, 1H), 3.68 (dd, J =

11.0, 5.5 Hz, 1H), 3.51 (dd, J = 8.0, 1.9 Hz, 1H), 3.48 (s, 3H), 3.46 (s, 3H), 1.45 (s, 3H), 1.37 (s, 3H). ^{13}C NMR (126 MHz, $CDCl_3$) δ 110.09, 103.70, 78.14, 77.17, 73.05, 64.17, 61.69, 55.10, 54.41, 27.34, 26.82.

2-azido-6-*O*-benzoyl-2-deoxy-3,4-*O*-isopropylidene-aldehydo-D-glucose dimethyl acetal (**4**):[4] Compound **3** (3.79 g, 13.0 mmol) was dissolved in dry pyridine (45 mL) and cooled to -23 °C with a dry ice/CCl_4 bath. Benzoyl chloride (1.6 mL, 13.78 mmol) was added, and the solution was stirred at -23 °C for 10 minutes. It was then stirred at -12–0 °C for 5 h. The reaction mixture was concentrated *in vacuo* and purified on silica gel, eluting with hexanes:EtOAc 3:1 to give **4** as a thick oil (4.47 g, 87% yield). ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.05 (dd, J = 8.4, 1.2 Hz, 2H), 7.58 (tt, J = 7.4, 1.2 Hz, 1H), 7.45 (t, J = 7.9 Hz, 2H), 4.64 (d, J = 8.0 Hz, 1H), 4.62 (dd, J = 11.9, 2.8 Hz, 1H), 4.42 (dd, J = 11.9, 6.6 Hz, 1H), 4.28 (dd, J = 7.7, 1.8 Hz, 1H), 4.08 (t, J = 7.8 Hz, 1H), 4.00 (ddd, J = 8.3, 6.5, 2.8 Hz, 1H), 3.53 (dd, J = 8.0, 1.8 Hz, 1H), 3.46 (s, 3H), 3.44 (s, 3H), 2.90 (s, 1H,) 1.47 (s, 3H), 1.39 (s, 3H). ^{13}C NMR (126 MHz, $CDCl_3$) δ 167.47, 133.55, 129.92, 129.67, 128.65, 110.44, 103.51, 78.66, 76.30, 72.73, 67.45, 61.74, 54.84, 54.28, 27.34, 26.86.

2-azido-6-*O*-benzoyl-2-deoxy-3,4-*O*-isopropylidene-5-*O*-mesyl-aldehydo-D-glucose dimethyl acetal (**5**):[4] Compound **4** (256 mg, 0.65 mmol) was dissolved in pyridine (3 mL) and cooled to 0 °C. Mesyl chloride (360 μL, 4.55 mmol) was added dropwise to the solution and the reaction mixture was stirred at 0 °C for 3 h until TLC showed complete conversion. The reaction mixture was concentrated under vacuum, the residue was dissolved in chloroform, washed with 2 N HCl twice and then sat. $NaHCO_3$. The water phases were back extracted with chloroform. The combined organic phase was concentrated to give **5** as a colorless oil (306 mg, 100% yield). ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.07 (m, 2H), 7.59 (t, J = 7.4 Hz, 1H), 7.46 (t, J = 7.7 Hz, 2H), 5.07 (td, J = 6.7, 2.8 Hz, 1H), 4.78 (dd, J = 12.6, 2.9 Hz, 1H), 4.62 (d, J = 7.6 Hz, 1H), 4.47 (dd, J = 12.6, 6.8 Hz, 1H), 4.37–4.32 (m, 2H), 3.48 (s, 3H), 3.47 (s, 3H), 3.42 (dd, J = 7.6, 2.0 Hz, 1H), 3.08 (s, 3H), 1.48 (s, 3H), 1.43 (s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 166.02, 133.48,

129.81, 129.32, 128.60, 111.03, 103.55, 79.13, 77.60, 75.14, 63.51, 61.55, 55.38, 54.24, 39.05, 37.59, 27.06, 26.71, 15.09.

5,6-anhydro-2-azido-2-deoxy-3,4-*O*-isopropylidene-aldehydo-D-glucose dimethyl acetal (**6**):[4] Compound **5** (264 mg, 0.56 mmol) was dissolved in a mixture of DCM (4 mL) and MeOH (4 mL) and cooled to 0 °C. Potassium carbonate (360 mg, 2.80 mmol) was added to the solution as a solid and the reaction was warmed to r.t. and stirred for 5 h until TLC showed complete conversion. Diethyl ether was dumped into the solution and the mixture was filtered. The filtrate was washed with brine and the aqueous phase was extracted with DCM for three times. The combined organic phase was dried over Na_2SO_4 and concentrated to get the crude as a colorless oil. Silica gel chromatography was used to purify the product with hexanes:EtOAc 4:1 and dried down to give **6** as a colorless oil (136 mg, 89% yield). ^{1}H NMR (500 MHz, $CDCl_3$) δ 4.59 (d, J = 7.4 Hz, 1H), 4.19 (dd, J = 8.3, 2.6 Hz, 1H), 3.97 (dd, J = 8.3, 5.0 Hz, 1H), 3.49 (s, 3H), 3.48 (s, 3H), 3.21 (dd, J = 7.4, 2.6 Hz, 1H), 3.03 (ddd, J = 5.0, 4.1, 2.6 Hz, 1H), 2.80 (dd, J = 5.2, 4.1 Hz, 1H), 2.71 (dd, J = 5.2, 2.6 Hz, 1H), 1.44 (s, 3H), 1.39 (s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 110.58, 104.07, 77.55, 76.52, 61.52, 56.24, 54.68, 51.02, 44.05, 26.94, 26.84.

2-azido-2,5,6-trideoxy-5,6-epithio-3,4-*O*-isopropylidene-aldehydo-D-glucose dimethyl acetal (**7**):[4] Compound **6** (2.14 g, 7.83 mmol) was dissolved in MeOH (40 mL) and thiourea (1.80 mg, 23.6 mmol) was added into the solution. The reaction mixture was heated to 60 °C and stirred for 3 h until TLC showed complete conversion. The reaction mixture was concentrated to get the crude. Silica gel chromatography was used to purify the product with hexanes:EtOAc 8:1 and dried down to give **7** as a colorless oil (2.05 g, 91% yield). ^{1}H NMR (400 MHz, $CDCl_3$) δ 4.61 (d, J = 8.0 Hz, 1H), 4.23 (dd, J = 8.0, 1.7 Hz, 1H), 3.50 (s, 3H), 3.48 (s, 3H), 3.46–3.44 (m, 1H), 3.38 (dd, J = 8.0, 1.7 Hz, 1H), 2.89 (ddd, J = 8.7, 6.0, 5.1 Hz, 1H), 2.58 (dd, J = 6.0, 1.5 Hz, 1H), 2.31 (dd, J = 5.1, 1.5 Hz, 1H), 1.46 (s, 3H), 1.45 (s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 110.49, 103.77, 82.25, 79.39, 61.54, 55.37, 54.53, 33.77, 27.25, 26.83, 23.77.

6-*O*-acetyl-5-*S*-acetyl-2-azido-2-deoxy-3,4-*O*-isopropylidene-5-thio-aldehydo-D-glucose dimethyl acetal (**8**):[4] Compound **7** (120 mg, 0.41 mmol) was dissolved in a mixture of acetic anhydride (3.8 mL) and acetic acid (0.6 mL), and potassium acetate (186 mg, 2.07 mmol) was added to the solution. The reaction mixture was heated to 160 °C and stirred for 10 h until TLC showed complete conversion. The reaction was concentrated under vacuum and dissolved in chloroform, then washed with 2 N HCl, sat. $NaHCO_3$. The aqueous phases were back extracted with chloroform. The combined organic phase was dried over Na_2SO_4 and concentrated to get the crude as a colorless oil. Silica gel chromatography was used to purify the product with hexanes:EtOAc 8:1 and dried down to give **8** as a colorless oil (138 mg, 85% yield). ^{1}H NMR (400 MHz, $CDCl_3$) δ 4.57 (d, J = 8.0 Hz, 1H), 4.38 (dd, J = 11.5, 4.3 Hz, 1H), 4.32 (dd, J = 11.5, 5.0 Hz, 1H), 4.19 (dd, J = 8.5, 7.4 Hz, 1H), 4.09 (dd, J = 7.4, 1.8 Hz, 1H), 3.92 (ddd, J = 8.4, 4.9, 4.3 Hz, 1H), 3.49 (s, 3H), 3.44 (s, 3H), 3.36 (dd, J = 8.0, 1.8 Hz, 1H), 2.36 (s, 3H), 2.07 (s, 3H), 1.45 (s, 3H), 1.42 (s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 193.72, 170.72, 110.59, 101.96, 78.31, 76.45, 64.01, 62.57, 55.83, 55.62, 44.96, 30.80, 27.27, 26.91, 20.93.

1,3,4,6-tetra-*O*-acetyl-2-azido-2-deoxy-5-thio-α-D-glucose (**9**):[4] Compound **8** (100 mg, 0.25 mmol) was dissolved in a mixture of acetic acid (2.1 mL) and 2 N HCl aqueous solution (0.9 mL) and the reaction mixture was heated to 40 °C and stirred for 18 h until TLC showed complete conversion. The reaction was concentrated under vacuum and the residue was dissolved in pyridine. Acetic anhydride was added dropwise at 0 °C and the reaction mixture was warmed to r.t. overnight. The reaction was concentrated under vacuum to get the crude as a brown oil. The product was purified by silica gel chromatography and eluted by hexanes:EtOAc 4:1, and further dried down to give **9** as a colorless oil (92 mg, 92% yield). ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.11 (d, J = 3.1 Hz, 1H), 5.40 (dd, J = 10.2, 0.4 Hz, 1H), 5.32 (dd, J = 10.7, 9.5 Hz, 1H), 4.40 (dd, J = 12.2, 4.9 Hz, 1H), 4.04 (dd, J = 12.2, 3.1 Hz, 1H), 3.89 (dd, J = 10.2, 3.1 Hz, 1H), 3.56 (ddd, J = 10.7, 4.9, 3.0 Hz, 1H), 2.19 (s, 3H), 2.11 (s, 3H), 2.07 (s, 4H), 2.06 (s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 170.47, 169.73, 169.54, 168.69, 149.15, 136.56, 71.72, 64.90, 60.92, 39.83, 20.98, 20.60, 20.59, 20.50.

1,3,4,6-tetra-*O*-acetyl-2-amino-2-deoxy-5-thio-α-D-glucose hydrochloride (**10**): Compound **9** (95 mg, 0.24 mmol) was dissolved in MeOH (2.5 mL) and Pd/C (10% loading, 30 mg) was added. HCl aqueous solution (12 N, 27 μL) was added at 0 °C and the reaction was stirred under 1 atm of H_2 at r.t. for 1 h until TLC showed complete conversion. The reaction mixture was filtered through a pad of Celite® and the filtrate was concentrated to give **10** as a white solid (84 mg, 86% yield).

1,3,4,6-tetra-*O*-acetyl-2-[[(2E)-1-oxo-4-chloro-2-buten-1-yl]amino]-2-deoxy-5-thio-α-D-glucose (**ES1**): (*E*)-4-chloro-2-butenoic acid (synthesized as previously described[11])(24 mg, 0.20 mmol), dry DCM (1 mL), and one drop of dry DMF were combined and cooled to 0 °C. Oxalyl chloride (28 μL, 0.32 mmol) was added dropwise. The solution was stirred at 0 °C for 20 minutes and then at r.t. for 2.5 h. The reaction mixture was concentrated in vacuo and the residue was cooled to 0 °C. A solution of 10 (16 mg, 0.040 mmol), Et_3N (6 μL, 0.040 mmol), and dry DCM (1.0 mL) was prepared, and slowly added to the residue. The mixture was stirred for 10 minutes at 0 °C and then 2 h at r.t. The solution was concentrated in vacuo, taken up in EtOAc (8 mL), and washed with saturated $NaHCO_3$ (8 mL) and water (8 mL). The organic layer was dried over Na_2SO_4 and concentrated in vacuo. The crude product was purified by column chromatography on silica gel (hexanes/EtOAc, 2:1), and dried down to give **ES1** as a white solid (11.6 mg, 62% yield). ^{1}H NMR (400 MHz, CDCl3) δ 6.90 (dt, J = 14.9, 5.7 Hz, 1H), 5.98 (m, 2H), 5.87 (d, J = 8.8 Hz, 1H), 5.40 (dd, J = 10.8, 9.6 Hz, 1H), 5.22 (dd, J = 10.9, 9.6 Hz, 1H), 4.71 (ddd, J = 11.1, 8.8, 3.1 Hz, 1H), 4.36 (dd, J = 12.1, 5.0 Hz, 1H), 4.15 (dd, J = 5.6 Hz, 1.5 Hz, 2H), 4.05 (dd, J = 12.1, 3.1 Hz, 1H), 3.49 (ddd, J = 10.8, 4.8, 3.3 Hz, 1H), 2.19 (s, 3H), 2.07 (s, 3H), 2.05 (s, 3H), 2.02 (s, 3H). ^{13}C NMR (101 MHz, CDCl3) δ 171.99, 170.71, 169.28, 168.80, 164.17, 139.85, 125.04, 72.79, 71.90, 71.57, 61.23, 55.51, 42.78, 39.91, 21.26, 20.80, 20.78, 20.66. Theoretical m/z calculated for $C_{18}H_{24}ClNNaO_9S$ $[M+Na]^+$: 488.0753. HRMS found: 488.0754. 95% purity by UHPLC.

1,3,4,6-tetra-*O*-acetyl-2-[[(2*E*)-1-oxo-4-fluoro-2-buten-1-yl]amino]-2-deoxy-5-thio-D-glucose (**ES2**): **ES2** was synthesized according to the procedure for **ES1**, using (*E*)-4-fluoro-2-butenoic acid in the first

step (6.3 mg, 55% yield). ^{1}H NMR (400 MHz, CDCl3) δ 6.89 (ddt, J =24.5, 15.4, 3.4 Hz, 1H), 5.98 (m, 2H), 5.86 (d, J = 8.7 Hz, 1H), 5.41 (t, J = 10.2 Hz, 1H), 5.23 (t, J = 10.2 Hz, 1H), 5.05 (ddd, J = 46.2, 3.4, 2.2 Hz, 2H), 4.72 (ddd, J = 11.2, 8.9, 3.1 Hz, 1H), 4.37 (dd, J = 12.1, 4.9 Hz, 1H), 4.06 (dd, J = 12.1, 3.0 Hz, 1H), 3.51 (ddd, J = 10.6, 4.6, 3.5 Hz, 1H), 2.19 (s, 3H), 2.08 (s, 3H), 2.05 (s, 3H), 2.03 (s, 3H). ^{13}C NMR (101 MHz, CDCl3) δ 171.93, 170.72, 169.30, 168.83, 164.25, 139.78 (d, J = 15.3 Hz), 121.97 (d, J = 10.1 Hz), 81.34 (d, J = 171.9 Hz), 72.78, 71.87, 71.58, 61.25, 55.51, 39.92, 21.24, 20.79, 20.78, 20.66. ^{19}F NMR (376 MHz) δ - 125.65. Theoretical m/z calculated for $C_{18}H_{24}FNNaO_9S$ $[M+Na]^+$: 472.1048. HRMS found: 472.1049. 95% purity by UHPLC.

1,3,4,6-tetra-*O*-acetyl-2-[[1-oxo-3,4-epoxybutane-1-yl]amino]-2-deoxy-5-thio-D-glucose (**ES3**): Oxiraneacetic acid (22 mg, 0.20 mmol), dry THF (1 mL), N,N,N′,N′- tetramethylfluoroformamidinium hexafluorophosphate (TFFH, 66 mg, 0.25 mmol), and DIPEA (43.5 μL, 0.25 mmol) were combined and stirred at r.t. for 2.5 h, and then compound 10 (10 mg, 0.025 mmol) was added. After another 8 h stirring at r.t. the solution was partially concentrated in vacuo, leaving a small amount of liquid in the flask, taken up with EtOAc (6 mL), and washed with water (6 mL). The organic layer was dried over Na_2SO_4 and concentrated. The crude product was purified by column chromatography on silica gel (hexanes/EtOAc, 1:2) (10.2 mg, 91% yield). ^{1}H NMR (400 MHz, CDCl3) δ 6.18 (d, J = 9.1 Hz, 0.5H), 6.11 (d, J = 8.8 Hz, 0.5H), 6.03 (d, J = 3.1 Hz, 0.5H), 5.96 (d, J = 3.1 Hz, 0.5H), 5.38 (t, J = 10.4 Hz, 1H), 5.23 (m, 1H), 4.64 (m, 1H), 4.37 (dd, J = 11.8, 5.2 Hz, 1H), 4.05 (dd, J = 12.1, 3.1 Hz, 1H), 3.51 (m, 1H), 3.13 (m, 1H), 2.84 (m, 1H), 2.65 (m, 1H), 2.55 (dd, J = 4.7, 2.6 Hz, 0.5H), 2.51 (dd, J = 4.6, 2.7 Hz, 0.5H), 2.20 (s, 3H), 2.08 (s, 3H), 2.05 (s, 3H), 2.05 (s, 3H). ^{13}C NMR (126 MHz, CD3OD) δ 173.03, 172.66, 172.39, 172.15, 171.83, 171.76, 171.26, 170.80, 73.61, 73.58, 73.52 (2C), 72.81, 72.71, 62.38, 56.14, 56.07, 49.80, 49.74, 47.31 (2C), 40.87, 40.82, 40.03, 39.92, 20.60, 20.58, 20.53 (2C), 20.49 (2C). Theoretical m/z calculated for $C_{18}H_{25}NNaO_{10}S$ $[M+Na]^+$: 470.1091. HRMS found: 470.1081.

1,3,4,6-tetra-*O*-acetyl-2-[[(2E)-1-oxo-2,4-pentadiene-1-yl]amino]-2-deoxy-5-thio-D-glucose (**ES4**): Compound **ES4** was synthesized according to the procedure for **ES1** using 2,4-pentadienoic acid in place of (*E*)-4-chloro-2-butenoic acid (9.3 mg, 40% yield). ^{1}H NMR (400 MHz, CDCl3) δ 7.18 (dd, J = 15.0, 11.0 Hz, 1H), 6.39 (dt, J = 16.9, 10.5 Hz, 1H), 5.97 (d, J = 2.9 Hz, 1H), 5.76 (m, 2H), 5.59 (d, J = 16.9 Hz, 1H), 5.47 (d, J = 10.2 Hz, 1H), 5.41 (t, J = 10.2 Hz, 1H), 5.22 (t, J = 10.2 Hz, 1H), 4.73 (ddd, J = 11.0, 8.9, 3.0 Hz, 1H), 4.36 (dd, J = 12.1, 4.9 Hz, 1H), 4.05 (dd, J = 12.1, 3.0 Hz, 1H), 3.49 (m, 1H), 2.18 (s, 3H), 2.07 (s, 3H), 2.04 (s, 3H), 2.01 (s, 3H). ^{13}C NMR (101 MHz, CDCl3) δ 172.02, 170.73, 169.28, 168.79, 165.30, 142.83, 134.56, 125.65, 123.30, 72.98, 71.97, 71.60, 61.27, 55.43, 39.93, 21.26, 20.81, 20.68. Theoretical m/z calculated for $C_{19}H_{25}NNaO_9S$ $[M+Na]^+$: 466.1142. HRMS found: 466.1152. 95% purity by UHPLC.

3.2.1.3 Synthesis of **UDP-ES1**

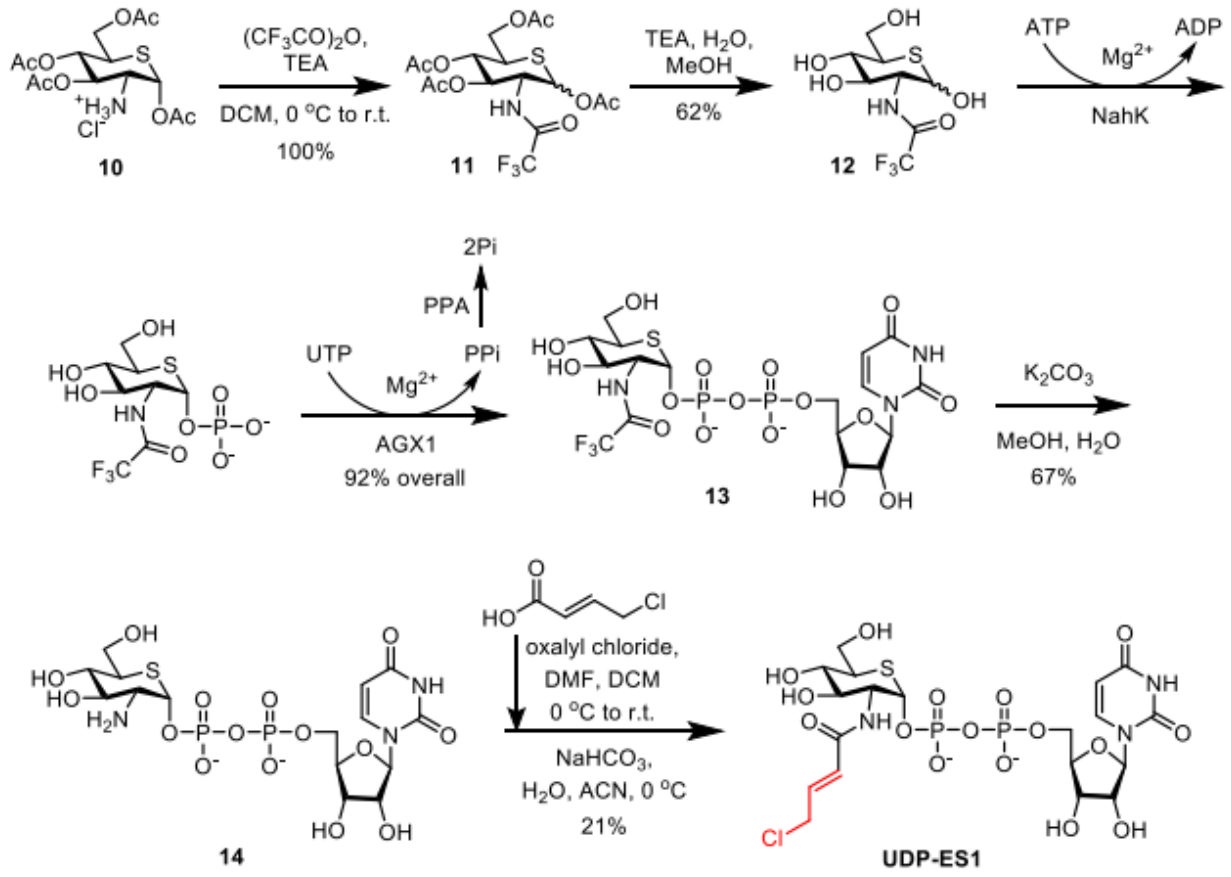


*Figure 3.3 Synthetic scheme for **UDP-ES1***

1,3,4,6-tetra-*O*-acetyl-2-deoxy-2-trifluoroacetamido-5-thio-α-D-glucose (**11**): Compound **10** (78 mg, 0.20 mmol) was dissolved in DCM (2 mL) and triethylamine (60 µL, 0.43 mmol) was added. Trifluoroacetic anhydride was added dropwise at 0 °C and the reaction mixture was warmed to r.t. overnight. The mixture was diluted with DCM and washed with sat. $NaHCO_3$. The aqueous phase was extracted with DCM twice and the combined organic phase was dried over Na_2SO_4 and concentrated to give **11** as a pale yellow oil (89 mg, 100% yield). ^{1}H NMR (400 MHz, $CDCl_3$) δ 6.66 (d, J = 8.7 Hz, 1H), 6.04 (d, J = 3.1 Hz, 1H), 5.45 (dd, J = 10.8, 9.6 Hz, 1H), 5.28 (dd, J = 10.8, 9.5 Hz, 1H), 4.66 (ddd, J = 11.3, 8.6, 3.1 Hz, 1H), 4.41 (dd, J = 12.2, 4,9 Hz, 1H), 4.09 (dd, J = 12.2, 3.1 Hz, 1H), 3.55 (ddd, J = 10.8,

4.8, 3.0 Hz, 1H), 2.24 (s, 3H), 2.11 (s, 3H), 2.09 (s, 3H), 2.08 (s, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 171.76, 170.50, 169.09, 168.52, 71.84, 71.30, 71.00, 60.86, 55.86, 39.87, 20.96, 20.62, 20.48, 20.38.

2-deoxy-2-trifluoroacetamido-5-thio-D-glucose (**12**): Compound **11** (113 mg, 0.21 mmol) was dissolved in a mixture of MeOH:H_2O:TEA (5:2:1, 2 mL). The reaction mixture was stirred at r.t. for 2 h and was then concentrated *in vacuo*. The product was purified by silica gel chromatography and eluted by DCM:MeOH 9:1, and further dried down to give **12** as a white solid (38 mg, 62% yield). ^{1}H NMR (400 MHz, D_2O) δ 7.94 (d, J = 8.1 Hz, 1H), 6.00–5.98 (dd, J = 3.3, 1.3 Hz, 1H), 5.96 (d, J = 8.1 Hz, 1H), 5.43 (dd, J = 7.5, 2.7 Hz, 1H), 4.37 (dd, J = 3.3, 1.0 Hz, 2H), 4.28 (d, J = 5.0 Hz, 2H), 4.23 (ddt, J = 11.6, 6.0, 2.7 Hz, 1H), 3.97 (dd, J = 12.1, 5.3 Hz, 1H), 3.87 (dd, J = 12.0, 3.1 Hz, 1H), 3.70–3.61 (m, 2H), 3.38 (d, J = 4.3 Hz, 1H), 3.23 (dd, J = 6.9, 2.9 Hz, 1H). ^{13}C NMR (101 MHz, D_2O) δ 164.45, 158.27, 141.49, 102.77, 88.59, 83.08, 77.02, 73.81, 73.40, 69.68, 64.99, 59.94, 59.16, 43.82, 23.33.

Chemoenzymatic synthesis of UDP-5SGlcNTFA (**13**): The reaction was set up with 100 mM Tris-HCl (pH 8.0), 10 mM $MgCl_2$, 12 mM ATP, 12 mM UTP, 10 mM **12**, 0.5 mg/mL NahK, 0.75 mg/mL AGX1, and 1 U/mL PPA (pyrophosphatase from baker's yeast, 50 U/mL stock, Sigma #I1643-100UN) in a total volume of 0.5 mL. The mixture was incubated at 37 °C for 36 h and the reaction was concentrated under vacuum to get the crude as a yellow solid. The crude product was purified by silica gel chromatography, eluting by EtOAc:MeOH:H_2O 4:2:1 and further dried down to give **13** as a white solid (3.3 mg, 92% yield). ^{1}H NMR (400 MHz, D_2O) δ 7.95 (d, J = 8.1 Hz, 1H), 6.00–5.96 (m, 2H), 5.43 (dd, J = 7.4, 2.2 Hz, 1H), 4.37 (t, J = 3.9 Hz, 2H), 4.34–4.26 (m, 2H), 4.22 (d, J = 13.6 Hz, 2H), 3.99 (dd, J = 12.0, 5.3 Hz, 1H), 3.93–3.84 (m, 2H), 3.76 (d, J = 4.7 Hz, 1H), 3.45-3.37 (m, 1H). Theoretical m/z calculated for $C_{17}H_{22}F_3N_3NaO_{16}P_2S$ $[M-2H+Na]^-$: 698.0051. HRMS found: 698.0046.

Uridine 5'-diphospho-2-amino-2-deoxy-5-thio-α-D-glucose (**14**): **13** (5.6 mg, 8.3 μmol) was dissolved in a mixture of MeOH:H_2O (2:1, 0.9 mL) and K_2CO_3 (50 mg, 0.36 mmol) was added. The reaction

mixture was stirred at r.t. overnight. The mixture was concentrated *in vacuo* and purified on silica gel, eluting with EtOAc:MeOH:H_2O 1:1:1. The product was dried down to give **14** as a powder (3.0 mg, 67% yield). ^{1}H NMR (400 MHz, D_2O) δ 7.94 (d, J = 8.1 Hz, 1H), 6.04–5.92 (m, 2H), 5.43 (dd, J = 7.5, 2.7 Hz, 1H), 4.37 (dd, J = 3.3, 1.0 Hz, 2H), 4.28 (m, 1H), 4.24 (m, 1H), 3.97 (dd, J = 12.1, 5.3 Hz, 1H), 3.87 (dd, J = 12.0, 3.1 Hz, 1H), 3.70–3.61 (m, 2H), 3.41–3.34 (m, 2H), 3.23 (dt, J = 9.9, 3.0 Hz, 1H).

Uridine 5′-diphospho-2-deoxy-2-[[(2*E*)-1-oxo-4-chloro-2-buten-1-yl]amino]-5-thio-α-D-glucose (**UDP-ES1**): Compound **15** (22.2 mg, 0.19 mmol), dry DCM (1 mL) and dry DMF (two drops) were combined and cooled to 0 ºC. Oxalyl chloride (26 µL, 0.30 mmol) was added to the solution, which was stirred at 0 ºC for 20 min, then at r.t. for 3 h. A solution of **14** (7 mg, 0.012 mmol) and $NaHCO_3$ (125 mg, 1.5 mmol) in ACN:H_2O 1:1 (1 mL) was prepared and cooled to 0 ºC. The reaction mixture of **15** was briefly concentrated *in vacuo* and the yellow residue was taken up in dry ACN (0.3 mL). This was dropped into the stirred **14** mixture, and the reaction was stirred at 0 ºC for 4 h. The mixture was then concentrated *in vacuo*, taken up in 5 mL water, and washed with 5 mL DCM. The aqueous layer was separated on silica gel eluting the desired product with EtOAc:MeOH:H_2O 7:2:1. The fraction was dried down to give 5S**GEP1** as a white solid (1.7 mg, 21% yield). ^{1}H NMR (400 MHz, D_2O) δ 7.95 (d, J = 8.1 Hz, 1H), 6.86 (dt, J = 15.4, 6.1 Hz, 1H), 6.39 (d, J = 15.3 Hz, 1H), 5.97 (m, 2H), 5.36 (d, J = 7.3 Hz, 1H), 4.39–4.17 (m, 10H), 4.02–3.89 (m, 2H), 3.77 (m, 1H). ^{31}P NMR (162 MHz, D_2O) δ -11.24 (d, J = 22.2 Hz), -12.97 (d, J = 22.2 Hz). Theoretical m/z calculated for $C_{19}H_{26}ClN_3O_{16}P_2S$ $[M-2H]^{2-}$: 340.5104. HRMS found: 340.5074.

3.2.2 Biochemical and mass spectrometric experiments

Cloning, expression, and purification of one-pot enzymes Nahk_ATCC15697 and AGX1. Nahk_ATCC15697 was amplified from genomic DNA of *Bifidobacterium longum* subsp. *infantis* ATCC#15697 (ref. [12]) using the primers listed in Table S1 and cloned into pET28a(+) vector to generate an N-terminal His_6-tagged fusion protein. AGX1 (ref. [13]) was similarly cloned from human AGX1 gene

(Harvard PlasmID #Hscd00329319) using the primers in Table S1 into pET28a(+) vector to generate an N-terminal His_6-tagged fusion protein. *E. coli* BL21(DE3) transformant containing the plasmid of pET28a-Nahk_ATCC15697 or pET28a-AGX1 was grown in LB medium with Kanamycin at 37 ^oC until OD_{600} reached 0.8. Isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 0.1 mM to induce protein expression at 16 ^oC overnight. The cells were harvested by centrifugation and the cell pellet was resuspended in TBS buffer (20 mM Tris pH 8.0, 150 mM NaCl) and lysed by ultra-high-pressure cell disrupter (Emulsiflex-C5). The lysate was clarified by centrifugation and the supernatant was purified on Ni-NTA column. The desired protein (eluted in TBS buffer with 250 mM imidazole) was further purified by size-exclusion chromatography (Superdex 200 increase 10/300, GE Healthcare) with buffer (TBS containing 0.5 mM tris(hydroxypropyl)phosphine (THP)). The fractions containing the purified enzyme were concentrated, and the final product was stored at -80 ^oC in small aliquots. The purity of proteins was examined by SDS-PAGE gel.

NUP62 protein expression and purification. Human NUP62 expression plasmid in pET21a vector (a kind gift from Dr. Suzanne Walker's lab) was transformed into *E. coli* BL21(DE3) competent cells and expressed as a fusion protein containing a C-terminal His_6-tag. The transformant was grown at 37 °C in LB medium. After OD_{600} reached 0.4, the culture was then induced with 0.2 mM IPTG for 3 h. Cells were pelleted and resuspended in lysis buffer (50 mM Tris pH 8.0, 10 mM EDTA, 0.5 M NaCl, and 1 mM PMSF). The cell suspension was lysed with ultra-high-pressure cell disrupter followed by centrifugation at 20,500 g at 4 °C for 30 min. The collected pellets containing the inclusion body of NUP62 were then washed twice with 30 mL detergent solution (1.1 M urea, 2% Triton X-100) and resuspended in 30 mL denaturing buffer (8 M urea, 50 mM Tris pH 8.0, 1 mM EDTA, 2 mM DTT) and incubated at r.t. to dissolve the inclusion body. The lysate was then centrifuged at 16,100 g at 20 °C for 20 min to remove the unbroken pellets. The supernatant was centrifuged at 50,000 g at 20 °C for 15 min to further clear up the lysate. Proteins were dialyzed against 2 M urea in 20 mM Tris pH 7.5 for 4 h followed by 20 mM Tris pH 7.5 overnight at 4 °C. Following dialysis, samples were centrifuged at 16,100 g at 4 °C for 20 min.

The supernatant containing NUP62 was concentrated and stored at -80 °C until use. The purity of NUP62 protein was examined by SDS-PAGE gel.

OGT purification. The full-length ncOGT and $OGT_{4.5}$ (ref. [14]) expression plasmids in pET24b vectors were kind gifts from Dr. Suzanne Walker's lab. Each plasmid was transformed into *E. coli* BL21(DE3) competent cells, and the transformant was grown at 37 °C in LB medium. After OD_{600} reached 1.0, the culture was induced by 0.2 mM IPTG at 16 °C for overnight. The cells were pelleted, re-suspended in TBS buffer supplemented with 1 mM PMSF, and lysed with ultra-high-pressure cell disrupter. After centrifugation, the supernatant was subjected to Ni-NTA column for affinity purification. The recombinant protein was subsequently eluted with TBS buffer containing 250 mM imidazole and 0.5 mM THP. The purified proteins were concentrated and stored in small aliquots at -80 °C.

Mutagenesis. $OGT_{4.5}$–C917S mutant was generated using the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent) according to the manufacturer's instructions. $OGT_{4.5}$ was used as the DNA template along with the primers listed in Table S1. The DNA sequence was verified by sequencing. The mutant protein was expressed and purified similarly as $OGT_{4.5}$ as mentioned above.

Radiolabeled assay to evaluate the time-dependent inhibition of OGT by UDP-ES1. Purified OGT (1 μM) was pre-incubated with indicated concentration of **UDP-ES1** at r.t. for 2.5 or 5 minutes. The mixture was then diluted 50-fold with reaction buffer (20 mM Tris pH 8.0, 150 mM NaCl, 0.5 mM THP) containing OGT substrates NUP62 (10 μM) and UDP-^{3}H-GlcNAc (100 μM, specific activity 0.092 Ci/mmol). The reactions were incubated at 37 °C for 0.5 h and then quenched by transferring the samples onto the nitrocellulose membrane and air-dried. The membranes were washed in PBS buffer for four times with 5 min each. The radioactivity on each membrane was counted by Tri-Carb 2900TR liquid scintillation analyzer (PerkinElmer). A reaction in the absence of OGT was set up as negative control, while another reaction without 5S**GEP1** served as no-inhibition control. Three independent experiments

were conducted for each condition. Data were analyzed using GraphPad Prism v5 (GraphPad Software, Inc.).

Intact protein mass spectrometry. To detect the covalent modification of **UDP-ES1** on OGT, purified OGT (10 μM wild-type $OGT_{4.5}$ or C917S mutant) was incubated in a 25 μL reaction containing 70 μM of 5S**GEP1** in the reaction buffer (10 mM Tris pH 8.0, 75 mM NaCl, 0.5 mM THP, 200 U/mL CIP alkaline phosphatase, and 30 mM $MgCl_2$) at r.t. for 2 h followed by C8 StageTip desalting.[15] Samples were SpeedVac dried and re-dissolved in 20 μL of 0.1% formic acid. MS analysis of intact protein was performed on ESI-Q-TOF Maxis II (Bruker) coupled with an ACQUITY UPLC (Waters). Samples were loaded onto a 500 μm × 200 mm PLRP column (PLRP-S, 10 μm, 1,000 Å, Varian). The injection volume was 7 μL with a flow rate of 300 nL/min. The mobile phases consisted of 0.1% TFA (solvent A) and 0.1% TFA in acetonitrile (solvent B). LC program: 5% B for 5 min, 5-40% B for 10 min, 40-70% B for 20 min, 70-95% B for 10 min, 95% B for 5 min, and 95-5% B for 1 min. MS analysis was operated in positive mode. End plate offset and capillary voltages were set at 400 and 4,000 V, respectively. The nebulizer was set at 0.5 bar. The interface heater temperature was set at 220 °C with the dry gas flow rate 4.0 L/min. Data were acquired using one full MS scan (m/z 500-3,000) with the scan rate at 1 Hz. LC-MS data were processed and analyzed using Compass Data Analysis software (version 4.3, Bruker). Deconvolution of the mass spectra was performed by maximum entropy algorithm and the parameters were as follows: mass range 50,000-100,000 Da, auto data point spacing, and the resolving power of 8,000. The increased mass resulting from 5S**GEP1** modification was determined by subtracting the original protein mass from the deconvoluted protein mass.

Cell culture and inhibitor treatment. MCF7, COS-7, and HEK293 cells were cultured in DMEM media (Corning) supplemented with 10% fetal bovine serum (FBS) (Sigma) and 0.8% penicillin-streptomycin. Cells were maintained in 5% CO2 at 37 °C and were seeded 20-24 h prior to inhibitor treatment at 50-65% confluence. Upon treatment, media was replaced with low glucose DMEM medium (1 g/L glucose,

5% FBS) containing Ac45SGlcNAc, **ES1**-**ES4**, BZX2, or DMSO at given concentration (final 1% DMSO). Cells were incubated for 13 h unless otherwise indicated. For post-incubation, medium was replaced with the inhibitor-free low glucose DMEM medium after inhibitor treatment and cells were incubated for another 8 h. For glucosamine treatment, 10 or 15 mM glucosamine was added to the COS-7 or MCF7 cells after 3 h of post-incubation, respectively. Cells were harvested by scrapping, washed by PBS, pelleted at 150 g (1,274 r.p.m., F241.5P rotor, Beckman Coulter) at 4 °C for 5 minutes, and stored at -80 °C until use.

Cell viability assay. MCF7 cells were seeded at 2,200 cells per well in a 96-well plate 23 h prior to inhibitor treatment. The cells reached 60% confluence upon treatment. Ac45SGlcNAc, **ES1**, or DMSO was added at varied concentrations in low glucose DMEM medium (1 g/L glucose, 5 % FBS) and pre-treated for 13 h. The medium was then replaced with inhibitor-free low glucose medium for 8 h post-incubation. The cell viability was measured by adding 50 μg/mL of resazurin[16] to each well and incubated for 12 h at 37 °C. The optical density was detected using an excitation wavelength of 560 nm and an emission wavelength of 590 nm in a microplate reader (Synergy H1 Hybrid, Biotek). Results from five different wells of each condition were used for quantitation. Data were analyzed using GraphPad Prism v5 and shown as mean values with error bars representing ± S.D. Statistical significance was determined using Student's *t*-test.

3.3 Results and discussion

3.3.1 Design and synthesis of covalent inhibitors **ES1-ES4**

Figure 3.4 Structures of the designed inhibitors

Previously I discussed the development of GlcNAc Electrophilic Probes[11] which covalently target C917 in the active site of OGT. We aim to adopt the same idea for the design of targeted covalent inhibitors. However, **GEP1** could not be applied as an inhibitor since the sugar transfer rate is faster compared to the covalent reaction rate, preventing the probe from remaining in the active site long enough to form covalent linkage with C917. Modifications on GEPs that could slow down the sugar transfer should be implemented. It has been reported that switching the oxygen in the pyranose ring in nucleotide sugar could decrease the sugar transfer rate of the corresponding glycotransferases by 95-99.8%.[4,17] We anticipated that replacing the oxygen with sulfur in the pyranose ring of GEPs could extend the time they stay in the active site to covalently inhibit OGT. Following the idea, we designed a panel of inhibitors **UDP-ES1**-**UDP-ES4**, each carried allyl chloride, allyl fluoride, epoxide, and conjugated double bonds side chains. (Figure 3.4) For cellular use, we also designed the prodrug form of each inhibitor, **ES1**-**ES4**, with UDP removed, and all hydroxyls protected by acetyl groups. The prodrugs could readily penetrate through the cell membrane and transformed to the corresponding UDP form through the hexosamine biosynthetic pathway.

We first started the synthesis of **ES1-ES4**. Although there is only one atom difference between 5-thio-GlcNAc and GlcNAc, the synthesis is complicated and time consuming. Starting with glucosamine, we first switched the free amine to the relatively stable azide protecting group. The pyranose ring was then opened, with hydroxyl and aldehyde groups protected as acetals. The acetal protecting 5 and 6-positions were selectively deprotected, leaving a primary alcohol and secondary alcohol. Taking advantage of the different reactivity of 5 and 6-position, we first protected 6-postion with benzyl then 5-position with mesyl. Under mild basic condition, an epoxide would be formed at 5 and 6-position. The epoxide could be replaced by episulfide with thiourea. After the replacement was finished, the other protecting groups were hydrolyzed, forming the 5-thio-pyranose. The azide could be reduced back to amine and ready to be attached with different electrophilic side chains, yielding **ES1-ES4**.

3.3.2 **ES1** is a potent, specific, and little toxic inhibitor in cells

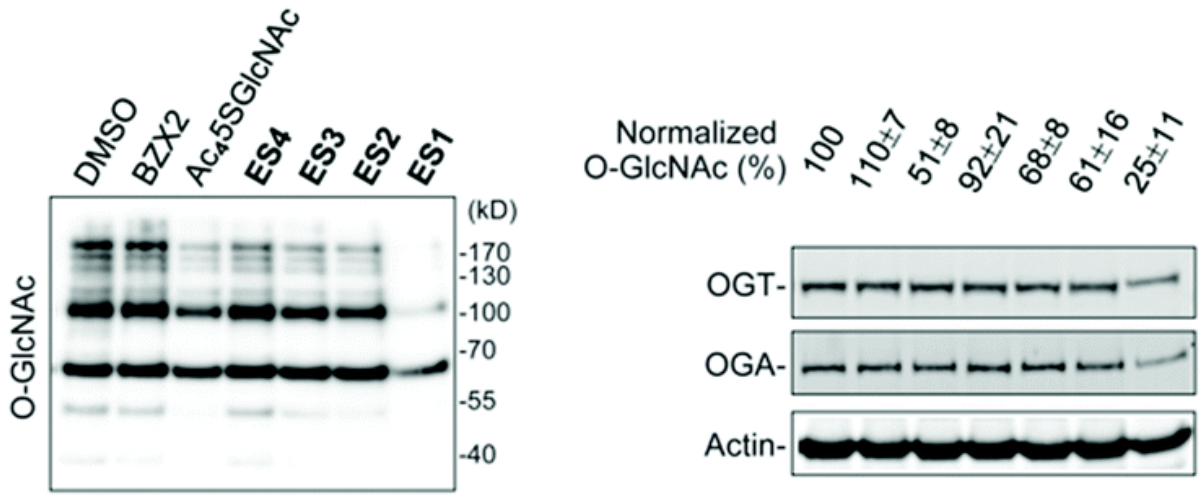


*Figure 3.5 **ES1** had best potency among a panel of inhibitors*

With the inhibitors synthesized, we first tested the cellular potency. Human breast cancer MCF7 cells were treated with **ES1-ES4**, $Ac_4$5SGlcNAc, one of the best OGT inhibitors, and BZX2, the only covalent inhibitor of OGT. The inhibitors were removed by washing to let the cells incubate in inhibitor-free media for several hours before testing the O-GlcNAcylation by western blot. From the results, **ES1, ES2, ES3**

and $Ac_4$5SGlcNAc showed significant suppression of O-GlcNAc level in cells while BZX2 and **ES4** had very minimal inhibition. (Figure 3.5) Specifically, **ES1** decreased the global O-GlcNAcylation level in MCF7 cells significantly more than all other inhibitors at the same concentration.

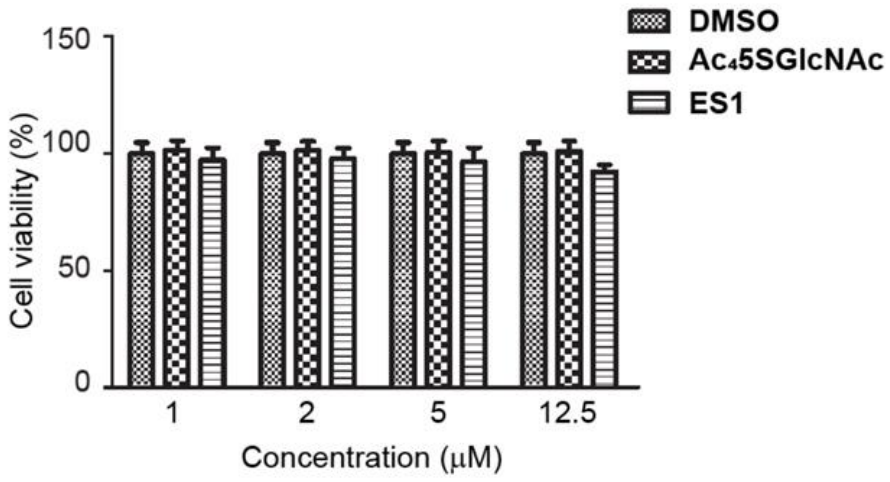


*Figure 3.6 **ES1** had minimal toxicity*

As the high potency of **ES1** was encouraging, we were also concerned about the toxicity and specificity of the inhibitor, especially that **ES1** was a covalent inhibitor that could permanently abolish the activity of OGT. At the same concentration that **ES1** could significantly decrease the O-GlcNAcylation level, we found that it had very minimal effect to the viability of MCF7 cells, indicating that **ES1** has very low toxicity at its functional concentration. (Figure 3.6) At the same time, we tested the influence of **ES1** on the biosynthesis of a few types of glycosylation, including N-glycans, O-glycans,[18] and a specific type of O-GlcNAcylation on EGF repeats of cell surface proteins[19] (Figure 3.7a) with the corresponding antibodies. At 15 μM which was sufficient for **ES1** to significantly inhibit OGT, N- and O-glycosylation were not affected. (Figure 3.7b) Treatment of **ES1** affected the glycosylation level of OGT substrate NUP62 but had minimal impact on EOGT substrate EGF. (Figure 3.7c)

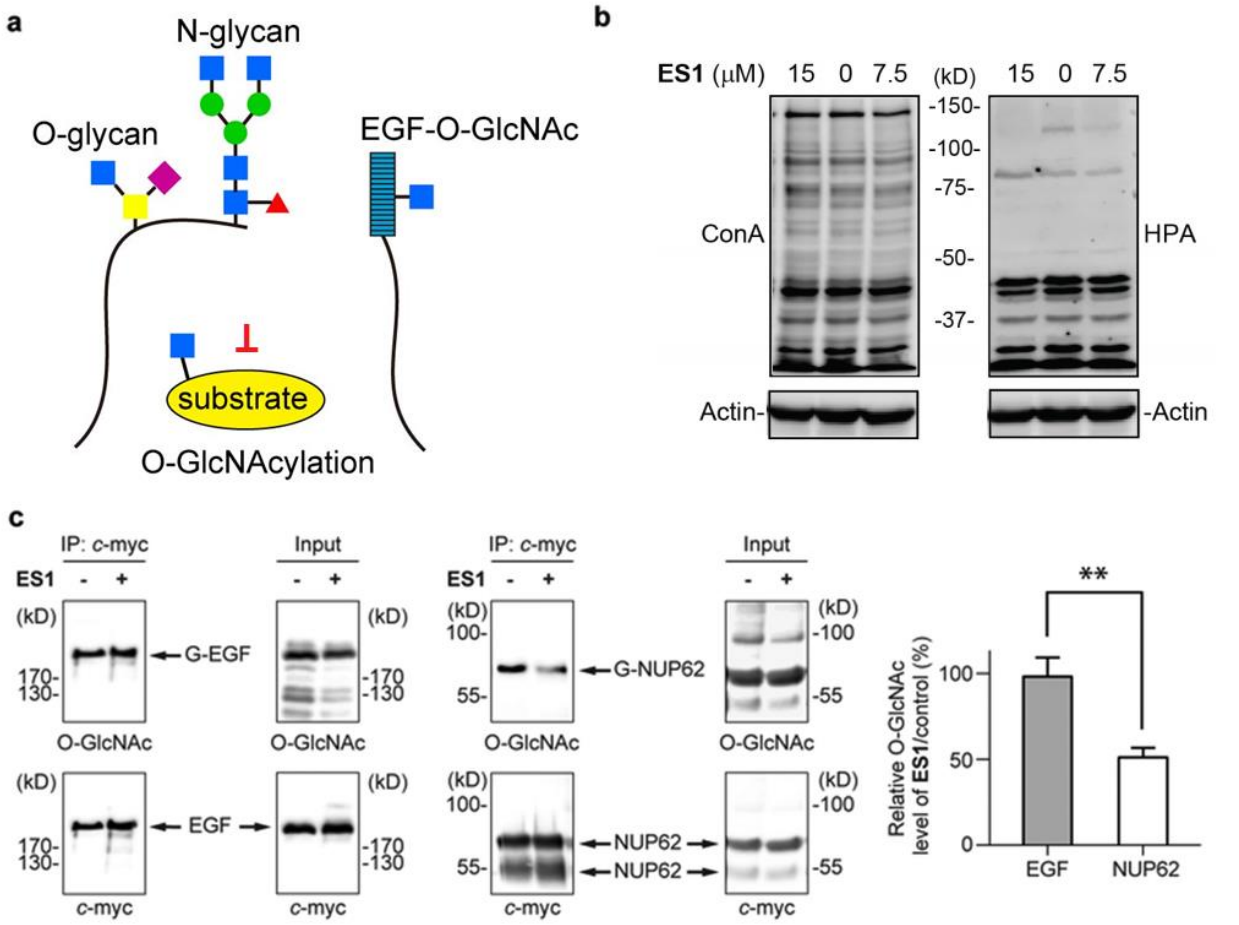


Figure 3.7 ***ES1*** *is a specific inhibitor for O-GlcNAcylation (a)* ***ES1*** *only inhibits O-GlcNAcylation but not other glycosylation (b)* ***ES1*** *did not affect N- and O-linked glycosylation (c)* ***ES1*** *did not affect the level of EGF repeats O-GlcNAcylation*

ES1 was proven to be a potent, specific, and not toxic covalent inhibitor. In order to characterize the *in vitro* inhibition of **ES1**, we followed to synthesize **UDP-ES1**.

3.3.3 Synthesis of **UDP-ES1**

The synthesis of **UDP-ES1** could follow the previous synthesis of **GEP1**. As the starting material **10** was hard to obtain, the low overall yield of traditional organic synthesis method with only ~10% overall yield to generate UDP-sugar could not be applied. In order to achieve high yielding attachment of UDP onto 5SGlcNAc, we adopted a previously reported one-pot three-enzyme chemoenzymatic synthesis, mimicking the cellular hexosamine biosynthetic pathway.[13] For the one-pot three-enzyme reactions, we

selected NahK for the attachment of α-phosphate onto the 1-position of the sugar, and AGX1 for the synthesis of UDP-5SGlcNAc from GlcNAc-1-phosphate, in presence of ATP. The third enzyme is inorganic pyrophosphatase to hydrolyze the diphosphate released from the reaction, pushing the reaction balance toward the final product.

In order to optimize reaction condition to obtain highest yield, I tested the reaction with different ATP, UTP, NahK, AGX1, and pyrophosphatase concentration, as well as reaction pH and time. With the optimized condition, compound **13** (Trifluoroacetylated UDP-5Sglucosamine) could be synthesized with 92% yield in 36 hours, compared to 10% yield in more than a week with traditional chemical synthesis. Following that, the trifluoroacetyl group was hydrolyzed for the attachment of allyl chloride side chain to get **UDP-ES1**.

3.3.4 **UDP-ES1** is a targeted covalent inhibitor of OGT *in vitro*

The inhibition pattern of **UDP-ES1** was first tested with kinetic assay. OGT pre-incubated with **UDP-ES1** was used for *in vitro* glycosylation of protein substrates with radio labeled UDP-GlcNAc. **UDP-ES1** was found to follow time-dependent OGT inactivation with inhibition efficiency of 0.06 $\mu M^{-1}min^{-1}$. (Figure 3.8a) Parallelly, intact protein MS data revealed that large excess **UDP-ES1** compared to OGT could only show one molecule attachment of the sugar, while the C917 mutant could not be modified by **UDP-ES1** under same condition, indicating that OGT could be modified by only one molecule of **UDP-ES1**. (Figure 3.8b) LC-MS/MS experiments mapped the modification site on C917. The *in vitro* data demonstrated that **UDP-ES1** was a targeted covalent inhibitor of OGT, targeting C917. (Figure 3.8c)

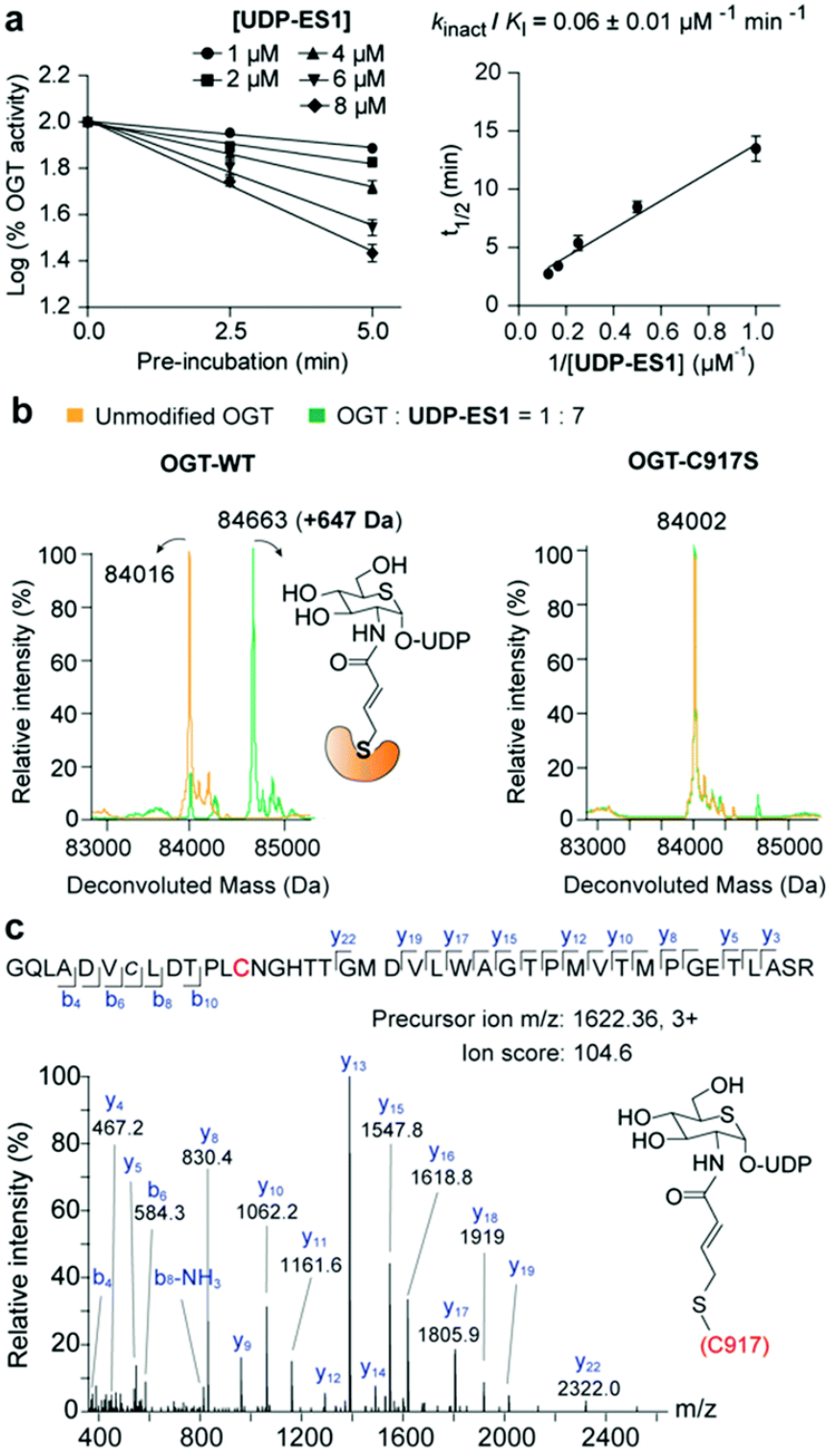


Figure 3.8 ***ES1*** *covalently inhibited OGT in vitro (a)* ***ES1*** *showed time-dependent inhibition of OGT (b) and (c) Intact protein MS and LC-MS/MS showed that* ***ES1*** *specifically targeted C917*

To obtain unambiguous evidence for the covalent inhibition, I solved the crystal structure of OGT in complex with **UDP-ES1** and CKII peptide at 2.5 Å resolution. From the structure, the density for **UDP-ES1** and CKII peptide could be seen clearly. Comparing the structure with apo form OGT, the overall conformation stayed unchanged, indicating that **UDP-ES1** would not affect the conformation of OGT. (Figure 3.9a) Detailed inspection into the active site revealed the clear electron density between the side chain of **UDP-ES1** and C917. (Figure 3.9b) Overlaying the structure with OGT:UDP-5SGlcNAc complex structure, no obvious conformational change of the UDP sugar was found, proving that the extra allyl chloride side chain would not affect how UDP-GlcNAc binds to OGT. (Figure 3.9c and 3.9d) Moreover, despite the presence of CKII peptide as a readily available substrate, the sugar was not transferred to the substrate. In comparison, when crystallized in the same condition with **GEP1**, the sugar has already been transferred to the CKII peptide. (Figure 3.9c and 3.9e) The crystal structure yielded the solid evidence for the targeted covalent inhibition of **UDP-ES1**.

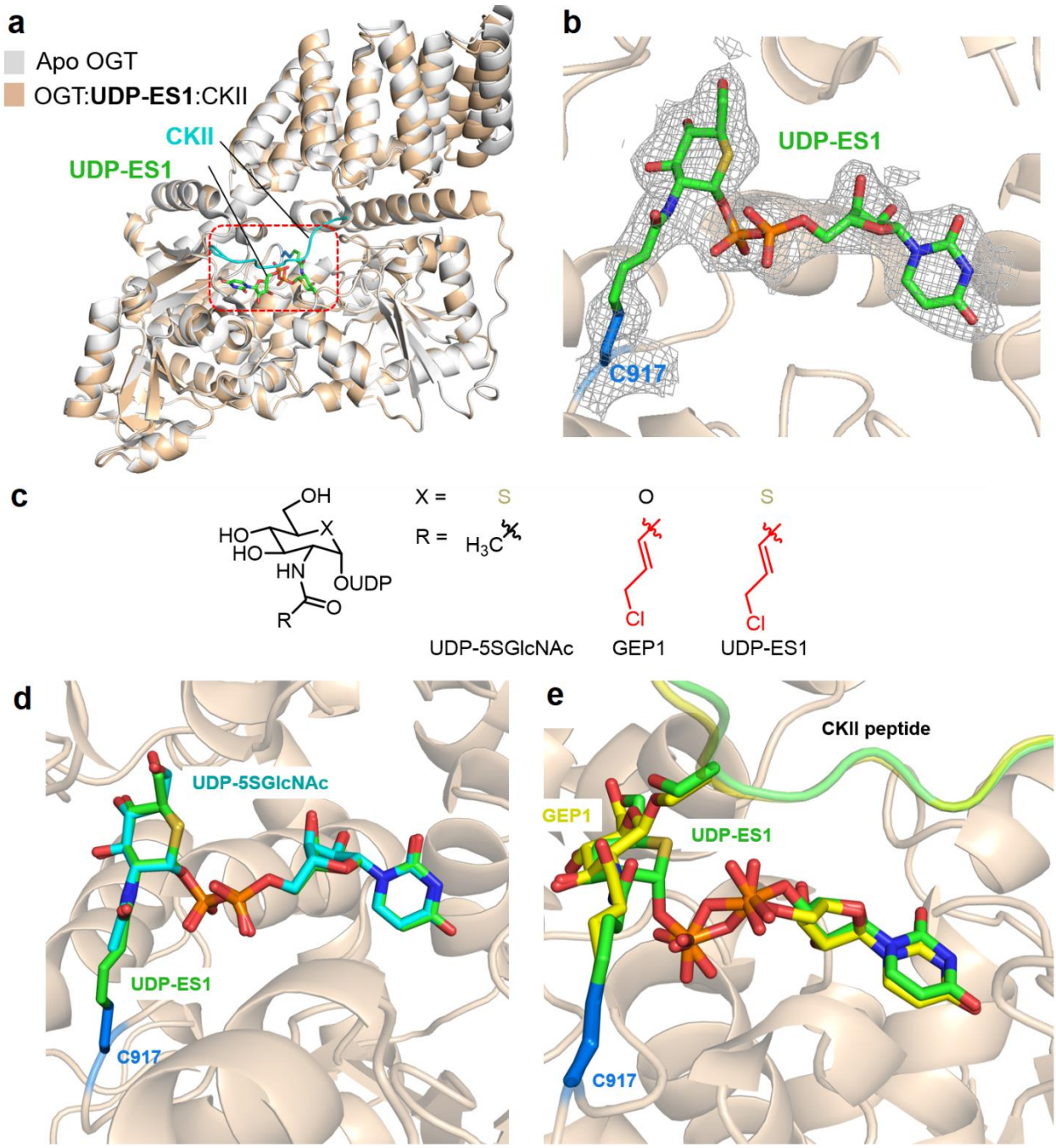


*Figure 3.9 Crystal structure of OGT in complex with **ES1** and CKII peptide (a) Superposition of OGT apo form structure (PDB 3PE3, white) and OGT:**ES1**:CKII peptide complex (wheat) (b) Electron density of the covalent linkage between **ES1** and C917 (c) Structures of UDP-5SGlcNAc, **GEP1**, and **ES1** (d) Superposition of OGT:5SGlcNAc:CKII peptide complex (PDB 4GYY, sugar shown in cyan sticks) and OGT:**UDP-ES1**:CKII peptide complex (sugar shown in green sticks) (e) Superposition of OGT:**GEP1**:CKII peptide complex (PDB 5VIF, sugar shown in yellow sticks) and OGT:**UDP-ES1**:CKII peptide complex (sugar shown in green sticks)*

3.4 Summary and future plan

In summary, we developed the cellular targeted covalent inhibitor of OGT, **ES1**. **ES1** could covalently inhibit OGT with permanent linkage onto C917. Cellular experiments indicated that **ES1** was more potent than one of the best inhibitors of OGT, 5SGlcNAc. At the same time, **ES1** inhibition is specific to O-GlcNAcylation with minimal toxicity to the cell. **UDP-ES1** as the *in vitro* form inhibitor, followed the inhibition pattern as typical covalent inhibitors. MS and crystallization data also confirmed the targeted covalent inhibition of **UDP-ES1** onto OGT. Last but not the least, the designing idea of targeted covalent inhibitor was also valuable as it could guide the development of more potent inhibitors in the future.

For future plans, we aim to apply the strategy to develop OGT covalent inhibitors with better potency and specificity.

3.5 References

1. Worth, M., Li, H. & Jiang, J. Deciphering the Functions of Protein *O* -GlcNAcylation with Chemistry. *ACS Chem. Biol.* **12**, 326–335 (2017).

2. de Queiroz, R. M., Carvalho, E. & Dias, W. B. O-GlcNAcylation: The Sweet Side of the Cancer. *Front. Oncol.* **4**, 132 (2014).

3. Wang, L. *et al.* Suppressed OGT expression inhibits cell proliferation while inducing cell apoptosis in bladder cancer. *BMC Cancer* **18**, 1141 (2018).

4. Gloster, T. M. *et al.* Hijacking a biosynthetic pathway yields a glycosyltransferase inhibitor within cells. *Nat. Chem. Biol.* **7**, 174–181 (2011).

5. Borodkin, V. S. *et al.* Bisubstrate UDP-peptide conjugates as human O-GlcNAc transferase inhibitors. *Biochem. J.* **457**, 497–502 (2014).

6. Ortiz-Meoz, R. F. *et al.* A small molecule that inhibits OGT activity in cells. *ACS Chem. Biol.* **10**, 1392–1397 (2015).

7. Martin, S. E. S. *et al.* Structure-Based Evolution of Low Nanomolar O-GlcNAc Transferase Inhibitors. *J. Am. Chem. Soc.* **140**, 13542–13545 (2018).

8. Bauer, R. A. Covalent inhibitors in drug discovery: from accidental discoveries to avoided liabilities and designed therapies. *Drug Discov. Today* **20**, 1061–1073 (2015).

9. Jiang, J., Lazarus, M. B., Pasquina, L., Sliz, P. & Walker, S. A neutral diphosphate mimic crosslinks the active site of human O-GlcNAc transferase. *Nat. Chem. Biol.* **8**, 72–77 (2011).

10. Yan, R.-B., Yang, F., Wu, Y., Zhang, L.-H. & Ye, X.-S. An efficient and improved procedure for preparation of triflyl azide and application in catalytic diazotransfer reaction. *Tetrahedron Lett.* **46**, 8993–8995 (2005).

11. Hu, C.-W. *et al.* Electrophilic probes for deciphering substrate recognition by O-GlcNAc transferase. *Nat. Chem. Biol.* **13**, 1267–1273 (2017).

12. Li, Y. *et al.* Substrate Promiscuity of N-Acetylhexosamine 1-Kinases. *Molecules* **16**, 6396–6407 (2011).

13. Guan, W., Cai, L. & Wang, P. G. Highly Efficient Synthesis of UDP-GalNAc/GlcNAc Analogues with Promiscuous Recombinant Human UDP-GalNAc Pyrophosphorylase AGX1. *Chem. - Eur. J.* **16**, 13343–13345 (2010).

14. Lazarus, M. B., Nam, Y., Jiang, J., Sliz, P. & Walker, S. Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature* **469**, 564–567 (2011).

15. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).

16. Nakayama, G. R., Caton, M. C., Nova, M. P. & Parandoosh, Z. Assessment of the Alamar Blue assay for cellular growth and viability in vitro. *J. Immunol. Methods* **204**, 205–208 (1997).

17. Tsuruta, O., Shinohara, G., Yuasa, H. & Hashimoto, H. UDP-N-acetyl-5-thio-galactosamine is a substrate of lactose synthase. *Bioorg. Med. Chem. Lett.* **7**, 2523–2526 (1997).

18. Arenas, M. I. *et al.* Identification of N- and O-linked oligosaccharides in human seminal vesicles. *J. Androl.* **22**, 79–87 (2001).

19. Ogawa, M. & Okajima, T. Structure and function of extracellular O-GlcNAc. *Curr. Opin. Struct. Biol.* **56**, 72–77 (2019).

3.6 Supplementary information

Table 3.1 Data collection and refinement statistics (molecular replacement)

	GOP-5S9c-CKII peptide
Data collection	
Space group	*F*222
Cell dimensions	
a, *b*, *c* (Å)	137.7, 151.0, 198.9
a, *b*, *g* (°)	90.0, 90.0, 90.0
Resolution (Å)	50.0–2.55 (2.64–2.55)*
R_{merge}	9.4 (51.2)
I/s(*I*)	16 (2.5)
$CC_{1/2}$	99.2 (93.2)
Completeness (%)	99.8 (99.9)
Redundancy	13.3 (12.1)
Refinement	
Resolution (Å)	50.0–2.55
No. reflections	34,179
R_{work} / R_{free}	16.7 / 21.1
No. atoms	
Protein	5,563
Ligand	41
Water	191
B factors	40.01
Protein	40.21
Ligand	29.68
Water	36.38
R.m.s. deviation	
Bond length (Å)	0.008
Bond angle (°)	0.968

*Each structure was determined from one crystal.

*Values in parentheses are for highest-resolution shell.

Chapter 4 Deciphering the substrate recognition of OGA

The results of this chapter were published on:

Li, B.; **Li, H.**; Lu, L.; Jiang, J. Structures of Human O-GlcNAcase and Its Complexes Reveal a New Substrate Recognition Mode. *Nat. Struct. Mol. Biol.* **2017**, *24* (4), 362–369.

Li, B.*; **Li, H.***; Hu, C.-W.; Jiang, J. Structural Insights into the Substrate Binding Adaptability and Specificity of Human O-GlcNAcase. *Nat. Commun.* **2017**, *8* (1), 666. (*: equal contribution)

Contributions

This project is a collaborative project. I was involved in protein crystallization. Together with Dr. Baobin Li, I determined the crystal structure of human OGA in complex with five O-GlcNAcylated peptide substrates.

4.1 Introduction

O-GlcNAcase is the sole glycosidase responsible for the removal of O-GlcNAc moiety in cell. Same as OGT, OGA catalyzes the deglycosylation on thousands of substrates, from which many are involved in significant biological processes.

However, unlike the important regulatory roles of OGT in O-GlcNAc biology that we are beginning to understand, it is still under debate how important OGA regulation is, it has been investigated significant less. There are two challenges preventing the investigation into the substrate recognition of OGA. First, there is no conserved sequence motif for O-GlcNAcylation, making the prediction of substrates based on the sequence unavailable. On the other hand, crystal structures of enzyme-substrate complex could provide information in the binding mode and recognition pattern of the enzyme, while the crystal structure of human OGA has not been solved after more than 15 years of its discovery.

Some people thought that OGA was only involved in the sugar recognition and removal. First, unlike OGT, knock out of OGA is not lethal.[1] Additionally, it has been reported that human OGA catalyzes the removal of GlcNAc on a few O-GlcNAcylated proteins with similar catalytic efficiency.[2] This was an indication that OGA might only recognize the sugar despite the amino acid sequence flanking the O-GlcNAc site. In this way OGA is not regulating O-GlcNAc balance with the substrate-specific recognition pattern. On the other hand, the crystal structure of *Og*OGA, a bacterial homologue of human OGA, was determined in complex with O-GlcNAcylated peptide substrate.[3] Surprisingly, *Og*OGA formed interaction with the peptide in addition to the sugar. Mutations were made on human OGA corresponding to the interacting residues in *Og*OGA, and the mutations showed impaired binding with the substrate. This evidence supported the hypothesis that OGA not only recognizes the sugar, but also possesses the ability to discriminate its substrates.

We aim to determine the crystal structure of human OGA in complex with different substrates to investigate whether OGA possesses the substrate recognition, and how OGA recognizes its substrates.

OGA has long been considered as a therapeutic target for certain types of neurodegenerative diseases.[4] We have discussed how O-GlcNAcylation regulates Alzheimer's Disease (AD) and Parkinson's Disease. One important hypothesis for the progression of AD is that hyperphosphorylation of tau protein engages microtubule deconstruction. Tau is an important stabilizing microtubule associated protein, and its function is highly regulated by its phosphorylation level, which has been connected to the O-GlcNAcylation level.[5] It has been reported that the treatment of OGA inhibitor thiamet-G could significantly increase O-GlcNAcylation on tau at S400.[6] This increase of O-GlcNAc level in turn suppresses the phosphorylation on nearby sites, preventing the aggregation of tau. Treatment of thiamet-G has also been proven to decrease phosphorylation of tau protein in rat model.[7] The effect of thiamet-G disappeared within 24 hours, limiting the usage of thiamet-G as drug for AD.

Since OGA inhibitors could provide a solution for AD symptoms, the development of OGA inhibitors with better potency, specificity and pharmacodynamics is extremely important for the treatment of AD. This also emphasized the importance of the crystal structure of human OGA, which could guide the development of inhibitors targeting only human OGA other than its analogues.

4.2 Material and methods

Protein expression and purification.

Sequences for the full-length and truncated OGA (OGA_{cryst} or OGA60-704-del, based on the numbering of the full-length human protein) proteins were subcloned into a modified pET24b vector or pET-SUMO vector, respectively. The plasmids were transformed into Escherichia coli strain Rosetta(DE3), and the transformants were grown at 37 °C in Luria–Bertani (LB) medium. After OD600 reached 0.8, the culture was cooled to 16 °C and induced by 0.3 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). After 16 h, the cells were harvested, resuspended and homogenized with an ultra-high-pressure cell disrupter (Emulsiflex-C5, Canada) at 4 °C. The supernatant was subjected to Ni–NTA column (Qiagen) for affinity purification and was subsequently eluted with buffer containing 20 mM Tris (pH 8.0), 150 mM NaCl, and

250 mM imidazole. The eluted full-length OGA was further purified by size-exclusion chromatography (Superdex 200 increase 10/300, GE Healthcare) in buffer containing 20 mM Tris (pH 8.0), 150 mM NaCl, and 0.5 mM Tris(hydroxymethyl)phosphine (THP; EMD), and it was then concentrated to a concentration of 3 mg/ml for the in vitro assays. The eluted OGAcryst was digested by Sumo protease to remove the 6× His–SUMO tag and further purified by size-exclusion chromatography. The OGAcryst protein was concentrated to a concentration of 3 mg/ml for crystallization. In this construct, residues 401–552 were replaced by a 15-residue glycine–serine linker (GGGGSGGGGSGGGGS). Various OGA mutants were generated using the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent), according to the manufacturer's instructions, using the sequence for full-length hOGA or OGAcryst as the DNA template along with the primers listed in Supplementary Table 2. The DNA sequences were verified by sequencing. Mutant proteins were expressed and purified similarly as described above.

Crystallization.

All of the crystals were generated by mixing 1 μl of protein with an equal volume of reservoir solution and equilibration against 200 μl of reservoir solution using the hanging-drop vapor-diffusion method at 20 °C. Native OGA_{cryst} crystals (wild-type and D175N mutant) were obtained in the reservoir solution containing 0.032 M ammonium citrate tribasic (pH 7.0), 0.02 M MES monohydrate, 0.128 M potassium thiocyanate, 0.016 M imidazole, 0.002 M zinc sulfate heptahydrate, 12.8% wt/vol polyethylene glycol 3,350, 3.2% wt/vol polyethylene glycol monomethyl ether 2,000, and 5% wt/vol polyethylene glycol monomethyl ether 550. For the inhibitor complex, OGA_{cryst} was incubated with 2 mM thiamet-G (synthesized as reported previously[7]) for 1 h and co-crystallized in conditions similar to those as for the native protein. For the OGAcryst-glycopeptide complexes, the glycopeptides were prepared by solid-phase peptide synthesis. Native crystals (D175N mutant) were soaked for 30 min in reservoir solution containing 2 mM of glycopeptide. All the crystals were transferred into cryoprotectant solution containing

the respective mother liquor plus 10% (vol/vol) glycerol before being flash-frozen in liquid nitrogen for storage.

Data collection and structural determination.

All of the X-ray data were collected on the Life Sciences Collaborative Access Team (LS-CAT) beam lines 21-ID-G (for OGA_{cryst}, OGA_{cryst}-D175N, OGAcryst-D175N-p53 complex, OGAcryst-D175N-TAB1 complex and OGAcryst-D175N-α-crystallin complex) and 21-ID-D (for the OGA_{cryst}-thiamet-G, OGA_{cryst}-D175N-p53, OGA_{cryst}-D175N-ELK1, OGA_{cryst}-D175N-Lamin B1 complexes) (LS-CAT, Advanced Photon Source, Argonne National Laboratory, IL, USA). The wavelength for data collection was 0.9787 Å. All data sets were processed using the HKL2000 package[8]. The crystal of OGA_{cryst}-D175N-p53 complex also belonged to the space group of *P*21 with four molecules per asymmetric unit. All other crystals belonged to the same space group of *P*21, containing two molecules per asymmetric unit. The structures were solved by molecular replacement using *Cp*OGA as a search model (PDB 2CBJ)39. Iterative model building was performed in COOT[9] and further refined in PHENIX[10] and CCP4.[11] Final refinement statistics are summarized in Table 1. Structural figures were drawn using the program PyMOL.[12] The coordinates and structure factors have been deposited in the Protein Data Bank (PDB) under accession codes 5TKE, 5UN9, 5UN8, 5VVO, 5VVV, 5VVU, 5VVT, and 5VVX.

Analytical ultracentrifugation sedimentation equilibrium.

Sedimentation equilibrium studies were performed using a Beckman Model XL-A Analytical Ultracentrifuge in the Biophysics Instrumentation Facility at the University of Wisconsin–Madison. Samples for the sedimentation equilibrium experiments were diluted to the desired protein concentrations in 20 mM Tris-HCl (pH 7.0), 150 mM NaCl, and 0.5 mM THP. Data were collected at 20 °C and three rotor speeds for each protein. For full-length hOGA, sedimentation equilibrium was analyzed at initial

loading concentrations of 0.47 mg/ml and 0.14 mg/ml with rotor speeds of 4,000 r.p.m., 5,600 r.p.m., and 7,600 r.p.m. OGAcryst protein was analyzed at initial loading concentrations of 0.30 mg/ml and 0.095 mg/ml with rotor speeds of 6,000 r.p.m., 8,800 r.p.m., and 12,000 r.p.m. Data were analyzed at the facility by Dr. Darrell McCaslin, using approaches similar to that described by Laue[13].

OGA enzymatic assays.

Steady-state kinetics of wild-type and mutant hOGA were determined using the strategy of Macauley33. The fluorogenic substrate 4-methylumbelliferyl-N-acetyl-β-D-glucosaminide (4MU-NAG; Fisher) was used in this assay. Standard reaction mixtures (25 μl) contained 2 nM of enzyme in 50 mM NaH2PO4 (pH 6.5), 100 mM NaCl, 0.1 mg/ml BSA, and various substrate concentrations (10, 50, 150, 300, 600, and 1200 μM). The reactions were incubated at 37 °C for 10 min, and then quenched by the addition of 150 μl of 200 mM glycine–NaOH (pH 10.75). The fluorescence of liberated 4-methylumbelliferone was monitored with a Microplate Fluorescence Reader (Bio-Tek), with an excitation wavelength of 360 nm and an emission wavelength of 460 nm. The released 4-methylumbelliferone was quantified by the standard curve of 4-methylumbelliferone under the identical conditions. All the measurements were carried out in triplicate, and the Michaelis–Menten kinetic parameters were calculated by fitting the data in GraphPad Prism (GraphPad Software, La Jolla, CA). Relative-activity assays were performed in the presence of 2 nM enzyme and 200 μM 4MU-NAG at 37 °C for 10 min in triplicate. The fluorogenic measurement was determined as mentioned above. Glycopeptide competition assays were performed in triplicate using 4MU-NAG as the reporter substrate, using multi-substrate enzyme kinetics as previously reported.[14] The data were analyzed by GraphPad Prism using the following equation: $vi/v0 = ((1 + Km/S))/(1 + Km/S(1 + S'/Km'))$ to yield Km′ (the Michaelis constant of the competition substrate p53 glycopeptide).

4.3 Results and discussion

4.3.1 OGA is present as an unusual arm-in-arm homodimer

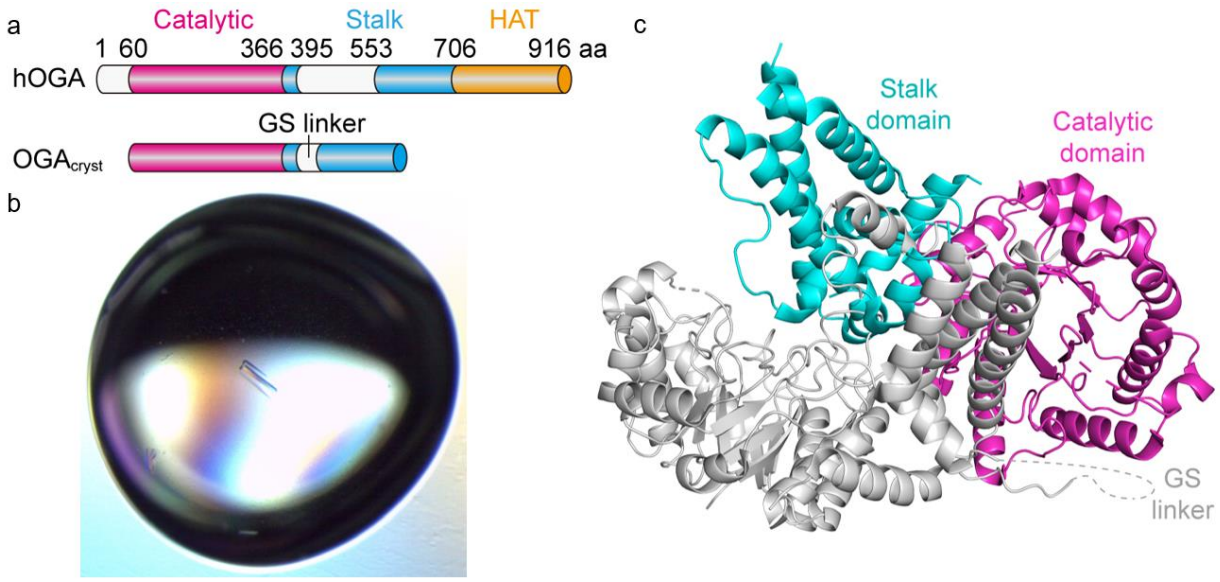


Figure 4.1 The first crystal structure of human OGA (a) Domain representation of human OGA and OGA_{cryst} (b) The crystal of OGA_{cryst} (c) The ribbon representation of OGA dimer structure. The catalytic domain and stalk domain of one monomer of OGA are shown in orange and blue ribbon, respectively. The other monomer is shown in white ribbon.

Human O-GlcNAcase is a multi-domain protein containing an N-terminal catalytic domain, a C-terminal pseudo histone acetal transferase (HAT) domain and a stalk domain in between.[15] (Figure 4.1a) The crystal structure of human O-GlcNAcase has been absent since the enzyme was identified more than fifteen years ago, due to its abundant structural disordered regions, including 60 amino acids in the N-terminus, a large portion of stalk domain (aa 401-554) and the HAT domain, based on the secondary structure prediction.[16] In collaboration with a postdoctoral researcher in the lab, I learned and participated in design and crystallization of a truncated construct of human OGA, OGA_{cryst} (OGA60-400-GS-554-704), by removing most of the disordered regions in the N and C-terminus, and replacing the loop in the intervening domain with a flexible GS linker. The activity of OGA_{cryst} was tested with the 4-methylumbelliferyl *N*-acetyl-β-D-glucosaminide (4MU-GlcNAc) fluorescent assay. In this assay,

catalytically active OGA could hydrolyze 4MU-GlcNAc into GlcNAc and methylumbelliferone, which had fluorescence in basic conditions. Activity test showed that OGA_{cryst} had similar activity as full length OGA (flOGA).

We solved the first crystal structure of human OGA, in parallel with two other research groups.[17–19] Our structure demonstrated that, surprisingly, OGA forms an unusual arm-in-arm homodimer in which the catalytic domain of one monomer is covered by the stalk domain of the other monomer. (Figure 4.1c) The catalytic domain (residues 60-366) formed the classic structure core of eight-stranded parallel β-sheet core surrounded by eight α-helices called the $(\beta/\alpha)_8$-barrel, providing the binding pocket for GlcNAc moiety. Six α-helices formed the stalk domain, with two helices extended from the catalytic domain, three helices bundled together forming the structure core of stalk domain, and one extended helical connected by a loop. The extended helix was called the "arm" in the arm-in-arm homodimer, forming extensive interactions with the stalk domain of the other monomer. The unique dimerization of human OGA was distinct from its bacterial homologues, BtGH84, *Cp*OGA, *Og*OGA, and *Tt*OGA.[4] (Figure 4.2) We found that when overlaying the structures of human OGA and bacterial homologues, the catalytic domain could be perfectly aligned, due to the shared $(\beta/\alpha)_8$-barrel. However, the stalk domains of different homologues could vary significantly compared to human OGA. Most importantly, the bacterial homologues present as monomers in their crystal structures, and the active sites were distinct with human OGA, which has a much narrower substrate binding cleft that could provide more substrate specificity.

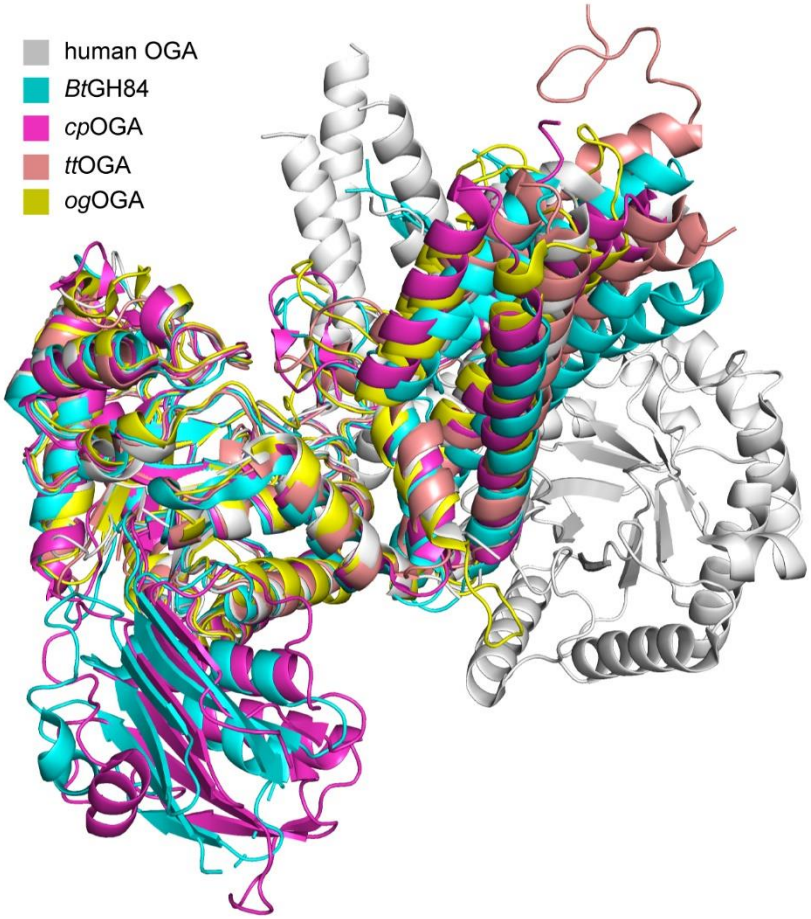


Figure 4.2 Superposition of OGA and bacterial homologues structures. Structures of human OGA, BtGH84, cpOGA, ttOGA, and ogOGA are shown in white, cyan, magenta, pink, and yellow ribbons respectively

Human OGA forms an extremely strong dimer in the crystal structure. Based on the PISA server[20] calculation, at least 20% of the surface area of OGA monomer was involved in the dimerization. (Figure 4.3) Additionally, numerous hydrogen bonds and salt bridges were found between the dimer interface, providing highly stable binding. The dimerization residues spread over the top surface of the catalytic domain, the stalk domain, and the arm helix interacting region. (Figure 4.3) The interactions strengthen the dimerization, and could play an important role in the activity, stability, and substrate recognition of human OGA.

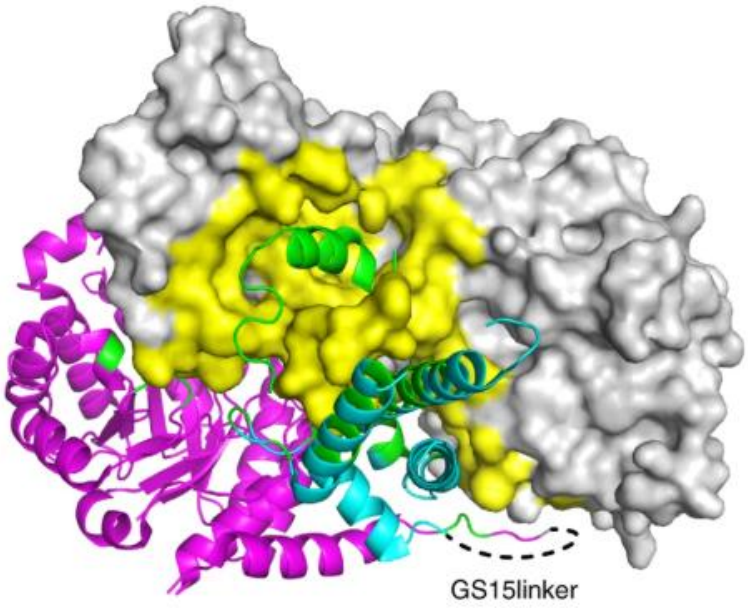
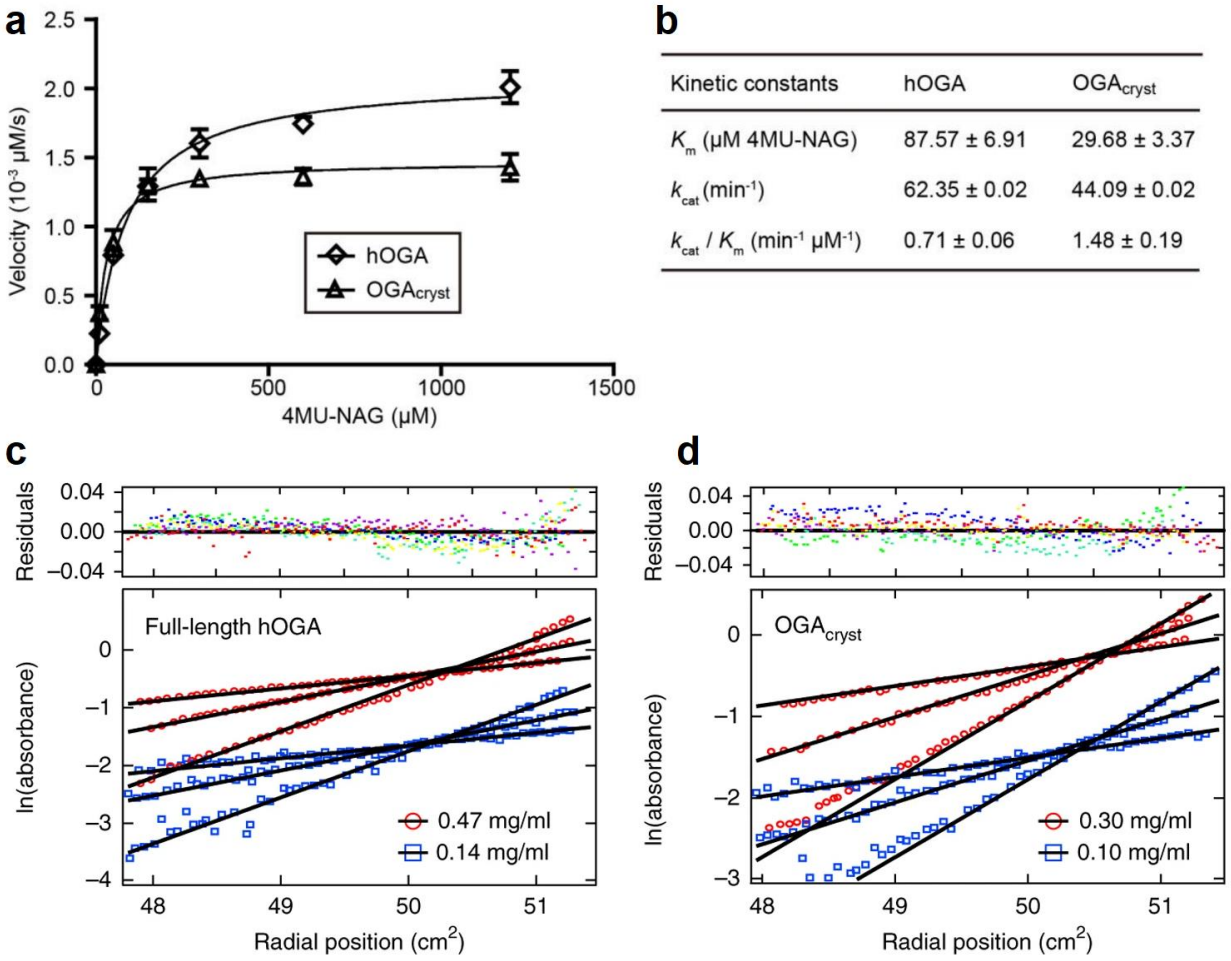


Figure 4.3 Indication of the dimer interface of human OGA. One monomer is shown in white surface. The catalytic domain and stalk domain are shown in magenta and cyan ribbons, respectively. The dimer interface is highlighted in yellow.

In order to test whether the dimerization is also present in solution for flOGA, analytical ultracentrifugation sedimentation equilibrium experiments were carried out. The results at different concentration and centrifugation speed showed that both flOGA and OGA_{cryst} exists as single species in solution, and only as dimers. (Figure 4.4a and 4.4b) The consistency between crystal and analytical ultracentrifugation experiments demonstrated that human OGA is a dimer in solution, and the crystal structure of OGA_{cryst} could represent flOGA.

This crystal structure not only served as the first published structure of human OGA, but also represented the first crystal structure of eukaryotes GH84 family glycosidases.



Kinetic constants	hOGA	OGA_{cryst}
K_m (µM 4MU-NAG)	87.57 ± 6.91	29.68 ± 3.37
k_{cat} (min^{-1})	62.35 ± 0.02	44.09 ± 0.02
k_{cat} / K_m (min^{-1} $µM^{-1}$)	0.71 ± 0.06	1.48 ± 0.19

Figure 4.4 OGA_{cryst} structure could represent full-length OGA in solution (a) Michaelis-Menten plots of hOGA and OGA_{cryst} (b) Kinetic parameters of hOGA and OGA_{cryst} (c) and (d) Analytical ultracentrifugation results of hOGA and OGA_{cryst}.

4.3.2 Structure of OGA in complex with thiamet-G could facilitate the development of OGA inhibitors

In order to obtain structure details about how OGA interacts with the GlcNAc moiety, we synthesized thiamet-G,[7] one of the best OGA inhibitor with K_i of 21 nM and determined the crystal structure of OGA in complex with thiamet-G at a resolution of 2.5 Å. (Figure 4.5a)

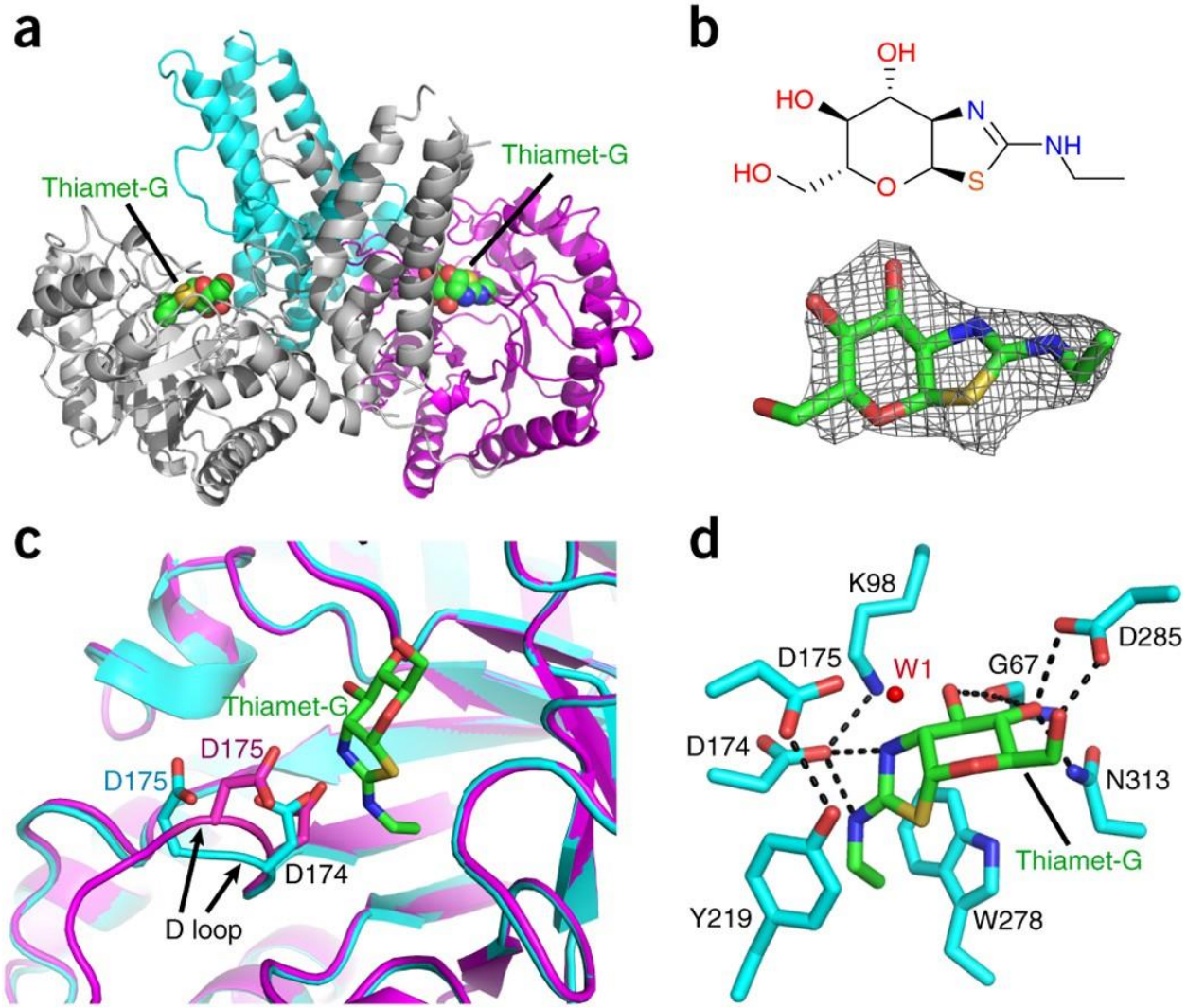


Figure 4.5 OGA in complex with an inhibitor thiamet-G (a) The ribbon representation of OGA in complex with thiamet-G. The catalytic domain and stalk domain of one monomer of OGA are shown in orange and blue ribbon, respectively. The other monomer is shown in white ribbon. Thiamet-G is shown in green spheres. (b) Structure of thiamet-G and the modeled structure of thiamet-G based on the electron density (c) Superposition of apo OGA (cyan) and OGA:thiamet-G complex (magenta) in the active site (d) Interactions between OGA active site and thiamet-G

From the structure, the electron density for thiamet-G could be clearly identified, (Figure 4.5b) and it occupied both active sites of OGA. (Figure 4.5a) Comparing the complex structure of OGA and thiamet-G with apo OGA structure, we found that there was minimal difference mainly in the active site loop (residues 172-182). In apo OGA structure, the loop was in a loose conformation, forming a wider active

site. On the other hand, with the binding of thiamet-G, the loop, especially residues D174 and D175 moved toward the sugar to form hydrogen bond interactions and stabilize the bound inhibitor. (Figure 4.5c) Previously, the catalytic mechanism of human OGA was proposed by crystal structures of bacterial homologues with structurally similar active site.[21] In the substrate facilitated mechanism, the proton transfer between D174-GlcNAc-D175 promoted the removal of GlcNAc from glycosylated substrate. D174 serving as a catalytic base, grabs the proton from the amide of the N-acetyl group, leading to formation of an oxazoline ring, and D175 provided a proton for the release of free substrate. In reverse, a water molecule was able to open the oxazoline ring with the help of D175N and recover the GlcNAc. In our structure, the residue locations were in agreement with this hypothesis, where D174 was close to the N-acetyl group and D175 near the anomeric carbon, with a water molecule in between, serving as the water for the hydrolysis. Aside from the residues that were already found to be important for catalysis, we identified that K98 and Y219 stabilized D174 and D175 in the process, respectively. According to the structure, Y219 and W278 also played important role in the stabilization of the oxazoline intermediate. (Figure 4.5d) Mutation of each of residues K98, D174, D175, Y219, and W278 led to the significant decrease of OGA activity. OGA-thiamet-G complex structure not only confirmed the catalytic mechanism of OGA, but also reveal novel residues of OGA responsible for sugar removal. The information was of great importance not only for the design of novel inhibitors of OGA for the study of OGA function and mechanism, but also for the development of treatment of neurodegenerative diseases.

4.3.3 Structures of OGA in complex with peptide substrates demonstrated that OGA is capable of recognizing its substrates

In order to understand the substrate recognition mode of OGA, we seek to solve the structures of OGA in complex with its peptide substrates. We first determined the crystal structure of OGA_{cryst}-D175N in complex with an O-GlcNAcylated peptide derived from p53 protein to 2.1 Å. P53 is an important tumor suppressor[22] and the O-GlcNAcylation on S149 could stabilize p53.[23] The peptide containing residues 144-154 with its O-GlcNAc on S149 was synthesized with commercial solid-phase peptide synthesis, and

soaked into the crystal of OGA_{cryst}-D175N to solve the complex structure. From the crystal structure, two molecules of p53 peptides were found to bind both active sites. (Figure 4.6a) Comparing the conformation of two peptides, we were surprised to find that although the sugars bound to the active sites in the same conformation, the peptides formed different binding modes in the substrate binding cleft which was formed by the catalytic domain of one monomer and the stalk domain of the other monomer. (Figure 4.6b) The different conformation of the two peptides might be contributed by the crystal packing. This unique binding mode was firstly discovered for glycosidases and significantly different from the homologues of OGA, as they do not have the same dimerization as human OGA. In another structure of *cp*OGA in complex with the exact same p53 O-GlcNAcylated peptide, the glycopeptide formed little interaction with *cp*OGA. (Figure 4.6c and 4.6d)

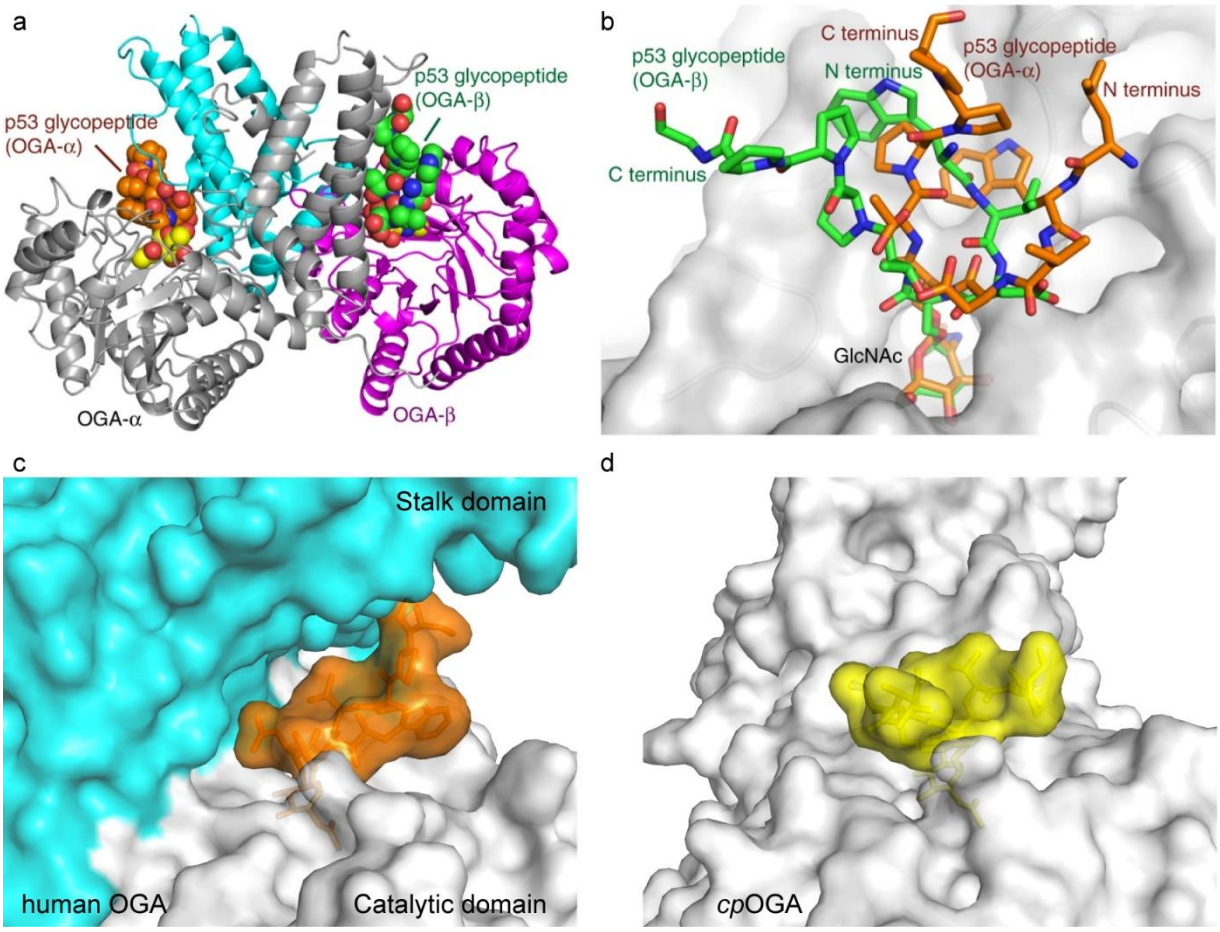


Figure 4.6 OGA bound to glycosylated p53 peptide different to the bacterial homologue (a) The ribbon representation of OGA in complex with p53. The catalytic domain and stalk domain of one monomer of OGA are shown in orange and blue ribbon, respectively. The other monomer is shown in white ribbon. Thiamet-G is shown in green sphere (b) Superposition of the p53 peptides in both substrate binding clefts (c) Surface representation of human OGA (white and cyan for catalytic and stalk domain, respectively) in complex with p53 peptide (orange) (d) Surface representation of cpOGA (white) in complex with p53 peptide (yellow)

Investigation into the detailed binding conformation of the p53 peptide led to important insight into the substrate recognition of OGA. First, the GlcNAc moiety attached to the peptide formed extensive hydrogen bond interactions with the catalytic pocket of OGA. (Figure 4.7a) Aside from the interactions contributed by the sugar, we were surprised to find additional interactions from the peptide of p53 with OGA. Hydrogen bonds were formed between D148, T150 from p53 and K648, W679 from OGA, while

hydrophobic interactions were identified between W146, V147, P152 from p53 and Y219, F223, V254, M622, F625, W645, W679 from OGA. (Figure 4.7b) The interacting residues on OGA were conserved among eukaryotes, indicating that these residues played an important role in the substrate recognition. The OGA-p53 complex structure not only served as the first crystal structure of human OGA complexed with a peptide substrate, but also proved that OGA could interact with its substrates through specific interactions, suggesting an ability to recognize its substrates through the substrate discrimination.

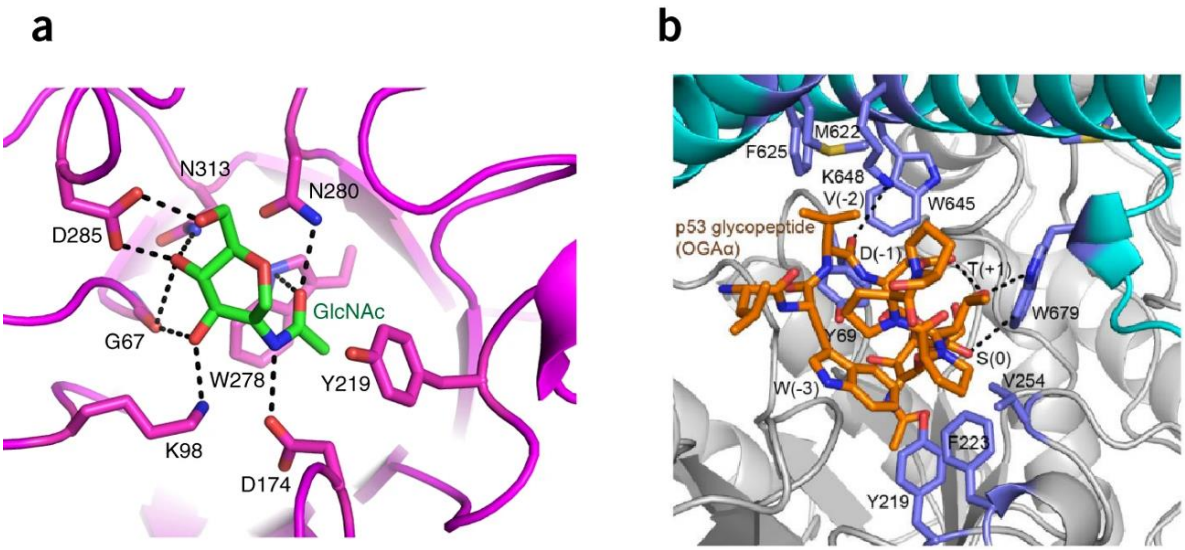


Figure 4.7 OGA formed specific interactions with p53 peptide (a) Interactions between O-GlcNAc moiety and OGA. OGA is shown in magenta ribbon and interacting residues shown in magenta sticks (b) Interactions between p53 peptide and OGA. OGA catalytic domain and stalk domain are shown in white and cyan ribbons, respectively. P53 peptide is shown in orange sticks.

With the encouraging finding that OGA recognizes its substrates, we seek to find evidences from more substrates. Next, we chose four more O-GlcNAcylated peptide substrates that were derived from characterized O-GlcNAcylation sites in the proteins: (a) α-crystallin B chain (FPT**S**TSLSPFYLR); (b) TAB1 (VPY**S**SAQS); (c) ELK1 (FWS**T**LSPI); and (d) Lamin B1 (KLSPSPSSRV**T**VS). (O-GlcNAc sites are highlighted with bold and underlined letters) Crystals of OGA_{cryst}-D175N were obtained and soaked with each of the four glycopeptides. We collected more than 100 datasets with different soaking

conditions and built models for all four peptides with electron density for 3-5 residues around the O-GlcNAc site.

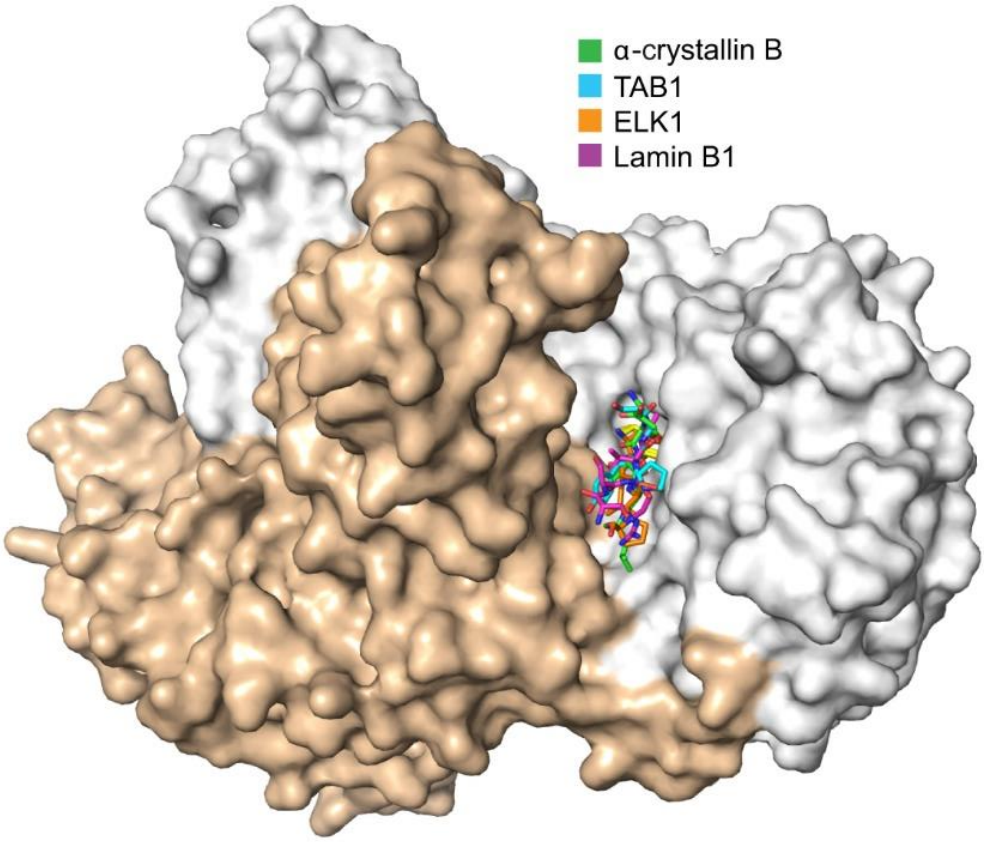


Figure 4.8 Overlay structure of OGA complexed with four glycopeptides. Glycopeptide substrates are derived from the following proteins: α-crystallin B chain, TAB1, ELK1, and Lamin B1. The two monomers of OGAcryst-D175N are shown in surface representation with white and wheat color, respectively. The glycopeptides are displayed in sticks with indicated colors.

From the solved structures, even though the peptide terminal residues lacked electron density, most residues adjacent to the O-GlcNAcylation site showed clear density, suggesting that while terminal residues could adopt flexible conformations, all peptides bound to OGA in the substrate cleft. Specifically, in each structure, the GlcNAc moiety of each glycopeptide were anchored by the same set of residues in the active site of OGA, showing identical conformation regardless of the peptide sequence connected to the sugar. (Figure 4.8)

Comparing the five structures of OGA complexed with different glycopeptides, including p53 peptide, we discovered that OGA bound to peptides with distinct sequences in a conserved manner, based on the conformation of peptide backbones. One slight variance was that p53 adopted a more curved conformation, which might be contributed by the intra-molecular hydrophobic interactions with in p53 peptide. Surprisingly, the peptides could adopt different directions in the substrate binding cleft, as α-crystallin and ELK1 bound to OGA in the reverse direction compared to the other three peptides.

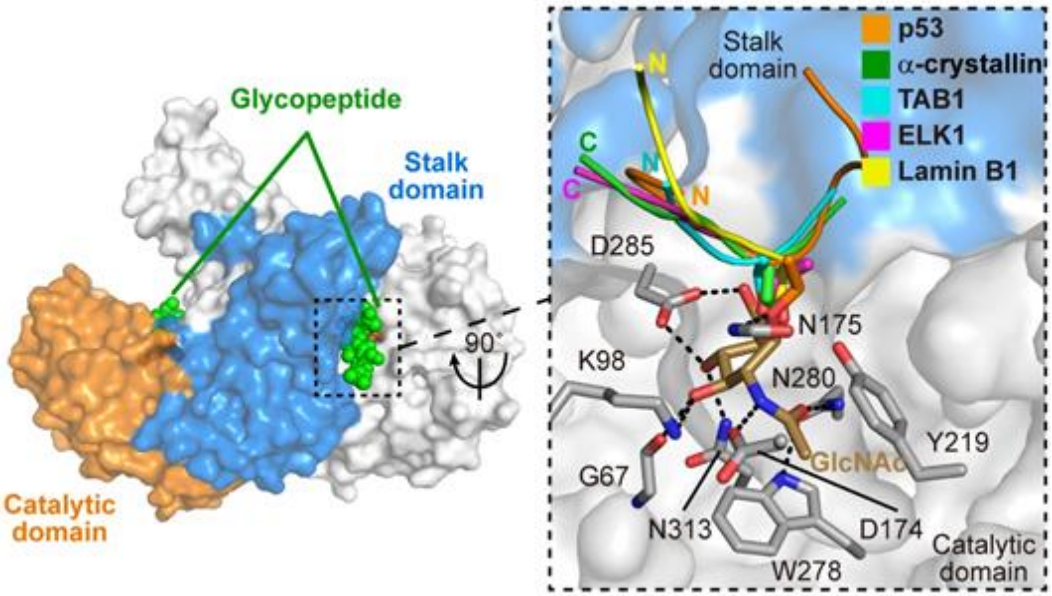


Figure 4.9 Glycopeptides adopted conserved and bi-directional conformations in the substrate binding cleft of OGA. OGA catalytic domain and stalk domain are shown as white and cyan surface, respectively. GlcNAc is demonstrated in brown sticks and the residues interacting with GlcNAc are demonstrated in white sticks. The peptides are shown in ribbon in different colors.

Detailed investigation into the binding of each peptide led to the finding that different peptides interacted with OGA in distinct interaction patterns, providing direct evidence that OGA is capable of recognizing and discriminating its substrates (Figure 4.10). For peptides O-GlcNAcylated on threonine (ELK1 and Lamin B1), the extra methyl groups were accommodated by the hydrophobic pocket formed by F223 and V254 from the catalytic domain and W679 from the stalk domain of the other monomer.

The structures gave insight into the general principle for OGA substrate recognition. The abundant and conserved interactions between the GlcNAc moiety and OGA catalytic site secure the binding of the glycopeptide within the substrate-binding cleft. These interactions serve as a prevalent driving force for OGA to selectively target O-GlcNAcylated substrates in the whole proteome. Enhanced selectivity can be achieved through substrate-specific interactions between OGA cleft surface residues and the substrate peptides. We have discussed the abundant interactions between OGA and p53 glycopeptide, and the other four peptides could form specific interactions with OGA as well. α-crystalline L44 binds to L141 of OGA through hydrophobic interaction. Although ELK1 P384 also bound similarly with L141 of OGA, it possessed two more hydrogen bonds between T381, S383 of the peptide and Y69, D177 of OGA. The binding of lamin B1 peptide, on the other hand, was stabilized by the hydrogen bond between S395 of lamin B1 and T626 of OGA, as well as the hydrophobic interaction between V398 and W679 of OGA. Last but not least, if the peptide bears intra-molecular interactions, it will provide additional stabilization energy for maintaining its ordered binding conformation. The knowledge obtained from this study will substantially advance understanding on the regulatory role of OGA in O-GlcNAc biology.

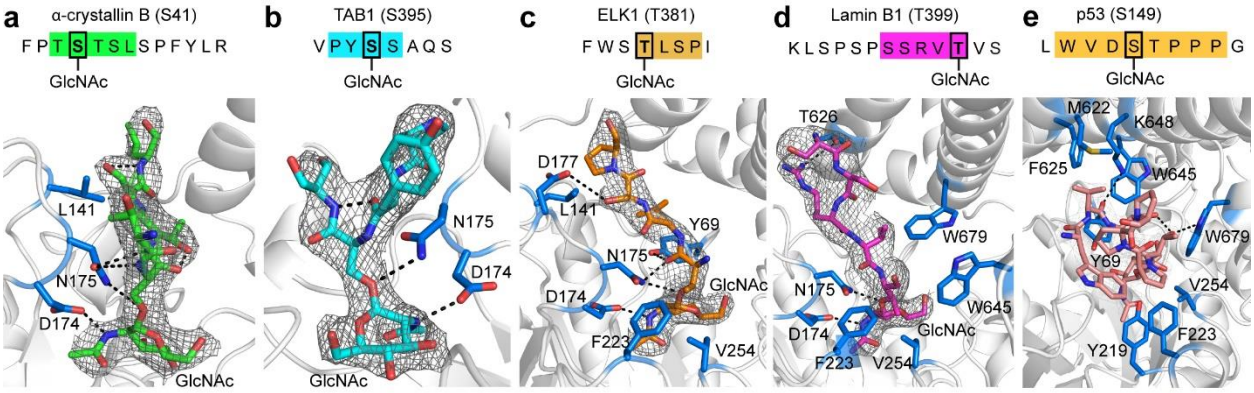


Figure 4.10 OGA formed specific interactions with distinct peptides. The sequences, conformations, and representative 2Fo–Fc electron density maps (gray) of four glycopeptides bound in the substrate-binding cleft of OGA, contoured at 1.0 σ. O-GlcNAcylated peptides: a α-crystallin B chain, b TAB1, c ELK1, and d Lamin B1. On the top of each panel, the glycopeptide sequence is displayed. The peptide residues observed in the crystal structure are highlighted with colored background and the O-GlcNAcylation site is highlighted by a black box. At the bottom of each panel, the binding conformation of each peptide is shown

in sticks with the same color as its highlighted sequence. The residues of OGA participating in the interactions with each peptide are shown in marine blue sticks and labeled with residue numbers. Hydrogen bonds are displayed as dashed lines

4.4 Summary and future plan

Here we present the first crystal structure of human OGA, in its apo form, in complex with an inhibitor and in complex with five different glycopeptides. The structure not only was the first crystal structure of human OGA, but also served as the only available structure of eukaryotic GH84 family glycosidase. Human OGA was found to be an unusual arm-in-arm homodimer where the substrate binding cleft was formed by the catalytic domain of one monomer and the stalk domain of the other. The dimerization of human OGA was absent in any of its bacterial homologues, providing the distinct binding and recognition mode of human OGA.

Based on the complex structure of human OGA with an inhibitor, thiamet-G, we confirmed the proposed mechanism of GlcNAc removal, and revealed novel residues that were important for the activity of OGA. The information could prove important for development of inhibitors to facilitate the study of OGA function, and to serve as a potential treatment for neurodegenerative diseases.

We also solved the structures of OGA complexed with a few glycopeptide substrates. Our structures demonstrated that different peptides were located in the substrate binding cleft in a conserved while bi-directional manner. Detailed investigation into the binding mode for each peptide showed that OGA was able to bind different substrate with distinct interaction pattern. It indicated that OGA could recognize its substrates through the specific interactions, and it could play an important role in the regulation of O-GlcNAc biology.

4.5 References

1. Keembiyehetty, C. Disruption of O-GlcNAc cycling by deletion of O-GlcNAcase (Oga/Mgea5) changed gene expression pattern in mouse embryonic fibroblast (MEF) cells. *Genomics Data* **5**, 30–33 (2015).

2. Shen, D. L., Gloster, T. M., Yuzwa, S. A. & Vocadlo, D. J. Insights into O-linked N-acetylglucosamine ([0-9]O-GlcNAc) processing and dynamics through kinetic analysis of O-GlcNAc transferase and O-GlcNAcase activity on protein substrates. *J. Biol. Chem.* **287**, 15395–15408 (2012).

3. Schimpl, M., Schüttelkopf, A. W., Borodkin, V. S. & van Aalten, D. M. F. Human OGA binds substrates in a conserved peptide recognition groove. *Biochem. J.* **432**, 1–12 (2010).

4. Worth, M., Li, H. & Jiang, J. Deciphering the Functions of Protein *O* -GlcNAcylation with Chemistry. *ACS Chem. Biol.* **12**, 326–335 (2017).

5. Gong, C.-X. & Iqbal, K. Hyperphosphorylation of microtubule-associated protein tau: a promising therapeutic target for Alzheimer disease. *Curr. Med. Chem.* **15**, 2321–2328 (2008).

6. Liu, F., Iqbal, K., Grundke-Iqbal, I., Hart, G. W. & Gong, C.-X. O-GlcNAcylation regulates phosphorylation of tau: a mechanism involved in Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 10804–10809 (2004).

7. Yuzwa, S. A. *et al.* A potent mechanism-inspired O-GlcNAcase inhibitor that blocks phosphorylation of tau in vivo. *Nat. Chem. Biol.* **4**, 483–490 (2008).

8. Otwinowski, Z. & Minor, W. [20] Processing of X-ray diffraction data collected in oscillation mode. in *Methods in Enzymology* vol. 276 307–326 (Elsevier, 1997).

9. Emsley, P. & Cowtan, K. *Coot* : model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

10. Adams, P. D. *et al. PHENIX* : building new software for automated crystallographic structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **58**, 1948–1954 (2002).

11. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **50**, 760–763 (1994).

12. *The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.*

13. Laue, T. M. [19] Sedimentation equilibrium as thermodynamic tool. in *Methods in Enzymology* vol. 259 427–452 (Elsevier, 1995).

14. Xie, D., Suvorov, L., Erickson, J. W. & Gulnik, S. Real-time measurements of dark substrate catalysis. *Protein Sci.* **8**, 2460–2464 (2008).

15. Gao, Y., Wells, L., Comer, F. I., Parker, G. J. & Hart, G. W. Dynamic O-glycosylation of nuclear and cytosolic proteins: cloning and characterization of a neutral, cytosolic beta-N-acetylglucosaminidase from human brain. *J. Biol. Chem.* **276**, 9838–9845 (2001).

16. Ishida, T. & Kinoshita, K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* **35**, W460-464 (2007).

17. Li, B., Li, H., Lu, L. & Jiang, J. Structures of human O-GlcNAcase and its complexes reveal a new substrate recognition mode. *Nat. Struct. Mol. Biol.* **24**, 362–369 (2017).

18. Elsen, N. L. *et al.* Insights into activity and inhibition from the crystal structure of human O-GlcNAcase. *Nat. Chem. Biol.* **13**, 613–615 (2017).

19. Roth, C. *et al.* Structural and functional insight into human O-GlcNAcase. *Nat. Chem. Biol.* **13**, 610–612 (2017).

20. Krissinel, E. Crystal contacts as nature's docking solutions. *J. Comput. Chem.* **31**, 133–143 (2010).

21. Gloster, T. M. & Vocadlo, D. J. Mechanism, Structure, and Inhibition of O-GlcNAc Processing Enzymes. *Curr. Signal Transduct. Ther.* **5**, 74–91 (2010).

22. Brady, C. A. & Attardi, L. D. p53 at a glance. *J. Cell Sci.* **123**, 2527–2532 (2010).

23. Yang, W. H. *et al.* Modification of p53 with O-linked N-acetylglucosamine regulates p53 activity and stability. *Nat. Cell Biol.* **8**, 1074–1083 (2006).

4.6 Supplementary information

Table 4.1 Data collection and refinement statistics (molecular replacement) for OGAcryst apo form, OGAcryst-thiamet-G complex and OGAcryst-D175N-p53 complex

	OGA_{cryst} (PDB 5TKE)	**OGA_{cryst}–thiamet-G (PDB 5UN9)**	**OGA_{cryst} (D175N)–p53 (PDB 5UN8)**
Data collection			
Space group	*P*21	*P*21	*P*21
Cell dimensions			
a, *b*, *c* (Å)	82.0, 96.6, 89.3	82.5, 96.1, 89.5	89.9, 95.4, 149.3
α, *β*, *γ* (°)	90.0, 115.0, 90.0	90.0, 115.0, 90.0	90.0, 96.9, 90.0
Resolution (Å)	50.0-2.5 (2.59-2.50)[a]	30.0-2.5 (2.54-2.50)	50.0-2.1 (2.18-2.14)
R_{merge}	8.7 (56.5)	6.3 (58.3)	8.2 (72.0)
$I/\sigma(I)$	19.5 (2.0)	23.0 (2.1)	18.4 (2.0)
$CC_{1/2}$	99.9 (79.1)	99.9 (78.5)	99.8 (78.1)
Completeness (%)	100.0 (100.0)	99.5 (95.1)	99.9 (100.0)
Redundancy	6.4 (6.2)	5.3 (5.2)	4.9 (4.5)
Refinement			
Resolution (Å)	50.0-2.5	30.0-2.5	45.0-2.1
No. reflections	44,255	44,310	139,265
R_{work} / R_{free}	18.2 / 23.7	21.0 / 26.0	18.4 / 22.9
No. atoms			
Protein	6,921	6,922	14,015
Ligand/peptide		32	305
Water	396	238	1247
B factors	43.40	43.38	30.02
Protein	43.41	43.49	29.28
Ligand/peptide		29.73	29.07
Water	42.68	40.93	38.56
R.m.s. deviations			
Bond lengths (Å)	0.008	0.011	0.009
Bond angles (°)	0.970	1.211	1.016

Table 4.2 Data collection and refinement statistics (molecular replacement) for OGAcryst-D175N apo form, OGAcryst-D175N in complex with four different peptide substrates

	OGA_{cryst} (D175N) **(PDB AAAA)**	**OGA_{cryst} (D175N)-□–crystallin B (PDB BBBB)**	**OGA_{cryst} (D175N)–TAB1 (PDB CCCC)**	**OGA_{cryst} (D175N)–ELK1 (PDB DDDD)**	**OGA_{cryst} (D175N)–Lamin B1 (PDB EEEE)**
Data collection					
Space group	*P*21	*P*21	*P*21	*P*21	*P*21
Cell dimensions					
a, *b*, *c* (Å)	82.4, 96.8, 89.1	83.1, 96.3, 89.8	82.5, 96.1, 88.9	82.4, 95.8, 89.3	82.9, 96.2, 89.7
α, *β*, *γ* (°)	90.0, 115.2, 90.0	90.0, 114.3, 90	90.0, 114.5, 90.0	90.0, 114.5, 90.0	90.0, 114.6, 90.0
Resolution (Å)	50.0–2.6 (2.64–2.60)[a]	50.0–2.7 (2.80–2.70)	50.0–2.7 (2.75–2.70)	50.0–2.8 (2.85–2.80)	50.0–2.9 (2.95–2.90)
R_{merge}	8.4 (63.1)	7.1 (93.5)	8.0 (68.0)	8.4 (72.2)	8.9 (69.0)
$I/□(I)$	21.9 (2.0)	18.8 (1.5)	22 (1.6)	17.8 (1.8)	20.4 (1.8)
$CC_{1/2}$	99.9 (83.3)	98.6 (58.7)	99.0 (70.7)	99.3 (73.1)	99.7 (74.5)
Completeness (%)	99.3 (93.3)	99.8 (98.6)	99.8 (97.3)	99.9 (99.7)	99.8 (98.9)
Redundancy	6.3 (5.7)	5.6 (5.5)	4.4 (3.5)	4.2 (4.1)	4.6 (4.5)
Refinement					
Resolution (Å)	50.0–2.6	50.0–2.8	50.0–2.7	50.0–2.8	50.0–2.9
No. reflections	38,896	31,917	34,865	31,269	28,990
R_{work} / R_{free}	21.0 / 26.0	19.6 / 25.6	20.7 / 27.5	21.1 / 29.1	20.0 / 28.3
No. atoms					
Protein	7,141	6,905	7,000	6,920	6,994
Ligand/peptide	0	75	90	71	96
Water	133	35	50	49	13
B factors	60.27	74.54	68.32	70.93	90.85
Protein	60.51	74.45	68.05	70.89	90.41
Ligand/peptide		85.10	88.49	81.79	125.59
Water	47.62	58.03	62.55	56.64	68.86
R.m.s. deviation					
Bond length (Å)	0.012	0.013	0.012	0.012	0.012
Bond angle (°)	1.622	1.706	1.592	1.636	1.536

Chapter 5 Conclusion and future perspectives

In an effort to investigate the substrate recognition of OGT, we developed an electrophilic probe (GEP) which could both label OGT and its substrate. Based on the ordered bi-bi mechanism of OGT, we applied the fluorescent probe to generate a fast readout of the relative ratio of **GEP1A** labeled OGT and protein substrate for OGT mutants, in order to investigate the importance of individual residues in protein and sugar binding. The **GEP1A** assay successfully generated accurate readout for known mutants of OGT with impaired sugar and protein binding. Moreover, we utilized the **GEP1A** assay to identify two residues in the TPR domain, N321 and N322, that are important for the binding and recognition of OGT substrates. **GEP1A** assay is a valuable tool that can efficiently identify OGT residues and structural features responsible for substrate recognition, which is not only significant for the investigation into mechanisms and functions of O-GlcNAcylation, but also provides targets for potential therapeutics for the many diseases O-GlcNAcylation has been linked to including cancer and diabetes.

Additionally, **GEP1** was also found to crosslink OGT and its substrates in certain conditions. The probe can overcome the challenges in OGT substrate detection and serve as a useful tool for identification of novel substrates of OGT.

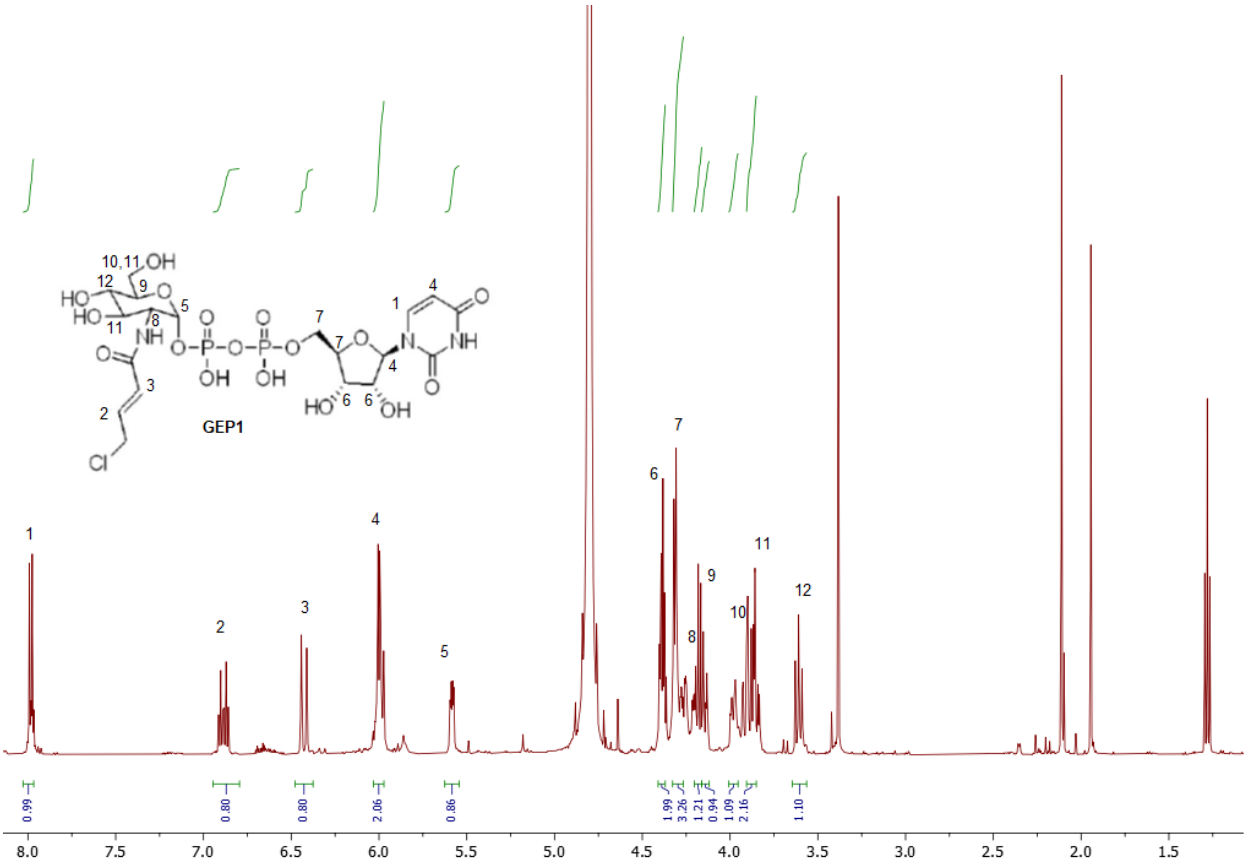
Adopting the same idea of GEPs, we developed a panel of targeted covalent inhibitors, **ES1**-**ES4**, for OGT by introducing a sulfur in the pyranose ring of GlcNAc to slow down the sugar transfer rate. **ES1** inhibited OGT with better potency compared to one of the best inhibitors, 5SGlcNAc, with low toxicity and good selectivity over other types of glycosylation. **ES1** served as the first reported target covalent inhibitor of OGT and the design rationale could provide a valuable tool for the development of better inhibitors.

Concurrently, we solved the first crystal structure of human OGA in its apo form, and it complexed with an inhibitor and select glycopeptide substrates. Human OGA served as a unique arm-in-arm homodimer and the substrate binding cleft was formed by the catalytic domain of one monomer and the stalk domain of the other. Based on the OGA-inhibitor crystal structure, the substrate-assisted catalytic mechanism was

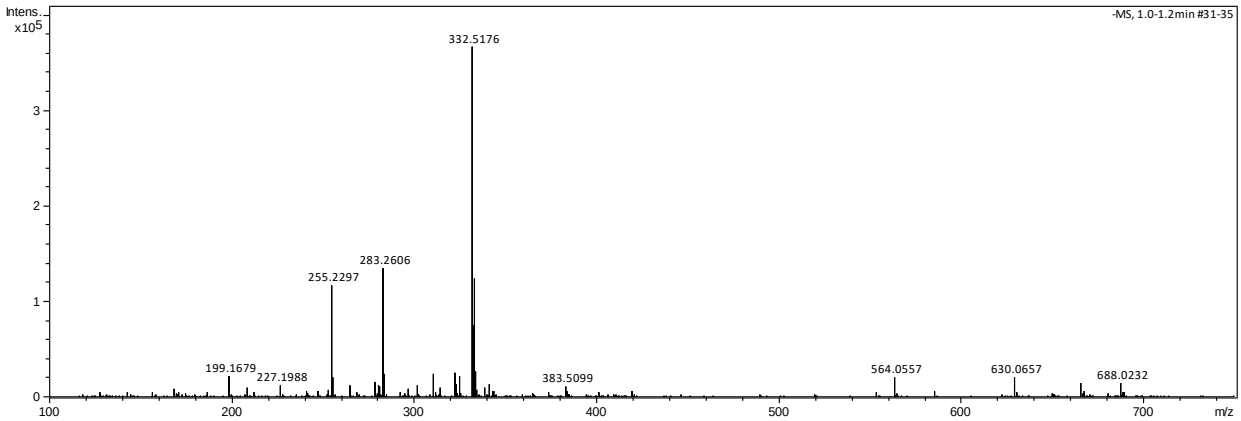
verified by the identification of the important residues involved in inhibitor binding. Moreover, we spotted novel residues important for binding the GlcNAc sugar, which could facilitate the design of more potent OGA inhibitors. Last but not least, we obtained the structures of OGA in complex with five distinct glycopeptide substrates. OGA binds its substrates in a conserved and bi-directional conformation. At the same time, by comparing the complex structures, we discovered that OGA was able to form specific interactions with different peptide substrates, supporting the hypothesis that OGA is capable of substrate discrimination, thus playing a significant role in O-GlcNAc biology. In summary, we proposed the substrate recognition mode of OGA, wherein the dominant interactions between OGA active site and GlcNAc moiety secure the binding of OGA and its substrates. In addition to the binding with sugar, OGA also forms substrate specific interactions to achieve enhanced selectivity and recognition. Lastly, the peptide can be stabilized in OGA substrate binding cleft through the intra-molecular interactions to maintain its ordered binding conformation.

Chapter 6 Supplementary figures

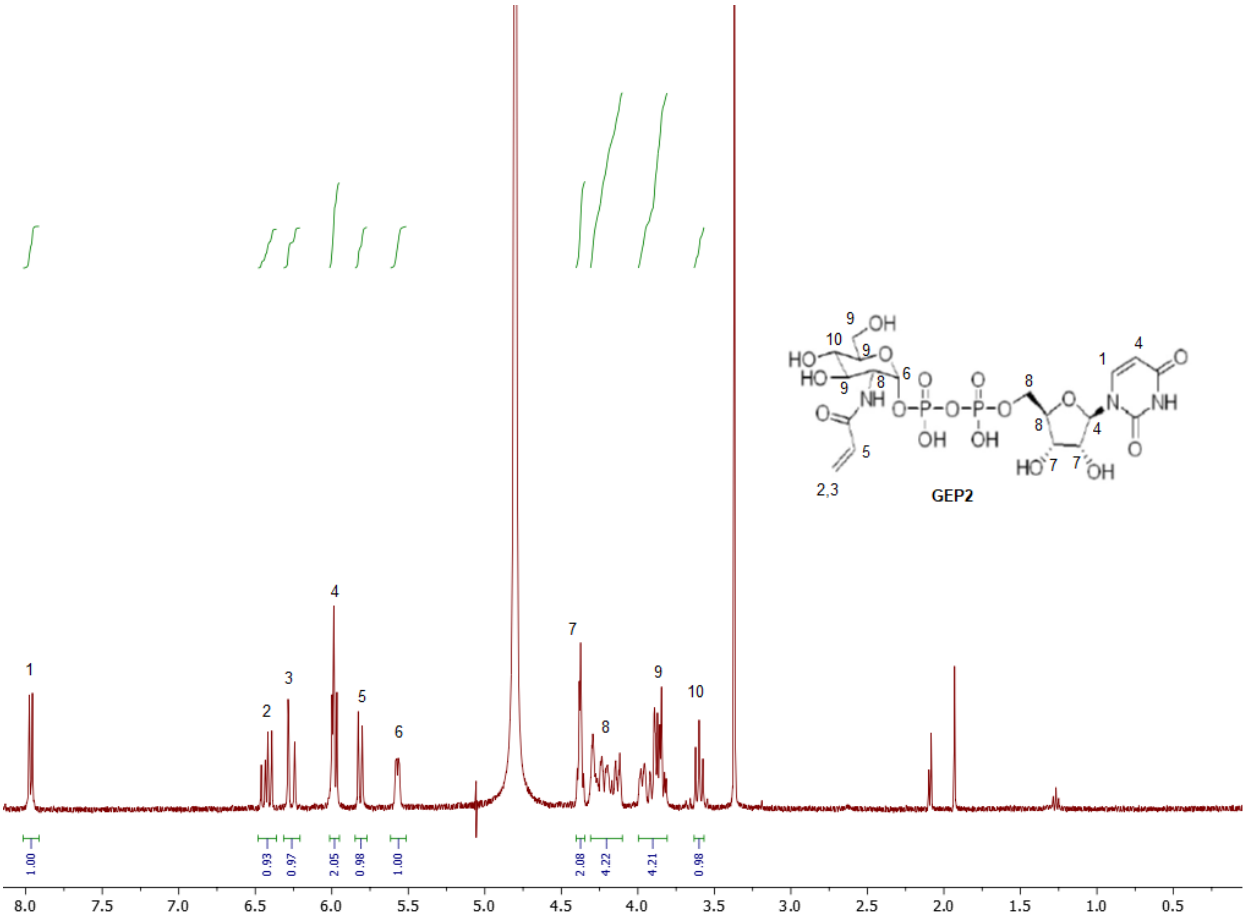
^{1}H NMR spectrum for **GEP1**:



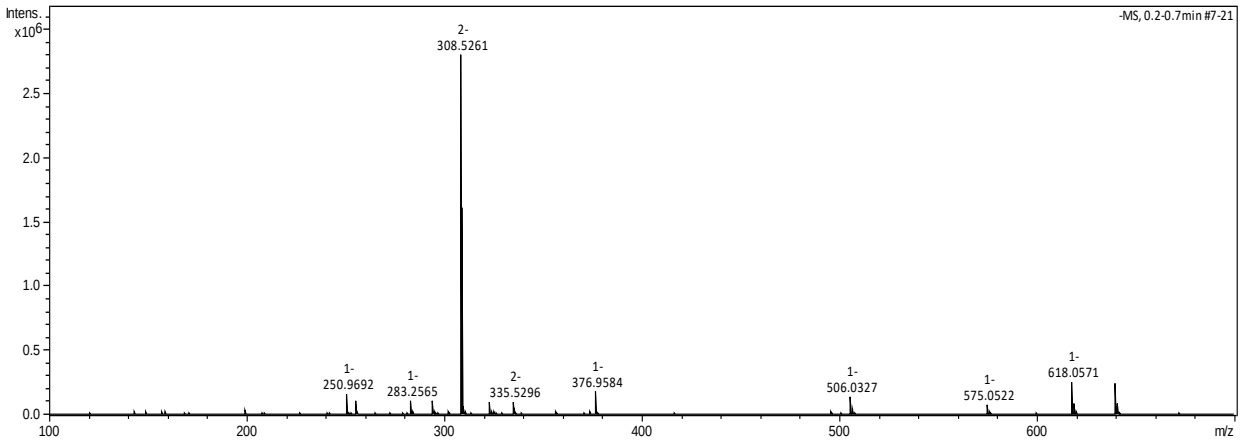
MS spectrum of **GEP1**. **GEP1** $[M]^{2-}$ = 332.522



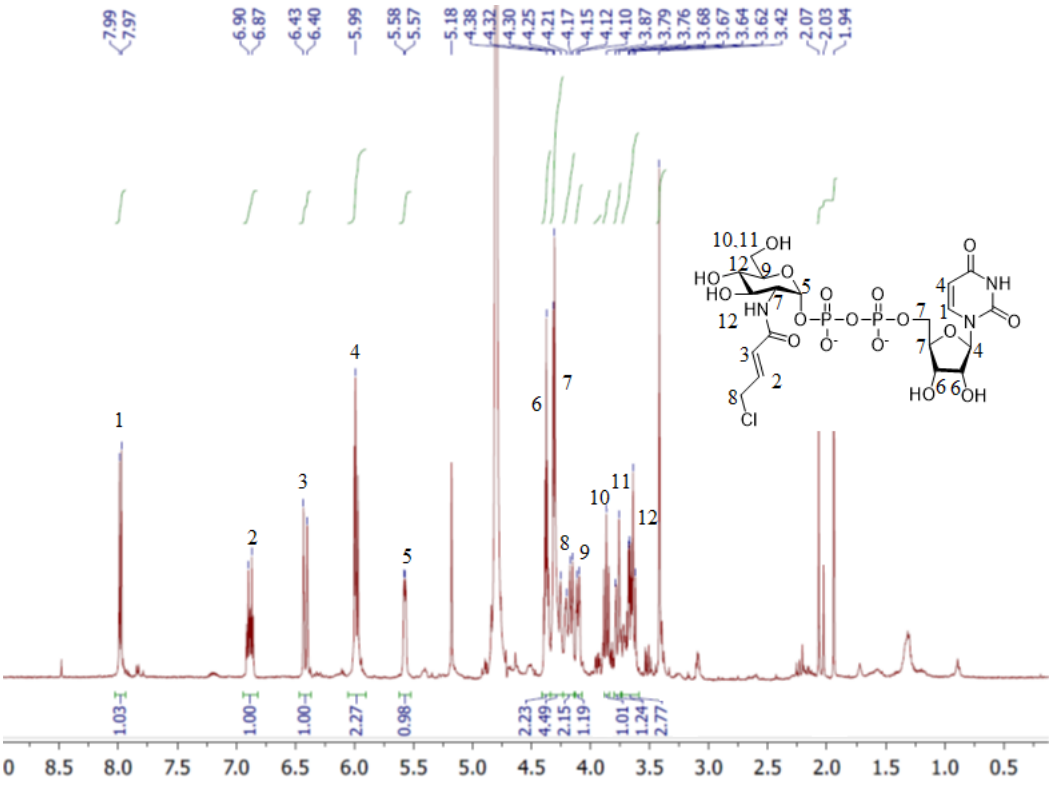
^{1}H NMR spectrum for **GEP2**:



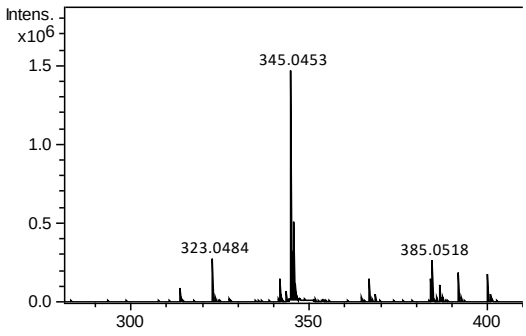
MS spectrum of **GEP2**. **GEP2** $[M]^{2-}$ = 308.534



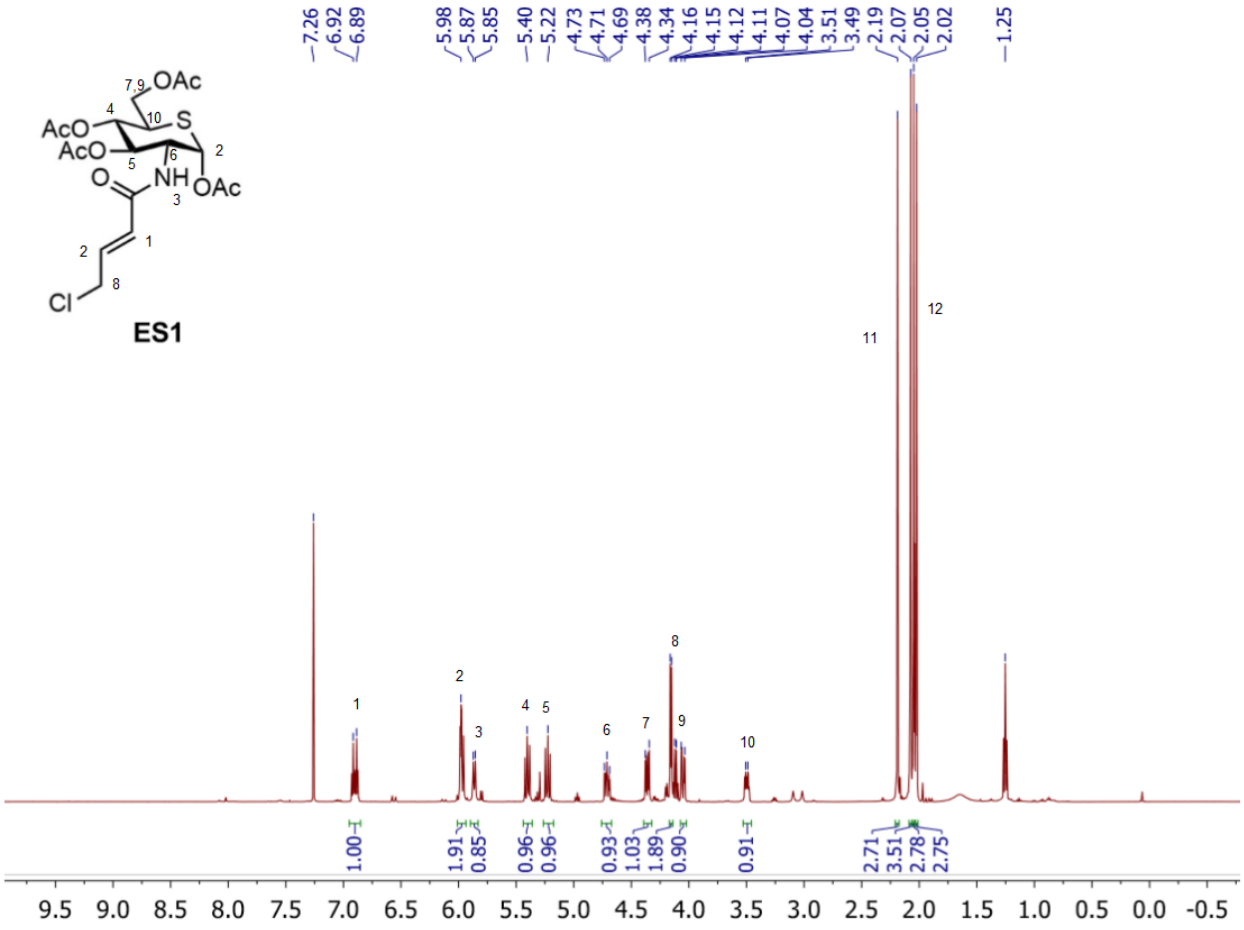
^{1}H NMR spectrum for **GEP1A**:



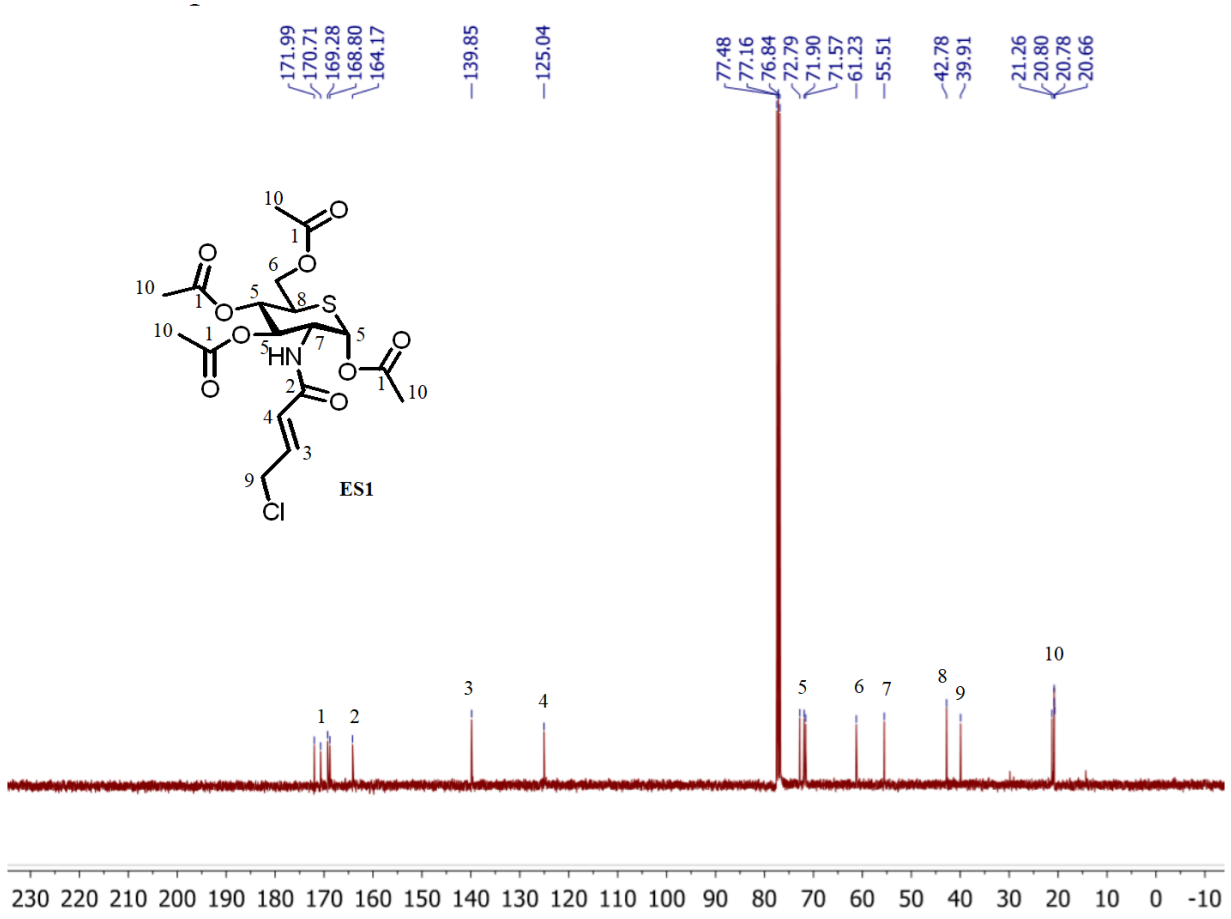
MS spectrum of **GEP1A**. **GEP1A** $[M]^{2-}$ = 345.0251



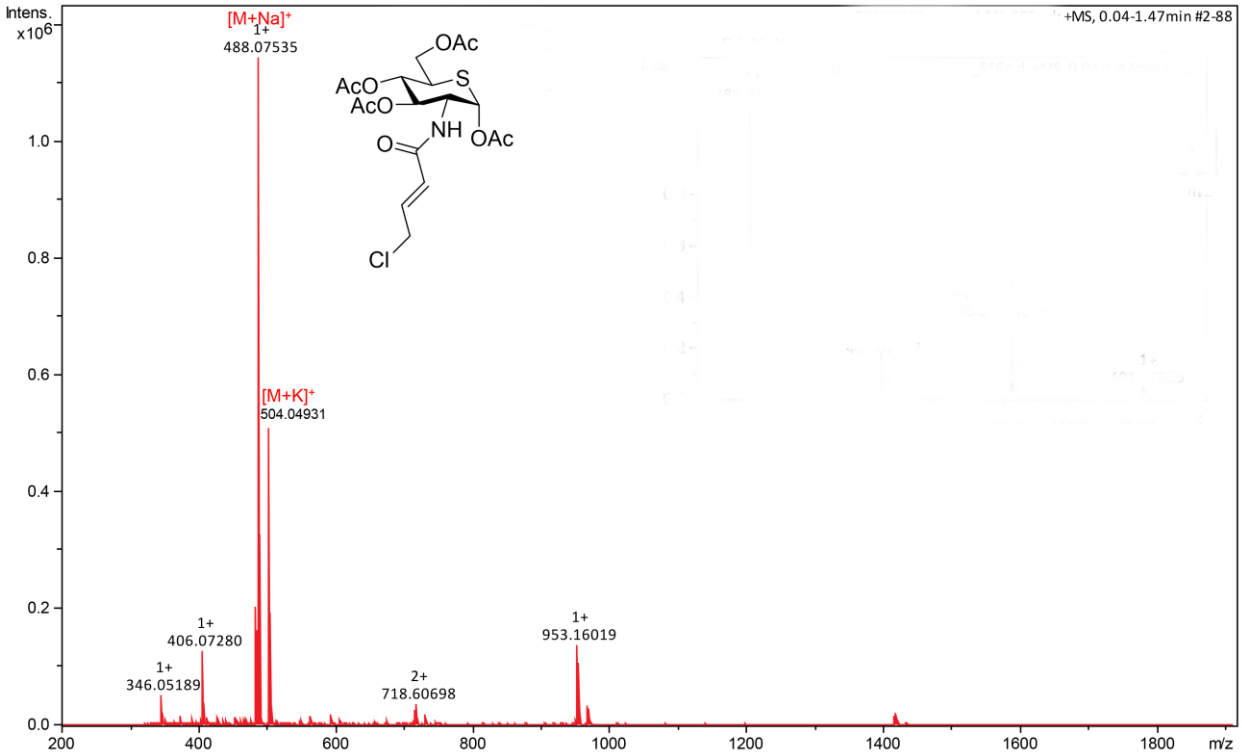
^{1}H NMR spectrum for **ES1**:



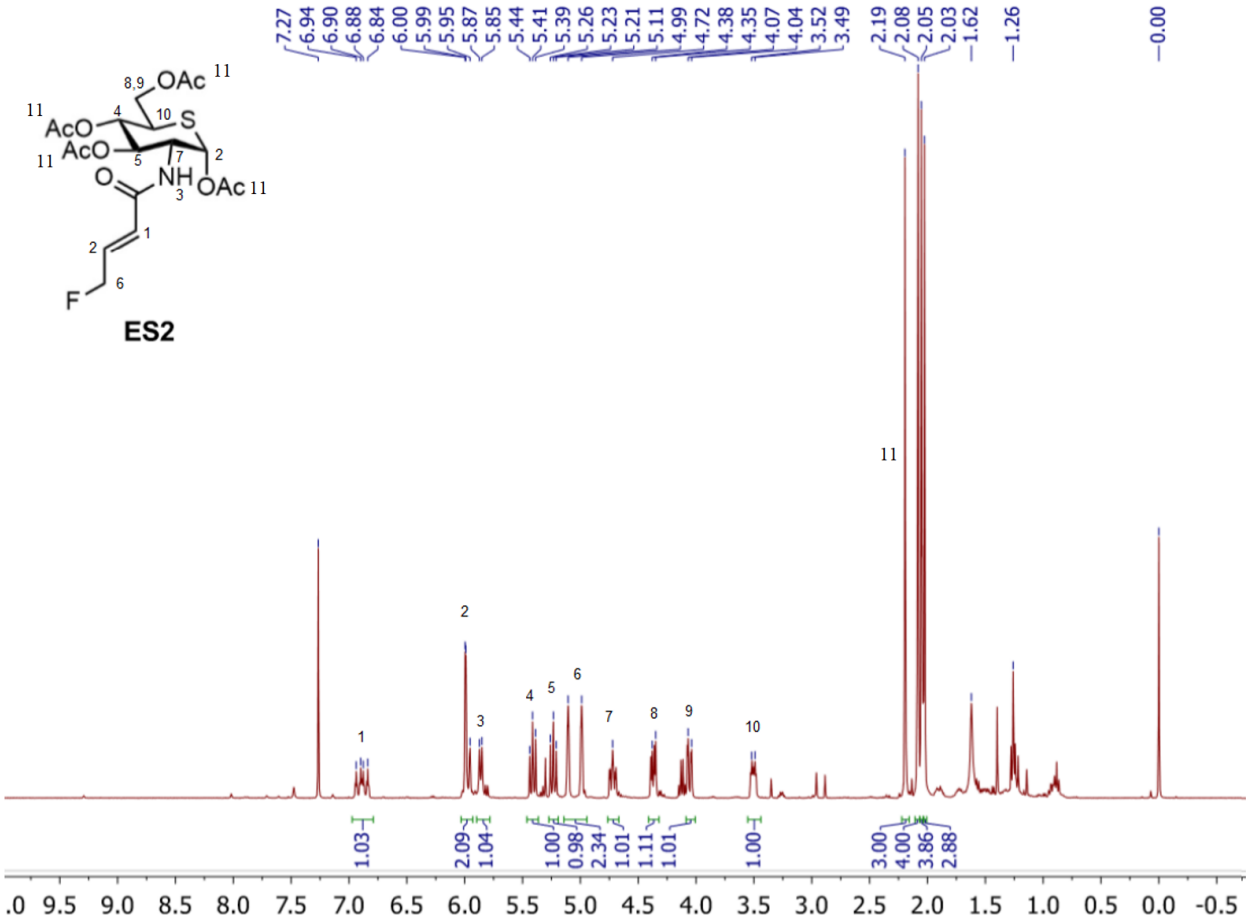
^{13}C NMR spectrum for **ES1**:



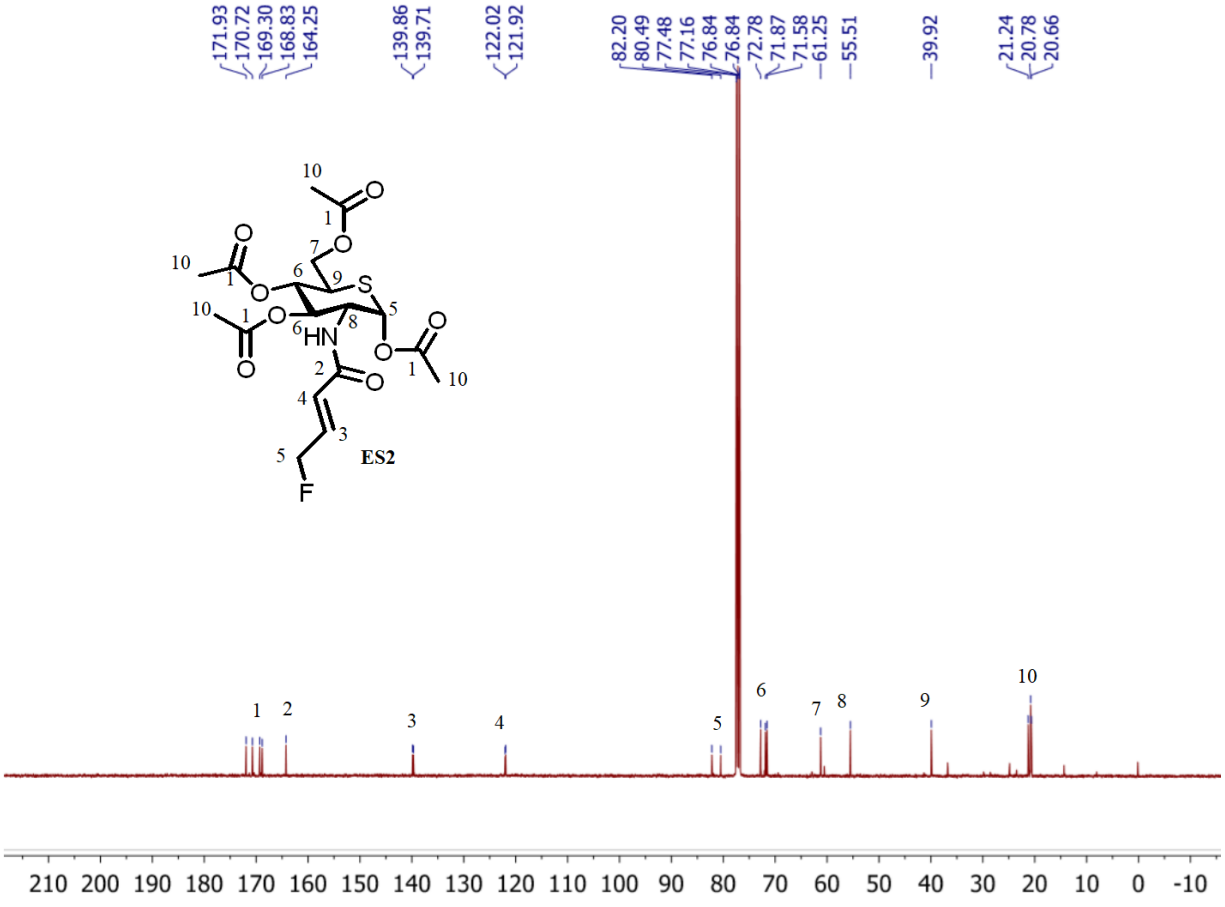
MS spectrum of **ES1**. **ES1** $[M+Na]^+$ = 488.0758; $[M+K]^+$ = 504.0497



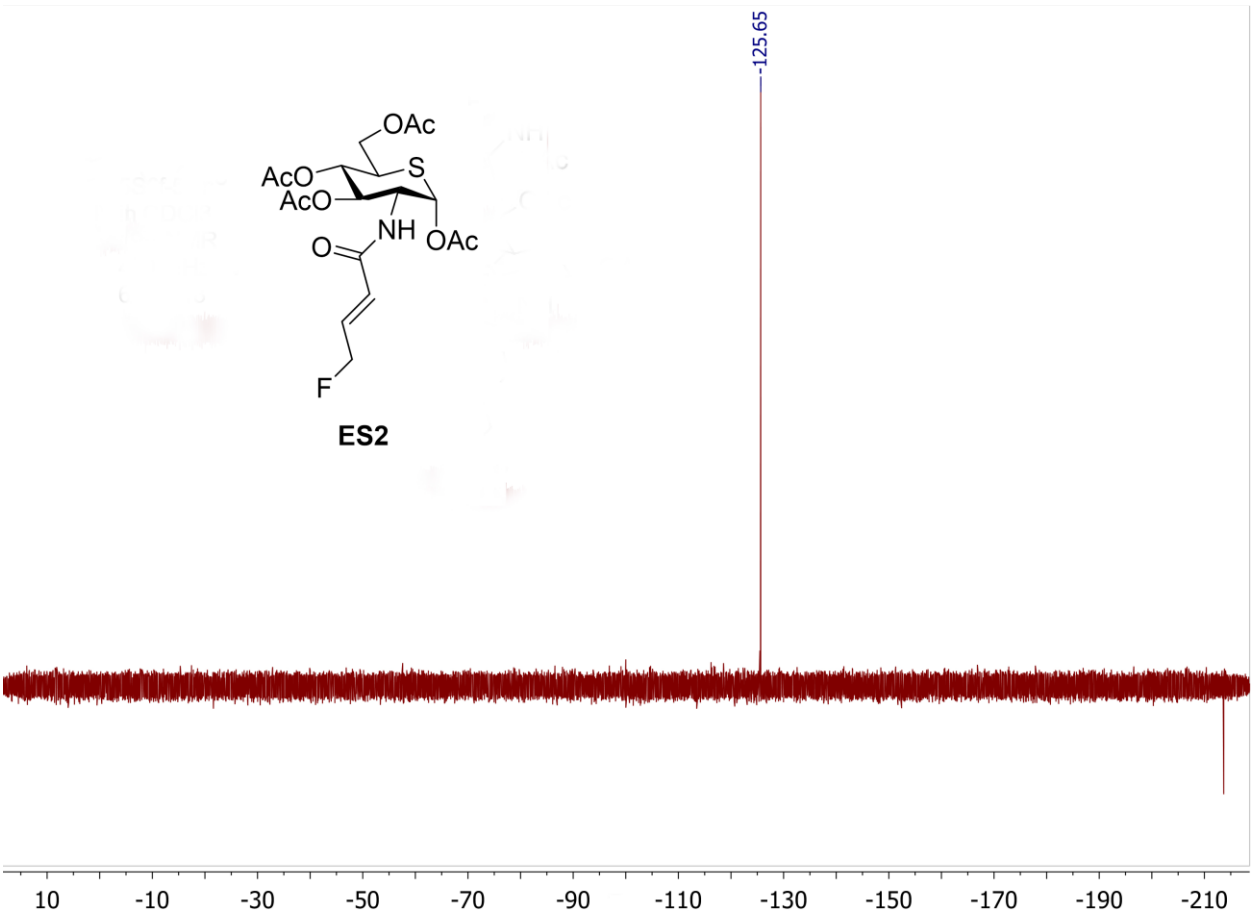
^{1}H NMR spectrum for **ES2**:



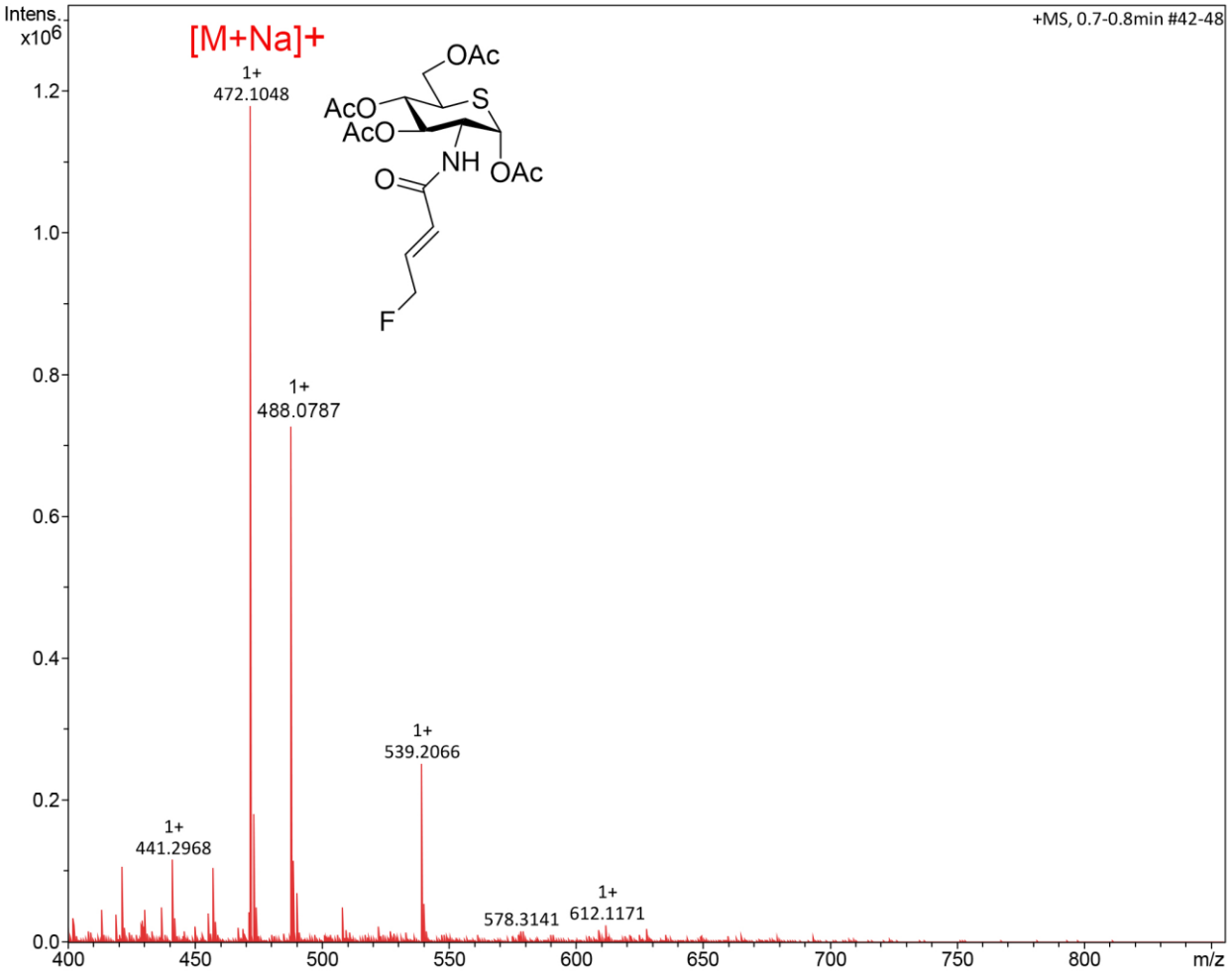
^{13}C NMR spectrum for **ES2**:



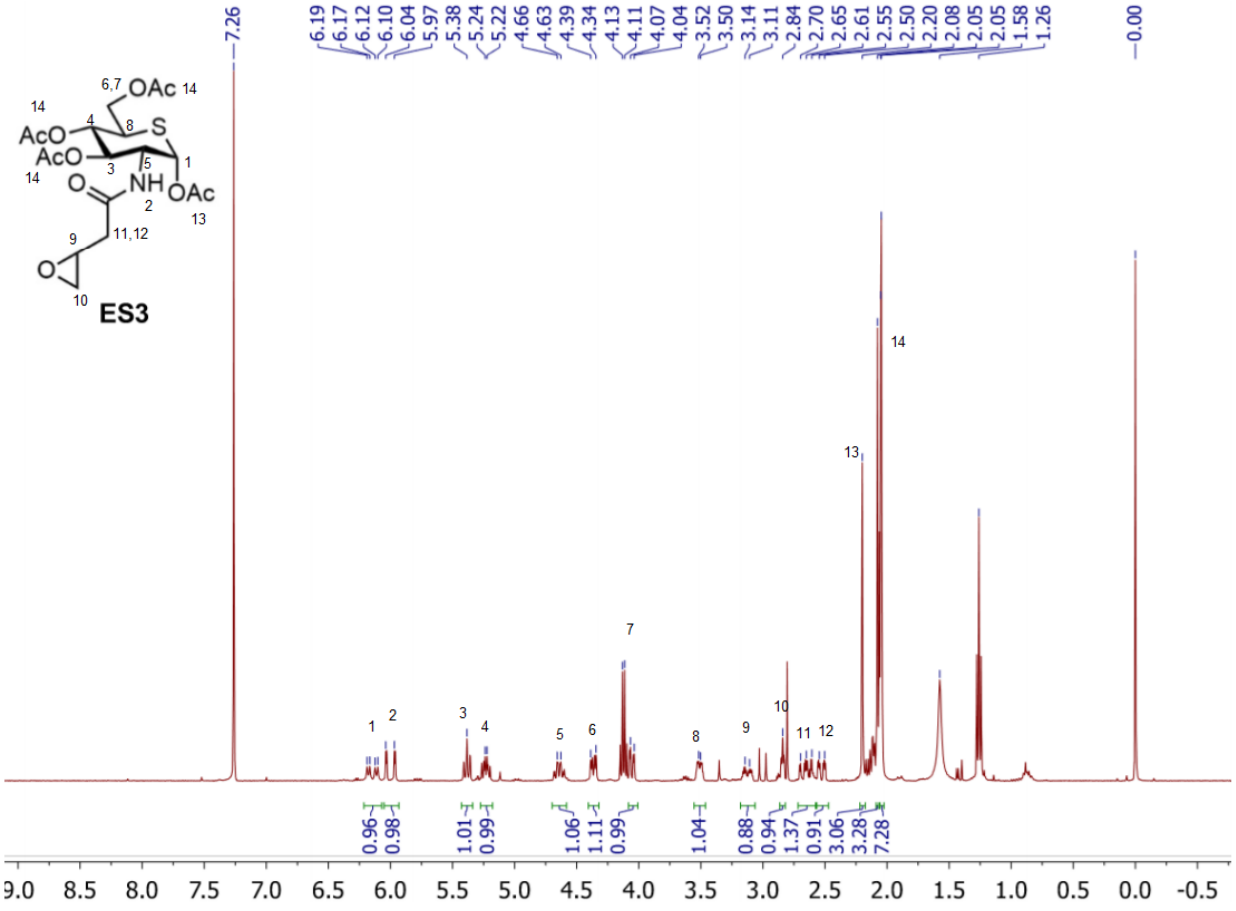
^{19}F NMR spectrum for **ES2**:



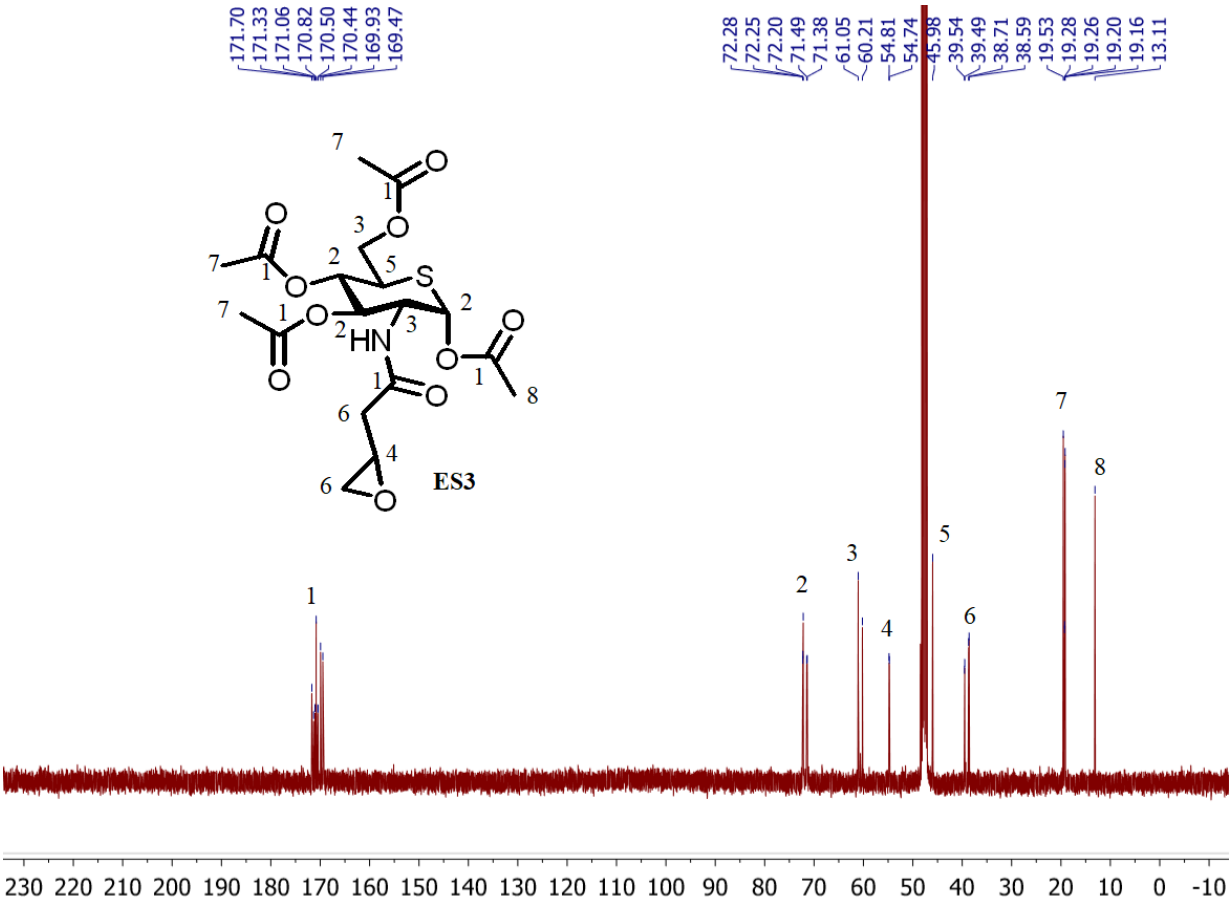
MS spectrum of **ES2**. **ES2** $[M+Na]^+$ = 472.1053; $[M+K]^+$ = 488.0793



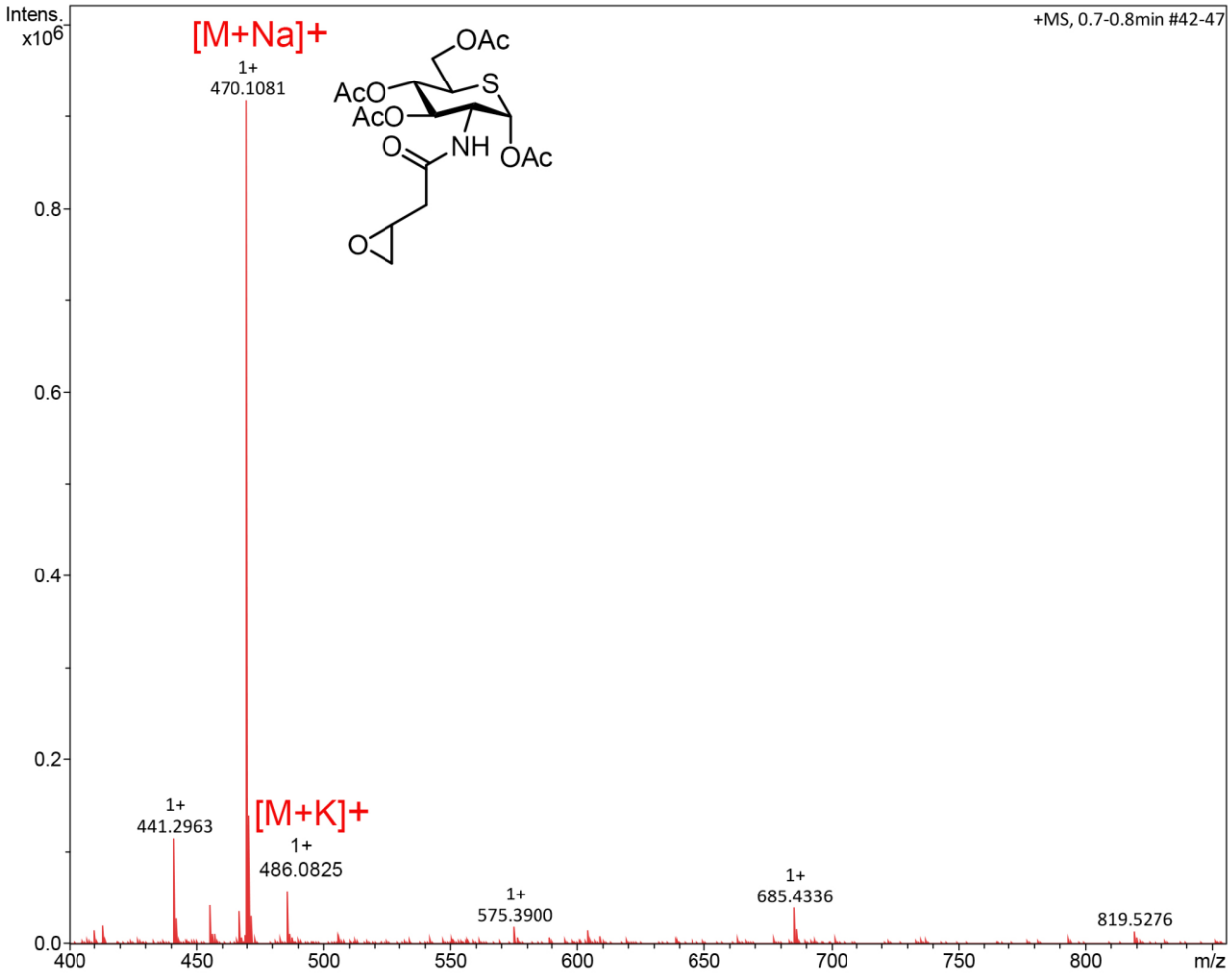
^{1}H NMR spectrum for **ES3**:



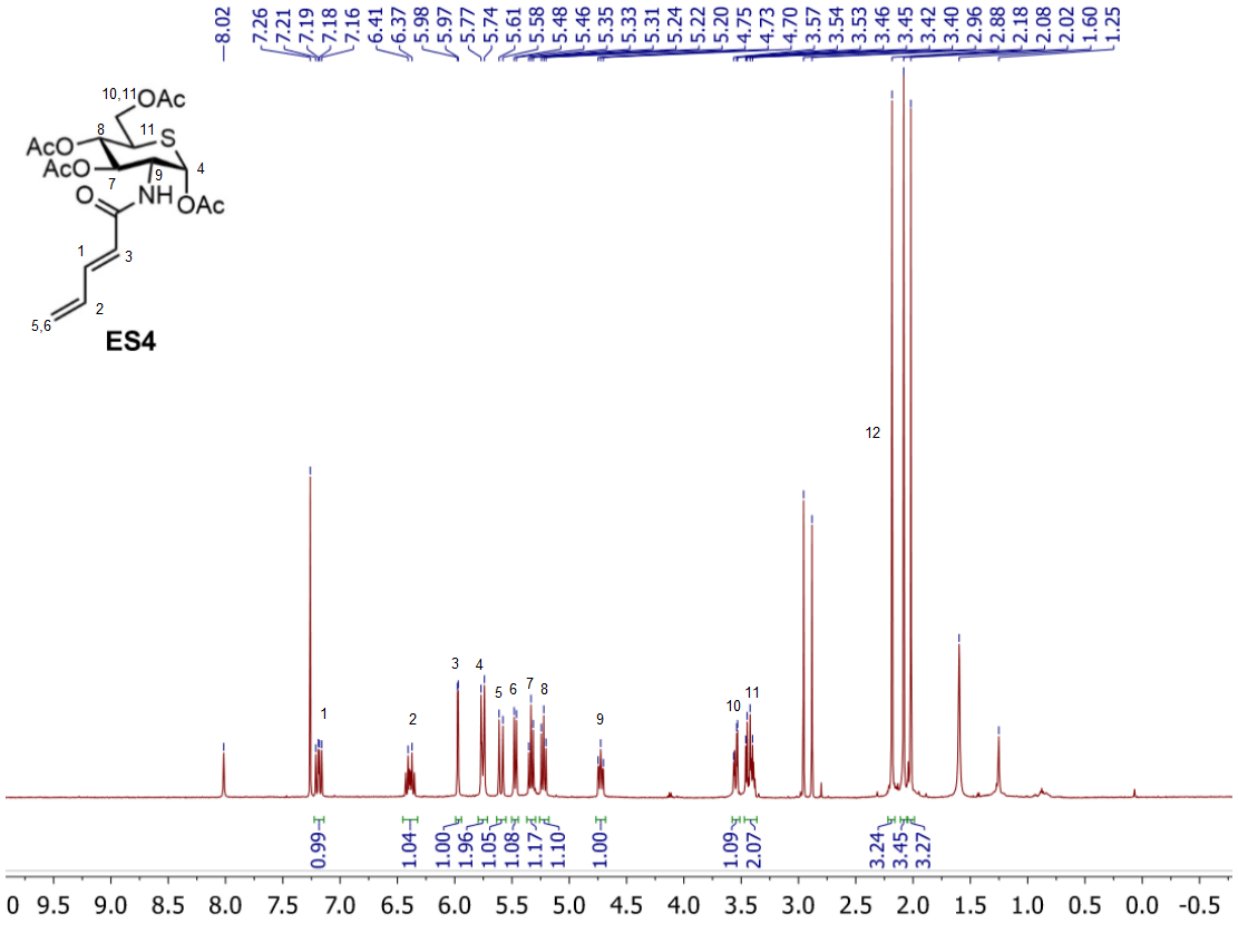
^{13}C NMR spectrum for **ES3**:



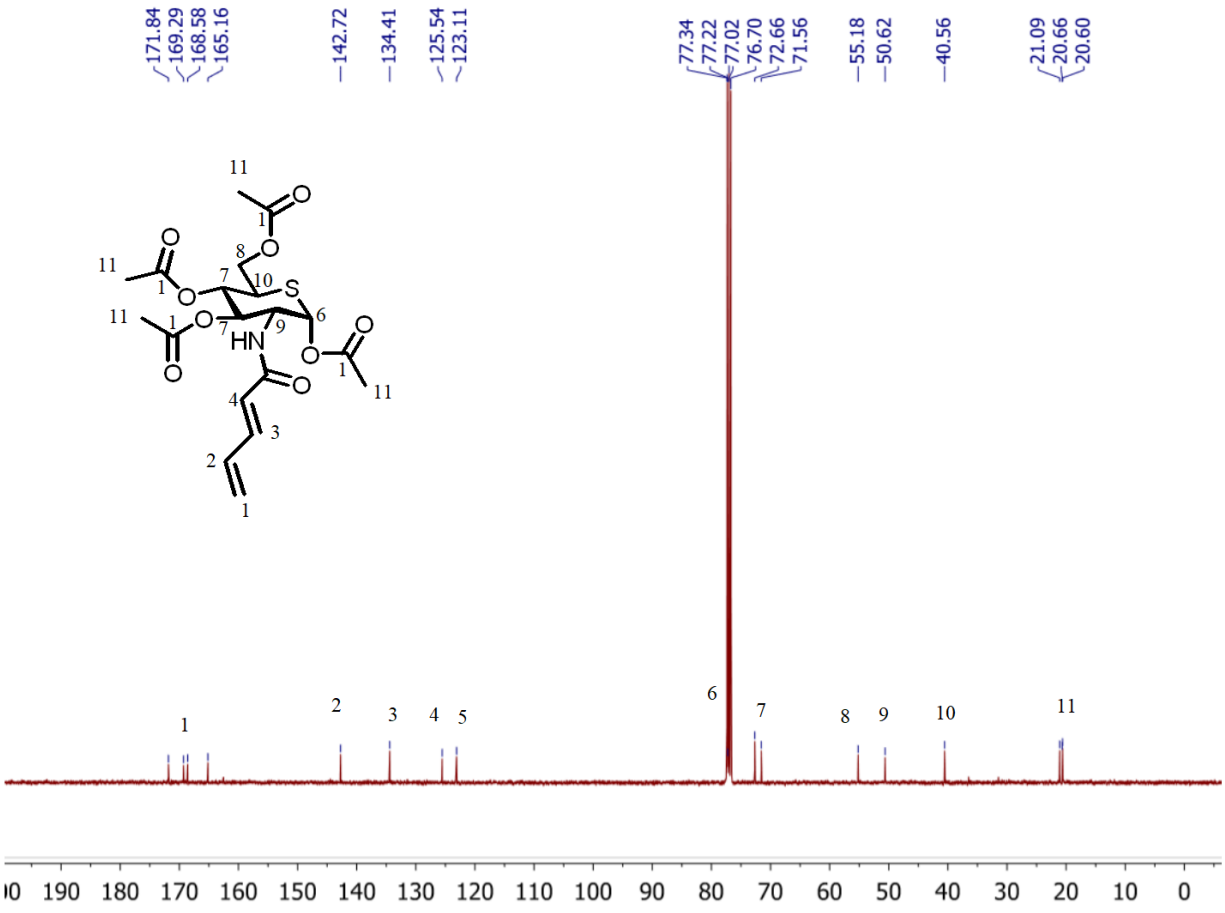
MS spectrum of **ES3**. **ES3** $[M+Na]^+$ = 470.1081; $[M+K]^+$ = 486.0836



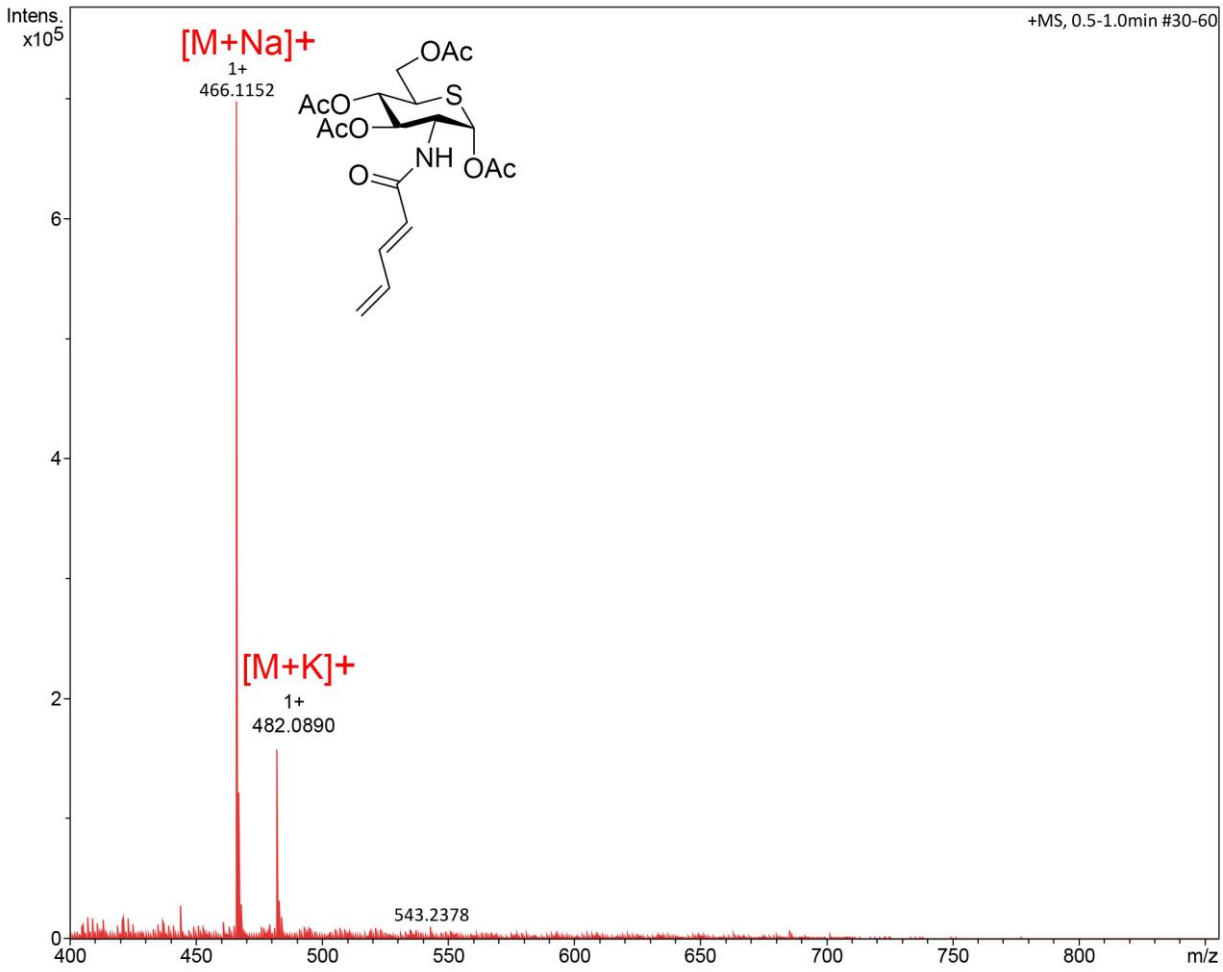
^{1}H NMR spectrum for **ES4**:



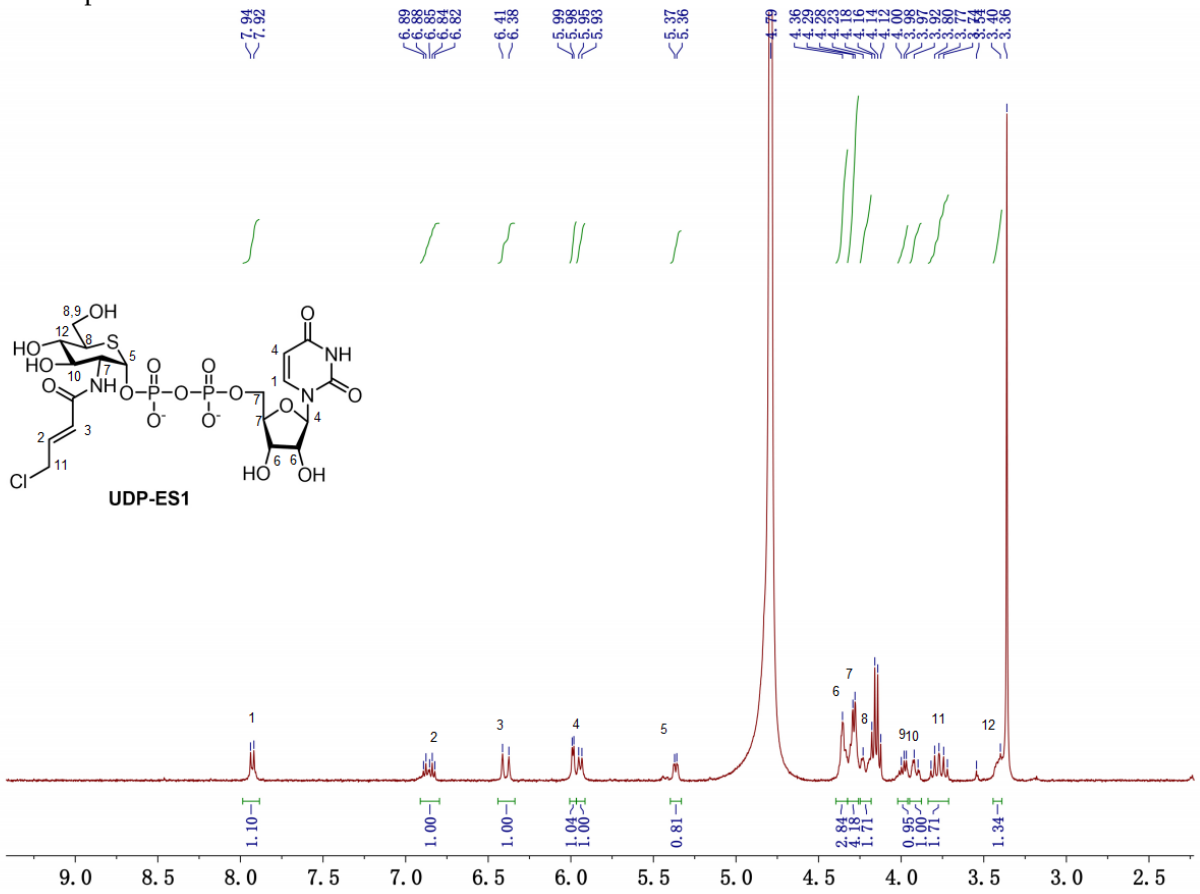
^{13}C NMR spectrum for **ES4**:



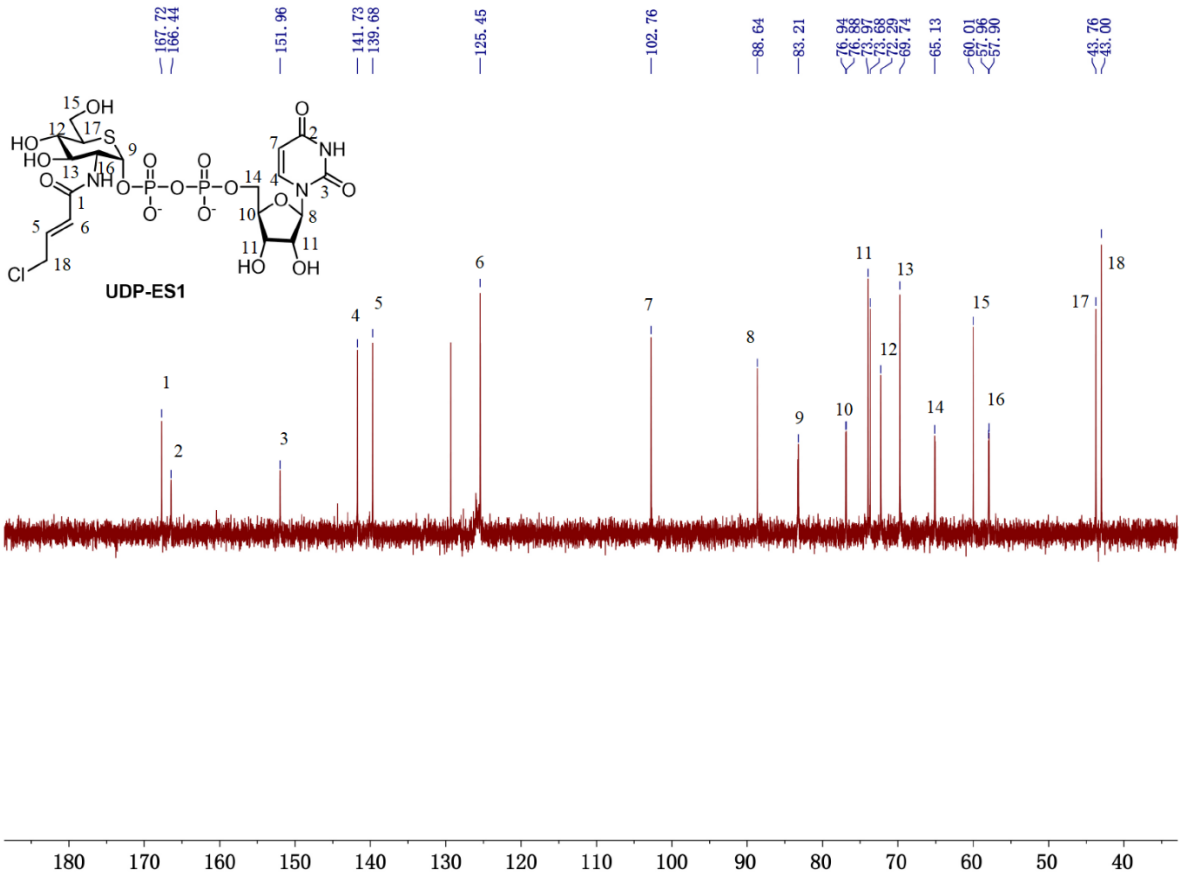
MS spectrum of **ES4**. **ES4** $[M+Na]^+$ = 466.1148; $[M+K]^+$ = 482.0887



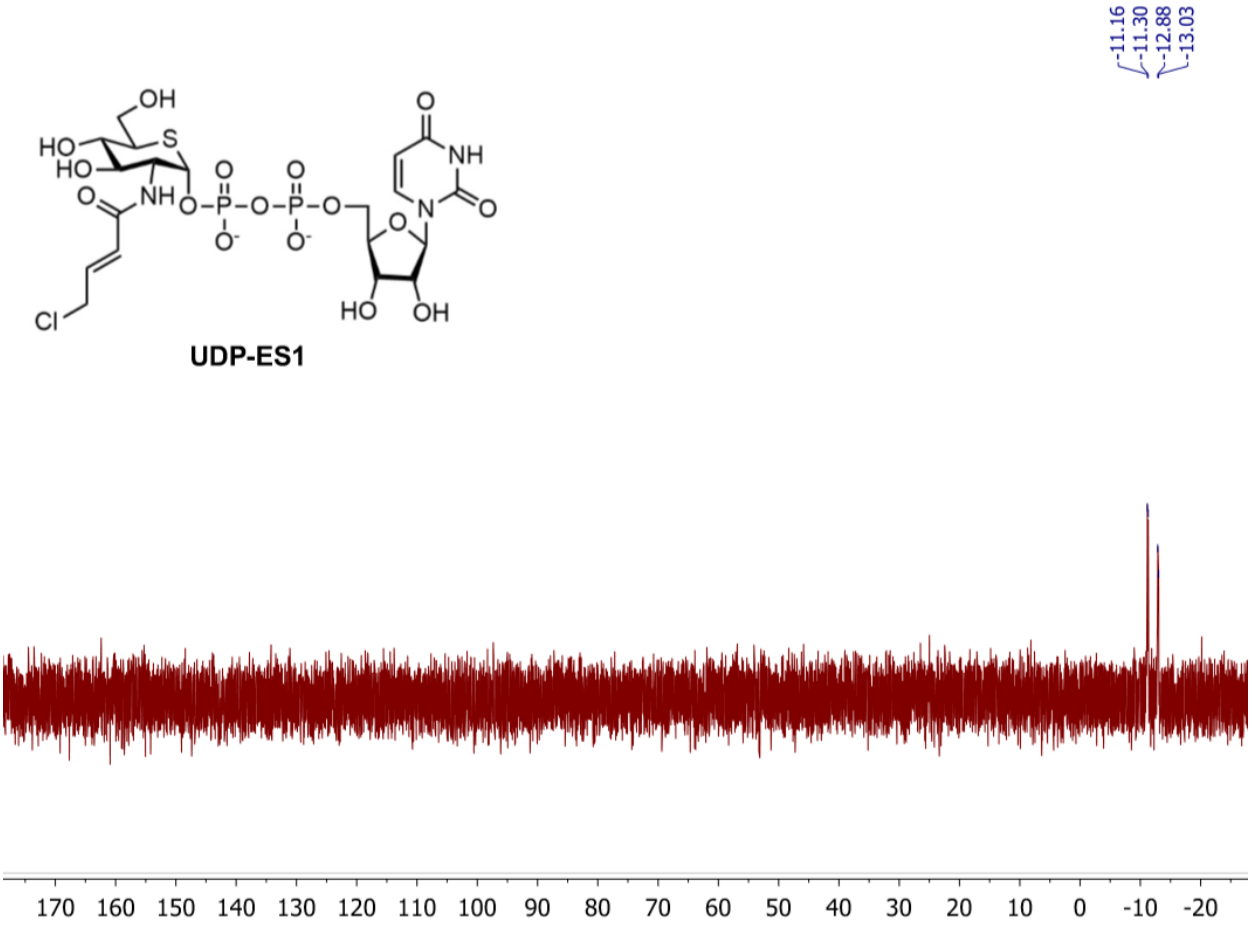
^{1}H NMR spectrum for **UDP-ES1**:



^{13}C NMR spectrum for **UDP-ES1**:



^{31}P NMR spectrum for **UDP-ES1**:



MS spectrum of **UDP-ES1**. **UDP-ES1** $[M-2H]^{2-}$ = 340.5104