

# DEVELOPMENT OF COUPLED ENZYME ASSAY AND *IN VITRO* SYNTHETIC BIOLOGY APPROACH FOR GLYCOSYLATION PATHWAY CHARACTERIZATION

The thesis presented to the

UNIVERSITY OF CAPE TOWN

In fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN CHEMISTRY

By

**ABDULLATEEF NASHED**

Supervisor: Professor Kevin J. Naidoo



Scientific Computing Research Unit

Department of Chemistry

University of Cape Town

July 2023

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

## DECLARATION

I declare that this dissertation titled Development of a Computational Platform for System Based Drug Discovery Demonstrated on Pneumococcal Sialidases is a presentation of my original research work done at the Scientific Computing Research Unit (SCRU), Department of Chemistry, University of Cape Town, South Africa. No part of this thesis has been submitted elsewhere for any other degree or qualification. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

|                     |
|---------------------|
| Signed by candidate |
|---------------------|

.....

Abdullateef Nashed

## ABSTRACT

Glycans play essential roles in living organisms that encompass a wide range of biological functions ranging from energy metabolism all the way through to intricate cell signalling pathways. They carry out these functions either directly by participating in metabolism and binding or indirectly by altering the structure and nature of their conjugates, for example by being the major post-translational modifier of proteins. The family of glycans in a biological organism (the glycome) correlates to the system's state. In this thesis the focus is the development of methods and measures to model the correlation of aberrant glycosylation of MUC1 that is associated with certain cancers. Central to this is an accurate MUC1 peptide monomer and associated glycosyltransferases (GTs) able to be deployed in specific construction. However, deconstructing the machinery of the GT-substrate interactome is challenging because of the complexity of glycan structures and the diversity of GTs acting on them. Two tool sets that can advance our understanding is the development of reliable kinetic assays and constructing biologically relevant substrate arrays. Presently used assays are disadvantaged by either being endpoint assays or continuous ones compromised by lack of specificity and sensitivity, or have high cost. Chemically synthesized glycan arrays often face insurmountable hurdles of complex carbohydrate chemistry. To address these challenges, the work undertaken in this thesis was firstly the development of a continuous enzymatic-coupled assay through the redesign of the pyruvate kinase/lactate dehydrogenase (PK/LDH)-coupled assay that lead to increasing the sensitivity and specificity of detection. Further the assay was adapted to accommodate all classes of GTs. Secondly, an *in vitro* synthetic biology approach was developed to construct a MUC1 O-glycans model peptide as a proof-of-concept of a seamless "one pot" glycopeptide synthesis strategy. Here a fusion protein, consisting of the MUC1 core peptide and a carrier protein, was designed for expression in *E. Coli*. The fusion protein elements were optimized to maximize solubility, facilitate without obstruction the *in vitro* enzymatic glycosylation of the peptide, and enable the recovery of the products through a simple one-step purification. The peptide in the fusion protein was optimized to undergo *in vitro* sequential enzymatic glycosylation that replicates the native pathway. The properties of the fusion proteins enabled investigation of the activities of the downstream GTs on the synthesized peptides in their fusion forms, which was used in the investigation of the site specificity of ST6GALNAC1, and its role in the biosynthesis of the breast cancer marker STn.

## ACKNOWLEDGEMENTS

I wish to thank the South African Research Chairs Initiative (SARChI) of the Department of Science and Technology (DST) and the National Research Foundation (NRF) for their generous funding, which made this research possible.

A special word of thanks goes to Prof. Kevin J. Naidoo, whose guidance and expertise have been instrumental in shaping this research. His mentorship has not only honed my research skills but also taught me the art of effective project management.

To Lydia Dreyer, I am truly grateful for your invaluable support in handling administrative matters throughout this journey. Your warmth and professionalism have made a significant difference.

I acknowledge and thank the Scientific Computing Research Unit (SCRU) members for the support I have received from each. Special appreciation goes to Mia Truter for her dedicated efforts in revising and proofreading the chapters of this thesis. Your keen eye for detail and constructive feedback have greatly improved the quality of this work.

I am also indebted to Prof. Ed Sturrock and Sylva Schwager for their generosity in sharing their lab and expertise with me. Their willingness to accommodate and collaborate has enriched my research experience beyond measure.

I owe an immeasurable debt of gratitude to my wonderful wife, Lienkie. Her unwavering patience and unconditional support have been my anchor throughout this challenging journey. Without her love and encouragement, I would not have had the strength to overcome the obstacles that came my way.

Last but certainly not least, a big thank you goes to my two boys, Jad and Rafi, who have been a constant source of inner strength, pushing me to stay focused on achieving the end result. Their presence in my life has been a driving force behind this accomplishment.

## ABBREVIATIONS

|           |                                                          |
|-----------|----------------------------------------------------------|
| ADP       | Adenosine Diphosphate                                    |
| AF        | Asialofetuin                                             |
| ATP       | Adenosine Triphosphate                                   |
| avGFP     | <i>Aequorea victoria</i> Green Fluorescent Protein       |
| Boc       | Tert-butyloxycarbonyl                                    |
| C1GALT1   | Core 1 $\beta$ 1,3-galactosyltransferase                 |
| C1GALT1C1 | C1GALT1-specific chaperone 1                             |
| CDP       | Cytidine Diphosphate                                     |
| GlcNAc    | <i>N</i> -Acetyl-D-glucosamine                           |
| CMK       | Cytidylate Kinase                                        |
| CMP       | Cytidine Monophosphate                                   |
| CTP       | Cytidine Triphosphate                                    |
| DoE       | Design of Experiments                                    |
| EGF       | Epidermal Growth Factor                                  |
| ER        | Endoplasmic Reticulum                                    |
| ESI-TOF   | Electrospray Ionization Time-of-Flight Mass Spectrometry |
| FBP       | Fructose-1,6-bisphosphate                                |
| Fomc      | 9-Fluorenylmethoxycarbonyl                               |
| FPIA      | Fluorescence Polarization Immunoassay                    |
| Fuc       | L-Fucose                                                 |
| Gal       | D-Galactose                                              |
| GalNAc    | <i>N</i> -Acetyl-D-galactosamine                         |
| GALNTs    | <i>N</i> -Acetylgalactosaminyltransferases               |
| GDP       | Guanosine Diphosphate                                    |
| GE        | Glycoenzyme                                              |

|       |                                               |
|-------|-----------------------------------------------|
| GFP   | Green Fluorescent Protein                     |
| Glc   | D-Glucose                                     |
| GlcA  | D-Glucuronic acid                             |
| GT    | Glycosyltransferase                           |
| GTP   | Guanosine Triphosphate                        |
| HPLC  | High-Performance Liquid Chromatography        |
| HTS   | High-Throughput Screening                     |
| IMAC  | Immobilized Metal Affinity Chromatography     |
| IPTG  | Isopropyl $\beta$ -D-1-thiogalactopyranoside  |
| LC    | Liquid Chromatography                         |
| LDH   | Lactate Dehydrogenase                         |
| Man   | D-Mannose                                     |
| MS    | Mass Spectrometry                             |
| MUC1  | Mucin Type 1                                  |
| NADH  | Reduced Nicotinamide Adenine Dinucleotide     |
| NDK   | Nucleoside diphosphate Kinase                 |
| NDP   | Nucleoside Diphosphate                        |
| OFAT  | One Factor at a Time                          |
| PEP   | Phosphoenolpyruvate                           |
| PK    | Pyruvate Kinase                               |
| PTMs  | Posttranslational Modifications               |
| sfGFP | Superfolder Green Fluorescent Protein         |
| Sia   | <i>N</i> -Acetylneuraminic acid (sialic acid) |
| SNFG  | Symbol Nomenclature for Glycans               |
| SPPS  | Solid Phase Peptide Synthesis                 |
| SPS   | Solution Phase Synthesis                      |
| ST    | Sialyl-Tn antigen                             |

|           |                                                                                |
|-----------|--------------------------------------------------------------------------------|
| ST3GAL1   | Sialyltransferase 3, N-acetylgalactosaminide $\alpha$ -2,3-sialyltransferase 1 |
| ST6GALNAC | Sialyltransferase 6, N-acetylgalactosaminide $\alpha$ -2,6-sialyltransferase   |
| STn       | Sialyl-Tn antigen                                                              |
| T         | Thomsen-Friedenreich antigen                                                   |
| TAE       | Trial and Error                                                                |
| TEV       | Tobacco Etch Virus                                                             |
| Tn        | T nouvelle antigen                                                             |
| UDP       | Uracil Diphosphate                                                             |
| UTP       | Uridine Triphosphate                                                           |
| Xyl       | D-Xylose                                                                       |

# TABLE OF CONTENTS

## Contents

|                                                                |    |
|----------------------------------------------------------------|----|
| Chapter 1 Introduction .....                                   | 1  |
| 1.1 Saccharides: Structures and Functions .....                | 2  |
| 1.2 Glycomics .....                                            | 6  |
| 1.3 Systems Glycobiology.....                                  | 10 |
| 1.4 Prospects and Promises of GE-Substrate Interactomics ..... | 12 |
| 1.5 Tools to Study GE-Substrate Interactome .....              | 14 |
| 1.6 Objectives.....                                            | 16 |
| Chapter 2: .....                                               | 19 |
| Kinetics and Glycosylation Reaction Assays .....               | 19 |
| Chapter Overview.....                                          | 20 |
| 2.1 Enzymatically catalyzed reactions.....                     | 20 |
| 2.2 Michaelis-Menten Kinetics .....                            | 24 |
| 2.3 Bi-Bi Reactions .....                                      | 29 |
| 2.4 Enzyme Assays .....                                        | 37 |
| 2.5 Currently used assays for GT .....                         | 40 |
| 2.6 Validity of an Enzymatic-Coupled Assay .....               | 44 |
| 2.7 Proposed Improvement to PK/LDH assay.....                  | 44 |
| Chapter 3.....                                                 | 47 |
| Assay Development.....                                         | 47 |
| 3.1 Chapter Overview .....                                     | 48 |
| 3.2 Optimization of NADH Signal .....                          | 50 |
| 3.3 Optimization of PK/LDH Coupling.....                       | 54 |
| 3.4 Coupling of Cytidylate Kinase (CMK).....                   | 66 |
| 3.5 Expression and Coupling of NDK.....                        | 69 |
| 3.6 Performance at Low pH.....                                 | 77 |
| 3.7 Applications .....                                         | 79 |
| Chapter 4.....                                                 | 83 |
| Substrate Specificity and In vitro Synthesis Method .....      | 83 |

|                  |                                                                              |     |
|------------------|------------------------------------------------------------------------------|-----|
| 4.1              | Chapter Overview .....                                                       | 84  |
| 4.2              | Mucin-Type O-Glycosylation.....                                              | 84  |
| 4.4              | Methods to Study GT Specificities .....                                      | 93  |
| 4.5              | Glycopeptide Synthesis Methods .....                                         | 94  |
| 4.6              | Short Peptide Synthesis .....                                                | 95  |
| 4.7              | Fusion Expression .....                                                      | 98  |
| 4.8              | Human GT Enzymes Expression.....                                             | 107 |
| Chapter 5: ..... |                                                                              | 110 |
|                  | In vitro Glycopeptide Synthesis.....                                         | 110 |
| 5.1              | Chapter Overview .....                                                       | 111 |
| 5.2              | Expression of The Fusion Proteins .....                                      | 113 |
| 5.3              | Validation of the Sequence Design.....                                       | 121 |
| 5.4              | Biological Relevance of Synthesized Structures.....                          | 127 |
| 5.5              | Preparative Tn Synthesis .....                                               | 128 |
| 5.6              | Preparative Core 1 (T antigen) Synthesis .....                               | 135 |
| 5.7              | Structure Confirmation via LC-MS .....                                       | 147 |
| Chapter 6.....   |                                                                              | 150 |
|                  | Application Case Study: Specificity of ST6GALNAC1 .....                      | 150 |
| 6.1              | Overview .....                                                               | 151 |
| 6.2              | STn in Cancer.....                                                           | 151 |
| 6.3              | Regulation of STn expression.....                                            | 153 |
| 6.4              | GALNTs relocation to the ER.....                                             | 155 |
| 6.5              | ST6GALNAC1 Core Peptide requirement.....                                     | 156 |
| 6.6              | Site Specificity of ST6GALNAC1 vs C1GALT1 .....                              | 158 |
| 6.7              | Summary of the site specificities and Regulation model of STn synthesis..... | 161 |
| Chapter 7.....   |                                                                              | 169 |
|                  | Conclusions .....                                                            | 169 |
| Appendix .....   |                                                                              | 172 |
| A.1              | <i>E. Coli.</i> Expression Plasmids.....                                     | 173 |
| A.2              | Materials and General Protocols .....                                        | 177 |
| A.3              | Expression of GTs .....                                                      | 183 |
| A.4              | Coupling assay protocol .....                                                | 187 |
| A.5              | Pooled standard deviation:.....                                              | 189 |
| A.6              | Batch Mode Simulation .....                                                  | 190 |

|      |                                                                 |     |
|------|-----------------------------------------------------------------|-----|
| A.7  | ADP Standard.....                                               | 193 |
| A.8  | Fed Batch Simulation.....                                       | 194 |
| A.9  | LC-MS Procedure: .....                                          | 195 |
| A.10 | General Protocol for Buffer Preparation and pH Adjustment ..... | 197 |
| A.11 | Aseptic Techniques and Cell Culture .....                       | 197 |
| A.12 | Statistical .....                                               | 197 |

# **Chapter 1**

## **Introduction**

# 1.1 Saccharides: Structures and Functions

## 1.1.1 Glycans

Saccharides comprise polyhydroxyketones, polyhydroxyaldehydes, or their derivatives. They are also called carbohydrates, which means the hydrates of carbon, a literal description of the generic chemical formula  $C_x(H_2O)_n$ . Naturally occurring monosaccharides can also carry modifications on one or more of the hydroxyl groups, such as substitution with an amino group or its amides, resulting in more complex structures that deviate from the generic formula. Saccharides are among the four most abundant macromolecules found in any biological system, alongside lipids, proteins, and nucleic acids. The term "Glycans" refers to saccharides at various levels of polymerization: monosaccharides, disaccharides, oligosaccharides, and polysaccharides. Saccharides can vary in polymerization homogeneity, being either homo- or heteropolymers. Furthermore, glycan structures can either be linear or branched, and may exhibit different degrees of branching, including di-, tri- and tetra-antennary structures. Glycans can also exist as conjugates with other molecules, such as proteins and lipids.

In humans, the entire spectrum of glycans is formed by a combination of mainly nine monosaccharides and their derivatives<sup>1</sup>. Each monosaccharide contains an anomeric carbon atom (reducing end) and multiple hydroxy groups (non-reducing end) (Figure 1.1). Each hydroxy group functions as a potential binding site for an anomeric carbon from another monosaccharide to form a new bond. This bond, known as glycosidic bond, can take  $\alpha$  or  $\beta$  anomeric configuration. Given the diversity of monosaccharides building blocks, the possible variety of sites and the anomeric configurations a glycosidic bond can have, and the potential polymerization possibilities, the number of theoretical structures of glycans may be close to infinite. In nature, these structures, despite the overwhelming diversity, are conserved and limited to core structures that help to classify any naturally occurring glycan. In addition to the glycan building monosaccharides, other sugar molecules such as fructose, ribose, deoxyribose, and ascorbic acid play important roles in cell functions such as energy metabolism, metabolic regulation, and cell signalling. However, these saccharides and their functions are out of the scope of this thesis.

Biosynthesis of glycans is regulated via two groups of enzymes: glycosyltransferases (GTs) that catalyze glycosidic bonds formation, and glycosidases that catalyze the hydrolysis of these

bonds. Glycans can bind to proteins and lipids to form glycoconjugates: glycoproteins and glycolipids, respectively. In fact, glycosylation is a major post translational modification (PTM) that takes place during protein maturation, and is determinant to protein folding and modulation of function. The assortment of glycans and glycoconjugates in a given system in a specific state is referred to as the glycome<sup>2,3</sup>.

The Symbol Nomenclature for Glycans (SNFG) is a standardized system for representing glycans. Before the SNFG's establishment, there was no universal nomenclature for glycans, hindering effective communication among researchers. In 2005, an international group of scientists proposed the SNFG, utilizing graphical symbols and colours to depict monosaccharides, linkages, and other structural features. This intuitive and widely adopted system has since facilitated comparative studies and accelerated research in glycobiology, immunology, and biomedicine, becoming a crucial tool for communicating and understanding glycan structures in the scientific community<sup>4</sup>.

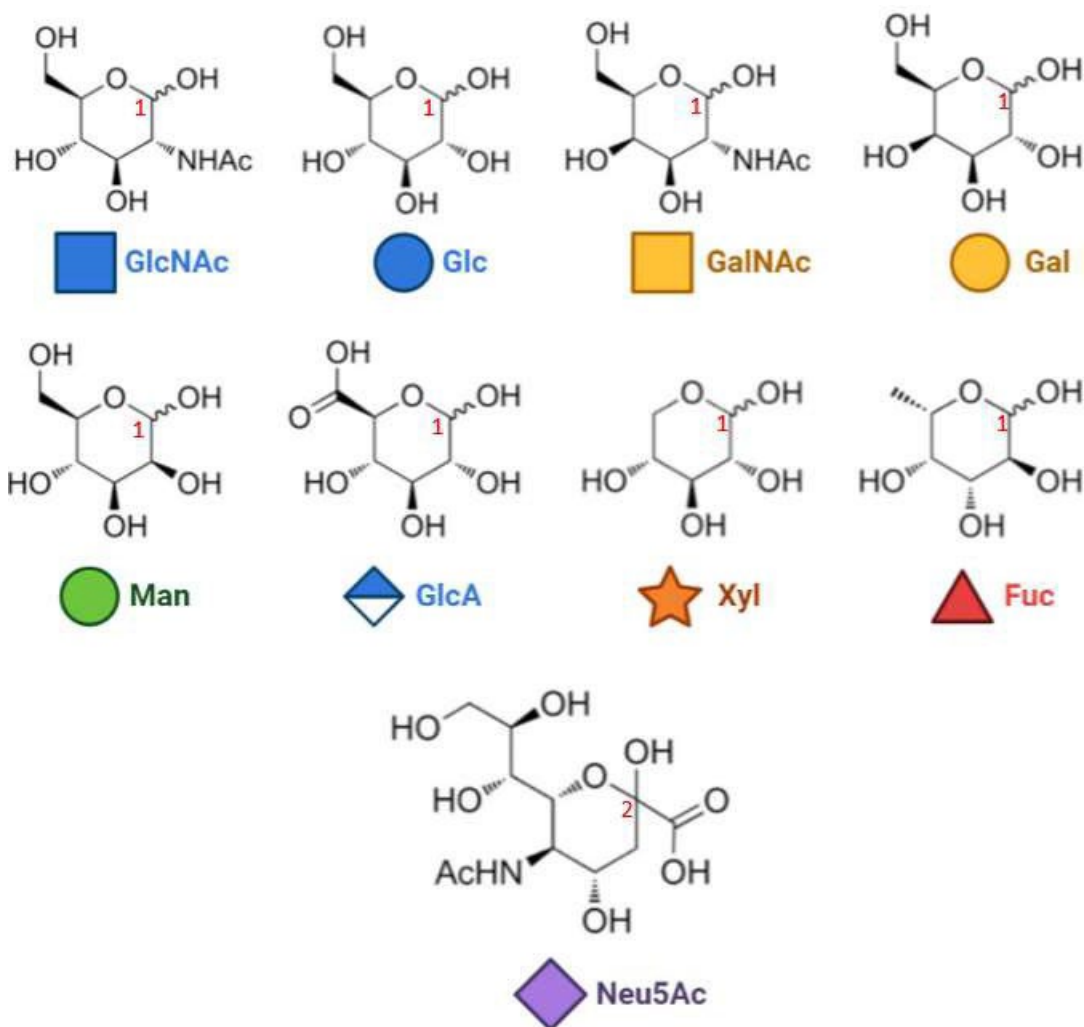


Figure 1.1. The structures of the nine major monosaccharides constituting the human glycome. Graphical symbols and abbreviations are presented according to The Symbol Nomenclature for Glycans (SNFG). Full names of the monosaccharides are as follows: GlcNAc: *N*-Acetyl-D-glucosamine, Glc: D-Glucose, GalNAc: *N*-Acetyl-D-galactosamine, Gal: D-Galactose, Man: D-Mannose, GlcA: D-Glucuronic acid, Xyl: D-Xylose, Fuc: L-Fucose, Sia: *N*-Acetylneuraminic acid (sialic acid). Anomeric carbon is labeled with its order in the ring for each monosaccharide.

### 1.1.2 Glycobiology

Glycobiology – the biology of sugars – is the study of structures, synthesis pathways, and functions of glycans and their glycoconjugates. The emergence of the field can be traced back to as early as the first half of the 20<sup>th</sup> century when the chemistry and structures of glycans were discovered. Later, the roles of these molecules in metabolism, particularly in energy storage and generation, became well-established.

The term glycobiology was first used in the late 1980s to describe the interdisciplinary field of studying glycans at the level of chemistry, biochemistry, cell biology, and molecular biology<sup>2</sup>. The rise of glycobiology was fueled by the expansion of information on the biological rules of glycans and the tight correlation between glycan structures and glycan biosynthesis pathways and diseases<sup>5,6</sup>. Most of the roles of glycans were discovered while investigating other physiological or pathological phenomena. For example, studying infectious agent – host system interactions led to the discovery that glycans on the cell surface and in the extracellular matrix of the host play roles in the pathogen attachment and invasion. In fact, many pathogens evolved specific glycosidases to break down specific host glycans to facilitate the infection. On the other hand, specific glycans were found on pathogens that are specialized in regulating the interaction with the host immune system. In a review on glycan functions, 49 unique functions of glycans were reviewed and categorized into four groups: Structural and modulatory roles, Extrinsic (interspecies) recognition of glycans, Intrinsic (intraspecies) recognition of glycans, and Molecular mimicry of host glycans<sup>5</sup>, highlighting the diverse biological functions of glycans.

The majority of glycans functions stem from the physical and chemical properties of their structures and/or the modulatory role of these structures on their conjugate partners. For example, cellulose, a polymer of (1→4)-β-D- glucopyranose, is responsible for rigidity, strength, and water regulation in plant cell walls<sup>7</sup>. One of the most ubiquitous functions of glycans is their role in protein PTM. The maturation of proteins occurs within the ER-Golgi axis, where precise glycosyltransferases (GTs) coordinate in a temporal and spatial manner to generate specific glycan structures on particular amino acids of the protein. It is well-established that the majority of membrane and secreted proteins are glycosylated on at least one amino acid<sup>8</sup>. Glycosylation of proteins regulates their functions via several mechanisms including: folding and degradation, water solubility, protection from proteases, adhesion/anti-adhesion properties, recognition by

glycan-binding proteins, immune recognition, and targeting the protein to trafficking and endocytosis<sup>5</sup>.

In humans, the extended structures of glycans conjugated to mucin proteins in normal epithelia provide protection from pathogens, pollutants, and pH changes. Additionally, these glycoproteins mediate cell-cell and cell-extracellular matrix interaction regulating cell adhesion and migration, tissue remodeling, and cell signaling. Changes to the glycan structures of mucins, as seen in infectious diseases and cancer, can disrupt the normal functions of these mucins, thereby facilitating disease progression<sup>9</sup>. The role of MUC1 glycosylation in breast cancer is covered in Chapter 6.

## **1.2 Glycomics**

Glycomics, a branch of glycobiology, describes the analysis of the structures and functions of glycans to understand their roles in biological systems. This includes studying glycan-protein interactions, glycan biosynthesis pathways, glycan modifications, and their impact on cellular signaling, immune responses, cell adhesion, and more<sup>5,8</sup>. Similar to proteomics, glycomics can be divided into structural and functional glycomics<sup>10</sup>. Structural glycomics focuses on glycan structures and their binding sites, while functional glycomics includes identification of the binding partners of glycans or glycoconjugates, as well as investigation of glycan differential expression in tissues and diseases (Figure 1.2).

### **1.2.1 Structural Glycomics**

The complexity of an organism's glycome exceeds that of its genome, transcriptome, or proteome (Figure 1.3). Recently, significant progress has been made in identifying and quantifying glycan structures. Advances in purification and recovery techniques, like enzymatic and chemical digestion of glycans, along with improved detection methods such as NMR and mass spectrometry, have accelerated the speed and throughput of glycan structure analysis. However, the primary focus remains on generating a "parts list" of glycans found in specific cells or tissues<sup>5</sup>. This limitation to partial profiling is caused by the specific purification and glycan release methods used for each class of glycans, which usually makes it difficult to recover and analyze the whole glycome of a given sample<sup>8</sup>.

To gain a complete understanding of glycans' biology, efforts are required to discover all glycan attachment sites on proteins and lipids. For example, of the 20 amino acids that constitute proteins, nine are known to undergo glycosylation in the mature protein. These amino acids are: asparagine, arginine, serine, threonine, tyrosine, tryptophan, cysteine, hydroxylysine, and hydroxyproline<sup>8</sup>. It is worth noting that both hydroxylysine, and hydroxyproline are products of the posttranslational hydroxylation of lysine and proline, respectively. Assignment of glycan structures to their native amino acids sites on each protein (glycoproteomics) remains a formidable task. However, such assignment has been performed on individual proteins with remarkable success using a combination of purification and MS techniques<sup>11-13</sup>. Current methods in glycoproteomics can be found in a review paper published in 2022 covering the major methods and limitations in the field<sup>14</sup>.

### **1.2.2 Functional Glycomics**

As mentioned earlier, the primary way glycans exert their functions is by modulating binding to partner molecules, usually proteins, either directly or by influencing the binding affinity of their conjugates. In the past, studying protein-glycan interactions was challenging due to time-consuming and low-throughput methods. However, recent advancements, such as glycan microarrays containing vast libraries of complex glycans, have brought about a revolution in glycobiology. These microarrays empower researchers to explore the binding specificities of glycans with various biomolecules, significantly advancing the field of glycobiology. Techniques like indirect or direct fluorescence-based assays were applied to the glycan microarray screening, facilitating the identification of specific interactions between glycans and glycan binding proteins (lectins), antibodies, toxins, viruses, and glycoenzymes (GEs)<sup>8,15</sup>.

Since glycoenzymes, comprising glycosidases and glycosyltransferases, play a crucial role in defining the glycome, understanding their specificity is a fundamental aspect of glycomics. Moreover, these enzymes serve as valuable resources for enzymatic glycan synthesis, which can reduce the cost of glycan microarray construction and enable the synthesis of glycans that cannot be synthesized using traditional carbohydrate chemistry. Researchers have shown significant interest in utilizing microarrays to investigate glycoenzyme activity. This approach enables the examination of enzyme specificity against a range of immobilized potential substrates<sup>16,17</sup>. This effort towards unravelling the regulation of glycan biosynthesis is the ultimate goal of

glycobiology. Understanding the regulation of glycan biosynthesis would provide valuable insights into the driving events of glycosylation involvement in pathogenesis. This knowledge can significantly enhance the diagnosis and therapy of diseases that rely crucially on glycosylation for their development.

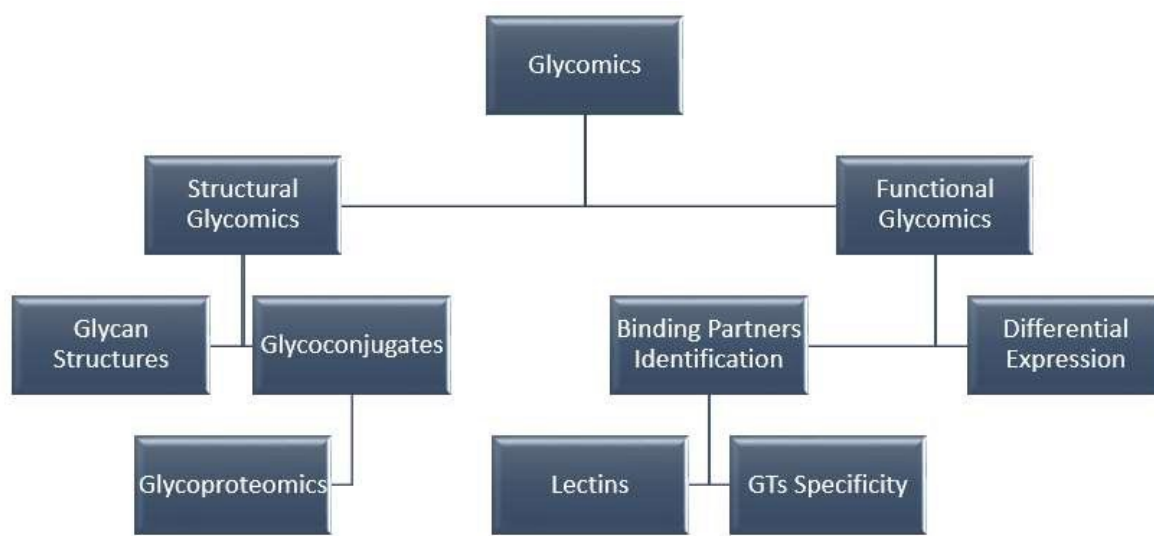


Figure 1.2. Levels of Glycomics. The field of glycomics studies the glycome from two key viewpoints: its structure and its function. Structural glycomics studies the glycans independent from their conjugates. This approach emerged due to technical limitations in isolating and detecting glycans in their conjugate's context. By overcoming these barriers, the study of glycan attachment sites on conjugate molecules, especially proteins, was enabled and led to the emergence of glycoproteomics. Functional glycomics investigates the functions of glycans, either at a low molecular level by investigating their binding partner, or at a higher level by investigating the association of different glycans with certain conditions (differential expression). Glycan microarrays are the main tools that enabled the high throughput screening of glycans interactions with other molecules. One significant application of these arrays is studying GTs-glycan interactome to study the specificity of GTs and ultimately decipher the biosynthesis pathways of glycans.

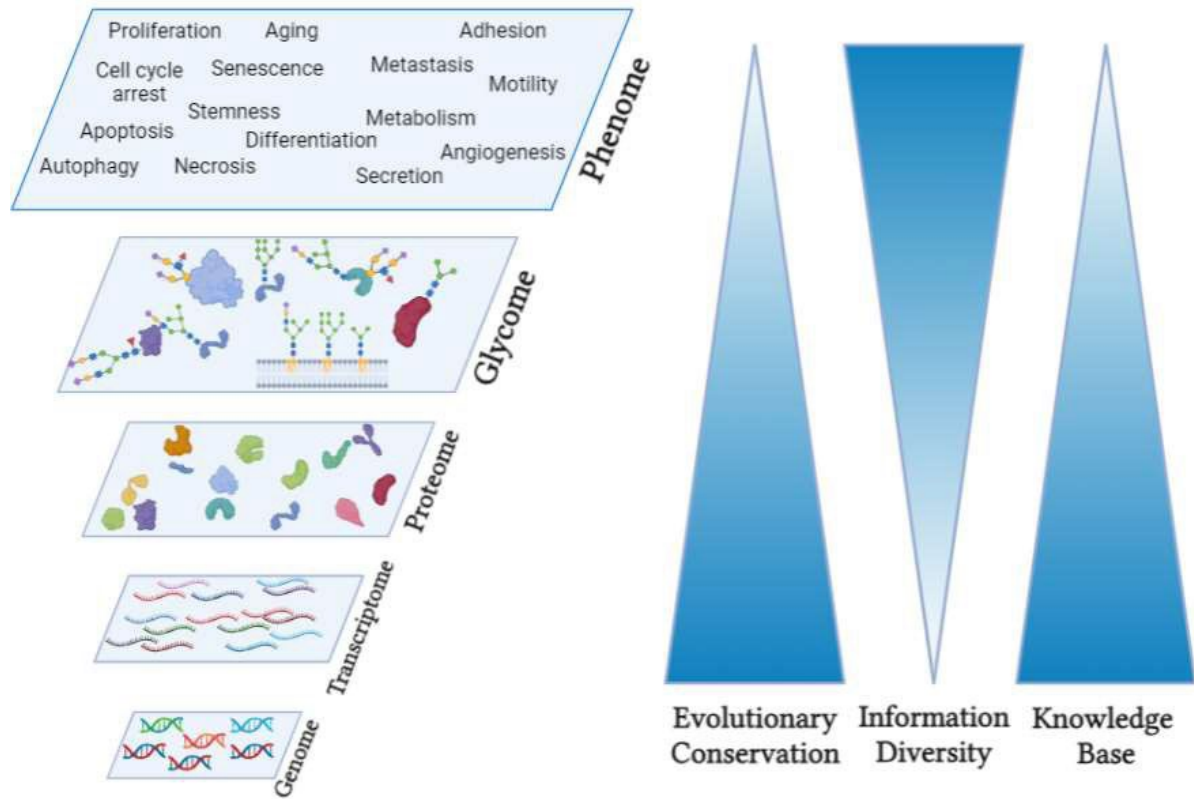


Figure 1.3. The Glycome in the context of the pan-omics view of a biological system. The regulatory network of a biological system is classified into multi-subnetworks, each consisting of a specific class of molecule as defined by the central dogma (shown on the left of the figure). The hierarchy represents the transformation of the information from one class of molecules to another from the genome to phenome direction. To the right is comparison of the different aspects of information among the molecule classes as adapted from (Varki, 2017)<sup>5</sup>

## 1.3 Systems Glycobiology

### 1.3.1 Glycosylation: a system-wide process

Glycosylation is a complex process that recruits and impacts a significant proportion of biological machinery<sup>18</sup>. The supply of glycosyltransferases and nucleotide sugar donors, representing glycosylation tools and building blocks, respectively, is initiated by two independent upstream pathways. The first step of the synthesis of sugar donors is the cellular uptake of monosaccharides: glucose, mannose, galactose, and fucose. These sugars undergo a series of modifications including epimerization, phosphorylation, and acetylation to produce the entire suite of nine saccharides that participate in glycan synthesis. The last step of the donor synthesis pathway is the transformation of the sugars into high-energy nucleotide sugar donors. The biosynthesis pathway of nucleotide sugar donors is tightly correlated to, and regulated by, the central metabolism pathways of sugars. Therefore, the crosstalk between glycosylation and energy metabolism is evident<sup>19</sup>.

There are more than 200 glycosyltransferases identified in humans. These enzymes are expressed in the endoplasmic reticulum (ER) and distributed between the ER and the different sacks of Golgi apparatus: cis, medial, and trans, ordered from the ER to the cell surface. The distribution of GTs is determined by the length of their transmembrane domains. The longer this domain is, the more likely the enzyme to localize towards the trans side of the Golgi<sup>20</sup>. The localization of GTs along the ER-Golgi axis is dynamic, and these enzymes are constantly trafficked in both directions via a complex but tightly regulated vesicle system. This trafficking system is regulated via signaling pathways involving cell membrane receptors and kinases that are also involved in cancer (Discussed in more detail in Chapter 6)<sup>21,22</sup>. In addition to localization, the expression levels of GTs (glycosyltransferases) are not static but rather dynamic and can be altered in response to changes in the overall functional and developmental state of the cell<sup>23</sup>. For instance, during the process of carcinogenesis (the development of cancer), the expression of GTs is modified in a manner that creates a unique profile associated with cancer. This altered GT expression profile in cancer samples can be utilized to classify different types and subtypes of cancer<sup>24</sup>. Finally, glycan modification can impact all known cellular functions via alteration of signaling pathways and modulation of interactions with the glycans and glycoconjugates partners as discussed above.

### 1.3.2 Systems glycobiology

Systems glycobiology, also referred to as the systems biology of glycans, is an interdisciplinary field that aims to understand the role of glycans and their regulation in a holistic and system-wide manner. It combines elements of glycomics with systems biology, a branch of biology that focuses on studying complex interactions and networks within biological systems. Glycomics, together with other omics disciplines, contributes to the comprehensive understanding of the molecular landscape of biological systems and is essential for advancing our knowledge of various diseases, including cancer, inflammation, and infectious diseases, where glycan structures and functions often play critical roles<sup>18,25</sup>.

Key aspects of systems glycobiology include network analysis, omics integration, computational modeling, and investigating functional consequences. Network analysis involves identifying and analyzing the intricate networks formed by glycans, glycoenzymes, glycan-binding proteins, and other biomolecules in a cellular context. Omics integration aims to combine glycomics data with other omics data (e.g., genomics, proteomics, transcriptomics) to gain a more comprehensive understanding of cellular processes and functions. Computational modeling focuses on developing computational models to simulate and predict the behavior of glycans and their interactions within biological systems in different states. Investigating functional consequences entails studying how changes in glycan structures impact cellular functions and contribute to disease development and progression<sup>18,25</sup>.

The complexity of the glycan structures surpasses that of proteins and nucleic acids. For instance, the nine monosaccharides in humans can theoretically be assembled into more than 15 million chemically possible tetrasaccharides<sup>18</sup>. Unlike proteins and nucleic acids, the structures of glycans are not directly encoded in the genome and their synthesis is not template-based, but rather determined by the orchestrating activities of glycoenzymes in a spatial and temporal fashion. The structures (and attachment sites) of glycans are therefore limited to the expression and the catalytic specificity of glycoenzymes.

While most glycoenzymes, including glycosyltransferases, have been identified and characterized<sup>26</sup>, the full spectrum of specificity remains incompletely understood. The current knowledge of these specificities is primarily derived from disease-specific models and is limited to only a few disease-specific glycosyltransferases tested across a handful of glycan structures.

However, specificity studies reveal that glycoenzymes demonstrate different degrees of promiscuity and multiple enzymes can catalyze the same glycosidic bond formation. These findings predict a complex system that deviates from the traditional view of one-enzyme, one-reaction<sup>18</sup>.

### **1.3.3 Glycoenzyme specificity: the missing link**

The lack of a complete picture of glycoenzyme-substrate specificity, as well as the lack of kinetic parameters for such reactions, has hindered the utilization of systems biology analysis tools in modeling glycan synthesis and functions. This is mainly due to incomplete understanding of the regulatory network as a whole, which arises as a result of inability to bridge the gap between glycomics and the other omics, specifically genomics and transcriptomics<sup>27</sup> i.e., the inability to link glycoenzyme genomics and transcriptomics to the final glycosylation landscape within the cell.

A comprehensive systems glycobiology model would address this gap by integrating a glycosylation biosynthesis network constructed from the glycoenzyme-substrate specificity data. Since glycoenzyme-substrate interactions have a holistic and system-wide nature, they will be collectively referred to as the "glycoenzyme (GE)-substrate interactome."

## **1.4 Prospects and Promises of GE-Substrate Interactomics**

### **1.4.1 Applications in Therapy**

The lack of understanding of the glycan biosynthesis network limits the therapeutic strategies of glycan-aberrant diseases to developing antibodies that bind the aberrant glycan and neutralize its effect<sup>28,29</sup>. Despite the success of antibody-based therapies, their applications are limited to targets that are displayed on the cell surface, as these antibodies cannot be taken up by cells. Furthermore, the deregulation of glycosylation associated with diseases does not adhere to a single-enzyme, single-glycan mechanism. Instead, it involves the alteration of a specific subset of GEs<sup>24</sup>. This unique GE signature, specific to each disease, is expected to modify a broad spectrum of glycans, some of which directly contribute to pathogenesis. Therefore, these modified glycans present potential therapeutic targets. Systems glycobiology models that are enabled by a comprehensive GE-substrate interactomics aid the identification of such targets.

Additionally, system-wide analyses help to identify key GEs driving abnormal glycosylation, which provides an opportunity to explore new therapeutic strategies that target drivers of abnormal glycosylation as opposed to downstream aberrancies.

Glycosylation can impair the efficacy of therapy by altering the binding interaction between therapy and target. This was discovered in a subgroup of cancer patients that showed resistance to treatment with antibodies targeting the programmed cell death ligand 1 (PD-L1). The resistance was found to be caused by the excessive glycosylation of PD-L1, which impaired antibody recognition. A similar example was found in breast cancer patients treated with trastuzumab, an anti-HER2 antibody. Sialylation of HER2 resulted in significant resistance to trastuzumab. This resistance was mitigated by combining the treatment with a sialidase, an enzyme that removes sialic acid from glycans<sup>29</sup>.

Understanding the underlying biosynthesis of glycans would improve therapeutic strategy in several ways. Instead of an antibody approach wherein the therapy can only target the cell-surface, small molecule inhibitors targeting the culprit glycans at the biosynthesis level would enable a wider range of glycan-therapy applications. The examples highlighted above also indicate that glycosylation pathways can be utilized as predictive biomarkers, including identification of patients that would respond to, or be resistant to, certain treatments. Glycosylation could therefore help pave the way toward personalized medicine.

#### **1.4.2 Applications in industry**

Antibodies as therapeutic agents suffer from batch-to-batch heterogeneity due to the variation in PTMs – mainly glycosylation – during production, which requires implementing intensive quality control procedures in the production process. Glycosylation heterogeneity of therapeutic antibodies is a major bottleneck impacting the cost and the efficacy of such therapeutics<sup>30</sup>. Therapeutic antibodies are usually produced in a host system, mainly eukaryotic cells, therefore glycosylation PTM is determined by the host glycosylation machinery, which can be altered by the growth conditions and other factors during the process. The conventional strategies to address this problem include implementing an upstream optimization of growth conditions and media composition to favour the desired glycosylation PTMs. These strategies are often applied in combination with downstream purification methods. Despite their efficiency in improving the

quality of antibodies, these methods still fall short in terms of yield and robustness. Currently, the effort is focused on implementing alternative upstream strategies that involve engineering the host cell's glycosylation pathways to produce more predictable glycosylation modifications<sup>30</sup>. This effort would benefit from the utilization of systems glycobiology tools and more so from a reliable GE-substrate interactomic data generated specifically for the antibody of interest, accounting for its unique peptide core sequence targeted by the host glycosylation machinery.

## 1.5 Tools to Study GE-Substrate Interactome

The *in vitro* assessment of GE-substrate interactome is a technologically demanding task. This is mainly due to the diversity of the enzymes and substrate pairs that need to be tested. To address this, an ideal method should be capable of high-throughput screening. The first challenge involves obtaining the glycoenzymes in their active and native forms. The second challenge is developing a measurement technology (assay) that can effectively monitor reactions involving all classes of glycoenzymes and their diverse substrates. The most significant challenge lies in the availability of glycan substrates. Advancements in chemical and biological technologies have facilitated enzyme expression and substrate synthesis. The accessibility of materials, along with improved detection methods, has increased the throughput of screening GT-substrate interactions.

The Gateway recombinational cloning technology has enabled the construction of an expression vector library encompassing all human glycosyltransferases. These vectors are designed for expression in human cell line hosts, preserving the post-translational modifications necessary for enzyme activities<sup>31</sup>. The expression constructs of this library were designed in the form of fusion proteins that include, in addition to the enzyme catalytic domains, regulatory elements to facilitate secretion and detection of the expressed protein. Additionally, the design enables the cleavage of these fused elements after expression to obtain the enzymes in their native forms.

Currently, several assays have been developed to study the kinetics of glycosyltransferases (GTs) as the major GEs involved in the glycan synthesis pathways. However, these methods have limitations, with some being endpoint assays, and others continuous but compromised in terms of specificity, sensitivity, or cost. Consequently, these current techniques are deemed inaccurate and impractical for use in high throughput screening (Chapter 2, Section 2.6). One assay, the

pyruvate kinase/ lactate dehydrogenase coupled assay<sup>32,33</sup> has the potential to overcome all the limitations mentioned above. However, this assay is underdeveloped, and its performance has not been optimized to accommodate all classes of GTs.

Substrates are usually chemically synthesized and tested as either oligosaccharide glycans or glycoproteins. Despite the challenges of chemical synthesis methods, significant progress has been made in generating a diverse range of oligosaccharide arrays akin to nucleic acids.

However, chemically synthesized arrays have their limitations, as certain structures are difficult to synthesize, and glycoarrays lack the glycoconjugate context, such as core proteins and lipids. The lack of glycoconjugate context compromises the relevance of information generated from these arrays, as exemplified by studies on GALNTs and other immunogenic glycoconjugates<sup>34</sup>. This need has led to the evolution of glycan arrays to glycoconjugate arrays, with a focus on glycoproteins, which dominate the cellular landscape.

## 1.6 Objectives

The general objective of this thesis is to design versatile tools that address key aspects crucial for advancing the acquisition of GT-substrate interactome data *in vitro*. These tools demonstrate applicability across all classes of GTs, ensuring improved biorelevance and enhanced accuracy. Moreover, they are tailored to facilitate high throughput formats, while remaining cost-effective and easy to implement, thereby ensuring simplicity in their application. Specific objectives are as follows:

The first objective is to develop a continuous enzymatic-coupled assay by redesigning the PK/LDH-coupled assay to increase the sensitivity and specificity of detection. Additional enzymatic steps were incorporated to accommodate all classes of GTs. The theory of enzymatic kinetics and the rationale of the design are covered in Chapter 2, while Chapter 3 covers the optimization, development, and validation procedures.

The second objective is to develop an *in vitro* synthetic biology approach for constructing MUC1 O-glycans as a proof-of-concept model. A fusion protein was designed, incorporating the MUC1 core peptide and a carrier protein, to be expressed in *E. Coli*. Other elements were optimized in the fusion constructs to facilitate affinity-based purification, maximize solubility, and minimize carrier interference with the core peptide. These features aim to enable *in vitro* enzymatic glycosylation and facilitate simple one-step purification. Chapter 4 covers the current methods used in glycan array synthesis and the rationale of the proposed *in vitro* synthetic biology method. The validation of the design and the construction of the glycopeptides are presented in Chapter 5.

A use case, presented in Chapter 6, aimed to investigate the site specificity of ST6GALNAC1 and its role in biosynthesis of a diagnostic and prognostic breast cancer marker, sialyl-Tn antigen.

## References

- 1 Kanie, Y. & Kanie, O. Addressing the glycan complexity by using mass spectrometry: In the pursuit of decoding glycologic. *Biochem. Compd* **5**, 3 (2017).
- 2 Varki A, Kornfeld S. Historical Background and Overview. In: Varki A, Cummings RD, Esko JD, et al., editors. *Essentials of Glycobiology* [Internet]. 4th edition. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2022. Chapter 1.
- 3 Kiessling, L. L. & Splain, R. A. Chemical approaches to glycobiology. *Annu Rev Biochem* **79**, 619-653 (2010).
- 4 Varki, A. *et al.* Symbol nomenclature for graphical representations of glycans. *Glycobiology* **25**, 1323-1324 (2015).
- 5 Varki, A. Biological roles of glycans. *Glycobiology* **27**, 3-49 (2017).
- 6 Ghazarian, H., Idoni, B. & Oppenheimer, S. B. A glycobiology review: carbohydrates, lectins and implications in cancer therapeutics. *Acta Histochem* **113**, 236-247 (2011).
- 7 McFarlane, H. E., Doring, A. & Persson, S. The cell biology of cellulose synthesis. *Annu Rev Plant Biol* **65**, 69-94 (2014).
- 8 Cummings, R. D. & Pierce, J. M. The challenge and promise of glycomics. *Chem Biol* **21**, 1-15 (2014).
- 9 Kufe, D. W. Mucins in cancer: function, prognosis and therapy. *Nat Rev Cancer* **9**, 874-885 (2009).
- 10 Lefkovits, I. Functional and structural proteomics: a critical appraisal. *J Chromatogr B Analyt Technol Biomed Life Sci* **787**, 1-10 (2003).
- 11 Harrison, R. *et al.* Glycoproteomic characterization of recombinant mouse  $\alpha$ -dystroglycan. *Glycobiology* **22**, 662-675 (2012).
- 12 Stalnaker, S. H. *et al.* Site mapping and characterization of O-glycan structures on  $\alpha$ -dystroglycan isolated from rabbit skeletal muscle. *J Biol Chem* **285**, 24882-24891 (2010).
- 13 Halim, A., Nilsson, J., Ruetschi, U., Hesse, C. & Larson, G. Human urinary glycoproteomics; attachment site specific analysis of N- and O-linked glycosylations by CID and ECD. *Mol Cell Proteomics* **11**, M111 013649 (2012).
- 14 Bagdonaite, I. *et al.* Glycoproteomics. *Nat. Rev. Methods Primers* **2**, 48 (2022).
- 15 Gao, C. *et al.* Glycan Microarrays as Chemical tools for identifying glycan recognition by immune proteins. *Front Chem* **7**, 833 (2019).
- 16 Peng, W., Nycholat, C. M. & Razi, N. Glycan microarray screening assay for glycosyltransferase specificities. *Methods Mol Biol* **1022**, 1-14 (2013).
- 17 Laurent, N., Voglmeir, J. & Flitsch, S. L. Glycoarrays' tools for determining protein-carbohydrate interactions and glycoenzyme specificity. *Chem Commun (Camb)*, 4400-4412 (2008).
- 18 Murrell, M. P., Yarema, K. J. & Levchenko, A. The systems biology of glycosylation. *Chembiochem* **5**, 1334-1347 (2004).
- 19 Mikkola, S. Nucleotide sugars in chemistry and biology. *Molecules* **25** (2020).
- 20 Fisher, P. & Ungar, D. Bridging the gap between glycosylation and vesicle traffic. *Front Cell Dev Biol* **4**, 15 (2016).
- 21 Liu, L., Doray, B. & Kornfeld, S. Recycling of Golgi glycosyltransferases requires direct binding to coatomer. *Proc Natl Acad Sci U S A* **115**, 8984-8989 (2018).
- 22 Gill, D. J., Chia, J., Senewiratne, J. & Bard, F. Regulation of O-glycosylation through Golgi-to-ER relocation of initiation enzymes. *J Cell Biol* **189**, 843-858 (2010).

- 23 Guo, H. & Pierce, J. M. in *Glycoscience: Biology and Medicine* (eds Tamao Endo *et al.*) 1-7 (Springer Japan, 2021).
- 24 Ashkani, J. & Naidoo, K. J. Glycosyltransferase gene expression profiles classify cancer types and propose prognostic subtypes. *Sci Rep* **6**, 26451 (2016).
- 25 Neelamegham, S. & Liu, G. Systems glycobiology: biochemical reaction networks regulating glycan structure and function. *Glycobiology* **21**, 1541-1553 (2011).
- 26 Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**, D490-495 (2014).
- 27 Aoki-Kinoshita, K. F. Glycome informatics: using systems biology to gain mechanistic insights into glycan biosynthesis. *Current Opinion in Chemical Engineering* **32**, 100683 (2021).
- 28 Brockhausen, I. & Melamed, J. Mucins as anti-cancer targets: perspectives of the glycobiologist. *Glycoconjugate J.* **38**, 459-474 (2021).
- 29 Berois, N., Pittini, A. & Osinaga, E. Targeting tumor glycans for cancer therapy: successes, limitations, and perspectives. *Cancers (Basel)* **14** (2022).
- 30 Edwards, E. *et al.* Strategies to control therapeutic antibody glycosylation during bioprocessing: synthesis and separation. *Biotechnol Bioeng* **119**, 1343-1358 (2022).
- 31 Moremen, K. W. *et al.* Expression system for structural and functional studies of human glycosylation enzymes. *Nat Chem Biol* **14**, 156-162 (2018).
- 32 Gosselin, S., Alhussaini, M., Streiff, M. B., Takabayashi, K. & Palcic, M. M. A continuous spectrophotometric assay for glycosyltransferases. *Anal Biochem* **220**, 92-97 (1994).
- 33 Ortiz-Soto, M. E., Reising, S., Schlosser, A. & Seibel, J. Structural and functional role of disulphide bonds and substrate binding residues of the human  $\beta$ -galactoside  $\alpha$ -2,3-sialyltransferase 1 (hST3Gal1). *Sci Rep* **9**, 17993 (2019).
- 34 Yoshimura, Y. *et al.* Products of Chemoenzymatic synthesis representing MUC1 tandem repeat unit with T-, ST- or STn-antigen revealed distinct specificities of anti-MUC1 antibodies. *Sci Rep* **9**, 16641 (2019).

# **Chapter 2:**

## **Kinetics and Glycosylation Reaction Assays**

## Chapter Overview

This chapter is an introduction to the enzymatic coupled assay designed in this thesis (Chapter 3) to assess the activities of glycosyltransferase. The chapter includes presentation of the thermodynamics and kinetics principles governing enzymatic reactions. The mechanisms and rate laws of the enzymatic Bi-Bi reactions, which represent the class of glycosyltransferase enzymes are demonstrated. The chapter also reviews the applications and limitations of the currently used assays to study glycosylation reactions, Enzymatic coupled assays, and their application in studying glycosyltransferase activity are covered in detail as they form the grounding principles of the assay developed in this thesis. Finally, the chapter provides a brief description of the development carried out in this thesis on the current enzymatic coupled assay, PK/LDH.

## 2.1 Enzymatically catalyzed reactions

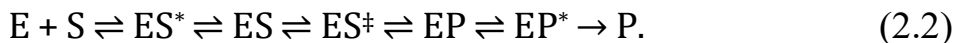
### 2.1.1 Overview

According to the transition state theory: For any chemical reaction to occur, the reactants must undergo transformation into a higher energy conformation before converting to products. This high energy conformation is called transition state. Thermodynamically, the transition state is quantified by an energy barrier that the reactants need to overcome called Gibbs free activation energy. Only the reactants that acquire enough energy to overcome Gibbs free activation energy succeed in transforming to transition state followed by conversion to products. Since the transition state is the highest energy state of the reaction species, transformation to this state is the slowest, and thus the rate-limiting step in the reaction. That implies that the higher the free activation energy the slower the reaction is. In fact, the rate of a reaction is a negative exponential function of the free activation energy as expressed in Eyring equation. Catalysts, for example, enzymes, increase the reaction rate by lowering the free activation energy<sup>1</sup>.

The catalytic function of enzymes can be mechanistically explained by studying a simple enzymatic reaction where a single substrate is converted to a single product, presented as:



Even for such a simple reaction, the mechanism is a multistep process involving multiple intermediates: ES, EP, and multiple transition states:  $ES^*$ ,  $ES^\ddagger$ , and  $EP^*$ . The mechanism can be presented as follows:



The reaction is initiated by binding of the substrate to the enzyme's active site in a rapidly equilibrated association to form ES complex. This binding requires activation of the molecules through a transition state  $ES^*$ . Upon binding, the substrate undergoes a conversion to product via an activation of ES complex to a short-lived transition state  $ES^\ddagger$ , followed by rapid conversion to product in bond form to the enzyme EP. The dissociation of the product from the enzyme then occurs at the final step (turnover step). Even the turnover step requires a transition into a higher energy conformation  $EP^*$ <sup>2</sup>. The diagram illustrated in Figure 2.1 shows the free energy profile of the reaction progress.

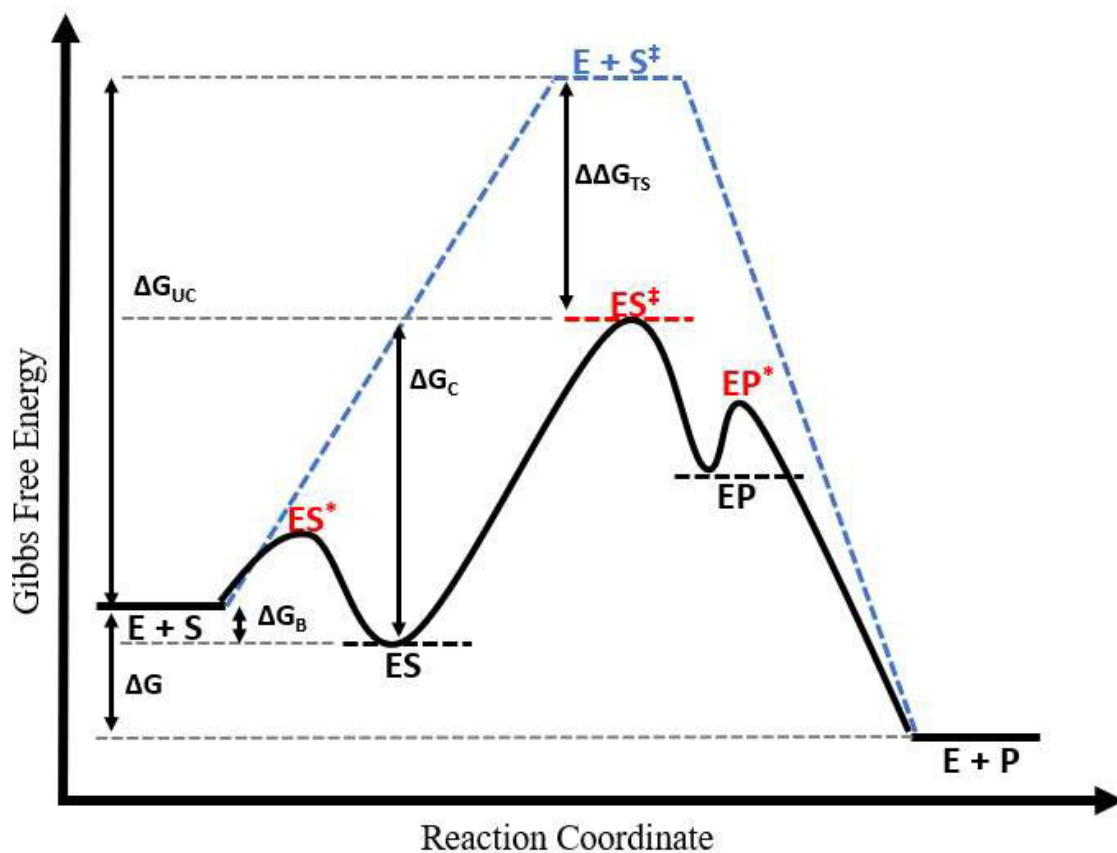
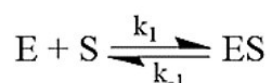


Figure 2.1: Gibbs free energy profile of the simple reaction  $S \rightarrow P$ : The dashed blue line represents the uncatalyzed reaction, whereas the bold line represents the enzymatically catalyzed reaction. High energy transition states are represented in red, reaction intermediates, substrates, and products in black. The enzymatic reaction involves the following steps: binding of the enzyme to the substrate to form ES complex and release of energy  $\Delta G_B$ , chemical transformation of the substrate in ES complex to produce the transition state  $ES^\ddagger$  by crossing the barrier of free energy of activation  $\Delta G_C$ , the final step is the release of the product P. The binding of the substrate and the release of the product requires activation into transition states,  $ES^*$  and  $EP^*$ , respectively. The uncatalyzed reaction involves a single step of chemical transformation to the transition state over a  $\Delta G_{UC}$  energy barrier. The reduction of the free activation energy via enzyme catalysis is represented as  $\Delta\Delta G_{TS}$ .

### 2.1.2 Equilibrium Thermodynamics

As mentioned, the enzymatic reaction presented in Figure 2.1 includes binding of the substrate to the enzyme in a rapid equilibrium. To understand the concept of equilibrium in the context of free energy, the principles of equilibrium thermodynamics need to be understood. If we take the equilibrium step in isolation, we can write the reaction as follows:



$k_1$  and  $k_{-1}$  are the rate constants of the forward and reverse directions. The rate of the forward direction is  $V_1 = k_1[E][S]$  and for the reverse direction  $V_2 = k_{-1}[ES]$ . The units of  $k_1$  and  $k_{-1}$  are  $M^{-1}.S^{-1}$  and  $S^{-1}$ , respectively. Any transformation requires overcoming an energy barrier that is called free activation energy. The rate of this transformation is a negative exponential function of the free activation energy i.e., the higher the barrier, the slower the transformation is. The free energy of the species defines its stability; the lower the free energy of a species is, the more stable it is. If we take the forward reaction above, the free activation energy associated with  $k_1$  is equal to the difference of the free energy between reactants  $E + S$  and  $ES^*$ . The reverse reaction associated with  $k_{-1}$ , requires crossing another barrier ( $ES$  to  $ES^*$ ). Despite having the same transition state, these two barriers are not the same. If the free energy of the reactants is higher than the complex  $ES$ , as in Figure 2.1, the barrier for the association is lower, therefore  $k_1$  is larger than  $k_{-1}$ . In any given moment, the value of the ratio  $Q = [ES]/[E].[S]$  determines to what extent a reaction proceeds in a certain direction from initial concentrations before no more work can be produced by the system. Assuming a reaction starts from initial conditions where  $ES$  is absent, then  $Q = 0$ , the reaction proceeds in  $k_1$  direction in a rate  $V_1$  and  $ES$  concentration builds, and so does  $V_2$  and  $Q$  up until  $Q = [ES]/[E].[S] = k_1/k_{-1} = K_{eq}$ , which is the equilibrium constant. The speed with which the equilibrium is reached is dependent on  $k_1$  and  $k_{-1}$ . The higher  $k_1$  and  $k_{-1}$  are, the lower the activation energy is, and the faster the equilibrium is attained. The ratio  $K_{eq} = k_1/k_{-1}$  determines the favourable direction of the equilibrium. To summarize,  $\Delta G_B$  determines  $K_{eq}$ , and the free activation energy determines the speed of the equilibrium. The equilibrium in

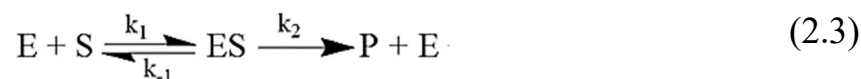
Figure 2.1 is fast with forward preference. The equilibrium constant, also referred to as the association constant, has the unit  $M^{-1}$ . Practically, another constant  $K_d = 1/K_{eq}$  is used to report binding affinity, called the dissociation constant and it has a dimension of  $M$ .

## 2.2 Michaelis-Menten Kinetics

### 2.2.1 Irreversible reaction

The behavior of enzymes as reaction regulators expands beyond a mere catalytic function. Unlike chemical catalysts, enzymes are proteins with evolutionarily optimized structures to promote specific binding and facilitate chemical reactions. The unique amino acid sequence at the binding sites and catalytic centers determines the specificity in substrate binding and reaction, and the flexible and dynamic structure promotes the chemical transformation and stabilization of the transitional state. Enzymes are also sensitive to the microenvironment and their activity is modulated in response to environment markers such as  $pH$ , temperature, and presence of other molecules and ions.

To study the kinetics of an enzymatically catalysed reaction, the simplified reaction above (2.1) will be used. From the graph above (Figure 2.1), the mechanism of the reaction can be reduced to two steps. The first step is the formation of the ES complex, which is incomplete (assumption 1), and exists in equilibrium. This can be explained by the nature of weak bonds that constitute this complex and represented by a small free energy change  $\Delta G_B$ . The second step is the conversion of the substrate to product from ES complex, which is associated with a large reduction in the free energy ( $\Delta G - \Delta G_B$ ). Therefore, this step progresses forward to completion, or near completion. Therefore, we can conveniently assume an irreversible conversion (assumption 2). The sum of the two steps of the global reaction can be represented as follows:



Where  $k_1$ ,  $k_{-1}$ , and  $k_2$  are the reaction constants of enzyme-substrate association, enzyme-substrate dissociation, and substrate to product conversion, respectively. The differential equations with respect to time for individual reactions are expressed as follows:

$$\frac{d[E]}{dt} = -k_1[E][S] + (k_{-1} + k_2)[ES] \quad (2.4)$$

$$\frac{d[S]}{dt} = -k_1[E][S] + k_{-1}[ES] \quad (2.5)$$

$$\frac{d[ES]}{dt} = k_1[E][S] - (k_{-1} + k_2)[ES] \quad (2.6)$$

$$\frac{d[P]}{dt} = k_2[ES] = v, \quad (2.7)$$

Where  $v$  is the overall reaction rate and is equal to the rate of product generation,  $[E]$ ,  $[S]$ , and  $[ES]$  are the concentrations of the reaction species at any given moment. From assumption 1 and by applying the mass conservation principle, and at any given moment we can write:

$$[E]_0 = [E] + [ES] \quad (2.8)$$

Where  $[E]_0$  is the starting concentration of the enzyme. The reaction (2.3) undergoes three phases: phase 1, the lag phase, is the time required to build up the ES complex. The rate of the reaction in the lag phase increases rapidly until an equilibrium balancing the formation and depletion of ES complex is reached. The length of the lag phase is dependent on the relative values  $k_1$ ,  $k_{-1}$ , and  $k_2$ . Enzymes are used in very small quantities compared to substrates i.e.,  $[E]_0 \ll [S]_0$  (assumption 3), which means that the equilibrium is reached while the substrate concentration remains practically unchanged. After the lag phase and given assumption 3, a middle phase characterized by a quasi-equilibrium is maintained for some time. This phase is called the steady state, in which  $d[ES]/dt = 0$  and  $[S] \approx [S]_0$ . This notion of insignificant change of initial substrate concentration during the steady state is the core assumption in experimental kinetics allowing measuring rates in predetermined concentrations. Steady state holds until a significant drop in  $[S]$  pushes the reaction to enter phase 3, which is characterized by exponential depletion of reaction rate as the substrate drops. Assuming that the reaction in a steady state

(assumption 4), the differential equations (2.4-2.8), with substitutions from (2-8) produce the rate at any moment during steady state can be written as in (2.9):

$$v = \frac{k_2 [E]_0 [S]_0}{\frac{k_{-1} + k_2}{k_1} + [S]_0} \quad (2.9)$$

$$K_m = \frac{k_{-1} + k_2}{k_1} \quad (2.10)$$

$$k_2 [E]_0 = V_{\max} \quad (2.11)$$

$$v = \frac{V_{\max} [S]_0}{K_m + [S]_0} \quad (2.12)$$

$V_{\max}$  is the velocity of the reaction when  $[E] = 0$ . This condition is defined as enzyme saturation with substrate, in other terms when  $[E]_0 = [ES]$ . The constant  $k_2$  is defined as the turnover number or the catalytic constant  $k_{\text{cat}}$ . The expression (2.12) is the Michaelis-Menten equation and  $K_m$  is the Michaelis-Menten constant<sup>3</sup>. From the relation (2.10),  $K_m$  is the ratio of substrate-enzyme complex dissociation constants  $k_{-1}$  and  $k_{\text{cat}}$  to the association constant  $k_1$ .

### 2.2.2 Meaning of $K_m$

$K_m$  represents the ratio of disappearance to the formation of ES, or the stability of the complex. From the nominator and denominator of (2.10): disappearance is the sum of dissociation from the enzyme ( $k_{-1}$ ) and catalytic transformation ( $k_{\text{cat}}$ ). Therefore, it can be used as an indicator of the affinity of the substrate to the enzyme, the smaller the value of  $K_m$  the higher the affinity is. From (2.12),  $K_m$  can also be defined as the concentration of the substrate when the reaction rate is half  $V_{\max}$  (Figure 2.2). When  $[S]_0 \ll K_m$ , (2.12) can be expressed as a linear relation,

$$v = \frac{V_{\max}}{K_m} \cdot [S]_0 = \frac{K_{\text{cat}}}{K_m} \cdot [E]_0 \cdot [S]_0 = K' \cdot [S]_0 \quad (2.13)$$

which is an expression for a first order reaction. We define  $K' = V_{\max}/K_m$  as the pseudo first order constant. When  $[S]_0 \rightarrow \infty$ ,  $v \rightarrow V_{\max}$  and the reaction rate is constant and independent of  $[S]_0$ , which defines a zero-order reaction (Figure 2.2)

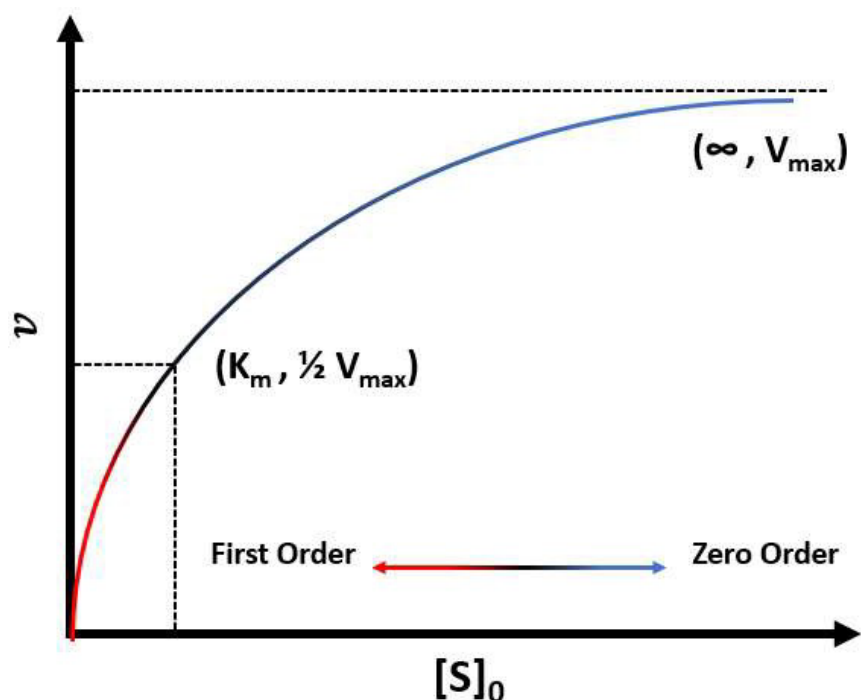


Figure 2.2. Direct Plot. The plot of the reaction rate  $v$  vs substrate concentration  $[S]$  for an enzyme follows Michaelis-Menten's kinetics produces a hyperbolic curve.  $V_{\max}$  is the rate when  $[S] \gg K_m$ . T these conditions, the reaction has a pseudo zero order. When  $[S] \ll K_m$ , the reaction has a pseudo first order. When  $[S] = K_m$  the reaction rate  $v = \frac{1}{2} V_{\max}$

Finally, the ratio  $k_{\text{cat}}/K_m$  is defined as the enzyme specificity constant to its substrate, and the catalytic efficiency of the enzyme.  $k_{\text{cat}}/K_m$  is a useful metric to evaluate the specificity of an enzyme to different substrate.

### 2.2.3 Reversible reaction

The assumption of irreversibility above was made to simplify the derivation and reduce the model in a way that enables definition of the kinetics properties of the forward reaction and individual substrates. However, enzymes in nature are capable of catalyzing the reaction in both directions to varying degrees. In a similar way to the derivation of an irreversible reaction, the derivation of a reversible reaction can be done for a reaction model modified from (2.3):



Following the same steps in deriving an irreversible reaction differential equation (2.4-2,7) and applying the mass conservation principle and the steady state assumption, the rate of a reversible reaction obeying Michaelis-Menten's assumptions is written as:

$$v = \frac{K_{mP}V_1[A] - K_{mA}V_2[P]}{K_{mA}K_{mP} + K_{mP}[A] + K_{mA}[P]} \quad (2.14)$$

$K_{mA}$  and  $K_{mP}$  are Michaelis-Menten constants, which are equal to  $(k_{-1} + k_2)/k_1$  and  $(k_{-1} + k_2)/k_2$ , respectively.  $V_1$  and  $V_2$  are maximum velocities in the forward and reverse direction, respectively. When either the substrate or the product is absent from the reaction mixture, the relation (2.14) returns the same irreversible relationship (2.12). This confirms the validity of the irreversible Michaelis-Menten equation in describing reversible reactions when only the substrate, or the product, is added to the reaction and initial rate conditions hold.

During reaction progress in one direction, the rate of the other direction increases gradually until an equilibrium is achieved when  $V_1 = V_2$  or  $v = 0$ . An important distinction must be made here between the dynamic steady state and static equilibrium. The steady state is when the concentration of the transition complex is relatively constant, whereas the static equilibrium is when  $v = 0$  and concentrations of all species are constant. The equilibrium is defined by the constant  $K_e$ :

$$K_e = \frac{[P]_e[E]_e}{[A]_e[E]_e} = \frac{[P]_e}{[A]_e} = \frac{k_1k_2}{k_{-1}k_{-2}} = \frac{K_{mP}V_1}{K_{mA}V_2} = \frac{K_p}{K_a} \quad (2.15)$$

The constants  $K_p$  and  $K_a$  are the dissociation constants of the product and the substrate, which are equal to  $k_2/k_{-2}$  and  $k_{-1}/k_1$ , respectively.  $[P]_e$  and  $[A]_e$  are the equilibrium concentrations.

#### 2.2.4 Meaning of $K_a$

$K_a$  defines the binding affinity of the substrate to the enzyme independent of the catalytic activity, whereas  $K_m$  is also a definition of the affinity, but taking the catalytic turnover in account.  $K_m = K_a$  when  $k_{cat} \ll k_{-1}$ , or what can be thermodynamically explained by a very high transition state barrier compared to the equilibrium activation barrier. In other words, binding of the substrate to the enzyme is thermodynamically more favourable than the catalytic transformation. This case is defined as a rapid enzymatic reaction.  $K_a$  also infers an inhibitory property of the substrate, and it is a fundamental concept in multi-substrate enzymatic reactions. In fact, the dissociation constants in multi-substrates reactions are referred to as inhibition constants  $K_i$ , and it defines the order of the substrates binding to the enzyme.

### 2.3 Bi-Bi Reactions

#### 2.3.1 Definition and Mechanisms

The simplification of single substrate-single product (Uni-Uni) model used to derive Michaelis-Menten's model helps to shed light on the mechanism of enzyme-substrate interaction and understanding the principle of steady state and the hyperbolic relationship between initial rate and substrate concentration. Most enzymes in nature catalyze reactions converting more than one substrate to more than one product<sup>4</sup>. Two substrates-two products (Bi-Bi reactions) are the most dominant type. The kinetics mechanisms of Bi-Bi reaction vary depending on two variables: the order of substrate binding and products release, and the existence of an inhibitory effect of one or more of the reacting species.

W. Wallace Cleland (1963)<sup>5</sup> described multi-substrate kinetics in detail and provided a nomenclature and a guideline to graphical and mathematical representation of these reactions (Figure 2.3). Despite the complexity and diversity of Bi-Bi reactions, they can be classified based on mechanism into two groups, sequential (single displacement) and Ping-Pong (double

displacement). In sequential reaction, both substrates bind to the enzyme, randomly or in a specific order, before any product is released. The release of products can also be ordered or random. During sequential binding/or releasing, the enzyme forms transitory complexes, such as EA and EP (Figure 2.3 a,b). Binding of both substrates to the enzyme forms a ternary complex before the conversion to product takes place (Figure 2.3 a,b). The ternary complex, also referred to as the central complex, undergoes catalytic transformation that is not represented by arrows, but rather by a dash (and parentheses in some texts) as the kinetics of the catalytic transformation is unmeasurable in kinetics studies. Ping-Pong reactions take place in two stages. Each can be considered as a Uni-Uni reaction with an independent catalytic step. In stage 1 the first substrate, commonly denoted A, binds to the enzyme and a central/catalytic complex is formed, EA-FP in Figure 2.3, c. During the catalysis, a functional group is exchanged between the enzyme and first substrate transferring the enzyme into active form, F. The enzyme remains in its active form in stage 2 until the chemical group is exchanged with the second substrate and the second product is released (Figure 2.3, c).

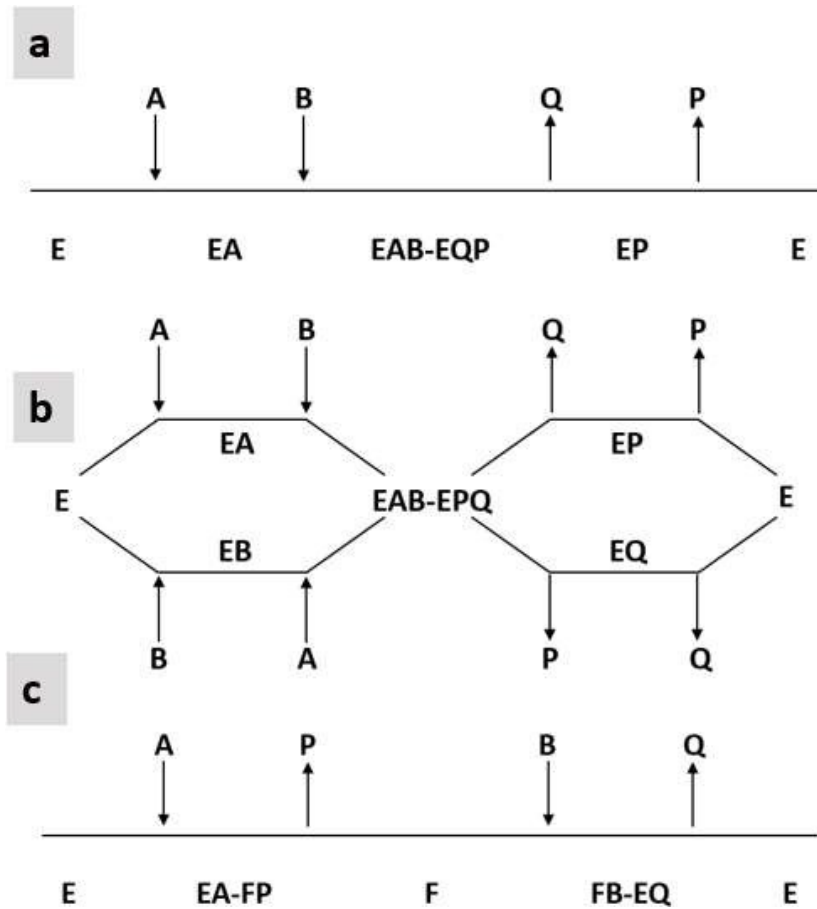


Figure 2.3: Bi-Bi kinetics mechanisms: kinetics mechanisms of sequential ordered (a), sequential random (b), and Ping-Pong Bi-Bi reactions. The lines represent the path the enzyme can take during the reaction course. Arrows with direction represent substrates and products binding and release. A and B are the substrates and P and Q are their associated products, respectively. EA, EB, EP, and EQ are transitory complexes. Central complexes are represented by a combination of substrate(s)-enzyme with their transformed isomers separated by a dash. In Ping-Pong mechanism, the enzyme E is activated to F and remain active form until transforming the second substrate to its product. Constants were omitted for simplification

### 2.3.2 Mathematical Expressions

Derivation of reaction rate equations for multi-substrate reactions, including Bi-Bi reactions, is complicated. Solving differential equations directly (section 2.2.1) results in intricately connected equations that are very difficult to solve. The other problem with this approach is that the constants in the differential equations are difficult to arrange in the final solution in the form of cofactors that show symmetry and represent kinetics properties. To solve this problem, a widely accepted method was developed by King and Altman (1956)<sup>6</sup> to derive equations for multi-substrate enzymatic reactions in the steady state. The method relies on solving a system of linear equations using the determinant method. The reaction is represented in a closed polygon where the different enzyme binding states constitute the nodes. The linear equations are extracted from the polygon and the terms and their coefficients in the numerator and the denominator of the solution are identified. The method is simplified and explained step by step in textbooks<sup>3</sup> and will not be covered in this thesis. However, for every reaction mechanism, King-Altman's method produces a reaction rate equation containing concentration terms and coefficients transformed as combinations of kinetics constants. The kinetics constants are: Michaelis-Menten constants, dissociation constants, maximum velocities for both directions, and equilibrium constant of the overall reaction. Kinetics constants in a multi-substrate reaction deviate from the simple mathematical expressions presented for single substrate in sections (2.2.1) and (2.2.2). The mathematical expressions are mechanism-dependent, and King-Altman method set rules do define them. Nevertheless, the functional definition of the kinetics constants remains the same. However, a distinction is made between Michaelis-Menten and dissociation constants. Michaelis-Menten constants define the interaction of the substrates and the products with the central complex, whereas dissociation complexes define the interaction with the free enzyme. Dissociation complexes are denoted as  $K_i$ , or inhibition constant, implying inhibitory function of the binding.

Despite the validity and systematic methodology of King-Altman derivation, reaction rate expressions produced by this method are still impractical for two reasons: Most of the enzymes are not fully characterized mechanistically and their kinetics models are uncertain. The second reason is that even for specified mechanisms, the model requires identification of a large suite of parameters, such as four  $K_m$  constants, four  $K_i$  constants, and two  $V_{max}$  for Bi-Bi reactions, which

is not attainable experimentally, for  $K_i$  values in particular. Practically, simplifications can be made for specific applications, for example when the concentration of one substrate (or product) is constant or added in saturating amounts, or when initial conditions permit the assumption of an irreversible reaction. Rate expressions in these special cases can be reduced to the expressions similar to 2.12 or 2.14. When simplification is not an option, or when the assumptions do not hold during the course of the simulation in case of modelling the reaction in a wide range of concentrations, generic equations defining higher level mechanisms have been used. The general aim for these methods is to reduce the parameters and generalize the expression. One example is generic rate laws proposed by (Rohwer et.al., 2006)<sup>4</sup>. The generic equation of this study assumes that all the reaction species binds with the enzyme in equilibrium and this binding is independent of the enzyme's prior binding status. While the numerator is uniform, the denominator consists of two factors each having a combination of terms including one substrate and its corresponding product i.e., the product that occupy the same site on the enzyme. The generic equation was tested against King-Altman rates expressions for ordered and Ping-Pong Bi-Bi mechanisms and the performance was similar in the selected range of kinetics parameters<sup>4</sup>.

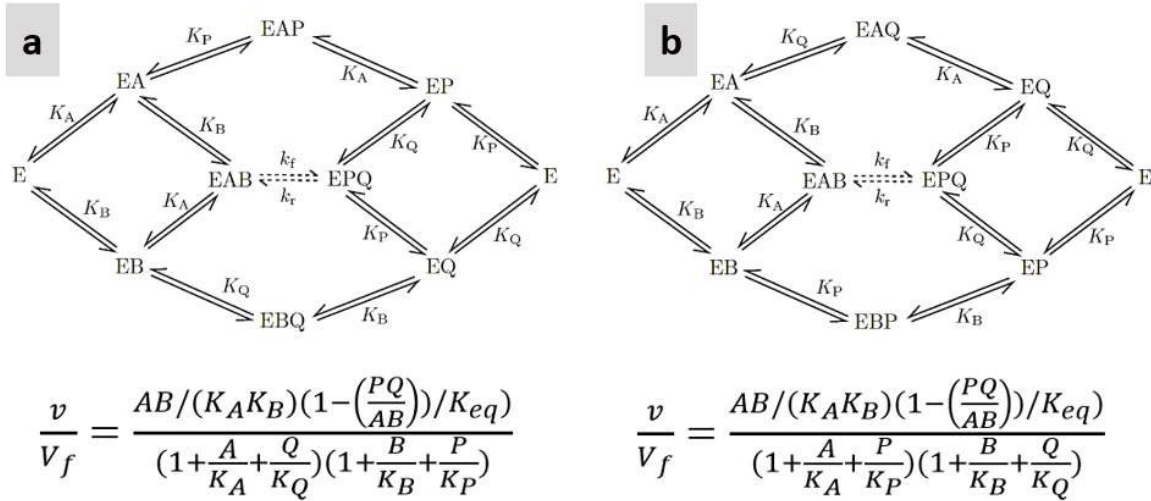


Figure 2.4: A schematic representation of the generic equation proposed by Rohwer et al., 2006)<sup>4</sup>. (a) represents the mechanism of which the substrate Q is the product of A and P is the product of B. (b) is the mechanism when P is the product of A and Q is the product of B. The rate expression for each case is presented under each scheme. A, B, P, and Q are the concentrations of the reactants.  $K_A$ ,  $K_B$ ,  $K_P$ , and  $K_Q$  are Michaelis-Menten constants.  $V_f$  is the maximum velocity in the forward direction.  $K_{eq}$  is the equilibrium constant for the overall reaction

Many other groups attempted to reduce or modify the parameterization of Bi-Bi and they all ended up with rate expressions that are left with at least one  $K_i$  parameter for one of the reactants or with complicated parameters defined by functions of the concentration of the reactants<sup>7,8</sup>.

Another approach is to reduce the mechanism to the following: reversibility, Michaelis-Menten conformity, and the number of central complexes<sup>9</sup>. This approach leads to a rate law as general as to include all reversible Bi-Bi reaction following sequential reactions (single central complex), and another one for Ping-Pong equations (two central complexes) 2.16 and 2.17, respectively.<sup>9</sup>

In both equations P is the product of substrate B and Q is the product of substrate A.

$$K_P \quad K_Q$$

In summary, there is no one solution to fit all applications due to the trade-off between accuracy on one hand, which requires more information on the detailed mechanism and parameters, and simplicity on the other hand, which entails the risk of over-generalization.

$$v = \frac{\frac{V_f AB}{K_A K_B} - \frac{V_r PQ}{K_P K_Q}}{1 + \frac{A}{K_A} + \frac{B}{K_B} + \frac{P}{K_P} + \frac{Q}{K_Q} + \frac{AB}{K_A K_B} + \frac{AP}{K_A K_P} + \frac{BQ}{K_B K_Q} + \frac{PQ}{K_P K_Q} + \frac{ABP}{K_A K_B K_P} + \frac{BPQ}{K_B K_P K_Q}} \quad (2.16)$$

$$v = \frac{\frac{V_f AB}{K_A K_B} - \frac{V_r PQ}{K_P K_Q}}{\frac{A}{K_A} + \frac{B}{K_B} + \frac{P}{K_P} + \frac{Q}{K_Q} + \frac{AB}{K_A K_B} + \frac{AP}{K_A K_P} + \frac{BQ}{K_B K_Q} + \frac{PQ}{K_P K_Q}} \quad (2.17)$$

### 2.3.3 Identifying Bi-Bi Mechanisms and Parameters Experimentally

Regardless of the mechanism of a Bi-Bi reaction and the mathematical expression used, kinetics parameters can be identified experimentally by adjusting the initial conditions to achieve pseudo-single-substrate conditions. Initial conditions of an infinite concentration of A or B and absence of P and Q transform any Bi-Bi expression to that in (2.12), where  $V_{\max} = V_f$ , and  $K_m = K_A$  if A is the variable substrate, or  $K_m = K_B$  if B is the variable substrate. The same transformation is valid for the reverse reaction at initial conditions of:  $A=B=0$  and P or Q is variable, and the other product is infinite. This transformation enables experimental identification of  $K_m$  values for all the reactants and  $V_f$  and  $V_r$  by alternating the conditions above. In addition to  $K_m$  and  $V_{\max}$  values, diagnostic approaches are used to identify the mechanism of a Bi-Bi reaction and calculate  $K_i$  values. These methods are more tedious than identifying  $K_m$  and  $V_{\max}$  and require generating series of dose responses of one substrate at different fixed concentrations of the other reactants and evaluate the intercepts and the patterns of the series in linear expressions. Details of these methods can be found in kinetics textbooks and will not be covered in this thesis<sup>3,10</sup>

### 2.3.3 GT Reactions

Glycosyltransferases (GTs) are examples of enzymes that catalyse reactions by Bi-Bi mechanisms. Classification of GTs based on their mechanisms was introduced in a review by Lairson *et al.*, (2008)<sup>11</sup>. GTs catalyse the transfer of a sugar moiety from an activated sugar nucleotide to an acceptor. The exact mechanism of which this transfer occurs is nucleophilic substitution in which the acceptor is the nucleophile, the sugar moiety is the electrophile, and the phosphate on the nucleotide is the leaving group. GTs can be classified into retaining or inverting enzymes with regard to the stereochemical confirmation of the product. Inverting GTs catalyze an  $S_N2$  reaction where both the nucleophile (acceptor) and the electrophile (nucleotide sugar donor) exist simultaneously in the enzyme's catalytic site prior to the catalytic transformation. The transformation is characterized by formation of an oxocarbenium ion-like transition state. This mechanism fits the definition of a single displacement mechanism of Bi-Bi reaction, which can be either random or sequential. As for the retaining GTs, the nucleotide sugar binds to the enzyme first then the phosphate group leaves before the acceptor binds and the catalytic conversion takes, which describes a double displacement mechanism. This double displacement mechanism is called  $S_N1$  nucleophilic substitution and fits the definition of a Ping-Pong Bi-Bi reaction. Typically,  $S_N1$  reactions result in a racemic mixture of the product. However, in retaining GTs, only one enantiomer with the original optical configuration of the nucleotide-sugar is produced. The inversion is prevented in this mechanism by the binding of anomeric carbon of the donor's sugar with the enzyme on the opposite site of the phosphate group. This binding blocks the access of the nucleophile and restricted to the side of the leaving phosphate group<sup>11</sup>. Sialyltransferases are inverting glycosyltransferase enzymes with oxocarbenium ion-like transition state. However, the exact mechanism is not clear. For example, investigating ST3GAL enzymes suggested either random or sequential Bi-Bi mechanism with inhibitory effect of CMP<sup>12</sup>

## 2.4 Enzyme Assays

Enzyme assays are analytical methods that aim to detect one of the reacting species in an enzymatic reaction, which can be either a substrate or a product. Monitoring the depletion of a substrate or generation of a product over time is used to assess enzyme activity. Depending on the aim of the experiment, enzyme activities can be tested to study enzyme behavior in different conditions, such as: different initial concentrations of one or more of the reacting species, different  $pH$ , different temperature, or in presence or absence of a modulator or a cofactor. In addition to studying enzymes behaviors, some assays can also be used to detect and quantify an undetectable substrate.

Enzyme assays are fundamental tools in enzymology and drug discovery. One main application in enzymology is identifying kinetics parameters and reaction mechanisms of an enzyme (section 2.3.3). In early drug discovery stages, enzyme assays are optimized in a high throughput screening (HTS) format to screen libraries of inhibitor candidates. The assay in HTS is designed where the substrates are added at concentrations determined in relation to their  $K_m$  values and the targeted inhibitory mechanism<sup>13</sup>. The progress of the reaction with and without drug candidates (added at a single standardized dose) is recorded and analyzed. The candidates are then ranked according to their inhibitory effect, and those that show inhibition above a certain threshold are selected and their dose responses are studied. The potency of the candidates is measured by their half maximal inhibitory concentration ( $IC_{50}$ ). In a following stage, candidates that pass certain threshold of  $IC_{50}$  can be more intensively investigated in term of their inhibitory mechanism<sup>14</sup>.

### 2.4.1 Assay Types

Measuring the concentration of the reaction species in enzyme assays can be carried out in real-time as the reaction is progressing. Assays carried out in this format are called continuous assays. On the other hand, when the measurement requires terminating the reaction to develop a signal from the reaction species, the assay is called an endpoint assay. Continuous assays are more economic and deliver more information from one reaction mixture compared to endpoint assays. For instance, the main disadvantage of an endpoint assay is the difficulty in identifying the linear range of a progress curve, which correspond to the steady state. The process of investigating the progress curve requires repetitive sampling with short intervals, which can be very challenging with endpoint assays. On the other hand, continuous assays detect the progress curve in real time

and the linear range is identified in one reaction mixture. As mentioned above, HTS format detects the effect of the screened compounds on enzyme activity, which is assessed by detecting the slowing of the signal change associated with reaction progress. Some compounds interfere with the signal and produce inhibition-independent signal shift. These shifts can lead to false interpretation of inhibition (or activation). Continuous assays are superior to endpoint assays in detecting these false signals because of their ability to detect signal interference at initiation and throughout the progress and distinguish between true inhibition and background shift. In other words, continuous assays have an internal control capacity in each reaction mixture<sup>15</sup>. Another aspect of the enzyme assays is the signal used to detect the reporter species. The two most dominant modalities are absorbance and fluorescence. Each has its advantages and disadvantages. For instance, fluorescence is more sensitive, but with limited linearity at higher concentrations of the fluorophore. Conversely, the linearity of absorbance holds for higher ranges of the analyte but is less sensitive than fluorescence<sup>15,16</sup>. Therefore, preference is context-dependent, and it will be discussed more in Chapter 3 when the NADH signal is optimized.

#### **2.4.2 Assay Optimization and Performance**

Development of enzymatic assays requires investigating the performance in different combinations of variables (assay parameters). These parameters can be *pH*, time, temperature, buffers, additives, enzyme concentration, and ions. The list of assay parameters varies depends on the assay specifications and it is longer with complex assay, such as in enzyme-coupled assays. Three approaches are used to achieve that, Trial and Error (TAE), one Factor at a Time (OFAT), or Design of Experiments (DoE). In trial and error, an experienced researcher may decide to rationally start with one assembly of values of parameters and adjust the parameters according to the observed response. TAE is the least systematic and most intuitive approach. In OFAT approach, each parameter is optimized while the other parameters are constant. This process is systematic and delivers the best possible outcome but lengthy and it may require some iterations to achieve an optimum combination of parameters<sup>17</sup>.

Alternatively, when a large number of parameters with a wide range of values are to be optimized, DoE is the most practical and efficient approach. For *n* parameters to be optimized, DoE determines a small number of values to test for each parameter, typically two values. The number of the total possible assemblies is given as a factorial  $2^n$ . Even with minimizing the

number of values, testing the whole factorial remains impractical when the number of parameters is large. Therefore, DoE approach employs statistical methods to extract a subset from the assemblies' factorial that guarantee equal representation of all the parameters. This approach is called a fractional factorial approach. DoE is fast and practical, and it can be systematically applied to any assay without much prior knowledge of the effect of parameters on performance, where only two values for each parameter need to be determined<sup>17-20</sup>.

Assay performance is a broad term with a definition that differs with the intended application. Aspects of performance can vary, such as cost, biological relevance, accuracy, specificity, miniaturization adaptability, etc. For example, if the assay is designed to study enzymes kinetics at different ranges of concentrations with different substrates, sensitivity and linearity range of the signal are particularly emphasized aspects required for an accurate calculation of initial rates. In addition to aspects of performance, an optimum enzymatic assay can maintain its performance in a wide range of conditions with enough tolerance to assay parameters changes (robustness of the assay). This tolerance to change in a method's parameters enhances the reproducibility of the results when the assay is adapted across labs and with modified conditions.

While robustness is not quantifiable, it is a function of combination of performance aspects that can be quantified by metrics. A review by Inglese *et al.*, (2007)<sup>21</sup> summarized these metrics. One major metric is  $Z'$  score, which defines the statistically significant window of the assay between two measurements.  $Z'$  is calculated from the expression (2.18); where  $\sigma_{\max}$  and  $\sigma_{\min}$  are the standard deviation of a high and low signal measurements, respectively,  $\mu_{\max}$  and  $\mu_{\min}$  are the means of these measurements. The closer  $Z'$  to 1, the wider the significance window is with a lower dispersion of readings. This means that assays with  $Z'$  approaching 1 are less sensitive to outlier readings, and therefore more tolerant to variation in the method i.e., more robust. When  $Z' < 0$ , the distributions of the two values overlap and the assay is invalid in this window. More details on the interpretation of  $Z'$  can be found in the original paper<sup>22</sup>.

$$Z' = 1 - \frac{3\sigma_{\max} + 3\sigma_{\min}}{|\mu_{\max} - \mu_{\min}|} \quad (2.18)$$

## 2.5 Currently used assays for GT

The main challenge in developing enzyme assays for GT enzymes is that none of the substrates (donor and acceptor) or the products have a differentiating optical property, fluorescence or absorbance. This limits the effort of development to integrating isolation steps<sup>23</sup> and/or to chemically modify the substrates or the products. The modification can either take place before or after the reaction. Modifications that take place after the reaction, as well as isolation-based methods, typically requires termination of the reaction to develop the signal, which results in an endpoint format. Some assays are based on modifying the substrates by introducing a fluorophore or a mediator group, of which some succeeded to develop continuous assays. However, using chemically modified substrate compromises the biological significance of the assay and limits the applicability to specific class of enzymes or substrates. The methods and their limitations were presented in detail in a review by Wangner and Pensot, (2010)<sup>24</sup>. Since Wagner and Pensot review was published, no significant advancement on GT assays has been made.

Most of the assays used for GT have limitations of either being endpoint assays or using modified substrates in a continuous mode; nevertheless, there are some attractive approaches that have been developed to assess activities of GTs in real time using native substrates. One method was developed to exploit the change of *pH* caused by the proton release accompanied the GT reaction<sup>25</sup>. However, relying on *pH* changes requires allowance of *pH* change by eliminating the buffer. It is well established that *pH* is a determinant parameter for enzyme activity, and it is a regulatory mechanism in cells, since the *pH* gradient of the Golgi determines the activity of GTs. Therefore, variation in *pH* compromises the biorelevance, as well as the precision, of the assay. Another method based on mass spectrometry was introduced in 2005<sup>26</sup>. In this method, GT reaction of native substrates is subjected to ESI-TOF analysis in near real time by an automated process allowing for periodic injections from the reaction mixture. The method also requires parallel injection of the nucleotide product as an internal control for quantification. This injection is carried out simultaneously and independently of the reaction mixture via a complex instrumental configuration aided by an HPLC auto sampler. The method was successful in obtaining kinetics parameters for the selected GT reactions and was proved to be applicable in

HTS format. However, the instrumentation demand and the complexity of the method's parameters makes it very difficult to adapt and may compromise the robustness of the method.

The criteria of an ideal GT assay that overcomes the limitation above is as follows: an assay that is robust, continuous, uses native substrates, can be adapted to all classes of GT enzymes, can be miniaturized for HTS applications, and is easy to perform using common instruments. The bottleneck of the field lies in finding a detection method that does not require separation, does not interfere with the reaction, and can be performed using simple instruments. Two approaches have proven superior in the field in terms of meeting these criteria: Using a fluorescent tracer/sensor, and enzymatic coupled assays.

### **2.5.1 Biosensor-Based Assays**

The biosensor developed by Wongkongkatep *et al.*<sup>27</sup> is a fluorophore that binds specifically to pyrophosphate monoesters (nucleotide diphosphate) and has low affinity to diesters (sugar-nucleotide diphosphate). This binding enhances its fluorescence proportionally to the released diphosphate nucleotide<sup>27</sup>. In a similar approach<sup>28</sup> BellBrook laboratories developed the commercial assay, Transcreen™. In this assay, a tracer consisting of a nucleotide conjugated with a fluorophore that binds to an antibody was used. This binding slows down the rotation of the tracer due to the size increase of the complex. When a fluorophore is excited by a polarized light, the retention of polarization in the same plane for emitted light is negatively correlated to the molecule rotation speed. Therefore, the antibody-bound tracer retains the polarization better than the free tracer. The principle of the assay is that the nucleotide released from the reaction replaces the tracer and reduce the polarization of the emitted light. This method is called competitive fluorescence polarization immunoassay (FPIA). The main caveats of Transcreen™ assays are the cross-affinity of the antibody to the donor and limited linearity of detection due to the nature of the competitive binding between the tracer and the released nucleotide. These limitations can be identified, and the assay can be adjusted via a careful optimization. However, these optimized conditions may not hold when a broad range of substrate is used, such as in kinetics studies. Another major limitation of the assay is its high cost.

### 2.5.2 Enzyme-Coupled assay

To study the kinetics of the sialyltransferase, ST3GAL1, Ortiz-Soto *et al.* (2019)<sup>29</sup> modified a continuous enzyme-coupled assay that was originally developed by Gosselin *et al.* (1994)<sup>30</sup> and optimized by Brown *et al.* (2012)<sup>31</sup> as a generic NDP-based GT assay. This assay is based on coupling the NDP (UDP or GDP) released from glycosyltransferase reaction with pyruvate kinase (PK) directly in case of NDP-based donor reactions (reaction 2, Figure 2.5 a). In case of the CMP-based reactions, the coupling is carried out via another kinase, nucleoside monophosphate kinase (NMPK), that phosphorylates CMP to CDP in presence of excess ATP (reaction 2, Figure 2.5 b). The resulting influx of UDP or ADP to PK reaction converts phosphoenolpyruvate (PEP) to pyruvate (PK reaction, Figure 2.5 a and b), which is used to oxidize NADH by lactate dehydrogenase (LDH). The depletion of NADH can be measured in real time by monitoring the absorbance at 340 nm, and the progress of GT reaction can be calculated from the drop of NADH concentration in the reaction mixture (Last reaction in panel a and b, Figure 2.5).

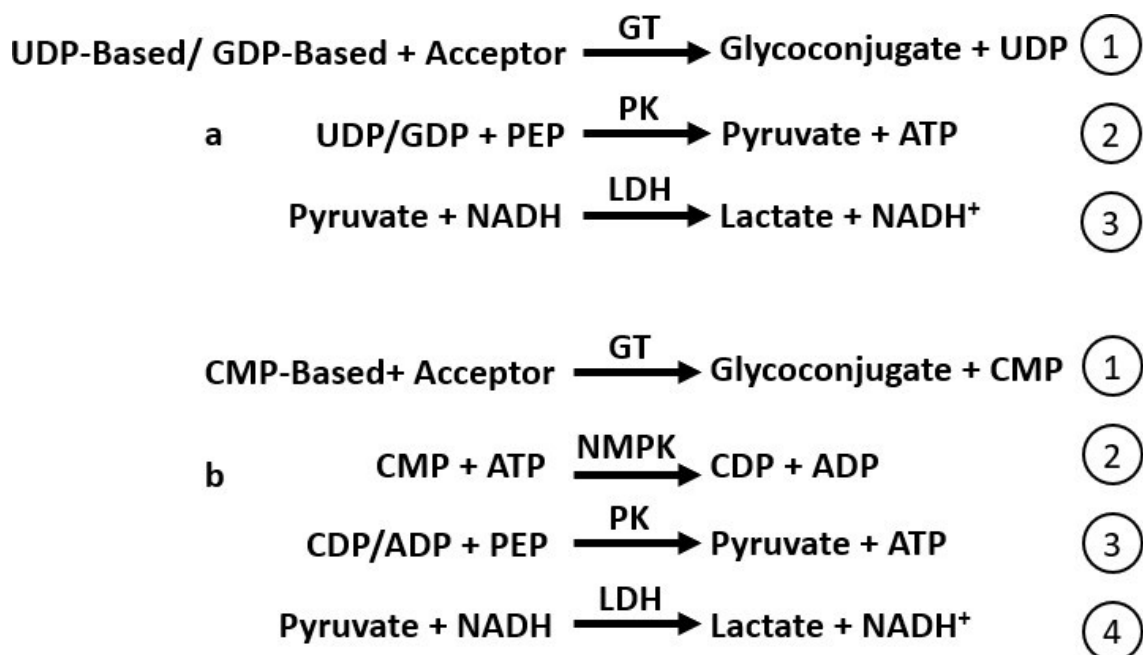


Figure 2.5: Enzymatic-coupled assays developed previously to study GTs kinetics. (a) NDP-sugar donor reaction. (b) CMP-sugar donor reactions. GT (glycosyl transferase), PK (pyruvate kinase), LDH (lactate dehydrogenase), NMPK (nucleoside monophosphate kinase)

The availability of the PK and LDH enzymes and their substrates, in addition to the simplicity of the spectrophotometry measurement and the potential broad utility of PK/LDH coupling assay, makes it an attractive approach to develop an optimum assay for GT enzymes. Despite the potential of the enzyme-coupled assay as a label-free continuous GT-generic assay, there has not been any effort to investigate and optimize its performance in a systematic way. Even the articles that used the assay did not include any details on its performance<sup>29</sup>. Furthermore, the potential of utilizing this assay in HTS has not been investigated thoroughly.

## 2.6 Validity of an Enzymatic-Coupled Assay

The main condition to the validity of enzyme-coupling assays as a tool to study an enzymatic reaction is that the reaction of interest (the non-observable reaction) must be the rate limiting step and the subsequent reactions (indicator reactions) must be sensitive and fast enough to detect the smallest amount of the product instantly. This condition assures that for any given initial concentration of the substrate of the non-observable enzyme (any given rate of that reaction), the rates for all the reactions in the chain are equal to that of the non-observable reaction. To achieve that, the enzymes in the subsequent reactions must be chosen to have low  $K_m$  and high  $k_{cat}$  values for their substrates, the intermediate products in the enzyme chain. This state of the system can be described as the system steady state. Another measure that can be taken is to increase the concentrations of the enzymes in the indicator reactions.

Other parameters to be considered when designing a coupled assay is the interaction between the substrates, products, and cofactors of each enzyme with the rest of the system and avoiding any unintended inhibitory effect. Other conditions, such as temperature and  $pH$  must be optimized to ensure that all enzymes operate at their optimal or acceptable level of activity.

## 2.7 Proposed Improvement to PK/LDH assay

The PK/LDH-coupled assay presented above was improved through a systematic bottom-up optimization approach. The optimization points were as follows: 1. Enhancement of the detection sensitivity and specificity of the detection of the reporter molecule, NADH, 2. Optimization of the PK/LDH coupled reaction in term of reactants concentrations, 3. Upgrade the assay to accommodate all classes of GTs, by integrating additional enzymes to PK/LDH that enable the coupling of the reactions of all classes of GTs to the assay, 4. The optimization process included assessment of the performance and determination of the assay window at each coupling step, a process that was aided by simulation and experiment. Details of the development process are presented in Chapter 3.

## References

- 1 Klotz, I. M. Free energy diagrams and concentration profiles for enzyme-catalyzed reactions. *J Chem Educ* **53**, 159-160 (1976).
- 2 Aledo, J. C., Lobo, C. & del Valle, A. E. Energy diagrams for enzyme-catalyzed reactions: Concepts and misconcepts. *Biochem. Mol. Biol. Educ.* **31**, 234-236 (2003).
- 3 Bisswanger, H. (2017). From Reaction Order to the Michaelis–Menten Law: Fundamental Relationships of Enzyme Kinetics. In *Enzyme Kinetics*, H. Bisswanger (Ed.) 55-86 (2017).
- 4 Rohwer, J. M., Hanekom, A. J., Crous, C., Snoep, J. L. & Hofmeyr, J. H. S. Evaluation of a simplified generic bi-substrate rate equation for computational systems biology. *Iee P Syst Biol* **153**, 338-341 (2006).
- 5 Cleland, W. W. The kinetics of enzyme-catalyzed reactions with two or more substrates or products. I. Nomenclature and rate equations. *Biochim Biophys Acta* **67**, 104-137 (1963).
- 6 King, E. L. & Altman, C. A Schematic Method of deriving the rate laws for enzyme-catalyzed reactions. *J. Phys. Chem.* **60**, 1375-1378 (1956).
- 7 Mitchell, D. A. & Krieger, N. Looking through a new lens: Expressing the Ping Pong bi equation in terms of specificity constants. *Biochem Eng J* **178** (2022).
- 8 Lee, L.W., Yin, L., Zhu, X. M., and Ao, P. Generic enzymatic rate equation under living conditions. *J. Biol. Syst.* **15**, 495-514 (2007).
- 9 Imperial, S. & Centelles, J. J. Enzyme kinetic equations of irreversible and reversible reactions in metabolism. *J. Biosci. Med.* **Vol.02No.04**, 6 (2014).
- 10 Cornish-Bowden, A. *Fundamentals of enzyme kinetics*. (John Wiley & Sons, 2013).
- 11 Lairson, L. L., Henrissat, B., Davies, G. J. & Withers, S. G. Glycosyltransferases: structures, functions, and mechanisms. *Annu Rev Biochem* **77**, 521-555 (2008).
- 12 Gupta, R., Matta, K. L. & Neelamegham, S. A systematic analysis of acceptor specificity and reaction kinetics of five human  $\alpha(2,3)$ sialyltransferases: Product inhibition studies illustrate reaction mechanism for ST3Gal-I. *Biochem Biophys Res Commun* **469**, 606-612 (2016).
- 13 Davies, G., Semple, H., McCandless, M., Cairns, J. & Holdgate, G. A. High-throughput mechanism of inhibition. *SLAS Discov* **26**, 248-256 (2021).
- 14 Brooks, H. B. *et al.* in *Assay Guidance Manual* (eds S. Markossian *et al.*) (2004).
- 15 Simeonov, A. *et al.* Quantitative high-throughput screen identifies inhibitors of the *Schistosoma mansoni* redox cascade. *PLoS Negl Trop Dis* **2**, e127 (2008).
- 16 Simeonov, A. & Davis, M. I. in *Assay Guidance Manual* (eds S. Markossian *et al.*) (2004).
- 17 Onyeogaziri, F. C. & Papanephytous, C. A general guide for the optimization of enzyme assay conditions using the design of experiments approach. *SLAS Discov.* **24**, 587-596 (2019).
- 18 Cowan, K. J. *et al.* Utilizing design of experiments to characterize assay robustness. *Bioanalysis* **4**, 2127-2139 (2012).
- 19 Chen, X. C., Zhou, L., Gupta, S. & Civoli, F. Implementation of design of experiments (DOE) in the development and validation of a cell-based bioassay for the detection of anti-drug neutralizing antibodies in human serum. *J Immunol Methods* **376**, 32-45 (2012).

- 20 Joelsson, D., Moravec, P., Troutman, M., Pigeon, J. & DePhillips, P. Optimizing ELISAs for precision and robustness using laboratory automation and statistical design of experiments. *J Immunol Methods* **337**, 35-41 (2008).
- 21 Inglese, J. *et al.* High-throughput screening assays for the identification of chemical probes. *Nat Chem Biol* **3**, 466-479 (2007).
- 22 Zhang, J. H., Chung, T. D. & Oldenburg, K. R. A Simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J Biomol Screen* **4**, 67- 73 (1999).
- 23 Mohideen, F. I. & Kwan, D. H. A “biphasic glycosyltransferase high-throughput screen” identifies novel anthraquinone glycosides in the diversification of phenolic natural products. *Int. J. Biol. Chem. (N. Y., NY, U. S.)* **299**, 102931 (2023).
- 24 Wagner, G. K. & Pesnot, T. Glycosyltransferases and their assays. *Chembiochem* **11**, 1939-1949 (2010).
- 25 Deng, C. & Chen, R. R. A pH-sensitive assay for galactosyltransferase. *Anal Biochem* **330**, 219-226 (2004).
- 26 Yang, M., Brazier, M., Edwards, R. & Davis, B. G. High-throughput mass-spectrometry monitoring for multisubstrate enzymes: determining the kinetic parameters and catalytic activities of glycosyltransferases. *Chembiochem* **6**, 346-357 (2005).
- 27 Wongkongkatep, J., Miyahara, Y., Ojida, A. & Hamachi, I. Label-free, real-time glycosyltransferase assay based on a fluorescent artificial chemosensor. *Angew Chem Int Ed Engl* **45**, 665-668 (2006).
- 28 Lowery, R. G. & Kleman-Leyer, K. Transcreeper™: screening enzymes involved in covalent regulation. *Expert Opin. Ther. Targets* **10**, 179-190 (2006).
- 29 Ortiz-Soto, M. E., Reising, S., Schlosser, A. & Seibel, J. Structural and functional role of disulphide bonds and substrate binding residues of the human  $\beta$ -galactoside  $\alpha$ -2,3-sialyltransferase 1 (hST3Gal1). *Sci Rep* **9**, 17993 (2019).
- 30 Gosselin, S., Alhussaini, M., Streiff, M. B., Takabayashi, K. & Palcic, M. M. A continuous spectrophotometric assay for glycosyltransferases. *Anal Biochem* **220**, 92-97 (1994).
- 31 Brown, C., Leijon, F. & Bulone, V. Radiometric and spectrophotometric *in vitro* assays of glycosyltransferases involved in plant cell wall carbohydrate biosynthesis. *Nat Protoc* **7**, 1634-1650 (2012).
- 32 Eilertsen, J. & Schnell, S. A kinetic analysis of coupled (or auxiliary) enzyme reactions. *Bull Math Biol* **80**, 3154-3183 (2018).

# **Chapter 3**

## **Assay Development**

### 3.1 Chapter Overview

Glycosyltransferase enzymes (GTs) catalyze the transfer of sugar moieties from activated nucleotide donor to a glycosyl acceptor molecule producing free nucleotide and a glycoconjugate. These reactions are termed Bi-Bi reactions as they involve two substrates and two products. Despite the diversity of these reactions in terms of the acceptors, the donors can be classified into either nucleoside diphosphate (NDP)-based, in the form of a UDP/GDP-based donor, or nucleoside monophosphate (NMP)-based, in the form of a CMP-based donor. CMP-based reactions are specific to sialyltransferases, GDP-based reactions are specific to fucosyltransferases and mannose, and UDP-based reactions constitute the rest of GT's reactions<sup>1</sup>. A few assays have been developed to study the kinetics of GTs. All these methods suffer from disadvantages of either being endpoint assays or continuous but with compromised specificity, sensitivity, or cost, which makes the current methods inaccurate and/or impractical to deploy in high throughput screening (Chapter 2, Section 2.6). In this thesis, a PK/LDH-coupled assay developed previously<sup>2-4</sup> was improved to increase the sensitivity and specificity of detection, accommodate all classes of GTs, and to accurately determine and adjust the assay window. This was achieved by implementing a systematic bottom-up optimization approach, incorporating optimized nucleoside kinases CMK and NDK, and using the fluorescence property of NADH for quantification instead of absorbance. Simulation of the coupled enzymes system was also carried out in batch and fed batch modes to explore the margins of the assay performance and aid the optimization. The optimized assay design (Figure 3.1) provides an accessible tool to study glycosyltransferase activity and can adapt a high throughput format for drug screening.

The chapter presents the optimization and development process of the assay in a bottom-up fashion including NADH signal optimization, optimization of the PK/LDH coupling step, expression and coupling of CMK, expression and coupling of NDK, and performance evaluation at each step. The chapter also includes an example of deploying the assay in studying kinetics parameters of ST3GAL.

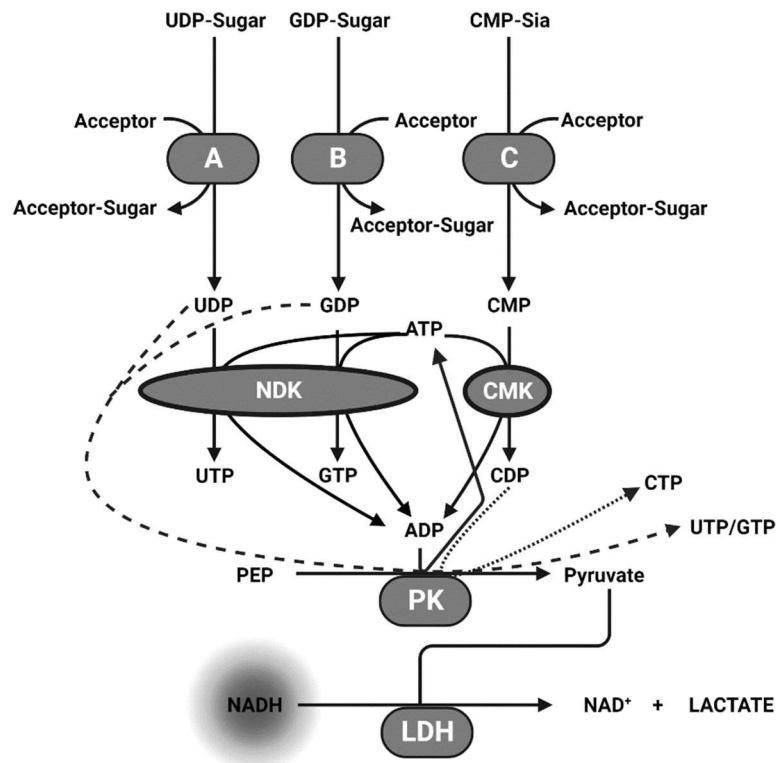


Figure 3.1: Schematic representation of GTs enzymatic-coupled assay. A, B, and C are the three classes of glycosyltransferase enzymes in human catalyzing the transfer of sugar moieties from UDP-, GDP-, and CMP-based donor, respectively. Nucleoside diphosphate Kinase (NDK) and Cytidylate Kinase (CMK) are the kinases used to phosphorylate released Nucleotides by ATP and release ADP. PK and LDH are Pyruvate Kinase and Lactate Dehydrogenase enzymes that couple released UDP to NADH depletion monitored by measuring the drop in fluorescence signal of NADH. Dashed arrows represent alternative slow routes of direct NDPs phosphorylation by PK (non-committed step). Dotted arrow represents incomplete (slow) route of CDP phosphorylation by PK

## 3.2 Optimization of NADH Signal

### 3.2.1 Limitations of absorbance

The aim of this optimization step is to maximize the sensitivity and specificity of the PK/LDH coupled reaction in detecting and reporting ADP in real time and to maintain linear performance for a defined range of ADP concentration. While the specificity of detection is limited to the affinity of PK and LDH to other substrates, the sensitivity can be maximized by supplying PK, LDH, PEP, and NADH in abundance. In addition to sensitivity enhancement, maximizing the turnover rate of PK/LDH warrants that the reaction remains the rate-limiting step, hence the validity of the assay in the defined window of ADP concentration is maintained.

NADH is the assay's reporter molecule and its depletion is proportional to ADP, which in turn is proportional to the nucleotide released from GT reaction (Figure 3.1). The concentration of NADH is conventionally determined by measuring the absorbance at 340 nm. Measuring NADH absorbance for quantification is a limiting factor to the sensitivity and specificity in PK/LDH coupled assays. One reason for this is that the absorbance of an analyte in a solution is proportional to the optical path length. In miniaturized assays (small volume setups), such as in 384-well plate format, the short path length severely compromises the strength of absorbance signal and impairs the sensitivity of the assay<sup>5</sup>. The other reason is that absorbance is susceptible to interference from other molecules that have an overlapping absorbance spectrum, either from particulates in the sample that cause light scattering, or from the well material itself, which results in poor signal to background ratio. Interference is particularly problematic in drug screening due to the difficulty in identifying and controlling for this interference for all screened molecules<sup>5,6</sup>. Using NADH as a reporter in PK/LDH-coupled assays imposes another challenge, which is measuring the depletion of NADH from initial high concentration. Most measuring technologies, such as absorbance and fluorescence, suffer from deviation from its linearity at higher concentration of the analyte, and require careful optimization of the instrument and reaction conditions to guarantee a linear correlation between NADH and the recorded signal in the desired assay window.

To overcome the shortcoming of absorbance, researchers tend to shift the measuring technology to fluorescence where possible<sup>7</sup>. Fortunately, this is possible with NADH due to its fluorescence

property<sup>8</sup>. The sensitivity of fluorescence is 1000 times more than that of absorbance<sup>9</sup>, therefore, it is the technology of choice for miniaturized assays. However, fluorescence is not a complication-free choice. Fluorophores are prone to quenching and interference by the assay component, similar to what was discussed above with absorbance, which may lead to false signals. One big advantage that continuous assays offer over endpoint counterparts is that these artifacts can be identified and quantified for each sample. For example, when screening a compound library using an HTS continuous assay, the progress of the signal is monitored as soon as the reaction is initiated by adding a mixture of the enzyme and tested compound to the reaction mixture. Any deviation from the control's initial signal (at time 0) can be easily identified and quantified and subtracted from the actual signal<sup>7</sup>.

### **3.2.2 Optimization of Fluorescence signal**

In this thesis, the aforementioned limitations related to NADH quantification were improved by optimizing a protocol to acquire fluorescence at excitation/emission wavelengths of 340/460 nm. The reaction volume was reduced to 30  $\mu$ l in a 384-well plate. Fluorescence linearity and stability for a serial dilution of NADH over 30 minutes' time course were investigated in real time. The instrument's parameters were optimized to maximize the performance for a selected dynamic window within the linear range. Optimization of NADH signal was used as a foundation to optimize the coupling of PK and LDH and produce a robust detection of ADP in a defined concentration range.

The initial NADH concentration added to the detection mixture must maintain a linear correlation with its fluorescence in a broad and high enough range. A broader range of linearity is essential to increase the span of the GT progress curve, so that it can be reliably recorded. In addition to linearity, the optimal range must be high enough to maintain a rapid and sensitive LDH reaction. Fluorescence linearity must also be assessed in the context of the reaction mixture to account for possible interference from other components. To achieve that, a serial dilution of NADH starting from 300  $\mu$ M as the highest concentration was prepared in 1 X assay reaction buffer (Appendix, A.2.1) and 30  $\mu$ l of each concentration was dispensed in triplicate to a black 384-well plate and the fluorescence was recorded in kinetics mode using EnVision microplate

reader at excitation/emission wavelengths of 340/460. Default optical parameters of the microplate reader were used in this preliminary test. A linear signal was detected for a range of NADH from 0-150  $\mu\text{M}$  with an ideal linearity below 100  $\mu\text{M}$ . As mentioned above, the aim is to maximize NADH concentration. Therefore, a 100  $\mu\text{M}$  NADH was selected as an optimal concentration. Next, the optical parameters of the instrument: measurement height, number of flashes, photomultiplier tube gain, and excitation and emission light intensities were adjusted using EnVision's built-in optimization program to maximize the  $Z'$  score for a range between 75 and 100  $\mu\text{M}$  of NADH. Triplicates of wells contain blank, high concentration (100  $\mu\text{M}$ ), and low concentration (75  $\mu\text{M}$ ) were used in this process. The process of testing the linearity and adjusting the optical parameters was repeated one more time. Optimum parameters resulted in a  $Z'$  value of 0.6 for the range between 75  $\mu\text{M}$  and 100  $\mu\text{M}$  of NADH. The serial dilution experiment was repeated. Fluorescence was stable for all concentrations during the 30 minutes of recording (Figure 3.2, A), and the means of fluorescence were calculated over 30 minutes for each concentration. The fluorescence means were linearly correlated to concentration for a range up to 150  $\mu\text{M}$ , with perfect linearity up to 100  $\mu\text{M}$  (Figure 3.2, B).

Kinetics parameters of LDH were reported by the manufacturer, except for NADH. However, the curated values of  $K_m$  on the Brenda database show a median of 58  $\mu\text{M}$ . Therefore, depletion of NADH in a range of around 100  $\mu\text{M}$  should provide a relatively stable turnover rate, especially when LDH is provided in abundance. This hypothesis was tested in the next section by testing the performance of PK/LDH with different concentrations of NADH and by simulating the system in batch and fed batch modes.

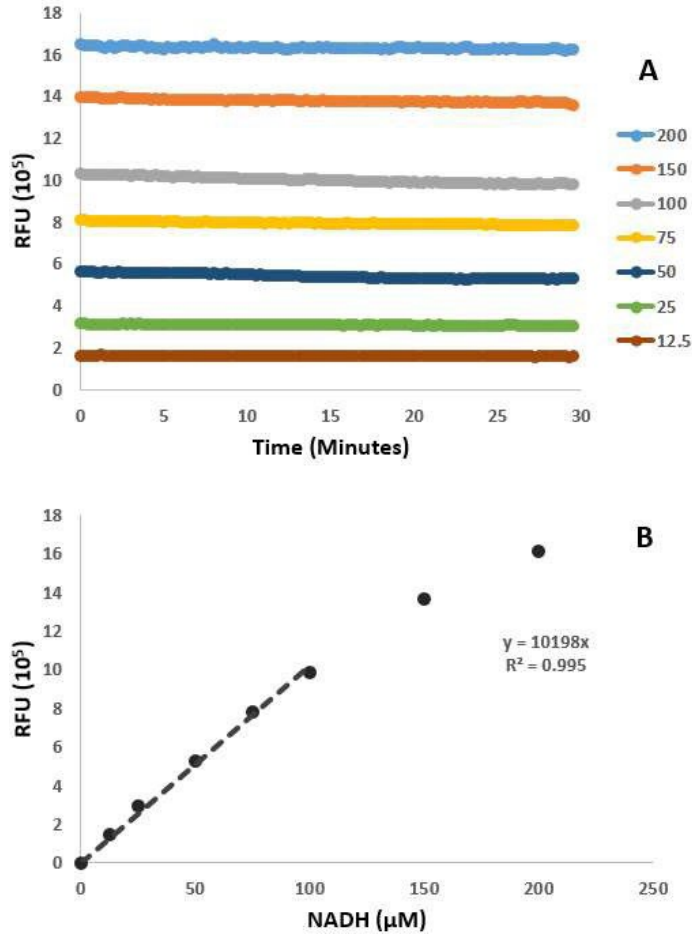


Figure 3.2. Linearity and stability of NADH fluorescence: A. Fluorescence of NADH serial dilution in PK/LDH reaction mix was recorded for 30 minutes. B. Means of fluorescence for each concentration during the recording period are plotted against concentration. Linear regression for the range 0-100  $\mu\text{M}$  with the equation and  $R^2$  value are shown. Concentrations were prepared in triplicate. Average standard deviation for each concentration was calculated from the standard deviations of each well using the pooled standard deviations method (Appendix, A.5). However, these values were small and not visible in the graph scale above.

### 3.3 Optimization of PK/LDH Coupling

The aim of this optimization step is to maximize the sensitivity of PK/LDH coupled reaction in detecting and reporting ADP in real time and maintain linear performance for a defined range of ADP concentration and ADP flux. While the specificity of detection is limited to the affinity of PK and LDH to other substrates, the sensitivity can be maximized by supplying PK, LDH, PEP, and NADH in abundance, i.e., enhancing the overall turnover rate. In addition to sensitivity enhancement, maximizing the turnover rate of PK/LDH warrants that the GT reaction remains the rate-limiting step, hence the validity of the assay in the defined window of ADP concentration (Chapter 2, Section 2.7). While the task is simple for PK reaction, the restraint on NADH concentration due to fluorescence saturation is the bottleneck of this optimization step.

#### 3.3.1 PK and LDH Used in The Assay

Pyruvate Kinase catalyses the final energy harvesting step of the glycolysis pathway by transferring phosphate from phosphoenolpyruvate to ADP to produce pyruvate and ATP. This reaction required  $Mg^{2+}$  ions and one monovalent cation (typically  $K^{1+}$ ) to facilitate substrate binding. Four isomers of PK are expressed in mammalian cells with type 2 being the predominant one. Type 2 PK exists as a dimer or tetramer (dimer of a dimer). The tetramer is significantly more active than the dimer due to the allosteric enhancement of PEP binding in the tetrameric structure, whereas ADP binding is independent of the polymerization status. This polymerization plays an important regulatory role in energy metabolism as its status is altered by feedback loops mediated by other metabolites and hormones, such as the glycolysis metabolite, fructose-1,6-bisphosphate (FBP)<sup>10</sup>. In addition to ADP, PK can phosphorylate other NDPs, such as GDP, UDP, and CDP, in the same order of increasing  $K_m$ . However, the relative value of  $K_m$  for GDP to ADP is reported to be higher in some publications<sup>11,12</sup> and in the same range (as medians) for the values curated in the Brenda database (Table 3.1). As for the reaction mechanism, a ternary complex BI-BI mechanism with no specific order of substrates binding or products leaving was reported<sup>13-15</sup> i.e., random Bi-Bi mechanism.

Lactate dehydrogenase, the last enzyme in the anaerobic glycolysis pathway, is an oxidoreductase that exists in all cells as a family of six isozymes (LDH1-LDH6). The enzyme exists in a homo- or hetero-tetramer of a combination of two subunits: A and B, also known as M

and H, for muscle and heart, respectively. The ratio of subunit B: subunit A in the tetramer increases in order from LDH1 to LDH5. LDH6 is a homotetramer of a unique subunit, C, and the isozyme is exclusively expressed in sperm and testis. LDH catalyses the reduction of pyruvate to lactate by NADH and the reverse oxidation by  $\text{NAD}^+$ . The favourable direction of the isoforms is determined by the ratio of subunit B: subunit A. The higher the ratio, the more the isoform is predisposed to prefer the reverse reaction from lactate to pyruvate<sup>16</sup>. The muscle isoform of LDH is isozyme 1, which is a tetramer of subunit A and has an optimum pyruvate to lactate preference, indicated by its remarkably low  $K_m$  values for NADH and pyruvate compared to  $\text{NAD}^+$  and lactate. The kinetics mechanism of LDH is ordered Bi-Bi with NADH as the leading substrate<sup>17,18</sup>.

Rabbit muscle LDH and PK are commercially available and commonly applied in coupling assays reported in literature<sup>2-4,19</sup>. The enzymes used in this thesis were obtained from Sigma-Aldrich and are both extracted from rabbit muscle and prepared as suspensions in ammonium sulfate buffer, pH 6.0. LDH is type 2 LDH a preference for the pyruvate to lactate direction. This preference is also evident from the low  $K_m$  values for NADH and pyruvate as reported on the Brenda database (Table 3.1). PK is also type 2 in a tetramer form, hence its high affinity to PEP, which can be inferred from the manufacturer's reported  $K_m$  for PEP (Table 3.1). High  $K_m$  values for ATP and pyruvate, 0.86 and 10 mM, respectively, as reported by the manufacturer, suggest that the reverse reaction is negligible. Specific activities for PK and LDH are 6000 and 12000 units/ml enzyme suspension, respectively. The unit is defined by the amount of the enzyme that converts 1  $\mu$  mole substrate to product per minute, typically tested near saturation conditions of the enzyme with its substrates. Therefore, every unit of PK and LDH added to 30  $\mu$ l of reaction mixture is expected to produce a  $V_{\max}$  value of 33.3 mM/min when the enzyme is saturated with its substrates. Given the low  $K_m$  values for their substrates, adding a few units of the enzymes per reaction should yield reactions fast enough to prevent the PK/LDH-coupled reaction from being the rate-limiting step in the assay. Concentrations of PK and LDH and the sensitivity to ADP concentration are investigated in section 3.3.2

### 3.3.1 PK/LDH Optimization and ADP Standard Curve

For the PK/LDH-coupled reaction to be successfully utilized in the assay, it must detect and report a broad variation in ADP flux readily and accurately. The detection must also persist for enough time, and sustain a large enough ADP concentration to enable monitoring of the progress curve of the reported reaction and calculate its rate. To achieve that, PK and LDH rates must remain high and apparently independent of the concentrations of their substrates during recording of the reaction progress. The rates are determined by enzyme concentration in relevance to  $k_{\text{cat}}$  (2.11), and substrates concentration in relation to  $K_m$  (Figure 2.2). In addition to enzymes and substrates concentrations, other parameters such as  $pH$ , ions, buffer type, and temperature are determinant to enzyme activity. These parameters are excluded from the optimization, and they are used according to previously published protocols<sup>2-4</sup>.

#### Parameters Selection

The focus of this optimization will only be on the concentrations of the reacting species and the enzymes. Therefore, four concentration parameters in PK/LDH coupling were only considered, which are concentrations of PK, NADH, PEP, and NADH. Generally, maximizing these parameters will achieve the goal of this optimization step. PK has high affinity to PEP ( $K_m = 0.07 \text{ mM}$ ). Therefore, a concentration of 2 mM was chosen, which approximates saturation conditions and falls in the upper range of previously published protocols. For the other three parameters: concentrations of PK, LDH and NADH will be optimized using the one-factor-at-a time (OFAT) approach (section 2.5.2)

NADH has a relatively high  $K_m$  value and adding it to near saturation is not possible due to the restraint of its fluorescence signal linearity. To make up for that, the amount of LDH in the assay can be increased. However, enzymes cost is a limiting factor to assay adoption, and they are usually prepared in buffers with other components that might interfere with the assay. Therefore, minimizing enzyme usage is a major goal for assay development.

## PK/LDH System Simulation in Batch mode

To explore the effect of NADH concentration on assay performance, a simulation model of the coupled assay in batch mode was designed (Appendix, A.6) and the simulation was performed using COPASI software. The choice of batch mode was made because it can be validated experimentally. In this simulation, the generic rate expression (2.16) for reversible Bi-Bi reaction with a single central complex was used. This was decided based on the reported mechanisms for the enzymes as mentioned in section (3.3.1). Initial concentrations were set as follows: one unit of each of PK and LDH (1:1 ratio), 2 mM of PEP, 2 mM of ATP, varying concentrations for NADH and ADP. Initial concentrations for pyruvate, lactate, and  $\text{NAD}^+$  were set to zero. ATP was added to account for CMK and NDK coupling. This will be presented later in the chapter.

ATP is a product of PK reverse reaction and, its inhibitory effect on the forward reaction must be factored in. The aforementioned conditions were used to simulate the response of the system (represented by NADH depletion) to addition of either 3 or 50  $\mu\text{M}$  of ADP. Simulation was repeated for four concentrations of NADH: 50, 100, 200, and 300  $\mu\text{M}$  (Figure 3.3). The results show that LDH was very responsive to PK flux, as indicated by the small lag between NADH depletion curves and ADP depletion, which was combined with a small spike in pyruvate between 0-2 seconds (data not shown). The results suggest that the effect of NADH concentration on LDH activity is insignificant in general. However, when high concentration of ADP is used (50  $\mu\text{M}$ ), decreasing the NADH concentration to 50  $\mu\text{M}$  increased the lag time of NADH depletion significantly, which can be explained by the linear substrate-dependency of the rate below  $K_m$  (Table 3.1 and Figure 2.2). There is significant overall lag in the system - more than 6 seconds - which, in reality, may be larger due to the impact of other factors such as diffusion lag time and suboptimal activities of the enzymes, and the lag time for each enzyme to reach its own steady state, which the mathematical model used here does not account for.

To reduce this lag time, and based on the finding of this test, the assumption of PK being the rate limiting step in the selected conditions was investigated. The time course simulation was repeated for 25  $\mu\text{M}$  ADP and 100  $\mu\text{M}$  NADH (average concentrations of what was tested above). Conditions for PEP, ATP, pyruvate, and  $\text{NAD}^+$  were kept unchanged from the previous test. LDH was fixed at 1 unit with PK tested at three concentrations: 1, 2 and 3 units. The results in

Figure 3.4 show that increasing the ratio of PK to LDH from 1:1 to 2:1 reduced the lag time significantly, 2.5 seconds for 2:1 compared to more than 6 seconds when 1:1 ratio was used. No additional improvement was detected in lag time when 3:1 ratio was tested (Figure 3.4). This suggests that PK to LDH should be used in a 2:1 ratio. Interestingly, most protocols use LDH at higher concentrations than PK, which contrasts with these simulation results.

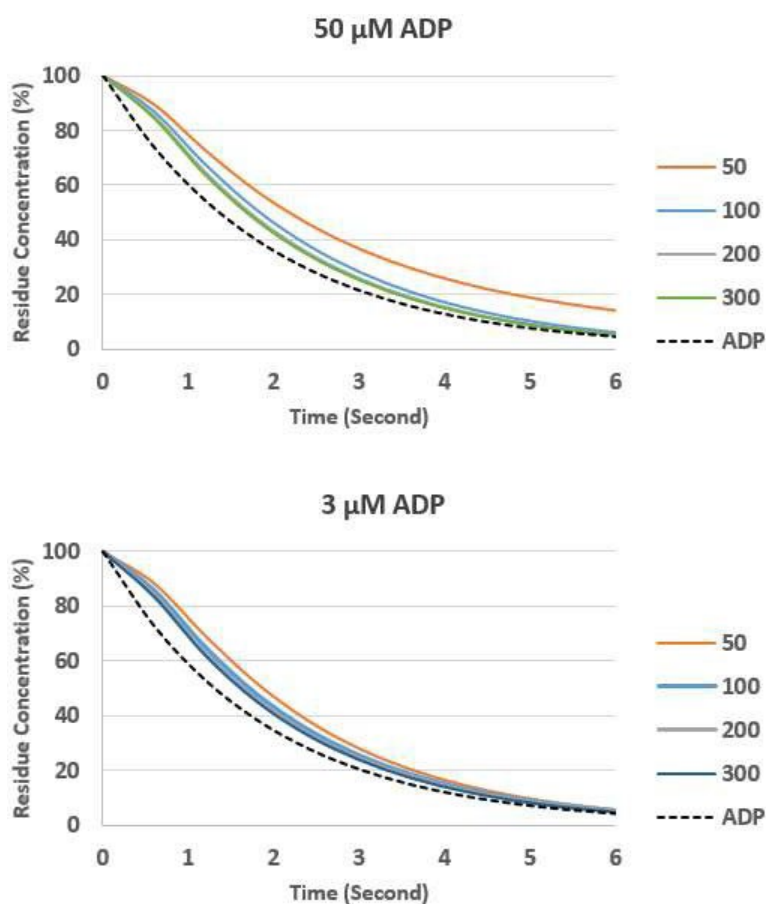


Figure 3.3. NADH concentration effect on PK/LDH performance. Time course simulation of PK/LDH system was performed for two concentrations of ADP supplied to the system in batch mode, 50  $\mu$ M (top) and 3  $\mu$ M (bottom). Coloured series show depletion of NADH for different initial concentrations of NADH (from 50 -300  $\mu$ M). Dashed lines are the depletion of ADP (chosen for 100  $\mu$ M initial concentration of NADH as ADP depletion was NADH-independent and curves of ADP depletion were overlapping for all NADH concentrations). Concentrations of Y axis are presented as percentage of the residue concentration (Appendix, A.6).

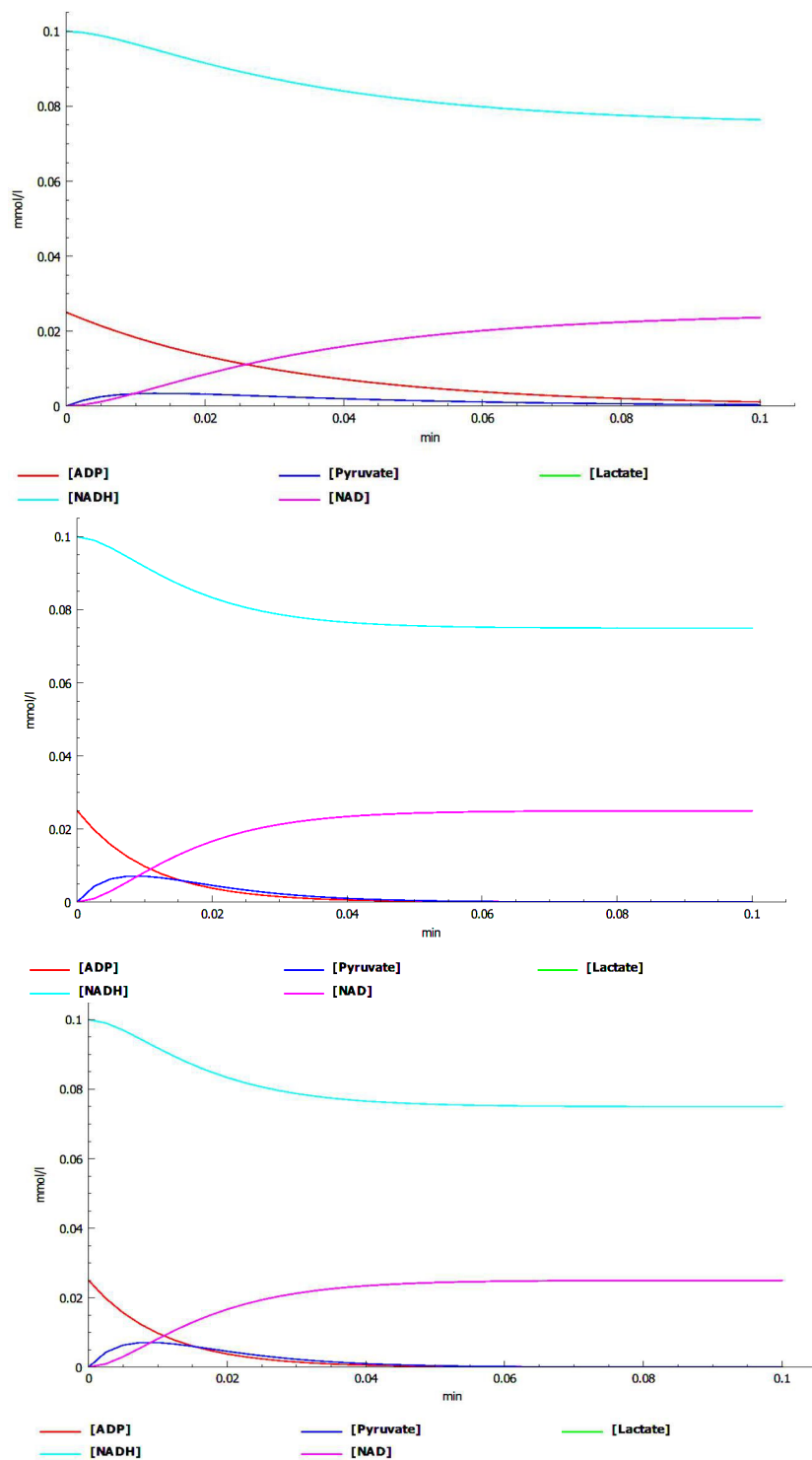


Figure 3.4. Effect of PK to LDH ratio on lag time. Time course simulation of PK/LDH system was performed for 25  $\mu$ M ADP supplied to the system in a batch mode. NADH was used at 100  $\mu$ M. 1 unit of LDH was used for all tests. PK was tested at three concentrations 1, 2 and 3 units, for the graphs from top to bottom, respectively. Other parameters were kept unchanged from Figure 3.2

## Experimental Validation of Parameters

Using the same parameters applied to the simulation with 6 units PK and 3 units LDH, the performance of the assay was tested with twofold serial dilution of ADP at different NADH concentrations as explained in the Appendix (A.7) (Figure 3.5). Due to the experimental limitation, the progress cannot be recorded instantly after initiation of the reaction and a delay of ~15 seconds is inevitable. Therefore, the lag time frame observed in simulation cannot be investigated experimentally. Taking this limitation into account, the reactions for all tested combinations were complete or at near completion by the time the signal recording started. Some lag was observed with 50  $\mu\text{M}$  NADH when ADP was used at high concentrations, which was observed in the simulation (Figure 3.3, upper graph) and explained by the concentration dependency of rates at and below  $K_m$ .

Progress curves in Figure 3.5 were all plateaued at stable fluorescence signals. To investigate the correlation between fluorescence and ADP concentration, the signals for each plateau were averaged and plotted against ADP concentration (Figure 3.6, A). Signals showed linear correlation with the concentrations except when ADP concentration approaches that of NADH in mixture. When excluding the nonlinear ranges, the slopes were independent of NADH concentration and that was confirmed from the linear regression at 100 and 50  $\mu\text{M}$  of NADH (Figure 3.6, C and D). Interestingly, the slope was the same as that of the NADH standard curve (Figure 3.2, B). This proves that the drop in both the NADH concentration and the fluorescence signal corresponds to ADP concentration perfectly, confirming complete depletion of ADP (in the linear regions).

From Figure 3.6, A, we see loss of linearity at low concentrations of ADP when NADH is used at high concentrations, which indicates loss of sensitivity as NADH increases. To further investigate that, the  $Z'$  score from equation (2.18) was used to evaluate the signal window between each concentration and the background (Figure 3.6, B). We can see from the  $Z'$  values that sensitivity improves when NADH decreased. The loss of sensitivity can be explained by the loss of fluorescence linearity at high concentrations of NADH, as it was discussed in the NADH fluorescence optimization section. The other reason is the dispensing error when the reaction mixture is added to the wells to initiate the reaction. The variation in background signal due to the dispensing error is more significant at high concentrations of NADH, especially when this

variation is accompanied by a small signal (at low ADP concentrations) i.e., low signal to background ratio and lower Z' score. Fortunately, this error and loss of sensitivity impacts the assay when used in endpoint format only, because continuous assay measures the change of signal over time from each well rather than the absolute readout. The principle of internal control in continuous assay is discussed in detail in Chapter 2 (section 2.5.1).

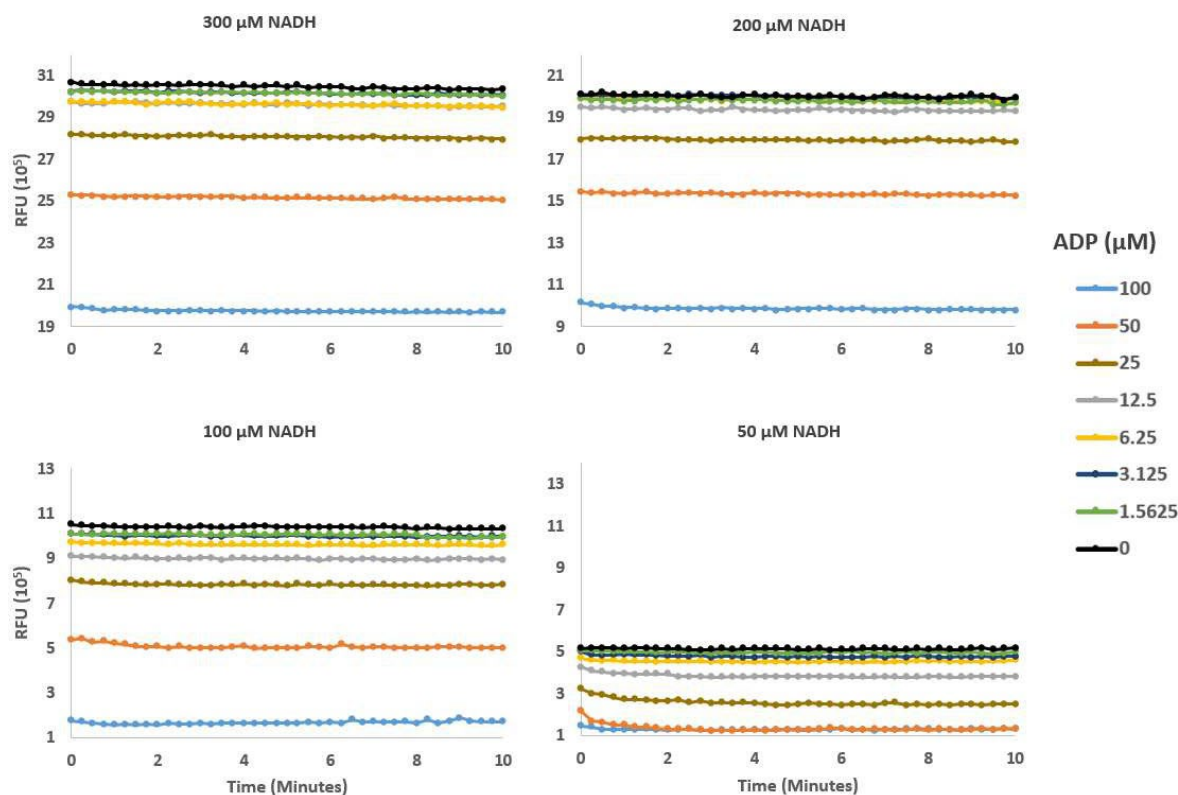


Figure 3.5. Time course of PK/LDH coupled assay with serial dilution of ADP. Time course of PK/LDH reaction mixture with a twofold serial dilution of ADP was performed at different concentrations of NADH. Triplication for each concentration of ADP was prepared and the curves shown are the means.

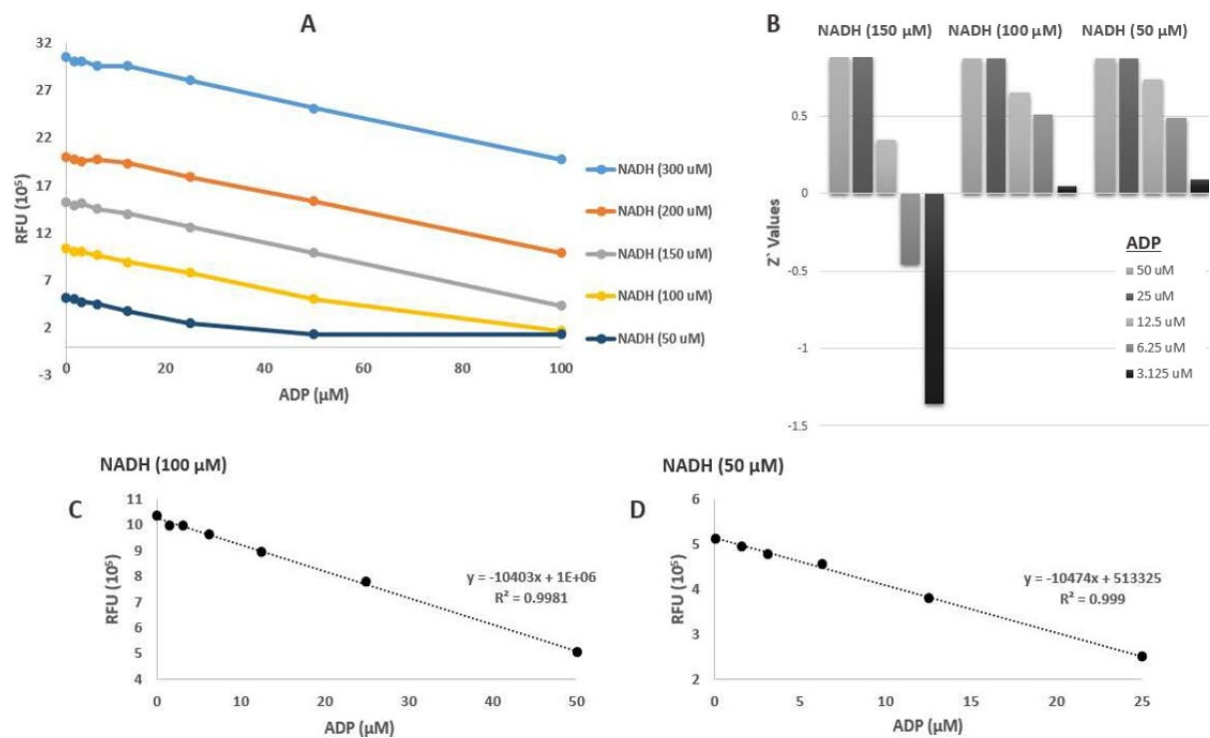


Figure 3.6. NADH concentration effect on ADP detection. A. Correlation of fluorescence and initial ADP concentration in the assay, at different NADH concentrations. B. Z' scores of the signal window between fluorescence at a given concentration of ADP and background when no ADP was added to the reaction mixture. C and D. linear regression of the linear regions from ADP titration curves from panel A at 100  $\mu$ M and 50  $\mu$ M of NADH in panel C and D, respectively.

## Fed Batch Simulation

The aim of PK/LDH system is to detect a flux of ADP. Therefore, the response of PK/LDH system to this flux needs to be evaluated. This is difficult to perform experimentally as there is no common method to generate controlled flux in this small scale of volume. Therefore, a simulation of PK/LDH response to a fed batch system when ADP is supplied in constant rate was performed (Appendix, A.7). Using COPASI, and with NADH concentration of 100  $\mu\text{M}$  and the same other optimized conditions for the assay, a parameter scan was performed by recording the rates of the reaction species in response to changing ADP flux. The rates of reaction species were recorded at 2 and 4 minutes as this is the reasonable time frame to experimentally record reaction progress (Figure 3.7). The results show that a constant overlapping between the rate of  $\text{NAD}^+$ , as another marker of the LDH progress, and ADP flux persists up to 20  $\mu\text{M}/\text{min}$  when the 4 minutes was the recording time point, and for a higher range for 2 minute as the recording time point. At higher rates, the assay does not keep up with ADP flux and accumulation of pyruvate takes place due to depletion of NADH to a level that impairs NADH rate.

This upper boundary estimation is very useful for preliminary design of experiment. For example, if one plans to use the assay for detecting ADP flux of an enzyme in a specific time frame, the rate allowance from the fed batch simulation can be used to estimate the highest  $V_{\text{max}}$  allowed. From this  $V_{\text{max}}$  limit, enzyme concentration limit can be calculated if  $k_{\text{cat}}$  of the enzyme is known.

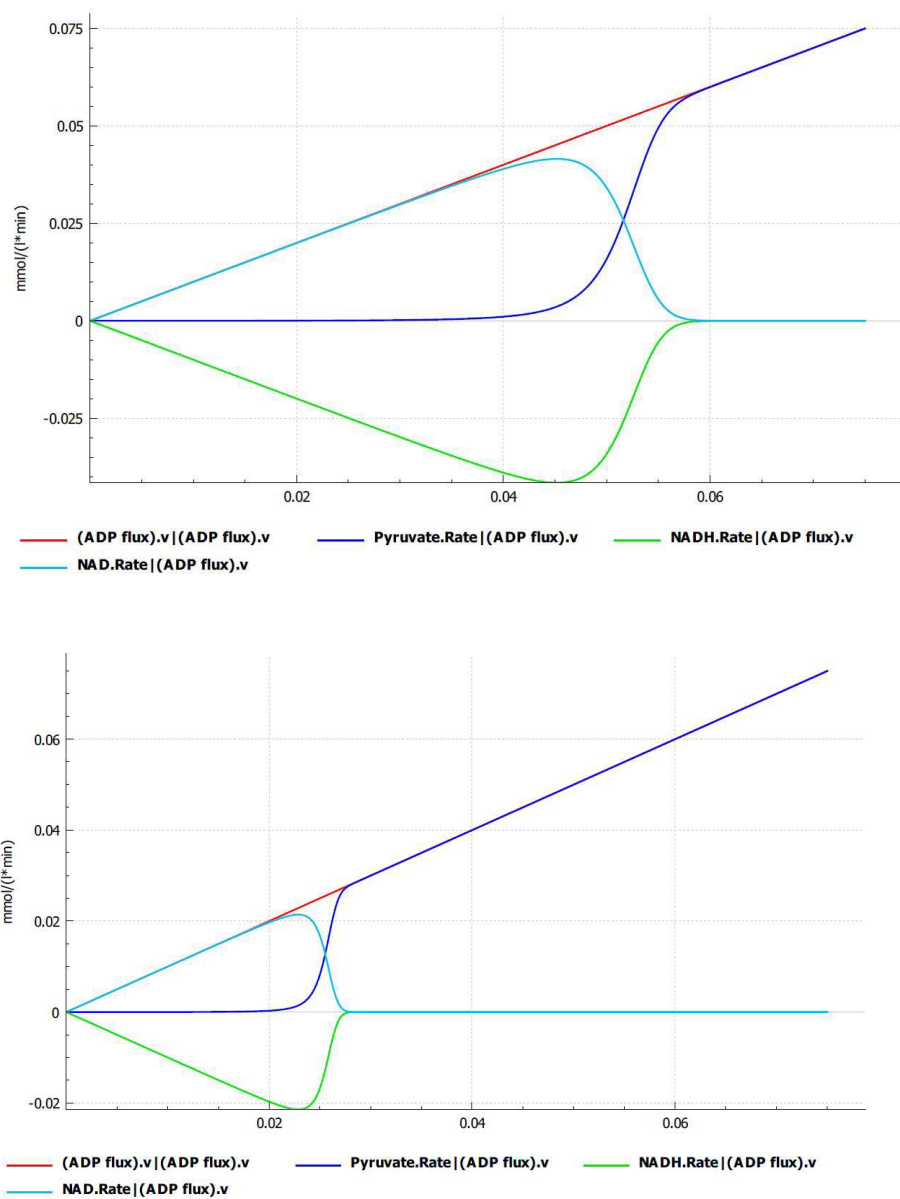


Figure 3.7. ADP flux scan in fed batch simulation of PK/LDH reaction. Rates of reaction species were recorded at 2 minutes (top) and 4 minutes (bottom) over a range of ADP flux in a fed batch simulation of PK/LDH reaction. Optimized parameters and 100  $\mu\text{M}$  NADH were used in this test.

To summaries, based on the optimization done above, the following conditions will be used as optimal parameters for PK/LDH assay: PK:LDH ratio of 2:1. The amount can be adjusted depending on the batch-to-batch variation of the two enzymes. A quick standard titration of ADP with different amounts of the two enzymes in 2:1 ratio can be performed with every new batch. The rest of the parameters are fixed at: 2 mM of PEP, 0.1 mM NADH (with a tolerance range of 0.05-0.15 mM), 50 mM of KCl, 10 mM of MnCl<sub>2</sub>, 10 mM of MgCl<sub>2</sub>, and 1 mg/ml of BSA. The assay also tolerates 2 mM of ATP when needed.

### 3.4 Coupling of Cytidylate Kinase (CMK)

The purpose of this enzyme is to couple CMP that is released from Sialyltransferase enzymes to PK/LDH. This step has been accomplished previously by using nucleoside monophosphate kinase enzyme (NMPK)<sup>2,4</sup>. NMPK phosphorylates CMP to CDP via ATP and releases ADP as product, which can be detected by the PK/LDH system. However, the baker's yeast NMPK, is not easily sourced commercially and does not have a complete kinetics parameters profile. Additionally, the coupling step has not been characterized or optimized before. In this thesis an alternative kinase, cytidylate kinase (CMK), was expressed and coupled with PK/LDH. The activity of CMK was optimized using PK/LDH. The result is a new robust assay, CMK/PK/LDH, capable of coupling CMP to PK/LDH.

#### 34.1 Expression of CMK

As mentioned above, the baker's yeast NMPK (EC 2.7.4.4)<sup>2</sup> is not characterized. Therefore, cytidylate kinase (EC 2.7.4.25) from *E. Coli*. was found to be an excellent alternative with low  $K_m$  values for ATP and CMP (Table 3.1). Additionally, CMK has a much lower  $K_m$  value for CMP than for other nucleotides, making it highly specific and minimizing the chance of unwanted side reactions. Finally, CMP, being an *E. Coli*. enzyme, has a greater chance of successful expression in an *E. Coli*. host. The expression plasmid, containing a C-terminal His-tagged CMK sequence, was designed and the enzyme was expressed in *E. Coli*. BL21 (DE3) strain and purified by IMAC as explained in the Appendix (A.1.1, A.2.2- A.2.4) (Figure 3.8).

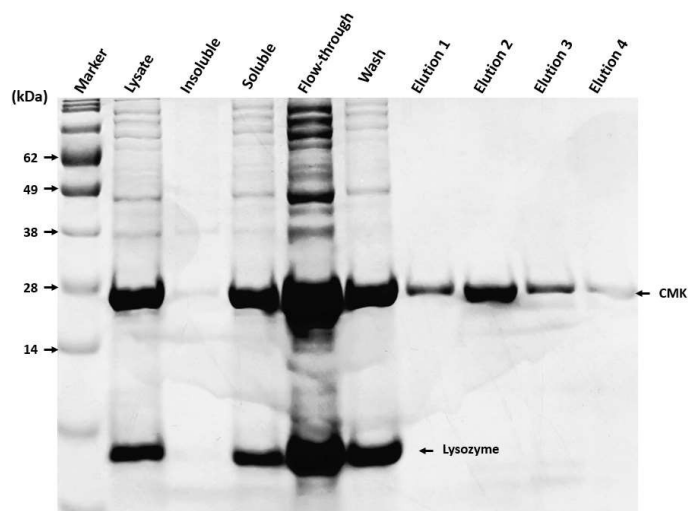


Figure 3.8. SDS-PAGE gel of CMK Purification steps. Purification of CMK was performed using Immobilized metal affinity chromatography (IMAC). The lanes show fractions from the purification process. Elution fractions containing pure enzyme were pooled.

### 3.4.2 Activity of CMK

Enzyme titration was carried out using the optimized PK/LDH assay with addition of CMP to a concentration of 60  $\mu\text{M}$  in the final reaction. A serial dilution of CMK stock was tested (Figure 3.9, A). The ideal concentration is the lowest concentration of the enzyme to achieve full depletion of CMP. This concentration was identified to be 390 nM. Next, CMK was coupled to PK/LDH by adding CMK to the PK/LDH optimized mixture at a final concentration of 390 nM. The response of CMK/PK/LDH to a serial dilution of CMP was tested in the same fashion as ADP (Figure 3.5 and Figure 3.6). Linear regression and Z' scores were calculated as described above for ADP (Figure 3.9, B, C, and D)

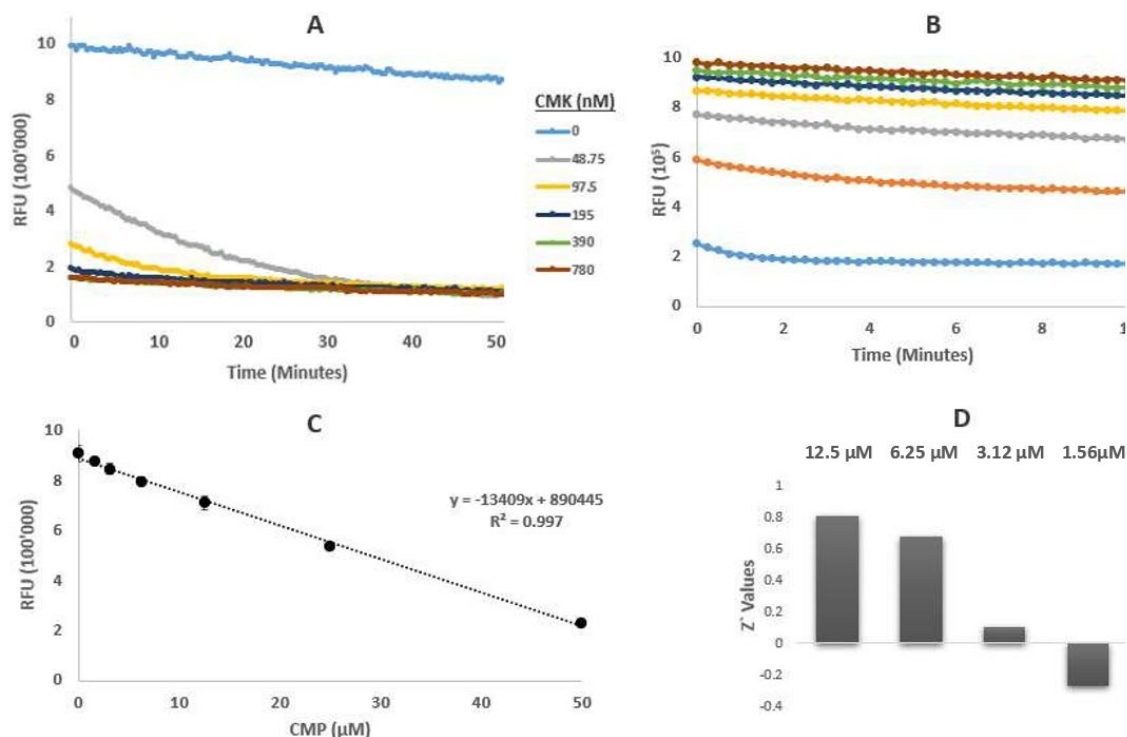


Figure 3.9. Activity of CMK. A. CMK enzyme was titrated with PK/LDH reaction mixture contains 60  $\mu\text{M}$  CMP, the fluorescence was recorded over 50 minutes. B. A serial dilution of CMP was tested with the coupled CMK/PK/LDH, CMK was used at 390 nM and reactions were initiated by adding the mixture to wells containing CMP dilutions. C. Linear regression of fluorescence signals. Three fluorescence readings between 1.75 and 2.25 minute from each curve of panel B were averaged and plotted against CMP concentration. Linear regression was performed. D. Z' values for the three readings from each concentration were calculated (relative to background signal). Triplication for each concentration of CMP was prepared and the curves shown in panels A and B are the means.

The results confirm the validity of the CMK/PK/LDH system in detecting CMP in a linear manner. However, the slope was slightly steeper than that of the ADP titration and NADH fluorescence signal (Figure 3.6, C and Figure 3.2, B). One explanation could be that the CDP generated from the CMK reaction is further phosphorylated by PK (Figure 3.1, dotted line). This possibility is inferred from the fact that CDP is a substrate for PK. However, the specificity of PK to CDP is low, as inferred from the high  $K_m$  value of 6.8 mM (Table 3.1). Since the range of

CDP generated in this test spans only 0.05 mM, which is much lower than the  $K_m$ , only a fraction of the generated CDP is phosphorylated to CTP. This can be the explanation of the increase of the slope by only 30 % when compared with ADP and NADH slopes above i.e., for each molecule of CMP, 1.3 molecules of NADH were depleted as a result of the coupling assay. If the phosphorylation of CDP was complete, a 100 % increase of the slope would have been observed, indicating that for each molecule of CMP, two molecules of NADH are depleted. Since the phosphorylation of CDP to CTP is slow, a lag time is expected to complete this phosphorylation (or to achieve a near depletion concentration). Indeed, this lag time was detected in the serial dilution experiment (Figure 3.9, B). This lag was more obvious at higher concentrations i.e., with 25 and 50  $\mu$ M CMP. Overall, since this linearity holds in the tested range, and the lag can be accounted for as an intrinsic limitation of the assay, no more optimization was performed.

### 3.5 Expression and Coupling of NDK

As mentioned in the introduction to this chapter, there are two more classes of GTs, besides silyltransferases, that use GDP or UDP nucleotide sugars as donors. The PK/LDH system was used previously to assay these reactions by directly coupling the nucleotide-generating reaction with PK<sup>2,3</sup>. However, as reviewed in section 3.3.1, PK has high  $K_m$  values for UDP and GDP. This low affinity is expected to impair the sensitivity of PK/LDH in detecting those nucleotides. In one research papers<sup>20</sup>, GDP and UDP were phosphorylated by another enzyme, NDP kinase from baker's yeast. This phosphorylation is done via ATP, and it releases ADP in equivalent amount to GDP/UDP. The ADP is then coupled with other enzymes in an endpoint format assay for detection. Here we first confirmed the limitation of the direct coupling of UDP to PK/LDH and then expressed another kinase, nucleoside-diphosphate kinase (NDK). The activity of NDK was optimized using PK/LDH and its performance was evaluated. The result is a new robust assay, NDK/PK/LDH, capable of coupling UDP to PK/LDH more efficiently than direct phosphorylation by PK.

### 3.5.1 Expression of NDK

In order to select an optimum kinase, The Brenda databases was used to survey the kinetics parameters for all nucleoside-diphosphate kinase enzymes (NDK, EC 2.7.4.6) with criteria set as follows: (1) low  $K_m$  value for UDP/GDP and this value must be lower than that of PK to guarantee that NDK reaction is the committed step, (2) NDK enzymes are promiscuous and are capable of phosphorylating other nucleoside-diphosphates. Therefore, a higher  $K_m$  for ADP than that of PK is essential to avoid capturing ADP and preventing ADP access to PK is important, (3) low  $K_m$  for ATP, (4) high  $K_m$  for UTP/GTP to minimize the possibility of phosphorylation of ADP with UTP/GTP, (5) easily expressed in *E. Coli*. i.e., lacking eukaryotic post-translational modifications requirement. The survey led to the selection of *Acanthamoeba polyphaga Mimivirus* orthologue with  $K_m$  values shown in Table 3.1. The *Mimivirus* NDK has a  $K_m$  value for UDP tenfold lower than that of PK and 3-4 times higher for ADP than that of PK. It also has an extremely low  $K_m$  for ATP and relatively high  $K_m$  for UTP and GTP with values of 0.1 and 0.83, respectively (data not shown in Table 3.1 and can be found on the Brenda website). The expression plasmid containing the C-terminal His-tagged CDK sequence was designed, and the enzyme was expressed in *E. Coli*. BL21 (DE3) strain and purified by IMAC as explained in the general protocols (Appendix: A.1.1, A.2.2- A.2.4) (Figure 3.10).

Table 3.1

## Kinetics Parameters of The Assay Enzymes

| Enzyme     | $K_M$ (mM) |                |            |      |       |        |          |        |         |                  |         |
|------------|------------|----------------|------------|------|-------|--------|----------|--------|---------|------------------|---------|
|            | ATP        | ADP            | GDP        | UDP  | CDP   | CMP    | Pyruvate | PEP    | NADH    | NAD <sup>+</sup> | Lactate |
| <b>LDH</b> | n/a        | n/a            | n/a        | n/a  | n/a   | n/a    | 0.13*    | 3.50** | 0.058** | 0.23**           | 10**    |
| <b>PK</b>  | 0.86#      | 0.24** - 0.30# | 0.26**     | 1**  | 6.8** | n/a    | 10#      | 0.07#  | n/a     | n/a              | n/a     |
| <b>CMK</b> | 0.038*     | n/a            | n/a        | n/a  | n/a   | 0.035* | na       | na     | na      | na               | n/a     |
| <b>NDK</b> | 0.005*     | 0.77*          | 0.02-0.3** | 0.1* | 0.05* | n/a    | n/a      | n/a    | n/a     | n/a              | n/a     |

# Supplier

\* From Brenda for the same species orthologue used in the assay (excluding mutant)

\*\* Median from Brenda of all reported values (All species, excluding mutant), this value is calculated when there is no information about the orthologue used in the assay

n/a: No available values

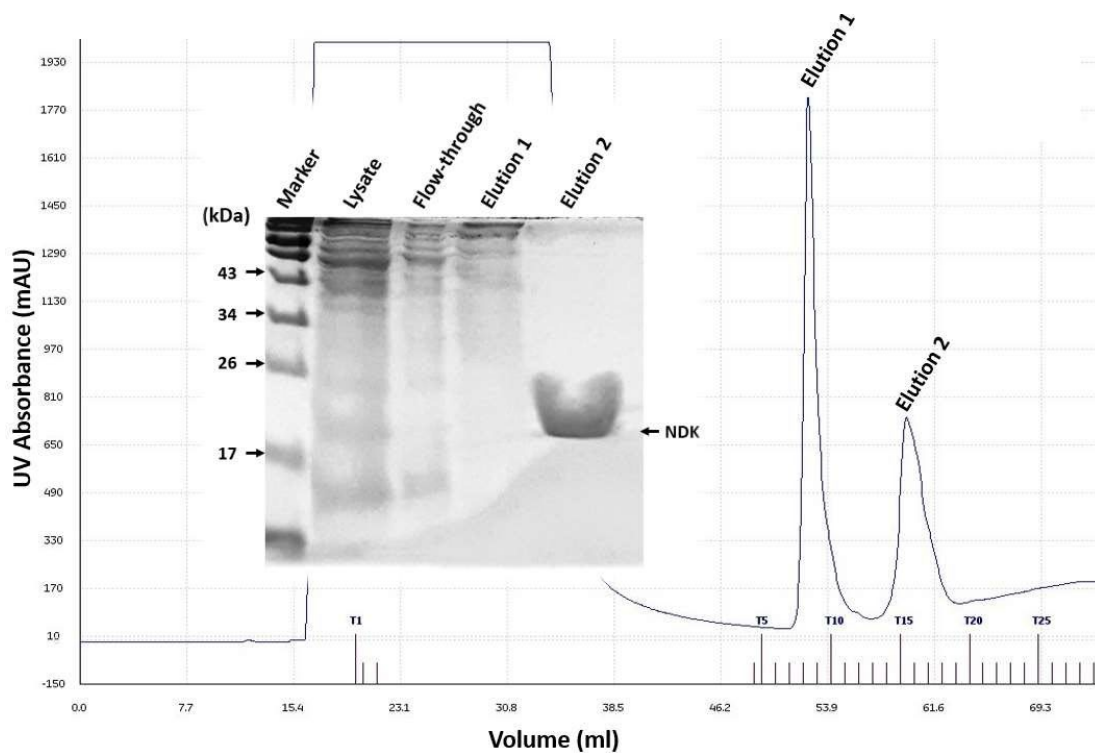


Figure 3.10. SDS-PAGE gel of the purification steps of NDK. Purification of NDK was performed using IMAC. The lanes show fractions from the purification process. Elution fraction 2 contains pure enzyme. The back figure shows the chromatogram of the purification with two elution bands; Elution 1 band contains nonspecifically bound proteins, while Elution 2 band contains the pure enzyme.

### 3.5.2 Activity of NDK

First, a comparison of the activity of NDK with that of PK in phosphorylating UDP was carried out. A reaction mixture of PK/LDH was prepared. The mixture contains UDP added at an amount that yields 50  $\mu\text{M}$  in the final reaction. The concentration of NADH was increased in the mixture from 100 to 125  $\mu\text{M}$  (as calculated in the final reaction). This modification of NADH was implemented after observing an initial drop of fluorescence in a trial (data not shown here). The reaction was initiated by adding the reaction mixture to wells containing a serial dilution of NDK in triplicates and fluorescence were recorded over 50 minutes (Figure 3.11, top panel). All combinations that contain NDK, in any concentration, produced an initial drop of fluorescence corresponding to a concentration of NADH equivalent to 50 $\mu\text{M}$ . This was calculated from the NADH titration slope (figure 3.2, B). The completion of the reaction that lacks NDK took around 5 minutes, indicated from the time point when it plateaued. This result suggests that UDP phosphorylation in NDK-containing reactions was complete when the fluorescence recording started, whereas it lagged in absence of NDK.

An unexpected continuation of fluorescence drop was observed in presence of NDK, in a dose-dependent rate. NDK titration was repeated in absence of UDP (Figure 3.11 bottom panel), and the same pattern of dose-dependent fluorescence decline was observed. Despite that, titration in absence of UDP resulted in initial fluorescence readings equivalent to the initial concentration of NADH (as calculated from NADH standard curve), which confirms that the initial drop in the previous experiment (titration with UDP) was UDP dependent.

To investigate the UDP dose dependency of NDK activity, three reaction mixtures were prepared: without NDK, with 500 nM NDK, and with 1000 nM NDK. A higher concentration of NADH (140  $\mu\text{M}$ ) was added to NDK containing mixtures to make up for fluorescence depletion during pre-initiation incubation. The mixtures were used to test the activity with a serial dilution of UDP (Figure 3.12, A, B, and C). The results confirmed the time lag in absence of NDK at all concentrations of UDP (more obvious at higher UDP concentrations). Both concentrations of NDK produced the background signal decline and the NDK dose dependency of this decline was also confirmed. The decline rate was constant for a given NDK concentration, and independent of UDP, indicated from the parallel lines in each set. The more extreme case, with 1000 nM

NDK, was used to plot the fluorescence against UDP concentration. The values were sampled at the 4-minute time point (Figure 3.12, D). The linear regression shows a slope value equivalent to that of ADP titration (Figure 3.6 C, D). This confirms that, when corrected for the background, the fluorescence is perfectly responsive to UDP dose, and the coupling in the NDK/PK/LDH system conserved the molar equivalence from UDP to ADP to NADH. The only impact of the background decline is on shortening the recordable time of the assay. However, from (Figure 3.12, C), even with extreme decline, a dose of UDP, up to 25  $\mu\text{M}$  maintained its linearity for 12 minutes and this time was around 5 minutes with 50  $\mu\text{M}$  UDP. For practical application, lower concentrations of NDK can be safely used. In (Figure 3.11, top panel) using 50 nM of NDK was sufficient to complete the phosphorylation of UDP instantly with much lower background decline compared to those of 500 and 100 nM.

As for the background fluorescence decline, it has no impact on the validity of the assay as proven from the linear regression of UDP vs corrected fluorescence signal. This background can be explained by any possibility of a slow NDK-dependent or NDK stock buffer-dependent reaction. To identify the cause of this background, some strategies can be followed, for example, different expression batches of NDK can be tested and each component of the reaction (except for NADH) can be investigated in terms of its dose impact on the signal.

In conclusion, the activity of PK is insufficient to provide sensitive detection of UDP. NDK provided an alternative route for UDP phosphorylation, which is faster and more sensitive than the direct phosphorylation via PK. The reactivity with GDP was not covered in this thesis. However, the fluctuation of reported  $K_m$  values for GDP (Table 3.1) creates uncertainty regarding whether NDK is needed to boost its coupling to PK, or whether it can be coupled directly. The same investigation carried out with UDP can be applied to GDP.

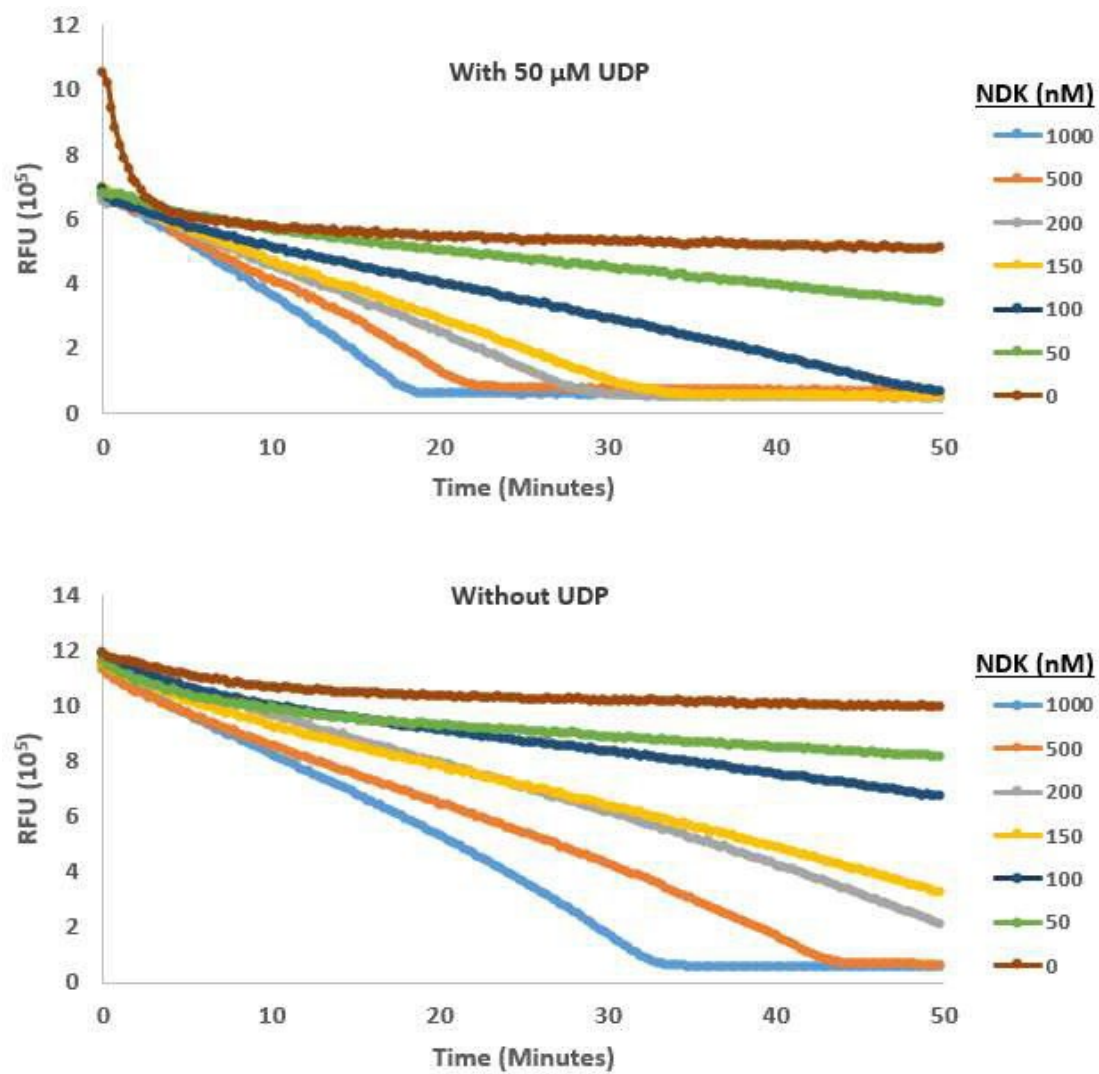


Figure 3.11. Titration of NDK. NDK was titrated in PK/LDH optimized reaction mixture with modifying NADH concentration from 100 to 125  $\mu$ M. Titration was performed with UDP (top) and without (bottom). Triplication for each concentration of NDK was prepared and the curves shown are the means.

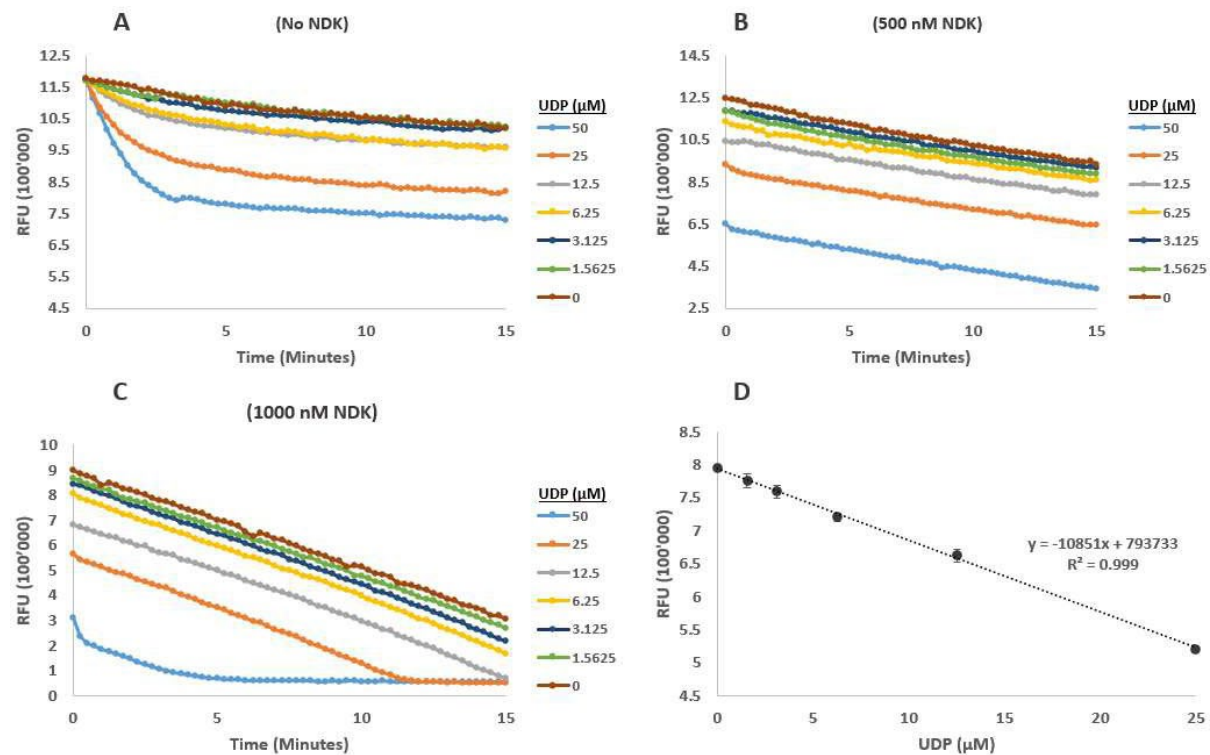


Figure 3.12. Activity of NDK. A dose response of the coupling assay to a serial dilution of UDP was performed in a reaction mixture without NDP (panel A), with 500 nM NDK (panel B), or with 1000 nM (Panel C). A linear regression of fluorescence plot against UDP concentration sampled at 4 minute was performed (Panel D)

### 3.6 Performance at Low *pH*

Glycosyltransferase enzymes are *pH* sensitive, and *pH* plays a regulatory role. For example, they can localize across different Golgi compartments along the Golgi cis-trans axis, and this localization entails operation at different *pH*<sup>21,22</sup>. Therefore, *pH* might have a regulatory effect on GTs activities. One aspect of the ideal assay designed to study GTs is that it should maintain its performance in a physiologically relevant range of *pH*. The assay thus far was optimized at *pH* = 7.4. To test if the assay can maintain the linear detection of CMP/UDP at a low *pH* of 6.3, both assay configurations, CMK/PK/LDH and NDK/PK/LDH were tested with serial dilutions of CMP and UDP, respectively. All parameters were kept the same but with using 500 mM MES buffer at *pH* 6.3 instead of HEPES buffer used above (Figure 3.13). The results confirm the validity of the assay at *pH* 6.3. The slopes in both configurations were comparable to those at *pH* 7.4 (Figure 3.9, C) and (Figure 3.12, D).

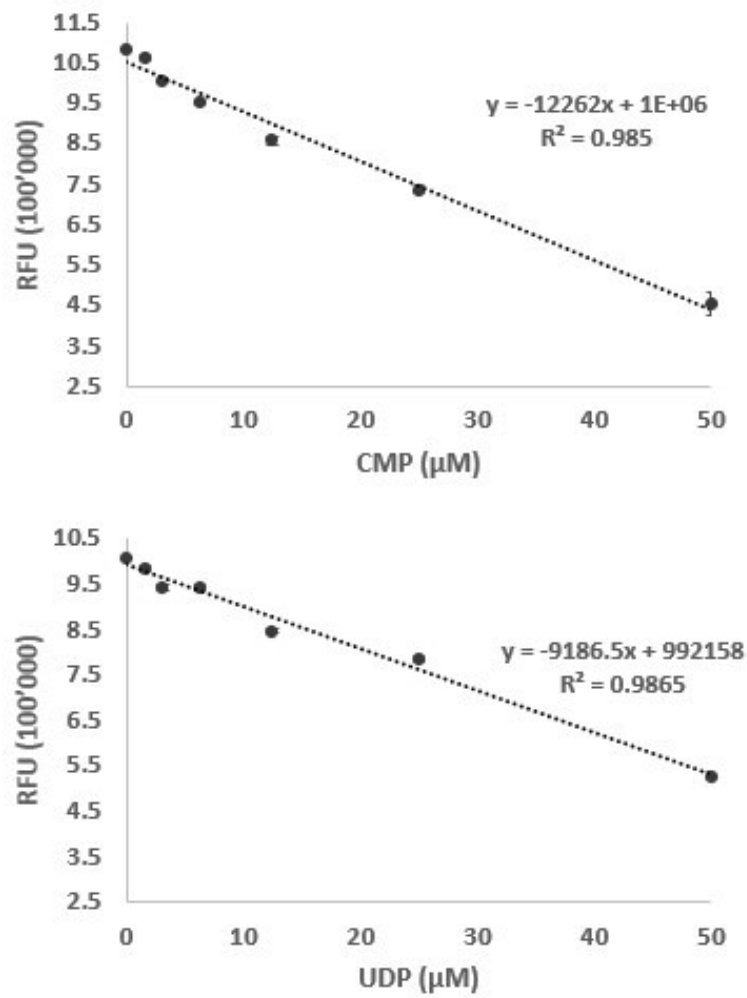


Figure 3.13. Performance of the assay at low *pH*. The assay was used in either CMK/PK/LDH or NDK/PK/LDH configuration to generate standard curves of CMP (top) and UDP (bottom), respectively. Same conditions optimized above were used with replacing HEPES buffer with MES *pH* 6.3.

### 3.7 Applications

Enzyme assays and their applications are covered in detail in section 2.5 of Chapter 2. In this Chapter, The CMK/PK/LDH configuration was used to identify the kinetics parameters of ST3GAL1. The NDK/PK/LDH assay was applied to monitoring glycopeptide synthesis and confirm substrate identity of the products in Chapter 5 and Chapter 6.

The human sialyltransferase enzyme, ST3GAL1, was expressed in the human kidney cell line, HEK 293, as explained in the Appendix (A.3). The enzyme was titrated using a fixed concentration of CMP-Sia and the acceptor Galacto-N-biose. The enzyme was titrated with the substrates added to the reaction in abundance (1 mM of each) to optimize the upper range of the signal. Enzyme concentration was chosen from the titration trial so that it produces a reaction slow enough to be recorded, and that was determined to be 226 nM. Kinetics were performed by alternating the saturation with one substrate and varying the other one. The reactions were performed simultaneously for all concentrations in triplicate. The rates were calculated by plotting the progress curve of each reaction. The linear range of each progress curve, including the background, was identified (representing the steady state), and the slope of the background signal was subtracted from the slope of each progress curve. These background-corrected slopes were used to calculate the rates by converting the slope expressed as RFU/min to rates in mM/min using the standard curve of CMP above. Rates were fitted to the Michalis-Menten curve via GraphPad Prism 8 software, which also calculates  $K_m$  and  $k_{cat}$  values from the fitting. The  $K_m$  values reported for ST3GAL1 in the Brenda database are 0.01-0.25 mM for the acceptor, and from 0.01-0.02 mM for CMP-SA. similar values were obtained in this thesis.

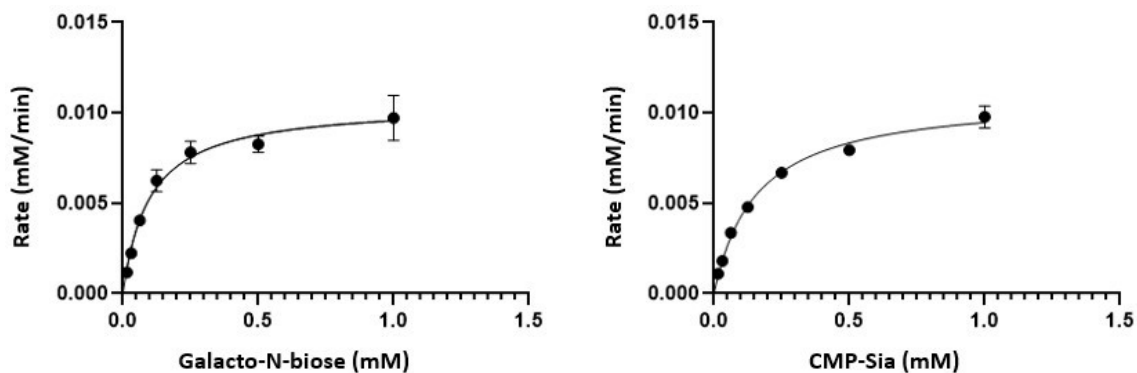


Figure 3.14. Michaelis-Menten fitting of ST3GAL1 initial rate correlation with substrate concentration. ST3GAL1 was used at a concentration of 226 nM. In the left panel the rates were calculated at a fixed concentration of CMP-Sia at 1 mM and varying concentrations of the acceptor, Galacto-N-biose. In the right panel, a fixed concentration of the acceptor at 1 mM was used with a varying concentration of the donor.  $K_m$  and  $k_{cat}$  values were: 0.15 mM and 48  $\text{min}^{-1}$  for the donor and 0.098 mM and 47.5  $\text{min}^{-1}$  for the acceptor.

## References

- 1 Lairson, L. L., Henrissat, B., Davies, G. J. & Withers, S. G. Glycosyltransferases: structures, functions, and mechanisms. *Annu Rev Biochem* **77**, 521-555 (2008).
- 2 Gosselin, S., Alhussaini, M., Streiff, M. B., Takabayashi, K. & Palcic, M. M. A continuous spectrophotometric assay for glycosyltransferases. *Anal Biochem* **220**, 92-97 (1994).
- 3 Brown, C., Leijon, F. & Bulone, V. Radiometric and spectrophotometric *in vitro* assays of glycosyltransferases involved in plant cell wall carbohydrate biosynthesis. *Nat Protoc* **7**, 1634-1650 (2012).
- 4 Ortiz-Soto, M. E., Reising, S., Schlosser, A. & Seibel, J. Structural and functional role of disulphide bonds and substrate binding residues of the human  $\beta$ -galactoside  $\alpha$ -2,3-sialyltransferase 1 (hST3Gal1). *Sci Rep* **9**, 17993 (2019).
- 5 Inglese, J. *et al.* High-throughput screening assays for the identification of chemical probes. *Nat Chem Biol* **3**, 466-479 (2007).
- 6 Simeonov, A. & Davis, M. I. in *Assay Guidance Manual* (eds S. Markossian *et al.*) (2004).
- 7 Simeonov, A. *et al.* Quantitative high-throughput screen identifies inhibitors of the *Schistosoma mansoni* redox cascade. *PLoS Negl Trop Dis* **2**, e127 (2008).
- 8 Cannon, T. M. *et al.* Characterization of NADH fluorescence properties under one-photon excitation with respect to temperature, pH, and binding to lactate dehydrogenase. *OSA Contin* **4**, 1610-1625 (2021).
- 9 Gomes, A. J., Lunardi, C. N., Rocha, F. S. & Patience, G. S. Experimental methods in chemical engineering: Fluorescence emission spectroscopy. *Can J Chem Eng* **97**, 2168-2175 (2019).
- 10 Israelsen, W. J. & Vander Heiden, M. G. Pyruvate kinase: Function, regulation and role in cancer. *Semin Cell Dev Biol* **43**, 43-51 (2015).
- 11 Plowman, K. M. & Krall, A. R. A kinetic study of nucleotide interactions with pyruvate kinase. *Biochemistry* **4**, 2809-2814 (1965).
- 12 Hohnadel, D. C. & Cooper, C. The effect of structural alterations on the reactivity of the nucleotide substrate of rabbit muscle pyruvate kinase. *FEBS Lett* **30**, 18-20 (1973).
- 13 Ainsworth, S. & MacFarlane, N. A kinetic study of rabbit muscle pyruvate kinase. *Biochem J* **131**, 223-236 (1973).
- 14 Seeholzer, S. H., Jaworowski, A. & Rose, I. A. Enolpyruvate: chemical determination as a pyruvate kinase intermediate. *Biochemistry* **30**, 727-732 (1991).
- 15 Mazurek, S., Grimm, H., Boschek, C. B., Vaupel, P. & Eigenbrodt, E. Pyruvate kinase type M2: a crossroad in the tumor metabolome. *Br J Nutr* **87 Suppl 1**, S23-29 (2002).
- 16 Forkasiewicz, A. *et al.* The usefulness of lactate dehydrogenase measurements in current oncological practice. *Cell Mol Biol Lett* **25**, 35 (2020).
- 17 Shoemark, D. K., Cliff, M. J., Sessions, R. B. & Clarke, A. R. Enzymatic properties of the lactate dehydrogenase enzyme from *Plasmodium falciparum*. *FEBS J* **274**, 2738-2748 (2007).
- 18 Zewe, V. & Fromm, H. J. Kinetic studies of rabbit muscle lactate dehydrogenase. II. mechanism of the reaction. *Biochemistry* **4**, 782-792 (1965).
- 19 Swift, R. P. *et al.* The NTP generating activity of pyruvate kinase II is critical for apicoplast maintenance in *Plasmodium falciparum*. *Elife* **9**, e50807 (2020).

- 20 Kumagai, K., Kojima, H., Okabe, T. & Nagano, T. Development of a highly sensitive, high-throughput assay for glycosyltransferases using enzyme-coupled fluorescence detection. *Anal Biochem* **447**, 146-155 (2014).
- 21 Frappaolo, A., Karimpour-Ghahnavieh, A., Sechi, S. & Giansanti, M. G. The close relationship between the Golgi trafficking machinery and protein glycosylation. *Cells* **9** (2020).
- 22 Fisher, P. & Ungar, D. Bridging the gap between glycosylation and vesicle traffic. *Front Cell Dev Biol* **4**, 15 (2016).

# **Chapter 4**

## **Substrate Specificity and In vitro Synthesis Method**

## 4.1 Chapter Overview

The structures of glycans are mainly determined by the specificity of glycosyltransferases (GTs) participating in their biosynthesis. While GTs are highly specific to their donors, their specificities to the acceptors vary and are not fully characterized. With over 200 GTs<sup>1</sup>, and the abundance of glycosylation as a major posttranslational modification of proteins<sup>2</sup>, there is a need for universal tools to systematically characterize GT activities on proteins. In addition to kinetic assays (addressed in Chapters 2-3), Chapters 4-5 focus on developing a substrate synthesis tool with a use case study in Chapter 6.

This chapter demonstrates the attributes of GT- substrate specificity through the Mucin-type O-glycosylation pathway, highlighting the significance of glycopeptide substrate structure in determining specificity. The chapter presents the current methods for studying substrate specificity and discusses their limitations. The review in this chapter covers the utilization of short glycopeptide substrates and methods for their synthesis. Finally, the chapter concludes with an argument on the advantages of using recombinant peptides expressed in *E. Coli.* over the chemical synthesis approach, along with a novel synthesis method. The proposed method involves designing a fusion protein containing MUC1 tandem repeat to enable sequential *in vitro* synthesis of the substrates involved in the MUC1 glycosylation pathway. The chapter provides a detailed explanation of the design rationale, including the selection of the expression carrier protein, the linkers used, and the peptide sequence frame.

## 4.2 Mucin-Type O-Glycosylation

While the specificity of GTs is strict for the donors, it can be relaxed or strict for the acceptor, depending on the enzyme. An example of relaxed specificities is fucosyltransferases 3-7 that catalyze the fucosylation of a terminal N-acetyllactosamine regardless of the glycan structure<sup>3</sup>. At the other end of the spectrum, some GTs are highly specific to their substrate, such as polypeptide GalNAc-transferase (GALNTs) family. To demonstrate substrate specificity and as a development case model, the synthesis pathway of O-glycosylation of MUC1 will be studied.

O-linked glycans are synthesized by adding saccharides to a core peptide, at serine or threonine residues. O-linked glycans are categorized by the first sugar linked to ser/thr:  $\alpha$ -linked O-GalNAc,  $\alpha$ - or  $\beta$ -linked O-galactose,  $\alpha$ - or  $\beta$ -linked O-glucose,  $\alpha$ -linked O-fucose,  $\alpha$ -linked O-mannose,  $\beta$ -linked O-GlcNAc and  $\beta$ -linked O-xylose. Among these O-linked glycans,  $\alpha$ -linked O-GalNAc, also known as mucin-type O-linked glycan, is the most dominant type. Although GalNAc O-linked glycosylation is prominent in mucins, it is also found in other proteins, such as P-selectin glycoprotein ligand 1 (PSGL-1). Indeed, a recent study found that around 83% of all secreted proteins are O-GalNAc glycosylated<sup>4</sup>. The biosynthesis of GalNAc O-linked glycan takes place in five steps: initiation, cores 1 and 3 synthesis, branching (including cores 2 & 4 synthesis), elongation, and sialylation. A schematic representation of the pathway is demonstrated in Figure 4.1

#### **4.2.1 Initiation**

Biosynthesis of GalNAc O-linked glycan is initiated by addition of GalNAc to serine or threonine to form Thomsen-nouvelle (Tn) antigen. This reaction is catalyzed by a large family of enzymes, polypeptide N-acetylgalactosaminyltransferases (GalNAc-Ts), also known as GALNTs. There are 20 GALNTs isozymes in human divided into two families, I and II, based on their peptide substrate preference. While family I (peptides-preferring) GALNTs glycosylate unglycosylated peptides and some glycosylated peptides, Family II (Glycopeptide-preferring) GALNTs act on O-GalNAc glycosylated peptides. GALNT 1-3 are examples of Family I, and GALNT4 of Family II. GALNTs families I and II are further divided based on their amino acid sequence similarity into subfamilies, Ia-g and IIa-b, respectively. Members of the same family have similar substrate affinity, but they are differently expressed in tissues<sup>5,6</sup>. The affinity of GALNTs towards O-GalNAc glycosylated substrates will be covered in detail in section 4.3.

#### **4.2.2 Synthesis of Core 1 and Core 3**

The sequential addition of one or two sugars to Tn antigen produces diverse glycan core structures. There are eight different core structures. While 1-4 are the main ones, 5-8 are rare and will not be covered in this review. The biosynthesis of cores 2 and 4 are subsequent to core 1 and 3, respectively.

Core1 (Gal $\beta$ 1-3GalNAc-O-Ser/Thr), also known as Tomsen-Friedenreich (T) antigen, is synthesized by adding galactose to Tn and forming glycosidic  $\beta$ 1-3 bond between galactose and GalNAc. This reaction is catalysed by Core 1  $\beta$ 1-3 galactosyltransferase 1 (C1GALT1), also known as T synthase that requires the chaperone C1GALT1C1 (COSMC), which provides stability and proper folding for T synthase<sup>7</sup>.

Core 3 is synthesized from Tn antigen by addition of GlcNAc, forming a glycosidic  $\beta$ 1-3 bond between GlcNAc and GalNAc on Tn. This reaction is catalysed by core 3  $\beta$ 1-3 *N*-acetylglucosaminetransferase (C3GnT), also known as B3GNT6, or core 3 synthase. Expression of core 3 structures are mostly limited to respiratory and gastrointestinal tract, where MUC2 is the dominant mucin in these tissues<sup>6</sup>. Studies on their role in cancer suggest a tumor suppression role of B3GNT6 and core 3 structures in colon cancer<sup>8</sup>.

#### 4.2.3 Branching and Elongation

Both core 2 and core 4 are products of a branching of core 1 and core 3, respectively. This branching occurs by adding GlcNAc to the GalNAc in core 1 and 3, forming a glycosidic  $\beta$ 1-6 bond. The branching is catalysed by a family of enzymes called  $\beta$ 1-6 *N*-acetylglucosaminyltransferases. There are three enzymes responsible for the synthesis of core 2 from core 1, GCNT1, GCNT3, and GCNT4. While GCNT1 and GCNT4 are specialized in core 2 synthesis, GCNT3 catalyses Core 4 as well. Another member of the same family, GCNT2, catalyses the same reaction but with longer linear structures as acceptors<sup>6</sup>.

Elongation is characterized by adding repetitions of GlcNAc, GalNAc, and Gal to the cores. The most common elongation is the addition of Gal $\beta$ 1-4GlcNAc (*N*-acetylglactoseamine) in alteration to the cores to form poly *N*-acetylglactoseamine units. The process is catalysed by two groups of enzymes  $\beta$ 1-3 GlcNAc-transferases and  $\beta$ 1-4 Gal-transferases. Another common elongation takes place on carbon 3 of galactose residue in core 1 and core 2 by addition of GlcNAc. This reaction is catalysed by a  $\beta$ 1-3 *N*-acetylglucosaminyltransferase (elongation  $\beta$ 3GnT).

#### 4.2.4 Sialylation

Tn and Core 1 (T) can be sialylated by ST6GALNAC1-4 by forming an  $\alpha$ 2-6 glycosidic bond between sialic acid and the first GalNAc of the glycan to produce STn and ST, respectively. ST6GALNAC1 is specific to Tn as a receptor, whereas ST6GALNAC2 is specific to Core 1. However, this specificity was only detected *in vivo*, and it will be covered in detail in Chapter 6. A member of  $\alpha$ 2-3 sialyltransferases family, ST3GAL1, is responsible for sialylation of Gal $\beta$ 1-3 on core 1. Sialylation terminates further extension of the glycan chain<sup>6</sup>.

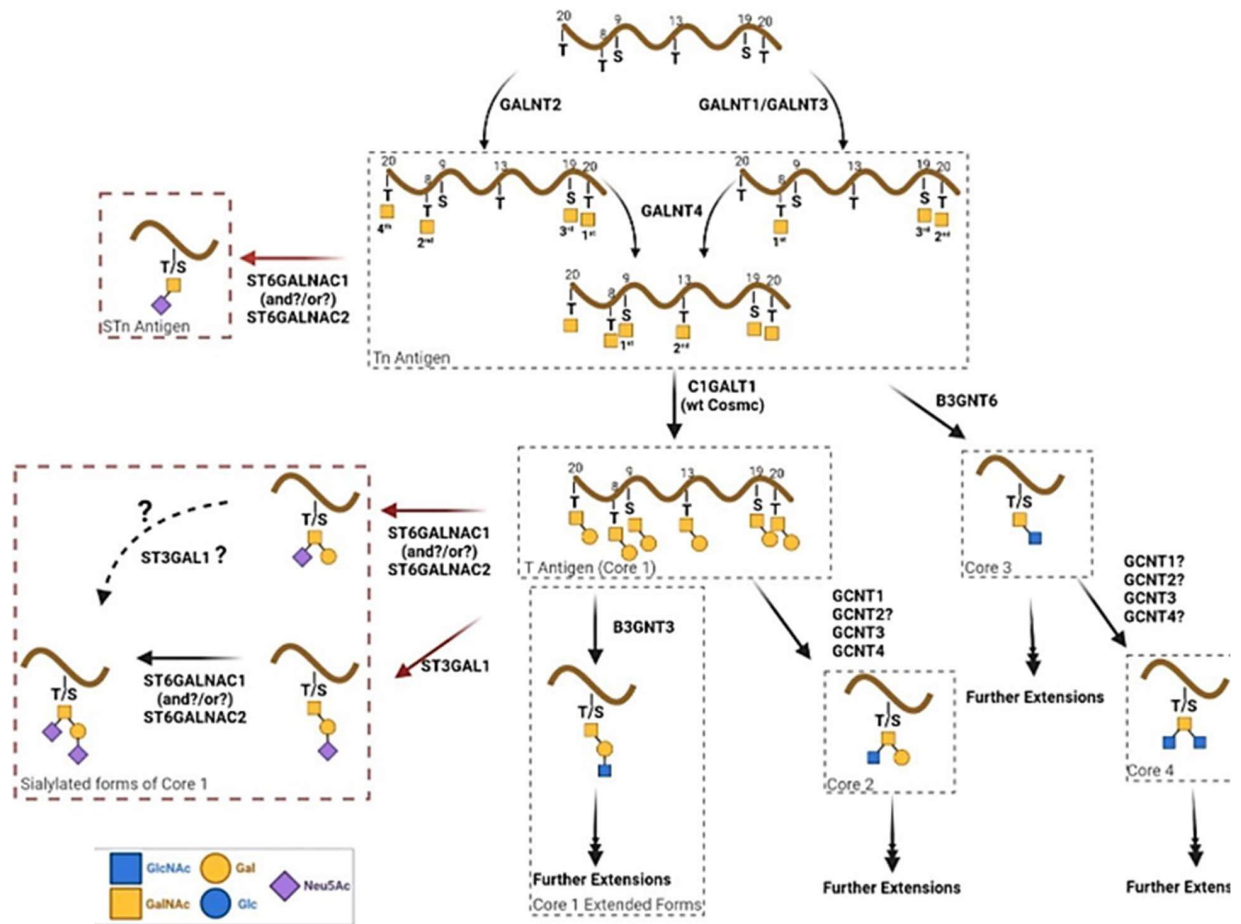


Figure 4.1. MUC1 O-glycosylation pathway as extracted from literature. Curved line represents MUC1-N tandem repeat, TAP24, with the potential glycosylation sites numbered. Sequential glycosylation by GALNTs is represented by the order of GalNAc moieties added to the core peptide. Red arrows represent sialylation (termination) pathways. Dashed arrows and question marks indicate uncertainty. Reactions downstream of T antigen are presented on a single site as no core peptide site specificity was reported for these reactions.

### 4.3 Polypeptide GalNAc-Transferase (GALNTs)

Polypeptide GalNAc-transferase (GALNTs) catalyze the transfer of GalNAc from UDP-GalNAc donor to T/S residues in a protein substrate. They are responsible for the initiation of the mucin-type O-glycan synthesis and the specification of the site of glycosylation on the core peptide. All GALNTs contain two domains: catalytic domain, and lectin-like binding domain (sugar binding domain), spanned by a flexible linker. The catalytic domain also contains two sites that play roles in its substrate specificity: peptide-binding site and glycopeptide-binding site. There are 20 discovered isoforms of GALNTs in humans that demonstrate different specificities to their substrates.

The substrate specificity of GALNT isoforms is determined by combinations of the following structural properties of the enzymes: Specificity of the peptide-binding site in the catalytic domain, specificity of glycopeptide-binding site in the catalytic domain, specificity of the lectin-like binding domain, and finally the properties of the flexible linker. These structural properties determine the preference of the enzyme to its substrate glycosylation site. The outcome of these different combinations is three substrate recognition mechanism: (1) recognition of the core peptide sequence at the glycosylation site by the catalytic domain, (2) recognition of the neighbouring glycosylated amino acid ( $\pm 1-3$  amino acids from the acceptor T/S) by the catalytic domain, (3) recognition of the distant glycosylated amino acid ( $\pm 6-15$  amino acids from the acceptor T/S) by the lectin-like binding domain. The length and flexibility of the linker determines the outreach of the lectin binding domain to its recognition site<sup>5,9</sup>. Details on site specificity of four members of GALNT family are presented in the next section. These members are well characterized and are utilized in this thesis to synthesize glycopeptide substrates using MUC1 tandem repeat as a core peptide. All the peptides' sequences mentioned in the following review and their alignment are presented in Figure 4.2.

### 4.3.1 GALNT1

The site specificity of GALNT1 was studied extensively by Wandall *et al.*, (1997)<sup>10</sup> using fragments of MUC1 peptide: MUC1a and MUC1b and TAP-24 (Figure 4.2). Only MUC1a and TAP-24 were reactive. The reaction of GALNT1 with TAP-24 over 24 hours was monitored by HPLC and mass spectrometry after endo-Asp digest that cleaved TAP-24 into 1-12 and 13-24 fragments. The results of the study concluded that GALNT1 glycosylated TAP-24 sites in the following order:



Sites 1,2 were completely glycosylated within 3.2 hours, whereas site 3 was only partly glycosylated after 24 hours. This was also confirmed in other studies<sup>11,12</sup>.

### 4.3.2 GALNT2

The same study by Wandall *et al.*, (1997), showed that GALNT2 was reactive with MUC1b and TAP-24 but no reactivity was observed with MUC1a. The same pipeline of investigation with HPLC and mass spectrometry revealed that GALNT4 could glycosylate TAP-24 sites in the following order:



Sites 1,2, and 3 were fully glycosylated within 24 hours. Site 4 took longer. This was also confirmed by later studies<sup>11-13</sup>. A recent study by Coelho *et al.*, (2022)<sup>14</sup>, investigated the mechanism of directed specificity of GALNT2 in a MUC1 sequence consisting of 4 tandem repeats. According to this study the glycosylation by GALNT2 starts at T20 followed by T8, and those 2 sites were fully glycosylated very quickly in all 4 repeats. This was demonstrated to be accomplished by direct recognition of T20 via the catalytic domain. Binding of the lectin-like domain of the enzyme to glycosylated T20 then facilitates the glycosylation of T8. In the Coelho study, all T20 sites on the four repeats were glycosylated instantly, unlike the T20 at position 1 in TAP-24 in the other studies. The reason for this can be deduced: the direct recognition by the

catalytic domain was impaired in TAP-24 due to the truncation of the peptide at the N-terminal, which may have impaired the affinity of the catalytic domain to this site. Fast glycosylation was also observed for 70 % of S19 indicating that S19 in one of the 4 tandem repeats took much longer to be fully glycosylated. However, after 24 hours, all the three sites, T20, T8 and S19 were glycosylated in all four repeats. Further investigation of the role of the lectin binding domain found that, the quick glycosylation of S19 in three tandem repeats was facilitated by the binding of the lectin-like binding domain to the T8 in the next tandem repeat (in C terminal direction), where the last repeat's S19 (the C terminal repeat) lacked C terminal neighbouring T8.

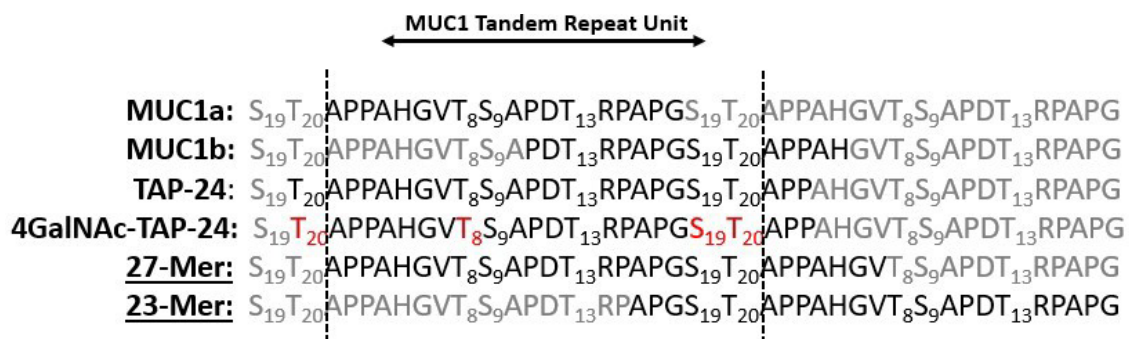


Figure 4.2. Sequences of MUC1 peptides used in literature to study GALNTs activities. All sequences were aligned with two consecutive tandem repeats of the VNTR region of MUC1-N domain. The sequence included in the peptide is in bold, whereas gray regions are only included for alignment. Vertical lines mark the 20 amino acids unit of the tandem repeat of which the order of the amino acids was assigned. 27-mer and 23-mer are the peptides designed in this thesis. Residues coloured in red are GalNAc-glycosylated

### 4.3.3 GALNT3

The Wandall study of GALNT3 concluded a similar mechanism of GALNT1 and the same results were reported in the Coelho study. The order of glycosylation was reported as follows:



According to the Coelho study, the mechanism of GALNT3 depends on direct recognition of T8 by the catalytic domain followed by binding of the lectin-like domain with glycosylated T8. This binding assists T20 and S19 glycosylation. The Coelho study also found that GALNT3 showed a very slow glycosylation rate at S19 compared to GALNT2, which was explained by the conformational changes at S19 when T20 is glycosylated. This conformational change inhibits the binding of GALNT3 but has less effect on GALNT2. This observation comes in agreement with Wandall *et al.*, 1997 study where GALNT2 was able to fully glycosylate S19 after 24 hours whereas GALNT1 and GALNT3 failed to do so.

From both GALNT1 and GALNT3 products' structural study, one can easily conclude that the failure in glycosylating T20 at position 1 in TAP-24 (T20') is due to the lack of glycosylated T8 on its N terminal direction (left to T20').

### 4.3.4 GALNT4

Unlike GALNT1-3, GALNT4 is incapable of glycosylating naked peptides, and it functions as a complementary enzyme for pre-glycosylated peptides. Among characterized GALNTs, GALNT4 is the only enzyme capable of glycosylating S9 and T13. The enzyme was identified and cloned by Bennett *et al.*, (1998) <sup>13</sup>, and the activity was tested against MUC1a, MUC1b, TAP24, and 4GalNAc-TAP24, which is the product of GALNT1 and GALNT2 glycosylation (Figure 4.2). GALNT4 showed a negligible reactivity with all forms of naked peptides except for 4GalNAc-TAP24 (Figure 4.2), which concluded the pre-glycosylation condition for GALNT4 activity. A follow-up intensive investigation of site specificity and the role of pre-glycosylation sites in the substrate<sup>11</sup> concluded that: (1) GALNT4 is incapable of glycosylating naked peptides, (2) Glycosylation of T13 in TAP-24 by GALNT4 requires pre glycosylation at T20' (the first T20 in TAP24) and/or Glycosylation at T8, i.e., either of them is needed, (3) Glycosylation of S9

requires both T20' and T8, (4) T20' is essential, whereas lack of T8 can be compensated for by increasing reaction time, (5) GALNT4 is also capable of glycosylating S19 if any of the N terminal T13 or T8 are glycosylated.

The more comprehensive structural study by Coelho *et al.*, 2022 agreed with previous studies. In Coelho study, S9 and T13 in the N terminal repeat of the 4 tandem repeats MUC1 peptide were not fully glycosylated after 48 hours and the structural analysis showed that S9 remained unglycosylated at all, while T13 remained partly glycosylated in that N repeat. Mutating the lectin binding domain and further mechanism study revealed the requirement of the lectin-like domain to facilitate the glycosylation which operates at long range C terminal direction (N-lectin domain-catalytic domain -C). The suggested mechanism is that glycosylation of S19 and/or T20 is required for the lectin-like binding domain to assist glycosylation of S9 and T13. These results also explain the reason for adding an extra T20 to the N terminal of TAP-24 design in previous studies.

## 4.4 Methods to Study GT Specificities

The main challenge in glycobiology is to decipher the complexity of the glycoenzymes-substrates interactome; many approaches have been followed and technologies have been developed to address this challenge. While some approaches focus on increasing the screening throughput of these interactions<sup>15</sup>, others focus on delivering more detailed and precise information<sup>10-14</sup>. The methods can be classified based on the following attributes: validity, relevance, throughput, standardization, and cost. Validity of the method describes the accuracy in measuring the intended dependent variable. In the context of enzyme-substrate interaction, this dependent variable is reactivity in the form of reaction rate or substrate conversion. Relevance, however, describes how representative the design is of the biological, physiological, or clinical phenomenon being tested. Throughput defines the multiplexity of measurements being acquired simultaneously. Standardization defines the uniformity of the method and its output. Examples of standardized methods are nucleic acid measurement methods used in arrays and sequencing. In this chapter the assay methods will not be covered as they are covered in detail in Chapter 2, and the focus is only on the synthesis of glycans or glycoconjugates and expression of glycoenzymes.

### 4.4.1 Current Methods

The advancement of chemical and biological technologies accelerated the process of enzyme expression and substrate synthesis. Materials accessibility combined with the advancement of detection methods have increased the throughput of screening GT-substrate interactions. For instance, Gateway recombinational cloning technology<sup>16</sup> enabled the construction of an expression vectors library covering all human glycoenzymes (GEs). These vectors were designed to express the enzymes in a human cell line host, which preserves the posttranslational modifications required for enzyme activities<sup>17</sup>. Another approach is to express glycoenzymes using an *E. Coli.* cell-free expression system in a high throughput fashion that bypasses purification steps<sup>18</sup>. However, the *E. Coli.* cell-free expression system lacks the posttranslational mechanisms needed for proper folding of human enzymes, and its success was only demonstrated in studying *E. Coli.* enzymes.

As for the substrates, they are either tested as oligosaccharides glycans or as conjugates in the form of glycoproteins. Despite the difficulties entailed by chemical synthesis methods, they

achieved significant success in producing a broad and diverse spectrum of oligosaccharides arrays parallel to that of nucleic acids. That was enabled by improving the chemistry of synthesis and immobilization to systemize the process<sup>19,20</sup>. Despite the utility of chemically synthesized arrays, they suffer from limitations. For example, some structures are difficult to synthesize chemically, and glycoarrays lack the glycoconjugate context i.e., core proteins and lipids. The specificity example of GALNTs presented in section 4.3 of this thesis and other immunogenic studies on glycoconjugates<sup>21</sup> suggest that the lack of the glycoconjugate context compromises the relevance of the information generated from these arrays<sup>22</sup>. This notion necessitated the evolution of glycan arrays to glycoconjugate arrays. Amongst all glycoconjugates, glycoproteins are the most dominant in the cell and therefore, the focus of this thesis will be on the synthesis of glycopeptides.

## 4.5 Glycopeptide Synthesis Methods

Chemoenzymatic synthesis is the standard method to produce glycopeptides. In chemoenzymatic synthesis, the core peptide is chemically synthesized, and the glycan is built enzymatically one sugar moiety at a time. Advancements of the knowledge of enzymes specificities and mechanisms enabled the synthesis of more complex glycopeptides and increased the dimensions of the glycopeptide arrays. For example, Yoshimura *et al.*, (2019)<sup>21</sup> produced an array of 20 glycopeptides covering the glycan structures Tn and T antigens and their sialylated forms, STn and ST, respectively. These glycans were alternated for all the 5 possible glycosylation sites of MUC1 tandem repeat. The method used in the Yoshimura study is chemical synthesis of glycopeptides carrying the first GalNAc residue, followed by enzymatic glycosylation to produce the other structures.

Another approach, synthetic biology, was used to develop genetically engineered human embryonic kidney (HEK) cells. The endogenous glycosylation pathways in these cells were rationally modified to synthesize specific glycan structures by knock-in or knock-out of specific GTs. The result was a library of over 100 isogenic HEK cells, each specialized in producing a specific glycan structure. When these engineered cells are used to express a reporter core protein, such as MUC1-N, the expressed reporter is displayed on the surface of the cell with the unique

glycan structure of the cell<sup>22</sup>. The main utility of the isogenic HEK cells library is assessing binding with antigens and lectins in a biologically relevant context<sup>22,23</sup>. Despite the utility and the biological relevance of the isogenic HEK cells library, the endogenous machinery of the cells cannot be fully controlled and optimized to avoid side reactions and incomplete glycosylation. Additionally, the pathways can only be modified to a certain degree that is limited to the minimal cell survival needs of these glycosylation pathways. These limitations result in heterogeneity in the expressed structures and leave some structures unattainable by this method. Therefore, for kinetics parameterization and mechanistic studies, diverse pure well-identified structures are needed, and that makes *in vitro* methods indispensable for such applications.

*In vitro* methods of glycopeptide synthesis start with the synthesis of the core peptides. There are three available options to achieve that: chemical synthesis, recombinant expression, and cell free translation. Each method with its limitations is presented in the next section, except for the cell-free system.

## 4.6 Short Peptide Synthesis

Short peptides have been produced as pharmaceutical and food preservative agents for over a century<sup>24,25</sup>. They can be synthesized chemically or biologically. The main method of biologically synthesized peptides is expression as recombinant peptides in various host cells, with *E. Coli.* being the first choice. In 1982, the FDA approved the first recombinant peptide expressed in *E. Coli.*, human insulin, under the commercial brand name Humulin by Eli Lilly<sup>24</sup>.

### 4.6.1 Chemical Synthesis

Solid phase peptide synthesis (SPPS) is the most common chemical synthesis technology. In SPPS, the peptide is anchored to a resin support via C terminus binding and amino acids with temporary protecting groups on the  $\alpha$  amino group and side chains are added one by one (Figure 4.3). The growth of the chain takes place in the C to N terminal direction. Boc and Fmoc protecting groups are commonly used and can be removed with trifluoroacetic acid or 20% piperidine in dimethylformamide. Each cycle of amino acid addition involves three steps: (1) deprotection of the  $\alpha$  amino group of the last added amino acid, (2) washing the resin, (3) then addition of the next protected amino acid (Figure 4.3). One limitation of SPPS is incomplete coupling and deprotection reactions, which leads to accumulation of byproducts. Aggregation of

growing peptides can also be a bottle neck that hinders SPPS process. The cost of the resin and equipment needed for automation is another limiting factor of adapting SPPS in research laboratories. In addition to SPPS, solution phase synthesis (SPS) is used to coupling single amino acids in solution. Fragment condensation is another method of synthesis used for longer peptides, which typically is a hybrid approach between SPPS and SPS when fragments are synthesized using SPPS and ligated using SPS. Microwave-assisted SPPS has been developed to improve synthesis, allowing for higher yields and lower degrees of racemization in a shorter time<sup>26</sup>.

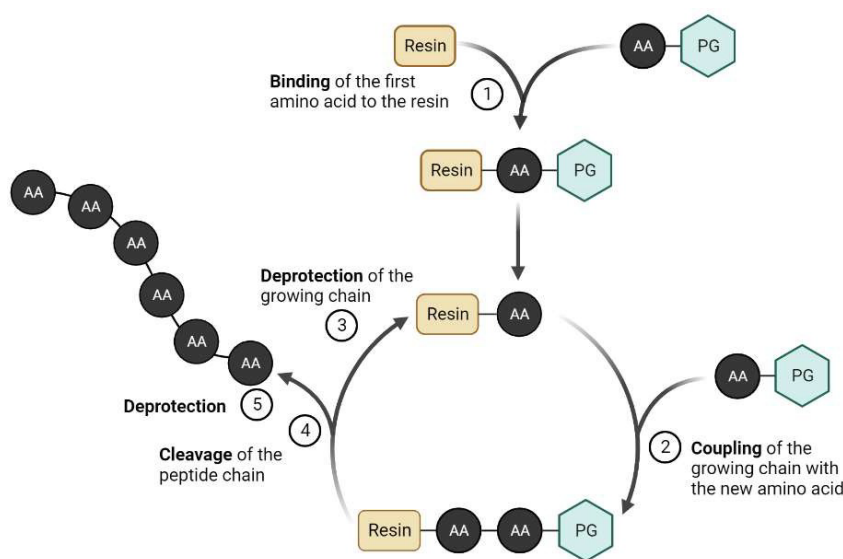


Figure 4.3. Solid Phase Peptide Synthesis (SPS). Steps 1-3 show the cycle of chain growth. Steps 4-5 are termination of synthesis. AA: amino acid, PG: protecting group.

#### 4.6.2 Recombinant Expression of Peptides

Due to the high cost and the length and complexity of the chemical synthesis steps, recombinant expression of peptides in *E. Coli.* or yeast is an appealing alternative as an economic, environment-friendly, and accessible technology. However, some factors such as: intrinsic disorder, susceptibility to aggregation and proteolysis, toxicity to the host cell, and the need for purification strategy, pose difficulties to the success of the recombinant expression approach. Most short peptides are characterized by their high intrinsic disorder, which refers to the lack of a defined three-dimensional structure in a protein or peptide under physiological conditions. It has been shown that highly disordered proteins or peptides can be challenging to express in large quantities due to their increased propensity for aggregation, misfolding, and degradation<sup>27</sup>. The amino acid composition of a peptide can affect its expression success. For example, the presence of hydrophobic or charged amino acids may affect the solubility and stability of the peptide, which makes them more susceptible to the host cell's proteases. Some short peptides can be toxic to the host cell, such as the case of antimicrobial peptides. Expression of such peptides inhibits cell growth and reduces the yield significantly. The crude product of recombinant expression is more complex compared to chemical synthesis product, due to presence of other proteins, nucleic acids, and other media and cell metabolites<sup>25</sup>. Finally, molecular biology tools that are routinely used to detect protein during and after purification, such as UV absorbance at 280 nm, SDS-PAGE gel separation, and other staining and detection methods, are not always applicable to short peptides.

To overcome these limitations in expression and purification of short peptides, short peptides are expressed in the form of fusion proteins. The fusion construct includes a carrier protein fused to the peptide. Carriers are selected for their solubility and folding enhancing properties, and they incorporate some form of affinity tag to facilitate the purification<sup>24</sup>. Some carriers were developed to overcome the solubility and toxicity problem by shielding the peptide in preventing its aggregation and cytotoxic effect<sup>28</sup>.

Overall, despite the lower environmental impact and production cost of recombinant peptide expression compared to chemical methods, recombinant peptide production entails long and complex development and scaling up phases, which includes: the selection of the expression host, design of the plasmid and fusion construct, expression conditions, and purification

strategies. These steps need to be tailored to account for the specific intrinsic properties of each peptide and for the target production scale and downstream applications. Once the optimum design is accomplished, recombinant expression outperforms chemical synthesis in the production phase.

## 4.7 Fusion Expression

As mentioned above, a multi steps optimization process is required to achieve a successful recombinant peptide expression. Thus, the design of fusion protein expression requires careful selection of the host cell, plasmid, carrier protein, affinity tags (if required), and linkers. Other technical steps, such as temperature, induction methods and choice of media will not be covered as these factors are subject to a trial-and-error approach and cannot always be predetermined in the design phase. Some of the following parameters can be overwhelmingly numerous. Therefore, the emphasis will be on the parameters selected for this study.

### 4.7.1 Host and Plasmids

Selection of the host is the first step in the design pipeline. The host cell determines the compatibility of the designed fusion to the expression machinery and the choice of plasmid and its elements. Host cells are selected for their unique properties that are required in the performed expression. For example, eukaryotic cells, such as the human embryonic kidney cell (HEK 293) is the host choice when eukaryotic posttranslational modification are required in the protein.

*E. Coli.* is a commonly used host for recombinant protein expression due to its cost-effectiveness, rapid growth, and productivity. However, it lacks proper post-translational modifications (PTMs) and can face challenges like inclusion body formation, codon bias, and endotoxin issues. To overcome these limitations, techniques such as fusion tags, cofactor supplementation, and co- expression with chaperones have been employed to prevent inclusion body formation and improve soluble expression<sup>29</sup>. Many strains of *E. Coli.* have been developed, each with specific application and limitations<sup>30</sup>. The system used in this thesis is BL21 (DE3) strain with pET plasmid system.

BL21 (DE3) is a commonly used *E. Coli.* strain for recombinant protein expression. It lacks OmpT and Lon proteases, resulting in reduced proteolytic degradation of heterologous proteins.

The strain carries the DE3 lysogeny system in its genome, enabling T7 RNA polymerase expression upon isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG) induction. The T7 RNA polymerase facilitates high-level protein expression by driving transcription from T7 promoter upstream of the target gene sequence in the pET plasmids. The DE3 system in BL21 (DE3) contains lac promoter regulated by the bacterial RNA polymerase, and lac operator that normally inhibits T7 polymerase expression. However, this inhibition is reversed upon induction with lactose or its analogue, IPTG. Consequently, the T7 RNA polymerase is transcribed, and its binding to the T7 promoter on the expression vector initiates transcription of the downstream target gene. Target gene transcription is also inhibited by another lac operator, which required IPTG to remove this inhibition. This inducible expression is important to maximize cell growth before channelling the cell metabolism into protein expression<sup>31</sup>.

#### **4.7.2 Carrier Proteins**

Thioredoxin (Trx) and glutathione transferase (GST) are the two most popular carriers<sup>32</sup>, which are selected for their high solubility and their chaperone-like function. GST's larger size can negatively impact yield by creating an unfavourable peptide-to-carrier ratio. However, its intrinsic affinity to glutathione can be deployed in the purification process. Trx-fused peptides on the other hand requires additional affinity tags, such as His-tag. Optimization of soluble expression usually requires expression at low temperature to slow cell metabolism and expression rate, allowing enough time for protein maturation and folding. Other carrier proteins are designed to express the fusion protein in an insoluble form aggregated in inclusion bodies. This strategy is believed to provide full protection of the peptide from proteolytic activities of the cell and prevent the cytotoxic effect of the peptide on the cell. Examples of such carriers include amidophosphoribosyltransferase (PurF)<sup>32</sup> and ketosteroid isomerase tag (KSI-tag)<sup>14</sup>.

#### **Superfolder GFP**

Green Fluorescent Protein (GFP) was initially cloned from *Aequorea victoria* (avGFP) and gained prominence due to its ability to emit green light when exposed to specific wavelengths. The crystal structure of GFP reveals a barrel or  $\beta$ -can structure consists of 11  $\beta$ -sheets and an  $\alpha$ -helix inside with two helical structures on each side of the can<sup>33</sup>. avGFP has limitations

in terms of folding efficiency and brightness. To address these issues, modifications were made to avGFP (F64L and S65T), resulting in enhanced-GFP (eGFP) with improved brightness and excitation efficiency.

Further advancements led to the creation of superfolder GFP (sfGFP) through molecular evolution strategies. sfGFP exhibits several beneficial mutations: (1) mutations that enhanced its folding efficiency (F99S, M153T, and V163A), (2) mutation that stabilized the  $\beta$ -strands (S30R), (3) mutations that eliminated hydrophobic residues at the dimer interference<sup>34</sup>. The improvements in sfGFP resulted in minimized dimerization and aggregation, resistance to denaturation, and enhanced folding (faster folding kinetics)<sup>35</sup>. When sfGFP is fused with other proteins, the fusion construct exhibited these improved properties, hence, sfGFP overcame the folding defects observed in avGFP and other GFP variants when fused with other proteins, making it highly suitable for biotechnology and protein expression applications<sup>17,35</sup>.

#### **4.7.4 Linkers**

The most employed strategy for constructing fusion proteins is the end-to-end fusion of genes, which involves linking the fusion components via peptide linkers. The properties of linker sequences, such as length, amino acid composition, and hydrophobicity, have been extensively investigated. Suitable linkers are crucial for providing stable linkage between fusion partners without affecting their original biological functions. The flexibility of the linker plays a significant role in determining the relative orientation of the fused proteins. A study by Li *et al.*, (2016)<sup>36</sup> aimed to explore the impact of linker flexibility on fusion proteins. Researchers in this study created a linker library using combinations of two elementary units: a flexible unit (GGGGS), and a rigid unit (EAAAK). The investigation revealed that linkers with more rigid units exhibited characteristics such as helical conformation stabilized by hydrogen bonding, and reduced distance fluctuation between the N- and C-termini of the linker. Desired properties in a linker can be achieved by optimizing a combination of the rigid and flexible units.

### **4.7 In vitro Synthetic Biology Approach**

A fusion protein consists of MUC1 tandem repeat fused with sfGFP was designed to study MUC1 O-glycosylation pathway. The fusion construct is optimized to synthesize the glycopeptides of the pathway following an *in vitro* synthetic biology approach. In this approach,

the current knowledge of the pathway is used to reconstruct the pathway *in vitro*. The synthesis method aims to mimic the temporal regulation of glycan growth by generating the substrates in the same sequence as in the cell. This sequential synthesis along with the functional features of the fusion proteins enables investigation of enzymes specificities and measure their kinetics parameters, while simultaneously constructing an array of biologically relevant glycopeptides. Peptide and fusion construct features that enable this approach are explained in the following sections in detail. The GALNTs used in this study are GALNT1, GALNT2, and GALNT4. These enzymes were selected because they are well characterized, and they cover all known variation of mechanisms; different sites affinity and different direction of lectin-like binding domain. Additionally, these three enzymes are capable, when combined, of glycosylating all the five unique sites on MUC1 tandem repeat.

#### **4.7.1 Peptide Design**

The purpose of this design is to quantify enzyme specificities and synthesize biologically relevant glycopeptides that capture the structural aspects in their native biological context. This is particularly important for GALNTs and their products due to their rigorous specificity. Therefore, the optimum sequence must account for all the determinant variables of GALNTs specificities, and each variable is only represented once in the sequence. These variables are explained in detail in section 4.3. To achieve that, the following criteria was set for an optimum peptide sequence: (1) includes all the five unique potential glycosylation sites of MUC1 tandem repeat, and each site is only represented once in the sequence, (2) none of the sites are terminal and there is sufficient extension of the tandem repeat sequence in both directions from the glycosylation site to account for the specificity of the peptide-binding site of the catalytic domain, (3) the frame of the sequence must be optimized for the position of each site in relation to the rest of the sites to account for the direction specificity of the lectin-like binding domain in GALNTs.

The sequence of TAP24, as the standard peptide design used in previous studies is inadequate for our purpose because it does not satisfy any of the criteria above. The additional T at the N terminus is repeated in the sequence and the relative directions of the sites are suboptimal. For

example, the N terminal T20 and S19 are difficult to glycosylate as explained earlier in section 4.3 (Figure 4.4)

Another peptide designed in this project is the 27-mer peptide, which was obtained by removing the N terminal T from TAP24 and expanding the C terminal sequence until before the next T8 (Figure 4.2). The 27-mer peptide satisfies the first and second criteria. However, it lacks S19 and/or T20 left to S9 and T13, which are required for GALNT4 glycosylation of S9 and T13 (section 4.3 and Figure 4.5). Nonetheless, this peptide was synthesized and used as a control in this thesis to demonstrate the effect of the frame on the activity.

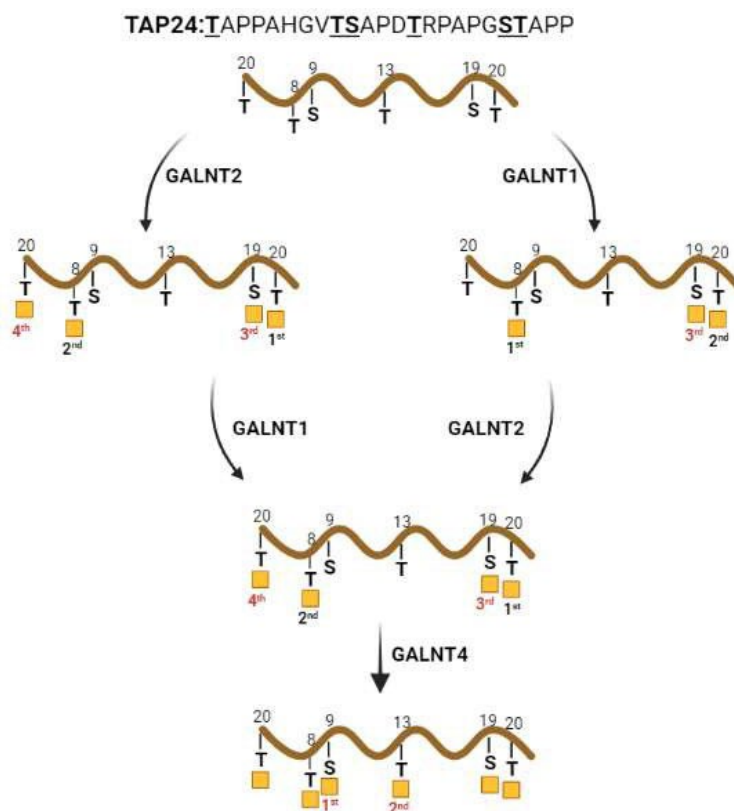


Figure 4.4. Glycosylation of TAP24 peptide with GALNT1,2 and 4. Two pathways are presented according to the curated mechanisms from literature, GALNT1-GALNT2-GALNT4 to the right, and GALNT2-GALNT1-GALNT4 to the left. The order of the newly added GalNAc is shown (if established in literature). The order label is coloured in red to indicate incomplete reaction at the site.

To address the limitation of the 27-mer, an optimum sequence was obtained by shifting the frame of the sequence to start with S19 and T20 from the N terminal end of the peptide (Figure 4.2). Despite being an unusual frame compared to the designs commonly used in literature, this sequence is expected to achieve the following (Figure 4.6): (1) 23-mer reaction with GALNT1 or GALNT2 glycosylates a single site only, which is the first site recognized by the catalytic domain. Data obtained from these reactions reflect the kinetic parameters of the catalytic activity in isolation i.e., independent from the lectin-like binding domain activity. (2) Glycosylation of S19 in the 23-mer peptide is expected to be complete, unlike that for both TAP24 and the 27-mer peptides. T8 in the 23-mer is at C terminal direction relative to S19, which is the requirement for GALNT2 lectin-like binding domain assisted glycosylation of S19. (3) The product of 23-mer peptide glycosylation by GALNT1 and GALNT2 in combination, or the sequential reaction of GALNT1 followed by GALNT 2 (the right-side pathway in Figure 4.6), is an optimum substrate for GALNT4. Glycosylated S19, T20, and T8 (in the product of GALNT1 and GALNT2) are at the N terminal side to S9 and T13. This alignment provides the perfect structure required by GALNT4 lectin-like binding domain assisted glycosylation of S9 and T13.

In conclusion the 23-mer peptides satisfies the criteria above and it is expected to outperform TAP24 and the 27-mer by achieving fast and complete glycosylation of the 5 sites using GALNT1, GALNT2 and GALNT4.

**27-mer:** APPAHGVTSAPDTRPAPGSTAPPAHGV

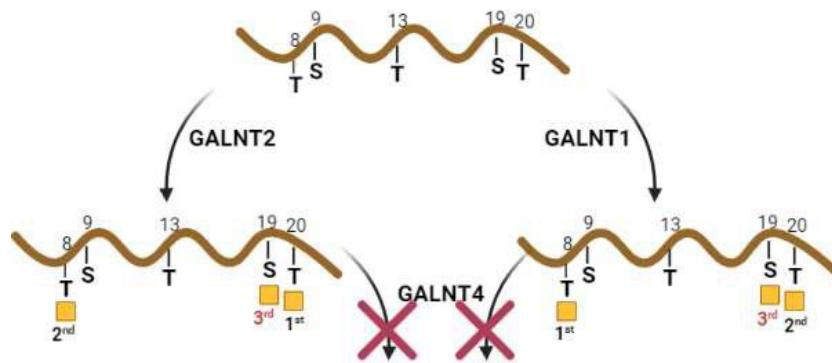


Figure 4.5. Glycosylation of 27-mer peptide with GALNT1,2 and 4. Lack of S19 and /or T20 at N terminal direction to S9 and T13 impairs GALNT4 reactivity. The order of the newly added GalNAc a result of the reaction is shown (if established in literature). The order label is coloured in red to indicate incomplete reaction at the site.

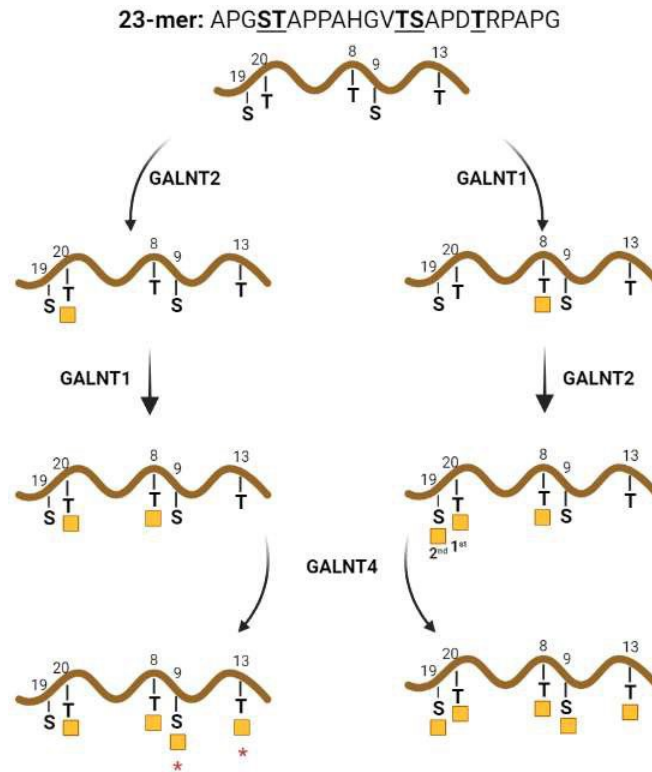


Figure 4.6. Glycosylation of 23-mer peptide with GALNT1,2 and 4. Two pathways are presented according to the curated mechanisms from literature, GALNT1-GALNT2-GALNT4 to the right, and GALNT2-GALNT1-GALNT4 to the left. The order of the newly added GalNAc is shown (if established in literature). The red asterisks indicate incomplete reaction at the site.

#### 4.7.2 Fusion Protein Design

Since no posttranslational modifications on the expressed peptide are required, *E. Coli.* strain BL21 (DE3) was selected as an expression host. The main application of the fusion protein is to participate in glycosylation reactions. Therefore, solubility of the fusion protein with minimum interaction between the carrier are the main requirements of the design. Among all the reviewed carrier proteins, sfGFP was selected for its folding efficiency, minimized dimerization, and enhanced solubility (section 4.7.2). To minimize the possibility of interaction with the peptide, sfGFP was separated from the peptide with linker (Linker 2) that contains three rigid units, and three flexible units (from N to C direction). The map of the fusion protein is illustrated in Figure 4.7. Another rigid linker (Linker 1) was also used to separate His tag from sfGFP in the N terminus.

The peptide was designed to be cleaved in its native sequence from the carrier protein using tobacco etch virus (TEV) protease. The recognition sequence of the TEV protease includes the first (N terminal) amino acid of the peptide, imposing a limitation on the amino acid allowed at this position (P1'). Fortunately, TEV protease has high cleavage efficiency with alanine in P1' position in TEV protease sequence<sup>37</sup>, hence the frames of both 23-mer and 27-mer peptides were selected to start with sequence (Figure 4.2). As mentioned above, Linker 2 includes flexible region adjacent to TEV protease site. This was designed to provide flexibility that might be required for TEV protease catalytic activity, which can reflect on enhancing the cleavage efficiency. Finally, an N terminal hexa-histidine-tag(6xHis-tag) was added for purification. Full sequence of the fusion protein is demonstrated in the Appendix. The fusion protein sequence was synthesized and inserted into pET-21 b (+).

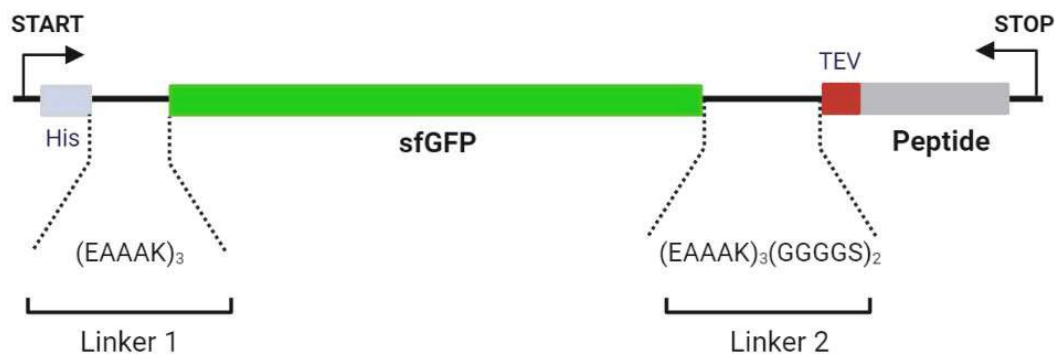


Figure 4.7. Fusion Protein Design. Template of the fusion protein used for both 27-mer and 23-mer peptide expression.

## 4.8 Human GT Enzymes Expression

All human GT enzymes were expressed in HEK293 cells by transfecting the cells with pGEn2 expression vectors established by Moremen *et al.*, (2018)<sup>17</sup>. The enzymes were expressed with a secretion signal peptide and sfGFP tag. The tag is cleavable by TEV protease enabling obtaining the enzymes in their native forms. Expressing human GTs in HEK293 preserve the native posttranslational modifications, particularly the glycan, required for proper folding and function of the enzymes.

## References

- 1 Gupta, R. *et al.* Global analysis of human glycosyltransferases reveals novel targets for pancreatic cancer pathogenesis. *Br J Cancer* **122**, 1661-1672 (2020).
- 2 Schjoldager, K. T., Narimatsu, Y., Joshi, H. J. & Clausen, H. Global view of human protein glycosylation pathways and functions. *Nat Rev Mol Cell Biol* **21**, 729-749 (2020).
- 3 Rini, J. M. & Esko, J. D. in *Essentials of Glycobiology* (eds A. Varki *et al.*) 65-75 (2015).
- 4 Steentoft, C. *et al.* Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J* **32**, 1478-1488 (2013).
- 5 Revoredo, L. *et al.* Mucin-type O-glycosylation is controlled by short- and long-range glycopeptide substrate recognition that varies among members of the polypeptide GalNAc transferase family. *Glycobiology* **26**, 360-376 (2016).
- 6 Brockhausen, I., Wandall, H. H., Hagen, K. G. T. & Stanley, P. in *Essentials of Glycobiology* (eds A. Varki *et al.*) 117-128 (2022).
- 7 Ju, T. & Cummings, R. D. A unique molecular chaperone Cosmc required for activity of the mammalian core 1  $\beta$  3-galactosyltransferase. *Proc Natl Acad Sci U S A* **99**, 16613-16618 (2002).
- 8 Bergstrom, K. *et al.* Core 1- and 3-derived O-glycans collectively maintain the colonic mucus barrier and protect against spontaneous colitis in mice. *Mucosal Immunol* **10**, 91-103 (2017).
- 9 de Las Rivas, M., Lira-Navarrete, E., Gerken, T. A. & Hurtado-Guerrero, R. Polypeptide GalNAc-Ts: from redundancy to specificity. *Curr Opin Struct Biol* **56**, 87-96 (2019).
- 10 Wandall, H. H. *et al.* Substrate specificities of three members of the human UDP-N-acetyl- $\alpha$ -D-galactosamine: Polypeptide N-acetylgalactosaminyltransferase family, GalNAc- T1, -T2, and -T3. *J Biol Chem* **272**, 23503-23514 (1997).
- 11 Hassan, H. *et al.* The lectin domain of UDP-N-acetyl-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase-T4 directs its glycopeptide specificities. *J Biol Chem* **275**, 38197-38205 (2000).
- 12 Wandall, H. H. *et al.* The lectin domains of polypeptide GalNAc-transferases exhibit carbohydrate-binding specificity for GalNAc: lectin binding to GalNAc-glycopeptide substrates is required for high density GalNAc-O-glycosylation. *Glycobiology* **17**, 374-387 (2007).
- 13 Bennett, E. P. *et al.* Cloning of a human UDP-N-acetyl- $\alpha$ -D- Galactosamine:polypeptide N-acetylgalactosaminyltransferase that complements other GalNAc-transferases in complete O-glycosylation of the MUC1 tandem repeat. *J Biol Chem* **273**, 30472-30481 (1998).
- 14 Coelho, H. *et al.* Atomic and specificity details of mucin 1 O-glycosylation process by multiple polypeptide GalNAc-transferase isoforms unveiled by NMR and molecular modeling. *JACS Au* **2**, 631-645 (2022).
- 15 Voglmeir, J., Sardzik, R., Weissenborn, M. J. & Flitsch, S. L. Enzymatic glycosylations on arrays. *OMICS* **14**, 437-444 (2010).
- 16 Reece-Hoyes, J. S. & Walhout, A. J. M. Gateway recombinational cloning. *Cold Spring Harb Protoc* **2018**, pdb top094912 (2018).
- 17 Moremen, K. W. *et al.* Expression system for structural and functional studies of human glycosylation enzymes. *Nat Chem Biol* **14**, 156-162 (2018).

- 18 Ban, L. *et al.* Discovery of glycosyltransferases using carbohydrate arrays and mass spectrometry. *Nat Chem Biol* **8**, 769-773 (2012).
- 19 Blixt, O. *et al.* Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proc Natl Acad Sci U S A* **101**, 17033-17038 (2004).
- 20 Fukui, S., Feizi, T., Galustian, C., Lawson, A. M. & Chai, W. Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions. *Nat Biotechnol* **20**, 1011-1017 (2002).
- 21 Yoshimura, Y. *et al.* Products of chemoenzymatic Ssynthesis representing MUC1 tandem repeat unit with T-, ST- or STn-antigen revealed distinct specificities of anti-MUC1 antibodies. *Sci Rep* **9**, 16641 (2019).
- 22 Narimatsu, Y. *et al.* An atlas of human glycosylation pathways enables display of the human glycome by gene engineered cells. *Mol Cell* **75**, 394-407 e395 (2019).
- 23 Bull, C., Joshi, H. J., Clausen, H. & Narimatsu, Y. Cell-based glycan arrays: A practical guide to dissect the human glycome. *STAR Protoc* **1**, 100017 (2020).
- 24 Wegmueller, S. & Schmid, S. Recombinant peptide production in microbial cells. *Curr Org Chem* **18**, 1005-1019 (2014).
- 25 Wang, L. *et al.* Therapeutic peptides: current applications and future directions. *Signal Transduct Target Ther* **7**, 48 (2022).
- 26 Chandrudu, S., Simerska, P. & Toth, I. Chemical methods for peptide and protein production. *Molecules* **18**, 4373-4388 (2013).
- 27 Argudo, P. G. & Giner-Casares, J. J. Folding and self-assembly of short intrinsically disordered peptides and protein regions. *Nanoscale Adv* **3**, 1789-1812 (2021).
- 28 Ishida, H., Nguyen, L. T., Gopal, R., Aizawa, T. & Vogel, H. J. Overexpression of antimicrobial, anticancer, and transmembrane peptides in *Escherichia coli* through a calmodulin-peptide fusion system. *J Am Chem Soc* **138**, 11318-11326 (2016).
- 29 Tripathi, N. K. & Shrivastava, A. recent developments in bioprocessing of recombinant proteins: expression hosts and process development. *Front Bioeng Biotechnol* **7**, 420 (2019).
- 30 Joseph, B. C. *et al.* An overview of the parameters for recombinant protein expression in *Escherichia coli*. *J. Cell Sci. Ther.* **6**, 1-7 (2015).
- 31 Mierendorf, R. C., Morris, B. B., Hammer, B. & Novy, R. E. Expression and purification of recombinant proteins using the pET system. *Methods Mol Med* **13**, 257-292 (1998).
- 32 Li, Y. Carrier proteins for fusion expression of antimicrobial peptides in *Escherichia coli*. *Biotechnol Appl Biochem* **54**, 1-9 (2009).
- 33 Yang, F., Moss, L. G. & Phillips, G. N., Jr. The molecular structure of green fluorescent protein. *Nat Biotechnol* **14**, 1246-1251 (1996).
- 34 Pedelacq, J. D. & Cabantous, S. Development and applications of superfolder and split fluorescent protein detection systems in biology. *Int J Mol Sci* **20**, 3479 (2019).
- 35 Pedelacq, J. D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**, 79-88 (2006).
- 36 Li, G. *et al.* Construction of a linker library with widely controllable flexibility for fusion protein design. *Appl Microbiol Biotechnol* **100**, 215-225 (2016).
- 37 Kapust, R. B., Tozser, J., Copeland, T. D. & Waugh, D. S. The P1' specificity of tobacco etch virus protease. *Biochem Biophys Res Commun* **294**, 949-955 (2002).

# **Chapter 5:**

## **In vitro Glycopeptide Synthesis**

## 5.1 Chapter Overview

An *in vitro* synthetic biology approach was proposed in Chapter 4 to construct MUC1 O-glycans. A fusion protein, consisting of the MUC1 core peptide and a carrier protein, was designed for expression in *E. Coli*. The fusion protein elements were optimized to maximize solubility, enhance the *in vitro* enzymatic glycosylation of the peptide, and enable the recovery of the products through a simple one-step purification. The peptide in the fusion protein is designed to undergo sequential enzymatic glycosylation *in vitro*. Glycosyltransferases, constituting the biosynthesis pathway, were used in a stepwise fashion for the glycosylation process, which enabled the synthesis of the biologically and clinically relevant glycopeptide structures in the pathway. The fusion is designed to minimize interaction between the peptide and the carrier protein, while maximizing solubility. Therefore, affinity and kinetics characterization of the glycopeptides is possible in their fusion form. Moreover, the design also enables the cleavage of the synthesized glycopeptides from their fusion protein at any step of the synthesis, when needed in their native forms. For example, this can be done to perform structure confirmation by LC-MS (Figure 5.1).

The chapter presents the validation of the design. Two constructs were used: the 23-mer fusion protein, as an optimum sequence and the 27-mer fusion protein as an example of a suboptimal frame design (section 4.7.1). Expression in *E. Coli* and purification were optimized and cleavage of the tag and recovering the native peptide were performed. Activity tests of different GALTs in different combinations were conducted on the fusion proteins and all cleavage products, confirming the activity of the peptide in both its fused and native forms. The inertness of the tag was also confirmed using this activity test. The preparative synthesis was optimized, and the completion of each synthesis step was confirmed by performing a series of activity tests. Three glycopeptides were produced as proof of concepts: two represents different saturation of Tn antigen, (GalNAc)<sub>3</sub> 23-mer fusion and (GalNAc)<sub>5</sub> 23-mer fusion, and one represent core 1 in its full saturation form, (Gal-GalNAc)<sub>5</sub> 23-mer fusion. Final confirmation of the synthesized structures was carried out by performing LC-MS analysis on the purified peptides. The *in vitro* synthetic biology method developed in this thesis provides a time and cost-effective tool to produce a wide range of glycopeptides for binding and kinetics studies.

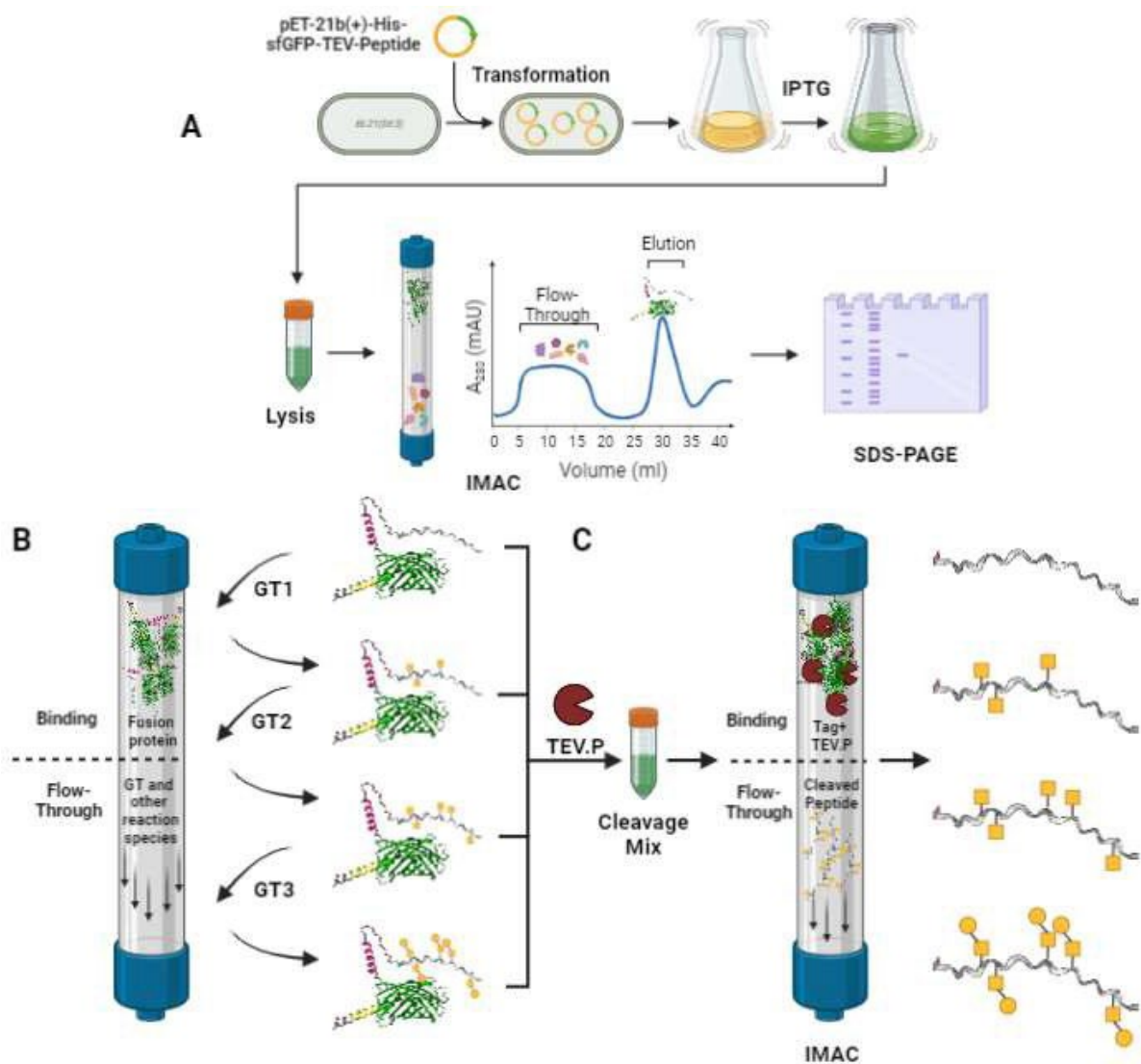


Figure 5.1. Overall Glycopeptides synthesis pipeline. The steps of the process are demonstrated: A. Expression of fusion protein in *E. coli*. and IMAC purification. B. The Sequential glycosylation of peptides where GT1, GT2, and GT3 are glycosyltransferase reactions. C. Removal of the tag and recovery of pure peptides at any stage of the sequential glycosylation.

## 5.2 Expression of The Fusion Proteins

### 5.2.1 Structure prediction

The neural network-based model, AFold<sup>1</sup> was used to predict the structure of the fusion protein using the full sequence of the construct (Figure 5.2). A lighter version of the software was used via the web interface, Colab Network<sup>1</sup>, to perform the structure prediction. The structure shows both rigid regions of the linkers predicted with confidence to have an  $\alpha$ -helix structures. As was intended from the design, the rigid linkers extend away from the  $\beta$ -can structure of sfGFP. The prediction model shows low confidence structures for the peptide, TEV sequence, and the flexible region of linker 2. This is also anticipated as these sequences do not conform into secondary or tertiary structures when they are in isolation. This result suggests flexibility in the linker around TEV protease site, which was intended by design to improve the cleavage efficiency. The rigid portion of linker 2 also provides sufficient separation of the peptide from sfGFP to prevent unwanted interaction that might interfere with the enzymatic glycosylation.

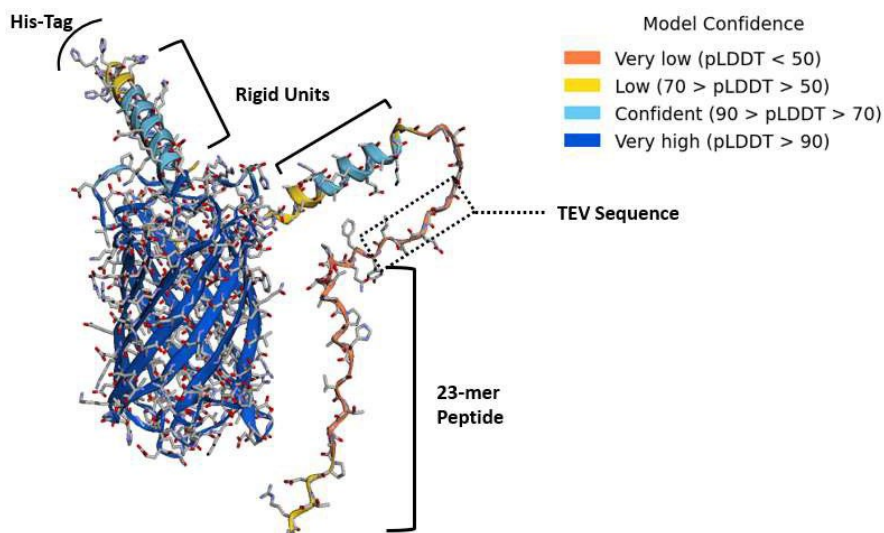


Figure 5.2. AFold structure prediction of the fusion protein. The full sequence of the fusion protein was used to predict the structure via the web interference, Colab Network. The structure shows the His-tag at the N terminal of the fusion attached to the  $\beta$ -can structure of sfGFP via Linker 1. Linker 2 (with both its rigid and flexible regions), TEV site, and the peptide are at the C terminal.

### 5.2.2 Fusion Proteins Expression

Expression was carried out for 23-mer and 27-mer fusion protein by transforming BL21 (DE3) cells with the corresponding transcription plasmid (Appendix, A.1.3). Optimization was carried out on the 23-mer fusion protein, and the optimum protocol was applied to the 27-mer fusion protein. Transformation and expression were carried out using the general protocols (Appendix, A.2.2, A.2.3, and A.2.4). Expression was carried out overnight in LB media at 37 °C. Purification of the fusion protein was carried out using IMAC purification (Appendix, A.2.6). The chromatogram of IMAC purification shows multiple overlapping elution peaks (the background of Figure 5.3). Investigating the pooled elution fractions from the IMAC column using SDS-PAGE gel, showed a major band corresponds to the expected molecular weight of the 23-mer fusion protein (~ 34 kDa). However, two minor bands at 30 and 20 kDa were also observed (trial gel not shown). The fractions were filtrated using a filter with molecular weight cutoff at 30 kDa. Five cycles of filtration and dilution with IMAC binding buffer (A.2.1) were performed, and the fraction above the filter was collected at the end of the procedure. SDS-PAGE was performed on the fraction before filtration and after filtration from the fraction above the filter (Figure 5.3). The results of SDS-PAGE show that MWCO filtration did not remove the impurities.

IMAC was then performed on the fraction above the filter obtained from MWCO filtration with a broader elution gradient to get better separation (Figure 5.4). Separation of two peaks was achieved using this method. Both fractions were green indicating presence of sfGFP. The SDS-PAGE was performed on the elution peaks confirmed that both peaks contain the 20 kDa band, whereas the 30 kDa band was only detected in the first peak (Figure 5.4). These results suggested that the observed impurities are proteins that are aggregated with the fusion protein. The suspected aggregation is not uniformed, and it impacts the protein affinity to the IMAC column.

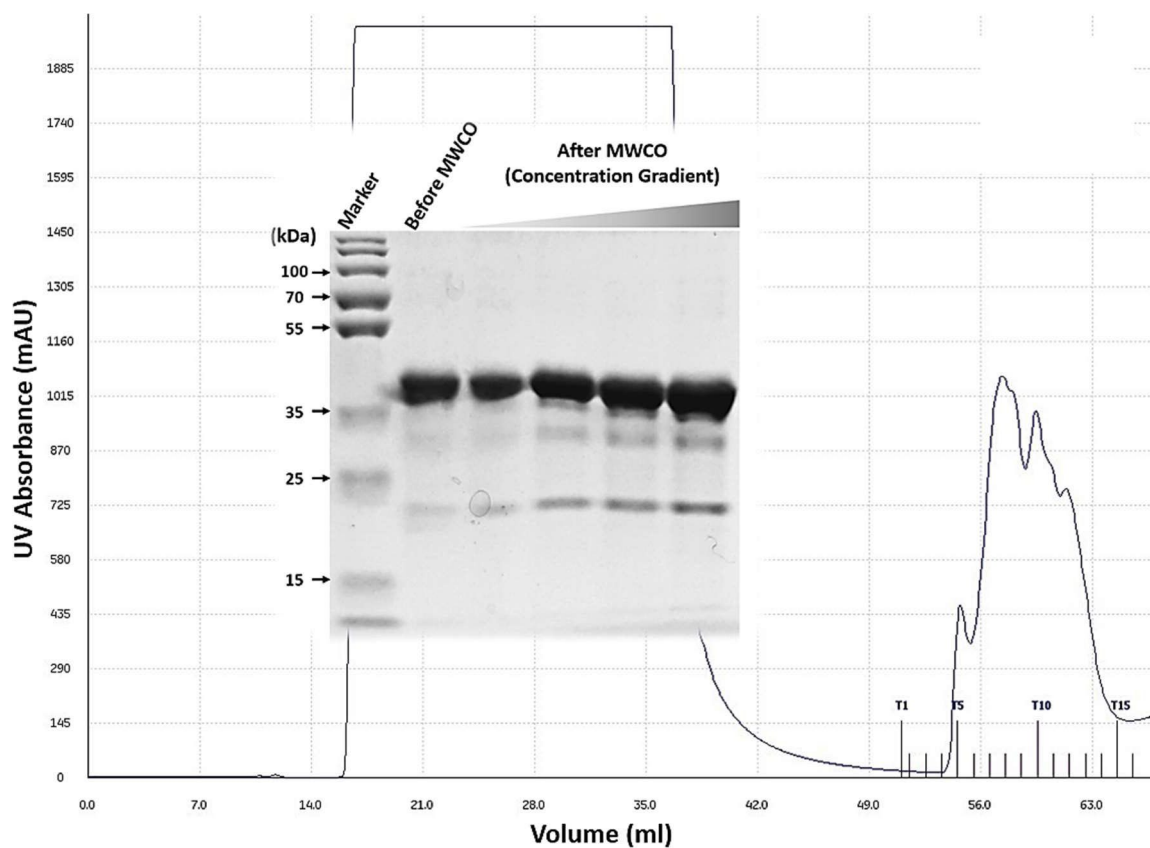


Figure 5.3. Purification of 23-mer fusion protein. Cell lysate of *E. Coli*. transformed with the 23-mer carrying plasmid were purified by IMAC (the chromatogram in the background). Elution peaks are shown after 50 ml of the buffer run. The elution peaks were all pooled and filtrated by molecular weight cutoff (MWCO) filter and the fraction above the filter were collected. SDS-PAGE Gel results show the elution fraction before filtration and after filtration with a gradient concentration of the total protein loaded in the gel.

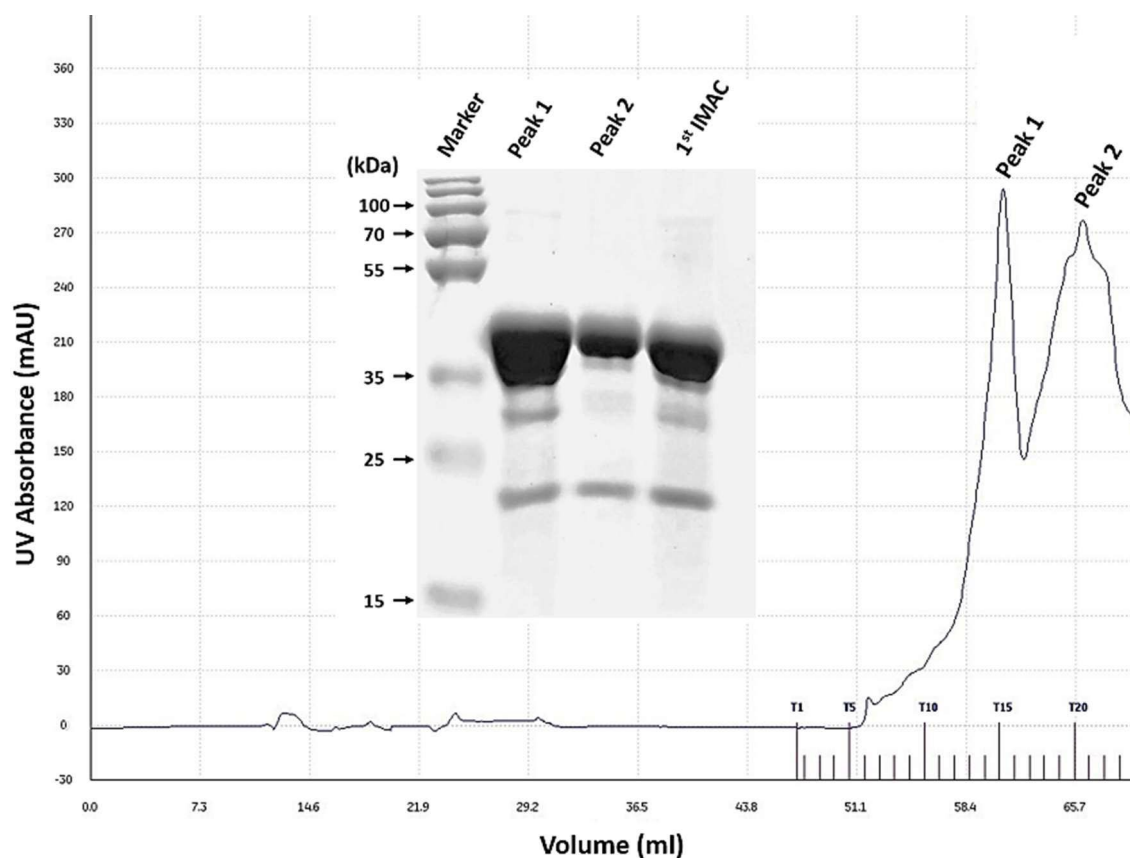


Figure 5.4. Second IMAC purification. The fraction above the filter obtained from the MWCO filtration was used in IMAC purification, but with a broader elution gradient. The fractions corresponding to the two peaks observed on the chromatogram were detected on SDS-PAGE.

To investigate the possibility of aggregation, fractions corresponding to both peaks were mixed again, buffer was exchanged to the standard freezing buffer, and size exclusion (SEC) was performed. The protocols used for buffer exchange and SEC are explained in the appendix (A.2.7, and A.2.8). The chromatogram of the size exclusion experiment (Figure 5.5) showed two peaks. Fractions corresponding to these two peaks were collected. Both fractions were green, which indicates the presence of sfGFP. SDS-PAGE was performed on both fractions (Figure 5.5). The gel results showed that while the first peak contains the same impurities detected before, the second peak contained mostly a pure fusion protein, with a minor amount of the 30 kDa impurity. Since larger molecules elute first from size exclusion column, the first peak in this separation indicates aggregation as expected.

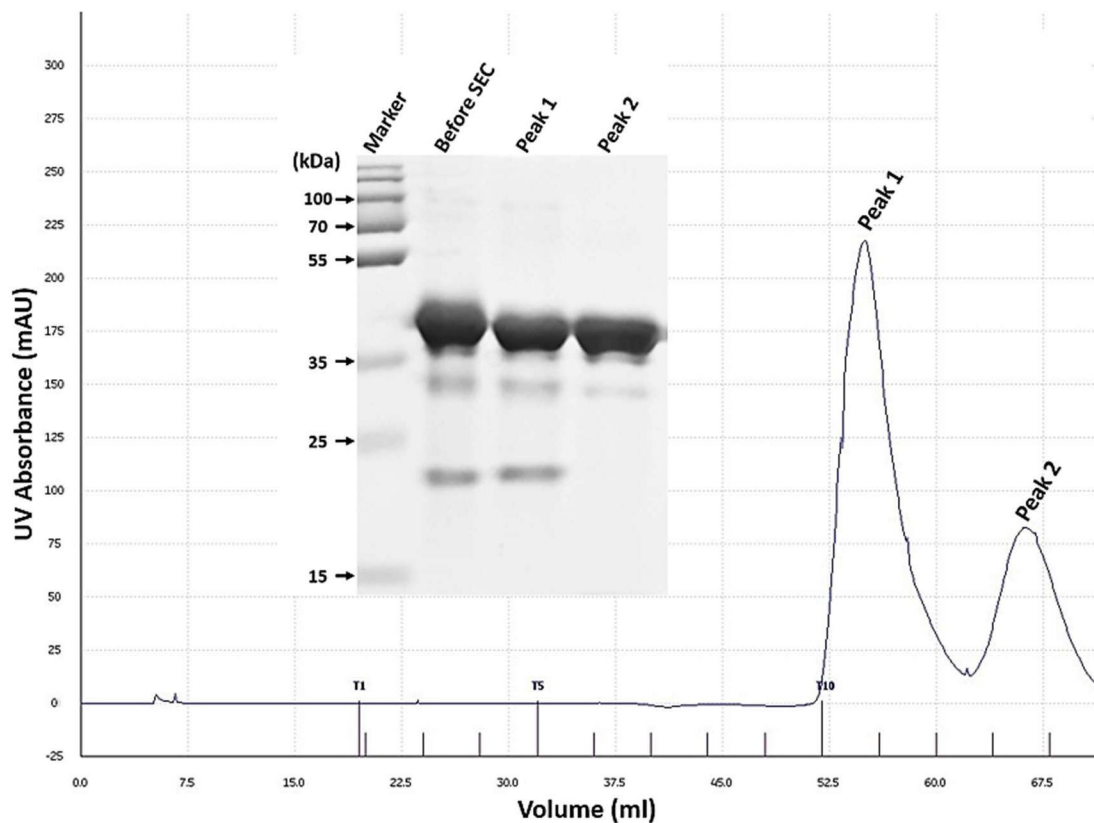


Figure 5.5. Size exclusion Separation of 23-mer fusion protein. The two peaks detected during the separation are shown in the chromatogram. Fractions correspond to the two peaks were detected on SDS-PAGE and compared with the solution before SEC.

The experiments above suggest that the proteins detected at 20 and 30 kDa are aggregated with the fusion protein, indicated by their presence in the first SEC peak, which results in a larger molecular weight than that of the pure fusion protein recovered from peak 2 of the SEC purification. The source of these impurities can either be endogenous proteins or a cleaved version of the fusion protein. However, obtaining a truncated version of the fusion protein with a molecular weight of 20 kDa entails a truncation, or cleavage, of the fusion protein within the sfGFP sequence, which is unlikely, given the reported stability of the sfGFP expression. It is impossible to determine with certainty the source and the identity of these proteins from the data above. Alternatively, general measures can be taken to mitigate the possibilities of aggregation,

truncated expression, or nonspecific proteolysis. These measures can be applied to the process at all steps: media preparation, expression induction, and protein extraction and recovery.

To mitigate the risk of misfolding, the expression was repeated at lower temperature, 16 °C and using terrific broth media (TB) instead of LB to support the growth at lower temperature. The concentration of IPTG used for expression induction was also reduced to half of the concentration used in the first trial. To mitigate the risk of aggregation during the recovery step, the cell lysis buffer was supplemented with 10 % (v/v) glycerol and more volume of the lysis buffer was used per cell pellet to dilute the protein extract and minimize the interaction with endogenous proteins. Both constructs, 23-mer and 27-mer fusion proteins were expressed and purified under these conditions. The new conditions achieved high purity for both fusion proteins (Figure 5.6). However, the yield was lower than that of the first trial, 20 mg of fusion protein from 250 ml culture compared to 60 mg obtained from the same volume of culture in the first trial.

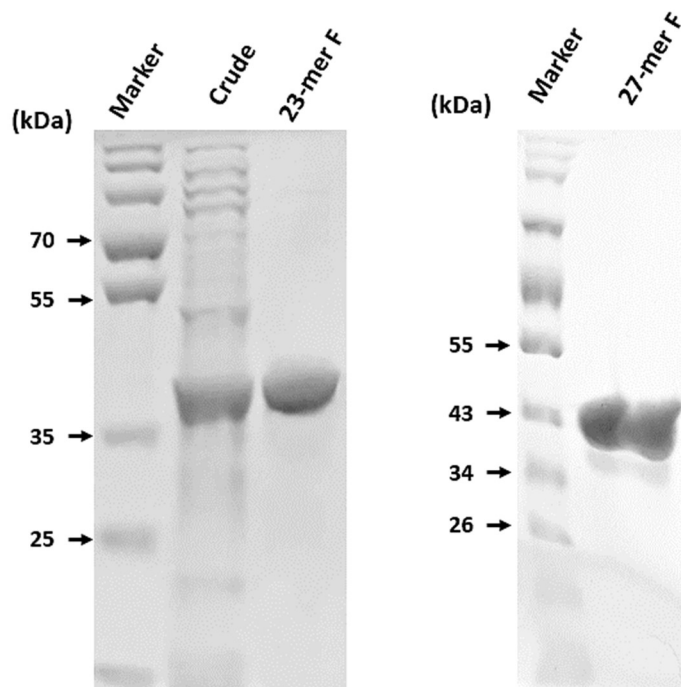


Figure 5.6. SDS-PAGE gel results of purified fusion proteins. Both 27-mer fusion (right) and 23-mer fusion (left) were expressed and purified using the optimized conditions. Purification was carried out using IMAC. Both fusion proteins have molecular weight around 34 kDa

### 5.2.3 Cleavage of the Tag

TEV protease was expressed following the protocol explained in the Appendix (A.2.9). The buffer of the purified 23-mer fusion protein was exchanged to TEV buffer (A.2.1) and a concentration ratio of 1:5 of the enzyme to the fusion protein was used to perform the cleavage. The reaction mixture was incubated overnight at room temperature and was then subjected to IMAC purification at the end of the incubation period. Both TEV protease and the carrier protein in the fusion construct are His-tagged. Therefore, only cleaved peptide is expected to flow through the column while both the carrier protein and the enzyme are captured on the column. Fractions contain the flow-through and all elution peaks were collected and visualized via SDS-PAGE (Figure 5.7). Gel results confirms the following (1) Elution peak 2 (E2) contains the His-tagged carrier with traces of TEV protease, (2) TEV protease comprises peak 3 (E3), (3) the size of the fusion protein is larger than the cleaved carrier indicating successful removal of the peptide. Unfortunately, the peptide itself cannot be detected with SDS-PAGE due to its small size.

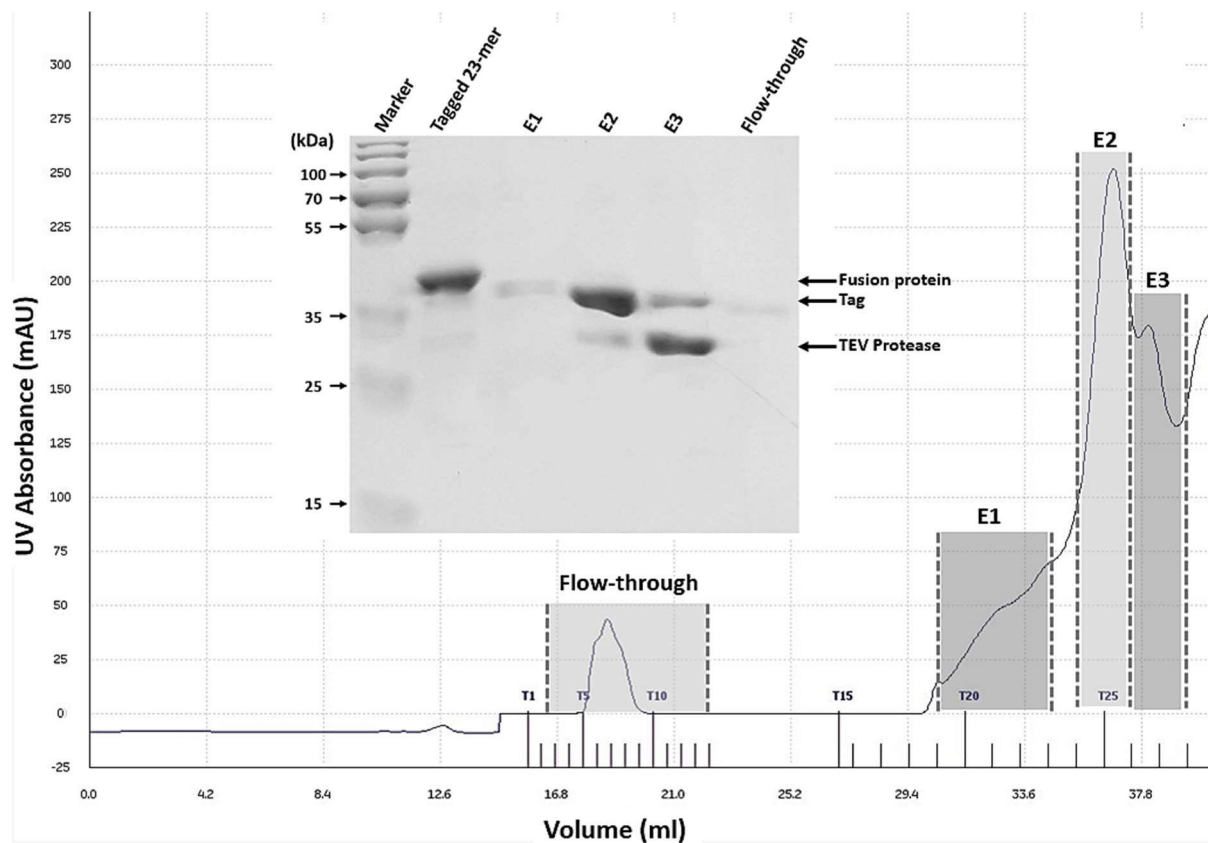


Figure 5.7. IMAC purification of the TEV protease cleavage reaction of the 23-mer fusion protein. The cleavage mixture of the 23-mer fusion protein by TEV protease was purified by IMAC and the fractions were detected on SDS-PAGE. The samples loaded on the gel are indicated on the chromatogram.

## **5.3 Validation of the Sequence Design**

### **5.3.1 Activity of GALNTs with the Fusion Proteins**

The sequence of the 23-mer peptide was proposed as an improved design to that of the 27-mer and TAP24 peptides (Chapter 4, section 4.7.1). To test the hypothesis, the glycosylation of the 23-mer and the 27-mer peptides with GALNT1 (T1), GALNT2 (T2), and GALNT4 (T4) were investigated. The progress of each reaction was monitored using NDK-PK-LDH coupled assay (Chapter 3, Section 3.5.2). Both peptides were used in their fusion form.

The 27-mer fusion protein was tested at concentration of 43  $\mu$ M with UDP-GalNAc as a donor at concentration of 2 mM. GALNT enzymes were added in combinations as follows: T1, T2, T4, T1+T2, T1+T4, T2+T4, T1+T2+T4. In all combinations all the enzymes were added at a final concentration of 220 nM. The assay was carried out using the general protocol (Appendix, A.4) with the enzymes being the initiation component (Figure 5.8). The same procedure was followed with 23-mer fusion protein (Figure 5.9)

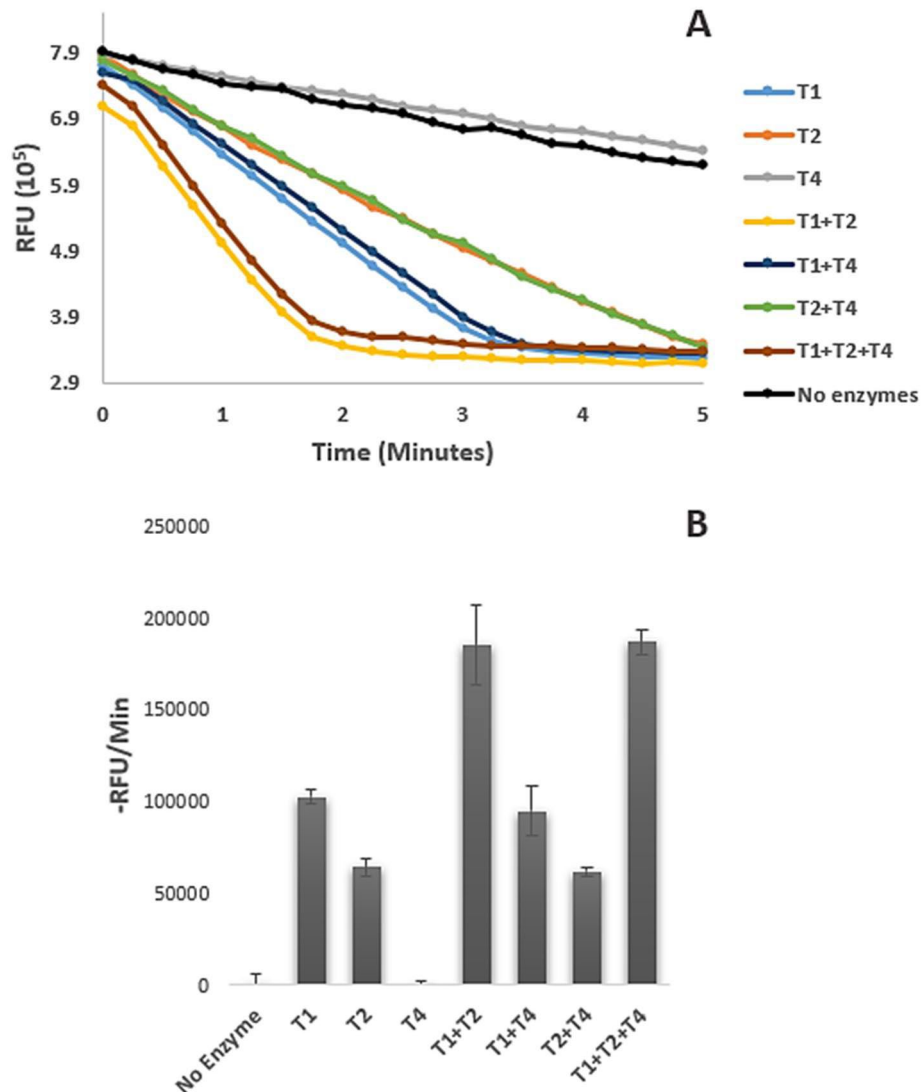


Figure 5.8. Activity of GALNTs with 27-mer fusion protein. Activities of GALNTs were tested with the 27-mer fusion protein at uniform concentrations. GALNT 1,2, and 4 are referred to as T1, T2, and T4, respectively. Reactions were monitored using NDK/PK/LDH assay. A. Progress curves are presented as the average of a triplicate for each reaction. B. The rates were calculated from the slope of each progress curve using the linear region of the curve. The mean rate for each reaction is presented with its standard deviation displayed as error bars.

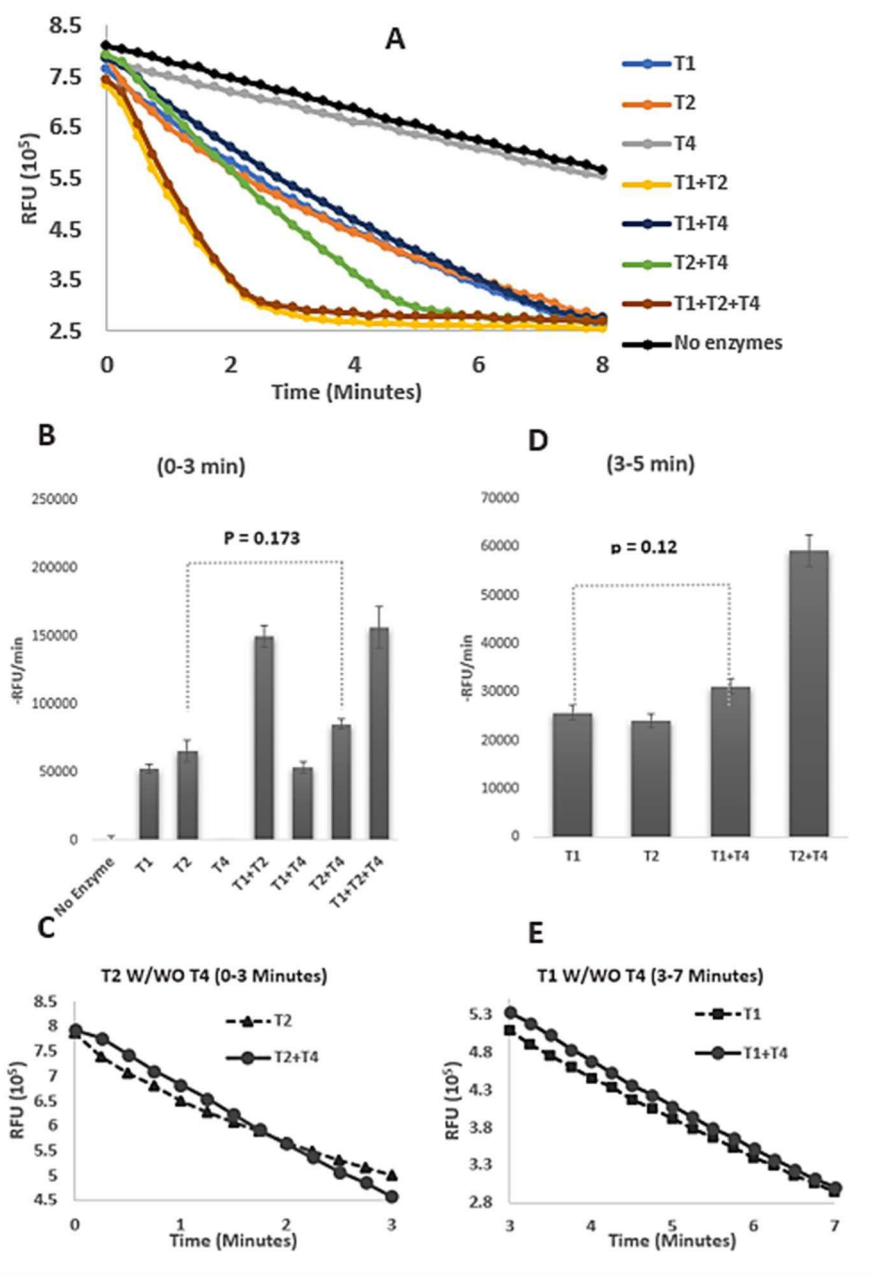


Figure 5.9. Activity of GALNTs with 23-mer fusion protein. Activities of GALNTs were tested with the 23-mer fusion protein at uniform concentrations. Reactions were monitored using NDK/PK/LDH assay. A. Progress curves are presented as the average of a triplicate for each reaction. B. The rates were calculated from the slope of each progress curve using the linear regions, between 0-3 minutes, B, and between 3-5 minutes, D. The mean rate for each reaction is presented with its standard deviation displayed as error bars. Curves in panels C and E are from panel A and are presented in isolation. Panel C shows progress curves of GALNT2 with and without GALNT4 between 0-3 minutes. Panel E shows the progress curves of GALNT1 with and without GALNT4 between 3-7 minutes. These two panels are correspondent to the T-test indicated in panels B and D.

The main observation from both experiments is that GALNT4 on its own does not react with either of the peptides. This observation agrees with previous studies discussed in Chapter 4. In the case of the 27-mer peptide, GALNT4 did not show any activity even when combined with GALNT1, GALNT2, or GALNT1 + GALNT2. That was also predicted because the 27-mer peptide lacks S19 and T20 residues on the N terminal direction of GALNT4 target sites, S9 and T13, (Figure 4.5). On the other hand, in case of the 23-mer peptide, GALNT4 showed significant activity when combined with GALNT2; this activity was more pronounced after 3 minutes (Figure 5.9, B and D). No significant GALNT4 activity was detected when combined with GALNT1. These results were also predicted when both peptides were designed; GALNT1 glycosylates T8 residue, whereas GALNT2 glycosylates T20 residue, which is used by GALNT4 via the lectin-like domain assisted glycosylation mechanism (the secondary mechanism) (Figure 4.6). Unfortunately, the reactions were too fast to detect the effect of GALNT1 and GALNT2 combined on GALNT4. However, GALNT1 and GALNT2 in combination are expected to further enhance GALNT4 activity by the glycosylated S19 (in addition to T20), which further enhances the secondary mechanism of GALNT4.

Another observation can be made is that GALNT1 and GALNT2 when added alone to the 23-mer peptide, their reactions slowed down after 3 minutes, whereas when reacted with the 27-mer peptide, the rate remained apparently constant throughout the course of the assay. This can be explained by the secondary mechanism in the 27-mer peptide (Figure 4.5) in which the 27-mer design enables both GALNT1 and GALNT2 of glycosylating two sites back-to-back, whereas the 23-mer gives access to only one site for each enzyme and the secondary mechanism is not possible with this design. This limited access to one site results in depletion of the sites, which manifests in the observed slowing rate.

Another interesting conclusion can be made when comparing the reaction rates of GALNT1 and GALNT2 in the 23-mer peptide case and in the 27-mer peptide case. As established above, GALNT1 and GALNT2 participate in the reaction with the 23-mer peptide via one mechanism only, which is the direct recognition via the catalytic domain (direct glycosylation). Whereas with the 27-mer, additional mechanism, the secondary mechanism, is also active (in addition to the direct glycosylation). The activities of GALNT1 and GALNT2 with the 23-mer peptide did not show statistically significant difference. However, in case of the 27mer peptide, GALNT1

was significantly more active than GALNT2. Since no difference in the activity can be attributed to the direct mechanism (from the 23-mer case), the observed difference in activities with the 27-mer case is attributed to the secondary mechanism. This suggests that the secondary mechanism in GALNT1 is significantly more efficient than that of GALNT2.

### **5.3.2 Activity with the Cleavage Products**

To assess the activity of GALNTs with the TEV protease cleavage products of the 23-mer fusion protein, the fractions obtained from the IMAC performed on the product of the previous cleavage reaction were used (section 5.2.3) and (Figure 5.7). Each reaction was prepared in triplicate. The reactions were carried out using a mixture of the three GALNTs with different fractions.

Concentrations of the donor and the enzymes were used as follows: 1 mM for UDP-GalNAc, 220 nM for each of GALNT1, GALNT2, and GALNT4. As for the acceptors, three acceptors were used: the intact fusion protein, the tag fraction (E2), and the flowthrough, which is expected to contain the peptide. Since there was no accessible method to quantify the peptide in the flowthrough, assumption of complete cleavage and full recovery of the peptide was used to calculate the concentration of the peptide in the flowthrough. This concentration was estimated to be around 72  $\mu$ M after combining all the flow-through volume in one solution. The estimated concentration of 72  $\mu$ M was used as 10X of the concentration in the final reaction, and the same concentration was prepared for both the intact fusion protein and the cleaved tag.

Progress curves of the reaction of GALNTs mixture with the different acceptors confirm that the enzymes did not react with the tag. Activity was detected with the flowthrough as an acceptor confirming the presence of the peptide. Activities (indicated from the slope of the progress curve) with the intact fusion protein and with the flow-through fraction were approximately equal (Figure 5.10). This indicates the validity of the assumption of complete cleavage and full recovery of the peptide. This experiment confirms that the tag, including carrier protein and linkers, is non-reactive with (inert to) GALNTs tested in this experiment. Since GalNAc conjugates are the precursors of the downstream O-glycosylation pathway, this inertness of the tag extends to the entire pathway. Furthermore, from the activity tests performed above (Figure 5.8 and Figure 5.9), it was concluded that fusion proteins for both the 23-mer and the 27-mer produced the same reactivity patterns expected from the native naked peptides (without the fusion partners) as predicted from literature and intended by design. These results provided

evidence of the lack of interference of the tag with the peptide. The lack of interference and the inertness confirmed in this section collectively validate the relevance of the fusion protein as a substrate for specificity assessment of GALNTs.

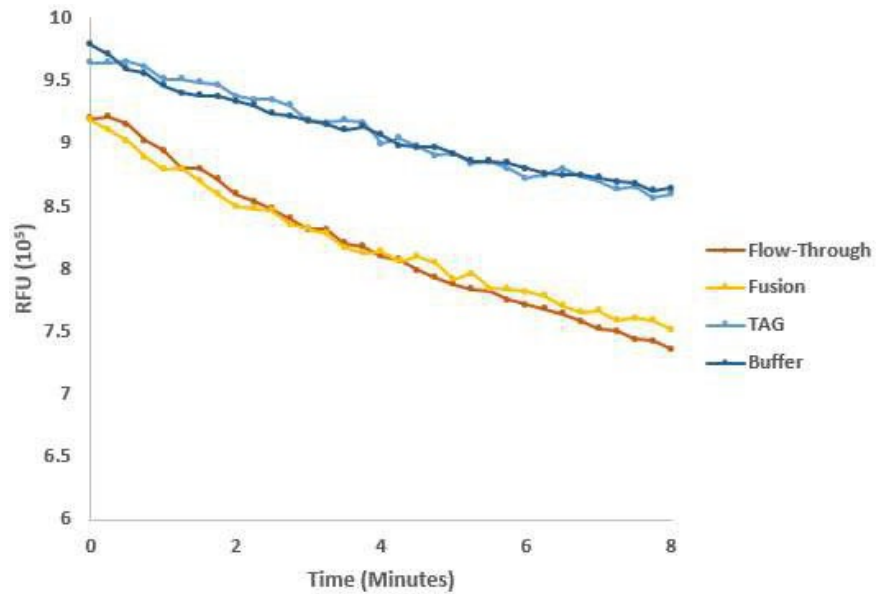


Figure 5.10. Activity of GALNT1, 2, and 4 combined with the cleavage products of fusion protein. The activity of GALNTs was tested with the intact fusion protein, the tag, and the flow-through fraction.

## 5.4 Biological Relevance of Synthesized Structures

The activity test above, in agreement with the literature, showed that a combination of GALNT1 and GALNT2 achieves full and fast glycosylation of S19, T20, and T8 residues. These two enzymes are ubiquitous, and they are expressed in almost all tissues<sup>2</sup>. Glycosylation of S9 and T13 however is specific to GALNT4. It has been reported that when MUC-N in breast milk is compared to that from breast cancer cells, sites on cancer cells' MUC1 were almost fully glycosylated. On the other hand, analysis of breast milk MUC1-N using mass spectrometer and Edman sequencing revealed that around 50 % of the sites were glycosylated, with preference for S19, T20, and T8 residues i.e., GALNT1 and GALNT2 sites<sup>2</sup>. It was demonstrated from the GALNTs review in Chapter 4 that any of GALNT1, GALNT2 or GALNT3 on its own can glycosylate S19, T20 and T8 completely in MUC1-N (except in one of the terminal tandem repeats) i.e., there is abundance of catalytic machinery to glycosylate these sites. However, S9 and T13 are only glycosylated by GALNT4 (GALNTs review in Chapter 4), which requires glycosylation of the other sites first. Therefore, glycosylation of S9 and T13 is expected to be slower even if all enzymes present together, which was confirmed from the activity test performed above (Figure 5.9). The activity of GALNT4 requires binding of its lectin-like binding domain to GalNAc-S and/or GalNAc-T structures at S19 and T20, respectively. After glycosylation of S19 and T20 by GALNT1,2 or 3, competition on these sites takes place between GALNT4 and core 1 synthase, C1GALT1. Formation of core 1 structures on S19 and T20 impairs GALNT4 binding to these sites and therefore S9 and T13 remain unglycosylated. This explains the half saturation of glycosylation sites in normal breast tissue when core 1 synthesis is fast<sup>2</sup>. In some cancers however, cosmc, the chaperone protein required for the expression of C1GALT1 in its active form, is mutated, which causes impaired activity of core 1 synthesis<sup>3</sup>.

This impaired activity of core 1 synthesis in breast cancer preserves GalNAc-S and GalNAc-T structures on S19 and T20 sites allowing GALNT4 to saturate the rest of MUC1 sites. In fact, one of the early antibodies, VU-2-G7, directed to the glycosylated T13, showed exceptional binding with breast cancer tissue<sup>4</sup>. Recent evidence suggested that the loss of core 1 synthesis efficiency is not the only explanation for the over expression of Tn in cancer. GALNTs in cancer are trafficked from Golgi to the endoplasmic reticulum (ER). This translocation enhances Tn formation due to eliminating the competition on the glycosylation sites with C1GALT1 and

allowing more time for GALNTs to saturate the sites on the target peptide<sup>5,6</sup>.

Based on the discussion above, differentiation in glycosylation pattern between normal MUC1 and cancer associated MUC1 needs to be represented in any model aims to study O-glycosylation pathway. Therefore, two substrates representing Tn antigen were synthesized: (GalNAc)<sub>3</sub>23-mer and (GalNAc)<sub>5</sub>23-mer glycopeptides. The former represents the underglycosylated MUC1 observed in normal tissue and the latter represents cancer associated MUC1. In addition to their cancer vs normal representation, these two substrates represent a chronological order in the biosynthesis pathway and the dominance of each of them over the other determines the fate of the downstream structures (Figure 5.11).

## 5.5 Preparative Tn Synthesis

Expression of the 23-mer fusion was repeated under the optimized conditions discussed above at a scale of one litre of culture. The fusion protein was purified using IMAC and quantified using Bradford assay. The procedure yielded 96 mg of fusion protein, of which 6 mg was the peptide. The fusion protein was glycosylated according to the diagram in Figure 5.11. First glycosylation was via GALNT1 and GALNT2 combined. The glycosylation product was then purified by IMAC, validated and a portion from it was glycosylated with GALNT4.

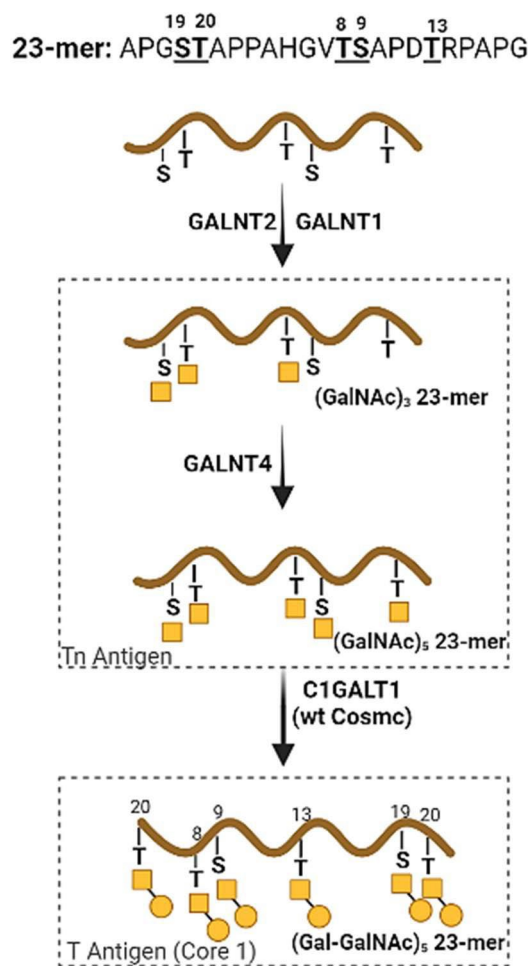


Figure 5.11. Synthesis of T and Tn antigens. Two forms of Tn antigens were synthesized: the underglycosylated form, (GalNAc)<sub>3</sub> 23-mer F, and the fully glycosylated form, (GalNAc)<sub>5</sub> 23-mer F. A combination of GALNT1 and GALNT2 were used to produce the former, and GALNT1,2, and 4 were used to produce the latter. T antigen was produced by glycosylating (GalNAc)<sub>5</sub> 23-mer F by C1GALT1 (core 1 synthase)

### 5.5.1 Glycosylation Reaction Conditions

Glycosylation of the fusion protein with GALNT1/2 and with GALNT4 underwent an optimization process with the aim to maximize the yield and minimize the loss by aggregation of the proteins during incubation. All reported protocols were performed on chemically synthesized peptides<sup>7-10</sup>. However, some insights were concluded from comparing all these methods, such as, refreshing the reaction by addition of enzymes and donor halfway into the incubation, adding non-denaturing detergents, such as triton-X 100, and incubation at an optimum temperature of 37°C. There are two aspects to the optimization process, maximizing the rate to drive the reaction to completion, and stabilizing the proteins (which are the enzymes and the fusion proteins) to endure the incubation period at high concentrations of the reactants. Increasing the rate can be theoretically achieved by increasing the concentrations of the substrates, above their  $K_m$  values. However, there are limitations to that in this synthesis. One limitation is the cost of UDP-GalNAc, which prohibits using it lavishly. The other limit is that the peptide is in its fusion protein form, which limits the concentration that can be tolerated before aggregation.

Fortunately, UDP-GalNAc has  $K_m$  values for GALNTs in a range below 100  $\mu\text{M}$ , as reported in the Brenda database. As for the fusion protein, a stock solution of 717  $\mu\text{M}$  was shown to be stable during freeze-thaw cycles. Therefore, this was used as a working solution.

A few iterations of optimization led to the following general protocol, which is optimized for solubility and complete glycosylation: the fusion protein concentration is greater than 400  $\mu\text{M}$ , and the donor is supplied in two batches. The first batch is supplied at a 2:1 ratio of the concentration of potential glycosylation sites, and the second batch is supplied after 12 hours at a ratio of 1:1 of the concentration of potential glycosylation sites. For example, if the final concentration of the fusion protein is 400  $\mu\text{M}$ , and three sites are potentially to be glycosylated, the first batch of UDP-GalNAc is at 2400  $\mu\text{M}$ , and the second batch is at 1200  $\mu\text{M}$ . Enzymes are also added in two batches, first batch is at 1  $\mu\text{M}$ , and half of this after 12 hours. The buffer is 25 mM Tris-HCl contains: 50 mM NaCl, 10 % (v/v) glycerol, 0.25% (v/v) TritonX-100, and 10 mM  $\text{MnCl}_2$ , pH 7.4.

### 5.5.2 Purification, Desalting and Storage

Glycosylation by GALNT1 and GALNT2 was performed using the optimized conditions on the fusion protein. The product was purified and a portion of it was used to perform the glycosylation with GALNT4. A single step IMAC purification followed by desalting was used to purify the products after each reaction. The protocols used in this process are the general protocols (A.2.6 and A.2.7).

For GALNT1 and GALNT2 reactions, 80 % (24 mg from 30 mg) of the initial amount of the fusion protein was obtained after glycosylation. A similar value was obtained from the GALNT4 reaction (13 mg from 16.5 mg). The chromatogram of (GalNAc)<sub>3</sub>23-mer fusion protein synthesis and the SDS-PAGE gel results for both Tn antigen substrates are shown in Figure 5.12. SDS-PAGE gel could not differentiate between (GalNAc)<sub>3</sub>23-mer and (GalNAc)<sub>5</sub>23-mer fusion proteins, whereas both products showed a slightly higher band compared to the naked peptide fusion protein. The purity of both products was also confirmed from the gel results.

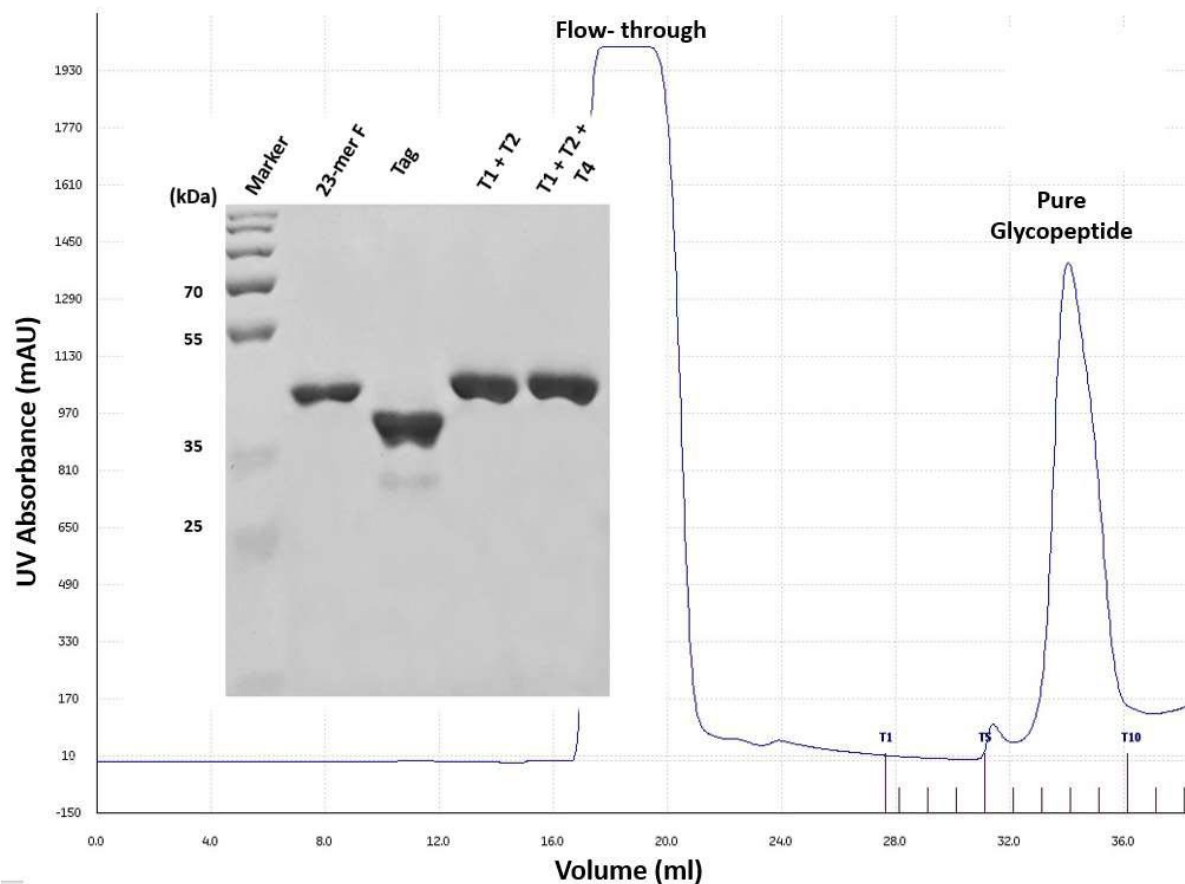


Figure 5.12. Purification of Tn antigen products obtained from the preparative synthesis. The product of the preparative glycosylation of the fusion protein with GALNT1/GALNT2 was purified with IMAC (the background chromatogram). Elution fractions were visualized by SDS-PAGE. SDS-PAGE gel results also include the product of GALT4 preparative synthesis.

### 5.5.3 Confirmation of reaction completion

To confirm the completion of GALNTs reactions, the activity of the enzymes was tested with their corresponding products from the preparative synthesis using NDK/PK/LDH coupled assay (Figure 5.13). The results show that GALNT1 and GALNT2, combined, did not react with their product, (GalNAc)<sub>3</sub> 23-mer fusion, whereas they reacted with the naked peptide fusion protein (Figure 5.13, A). This observation suggests that the product of the preparative glycosylation with GALNT1/2 lacks glycosylation sites to incorporate more GalNAc, hence, the synthesis was complete.

The GALNT4 reaction was tested with GALNT1/2 product, (GalNAc)<sub>3</sub> 23-mer fusion, and compared to GALNT4 product, (GalNAc)<sub>5</sub> 23-mer fusion. The reaction was too slow to be detected in real time. Therefore, an endpoint assay was performed to measure the UDP released after 4 hours of incubation (Figure 5.13, B). The fluorescence signal from each reaction was normalized to the signal from a control mixture containing all the components of the reaction but lacking UDP-GalNAc. Cleaved tag was used as an acceptor in the control reaction. Normalized signals were converted to concentration of UDP by using a standard curve similar to that in (Figure 3.12, D). The results showed no significant difference in the UDP detected between the reaction mixtures containing the naked peptide fusion protein and the mixture containing (GalNAc)<sub>5</sub> 23-mer fusion protein. In contrast, a twofold UDP increase was detected in the reaction mixture containing (GalNAc)<sub>3</sub> 23-mer fusion as an acceptor. A small amount of UDP was released independently from the reaction, as compared with a mixture that does not contain UDP-GalNAc (No-donor control, not shown in the results). This can be explained by nonspecific hydrolysis of UDP-GalNAc, or UDP present in the UDP-GalNAc stock as impurity. However, this amount was only around 0.5 % of the UDP-GalNAc added to the reactions. Absence of additional UDP in (GalNAc)<sub>5</sub> 23-mer F reaction, suggest lack of unglycosylated S9 and T13, hence, the reaction was complete.

If the UDP from hydrolysis/impurity is accounted for, the UDP released from the reaction with (GalNAc)<sub>3</sub> 23-mer F can be estimated at 10  $\mu$ M of UDP after 4 hours (Figure 5.13, B). Given that the substrate has two potential glycosylation sites, S9 and T13, and the starting concentration of the substrate is 10  $\mu$ M, a rough estimation leads to the conclusion that half of the sites were glycosylated in 4 hours under the test conditions. Furthermore, comparing the test conditions

with the preparative synthesis that provides more saturation with all the substrates and was incubated for 24 hours, offers further evidence of sufficient conditions for glycosylation completion in the preparative synthesis.

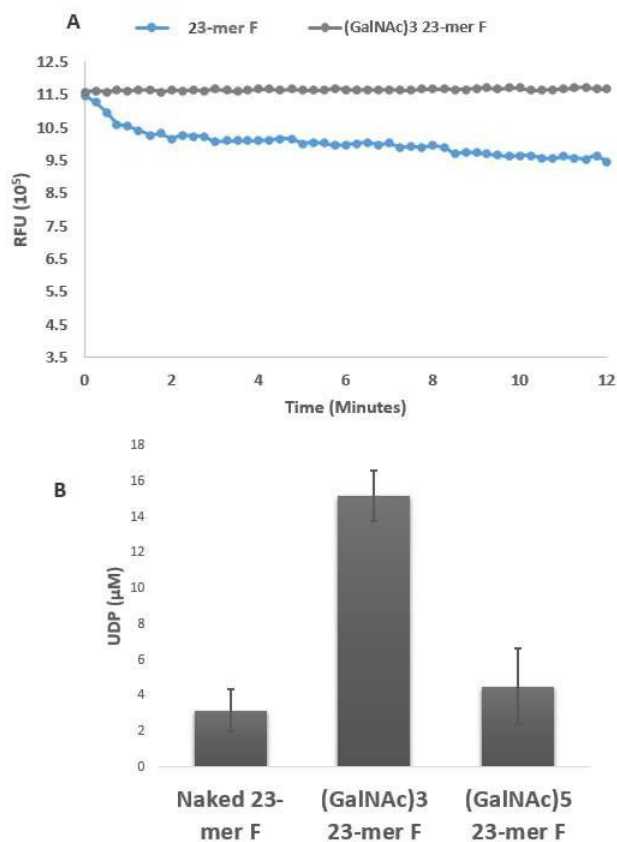


Figure 5.13. Confirmation of GALNTs reactions completion. Each reaction mixture contains 10  $\mu$ M acceptor substrate, 1 mM UDP-GalNAc, 640 nM enzymes. A. Real-time activity test carried out using NDK/PK/LDH assay. The activity of a combination of GALNT1 and GALNT2 was tested with each one of the acceptors: Tag, 23-mer fusion protein, and the product of GALNAC1/2 preparative synthesis, (GalNAc)<sub>3</sub>23-mer F. The reactions were prepared in triplicate and the signal is normalized to the tag reaction and averaged. B. Endpoint assay using NDK/PK/LDH to detect the UDP released from the reactions of GALNT4 with the substrates shown in the results and a No-donor control with the naked 23-mer F. The reaction mixtures were prepared in the assay buffer (1X) and after 4 hours, a mixture of PEP, ATP, NADH, and assay enzymes in 1 X reaction buffer were added to final concentration of 1 X from each component (Appendix, A.4). The signals from a triplicate of each reaction were normalized to the control and averaged. UDP concentration was calculated from the fluorescence using a standard curve similar to that obtained in Figure 3.12, D.

## 5.6 Preparative Core 1 (T antigen) Synthesis

The expression of C1GALT1 (core 1 synthase) in its active form requires the coexpression of the molecular chaperone C1GALT1C1, also known as cosmc. Cosmc functions as a chaperone protein that enhances the proper folding and solubility of C1GALT1<sup>11</sup>. In the absence of cosmc or when it is mutated, C1GALT1 aggregates and is transported to the cytoplasm, where it undergoes ubiquitination and degradation in the proteasome<sup>11,12</sup>. C1GALT1 and its chaperone play a central role in cancers, and the associations of their product (core 1) with different cancer subtypes and outcomes are well-established<sup>13,14</sup>. Furthermore, core 1 is a central glycan structure that serves as the precursor for all other downstream structures (Figure 4.1). Therefore, it is reasonable to include Core 1 in every *in vitro* modeling of the O-glycosylation pathway.

### 5.6.1 Core 1 synthase expression

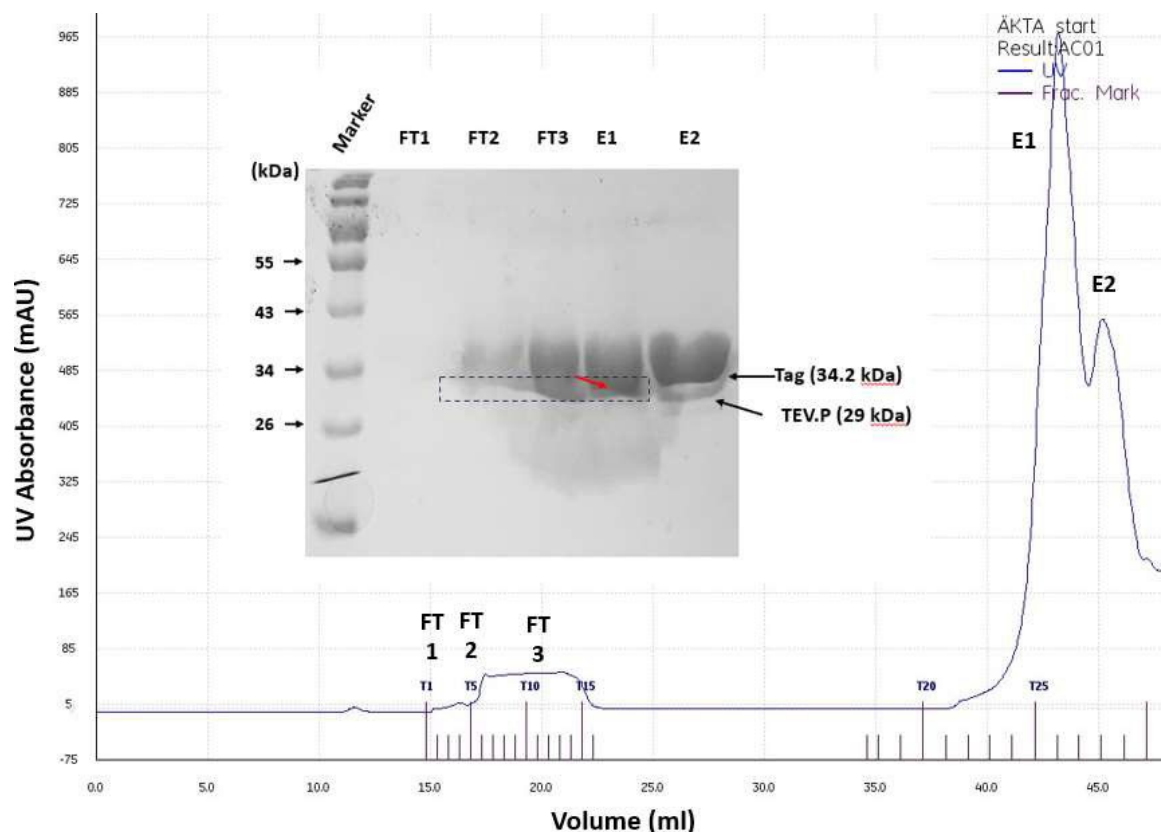
All GT enzymes used in this thesis were obtained from the vector library constructed by Moremen *et al*<sup>14</sup>. The pGen2 plasmids containing the enzymes, including C1GALT1 and C1GALT1C1, feature a signaling sequence (TCM) that directs the fusion protein for secretion. These fusion proteins include affinity tags and detection markers, such as Avi-tag, His-tag, and sfGFP, which were separated from the catalytic domain by a TEV protease cleavage site. In the Moremen paper, the expression and secretion of C1GALT1 from the pGen2 construct were reported to be very low. The low secretion of C1GALT is suggested to be due to insufficient endogenous chaperone (C1GALT1C1) levels required for proper folding. As a result, the absence or suboptimal levels of the chaperone may lead to misfolding, potentially causing protein aggregation and precipitation within the cells. To solve this problem, cells in a previous study<sup>11</sup> were cotransfected with C1GALT1 and its chaperone using pcDNA3.1 plasmid. To follow a similar strategy, pcDNA3.1 plasmid containing the full sequence of cosmc protein was designed and used in combination with pGen2 containing C1GALT1 to cotransfect HEK293 cells. However, the secretion of C1GALT1 did not improve and remained undetectable. It is worth noting that the Moremen library is optimized for preparative scale expression, whereas pcDNA3.1 expression is low and intended for physiological studies, which may explain the failure of pcDNA3.1 to match the expression level of pGen2.

Moremen's library also includes a C1GALTC1 construct in the pGen2 vector. However, a concern with this design is that the chaperone fusion protein, designed with a TCM signaling peptide for secretion, and the enzymes produced from this vector lack their transmembrane domains to enhance their solubility. The secretion signaling and lack of transmembrane domain may interfere with the function of the chaperone when it is expressed and secreted instead of being properly localized in the endoplasmic reticulum (ER) to assist with C1GALT1 folding. Despite these challenges, transfection was attempted.

First, C1GALT1C1 was expressed alone to confirm the expression level and identify the molecular weight of the protein on SDS-PAGE gel (Figure 5.14). Expression, cleavage, and purification were carried out according to the protocol in the Appendix (A.3). Purification of the cleaved protein resulted in two elution peaks. The SDS-PAGE gel revealed that the second peak consists of the cleaved Tag and TEV protease, which is consistent with the previous results from other enzymes expressed from the same vector. However, the first peak showed a band with a molecular weight overlapping with both the TEV protease and the tag. The theoretical molecular weights of C1GALT1C1, the Tag, and TEV protease are 33.4, 34.2, and 29 kDa, respectively. Therefore, it was difficult to identify the bands based on the molecular weight. Hypothetically, after cleaving the Tag, the protein is expected to appear in the flow-through, which surprisingly was not the case in this experiment. Investigating the sequence of the expressed protein revealed a histidine-rich region at its N terminus (Figure 5.14). This His-rich region was believed to be the reason for the unexpected high affinity to the column and the appearance of the protein in the elution fraction.

The coexpression was carried out with C1GALT1 and C1GALT1C1 by transfecting the cells with both vectors in one step. Transfection with C1GALT1 alone was carried out in parallel as a control experiment. Cells were incubated for 6 days and examined under a fluorescence microscope. Both experiments showed high expression of GFP-tagged proteins (Figure 5.15). The media from each experiment was purified using IMAC. No secretion was detected in the media from the control experiment (Figure 5.16, A), whereas a high elution peak was detected from the coexpression experiment indicating large quantity of secreted proteins (Figure 5.16.B). The lack of secretion from the C1GALT1 control experiment suggests that the fusion protein was not secreted to the media, despite the high level of sfGFP detected in the cells indicating a proper

folding of sfGFP. However, it has been reported that sfGFP is capable of proper folding even if its fusion partner is misfolded. Therefore, no conclusion on the folding state of the enzyme can be made from that of the sfGFP. TEV protease cleavage was carried out on the protein purified from the coexpression experiment and a second IMAC was performed on the cleavage product (Figure 5.16, C). SDS-PAGE results showed a large band in the elution fraction with a molecular weight similar to what was detected in the C1GALT1C1 experiment. A combination of the cleaved tags and C1GALT1C1 is a reasonable explanation for this band. However, from the flow-through fractions, two bands were detected: a smaller one corresponding to the molecular weight of C1GALT1C1 and a larger one with a molecular weight around 40 kDa. The theoretical molecular weight of C1GALT1 is around 36 kDa, which is 3 kDa larger than C1GALT1C1, suggesting that the second band was C1GALT1. The protein was quantified using Bradford assay and the percentage of the 40 kDa band was calculated (using ImageJ software) to be around 48 % from the total protein in this preparation.



GRIGHGNRMHHHEHHHLQAPNKEDILKISEDERMELSKSFRVYCIILVKPKDVSLWAAVKETWTKHCDKA  
EFFSSENVKVFESINMDTNDMWLMMRKAYKYAFDKYRDQYNWFFLARPTTFAIENLKYFLKKDPSQPF  
YLGHTIKSGDLEYVGMEGGIVLSVESMKRLNSLLNIPEKCPEQGGMIWKISEDKQLAVCLKYAGVFAENAE  
DADGKDVFNKTSVGLSIKEAMTYHPNQVVEGCCSDMAVTFNGLTPNQMHVMMYGVYRLRAFGHIFND  
ALVFLPPNGSDND

Figure 5.14. Purification of C1GALT1C1 after the Tag cleavage. C1GALT1C1 was expressed in HEK293 cells according to the general protocol (Appendix, A.3). The enzyme was purified and cleaved with TEV protease. After cleavage, IMAC was carried out on the reaction mixture. The background chromatogram shows the Flow-through fractions designated FT1, FT2, FT3 that were collected separately. Two elution fraction E1 and E2 were also collected separately. SDS-PAGE results show bands consistent with the tag and TEV protease bands detected from other experiments in the E2 peak. FT2, FT3, and E1 all show the same band (indicated in the red arrow) with a molecular weight indistinguishable from those of TEV and the tag. The sequence of C1GALT1C1 is shown below the chromatogram with the His-tag-like sequence highlighted in red.

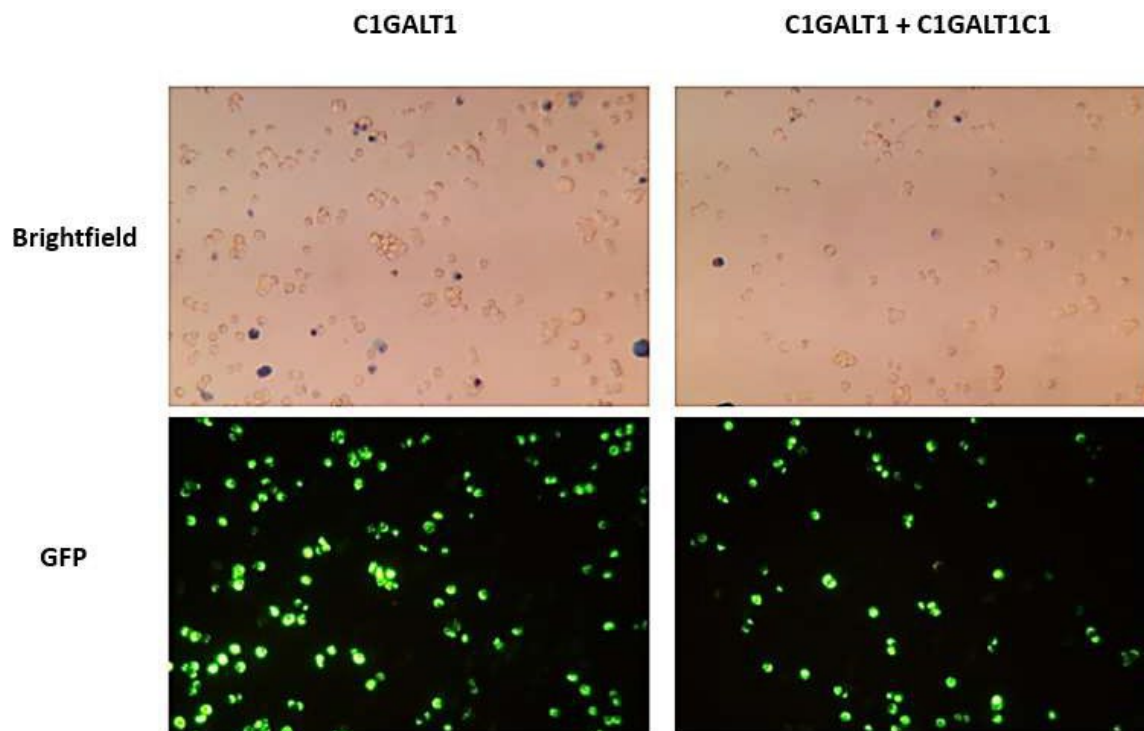


Figure 5.15. Expression of sfGFP-tagged C1GALT1 and C1GALT1 in combination with C1GALT1C1. HEK293 were transfected with C1GALT1 expression vector solely, or in combination with C1GALT1C1 expression vector. Images using fluorescence microscope were taken in two modes, brightfield and fluorescence after 6 days intubation. Cells were diluted with trypan blue dye to detect dead cells. Blue cells visualized in brightfield mode represent dead cells. Green fluorescence represents sfGFP accumulated in the cells.

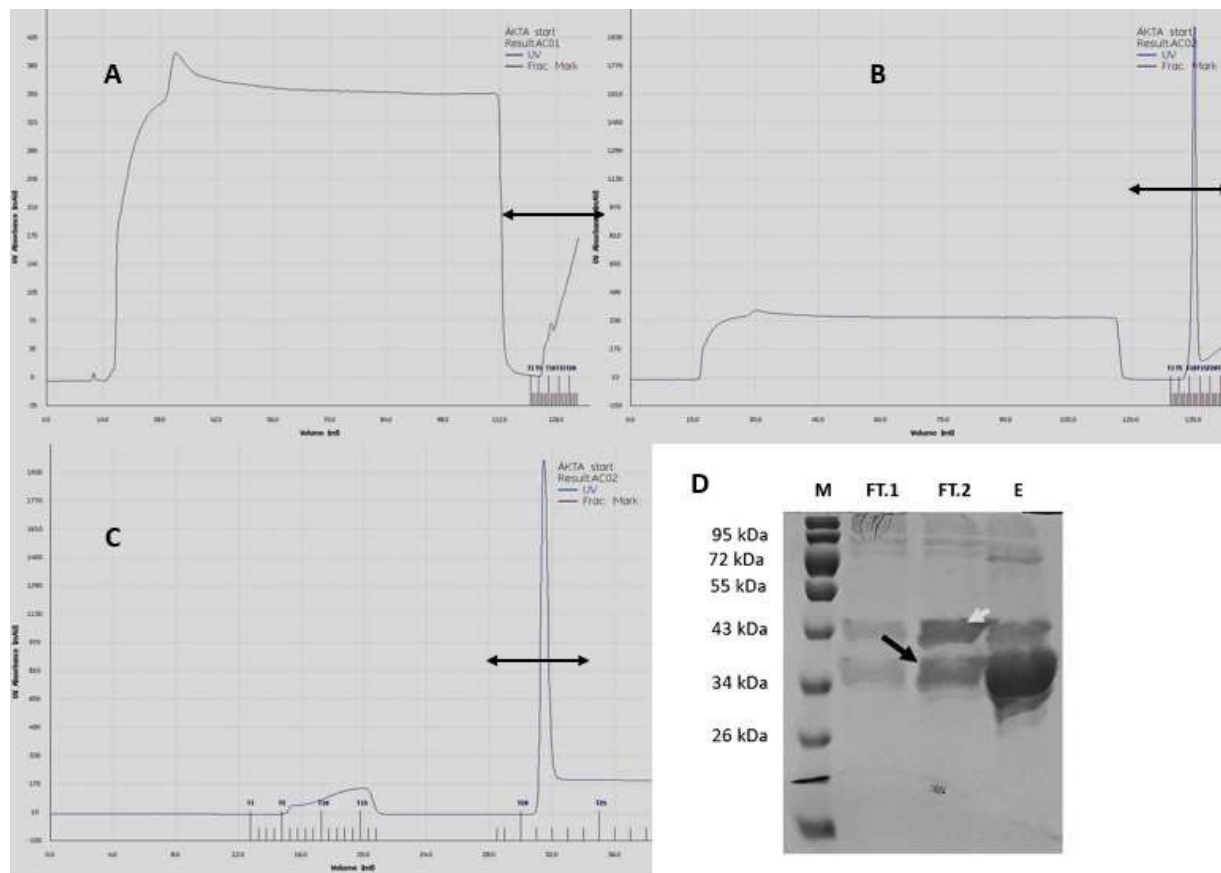


Figure 5.16. Purification of C1GALT1 and C1GALT1C1 coexpression products. Panels A and B are the chromatograms of media purification by IMAC from C1GALT1 and C1GALT1 in combination with C1GALT1C1, respectively. Panel C is the chromatogram of the IMAC performed on TEV protease cleavage of the coexpression experiment. Arrows on the chromatogram indicate the elution phase. Panel D is the SDS-PAGE performed on the fractions collected from the TEV protease reaction. The white arrow indicates C1GALT1, and the black arrow indicates C1GALT1C1

### 5.6.2 Testing activity of C1GALT1

The activity of the protein purified from the coexpression experiment was tested with different substrates: naked 23-mer F, (GalNAc)<sub>3</sub> 23-mer F, and (GalNAc)<sub>5</sub> 23-mer F. The conditions of the test and the concentrations of the reaction components are explained in Figure 5.17. The enzyme showed specific activity with both glycopeptides and the reaction was apparently complete after 5 minutes for both. The amount of UDP calculated after reaction completion was around 30 and 55  $\mu$ M for (GalNAc)<sub>3</sub> 23-mer F and (GalNAc)<sub>5</sub> 23-mer F, respectively. Since the substrates were added at 15  $\mu$ M, the concentration of the available sites can be calculated as 45 and 75  $\mu$ M for (GalNAc)<sub>3</sub> 23-mer F and (GalNAc)<sub>5</sub> 23-mer F, respectively. Comparing that to the concentration of UDP suggests that the reaction was not complete at 5 minutes (75% completion), and further incubation is needed for a complete reaction. However, this result confirms that the protein detected from the coexpression was in fact C1GALT1. The activity of the enzyme was also confirmed, and it was proportional to the acceptor's glycosylation sites. Additionally, this result presents further confirmation of the differential GalNAc saturation on the two substrates used to test the activity, hence, it provides further evidence of identities and structures of (GalNAc)<sub>3</sub> 23-mer F and (GalNAc)<sub>5</sub> 23-mer F.

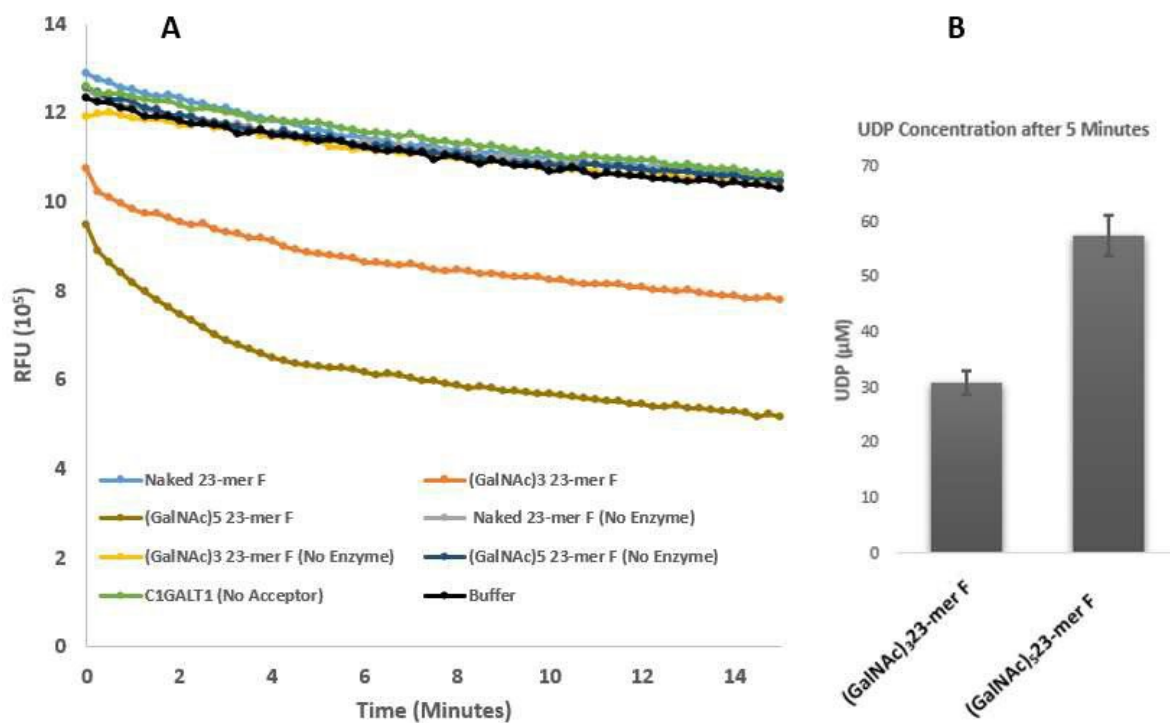


Figure 5.17. Activity of C1GALT1. The activity of C1GALT1 was tested with different acceptors. Enzyme was used at a concentration of  $0.5 \mu\text{M}$ , all acceptor substrates were added at a concentration of  $15 \mu\text{M}$ , UDP-Gal at a concentration of  $800 \mu\text{M}$ . The activity was measured using NDK/PK/LDH assay, acceptor substrates were used to initiate the reactions. A. Progress curves of the reactions. B. Concentration of UDP released in the reaction mixture for the two substrates after 5 minutes. The concentration was calculated from converting the UV signal drop from the control using UDP standard curve.

### 5.6.3 Preparative Core 1 synthesis

The preparative synthesis of core 1 was carried out for 8 hours using the same buffer and conditions optimized for GALTs preparative reactions. Final reaction mixture contains: 160  $\mu$ M fusion protein, around 1.6 mM UDP-Gal, 1.2  $\mu$ M enzyme, 50 mM HEPES, 10 mM  $\text{MnCl}_2$ , 10 % glycerol, and 0.2% Triton X-100, pH 7.4. After incubation, the reaction mixture was purified by IMAC and SDS-PAGE was performed on the products (Figure 5.18, B). The SDS- PAGE gel shows a band slightly higher for the core 1 product compared to the  $(\text{GalNAc})_5$  23- mer suggesting an increase of the molecular weight because of the glycosylation. The activity test was performed on the purified product as well as the  $(\text{GalNAc})_5$  23-mer, using the same conditions and concentrations used in section 5.6.2. No reactivity was detected from the purified product within the first 5 minutes of the reaction, whereas the  $(\text{GalNAc})_5$  23-mer F showed a fast reaction phase similar to what was detected above. Expanding the incubation time to 50 minutes (Figure 5.18, A) revealed a slow phase of glycosylation after 5 minutes for the  $(\text{GalNAc})_5$  23-mer, as indicated by the difference in slope between the reaction containing  $(\text{GalNAc})_5$  23-mer as an acceptor and the control. The slope of the product's , designated as  $(\text{Gal-GalNAc})_5$  23-mer, reaction remained identical to that of the control during the entire 50 minutes. To confirm the presence of core 1 structure, ST3GAL1, the sialyltransferase specific to core1 was tested with the  $(\text{GalNAc})_5$ 23-mer fusion protein and with the product of C1GALT1 preparative reaction (Figure 5.19). The results showed reactivity was only detected with C1GALT1 product. These results suggest a complete glycosylation by C1GALT1 and confirm core 1 structure.

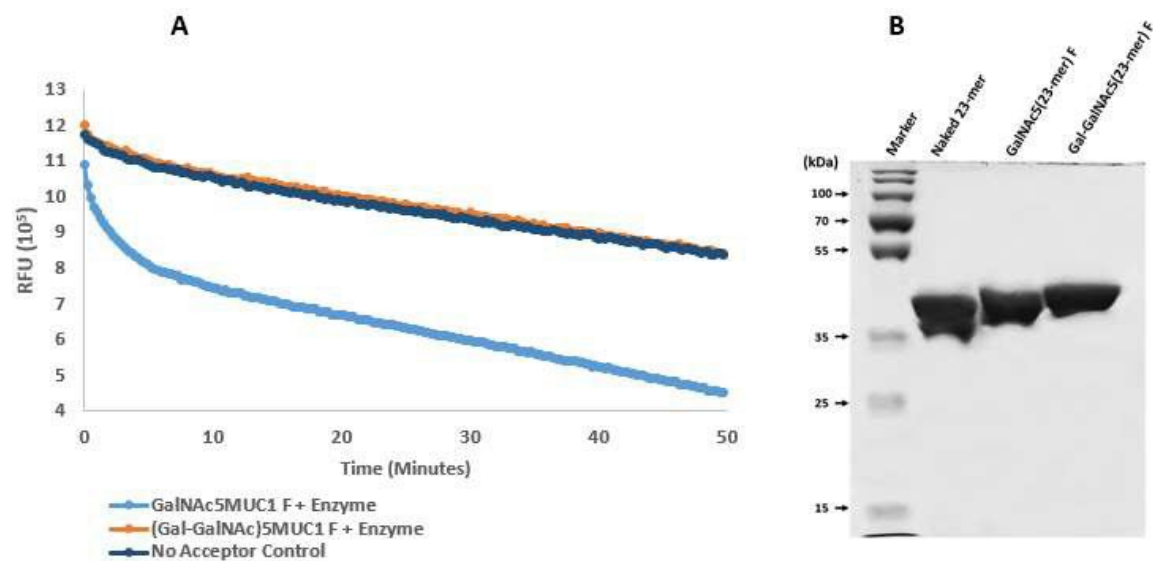


Figure 5.18. Product of Core 1 preparative synthesis. A. activity of the product compared to Tn antigen and a control contains the tag as an acceptor. B. SDS-PAGE of performed on Tn and T antigens.

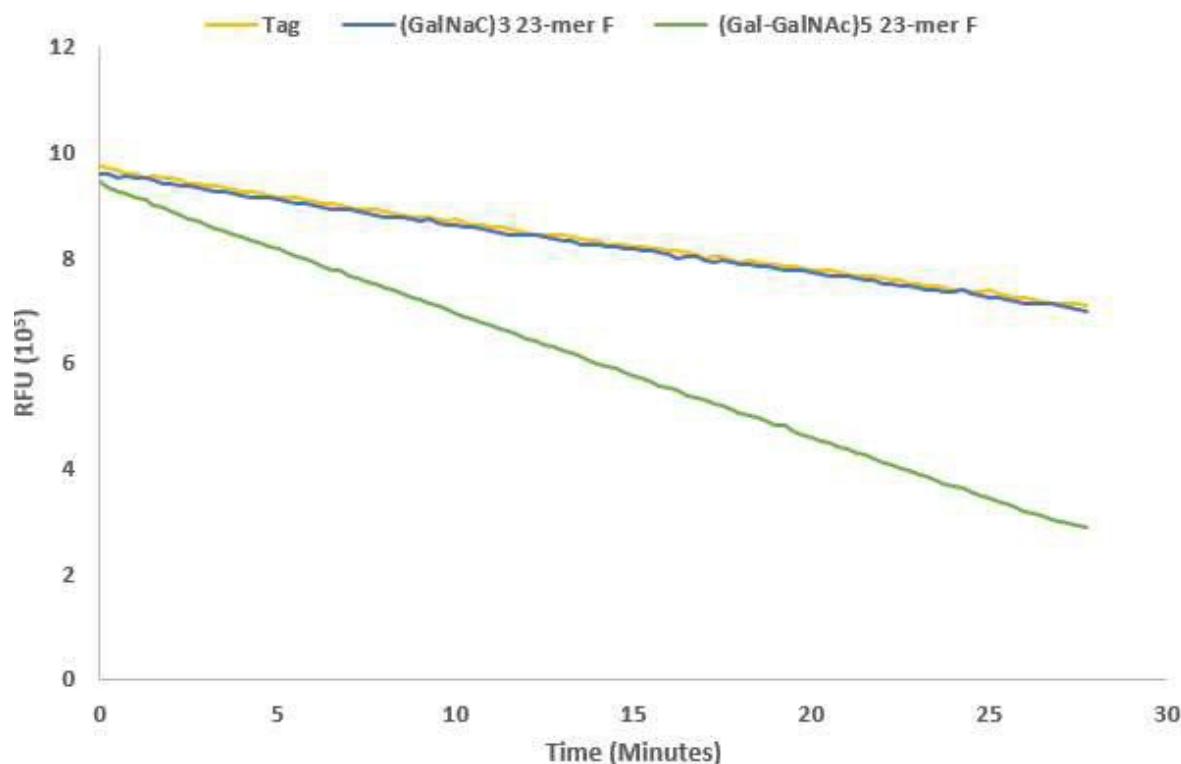


Figure 5.19. Activity of ST3GAL1 with C1GALT1 product. The activity of St3GAL1 was tested with different acceptors. Enzyme was used at a concentration of 0.25  $\mu$ M, all acceptor substrates were added at a concentration of 25  $\mu$ M, CMP-Sia at a concentration of 1000  $\mu$ M. The activity was measured using CMK/PK/LDH assay, acceptor substrates were used to initiate the reactions. Progress curves were normalized to the progress curve of the reaction with the tag as an acceptor

Finally, the TEV protease cleavage efficiency was tested on all the products (Figure 5.20). A ratio of 1:7 of TEV protease to fusion protein was sufficient to cleave all the proteins over night at room temperature.

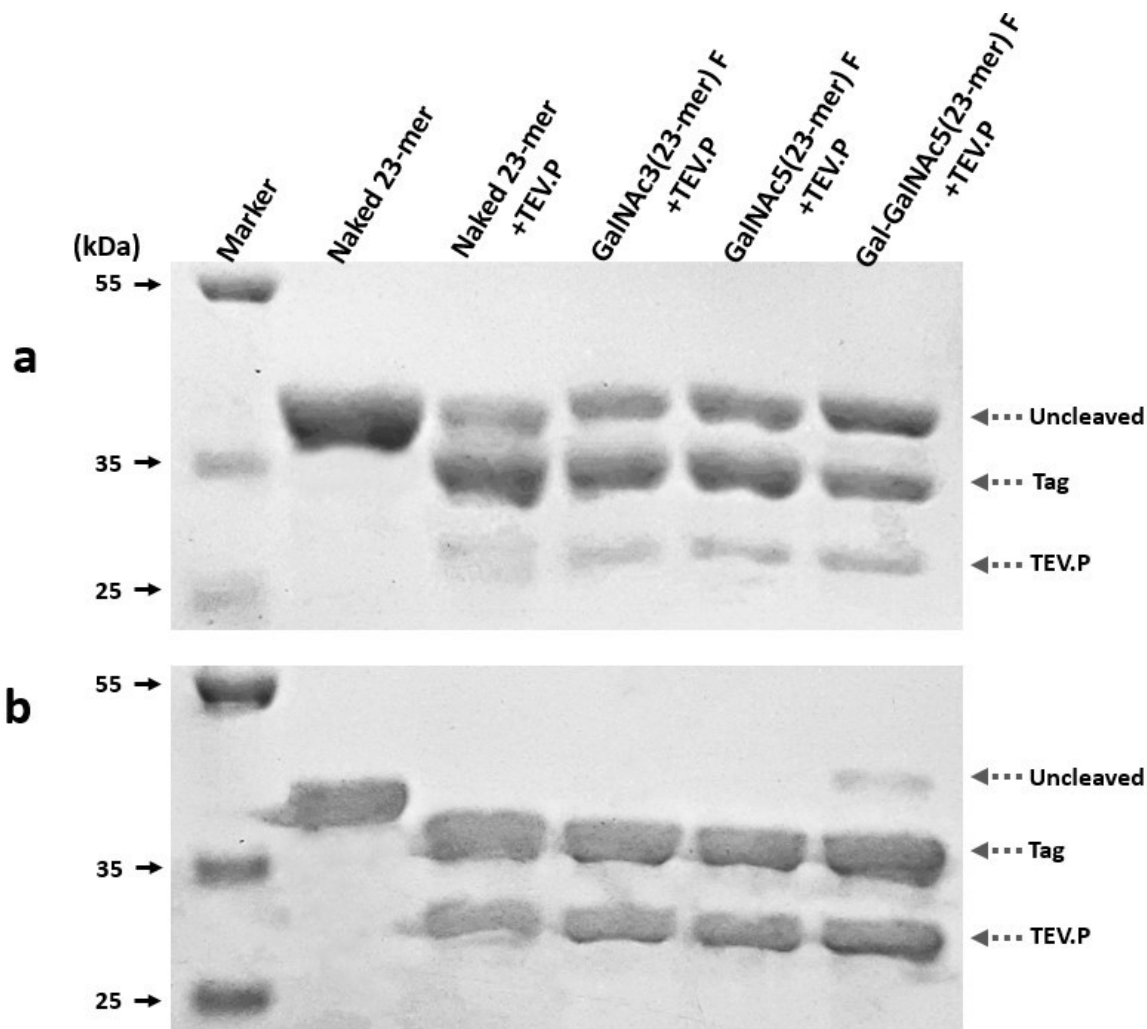
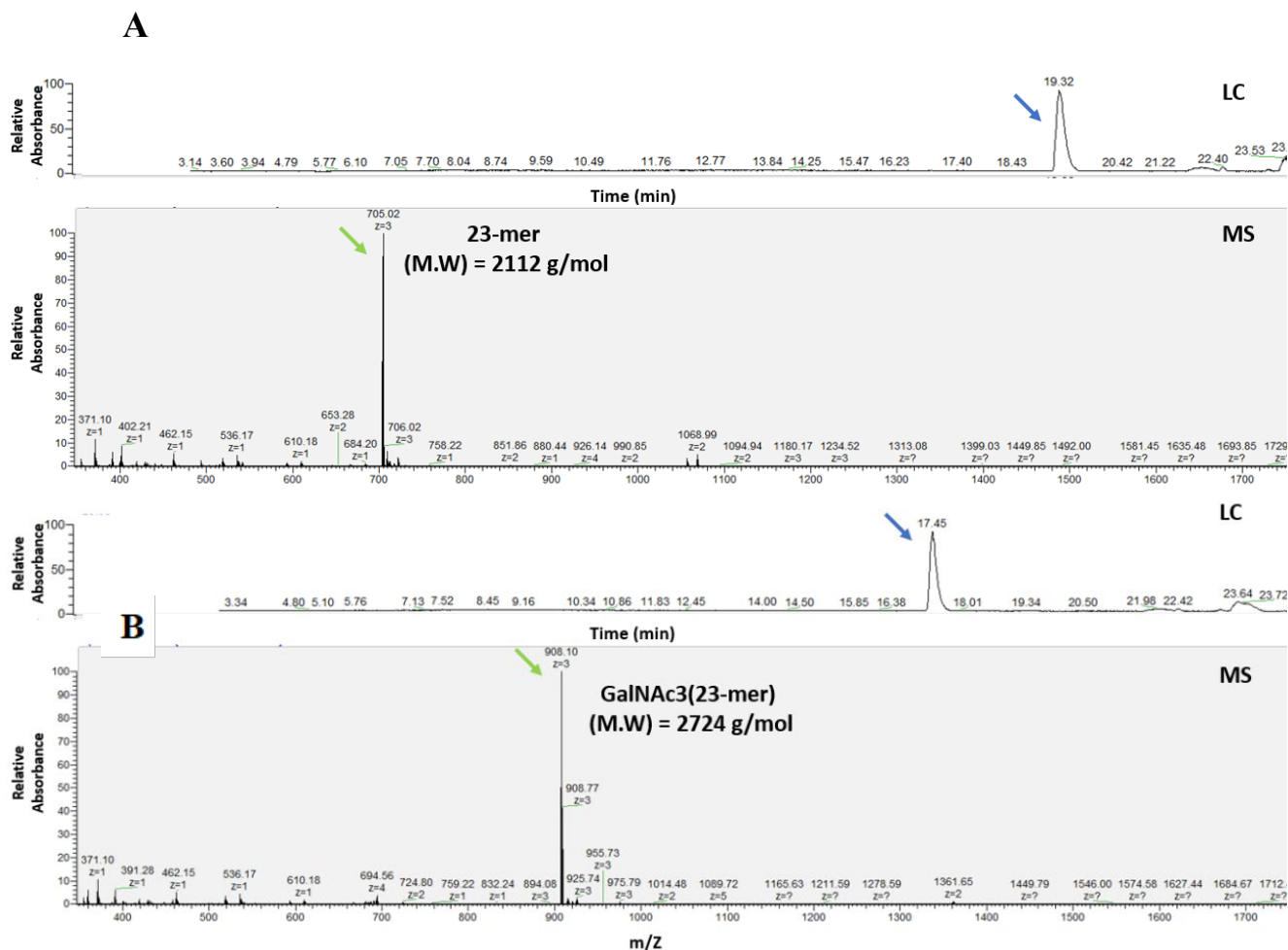


Figure 5. 20. Optimization of TEV protease cleavage. Two ratios of TEV protease to protein were tested: 1:15 and 1:7, Panels a, and b, respectively. Cleavage was carried out in TEV protease buffer (A.2.1) over night at room temperature.

## 5.7 Structure Confirmation via LC-MS

The products of the cleavage with TEV protease performed in the previous section were purified with spinning the reaction mixture on 10 kDa molecular weight cutoff filters. The flow through fractions were collected and analyzed by LC-MS as described in detail in A.9.



ate

**C**

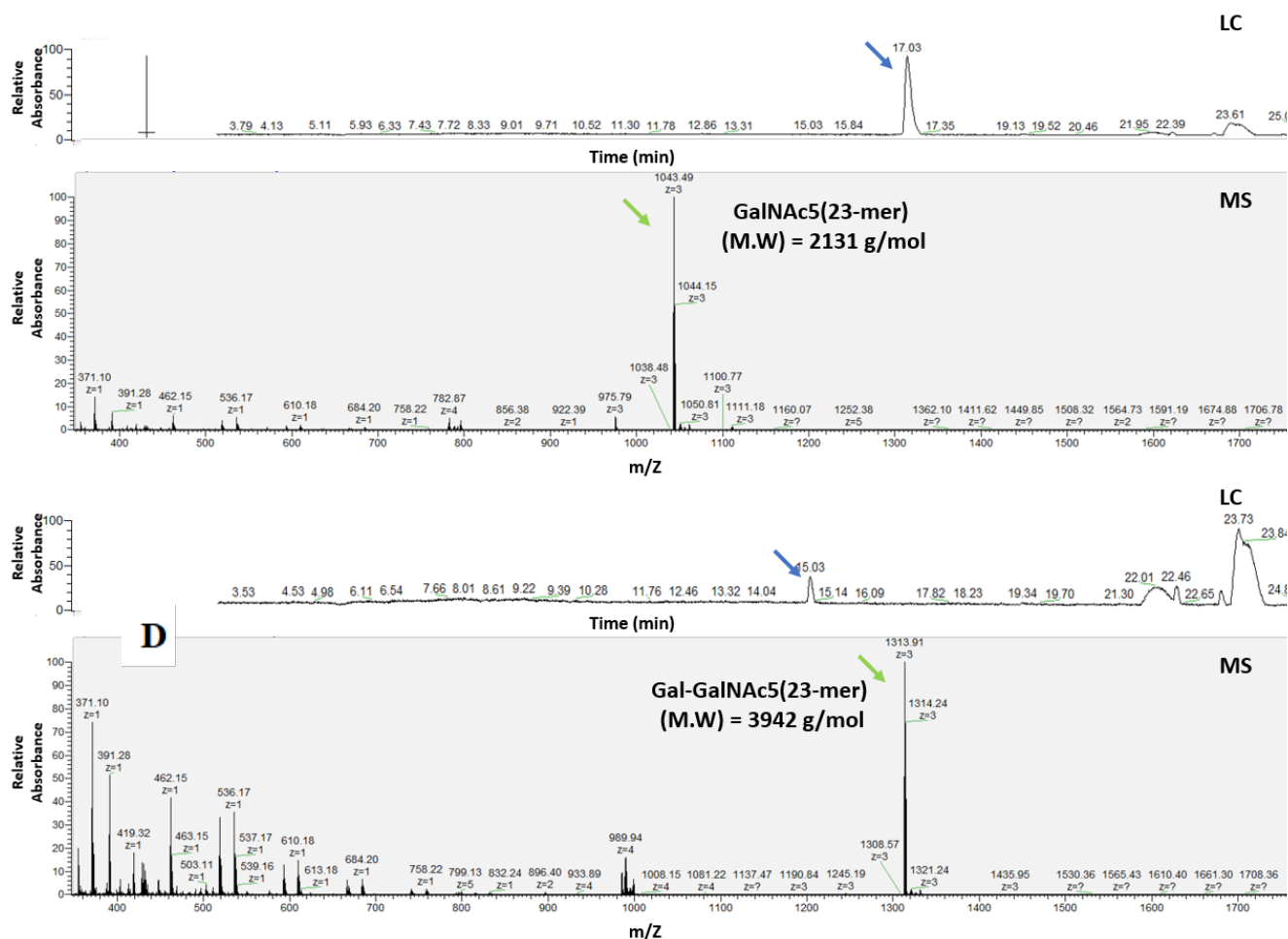


Figure 5.21. LC-MS results for the cleaved and purified glycopeptides. Panels A, B, C, and D show the chromatograms and the mass spectra of the naked 23-mer, GalNAc3(23-mer), GalNAc5(23-mer), and Gal-GalNAc5(23-mer) peptides, respectively. The elution peaks on the chromatograms are labeled with the green arrows and the  $m/z$  peaks are labeled with the blue arrows.

The results of the LC-MS analysis (Figure 5.21) confirm the purity and the molecular weight of the products. The LC spectra (performed on reverse phase protocol) conform increasing the hydrophilicity by increasing the monosaccharides moieties attached to the peroxides. This is concluded by the observed reduction in the retention time of the eluted product as the monosaccharide number increases.

## References

- 1 Jumper, J. *et al.* Highly accurate protein structure prediction with AFold. *Nature* **596**, 583-589 (2021).
- 2 Hanisch, F.-G. & Müller, S. MUC1: the polymorphic appearance of a human mucin. *Glycobiology* **10**, 439-449 (2000).
- 3 Ju, T., Aryal, R. P., Kudelka, M. R., Wang, Y. & Cummings, R. D. The Cosmc connection to the Tn antigen in cancer. *Cancer Biomark* **14**, 63-81 (2014).
- 4 Ryuko, K. *et al.* Characterization of a new MUC1 monoclonal antibody (VU-2-G7) directed to the glycosylated PDTR sequence of MUC1. *Tumour Biol* **21**, 197-210 (2000).
- 5 Gill, D. J. *et al.* Initiation of GalNAc-type O-glycosylation in the endoplasmic reticulum promotes cancer cell invasiveness. *Proc Natl Acad Sci U S A* **110**, E3152-3161 (2013).
- 6 Gill, D. J., Chia, J., Senewiratne, J. & Bard, F. Regulation of O-glycosylation through Golgi-to-ER relocation of initiation enzymes. *JCB* **189**, 843-858 (2010).
- 7 Wandall, H. H. *et al.* The lectin domains of polypeptide GalNAc-transferases exhibit carbohydrate-binding specificity for GalNAc: lectin binding to GalNAc-glycopeptide substrates is required for high density GalNAc-O-glycosylation. *Glycobiology* **17**, 374-387 (2007).
- 8 Wandall, H. H. *et al.* Substrate specificities of three members of the human UDP-N-acetyl- $\alpha$ -D-galactosamine:Polypeptide N-acetylgalactosaminyltransferase family, GalNAc- T1, -T2, and -T3. *J Biol Chem* **272**, 23503-23514 (1997).
- 9 Coelho, H. *et al.* Atomic and Specificity Details of Mucin 1 O-Glycosylation Process by Multiple Polypeptide GalNAc-Transferase Isoforms Unveiled by NMR and Molecular Modeling. *JACS Au* **2**, 631-645 (2022).
- 10 Hassan, H. *et al.* The lectin domain of UDP-N-acetyl-D-galactosamine: polypeptide N-acetylgalactosaminyltransferase-T4 directs its glycopeptide specificities. *J Biol Chem* **275**, 38197-38205 (2000).
- 11 Ju, T. & Cummings, R. D. A unique molecular chaperone Cosmc required for activity of the mammalian core 1  $\beta$  3-galactosyltransferase. *Proc Natl Acad Sci U S A* **99**, 16613-16618 (2002).
- 12 Wang, L. *et al.* Therapeutic peptides: current applications and future directions. *Signal Transduct Target Ther* **7**, 48 (2022).
- 13 Xiang, T., Qiao, M., Xie, J., Li, Z. & Xie, H. Emerging roles of the unique molecular chaperone Cosmc in the regulation of health and disease. *Biomolecules* **12**, 1732 (2022).
- 14 Burchell, J. M., Beatson, R., Graham, R., Taylor-Papadimitriou, J. & Tajadura-Ortega, V. O-linked mucin-type glycosylation in breast cancer. *Biochem Soc Trans* **46**, 779-788 (2018).

# **Chapter 6**

## **Application Case Study: Specificity of ST6GALNAC1**

## 6.1 Overview

This chapter is a use case of the glycopeptides synthesized in Chapter 5. The specificities of ST6GALNAC1 to different substrates were investigated. Comparison of the specificities confirmed that the inclusion of the core peptide in the acceptor structure is a prerequisite for the enzymatic activity. This result was in line with the previously reported studies on ST6GALNAC1 specificity<sup>1,2</sup>. To leverage this finding, the differential specificity of ST6GALNAC1 to the two forms of the Tn antigen, the (GalNAc)<sub>3</sub> 23-mer and the (GalNAc)<sub>5</sub> 23-mer fusion proteins was tested and compared with that of C1GALT1. The results revealed site preference in the case of sialylation of Tn antigen by ST6GALNAC1, whereas no such preference was observed in the case of glycosylation of Tn by C1GALT1. These novel findings were used to provide an updated insight into the regulatory mechanism involved in truncation of MUC1 and expression of the sialylated Tn (STn) observed in breast cancer.

## 6.2 STn in Cancer

Sialylation, the addition of sialic acid units to oligosaccharides and glycoproteins, is involved in crucial biological processes including embryonic and neurodevelopment, oncogenesis, and immune responses. Sialylation regulates these processes primarily through adhesion and cell signaling mechanisms<sup>3</sup>. In this chapter, the role of sialylation is studied through the lens of MUC1 glycosylation in breast cancer. MUC1 is a transmembrane mucin that consists of two subunits, an extracellular heavily GalNAc O-glycosylated subunit, MUC1-N, and an intracellular subunit MUC1-C. MUC1 is expressed in the epithelia of mammary glands, female reproductive tract, lung, kidney, stomach, gall bladder and pancreas<sup>4</sup>. The primary role of MUC1 is protecting the underlying epithelium from pathogens, pollutants, and pH changes. MUC1 exerts these functions via its MUC1-N subunit. In addition to protection, MUC1 serves as a sensor, detecting signals from the surrounding microenvironment and transducing them via modulating other receptors on the cell surface or via its exclusive signal transduction pathway involving the MUC1-C subunit<sup>5</sup>.

In breast cancer, MUC1 is overexpressed and this overexpression correlates to poor outcome<sup>6,7</sup>. In addition to the expression level, Tumor associated MUC1, TA-MUC1, is characterized by shorter glycan structures, Tn and T (core 1) antigens. Sialylated forms of these short glycans,

STn and ST are the main glycan structures found in TA-MUC1<sup>5</sup>. The biosynthesis of GalNAc O-glycosylation is explained in detail in Chapter 4. The truncation of MUC1 glycans is explained by the loss of the activity of core 2 synthase enzymes (GCNTs) and the sialylation of Tn and T antigens, which terminates the elongation and the branching of the glycans. Sialylation of Tn and T is catalyzed by ST6GALNAC1 and ST3GAL1, respectively<sup>8</sup>. From these sialylated products, STn has been found to be overexpressed predominantly in all epithelial cancers. For example, around 90 % of colon cancer samples were reported to have high expression of STn, this percentage is around 50 % in breast cancer, and the lowest, around 10 %, was detected in hepatic cancer<sup>9</sup>.

### **6.2.1 STn as a Marker in Breast Cancer**

A study carried out on breast cancer tissue samples by performing immunohistochemistry staining found that STn was exclusively expressed in stage 3 and metastatic cancers<sup>10</sup>. In a similar study, STn expression was found to be associated with HER2 expression, ER negativity (associated with aggressiveness) and axillary node metastasis<sup>11</sup>. Such correlations were also confirmed by another study but disagreed with the association with lymph node metastasis status<sup>12</sup>. Recently, a study carried out on cells isolated from pathological nipple discharge found that expression of STn has a high breast cancer diagnostic specificity (95 %) and HER2 association<sup>13</sup>. However, one study did not find any significant correlation with any of the prognostic markers including lymph node and hormone receptors expression<sup>14</sup>. Despite the common consensus on the diagnostic value of STn, its association with other prognosis markers in breast cancer is inconsistent across studies<sup>9</sup>. Improvement on the utilization of a marker requires a better understanding of the underlying regulatory mechanisms of its expression and its downstream effect on disease progression. Construction of the regulatory network enables a more accurate holistic (system) approach to classification, which requires upgrading the classification model from one marker at a time to a module-based classification. In this chapter, the Tn glycopeptides models produced in Chapter 5 will be used to study the synthesis of STn and upgrade its regulatory model.

## 6.2.2 STn and Cancer Progression

The association between STn and cancer development is well established. In a study investigating the downstream effect of STn, stable overexpression of ST6GALNAC1 in MDA-MB-231 breast cancer cell lines was used to overexpress STn<sup>15</sup>. The study found that overexpression of STn enhanced cell mobility *in vitro*. When STn<sup>+</sup> cells were injected in mice, the tumors formed in these mice grew faster and larger than those injected with STn<sup>-</sup> cells. Furthermore, even within the tumors generated from STn<sup>+</sup> cells, STn expression was higher in the less dense regions of the tumor. These regions are associated with increased cell mobility. Both observations suggest that the oncogenic effect of STn is migration and proliferation. The association between STn and cancer progression was also confirmed in other cancer types. However, the detailed mechanisms are not well understood. Evidence from all reported studies suggests that the mechanisms with which STn aggravates cancer involve alteration of the binding affinity of STn-bearing MUC1 to lectin molecules, such as selectins, siglecs, and galectins. Binding with lectins regulates cell–cell and cell–matrix interactions, and may influence cancer progression<sup>16</sup>

## 6.3 Regulation of STn expression

### 6.3.1 Role of ST6GALNAC1

One of the early studies that identified ST6GALNAC1 as the major regulator of STn expression is the study by Sewell *et al.* (2006)<sup>17</sup>. This study confirmed the correlation between STn expression and the mRNA levels of ST6GALNAC1, not ST6GALNAC2, in samples from primary breast cancer tissues. In the same study, immunohistostaining revealed that ST6GALNAC1 protein is distributed evenly across Golgi sacks without preference to either cis or trans Golgi. The role of ST6GALNAC1 in STn expression was confirmed in the same study by using the Chinese Hamster Ovary (CHO) cells to overexpress MUC1, either alone or in combination with ST6GALNAC1. Profiling O-glycans on MUC1 extracted from the coexpression cells showed that STn made up 83 % of the total glycans, whereas no STn was detected on MUC1 when expressed alone. Interestingly, the study found that ST6GALNAC1 reduced the site occupancy of the five unique sites on MUC1 tandem repeat by 25%. The role of ST6GALNAC1 in the overexpression of STn was also confirmed in other studies<sup>2,15,18</sup>. Although

in some of these studies, an *in vitro* activity of ST6GALNAC2 in sialylating Tn antigen was observed<sup>1</sup>, this activity was not confirmed in cell models<sup>2</sup>, which was in agreement with the Sewell study mentioned above. Therefore, it is widely accepted that ST6GALNAC1 and not ST6GALNAC2 is the sialyltransferase responsible for STn expression. The role of ST6GALNAC1 in synthesizing STn was also confirmed in different cancers, such as prostate cancer<sup>19</sup>.

*In vitro* investigation of the substrate specificity of ST6GALNAC1 (and ST6GALNAC2) confirmed the contribution of the amino acid sequence of the glycopeptide substrate to the enzyme specificity. These studies also confirmed that the peptide core of the substrate is a prerequisite for the enzyme activity, given that the enzymes did not show any activity with the benzylated GalNAc as an acceptor<sup>1,2</sup>. Despite the confirmation of the contribution of the amino acid sequence to the activity, the site specificity of ST6GALNAC1 has not been investigated at the time of writing this thesis.

### **6.3.2 The role of C1GALT1 and its Chaperone**

As it was reviewed in Chapter 4 (Figure 4.1), Tn antigen is the precursor of the glycan structures observed in healthy tissues. Branching and elongation of Tn result in the formation of the core structures, mainly core 1, core 2, and their extended forms in MUC1 glycoprotein. The first step of the biosynthesis of these healthy glycans is the galactosylation of Tn by C1GALT1 (core 1 synthase). In Chapter 5 (Section 5.4), the role of C1GALT1 and its chaperone C1GALT1C1 (cosmc) was presented in the context of its competition with GALNT4 and the impact of this competition on the occupancy of S9 and T13 sites. In a similar mechanism, glycosylation of Tn by C1GALT1 competes with the sialylation of the same substrate by ST6GALNAC1. Therefore, the loss of C1GALT1 activity has been found to correlate to STn overexpression. Expression of C1GALT1 in an active form requires coexpression with its chaperone C1GALT1C1 (cosmc)<sup>20</sup>, and all studies addressing C1GALT1 loss of function point to cosmc as the main cause of this loss of function.

Mutations in cosmc were identified as a cause for C1GALT1 loss of activity in colon cancer and melanoma cell lines<sup>21</sup>. However, a comprehensive study investigated the presence of mutations of the cosmc gene in tissue samples from different cancers: colon, breast, gastric, lung,

hepatocellular, and acute leukemia. The study did not detect any mutation in any of the tested samples<sup>22</sup>. A reasonable conclusion was made by the authors of this study that mutations in cosmc might be rare events and those detected in the cell lines studies are not naturally occurring and might have arisen during the subculture of the cell lines. In addition to mutations, hypermethylation of the cosmc gene was identified as another mechanism of cosmc malfunction. For example, a study found that hypermethylation was detected in 40 % of pancreatic cancer samples and this hypermethylation was correlated to the loss of C1GALT1 expression and the overexpression of Tn and STn glycans<sup>23</sup>. The transcriptional regulation of C1GALT1 was studied by Zeng *et al.*, 2015<sup>24</sup>. This study identified similar elements in the promoter sequence in both C1GALT1 and cosmc genes. These promoters were found to be a transcription target for the Krüppel-like transcription factors, including SP1/SP3, which has been found to be elevated in cancer cells<sup>25</sup>. The results presented above suggest that C1GALT1 expression is not the main driving factor of elevated STn found in cancer, particularly in breast cancer.

## **6.4 GALNTs relocation to the ER**

In a study by Gill *et al.*, 2010<sup>26</sup>, the cellular localization of multiple GALNTs: GALNT1, GALNT2, GALNT3, GALNT4, and GALNT6, was studied in HeLa cells and the human fibroblast cells, W138. The study found that all the GALNTs tested were redistributed from the Golgi to the endoplasmic reticulum (ER) when the non-receptor tyrosine kinase, Src was stimulated with EGF (Epidermal Growth Factor,). This redistribution was found to be mediated via the retrograde trafficking mechanism that is mediated by COP-1 vesicles. The study found that redistribution of GALNTs to the ER, in Src activated cells, produced MUC1 with higher O-glycans density. In a following study<sup>27</sup>, the same group found that overexpression of GALNT2 in the ER induced the expression of Tn by 7-10 folds compared to 3 folds when GALNT2 was overexpressed in the Golgi. The localization of GALNT1 and GALNT2 in the ER was found to induce Tn expression in multiple human cancer cell lines and this mechanism was inhibited when Src was deactivated. The effect of C1GALT1 deactivation was compared to that of GALNT2 ER localization on Tn expression. Total cellular Tn levels were elevated in both cases. However, Tn was mainly distributed on the cell surface in C1GALT1 deactivated cells, whereas the overexpression of Tn was detected on cell surface but mainly in the ER when GALNT2 was relocated to the ER. T antigen disappeared in C1GALT1 deactivated cells, whereas ER-expressed GALNT2 cells showed

elevated T antigen. However, this T elevation was more pronounced when GALNT4 was overexpressed in the Golgi compared to when it was overexpressed in the ER. Collectively, these results suggest that localization of GALNTs in the ER results in overexpression of Tn in the ER and that a fraction of this Tn transits to the cell surface without modification, while another fraction transits with modification to T antigen. In patient samples, the study found that the mean expression of Tn in breast cancer tissue samples was 4.5 folds higher than in normal tissues. From the samples with high Tn expression, 70 % of the samples showed ER localization of GALNTs (inferred indirectly from ER localization of Tn), whereas no significant loss of C1GALT1 was detected in these samples.

## **6.5 ST6GALNAC1 Core Peptide requirement**

In this thesis, the human ST6GALNAC1 was expressed in HEK 293 cells as described in the general protocol (A.2.5). Expression results including the purity of the enzyme as indicated from the SDS-PAGE gel are shown in the appendix (A 3.5). To test the activity of ST6GALNAC1, the enzymatic-coupled assay CMK/PK/LDH was used to titrate the enzyme following the general assay protocol (A.4). Both the donor and the acceptor were used at 500  $\mu$ M and a twofold serial dilution of the enzyme was tested in triplicate. The acceptor and donor were added to the reaction mixture (Master mixture, A.4) and the enzyme initiated the reaction. The titration was first performed with the benzylated GalNAc as an acceptor. In this trial, no significant activity was detected in any of the tested concentrations of ST6GALNAC1. The titration was repeated using the same conditions but with asialofetuin (AF) as an acceptor (Figure 6.1). AF was used at a concentration of 5 mg/ml, which is equivalent to 0.1 mM of the glycoprotein<sup>28</sup>. A significant activity was detected with AF at a concentration of the enzyme as low as 130 nM, despite using a lower concentration of AF compared to the benzylated GalNAc. These results agreed with the previous studies<sup>1,2</sup>. The results confirmed that the enzyme was expressed in an active form. Asialofetuin is a desialylated form of the fetuin protein extracted from bovine blood. The composition of the glycans on this protein is a mixture of O- and N- glycans<sup>28</sup>. AF is a commonly used acceptor with many sialyltransferases. However, the lack of specific composition due to the batch-to-batch variation makes AF an invalid substrate for quantification of the activity. The reactivity of ST6GALNAC1 with AF, but not with the benzyl-GalNAc, suggests that the enzyme required a peptide core in its acceptor as it was mentioned above.

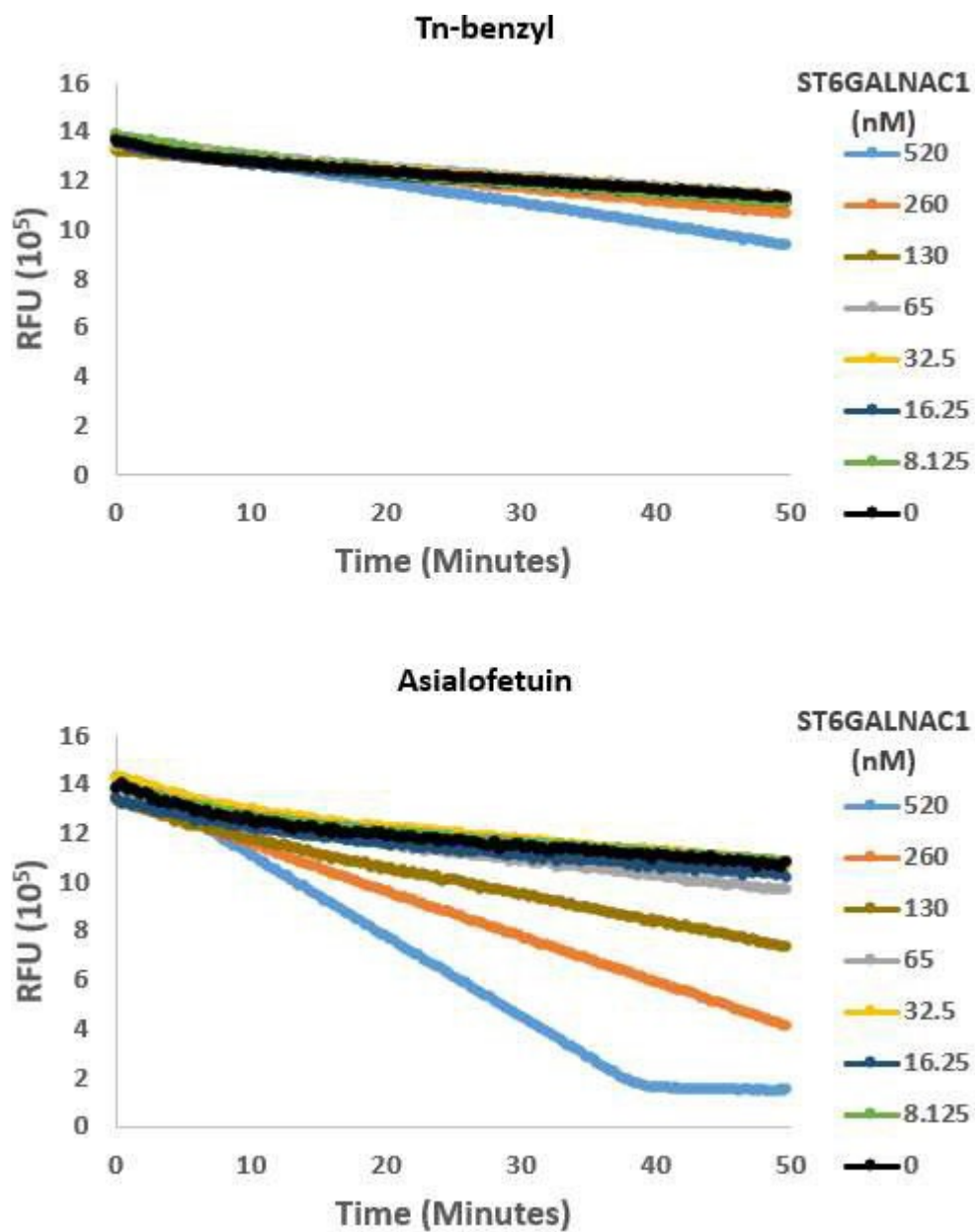


Figure 6.1. Titration of ST6GALNAC1. ST6GALNAC1 was titrated with two substrates: benzylated GalNAc (Tn-benzyl) and Asialofetuin as acceptors. Reaction preparation is explained in the text (section 6.5).

## 6.6 Site Specificity of ST6GALNAC1 vs C1GALT1

From the review on STn regulation above, a conclusion can be made that the competition between ST6GALNAC1 and C1GALT1 on Tn is the main factor determining the abundance of STn over core 1 and its extended glycans. However, the site preference of these enzymes has not been investigated before. To test that, the two Tn glycopeptides that were produced in the preparative synthesis in Chapter 5: the (GalNAc)<sub>3</sub> 23-mer and the (GalNAc)<sub>5</sub> 23-mer fusion proteins were used. These two substrates are representatives of the differential sites' occupancy observed in normal and cancer tissues. The activity test was performed for both enzymes with each of the following substrates as acceptors: the (GalNAc)<sub>3</sub> 23-mer fusion protein, the (GalNAc)<sub>5</sub> 23-mer fusion protein, and the cleaved tag (fusion protein without the peptide). Acceptors were added to a final concentration of 25 µM. Both enzymes were used at a final concentration of 250 nM. The final concentration of the donor, CMP-Sia, in the reaction was 2 mM. The reaction mixture was prepared according to the general protocol (A.4), which excluded the enzymes and the acceptors. The enzymes and the acceptors were used to initiate the reaction. Reactions were prepared in triplicate. The progress curves were normalized to the progress curve of the reaction that contains the tag as a substrate. The results show that for both enzymes, the glycosylation rate was higher with the (GalNAc)<sub>5</sub> 23-mer fusion protein compared to that with the (GalNAc)<sub>3</sub> 23-mer fusion protein (Figure 6.2, A, and B). C1GALT1 reactions showed a rapid progress phase within the first minute of the reaction and a slow plateauing afterward, indicating depletion of the substrate. ST6GALNAC1 reactions were significantly slower, inferred from the moderate slope of the progress curves. The slopes of the progress curves were calculated from the linear regression. Only the rapid phase of the progress curves of C1GALT1 reactions were used in the linear regression (Figure 6.2, B). The rates were calculated from the linear regression by converting the slopes to rates from CMP standard curve (Figure 3.9, C). The rates were presented in bars and normalized to the number of sites occupied with GalNAc in the corresponding acceptor (Figure 6.2, C and D).

The rate difference between the two acceptors is attributed to a combination of two factors: the molar equivalent increment (concentration increment) resulting from addition of more GalNAc residues in the (GalNAc)<sub>5</sub> 23-mer compared to (GalNAc)<sub>3</sub> 23-mer, and the differential specificity to the additional GalNAc residues. From the Michaelis-Menten curve (Figure 2.2) the rate

increment is linear with the substrate concentration at substrate concentration range smaller than  $K_m$ . The rate increment declines as the substrate concentration increases. Therefore, the site normalized rate is expected to decline in (GalNAc)<sub>5</sub> 23-mer compared to (GalNAc)<sub>3</sub> 23-mer. The site-normalized rate for ST6GALNAC1 was twofold higher with (GalNAc)<sub>5</sub> 23-mer compared to (GalNAc)<sub>3</sub> 23-mer. This elevation can only be explained by the additional specific reactivity accompanying the addition of GalNAc residues at S9 and T13. C1GALT1 showed no significant difference between the two acceptors. This can be explained by one of two possibilities: 1. No additional reactivity was attributed to the addition of GALNAc at S9 and T13, and the concentrations of the substrates fall significantly below  $K_m$ , therefore the rate increment is a linear function of the number of sites (concentration), 2. The decline in rate per site was balanced with an increase in the reactivity per site. Since the  $K_m$  value is not determined, no conclusion can be made from this result.

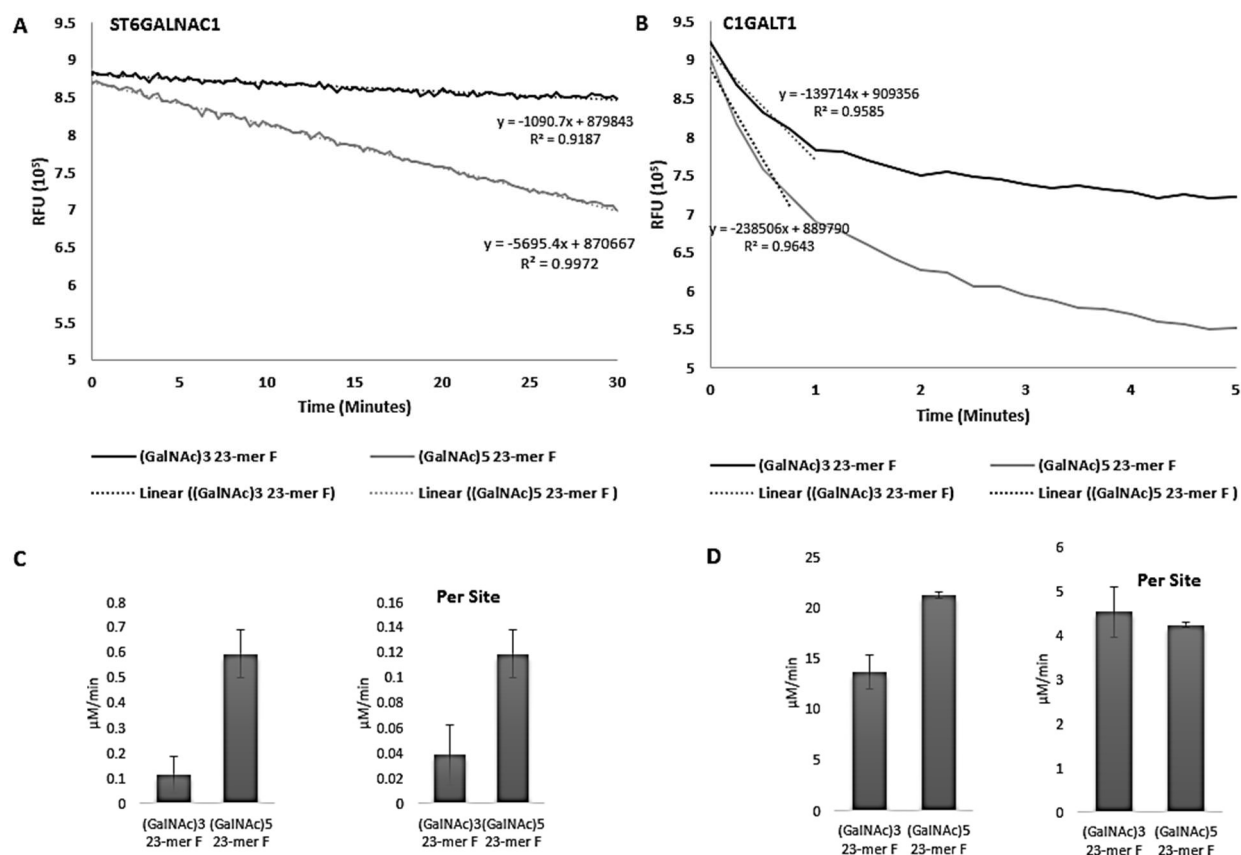


Figure 6.2. Substrate specificity of C1GALT1 and ST6GALNAC1. The activity of both enzymes was tested with the two Tn acceptors as explained in the text (section 6.6). A, and B, are the progress curves of ST6GALNAC1 and C1GALT1 with the two acceptors, respectively. Progress curves were normalized to a control reaction containing the tag as an acceptor. The bars in C, and D represent the rates calculated from the linear regression of the corresponding progress curves and by converting the slope to rate using CMP standard curve (Figure 3.9, C). The rates per site were calculated by dividing the rates for each acceptor by the number of GalNAc groups in the structure.

## **6.7 Summary of the site specificities and Regulation model of STn synthesis**

The site specificity of the different GTs, based on the results obtained in Chapter 5 and Chapter 6, is summarized in Figure 6.3. As it was discussed in section 5.4, The sites on MUC1 can be classified into two groups: Group 1, which includes S19, T20, and T8 sites, and Group 2, which includes S9 and T13. Group 1 sites are glycosylated by a group of GALNTs, including GALNT1 and GALNT2, expressed in abundance in all tissues, including normal and tumor tissues.

However, S9 and T13, in Group 2, are associated with the high-density glycosylation pattern observed in cancer. From what is currently known about GALNTs specificities, GalNAc glycosylation of Group 2 sites is: dependent on GalNAc occupancy of the sites in Group 1, slow, and limited to the activity of GALNT4. Both ST6GALNAC1 and C1GALT1 can glycosylate any GalNAc residue on any site independent from the status of the occupancy of the other sites.

Based on the results in the previous section, C1GALT1 showed no conclusive preference between Group 1 and Group 2, whereas ST6GALNAC1 showed a significant preference to Group 2.

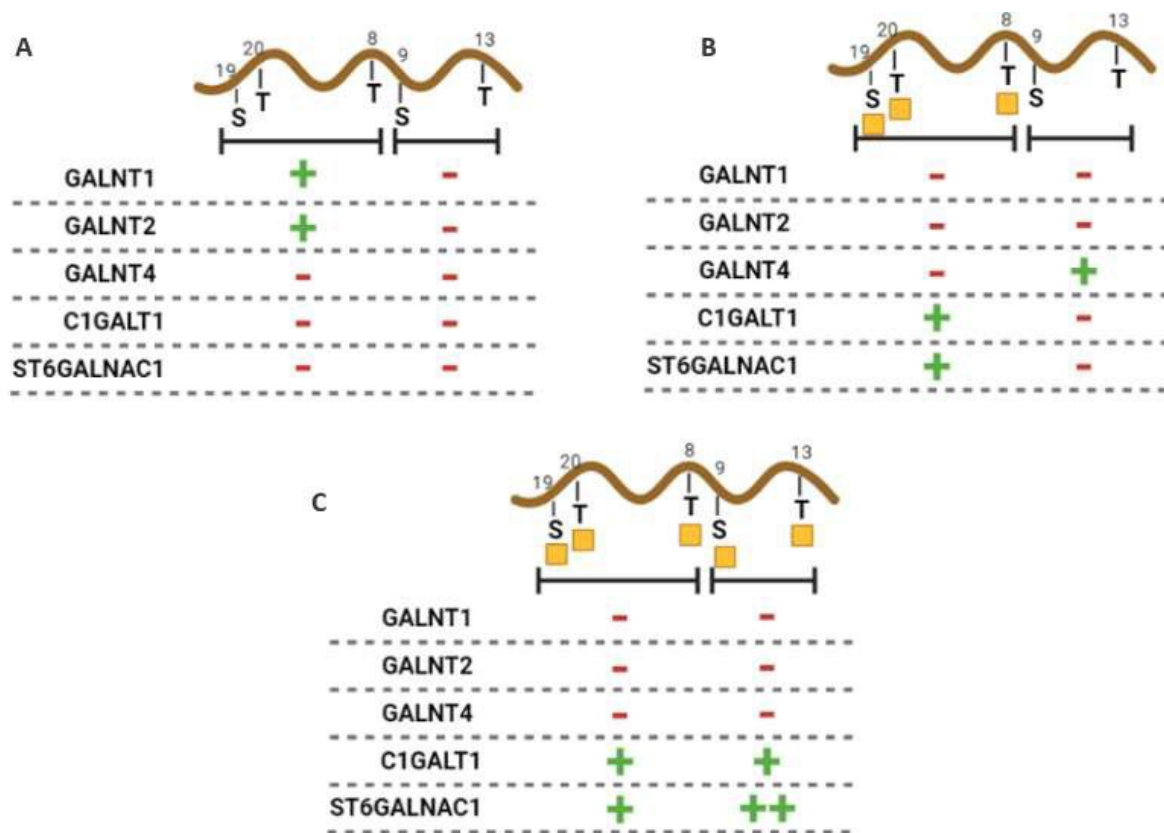


Figure 6.3. Summary of all the specificity tests done in chapter 5 and 6. The five unique sites of MUC1-N are categorized into two groups: group 1 consists of S19, T20 and T8, and group 2 consists of S9 and T13. Group 1 sites are characterized by the abundance of the catalytic activity of GALNTs required for their occupancy with GalNAc, whereas group 2 sites are cancer specific. Specificity of each enzyme to a given site is represented by (-), (+), (++) to indicate lack of reactivity, presence of reactivity, or excess specificity, respectively. A, B, and C, represent the precursor substrates tested in the context of Tn, STn, and T synthesis, respectively.

From the summary of the site specificity represented in Figure 6.3, and the review on STn regulation, the expression of STn is determined by the ratio of the activity of ST6GALNAC1 to C1GALT1, and by a novel proposed mechanism based on the saturation status of the sites by GalNAc. In normal cells, when ST6GALNAC1 expression is low, C1GALT1 expression is high, and GANTs are Golgi-localized, Tn synthesis takes place in the Golgi where the glycosylated Group 1 sites are readily galactosylated by C1GALT1 to produce core 1. This fast transformation

of these sites to core 1, prevents the saturation of the sites in Group 2 by GALNT4 (or GALNT4-like members). The prevalence of C1GALT1 activity compared to ST6GALNAC1 activity confers a competitive advantage to T expression over STn, hence, the low occupancy and absence of STn in normal cells (Figure 6. 4)

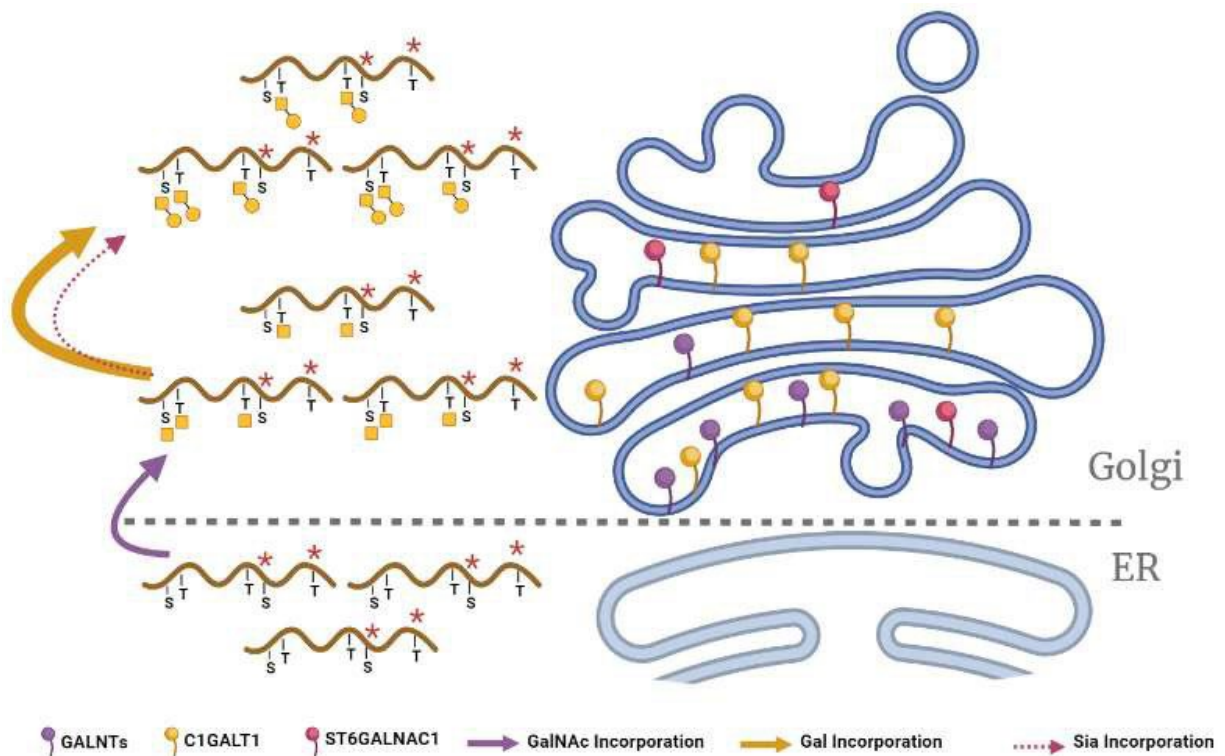


Figure 6.4. Expression of T and STn in normal tissue. Normal tissues are characterized by high expression of C1GALT1, low expression of ST6GALNAC1 and Golgi distribution of GALNTs. This expression signature gives an advantage to the synthesis of T core over STn. C1GALT1 competes with GALNTs and impairs sites saturation with GalNAc. The arrow's width represents the dominance of the reaction relative to the other reactions in the pathway. S9 and T13 are labeled with red asterisks

In the case of cancer tissues with overexpression of ST6GALNAC1, Golgi-localized GALNTs and normal expression of C1GALT1. The expression of STn glycan is explained by the competitive advantage of ST6GALNAC1 over C1GALT1. However, both ST6GALNAC1 and C1GALT1 compete with the glycopeptide-preferring members of GALNTs, such as GALNT4, which results in reduction of the site occupancy (Figure 6.5). This observation was pointed out in the Sewell *et al.* (2006)<sup>17</sup> study.

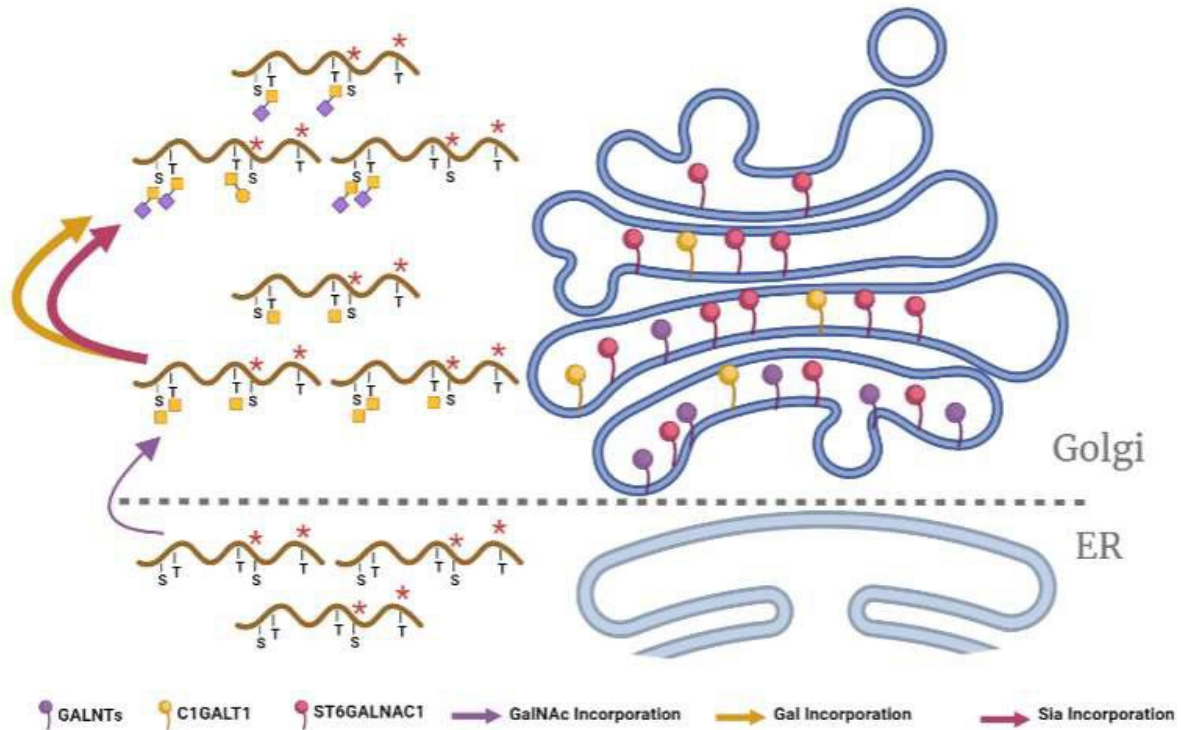


Figure 6.5. The role of ST6GALNAC1 overexpression in STn synthesis in cancer. The model represents the effect of high ST6GALNAC1 to C1GALT1 expression ratio without Golgi to ER redistribution of GALNTs. ST6GALNAC1 competes with C1GALT1 and GALNT4 resulting in expression of STn but with reduced site occupancy, as explained in the text above. The width of the arrow represents the dominance of the reaction relative to the other reactions in the pathway. S9 and T13 are labeled with red asterisks

The correlation between Golgi to ER redistribution of GALNTs and Tn expression is investigated thoroughly in the two papers presented in section 6.4<sup>26,27</sup>. However, in both studies, the impact of this relocation on STn was not the scope of the investigation. Additionally, no site specification was investigated and only GALNT1- and GALNT2- induced redistribution models were used. Therefore, the effect of S9 and T13 sites occupancy by GalNAc on STn expression has not been investigated. Based on the site specificity of ST6GALNAC1, the occupancy of S9 and T13 by GalNAc is favoured by ST6GALNAC1 over the other sites. This suggests that the saturation with GalNAc, associated with GALNTs ER localization, enhances the synthesis of STn independent from the level of ST6GALNAC1 expression (Figure 6.6).

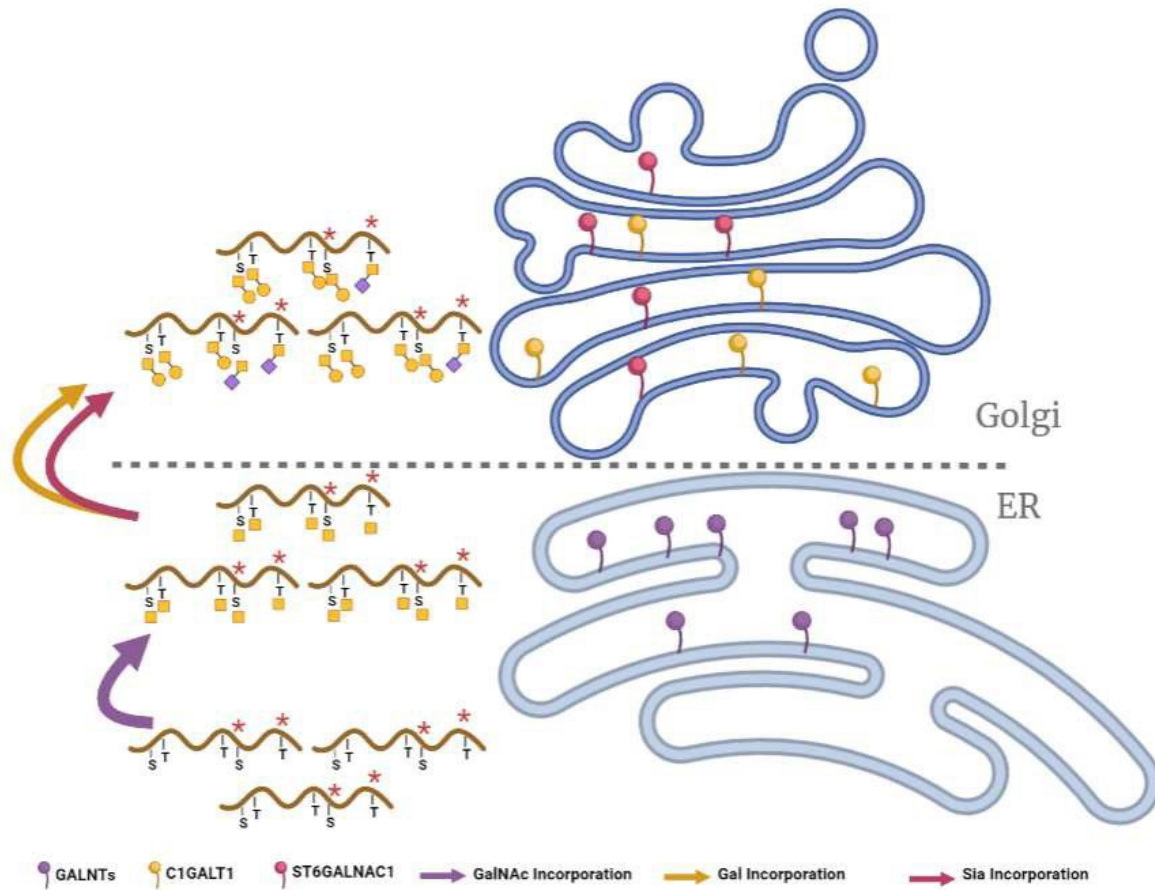


Figure 6.6. Proposed Model based on the site preference: The saturation with GalNAc, associated with GALNTs ER localization, enhances the synthesis of STn independent from the level of ST6GALNAC1 expression. Site preference of Tn sialylation by ST6GALNAC1 is expected to be towards S9 and T13. The arrow's width represents the dominance of the reaction relative to the other reactions in the pathway. S9 and T13 are labeled with red asterisks

## References

- 1 Kono, M. *et al.* Redefined substrate specificity of ST6GalNAc II: a second candidate sialyl-Tn synthase. *Biochem Biophys Res Commun* **272**, 94-97 (2000).
- 2 Marcos, N. T. *et al.* Role of the human ST6GalNAc-I and ST6GalNAc-II in the synthesis of the cancer-associated sialyl-Tn antigen. *Cancer Res* **64**, 7050-7057 (2004).
- 3 Li, F. & Ding, J. Sialylation is involved in cell fate decision during development, reprogramming and cancer progression. *Protein Cell* **10**, 550-565 (2019).
- 4 Gendler, S. J. & Spicer, A. P. Epithelial mucin genes. *Annu Rev Physiol* **57**, 607-634 (1995).
- 5 Beckwith, D. M. & Cudic, M. Tumor-associated O-glycans of MUC1: Carriers of the glyco-code and targets for cancer vaccine design. *Semin. Immunol.* **47**, 101389 (2020).
- 6 Jing, X., Liang, H., Hao, C., Yang, X. & Cui, X. Overexpression of MUC1 predicts poor prognosis in patients with breast cancer. *Oncol Rep* **41**, 801-810 (2019).
- 7 Nath, S. & Mukherjee, P. MUC1: a multifaceted oncoprotein with a key role in cancer progression. *Trends Mol Med* **20**, 332-342 (2014).
- 8 Dalziel, M. *et al.* The relative activities of the C2GnT1 and ST3Gal-I glycosyltransferases determine O-glycan structure and expression of a tumor-associated epitope on MUC1. *JBC* **276**, 11007-11015 (2001).
- 9 Julien, S., Videira, P. A. & Delannoy, P. Sialyl-tn in cancer: (how) did we miss the target? *Biomolecules* **2**, 435-466 (2012).
- 10 Ghosh, S. K., Pantazopoulos, P., Medarova, Z. & Moore, A. Expression of underglycosylated MUC1 antigen in cancerous and adjacent normal breast tissues. *Clin Breast Cancer* **13**, 109-118 (2013).
- 11 Soares, R., Marinho, A. & Schmitt, F. Expression of sialyl-Tn in breast cancer. Correlation with prognostic parameters. *Pathol Res Pract* **192**, 1181-1186 (1996).
- 12 Leivonen, M., Nordling, S., Lundin, J., von Boguslawski, K. & Haglund, C. STn and prognosis in breast cancer. *Oncology* **61**, 299-305 (2001).
- 13 Xu, F. *et al.* Diagnostic value of sialyl-Tn immunocytochemistry in breast cancer presenting with pathological nipple discharge. *Cancer Med* **10**, 1783-1790 (2021).
- 14 Imai, J., Ghazizadeh, M., Naito, Z. & Asano, G. Immunohistochemical expression of T, Tn and sialyl-Tn antigens and clinical outcome in human breast carcinoma. *Anticancer Res.* **21**, 1327-1334 (2001).
- 15 Julien, S. *et al.* ST6GalNAc I expression in MDA-MB-231 breast cancer cells greatly modifies their O-glycosylation pattern and enhances their tumourigenicity. *Glycobiology* **16**, 54-64 (2006).
- 16 Munkley, J. The Role of Sialyl-Tn in Cancer. *Int J Mol Sci* **17**, 275 (2016).
- 17 Sewell, R. *et al.* The ST6GalNAc-I sialyltransferase localizes throughout the Golgi and is responsible for the synthesis of the tumor-associated sialyl-Tn O-glycan in human breast cancer. *J Biol Chem* **281**, 3586-3594 (2006).
- 18 Pinho, S. *et al.* Biological significance of cancer-associated sialyl-Tn antigen: Modulation of malignant phenotype in gastric carcinoma cells. *Cancer Lett* **249**, 157-170 (2007).
- 19 Munkley, J. *et al.* The androgen receptor controls expression of the cancer-associated sTn antigen and cell adhesion through induction of ST6GalNAc1 in prostate cancer. *Oncotarget* **6**, 34358-34374 (2015).

- 20 Ju, T. & Cummings, R. D. A unique molecular chaperone Cosmc required for activity of the mammalian core 1  $\beta$  3-galactosyltransferase. *Proc Natl Acad Sci U S A* **99**, 16613-16618 (2002).
- 21 Ju, T. *et al.* Human tumor antigens Tn and sialyl Tn arise from mutations in Cosmc. *Cancer Res* **68**, 1636-1646 (2008).
- 22 Yoo, N. J., Kim, M. S. & Lee, S. H. Absence of COSMC gene mutations in breast and colorectal carcinomas. *APMIS* **116**, 154-155 (2008).
- 23 Radhakrishnan, P. *et al.* Immature truncated O-glycophenotype of cancer directly induces oncogenic features. *Proc Natl Acad Sci U S A* **111**, E4066-4075 (2014).
- 24 Zeng, J. *et al.* Promoters of human Cosmc and T-synthase genes are similar in structure, yet different in epigenetic regulation. *J Biol Chem* **290**, 19018-19033 (2015).
- 25 Li, L. & Davie, J. R. The role of Sp1 and Sp3 in normal and cancer cell biology. *Annals of Anatomy - Anatomischer Anzeiger* **192**, 275-283 (2010).
- 26 Gill, D. J., Chia, J., Senewiratne, J. & Bard, F. Regulation of O-glycosylation through Golgi-to-ER relocation of initiation enzymes. *JCB* **189**, 843-858 (2010).
- 27 Gill, D. J. *et al.* Initiation of GalNAc-type O-glycosylation in the endoplasmic reticulum promotes cancer cell invasiveness. *Proc Natl Acad Sci U S A* **110**, E3152-3161 (2013).
- 28 Spiro, R. G. Studies on fetuin, a glycoprotein of fetal serum. I. Isolation, chemical composition, and physiochemical properties. *J Biol Chem* **235**, 2860-2869 (1960).

# **Chapter 7**

## **Conclusions**

The main objective of this work was concisely outlined in the introduction to this thesis as follows: "design versatile tools that address key aspects crucial for advancing the acquisition of GT-substrate interactome data *in vitro*. These tools demonstrate applicability across all classes of GTs, ensuring improved biorelevance and enhanced accuracy. Moreover, they are tailored to facilitate high throughput formats, while remaining cost-effective and easy to implement, thereby ensuring simplicity in their application."

The specific objective of developing a continuous enzymatic assay was achieved by optimizing the detection method of the PK/LDH. This was achieved firstly, by using the fluorescence property of NADH for detection, and secondly, by optimizing the NADH concentration and the detector's parameters to maximize sensitivity and linearity of detection. Incorporation of the nucleoside kinases CMK and NDK enabled the coupling of CMP and UDP, the byproducts of the glycosylation reactions, into the PK/LDH assay. The result was two assay formats: CMK/PK/LDH and NDK/PK/LDH. These two formats were proven effective in sensitively monitoring both CMP- and UDP-based reactions. The original protocol developed to monitor UDP-based reactions relied on detecting UDP via direct phosphorylation by PK. Here, it was demonstrated that this approach suffered from lag time in detection, which was mitigated in this work by incorporating the NDK enzyme. The assay was utilized in the *in vitro* synthesis optimization and in the specificity studies.

The real-time detection capacity of the assay enabled the identification of the phase shift in the glycosylation reactions associated with the sequential activities of GALNTs and revealed the dependency of GALNT4 on pre-glycosylation by GALNT1 and GALNT2.

An intrinsic limitation of the assay was observed when used in its CMK/PK/LDH format, which was hypothesized to be a secondary phosphorylation reaction of CDP to CTP by PK. This hypothesis was based on two observations: a dose-dependent lag time and additional NADH depletion detected when CMK titration was performed. However, fluorescence was linear with CMK titration, and the performance of the assay was validated by reproducing the kinetics parameters of ST3GAL1. Future work would include testing the assay's performance with GDP, with and without NDK.

The second objective of developing an *in vitro* synthesis system was also achieved. The yield of the fusion protein was high (80 mg of fusion protein from a litre of culture). All the intended properties of the fusion proteins were validated, including solubility, inertness of the tag, cleavage efficiency of the tag, and purification efficiency. These features enabled glycosylation of the peptide in its fusion form, which was utilized in the glycopeptide synthesis and specificity studies. The peptide design enabled the construction of a collection of MUC1 glycopeptides relevant to the native pathway and representative of the glycans in both normal and breast cancer-associated MUC1. The specificity of GALNTs measured in this work using the designed peptide fusion protein and the enzymatic coupled assay were in agreement with the reported results.

Using the peptides in their fusion forms limited the concentration that could be tested. Purification and accurate quantification of the peptides after cleavage from their fusion proteins would be beneficial to perform full kinetics studies by permitting the use of higher concentrations. However, the focus of this work was on proving the feasibility of the design, the evidence gathered in this work was considered sufficient. One important future application of the biosynthesis method is the performance of a one-pot synthesis by combining the glycosylation steps in one reaction.

Finally, the tools developed in this work were employed to investigate the site specificity of ST6GALNAC1. This was achieved by testing the reactivity of the enzyme with Tn glycopeptides with different site occupancy by GalNAc. The differential reactivity suggested a preference of ST6GALNAC1 for the cancer-associated occupancy state of glycosylation. This novel finding, when validated in cells and, provides new insights into the regulatory system of glycosylation. The site specificity results of C1GALT1 were not conclusive, and full kinetics investigation is needed.

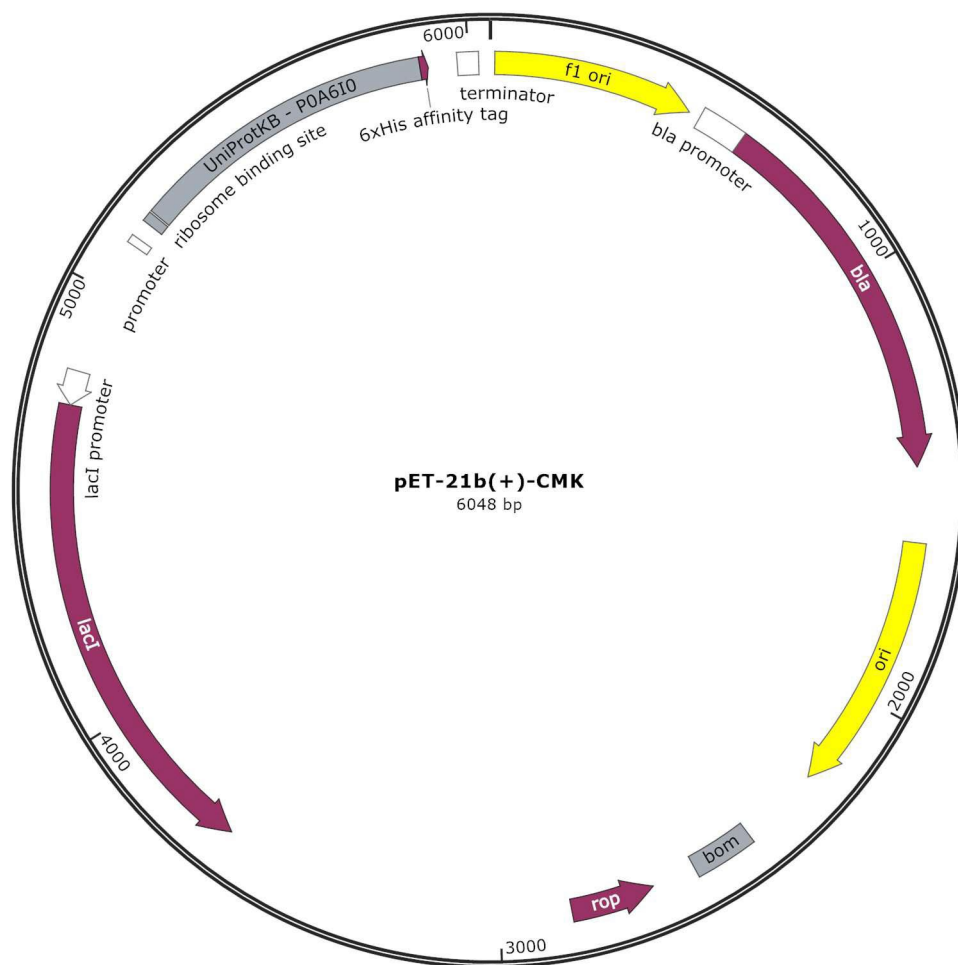
# Appendix

## A.1 *E. Coli.* Expression Plasmids

The gene inserts for all *E. Coli.* expression plasmids were designed and gene insertion was done at BIOMATIK (Ontario, Canada)

### A.1.1 Cytidylate Kinase (CMK)

#### Plasmid Map



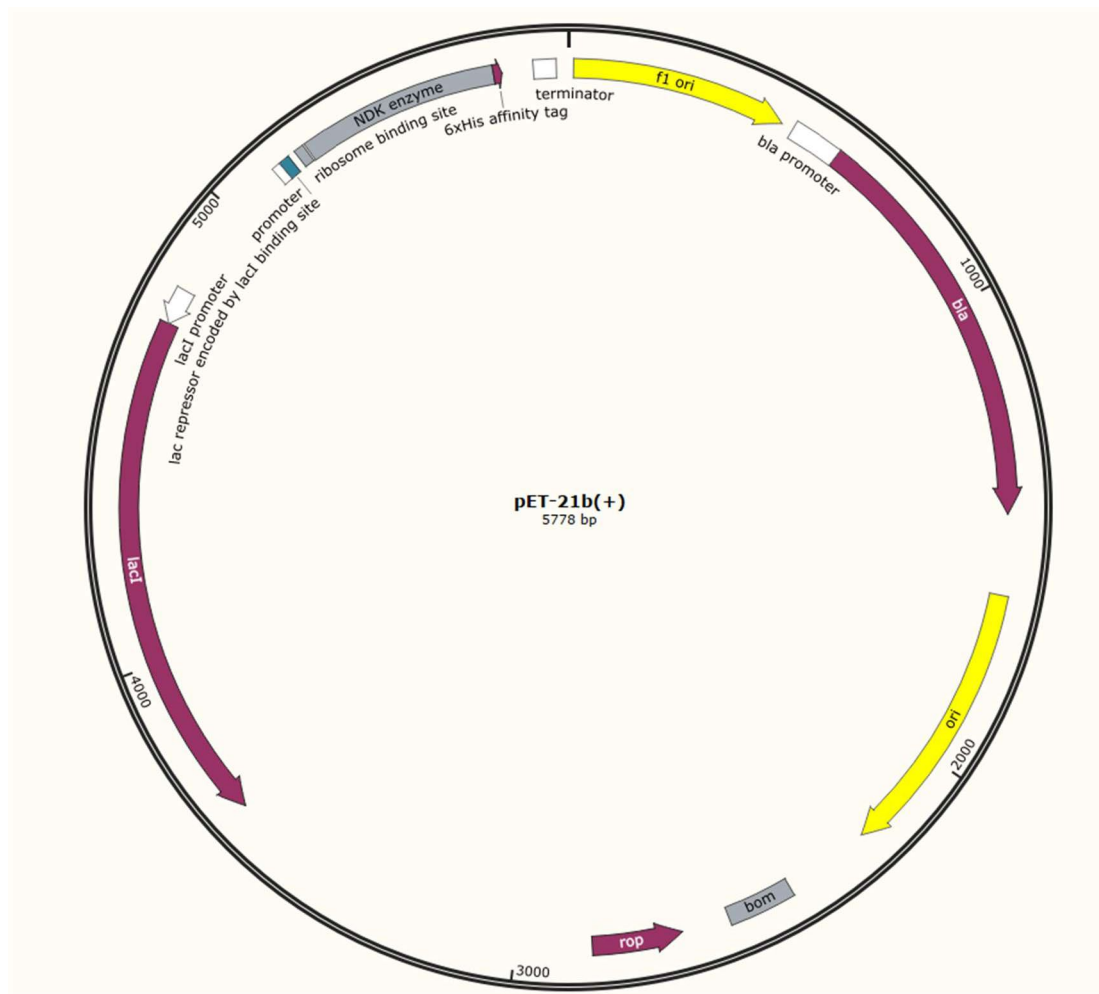
#### Sequence

MTAIAPVITIDGPSGAGKGTLCkAMAEALQWHLLDSGAIYRVLALAALHHHVDVASED  
ALVPLASHLDVRFVSTNGNLEVILEGEDVSGEIRTQEVANAASQVAAPRVREALRRQ

RAFRELPLGIADGRDMGTVVFPDAPVKIFLDASSEERAHRRMLQLQEKGFSVNFERLLA  
EIKERDDDRDRNRAVAPLVPAADALVLDSTTLSIEQVIEKALQYARQKLALA-**HHHHHH**

### A.1.2 Nucleoside Diphosphate Kinase (NDK)

#### Plasmid Map

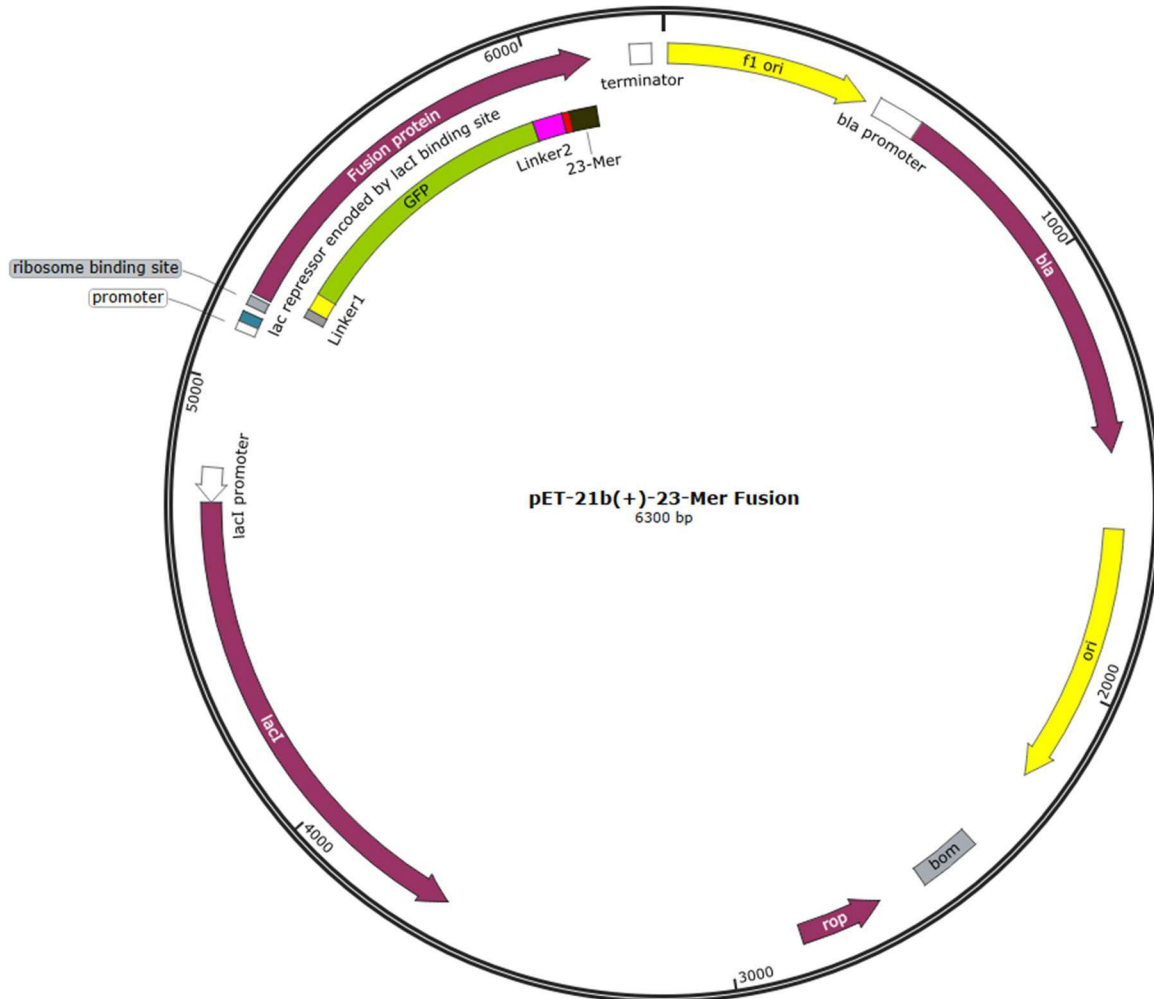


#### Sequence

MQRTLVLIKPDAFERSLVAEIMGRIEKKNFKIVSMKFWSKAPRNLEQHYKEHSEQSYFN  
DNCDFMVSGPIISIVYEGTDAISKIRRLQGNILTPGTIRGDLANDIRENLIHASDSESDAVD  
EISIWFPETKMETDNLE-**HHHHHH**

### A.1.3 23-mer Fusion Proteins

#### Plasmid Map (23-mer fusion)



#### 27-mer Fusion Protein Sequence\*

MHHHHHHHHH(EAAAK)<sub>3</sub>MSKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGK  
 LTLKFICTTGKLPVPWPTLVTTLTYGVCFSRYPDHMKRHDFFKSAMPEGYVQERTISF  
 KDDGTYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFNShNVYITADKQKN  
 GIKANFKIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDH  
 MVLLEFVTAAGITHG(EAAAK)<sub>3</sub>(GGGGS)<sub>2</sub>ENLYFQ/APPAHGVTSAPDTRPAPGSTAPPA

HGV

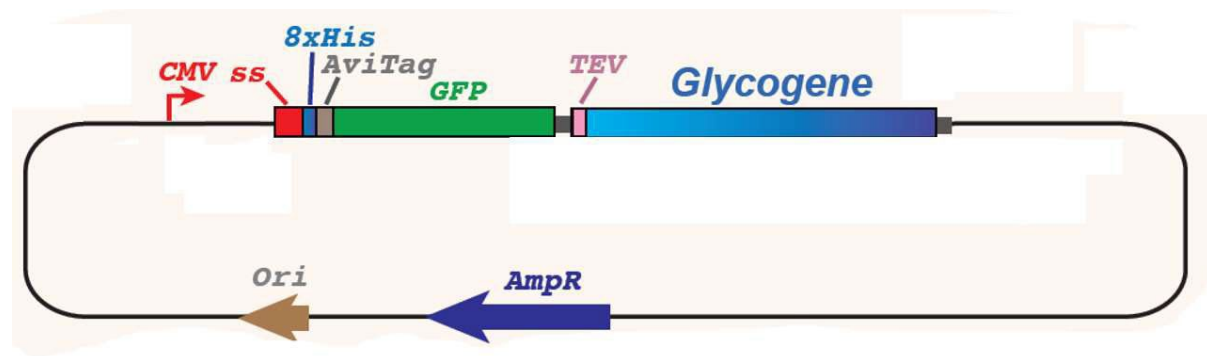
### 23-mer Fusion Protein Sequence\*

MHHHHHHHHH(EAAAK)<sub>3</sub>MSKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGK  
LTLKFICTTGKLPVPWPTLVTTLTYGVCFSRYPDHMKRHDFFKSAMPEGYVQERTISF  
KDDGTYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFNShNVYITADKQKN  
GIKANFKIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDH  
MVLLEFVTAAGITHG(EAAAK)<sub>3</sub>(GGGS)<sub>2</sub>ENLYFQ/APGSTAPPAHGVTSAPDTRPAPG

\*23/27-mer MUC1 F amino acid sequence. This sequence consists of, from N to C terminals: Methionine (red font), Histidine-tag (grey highlight), rigid linker 1 (yellow highlight), superfolder GFP (green highlight), rigid-flexible mixed linker 2 (purple highlight), a TEV site (red highlight), and the amino acid sequence of MUC1 tandem repeat (underlined).

#### A.1.4 GT plasmids

All expression vectors for GTs used in the thesis were purchased from the plasmid repository DNASU. These expression vectors are the pGen2 vectors constructed by Moremen *et al.* (2018) (section 4.8). The general map of the expression vectors is shown below:



The elements shown in the figure are as follows: CMV - Cytomegalovirus Promoter, ss - secretion signal peptide, 8xHis - histidine tag, Avi-tag - Avidin tag, GFP - Superfolder Green Fluorescent Protein, TEV - Tobacco Etch Virus Protease Recognition Site, ampR - Ampicillin Resistance Gene, ori - Origin of Replication.

## A.2 Materials and General Protocols

### A.2.1 Materials

#### Chemicals

Yeast extract (Cat. No. HG000BX6), bacteriological agar (Cat. No. HG000BX1), and tryptone (Cat. No. HG000BX4) were purchased from biolab (Merck, Modderfontein, South Africa). Polyethylenimine (PEI), Linear, MW 25000, Transfection Grade (cat. No. 23966-1) was purchased from Polysciences (USA). L-Lactic dehydrogenase (LDH, Cat. No. L2500), pyruvate kinase (PK, Cat. No. P1506), Cytidine 5'-monophosphate (CMP, Cat. No. C1006), bovine serum albumin (BSA, Cat. No. A3059), N-(2-Hydroxyethyl) piperazine-N'-(2-ethanesulfonic acid), 4-(2-Hydroxyethyl) piperazine-1-ethanesulfonic acid (HEPES, Cat. No. H3375), (Gal)<sub>2</sub>(GlcNAc)<sub>2</sub>(Man)<sub>3</sub>(GlcNAc)<sub>2</sub>Fuc (NA2F-Glycan, Cat. No. N4540),  $\beta$ -D-Gal-(1 $\rightarrow$ 4)-D-GlcNAc, (N-Acetyl-D-lactosamine, Cat. No. A7791), CMP-sialic acid (CMP-Neu5Ac, Cat. No. C1006), phosphor(enol)pyruvic acid monopotassium salt (PEP-K, Cat. No. 860077), Tris base (Trizma, Cat. No. T1503), isopropyl  $\beta$ -D-thiogalactoside (IPTG, Cat. No. 16758), protease inhibitor (cOmplete, Cat. No. 11873580001), lysozyme (Cat. No. 10837059001), adenosine 5'-triphosphate disodium salt trihydrate (ATP, Cat. No. 10519979001),  $\beta$ -Nicotinamide adenine dinucleotide, reduced disodium salt hydrate (NADH, Cat. No. 10128023001), Valproic acid sodium salt (VPA, Cat. No. P4543), glycerol (Cat. No. G5516), KCl, MgCl<sub>2</sub>, MnCl<sub>2</sub>, ampicillin, and other common salts and reagents were all purchased from Sigma Aldrich (South Africa).

#### Common Buffers

- Standard Freezing Buffer: 20 mM Tris-HCl, 150 mM NaCl and, 10% (v/v) Glycerol, pH 7.6
- TEV buffer: 50 mM Tris-HCl, 0.5 mM EDTA, 1 mM DTT, pH 8.0.
- His-Tag Binding Buffer: 20 mM Tris-HCl, 5 mM Imidazole, 500 mM NaCl, 0.05% (w/v) sodium azide, pH 7.9.
- His-Tag Elution Buffer: 20 mM Tris-HCl, 500 mM Imidazole, 500 mM NaCl, 0.05% (w/v) sodium azide, pH 7.9.
- NiSO<sub>4</sub>: 100 mM
- Strip Buffer: 20 mM Tris-HCl, 100 mM EDTA, 500 mM NaCl, pH 7.9

- LB Media: 10 g/l Tryptone, 5 g/l yeast extract, 10 g/l NaCl.
- LB Agar: 15 g agar in 1 l LB media. Antibiotics were added when needed.
- *E. Coli*. Lysis Buffer: 1 tablet of protease inhibitor and 20 mg of lysozyme added to 10 ml of His-tag binding buffer.
- Assay Reaction Buffer (10X): 500 mM HEPES buffer, 500 mM KCl, 100 mM MnCl<sub>2</sub>, 100 mM MgCl<sub>2</sub>, and 10 mg/ml BSA, pH 7.4.

### Cells and Tissue Culture

- BL21(DE3): (Sigma Aldrich, Cat. No. CMC0014)
- DH5 $\alpha$ : Cells were a gift from E. Sturrock (University of Cape Town, South Africa)
- HEK293F: Cells were a gift from E. Sturrock (University of Cape Town, South Africa)
- All plasticware and glassware used in cell culture are bought from Sigma Aldrich.
- HEK293F Freestyle™ 293 Expression Medium (Thermo Fisher Scientific, Cat. No. 12338026)

#### A.2.2 Producing Chemically competent cells

This protocol was used to produce chemically competent *E. Coli*. cells from both strain DH5-  $\alpha$  and BL21 (DE3). DH5- $\alpha$  cells were used to store the plasmids, including pGen2 plasmids containing glycosyltransferase enzymes, and pET 21-b (+) containing CMK, NDK, 23- mer peptide, and 27-mer peptide. BL21 (DE3) was used for expression of CMK, NDK, 23-mer peptide, and 27-mer peptide cloned into pET 21-b (+) plasmid. The protocol of producing chemically competent cells is as follows:

*E. Coli*. cells were streaked onto plain LB agar plates and grown overnight at 37 °C. A single colony was picked and inoculated into 5 ml LB media, then incubated at 37 °C with shaking at 200 rpm until slightly cloudy (3-4 hours). The mixture was then centrifuged at 4250 RCF for 5 minutes at room temperature, and the supernatant was discarded. The cells were subsequently suspended in 3 ml of cold 50 mM CaCl<sub>2</sub> and left on ice for 30 minutes. After that, another round of centrifugation at 4250 RCF for 5 minutes at 4°C followed, and the supernatant was again discarded. The cells were then suspended in 1 ml of cold (on ice) CaCl<sub>2</sub>. Finally, the suspension was stored at 4 °C and used within 3 days.

### **A.2.3 Heat Shock Transformation**

This is a general protocol used to transform any *E. Coli.* strain with a plasmid:

0.5 µl of plasmid stock was added to 80 µl of competent cells in an Eppendorf tube. The mixture was then placed on ice for 30 minutes. Subsequently, heat shock was performed at 42 °C for 45 seconds, followed by placing the tube on ice for 5 minutes. After that, 900 µl of LB media was added, and the mixture was incubated on a shaker at 37 °C for 1 hour. The cells were then spun down at 12800 RCF for 4 minutes at room temperature, and 850 µl of the supernatant was removed. The cells were suspended in the remaining supernatant. A spread plate of the suspended cells was made on an LB + antibiotic agar plate, which was then incubated overnight at 37 °C. Colonies that formed overnight indicated that the desired plasmid had been taken up. It should be noted that leaving the plate in the incubator for longer could lead to the emergence of satellite colonies, which possess antibiotic resistance but do not contain the desired plasmid. Plasmid propagation and quantification was performed by growing the transformed bacteria overnight with the antibiotic corresponding to the resistance gene in the plasmid. Cells were harvested and plasmid was purified using the GeneJET Plasmid Maxiprep Kit (Cat. No. K0491) purchased from Thermo Fisher (South Africa). Purification steps were followed according to the manufacturer's protocol and quantification was performed using NanoDrop® ND-1000 UV-Vis Spectrophotometer by measuring the absorbance at 260 nm.

### **A.2.4 General Bacterial Protein Expression Protocol**

The protocol below is general, and modifications are indicated in the chapter when implemented: A single colony of transformed *E. Coli.* BL21 was inoculated into 5 ml LB+ antibiotic in a 15 ml tube and grown overnight at 37 °C with shaking at 150 rpm. The next day, the 5 ml culture was transferred into a batch culture flask - 2.5 L flask with 500 mL LB + antibiotic and incubated with shaking at 37°C and 150 rpm until an OD600 of 0.4-0.6 had been reached. The culture was then made up to 0.5 mM with IPTG and incubated with shaking overnight at 37 °C. On the following day, the 500 ml culture was transferred into 50 mL falcon tubes and spun down at 4250 RCF at room temperature. The cleared media was poured out and decontaminated before being discarded as waste, and the pellet was kept for purification, which could be done immediately or stored at -20 degrees until ready to use.

### **A.2.5 Mammalian Protein Expression Using HEK 293 cells**

Starting with a healthy cell culture with >95% viability, a cell count was taken, and the cells were diluted with Freestyle media to a concentration of 1 million cells per ml in batches of 100 ml cultures. The cultures were then incubated for 24 hours. After the incubation period, the cells were counted, spun down at 450 RCF for 5 minutes at room temperature, and resuspended in 50 mL Freestyle media to achieve a final cell count of 2.5 – 3.0 million cells per ml. It is important to note that if the culture did not reach 2.5 million/ml after 24 hours, it might indicate that the culture was not healthy enough for transfection.

The fresh cell suspension was incubated for 30 minutes before the dropwise addition of 150-300 µg of plasmid. Following this step, the culture was incubated for an additional 5 minutes before 450 µg of PEI was added dropwise, equivalent to 900 µl of 0.5 mg/mL PEI, which was made up with 1 mg/ml PEI diluted 1:1 with Freestyle media.

The culture was further incubated for 24 hours, and then 50 ml Freestyle media containing 4.4 mM valproic acid was added to achieve a total culture volume of 100 ml containing 2.2 mM valproic acid (36.56 mg/ml).

Samples of the culture were taken every 24 hours to determine when it was ready for purification. This determination was based on taking fluorescence readings of the cell suspension and of the clarified media (without cells). The ratios of the secreted to the total GFP were calculated to determine the secretion efficiency. When no more increase in fluorescence was detected, and when cell viability fell below 50%, the expression was terminated, and the media was collected.

Once the culture reached the appropriate stage for harvesting, the cells were spun down at 450 RCF for 10 minutes at room temperature, and the media was loaded directly onto an IMAC column for purification after filtering through a 0.45 µm sterile filter.

### A.2.6 IMAC Protocol

The Immobilized Metal Affinity Chromatography (IMAC) was performed throughout the project to purify all the proteins with a terminal poly-histidine tag. All protein purifications, including IMAC, size exclusion, and buffer exchange, were performed using the protein liquid chromatography (FPLC) system, ÄKTA Start. The columns used for IMAC are the 1 ml HiTrap® Chelating High-Performance columns (Cat. No. 17-0408-01), purchased from Cytiva (USA).

Cell lysate from bacterial expression: Cells were collected by spinning the culture media at 4250 RCF for 10 minutes at room temperature. The cells were then suspended in the *E. Coli*. Lysis Buffer described above. Cells were disrupted by sonication, spun down, and filtrated using a 0.45 µm filter. Media from mammalian protein expression experiments were used directly as explained above. The IMAC purification protocol was performed as follows:

After washing the column with ultrapure water, it was loaded with the 100 mM nickel sulfate solution (NiSO<sub>4</sub>). The excess nickel sulfate was washed, and the column was equilibrated with an excess flow of the His-Tag binding buffer. Protein crude extract was loaded into the column at a flow rate of 1 ml/min. After completion of sample loading, the nonspecifically bound proteins were washed by running a volume of the His-Tag binding buffer equal to 10 times the volume of the column. Proteins were eluted by running a gradient of the His-Tag elution buffer over a volume that was optimized for each case. For instance, the volume of the gradient was increased when a crude extract with a complex protein composition was purified, such as the *E. Coli*. lysates. The fraction collector accessory was used to collect fractions throughout the purification steps. Finally, columns were stripped from the nickel at the end of each procedure using the Strip Buffer, washed with water, and preserved in 20% (v/v) ethanol for the next use. To avoid any possibility of cross-contamination, an IMAC column was dedicated to each enzyme.

### **A.2.7 Buffer Exchange (Desalting)**

Buffer exchange took place throughout the project for different purposes. For example, to transfer purified proteins from IMAC buffers to the Standard Freezing Buffer for storage. Buffer exchange was performed using ÄKTA Start on the HiTrap® Desalting Column (Cat. No. 17-1408-01), purchased from Cytiva (USA). The column was equilibrated using the desired buffer, the Standard Freezing Buffer, and the protein sample was loaded at volume of less than 2 ml at a time. Elution fractions were collected and visualized by SDS-PAGE

### **A.2.8 Size Exclusion protocol**

Similar to the buffer exchange protocol, Size Exclusion was performed using HiPrep 16/60 Sephacryl S-100 HR column (Cat. No. 17116501) that has a separation range of 5 000–250 000 Da. The column was purchased from Cytiva (USA). The procedure was performed according to the manufacturer's protocol. Briefly, the column was equilibrated with the desired buffer and the sample was loaded and the elution fractions were collected using the fraction collector. SDS-PAGE was performed on the fractions to identify the protein's size and purity.

### **A.2.9 Expression of TEV Protease**

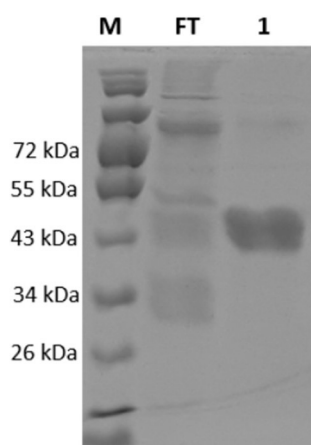
The *E. Coli.* strain BL21(DE3)-RIL transformed with TEV protease vector (pRK793) was used to express and purify TEV protease following the protocol published previously<sup>1</sup> with modification. The transformed strain was grown overnight at 37 °C on LB agar supplemented with 100 mg/l ampicillin and 30 mg/l chloramphenicol. A single colony was picked up to prepare a starter culture overnight at 37 °C and 200 rpm in 10 ml LB supplied with both antibiotics. Inoculation at a ratio of 1:100 of the starter in fresh LB supplemented with antibiotics was carried out and optical density (OD600) was monitored until it reached 0.6 (3-4 hours at 37 °C and 200 rpm).

IPTG was then added to a final concentration of 1 mM to induce the expression and incubation temperature was dropped from 37 to 30 °C. The expression took place for a period of 12 hours. Cells were harvested at the end of the expression by centrifugation and pellets were stored at -20 °C. Aliquots of 1 mg/ml of TEV protease were prepared in 50 mM Tris-HCl, supplied with 30 % (v/v) glycerol, pH 7.4. The aliquots were stored in -80 °C.

## A.3 Expression of GTs

All glycosyltransferase genes used in this thesis were expressed according to the standard protocol above (A.2.5). Tagged enzymes were cleaved with TEV protease and the products of the cleavage reactions were purified with IMAC. SDS-PAGE gel results (of the purification step after TEV protease cleavage) for reach of the enzymes are presented in this section. The purity is calculated from the band intensity using ImageJ software.

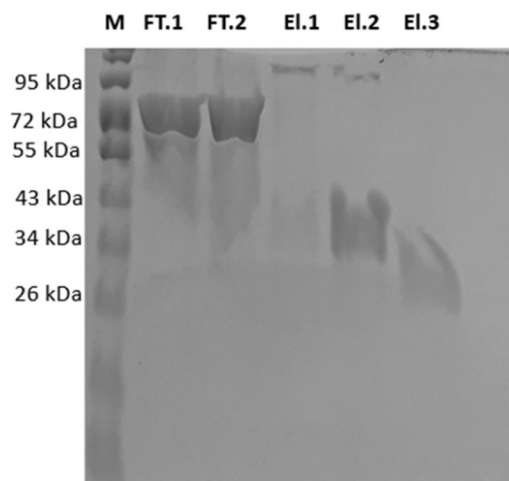
### A.3.1 ST3GAL1



M: Molecular Marker, FT: flow-through, 1: the fraction that contains the pure and cleaved enzyme.

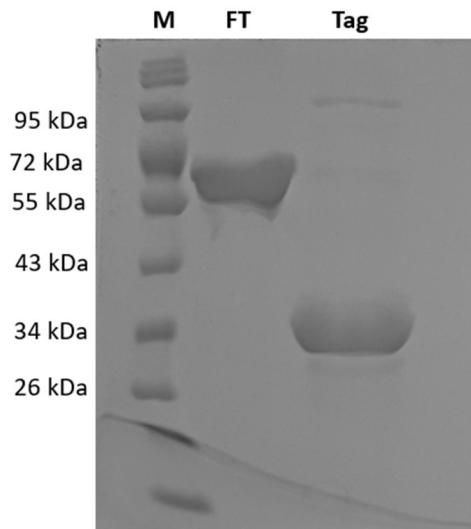
Purity was calculated at 93 %. Final yield was 1.2 mg from a 100 ml expression batch.

### A.3.2 GALNT1



M: Molecular Marker, FT.1 and FT.2: flow-through fractions that contain the pure and cleaved enzyme. El.1: The elution fraction contains traces of the tagged enzyme. El.2 and El.3 are the elution fraction containing the cleaved tag and TEV protease, respectively. Purity was calculated at 94 % (from FT.2). The final yield was 3.35 mg from a 100 ml expression batch.

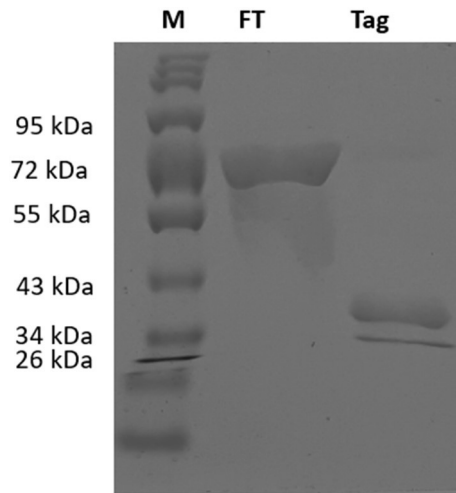
### A.3.3 GALNT2



M: Molecular Marker, FT: flow-through fraction that contains the pure and cleaved enzyme.

Tag: The elution fraction contains the tag and traces of the tagged enzyme. Purity was calculated at 99 %. The final yield was 11.1 mg from a 100 ml expression batch.

#### A.3.4 GALNT4

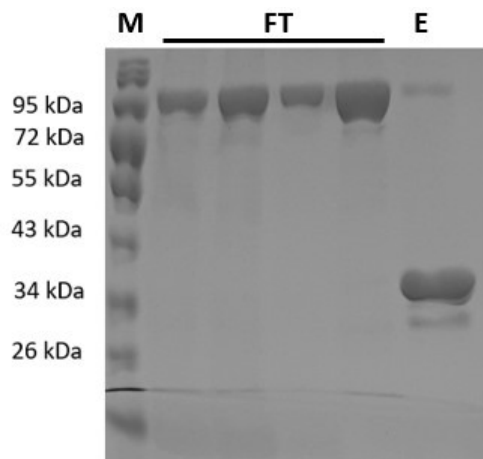


M: Molecular Marker, FT: flow-through fraction that contains the pure and cleaved enzyme.

Tag: The elution fraction contains the cleaved tag (top band) and TEV protease (bottom band).

Purity was calculated at 97 %. The final yield was 0.48 mg from a 100 ml expression batch.

#### A.3.5 ST6GALNAC1



M: Molecular Marker, FT: flow-through fractions that contain the pure and cleaved enzyme. E:

The elution fraction containing the cleaved tag (top band) and TEV protease (bottom band).

Purity was calculated at 95-97 %. The final yield was 1 mg from a 100 ml expression batch.

## A.4 Coupling assay protocol

### Solutions

The general reaction components are prepared as follows:

- 1- Assay Reaction Buffer (10X): 500 mM HEPES buffer, 500 mM KCl, 100 mM  $\text{MnCl}_2$ , 100 mM  $\text{MgCl}_2$ , and 10 mg/ml BSA, pH 7.4.
- 2- NADH (10X): The concentration of NADH (10X) varies depending on the requirement of the assay. However, the final concentration in the reaction was optimized to be between 100 to 150  $\mu\text{M}$ .
- 3- PEP (10X): 20 mM.
- 4- ATP (10X): 20 mM.
- 5- Donor (10X): varies.
- 6- GT enzymes (10 X): varies.
- 7- Acceptor (10X): varies.
- 8- PK: varies.
- 9- LDH: varies.
- 10- CMK/NDK (10X): Titration of each batch is needed to determine the optimum concentration.

### **Master mixture**

Below is an example of a master mixture for 30 reactions:

| COMPONENT      | 1 REACTION             | 30 REACTIONS |
|----------------|------------------------|--------------|
| BUFFER (10X)   | 3                      | 90           |
| WATER          | Complete to 30 $\mu$ l | V            |
| PEP (10X)      | 3                      | 90           |
| NADH (10X)     | 3                      | 90           |
| LDH            | Eq to 3 units          | V            |
| PK             | Eq to 6 units          | V            |
| CMK/NDK (10X)  | 3                      | 90           |
| ATP (10X)      | 3                      | 90           |
| DONOR (10X)    | 3                      | 90           |
| ACCEPTOR (10X) | 3                      | 90           |
| GT ENZYMES     | 3                      | 90           |
| TOTAL          | 30                     | V            |

**V: Variable volumes**

### **Procedure**

- 1- Prepare solutions as mentioned above.
- 2- Prepare the master mixture by combining all the volumes required for the desired number of reactions, excluding the component intended for reaction initiation.
- 3- Incubate the master mixture at 37 °C for 10 minutes for NDK-PK-LDH and 15 minutes for CMK-PK-LDH assay.
- 4- Distribute the reaction initiation component to wells and add the reaction mixture to a final volume of 30  $\mu$ l per well.
- 5- Read the plate using the optimized NADH fluorescence protocol.

The protocol can be modified to include a mixture of enzymes, drugs, or any other components. However, any modification must be accompanied by an adjustment of the water volume to maintain a reaction volume of 30 µl per well.

## **A.5 Pooled standard deviation:**

This method was used to calculate the “average” standard deviation for independent wells, each with its own standard deviation, such as the experiment performed in sections 3.2.2 and 3.3.1. The assumption in this experiment was that the multiple groups of measurements might differ in their means (due to dispensing errors) but have the same standard deviation (caused by factors like signal drop or machine error, which are consistent across all wells), the following method can be used to calculate the pooled standard deviation. The mean of all the groups can be obtained from calculating the mean of means, considering adjustments for the sample size of each group if they are not equal. The standard deviation can be averaged using the formula below:

$$\text{Average S.D.} = \sqrt{((n_1-1)s_1^2 + (n_2-1)s_2^2 + \dots + (n_k-1)s_k^2) / (n_1+n_2 + \dots + n_k - k)}$$

where:  $n_k$ : Sample size for the  $k$ th group,  $s_k$ : Standard deviation for the  $k$ th group,  $k$ : Total number of groups. Alternatively, if the sample sizes are equal for all groups ( $n_1 = n_2 = \dots = n_k = n$ ), the formula simplifies to:

$$\text{Average S.D.} = \sqrt{(s_1^2 + s_2^2 + \dots + s_k^2) / k}$$

where:  $s_k$ : Standard deviation for the  $k$ th group,  $k$ : Total number of groups

## A.6 Batch Mode Simulation

PK/LDH reaction was simulated using COPASI 4.4 (build 278) software as explained in section 3.3.2. The purpose of this simulation is to study the effect of NADH and PK/LDH ratio on ADP detection by the system. Reactions were set as shown in the figure:

| # | Name         | Reaction                        | Rate Law                         |
|---|--------------|---------------------------------|----------------------------------|
| 1 | PK forward   | ADP + PEP = Pyruvate + ATP      | Reversible Ternary Complex Bi-Bi |
| 2 | LDH forward  | NADH + Pyruvate = Lactate + NAD | Reversible Ternary Complex Bi-Bi |
|   | New Reaction |                                 |                                  |

The rate law was adapted from the expression 2.16. The parameters of the law were defined as shown in the figure:

**Function** Reversible Ternary Complex Bi-Bi

Details Notes Annotation RDF Browser

Formula
 
$$\frac{V_f \cdot A \cdot B}{K_m A \cdot K_m B} - \frac{V_r \cdot P \cdot Q}{K_m P \cdot K_m Q}$$

$$1 + \frac{A}{K_m A} + \frac{B}{K_m B} + \frac{P}{K_m P} + \frac{Q}{K_m Q} + \frac{A \cdot B}{K_m A \cdot K_m B} + \frac{A \cdot P}{K_m A \cdot K_m P} + \frac{B \cdot Q}{K_m B \cdot K_m Q} + \frac{P \cdot Q}{K_m P \cdot K_m Q} + \frac{A \cdot B \cdot P}{K_m A \cdot K_m B \cdot K_m P} + \frac{B \cdot P \cdot Q}{K_m B \cdot K_m P \cdot K_m Q}$$

Function Type ☒ reversible ☐ irreversible ☐ general

Parameters
 

| Name | Description | Unit                     |
|------|-------------|--------------------------|
| A    | Substrate   | mmol/l                   |
| B    | Substrate   | mmol/l                   |
| Vf   | Parameter   | mmol/(l*min) or mmol/min |
| KmA  | Parameter   | mmol/l                   |
| KmB  | Parameter   | mmol/l                   |
| Vr   | Parameter   | mmol/(l*min) or mmol/min |
| P    | Product     | mmol/l                   |
| Q    | Product     | mmol/l                   |
| KmP  | Parameter   | mmol/l                   |
| KmQ  | Parameter   | mmol/l                   |

The parameters were adjusted for each reaction as follows:

#### For PK reaction

| Role      | Name | Mapping   | Value | Unit         |
|-----------|------|-----------|-------|--------------|
| Substrate | A    | ADP       |       | mmol/l       |
| Substrate | B    | PEP       |       | mmol/l       |
| Parameter | Vf   | --local-- | 33    | mmol/(l*min) |
| Parameter | KmA  | --local-- | 0.3   | mmol/l       |
| Parameter | KmB  | --local-- | 0.07  | mmol/l       |
| Parameter | Vr   | --local-- | 2.12  | mmol/(l*min) |
| Product   | P    | Pyruvate  |       | mmol/l       |
| Product   | Q    | ATP       |       | mmol/l       |
| Parameter | KmP  | --local-- | 10    | mmol/l       |
| Parameter | KmQ  | --local-- | 0.86  | mmol/l       |

#### For LDH reaction

| Role      | Name | Mapping   | Value | Unit         |
|-----------|------|-----------|-------|--------------|
| Substrate | A    | NADH      |       | mmol/l       |
| Substrate | B    | Pyruvate  |       | mmol/l       |
| Parameter | Vf   | --local-- | 33    | mmol/(l*min) |
| Parameter | KmA  | --local-- | 0.058 | mmol/l       |
| Parameter | KmB  | --local-- | 0.13  | mmol/l       |
| Parameter | Vr   | --local-- | 0.028 | mmol/(l*min) |
| Product   | P    | Lactate   |       | mmol/l       |
| Product   | Q    | NAD       |       | mmol/l       |
| Parameter | KmP  | --local-- | 10    | mmol/l       |
| Parameter | KmQ  | --local-- | 0.23  | mmol/l       |

The Maximum forward velocities values Vf were calculated from the number of units added to the reaction for each of PK and LDH (as provided by the manufacturer). Vr values were calculated from the relation:  $K_{eq} = \frac{V_f K_{mP} K_{mQ}}{V_r K_{mA} K_{mB}}$ .  $K_{eq}$  were obtained from curated values in literature <sup>2</sup>,  $K_m$  values were obtained from Table 3.1.

The simulation was performed in a time course batch mode for initial concentrations of the reaction's species that are defined in section 3.3.2 and Figure 3.3. Simulation was run for 0.1 min, and concentrations of the species were collected for 40 intervals within the simulation duration. Data were exported and plotted for each ADP concentration as one set (Figure 3.3). Concentrations of NADH were converted to "Residue percentage" as shown in Figure 3.3. This was calculated from the formula:  $\text{Residue (\%)} = 100 \cdot (A - A_t) / (A_0 - A_t)$ ; A is the concentration at a given point of time,  $A_t$  is the terminal concentration after complete depletion (equal to NADH initial concentration - ADP initial concentration),  $A_0$  is the initial concentration.

## A.7 ADP Standard

ADP standard curves with different NADH concentration were obtained by performing a procedure that was modified from A.4 as follows:

### Solutions

- 1- Assay Reaction Buffer (10X): 500 mM HEPES buffer, *pH* 7.4, 500 mM KCl, 100 mM MnCl<sub>2</sub>, 100 mM MgCl<sub>2</sub>, and 10 mg/ml BSA.
- 2- NADH (10X) solutions: Five solutions were prepared of NADH (10X) at 3, 2, 1.5, 1, and 0.5 mM.
- 3- PEP (10X): 20 mM.
- 4- ATP (10X): 20 mM.
- 5- ADP (10 X) solutions: a serial dilution of ADP was prepared at 10X of the final concentrations shown in Figure 3.5.
- 6- PK and LDH: 6 units PK and 3 units LDH per reaction

### Master mixture

Below is an example of a master mixture for 30 reactions:

| COMPONENT        | 1 REACTION        | 30 REACTIONS   |
|------------------|-------------------|----------------|
| BUFFER (10X)     | 3                 | 90             |
| WATER            | Complete to 30 µl | V              |
| PEP (10X)        | 3                 | 90             |
| NADH (10X)       | 3                 | 90             |
| LDH              | Eq to 3 units     | V              |
| PK               | Eq to 6 units     | V              |
| SERIAL ADP (10X) | 3                 | Added to wells |
| TOTAL            | 30                | V              |

**V: Variable volumes**

## **Procedure**

- 6- Prepare solutions as mentioned above.
- 7- Prepare the master mixture by combining all the volumes required for the desired number of reactions, excluding ADP.
- 8- Incubate the master mixture at 37 °C for 10 minutes.
- 9- Distribute the 10X ADP serial solutions to wells in triplicate. Add 27 µl from the reaction mixture to each well for reaction initiation.
- 10- Read the plate using the optimized NADH fluorescence protocol.

Reaction master mixtures contain different concentrations of NADH were prepared. Three replicates of a serial dilution of ADP 10X solutions were added to the wells (3 µl per well). Reaction master mixtures were added to the well simultaneously (27 µl) using a multi-channel pipette. The fluorescence was recorded over a period of time (30 minutes). Average signals for each replicate over 10 minutes were calculated (and the standard deviations) using Excel software. The means of the means were calculated and the standard deviations were pooled as explained in A.5.

Z score values were calculated from the three replicates of fluorescence signal measured for each concentration of ADP compared to the three replicates of the control (without ADP). The fluorescence with ADP was used as a positive signal and the control (with no ADP added) was used as the negative signal. The calculations were repeated for different concentrations of NADH. The aim of this is to show the effect of NADH concentration on the sensitivity of detecting ADP. Z scores were calculated using the means of the means and pooled standard deviations.

## **A.8 Fed Batch Simulation**

The model used in A.6 was modified by adding another reaction, ADP flux, as follows:

| # | Name         | Reaction                        | Rate Law                         |
|---|--------------|---------------------------------|----------------------------------|
| 1 | PK forward   | ADP + PEP = Pyruvate + ATP      | Reversible Ternary Complex Bi-Bi |
| 2 | LDH forward  | NADH + Pyruvate = Lactate + NAD | Reversible Ternary Complex Bi-Bi |
| 3 | ADP flux     | NP -> ADP                       | Constant flux (irreversible)     |
|   | New Reaction |                                 |                                  |

ADP flux is a rate zero reaction with constant rate. The rest of the reaction conditions were set the

same as in A.6. Initial concentration of NADH was set to 100  $\mu\text{M}$ . A parameter scan was performed by recording the rates of the reaction species in response to changing ADP flux. The rates of reaction species were recorded at 2 and 4 minutes. The parameter scan setup is demonstrated in the figure below:

**Scan** ☒ Intervals ☐ Values

Object: (ADP flux).v

Intervals: min max

5000 1e-6 0.075

☐ logarithmic

**Task** Time course

☒ Continue from Current State ☐ output during subtask execution ☐ Continue on Error

The output was set to record the rates of all the species at 4 and 2 minutes (Figure 3.7)

## A.9 LC-MS Procedure:

LCMS analysis was conducted with a Q-Exactive quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific, USA) coupled with a Dionex Ultimate 3000 nano-UPLC system. Data was acquired using Xcalibur v4.1.31.9, Chromeleon v6.8 (SR13), Orbitrap MS v2.9 (build 2926) and Thermo Foundations 3.1 (SP4). Peptides were dissolved in a 0.1% (v/v) Formic acid (FA), 2% (v/v) Acetonitrile (ACN) to a final concentration of 10 nM and loaded on a C18 trap column (PepMap100, 9027905000, 300  $\mu\text{m} \times 5 \text{ mm} \times 5 \mu\text{m}$ ). Approximately 50 fmol of peptide was injected per sample. Samples were trapped onto the column and washed for 3 minutes before the valve was switched and peptides eluted onto the

analytical column as described hereafter. Chromatographic separation was performed on a ReproSil-Pur 120 C-18-AQ column (Dr Maisch, r119.aq.n150.075, 75  $\mu\text{m} \times 15 \text{ cm} \times 1.9 \mu\text{m}$ ) as described below. The solvent system employed was solvent A: LC water (Burdick and Jackson BJLC365), 0.1% (v/v) FA and solvent B: ACN, 0.1% (v/v) FA. The multi-step gradient for peptide separation was as follows: time change 3 - 4.4 min, gradient change 2 – 5 % Solvent B, flow rate change to 300 nL/min and time change 0.1 min, time change 4.5 - 20 min, gradient change 5 – 18% Solvent B, time change 20 - 23 min, gradient change 18 – 30% Solvent B, time change 23 - 25 min, gradient change 30 – 80% Solvent B. The flow rate was increased to 450 nL/min and the gradient was then held at 80% Solvent B for 5 minutes before returning to 2% Solvent B for 5 minutes. All data acquisition was obtained using Proxeon stainless steel emitters (Thermo Fisher TFES523). The mass spectrometer was operated in

positive ion mode with a capillary temperature of 320°C. The applied electrospray voltage was 1.95 kV. Details of data acquisition are shown in the table below. The Xalibur Qual Browser was used to determine MS1 and MS2 m/z distributions from the various samples.

Data Acquisition Details:

|                                |                        |
|--------------------------------|------------------------|
| <b>Full Scan</b>               |                        |
| Resolution                     | 70000 (@ m/z 200)      |
| AGC target value               | 3e6                    |
| Scan range                     | 350-1750 m/z           |
| Maximal injection time (ms)    | 100                    |
| <b>Data-dependent MS/MS</b>    |                        |
| Inclusion                      | Off                    |
| Resolution                     | 35000                  |
| AGC target value               | 1e5                    |
| Maximal injection time (ms)    | 100                    |
| Loop Count                     | 9                      |
| Isolation window width (m/z)   | 2                      |
| <b>NCE (%)</b>                 | <b>27</b>              |
| <b>Data-dependent Settings</b> |                        |
| Underfill ratio (%)            | 1                      |
| Charge exclusion               | Unassigned, 1, 5-8, >8 |
| Peptide match                  | preferred              |
| Exclusion isotopes             | on                     |
| Dynamic exclusion (s)          | 20                     |

## **A.10 General Protocol for Buffer Preparation and pH Adjustment**

Buffers were prepared by dissolving the solid components in 80% of the final volume of autoclaved water and adding the liquid components, such as glycerol and triton X-100. The *pH* was adjusted using GLP21 *pH* meter. The *pH* meter was calibrated with the standard buffers (Giorgio Bormac s.r.l, Cat. No. IBOR51100033, IBOR51100043, and IBOR51100073 for *pH* values of 4.01, 7, and 9.21, respectively) the adjustment was performed with concentrated HCl or NaOH while stirring. After the desired *pH* is achieved, water was added to obtain the desired volume of the solution.

## **A.11 Aseptic Techniques and Cell Culture**

Solutions were sterilized in two ways, either by autoclave at a temperature of 120 °C for 30 minutes with saturated steam under pressure of 15 psi. or by performing filter sterilization using 0.22 µm filter. Media for bacterial culture were autoclaved, whereas reagents such as PEI and VPA were filter sterilized. All mammalian cell culture procedures were performed in the biosafety cabinet. Procedures with bacteria including agar plates streaking, liquid culture seeding, transformation, and competent cells preparation were performed near flame source, which was also used to sterilize the loops used for colony pickup. The tips, tubes, and flasks used in cell culture were all autoclaved and dried in a 60 °C oven.

## **A.12 Statistical**

All enzymatic activity tests throughout the thesis were performed using three independent replicates as explained in the protocol in A.7. the source of error introduced during this replication is limited to the dispensing error when the induction substrate is aliquoted to the wells and when the master mixture is added to initiate the reaction. Shapiro-Wilk normality tests were performed when relevant in the figures:

-3.6 A, C, and D

-3.9 A, C, and D

-3.12, D

-3.13

-3.14

-5.8, B

- 5.9, B and D

-5.13, B

-5.17, B

6.2, C, and D

The normality test was performed using GraphPad Prism 8 software. The null hypothesis (that the samples are normally distributed) cannot be rejected when P values are above 0.05 and the samples were designated normally distributed.

The Statistical significance of the comparisons made in figures:

- 5.9 B, and D

- 5,13, B

- 5.17, B

- 6.2, C, and D

Was evaluated using t-test. The parametric t-test was performed using GraphPad Prism 8 software using the following parameters.

**Experimental Design:**

- Experimental design: Paired
- Assume Gaussian distribution: Yes.
- Choose test: Paired t-test

**Options:**

- two-tail P value
- confidence level 95%.

All bar results for activity tests throughout the thesis are represented as means  $\pm$  SD, n = 3.

## References

- 1 Kapust, R. B. *et al.* Tobacco etch virus protease: mechanism of autolysis and rational design of stable mutants with wild-type catalytic proficiency. *Protein Eng* **14**, 993-1000 (2001).
- 2 Imperial, S. & Centelles, J. J. Enzyme kinetic equations of irreversible and reversible reactions in metabolism. *J. Biosci. Med.* **02**, 24-29 (2014).