# NOVEL PROTEIN ENGINEERING APPROACHES FOR CHEMOENZYMATIC SYNTHESIS OF GLYCAN POLYMERS

By

**CHANDRA KANTH BANDI**

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Chemical and Biochemical Engineering

Written under the direction of

Shishir P.S. Chundawat

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2021

# Abstract of the Dissertation

Novel protein engineering approaches for chemoenzymatic synthesis of glycan polymers

By CHANDRA KANTH BANDI

Thesis Director: Shishir P.S. Chundawat

Glycans or carbohydrates are the most abundant class of biomolecules on the planet, but we are far from elucidating their role in the design and regulation of biological ecosystems spanning from organismal- to protein- level. Nearly every human cell-pathogen interaction and immune system related disorder or disease involves protein-glycan and glycan-glycan interactions at the molecular level. With the increase healthcare costs and greater emphasis by the pharmaceutical industry towards developing more effective biological drugs it becomes imperative to unravel the “glycan code”. Unfortunately, glycan synthesis is not template encoded and is the primary reason for the lack of efficient synthetic tools for widespread application. In nature, glycans are synthesized using membrane associated glycosyl transferases that are associated with several limitations including poor expression, solubility, and the need for expensive donor sugars for large scale synthesis. Alternatively, glycosyl hydrolases (GHs) can be reverse engineered and modulated to synthesize sugar polymers using their inherent transglycosylation property. However, transglycosylation pathway of GHs suffers from low yields because of enzyme’s natural preference towards competing water molecules leading to hydrolysis. Fortunately, protein

engineering approaches such as rational engineering and directed evolution can address these challenges to shift the equilibrium towards transglycosylation. In this work, two protein engineering strategies that have been developed that can used engineer glycosyl hydrolysis for efficient glycan synthesis. The first approach is structure guided rational approach where a non-catalytic carbohydrate binding domain is fused to inactive GH family 5 enzymes to rescue the transglycosylation activity. The second approach is high throughput screening development to identify mutant variants that can synthesize glycan using azido-sugars. Here, an azide biosensing toolkit was developed that can be used for directed evolution based glycosynthase engineering. Both these engineering approaches can invariably be applied to many different glycosyl hydrolase families to generate potent transglycosidases.

**This work is dedicated to my mother and my father who have provided me courage and support throughout my education.**

# Acknowledgements

Thank you Mother and Father

I would like to express my sincere gratitude to my advisor, Dr. Shishir Chundawat for providing me an opportunity to work in his lab on this amazing project. I will be grateful to him for providing me with ample support, guidance, confidence, and freedom to explore, invent and grow as an independent researcher. I would like to thank my dissertation committee members Dr. Charles Roth, Dr. Haoran Zhang and Dr. Heather Mayes for their timely guidance and suggestions. This work was funded by Rutgers School of Engineering (SoE) and National Science Foundation (NSF).

I would like to extend my heartfelt appreciation to Bhargava Nemmaru who has been my friend and an amazing colleague from my first day in this lab. I will cherish all the wonderful moments we had setting up and organizing the lab, building, and troubleshooting protocols, brainstorming research ideas, and going to conferences along with Dr. Chundawat. I would like to thank my PhD colleagues, Markus Hackl, Dharanidharan Jayachandran, Aron Gyorgypal and Mohit Kumar for providing continued support and help to complete my work.

My sincere thanks to my colleagues, Ayushi Agrawal, Antonio Goncalves, Madhura Kasture and Namratha Subash for their timely help and support and to accomplish this work. I will especially reminisce all the amazing experiences and moments I had while working with Ayushi on flow cytometer and FACS experiments, with Antonio on time course activity assays and with Madhura on generating and handling numerous constructs. I would like to thank Shraddha Gupta, Navya Sahithi for their help on bioseparations and supercharging project. I would like to thank all my undergraduate colleagues, Alina Thokkadam, Neelan Sivaneri, Wen-Chen Chen, Khushbu Patel,

Deepika Saravana, Jorge Tapia and Margaux Thiry for all the experimental help in these projects. I will be thankful to all my masters and undergraduate colleagues who helped me grow as an individual and a mentor.

I would like to thank Dr. Heather Mayes and Tucker Burgin for their useful insights and technical support for the project. I would like to gratefully acknowledge Dr. Sai Venkatesh Pingali for help with SAXS experiments and data analysis. I would like to thank Kyle S. Skalenko for help with troubleshooting azide promoter system. I would like to thank Dr. Brian Fox for providing the CelE and CelE-CBM3a constructs.

Finally, I would like to express my deepest thanks and love to my father, Mr. Raja Gopal Bandi and my mother, Mrs. Hema Latha Bandi for supporting and motivating me throughout my life and education. I would like to thank my friends Sreenath Yeluri, Varsha Rane and Pranav Ramesh. I am glad I had such amazing roommates, Pruthvi Jujjavarapu and Bharath Kumar. Thank you very much for being patient, for the splendid food, emotional talks, playing FIFA, Age of empires and making it a pleasant home to come back from work.

**THANK YOU!!!**

# Table of Contents

**Abstract of the Dissertation** ..... **ii**

**Acknowledgements** ..... **v**

**Table of Figures** ..... **xi**

**Chapter 1: Introduction** ..... **1**

1.1 Glycans and their importance ..... 1

1.2 Synthesis of glycans ..... 3

1.2.1 Chemical synthesis ..... 3

1.2.2 Enzymatic synthesis ..... 4

1.3 Glycosidase engineering for glycans synthesis ..... 7

1.3.1 Sequence/structure guided rational engineering ..... 7

1.3.2 Selective screening strategies guided randomized engineering ..... 11

1.4 Problem statement ..... 14

1.5 Thesis Objectives ..... 15

**Chapter 2: Sequence and structure guided engineering of GH5 family transglycosidases** ..... **16**

2.1 Introduction ..... 16

2.2 Experimental outline and design ..... 22

2.3 Methods ..... 22

2.3.1 Generation of plasmid DNA constructs ..... 22

2.3.2 Protein expression and purification ..... 23

2.3.3 Protein characterization ..... 24

2.3.4 End point enzyme activity assay ..... 25

2.3.5 Time course enzyme kinetic assay ..... 26

2.3.6 Quantitative Thin Layer Chromatography (TLC) analysis ..... 27

2.3.7 Small angle X-ray scattering (SAXS) analysis ..... 28

2.4 Results and discussion ..... 29

2.4.1 Preliminary efforts to engineer CelE into a glycosynthase ..... 29

2.4.2 CBM3a recovers CelE nucleophile mutants transglycosylation and hydrolytic activity ..... 31

2.4.3 Short chain cello-oligosaccharide synthesis by CelE-E316G-CBM3a ..... 34

2.4.4 Improved glucan synthase activity unique to CBM3a amongst other Type-A/B family CBMs ..... 38

2.4.5 Kinetic analysis of transglycosylation by CelE-E316G-CBM3a ..... 41

2.4.6 Small angle X-ray scattering analysis of CelE-E316G-CBM3a ..... 46

2.4.7 Inter-domain protein dynamic study through linker modifications ..... 49

2.4.8 Mechanistic analysis of transglycosylation by CelE-E316G-CBM3a ..... 52

2.4.9 CBM3a engineering to facilitate transglycosylation in other GH family 5 enzymes. ..... 56

**Chapter 3: Development of high-throughput screening technique to facilitate directed evolution of glycosynthases ..... 60**

3.1 Introduction ..... 60

3.2 Experimental outline and design ..... 62

3.3 Methods ..... 63

3.3.1 Bacterial strain engineering: ..... 63

3.3.2 Design and construction of pCyn vectors: ..... 65

3.3.3 Induction of pCyn-v1/v2-GFP expression: ..... 66

3.3.4 Induction of pCyn-v2/v3/v4-GFP expression: ..... 66

3.3.5 Induction of pCyn-v2/v5/v6/v7/v8-GFP expression: ..... 67

3.3.6 Bulk GFP fluorescence measurement: ..... 67

3.3.7 Flow cytometer data acquisition and analysis: ..... 67

3.4 Results and discussion ..... 68

3.4.1 cynTSX operon ..... 68

3.4.2 Azide is a gratitious inducer for cynTSX operon ..... 70

3.4.3 Engineering GFP promoter ..... 71

3.4.4 CynR constitutive promoter engineering ..... 75

3.4.5 Application of synthetic azide promoter towards heterologous protein expression and glycosynthase screening ..... 78

**Chapter 4: Conclusions and recommendations for future work ..... 81**

4.1 Conclusions ..... 81

4.2 Recommendation for future work ..... 85

**Appendix ..... 86**

A1.1 Supplementary methods ..... 86

A1.2 Supplementary information ..... 88

A1.3 Collaborative projects ..... 98

A1.3.1 Work with Madhura Kasture ..... 98

A1.3.2 Work with Ayushi Agrawal ..... 100

**References ..... 103**

# Table of Figures

**Figure 1. Glycans are an integral part of all biological systems. (A)** Glycans are one of the four building blocks of life (DNA/RNA, proteins, lipids are other three). **(B)** Glycans regulate many biological process such as cell-cell interaction, antibody and hormone regulation, pathogen detection. ..... 1

**Figure 2. Chemical synthesis of oligosaccharides using protecting groups.** Promoter facilitates the departure of leaving group and nucleophile group attacks the anomeric carbon. Depending on the electron withdrawing potential of protecting group (P), the product stereochemistry is controlled. ..... 4

**Figure 3. Mechanism of β-glycosyl hydrolases for glycosidic bond hydrolysis.** ..... 6

**Figure 4. Rational protein engineering strategies for improving transglycosidase activity of glycosyl hydrolases** ..... 8

**Figure 5. Leading examples of leaving group based low to high-throughput based screening strategies to enable directed evolution of glycan synthesizing enzymes.** ..... 14

**Figure 6. Overview of generic $S_N2$ based two-step mechanism employed by retaining glycosyl hydrolases and corresponding nucleophile mutants (e.g., glycosynthases) to facilitate both hydrolysis and transglycosylation type reactions.** Here, **(A)** depicts the glycosylation step initiated by the enzyme catalytic nucleophile residue that results in the formation of a Glucosyl-Enzyme Intermediate (GEI). In the presence of water acting as a nucleophile, hydrolysis (and subsequent deglycosylation of the GEI) takes place for most native wild-type glycosidases. In the presence of a suitable glycosyl acceptor group, transglycosylation can take place as well for some native or engineered glycosidases. **(B)** Mutations at the nucleophile site to develop engineered transglycosidases mutants, like glycosynthases, can often facilitate efficient synthesis of

glycans like oligosaccharides using simple activated donor sugars that mimic the GEI conformation. .....17

**Figure 7. Nucleophilic mutations of CelE did not yield an active glycosynthase. (A)** SDS-PAGE protein characterization for molecular weight and purity. (B) Chemical rescue activity assay of CelE and its nucleophilic mutants on pNP-CB in the presence of 1M sodium azide. (C) Glycosynthase reaction testing using CelE nucleophile mutants on α-cellobiosyl fluoride and cellobiose sugars. The two sugars have similar retention factor (Rf) on TLC plate.....30

**Figure 8. Carbohydrate binding modules (CBMs) aid in recovery of transglycosylation activity of nucleophilic mutants belonging to a family 5 glycosyl hydrolase (CelE).** (A) The relative enzyme activities were estimated by measuring the release of pNP from pNP-Cellobiose (pNPC) as starting substrate. Here, 500 picomoles of each purified protein was incubated along with 2 µmoles of pNPC and the reaction was run for 4.5 h at 60°C reaction temperature. All reactions were run in triplicates with error bars representing 1$\sigma$ from reported mean value. Mutation of catalytic nucleophile (E316) abrogated the activity of only the CelE catalytic domain (red color), while the E316G and E316S mutants of CBM3a tagged CelE showed ~20% pNP-release activity. (B) TLC image confirms the formation of longer chain oligosaccharides in the reaction mixture and the improved transglycosylation to hydrolysis product ratio (T/H) observed for the mutants. .....32

**Figure 9. Intermolecular interactions of exogenously added CBM3a domain along with CelE-E316G catalytic nucleophile mutant does not increase transglycosylation activity.** Here, 1 nanomoles of each purified protein was incubated along with 1.5 µmoles of pNP-Cellobiose (pNPC) and the reaction was run for 4 h at 60°C reaction temperature. Different concentrations of purified CBM3a domain alone was spiked into the reaction mixture at 0.1-10 times the relative molar ratio of the added CelE or CelE-CBM enzyme

construct. All reactions were run in duplicate with error bars representing 1σ from reported mean value. Total pNP released after 4 h reaction time is shown here. ..... 33

**Figure 10. Effect of protein concentration and substrate concentration on transglycosylation catalytic activity of CelE and CelE-CBM3a based enzyme constructs.** A) 10-to-2500 picomoles of each purified protein was incubated along with fixed 1.5 μmoles of pNP-Cellobiose (pNPC) and the reaction was run for 4 h at 60°C reaction temperature. B) 1000 picomoles of each purified protein was incubated along with 0.15-to-11.25 μmoles of pNP-Cellobiose (pNPC) and the reaction was run for 3 h at 60°C reaction temperature. Here, $K_m$ and $V_{max}$ was calculated by fitting of standard Michaelis-Menten model to data shown in (B). All reactions were run in triplicates (to measure pNP release activity) with error bars representing 1σ from reported mean value. ..... 34

**Figure 11. Biochemical analysis of the reaction mixture indicate the formation of β(1,4) glucosides. (A)** The products formed by the mutant enzymes were completely hydrolyzed when supplemented with wild type CelE enzyme indicating that the glycosidic linkage between the products is β-retaining. **(B)** Activity assays of various enzyme on substrates with different linkages (e.g., β(1,4), β(1,3) and α(1,4)) further suggested that the transglycosylation reaction products contained β(1,4) linkages. ..... 36

**Figure 12. Tandem mass spectrometric characterization using Me-FRAGS technique provided structural confirmation for the formation of cellotriose and cellotetraose as reaction products for E316G mutant of CBM3a-CelE.** Here, 2 nanomoles of each purified protein was incubated along with 1.5 μmoles of pNPC and the reaction was run for 7 h, respectively, at 60°C reaction temperature. All samples were run in duplicate. Total ion chromatograms (A) indicated the presence of glucose, cellobiose, cellotriose, and cellotetraose. ..... 37

**Figure 13. CBM3a enhances transglycosylation reaction rate of CelE nucleophile mutant by nearly two orders of magnitude unlike other CBM families. (A)** CBM1

(Type A) and CBM17 (Type B), which have similar binding affinity towards cellobiose, were shown to not significantly impact the pNP-release activity of E316G mutant of CelE. End-point kinetic assays performed by measuring the pNP released over time showed an increased activity was seen only for the CBM3a tagged E316G CelE mutant. **(B)** The specific activities of all enzymes were estimated using the kinetic data for initial activity time points to show that there was a nearly 60-fold increase in the specific activity of CBM3a tagged E316G mutant compared to the CelE-E316G control. .....39

**Figure 14. Comparing crystal structures of CBM1, CBM17, and CBM3a.** The classical cellulose binding sites/surfaces that could target binding to cellobiose (or pNP-cellobiose) are shown here. Additionally, the hydrophobic cleft in CBM3a on the opposite side of the classical binding surface is also illustrated here. The additional hydrophobic cleft is hypothesized to interact with the pNP-cellobiose substrate to facilitate transglycosylation in CelE-E316G-CBM3a. .....40

**Figure 15. CelE-E316G-CBM3a utilizes a complex multiple-step reaction cascade, based on analyzed reaction products, to efficiently convert pNPC to mixture of complex glyco-products. (A)** Hypothesized reaction cascade scheme is shown here for all major reaction products observed with wild-type and E316G nucleophile mutants of CelE-CBM3a starting with pNPC as initial substrate. Glucose as product was only seen for wild-type enzymes (CelE-Wt and CelE-CBM3a-Wt) and is therefore not shown in cascade here. The blue circles indicate the glucosyl residue and yellow circle indicates the para-nitrophenyl group. **(B)** Representative TLC analysis after Orcinol staining for duplicate reaction mixtures of CelE, CelE-E316G, CelE-CBM3a-Wt, and CelE-CBM3a-E316G with pNP-cellobiose as starting substrate for varying reaction times. .....42

**Figure 16. CelE-E316G-CBM3a utilizes a complex multiple-step reaction cascade, based on analyzed reaction products, to efficiently convert pNPC to mixture of complex glyco-products.** Detailed time course of fractional concentration profiles for

various reaction products observed in the reaction mixtures as analyzed by quantitative thin layer chromatography (TLC) are shown here. Data is provided for CelE-Wt (orange square), CelE-CBM3a-Wt (blue triangle), CelE-E316G (red circle), and CelE-E316G-CBM3a (green inverted triangle) constructs. Error bar indicates one standard deviation from reported mean values from two biological replicates. Here, y-axis denotes fractional concentration of Glucose, pNPG: pNP-glucose, cellobiose, pNPG2: pNP-cellobiose, Oligosaccharide, pNPG3: pNP-cellotriose, and pNPG4: pNP-cellotetraose. All components were estimated as pNPG2 (or pNP-CB) equivalent using pNPG2 as standard. The most abundant product for wild type enzymes was cellobiose while for CelE-E316G-CBM3a construct the transglycosylation products are mostly dominant. .....43

**Figure 17. CelE-E316G-CBM3a is an efficient transglycosidase with very high T/H ratio. (A)** Transglycosylation and hydrolysis products observed during the enzymatic reaction on pNP-CB. **(B)** The transglycosylation/hydrolysis product (T/H) ratio for the four constructs indicate that the CelE-E316G-CBM3a is a highly efficient transglycosidase with a ~140-fold higher T/H ratio as compared to wild type counterpart as the reaction proceeds to equilibrium. During the initial rate, there were no hydrolyzed products observed and only oligosaccharides-based products were seen. For these time points the T/H were not computed (indicated in green box). CelE-E316G mutant is not active and shows no transglycosylation or hydrolysis products and hence the T/H ratio was not computed (red box). Error bar indicates one standard deviation from reported mean values from biological replicates as highlighted in methods section. .....45

**Figure 18. Small angle X-ray scattering indicates the presence of dynamic flexibility between the CBM3a and CelE-E316G domains. (A)** Original SAXS profiles and **(B)** analyzed p(r) distribution profiles of CelE-E316G and CelE-E316G-CBM3a enzyme with and without pNP-cellobiose are shown here. The distance distribution function (p(r) curves) for each sample were generated from original SAXS data to estimate the real

space radius of gyration (Rg) and maximum dimension (Dmax). **(C)** Kratky plots for CelE-CBM3a-E316G mutant enzyme with and without pNP-cellobiose (pNP-CB) highlights the presence of varying degrees of flexibility in the presence of pNP-CB substrate. **(D)** Ensemble Optimization Method (EOM) analysis performed on SAXS profiles using the ATSAS package predicted an ensemble of structures with different inter-domain distances between CelE and CBM3a. A representative subset of EOM generated structures which had close interaction between the CelE (as Connolly surface model) and CBM3a (as ribbon models) domains are highlighted here. .....48

**Figure 19. Inter-domain interactions of CelE and CBM3a domains critical for observed transglycosylation activity is facilitated by optimum linker size and sequence structure**. **(A)** Biochemical assays to verify the critical role of inter-domain interactions facilitating transglycosylation activity were performed by varying the size of the linker between the CelE and CBM3a domains. A very short linker (6aa linker) showed reduced activity as compared to the original 42aa native linker. Furthermore, there appeared to be an optimal linker length (21aa) that strengthened the interactions between the two domains to further improve transglycosylation activity compared to the native 42aa linker. **(B)** Flexibility of the 42aa long linker was varied to become either highly flexible (F1 and F2) or highly rigid (R1 and R2) and the corresponding transglycosylation activities of the resultant mutants are highlighted here. The native linker of CBM3a plays a very important role in aiding the inter-domain interactions unlike other similar sized linkers of varying sequence structure. .....51

**Figure 20. Possible mechanisms facilitating retaining transglycosylation reaction in CelE-E316G-CBM3a. (A)** A potential pseudo nucleophile promoting $S_N2$ mechanism. **(B)** The aromatic residues in the active site create strain on the donor sugar and promote $S_Ni$-like mechanism .....53

**Figure 21. Impact of critical amino acid residues identified within the hypothesized CelE-E316G-CBM3a enzyme-substrate complex that impact transglycosylation activity. (A)** Active site cleft of native CelE enzyme is shown here. The catalytic nucleophile (E316 – red), catalytic acid/base (E193 – green), negative subsite (W349 – white) and positive subsites (Y270, Y273, W203 and H268 – blue) are highlighted. **(B)** Key residues within the hydrophobic cleft of CBM3a are illustrated here (R407, Y458, Y409, E521, and T509 – orange). **(C)** Relative activity of single alanine point mutations of the key identified residues of CelE and CBM3a with native nucleophile as compared to CelE-CBM3a-wt is present here as bar graphs. **(D)** Activity of the double mutants containing alanine mutation of key residues with nucleophile also mutated to glycine (E316G) are shown as relative activity compared to CelE-E31G-CBM3a. .....55

**Figure 22. MAFFT alignments of CelE and CBM3a domains using Pfam database.** The sequence logo was generated using Geneious R11 software and several major conserved residues relevant to this study are highlighted here. .....58

**Figure 23. Activity of homologous GH5 family protein (GH5-gene7) on pNP-cellobiose.** GH5-gene7 belongs to the same GH5_4 subfamily as CelE enzyme. Here, 300 pmoles of the purified GH5g7 enzyme, with or without tethered CBM3a (analogous to CelE-CBM3a constructs), were reacted with 0.8 µmoles of pNP-cellobiose (pNP-CB) for 24 h at 40°C reaction temperature. All reactions were run in triplicates with error bars representing 1$\sigma$ from reported mean value. The relative amounts of pNP released **(A)** and TLC analysis by Orcinol staining **(B)** to evaluate the products formed in each reaction mixture are shown below. The 3D homologous structure of GH5-gene7 was docked along with the CelE structure to highlight how closely the two GH5_4 catalytic domains overlap structurally **(C)**. CBM3a is docked adjacent to the CelE catalytic domain based on EOM analysis of SAXS data.....59

**Figure 24. Azide can be a gratuitous inducer for cynTSX operon.** (**A)** Native cynTSX operon is present in E. coli genome to facilitate cell growth in a cyanate-rich environment. Cyanate molecule binds to the cynR repressor protein which regulates the downstream protein expression of three essential genes (cynT, cynS, cynX). (**B)** Molecular docking of cyanate and azide ions in the binding pocket of cynR protein is shown. Since azide is a structural homolog for cyanate, it also binds tightly in the cynR binding pocket and in the same location as cyanate. ..... 69

**Figure 25. The cyn operator is inducible by azide for biosensing. (A)** Plasmid map design of pCyn-v1-GFP containing the native cynR and cyn operator region cloned upstream of a GFP reporter gene. The native version (v1) was used first to monitor GFP expression using azide and cyanate as inducers. **(B)** Flow cytometer analysis of GFP fluorescence signal confirms cyanate and azide inducible heterologous in vivo protein expression using native (BW25113-wt) and engineered ΔcynS knockout (BW25113-sCB1) E. coli strains. ..... 71

**Figure 26. Engineering GFP promoter region for improved expression.** (**A)** Illustration of the promoter regions of pCyn-v1-GFP, pCyn-v2-GFP, pCyn-v3-GFP, pCyn-v4-GFP. Fluorescence measurements of lysates from BW25113-Wt cells containing plasmids pCynv2-GFP (blue square), pCyn-v3-GFP (red triangle), and pCyn-v4-GFP (green circle) and induced with 1 mM sodium azide, or uninduced cells for 2 hours at 37°C **(B)** and for various time points **(C).** Error bars indicate one standard deviation from reported mean values from three biological replicates ..... 73

**Figure 27.Characterization of pCyn-v2-GFP. (A)** Fluorescence plot for cell lysate of BW25113-wt strain containing pCyn-v2-GFP and induced with increasing concentrations of sodium azide **(B)** E. coli cells incubated with varying concentrations of sodium azide to study the inhibitory effect of azide on cell growth. Cells were grown in LB media at 37°C for 18 hours and the OD of the cells were measured at various intermittent time points

after introduction of sodium aide at time t=0. Error bars indicate one standard deviation from reported mean values from three biological replicates. (**C)** Cell lysate GFP fluorescence from E.coli BW25113 cells induced using 1 mM sodium azide but at different starting cell optical density ($OD_{600}$). Cells were grown until the desired OD in LB media and induced with 1 mM sodium azide and grown then for 4 hours with measurements taken at 1 hour (grey square, left axis) and 4 hours (red square, right axis), respectively. The GFP fluorescence on the y-axis is normalized with the measured OD after induction for specified culture times. Error bars indicate one standard deviation from reported mean values from three biological replicates .....74

**Figure 28. cynR constitutive promoter engineering.** Illustration of the promoter regions of pCyn-v5-GFP, pCyn-v6-GFP, pCyn-v7-GFP, pCyn-v8-GFP. .....76

**Figure 29. Strain engineering and promoter design co-optimization to increase reporter GFP protein expression upon induction by azide.** Bar graphs representing GFP fluorescence intensities from cell lysates of engineered strains containing three distinct plasmid variants (v2, v5, v6, v7, v8) is shown. The cells were induced with varying amounts of azide (10 μM, 100μM, 1000μM). .....77

**Figure 30. Application of synthetic azide promoter towards heterologous protein expression and biorthogonal chemical-biology. (A)** E. coli BW25113-Wt containing pCyn-v2-GFP plasmid was induced with varying azide concentrations for 24 hours and GFP fluorescence of cell lysate was measured at various time points. **(B)** E. coli BL21 containing pEC-GFP plasmid was induced with different IPTG concentrations for 24 hours and GFP fluorescence of the cell lysate was measured at different time points. **(C)** E. coli BW25113-Wt cells with pCyn-v2-GFP plasmid was incubated with 1-azido-β-D-glucopyranosyl azide (1-Glc-$N_3$) and 2-Deoxy-2-azido-β-D-glucopyranosyl azide (2-Glc-$N_3$) and GFP fluorescence of respective cell lysates are shown as bar graphs. The inset depicts the flow cytometry analysis data for cells incubated with 1-Glc-N3 for 4 hours. Error

bar indicates one standard deviation from reported mean values from three biological replicates. ..... 79

# Chapter 1: Introduction

## 1.1 Glycans and their importance

Glycans are naturally occurring carbohydrate biomolecules that are composed of monosaccharides linked via glycosidic bonds to either proteins, lipids, or other carbohydrates to form diverse range of structures that differ subtly in their stereochemistry and regioselectivity. Glycans are ubiquitous in nature present in all the living cells. They are one of the major building blocks of life along with other biomolecules like nucleic acids (DNA/RNA), proteins, and lipids (**Figure 1A**). Glycans play an important role in regulating a plethora of biological processes required for the development, growth, and survival of an organism (**Figure 1B**). Some of the biological role of glycans include (i) protective role

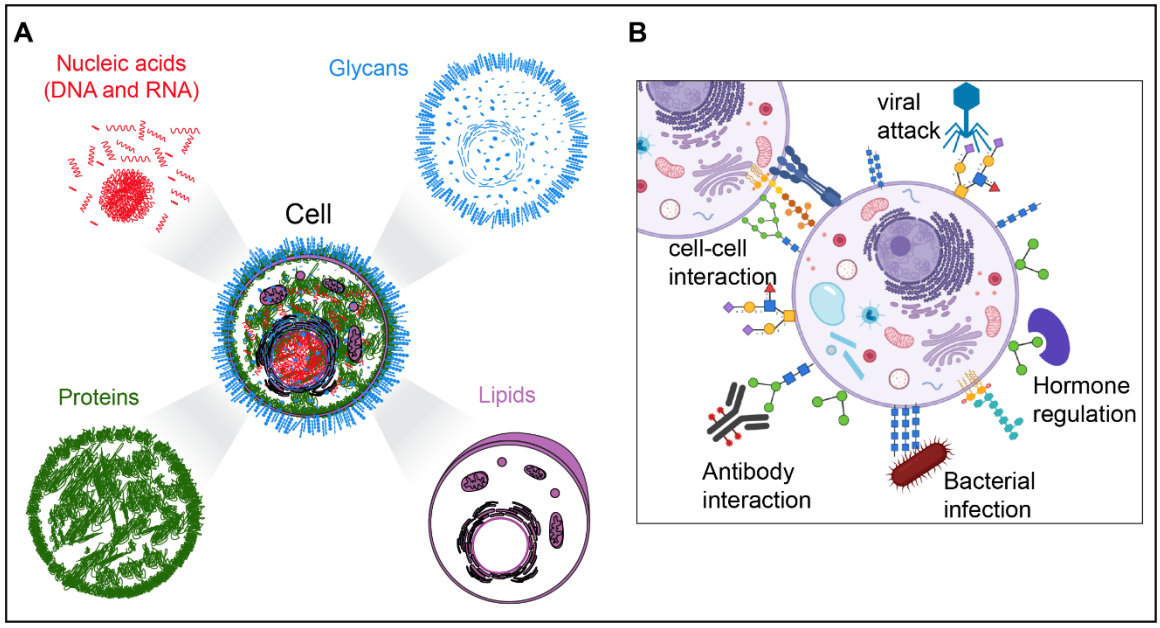



***Figure 1. Glycans are an integral part of all biological systems. (A)*** *Glycans are one of the four building blocks of life (DNA/RNA, proteins, lipids are other three).* ***(B)*** *Glycans regulate many biological process such as cell-cell interaction, antibody and hormone regulation, pathogen detection.*

on cell surface forming a physical barrier to shield the underlying biomolecules, (ii) stabilizing role for protein folding and trafficking (iii) modulatory role in enzymatic reactions and antibody recognition (iv) functioning as intrinsic ligands for cell-cell sensing/interactions (v) extrinsic ligands for symbiosis and pathogenesis [1]. Although glycans play a defining role in the design and regulation of biological process in nearly all living systems, we are far from elucidating their fundamental structure-function relationship with other biomolecules. Our understanding of glycome is very limited when compared to nucleic acids and protein research primarily due to the lack of efficient tools to synthesize them. There is a great need to develop tools that can be used by researchers to synthesize glycans reproducibly and economically in the laboratory akin to DNA or peptide synthesizers, PCR machines and in-vivo protein expression systems. Synthesis of designer oligosaccharides and glycoconjugates is critical for unraveling the structural and functional properties of glycans. This has become an urgent requirement due to the rapid increase in the application of glyco-engineering for improving drug efficacy and immunogenicity [2,3]. The glycoengineering approach to manipulate the native glycan sequence/structure provides a promising avenue to design and improve biotherapeutics particularly for time-sensitive viral diseases such coronaviruses that are often glycosylated to evade the host immune system. The recent COVID-19 pandemic caused by betacoronavirus SARS-CoV-2 emphasizes this issue as the glycans on spike protein (S-protein) shields the virus from the neutralizing antibodies. Therefore, therapeutic development efforts have been targeted towards glycoengineering these S-protein glycan shields to present α-gal epitopes that can improve anti-viral immune response [4]. The precise glycan pattern of these engineered antigens and antibodies are critical for their efficacious function and overall safety.

## 1.2 Synthesis of glycans

Glycans are synthesized naturally during cellular pathways like cell membrane synthesis, post translational modification, and metabolites production. Obtaining designer glycans from these natural sources is challenging because glycan synthesis is a template independent pathway that is carried by complex interplay of multiple enzymes which is not thoroughly understood. Chemical strategies and engineered enzymatic routes can address this challenge and synthesize desired glycans.

### 1.2.1 Chemical synthesis

Chemical synthesis offers promising solution to synthesize diverse range of oligosaccharides. It involves the glycosidic bond formation between a donor monosaccharide containing a leaving group and an acceptor molecule containing a nucleophilic group (**Figure 2**). A promoter molecule is also added to the reaction to facilitate the departure of the leaving group on donor monosaccharide. However, this deceptively simple approach is intertwined with problems of regioselectivity and stereospecificity. The promoter activated donor monosaccharide can form a glycosidic linkage with all available nucleophilic groups on the acceptor molecule leading to undesired products. A clever strategy to achieve control over glycosylation selectivity and specific involves using protection groups on the substrate sugars. The protecting groups on acceptor molecules prevents undesired linkages whereas the protecting groups on the donor molecule can control the stereochemistry of the products [5] (**Figure 2**). This strategy has been extensively used to developed automated solution-phase and solid-phase methods to synthesize various bioactive oligosaccharides such as sulfated glycosaminoglycans and human milk oligosaccharides [6–8]. However, the reaction conditions and protection/deprotection groups must be optimized for every glycoside

synthesized. A universal strategy is yet to be developed to synthesize glycans with requisite stereochemistry and rapid progress is being made on this front.

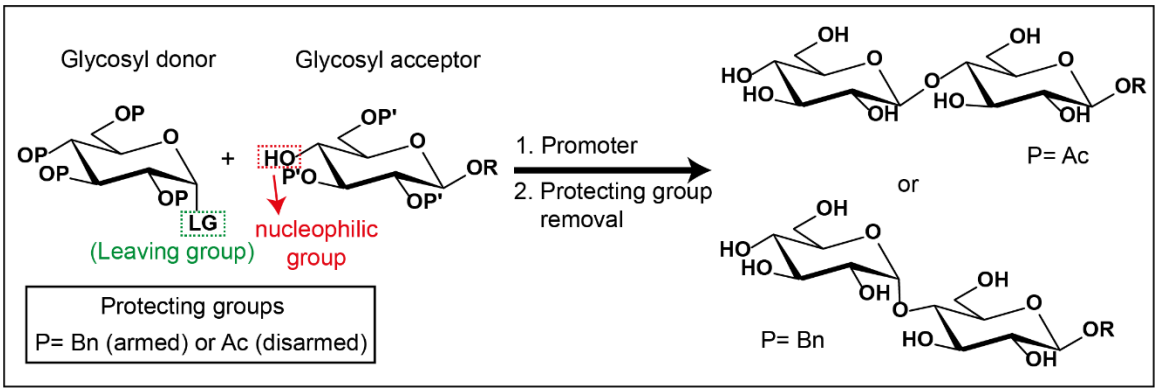



***Figure 2. Chemical synthesis of oligosaccharides using protecting groups.*** *Promoter facilitates the departure of leaving group and nucleophile group attacks the anomeric carbon. Depending on the electron withdrawing potential of protecting group (P), the product stereochemistry is controlled.*

### 1.2.2 Enzymatic synthesis

Enzymatic methods provide an economical solution to the expensive multi-step chemical synthetic procedures. Enzymes carry out reactions in aqueous solutions at higher efficiency with exquisite control over the stereochemistry of synthesized products [9]. Over the years of evolution process, Nature has leveraged a large repertoire of enzymes that can synthesize glycans. However, only a miniscule of them have been explored and characterized until today. This has limited the broad application of enzymes for producing oligosaccharides. Expanding the list of natural enzymes or engineered enzymes that can efficiently synthesize desired glycans has garnered attention from the research communities across the world.

Glycosyl transferases (GTs) are enzymes that are primarily responsible for biosynthesis of most cellular glycans and glycoconjugates. They facilitate the transfer of glycosyl moiety from nucleotide-sugar donors to either glycone or aglycone acceptor groups [10]. Unlike the synthetic procedures, glycosyl transferase enable high precision glycoside synthesis

in a single step without any protection groups. Glycosyl transferases reactions are nucleophilic substitution of the leaving group at the anomeric carbon of donor sugar. Based on the sequence similarity, GTs have been classified into 111 families in carbohydrate active enzyme (CAZy) database [11]. These enzymes once characterized can be obtained by cloning and expression in mammalian hosts. However, number of characterized GTs are very less and among the ones that are characterized, many GTs are typically membrane bound proteins that express poorly in E. coli, have limited solubility/stability, limited substrate specificity and requires recycling of expensive nucleotide donor sugar reagents that limits scaling up in-vitro synthetic routes [12]. *In-vivo* synthesis can address some of these challenges by either transferring or modifying the glycosylation biosynthetic pathways from desired eukaryotic or prokaryotic systems (like *Campylobacter jejuni* or Chinese Hamster Ovary Cells) into genetically tractable and industrially relevant expression systems [13–15]. Nevertheless, considering that *in-vivo* glycosylation is species-specific and produces a complex milieu of glycan isoforms due to the compounding presence of native carbohydrate-active enzymes and other cell growth related parameters, it is extremely challenging to synthesize designer glycans of interest in-vivo using GTs.

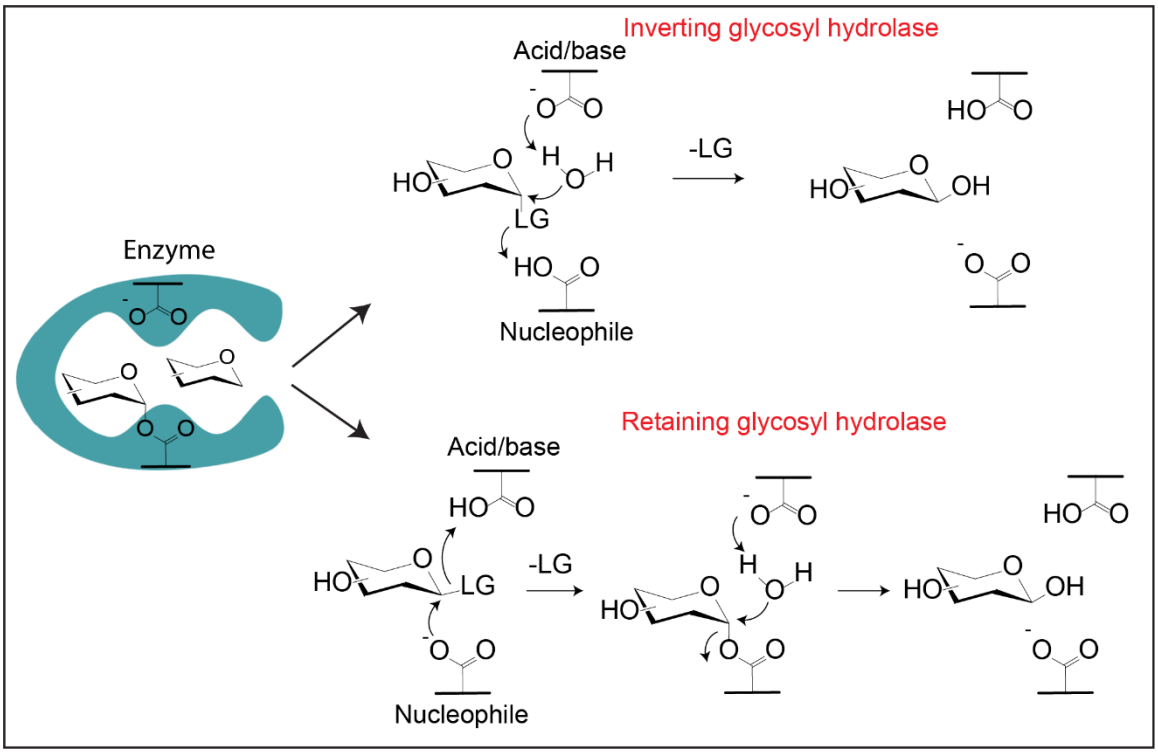



***Figure 3. Mechanism of β-glycosyl hydrolases for glycosidic bond hydrolysis.***

Another enzyme family that can be used for synthesizing oligosaccharides is Glycosyl hydrolases (GH) or glycosidases. As the name implies, glycosyl hydrolases are nature's antipode of GTs that catalyzes hydrolysis of glycosidic linkages using a retaining or inverting mechanism as illustrated in **Figure 3**. While most native GH enzymes catalyze hydrolysis, many GHs can also produce oligosaccharides following a competing transglycosylation mechanism if a suitably localized glycosyl acceptor group is present instead of a water molecule within the enzyme active site. Unlike GTs, GHs are more abundantly available in genomic databases, are often better characterized, express readily using *E. coli*, and have promiscuous substrate specificity that makes them more attractive biocatalysts to further tailor for bespoke glycans synthesis. Unfortunately, transglycosylation pathway suffers from low yields since the transglycosylation product is prone to subsequent hydrolysis by native GH hydrolytic mechanism.

All these limitations have restricted the widespread application of glycosyltransferases and glycosyl hydrolases for *in-vivo* and *in-vitro* synthesis of designer glycans. Fortunately, protein engineering of GTs and GHs can circumvent these challenges and achieve near-theoretical product yields. The engineered enzymes can be used in a chemoenzymatic process where the donor sugars can be synthesized using chemical routes and glycan polymers are synthesized using enzymes. Over the past few years, notable work has been done on rational engineering and directed evolution of glycosyl transferases to achieve (i) enhanced heterologous expression in bacterial or yeast expression system, (ii) increased enzyme half-life, (iii) thermostability, (iv) higher specific activity, (v) minimal substrate hydrolysis, and (vi) improve or alter regio- or stereo- specificity of substrate [16]. This work focuses on protein engineering strategies applied to glycosyl hydrolases to enable efficient glycan synthesis.

## 1.3 Glycosidase engineering for glycans synthesis

### 1.3.1 Sequence/structure guided rational engineering

The transglycosylation mechanism of GH enzymes provides an interesting approach to synthesize glycans. To circumvent the limitation of low transglycosylation efficiency of native glycosidases, many researchers employed process engineering where the reaction conditions and substrate concentrations are often varied to shift the equilibrium towards transglycosylation [17]. Usvalmapi et al [18] altered the reaction equilibrium of GH29 α-fucosidase derived from *Aspergillus niger* by incubating with high concentrations of lactose and fucose to synthesize 1-fucosyllactose as the major transfucosylation product. Although this approach did not require use of any activated sugars, decreasing substrate concentrations with prolonged reaction times and intrinsic hydrolysis activity of native enzyme results in poor transglycosylation efficiency. Therefore, utilizing classical reaction engineering methods alone to improve transglycosylation yield can prove to be inefficient

and is therefore coupled with protein engineering methods to design more efficient transglycosidases. Recent strides towards protein engineering strategies guided by protein sequence and structure have enabled the improvement of the transglycosylation/hydrolysis (T/H) ratio of mutant or engineered transglycosidases. The common strategies employed for engineering transglycosidases are summarized as follows (**Figure 4**); (i) active site non-catalytic residue mutations, (ii) loop engineering near active site, (iii) catalytic residues (nucleophile and acid/base) mutation.

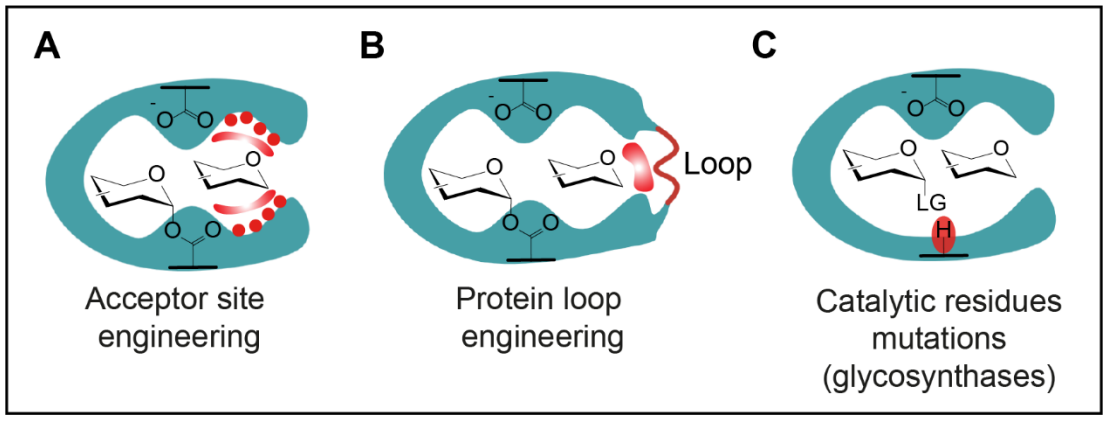



***Figure 4. Rational protein engineering strategies for improving transglycosidase activity of glycosyl hydrolases***

For CAZymes with solved structures or reliable homology models, active site modifications can be easily predicted using substrate docking and molecular dynamics simulations. Amino acid mutations that increase acceptor sugar group binding interactions can be identified to facilitate the acceptor attack of the enzyme-substrate complex. Additionally, predictions based on sequence similarity to existing homologous transglycosidases would narrow the search space for computational tools. Lundemo et al observed that mutating an asparagine residue (N220) in the acceptor site of GH1 β-glucosidase from *Thermotoga neapolitana* to more hydrophobic residues (N220F, N220W) resulted in upto 8-fold increase in the T/H ratio [19]. The asparagine residue in the wild type enzyme promoted water mediated interactions. Therefore, making the site more hydrophobic for mutant

enzymes facilitated deglycosylation using acceptor sugar hydroxyl group attack. Likewise, the active site residue (E361) of GH 42 β-galactosidase was mutated by Strazzulli et al to a smaller and less hydrophilic glycine residue to increase the transglycosylation efficiency by 177-fold [20]. In another study, Tran et al noticed that the hydrolytic activity of another GH1 transglucosidase (Os9BGlu31) enzyme was increased when Leu241 was mutated to a hydrophilic aspartic acid [21]. The observations made by Lundemo et al, Strazzulli et al and Tran et al suggests a direct relationship between the active site hydrophobicity and enzyme transglycosylation efficiency. However, the complex architecture of the enzyme active site refutes this overly simplistic prediction as seen by Tran et al where a more hydrophobic residue (W243) was replaced with a hydrophilic asparagine residue (W243N) to improve the transglycosylation efficiency of Os9BGlu31. Surprisingly, while the L241D mutation had initially increased hydrolysis activity of Os9BGlu31, this mutation also increased the transglycosylation rate when coupled with W243N by creating a positive epistasis. These non-linear effects of mutations on engineered enzyme activity makes it difficult for the researchers to use classical rational methods alone to engineer enzymes.

Nevertheless, mutations that shield the substrate-binding pocket and provide favorable interactions for acceptor/donor sugar binding are often good starting targets for engineering more efficient transglycosidases. A recent study comparing two enzymes of cycloalternan (CA) metabolic pathway in a foodborne pathogen *Listeria monocytogenes* identified key parameters that confer transferase versus hydrolase activity to these enzymes [22]. The CA forming enzyme (LmCAFE) and CA degrading enzyme (LmCADE) use similar catalytic apparatus to catalyze either synthesis or hydrolysis of the α(1,3) glucan linkages, respectively. Comparative analysis of these two enzymes revealed distinct structural features such as conformational changes in the loop near the active site and a non-catalytic loop domain that promoted acceptor sugar binding in the LmCAFE

enzyme active site. The active site loop containing a hydrophobic tryptophan residue (W430) assumes a deeper conformation in LmCAFE enzyme to shield the active site and allowed for CH-π substrate stacking. Whereas a shallow conformation of the loop in LmCADE promotes hydrolysis. A similar loop engineering approach was used by Jamek et al to improve the activity of β-N-Acetylhexosaminidase (HEX1) to synthesize lacto-N-triose II from lactose and chitobiose [23]. Sequence alignment of closely related GH20 family enzymes revealed key loop residues that were introduced in HEX1 enzyme to provide a 9-fold increase in transglycosylation activity. The work by Light et al also noted the presence of carbohydrate binding domain (CBM) in LmCAFE that provides additional non-catalytic interactions that orients and increases the effective substrate concentration near the catalytic domain for efficient transferase activity.

The presence of intact catalytic nucleophile residue hinders the transglycosylation efficiencies of engineered enzymes as the synthesized products are prone to subsequent hydrolysis, especially during longer reaction times. Employing classical reaction engineering methods to optimize biosynthesis conditions such as pH, temperature, and solvent composition can partly address these limitations. Along with engineering the active site of GH1 glucosidase, Lundemo et al [19] eliminated mutant hydrolytic activity without affecting transglycosylation activity by using high pH reaction conditions. Alternatively, since the early 2000s, a subset of engineered transglycosidases called glycosynthases (GS) have been designed by mutating the catalytic nucleophile of glycosyl hydrolase to a smaller non-nucleophilic residue such as alanine, serine, or glycine [24]. These mutant enzymes catalyze the glycosidic bond formation using modified activated donor sugars which structurally resembles the enzyme-donor sugar transition state intermediate. The resultant products are not hydrolyzed further as these glycosynthase enzymes lack a catalytic nucleophile residue to initiate the hydrolysis reaction step. This approach has

enabled engineering multiple GH family enzymes involving both retaining and inverting mechanisms to synthesize several oligosaccharides and polysaccharides [24,25]. Of late, chitinases from GH 18 [26–28] and α-galactosidases from GH 97 [29] were successfully engineered to glycosynthases by mutating their catalytic residues. Apart from the nucleophilic residues, the catalytic acid/base residues were subjected for mutagenesis to generate thioglycoligases that can synthesize thiol-containing glycoconjugates [30]. Here, catalytic acid/base mutant (E314A) of *Streptomyces plicatus* GH20 hexosaminidase utilized GlcNAc and GalNAc donors and coupled them to thio-containing acceptor groups. The general base is mutated to alanine to prevent the deprotonation of water molecule during the de-glycosylation step while allowing attack by a low pKa thio-group of the acceptor sugar. Combining the catalytic residue mutations with other active site or loop engineering approaches can be very powerful in rendering highly efficient transglycosidases, particularly with increased availability of solved glycosyl hydrolase structures.

### 1.3.2 Selective screening strategies guided randomized engineering

There are currently a vast majority of glycosyl hydrolases and transferases that are yet to be structurally or functionally characterized which has impeded the use of above-mentioned rational approaches for engineering CAZymes. Directed evolution is an alternative strategy for engineering enzymes where the parent enzyme template sequence is iteratively mutated and rapidly screened until a mutant with desired functionality is identified [31]. The crucial process in the directed evolution methodology is the screening strategy that is applied to identify the desired improved constructs from a pool of redundant or inactive variants. The degree of screening strategies ranges from low-throughput to ultra-high throughput which depends on the detection principle, sensitivity, and instrumentation. Till date, the screening strategies developed for glycosyl hydrolase can

be broadly classified into two types; (i) products based, and (ii) by-product (or donor leaving group) based. The products-based screening methodology relies on chemically tagging the donor and/or acceptor sugar with a fluorophore or suitable tag that would have a distinct physicochemical function upon glycoside product formation. The resultant product with the label can act as a trigger for transcription of a reporter gene inside the cell [32] or the label molecule can be cleaved using an enzyme specific to the formed product after which the cleaved label is detected using absorbance, colorimetric, or fluorescence measurements. Armstrong et al [33] used fluorogenic and chromogenic substrates to develop a screening strategy for identifying donor specificity and acceptor specificity for glycosynthase enzymes. These substrate-based methods are very specific to the product of interest but are highly efficient and selective in identifying improved mutants for a specific enzyme.

Alternatively, screening of glycosynthases have been done in the past by detecting the by-products of the glycosylation reaction [34]. The by-products of the glycosylation reaction are mostly the leaving groups attached to donor sugars and therefore, focus has been on developing strategies to identify these leaving groups. The most common leaving groups used for these reactions include the phosphate, fluoride, azide, pNP where each group has an exclusive detection strategy developed (**Figure 5**). The phosphate ion complexes with molybdate to form molybdenum blue that has strong absorbance at 655 nm [35]. A low-throughput 96-well plate assay was developed by Macdonald et al [35] to identify new glycoside phosphorylases from GH 94 and GH 149 by monitoring the release of inorganic phosphate based on the formation of molybdenum blue. When glycosyl fluoride is used as a donor sugar for many glycosynthase reaction it releases a fluoride anion which decreases the pH of the reaction mixture. Therefore, pH indicators such as bromocresol purple, methyl red, and bromophenol blue were used to measure the

reduction in cellular pH to quantity the intracellular expressed glycosynthase activity. For detection method of fluoride, two new chemosensor assays consisting of silyl ether of fluorogenic methylumbelliferone or chromogenic p-nitrophenol were used for engineering *Bacillus licheniformis* 1,3-1,4-β-glucanase [36] and β-glucosynthases from *Rhizobium radiobacter* and *Micrococcus antarcticus* [37].

Likewise, to detect azide ion released when using glycosyl azides as donor sugars, a click chemistry based fluorescent quenching method has been developed [38]. The basis of the screening strategy is the difference in fluorescence signal of the click chemistry product formed with released inorganic azide and fluorescent cyclooctyne as compared to click product of substrate glycosyl azides and fluorescent cyclooctyne. Recently, a p-nitrophenol (pNP) based biosensor was developed to engineer hydrolytic enzymes [39], but this system could be readily used for transglycosidase/glycosynthase screening as well. These leaving groups-based detection strategy are not substrate or enzyme specific and hence could be used applied universally for various families of glycosyl hydrolases. However, these techniques should be supported with additional product characterization for glycan synthesis since both hydrolysis and glycosylation can lead to the formation of detected by-products. In summary, both the screening methods have their respective advantages and should be used based on the enzyme template used for engineering.

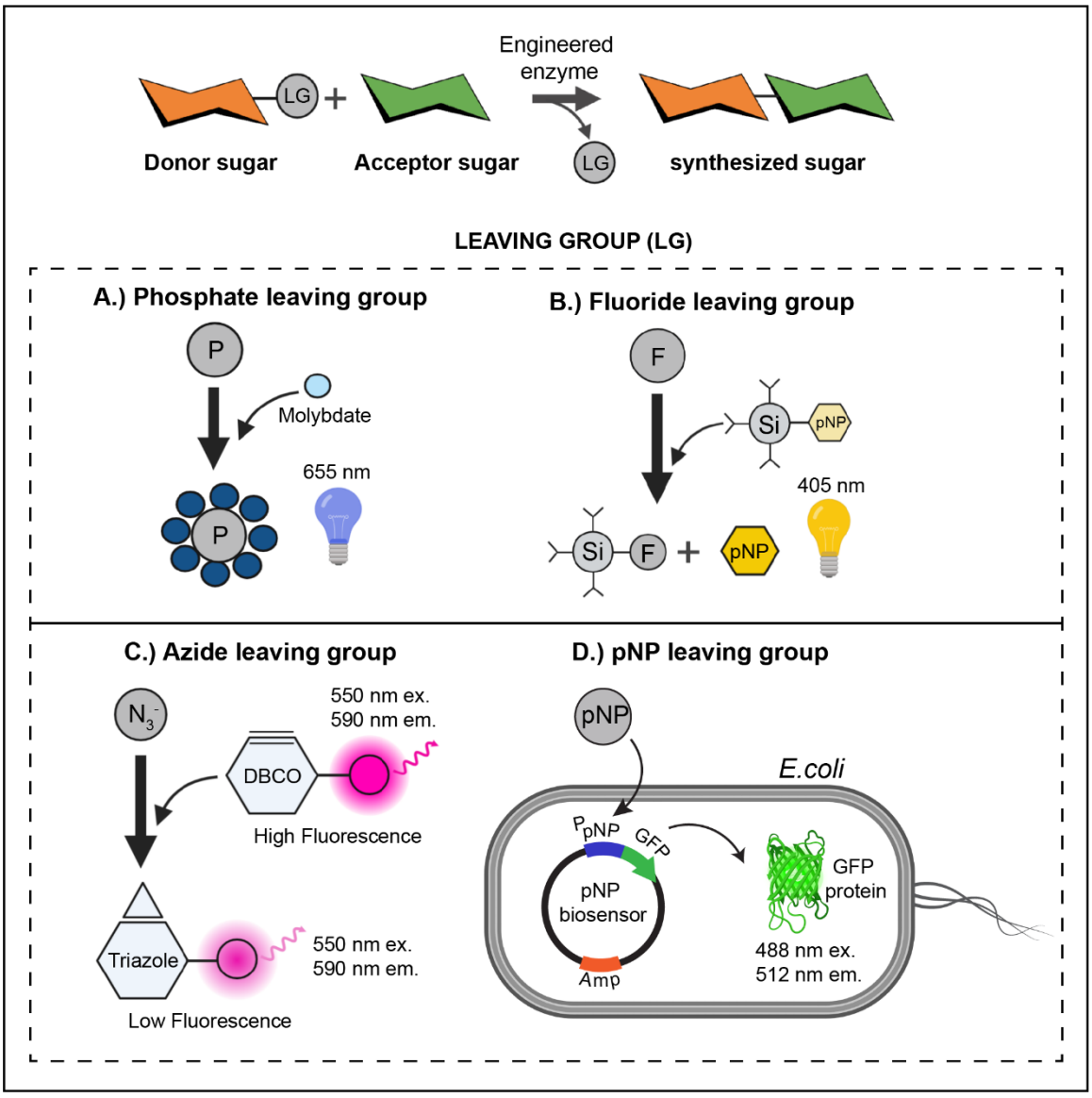



***Figure 5. Leading examples of leaving group based low to high-throughput based screening strategies to enable directed evolution of glycan synthesizing enzymes.***

## 1.4 Problem statement

Glycan synthesis mediated by engineered enzymes provide an optimistic solution to overcome current challenges of low yields and undesired products. However, there is a shortage of protein engineering tools to design and identify engineered enzymes. The current protein engineering strategies are very system specific and often fail when tried

with a new system. Novel protein strategies are required to be formulated to can be universally used across various families of enzymes.

## 1.5 Thesis Objectives

This thesis is aimed to contribute towards populating the protein engineering toolbox for efficiently engineering enzymes for glycan synthesis. The work is divided into two objectives.

**Objective 1: Sequence and structure guided engineering of GH5 family transglycosidases**

This objective focuses on developing a new rational engineering approach where a non-catalytic auxiliary domain rescues the transglycosylation activity of catalytically inactive glycosyl hydrolase enzyme. A detailed kinetic and mechanistic analysis of the enzyme is performed. Broader application of the developed strategy to other homologous enzymes is also shown.

**Objective 2: Development of high-throughput screening technique to facilitate directed evolution of glycosynthases**

This objective focuses on developing a universal cell based screening approach for directed evolution of glycosynthase enzymes. This strategy is based on identifying the azide by-product of glycosynthase reaction using an azide biosensor. An azide inducible promoter with a GFP reporter gene is designed and optimized for improved GFP expression.

# Chapter 2: Sequence and structure guided engineering of GH5 family transglycosidases

## 2.1 Introduction

Glycosyl hydrolases (GHs) are grouped into various families based on amino-acid sequence similarity,[40] currently numbering at 165+ families as curated on the Carbohydrate-Active enZyme (CAZyme) database [11,41] and still growing with newly sequenced genomes and metagenomes becoming readily available. Another important classification of enzymes of each GH family is as either a retaining or inverting enzyme sub-class, depending on whether the stereochemistry at the anomeric carbon is preserved between the reactant and product [42]. As first proposed by Koshland for glycosidases, retaining enzymes require a classical two-step double displacement hydrolysis mechanism (i.e., $S_N2$) with a glycosyl-enzyme intermediate (GEI) formed via a covalent bond between the cleaved substrate and the protein in the alternate orientation (e.g., GEI for retaining enzyme will have an $\alpha$-bond, as opposed to the $\beta$-orientation of the reactant and product) [43]. Retaining GHs that show significant transglycosylation reactivity are also thought to utilize a similar $S_N2$ type general mechanism, except that rather than nucleophilic attack by water, the attack is instigated by a hydroxyl oxygen on an acceptor sugar placed adjacent to the GEI complex (**Figure 6**). However, mutant GHs like glycosynthases (GSs) are necessary to achieve higher transglycosylation yields over native enzymes. Specifically, the retaining GH nucleophilic residue is often mutated to one that can no longer accept a proton, such as mutation from wild-type aspartate or glutamate residue to either alanine, glycine, serine, or even cysteine [44]. It is likely that the size of the side chains used in place of the wild-type nucleophile is an important factor in proper

substrate positioning and impacts the choice of activated donor sugar, but these structure-function relationships are not well understood.

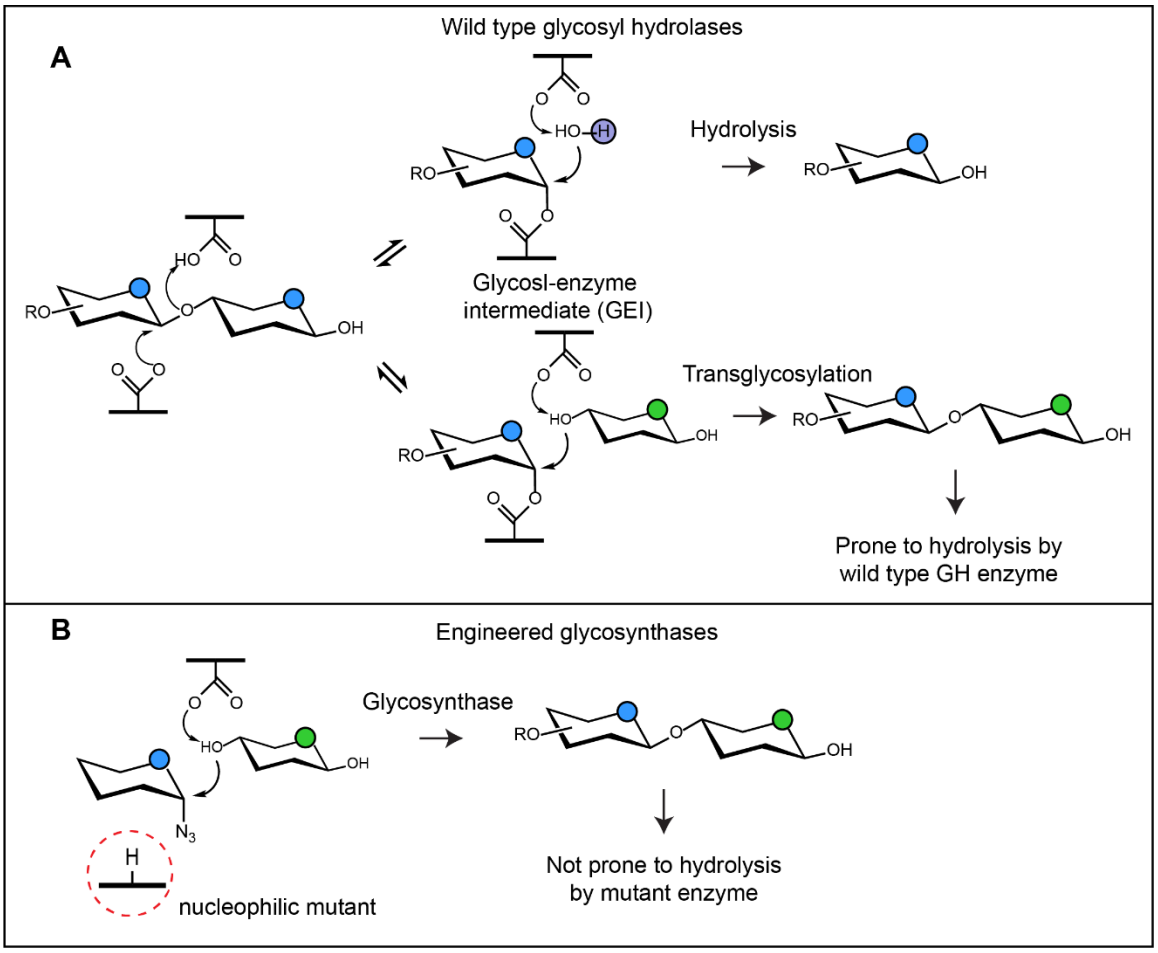



***Figure 6. Overview of generic $S_N2$ based two-step mechanism employed by retaining glycosyl hydrolases and corresponding nucleophile mutants (e.g., glycosynthases) to facilitate both hydrolysis and transglycosylation type reactions.*** *Here,* ***(A)*** *depicts the glycosylation step initiated by the enzyme catalytic nucleophile residue that results in the formation of a Glucosyl-Enzyme Intermediate (GEI). In the presence of water acting as a nucleophile, hydrolysis (and subsequent deglycosylation of the GEI) takes place for most native wild-type glycosidases. In the presence of a suitable glycosyl acceptor group, transglycosylation can take place as well for some native or engineered glycosidases.* ***(B)*** *Mutations at the nucleophile site to develop engineered transglycosidases mutants, like glycosynthases, can often facilitate efficient synthesis of glycans like oligosaccharides using simple activated donor sugars that mimic the GEI conformation.*

In addition, the role of other residues within the active site or its vicinity on the mutant GH or GS activity and the actual reaction mechanism followed is even less understood. To perform glycan synthesis, GSs or mutant transglycosidases are supplied with activated

donor sugars whose structures mimic the GEI. Like the native GEI, these donor sugars have the opposite stereochemistry at the anomeric center from the native reactant, and the anomeric carbon is bonded to a good electrophilic leaving group [44]. Nucleophilic-site mutations introduced into retaining GHs to create more efficient transglycosidases like GSs were thought originally to not impact the $S_N2$ reaction mechanism. However, a seminal study by Davis and co-workers has challenged this paradigm showing that a nucleophile mutant of a GH family 1 β-glycosidase follows an $S_Ni$-like mechanism to synthesize β-glycosides in the presence of activated β-donors like p-nitrophenyl (pNP) based glycosides [45]. Currently there are no other reports in the literature of β-retaining enzymes that follow a front-facing $S_Ni$-like mechanism. It is also unclear if other CAZyme domains found often associated with GHs would impact such front-facing reaction mechanism in engineered β-retaining GH enzymes, analogous to the critical role of lectin-like domains identified in the case of multidomain GTs that perform the glycosyl transfer step with nucleotide donor sugars using a classical $S_Ni$ type mechanism [46].

GHs are often found naturally associated with non-catalytic auxiliary domains, like CBMs, that specifically recognize and bind to carbohydrates. Ever since the identification and isolation of the first CBM from *Trichoderma reesei* over three decades ago,[47] multiple studies have identified a tremendous variety of CBMs with defined structures and functions found associated with CAZymes isolated from diverse ecosystems [48]. CBMs are thought to increase the catalytic efficiency of CAZymes such as GHs, GTs, and other carbohydrate-modifying enzymes like carbohydrate esterases (CEs) by mostly overcoming substrate diffusion limitations. CBMs have been shown to particularly increase the catalytic activities of cellulases and hemicellulases on various insoluble polysaccharides like cellulose and arabinoxylan found in plant biomass mostly via increasing the effective substrate concentration near the catalytic domain [48]. CBMs have

also been hypothesized to disrupt or modify the polysaccharide structure to improve catalytic activity via unknown non-hydrolytic mechanisms,[49] but this latter hypothesis is yet to be conclusively tested. While removal of CBMs from full-length GHs has been shown to cause significant decrease in activity towards mostly larger polysaccharides,[50–52] CBMs are not thought to impact the catalytic domain activity towards soluble substrates like model pNP-glycosides or smaller oligosaccharides. For example, proteolytic cleavage of CBM from *T. reesei* full-length cellulases dramatically reduced the specific activity of proteolyzed enzymes towards insoluble microcrystalline cellulose, but showed no difference in activity compared to full-length enzymes acting on activated soluble substrates like 4-Methylumbelliferyl-β-cellotrioside [53]. A recent study showed that appending a multifunctional GH 5 family enzyme, with activity on cellulose/xylan/mannan, to different CBMs that specifically targets each polysaccharide can improve overall activity towards each respective substrate [54]. This study also indicated the significance of the relative orientation of the two domains and the linker peptide properties on the full-length CAZyme activity. In other non-hydrolytic polysaccharide degrading enzymes, such as lytic polysaccharide monooxygenases (LPMO), the CBMs also play a critical role in orienting the catalytic modules towards their respective substrates, but non-specific CBM-driven binding can sometimes result in lowering overall activity as well. For example, while several CBMs have been shown to improve the activity of lytic polysaccharide monooxygenases, some have had a deleterious effect on LPMO activity as well [55]. A similar deleterious effect of CBM on xylan esterase activity was also observed where the catalytic domain alone had higher specific activity than the native full-length enzyme [56]. In summary, CBMs actively modulate the hydrolytic activity of glycosyl hydrolases as well as several classes of CAZymes. On the other hand, the effect of CBMs on transglycosylation activity of GHs or GSs is still not very well understood.

Since the transglycosylation reaction step involves nucleophilic attack by an acceptor sugar moiety on the GEI, the presence of a CBM close to the product binding active site could in principle increase (or decrease) transglycosylation efficiency by providing additional substrate-binding interactions. However, a very limited number of studies have explored the role of CBMs on transglycosylation activity of GHs/GSs and none have so far explored the possibility of the reaction mechanism directly employing the CBM. For example, CBM14 was found to provide an extended product binding site for GH18 chitinase that facilitated the transglycosylation driven synthesis of longer chito-oligosaccharides likely via an improved substrate-diffusion analogous mechanism as seen for native CAZymes [57]. On the other hand, the presence of CBM near the product binding site could sometimes be detrimental to transglycosidase activity due to steric hindrance as was observed in the case of CBM32 appended to C-terminus of GH5 mannanase [58]. In the latter case, it was hypothesized that CBM32 sterically hinders the attack of the intermediate GEI complex by manno-oligosaccharide acceptor sugars thereby promoting higher hydrolysis yields. In both these studies, the purported role of CBMs on transglycosylation was hypothesized but no detailed mechanistic analysis was performed to systematically test these hypotheses. In another instance, a CBM11 was appended to the GS mutant of β(1,3 -1,4)-glucanase belonging to GH16 for synthesis of mixed-linkage glucans [59]. However, while the presence of CBM11 increased the degree of polymerization of insoluble products formed, the overall catalytic activity of the CBM-GS chimera was found to be lower than the GS catalytic domain alone. The slower reaction rates were hypothesized to enable the CBM to dissociate and then re-associate the intermediate products in the active site in a timely manner leading to the formation of longer polysaccharides. However, the conformational dynamics of the CBM and GS/GH domain interactions would be specific to the properties of each protein domain and it is

still unclear how exactly the mechanical properties of the linker would impact the inter-domain interactions to facilitate the transglycosylation step. In summary, while these preliminary studies suggest that CBMs can impact transglycosylation reactions, we are still far from fully understanding the actual mechanisms utilized by multi-domain CBM-based CAZymes to facilitate synthesis of bespoke glycans.

Here, we report the influence of CBMs on the transglycosylation activity of a β-retaining chimeric transglycosidase enzyme design for the synthesis of β-1,4-glucan oligosaccharides (i.e., cellodextrins) that surprisingly followed a poorly understood synthase mechanism even after mutation of the nucleophilic active site instead of the classical $S_N2$-type mechanism seen for the native enzyme. The native GH scaffold chosen for this study belonged to a native β-retaining family 5 cellulase, called CelE from a well-known CBP microbe *Clostridium thermocellum*, with characteristic catalytic nucleophile and acid/base residues. Unsurprisingly, the nucleophile site mutation (to Alanine) of CelE's CD domain alone abrogated enzyme activity on soluble substrates like pNP-β-D-cellobiose in line with the expected role of a true nucleophile residue on the catalytic turnover of a β-retaining enzyme following an $S_N2$ type mechanism. However, fusion of these catalytically-inactive CelE nucleophile mutants to a *C. thermocellum* CBM3a domain associated with its native linker converted the inactive CD into an active transglycosidases that produced cellodextrin based glycosylation products. These studies also revealed, for the first-time, the unexpected involvement of CBMs and optimum linker domain in the actual catalytic reaction step of chimeric CAZymes. Through detailed biochemical characterization of several engineered enzyme constructs, along with complementary structural analyses, we also provide preliminary supporting evidence for previously unknown competing reaction mechanisms (e.g., $S_Ni$ vs. $S_N2$ mechanism) utilized by such CAZymes.

## 2.2 Experimental outline and design

A carbohydrate binding domain (CBM3a) was found to recover the transglycosylation activity of inactive CelE-E316G mutant enzyme using either $S_N2$ or $S_Ni$-like mechanism. The following experimental plan led us to this conclusion.

Step 1: Nucleophile mutations of CelE enzyme was created and tested for glycosynthase (GS) activity. But no GS activity was observed.

Step 2: The mutant CelE domains were fused with various carbohydrate binding domains, CBM3a, CBM1 and CBM17 and we serendipitously found that CBM3a fused CelE-E316G showed transglycosylation (TG) activity.

Step 3: To understand why and how CBM3a improves TG activity of CelE-E316G, we modified linker region and mutated multiple critical residues of CelE and CBM3a. These revealed that the mechanism could be either $S_N2$ or $S_Ni$-like mechanism.

## 2.3 Methods

### 2.3.1 Generation of plasmid DNA constructs

The wild type plasmids (pEC-CelE and pEC-CelE-CBM3a) were gifted by Dr. Brian G. Fox (University of Wisconsin Madison). The single point or double point mutations in the catalytic domain (CelE) and carbohydrate binding domain (CBM3a) genes were introduced using site directed mutagenesis (SDM, **Appendix 1.1.1**). Sequence and Ligation-Independent Cloning (SLIC) protocol was used to modify (insert or delete) large DNA sequences (**Appendix 1.1.2**). The plasmid DNA constructs used in this study are tabulated in **Table 1**.

***Table 1. Major plasmids used in the study and cloning techniques used to generate respective plamsid***

| Plasmid DNA | Cloning technique |
|---|---|
| pEC-CelE | Donated |
| pEC-CelE-CBM3a | Donated |
| pEC-CelE-E316G | SDM |
| pEC-CelE-E316G-CBM3a | SDM |
| pEC-CelE-CBM17 | SLIC |
| pEC-CelE-CBM1 | SLIC |
| pEC-CelE-CBM3a-21aa | SLIC |
| pEC-CelE-CBM3a-11aa | SLIC |
| pEC-CelE-CBM3a-6aa | SLIC |
| pEC-CelE-CBM3a-F1 | SLIC |
| pEC-CelE-CBM3a-F2 | SLIC |
| pEC-CelE-CBM3a-R1 | SLIC |
| pEC-CelE-CBM3a-R2 | SLIC |

### 2.3.2 Protein expression and purification

The successfully cloned and sequence verified plasmids were transformed into *E. coli* BL21 cells for protein expression. Transformed colonies were inoculated into 25 ml of LB media supplemented with kanamycin (50 µg/ml) and grown at 37°C for 16 hrs. The 25 ml of overnight grown culture was transferred to 500 ml of fresh LB media with kanamycin and incubated at 37°C until mid-exponential phase ($OD_{600}$ 0.4-0.6) was reached. At this point, protein expression was induced by adding either 0.5 mM IPTG and protein induction was carried out at either 37°C for 4 - 6 hrs or 25°C for 20 - 24 hrs. The cells were harvested

by centrifugation at 8000 g for 15 mins at 4°C and the cell pellets were stored in -80°C until purification process. The cell pellet was re-suspended in cell lysis buffer (20 mM sodium phosphate, 500 mM NaCl, 20% glycerol; pH 7.4) supplemented with lysozyme (10 µg/ml) and protease inhibitor cocktail (EDTA, E-64 and benzamidine). The cells were lysed by sonication at 4°C for 5 min with 10-s on-bursts and 30-s off periods. After sonication, the cell debris was separated by centrifuging at 12000 g for 60 min at 4°C and the supernatant containing the soluble protein of interest was collected and purified by immobilized metal affinity chromatography (IMAC). The affinity purification was carried out using one of the following methods (i) Bio-rad NGC chromatography system with $Ni^{+2}$-NTA based column (GE Healthcare) (ii) Magnetic purification using Ni-charged MagBeads (Genscript). The desired protein eluted in 100% IMAC B buffer was buffer exchanged into storage buffer (10 mM MES pH 6.5) using PD-10 desalting columns (GE Healthcare). The purified protein was finally aliquoted into 100 µl and 1 ml aliquots and flash frozen in liquid $N_2$ and stored at -80°C.

### 2.3.3 Protein characterization

The eluted proteins were characterized for their purity and concentration. The protein concentration was estimated by measuring absorbance at 280 nm in a spectrophotometer. Briefly, 4 µl of protein was spotted on a SpectraDrop Micro-Volume Microplate and covered with the top slides. The absorbance ($A_{280nm}$) values were measured and Beer-Lambert law ($A=\varepsilon LC$) was used to calculate the protein concentration where the path length (L) was 0.1 cm and the extinction coefficient (ε) was obtained from Geneious R10.1 bioinformatics software.

The molecular weight (M.W.) and purity of all the enzymes was confirmed by SDS-PAGE based densitometric analysis using 4-20% gradient polyacrylamide gels. Samples were

prepared by diluting 1-part protein with 1-part 2X Laemmli buffer (i.e., 95% Laemmli sample buffer + 5% β-mercaptoethanol) and denatured at 95°C for 5 minutes. Then, 10-20 µl of the denatured sample was loaded alongside protein ladder and the gel was run at 200 V for 30-40 minutes. The gel was stained in Coomassie blue staining solution (50% ethanol, 40% water, 10% acetic acid, 0.1% Coomassie brilliant blue R 250) by microwaving for 45 seconds and incubating at room temperature for 10 minutes. The gel destaining was performed in destaining buffer (50% water, 40% methanol and 10% acetic acid) by microwaving for 45 seconds and incubating at room temperature for 10 minutes, followed by overnight destaining in DI water. The destained gel was imaged using Gel Doc EZ Imager and image analysis was done using Image Lab software. The purified protein molecular weight was estimated by comparing with the bands in protein ladder and matched to predicted translational products using bioinformatics software. In the samples where multiple bands were observed, the densitometric analysis was applied to determine the purity of the sample based on the band intensities.

### 2.3.4 End point enzyme activity assay

The activity assays of CelE-CBM3a proteins on 4-nitrophenyl-β-D-cellobiose (pNP-cellobiose or pNP-CB) were performed in 200 µl reaction volume in three replicate reactions. 500 pmol of protein was reacted with 2 µmoles of pNP-CB in 50 mM MES pH 6.5 buffer at 60 °C for 7 hours with shaking at 400 rpm. 10 µl of reaction mixture was taken at 3 hr (or 4 hr) and 7 hr and added to a clear flat bottom microplate containing 100 µl of 0.1M NaOH and 90 µl of DI water. The absorbance of the microplate at 405 nm was measured in a spectrophotometer and pNP amounts in each well was calculated with the help of a calibration curve built for pNP standards and $\lambda_{405}$ absorbance. Thin layer chromatography (TLC) analysis was also performed at every sampled interval. From the reaction mixture, 1 µl was added to the silica coated glass TLC plates and run in a closed

TLC chamber saturated with the mobile phase. The mobile phase used was ethyl acetate: 2-propanol: acetic acid: water (at 3:2:1:1 v/v ratios). Known standards such as pNP, pNP-cellobiose, glucose and cellobiose were also run to determine the unknown detected spots in reaction sample based on retention factor ($R_f$) value. The plate was epi-illuminated and directly imaged under UV light at wavelength λ=305 nm to visualize pNP and pNP-containing compounds. The plates were then immediately sprayed with visualization solution containing 0.1% orcinol dye in 10% $H_2SO_4$ in ethyl alcohol, then dried and heated at 100 °C for 15 min to visualize reducing sugars and acid-labile sugars. The plates were then imaged GelDoc EZ imager (BioRad) and analyzed using Image Lab software.

### 2.3.5 Time course enzyme kinetic assay

Kinetic activity of purified enzymes on pNP-cellobiose was performed by measuring the substrate and product concentrations at various time points. The 150 µl reaction mixture consisted of 1.5 µmoles of pNP-CB and 100 pmoles of protein in MES buffer (50mM, pH 6.5). Three replicates of individual reaction mixtures for every time point was setup in 200 µl PCR tubes in the following order: (i) 70 µl of DI water (ii) 30 µl of 250 µM MES buffer (pH 6.5) (iii) 30 ul of 50 mM pNP-CB (iv) 20 ul of 5 pmol/µl of protein was added using multi-channel pipettes. Protein sub-stocks with 5 pmol/µl concentration were prepared by diluting the protein stocks in 10 mM MES (pH 6.5). The tubes were kept at room temperature while adding the reagents. The PCR tubes were immediately transferred to thermomixers (preheated to 60°C) after the addition of proteins. For each time point, PCR tubes were recovered and the reaction was quenched by denaturing the protein at 95°C for 5 minutes. The denatured tubes were centrifuged and 10 µl of supernatant was added to microplate well containing 100 µl of 0.1M NaOH and 90 µl of DI water to measure the pNP absorbance at 405 nm. Simultaneously, 8 pNP standards with known concentration between 0.25 mM and 10 mM were used to build the calibration curve between pNP

concentration and $\lambda_{405}$ absorbance. The unknown pNP absorbance was estimated using the calibration curve. Samples were collected for 8 proteins and 1 reaction blank for 17 time points spanning from 0 hr to 42 hrs. The concentrations of other reactants and products in the reaction mixture was measured using quantitative thin layer chromatography.

### 2.3.6 Quantitative Thin Layer Chromatography (TLC) analysis

Reaction mixtures were analyzed by TLC using Analtech P21521 Silica Gel GHLF TLC plates for quantitative analysis of all hydrolysis and transglycosylation reaction products. The mobile phase used for TLC was ethyl acetate: 2-propanol: acetic acid: water (at 3:2:1:1 v/v ratios). 1 µl of the reaction mixture from each time point was spotted on the TLC plates. Several standards were also run on the TLC plate to determine the unknown detected spots in reaction sample based on retention factor ($R_f$) value. In every TLC plate, one lane was left empty for spotting quantification standards after the TLC development. The TLC plate was kept in the dessicator until the spots were completely dried after which they were placed in the TLC chamber (pre-saturated with mobile phase). After the TLC plate was developed (mobile phase reached the top), the plate was recovered and air dried. In the empty lane, pNP-cellobiose standards of known concentrations ranging from 0.25 mM to 10 mM were spotted. The plate was epi-illuminated and directly imaged under UV light at wavelength λ=305 nm to visualize pNP and pNP-containing compounds. The plates were then immediately sprayed with visualization solution containing 0.1% orcinol dye in 10% $H_2SO_4$ in ethyl alcohol, then air-dried and heated at 100 °C for 15 min to visualize reducing sugars and acid-labile sugars. The plates were then imaged GelDoc EZ imager (BioRad) and spot intensity was quantified using GelDoc EZ imaging software. For quantitative analysis of kinetic assay reaction mixtures, a standard curve was built using the pNP-CB standards and respective standard spot intensities. The standard curve

was then used to estimate the absolute concentration (as pNP-CB equivalents) of all other hydrolysis or transglycosylation products seen on the TLC plate (e.g., glucose, pNP-glucose, pNP-cellotriose, pNP-cellotetraose, cellobiose, cellotriose, cellotetraose, and cellopentaose). Fractional concentrations for each product reported here was normalized to the total observed residual substrate and formed products concentration for each reaction mixture. Final results reported here for each reaction condition was based on two biological replicates.

### 2.3.7 Small angle X-ray scattering (SAXS) analysis

Purified protein samples for SAXS analysis were diluted in 50 mM MES pH 6.5 buffer to get a final concentration of 1 mg/ml and 2.5 mg/ml. The reaction samples were prepared to consist of protein at 1 mg/ml and 2.5 mg/ml and substrate (pNP-CB) at 10 mM as final concentration in 50 mM MES pH 6.5 buffer. All the samples mixtures were prepared on ice and instantly frozen using liquid nitrogen and shipped on dry ice to Oak Ridge National Laboratory (ORNL) for SAXS measurements. SAXS experiments were carried out on a Rigaku BioSAXS 2000 instrument equipped with a Pilatus 100K detector (Rigaku Americas) at Oak Ridge National Laboratory (ORNL). Silver behenate were used to calibrate for sample-to-detector distance as well as beam center. Sample volumes of ~80-100 µl was loaded into a Julabo temperature-controlled 96-well plate. An automatic sample loader pumped samples from the 96-well plate into the Julabo temperature-controlled flow cell for measurement. The 2D images obtained were reduced using instrument reduction software to 1D curves, $I(Q)$ vs. $Q$, where $Q$ the wave-vector is a function of scattering angle, $2\theta$ and wavelength, $\lambda$. An accompanying buffer was measured for each sample and subsequently, the buffer background was subtracted. The resulting scattering profile representative of the sample was used for data analysis. The pair distance distribution function P(r) of the protein was calculated from the indirect

Fourier transform of *I(Q)* and further a low-resolution volume envelop using a dummy atom model was obtained using the ATSAS package, GNOM and DAMMIN, respectively [60]. Ensemble optimization method (EOM) module of ATSAS package was used to generate the model structures. The crystal structures of CelE (PDB ID: 4IM4) and CBM3a (PDB ID: 1NBC) were used as input structures for EOM analysis along with the experimental scattering data.

## 2.4 Results and discussion

### 2.4.1 Preliminary efforts to engineer CelE into a glycosynthase

The catalytic nucleophile (E316) and catalytic acid/base residue (E193) of CelE enzyme was identified from the UniProt database (UniProt ID: P10477). In order to engineer CelE into glycosynthase, the catalytic nucleophile was mutated to an Alanine (E316A), Serine (E316S), or Glycine (E316G) residue. The resultant CelE constructs were cloned into a T5 promoter plasmid with N-terminal his tag and transformed into *E.coli* expression cells for protein expression and purification. A total amount of 50-100 mg protein with high purity (>90%) was obtained from 1 L of cell culture (**Figure 7**). The purified nucleophile mutant proteins were tested for recovered hydrolytic activity on pNP-β-D-cellobiose (soluble) substrates in the presence of exogenously added chemical rescue agents such as azide and formate ions. The reactivation experiments revealed that hydrolytic activity for none of the nucleophilic mutants of CelE could be chemically rescued by either azide or formate (**Figure 7**). These results suggest that it would be likely not possible to form a true glycosynthase from CelE, based on previous reports that suggest a strong correlation between chemical rescue and glycosynthase activity [61]. Nevertheless, CelE nucleophilic mutants were tested for their glycosynthase activity using α-cellobiosyl fluoride as donor and cellobiose as acceptor (**Figure 7**). Orcinol dye stained TLC analysis of the reaction

mixtures did not show any hydrolyzed or higher molecular weight products confirming that the CelE-E316A/S/G mutants are not glycosynthase mutants.

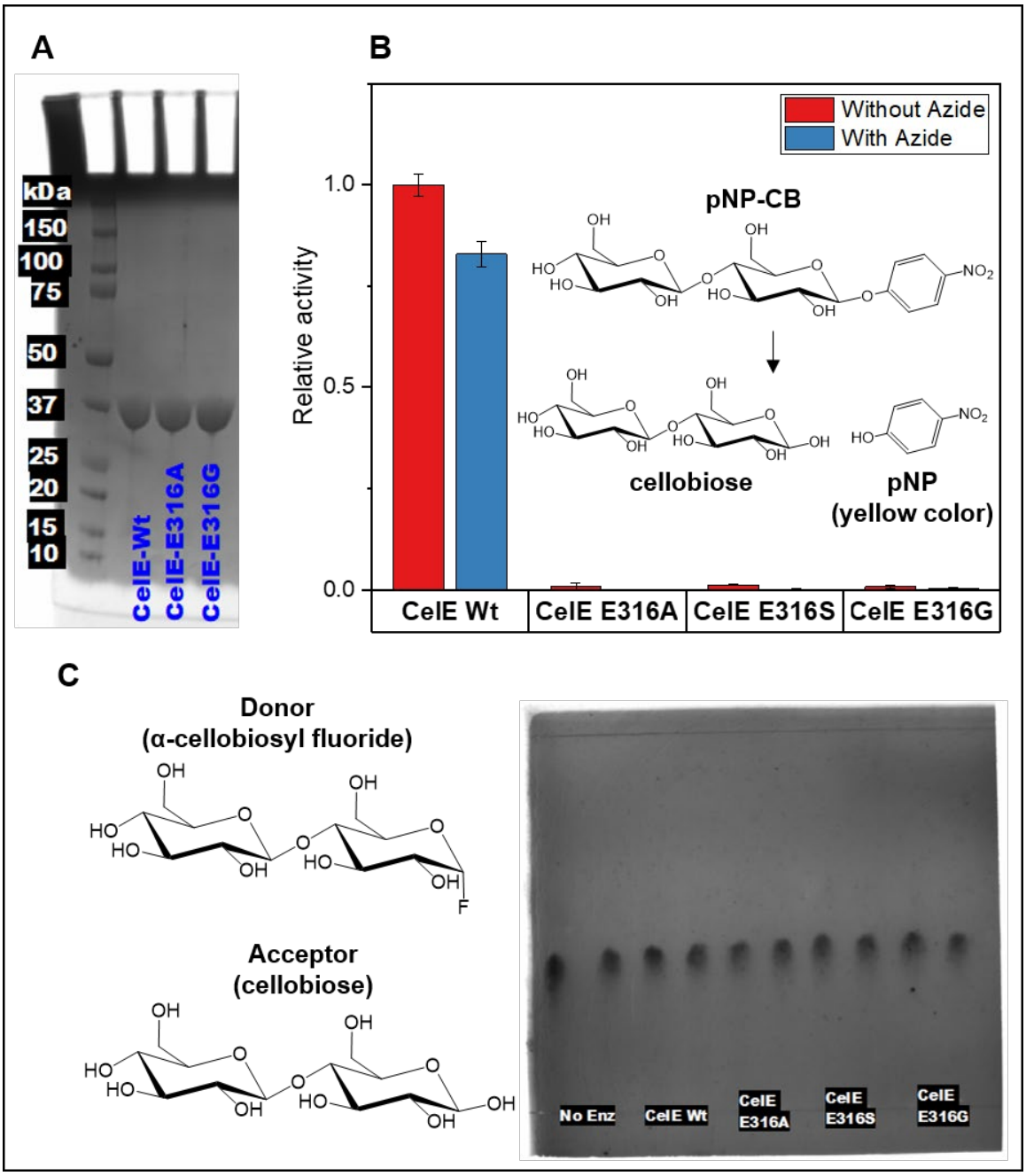



***Figure 7. Nucleophilic mutations of CelE did not yield an active glycosynthase. (A)*** *SDS-PAGE protein characterization for molecular weight and purity. (B) Chemical rescue activity assay of CelE and its nucleophilic mutants on pNP-CB in the presence of 1M sodium azide. (C) Glycosynthase reaction testing using CelE nucleophile mutants on α-cellobiosyl fluoride and cellobiose sugars. The two sugars have similar retention factor (Rf) on TLC plate.*

### 2.4.2 CBM3a recovers CelE nucleophile mutants transglycosylation and hydrolytic activity

The wild type CelE and the nucleophilic mutants were then fused on the C-terminus to a carbohydrate binding domain, CBM3a from *C. thermocellum* using a 42 amino acid flexible linker from the *C. thermocellum* scaffoldin protein CipA (Cthe_3077). Similar to the CelE mutants, the purified CelE-CBM3a proteins also did not show any rescue of hydrolytic activity in the presence of azide and formate ions (**Figure 8A**). However, surprisingly, we observed that the serine and glycine mutants of CelE-CBM3a nucleophile mutants showed significant release of the pNP leaving group upon long periods of incubation with pNP-CB even without any exogenous nucleophiles (**Figure 8A**). The orcinol stained TLC analysis of the reaction mixtures indicated the formation of cello-oligosaccharides with a degree of polymerization or DP≥3 which are the transglycosylation products. These reaction products were not observed when the non-tethered CelE-E316G and CelE-E316S mutants were reacted with pNP-CB. The CBM3a domain fused to the CelE-E316S/G nucleophile mutants significantly ($p<0.001$, based on pNP release data shown in **Figure 8A**) improved hydrolytic and/or transglycosylation activity of the inactive catalytic domains towards pNP-CB substrate, compared to CelE-E316S/G domains alone. This was a surprising result because type A CBM domains like CBM3a are generally thought to increase the relative substrate concentrations of insoluble substrates and have less/no effect on soluble substrates. Hence, to identify the mechanism through which CBM3a rescues activity, a series of experiments were performed.

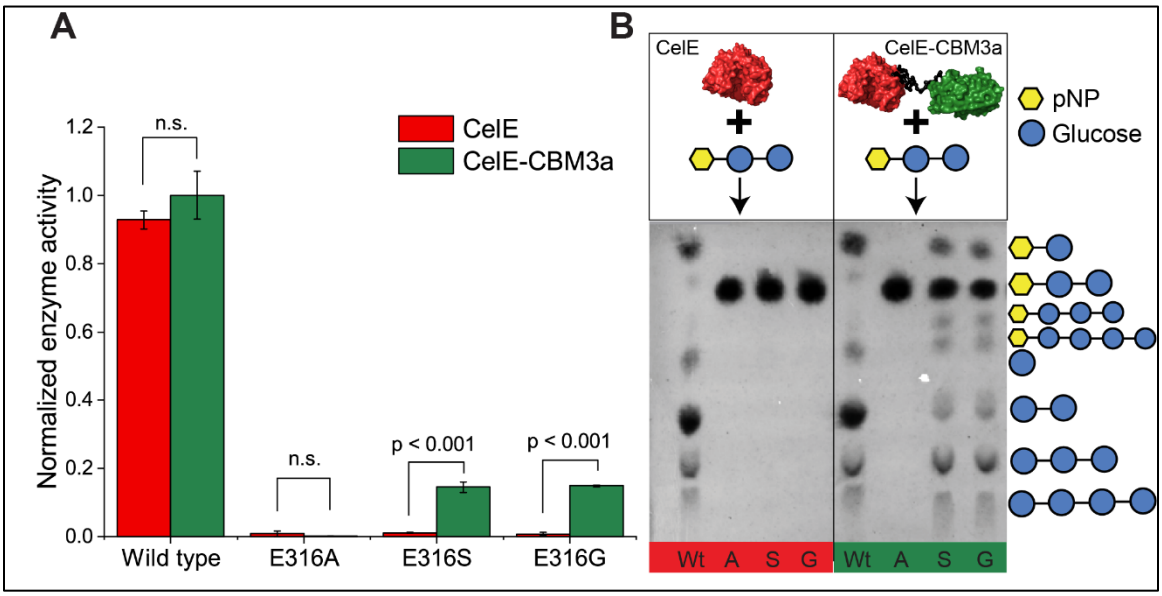



***Figure 8. Carbohydrate binding modules (CBMs) aid in recovery of transglycosylation activity of nucleophilic mutants belonging to a family 5 glycosyl hydrolase (CelE).*** *(A) The relative enzyme activities were estimated by measuring the release of pNP from pNP-Cellobiose (pNPC) as starting substrate. Here, 500 picomoles of each purified protein was incubated along with 2 µmoles of pNPC and the reaction was run for 4.5 h at 60°C reaction temperature. All reactions were run in triplicates with error bars representing 1$\sigma$ from reported mean value. Mutation of catalytic nucleophile (E316) abrogated the activity of only the CelE catalytic domain (red color), while the E316G and E316S mutants of CBM3a tagged CelE showed ~20% pNP-release activity. (B) TLC image confirms the formation of longer chain oligosaccharides in the reaction mixture and the improved transglycosylation to hydrolysis product ratio (T/H) observed for the mutants.*

Firstly, to test if CBM3a plays a catalytic role or non-specific interactions between CBM3a and CelE domains results in the observed activity, exogenous CBM3a was added to the reaction mixtures at varying ratios with respect to catalytic enzyme. (**Figure 9**). Addition of exogenous CBM3a to CelE-E316G did not show any impact on the transglycosylation activity (**Figure 9**), which suggests that the subtly-timed and highly dynamic inter-domain interactions for the full-length chimeric E316G nucleophile mutants of CelE-CBM3a is likely very critical to recover transglycosylation activity. Also, increasing the added enzyme concentration over >2 orders of magnitude or the substrate concentration in the reaction wells did not have a significant effect on the pNP-release activity of CelE-E316G unlike the CBM3a tethered CelE-E316G construct (**Figure 10**). This suggested that the pNP

release rate for CelE-E316G was not limited by the relative enzyme to substrate

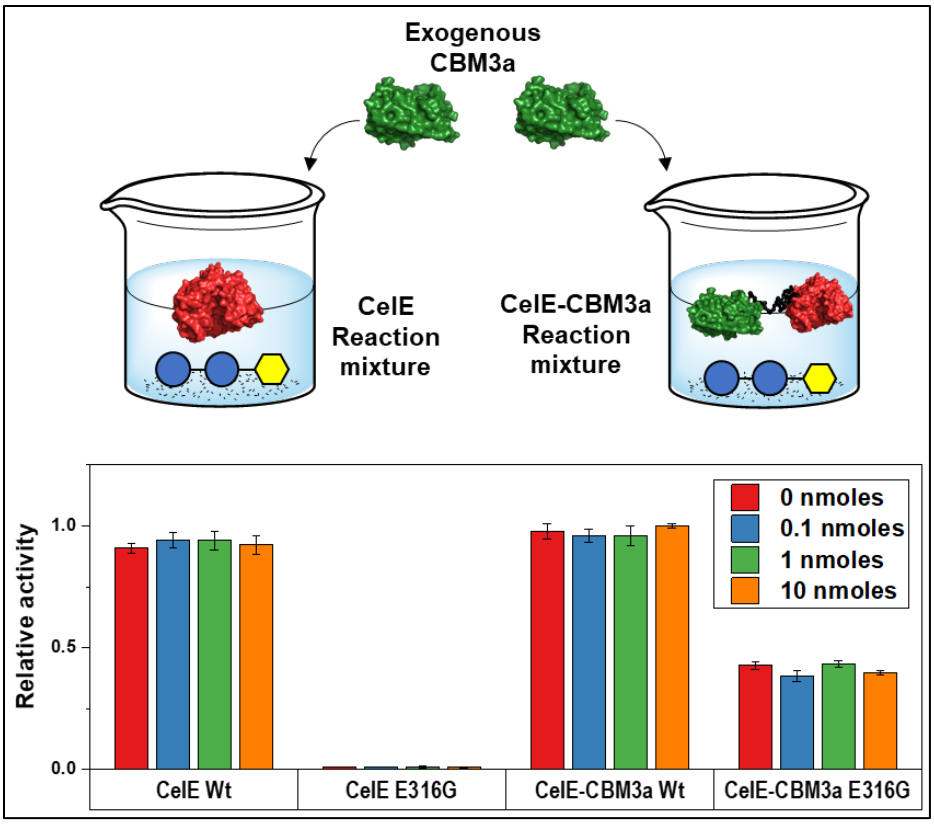



***Figure 9. Intermolecular interactions of exogenously added CBM3a domain along with CelE-E316G catalytic nucleophile mutant does not increase transglycosylation activity.*** *Here, 1 nanomoles of each purified protein was incubated along with 1.5 µmoles of pNP-Cellobiose (pNPC) and the reaction was run for 4 h at 60°C reaction temperature. Different concentrations of purified CBM3a domain alone was spiked into the reaction mixture at 0.1-10 times the relative molar ratio of the added CelE or CelE-CBM enzyme construct. All reactions were run in duplicate with error bars representing 1σ from reported mean value. Total pNP released after 4 h reaction time is shown here.*

concentrations in our assay conditions or the accessibility of substrate to enzyme active site. This very unusual behavior in β-(1,4)-glucan synthase activity could be clearly attributed to the presence of the tethered CBM3a domain. The CBM3a domain could have either potentially increase the local concentration of substrate around the CelE domain or could have facilitated inter-domain interactions between CBM3a-CelE to somehow play an active role in either stabilizing the substrate and/or product in the catalytic turnover step. For both these scenarios, we hypothesized that there were likely additional CBM3a

residues participating in the catalytic turnover of the E316G mutant catalytic domain to result in β-retaining gluco-oligosaccharide products formed from a starting β-substrate (i.e., pNP-CB).

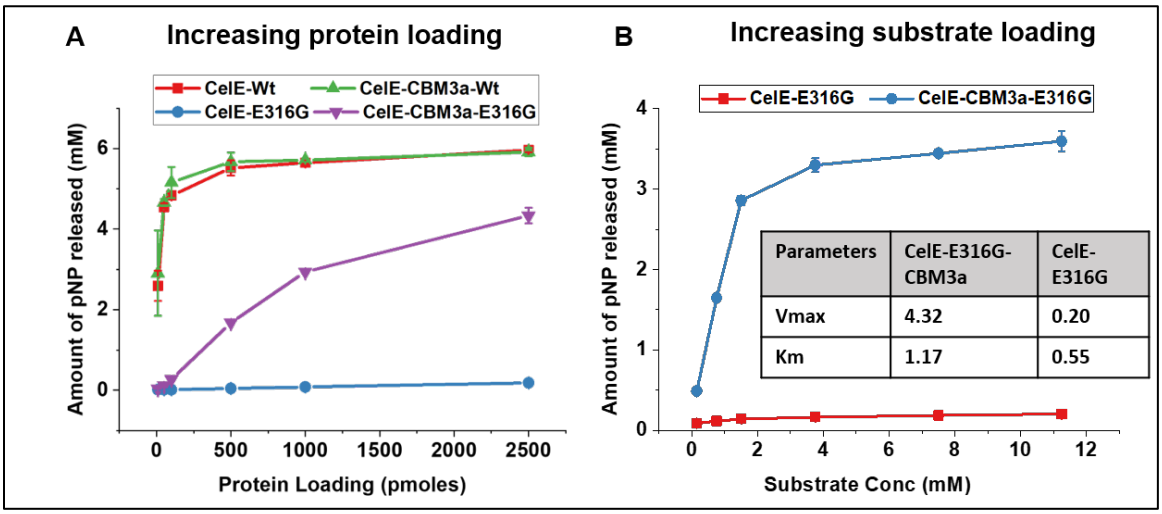



***Figure 10. Effect of protein concentration and substrate concentration on transglycosylation catalytic activity of CelE and CelE-CBM3a based enzyme constructs.*** *A) 10-to-2500 picomoles of each purified protein was incubated along with fixed 1.5 µmoles of pNP-Cellobiose (pNPC) and the reaction was run for 4 h at 60°C reaction temperature. B) 1000 picomoles of each purified protein was incubated along with 0.15-to-11.25 µmoles of pNP-Cellobiose (pNPC) and the reaction was run for 3 h at 60°C reaction temperature. Here, $K_m$ and $V_{max}$ was calculated by fitting of standard Michaelis-Menten model to data shown in (B). All reactions were run in triplicates (to measure pNP release activity) with error bars representing 1σ from reported mean value.*

### 2.4.3 Short chain cello-oligosaccharide synthesis by CelE-E316G-CBM3a

Both the CelE-E316G-CBM3a and CelE-E316S-CBM3a enzymes show activity towards pNP-CB and show similar product profiles as observed through TLC analysis (**Figure 8**). Hence, CelE-E316G-CBM3a construct was chosen for all future experiments to characterize the products and study the mechanism. For the TLC analysis, pure standards like pNP, pNP-glucose, pNP-cellobiose, glucose, cellobiose, cellotriose, cellotetraose were run simultaneously with the reaction mixture samples. The retention factors of the standards were used to identify the reaction products. The two dominant equilibrium reaction products were identified to be cellotriose (DP3) and cellotetraose (DP4) by

comparing the TLC retention factors of the eluting compounds compared to their respective standards (**Figure 8B**). The UV image of the TLC plate prior to orcinol staining also confirmed the presence of pNP containing higher molecular weight sugar polymers like pNP-cellotetraose. Here, the TLC analysis can provide an approximate prediction on the degree of polymerization of the products and not the glycosidic linkages between them. To predict the glycosidic linkage, the reaction mixtures after incubation with CelE-E316G-CBM3a were reacted with CelE wild type enzyme. The oligomeric products synthesized by the E316G nucleophile mutants of CelE-CBM3a were fully hydrolyzed upon the addition of wild type CelE enzyme (**Figure 11A**). This suggested that the glycosidic linkage formed is likely β(1,4) and the transglycosylation products formed can be recognized by the native CelE enzyme active site and fully hydrolyzed into cellobiose alone. Since, CelE belongs to very large family of multifunctional GH5 enzymes that predominantly catalyze hydrolysis of β(1,4) linked sugars, we expected the synthesized higher DP products to be β-retaining as well like pNPC substrates. Because of the multifunctional nature of these GH5 enzymes, we also tested activity of all enzyme constructs towards other glycosidic linkages. The native enzymes CelE, CelE-CBM3a and their nucleophilic mutants were reacted with polysaccharides such as β(1,4) phosphoric acid swollen cellulose (PASC), β(1,3) glucan and α(1,4) starch. However, the native enzyme or its nucleophile mutants, either with or without the CBM3a domain, showed no activity towards the substrates with β(1,3) and α(1,4) linkages (**Figure 11B**) implying that the higher DP products formed were likely not β(1,3) or α(1,4) linked gluco-oligosaccharides. These initial observations regarding the exact molecular weight and type of glycosidic linkage of the higher DP oligosaccharide products inferred from the earlier biochemical assays were also corroborated by conducting detailed tandem mass

spectrometric characterization of the reaction mixture products for E316G nucleophile mutants of CelE-CBM3a.

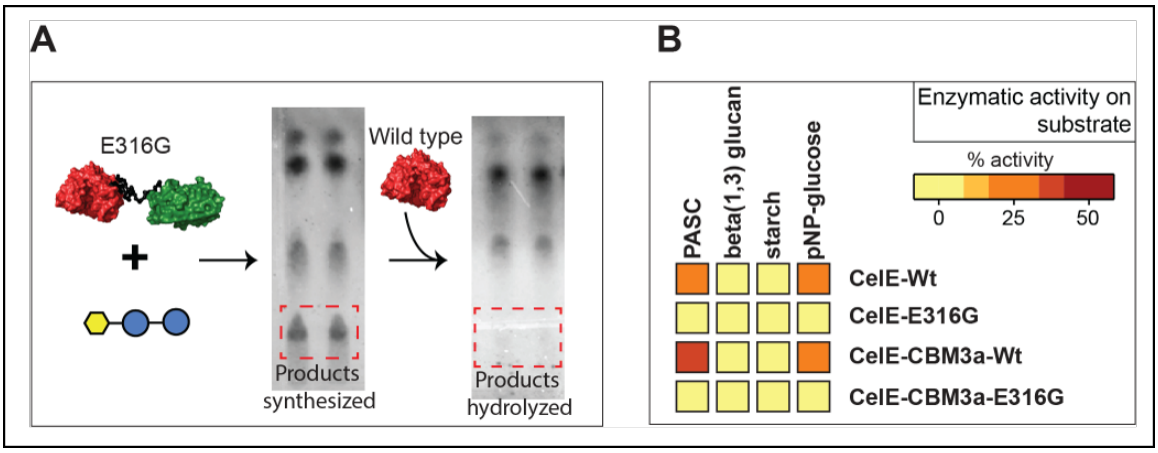



***Figure 11. Biochemical analysis of the reaction mixture indicate the formation of β(1,4) glucosides. (A)*** *The products formed by the mutant enzymes were completely hydrolyzed when supplemented with wild type CelE enzyme indicating that the glycosidic linkage between the products is β-retaining.* ***(B)*** *Activity assays of various enzyme on substrates with different linkages (e.g., β(1,4), β(1,3) and α(1,4)) further suggested that the transglycosylation reaction products contained β(1,4) linkages.*

Briefly, the reaction mixture compounds were all derivatized with methylated free radical activated glycan sequencing reagent (Me-FRAGS) followed by electrospray injection (ESI) tandem mass spectrometric ($MS^n$) analysis to get the exact molecular weights and structural linkage information for the complex sugar-based products. Me-FRAGS derivatized products were directly infused into the ESI-MSn to obtain the MS1 total ion chromatogram profile of the reaction mixture. **Figure 12** shows the representative full ion scan and MS1 data for E316G nucleophile mutants of CelE-CBM3a clearly showed peaks for compounds corresponding with the expected molecular weights (MW) for cellotriose and cellotetraose. Furthermore, MS2 spectra obtained by collision induced dissociation (CID) of selected ions further confirmed the presence of signature peaks of β(1,4) linked glucose sugars.

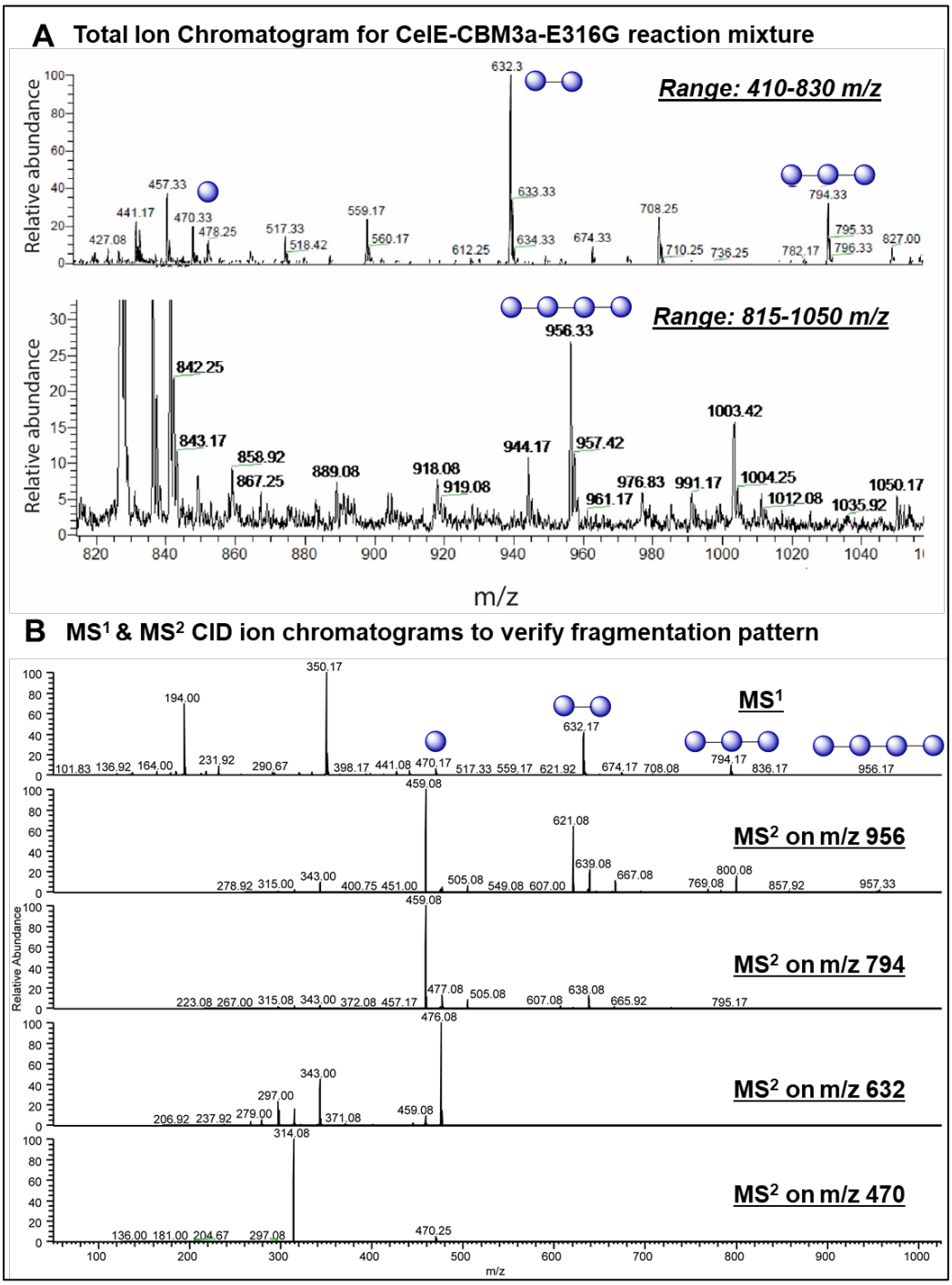



***Figure 12. Tandem mass spectrometric characterization using Me-FRAGS technique provided structural confirmation for the formation of cellotriose and cellotetraose as reaction products for E316G mutant of CBM3a-CelE.*** *Here, 2 nanomoles of each purified protein was incubated along with 1.5 µmoles of pNPC and the reaction was run for 7 h, respectively, at 60°C reaction temperature. All samples were run in duplicate. Total ion chromatograms (A) indicated the presence of glucose, cellobiose, cellotriose, and cellotetraose.*

### 2.4.4 Improved glucan synthase activity unique to CBM3a amongst other Type-A/B family CBMs

By comparing the activity of CelE-316G and CelE-E316G-CBM3a and based on the prior knowledge about carbohydrate binding domains, the peculiar glucan synthase activity results could be associated with increased local substrate concentration due to interactions of the carbohydrate-binding motifs of the CBM with the pNPC substrate, via sugar-$\pi$, $\pi-\pi$, and hydrogen bonding interactions [62,63], that drive up the reaction rate of CelE-E316G domain. However, this was unlikely the reason responsible for the observed synthase activity since the reaction mixtures were homogenously well mixed and also the pNPC substrate has high water solubility (~50 g/L). Also, no significant increase in activity of CelE-E316G with increasing pNP-CB concentration suggests that local substrate concentration is not the limiting factor. Nevertheless, we hypothesized that if the increased substrate concentration is driven by the binding of pNPC towards CBM3a, this behavior could be possibly extended to other CBMs with similar affinity towards pNP-CB or cellobiosyl-like substrates (e.g., cellulose). Representative Type-A (CBM1 from *Trichoderma reesei*) and Type-B (CBM17 from *Clostridium cellulovorans*) CBMs were fused instead of CBM3a to CelE-E316G and tested for pNP release activity towards pNP-CB. The reactions were run for 45 hours total (**Figure 13**) and the total pNP release was continuously monitored and is shown plotted in **Figure 13A** for multiple time points for all tested enzyme constructs. The wild type CelE-CBM constructs (i.e., with no E316G mutation) reached equilibrium within 2 hours of reaction at 60°C with comparable measured specific activities based on overall pNP release rate (**Figure 13B**). While the presence of CBMs did not largely impact the activity of CelE-WT domain, it was observed that only the CBM3a domain contributed to a ~60-fold increase (molar basis based on overall pNP release rate) in the specific activity of CelE-E316G catalytic domain. The

CBM1 or CBM17 caused only a marginal 2 to 3-fold improvement in specific activity compared to the CelE-E316G catalytic domain alone.

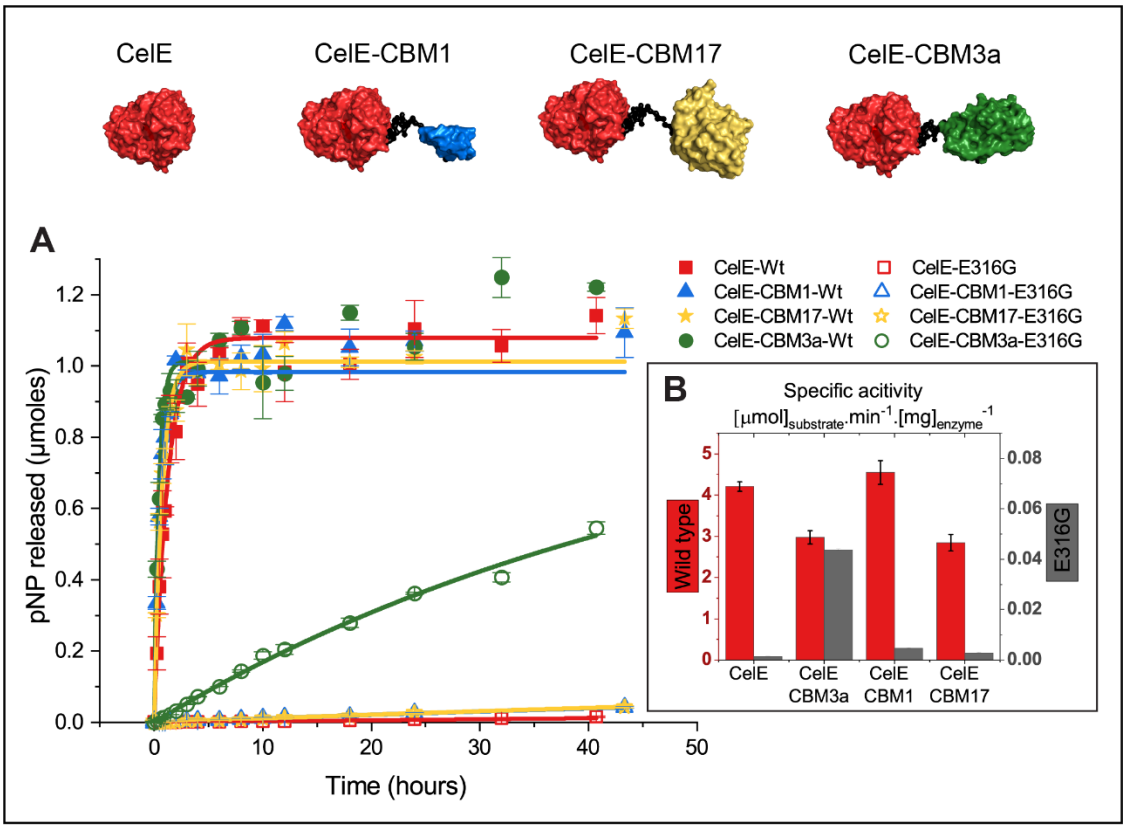



***Figure 13. CBM3a enhances transglycosylation reaction rate of CelE nucleophile mutant by nearly two orders of magnitude unlike other CBM families. (A)*** *CBM1 (Type A) and CBM17 (Type B), which have similar binding affinity towards cellobiose, were shown to not significantly impact the pNP-release activity of E316G mutant of CelE. End-point kinetic assays performed by measuring the pNP released over time showed an increased activity was seen only for the CBM3a tagged E316G CelE mutant.* ***(B)*** *The specific activities of all enzymes were estimated using the kinetic data for initial activity time points to show that there was a nearly 60-fold increase in the specific activity of CBM3a tagged E316G mutant compared to the CelE-E316G control.*

Although no direct measurements to characterize binding affinity of pNP-CB or cellobiose to CBMs was conducted here, the crystalline cellulose binding affinities for both CBM3a and CBM1 are very similar and in the µM affinity range [64,65]. CBM17 has been shown to bind specifically to both soluble sugars and insoluble amorphous cellulose, albeit with much weaker affinities in the mM-to-µM range [64,66]. Clear differences observed in the activities of CelE-E316G fused to different CBMs with similar binding characteristics

suggest that the observed β-glucan synthase activity seen was not necessarily driven primarily by increased local substrate concentration. Instead, there are likely additional subtle and highly specific interactions between the CBM3a and CelE domains that are critical to recovering catalytic activity for the CelE-E316G nucleophile mutant. Closer inspection of the available crystal structures of the three CBMs studied here (CBM1 PDB ID: 1CBH; CBM17 PDB ID: 1J84; CBM3a PDB ID: 1NBC) revealed the presence of an additional hydrophobic cleft (**Figure 14**) in CBM3a which might assist in substrate and/or product stability in synergy with the CelE active site as scaffold for synthase activity. The original 42aa linker between the CelE and CBM3a is moderately flexible with multiple threonine residues that likely facilitates interactions between the two domains. Additionally, since the linker is native to CBM3a from a cellulosomal bacterial system, it is likely that the linker dynamically folds to form compact structures facilitated by CBM3a specific-interactions not possible with other CBMs (i.e., CBM1, CBM17). This could also explain why CBM3a alone gave the highest measured β-glucan synthase activity over CBM1 and CBM17.

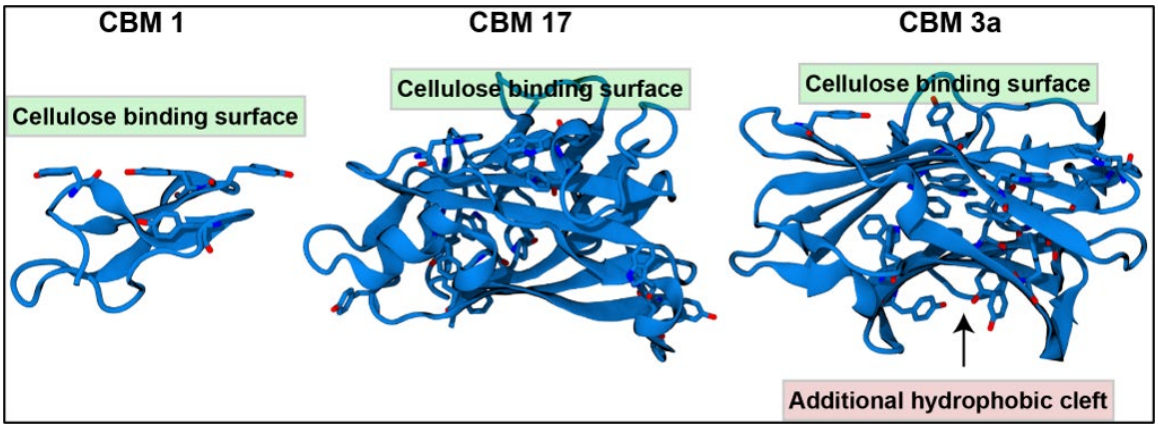



***Figure 14. Comparing crystal structures of CBM1, CBM17, and CBM3a.*** *The classical cellulose binding sites/surfaces that could target binding to cellobiose (or pNP-cellobiose) are shown here. Additionally, the hydrophobic cleft in CBM3a on the opposite side of the classical binding surface is also illustrated here. The additional hydrophobic cleft is hypothesized to interact with the pNP-cellobiose substrate to facilitate transglycosylation in CelE-E316G-CBM3a.*

### 2.4.5 Kinetic analysis of transglycosylation by CelE-E316G-CBM3a

In addition to monitoring pNP release, the total product profiles of transglycosylation and hydrolysis products for the enzyme constructs were characterized as shown in **Figure 15A** and **Figure 16**. The reactions were carried out at 60 °C for 42 hrs and samples were collected at various time points; 15 mins, 30 mins, 45 mins, 60 mins, 1.5 hrs, 2 hrs, 3 hrs, 4 hrs, 6 hrs, 8 hrs, 12 hrs, 16 hrs, 24 hrs, 36 hrs, 42 hrs. Smaller sampling time intervals during the early reaction times was taken to capture the initial reaction kinetics. For every time point, a TLC analysis was performed with UV imaging and orcinol imaging. The TLC data was used to build the kinetic profiles for all the reaction products observed for every time point (**Figure 15B**). **Figure 16** shows the time course profile of all the hydrolysis and transglycosylation products observed in the reaction mixtures. As seen in **Figure 16**, CelE-E316G nucleophile mutant showed no significant change in reaction product profile over the entire duration, whereas wild-type CelE-CBM3a showed formation of some transglycosylation products within the first 15-30 mins of the reaction that mostly disappeared as the reaction tended towards equilibrium giving rise to predominantly cellobiose as final hydrolysis product. On the other hand, CelE-E316G-CBM3a clearly showed a significantly higher concentration of multiple cellodextrins (along with major intermediate products like pNP-cellotriose and pNP-cellotetraose; see **Figure 16**) that increased in yield as the reaction slowly tended towards equilibrium. The fractional oligosaccharide product concentration profile for CelE-E316G-CBM3a shows that the majority of the products are indeed transglycosylation type products (~50%), while the remainder of the product fraction is unreacted pNP-cellobiose substrate along with only minor fractional concentration of cellobiose (<5%). This is unlike the product profile seen for wild-type enzymes that resulted in mostly cellobiose (>90%). Since measurement of various oligosaccharides requires the use of suitable chromatographic techniques (e.g.,

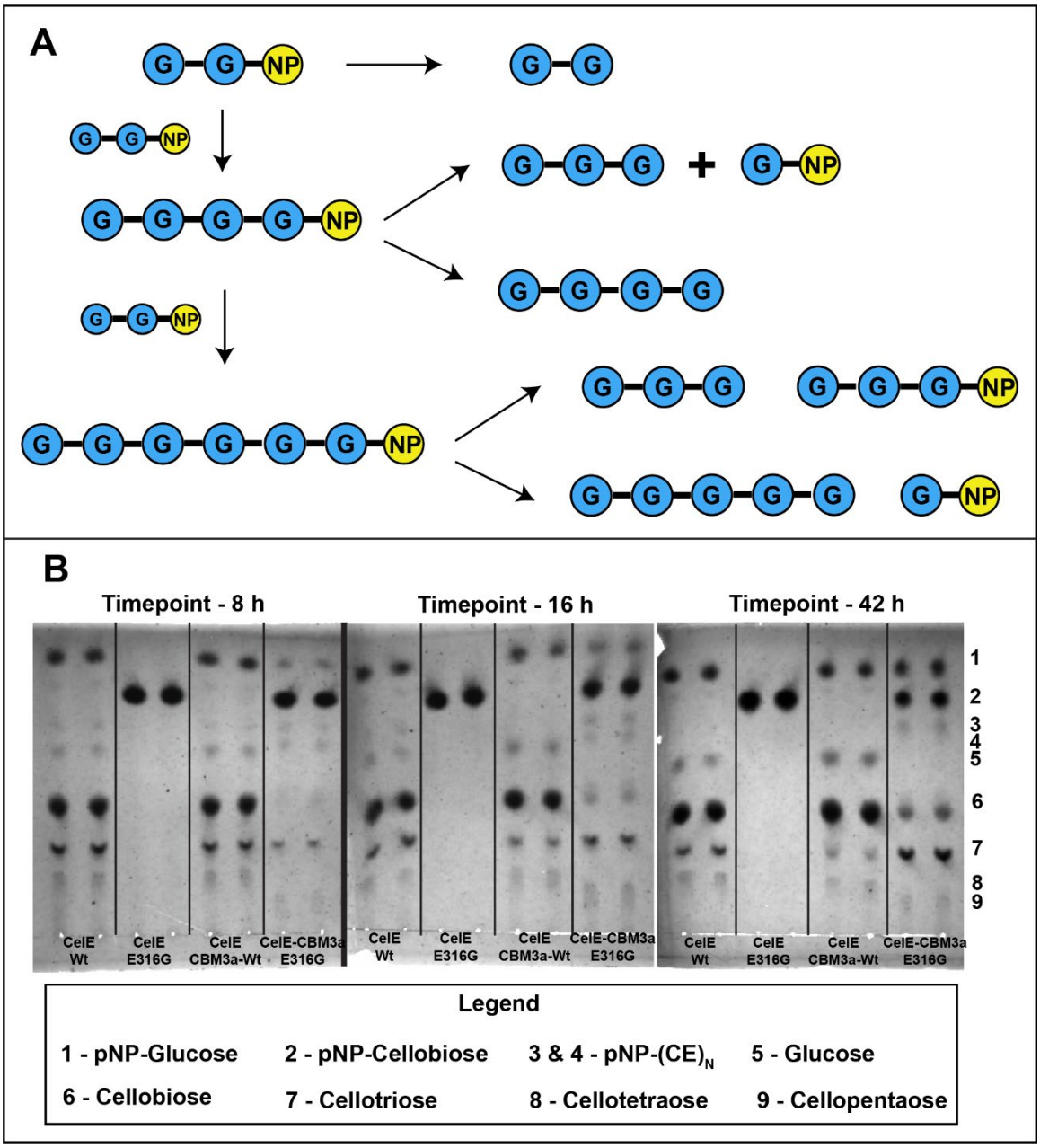



***Figure 15. CelE-E316G-CBM3a utilizes a complex multiple-step reaction cascade, based on analyzed reaction products, to efficiently convert pNPC to mixture of complex glyco-products. (A)*** *Hypothesized reaction cascade scheme is shown here for all major reaction products observed with wild-type and E316G nucleophile mutants of CelE-CBM3a starting with pNPC as initial substrate. Glucose as product was only seen for wild-type enzymes (CelE-Wt and CelE-CBM3a-Wt) and is therefore not shown in cascade here. The blue circles indicate the glucosyl residue and yellow circle indicates the para-nitrophenyl group.* ***(B)*** *Representative TLC analysis after Orcinol staining for duplicate reaction mixtures of CelE, CelE-E316G, CelE-CBM3a-Wt, and CelE-CBM3a-E316G with pNP-cellobiose as starting substrate for varying reaction times.*

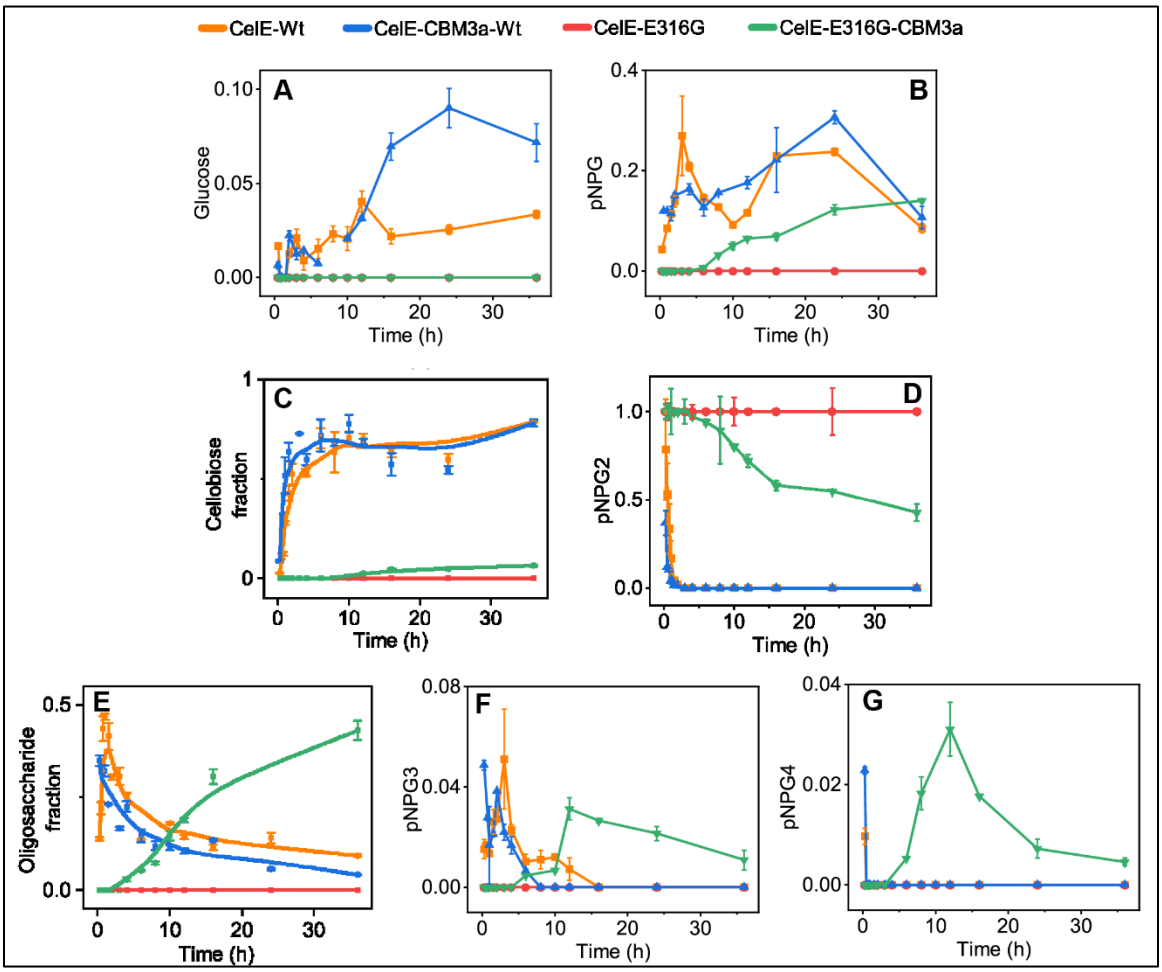



***Figure 16. CelE-E316G-CBM3a utilizes a complex multiple-step reaction cascade, based on analyzed reaction products, to efficiently convert pNPC to mixture of complex glyco-products.*** *Detailed time course of fractional concentration profiles for various reaction products observed in the reaction mixtures as analyzed by quantitative thin layer chromatography (TLC) are shown here. Data is provided for CelE-Wt (orange square), CelE-CBM3a-Wt (blue triangle), CelE-E316G (red circle), and CelE-E316G-CBM3a (green inverted triangle) constructs. Error bar indicates one standard deviation from reported mean values from two biological replicates. Here, y-axis denotes fractional concentration of Glucose, pNPG: pNP-glucose, cellobiose, pNPG2: pNP-cellobiose, Oligosaccharide, pNPG3: pNP-cellotriose, and pNPG4: pNP-cellotetraose. All components were estimated as pNPG2 (or pNP-CB) equivalent using pNPG2 as standard. The most abundant product for wild type enzymes was cellobiose while for CelE-E316G-CBM3a construct the transglycosylation products are mostly dominant.*

HPLC or TLC) which can prove to be tedious especially when dealing with a large number of mutants. Here, we note that total pNP release is directly correlated to the total oligosaccharide formation seen for the CelE-E316G-CBM3a construct. This allows easy high-throughput spectrophotometric based monitoring and rapid screening of mutants with

improved transglycosylation prior to detailed TLC based analysis. The transglycosylation efficiency of a transglycosidase is indicated as a ratio of total transglycosylation product to hydrolysis products (T/H). For an efficient transglycosidase, T/H ratio should be higher as compared to native hydrolytic enzyme. **Figure 17** shows the T/H product ratio for the CelE-E316G-CBM3a construct and respective controls. The wild type CelE constructs with or without CBM3a have similar T/H profiles. The initial products that are formed in the reaction using wild-type enzymes are transglycosylation products that nearly instantly get hydrolyzed into cellobiose as the reaction proceeds to equilibrium. This can be clearly seen as the T/H ratio decreases rapidly to less than 0.2 within few hours of reaction. No products were seen for CelE-E316G mutant and hence the T/H ratio could not be computed for this construct. On the other hand, the presence of CBM3a tethered to CelE-E316G facilitated the transglycosylation activity of CD to form transglycosylation products initially (<10 hours) and virtually no hydrolysis products during this period. Marginal amounts of hydrolyzed products are subsequently observed as the reaction proceeds beyond 10 hrs. Ultimately, as the reaction proceeded to equilibrium, the CelE-E316G-CBM3a construct gave T/H product ratio ranging between 40 to 140-fold higher than the values obtained for the CelE-CBM3a-Wt control. The shift in T/H ratio from <0.2 for wild type enzyme to >6 for CelE-E316G-CBM3a is observed because of nucleophilic mutation and fusion of a CBM3a domain. This methodology could potentially serve as a general design principle to engineering other GH5 family enzymes into transglycosidases.

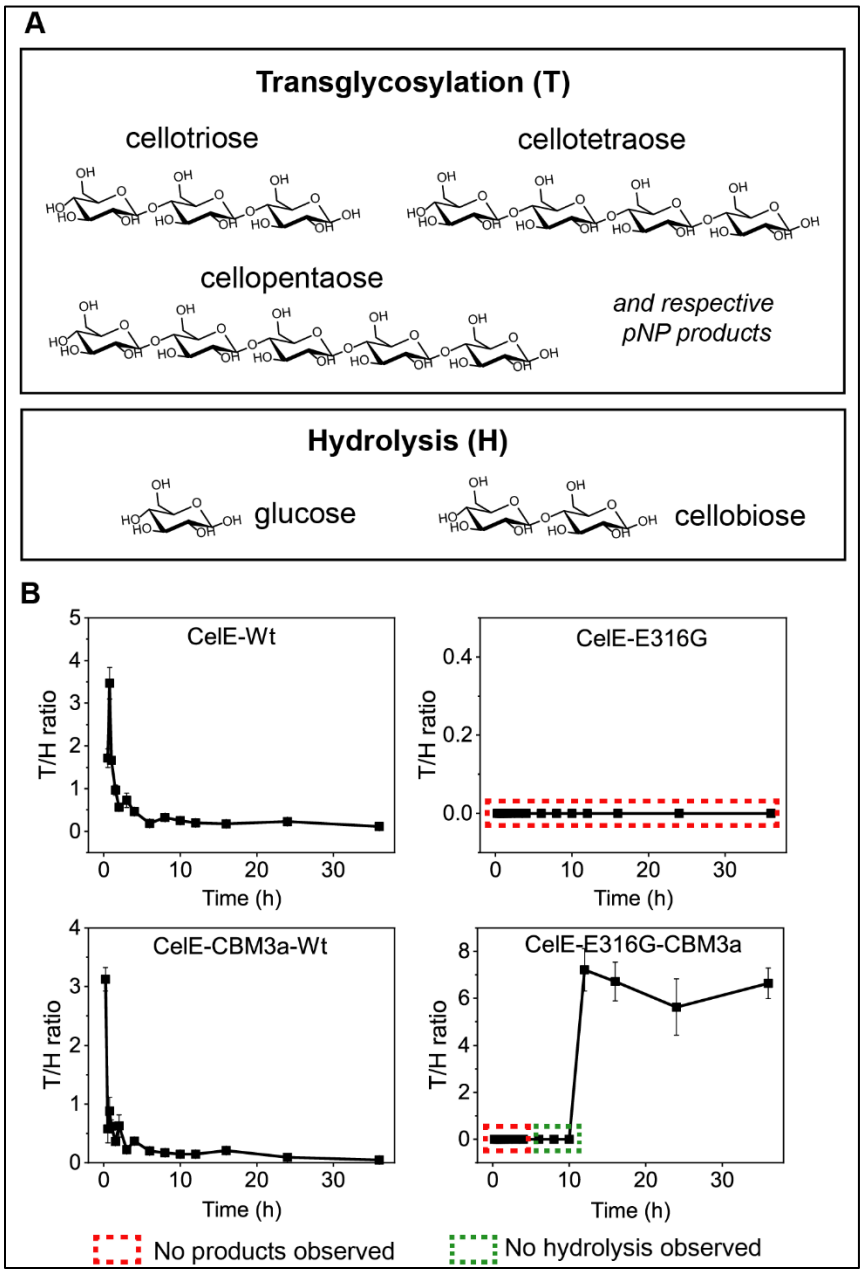



***Figure 17. CelE-E316G-CBM3a is an efficient transglycosidase with very high T/H ratio. (A)*** *Transglycosylation and hydrolysis products observed during the enzymatic reaction on pNP-CB.* ***(B)*** *The transglycosylation/hydrolysis product (T/H) ratio for the four constructs indicate that the CelE-E316G-CBM3a is a highly efficient transglycosidase with a ~140-fold higher T/H ratio as compared to wild type counterpart as the reaction proceeds to equilibrium. During the initial rate, there were no hydrolyzed products observed and only oligosaccharides-based products were seen. For these time points the T/H were not computed (indicated in green box). CelE-E316G mutant is not active and shows no transglycosylation or hydrolysis products and hence the T/H ratio was not computed (red box). Error bar indicates one standard deviation from reported mean values from biological replicates as highlighted in methods section.*

### 2.4.6 Small angle X-ray scattering analysis of CelE-E316G-CBM3a

In the previous section, we concluded that observed glucan synthesis activity in CelE-E316G-CBM3a may not be due to increased local concentration of pNP-CBM susbrate. Furthermore, we hypothesized that specific inter-domain interactions between CelE-E316G and CBM3a due to the flexible native 42aa linker could exist that likely facilitate transglycosylation activity for the chimeric E316G nucleophile mutants of CelE-CBM3a. Although crystal structures are available for individual domains of CelE and CBM3a, there is no crystal structure of full length CelE-CBM3a which could be used to decipher any potential inter-domain interactions. While obtaining the crystal structure could be time consuming for multi-domain proteins, small angle X-ray scattering (SAXS) can be used to study the interactions between the two domains. The compactness/flexibility of the domains and conformer ensemble populations could be predicted using SAXS experiments in the presence and absence of substrate. SAXS measurements were conducted for CelE-CBM3a and CelE-E316G-CBM3a in the presence and absence of pNPC substrate. The SAXS profiles of CelE-E316G-CBM3a, after solvent background subtraction, are shown as double log plot in **Figure 18A** and the p(r) distribution plots are illustrated in **Figure 18B**. The pair distance distribution function, p(r) is calculated by Fourier transformation of the scattering data and can be used to estimate the maximum dimension of the protein complex and the radius of gyration (Rg) of the protein (**Table 2**). In the **Figure 18A**, the scattering intensity for the two domain CelE-E316G-CBM3a protein alone appears to have a sharp drop at the mid-q to high-q range, while the same protein samples quenched with excess pNPC substrate showed a more gradual drop. These results indicate that the ensemble is composed of a limited set of conformations for the protein only samples, however, the samples which have both the protein-substrate display a large number of dynamic conformations. The asymptotic behavior of SAXS intensity

decay in the Porod regime following the Guinier region provides structural information about the shape of the scattering sample.

***Table 2. Real space radius of gyration (Rg) and Dmax for CelE-E316G and CelE-E316G-CBM3a protein constructs, with or without added increasing pNP-cellobiose substrate concentrations (i.e., 0, 258, 644 moles pNPC per mole protein), were estimated using the p(r) function fitted to the SAXS data.***

| Sample Type | $R_g$ (Å) | $D_{max}$ (Å) |
|---|---|---|
| His-CD (or CelE-E316G) | 28 ± 1 | 102 |
| His-CD-Linker-CBM3a (No substrate) | 44 ± 1 | 154 |
| His-CD-Linker-CBM3a: Substrate (1:258) | 43 ± 2 | 155 |
| His-CD-Linker-CBM3a: Substrate (1:644) | 45 ± 1 | 157 |

The Kratky plot (i.e., $q^2R_gI(q)/I_0$ vs. $qR_g$ plot) can qualitatively assess the compactness and degree of unfolding of multidomain proteins. The Kratky plots for two domain proteins shown in **Figure 18C** demonstrate that CelE-E316G-CBM3a exists as a compact structure in solution in the absence of any substrate. Upon addition of the pNPC substrate, the inter-domain flexibility or 'breathing' motion between two domains increases compared to no substrate samples, indicative of a multi-domain protein with a highly flexible linker domain. Furthermore, the dynamic motion between the two domains increases as the concentration of pNPC substrate added relative to protein is increased. This suggests that either; (i) initially a more compact CelE-E316G-CBM3a opens up to make the active site more accessible and the flexibility between the two domains enable the domains to come closer to stabilize substrate interactions and facilitate synthase activity, and/or (ii) the energy released during the slow transglycosylation reaction taking place causes the dynamic flexibility between two domains. In either case, a subset of

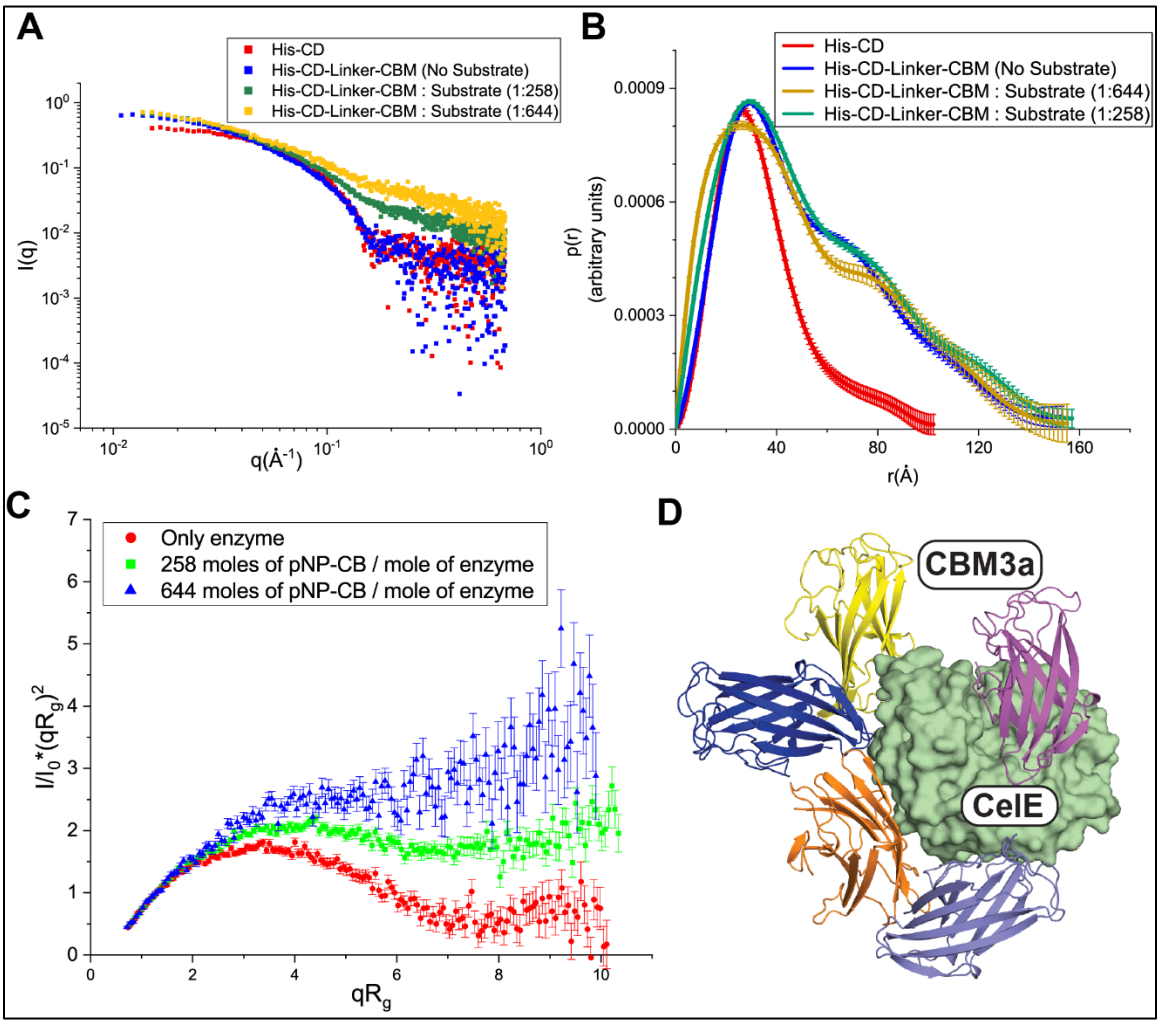



***Figure 18. Small angle X-ray scattering indicates the presence of dynamic flexibility between the CBM3a and CelE-E316G domains. (A)*** *Original SAXS profiles and* ***(B)*** *analyzed p(r) distribution profiles of CelE-E316G and CelE-E316G-CBM3a enzyme with and without pNP-cellobiose are shown here. The distance distribution function (p(r) curves) for each sample were generated from original SAXS data to estimate the real space radius of gyration (Rg) and maximum dimension (Dmax).* ***(C)*** *Kratky plots for CelE-CBM3a-E316G mutant enzyme with and without pNP-cellobiose (pNP-CB) highlights the presence of varying degrees of flexibility in the presence of pNP-CB substrate.* ***(D)*** *Ensemble Optimization Method (EOM) analysis performed on SAXS profiles using the ATSAS package predicted an ensemble of structures with different inter-domain distances between CelE and CBM3a. A representative subset of EOM generated structures which had close interaction between the CelE (as Connolly surface model) and CBM3a (as ribbon models) domains are highlighted here.*

ensemble of protein conformations likely conform close proximity of the two domains. An ensemble of conformations taken by multi-domain protein could be readily generated using the ensemble optimization method (EOM). EOM analysis of the experimental SAXS

profiles was used to predict the model shapes and the fraction of conformations that exhibit close proximity of CelE and CBM3a was determined. Multiple conformations were observed where the CBM3a domain was very close to the native CelE active site cleft (**Figure 18D**). The presence of a long flexible linker chain (42 aa) likely facilitated the dynamic interaction of CBM3a close to the CelE substrate binding site cleft.

### 2.4.7 Inter-domain protein dynamic study through linker modifications

The 42 amino acid linker region between the CelE and CBM3a domains majorly comprises of glycine, threonine, proline and serine residues. The presence of glycine, threonine and serine residues makes the linker flexible and the proline residue impart rigidity to facilitate dynamic folding and unfolding of the two domains. To bolster our previous conclusion that local substrate concentration is not the reason for activity seen in CelE-E316G-CBM3a and to test our hypothesis of inter-domain protein interactions driving the catalytic activity of CelE-E316G-CBM3a, the linker region was further engineered. The CelE active is oriented as polar opposite to CBM3a and the long 42 aa native linker enables the CBM3a domain to curl up and reach the CelE active site to facilitate activity. Hence, the native linker was serially truncated to various lengths of 6 aa, 11 aa and 21 aa to minimize the inter-domains interactions of CelE active site and CBM3a. **Figure 19A** highlights the impact of linker truncation on the activity of CelE-E316G-CBM3a. The shortest linker (6 aa) resulted in a significant decrease in the relative pNP-release activity, while the 21 aa linker showed higher activity compared to even the original native 42aa linker. An optimal linker length might help stabilize the inter-domain interactions by reducing the number of extended conformations and leading to more compact structures. In the control experiments for wild-type CelE and CelE-CBM3a constructs with no nucleophile site mutations, the impact of the linker length was insignificant on the wild type enzyme pNPC hydrolytic activity indicating that the linker truncations did not likely affect protein folding.

In addition to the linker length analysis, the flexibility of the linker was also modified by altering the 42aa linker sequence composition to make the linker either highly flexible or highly rigid compared to the native linker for the same linker length (e.g., more flexible (F2) > F1 >> Native Linker >> R1 > (R2) more rigid). The flexible linker F1 was designed by replacing the proline residue in native linker with pro-flexible glycine residue [67]. Similarly, the threonine in F1 was further replaced with glycine and serine residues to design F2 with increased flexibility. On the other hand, to increase the rigidity, linker R1 was generated by mutating some glycine and serine residues with proline. The more rigid linker R2 was designed completely different to other linkers. The entire native linker was replaced with a 41 aa alpha helix sequence [68]. During the protein folding stage of translation process, the linker folds into 7 alpha helices which does not allow for any proximal interactions of CelE and CBM3a. The relative activity profiles of these highly flexible and rigid proteins are shown in **Figure 19B**, which confirms that the extreme ranges of linker flexibility deleteriously impact the CBM-CD domains from forming a productive complex necessary for achieving optimal synthase activity. Though, it seems clear that the more flexible linkers (F1 and F2) give marginally higher pNP-release activity activity than the more rigid linkers (R1 and R2).

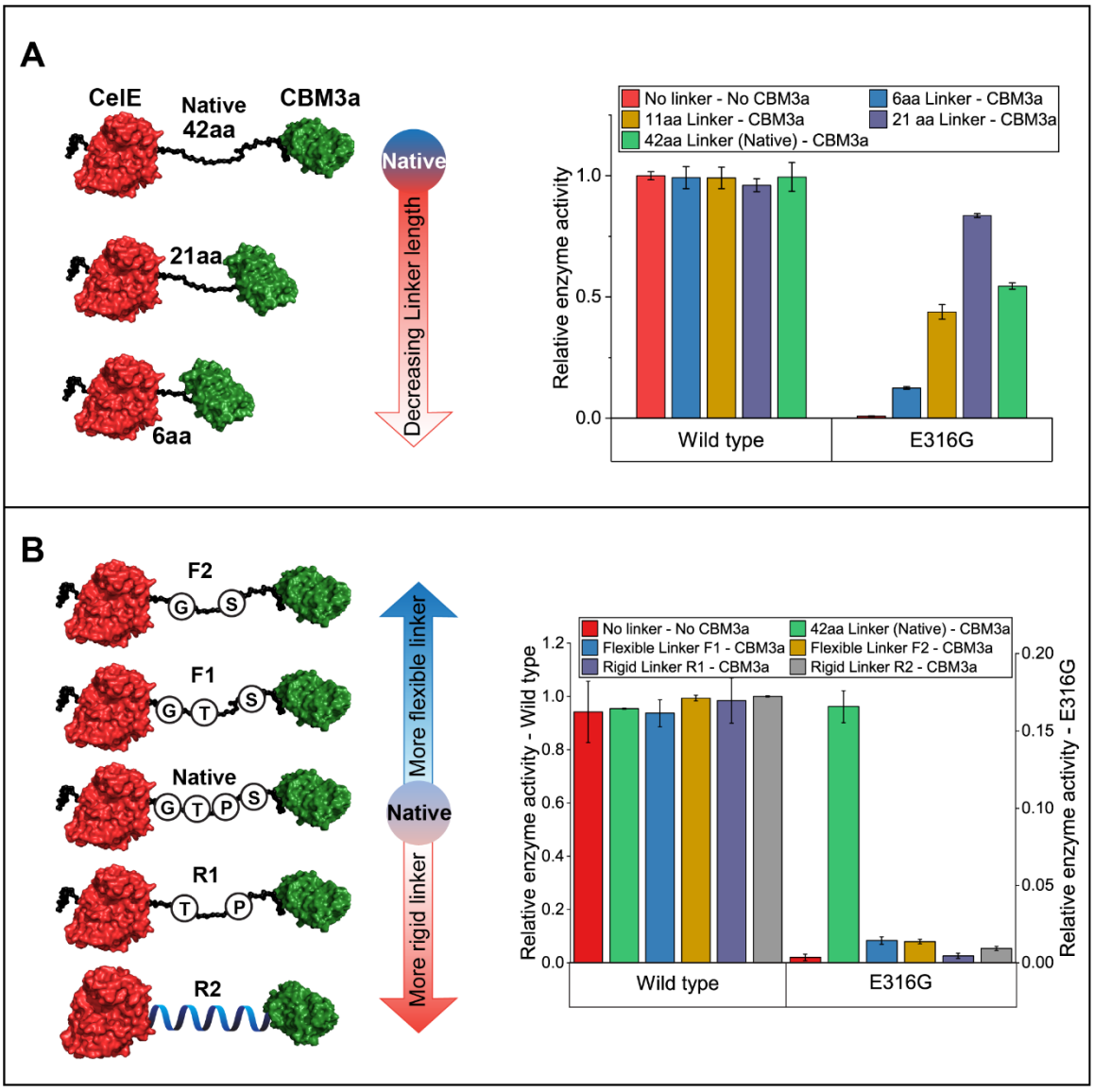



***Figure 19. Inter-domain interactions of CelE and CBM3a domains critical for observed transglycosylation activity is facilitated by optimum linker size and sequence structure****. (A) Biochemical assays to verify the critical role of inter-domain interactions facilitating transglycosylation activity were performed by varying the size of the linker between the CelE and CBM3a domains. A very short linker (6aa linker) showed reduced activity as compared to the original 42aa native linker. Furthermore, there appeared to be an optimal linker length (21aa) that strengthened the interactions between the two domains to further improve transglycosylation activity compared to the native 42aa linker.* ***(B)*** *Flexibility of the 42aa long linker was varied to become either highly flexible (F1 and F2) or highly rigid (R1 and R2) and the corresponding transglycosylation activities of the resultant mutants are highlighted here. The native linker of CBM3a plays a very important role in aiding the inter-domain interactions unlike other similar sized linkers of varying sequence structure.*

### 2.4.8 Mechanistic analysis of transglycosylation by CelE-E316G-CBM3a

While the intra-protein interactions between CelE and CBM3a domains could be attributed to observed activity in CelE-E316G-CBM3a, the underlying catalytic mechanism is not known especially because of lacking the catalytic nucleophile to initiate the reaction. Formation of typical transglycosylation product intermediates, though mostly short-lived once the reaction reaches equilibrium, for wild-type β-retaining GHs to form β-retaining products like pNP-cellotetraose or cellotetraose necessitates the formation of a covalent glycosyl enzyme intermediate (GEI). However, in our study the mutation of CelE catalytic nucleophile (E316) to Ala, Ser, Gly eliminates the covalent GEI formation at 316 residue position. Considering all experimental results discussed earlier, only β-transglycosylation products are synthesized from a β-pNP-CB substrate. This could be possible through two mechanisms (i) $S_N2$ type mechanism where a pseudo nucleophile exists in the CelE active site that inherits the role of native catalytic nucleophile (E316) forming a GEI. All retaining GH5 enzymes operate with an $S_N2$ type mechanism function with the help of two catalytic residues; the nucleophile and acid/base residues (**Figure 20A**). (ii) $S_Ni$-like mechanism where the hydrophobic pocket on the underside of CBM3a coordinates with CelE active site to provide a suitable substrate binding scaffold and facilitate the departure of the pNP leaving group. The $S_Ni$-like mechanism is a substrate assisted mechanism where the protein does not actively form any covalent bond with the substrate and only aids in creating bond strains through non-covalent binding interactions between the protein active-site scaffold and the substrates (**Figure 20B**).

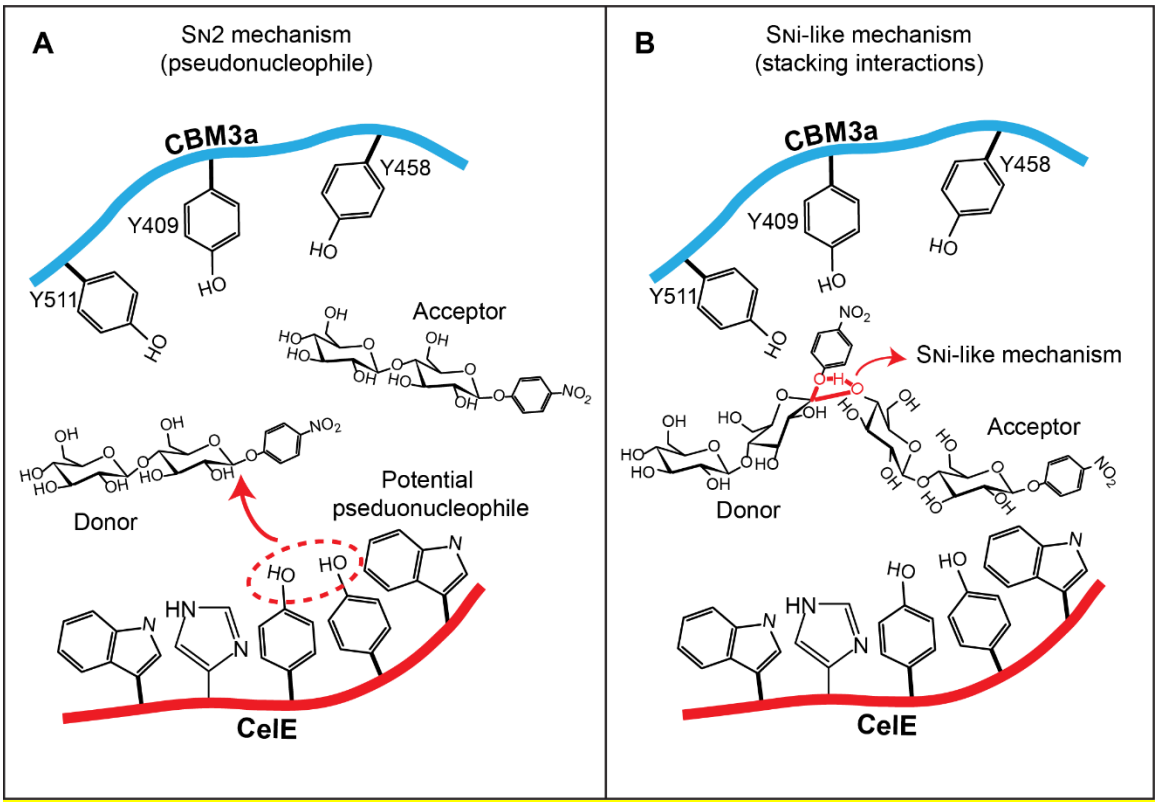



***Figure 20. Possible mechanisms facilitating retaining transglycosylation reaction in CelE-E316G-CBM3a. (A)*** *A potential pseudo nucleophile promoting $S_N2$ mechanism.* ***(B)*** *The aromatic residues in the active site create strain on the donor sugar and promote $S_Ni$-like mechanism*

To unravel the possible mechanism employed by CelE-E316G-CBM3a, the protein model based on a structure predicted using SAXS and the available crystal structures CelE (PDB ID: 4IM4) and CBM3a (PDB ID: 1NBC) was therefore analyzed further to explore the role of key CD/CBM residues on the observed transglycosylation reaction (**Figure 21A and Figure 21B**). An initial scan of CelE active site did not identify any other residues with nucleophilic potential (apart from E316) and at a favorable distance to catalytic acid/base (E193) to facilitate SN2 mechanism. The CelE substrate binding cleft harbors a series of aromatic amino acids (Y270, Y273, W203, W349, H268) and active acid/base residue (E193) that could potential participate in the reaction (**Figure 21A**). On the other domain, additional hydrophobic cleft on the underside of CBM3a includes many aromatic residues could potentially be involved in facilitating the transglycosylation reaction through a poorly

understood reaction mechanism (**Figure 21B**). The aromatic residues are known to be often involved in sugar-aromatic stacking interactions. For these reasons, further targets for mutational studies aimed at disrupting the residues in the CelE active site and CBM3a hydrophobic pocket. All the identified residues (five on CBM3a and six on CelE domains) were mutated to alanine, creating double mutant constructs (along with CelE-E316G included in each case). Similarly, twelve individual single mutations were introduced while keeping the catalytic nucleophile of CelE intact. These additional twelve controls were used to account for any potential activity loss due to active site structural changes or protein misfolding. While all the control samples (with single point mutations on CBM3a alone) showed identical activity as the wild type CelE-CBM3a enzyme (**Figure 21C**), the double mutations introduced resulted in a significant loss in pNP-release activity compared to CelE-E316G-CBM3a (**Figure 21D**). Mutation of each CBM3a residue decreased the pNP release activity by at least 50% when compared to CelE-E316G-CBM3a, which was also directly correlated to the higher molecular weight transglycosylation products formed. All the aforementioned residues of CBM3a together could contribute to stable docking of pNPC substrate or pNP leaving group along with the CelE substrate binding site scaffold to a certain extent and no mutation alone caused a complete loss of transglycosylation activity.

The single point mutation of the active site residues of CelE alone (**Figure 21C**), which were in close vicinity to the native GEI transition state complex, further reduced the pNP-release activity by 90% or even higher (e.g., Y270A, W349). This was not unexpected because these residues are critical for docking and strain promotion in the substrate. Interestingly, mutation of residue Y273 does not affect the hydrolytic activity of CelE-CBM3a (**Figure 21C**). Mutating the tyrosine residues at Y273 and Y270 position quenches

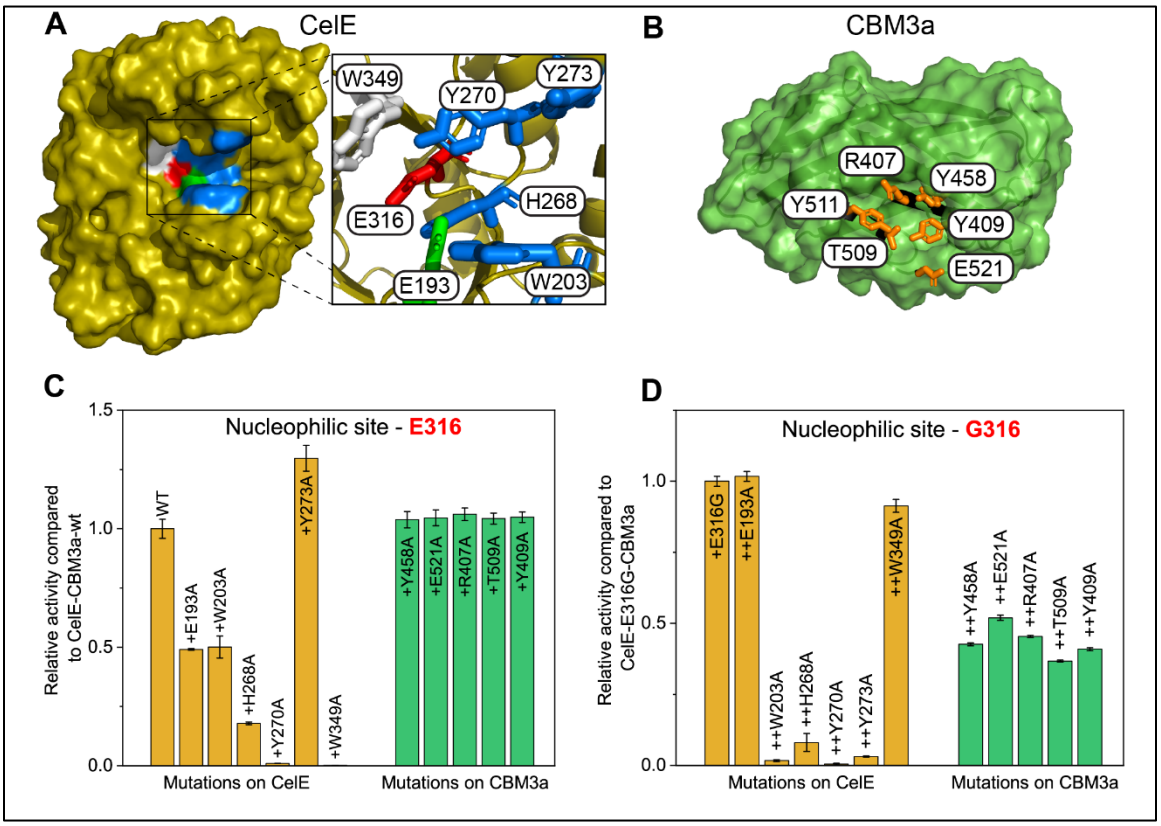



***Figure 21. Impact of critical amino acid residues identified within the hypothesized CelE-E316G-CBM3a enzyme-substrate complex that impact transglycosylation activity. (A)*** *Active site cleft of native CelE enzyme is shown here. The catalytic nucleophile (E316 – red), catalytic acid/base (E193 – green), negative subsite (W349 – white) and positive subsites (Y270, Y273, W203 and H268 – blue) are highlighted.* ***(B)*** *Key residues within the hydrophobic cleft of CBM3a are illustrated here (R407, Y458, Y409, E521, and T509 – orange).* ***(C)*** *Relative activity of single alanine point mutations of the key identified residues of CelE and CBM3a with native nucleophile as compared to CelE-CBM3a-wt is present here as bar graphs.* ***(D)*** *Activity of the double mutants containing alanine mutation of key residues with nucleophile also mutated to glycine (E316G) are shown as relative activity compared to CelE-E31G-CBM3a.*

the activity of CelE-E316G-CBM3a suggesting that one of the tyrosine residues could be a pseudo-nucleophile. In other GH families like GH33, tyrosine has been shown function as a catalytic nucleophile where they facilitate transglycosylation of sialic acid sugars [69]. Since retaining enzymes operating with an $S_N2$ type mechanism require two functioning catalytic residues; the nucleophile and acid/base residues, we performed additional mutational studies to study the impact of the acid-/base residue (**Figure 21D**). As

expected, mutating the acid/base residue (E193) to alanine abrogated the pNP-release hydrolytic activity of CelE-E193A by ~50% (compared to control enzyme construct with an intact catalytic nucleophile or CelE-CBM3a) confirming the likely participation of this predicted acid/base residue in the retaining mechanism of native CelE. However, introducing the E193A mutation into CelE-E316G-CBM3a construct did not cause any change in pNP-release activity suggesting that the traditional acid/base residue does not likely participate in the catalytic reaction for our chimeric glucan synthase mutants. Non-participation of catalytic acid/base residue in CelE-E316G-CBM3a activity suggests that $S_N2$ type reaction would be less likely. However, similar results regarding the non-participatory role of traditional acid/base residues was observed in another GH family 1 transglycosidase [45]. The enzyme was catalytically functional even after the mutation of both catalytic nucleophile and acid/base residue. Quantum mechanics/molecular mechanics (QM/MM) simulation revealed the use of SNi-like mechanism to facilitate transglycosylation where transition state was stabilized by tyrosine residues in the active site. Previously assigned role of pseudo-nucleophile for Y270 and Y273 could also be attributed to play a transition state stabilizing role for a $S_Ni$-like mechanism. While clearly all these residues play a key role in facilitating transglycosidase activity facilitated by the CBM3a domain, the actual reaction mechanism of glycosyl transfer is still debatable and would require additional experiments for a definitive proof.

### 2.4.9 CBM3a engineering to facilitate transglycosylation in other GH family 5 enzymes.

The residues identified through biochemical assays played a key role in the activity of CelE-E316G-CBM3a. MAFFT (**M**ultiple **A**lignment using **F**ast **F**ourier **T**ransform) sequence alignment of the CelE protein sequence and CBM3a protein sequence to protein sequences of their respective GH/CBM families showed that these interacting residues

are highly conserved throughout these families (**Figure 22**), and we hypothesize that similar mechanisms are likely to operate in other GH-CBM chimeric constructs. Preliminary experimental studies have indeed confirmed that fusion of the CBM3a-42aa linker domain to nearly a dozen other phylogenetically related GH5 families (and all with corresponding nucleophile site mutations to glycine) resulted in active transglycosidases as seen here with CelE-E316G-CBM3a (Collaborative work with Ms. Madhura Kasture, **Appendix A1.3.1**). Representative results for one other homologous GH5 family protein [70] fused to CBM3a are shown in **Figure 23**. This study suggests that CBM3a fusion engineering can be performed to other homologous GH-CBM chimeric enzymes to engineer efficient transglycosidases.

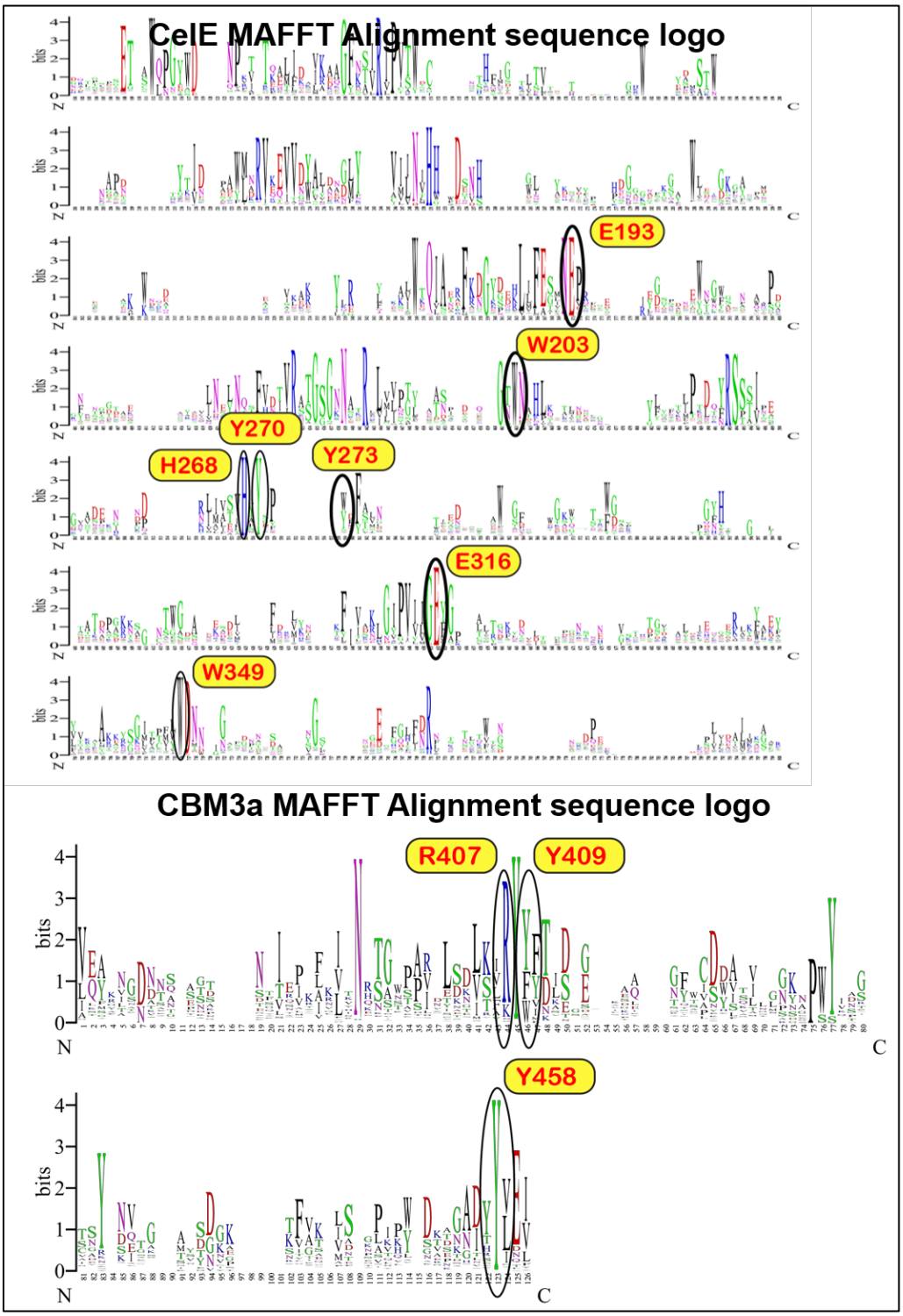



***Figure 22. MAFFT alignments of CelE and CBM3a domains using Pfam database.*** *The sequence logo was generated using Geneious R11 software and several major conserved residues relevant to this study are highlighted here.*

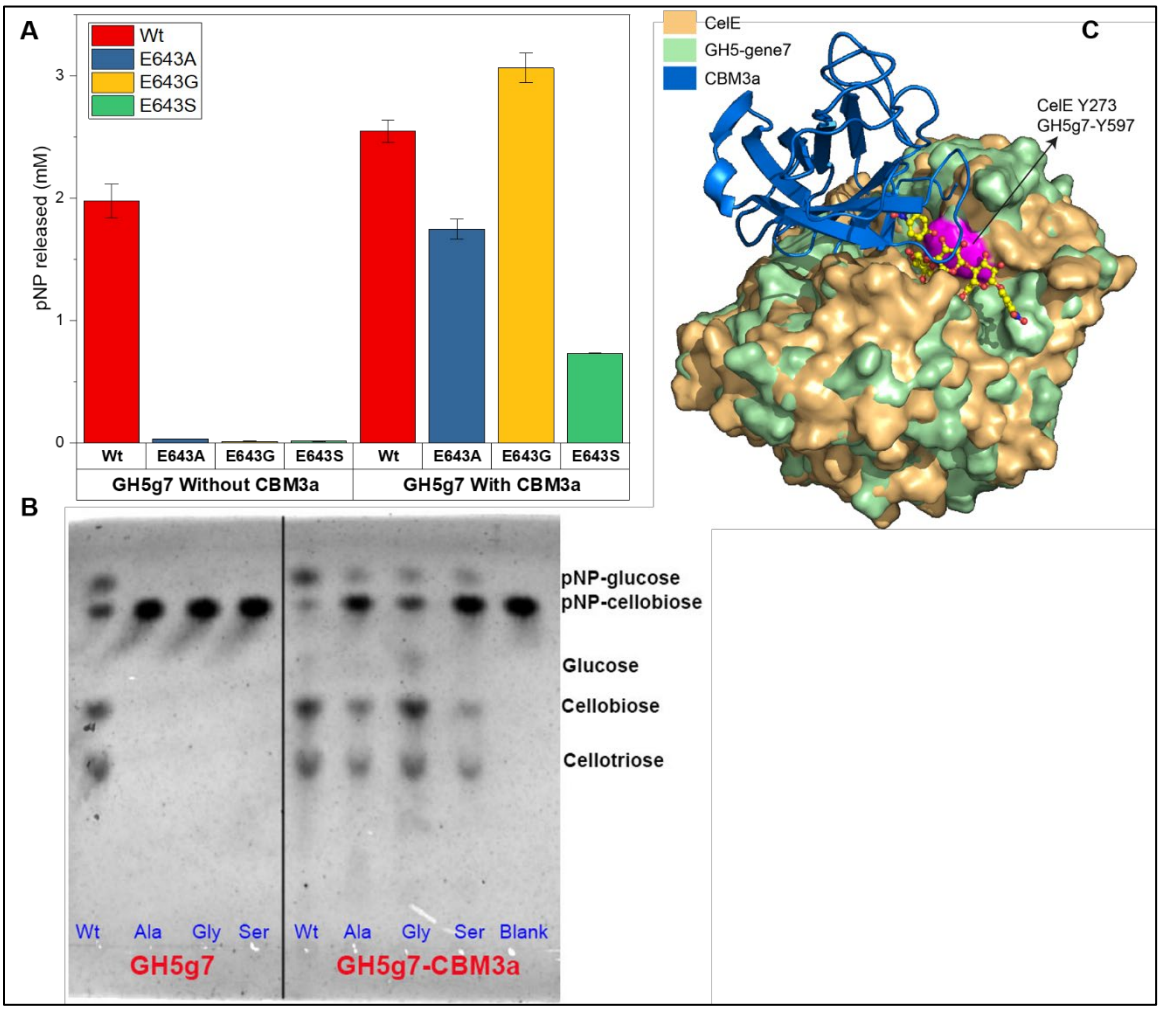



***Figure 23. Activity of homologous GH5 family protein (GH5-gene7) on pNP-cellobiose.*** *GH5-gene7 belongs to the same GH5_4 subfamily as CelE enzyme. Here, 300 pmoles of the purified GH5g7 enzyme, with or without tethered CBM3a (analogous to CelE-CBM3a constructs), were reacted with 0.8 µmoles of pNP-cellobiose (pNP-CB) for 24 h at 40°C reaction temperature. All reactions were run in triplicates with error bars representing 1$\sigma$ from reported mean value. The relative amounts of pNP released* ***(A)*** *and TLC analysis by Orcinol staining* ***(B)*** *to evaluate the products formed in each reaction mixture are shown below. The 3D homologous structure of GH5-gene7 was docked along with the CelE structure to highlight how closely the two GH5_4 catalytic domains overlap structurally* ***(C)****. CBM3a is docked adjacent to the CelE catalytic domain based on EOM analysis of SAXS data.*

# Chapter 3: Development of high-throughput screening technique to facilitate directed evolution of glycosynthases

## 3.1 Introduction

Glycosynthases (GS) are nucleophilic mutants of glycosyl hydrolases that are devoid of hydrolytic activity but can synthesize glycan polymers using modified donor sugars. The first glycosynthase mutant was engineered from GH family 1 *Agrobacterium* β-glucosidase by mutating the glutamic acid nucleophile to an alanine residue [71]. Using α-glucosyl fluoride as a donor sugar, this glycosynthase mutant was successful in synthesizing various disaccharides and trisaccharides consisting of glucose and galactose sugars. This remarkable work has inspired many studies across the world to design efficient glycosynthases from various GH families. The general strategy adopted for engineering GHs to GSs is (i) mutating the catalytic nucleophile to a small non-polar residue (ii) testing the rescue of hydrolytic activity using a small external nucleophile like azide and formate (iii) performing glycosynthase activity assays. However, out of 166 GH families classified, only limited enzymes from 17 different GH families have been engineered to glycosynthases. This is because, identification of active and efficient glycosynthases has been difficult owing to various reasons such as; (a) The potential target candidates available to substitute for catalytic nucleophile for glycosynthase activity is not consistent between different GSs. Therefore, a large mutant library for each GH enzyme must be cloned, expressed and tested for chemical rescue which makes the process cumbersome [24,44] (b) Multiple GHs mutants that are identified as positive glycosynthases, do not show chemical rescue using azide and formate [72]. Using chemical reactivation experiments as a lone screening method makes the process unreliable to engineer

efficient glycosynthases. (c) Finally, additional active site engineering is required to over come the slower catalytic rate of glycosynthases, and the structural features required for efficient glycosynthases are currently unknown.

For these reasons, directed evolution approach can prove advantageous to generate large library sets and screen for positive constructs in-vivo. An essential requirement for the directed evolution approach is the need for suitable high throughput screening strategies which are lacking for glycosynthase enzymes. Withers groups (UBC, Canada) first developed a two-plasmid HTS method where one plasmid contains the GS gene while the other contains a screening enzyme that only releases a fluorophore from the GS reaction product and not the reactants [73,74]. Another approach was based on chemical complementation using a yeast three-hybrid system where the GS activity was linked to the transcription of *LEU2* reporter gene, making cell growth dependent on product formation [32]. Both these approaches are highly specific to individual GS family and have narrow applicability to screen for novel substrate specificity. Hence, efforts towards universal method to screen GS libraries using glycosyl fluoride as the sugar donor led to the development of a pH based assay [34]. Here, hydrofluoric acid, a by-product of GS reaction was detected by a pH sensitive color indicator. In another study, a chemical probe that reacts specifically to the fluoride anion to generate a fluorophore has been developed to screen small GS libraries [36,75]. These techniques can screen small GS libraries ($10^2$-$10^4$ mutants) and use thermolabile glycosyl-fluoride donor sugars which cannot be used for engineering thermophilic enzymes. To increase the probability of finding rare GS mutants, screening techniques capable of handling much larger library ($10^6$-$10^{10}$ mutants) are necessary. Also, to reduce false positive due to self-hydrolysis of donor substrate, more stable substrates like glycosyl azides needs to be used. The first screening method to identify the by-products (azide ion) of glycosynthase reactions using glycosyl azide

donors was developed in Chundawat group (Rutgers University, USA) (**Appendix 1.3.2**). The screening technique uses copper free click chemistry reaction between azide and cycloalkyne to form a triazole product [38]. The triazole compound of the by-product (azide ion) had lower fluorescence than the triazole compound of substrate (glycosyl azide). This difference in fluorescence was used for screening GH29 fucosynthase enzyme and identify mutants with improved activity. However, this technique uses a negative selection of improved mutants with low fluorescence signal against a high fluorescence background from wild type enzyme and this decreases the screening efficiency. Hence, there is a need to develop a simpler positive selection technique for *in vivo* azide detection to enable glycosynthase screening.

In this chapter, we report the development of an *in vivo* cell based biosensing toolkit to selectively detect azide ions which can be used to screen for glycosynthases. Briefly, a native *Escherichia coli* cyanate operon was engineered to generate a synthetic promoter plasmid that is selectively inducible by azide ions. The tunable expression of model green fluorescence protein (GFP) was shown using azide based promoter induction and compared to the standard *E. coli* lactose (*lac)* operon. Finally, the biosensing potential of this promoter to selectively detect inorganic azide ions, with respect to organic glycosyl azides, is showcased as a proof-of-concept glycobiology application.

## 3.2 Experimental outline and design

A *E.coli* based biosensor which can detect azide ions was developed. This biosensor can be used an alternative protein expression system and as a tool for glycosynthase engineering. The following experimental plan was followed to develop the biosensing system.

Step 1: A cyanate inducible operon in *E.coli* was identified and used to test for azide induction.

Step 2: The native promoter was not very efficient. Hence, promoter engineering was done to improve the induction strength of azide ion.

Step 3: The optimized azide inducible promoter was tested for protein expression levels and was compared to lac promoter expression levels. Also, this system did not show protein expression with organic azides indicating that this can be a good tool for glycosynthase engineering.

## 3.3 Methods

### 3.3.1 Bacterial strain engineering:

All the chemicals, reagents, and solvents were purchased from Fisher Scientific and Sigma-Aldrich and used without purification. *E. coli* strain used for cloning was E. cloni® 10G (Lucigen, WI, USA) and for protein expression using lac promoter was BL21-CodonPlus-RIPL [λDE3] (Stratagene, Santa Clara, CA, USA). Phusion 2X high fidelity PCR master mix (0.04 U/μL Phusion DNA polymerase, 400 μM dNTPs, Phusion 2XHF buffer, 3 mM MgCl2) was purchased from Thermo Fisher Scientific (USA) and restriction enzymes were procured from New England BioLabs Inc. (USA). The primers used for site directed mutagenesis (SDM), sequence and ligation independent cloning (SLIC), and sequencing reactions were obtained from Integrated DNA Technologies, Inc (USA). Successfully cloned plasmids were isolated from E. cloni® 10g cells using IBI Scientific (USA) plasmid extraction kit and sequences were confirmed through Sanger sequencing performed by Genscript Inc. (NJ, USA). The carbohydrate substrates used in reported assays were purchased from Synthose Inc, Canada.

The *E.coli* strains used for protein expression using *cyn* promoter were constructed using the protocol and reagents outlined in Datsenko 2000 [76]. Briefly, strains sCB1 (BW25113 *cynS*::FRT) and sKS3 (BW25113 *cynX*::FRT) were constructed by first streaking out the strains JW0331 (BW25113 *cynS*::Kan) and JW0332 (BW25113 *cynX*::Kan)[77] respectively from the Keio collection onto LB-agar kanamycin (50 µg/ml) plate. The kanamycin marker was then removed by transforming the individual strains with pCP20, following the protocol outlined in Datsenko 2000, and curing the strain of the plasmid. The final strain was diagnosed by PCR and for loss of kanamycin resistance.

Strain sKS4, BW25113 *cynR,cynTSX*::FRT, was constructed by first transforming the Keio parent strain, BW25113, with pKD46 and plated onto carbenicillin (100 µg/mL) plates. 5 mL overnights of BW25113 + pKD46 in LB + carbenicillin (100 µg/mL) were grown at 30°C and then back-diluted 1:100 into 50 mL LB + carbenicillin (100 µg/mL) + 0.2% L-arabinose. The 50 mL culture was grown to an OD600 of 0.8, at 30°C, and cells were washed 4 times with 50 mL of ice-cold water. The final cell pellet was resuspended 1:250$^{th}$ the starting culture volume with fresh water and sat on ice until ready to electroporate. The linear DNA fragment, that was used to knockout the *cynR* gene and *cynTSX* operon, was synthesized by amplifying the kanamycin gene from pKD4 using primers k1 and k2. The linear DNA fragment was checked by gel electrophoresis for purity and was then cleaned-up and concentrated using Qiagen's PCR clean-up kit, the product was eluted in water. The intermediate strain sKS1 (BW25113 *cynR*, *cynTSX*::Kan) was made by mixing 100 ng of linear DNA with 50 µL electrocompetent BW25113 + pKD46 cells, shocking immediately with 1.8 kV (with a pulse constant of 5.2ms), and recovered in 900 µL of SOC at 37°C for 3 hours. After recovery, cells were spun down at 10,000xg for 2 minutes at room temperature, resuspended to 100 µL, plating solution onto kanamycin (50 µg/mL) plates, and grown at 37°C overnight. sKS1 was transformed with pCP20 to remove the kanamycin

selection marker, following the protocol in Datsenko 2000, making the final strain sKS4 (BW25113 *cynR,cynSTX*::FRT). sKS4 was diagnosed by loss of kanamycin resistance and PCR.

### 3.3.2 Design and construction of pCyn vectors:

The cloning of the plasmid constructs used in this study were performed using Sequence and Ligation-Independent Cloning (SLIC) protocol as outlined in Stevenson et.al [78]. Briefly, to create the pCyn-v1-GFP, the gene fragment consisting of native cynR and cyn promoter/operator region (gblock1) was custom synthesized from Genscript Inc, USA. The GFP gene fragment was taken from pEC-GFP plasmid available at Chundawat lab. Both the gene fragments were cloned into the parent plasmid, ptrc99a while getting rid of the intrinsic lac promoter using the primers p1-p4 and following the SLIC protocol (**SI Table S2A**). To generate pCyn-v2-GFP, an optimized promoter region was designed in-silico and the corresponding DNA fragment (gblock2) was custom synthesized and pCynv-1-GFP was used as starting DNA with primers used being p5-p8. The spacer region of 100 bp from the pCyn-v2-GFP was removed using primers p9-p10 to get pCyn-v3-GFP while an additional random sequence of 900 bp (gblock3) was added using p11-p14 to generate pCyn-v4-GFP. Further, pCyn-v5-GFP was constructed from pCyn-v2-GFP by site directed mutagenesis of the cynR constitutive promoter using primers p15-p16. In the end, the pCyn-v5-GFP was the parent DNA used to create pCyn-v6-GFP, pCyn-v7-GFP and pCyn-v8-GFP using the primers p17-p18, p19-p20 and p21-p22, respectively. All the constructs were diagnosed using Sanger sequencing and the sequence verified plasmids were preserved at -80°C and their corresponding transformed E. cloni cells were stored in 15% glycerol stocks at -80°C.

### 3.3.3 Induction of pCyn-v1/v2-GFP expression:

The pCyn-GFP plasmid constructs were transformed into BW25113 strains (wt, sCB1) and individual colonies were obtained on LB agar plates supplemented with 100 µg/ml Carbenicillin antibiotic. The transformants were inoculated into 10 ml of LB media with carbenicillin and grown at 37°C for 16 hrs. The 10 ml of overnight grown culture was transferred to 200 ml of fresh LB media with carbenicillin and incubated at 37°C until mid-exponential phase (OD 0.4-0.6) was reached. At this point, the 200 ml culture was split into six 25 ml falcon tubes. Two of the tubes were induced with 1 mM sodium azide and two tubes with 1 mM cyanate while the remaining two tubes were used as no induction control. The cultures were placed in the 37°C shaking incubator and 2 ml of sample for bulk fluorescence measurement and 200 ul of sample for flow cytometer runs were collected at every time point.

### 3.3.4 Induction of pCyn-v2/v3/v4-GFP expression:

The pCyn-GFP plasmid constructs were transformed into BW25113 strains (wt, sCB1, sKS3, sKS4) and individual colonies were obtained on LB agar plates supplemented with 100 µg/ml Carbenicillin antibiotic. The transformants were inoculated into 10 ml of LB media with carbenicillin and grown at 37°C for 16 hrs. The 10 ml of overnight grown culture was transferred to 200 ml of fresh LB media with carbenicillin and incubated at 37°C until mid-exponential phase (OD 0.4-0.6) was reached. At this point, the 200 ml culture was split into six 25 ml falcon tubes. Three of the tubes were induced with 1 mM sodium azide and remaining three tubes were used as no induction control. The cultures were placed in the 37°C shaking incubator and 2 ml of sample was collected at every time point for bulk fluorescence measurement and 200 ul was sampled to measure OD600.

### 3.3.5 Induction of pCyn-v2/v5/v6/v7/v8-GFP expression:

The pCyn-GFP plasmid constructs were transformed into BW25113 strains (wt, sCB1, sKS3, sKS4) and individual colonies were obtained on LB agar plates supplemented with 100 µg/ml Carbenicillin antibiotic. The transformants were inoculated into 10 ml of LB media with carbenicillin and grown at 37°C for 16 hrs. Then, 400 µl of overnight grown culture was transferred to 20 ml of fresh LB media with carbenicillin and incubated at 37°C until an OD600 of 0.3-0.4 was reached. At this point, 1 ml of culture was added to 12 wells in a 96 deep well plate. Three wells each were induced with 10 µM, 100 µM 1000 µM of sodium azide and three wells were left uninduced. The 96 well plate was placed in the 37°C shaking incubator for 2 hours. 200 ul of the sample was used for OD600 measurements and the residual sample was used for bulk GFP fluorescence measurement.

### 3.3.6 Bulk GFP fluorescence measurement:

The samples collected from each experiment were first centrifuged to remove the LB media. The pelleted cells were resuspended in 250 µl of BPER reagent (Bacterial Protein Extraction Reagent) and incubated at room temperature for 10 min to lyse the cells. The resultant samples were centrifuged and 200 ul of the supernatant was used to measure GFP fluorescence in a black opaque bottom 96 well plate using a spectrophotometer.

### 3.3.7 Flow cytometer data acquisition and analysis:

For measuring the GFP expression in individual cells, flow cytometer analysis was performed as reported previously by the Chundawat lab [38]. Briefly, at each timepoint for a given induction experiment, 200 µl of cells were collected. The cells were centrifuged to remove the LB media components, resuspended in 1x PBS buffer (phosphate buffered saline) and run through the Guava® easyCyte™ flow cytometer to measure the

fluorescence distribution at 488 nm excitation and 525 nm emission. Guavasoft 3.3 software was used for gating live cells based on forward scatter (FSC) and side scatter (SSC) and the fluorescence associated with each cell was collected for 10,000 cells per sample and analyzed using FlowJo software.

## 3.4 Results and discussion

### 3.4.1 cynTSX operon

The *E. coli* genome consists of several operons (~700 operons) performing specific functions essential for bacterial metabolism and survival in harsh environments [79]. In particular, the cyanate or *cyn* operon enables *E. coli* cells to survive in cyanate-rich environments [80]. This gene had evolved in archaea and cyanobacteria for energy production and nitrogen assimilation during the early history of single-celled life in extreme marine environments [81]. The *cyn* operon, analogous to the *lac* operon, is comprised of three structural and functional genes; *cynT*, *cynS*, and *cynX* (**Figure 24A**) that encode carbonic anhydrase, cyanate hydratase, and cyanate transporter proteins, respectively. These genes catalyze the bicarbonate-dependent decomposition of cyanate ions into carbon dioxide and ammonia. Upstream of *cynTSX* genes, *cynR* repressor gene encoding cynR protein is present for regulatory purposes. The operon is under tight negative regulation of the cynR repressor protein which upon binding to the operator region results in unfavorable DNA bending to prevent transcription of the *cynTSX* genes. The *cyn* operon is only activated when a cyanate molecule binds to the allosteric repressor protein also bound to the promoter/operator region. Exogenous cyanate can bind to the cynR inducer binding domain and cause conformational changes in the binding domain thereby reducing the DNA bend in the operator region to facilitate transcription [80,82,83].

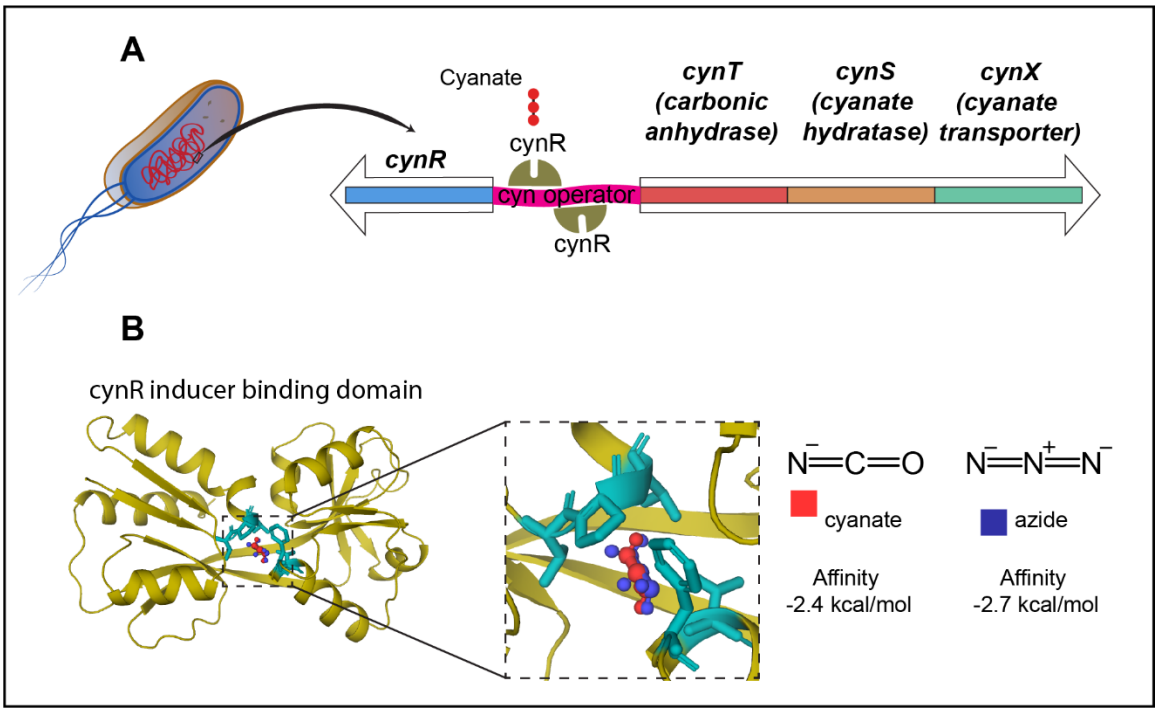



***Figure 24. Azide can be a gratuitous inducer for cynTSX operon.*** *(**A**) Native cynTSX operon is present in E. coli genome to facilitate cell growth in a cyanate-rich environment. Cyanate molecule binds to the cynR repressor protein which regulates the downstream protein expression of three essential genes (cynT, cynS, cynX). (**B**) Molecular docking of cyanate and azide ions in the binding pocket of cynR protein is shown. Since azide is a structural homolog for cyanate, it also binds tightly in the cynR binding pocket and in the same location as cyanate.*

Since cyanate is a linear molecular ion that is structurally homologous to azide, we hypothesized that azide could also bind to cynR and trigger the *cyn* operon. To confirm if azide can bind with similar binding affinity to cynR, Autodock vina [84] was used to perform docking simulations for both cyanate and azide in the binding pocket of cynR. The binding orientation of the docked ligands revealed a good overlap in the binding site with similarly predicted binding affinities (**Figure 24B**). This suggested that azide could indeed function as a gratuitous inducer for the native *cyn* operon.

### 3.4.2 Azide is a gratitious inducer for cynTSX operon

The regulatory segment of cynTSX operon is composed of the cyn operator and *cynR* coding gene. The cyn operator consists of the cynTSX promoter and cynR promoter with their respective -35 and -10 regions and ribosomal binding sites. The regulatory segment was subcloned into a plasmid vector with ampicillin resistance and with green fluorescent protein (*gfp*) as the reporter gene. The regulatory segment was placed upstream of the *gfp* reporter gene to facilitate GFP expression using either cyanate or azide ions as inducers (**Figure 25A**). The resultant plasmid (called pCyn-v1-GFP) was transformed into *E. coli* BW25113 wildtype (Wt) strain and induced with 1 mM of sodium cyanate or sodium azide. The supernatant of the lysed cells after induction was analyzed for fluorescence using a spectrophotometer for GFP expression. In addition, the fluorescence of individual intact cells was analyzed through a flow cytometer and **Figure 25B** presents the resultant histogram plots of the induced cells fluorescence after 19 hours of induction at 37°C. Compared to the control (uninduced cells), the azide induced cell lysate had a nearly 20-fold increase in GFP fluorescence while cyanate did not result in any significant GFP expression. The absence of GFP expression in cyanate induced cells was likely due to the breakdown of cyanate by the endogenous cyanate hydratase enzyme encoded by the native *E. coli* BW25113-Wt genome. Hence, we next tested the induction capacity of the synthetic promoter using a Δ*cynS* knockout strain (BW25113-sCB1) to clearly show GFP expression upon induction by cyanate (**Figure 25B**). However, the amount of protein expressed based on total cell lysate GFP fluorescence was very low suggesting that the native *cyn* promoter strength was quite poor. This is not surprising since the native *cyn* promoter regulates associated CynTSX proteins expression required to overcome cyanate toxicity and that seldom requires high protein expression yields. The native operator region also contains suboptimal -10 sequence and Shine-Delgarno (SD)

sequences (or ribosomal binding site; RBS) which could play a role in poor expression strength.

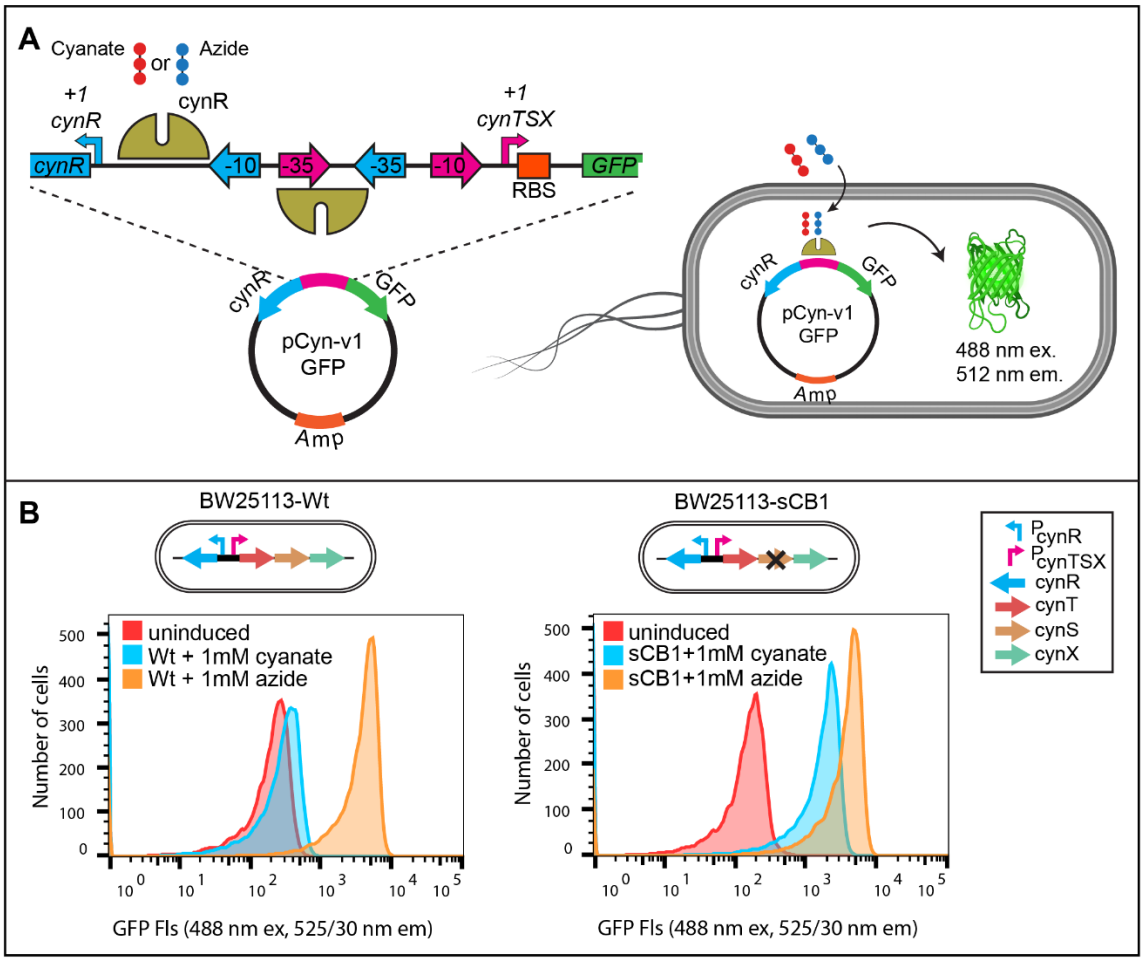



***Figure 25. The cyn operator is inducible by azide for biosensing. (A)*** *Plasmid map design of pCyn-v1-GFP containing the native cynR and cyn operator region cloned upstream of a GFP reporter gene. The native version (v1) was used first to monitor GFP expression using azide and cyanate as inducers.* ***(B)*** *Flow cytometer analysis of GFP fluorescence signal confirms cyanate and azide inducible heterologous in vivo protein expression using native (BW25113-wt) and engineered ΔcynS knockout (BW25113-sCB1) E. coli strains.*

### 3.4.3 Engineering GFP promoter

To improve the native promoter strength for enabling higher inducible protein expression levels, the native *cyn* promoter was next engineered to contain consensus -10 and SD sequences. The *cynR* gene, that is negatively regulated by *cyn* promoter, was placed under control of an independent constitutive promoter[85] (see SI Text for sequence

information). Additionally to making these modifications, the promoters were separated by inserting a random DNA spacer sequence of 100 bp (pCyn-v2-GFP) and 1000 bp (pCyn-v4-GFP) to avoid any interference and steric hinderances between the two promoters (**Figure 26**). A control construct that contained no spacer sequence (pCyn-v3-GFP) was also generated. The modified constructs were individually transformed into *E. coli* cells and tested for GFP expression using sodium azide as inducer. All three engineered constructs showed a 30-fold increase in the fluorescence signal in the cell lysate, indicating a significant increase in GFP expression as compared to the proof-of-concept pCyn-v1-GFP version 1 promoter (**Figure 26**). The pCyn-v4-GFP construct with longest spacer region showed around 30% increased fluorescence when compared to the no spacer pCyn-v3-GFP control. However, significant leaky GFP expression was seen for the no induction controls of v4 construct (**Figure 26B**). The extra-long spacer sequences could have potentially allowed undesirable interactions in the plasmid causing a change in the DNA bending properties for CynR protein.

Decreasing the length between the promoters largely reduced the leaky expression under no induction while maintaining the largely improved GFP yield upon azide induction (**Figure 26B**). For all subsequent work, pCyn-v2-GFP design was chosen to keep the promoters at optimal distance from each other. The nearly 600-fold increase in GFP fluorescence observed was now promising to be able to utilize the pCyn-v2-GFP engineered design for autonomously sensing azide ions. To determine the minimal azide amount required to induce GFP protein expression, we induced cells transformed with the pCyn-v2-GFP plasmid with varying concentrations of azide ions. The maximum amount of azide used for induction was limited to 5 mM since higher amounts had a negative

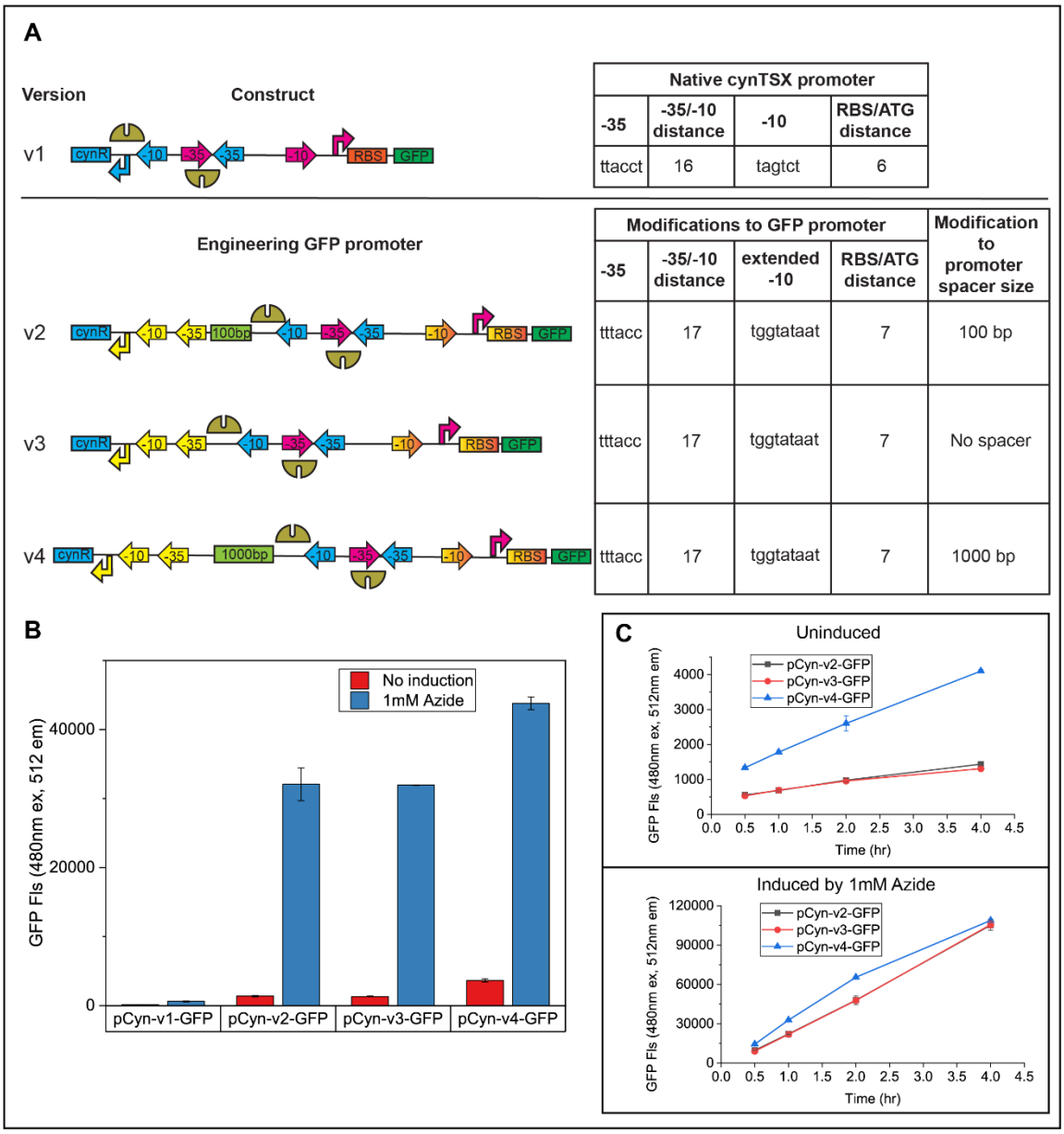



***Figure 26. Engineering GFP promoter region for improved expression.*** *(**A**) Illustration of the promoter regions of pCyn-v1-GFP, pCyn-v2-GFP, pCyn-v3-GFP, pCyn-v4-GFP. Fluorescence measurements of lysates from BW25113-Wt cells containing plasmids pCynv2-GFP (blue square), pCyn-v3-GFP (red triangle), and pCyn-v4-GFP (green circle) and induced with 1 mM sodium azide, or uninduced cells for 2 hours at 37°C **(B)** and for various time points **(C).** Error bars indicate one standard deviation from reported mean values from three biological replicates*

influence on bacterial cell growth (**Figure 27B**). GFP fluorescence of cell lysate showed a strong correlation in protein expression as a function of inducer dosage. GFP expression increased until 1 mM azide concentration induction after which there was reduction in the

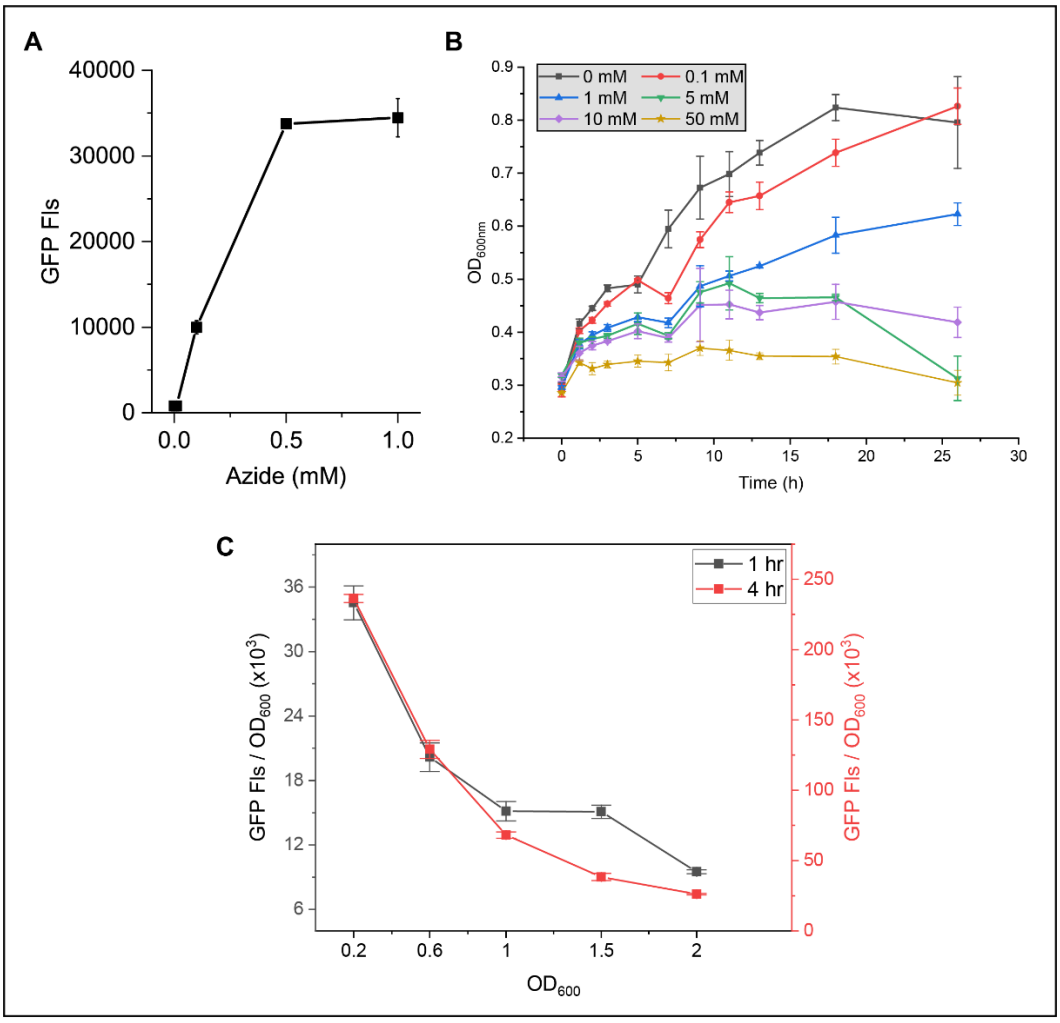


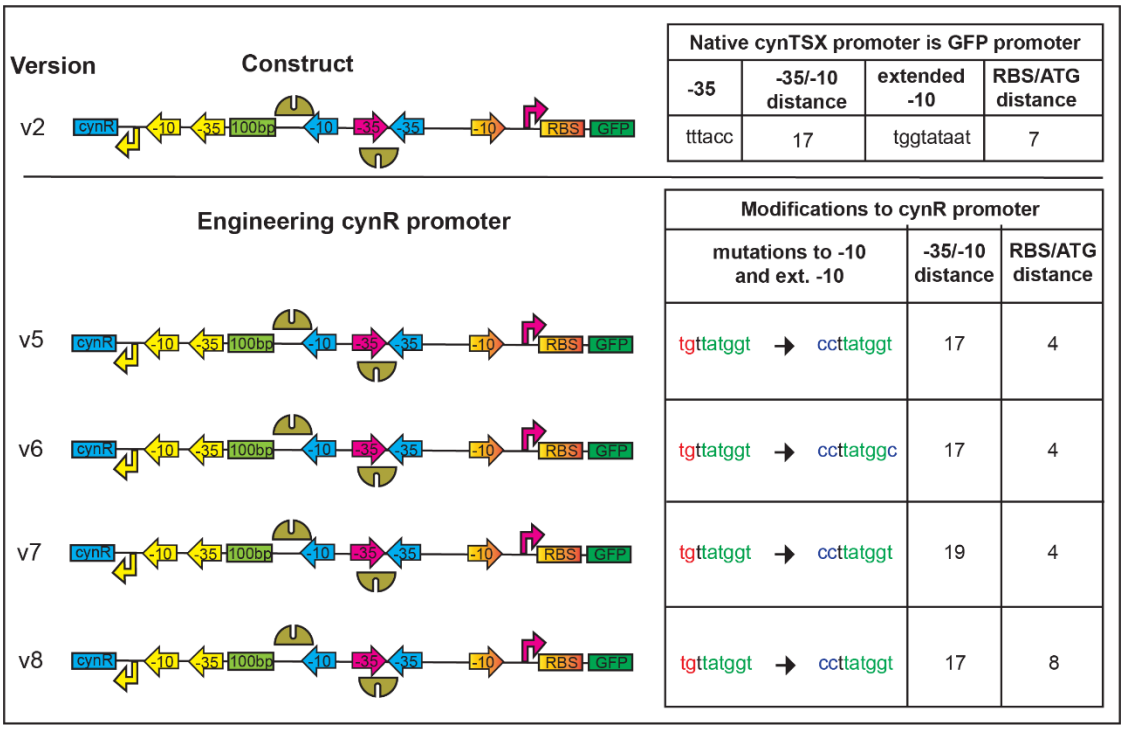

***Figure 27.Characterization of pCyn-v2-GFP. (A)*** *Fluorescence plot for cell lysate of BW25113-wt strain containing pCyn-v2-GFP and induced with increasing concentrations of sodium azide* ***(B)*** *E. coli cells incubated with varying concentrations of sodium azide to study the inhibitory effect of azide on cell growth. Cells were grown in LB media at 37°C for 18 hours and the OD of the cells were measured at various intermittent time points after introduction of sodium aide at time t=0. Error bars indicate one standard deviation from reported mean values from three biological replicates. (**C**) Cell lysate GFP fluorescence from E.coli BW25113 cells induced using 1 mM sodium azide but at different starting cell optical density ($OD_{600}$). Cells were grown until the desired OD in LB media and induced with 1 mM sodium azide and grown then for 4 hours with measurements taken at 1 hour (grey square, left axis) and 4 hours (red square, right axis), respectively. The GFP fluorescence on the y-axis is normalized with the measured OD after induction for specified culture times. Error bars indicate one standard deviation from reported mean values from three biological replicates*

GFP signal for 5 mM inducer concentration likely due to cellular toxicity of azide (**Figure 27A**). The lowest tested azide concentration (10 μM) showed marginal GFP expression indicating that the detection limit for this synthetic promoter would be around 10 μM. The

bacterial growth phase during the time of induction also played an important role and we observed that induction at early exponential phase yielded higher expressed protein amounts (**Figure 27C**).

### 3.4.4 CynR constitutive promoter engineering

CynR protein acts as a repressor for the promoter by causing a bend at the -35 site hindering the binding of RNA polymerase. The cynR protein consists of two domains: an inducer binding domain and a DNA binding domain. The inducer binding domain binds to the inducer (cyanate or azide) which decreases the bend at the promoter site to facilitate transcription. For regulating protein expression levels, along with the inducer amounts the basal level of cynR protein present is also critical. We therefore further engineered the cynR constitutive promoter to adjust the background level of repressor protein expressed. Four different constructs (v5, v6, v7 and v8) were created by introducing mutations at the promoter -10 site, regions between -10 and -35 site, and between RBS and translation start site (**Figure 28**). The expression strength of these modified promoters was tested in the wild type (BW2113) and several additional knockout strains (sCB1, sKS3, sKS4) by monitoring GFP fluorescence in cell lysates. The knockout strains sKS3 and sKS4 were generated to remove the native *cynX* gene and *cyn* operon (**Table 3**), respectively, to minimize the export of azide ions using native transporter proteins and reduce interference from the endogenous *cyn* operon. The pCyn-v8-GFP construct, which had reduced efficiency at -10 site but optimal length between RBS and translation start site, gave about 120-160 fold higher reporter GFP fluorescence as compared to the native promoter, with the maximum fold increase observed in *ΔcynX* (sKS3), *ΔcynR*, and *ΔcynTSX* (sKS4) knockout strains (**Figure 29D**). On the other hand, while other designs




| Native cynTSX promoter is GFP promoter | | | |
|---|---|---|---|
| -35 | -35/-10 distance | extended -10 | RBS/ATG distance |
| tttacc | 17 | tggtataat | 7 |

| Modifications to cynR promoter | | |
|---|---|---|
| mutations to -10 and ext. -10 | -35/-10 distance | RBS/ATG distance |
| tgttatggt → ccttatggt | 17 | 4 |
| tgttatggt → ccttatggc | 17 | 4 |
| tgttatggt → ccttatggt | 19 | 4 |
| tgttatggt → ccttatggt | 17 | 8 |

***Figure 28. cynR constitutive promoter engineering.*** *Illustration of the promoter regions of pCyn-v5-GFP, pCyn-v6-GFP, pCyn-v7-GFP, pCyn-v8-GFP.*

***Table 3. Genotype of the bacterial knockout strains generated in this study.***

| Strain name | Genotype |
|---|---|
| BW25113-wt | Wild type |
| BW25113-sCB1 | BW25113 *cynS*::FRT |
| BW25113-sKS3 | BW25113 *cynX*::FRT |
| BW25113-sKS4 | BW25113 *cynR,cynSTX*::FRT |

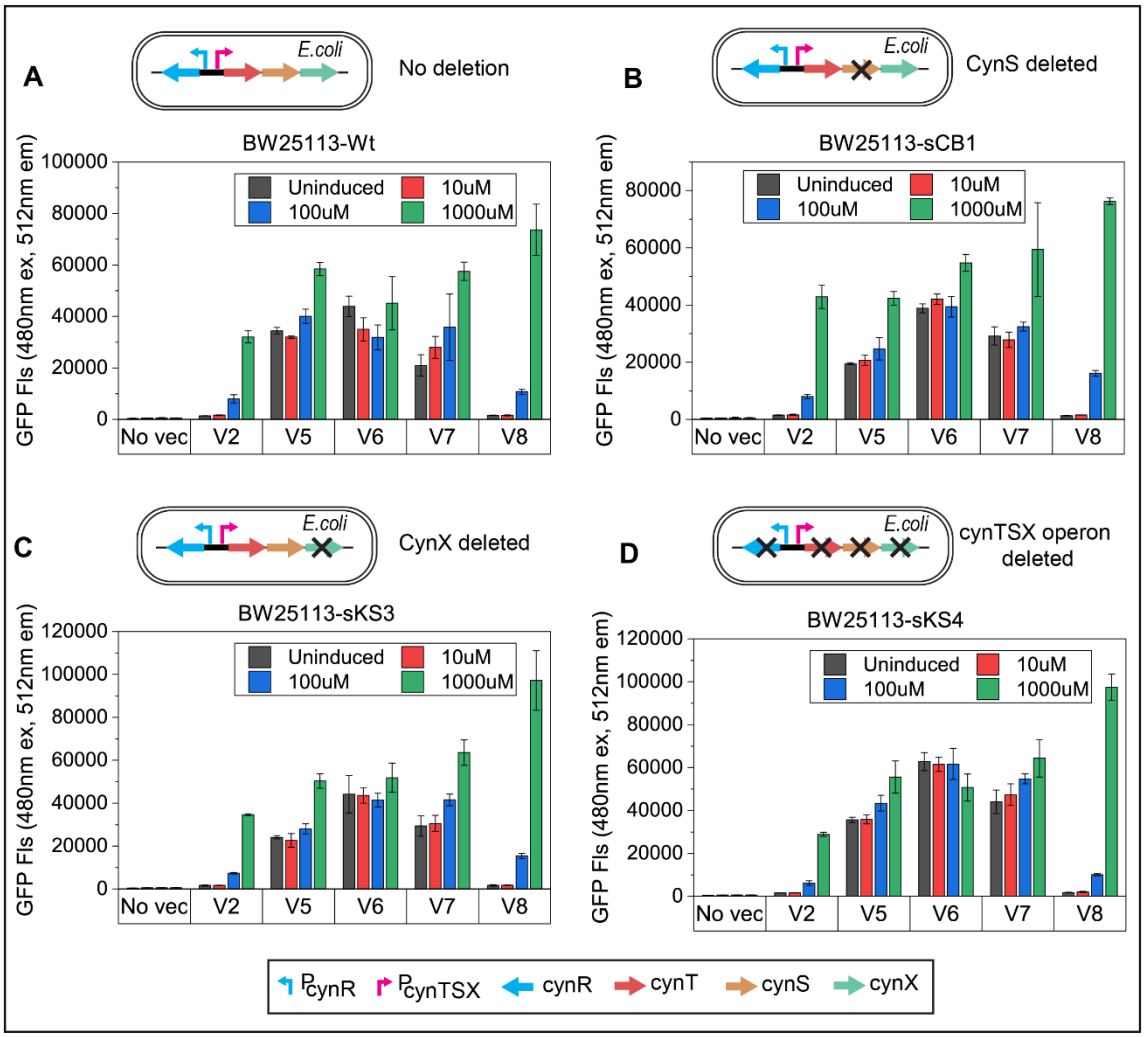



***Figure 29. Strain engineering and promoter design co-optimization to increase reporter GFP protein expression upon induction by azide.*** *Bar graphs representing GFP fluorescence intensities from cell lysates of engineered strains containing three distinct plasmid variants (v2, v5, v6, v7, v8) is shown. The cells were induced with varying amounts of azide (10 μM, 100μM, 1000μM).*

(v5, v6, v7) showed comparable fluorescence with respect to the v2 design, there was undesirable leaky background GFP expression seen even without inducer addition (**Figure 29**). The optimized plasmid designs (pCyn-v2-GFP and pCyn-v8-GFP) could now function as highly tunable synthetic biosensors for rapid *in vivo* detection of azide ions. The amount of GFP expressed within 2 hours of 1 mM azide induction in BW25113-Wt strain was estimated to be 0.4 mg and 0.9 mg from 1 ml culture for pCyn-v2-GFP and

pCyn-v8-GFP, respectively, based on the protein concentration and GFP fluorescence calibration curve.

### 3.4.5 Application of synthetic azide promoter towards heterologous protein expression and glycosynthase screening

The heterologous protein expression efficiency of the engineered promoter (pCyn-v2-GFP) was compared against the *E. coli lac* operon system [86,87]. The synthetic *cyn* promoter ($P_{cyn}$) in BW25113-Wt (**Figure 30A**) and *lac* promoter ($P_{lac}$) in BL21 cells (**Figure 30B**) were induced using sodium azide and isopropyl-β-D-1-thiogalactopyranoside (or IPTG), respectively, at various concentrations ranging between 0.01 mM to 1 mM. The amount of protein expressed in the $P_{lac}$ system was higher during the early time points after induction and at the lowest inducer concentrations. Even at lower IPTG concentrations (0.01 mM and 0.1 mM), GFP expression increased with induction time and reached a maximum after 24 hours of induction. Whereas we see a relatively lower GFP expression at similar low azide concentrations (**Figure 30A** and **Figure 30B**). However, we see a highly tunable reporter gene expression during azide induction as the amount of GFP expressed was closely proportional to inducer concentration, unlike IPTG. Also, the maximum GFP expression achieved after 24 hours was at least two-fold higher for the *cyn* promoter than *lac* promoter at inducer concentrations higher than 0.5 mM (**Figure 30A and Figure 30B**). These results showcase the utility of the novel engineered plasmid harboring the $P_{cyn}$ promoter as an alternative system for heterologous protein expression and other biotechnology applications.

In addition to application as a heterologous protein production system, the developed azide promoter can be useful for screening glycosynthases. But for that purpose, the proposed system should selectively identify inorganic azide amongst other chemically

linked organic azides (substrates). Azide ion is released as a by-product of the glycosynthase reaction [88] (for eg. Fucosynthase reaction by GH 29 family enzyme) and this release of azide can be potentially identified using the proposed *cyn* promoter system.

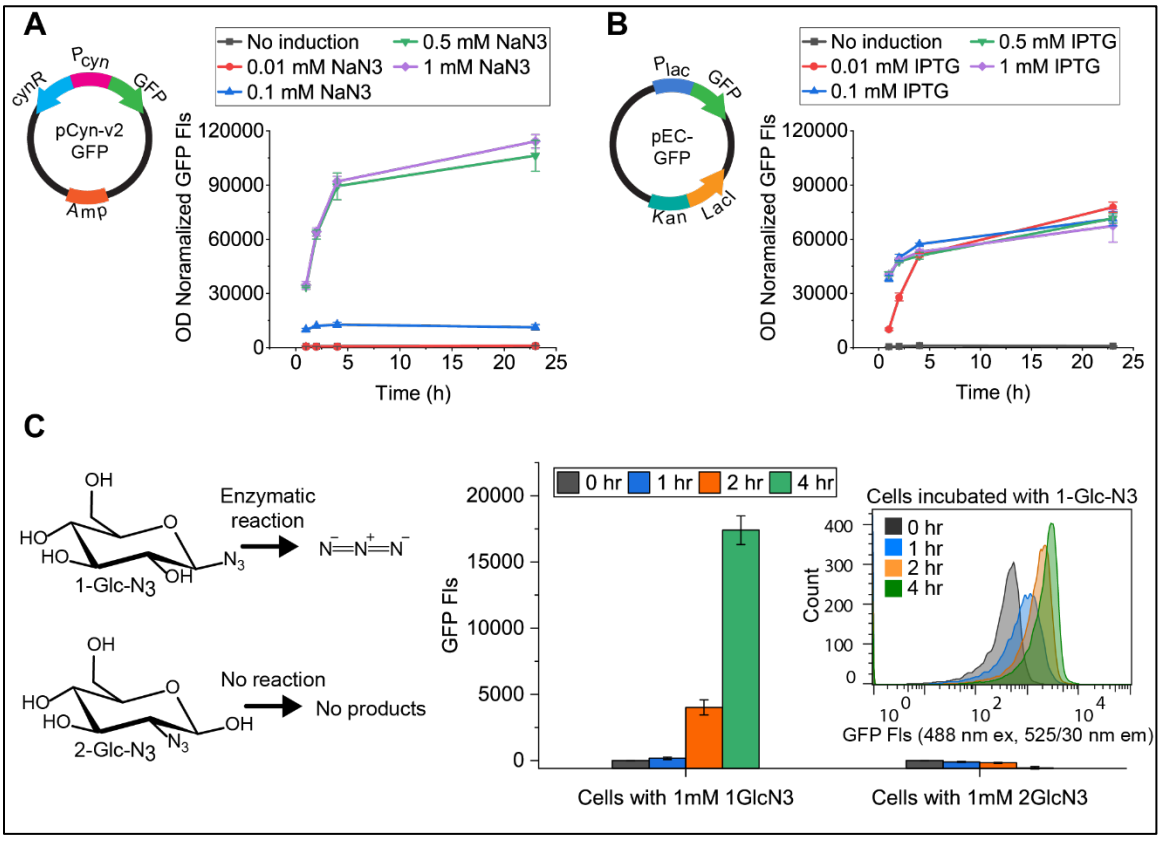



***Figure 30. Application of synthetic azide promoter towards heterologous protein expression and biorthogonal chemical-biology. (A)*** *E. coli BW25113-Wt containing pCyn-v2-GFP plasmid was induced with varying azide concentrations for 24 hours and GFP fluorescence of cell lysate was measured at various time points.* ***(B)*** *E. coli BL21 containing pEC-GFP plasmid was induced with different IPTG concentrations for 24 hours and GFP fluorescence of the cell lysate was measured at different time points.* ***(C)*** *E. coli BW25113-Wt cells with pCyn-v2-GFP plasmid was incubated with 1-azido-β-D-glucopyranosyl azide (1-Glc-$N_3$) and 2-Deoxy-2-azido-β-D-glucopyranosyl azide (2-Glc-$N_3$) and GFP fluorescence of respective cell lysates are shown as bar graphs. The inset depicts the flow cytometry analysis data for cells incubated with 1-Glc-N3 for 4 hours. Error bar indicates one standard deviation from reported mean values from three biological replicates.*

Carbohydrate-Active enZymes (CAZymes) such as glycosyl hydrolases and transglycosidases have been shown to release azide from azido-based sugars during glycosidic bond hydrolysis [89] and synthesis, [88–92] respectively. Multiple glycosyl

azides such as galactosyl-, glucosyl-, and mannosyl-azides have been reported to be hydrolyzed by galactosidases, glucosidases, and mannosidases, respectively.[89] Similarly, engineered glycosidases (i.e., transglycosidases and glycosynthases) have been shown to use azido-hexoses and other N-acetyl derivatives as donor sugars for oligosaccharides synthesis [88,91]. β-glucosidases from *Aspergillus* sp. and *Agrobacterium* sp. belonging to glycosyl hydrolase (GH) families GH3 and GH1 were reported to actively release azide ion from glucosyl azides.[90] Similarly, homologous β-glucosidases belonging to the GH1 and GH3 families present in native *E. coli* can potentially also show similar substrate specificity to release azide ions when incubated with glucosyl azides.

Here, we incubated *E. coli* BW25113-Wt cells containing the engineered pCyn-v2-GFP plasmid with 1-azido-β-D-glucopyranosyl azide (1-Glc-$N_3$) for 4 hours. As illustrated in **Figure 30C**, a small change in GFP fluorescence was seen in the first hour after which a rapid increase in fluorescence was observed owing to the release of azide when using 1-Glc-$N_3$. To verify if the detected GFP fluorescence is due to release of azide ions and not due to presence of azido-glucose, we also incubated cells with 2-Deoxy-2-azido-β-D-glucopyranosyl azide (2-Glc-$N_3$) analogous to the 2-deoxy-2-fluoro glycoside based β-glucosidase inhibitors [93]. No change in fluorescence was detected in the control samples since 2-Glc-$N_3$ is not hydrolyzed by β-glucosidases and hence cannot release free azide ions for induction of the engineered promoter. The preferential sensitivity of the designed promoter towards azide ions can therefore be used for *in vivo* glycoengineering to identify efficient CAZymes for glycosidic bonds synthesis and/or hydrolysis [38].

# Chapter 4: Conclusions and recommendations for future work

## 4.1 Conclusions

### 4.1.1 Transglycosylation engineering in GH5 family enzymes

GHs and GTs are the two major classes of enzymes that are known to catalyze the hydrolysis and/or synthesis of glycosidic linkages between carbohydrate moieties. Although both these enzyme classes are functionally different, their mode of action on glycosidic bonds follow similar basic design principles (i.e., donor glycone with a good leaving group is 'activated' in the enzyme active site to generate a short-lived intermediate or stable GEI and next a suitable acceptor molecule is attached to the activated/intermediate donor glycone group). This could explain why gene annotation/classification of these enzymes based on sequence similarities and protein folds often results in significant overlap between the two enzyme classes [11,94,95]. However, few studies have explored the role of CBMs on transglycosylation activity of GHs and none have so far explored the possibility of the reaction mechanism directly employing CBMs [57–59]. While CBMs or lectin-like domains have been never reported previously to directly participate in the catalytic reaction step for glycosidic bonds synthesis by GHs, these domains have been shown to play important roles in the functioning of several GTs. GTs are often associated with one or multiple CBMs/lectins-like domains on either the N- or C-terminal ends, where these domains participate in substrate recognition and in some cases participate in active site cleft formation to facilitate biocatalysis. For example, GTs responsible for mucin biosynthesis, namely N-acetyl galactosamine transferases (GalNAc-T1 [96] and GalNAc-T2 [46]), belonging to the GT27 family have a CBM13 domain tethered to the C-terminus of the catalytic domain. The two domains were

shown to dynamically interact with each other forming an active site cleft to facilitate addition of a N-acetyl-galactosamine donor group to an acceptor polypeptide chain [46]. The linkers in multi-domain GTs have also co-evolved with CBMs/lectin-like domains to stabilize these 'active' inter-domain conformations that enable efficient glycosidic bonds synthesis. But, we currently have a poor understanding of the dynamic interplay between multiple domains due to the inherent lack of available crystallographic data or the subtle impact of such dynamic domain interactions on overall catalytic turnover rates. The linker regions between most GHs and CBMs also tend to be much more flexible than GTs to allow for adjustments to complex structural motifs often found in polysaccharide and oligosaccharide type substrates (e.g., in plant cell walls) [97,98]. Nevertheless, multi-domain interactions are widely extant in CAZymes, and therefore likely play a significant but poorly understood role in modulating both GTs synthase and GHs hydrolase (and transglycosidase) activity [99]. Better understanding how similar GH-CBM chimeras allow subtle fine-tuning of transglycosylation mechanism that allow selective formation of glyco-synthase versus glyco-hydrolase based reaction products could help improve cellulosic biofuel production using engineered CBP microbes. Considering that most cellulolytic bacterial CAZymes are often appended to CBMs [100], it is likely that engineered cellulosomal enzyme complexes could utilize similar mechanisms to generate cello-oligosaccharides for synergistic uptake by CBP bacteria (like *C. thermocellum*) could allow for more efficient fermentation of cellulosic biomass into ethanol [101].

Here, we reported novel chimeric CBM-based enzyme designs that provide an unexplored route for chemoenzymatic synthesis of bespoke glycans like cellodextrins and pNP-based oligosaccharides. We have found that, similar to multi-domain GTs, engineered chimeric GH 5 catalytic domain scaffolds (in the absence of any true nucleophile residue) tethered by a linker to CBM-like domains can dynamically interact to form ‘active’ GH-CBM

complexes that facilitated synthesis of glucose based oligosaccharides from activated donor sugar monomers. We observed that the total transglycosylation to hydrolysis (T/H) product ratio for the CelE-E316G-CBM3a construct was 40 to 140-fold higher than the values obtained for the control wild-type enzymes. Clearly our chimeric mutant is a highly efficient transglycosidase based on T/H metrics established for other CAZyme systems [102,103]. Previous studies on another glucosidase have reported similar order of magnitude (70-fold) increase in T/H ratio but only after extensive mutagenesis and directed evolution of the wild-type enzymes [104], which further highlights the advantages offered by our current approach. This approach could be used to also target highly efficient synthesis of other non-cellulosic glycan polymers using readily accessible pNP-glycosides or naturally available aromatic leaving group based glycosides as substrates [105]. However, future work will need to also focus on fully unraveling the reaction mechanism and showcasing how similar CBM-/lectin-like domain assisted mechanism could be prevalent in other GH families as well.

In the case of CelE-E316G-CBM3a construct, the CBM3a domain is speculated to stabilize the leaving group (pNP) near the active site substrate binding cleft of CelE allowing for the first transglycosylation reaction to take place forming pNP-cellotetraose or similar pNP based oligosaccharides. While, similar transglycosylation products are likely also formed for native CelE-CBM3a, these products are nearly instantly degraded to produce the equilibrium dominant products (i.e., pNP and cellobiose). In the absence of the native nucleophilic residue (E316), significantly higher concentrations of transglycosylation products are seen even as the reaction tends towards equilibrium. Here, the rate-limiting step is hypothesized to be the formation of the active CelE-E316G-CBM complex that facilitates the synthesis of the higher molecular weight transglycosylation products. This hypothesis is consistent with the critical importance of

the linker identified that likely facilitates highly concerted dynamic, but poorly understood, interactions between the GH, CBM, and substrate. Since $S_N$i-like/$S_N$1 and $S_N$2 type mechanisms are often seen to operate in retaining and inverting GTs [10], respectively, it is likely that similar mechanisms could operate in GHs as well. However, only $S_N$1 and $S_N$2 type mechanisms have been reported for GHs, but with one recent exception [45]. However, it is also possible that there is a likely pseudo-nucleophile residue that participates in a conventional $S_N$2-like reaction mechanism (i.e., after mutation of E316G). Nevertheless, our findings suggest that researchers should more closely investigate the role of 'true' nucleophilic residues for each specific GH families, which is likely substrate-dependent as well.

### 4.1.2 Development of azide biosensor toolkit for protein expression and glycosynthase screening.

In summary, we have demonstrated the design and engineering of an azide inducible promoter system for *E. coli*. This azide promoter system allows tunable expression by varying the inducer (azide ion) concentration and outperforms conventional lactose/IPTG based system for heterologous reporter GFP expression. Additionally, the developed toolkit functions as a biosensor for detecting the presence of azide ions inside living cells. This would allow employing this tool for engineering CAZymes such as glycosyl hydrolases which use azido-sugars as substrates and autonomously monitor released azide ions upon substrate hydrolysis or transglycosylation. Furthermore, this biosensing system can be evolved for use with other prokaryotic or eukaryotic cell protein expression systems by replacing the current -10 and -35 sequences with specific target RNA polymerase recognition sequence. Adaptation of an azide specific promoter system for

other cell types would be beneficial overall for diverse drug development and chemical-biology focused research communities

## 4.2 Recommendation for future work

Carbohydrate binding domain (CBM3a) helps recover the transglycosylation activity of mutant CelE-E316G through a mechanism that could be either $S_N2$ or $S_Ni$-like. To decode the exact mechanism involved, inhibitor-based studies should be performed. Incubating the CelE-E316G-CBM3a with an inhibitor would reveal any potential nucleophiles by forming an enzyme-substrate complex that can be detected by mass spectrometer. The CBM3a engineered transglycosylation was successfully extended to other homologous GH5 family enzymes. To show the widespread application of this approach, CBM3a should be fused with other GH family enzymes and tested for transglycosylation activity.

The universal high-though put azide ion biosensing screening method has been shown to selectively identify by-products of GS reaction. This system could be tested for GH29 fucosynthase engineering using a dual plasmid *E.coli* system where fucosynthase is controlled by lac operator and GFP reporter gene is induced by azide ion. An active GH29 fucosynthase construct will release azide ion during the glycosynthase reaction and azide ion induces GFP expression that can be detected using flow cytometer. The dual plasmid system can be introduced in *E.coli* and directed evolution on fucosynthase gene can be performed to identify improved variants.

# Appendix

## A1.1 Supplementary methods

### A1.1.1 Site directed mutagenesis

The protocol followed for site directed mutagenesis was developed by Stratagene. Polymerized chain reaction was performed using 1X Phusion master mix, template DNA and 0.5 µM of two complementary primers. After 20 cycles, the reaction mixtures were incubated with Dpn1 in 1X cut smart buffer at 37°C for 1 hr and transformed into E. cloni 10G competent cells and plated on LB agar plates supplemented with Kanamycin (50 µg/ml) to get selectively transformed colonies. Plasmids were isolated from individual transformed colonies and sequenced to confirm the nucleotide identity at the mutational site.

### A1.1.2 Sequence and ligation independent cloning

The protocol used for sequence and ligation independent cloning was adapted from Li et.al [106] and Stevenson et.al [78]. PCR amplifications of vector DNA and insert DNA were performed using their corresponding SLIC primers. The vector and insert PCR products after verification for amplification using DNA gel electrophoresis were washed to remove the unreacted nucleotides (dNTPs). The reaction mixture containing 0.025 pmol of vector and 0.0625 pmol insert purified products (2.5:1 insert: vector ratio) were digested using Dpn1 at 37°C for 1 hour. Dpn1 digested products were reacted with 1.5U of T4 DNA polymerase at 25°C for 5 mins and immediately placed on ice. Entire reaction mixture was transformed into E. cloni 10g cells and plated to get the colonies. The transformant colonies from the plate were screened using PCR amplification for identifying constructs

with correct insert size observed in agarose gel electrophoresis. Plasmids from the selected colonies were isolated and sequenced to verify the nucleotide sequence.

### A1.1.3 Transformation of plasmid DNA

50 µl of the competent cells were taken in pre-cooled sterile microtubes. The cells were incubated on ice for 10 mins. 50 ng of the plasmid DNA was added to the tube and incubated on ice for 10 mins. The cells were then subjected to heat shock at 42 °C for 45 secs and immediately incubated on ice for 10 minutes. After incubation, 250 µl of SOC media was added and the cultures were incubated at 37 °C for 2 hours. 100 µl of cells were then plated on respective LB agar plates.

### A1.1.4 Flow cytometry analysis

Before running the samples, perform two cleaning cycles. QC bottle and the beads are stored in the fridge. It is necessary for the beads to be at room temperature before staining them (at lower temperatures, beads tend to clump and give inaccurate results). Beside the flow cytometer, there are two bottles labeled as ICF and empty. The bottle labeled with ICF must be filled with ICF (Instrument Cleaning Fluid) almost till the top and the bottle labeled with empty must be emptied before starting the cleanup. Now open the software on the desktop Guavasoft 3.3. Select the cleaning option under essential tools and click on start cleaning. As soon as you click on start cleaning, a tray message dialogue box will appear, and the tray will open in the instrument. Set the tray of the instrument the same way it shows in the software. Select the rest position of the sip. Always make sure there is a blue tube full of DI water in rest position of the sip otherwise the SIP will dry, and air will enter the system. The blue tubes will be filled with DI water, the purple and yellow tubes with ICF. All the tubes are supposed to be filled till the very top. Select the rest position in the template as the well where the SIP has to rest and press Start. After two

cleaning cycles, go to easyCheck, a dialog box will open (Note: The number of particles detected in QC should be close to 50,000 particles/ml). Prepare the QC sample as follows: Add 12 ul beads (which should be at room temperature) to 188 ul diluent and mix them well. Vortex at least for 10 seconds. Transfer the beads to one of the wells in the plate. Before transferring, vortexing and pipette mixing is a must. First transfer the diluent to the plate and then mix and vortex the beads and immediately add and run it immediately by pressing on OK (If the beads stay in the plate for long, they will settle down and give misleading results). Now click on start, a dialog box will open. Make sure to keep the plate in the correct orientation or the SIP might break. Select the well which contains beads. Click OK. QC will start. Three replicates will be done. The numbers we get in average must be in the range of numbers provided on the label of the flow cytometer. Eject the tray after QC and take out the plate, which was used for QC. Close the window. Open InCyte and a new window will appear on the screen. Click on Edit Worklist and select the wells where your samples are located. For the DO CLEAN option, click at the end of every 24 samples in case you are running more than 24 samples. Click on Run Worklist. Prepare for acquisition windows. Now adjust the settings (gates, gain, fluorescence wavelengths, sample flow rate). Make sure that during adjust settings, the SIP does not consume the entire volume of the sample inside the well because that might lead to air entering the system which might spoil Guava flow cytometer. As soon as you are done adjusting the settings, click on Run Worklist which will run all the samples. Wait as the samples are being run on the flow cytometer. Clean thrice once you are done using the flow cytometer. Collect the fluorescence data and export the file and analyze using easyCyte.

## A1.2 Supplementary information

**Supplementary Text S1. Protein sequences of all major constructs used for CelE-CBM3a study.**

>> CelE-Wt

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAG

>> CelE-E316A

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGAFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAG

>> CelE-E316S

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGSFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAG

>> CelE-E316G

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNAPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGGFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAG

>> CelE-Wt-CBM3a-42aaLinker

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE- E316A-CBM3a-42aaLinker

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGAFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE- E316S-CBM3a-42aaLinker

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGSFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE- E316G-CBM3a-42aaLinker

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGGFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE-Wt-CBM1-42aaLinker

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTLKPGPTQSHYGQCGGIGYSGPTVCASGTTC
QVLNPYYSQCL

>> CelE-Wt-CBM17-42aaLinker

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA

SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTLKSQPTAPKDFSSGFWDFNDGTTQGFGVNP
DSPITAINVENANNALKISNLNSKGSNDLSEGNFWANVRISADIWGQSINIYGDTKLTMD
VIAPTPVNVSIAAIPQSSTHGWGNPTRAIRVWTNNFVAQTDGTYKATLTISTNDSPNFNT
IATDAADSVVTNMILFVGSNSDNISLDNIKFTK

>> CelE-Wt-CBM3a- 6aaLinker

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKVSGNLKV
EFYNSNPSDTTNSINPQFKVTNTGSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGS
NGSYNGITSNVKGTFVKMSSSTNNADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNY
TQSNDYSFKSASQFVEWDQVTAYLNGVLVWGKEPG

>> CelE-Wt-CBM3a-11aaLinker

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTVS
GNLKVEFYNSNPSDTTNSINPQFKVTNTGSSAIDLSKLTLRYYYTVDGQKDQTFWCDH
AAIIGSNGSYNGITSNVKGTFVKMSSSTNNADTYLEISFTGGTLEPGAHVQIQGRFAKND
WSNYTQSNDYSFKSASQFVEWDQVTAYLNGVLVWGKEPG

>> CelE-Wt-CBM3a-21aaLinker

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATVSGNLKVEFYNSNPSDTTNSINPQFKVTNTGSSAIDLSKLTLRYYYTVDGQ
KDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTNNADTYLEISFTGGTLEPGAHV
QIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTAYLNGVLVWGKEPG

>> CelE-Wt-CBM3a-Flex1

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATGTKGATGTNT
ATGTKSATATGTRGSVGTNTGTNTGANTGVSGNLKVEFYNSNPSDTTNSINPQFKVTN

TGSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSST
NNADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVT
AYLNGVLVWGKEPG

>> CelE-Wt-CBM3a-Flex2

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNASGGKGATGGN
TAGGTKSATAGGSRGSVGGNSGTNGGANGGVSGNLKVEFYNSNPSDTTNSINPQFKV
TNTGSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSS
STNNADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQ
VTAYLNGVLVWGKEPG

>> CelE-Wt-CBM3a-Rig1

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKPATPTNT
PTPTKPATATPTRPPVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE-Wt-CBM3a-Rig2

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLGGSAEAAAKAAE
AAAKAAEAAAKAAEAAAKAAEAAAKASGGVSGNLKVEFYNSNPSDTTNSINPQFKVTN
TGSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSST
NNADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVT
AYLNGVLVWGKEPG

>> CelE-Wt-CBM3a-Y458A

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN

NADTALEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE-Wt-CBM3a-E521A

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKAPG

>> CelE-Wt-CBM3a-R407A

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLAYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE-Wt-CBM3a-T509A

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVAA
YLNGVLVWGKEPG

>> CelE-Wt-CBM3a-R409A

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYAYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN

NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE-Y273A-CBM3a

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPAFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE-Y270A-CBM3a

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAASPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE-H268A-CBM3a

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIAAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE-W203A-CBM3a

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEAMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWWDNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN

NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

>> CelE-W349A-CBM3a

GMRDISAIDLVKEIKIGWNLGNTLDAPTETAWGNPRTTKAMIEKVREMGFNAVRVPVTW
DTHIGPAPDYKIDEAWLNRVEEVVNYVLDCGMYAIINVHHDNTWIIPTYANEQRSKEKLV
KVWEQIATRFKDYDDHLLFETMNEPREVGSPMEWMGGTYENRDVINRFNLAVVNTIRA
SGGNNDKRFILVPTNAATGLDVALNDLVIPNNDSRVIVSIHAYSPYFFAMDVNGTSYWG
SDYDKASFTSELDAIYNRFVKNGRAVIIGEFGTIDKNNLSSRVAHAEHYAREAVSRGIAV
FWADNGYYNPGDAETYALLNRRNLTWYYPEIVQALMRGAGVEGLNATPTKGATPTNT
ATPTKSATATPTRPSVPTNTPTNTPANTPVSGNLKVEFYNSNPSDTTNSINPQFKVTNT
GSSAIDLSKLTLRYYYTVDGQKDQTFWCDHAAIIGSNGSYNGITSNVKGTFVKMSSSTN
NADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDYSFKSASQFVEWDQVTA
YLNGVLVWGKEPG

**Supplementary Text S2. Sequence for promoter regions used for Azide biosensor study**

>> Promoter region for pCyn-v1-GFP

TCGCAACCTATAAGTAAATCCAATGGAACTCATCATAAATGAGACTTTTACCTTATGA
CAATCGGCGAGTAGTCTGCCTCTCATTCCAGAGACAGACAGAGGTTAACGATG

>> Promoter region for pCyn-v2-GFP

GGTTCCTCATTACCGTTATCATATGAACACACCATAACAAAGATGCATGCAGCTGTC
TAAATCCCGCGGCCATGGCGGCCGGGAGCATGCGACGTCGGGCCCAATTCGCCC
GATCTTAATGAATGGCCGGAAGAGGTACGGACGCGATATGCGGGGGTGAGAGGG
CAAATAGGCAGGTTCGCCTTCGTCACGCTAGGAGGCAATTCTATAAGGATCCTCGC
AACCTATAAGTAAATCCAATGGAACTCGTCAGAAATGAGACTTTTACCTTATGACAA
TCGGCTGGTATAATGCCTCTACTTCCAGAGACAGACATAAGGAGATTACGCATG

>> Promoter region for pCyn-v3-GFP

GGTTCCTCATTACCGTTATCATATGAACACACCATAACAAAGATGCATGCAGCTGTC
TAAATCCCGCGGCCATGGCGGCCGGGAGCATGCGACGTCGGGCCCAATTCGGGA
TCCTCGCAACCTATAAGTAAATCCAATGGAACTCGTCAGAAATGAGACTTTTACCTT
ATGACAATCGGCTGGTATAATGCCTCTACTTCCAGAGACAGACATAAGGAGATTAC
GCATG

>> Promoter region for pCyn-v4-GFP

GGTTCCTCATTACCGTTATCATATGAACACACCATAACAAAGATGCATGCAGCTGTC
TAAATCCCGCGGCCATGGCGGCCGGGAGCATGCGACGTCGGGCCCAATTCGCCC
GATCTTAATGAATGGCCGGAAGAGGTACGGACGCGATATGCGGGGGTGAGAGGG
CAAATAGGCAGGTTCGCCTTCGTCACGCTAGGAGGCAATTCTATAAGATTGGACCG
TACGCATGTCAAACTGCTGGCGAACCGCGATTCCACGACCGGTGCACGATTTAACT
ACGCCGACGTGACGACATTCCTGCTAATGCCTCGCCCGCCGGACCGCCCTCGTGA
TGGGGTAGCTGGGCATGACCTTGTGACATATAACGAGAGTCTACTTGTTTAATCATC

TCACGGCGAAAGTCGGGGGGACAGCAGCCGCTGCAGACATTATACCGCAACTACA
CCAAGCTGAGATAACTCCGTAGTTGACTACGCATCCCTCTAGGCCTTACTTAACCG
GATACAGTGACTTTGACAGGTTTGTGGGCTACAGCAATCACTTGCATAGCTGCGTA
TGGAGGAAGCAACTCTTGGGTGTTAGTATGTTGACCCCTGTATTAGGGATGCGGGT
AGTAGATGTGGGCAGAGACACCCAGGTCAAGTACACGACCCTCTCGTAGGAGGTG
TTCCAGATCACCATACCACCATACCATTCGAGCATGGCACTATGTACGCTGTCCCC
ATTCTGGTAGTCATCATCCCTATCACGGTTTCGAGTGACTGGTGACGGATATCCCC
CACGAATGGAGATCTTATTCACAGTCGGTCACATTGGAGTGCTCCTTGACTAATCA
GCTTGGCCAGGTCTGTTGGGCCTCCGTGCCCCGAGTTTCGGCGCTGTGCTGCCGA
GAGTCGGCCATTGTCATTGGGGCCTCACTTGTGGATACCCCGACCTATTTTGACGG
GACCACTCGCGGTAGTCGTTGGGCTTATGCACCGTGAAGTCCTCCGCCGGCCTCC
CCCCTACAAAAGATGATAAGCTCCGGCAAGCAATATTGAACAACGCAAGGATCGGC
GATATAAACAGAGAAACGGCTGATTACTCTTGTTGGTGTGGTATCGCTAAACTGGG
ATCCTCGCAACCTATAAGTAAATCCAATGGAACTCGTCAGAAATGAGACTTTTACCT
TATGACAATCGGCTGGTATAATGCCTCTACTTCCAGAGACAGACATAAGGAGATTAC
GCATG

>> Promoter region for pCyn-v5-GFP

GGTTCCTCATTACCGTTATCATATGAACACACCATAAGGAAGATGCATGCAGCTGTC
TAAATCCCGCGGCCATGGCGGCCGGGAGCATGCGACGTCGGGCCCAATTCGCCC
GATCTTAATGAATGGCCGGAAGAGGTACGGACGCGATATGCGGGGGTGAGAGGG
CAAATAGGCAGGTTCGCCTTCGTCACGCTAGGAGGCAATTCTATAAGGATCCTCGC
AACCTATAAGTAAATCCAATGGAACTCGTCAGAAATGAGACTTTTACCTTATGACAA
TCGGCTGGTATAATGCCTCTACTTCCAGAGACAGACATAAGGAGATTACGCATG

>> Promoter region for pCyn-v6-GFP

GGTTCCTCATTACCGTTATCATATGAACACGCCATAAGGAAGATGCATGCAGCTGT
CTAAATCCCGCGGCCATGGCGGCCGGGAGCATGCGACGTCGGGCCCAATTCGCC
CGATCTTAATGAATGGCCGGAAGAGGTACGGACGCGATATGCGGGGGTGAGAGG
GCAAATAGGCAGGTTCGCCTTCGTCACGCTAGGAGGCAATTCTATAAGGATCCTCG
CAACCTATAAGTAAATCCAATGGAACTCGTCAGAAATGAGACTTTTACCTTATGACA
ATCGGCTGGTATAATGCCTCTACTTCCAGAGACAGACATAAGGAGATTACGCATG

>> Promoter region for pCyn-v7-GFP

GTTCCTCATTACCGTTATCATATGAACACACCATAAGGAAAAGATGCATGCAGCTGT
CTAAATCCCGCGGCCATGGCGGCCGGGAGCATGCGACGTCGGGCCCAATTCGCC
CGATCTTAATGAATGGCCGGAAGAGGTACGGACGCGATATGCGGGGGTGAGAGG
GCAAATAGGCAGGTTCGCCTTCGTCACGCTAGGAGGCAATTCTATAAGGATCCTCG
CAACCTATAAGTAAATCCAATGGAACTCGTCAGAAATGAGACTTTTACCTTATGACA
ATCGGCTGGTATAATGCCTCTACTTCCAGAGACAGACATAAGGAGATTACGCATG

>> Promoter region for pCyn-v8-GFP

GGAGAGTTCCTCATTACCGTTATCATATGAACACACCATAAGGAAGATGCATGCAG
CTGTCTAAATCCCGCGGCCATGGCGGCCGGGAGCATGCGACGTCGGGCCCAATTC
GCCCGATCTTAATGAATGGCCGGAAGAGGTACGGACGCGATATGCGGGGGTGAGA
GGGCAAATAGGCAGGTTCGCCTTCGTCACGCTAGGAGGCAATTCTATAAGGATCCT
CGCAACCTATAAGTAAATCCAATGGAACTCGTCAGAAATGAGACTTTTACCTTATGA
CAATCGGCTGGTATAATGCCTCTACTTCCAGAGACAGACATAAGGAGATTACGCAT
G

## A1.3 Collaborative projects

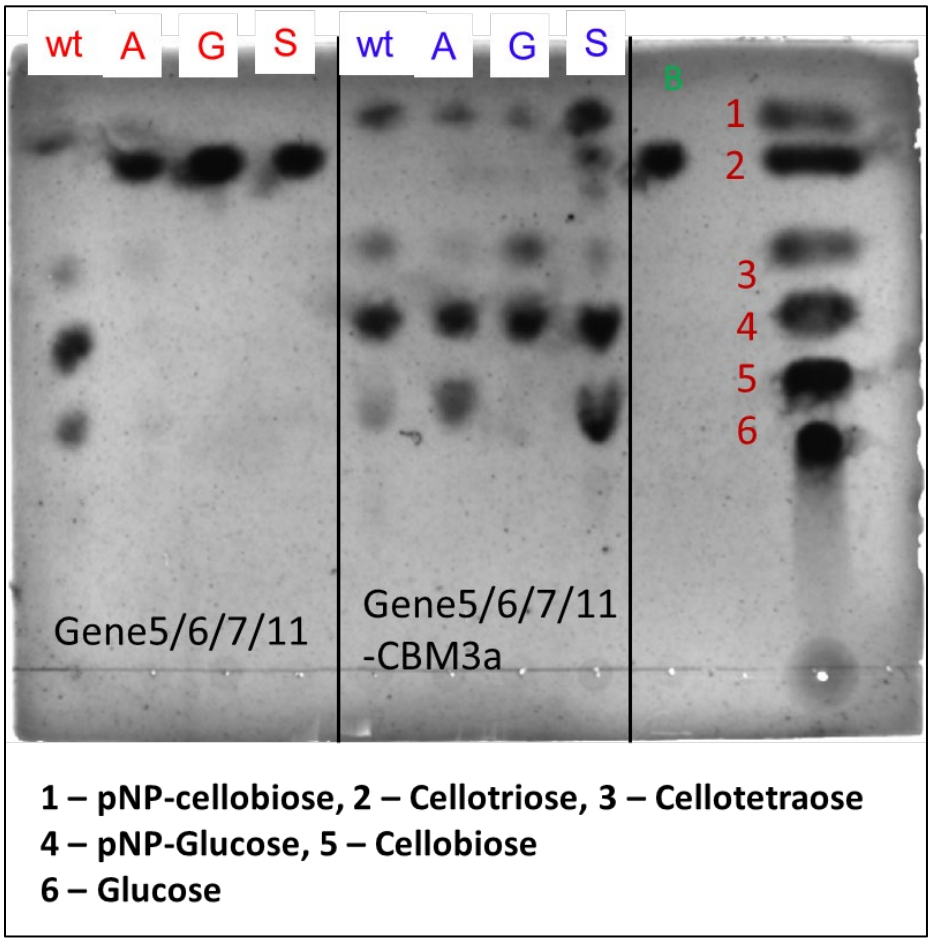
### A1.3.1 Work with Madhura Kasture

The CBM3a assisted transglycosylation engineering as observed for CelE-E316G-CBM3a was tested for 15 homologous enzymes from GH5 family. All the representative 15 genes with and without CBM3a were cloned, expressed and purified. The purified enzyme activity assays were performed using pNP-CB as substrate and TLC analysis was performed to analyze the reaction products. It was observed that CBM3a assisted transglycosylation was seen across various genes suggesting that this could be used as a general protein engineering strategy for GH-CBM chimeras.

***Appendix Table 1. GH 5 family enzymes homologous to CelE that were tested for CBM3a engineering for transglycosylation.***

| Sample No | GH family | Bacteria |
|---|---|---|
| Gene1 | GH5_4 | *Ruminococcus flavefaciens* |
| Gene2 | GH5_unknown | *Treponema sp. JC4* |
| Gene3 | GH5_4 | *Cellulosilyticum ruminicola* |
| Gene4 | GH5_unknown | *Ruminococcus albus* |
| Gene5 | GH5_4 | Uncultured bacterium |
| Gene6 | GH5_4 | Uncultured bacterium |
| Gene7 | GH5_4 | *Clostridium acetobutylicum* |
| Gene8 | GH5_4 | *Polyplastron multivesiculatum* |
| Gene9 | GH5_unknown | *Eubacterium siraeum* |
| Gene10 | GH5_4 | uncultured *Butyrivibrio sp* |

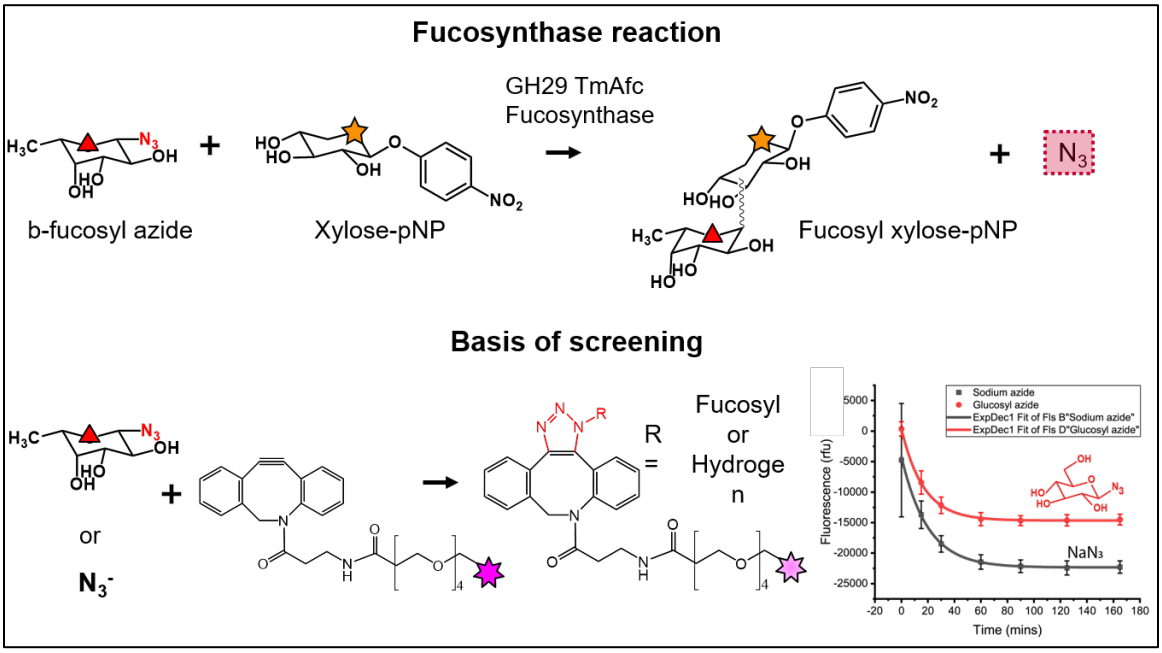
| | | |
|---|---|---|
| Gene11 | GH5_4 | Uncultured bacterium |
| Gene12 | GH5_4 | *Clostridium cellulovorans* |
| Gene13 | GH5_4 | *Paenibacillus pabuli* |
| Gene14 | GH5_2 | *Salipaludibacillus agaradhaerens* |
| Gene15 | GH5_unknown | *Eubacterium cellulosolvens* |



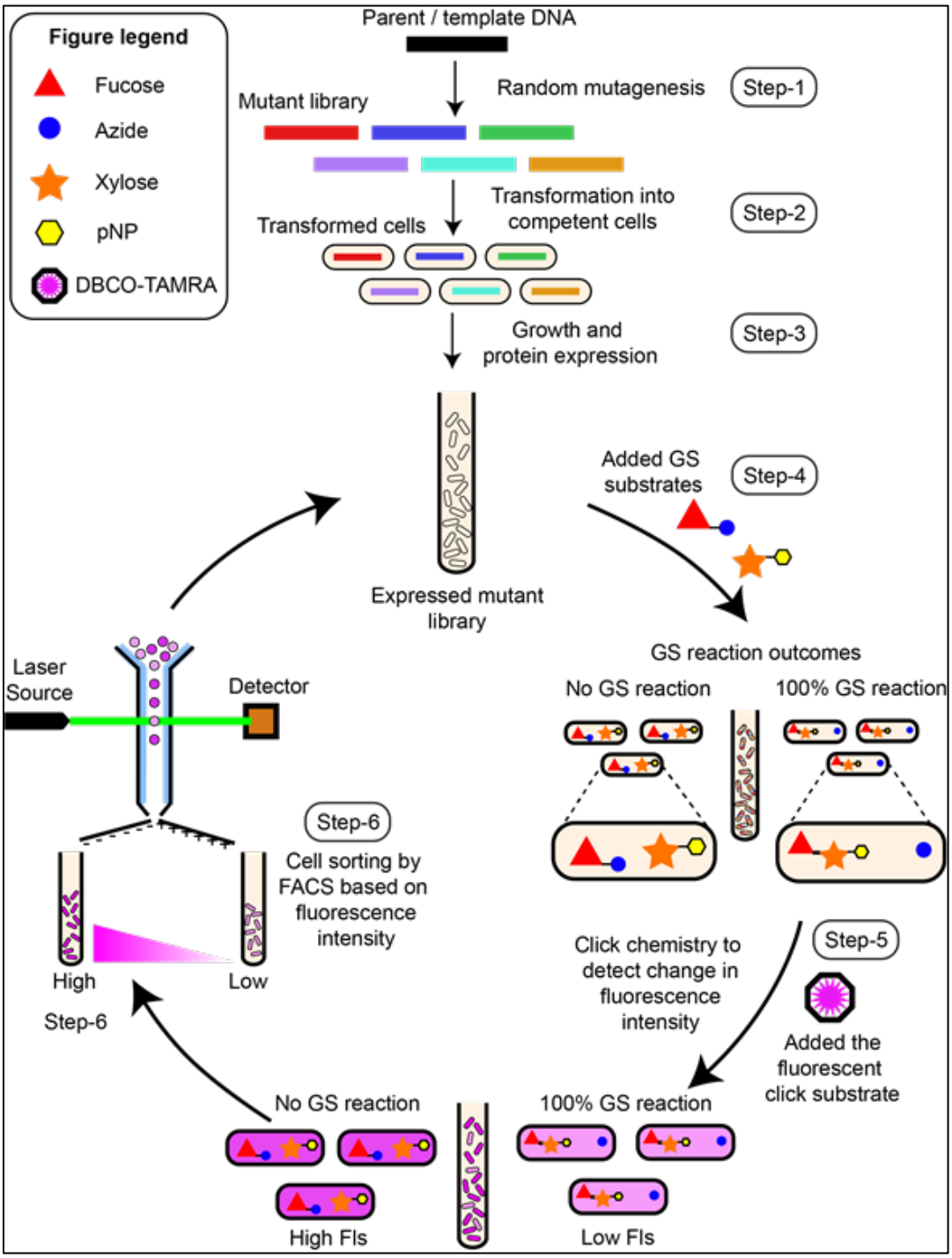

***Appendix Figure 1. Representative TLC analysis of reaction mixtures which showed CBM3a assisted transglycosylation. Here, Gene5, Gene6, Gene7 and Gene11 showed similar results and hence a sample TLC image is shown here.***

### A1.3.2 Work with Ayushi Agrawal

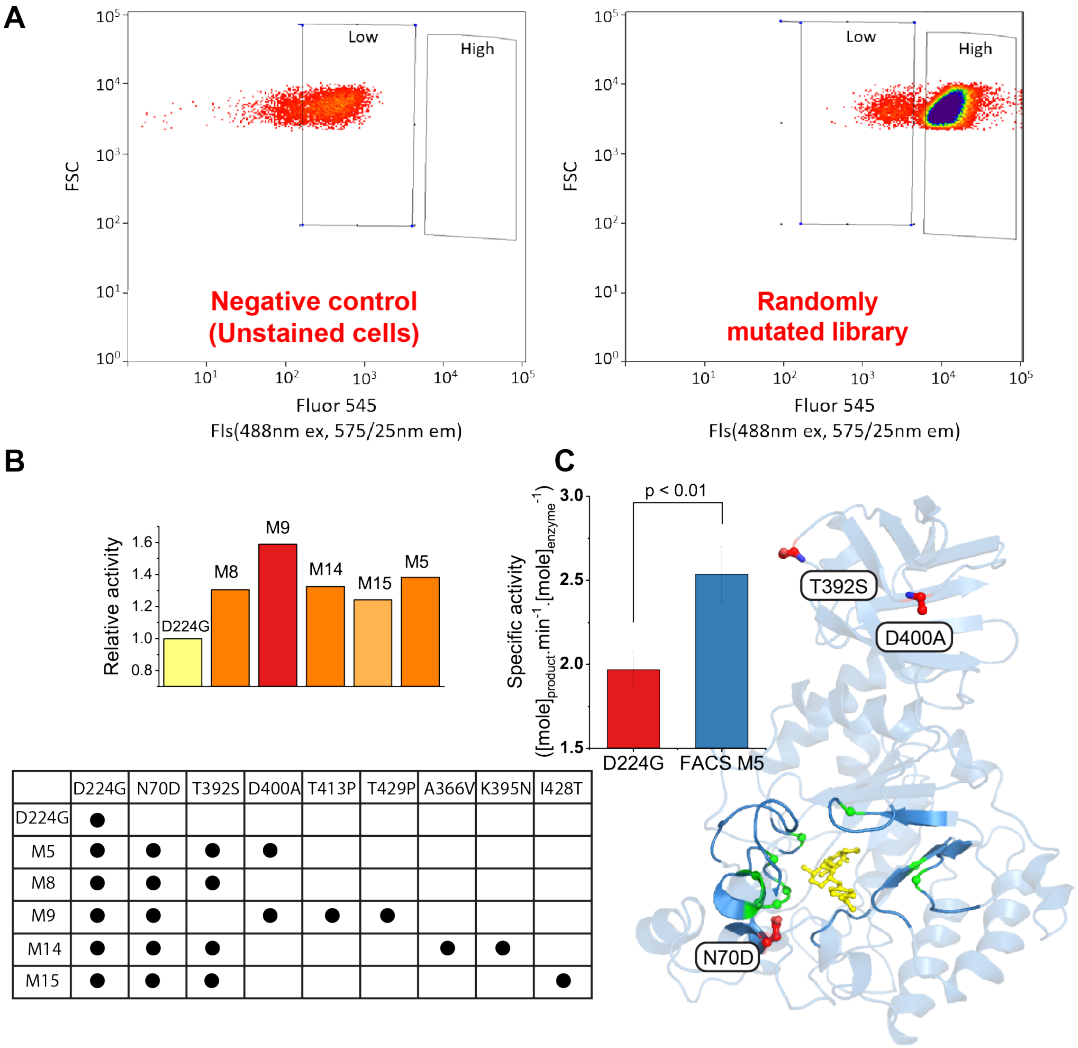
A fluorescence-based approach employing click-chemistry for the selective detection of glycosyl azides as substrates versus free inorganic azides as reaction products to facilitate ultrahigh-throughput *in-vivo* single cell-based assay of glycosynthase activity was developed. This assay was developed based on the distinct differences observed in relative fluorescence intensity of the triazole-containing fluorophore product formed during the click-chemistry reaction of organic glycosyl azides versus inorganic azides. The GH29 fucosynthase enzyme was subjected to random mutagenesis and this screening method was used to identify improved variant.




***Appendix Figure 2. Click chemistry based fluorescent method to selectively detect azide ions.***




***Appendix Figure 3. Overview of directed evolution of GH29 fucosynthase enzyme***




| | D224G | N70D | T392S | D400A | T413P | T429P | A366V | K395N | I428T |
|---|---|---|---|---|---|---|---|---|---|
| D224G | ● | | | | | | | | |
| M5 | ● | ● | ● | ● | | | | | |
| M8 | ● | ● | ● | | | | | | |
| M9 | ● | ● | | ● | ● | ● | | | |
| M14 | ● | ● | ● | | | | ● | ● | |
| M15 | ● | ● | ● | | | | | | ● |

***Appendix Figure 4. FACS sorting data for GH29 random mutant library.***

# References


1. Varki A, Gagneux P: **Biological Functions of Glycans**. In *Essentials of Glycobiology*. Edited by Varki A, Cummings R., Esko J, et.al. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2017:Chapter 7.

2. Abdel-Motal UM, Wang S, Awad A, Lu S, Wigglesworth K, Galili U: **Increased immunogenicity of HIV-1 p24 and gp120 following immunization with gp120/p24 fusion protein vaccine expressing α-gal epitopes**. *Vaccine* 2010, **28**:1758–1765.

3. Elliott S, Lorenzini T, Asher S, Aoki K, Brankow D, Buck L, Busse L, Chang D, Fuller J, Grant J, et al.: **Enhancement of therapeutic protein in vivo activities through glycoengineering**. *Nat Biotechnol* 2003, **21**:414–421.

4. Galili U: **Amplifying immunogenicity of prospective Covid-19 vaccines by glycoengineering the coronavirus glycan-shield to present α-gal epitopes**. *Vaccine* 2020, **38**:6487–6499.

5. Mootoo DR, Konradsson P, Udodong U, Fraser-Reid B: **"Armed" and "Disarmed" n-Pentenyl Glycosides in Saccharide Couplings Leading to Oligosaccharides**. *J Am Chem Soc* 1988, doi:10.1021/ja00224a060.

6. Codée JDC, Van Den Bos LJ, Litjens REJN, Overkleeft HS, Van Boom JH, Van Der Marel GA: **Sequential one-pot glycosylations using 1-hydroxyl and 1-thiodonors**. *Org Lett* 2003, **5**:1947–1950.

7. Joseph AA, Pardo-Vargas A, Seeberger PH: **Total Synthesis of Polysaccharides by Automated Glycan Assembly**. *J Am Chem Soc* 2020, **142**:8561–8564.

8. Seeberger PH: **Automated oligosaccharide synthesis**. *Chem Soc Rev* 2008,

**37**:19–28.

9. Krasnova L, Wong CH: *Understanding the Chemistry and Biology of Glycosylation with Glycan Synthesis*. 2016.

10. Lairson LL, Henrissat B, Davies GJ, Withers SG: **Glycosyltransferases: Structures, Functions, and Mechanisms**. *Annu Rev Biochem* 2008, **77**:521–555.

11. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B: **The carbohydrate-active enzymes database (CAZy) in 2013**. *Nucleic Acids Res* 2014, **42**:490–495.

12. Boltje TJ, Buskas T, Boons GJ: **Opportunities and challenges in synthetic oligosaccharide and glycoconjugate research**. *Nat Chem* 2009, **1**:611–622.

13. Yang Z, Wang S, Halim A, Schulz MA, Frodin M, Rahman SH, Vester-Christensen MB, Behrens C, Kristensen C, Vakhrushev SY, et al.: **Engineered CHO cells for production of diverse, homogeneous glycoproteins.** *Nat Biotechnol* 2015, **33**:842–844.

14. Hamilton SR, Davidson RC, Sethuraman N, Nett JH, Jiang Y, Rios S, Bobrowicz P, Stadheim TA, Li H, Choi B-K, et al.: **Humanization of yeast to produce complex terminally sialylated glycoproteins.** *Science* 2006, **313**:1441–3.

15. Valderrama-Rincon JD, Fisher AC, Merritt JH, Fan Y-Y, Reading CA, Chhiba K, Heiss C, Azadi P, Aebi M, DeLisa MP: **An engineered eukaryotic protein glycosylation pathway in Escherichia coli.** *Nat Chem Biol* 2012, **8**:434–6.

16. McArthur JB, Chen X: **Glycosyltransferase engineering for carbohydrate synthesis**. *Biochem Soc Trans* 2016, **44**:129–142.

17. Kono H, Waelchli MR, Fujiwara M, Erata T, Takai M: **Transglycosylation of cellobiose by partially purified Trichoderma viride cellulase**. *Carbohydr Res* 1999, **319**:29–37.

18. Usvalampi A, Ruvalcaba Medrano M, Maaheimo H, Salminen H, Tossavainen O, Frey AD: **Production and characterization of Aspergillus niger GH29 family α-fucosidase and production of a novel non-reducing 1-fucosyllactose**. *Glycoconj J* 2020, **37**:221–229.

19. Lundemo P, Karlsson EN, Adlercreutz P: **Eliminating hydrolytic activity without affecting the transglycosylation of a GH1 β-glucosidase**. *Appl Microbiol Biotechnol* 2017, **101**:1121–1131.

20. Strazzulli A, Cobucci-Ponzano B, Carillo S, Bedini E, Corsaro MM, Pocsfalvi G, Withers SG, Rossi M, Moracci M: **Introducing transgalactosylation activity into a family 42 β-galactosidase**. *Glycobiology* 2017, **27**:425–437.

21. Tran LT, Blay V, Luang S, Eurtivong C, Choknud S, González-Diáz H, Ketudat Cairns JR: **Engineering faster transglycosidases and their acceptor specificity**. *Green Chem* 2019, **21**:2823–2836.

22. Light SH, Cahoon LA, Mahasenan K V., Lee M, Boggess B, Halavaty AS, Mobashery S, Freitag NE, Anderson WF: **Transferase Versus Hydrolase: The Role of Conformational Flexibility in Reaction Specificity**. *Structure* 2017, **25**:295–304.

23. Jamek SB, Muschiol J, Holck J, Zeuner B, Busk PK, Mikkelsen JD, Meyer AS: **Loop Protein Engineering for Improved Transglycosylation Activity of a β-N-Acetylhexosaminidase**. *ChemBioChem* 2018, **19**:1858–1865.

24. Cobucci-Ponzano B, Strazzulli A, Rossi M, Moracci M: **Glycosynthases in Biocatalysis**. *Adv Synth Catal* 2011, **353**:2284–2300.

25. Hayes MR, Pietruszka J: **Synthesis of glycosides by glycosynthases**. *Molecules* 2017, **22**.

26. Alsina C, Faijes M, Planas A: **Glycosynthase-type GH18 mutant chitinases at the assisting catalytic residue for polymerization of chitooligosaccharides**. *Carbohydr Res* 2019, **478**:1–9.

27. Tong X, Li T, Li C, Wang LX: **Generation and Comparative Kinetic Analysis of New Glycosynthase Mutants from Streptococcus pyogenes Endoglycosidases for Antibody Glycoengineering**. *Biochemistry* 2018, **57**:5239–5246.

28. Shivatare SS, Huang LY, Zeng YF, Liao JY, You TH, Wang SY, Cheng T, Chiu CW, Chao P, Chen LT, et al.: **Development of glycosynthases with broad glycan specificity for the efficient glyco-remodeling of antibodies**. *Chem Commun* 2018, **54**:6161–6164.

29. Okuyama M, Matsunaga K, Watanabe KI, Yamashita K, Tagami T, Kikuchi A, Ma M, Klahan P, Mori H, Yao M, et al.: **Efficient synthesis of α-galactosyl oligosaccharides using a mutant Bacteroides thetaiotaomicron retaining α-galactosidase (BtGH97b)**. *FEBS J* 2017, **284**:766–783.

30. Tegl G, Hanson J, Chen HM, Kwan DH, Santana AG, Withers SG: **Facile Formation of β-thioGlcNAc Linkages to Thiol-Containing Sugars, Peptides, and Proteins using a Mutant GH20 Hexosaminidase**. *Angew Chemie - Int Ed* 2019, **58**:1632–1637.

31. Cobb RE, Chao R, Zhao H: **Directed evolution: Past, present, and future**. *AIChE J* 2013, **59**:1432–1440.

32. Lin H, Tao H, Cornish VW: **Directed evolution of a glycosynthase via chemical complementation.** *J Am Chem Soc* 2004, **126**:15051–9.

33. Armstrong Z, Liu F, Chen HM, Hallam SJ, Withers SG: **Systematic Screening of Synthetic Gene-Encoded Enzymes for Synthesis of Modified Glycosides**. *ACS Catal* 2019, **9**:3219–3227.

34. Ben-David A, Shoham G, Shoham Y: **A Universal Screening Assay for Glycosynthases: Directed Evolution of Glycosynthase XynB2(E335G) Suggests a General Path to Enhance Activity**. *Chem Biol* 2008, **15**:546–551.

35. Macdonald SS, Armstrong Z, Morgan-Lang C, Osowiecka M, Robinson K, Hallam SJ, Withers SG: **Development and Application of a High-Throughput Functional Metagenomic Screen for Glycoside Phosphorylases**. *Cell Chem Biol* 2019, **26**:1001-1012.e5.

36. Andrés E, Aragunde H, Planas A: **Screening glycosynthase libraries with a fluoride chemosensor assay independently of enzyme specificity: identification of a transitional hydrolase to synthase mutant**. *Biochem J* 2014, **458**:355–363.

37. Hayes MR, Bochinsky KA, Seibt LS, Elling L, Pietruszka J: **Development of a colourimetric assay for glycosynthases**. *J Biotechnol* 2017, **257**:162–170.

38. Agrawal A, Bandi CK, Burgin T, Woo Y, Mayes HB, Chundawat SPS: **Click-chemistry enabled directed evolution of glycosynthases for bespoke glycans synthesis**. *bioRxiv* 2020, doi:10.1101/2020.03.23.001982.

39. Jha RK, Strauss CEM: **Smart Microbial Cells Couple Catalysis and Sensing to Provide High-Throughput Selection of an Organophosphate Hydrolase**. *ACS Synth Biol* 2020, **9**:1234–1239.

40. Henrissat B, Davies G: **Structural and sequence-based classification of glycoside hydrolases**. *Curr Opin Struct Biol* 1997, doi:10.1016/S0959-440X(97)80072-3.

41. Cantarel BI, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics**. *Nucleic Acids Res* 2009, **37**:233–238.

42. Davies G, Henrissat B: **Structures and mechanisms of glycosyl hydrolases**. *Structure* 1995, **3**:853–859.

43. KOSHLAND DE: **STEREOCHEMISTRY AND THE MECHANISM OF ENZYMATIC REACTIONS**. *Biol Rev* 1953, doi:10.1111/j.1469-185X.1953.tb01386.x.

44. Cobucci-Ponzano B, Moracci M: **Glycosynthases as tools for the production of glycan analogs of natural products**. *Nat Prod Rep* 2012, **29**:697.

45. Iglesias-Fernández J, Hancock SM, Lee SS, Khan M, Kirkpatrick J, Oldham NJ, McAuley K, Fordham-Skelton A, Rovira C, Davis BG: **A front-face "SNi synthase" engineered from a retaining "double-SN2" hydrolase**. *Nat Chem Biol* 2017, **13**:874–881.

46. Lira-Navarrete E, de las Rivas M, Compañón I, Pallarés MC, Kong Y, Iglesias-Fernández J, Bernardes GJL, Peregrina JM, Rovira C, Bernadó P, et al.: **Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein O-glycosylation**. *Nat Commun* 2015, **6**:6937.

47. Van Tilbeurgh H, Tomme P, Claeyssens M, Bhikhabhai R, Pettersson G: **Limited proteolysis of the cellobiohydrolase I from Trichoderma reesei: Separation of functional domains**. *FEBS Lett* 1986, **204**:223–227.

48. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ: **Carbohydrate-binding modules: fine-tuning polysaccharide recognition**. *Biochem J* 2004, **382**:769–781.

49. Din N, Gilkes NR, Tekant B, Miller RC, Warren RAJ, Kilburn DG: **Non-Hydrolytic Disruption of Cellulose Fibres by the Binding Domain of a Bacterial Cellulase**. *Nat Biotech* 1991, **9**:1096–1099.

50. Burstein T, Shulman M, Jindou S, Petkun S, Frolow F, Shoham Y, Bayer EA, Lamed R: **Physical association of the catalytic and helper modules of a family-9 glycoside hydrolase is essential for activity**. *FEBS Lett* 2009, **583**:879–884.

51. Gal L, Gaudin C, Belaich A, Pages S, Tardif C, Belaich JP: **CelG from Clostridium cellulolyticum: a multidomain endoglucanase acting efficiently on crystalline cellulose.** *J Bacteriol* 1997, **179**:6595–601.

52. Gaudin C, Belaich A, Champ S, Belaich J-P: **CelE, a Multidomain Cellulase from Clostridium cellulolyticum: a Key Enzyme in the Cellulosome?** *J Bacteriol* 2000, **182**:1910–1915.

53. TOMME P, TILBEURGH H, PETTERSSON G, DAMME J, VANDEKERCKHOVE J, KNOWLES J, TEERI T, CLAEYSSENS M: **Studies of the cellulolytic system of Trichoderma reesei QM 9414. Analysis of domain function in two cellobiohydrolases by limited proteolysis**. *Eur J Biochem* 1988, **170**:575–581.

54. Walker JA, Takasuka TE, Deng K, Bianchetti CM, Udell HS, Prom BM, Kim H, Adams PD, Northen TR, Fox BG: **Multifunctional cellulase catalysis targeted by**

**fusion to different carbohydrate-binding modules.** *Biotechnol Biofuels* 2015, **8**:220.

55. Crouch LI, Labourel A, Walton PH, Davies GJ, Gilbert HJ: **The contribution of non-catalytic carbohydrate binding modules to the activity of lytic polysaccharide monooxygenases**. *J Biol Chem* 2016, **291**:7439–7449.

56. Margolles-Clark E, Tenkanen M, Söderlund H, Penttilä M: **Acetyl xylan esterase from Trichoderma reesei contains an active-site serine residue and a cellulose-binding domain**. *Eur J Biochem* 1996, **237**:553–560.

57. Stockinger LW, Eide KB, Dybvik AI, Sletta H, Vårum KM, Eijsink VGHH, Tøndervik A, Sørlie M: **The effect of the carbohydrate binding module on substrate degradation by the human chitotriosidase**. 2015, **1854**:1494–1501.

58. Mizutani K, Fernandes VO, Karita S, Luís AS, Sakka M, Kimura T, Jackson A, Zhang X, Fontes CMGA, Gilbert HJ, et al.: **Influence of a mannan binding family 32 carbohydrate binding module on the activity of the appended mannanase.** *Appl Environ Microbiol* 2012, **78**:4781–7.

59. Codera V, Gilbert HJ, Faijes M, Planas A: **Carbohydrate-binding module assisting glycosynthase-catalysed polymerizations**. *Biochem J* 2015, **470**:15–22.

60. Franke D, Petoukhov M V., Konarev P V., Panjkovich A, Tuukkanen A, Mertens HDT, Kikhney AG, Hajizadeh NR, Franklin JM, Jeffries CM, et al.: **ATSAS 2.8: A comprehensive data analysis suite for small-angle scattering from macromolecular solutions**. *J Appl Crystallogr* 2017, **50**:1212–1225.

61. Zechel DL, Reid SP, Stoll D, Nashiru O, Warren RAJ, Withers SG: **Mechanism,**

**Mutagenesis, and Chemical Rescue of a β-Mannosidase from Cellulomonas fimi**. *Biochemistry* 2003, **42**:7195–7204.

62. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ: **Carbohydrate-binding modules: Fine-tuning polysaccharide recognition**. *Biochem J* 2004, **382**:769–781.

63. Boraston AB: **The interaction of carbohydrate-binding modules with insoluble non-crystalline cellulose is enthalpically driven**. *Biochem J* 2005, **385**:479–484.

64. Georgelis N, Yennawar NH, Cosgrove DJ: **Structural basis for entropy-driven cellulose binding by a type-A cellulose-binding module (CBM) and bacterial expansin**. *Proc Natl Acad Sci* 2012, **109**:14830–14835.

65. Linder M, Teeri TT: **The cellulose-binding domain of the major cellobiohydrolase of Trichoderma reesei exhibits true reversibility and a high exchange rate on crystalline cellulose**. *Proc Natl Acad Sci* 1996, **93**:12251–12255.

66. Boraston AB, Chiu P, Warren RAJ, Kilburn DG: **Specificity and Affinity of Substrate Binding by a Family 17 Carbohydrate-Binding Module from Clostridium cellulovorans Cellulase 5Aâ€** . *Biochemistry* 2000, **39**:11129–11136.

67. Ruiz DM, Turowski VR, Murakami MT: **Effects of the linker region on the structure and function of modular GH5 cellulases**. *Sci Rep* 2016, **6**:1–13.

68. Arai R, Ueda H, Kitayama A, Kamiya N, Nagamune T: **Design of the linkers which effectively separate domains of a bifunctional fusion protein**. *Protein Eng Des Sel* 2001, **14**:529–532.

69. Watts AG, Damager I, Amaya ML, Buschiazzo A, Alzari P, Frasch AC, Withers SG:

**Trypanosoma cruzi trans-sialidase operates through a covalent sialyl-enzyme intermediate: Tyrosine is the catalytic nucleophile**. *J Am Chem Soc* 2003, **125**:7532–7533.

70. Aspeborg H, Coutinho PM, Wang Y, Iii HB, Henrissat B: **Evolution , substrate specificity and subfamily classification of glycoside hydrolase family 5 ( GH5 )**. 2012,

71. Mackenzie LF, Wang Q, Warren RAJ, Withers SG: **Glycosynthases: mutant glycosidases for oligosaccharide synthesis**. *J Am Chem Soc* 1998, **120**:5583–5584.

72. Wilkinson SM, Liew CW, Mackay JP, Salleh HM, Withers SG, McLeod MD: **Escherichia coli glucuronylsynthase: An engineered enzyme for the synthesis of β-glucuronides**. *Org Lett* 2008, **10**:1585–1588.

73. Mayer C, Jakeman DL, Mah M, Karjala G, Gal L, Warren RA., Withers SG: **Directed evolution of new glycosynthases from Agrobacterium β-glucosidase: a general screen to detect enzymes for oligosaccharide synthesis**. *Chem Biol* 2001, **8**:437–443.

74. Kim Y-W, Lee SS, Warren RAJ, Withers SG: **Directed evolution of a glycosynthase from Agrobacterium sp. increases its catalytic activity dramatically and expands its substrate repertoire.** *J Biol Chem* 2004, **279**:42787–93.

75. Hayes MR, Bochinsky KA, Seibt LS, Elling L, Pietruszka J: **Development of a colourimetric assay for glycosynthases**. *J Biotechnol* 2017, **257**:162–170.

76. Datsenko KA, Wanner BL: **One-step inactivation of chromosomal genes in**

**Escherichia coli K-12 using PCR products**. *Proc Natl Acad Sci* 2000, **97**:6640–6645.

77. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H: **Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection**. *Mol Syst Biol* 2006, **2**.

78. Stevenson J, Krycer JR., Phan L, Brown AJ: **A Practical Comparison of Ligation-Independent Cloning Techniques**. *PLoS One* 2013, **8**:e83888.

79. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J: **Operons in Escherichia coli: Genomic analyses and predictions**. *Proc Natl Acad Sci U S A* 2000, **97**:6652–6657.

80. Sung Y-C, Fuchs JA: **Characterization of the cyn Operon in Escherichia coli K12* Young-chul Sung and**. *J Biol Chem* 1988, **263**:14769–14775.

81. Palatinszky M, Herbold C, Jehmlich N, Pogoda M, Han P, von Bergen M, Lagkouvardos I, Karst SM, Galushko A, Koch H, et al.: **Cyanate as an energy source for nitrifiers**. *Nature* 2015, **524**:105–108.

82. Sung Y-C, Fuchs JA: **The Escherichia coli K-12 cyn Operon Is Positively Regulated by a Member of the lysR Family**. 1992, **174**.

83. Lamblin A-FJ, Fuchs JA: **Functional Analysis of the Escherichia coli K-12 cyn Operon Transcriptional Regulation**. *J BACrERIOLOGY* 1994, **176**:6613–6622.

84. Trott,O., Olson AJ: **Autodock vina: improving the speed and accuracy of docking**. *J Comput Chem* 2019, **31**:455–461.

85. Stringer AM, Singh N, Yermakova A, Petrone BL, Amarasinghe JJ, Reyes-Diaz L,

Mantis NJ, Wade JT: **FRUIT, a Scar-Free System for Targeted Chromosomal Mutagenesis, Epitope Tagging, and Promoter Replacement in Escherichia coli and Salmonella enterica**. *PLoS One* 2012, **7**.

86. Studier FW, Moffatt BA: **Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes**. *J Mol Biol* 1986, **189**:113–130.

87. Eames M, Kortemme T: **Cost-Benefit Tradeoffs in Engineered lac Operons**. *Science (80- )* 2012, **336**:911–915.

88. Cobucci-Ponzano B, Conte F, Bedini E, Corsaro MM, Parrilli M, Sulzenbacher G, Lipski A, Dal Piaz F, Lepore L, Rossi M, et al.: **beta-Glycosyl azides as substrates for alpha-glycosynthases: preparation of efficient alpha-L-fucosynthases.** *Chem Biol* 2009, **16**:1097–108.

89. Bojarová P, Petrásková L, Ferrandi EE, Monti D, Pelantová H, Kuzma M, Simerská P, Křen V: **Glycosyl azides - An alternative way to disaccharides**. *Adv Synth Catal* 2007, **349**:1514–1520.

90. Müllegger J, Jahn M, Chen HM, Warren RAJ, Withers SG: **Engineering of a thioglycoligase: Randomized mutagenesis of the acid-base residue leads to the identification of improved catalysts**. *Protein Eng Des Sel* 2005, **18**:33–40.

91. Cobucci-Ponzano B, Zorzetti C, Strazzulli A, Carillo S, Bedini E, Corsaro MM, Comfort DA, Kelly RM, Rossi M, Moracci M: **A novel -d-galactosynthase from Thermotoga maritima converts -d-galactopyranosyl azide to -galacto-oligosaccharides**. *Glycobiology* 2011, **21**:448–456.

92. Fialová P, Carmona AT, Robina I, Ettrich R, Sedmera P, Přikrylová V, Petrásková-Hušáková L, Křen V: **Glycosyl azide—a novel substrate for enzymatic**

**transglycosylations**. *Tetrahedron Lett* 2005, **46**:8715–8718.

93. Rempel BP, Withers SG: **Covalent inhibitors of glycosidases and their applications in biochemistry and biology**. *Glycobiology* 2008, **18**:570–586.

94. Roston RL, Wang K, Kuhn LA, Benning C: **Structural determinants allowing transferase activity in SENSITIVE TO FREEZING 2, classified as a family I glycosyl hydrolase.** *J Biol Chem* 2014, **289**:26089–106.

95. Hidaka M, Honda Y, Kitaoka M, Nirasawa S, Hayashi K, Wakagi T, Shoun H, Fushinobu S: **Chitobiose phosphorylase from Vibrio proteolyticus, a member of glycosyl transferase family 36, has a clan GH-L-like (α/α)6 barrel fold**. *Structure* 2004, **12**:937–947.

96. Fritz TA, Hurley JH, Trinh L-B, Shiloach J, Tabak LA: **The beginnings of mucin biosynthesis: the crystal structure of UDP-GalNAc:polypeptide alpha-N-acetylgalactosaminyltransferase-T1.** *Proc Natl Acad Sci U S A* 2004, **101**:15307–12.

97. Currie MA, Cameron K, Dias FMV, Spencer HL, Bayer EA, Fontes CMGA, Smith SP, Jia Z: **Small angle x-ray scattering analysis of clostridium thermocellum cellulosome N-terminal complexes reveals a highly dynamic structure**. *J Biol Chem* 2013, **288**:7978–7985.

98. Receveur V, Czjzek M, Schulein M, Panine P, Henrissat B: **Dimension, shape, and conformational flexibility of a two domain fungal cellulase in solution probed by small angle x-ray scattering**. *J Biol Chem* 2002, **277**:40887–40892.

99. Albesa-Jové D, Guerin ME: **The conformational plasticity of glycosyltransferases**. *Curr Opin Struct Biol* 2016, doi:10.1016/j.sbi.2016.07.007.

100. Brumm PJ: **Bacterial genomes: what they teach us about cellulose degradation**. *Biofuels* 2013, **4**:669–681.

101. Lu Y, Zhang Y-HP, Lynd LR: **Enzyme-microbe synergy during cellulose hydrolysis by Clostridium thermocellum**. *Proc Natl Acad Sci* 2006, **103**:16165–16169.

102. Nordvang RT, Nyffenegger C, Holck J, Jers C, Zeuner B, Sundekilde UK, Meyer AS, Mikkelsen JD: **It All Starts with a Sandwich: Identification of Sialidases with Trans-Glycosylation Activity**. *PLoS One* 2016, **11**:e0158434.

103. Lundemo P, Adlercreutz P, Karlsson EN: **Improved transferase/hydrolase ratio through rational design of a family 1 β-glucosidase from Thermotoga neapolitana**. *Appl Environ Microbiol* 2013, **79**:3400–3405.

104. Kone FMT, Le Bechec M, Sine J-P, Dion M, Tellier C: **Digital screening methodology for the directed evolution of transglycosidases**. *Protein Eng Des Sel* 2008, **22**:37–44.

105. Gantt RW, Peltier-Pain P, Singh S, Zhou M, Thorson JS: **Broadening the scope of glycosyltransferase-catalyzed sugar nucleotide synthesis**. *Proc Natl Acad Sci* 2013, **110**:7648–7653.

106. Li MZ, Elledge SJ: **Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC**. *Nat Methods* 2007, **4**:251–256.