Site-selective Strategies for the Biological Investigation of Glycoproteins



A thesis submitted to the Board of the Faulty of Physical Sciences in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Maria Alexandra Gafitescu

University of Oxford, Department of Chemistry

St. Cross College

December 2021

Author's Declaration

The work described in this thesis was conducted between October 2016 and December 2020 at the Department of Chemistry, University of Oxford under the supervision of Prof. Ben G. Davis and Prof. Wyatt Yue. All work presented in this thesis is my own, except where otherwise stated, and has not been submitted for any other degree at this or any other university.

The following experiments were performed in collaboration:

The iSCAMS and native MS analyses described in Chapter 2 were performed by Dr. Weston Struwe and Dr. Di Wu, respectively, Department of Chemistry, Chemistry Research Laboratory, University of Oxford UK.

The iSCAMS analyses described in Chapter 4 were performed by Dr. Adam Fineberg, Department of Chemistry, Chemistry Research Laboratory, University of Oxford UK.

LC-MS/MS samples described in Chapter 2 and Chapter 3 were processed by Dr. Simon Nadal at the Advanced Proteomics Facility, Department of Biochemistry, University of Oxford, UK and by Mrs. Elisabete Pires at the Mass Spectrometry Research Facility, Department of Chemistry, University of Oxford, UK.

LC-MS/MS samples described in Chapter 4 were processed by Dr. Marjorie Fournier and analysed by Prof. Shabaz Mohammed at the Advanced Proteomics Facility, Department of Biochemistry, University of Oxford, UK

Expression analysis of HUVEC cells by RT-qPCR was performed by Dr. Naveed Akbar, Radcliffe Department of Medicine, University of Oxford, UK.

Molecular dynamics (MD) simulation was done by Dr. Simon Nadal, Department of Chemistry, Chemistry Research Laboratory, University of Oxford UK.

Histone H3, Histone H2B, Histone TEV-H3, were kindly donated by Dr. Jitka Dadova, Dr. Simon Nadal, Brian Josephson and Yue Yu, Department of Chemistry, Chemistry Research Laboratory, University of Oxford UK.

ΔpmHS2 was kindly donated by Dr. Liang Wu, The Rosalind Franklin Institute, UK whereas GYS and GBE, by Dr. Thomas McCorvie, Radcliffe Department of Medicine, University of Oxford, UK.

1-thio-saccharides different from the GlcNAc-SH and the GlcNAc-ethyl-SH were kindly donated by Dr. Mylene Richards, Department of Chemistry, Chemistry Research Laboratory, University of Oxford UK and by Dr. Elizabeth Grayson, Department of Chemistry, Durham University UK.

2,5-dibromohexadiamide (DBHDA) and N1,N6-bis(2-amino-2-oxoethyl)-2,5-dibromohexanediamide (DBDGla) were kindly donated by Dr. Sebastien Galan and Dr. Jitka Dadova, Department of Chemistry, Chemistry Research Laboratory, University of Oxford UK

Maria Alexandra Gafitescu

Abstract

Protein glycosylation is a post-translational modification that modulates numerous intracellular pathways and extracellular events. However, the study of glycosylated proteins is challenging since these are naturally produced as heterogeneous mixtures of different glycoforms. Moreover, isolation and characterisation of single glycoforms is not always feasible, and the chemical synthesis of such complex biomolecules is often impossible. Although several glycan-protein ligation methods have emerged in recent years, there is still significant progress to be made for the generation of true, homogeneous glycoform-mimics. This doctoral thesis aims at expanding the current toolbox used for the site-selective glycosylation of proteins with strategies that generate 'near-native' glycosidic linkages.

The 'tag-and-modify' approach has been widely recognized as a powerful method for the site-selective chemical modification of proteins. Herein, the application of this methodology for the synthesis of proteoglycan (PG)-mimics is described. A natural PG was homogeneously 'tagged' *via* enzymatic trimming of its glycosaminoglycan (GAG) chain, and a biocompatible radical ligation was developed to install a priming sugar. The minimally 'modified' PG was subjected to enzymatic GAG-polymerization and this allowed the unprecedented access to remodeled PG-analogues. Optimised conditions for the on-protein GAG-extension revealed interesting aspects pertaining to the activity of the employed GAG-synthase. Future work should reveal the utility of such methodology towards understanding the extent to which variously engineered GAGs affect the biological activity of their corresponding PG glycoforms.

In our work we also explored a cysteine (Cys)-based 'tag-and-modify' strategy for the generation of a homogeneous *N*-glycoprotein-mimic. We found that the presumed disulfide reshuffling of the native Cys-residues had impeded the site-specific chemical glycosylation at the inserted Cys-tag.

Lastly, our focus turned towards further investigating the protein that uses the high energy glucose (Glc)-donor UDP-Glc to initiate, *via* autoglucosylation, the biosynthesis of glycogen. Close inspection of both its self-extended glycoforms and the more complex glucosylation states generated by the

downstream enzymatic partners raised exciting new questions about the basic mechanisms of glycogenesis and glycogenolysis.

Acknowledgements

Firstly, I would like to thank my supervisor Prof. Ben Davis who believed in me when he gave me the exciting opportunity to work in his group and explore the intertwining worlds of chemistry and biology. I am deeply grateful not only for the continuous guidance and valuable advice but also for the patience and support during more challenging times of personal struggles. Similarly, I would like to express my gratitude to Prof. Wyatt Yue for his kind, unconditional help, and priceless insights – I feel honoured to have been able to learn from you both and to have grown under your watch.

I would also like to thank my collaborators (already mentioned above) for their important contributions, they all helped overcome challenges that seemed insurmountable beforehand. This made me realize how important it is to be part of a highly effective, supportive team and how by working together, a final common goal can be reached in a timely manner.

Thank you to the ISOTOPICS network, a Marie Sklodowska Curie action funded by the European Union, which provided the main financial support for my research as well as to the Rosalind Franklin Institute, the St. Cross College and the Department of Chemistry for providing additional funding during my final year.

Thank you to the Department of Chemistry and in particular to Elisabete Pires for putting up with my incessant questions and demands in the mass spectrometry area, and for being a patient listener and a very good friend.

Special thanks go to Dr. Khan Maola, Dr. Jitka Dadova and Dr. Mylene Richards, people who helped guide my first scientific endeavours of my PhD, and to Dr. Matt Bilyard, Dr. Sarah Faulkner, Sadra Hamedzadeh and Dr. Mateusz Imiolek, members of the group that brought fun in and outside the lab. I am also grateful to the people who took the time to proofread sections of this thesis: Brian, Dr. Tobi Arisa and Georgios Papadakis. I owe you all multiple drinks and Brian, I promise those won't be like the ardent spirits I offered you last time :D.

I extend my thanks once again to George who willingly agreed to continue one of the challenging but fascinating projects I have led during my PhD studies.

Finally, I am indebted to Tobi and Dr. Andrew Giltrap who helped me put into practice one of my last scientific ideas – without your help I would not have reached one of the most important milestones I had aimed for. In addition, thank you Tobi for your continuous encouragement and the beautiful surprises that helped uplifting me during more challenging times.

Thanks to all my friends from around the world – Nicoleta, Luna, Lucia, Linda and David – and to my family from Romania - Mom, Rodica, Mihai, Veronica and Brandusa - you were all instrumental in helping me complete my PhD and I will always remember it.

"It is not the critic who counts; not the man who points out how the strong man stumbles, or where the doer of deeds could have done them better. The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood; who strives valiantly; who errs, who comes short again and again, because there is no effort without error and shortcoming; but who does actually strive to do the deeds; who knows great enthusiasms, the great devotions; who spends himself in a worthy cause; who at the best knows in the end the triumph of high achievement, and who at the worst, if he fails, at least fails while daring greatly, so that his place shall never be with those cold and timid souls who neither know victory nor defeat."

Theodore Roosevelt, 1910

Table of Contents

Author Declaration i

Abstract iii

Acknowledgements v

Table of Contents viii

List of Abbreviations xii

Publications and Presentations xxii

1. Introduction 1

1.1 Protein glycosylation: a "sweet" refinement of the biological information 3

1.2 Access to homogeneously glycosylated proteins for robust glyco-biology studies is of paramount importance 8

1.2.1 Chemical synthesis of glycoproteins 9

1.2.2 Engineering glycosylation pathways 14

1.2.3 Direct enzymatic protein glycosylation and glycoprotein remodeling 16

1.2.4 Site-selective chemical glycosylation of proteins 19

1.3 Aims of this thesis 31

1.4 References 33

2. GAG ligation: thiyl-glycosylation allows direct construction of synthetic proteoglycans 45

2.1 Introduction 49

2.1.1 Natural heterogeneity impedes the structure-activity relationship studies of proteoglycans 49

2.1.2 Thiyl-glycosylation as a strategy for the direct construction of semi-synthetic, minimal PGs 51

2.1.3 Bacterial glycosyltransferases as protagonists of the PG-remodeling process with non-native GAG chains 53

2.2 Results and discussion 55

2.2.1 Extraction, purification and characterization of a natural PG: decorin 55

2.2.2 Development of a radical GAG ligation to obtain semi-synthetic PG-intermediates 62

2.2.3 Chemo-enzymatic synthesis of neoproteoglycans containing a heparosan GAG chain.....71

2.2.4 Optimization of the pmHS2-catalysed enzymatic extension and development of the Hep-remodeled decorin.....81

2.2.5 Decorin induces DNA transcription of the intercellular adhesion molecule 1 (ICAM-1) in human umbilical vein endothelial cells (HUVECs).....88

2.3 Conclusions and future perspectives.....91

2.4 Experimental section.....93

2.4.1 Plasmid DNA details and protein sequences.....93

2.4.2 Protein extraction/expression and characterisation.....95

2.4.3 Development of a free-radical hydrothiolation reaction to ligate unprotected glycans to the core decorin.....110

2.4.4 Chemo-enzymatic synthesis of neoproteoglycans containing a heparosan GAG chain.....127

2.4.5 PmHS2, ΔpmHS2 and GalT extend the primed GlcNAc-ethyl-S-decorin but not the GlcNAc-S-variant.....141

2.4.6 Quantification of the relative protein levels of Peg3, Bec1, LC3 and ICAM-1, VCAM-1, IL-1α 149

2.4.7 Chemical synthesis of GlcNAc-SH and GlcNAc-ethyl-SH.....150

2.5 References.....157

3. Towards a single glycoform of erythropoietin through site-selective chemical glycosylation 164

3.1 Introduction.....168

3.1.1 Natural heterogeneity impedes the structure-activity relationship studies of *N*-glycoproteins 168

3.1.2 A combined chemical mutagenesis/thia-Michael addition approach to access defined EPO glycofom-mimics.....172

3.2 Results and discussion.....177

3.2.1 Introduction of dehydroalanine to the MBP-fused EPO (MBP-EPO-1).....177

3.2.2 Introduction of dehydroalanine to the MBP-fused EPO which was engineered for reduced aggregation and disulfide shuffling (MBP-EPO-2).....189

3.2.3 Introduction of dehydroalanine to the MBP-fused EPO which was engineered for reduced aggregation (MBP-EPO-3).....198

3.3 Conclusions and future perspectives.....202

3.4 Experimental Section 205

3.4.1 Plasmid DNA details and protein sequences 205

3.4.2 Protein expression and characterisation 206

3.4.3 Formation of Dha-MBP-EPO-1 and GlcNAc-S-MBP-EPO-1 213

3.4.4 Formation of Dha-MBP-EPO-2 215

3.4.5 Formation of Dha-MBP-EPO-3 and GlcNAc-S-MBP-EPO-3 224

3.4.6 LC-MS/MS analysis of enzymatically digested GlcNAc-S-MBP-EPO-1, βME-S-MBP-EPO-2 and βME-S-MBP-EPO-3 225

3.4.7 Molecular Dynamics (MD) of EPO-1 227

3.5 References 228

4. Insights into the early stages of the glycogen biogenesis 235

4.1 Introduction 238

4.1.1 Glycogen: a readily accessible reserve for energy-depleting conditions 238

4.2 Results and discussion 242

4.2.1 Enzyme-compatible Suzuki-Miyaura cross-coupling (SMC) enabled mechanistic insights into the biocatalytic activity of GYG 242

4.2.2 ISCAMS analyses revealed dissociation of the autoglucosylated GYG-wt-nGlc 246

4.2.3 Intact protein LC-MS analyses revealed hydrolysis of the autoglucosylated GYG-wt-nGlc and the GYS-extended GYG-wt-(m+n)Glc polyglucosan chains in absence of UDP-Glucose 250

4.3 Conclusions and future perspectives 256

4.4 Experimental section 259

4.4.1 Plasmid DNA details and protein sequences 259

4.4.2 Protein expression and characterization 260

4.4.3 Autoglucosylation and glucosylation assays 265

4.4.4 Hydrolysis assays 273

4.5 Supplementary text and figures 281

4.5.1 Additional figures 281

4.5.2 Note on the use of GYS as a complex with GYG-wt 285

4.6 References 287

Appendix 291

1. General procedures for small molecule chemistry 293

2. General procedures for biological work and protein modification 294

3. Denaturing intact protein LC-MS 295

4. NMR spectra 298

5. Results of protein and peptide identifications obtained from the shotgun proteomics analysis of GYG-wt-0Glc 305

List of Abbreviations

aaRS	aminoacyl-tRNA synthetase
Ac	acetyl
Acm	acetamidomethyl
Aha	azidohomoalanine
Ahc	allylhomocysteine
AKT	protein kinase B
Ala	alanine
AllylCys	*S*-allylcysteine
AMP	adenosine monophosphate
AMPK	AMP-activated protein kinase
Asn	asparagine
Asp	aspartate
ATP	adenosine triphosphate
aq.	aqueous
Bac	bacillosamine
Beclin 1	Bec1
BSA	bovine serum albumin
BS3	bis(sulfosuccinimidyl)suberate
βME	β-mercaptoethanol
C	Celsius
CBM48	carbohydrate-binding module 48
CD	circular dichroism
CHO	Chinese hamster ovary
CID MS/MS	collision induced dissociation mass spectrometry fragmentation
ConA	concavalin A

Conc.	concentration
CS	chondroitin sulfate
ctTS	*Comamonas testosteroni* testosteronan synthase
CuAAC	copper(I) catalyzed 1,3-dipolar cycloaddition
Cys	cysteine
CV	column volumes
Da	Dalton(s)
DBDGla	*N*1, *N*6-bis(2-amino-2-oxoethyl)-2,5-dibromohexanediamide
DBHDA	2,5-dibromohexadiamide
DCM	Dichloromethane
ddH_2O	double distilled water
Dha	dehydroalanine
DMF	dimethylformamide
DMSO	dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DS	dermatan sulfate
DS-epi1/DS-epi2	dermatan sulfate epimerase 1 or 2
DSF	Differential Scanning Fluorimetry
DTT	dithiothreitol
E. coli	*Escherichia coli*
EB	elution buffer
EDTA	ethylenediaminetetraacetic acid
EGFR	endothelial growth factor receptor
EMS	ethyl methanesulfonate
EndoH	endoglycosidase H
EPO	erythropoietin
EPO-R	erythropoietin receptor
equiv.	equivalents
ER	endoplasmic reticulum

ERK	extracellular-regulated kinase
ESC	embryonic stem cell
ESI	electron spray ionisation
ETD	electron transfer dissociation
ERK1/2	extracellular signal-regulated kinases 1/2
FA	formic acid
FGE	fGly-generating enzyme
fGly	formylglycine
FOXO3a	forkhead box class O 3a
FPLC	fast protein liquid chromatography
Fuc	fucose
g (in context of masses)	Gram(s)
GAG	glycosaminoglycan
Gal	galactose
GalNAc	*N*-acetyl-D-galactosamine
GalT	bovine β(1,4)-galactosyltransferase
GBE	glycogen branching enzyme
GDE	glycogen debranching enzyme
GdnHCl	guanidine hydrochloride
GH	glycosyl hydrolase
G1P	glucose-1-phosphate
G6P	glucose-6-phosphate
Glc	glucose
GlcA	D-glucuronic acid
ΔGlcA	4,5-dehydro-D-glucuronic acid
GlcN	glucosamine
GlcNAc	*N*-acetyl-D-glucosamine
GlcNAc-SH	1-thio-*N*-acetylglucosamine
Gln	glutamine

Glu	glutamate
GluC	*Staphylococcus aureus* protease V8
GPC3	glypican-3
GPH	glycogen phosphorylase
GPI	glycosylphosphatidylinositol
GRB2	growth factor receptor-bound protein 2
GSD	glycogen storage disease
GSK-3β	glycogen synthase kinase-3β
GT	glycosyltransferase
GYG	glycogenin
GYS	glycogen synthase
h	hour(s)
HA	hyaluronic acid
Hag	homoallylglycine
HCD	higher-energy collision dissociation
Hh	Hedgehog
HEK	Human embryonic kidney
Hep	heparosan
Hex	hexose
HexNAc	*N*-acetyl-D-hexosamine
HGFR	hepatocyte growth factor receptor
HIF	hypoxia-inducible transcription factor
HIFα	hypoxia-inducible transcription factor alpha
HIFβ	hypoxia-inducible transcription factor beta
His_6	polyhistidine
Hpg	homopropargylglycine
HREs	hypoxia-response elements
HRMS	high resolution mass spectrometry
HS	heparin/heparan sulfate

Hyp	hydroxyproline
Hz	Hertz
IdoA	iduronic acid
Ile	isoleucine
Il10	interleukin-10
Il12b	interleukin-12b
iPrOH	isopropanol
IPTG	isopropyl-β-D-1-thiogalactopyranoside
IR	infrared
iSCAMS	interferometric scattering mass spectrometry
J	coupling constant
JAK2	Janus Kinase 2
KAHA	α-ketoacid-hydroxylamine
kDa	kilodalton(s)
KfiA	*E*. *coli K5 N*-acetyl-D-glucosaminyltransferase
KfiC	*E*. *coli K5* glucuronyltransferase
KfiB/KfiD	subunit B/D of the *E*. *coli K5* heparosan synthase
KI	Knock in
KO	Knock out
L	litre
LacNAc	*N*-acetyllactosamine
L-ArgHCl	L-arginine hydrochloride
LB	loading buffer
LC3	microtubule-associated proteins 1A/1B light chain 3B
LC-MS	liquid chromatography-mass spectrometry
Leu	leucine
LLO	lipid-linked oligosaccharide
LRMS	low resolution mass spectrometry
LRR	leucine-rich repeat

M	molar
Man	mannose
MBP	maltose binding protein
MD	molecular dynamics
MDBP	methyl 2,5-dibromopentanoate
MeCN	acetonitrile
MEK	mitogen-activated protein kinase kinase
MeOH	methanol
Met	methionine
mg	milligram(s)
MGAT3	mannosyl-glycoprotein β(1,4) GlcNAc-transferase
MHz	Megahertz
µL	microliter(s)
mL	millilitre(s)
mM	millimolar
mmol	millimole(s)
$MnCl_2$	manganese chloride
mol	mole(s)
m.p.	melting point
mRNA	messenger ribonucleic acid
MS	mass spectrometry
mTOR	mechanistic target of rapamycin
µM	micromolar
µmol	micromole(s)
NaCl	sodium chloride
NaHEPES	4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid sodium salt
Neu5Ac	*N*-acetylneuraminic acid
NCL	native chemical ligation

NGT	*N*-glycosyltransferase
nm	nanometre(s)
nM	nanomolar
nmol	nanomole(s)
NMR	nuclear magnetic resonance
OD	optical density
OGA	*O*-GlcNAcase
OGT	*O*-GlcNActransferase
OST	oligosaccharyltransferase
P	phosphorylation
Pac	phenoxyacetyl
*p*AcPhe	*p*-acetylphenylalanine
Pan-RTK	pan-receptor tyrosine kinase
pAzF	*p*-azidophenylalanine
PDB	Protein Data Bank
PDCD4	programmed cell death 4
PEG	polyethylene glycol
Peg3	paternally-expressed gene 3
PG	proteoglycan
PGM	phosphoglucomutase
PHD	prolyl hydroxylase domain
PI3K	phosphoinositide 3-kinase
*p*IBenzylCys	*p*-iodobenzylcysteine
Plk	N^6 –(propargyloxycarbonyl)-L-lysine
PLP	pyridoxal phosphate
*p*IPhe	*p*-iodophenylalanine
pmCS	*Pasteurella multocida* chondroitin synthase
pmHAS	*Pasteurella multocida* hyaluronic acid synthase
pmHS1/pmHS2	*Pasteurella multocida* heparosan synthase 1 or 2

ΔpmHS2	*N*-terminus truncated *Pasteurella multocida* heparosan synthase 2
PNGaseF	peptide-*N*-glycosidase F
polyLacNAc	poly-*N*-acetyl-lactosamine
ppm	parts per million
Pro	proline
PTM	post-translational modification
Pyl	pyrrolysyl
RAF	rapidly accelerated fibrosarcoma
RAS	rat sarcoma protein
RBCs	red blood cells
rEPO	recombinant erythropoietin
RMSF	root mean square fluctuation
RNA	ribonucleic acid
rpm	revolutions per minute
RPTPσ	receptor protein tyrosine phosphatase sigma
rt	room temperature
RT-qPCR	quantitative reverse transcription polymerase chain reaction
s	second
SAR	structure-activity relationship
sat.	saturated
SAU-X	sodium acetate urea, followed by the concentration of sodium chloride (X)
SDS-PAGE	sodium dodecyl sulfate – polyacrylamide gel electrophoresis
SEC	size exclusion chromatography
Sia	sialic acid
siRNAs	small interfering RNAs
Ser	serine
SLRP	small leucine-rich proteoglycan

SMC	Suzuki-Miyaura cross-coupling
SOC	super optimal broth with catabolite repression
SOS	son of Sevenless protein
SPPS	solid-phase peptide synthesis
STAT5	signal transducer and activator of transcription 5
sulfo-SMPB	sulfosuccinimidyl 4-(*N*-maleimidophenyl)butyrate
TCEP	tris(2-carboxyethyl)phosphine
TEAB	triethylammonium bicarbonate
TEV	*Tobacco etch* virus
TFA	trifluoroacetic acid
TFE	2,2,2-trifluoroethanol
TGF-β	transforming growth factor β
THF	tetrahydrofuran
Thr	threonine
Thz	thiazolidine
TIC	total ion count
TLC	thin layer chromatography
TLR 2/4	Toll-like receptor 2/4
Tm	melting temperature
TNFα	tumour necrosis factor alpha
TrisHCl	tris(hydroxymethyl)aminomethane hydrochloride
tRNA	transfer RNA
TTL	*Thermoanaerobacter thermohydrosulfuricus* lipase
Tyr	tyrosine
UAA	unnatural amino acid
UDP	uridine diphosphate
UDP-Gal	uridine diphosphate - galactose
UDP-Glc	uridine diphosphate - glucose
UDP-GlcA	uridine diphosphate - glucuronic acid

UDP-GlcNAc	uridine diphosphate - *N*-acetylglucosamine
UV	ultraviolet
Val	valine
Vazo-44	2,2'-azobis(2-(2-imidazolin-2-yl)propane)dihydrochloride
VEGFR2	vascular endothelial growth factor receptor 2
VHL	Von Hippel-Lindau protein
Xyl	xylose
Wnt	Wingless/Integrated
WISP-1	Wnt-1-induced secreted protein 1

Publications and Presentations

Publications

"GAG ligation: thiyl-glycosylation allows direct construction of synthetic proteoglycans", Maria A. Gafitescu, Georgios Papadakis, Mylene Richard, Charles Fehl, Elena Tobolkina, Arghya Modak, Elizabeth J. Grayson, Wei-Min Liu, Yassir Ahmed, Elisabete Pires, Simon Nadal, Naveed Akbar, Jeremy E. Turnbull, and Benjamin G. Davis; manuscript in preparation

"Palladium-mediated enzyme activation suggests multiphase initiation of glycogenesis", Matthew K. Bilyard, Henry Bailey, Lluis Raich, Maria A. Gafitescu, Takuya Machida, Javier Iglesias-Fernandez, Seung Seo Lee, Christopher D. Spicer, Carme Rovira, Wyatt. W. Yue and Benjamin. G. Davis (Nature, 2018, vol. 563, 235–240)

Presentations

ACS Fall 2019 National Meeting & Exposition - American Chemical Society, San Diego USA, 2019. Oral presentation: "Elucidating glycogen biosynthesis through palladium-mediated enzyme control. Maria A. Gafitescu, Matthew K. Bilyard, Henry Bailey, Lluis Raich, Takuya Machida, Javier Iglesias-Fernandez, Seung Seo Lee, Christopher D. Spicer, Carme Rovira, Wyatt. W. Yue and Benjamin. G. Davis."

Chapter 1

Introduction

Contents

1.1 Protein glycosylation: a “sweet” refinement of the biological information 3

1.2 Access to homogeneously glycosylated proteins for robust glyco-biology studies is of paramount importance 8

1.2.1 Chemical synthesis of glycoproteins 9

1.2.2 Engineering glycosylation pathways 14

1.2.3 Direct enzymatic protein glycosylation and glycoprotein remodeling 16

1.2.4 Site-selective chemical glycosylation of proteins 19

1.2.4.1 Cysteine incorporation: site-directed mutagenesis 20

1.2.4.2 Genetic unnatural amino acid incorporation: sense and nonsense codon reassignment 25

1.2.5 Aims of this thesis 31

1.3 References 33

1.1 Protein glycosylation: a "sweet" refinement of the biological information

Glycoproteins are ubiquitous in the biological world and as such, they are involved in a wide variety of physiological processes: cell-cell/cell-matrix adhesion,[1,2] intracellular signalling events,[3] host-pathogen interactions[4] and immune responses,[5] just to name a few. It emerges thus, that protein glycosylation, one of the most common posttranslational modifications (PTMs) (more than 50 % of all proteins are estimated to be glycosylated)[6] represents the main strategy that nature uses to increase its proteome diversity and thus convey additional layers of molecular information.[7] On proteins, the installed glycan structure can be imagined as a "sweet" modulation of the molecular message that has already undergone successive stages of refinement during transcription and translation, on the DNA-RNA-protein axis of biological information.[7,8]

There are mainly two classes of glycoproteins: the *N*-linked and the *O*-linked glycoproteins (Figure **1.1**).[9]

In *N*-glycoproteins, the glycan attachment takes place at the amide side chain of an asparagine (Asn) residue, part of the consensus sequence Asn-Xxx-Ser/Thr, where "Xxx" is any amino acid except proline (Pro).[10,11] In eukaryotes, all the *N*-glycans begin with a β-linked *N*-acetyl-glucosamine (GlcNAc) residue and they are classified into three general classes: oligomannose, complex and hybrid (Figure **1.1**, a). While they all share a core mannose (Man)-extended GlcNAc-structure, the Man-α(1,3)-[Man-α(1,6)]-Man-β(1,4)-GlcNAc-β(1,4)-GlcNAc-β1 pentasaccharide, they are differentially decorated with either Man residues, GlcNAc-initiated branches, or through combined Man and GlcNAc extensions.[12] The structural heterogeneity is further enhanced, primarily in the complex *N*-glycans and to a less extent in the hybrid ones, by GlcNAc-fucosylations, variable insertions of branching GlcNAc residues, sequential additions of *N*-acetyllactosamine (LacNAc) repeats and sporadic sialic acid (Sia) terminations.[13,14]

In *O*-glycoproteins, glycosylation occurs at the hydroxy group of a serine (Ser) or threonine (Thr)[15] and much more rarely of a tyrosine (Tyr) residue[16,17]. Many of the Ser/Thr-linked glycoproteins carry an initial α-linked *N*-acetyl-galactosamine (GalNAc) residue and in the case of the vast class of *O*-glycoproteins represented by the mucins (Figure **1.1**, b), this is further extended to display one of the four most predominant core structures: Gal-(β1,3)GalNAc-α1 (core 1), GlcNAc-β(1,6)-[Gal-β(1,3)]-GalNAc-α1 (core 2), GlcNAc-β(1,3)-GalNAc-α1 (core 3), GlcNAc-β(1,6)-[GlcNAc-β(1,3)]-GalNAc-α1 (core 4). Each of these core glycans may be further linked to galactose (Gal), branched by the addition of

GlcNAc residues, receive fucose (Fuc) or sulfate groups or be capped with Sia, to give mature, complex *O*-glycosylated proteins that are recognized by lectins or have antigenic properties.[18]

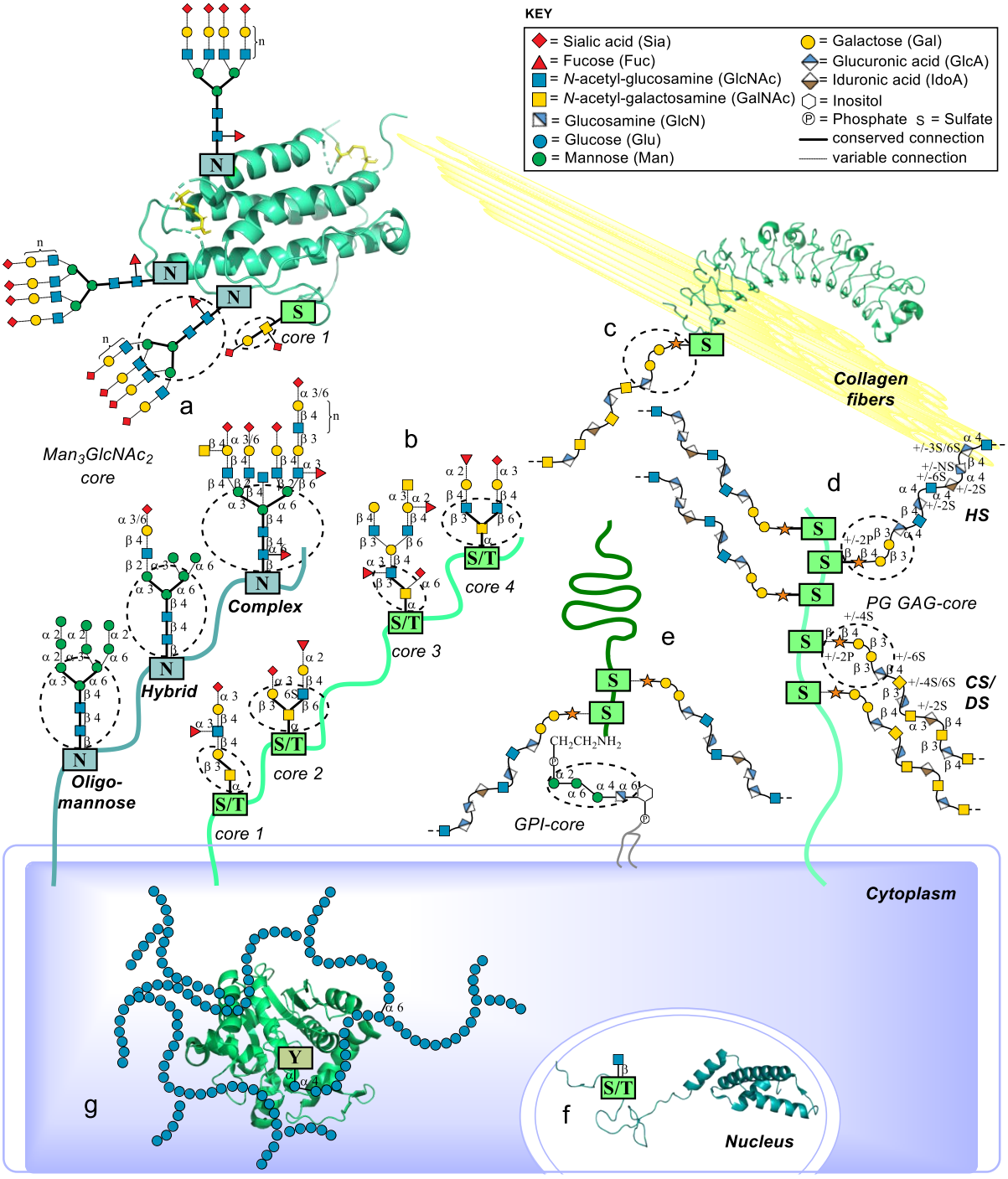


Figure 1.1. Selected classes of eukaryotic glycoproteins. The examples that are depicted herein are relevant in the context of this thesis and they include: soluble (EPO) and cell surface *N*-glycosylated proteins (**a**), soluble (EPO) and cell surface mucin-type *O*-linked glycoproteins (**b**), soluble (decorin), transmembrane (syndecans) and GPI-anchored (glypicans) PGs (**c**-**e**), *O*-glycosylated nuclear proteins (histones, **f**) and the cytoplasmatic, auto-glucosylating GYG protein (**g**). The core structures are highlighted by the dashed ellipses and the glycosylation sites are indicated by the following one-letter abbreviations: N (Asn), S (Ser), T (Thr), Y (Tyr).

Erythropoietin (EPO) is a soluble, glycosylated hormone that contains complex *N*-glycans.[19,20] The latter are characterized by varying antennary structures (bi-, tri- and tetra-) that carry one or multiple LacNAc repeats and display incomplete sialylation. In addition, this glycoprotein has a mucin-type *O*-linked saccharide (core 1) attached to its Ser126 residue which is predominantly disialylated and thereby converted into a mature, branched *O*-GalNAc glycan.[21]

Proteoglycans (PGs) constitute another major class of glycoproteins which are characterized by the presence of one or more linear glycosaminoglycan (GAG) chains (Figure **1.1**, c-e).[22] The chondroitin sulfate (CS), the dermatan sulfate (DS) and the heparin/heparan sulfate (HS) are *O*-linked GAG chains attached to the hydroxy group of a Ser, part of a Ser-Gly sequence, through a conserved tetrasaccharide core: glucuronic acid (GlcA)-β(1,3)-Gal-β(1,3)-Gal-β(1,4)-xylose (Xyl)-β1.[23] Based on the spectrum of biosynthetic glycosyltransferases available in the cell, this common motif is extended either with [GlcA-β(1,3)-GalNAc-β(1,4)] repeating disaccharide units to form a CS chain or with [GlcA-β(1,4)-GlcNAc-α(1,4)], in the case of a HS-containing PG.[23,24] Epimerization of one or more GlcA residues of the CS chain to iduronic acid (IdoA) generates the DS glycosaminoglycan.[22] All in all, these sugar polymers are sulfated to various degrees at their hexosamine and uronic acid repeating units, and the generated sulfation patterns determine the ability of a GAG to selectively bind specific ligands.[25-27]

Most of the PGs have an extracellular localization, with decorin being one of the most studied PGs, partially due to its promising and highly intricate signalling roles (Figure **1.1**, c).[28] Nevertheless, emerging studies are increasingly focused on studying cell-surface associated PGs as well, such as the transmembrane PGs syndecans (Figure **1.1**, d) and the membrane-anchored glypicans (Figure **1.1**, e).[24,29] Due to their combined structural features, glypicans can be categorised both as PGs as well as glycosylphosphatidylinositol (GPI)-anchored glycoproteins.[30] The latter have their phosphoethanolamine-*C*-terminus protein backbone bridged to inositolphospholipids embedded into the cell membrane, through a trimannosyl-nonacetylated glucosamine (GlcN) core: Man-α(1,2)-Man-α(1,6)-Man-α(1,4)-GlcN-α(1,6).[31]

While glycoproteins are predominantly found at the cell surface and in cellular secretions, there are examples of interesting nucleocytoplasmic glycoproteins that are currently drawing increased attention from the scientific community.[32] For instance, β-*O*-GlcNAcylation is one of the PTMs found on histone tails (Figure **1.1**, f).[33] Studies have shown that, depending on the targeted Ser/Thr residue(s), this epigenetic regulator either promotes gene transcription through chromatin relaxation or prevents transcription by way of chromatin compaction.[34,35]

Another unique glycoprotein example is glycogenin (GYG), a cytosolic glycosyltransferase able to autoglucosylate its Tyr195 residue through an uncommon α-*O*-glucose (Glc) linkage (Figure **1.1**, g).[36]

Thus, *via* its stepwise, Glc-α(1,4) polyglucosan extension, GYG effectively converts itself into a glycoprotein carrying a maltosaccharide chain that can reach a length of ~11 average Glc residues.[36,37] Further processing by the glycogen synthase (GYS) and the glycogen branching enzyme (GBE) generates a mature glycogen particle, through additional Glc-α(1,4) polymerization and Glc-α(1,6) branching, respectively.[38]

Over the years, numerous studies have shown that alterations in the glycan composition and architecture of glycoproteins often are correlated with significant cell-specific phenotypic changes. Thus, it has become increasingly clear that glycosylated proteins govern important signalling pathways, and the elicited biological events occur in a glycoform-dependent manner.

For instance, it appears that the mannosyl-glycoprotein β(1,4) GlcNAc-transferase (MGAT3), the enzyme that catalyses the transfer of bisecting GlcNAc-β(1,4) residues to the core of *N*-glycans in glycoprotein receptors, is responsible for inhibiting the development and migration of mammary tumor cells, both *in vitro* and in mice.[39] This, combined with additional glycosylation changes such as the altered GlcNAc-β(1,6) branching, core fucosylation or terminal sialylation of *N*-glycoproteins that were observed to occur in different tumour types, suggests that certain glycoforms are highly relevant in tumorigenesis and cancer progression.[40]

On the other hand, Aricescu *et al.*[41] demonstrated that PGs equipped with different GAG-chains (HS and CS) of similar length have opposing effects on the receptor protein tyrosine phosphatase sigma (RPTPσ), a cell surface receptor that is found in the axonal growth cones of neurons.[42-45] Specifically, while the HS-chain acts as a concatenator of RPTPσ oligomers thus promoting neuronal extension, the CS inhibits RPTPσ oligomerization thereby stalling the neuronal growth.

In addition to the crucial signalling roles, it is well established that certain glycan features ensure that only properly folded glycoproteins are being trafficked from the endoplasmic reticulum (ER) and/or the Golgi apparatus and inserted into the plasma membrane or secreted by the synthesizing cell. For instance, in the ER, the *N*-glycosylation serves as a "quality control" check during the folding process of *N*-glycoproteins.[46] If the protein is erroneously misfolded into a non-native conformation, the *N*-glycans are inefficiently displayed and this constitutes a molecular tag that directs them onto cytosolic degradation pathways, thereby impeding the trafficking of aberrant glycoproteins.

Furthermore, glycosylation has a profound impact on the metabolic fate and stability of the protein to which the glycan structure is attached to. Due to the steric hindrance that the glycan provides to the protein backbone and amino acids situated in close proximity, glycosylation impedes the exposure to the active site of proteolytic enzymes thereby increasing the stability of the protein.[47-49] In several cases, it was found that due to intermolecular steric repulsion, carbohydrates prevent disulfide crosslinking[50] and non-specific aggregation of the proteins to which they are attached,[51,52]

thus inhibiting their toxic accumulation into living cells. Lastly, glycosylation increases the solubility of numerous proteins,[53,54] while also improving their stability against exogenous physical factors such as pH fluctuations,[55] chemical denaturants[56,57] and temperature changes (heating[58] or freeze-thawing cycles[59]).

1.2 Access to homogeneously glycosylated proteins for robust glyco-biology studies is of paramount importance

Glycoproteins are produced as mixtures of glycoforms and are characterized by a variable degree of complexity in the attached glycan structures, at various sites of the peptide backbone.[60] While the biosynthesis of other biomolecules such as RNA and proteins occurs in a template-directed fashion, the biosynthesis of glycoproteins, and specifically the glycan portion, is not under genetic control.[8] The synthesis of each glycoprotein is a complex biological process dependent on multiple cellular parameters such as activation, localization and expression levels of different glycosyltransferases and glycosidases, as well as substrate/donor specificity or local substrate availability.[61] Consequently, the resulting structural heterogeneity of the glycoproteins that are being naturally produced is the cumulative result of all these physiological factors.

Firstly, the high molecular complexity and functional diversity imparted by carbohydrates to proteins are a consequence of the variability of the amino acid sites on the peptide backbone that can undergo glycosylation (macroheterogeneity).[60-62] Secondly, the structural diversity of glycoproteins and consequently their plastic "sweet talk" to other functional partners also derives from a glycan's inherent compositional variation, greatly expanded by additional structural features such as ring size, glycosidic linkage, anomeric configuration, and subunit modifications (e. g. sulfation, phosphorylation and acylation) (microheterogeneity).[60-62]

The natural heterogeneity of glycoproteins makes the direct assignment of observed biological activities to specific glycosylation profiles challenging, if not impossible. Moreover, many of these heterogeneous mixtures are extremely difficult to separate into pure, homogeneous glycoforms even when using extensive chromatographic techniques.[63] Therefore, there is an urgent need to develop novel and efficient methods that allow us to gain access to glycan-defined glycoproteins which can be further used in glycobiology studies.

Many strategies have been developed to obtain homogeneously glycosylated proteins and decipher their structure related biological activity *in vitro*, yet each method has its inherent limitations, and the optimal approach is usually glycoprotein specific. Such notable methods include total chemical synthesis of glycoproteins, recombinant production in organisms that have engineered glycosylation pathways, site-selective glycan-protein ligation strategies, enzymatic glycosylation of proteins and glycoprotein remodeling.[63-65] This introduction highlights emerging technologies that hold great promise in making a variety of glycan-defined glycoproteins, with a particular emphasis on those which

were developed to generate PG analogues, EPO glycoforms and mimics of thereof, and GYG glycoconjugates.

1.2.1 Chemical synthesis of glycoproteins

Since the development of the native chemical ligation (NCL) in 1994,[66] a method that enabled the chemical synthesis of proteins containing more than 50 amino acid residues, several research groups around the world have embarked on adapting and optimizing this strategy for the synthesis of homogenous glycoproteins. Thus, the chemical synthesis of a desired glycoform involves the ligation of several peptidic and glycopeptidic fragments that are most often obtained by the well-established Solid-Phase Peptide Synthesis (SPPS) methodology.[67] The classic NCL strategy involves the trans-thioesterification between a peptide carrying an *N*-terminal cysteine (Cys) and the adjacent peptide bearing a *C*-terminal thioester. The subsequent *S,N*-acyl shift allows for the native peptidic bond to be formed and the Cys-side chain to be released (Figure **1.2**, a.1).

Although NCL has emerged as the most powerful and frequently used ligation method,[68,69] not all glycoproteins contain cysteine residues and many that do, do not contain this amino acid at useful disconnection sites. Therefore, further extensions to NCL[70-72] (Figure **1.2**, a.2) and other ligation strategies such as the traceless Staudinger ligation,[73] serine/threonine ligation,[74] α-ketoacid-hydroxylamine (KAHA) ligation,[75] ligation at Glu/Asp[76] (Figure **1.2**, b) have been developed and efficiently applied when more appropriate.

However, independently from the chosen ligation strategy, additional difficulties may arise during the chemical synthesis of complex glycoproteins, bearing multiple glycosylated sites. These include the use of extensive protection regimes for both the saccharides and the amino acids during the synthesis of the glycosylated peptides, the acid and base lability of the latter and the inefficient ligation between disconnection points situated in close proximity to the bulky *N*-glycans.[77,78]

To allow incorporation of the intended glycan at the desired amino acid position, either a linear or a convergent glycopeptide assembly strategy can be employed (Figure **1.2**, b).[79,80] *N*-linked glycans are often installed using the convergent strategy and as such, after the SPPS synthesis of the peptide, the Lansbury aspartylation of a pre-synthesized anomeric glycosylamine[81] or the Staudinger ligation of a glycosylazide to an activated aspartate (Asp)-containing peptide,[82] allows for the desired *N*-glycopeptide to be obtained (Figure **1.2**, b.1).

In the linear, or so-called “cassette”-based strategy, appropriately protected, pre-glycosylated amino acids are used as building blocks during the SPPS of the glycopeptide (Figure **1.2**, b.2).[78] Due to the immense challenge in the attachment of α-selective *O*-linked glycans, the linear assembly is the

preferred approach by which most *O*-linked glycopeptides are being obtained. Nevertheless, optimization of the amide coupling conditions is often required to minimize the epimerization of the glycosylated serine residues or the β-elimination of the sugar moiety in the case of threonine derivatives.[83]

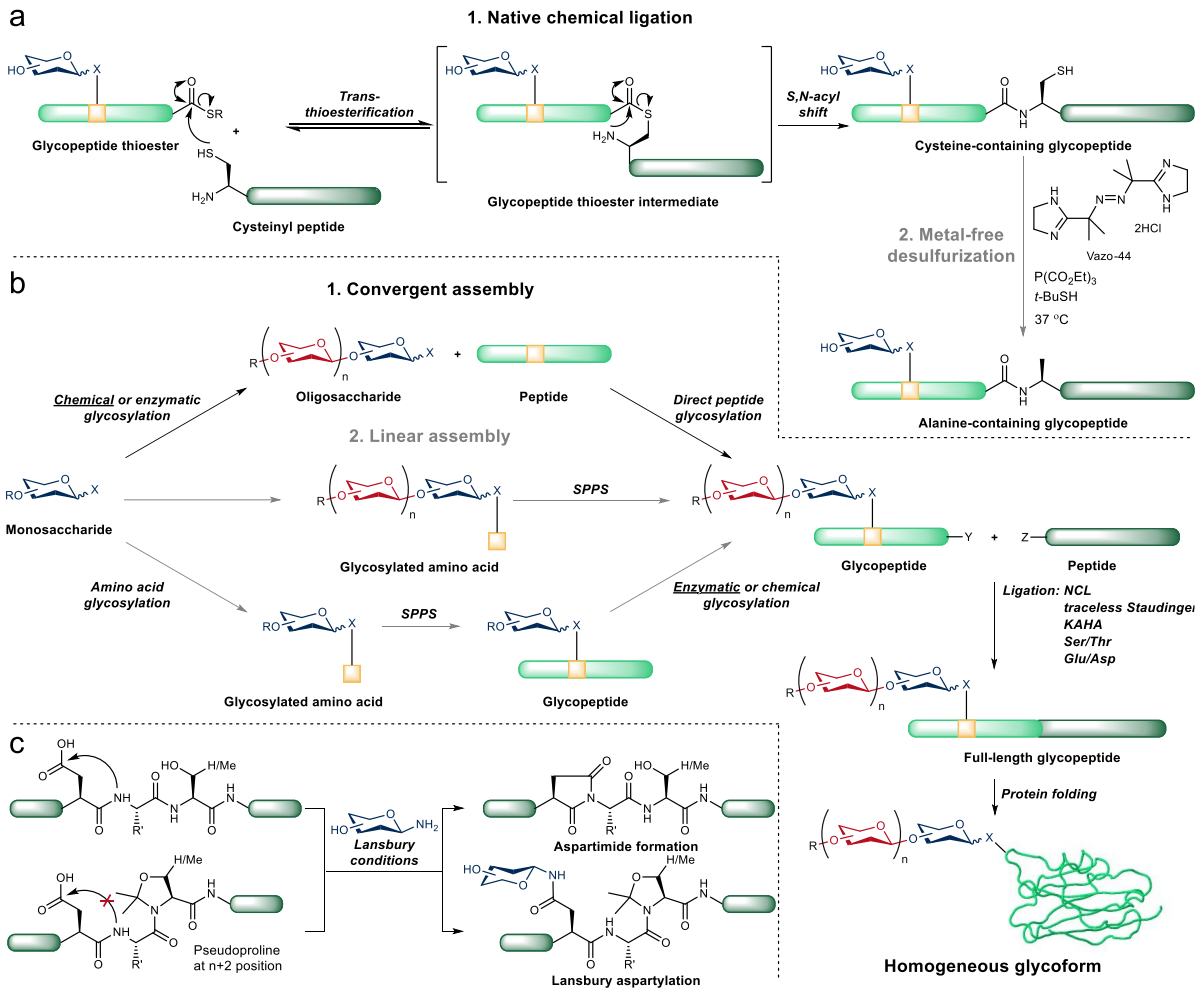


Figure 1.2. The chemical synthesis of homogeneous glycoproteins. The native chemical ligation (1)/metal-free desulfurization (2) combined strategy is one of the most evolved methods used to interconnect separately obtained (glyco)peptides (**a**). The glycopeptides can be synthesized *via* a convergent (1) or a linear (2) approach and when advantageous, glycosyltransferases can be used to generate the desired glycosylation features. In addition to NCL, other ligation methods may be used to obtain the desired full-length glycopeptide such as the traceless Staudinger, the KAHA, the Ser/Thr or the Glu/Asp ligation (**b**). The pseudoproline-based aspartylation method allows for complex *N*-glycopeptides to be efficiently obtained in a convergent fashion since this motif, temporary installed at the n+2 position, is able to suppress the undesired aspartimide formation (**c**).

Noteworthily, the synthesis of glycopeptides containing complex glycans may be streamlined by using appropriate glycosyltransferases. These may be employed either to extend a starting sugar and afford the desired oligosaccharide to be further used in the convergent approach,[84,85] generate an extensively glycosylated amino acid to be later utilized in the SPPS synthesis[86] or construct a complex glycopeptide from a simpler one, in which the substrate had already been incorporated during the linear assembly[87] (Figure **1.2**, b). Once the desired glycopeptides are obtained either by a linear or

convergent approach (or a combination of the two) these are ligated to the rest of the peptidic segments obtained by SPPS (and occasionally by bacterial expression)[70] to obtain the full-length glycoprotein linear sequence.

Due to non-trivial advancements that had to be developed to overcome challenges associated with the chemical synthesis of complex glycoproteins, it took more than 10 years of research until Danishefsky *et al.*[88] reported the first total synthesis of a single erythropoietin (EPO) glycoform (Figure **1.3**).

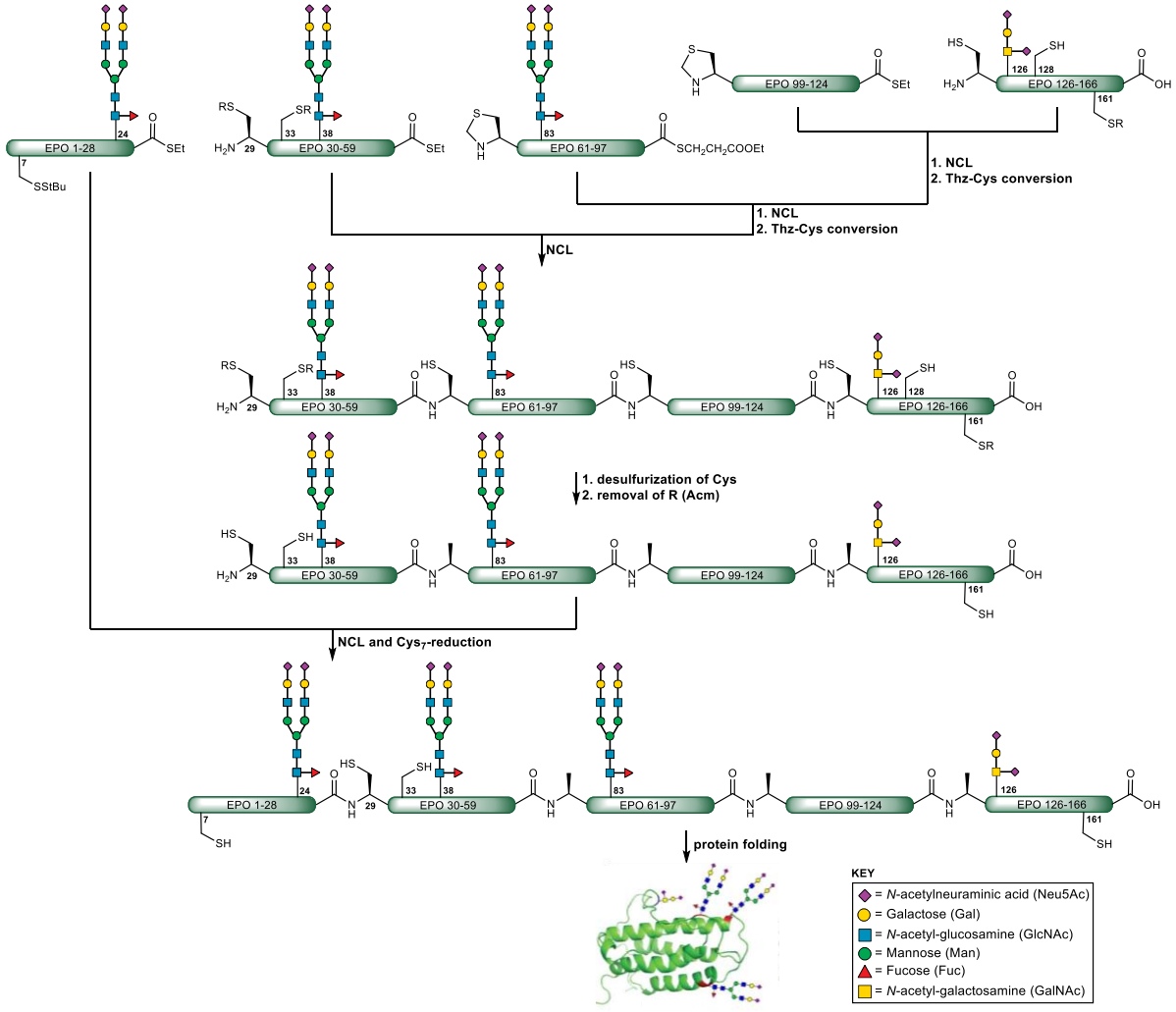


Figure 1.3. Chemical strategy developed by Danishefsky *et al.* to obtain a homogeneous EPO glycoform. The approach combined solid-phase peptide synthesis (SPPS) of desired (glyco)peptides which were joined by native chemical ligation, with appropriate Cys-protection and desulfurization methods to obtain the denatured EPO. Upon folding, the homogeneously glycosylated protein adopted its 3D active conformation (Thz, thiazolidine; Acm, acetamidomethyl).

In fact, their approach only became successful after reaching important synthetic milestones such as: i) the advancement of a metal-free dethiylation protocol which allowed the exploitation of alanine (Ala) residues as NCL disconnection points, more frequent and suitably placed in the EPO sequence than the Cys ones (Figure **1.2**, a.2);[89] ii) the synthesis of the *O*-linked glycopeptide through

development of the so called "cassette" strategy (Figure **1.2**, b.2);[90] iii) the synthesis of the anomeric NH_2-functionalized bi-antennary *N*-glycan by Kochetkov amination,[91,92] and iv) the convergent assembly of *N*-glycopeptides *via* pseudoproline-based Lansbury aspartylation[93] (Figure **1.2**, c).

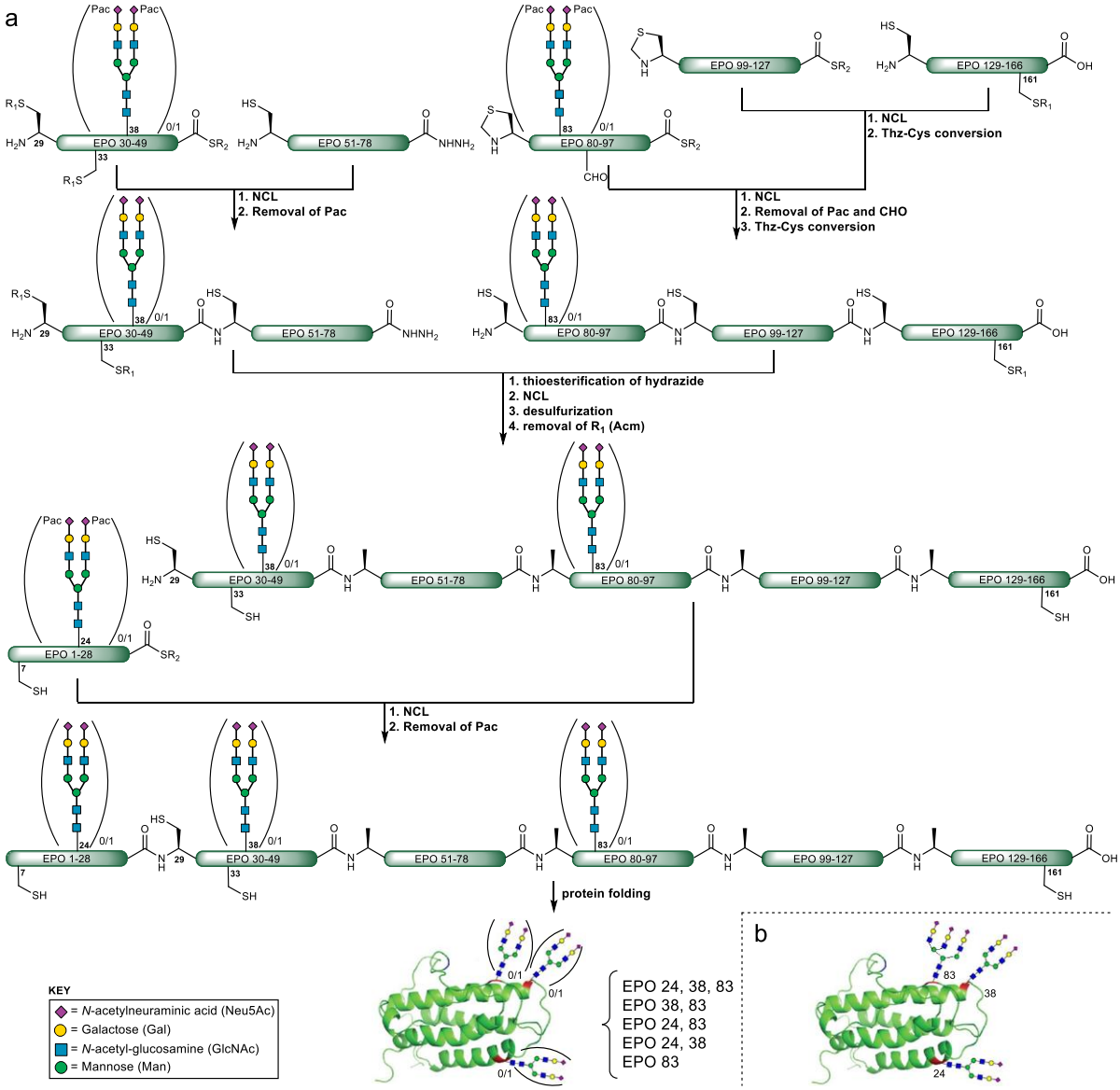


Figure 1.4. Chemical strategy developed by Kajihara *et al.* The group obtained a SAR of homogeneously glycosylated EPOs presenting bi-antennary *N*-glycans exclusively (**a**) or in combination with a tri-antennary oligosaccharide at Asn83 (**b**) (Thz, thiazolidine; Acm, acetamidomethyl; Pac, phenoxyacetyl).

Another chemical synthesis of homogeneous EPO glycoforms was later reported by Kajihara and co-workers,[94] and their elegant, flexible design allowed for the first structure-activity relationship (SAR) to be established among EPO species carrying one to three bi-antennary oligosaccharides (Figure **1.4**, a). The strategy was streamlined by several synthetic considerations such as: i) the use of non-fucosylated Asn-linked glycans which were readily extracted from egg yolk; ii) the protection of their terminal sialic acids from intramolecular hydrolysis with phenoxyacetyl (Pac) groups, and iii) the Gln78Ala substitution which helped avoid undesired intramolecular cyclization of this glutamine (Gln)

NCL disconnection point. Their SAR study was instrumental in determining that the maximum *in vitro* cell proliferation could be achieved only when the tri-glycosylated EPO was used, and that glycoforms bearing glycans at Asn38 and Asn83 had the highest *in vivo* activities among all the other di-glycosylated EPO derivatives that were tested. These findings were complemented by additional, recently published work which showed that a tri-antennary glycan placed at the Asn83 position caused a dramatic increase of the hematocrit levels in mice, similarly to the rEPO positive standard (Figure **1.4**, b).[95]

While both of these strategies are exquisite, labour-intensive, organic chemistry endeavours, they suffer from major drawbacks such as low yields and lengthy periods of time spent in overcoming synthetic challenges which are often structure-specific for any chosen glycoprotein. These may be lessened sometimes by strategically selecting intermediates which can be produced by a living organism, such as, in the EPO case, a large segment of its protein sequence (Figure **1.5**).

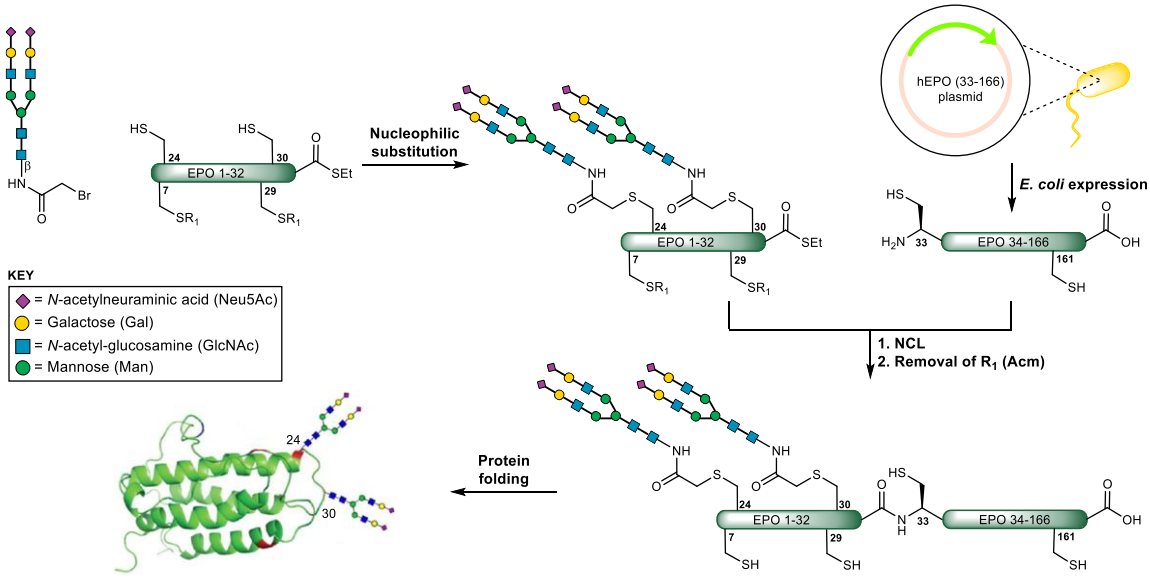


Figure 1.5. Semi-synthetic strategy developed by Kajihara et *al.* to obtain a homogeneously glycosylated EPO mimic presenting two bi-antennary *N*-glycans at amino acids 24 and 30. A 32-residue *N*-terminus segment of EPO was chemically synthesized and an additional Cys was inserted at position 30 in the amino acid sequence. Nucleophilic substitution of the bromoacetamidyl undecasaccharide by the selectively deprotected Cys24 and Cys30 afforded the desired bi-glycosylated peptide-thioester. The latter was joined by NCL to the remaining EPO linear sequence obtained by heterologous expression. Final folding afforded a glycoform analogue that presented one of the glycans at the unnatural 30th position in the amino acid sequence of EPO (Acm, acetamidomethyl; ➡ gene EPO fragment).

For example, Kajihara *et al.*[96] expressed a large EPO fragment (amino acids 33-166) in a bacterial host and as a result, high yields of this peptide, non-glycosylated at Asn38 and Asn83 were obtained (Figure **1.5**). This was subsequently ligated to a short glycopeptide in which the bi-antennary glycans had been installed at the 24th (natural) and the 30th (unnatural) amino acids using the haloacetamide method.[97] The obtained homogeneous glycoform-mimic was able to induce

proliferation of bone marrow cells despite having non-native linkages between the protein and the two oligosaccharides, one of which had been installed on a slightly shifted position (30 instead of 38).

To conclude, in all the cases in which EPO glycoforms derived by chemical synthesis were described, a final, *in vitro* folding step was required to obtain the individual glycoproteins in their active, 3D conformations. This is a problem that should not be overlooked given that the glycoform folding can be a potentially unsuccessful endeavour and consequently, it may lead to undesired aggregation.[94] On occasion, these types of results can be highly informative with respect to similar aggregation events occurring *in vivo* and thus be instrumental in identifying various pathological conditions caused by such phenomena. However, accessing single glycoform(-mimic)s faster and not through such a "*Sisyphus*-type" approach may prove more valuable in studying other important biological events, besides protein aggregation. To this end, the genetic or chemical manipulation of glycosylation pathways in various organisms (Section **1.2.2**), the glycan remodeling of heterogeneous glycoprotein mixtures (Section **1.2.3**) and the glyco-conjugation of appropriately tagged proteins (Section **1.2.4**) are some alternative appealing strategies that can be pursued.

1.2.2 Engineering glycosylation pathways

As mentioned, an alternative approach that is often utilized to generate glycoproteins with a lower degree of heterogeneity is to produce them in glycoengineered eukaryotic organisms.[64] These are expression hosts in which glycan-processing enzymes involved in the native glycosylation pathways are altered either chemically, by using small-molecule enzymatic inhibitors,[98] or genetically, using knockdown, nontargeted overexpression or targeted gene editing strategies.[99] The latter approaches encompass the manipulation of genes that encode for specific glycosyltransferases, hydrolases and other enzymes involved in the biosynthesis and transport of appropriately activated nucleotide sugar donors. Thus, by interfering at a genetic level, the biosynthetic glycosylation machinery of a chosen organism is tweaked to generate glycoproteins with defined, human-like glycosylation profiles (Figure **1.6**).

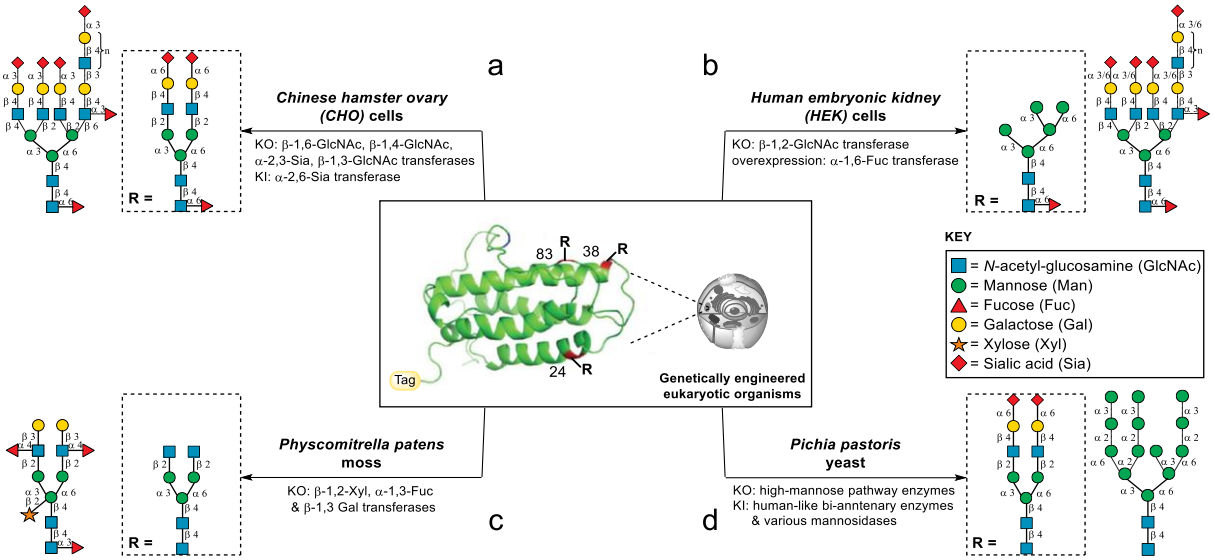


Figure 1.6. Major achievements in genetic, glycoengineering of cells to produce homogeneous EPO glycoforms. Humanized *CHO* cells afforded homogeneously glycosylated EPO carrying bi-antennary, sialylated *N*-glycans (**a**). A genetically engineered *HEK* cell line produced EPO species presenting core-fucosylated, curtailed oligomannose *N*-linked sugars (**b**). Mutagenized *Physcomitrella patens* moss generated truncated EPO *N*-glycoforms devoid of plant-based immunogenic epitopes (**c**). The yeast-type hypermannosylation of EPO was replaced with the human-like, complex *N*-glycosylated protein in an ambitious endeavour of genetic engineering in *Pichia pastoris* (**d**) (**R** = homogeneous *N*-glycosylation features obtained on EPO; KO = knock out; KI = knock in; the structures next to the dashed squares are representative, species-specific *N*-glycans for visual comparison).

The expression of EPO, one of the most investigated therapeutic glycoproteins,[100] in various mutant cell lines has generated several biotechnological products with various degrees of glycoform heterogeneity. For instance, the Chinese hamster ovary (CHO) cell line, a mammalian strain capable of human-like, heterogeneous glycosylation but lacking the Sia-α(2,6)-transferase activity,[101] was subjected to a combinatorial approach of knock out (KO) and knock in (KI) of various glycan-processing enzymes, to produce EPO mixtures with a more homogeneous glycosylation profile.[102] In this precise gene editing study, one of the generated clones lacked the enzymes responsible for the GlcNAc-β(1,6)- and the GlcNAc-β(1,4)-branching of *N*-glycans, as well as the GlcNAc-β(1,3)-transferase, the enzyme that through extension of Gal gives poly-*N*-acetyl-lactosamine (polyLacNAc) elongated chains. In addition, genetic knock out of the α(2,3)-sialylation in the cells, combined with knock in of the human Sia-α(2,6)-transferase activity led to the production of homogeneous EPO glycoforms bearing human-like, bi-antennary, sialylated *N*-glycans (Figure **1.6**, a).

Conversely, a stable Human embryonic kidney (HEK) cell line was generated in which the GlcNAc-β(1,2)-transferase activity had been knocked down using the ethyl methanesulfonate (EMS) mutagen.[103] This enzyme is responsible for the late processing step in which the biosynthesis of *N*-glycans can diverge from the oligomannose type, to generate the hybrid and the complex Asn-linked oligosaccharides. Therefore, the engineered cells are able to produce a single EPO glycoform, the key

intermediate specie bearing exclusively $Man_5GlcNAc_2$-glycans. Moreover, through the lentivirus-mediated gene transfer approach, the overexpression of Fuc-α(1,6)-transferase was achieved and the corresponding fully core-fucosylated, high-mannose EPO glycoform was obtained (Figure **1.6**, b).[104]

Nowadays, plants are increasingly being considered as factories to produce therapeutic glycoproteins.[105] Nevertheless, special consideration has to be given to efficiently eliminate certain plant-specific glycosylation patterns that are highly immunogenic in humans.[106,107] For instance, knock out of the glycosyltransferases that catalyse β(1,2)-xylosylation, α(1,3)-fucosylation and β(1,3)-galactosylation of *N*-glycoproteins in the *Physcomitrella patens* moss, afforded highly homogeneous EPO bearing the minimal $GlcNAc_2Man_5GlcNAc_2$-glycans (Figure **1.6**, c).[108] This approach not only established this plant as a viable bioreactor alternative for the expression of curtailed *N*-glycosylated proteins but it also confirmed that Fuc-α(1,4)-transferase does not modify a substrate lacking the terminal Gal-residues.

Lastly, a homogeneous, bi-antennary, sialylated EPO was efficiently expressed by a "humanized" *Pichia pastoris* yeast strain (Figure **1.6**, d).[109] This was extensively engineered to first abrogate the yeast-type hypermannosylation through gene knock out, then the pentasaccharide core glycoprotein was obtained by overexpression of several mannosidases and finally, the human-like *N*-glycans were re-constructed by knocking in the GlcNAc-, Gal-transferase activities and the biosynthetic enzymes necessary for optimal sialylation.

While undoubtedly tremendous progress has been made in the glycoengineering of various expression hosts to produce more homogeneous glycoproteins, these remain rather tedious, trial-and-error endeavours. One of the most notable challenges is that often, the exogeneous enzyme has to be co-expressed with a transmembrane domain for the desired ER/Golgi localization to be reached. Thus, it is the resulted chimera protein that, if appropriately constructed, will determine the correct sub-cellular retention and thus, the optimal enzymatic function.[99,110]

1.2.3 Direct enzymatic protein glycosylation and glycoprotein remodeling

Alternative strategies that can be exploited to make homogeneously glycosylated proteins make use of designated glycan-processing enzymes.[111] These can be divided into methods that comprise the direct glycosylation of the protein backbone through either sequential monosaccharide additions or *en bloc* transfer of complex oligosaccharides (Figure **1.7**, a and black arrows), and approaches aimed at the 'on-protein' glycan remodeling of natural, heterogeneous mixtures of glycoforms (Figure **1.7**, b and grey arrows).[64] The latter are based on the use of a glycosidase able to trim the heterogeneous species to generate a minimal glycoform bearing a residual sugar-tag. This can

then be recognized and extended by appropriate glycosyltransferases or mutant endoglycosidases to afford the desired, remodeled, homogeneous glycoform.[64,112]

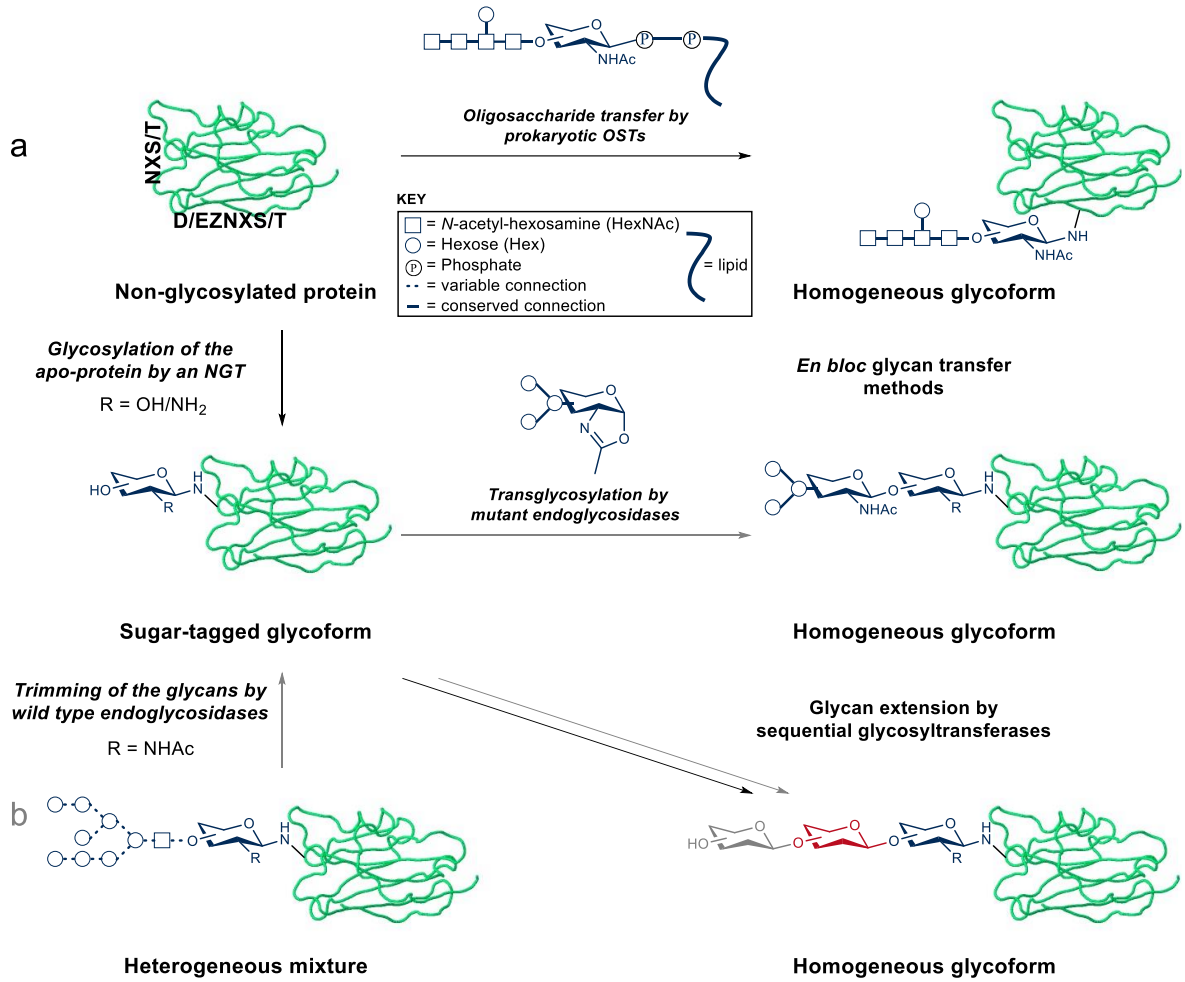


Figure 1.7. The concepts of direct enzymatic protein glycosylation and glycoprotein remodelling. A protein can be *en bloc* glycosylated or mono-glycosylated, by a prokaryotic oligosaccharyltransferase (OST) or a *N*-glycosyltransferase (NGT) respectively, if it contains the required sequon (**a** and black arrows, X and Z are any amino acid except Pro). A heterogeneous glycoprotein mixture can be subjected to enzymatic trimming to yield a homogeneous, sugar-tagged glycoform that can then be extended with a preassembled glycan oxazoline or by sequential additions of separate monosaccharides (**b** and grey arrows).

Recently, Wang *et al.*[113] developed a powerful chemoenzymatic strategy to remodel the unnatural, fully fucosylated, high-mannose EPO glycoform which had been expressed by a genetically engineered HEK cell line (Figure **1.6**, b). This excellent starting material was trimmed using the endoglycosidase H (EndoH) to afford a homogeneous GlcNAc-tagged acceptor, a suitable substrate for the D165A glycosynthase mutant derived from the endoglycosidase F3 (EndoF3-D165A) (Figure **1.8**, a). The EndoF3-D165A catalysed the transglycosylation of two bi-antennary oxazoline-activated glycans at position 38 and 83 and the resulted glycoform was further engineered by a second endoglycosidase-based glycosynthase (EndoA). The latter transferred an azido-functionalized, oxazoline-activated glycan at position 24 which could be used for additional chemical manipulations.

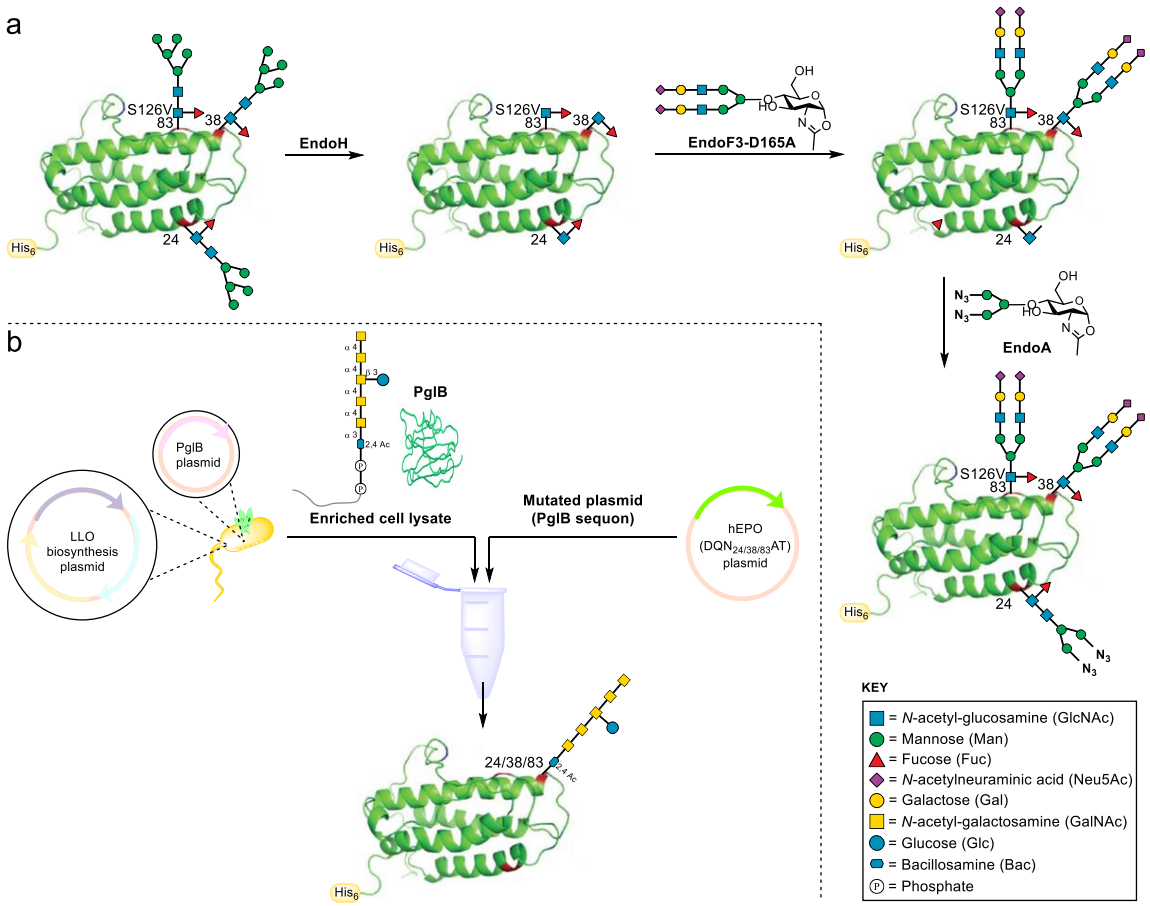


Figure 1.8. The concepts of glycoprotein remodeling and direct enzymatic glycosylation as applied to EPO. The unnatural, homogeneously glycosylated EPO was first trimmed by EndoH to afford a GlcNAc-tagged intermediate. This was first remodeled with two biantennary *N*-glycans at positions 38 and 83 using the EndoF3-D165A mutant and then with an azido-functionalized oligosaccharide using the EndoA enzyme to obtain a fully glycosylated EPO analogue (**a**). An OST-dependent glycosylation of EPO in glycoengineered *E. coli* enriched extracts afforded unnatural glycoforms presenting the characteristic *Campylobacter jejuni* PglB-transferred heptasaccharide (**b**) (mutated EPO gene, PglB gene).

The direct enzymatic glycosylation of proteins is an alternative approach by which specific *N*-linked glycoforms can be obtained. In this case, the site-specificity is ensured by a consensus amino acid sequence (NXS/T) which is characteristic to *N*-glycoproteins. Then, the Asn-linked simple glycan can be elongated stepwise using soluble glycosyltransferases, in a controlled, ordered fashion (Figure **1.7**, a and black arrows).[112]

Furthermore, for more complex glycoproteins, the process can be streamlined using membrane-bound oligosaccharyltransferases (OSTs), enzymes able to catalyse the *en bloc* transfer of an oligosaccharide to their acceptor sequon (D/EZNXS/T). For example, the concept was demonstrated by DeLisa *et al.*[114] who developed a cell-free glycoprotein synthesis platform to obtain a homogeneously glycosylated EPO bearing the unnatural heptasaccharide: GalNAc-α(1,4)-GalNAc-α(1,4)-[Glc-(β1,3)]-GalNAc-α(1,4)-GalNAc-α(1,4)-GalNAc-α(1,3)-2,4-diacetamido-2,4,6-trideoxyglucopyranose (diNAcBac)-β1 (Figure **1.8**, b). The strategy encompassed the development of the

glycoengineered *E. coli* cells which were equipped with the glycosylation machinery derived from the *Campylobacter jejuni*: the single-unit, prokaryotic OST PglB and the additional enzymes, responsible for the biosynthesis of the lipid-linked oligosaccharide (LLO) donors. Then, the *E. coli* cell lysate was enriched in these essential components and primed with the plasmid encoding for the acceptor EPO protein in which the three *N*-glycosylation sites had been individually mutated to the optimal OST sequon: DQNAT.

This is an excellent proof-of-principle exercise to demonstrate that a bacterial *N*-glycosylation system can be rapidly produced in a robust, high-yielding host and then used to glycosylate a therapeutic, eukaryotic protein in a tuneable, open environment. However, more progress needs to be made in order to homogeneously glycosylate proteins with eukaryotic-type *N*-glycans.[115] Current scientific endeavours are focused on glycoengineering *E. coli* cells to efficiently produce either eukaryotic OSTs complexes and LLOs[116] or bacterial *N*-glycosylation machineries with relaxed donor (LLO)[117] and acceptor (sequon) specificities[118,119].

1.2.4 Site-selective chemical glycosylation of proteins

The need to readily access homogeneous glycoproteins has been driving a new area of research, increasingly focused on finding chemo-selective strategies that allow the covalent linkage of a suitably functionalized glycan, to a pre-determined position of the expressed protein backbone: the so called "tag-and-modify" strategy.[120]

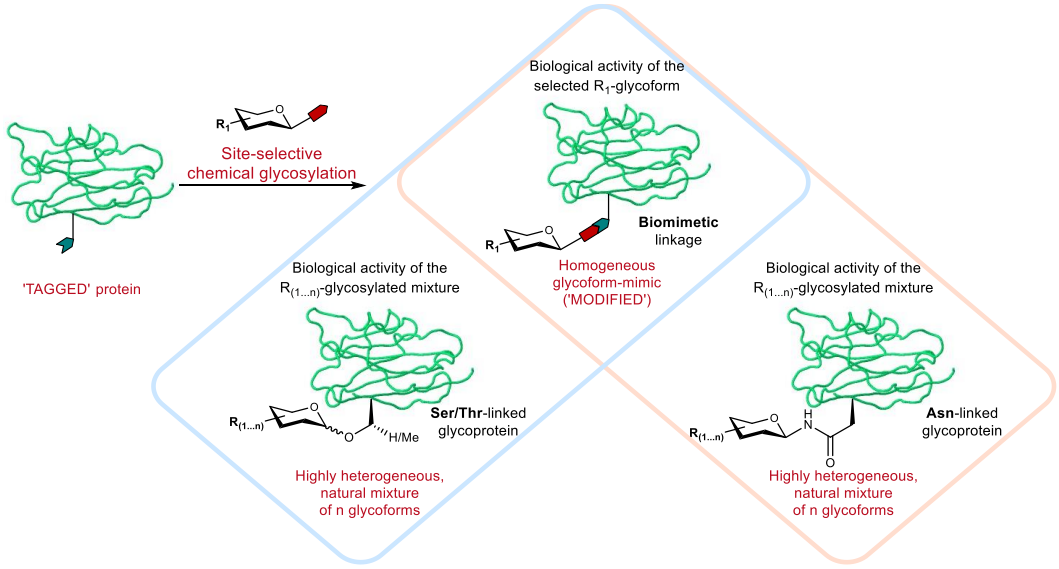


Figure 1.9. The "tag-and-modify" concept applied to generate homogeneous glycoform-mimics of both *O*- and *N*-linked heterogeneous glycoproteins.

This approach involves the initial installation of a uniquely reactive amino acid residue ("tag") in the protein that can then be glycosylated ("modified") by a derivatized oligosaccharide, *via* a benign

chemical ligation (Figure **1.9**). A plethora of methods have been described for the site-selective glycosylation of proteins and while unnatural, the resulted glycan-protein disconnections aim to closely resemble the native *N*- and *O*- linkages and thus provide optimal biomimicry.

1.2.4.1 Cysteine incorporation: site-directed mutagenesis

Due to its low abundance in proteins,[121] Cys is, among the 20 canonical amino acids, the most amenable one to be employed as a tag, for the site-selective chemical glycosylation of proteins. Cys can be straightforwardly introduced at the desired glycosylation site *via* site-directed mutagenesis and used as such or upon chemical/enzymatical derivatization, to attach the appropriately pre-functionalised glycan.

Non-derivatized Cys-tagged proteins for the generation of homogeneous glycoform-mimics

Owing to its highest nucleophilicity, the Cys thiolate (S^-) group formed at slightly basic pH, can react with dithiopyridyl-functionalyzed glycans,[122] glycosyl selenosulfide species,[123] methyl- or phenylthiosulfonate-derivatized sugars,[124,125] haloacetamidated saccharides[97] and glycosyl maleimides[126] (Figure **1.10**, a-e). In addition, the homolytic cleavage of the side chain thiol group generates the corresponding 'on-protein' thiyl radical (S•) that can selectively add to a radical acceptor such as a pre-synthesized, unsaturated glycosyl-derivative (Figure **1.10**, f).[127]

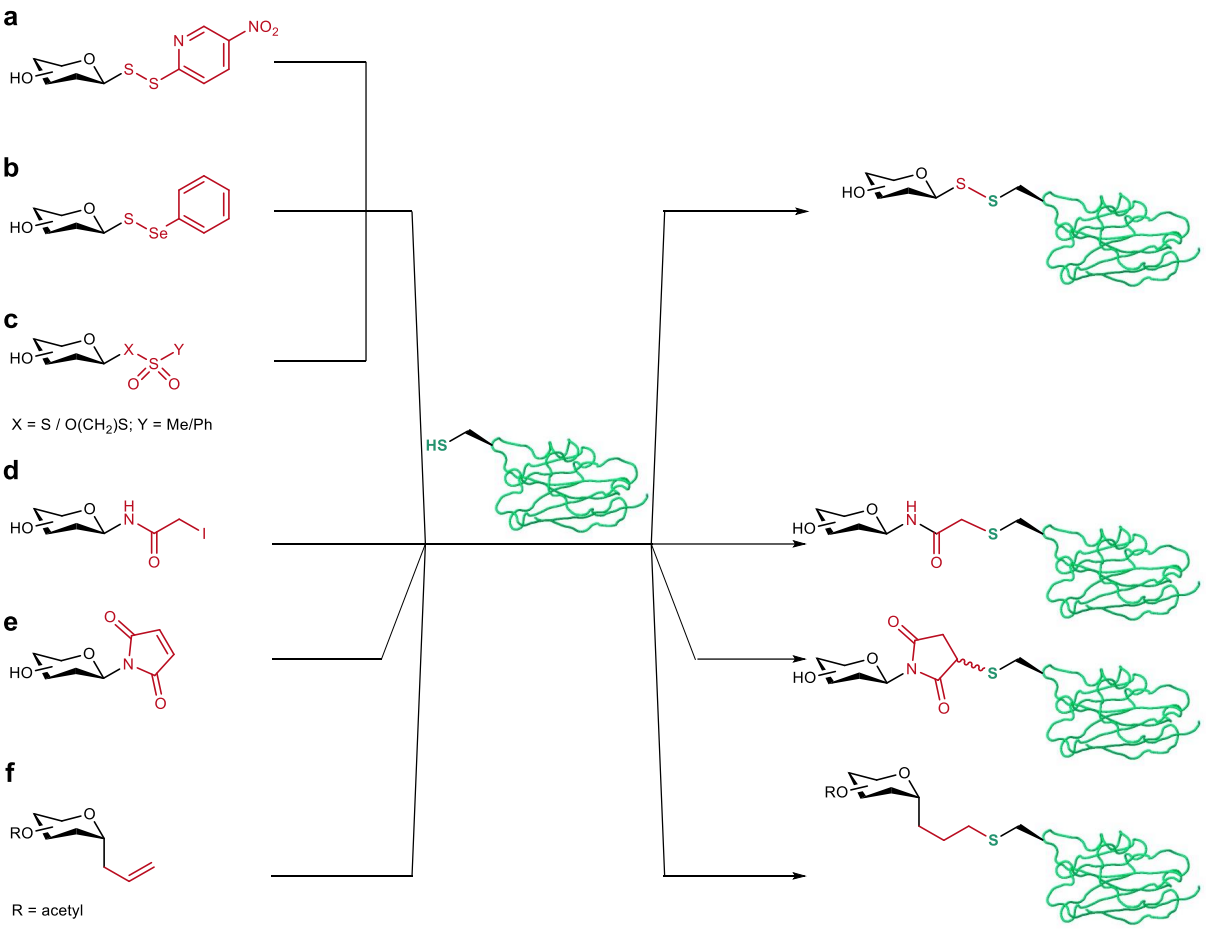


Figure 1.10. Non-derivatized cysteine-based methods for the site-selective glycosylation of proteins. Disulfide and α-thioether amide-linked glycoforms can be accessed using carbohydrates appropriately functionalized with various leaving groups for the S^--mediated nucleophillic substitution (**a**-**d**). The 1,4-conjugate addition of a thiolated protein to a maleimide-containing saccharide (**e**) and the radical addition of the thiyl-protein to an ene-tagged glycan (**f**) provides additional Cys-derived unnatural glyco-linkages.

For instance, Flitsch and co-workers[128] reported a systematic study in which each of the three *N*-glycosylation sites of EPO (Asn24, 38 and 83) was mutated to cysteine and the resulted Cys-tagged, non-glycosylated proteins were expressed in *E. coli* cells. Upon denaturation, the highly aggregated EPO variants were purified using a His-tag affinity column, reduced with dithiothreitol (DTT) and then subjected to oxidative refolding. After this, each of the three tagged proteins (Cys24-, 38- and 83-EPO) was reacted with a small, 1-iodoacetamido-β-monosaccharide (Figure **1.11**) and the strategy led to the formation of very simple EPO glycoform-mimics which only contained an additional thio-methylene in their *N*-glycosidic linkages. Often times, many of the native Cys-residues found in proteins are forming intramolecular disulfide bridges and in the case of EPO these are connected as follows: Cys7-Cys161 and Cys29-Cys33. Protein disulfides do not interfere with the site-selective modification of only the intended, Cys-tagged glycosylation site, as long as they are correctly formed during the expression of the protein or its subsequent *in vitro* refolding. In this study, the selective GlcNAc-installation was unequivocally demonstrated by collision induced dissociation mass spectrometry fragmentation (CID

MS/MS) only for the digested Cys83-EPO, whereas in the case of the other variants, containing three Cys-residues clustered in a small region of the protein sequence (Cys24…Cys29-Cys33 & Cys29-Cys33…Cys38), the modification site remained uncertain. In addition, although at pH 8.0 cysteine is significantly more reactive than other nucleophilic amino acids (e. g. lysine, histidine, or methionine residues) an off-target GlcNAc-installation was identified on the *N*-terminus polyhistidine-linked hexapeptide as well. The formation of the bi-glycosylated side-products was abrogated by adding imidazole, a competitor nucleophile, to the protein reaction mixtures.

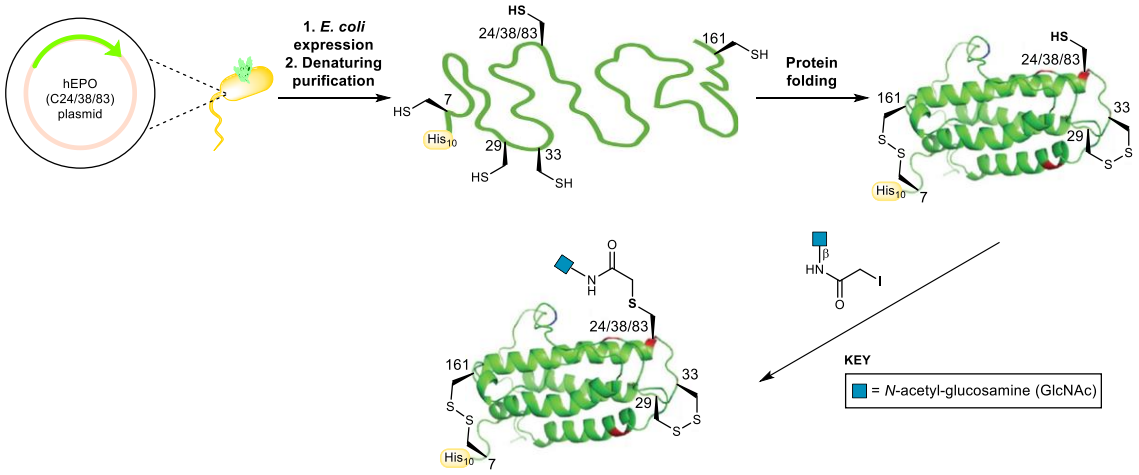


Figure 1.11. A non-derivatized cysteine-based method for the site-selective glycosylation of EPO. Cys-tagged variants were expressed in a bacterial host and upon purification in reducing and denaturing conditions, these were refolded. The disulfide re-bridging was assumed to have occurred as in the native protein, but this was partially demonstrated for the Cys83-EPO mutant upon reaction with the 1-iodoacetamido-β-GlcNAc (⟹ Asn-Cys mutated EPO gene).

Chemically derivatized Cys-tagged proteins for the generation of homogeneous glycoform-mimics

Additional methods that were developed for the site-selective chemical glycosylation of proteins involve further derivatization of the Cys-tag (Figure **1.12**).

The chemical mutation of Cys, a nucleophilic amino acid to dehydroalanine (Dha), an electrophilic Michael-acceptor offers the opportunity of inverting the polarity of the disconnection and the creation of novel, unnatural glycosidic linkages (Figure **1.12**, a). Davis and co-workers exploited the unique reactivity of the Dha-residue and upon conjugate addition of 1-thio-saccharides, *S*-linked mimics of *O*-glycosylated proteins were readily obtained.[129-131]

Moreover, the group also recognized the potential of the Dha-tagged protein to act as a radical acceptor and thus, they developed an unprecedented $C(sp^3)$–$C(sp^3)$ radical ligation strategy that allowed the synthesis of excellent *N*- and *O*-linked glyco-mimics.[132] Glycosylated alkyl radicals were formed from the corresponding precursor halides and were selectively added to the C_β of the Dha-

residue. Then, upon quenching with a suitable "radical hydride" source several *N*- and *O*-glycosylated proteins were generated. These were evaluated as potential substrates for various glycan-processing enzymes thereby revealing interesting aspects with regards to their plasticity.

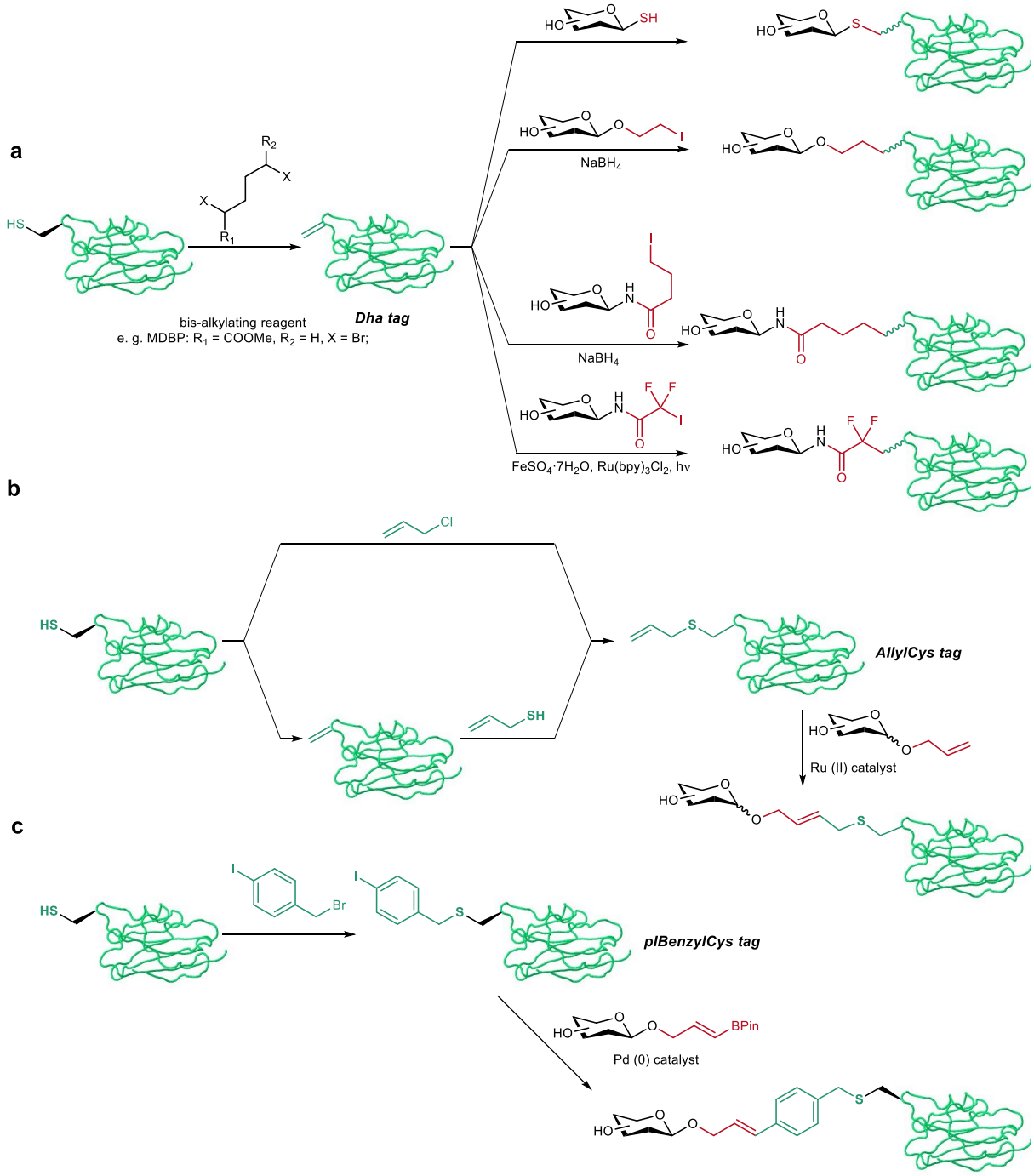


Figure 1.12. Chemically derivatized cysteine-based methods for the site-selective glycosylation of proteins. Conversion of a Cys-residue into the Dha tag generates a versatile chemical handle: the thia-Michael addition of a 1-thio-saccharide affords the corresponding *S*-linked glycoform-mimic whereas additions of glycosylated carbon-centred C• radicals give analogous *O*- and *N*-glycoproteins containing methylene-extended linkages exclusively (**a**). The installation of an allyl moiety *via* several chemical routes generates the AllylCys-tagged protein which in turn undergoes Ru-mediated olefin cross-metathesis with allyl glycosides (**b**). In a nucleophilic substitution, the Cys-tagged protein is functionalized with a *p*-iodobenzyl handle that can then be attached to glycosylated vinylboronic acids through a biocompatible Suzuki–Miyaura cross-coupling (**c**).

In recently published studies, Davis *et al.*[133] perfected the post-translational radical editing of proteins by developing a visible light, Ru(II)/Fe(II)-mediated photoredox route to generate, a different *N*-glycosylated carbon-centred radical. This, not only added efficiently to the Dha-tagged protein but it concomitantly allowed for a difluoro label to be incorporated into the unusual, glutamine (Gln)-linked glycoprotein.

An attractive chemical derivatization of the Cys-handle is represented by the *S*-allylcysteine (AllylCys) tag which can be installed either through thia-Michael addition of the allyl thiol to the Dha-tagged protein or *via* direct alkylation of the Cys-containing biomolecule with allyl chloride (Figure **1.12**, b).[134] The obtained AllylCys-tagged protein can then be readily coupled to an unsaturated saccharide, under Ru(II)-mediated olefin cross-metathesis reaction conditions.[135,136]

Finally, conversion of the Cys-tag into the *p*-iodobenzylcysteine residue (*p*IBenzylCys) affords a suitable protein substrate for the Suzuki-Miyaura cross-coupling to various boronic acids/esters (Figure **1.12**, c).[137] Among these, a glycosylated pinacolboronic ester gives, *via* the biocompatible Pd-mediated $C(sp^2)$-$C(sp^2)$ bond forming ligation strategy, semi-synthetic glycoproteins presenting a rather extended glycosidic linkage.

Enzymatically derivatized Cys-tagged proteins for the generation of homogeneous glycoform-mimics

By introducing the Cys-residue as part of a CXPXR consensus sequence (X represents any amino acid, except Pro) this can be selectively oxidized to an unusual formylglycine (fGly)-tag by the co-expressed fGly-generating enzyme (FGE).[138] For instance, this Cys post-translational modification was recently utilized to generate the aldehyde-tagged core protein of glypican-3 (GPC3), a proteoglycan (PG) expressed on the surface of living cells.[139] GPC3 naturally presents two serines (495 and 509) carrying two heparosan sulfate (HS) chains but after insertion of the FGE sequons, subsequent enzymatic derivatization of the Cys to fGly, and reaction with hydrazide-labeled glycosaminoglycans (GAGs) of various length and composition, several PG-hydrazone conjugates were obtained (Figure **1.13**). These proved to be instrumental in determining that GPC3 engineered with tetradecasaccharide HS or heparosan (Hep) GAG chains have the highest activity in stimulating the Wnt (Wingless/integrated) cascade and the most pronounced effect in inhibiting the Hedgehog (Hh) signalling pathway.

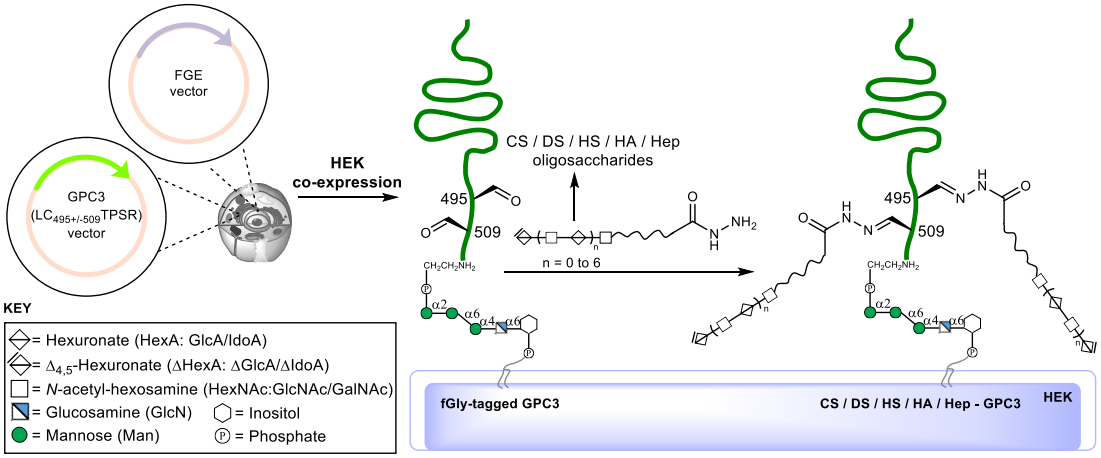


Figure 1.13. An enzymatically-derivatized cysteine-based method for the site-selective glycosylation of a PG. Li F. *et al.* co-expressed the GPC3 together with the FGE, an enzyme capable of the *in situ* oxidation of the Cys-residues 495 and 509 to fGly-tags. The generated electrophilic handles were reacted with hydrazide-functionalized GAGs to afford homogeneously glycosylated PG-mimics of various nature and length (CXPXR mutated GPC3 gene, FGE gene).

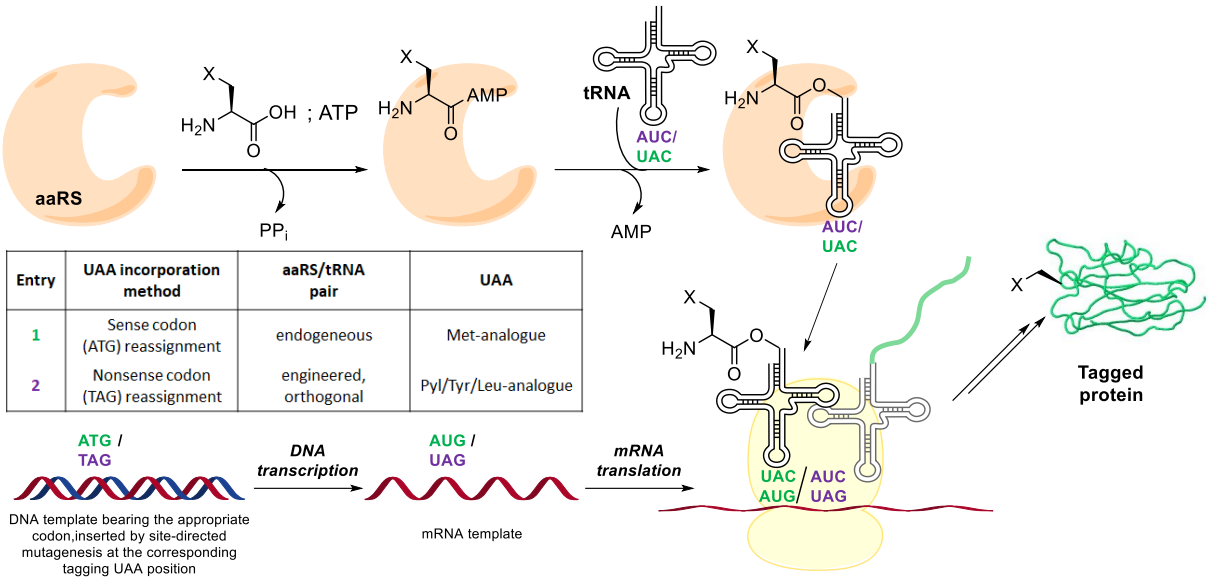
1.2.4.2 Genetic unnatural amino acid incorporation: sense and nonsense codon reassignment

Other methods that proved their utility in generating homogeneous glycoprotein-mimics are based on the programmed incorporation of unnatural amino acid (UAA) tags into recombinant proteins.[140] These strategies are particularly important when a Cys-based "tag-and-modify" strategy cannot be applied either because similarly reactive, native Cys residues present in the protein substrate do not allow for the glycosylation to occur in a regioselective manner, or the site-selective Cys-based glycosylation does not afford a glyco-conjugated protein with suitable biomimetic properties.

In cells, DNA is transcribed into messenger RNA (mRNA) and the latter is translated into proteins by ribosomes equipped with transfer RNA (tRNA) adaptors. Each tRNA molecule is charged with its cognate amino acid by a specific aminoacyl-tRNA synthetase (aaRS) and upon appropriate base pairing interactions between the mRNA codon and the anticodon of the aminoacyl-tRNA, the amino acid residue gets incorporated into the growing protein sequence (Figure **1.14**).[141-144]

There are mainly two approaches that have been employed to incorporate UAAs into the primary structure of a target protein. The first is based on reassigning a sense codon that naturally encodes for a rare, canonical amino acid (e. g. ATG for methionine, Met) to insert a structurally similar, non-canonical analogue (Figure **1.14**, entry 1). Thus, by exploiting the flexibility of the existing aaRS/tRNA pair, in an expression host unable to synthesize the encoded natural amino acid

(auxotrophic) but grown in Met-depleted media supplemented with the UAA, the global incorporation of the latter throughout the entire proteome of the organism is accomplished.[145]



Entry	UAA incorporation method	aaRS/tRNA pair	UAA
1	Sense codon (ATG) reassignment	endogeneous	Met-analogue
2	Nonsense codon (TAG) reassignment	engineered, orthogonal	Pyl/Tyr/Leu-analogue

Figure 1.14. Sense and nonsense codon reassignment technologies illustrated for the specific examples of global methionine replacement (green) and amber codon suppression (violet). A template DNA bearing the ATG/TAG codon is transcribed and translated. The tRNA charged by the corresponding aaRS with the UAA recognises the mRNA codon and ultimately, the expressed protein incorporates the tagging amino acid at the desired position (X = amino acid side chain carrying the tag; ATP, adenosine triphosphate; AMP, adenosine monophosphate; PP_i, pyrophosphate).

This method has proven to be particularly useful to introduce homoallylglycine (Hag), homopropargylglycine (Hpg) or azidohomoalanine (Aha), as replacements for Met, by mutating the desired site intended for the incorporation of the tagging UAA to the ATG codon (Figure **1.15**, a-c).[146,147]

For instance, the free-radical addition of 1-glycosyl thiols to the olefinic side chain of Hag-tagged proteins allowed for various *S*-linked protein glycoconjugates to be efficiently obtained (Figure **1.15**, a).[148] Most notably, this chemical glycosylation applied in a reversed manner to the previously described hydrothiolation conjugation (Figure **1.10**, f) generated a Qβ virus-like particle carrying 180 Glc-residues, one per each Hag-tagged monomer.

In addition, the copper-catalysed azide-alkyne cycloaddition (CuAAC), due to its excellent chemo-selectivity and fast kinetics under protein-friendly reaction conditions, has been successfully adapted to allow the conjugation of both azide- and alkyne-functionalized saccharides to Hpg- (Figure **1.15**, b)[149,150] and Aha-tagged proteins (Figure **1.15**, c),[151-153] respectively.

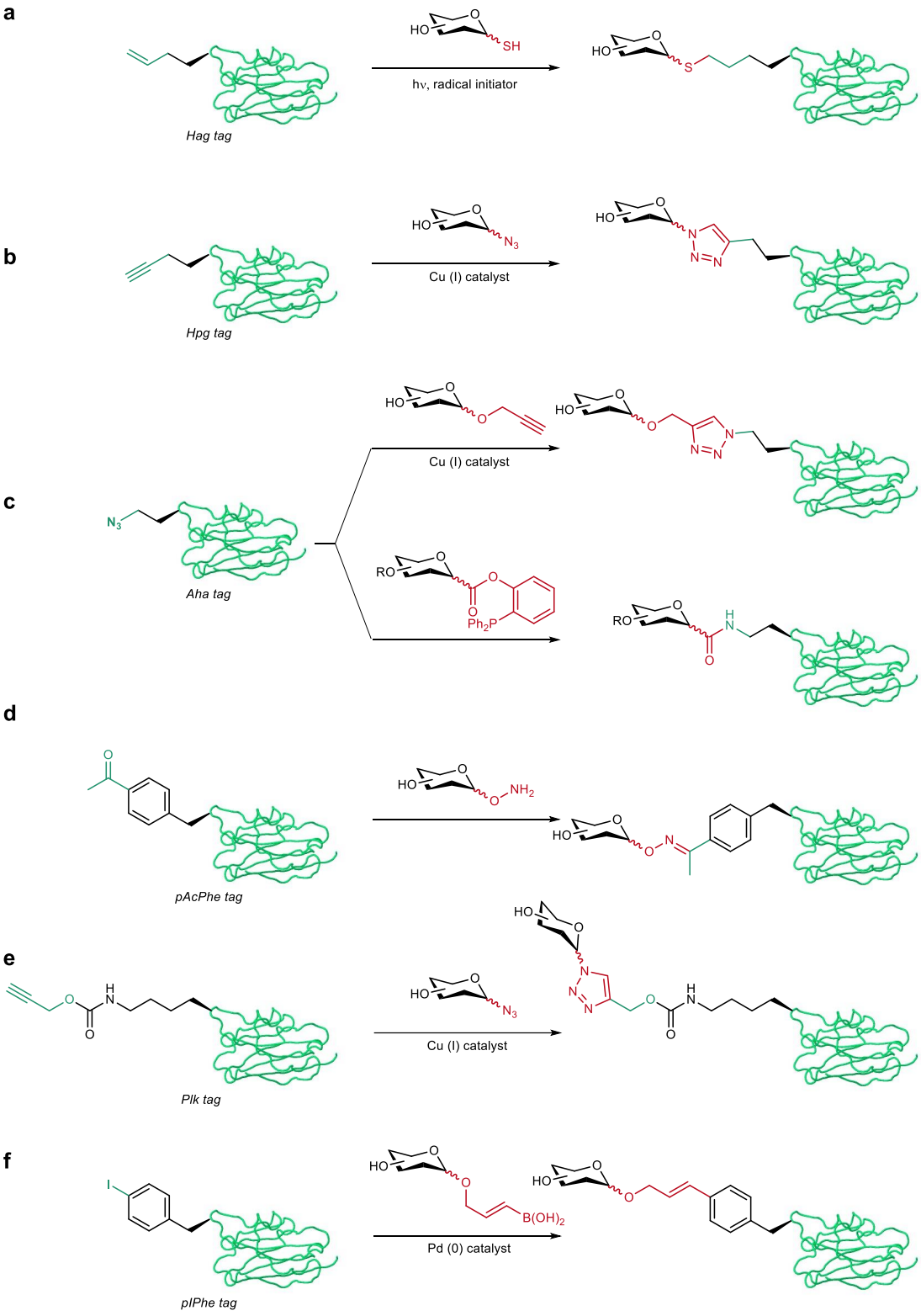


Figure 1.15. Unnatural amino acids as tags for the site-selective glycosylation of proteins. The thiol-ene radical addition to a Hag-tagged protein (**a**), the CuAAC to a protein containing the Hpg (**b**) or the Aha non-canonical amino acid and the "traceless" Staudinger ligation to the latter (**c**) are strategies that were used to generate homogeneous glycoform-mimics *via* the Met-codon reassignment method. Three approaches developed using the amber codon suppression technology allowed the synthesis of analogous glycoforms: condensation of a *p*AcPhe-tagged protein to a hydroxylamine sugar (**d**), the CuAAC of an azido-glycan to the Plk-tagged protein (**e**) and the Suzuki-Miyaura cross-coupling of a glycosylated vinylboronic acid to a *p*IPhe-containing protein (**f**).

Moreover, the introduction of the unnatural Aha residue into a protein generates the corresponding azide-tagged variant that can be an excellent starting material for the "traceless" Staudinger ligation with a phosphine-functionalised carbohydrate (Figure **1.15**, c).[154] Thus, the construction of reverse amide *N*-linked glycoproteins proceeds with moderate yields as long as Cys-dependent side-reactions are avoided.

In all of these great examples described so far, unless a protein does not contain any other Met residue, the site-specific incorporation of the UAA can only be achieved upon mutation of the native methionine residues to closely resembling natural amino acids (e. g. Isoleucine, Ile). However, this is not a universally applicable compromise since it may affect the structural and thus the biological activity of the (glyco)protein to which is being applied.

Since the tolerability of a protein to such extensive mutagenesis experiments is different in every case, a site-specific method for the incorporation of UAAs is generally preferred: the so called "genetic code expansion" technology (Figure **1.14**, entry 2).[140-142,155,156]

This second approach relies on the reassignment of a nonsense codon, typically a rarely occurring one, in the frequently employed *E coli.* expression host (the amber stop codon, TAG)[157] and on the use of an orthogonal aaRS/tRNA pair derived from a different species, of which selectivity towards the UAA is enhanced by directed evolution and rational design[142]. Accordingly, site-directed mutagenesis to TAG of the codon corresponding to the position intended for the UAA incorporation is followed by introduction of the resulted DNA template together with the orthogonal tRNA/aaRS encoding plasmid into the bacterial host. Subsequent proliferation of the transformed cells in UAA-containing media allows the expression of the desired, site-specifically tagged protein.

Over the past two decades, the genetic code expansion approach has become one of the most powerful strategies used not only to incorporate hundreds of unique UAAs that can be subsequently manipulated *via* bio-orthogonal chemical transformations but also to introduce post-translationally modified versions of natural amino acids.[156] For instance, nitrated[158] and sulfated Tyr,[159] phosphorylated Ser,[160] analogues of phosphorylated Tyr,[161,162] monomethylated,[163,164] dimethylated,[165] acetylated[158] and crotonylated Lys[166] have been instrumental in generating a platform of homogeneous, site-specifically modified proteins that can help elucidate the biological role of various, post-translationally regulated phenomena.

Despite the immense progress, so far, the *in vivo* genetic code expansion with glycosylated amino acids has not been achieved,[167] although several studies have reported the amber codon suppression incorporation of a glycosylated Thr,[168] Ser and Tyr[169] in an *in vitro* setting, using the corresponding enzymatically pre-charged, glycosylated aminoacyl-tRNAs.

It remains thus, that out of the myriad of UAAs that were successfully genetically encoded into proteins, a few demonstrated their applicability in the semi-synthesis of homogeneous glycoform-mimics *via* their bio-orthogonal ligation to suitably derivatized carbohydrates.

For instance, the TyrRS/tRNA^{Tyr} pair from the archaebacteria *Methanococcus jannaschii* was evolved by Schultz *et al.*[170] to incorporate the ketone-containing unnatural amino acid *p*-acetylphenylalanine (*p*AcPhe) into the Z domain of staphylococcal protein A. This electrophilic chemical handle was then targeted by unprotected saccharides bearing a hydroxylamino group at the reducing end to afford oxime-ligated glycoconjugates (Figure **1.15**, d).[171]

Another strategy to generate erythropoietin analogues was recently established by Rubini and co-workers[172] who used an engineered bacterial strain to express EPO proteins in which the three Asn glycosylation sites had been systematically substituted with the unnatural aminoacid N^6-(propargyloxycarbonyl)-L-lysine (Plk), by exploiting the pyrrolysyl (Pyl)RS/tRNA^{Pyl} orthogonal pair from *Methanosarcina barkeri* (Figure **1.15**, e and Figure **1.16**, a).[173] Thus, the resulted site-selectively alkyne-tagged protein scaffolds were conjugated to an azido-functionalized bi-antennary sialylated/non-sialylated glycan through copper-catalyzed azide–alkyne cycloaddition (CuAAC) to afford six biologically active, homogeneous EPO glycoform-mimics.

In a previous study,[174] the group had already employed the amber codon suppression technology to insert a *p*-azidophenylalanine (pAzF) handle at the three *N*-glycosylation sites and upon Staudinger-phosphite ligation, PEG chains of different length (750/2000 Da units) were installed (Figure **1.16**, b). The obtained soluble EPO derivatives showed increased stability against enzymatic proteolysis and a similar or higher *in vitro* activity to that of the rEPO.

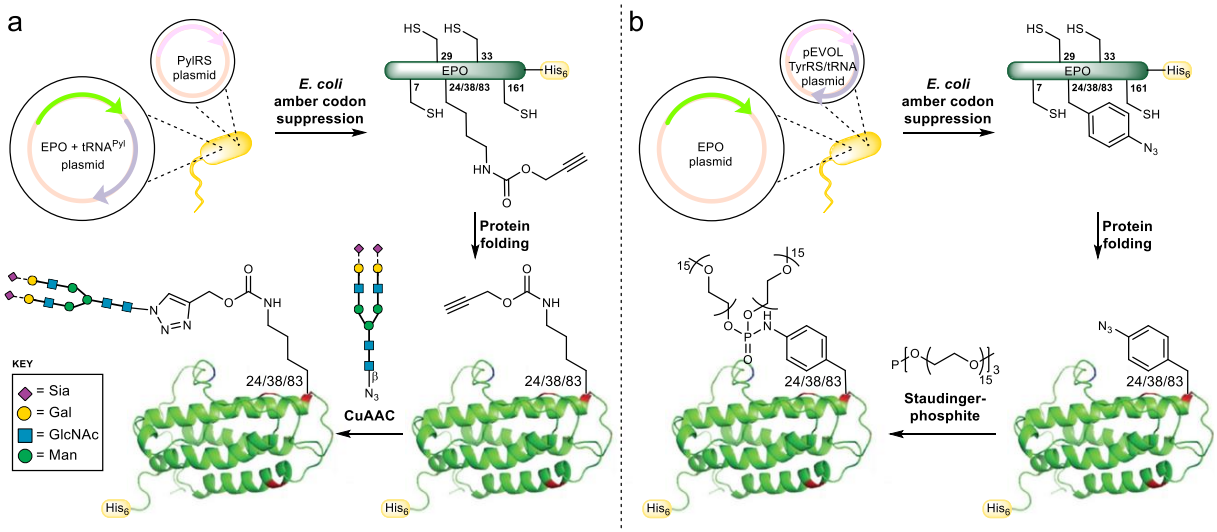


Figure 1.16. Semi-synthetic strategies developed by Rubini *et al.* to obtain homogeneous EPO analogues. The group employed CuAAC chemistry (**a**) and Staudinger-phosphite ligation (**b**) to install a bi-antennary glycan and PEG chains, respectively, at the amino acid position 24, 38 or 83 which had been appropriately tagged using the amber codon suppression approach (EPO gene, tRNA gene, Pyl/TyrRS gene).

Perhaps the most striking example in which the genetic code expansion methodology was instrumental in answering a complex but highly relevant biological question was when applied to decipher the early stages of the GYG-initiated glycogen biogenesis.[175] Davis and co-workers used a genetically tagged variant of GYG bearing the unnatural *p*-iodophenylalanine (*p*IPhe) in place of the catalytic Tyr195, and then developed a palladium-mediated ligation strategy to access catalytically active mimics of GYG-autoglucosylation states (Figure **1.15**, f and Figure **1.17**). Comparative analyses of synthetic GYG-intermediates revealed unusual triphasic kinetics (slow-fast-slow) for the initiation of glycogen biosynthesis corresponding to various stages of enzyme plasticity towards using non-glucose sugars. Moreover, the alignment of these results with crystallographic data suggested that kinetic phases may reflect transitions between inter- and intra-monomeric mechanisms.

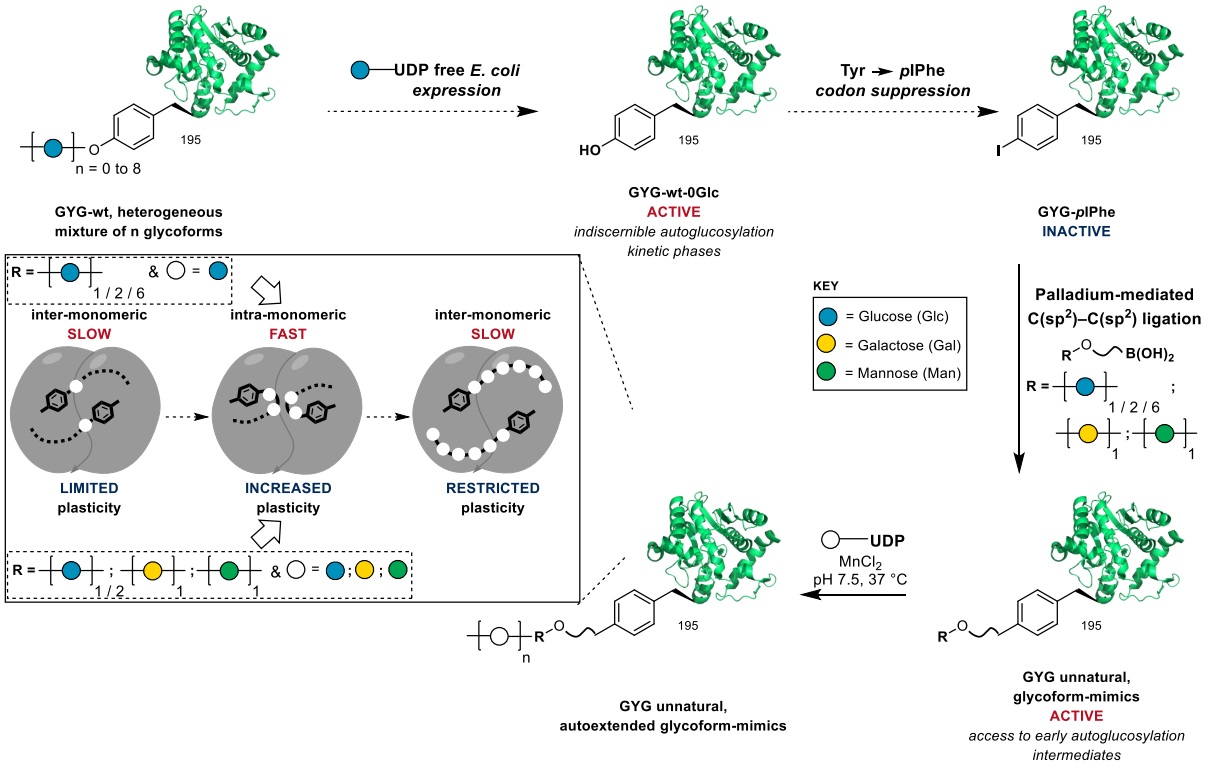


Figure 1.17. Enzyme-compatible, palladium-mediated C(sp²)–C(sp²) ligation enables mechanistic insights into the biocatalytic activity of GYG. The bacterial over-expression of GYG typically yields a heterogeneous mixture of partially glucosylated glycoforms (GYG-wt), whereas expression in an *E. coli* strain deficient in UDP-glucose pyrophosphorylase activity affords the non-glucosylated enzyme (GYG-wt-0Glc). The study of both these variants cannot provide additional information regarding the early autoglucosylation sub-phases of GYG. On the contrary, an inactive form was generated through the Tyr195 → *p*IPhe mutation and upon subjecting GYG-*p*IPhe to Pd-mediated cross-coupling reactions, catalytically active GYG intermediate states were obtained. The approach revealed a surprising triphasic auto-glucosylation mechanism and exposed GYG as both donor and acceptor tolerant (to various degrees) when incorporating non-glucose sugars (⌒ = allyl linker).

1.3 Aims of this thesis

This introduction has outlined the crucial and diverse roles that glycoproteins have in living organisms while highlighting the need for new tools that allow the elucidation of glycoform-dependent biological activities. How a certain glycosylation profile, selected from a hugely abundant repertoire of potential glycoforms, affects the interaction of a studied glycoprotein with its biological partners is notoriously challenging to determine.

Following a survey of existing methodologies for the generation of precisely glycosylated proteins and of closely resembling glycoform-mimics thereof, we identified the "tag-and-modify" strategy as a rather robust and generally applicable method to facilitate structure-activity relationship (SAR) studies of glycoproteins. However, although numerous such glycosylation methods have been developed in recent years, they often generate unnatural glycan attachments that suffer from poor mimicry of the original *N*- or *O*-linkages in the emulated glycoproteins.

Thus, the fundamental goal of this thesis is to contribute at developing site-selective glycan ligation strategies that would cause minimal structural perturbation and thus provide access to pre-determined protein glycosylation states with excellent functional biomimicry.

This thesis is essentially structured in three parts.

Chapter 2 describes the unprecedented direct access to hybrid PG glycoforms through grafting of a Hep chain onto decorin, to replace its naturally occurring chondroitin-/dermatan sulfate linear *O*-glycan. We anticipated that trimming of an isolated mixture of decorin glycoforms by apposite GAG lyases would afford a homogenised ene-tagged minimal glycoform. This would be an excellent acceptor for the radical thiyl-glycosylation and could generate a suitably GlcNAc-primed substrate for the Hep-polymerizing enzyme pmHS2. We reasoned that access to GAG-remodeled proteoglycan-analogues will benefit future studies aimed at expanding the current understanding of PG-governed biological events.

Chapter 3 focuses on the development of a "tag-and-modify" strategy for the generation of an erythropoietin glycoform-mimic. We aimed to investigate whether the Cys->Dha chemical mutagenesis method followed by the thia-Michael conjugate addition of a human-like glycan could be applied to a challenging, aggregation-prone protein such as EPO. Thus, emphasis has been placed on studying the selectivity of the chosen strategy for the Cys-mutated *N*-glycosylation site situated in close proximity to one of the natural disulfide bridges of this hormone.

Chapter 4 aims at looking deeper into the biosynthesis of glycogen, the main energy-storage molecule in a wide variety of organisms. For this, various analytical techniques were used to assess variations in the glycosylation extent of human glycogenin (GYG), the protein core of a nascent

glycogen particle, upon interaction with its downstream biosynthetic partners: the glycogen synthase (GYS) and the glycogen branching enzyme (GBE).

1.4 References

1. Gu J, Isaji T, Xu Q, et al. Potential roles of *N*-glycosylation in cell adhesion. *Glycoconj J*. 2012;29(8-9):599-607.

2. Bektas M, Rubenstein DS. The role of intracellular protein *O*-glycosylation in cell adhesion and disease. *J Biomed Res.* 2011;25(4):227-236.

3. Boscher C, Dennis JW, Nabi IR. Glycosylation, galectins and cellular signaling. *Curr Opin Cell Biol*. 2011;23(4):383-392.

4. Tan FYY, Tang CM, Exley RM. Sugar coating: Bacterial protein glycosylation and host–microbe interactions. *Trends Biochem Sci*. 2015;40(7):342-350.

5. Rudd PM, Elliott T, Cresswell P, Wilson IA, Dwek RA. Glycosylation and the immune system. *Science*. 2001;291(5512):2370-2376.

6. Apweiler R, Hermjakob H, Sharon N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta Gen Subj*. 1999;1473(1):4-8.

7. Marth JD. A unified vision of the building blocks of life. *Nat Cell Biol*. 2008;10(9):1015-1016.

8. Varki A. Biological roles of glycans. *Glycobiology*. 2017;27(1):3-49.

9. Reily C, Stewart TJ, Renfrow MB, Novak J. Glycosylation in health and disease. *Nat Rev Nephrol*. 2019;15(6):346-366.

10. Weerapana E, Imperiali B. Asparagine-linked protein glycosylation: From eukaryotic to prokaryotic systems. *Glycobiology*. 2006;16(6):91R-101R.

11. Zielinska DF, Gnad F, Wiśniewski JR, Mann M. Precision mapping of an *in vivo N*-glycoproteome reveals rigid topological and sequence constraints. *Cell*. 2010;141(5):897-907.

12. Stanley P, Taniguchi N, Aebi M. N-glycans. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

13. Varki A, Schnaar RL, Schauer R. Sialic acids and other nonulosonic acids. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

14. Stanley P, Cummings RD. Structures common to different glycans. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

15. Steen PVd, Rudd PM, Dwek RA, Opdenakker G. Concepts and principles of *O*-linked glycosylation. *Crit Rev Biochem Mol Biol*. 1998;33(3):151-208.

16. Smythe C, Caudwell FB, Ferguson M, Cohen P. Isolation and structural analysis of a peptide containing the novel tyrosyl-glucose linkage in glycogenin. *EMBO J*. 1988;7(9):2681-2686.

17. Halim A, Brinkmalm G, Rüetschi U, et al. Site-specific characterization of threonine, serine, and tyrosine glycosylations of amyloid precursor protein/amyloid β-peptides in human cerebrospinal fluid. *PNAS*. 2011;108(29):11848-11853.

18. Brockhausen I, Stanley P. *O*-GalNAc glycans. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

19. Falck D, Haberger M, Plomp R, et al. Affinity purification of erythropoietin from cell culture supernatant combined with MALDI-TOF-MS analysis of erythropoietin *N*-glycosylation. *Sci Rep*. 2017;7(1):1-10.

20. Skibeli V, Nissen-Lie G, Torjesen P. Sugar profiling proves that human serum erythropoietin differs from recombinant human erythropoietin. *Blood*. 2001;98(13):3626-3634.

21. Byeon J, Lim Y, Kim H, Suh J. Structural identification of a non-glycosylated variant at Ser126 for *O*-glycosylation site from EPO BRP, human recombinant erythropoietin by LC/MS analysis. *Mol Cells*. 2015;38(6):496-505.

22. Lindahl U, Couchman J, Kimata K, Esko JD. Proteoglycans and sulfated glycosaminoglycans. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

23. Karamanos NK, Piperigkou Z, Theocharis AD, et al. Proteoglycan chemical diversity drives multifunctional cell regulation and therapeutics. *Chem Rev*. 2018;118(18):9152-9232.

24. Iozzo RV, Schaefer L. Proteoglycan form and function: a comprehensive nomenclature of proteoglycans. *Matrix Biol*. 2015;42:11-55.

25. Smock RG, Meijers R. Roles of glycosaminoglycans as regulators of ligand/receptor complexes. *Open Biol.* 2018;8(180026):1-10.

26. Mizumoto S, Yamada S, Sugahara K. Molecular interactions between chondroitin–dermatan sulfate and growth factors/receptors/matrix proteins. *Curr Opin Struct Biol*. 2015;34:35-42.

27. Esko JD, H. Prestegard J, Linhardt RJ. Proteins that bind sulfated glycosaminoglycans. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

28. Neill T, Schaefer L, Iozzo RV. Decorin: a guardian from the matrix. *Am J Pathol*. 2012;181(2):380-387.

29. Couchman JR, Pataki CA. An introduction to proteoglycans and their localization. *J Histochem Cytochem*. 2012;60(12):885-897.

30. Filmus J, Capurro M, Rast J. Glypicans. *Genome Biol*. 2008;9(5):224.1-224.6.

31. Ferguson MAJ, Hart GW, Kinoshita T. Glycosylphosphatidylinositol anchors. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

32. West CM, Hart GW. Nucleocytoplasmic glycosylation. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

33. Sakabe K, Wang Z, Hart GW. β-*N*-acetylglucosamine (*O*-GlcNAc) is part of the histone code. *PNAS*. 2010;107(46):19915-19920.

34. Dehennaut V, Leprince D, Lefebvre T. *O*-GlcNAcylation, an epigenetic mark. Focus on the histone code, TET family proteins, and polycomb group proteins. *Front Endocrinol (Lausanne)*. 2014;5(155):1-7.

35. Lewis BA, Hanover JA. *O*-GlcNAc and the epigenetic regulation of gene expression. *J Biol Chem*. 2014;289(50):34440-34448.

36. Aon MA, Curtino JA. Protein-bound glycogen is linked to tyrosine residues. *Biochem J*. 1985;229(1):269-272.

37. Issoglio FM, Carrizo ME, Romero JM, Curtino JA. Mechanisms of monomeric and dimeric glycogenin autoglucosylation. *J Biol Chem*. 2012;287(3):1955-1961.

38. Roach PJ, Depaoli-Roach AA, Hurley TD, Tagliabracci VS. Glycogen and its metabolism: some new developments and old themes. *Biochem J*. 2012;441(3):763-787.

39. i. Narasimhan S. Control of glycoprotein synthesis. UDP-GlcNAc:glycopeptide β4-*N*-acetylglucosaminyltransferase III, an enzyme in hen oviduct which adds GlcNAc in β1-4 linkage to the β-linked mannose of the trimannosyl core of *N*-glycosyl oligosaccharides. *J Biol Chem*. 1982;257(17):10235-10242. ii. Song Y, Aglipay JA, Bernstein JD, Goswami S, Stanley P. The bisecting GlcNAc on *N*-glycans inhibits growth factor signaling and retards mammary tumor progression. *Cancer Res*. 2010;70(8):3361-3371.

40. Varki A, Kannagi R, Toole B, Stanley P. Glycosylation changes in cancer. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology*. 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

41. Coles CH, Shen Y, Tenney AP, et al. Proteoglycan-specific molecular switch for RPTPσ clustering and neuronal extension. *Science*. 2011;332(6028):484-488.

42. Tonks NK. Protein tyrosine phosphatases: from genes, to function, to disease. *Nat Rev Mol Cell Biol*. 2006;7(11):833-846.

43. Johnson KG, Van Vactor D. Receptor protein tyrosine phosphatases in nervous system development. *Physiol Rev*. 2003;83(1):1-24.

44. Duan Y, Giger RJ. A new role for RPTPσ in spinal cord injury: signaling chondroitin sulfate proteoglycan inhibition. *Sci Signal*. 2010;3(110):1-4.

45. Parodi A, Cummings RD, Aebi M. Glycans in glycoprotein quality control. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

46. Tannous A, Pisoni GB, Hebert DN, Molinari M. *N*-linked sugar-regulated protein folding and quality control in the ER. *Semin Cell Dev Biol*. 2015;41:79-89.

47. Carter CRD, Whitmore KM, Thorpe R. The significance of carbohydrates on G-CSF: differential sensitivity of G-CSFs to human neutrophil elastase degradation. *J Leukoc Biol*. 2004;75(3):515-522.

48. Raju TS, Scallon B. Fc glycans terminated with *N*-acetylglucosamine residues increase antibody resistance to papain. *Biotechnol Prog*. 2007;23(4):964-971.

49. Walker AK, Soo KY, Levina V, Talbo GH, Atkin JD. *N*–linked glycosylation modulates dimerization of protein disulfide isomerase family A member 2 (PDIA2). *FEBS J*. 2013;280(1):233-243.

50. Yi C, Wang L, Huang J, Pan K, Chen J, Liang Y. Glycosylation significantly inhibits the aggregation of human prion protein and decreases its cytotoxicity. *Sci Rep*. 2018;8(12603):1-14.

51. Lewis YE, Galesic A, Levine PM, et al. *O*-GlcNAcylation of α-synuclein at serine 87 reduces aggregation without affecting membrane binding. *ACS Chem Biol*. 2017;12(4):1020-1027.

52. Tams JW, Vind J, Welinder KG. Adapting protein solubility by glycosylation: *N*-glycosylation mutants of coprinus cinereus peroxidase in salt and organic solutions. *Biochim Biophys Acta, Protein Struct Mol Enzymol.* 1999;1432(2):214-221.

53. Lizak C, Fan Y, Weber TC, Aebi M. *N*-linked glycosylation of antibody fragments in *Escherichia coli*. *Bioconjugate Chem*. 2011;22(3):488-496.

54. The effect of carbohydrate on the structure and stability of erythropoietin*. *J Biol Chem*. 1991;266:23022-23026.

55. Rasheedi S, Haq SK, Khan RH. Guanidine hydrochloride denaturation of glycosylated and deglycosylated stem bromelain. *Biochemistry (Mosc)*. 2003;68(10):1097-1100.

56. Fatima A, Husain Q. A role of glycosyl moieties in the stabilization of bitter gourd (*Momordica charantia*) peroxidase. *Int J Biol Macromol*. 2007;41(1):56-63.

57. Solá RJ, Rodríguez-Martínez JA, Griebenow K. Modulation of protein biophysical properties by chemical glycosylation: biochemical insights and biomedical implications. *Cell Mol Life Sci*. 2007;64(16):2133-2152.

58. Tzeng S, Jiang S. Glycosylation modification improved the characteristics of recombinant chicken cystatin and its application on mackerel surimi. *J Agric Food Chem*. 2004;52(11):3612-3616.

59. Rademacher TW, Parekh RB, Dwek RA. Glycobiology. *Annu Rev Biochem*. 1988;57:785-838.

60. Schjoldager KT, Narimatsu Y, Joshi HJ, Clausen H. Global view of human protein glycosylation pathways and functions. *Nat Rev Mol Cell Biol*. 2020;21(12):729-749.

61. Rudd PM, Dwek RA. Glycosylation: heterogeneity and the 3D structure of proteins. *Crit Rev Biochem Mol Biol*. 1997;32(1):1-100.

62. Zhang P, Woen S, Wang T, et al. Challenges of glycosylation analysis and control: an integrated approach to producing optimal and consistent therapeutic drugs. *Drug Discov Today*. 2016;21(5):740-765.

63. Li C, Wang L. Chemoenzymatic methods for the synthesis of glycoproteins. *Chem Rev*. 2018;118(17):8359-8413.

64. Wang L, Lomino JV. Emerging technologies for making glycan-defined glycoproteins. *ACS Chem Biol*. 2012;7(1):110-122.

65. Li Y, Tran AH, Danishefsky SJ, Tan Z. Chemical biology of glycoproteins: from chemical synthesis to biological impact. *Methods Enzymol*. 2019;621:213-229.

66. Dawson PE, Muir TW, Clark-Lewis I, Kent SB. Synthesis of proteins by native chemical ligation. *Science*. 1994;266(5186):776-779.

67. Merrifield RB. Solid phase peptide synthesis. I. the synthesis of a tetrapeptide. *J Am Chem Soc*. 1963;85(14):2149-2154.

68. Kent S. Origin of the chemical ligation concept for the total synthesis of enzymes (proteins). *Biopolymers*. 2010;94(4):iv-ix.

69. Conibear AC, Watson EE, Payne RJ, Becker CFW. Native chemical ligation in protein synthesis and semi-synthesis. *Chem Soc Rev*. 2018;47(24):9046-9068.

70. Malins LR, Payne RJ. Recent extensions to native chemical ligation for the chemical synthesis of peptides and proteins. *Curr Opin Chem Biol*. 2014;22:70-78.

71. Kulkarni SS, Sayers J, Premdjee B, Payne RJ. Rapid and efficient protein synthesis through expansion of the native chemical ligation concept. *Nat Rev Chem*. 2018;2(4):1-17.

72. Saxon E, Armstrong JI, Bertozzi CR. A “traceless” staudinger ligation for the chemoselective synthesis of amide bonds. *Org Lett*. 2000;2(14):2141-2143.

73. Zhang Y, Xu C, Lam HY, Lee CL, Li X. Protein chemical synthesis by serine and threonine ligation. *Proc Natl Acad Sci USA*. 2013;110(17):6657-6662.

74. Bode JW, Fox RM, Baucom KD. Chemoselective amide ligations by decarboxylative condensations of *N*-alkylhydroxylamines and α-ketoacids. *Angew Chem Int Ed*. 2006;45(8):1248-1252.

75. Thomas GL, Hsieh YSY, Chun CKY, Cai Z, Reimers JR, Payne RJ. Peptide ligations accelerated by *N*-terminal aspartate and glutamate residues. *Org Lett*. 2011;13(18):4770-4773.

76. Fernández-Tejada A, Brailsford J, Zhang Q, Shieh J, Moore MAS, Danishefsky SJ. Total synthesis of glycosylated proteins. *Top Curr Chem*. 2015;362:1-26.

77. Kunz H. Synthesis of glycopeptides, partial structures of biological recognition components [new synthetic methods (67)]. *Angew Chem Int Ed Engl*. 1987;26(4):294-308.

78. Gamblin DP, Scanlan EM, Davis BG. Glycoprotein synthesis: an update. *Chem Rev*. 2009;109(1):131-163.

79. Wang L, Amin M. Chemical and chemoenzymatic synthesis of glycoproteins for deciphering functions. *Chem Biol*. 2014;21(1):51-66.

80. Cohen-Anisfeld ST, Lansbury PT. A practical, convergent method for glycopeptide synthesis. *J Am Chem Soc*. 1993;115(23):10531-10537.

81. Doores KJ, Mimura Y, Dwek RA, Rudd PM, Elliott T, Davis BG. Direct deprotected glycosyl–asparagine ligation. *Chem Commun*. 2006(13):1401-1403.

82. He Y, Hinklin RJ, Chang J, Kiessling LL. Stereoselective *N*-glycosylation by Staudinger ligation. *Org Lett*. 2004;6(24):4479-4482.

83. Zhang Y, Muthana SM, Barchi JJ, Gildersleeve JC. Divergent behaviour of glycosylated threonine and serine derivatives in solid phase peptide synthesis. *Org Lett*. 2012;14(15):3958-3961.

84. Li L, Liu Y, Ma C, et al. Efficient chemoenzymatic synthesis of an *N*-glycan isomer library. *Chem Sci*. 2015;6(10):5652-5661.

85. Liu L, Prudden AR, Capicciotti CJ, et al. Streamlining the chemoenzymatic synthesis of complex *N*-glycans by a stop and go strategy. *Nat Chem*. 2019;11(2):161-169.

86. Naruchi K, Hamamoto T, Kurogochi M, et al. Construction and structural characterization of versatile lactosaminoglycan-related compound library for the synthesis of complex glycopeptides and glycosphingolipids. *J Org Chem*. 2006;71(26):9609-9621.

87. Matsushita T, Hinou H, Fumoto M, et al. Construction of highly glycosylated mucin-type glycopeptides based on microwave-assisted solid-phase syntheses and enzymatic modifications. *J Org Chem*. 2006;71(8):3051-3063.

88. Wang P, Dong S, Shieh J, et al. Erythropoietin derived by chemical synthesis. *Science*. 2013;342(6164):1357-1360.

89. Wan Q, Danishefsky SJ. Free-radical-based, specific desulfurization of cysteine: a powerful advance in the synthesis of polypeptides and glycopolypeptides. *Angew Chem Int Ed*. 2007;46(48):9248-9252.

90. Kuduk SD, Schwarz JB, Chen X, et al. Synthetic and immunological studies on clustered modes of mucin-related Tn and TF *O*-linked antigens: the preparation of a glycopeptide-based vaccine for clinical trials against prostate cancer. *J Am Chem Soc*. 1998;120(48):12474-12485.

91. Wu B, Hua Z, Warren JD, et al. Synthesis of the fucosylated biantennary *N*-glycan of erythropoietin. *Tetrahedron Lett*. 2006;47(31):5577-5579.

92. Wu B, Tan Z, Chen G, et al. Mature homogeneous erythropoietin building blocks by chemical synthesis: The EPO 22–37 glycopeptide domain presenting the full *N*-linked dodecasaccharide. *Tetrahedron Lett*. 2006;47(46):8009-8011.

93. Wang P, Aussedat B, Vohra Y, Danishefsky SJ. An advance in the chemical synthesis of homogeneous *N*-linked glycopolypeptides by convergent aspartylation. *Angew Chem Int Ed*. 2012;51(46):11571-11575.

94. Murakami M, Kiuchi T, Nishihara M, et al. Chemical synthesis of erythropoietin glycoforms for insights into the relationship between glycosylation pattern and bioactivity. *Sci Adv*. 2016;2(1):1-12.

95. Maki Y, Okamoto R, Izumi M, Kajihara Y. Chemical synthesis of an erythropoietin glycoform having a triantennary *N*-glycan: significant change of biological activity of glycoprotein by addition of a small molecular weight trisaccharide. *J Am Chem Soc*. 2020;142(49):20671-20679.

96. Hirano K, Macmillan D, Tezuka K, Tsuji T, Kajihara Y. Design and synthesis of a homogeneous erythropoietin analogue with two human complex-type sialyloligosaccharides: combined use of chemical and bacterial protein expression methods. *Angew Chem Int Ed*. 2009;48(50):9557-9560.

97. Wong SY, Guile GR, Dwek RA, Arsequell G. Synthetic glycosylation of proteins using *N*-(β-saccharide) iodoacetamides: applications in site-specific glycosylation and solid-phase enzymic oligosaccharide synthesis. *Biochem J*. 1994;300(3):843-850.

98. Esko JD, Bertozzi C, Schnaar RL. Chemical tools for inhibiting glycosylation. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

99. Clausen H, Wandall HH, Steentoft C, Stanley P, Schnaar RL. Glycosylation engineering. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

100. Suresh S, Rajvanshi PK, Noguchi CT. The many facets of erythropoietin physiologic and metabolic response. *Front Physiol*. 2019;10(1534):1-20.

101. Lalonde M, Durocher Y. Therapeutic glycoprotein production in mammalian cells. *J Biotechnol*. 2017;251:128-140.

102. Čaval T, Tian W, Yang Z, Clausen H, Heck AJR. Direct quality control of glycoengineered erythropoietin variants. *Nat Commun*. 2018;9(1):1-8.

103. Reeves PJ, Callewaert N, Contreras R, Khorana HG. Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous *N*-glycosylation by a tetracycline-inducible

N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. *Proc Natl Acad Sci USA*. 2002;99(21):13419-13424.

104. Yang Q, Wang L. Mammalian α-1,6-fucosyltransferase (FUT8) is the sole enzyme responsible for the *N*-acetylglucosaminyltransferase I-independent core fucosylation of high-mannose *N*-glycans. *J Biol Chem*. 2016;291(21):11064-11071.

105. Montero-Morales L, Steinkellner H. Advanced plant-based glycan engineering. *Front Bioeng Biotechnol*. 2018;6(81):1-8.

106. Bencúrová M, Hemmer W, Focke-Tejkl M, Wilson IBH, Altmann F. Specificity of IgG and IgE antibodies against plant and insect glycoprotein glycans determined with artificial glycoforms of human transferrin. *Glycobiology*. 2004;14(5):457-466.

107. Bardor M, Faveeuw C, Fitchette A, et al. Immunoreactivity in mammals of two typical plant glyco-epitopes, core α(1,3)-fucose and core xylose. *Glycobiology*. 2003;13(6):427-434.

108. Parsons J, Altmann F, Arrenberg CK, et al. Moss-based production of asialo-erythropoietin devoid of lewis A and other plant-typical carbohydrate determinants. *Plant Biotechnol J*. 2012;10(7):851-861.

109. Nett JH, Gomathinayagam S, Hamilton SR, et al. Optimization of erythropoietin production with controlled glycosylation-PEGylated erythropoietin produced in glycoengineered *Pichia pastoris*. *J Biotechnol*. 2012;157(1):198-206.

110. Kellokumpu S, Hassinen A, Glumoff T. Glycosyltransferase complexes in eukaryotes: long-known, prevalent but still unrecognized. *Cell Mol Life Sci*. 2016;73(2):305-325.

111. Rini JM, Esko JD. Glycosyltransferases and glycan-processing enzymes. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

112. Li C, Wang L. Chemoenzymatic methods for the synthesis of glycoproteins. *Chem Rev*. 2018;118(17):8359-8413.

113. Yang Q, An Y, Zhu S, et al. Glycan remodeling of human erythropoietin (EPO) through combined mammalian cell engineering and chemoenzymatic transglycosylation. *ACS Chem Biol*. 2017;12(6):1665-1673.

114. Jaroentomeechai T, Stark JC, Natarajan A, et al. Single-pot glycoprotein biosynthesis using a cell-free transcription-translation system enriched with glycosylation machinery. *Nat Commun*. 2018;9(1):1-11.

115. Hershewe J, Kightlinger W, Jewett MC. Cell-free systems for accelerating glycoprotein expression and biomanufacturing. *J Ind Microbiol Biotechnol*. 2020;47(11):977-991.

116. Glasscock CJ, Yates LE, Jaroentomeechai T, et al. A flow cytometric approach to engineering *Escherichia coli* for improved eukaryotic protein glycosylation. *Metab Eng*. 2018;47:488-495.

117. Liu F, Vijayakrishnan B, Faridmoayer A, et al. Rationally designed short polyisoprenol-linked PglB substrates for engineered polypeptide and protein *N*-glycosylation. *J Am Chem Soc*. 2014;136(2):566-569.

118. Ihssen J, Haas J, Kowarik M, et al. Increased efficiency of *Campylobacter jejuni N*-oligosaccharyltransferase PglB by structure-guided engineering. *Open Biol*. 2015;5(140227):1-10.

119. Ollis AA, Zhang S, Fisher AC, DeLisa MP. Engineered oligosaccharyltransferases with greatly relaxed acceptor site specificity. *Nat Chem Biol*. 2014;10(10):816-822.

120. Chalker JM, Bernardes GJL, Davis BG. A "Tag-and-modify" approach to site-selective protein modification. *Acc Chem Res*. 2011;44(9):730-741.

121. Miseta A, Csutora P. Relationship between the occurrence of cysteine in proteins and the complexity of organisms. *Mol Biol Evol*. 2000;17(8):1232-1239.

122. Macindoe WM, Oijen AHv, Boons G. A unique and highly facile method for synthesising disulfide linked neoglycoconjugates: a new approach for remodelling of peptides and proteins. *Chem Commun*. 1998(7):847-848.

123. Boutureira O, Bernardes GJL, Fernández-González M, Anthony DC, Davis BG. Selenenylsulfide-linked homogeneous glycopeptides and glycoproteins: synthesis of human "hepatic Se metabolite A". *Angew Chem Int Ed*. 2012;51(6):1432-1436.

124. Gamblin DP, Garnier P, Ward SJ, Oldham NJ, Fairbanks AJ, Davis BG. Glycosyl phenylthiosulfonates (glyco-PTS): novel reagents for glycoprotein synthesis. *Org Biomol Chem*. 2003;1(21):3642-3644.

125. Davis BG, Maughan MaT, Green MP, Ullman A, Jones JB. Glycomethanethiosulfonates: powerful reagents for protein glycosylation. *Tetrahedron Asymmetry*. 2000;11(1):245-262.

126. Shin I, Jung HJ, Lee MR. Chemoselective ligation of maleimidosugars to peptides/protein for the preparation of neoglycopeptides/neoglycoprotein. *Tetrahedron Lett*. 2001;42(7):1325-1328.

127. Dondoni A, Massi A, Nanni P, Roda A. A new ligation strategy for peptide and protein glycosylation: photoinduced thiol-ene coupling. *Chemistry*. 2009;15(43):11444-11449.

128. Macmillan D, Bill RM, Sage KA, Fern D, Flitsch SL. Selective in vitro glycosylation of recombinant proteins: semi-synthesis of novel homogeneous glycoforms of human erythropoietin. *Chem Biol*. 2001;8(2):133-145.

129. Bernardes GJL, Chalker JM, Errey JC, Davis BG. Facile conversion of cysteine and alkyl cysteines to dehydroalanine on protein surfaces: versatile and switchable access to functionalized proteins. *J Am Chem Soc*. 2008;130(15):5052-5053.

130. Raj R, Lercher L, Mohammed S, Davis BG. Synthetic nucleosomes reveal that GlcNAcylation modulates direct interaction with the FACT complex. *Angew Chem Int Ed*. 2016;55(31):8918-8922.

131. Lercher L, Raj R, Patel NA, et al. Generation of a synthetic GlcNAcylated nucleosome reveals regulation of stability by H2A-Thr101 GlcNAcylation. *Nat Commun*. 2015;6(7978):1-9.

132. Wright TH, Bower BJ, Chalker JM, et al. Posttranslational mutagenesis: a chemical strategy for exploring protein side-chain diversity. *Science*. 2016;354(6312):597-608.

133. Josephson B, Fehl C, Isenegger PG, et al. Light-driven post-translational installation of reactive protein side chains. *Nature*. 2020;585(7826):530-537.

134. Lin YA, Chalker JM, Davis BG. Olefin metathesis for site-selective protein modification. *Chembiochem*. 2009;10(6):959-969.

135. Lin YA, Chalker JM, Davis BG. Olefin cross-metathesis on proteins: investigation of allylic chalcogen effects and guiding principles in metathesis partner selection. *J Am Chem Soc*. 2010;132(47):16805-16811.

136. Lin YA, Chalker JM, Floyd N, Bernardes GJL, Davis BG. Allyl sulfides are privileged substrates in aqueous cross-metathesis: application to site-selective protein modification. *J Am Chem Soc*. 2008;130(30):9642-9643.

137. Chalker JM, Wood CSC, Davis BG. A convenient catalyst for aqueous and protein Suzuki−Miyaura cross-coupling. *J Am Chem Soc*. 2009;131(45):16346-16347.

138. Appel MJ, Bertozzi CR. Formylglycine, a post-translationally generated residue with unique catalytic capabilities and biotechnology applications. *ACS Chem Biol*. 2015;10(1):72-84.

139. Wang W, Han N, Xu Y, et al. Assembling custom side chains on proteoglycans to interrogate their function in living cells. *Nat Commun*. 2020;11(1):1-13.

140. Dumas A, Lercher L, Spicer CD, Davis BG. Designing logical codon reassignment - expanding the chemistry in biology. *Chem Sci*. 2015;6(1):50-69.

141. de la Torre D, Chin JW. Reprogramming the genetic code. *Nat Rev Genet*. 2021;22(3):169-184.

142. Arranz-Gibert P, Vanderschuren K, Isaacs F. Next-generation genetic code expansion. *Curr Opin Chem Biol*. 2018;46:203-211.

143. Chin JW. Expanding and reprogramming the genetic code. *Nature*. 2017;550(7674):53-60.

144. Davis L, Chin JW. Designer proteins: applications of genetic code expansion in cell biology. *Nat Rev Mol Cell Biol*. 2012;13(3):168-182.

145. Budisa N, Minks C, Alefelder S, et al. Toward the experimental codon reassignment *in vivo*: protein building with an expanded amino acid repertoire. *FASEB J*. 1999;13(1):41-51.

146. van Hest, Jan CM, Kiick KL, Tirrell DA. Efficient incorporation of unsaturated methionine analogues into proteins *in vivo*. *J Am Chem Soc*. 2000;122(7):1282-1288.

147. Kiick KL, Saxon E, Tirrell DA, Bertozzi CR. Incorporation of azides into recombinant proteins for chemoselective modification by the staudinger ligation. *PNAS*. 2002;99(1):19-24.

148. Floyd N, Vijayakrishnan B, Koeppe JR, Davis BG. Thiyl glycosylation of olefinic proteins: *S*-linked glycoconjugate synthesis. *Angew Chem Int Ed*. 2009;48(42):7798-7802.

149. De Munari S, Schiffner T, Davis BG. A triply divergent reagent for glycoprotein synthesis. *Isr J Chem*. 2015;55(3-4):387-391.

150. Boutureira O, D'Hooge F, Fernández-González M, et al. Fluoroglycoproteins: ready chemical site-selective incorporation of fluorosugars into proteins. *Chem Commun*. 2009;46(43):8142-8144.

151. Kasteren SIv, Kramer HB, Jensen HH, et al. Expanding the diversity of chemical protein modification allows post-translational mimicry. *Nature*. 2007;446(7139):1105-1109.

152. van Kasteren SI, Gamblin DP, Davis BG, Kramer HB. Site-selective glycosylation of proteins: creating synthetic glycoproteins. *Nat Protoc*. 2007;2(12):3185-3194.

153. Merkel L, Beckmann HSG, Wittmann V, Budisa N. Efficient *N*-terminal glycoconjugation of proteins by the *N*-end rule. *ChemBioChem*. 2008;9(8):1220-1224.

154. Bernardes GJL, Linderoth L, Doores KJ, Boutureira O, Davis BG. Site-selective traceless staudinger ligation for glycoprotein synthesis reveals scope and limitations. *Chembiochem*. 2011;12(9):1383-1386.

155. Chin JW. Expanding and reprogramming the genetic code. *Nature*. 2017;550(7674):53-60.

156. Davis L, Chin JW. Designer proteins: applications of genetic code expansion in cell biology. *Nat Rev Mol Cell Biol*. 2012;13(3):168-182.

157. Alff-Steinberger C, Epstein R. Codon preference in the terminal region of *E. coli* genes and evolution of stop codon usage. *J Theor Biol*. 1994;168(4):461-463.

158. Neumann H, Hazen JL, Weinstein J, Mehl RA, Chin JW. Genetically encoding protein oxidative damage. *J Am Chem Soc*. 2008;130(12):4028-4033.

159. Schultz PG, Liu CC. Recombinant expression of selectively sulfated proteins in *Escherichia coli*. *Nat Biotechnol*. 2006;24(11):1436-1440.

160. Park H, Hohn MJ, Umehara T, et al. Expanding the genetic code of *Escherichia coli* with phosphoserine. *Science*. 2011;333(6046):1151-1154.

161. Serwa R, Wilkening I, Del Signore G, et al. Chemoselective staudinger-phosphite reaction of azides for the phosphorylation of proteins. *Angew Chem Int Ed*. 2009;48(44):8234-8239.

162. Xie J, Supekova L, Schultz PG. A genetically encoded metabolically stable analogue of phosphotyrosine in *Escherichia coli*. *ACS Chem Biol*. 2007;2(7):474-478.

163. Ai H, Lee JW, Schultz PG. A method to site-specifically introduce methyllysine into proteins in *E. coli. Chem Commun*. 2010;46(30):5506-5508.

164. Wang Y, Wu B, Wang Z, et al. A genetically encoded photocaged N^{ε}-methyl-L-lysine. *Mol Biosyst*. 2010;6(9):1557-1560.

165. Nguyen DP, Garcia Alai MM, Virdee S, Chin JW. Genetically directing ε-*N, N*-dimethyl-L-lysine in recombinant histones. *Chem Biol*. 2010;17(10):1072-1076.

166. Kim CH, Kang M, Kim HJ, Chatterjee A, Schultz PG. Site-specific incorporation of ε-*N*-crotonyllysine into histones. *Angew Chem Int Ed Engl*. 2012;51(29):7246-7249.

167. Zhang Z, Gildersleeve J, Yang Y, et al. A new strategy for the synthesis of glycoproteins. *Science*. 2004;303(5656):371-373.

168. Matsubara T, Iijima K, Watanabe T, Hohsaka T, Sato T. Incorporation of glycosylated amino acid into protein by an *in vitro* translation system. *Bioorganic Med Chem Lett*. 2013;23(20):5634-5636.

169. Fahmi NE, Dedkova L, Wang B, Golovine S, Hecht SM. Site-specific incorporation of glycosylated serine and tyrosine derivatives into proteins. *J Am Chem Soc*. 2007;129(12):3586-3597.

170. Wang L, Zhang Z, Brock A, Schultz PG. Addition of the keto functional group to the genetic code of *Escherichia coli*. *PNAS*. 2003;100(1):56-61.

171. Liu H, Wang L, Brock A, Wong C, Schultz PG. A method for the generation of glycoprotein mimetics. *J Am Chem Soc*. 2003;125(7):1702-1703.

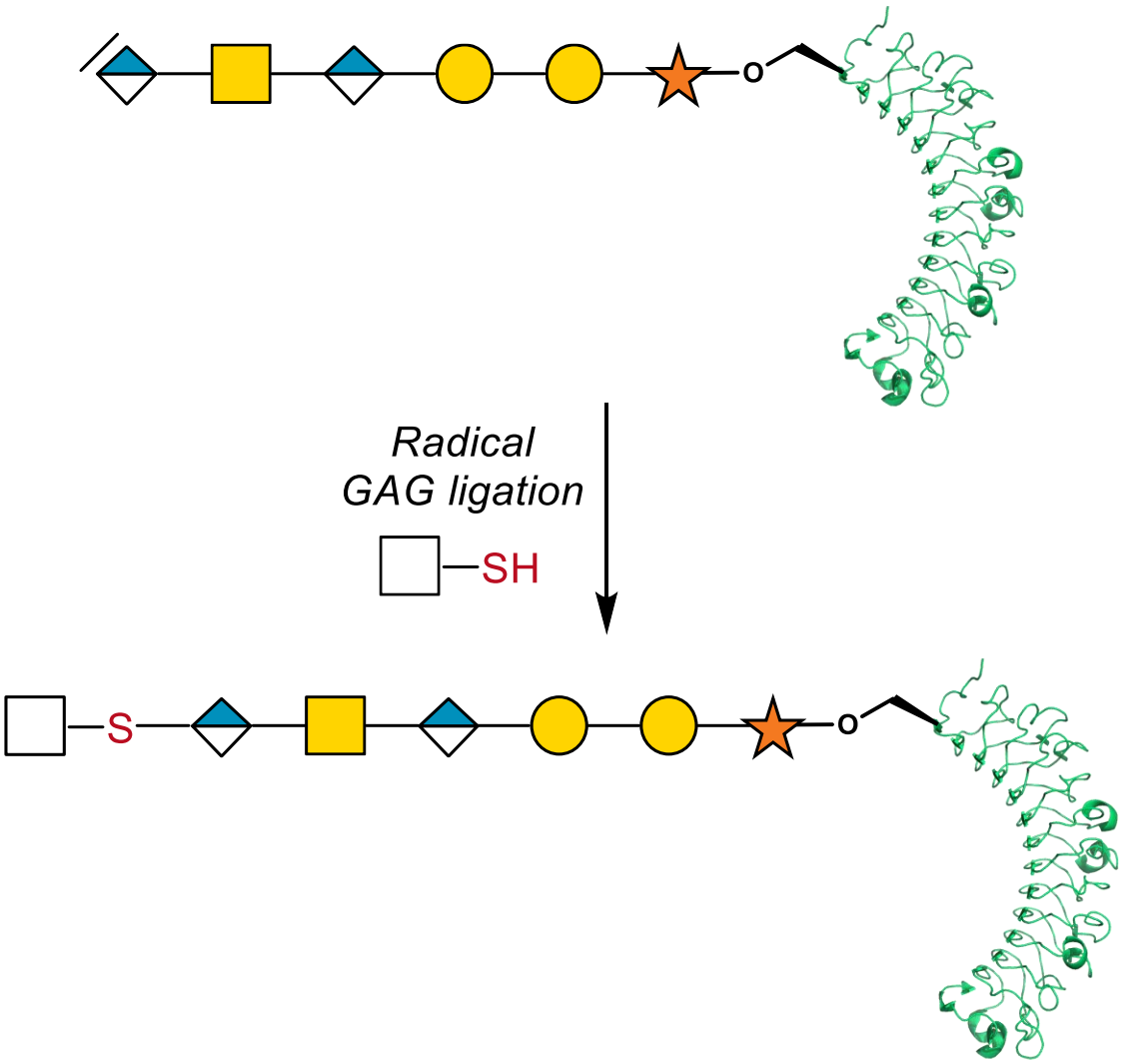
172. Streichert K, Seitz C, Hoffmann E, et al. Synthesis of erythropoietins site-specifically conjugated with complex-type *N*-glycans. *ChemBioChem*. 2019;20(15):1914-1918.

173. Nguyen DP, Lusic H, Neumann H, Kapadnis PB, Deiters A, Chin JW. Genetic encoding and labeling of aliphatic azides and alkynes in recombinant proteins *via* a pyrrolysyl-tRNA synthetase/tRNA(CUA) pair and click chemistry. *J Am Chem Soc*. 2009;131(25):8720-8721.

174. Hoffmann E, Streichert K, Nischan N, et al. Stabilization of bacterially expressed erythropoietin by single site-specific introduction of short branched PEG chains at naturally occurring glycosylation sites. *Mol BioSyst*. 2016;12(6):1750-1755.

175. Bilyard MK, Bailey HJ, Raich L, et al. Palladium-mediated enzyme activation suggests multiphase initiation of glycogenesis. *Nature*. 2018;563(7730):235-240.

Chapter 2



GAG ligation: thiyl-glycosylation allows direct construction of synthetic proteoglycans

Contents

2.1 Introduction 49

2.1.1 Natural heterogeneity impedes the structure-activity relationship studies of proteoglycans 49

2.1.2. Thiyl-glycosylation as a strategy for the direct construction of semi-synthetic, minimal PGs 51

2.1.3 Bacterial glycosyltransferases as protagonists of the PG-remodeling process with non-native GAG chains 53

2.2 Results and discussion 55

2.2.1 Extraction, purification and characterization of a natural PG: decorin 55

2.2.2 Development of a radical GAG ligation to obtain semi-synthetic PG-intermediates 62

2.2.3 Chemo-enzymatic synthesis of neoproteoglycans containing a heparosan GAG chain ... 71

2.2.4 Optimization of the pmHS2-catalysed enzymatic extension and development of the Hep-remodeled decorin 81

2.2.5 Decorin induces DNA transcription of the intercellular adhesion molecule 1 (ICAM-1) in human umbilical vein endothelial cells (HUVECs) 88

2.3 Conclusions and future perspectives 91

2.4 Experimental section 93

2.4.1 Plasmid DNA details and protein sequences 93

2.4.2 Protein extraction/expression and characterisation 95

2.4.2.1 Decorin variants: intact, GAG-trimmed and core decorin 95

2.4.2.2 pmHS2 104

2.4.2.3 Cys-TEV-H3 107

2.4.3 Development of a free-radical hydrothiolation reaction to ligate unprotected glycans to the core decorin 110

2.4.3.1 Optimization of the GlcNAc-addition to the core decorin 110

2.4.3.2 Negative control reactions for the thiyl-glycosylation 115

2.4.3.3 Radical hydrothiolation of decorin with various 1-thio-saccharides 118

2.4.3.4 Characterization of enzymatically digested GlcNAc-S-decorin by LC-MS/MS analysis124

2.4.3.5 Refolding of GlcNAc-S-decorin in the L-arginine containing buffer. 125

2.4.4 Chemo-enzymatic synthesis of neoproteoglycans containing a heparosan GAG chain .127

2.4.4.1 Formation of Dha-TEV-H3, GlcNAc-S-TEV-H3 and GlcAGlcNAc-S-TEV-H3 127

2.4.4.2 Optimization of the enzymatic GlcA-extension of GlcNAc-TEV-H3 by pmHS2 129

2.4.4.3 Formation of various GlcNAc-primed proteins as potential substrates for pmHS2 (GlcNAc-S-H2B, GlcNAc-S-H3, GlcNAc-S-TEV-H3 and GlcNAc-ethyl-S-TEV-H3)..... 131

2.4.4.4 pmHS2 GlcA-transferase activity on various GlcNAc-S-histone substrates (GlcNAc-S-H2B, GlcNAc-S-H3 and GlcNAc-S-TEV-H3)..... 133

2.4.4.5 Enzymatic extension of the GlcNAc-S-TEV-H3 substrate by pmHS2, formation of the GlcAGlcNAc-S-TEV-H3 intermediate and attempted on-protein Hep-polymerizations..... 134

2.4.4.6 Enzymatic extension of the GlcAGlcNAc-S-TEV-H3 substrate by pmHS2 and formation of the Hep-remodeled H3 neoproteoglycan 136

2.4.4.7 GalT enzymatic extension of GlcNAc-S-TEV-H3 with a Gal residue 137

2.4.4.8 ΔpmHS2-catalysed extension of GlcNAc-S-TEV-H3 and GlcNAc-ethyl-S-TEV-H3 with a GlcA residue (comparison to the pmHS2 activity) 138

2.4.4.9 pmHS2-catalysed extension of GlcAGlcNAc-S-TEV-H3 with a Glc residue..... 140

2.4.5 PmHS2, ΔpmHS2 and GalT extend the primed GlcNAc-ethyl-S-decorin but not the GlcNAc-S-variant 141

2.4.5.1 Enzymatic GlcA-extension attempt of GlcNAc-S-decorin by pmHS2 141

2.4.5.2 Enzymatic GlcA-extension attempt of GlcNAc-S-decorin by pmHS2 in presence of 0.5 M urea 142

2.4.5.3 Enzymatic Gal-extension attempt of GlcNAc-S-decorin by GalT..... 142

2.4.5.4 Enzymatic GlcA-extension attempt of GlcNAc-S-decorin by ΔpmHS2 142

2.4.5.5 Evaluation of the glycosyl transferase activities of pmHS2, ΔpmHS2 and GalT using the GlcNAc-ethyl-S-decorin substrate 143

2.4.5.6 PmHS2 GlcA-transferase activity on GlcNAc-ethyl-S-decorin at lower UDP-GlcA concentrations (0.5 mM and 0.05 mM) 145

2.4.5.7 Enzymatic extension of GlcNAc-ethyl-S-decorin with 1.7 mM UDP-GlcA, 1.7 mM UDP-GlcNAc and additional 0.17 mM UDP-GlcA..... 146

2.4.5.8 Enzymatic extension of GlcNAc-ethyl-S-decorin with 1.7 mM UDP-GlcA, 0.17 mM UDP-GlcNAc, 0.17 mM UDP-GlcA and additional 0.17 mM UDP-GlcNAc 146

2.4.5.9 Negative control reactions for the pmHS2 enzymatic extension of GlcNAc-ethyl-S-decorin 147

2.4.6 Quantification of the relative protein levels of Peg3, Bec1, LC3 and ICAM-1, VCAM-1, IL-1α 149

2.4.7 Chemical synthesis of GlcNAc-SH and GlcNAc-ethyl-SH 150

2.5 References 157

2.1 Introduction

2.1.1 Natural heterogeneity impedes the structure-activity relationship studies of proteoglycans

Proteoglycans (PG) are a diverse class of proteins essential to all kingdoms of life, presenting one or more complex glycosaminoglycan (GAG) chains.[1] GAGs are linear polysaccharides of variable length, composed of repeating disaccharide building blocks constructed from a hexosamine and an uronic acid residue (see Chapter **1**, figure **1.1**).[2] In particular, chondroitin sulfate (CS), dermatan sulfate (DS) and heparin/heparan sulfate (HS) are GAG chains linked to serine residues on the protein core through an initial xylose.[3] Notably, all GAGs with the exception of hyaluronic acid (HA) have increased complexity through random sulfation and/or epimerization at the local monosaccharide level (Figure **2.1**).[2]

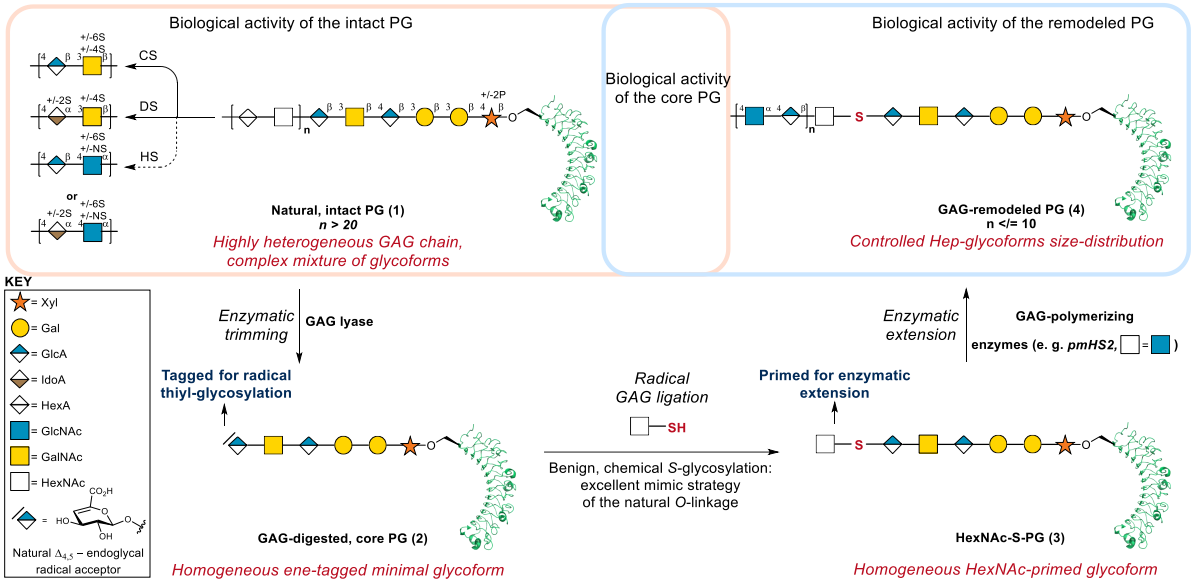


Figure 2.1. A hybrid chemo-enzymatic strategy was designed to generate proteoglycan (PG)-mimics with altered biological activity. A heterogeneous mixture of a natural PG (1) is enzymatically homogenized by GAG lyase treatment to a conserved hexasaccharide-containing protein core (2) which presents a reactive 'tag', the terminal $\Delta_{4,5}$-unsaturation. A biocompatible, radical ligation is then used to install a thiosugar-primer, generating a system amenable to PG-remodeling. Replacement of the *O* atom in the natural glycosidic linkage with another chalcogen such as the axially delivered *S* is a powerful strategy that allows access to the *N*-acetyl-hexosamine (HexNAc)-S-PG (3). The latter is a suitable substrate for GAG-polymerizing enzymes such as the heparosan synthase from type D *Pasteurella multocida* (pmHS2). In this way, PG-analogues that carry a heparosan (Hep) chain (4) can be efficiently synthesized and have the potential to exhibit a biological activity, imparted by the installed non-native GAG-chain (highlighted by the blue square), different from the one observed when testing the original PG (highlighted by the pink square). Any biological effect caused by both 1 and 4 may be attributed to the core protein of the studied PG (intersection of the two squares).

This diversity in composition presents an enormous range of structures and functions, making PGs extremely challenging to study. Templates do not exist to control the length, sulfation pattern, and stereochemistry of their GAG chains, unlike their protein components whose primary structures are defined by discrete genes. To date, no general methodology has been developed to allow the biological and structural studies of defined, soluble PG glycoforms, although several have been developed for the structure-based study of GAG chains.[3-6]

PGs signal both *via* protein-GAG interactions as well as *via* protein-protein interactions.[1] The latter are multifaceted and can be enhanced or diminished by the nature of the GAG chain. For instance, the small PG decorin interacts differently with the Wingless/integrated (Wnt) signalling pathway based on the composition of its single GAG chain.[7] The Wnt-1-induced secreted protein 1 (WISP-1) preferentially binds DS chains on decorin, but it can also bind CS chains with lower affinity, and the WISP-1 association with decorin is abolished upon GAG removal. Conversely, the Leu155-Val260 fragment of the decorin core protein binds transforming growth factor β (TGF-β) isoforms.[8] This interaction is enhanced upon GAG removal, suggesting that this decorin-cytokine binding is partially occluded by the GAG chain.[9]

The PG-controlled signalling pathways are extremely difficult to study due to the natural heterogeneity of these biomolecules in the tissue. In fact, it is currently nearly impossible to disentangle this complex network of biological interactions and precisely attribute observed binding properties to the core protein, the GAG chain, or the sugar modifications such as sulfation and epimerization.[2] Therefore, the ability to reprogram proteoglycans with defined GAG chains, either a glycoform of the natural CS/DS chain or a non-native GAG chain, such as heparosan (Hep) or hyaluronan (HA), would allow better control in the study of proteoglycan biology.

Existing strategies for the precise chemical synthesis of a GAG-containing glycoprotein involve the attachment of a synthetic or a semi-synthetic GAG chain to a model protein, to obtain a so called "neoproteoglycan".[10] In the context of this thesis, this term is used to describe any protein scaffold that is chemically or enzymatically engineered to bear a GAG chain but does not naturally present one, and it does not refer to the other types of PG-mimics such as the nano-GAG composites or the polymer–GAG complexes.[10]

In the first reported case of neoproteoglycan synthesis, Yamaguchi *et al.*[11] used the copper(I) catalyzed 1,3-dipolar cycloaddition (CuAAC) to link an alkyne-containing CS chain to the non-selectively azido functionalized bovine serum albumin (BSA). In a more recent example, Kohling and co-workers[12] employed radical thiol-ene chemistry to attach a 1-thio-hexa-hyaluronane to the *N*-pentenoyl lysine residue of a lipase from *Thermoanaerobacter thermohydrosulfuricus* (TTL). The unnatural amino acid reactive handle was efficiently incorporated into the structure of the protein using the amber codon

suppression technology but the harsh reaction conditions that were employed to install the GAG chain caused significant protein cleavage. Perhaps the most striking example so far, showcasing the GAG-dependent biological effects of extracellular PGs, was demonstrated by Hsieh-Wilson *et al.*[13] The group developed a new methodology by which the appendage of differently sulfated, chloro-functionalized HS chains to a HaloTag protein,[14] engineered for persistent transmembrane localization, was instrumental in understanding the role of PGs in driving embryonic stem cell (ESC) differentiation. In specific, the ESCs decorated with highly sulfated HS chains exhibited an accelerated exit from self-renewal and commitment to neural lineages, whereas cells equipped with HaloTag-bound undersulfated HS adopted the opposite phenotypic characteristics, such as delayed responses to differentiation cues and restricted differentiation potential.

All things considered, tremendous progress has been made in developing novel chemical strategies for the site-selective conjugation of certain GAG chains to non-native protein cores. Despite this, the resulted neoproteoglycans contain unnatural, glycan-protein disconnections that show limited structural resemblance to the peculiar *O*-linked tetrasaccharide (GlcA)-β(1,3)-Gal-β(1,3)-Gal-β(1,4)-Xyl-β1, the core structure found in all CS-/DS-/HS-containing PGs (see Chapter **1**, figure **1.1**). In addition, the effects that were observed upon testing such artificial PG surrogates might not have been a true reflection of the biological events governed by the natural, emulated PGs since the contribution that their native protein components have on activating the analysed signalling pathways was not evaluated.

Therefore, our goal was to access defined glycoforms of natural PGs, *via* a more general and efficient approach. We sought to develop a methodology that not only would be applicable to any soluble or membrane-bound PG but that it would also allow the biological investigation of the PG-governed interactions in a GAG-defined manner, without discarding the significant input imparted by the natural protein scaffold. Moreover, we aimed to obtain PG-mimics in which the conserved tetrasaccharide core structure of the natural PG would not be altered, and any inherent, structural changes caused by the ligation chemistry used to generate the PG-glycoconjugates would be minimum, so to ensure optimal biomimicry.

2.1.2. Thiyl-glycosylation as a strategy for the direct construction of semi-synthetic, minimal PGs

Traditional two-electron chemistry involves the use of a nucleophile and an electrophile which react to form a covalent bond, and it often requires reagents and reaction conditions that are not biocompatible and cannot be used in an aqueous environment.[15] On the contrary, free radicals can be

mildly generated in water and were successfully used in single-electron chemistry to chemoselectively glycosylate proteins.[16-18] The Davis group demonstrated that addition of unprotected *N*- and *O*-linked glycosidic alkyl radicals to the C_β of a dehydroalanine (Dha)-containing protein allowed access to *O*-glycosylated and *N*-glycosylated variants which were further extended or cleaved by relevant glycan-processing enzymes.[16] In a previous study, the group had also developed the free-radical addition of unprotected 1-thiyl glycosyl radicals to homoallylglycine (Hag) containing proteins and by this means, was able to obtain glycoconjugates of virus-like particles that displayed up to 180 glycans per molecular assembly.[18] Clear depictions of these two chemical strategies can be found in Chapter **1**, figures **1.12**, a and **1.15**, a.

Free radicals can be tolerant of aqueous conditions and are compatible with a plethora of functionalities present in PGs such as the unprotected sugars and the amino acid side chains.[19] Therefore, we reasoned that homolytic, single-electron chemistry can be utilized to generate 1-glycosyl thiyl radicals, which are sufficiently reactive, long-lived, and able to selectively target a natural, unsaturated glycan functionality of PGs generated during GAG removal (Figure **2.1** and **2.2**).[20]

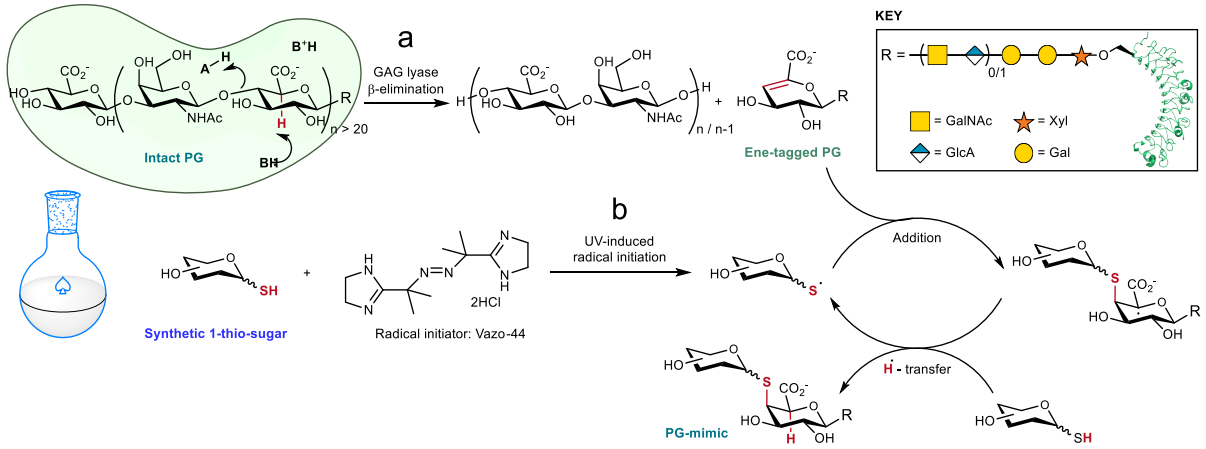


Figure 2.2. The designed chemo-enzymatic strategy allows for *S*-linked minimal PGs to be obtained. This combines biocompatible and selective methods that are enzymatic (**a**) for glycoside cleavage and chemical (**b**) for glycoside ligation (Vazo-44, 2,2'-azobis[2-(imidazolin-2-yl)propane] dihydrochloride). Their selectivity and aqueous compatibility uniquely allow direct application without the need for protecting groups.

In this strategy, a general method can be applied to obtain unsaturated, homogeneous, minimal PG-glycoforms and this entails the use of GAG lyases, enzymes able to depolymerize the glycan chain through a β-elimination mechanism (Figure **2.2**, a).[21,22] Accordingly, charge neutralization of the C_5 carboxyl group of the uronic residue followed by the abstraction of the relatively acidic H_5 proton causes elimination of the adjacent hexosamine and generates a double bond between the C_4 and the C_5 atom of the glucuronic acid (GlcA) ring.

Then, the regioselective anti-Markovnikov radical addition of a 1-thiyl-glycosyl at the C_4 of the enzymatically generated $\Delta_{4,5}$-endoglycal specie (ΔGlcA) gives the corresponding capto-dative stabilized

radical which can then be quenched by the hydrogen abstraction from a second thio-saccharide molecule (Figure **2.2**, b).[20] Previous studies in the Davis group performed on a model unsaturated monosaccharide (data not published) demonstrated that the kinetically-controlled axial attack of the thiyl radical onto the endocyclic C=C bond ($\Delta_{4,5}$) of the GlcA in its half-chair conformation is followed by the axial hydride delivery which is controlled by a pseudoanomeric effect.[23] While not natural, the resulting 1-4 glycosidic *S*-linkage adopts a unique conformation, and may provide good functional mimicry of the natural *O*-linkage.[24] Consequently, we envisioned that after the thiyl-glycosylation, designated GAG synthases will tolerate a substrate such as the primed HexNAc-S-PG and will further polymerize its glycosidic chain (Figure **2.1**).

2.1.3 Bacterial glycosyltransferases as protagonists of the PG-remodeling process with non-native GAG chains

Several enzymatic approaches have been developed over the years, for the *in vitro* production of GAG chains.[25-27] Moreover, rather than using the complex network of eukaryotic GAG synthases, various bacterial enzymes that synthesize GAGs were isolated, and their structures were optimized for increased activity and/or stability.[3,4,28,29]

The heparosan (Hep) synthases pmHS1 and pmHS2 from type D *Pasteurella multocida,* each with two active sites, efficiently synthesize a GAG chain containing the repeating disaccharide unit [GlcA-β(1,4)-GlcNAc-α(1,4)], starting from mono- or oligo-saccharides acceptors (Figure **2.3**, a).[30-32] The *Escherichia coli* strain K5 instead produces a highly-active heparosan synthetic complex, KfiA-KfiD.[33] While subunits KfiB and KfiD are only responsible of stabilizing this enzymatic complex, the subunits KfiA and KfiC are glycosyltransferases that link GlcNAc and GlcA respectively, to create a Hep chain.

Similarly to pmHS2, the hyaluronan (HA) synthase pmHAS isolated from type A *Pasteurella multocida*, contains two active sites that are able to process [GlcA-β(1,3)-GlcNAc-β(1,4)] as HA, a GAG chain that only varies from Hep in the regio- and stereochemistry of the *O*-linkages, respectively (Figure **2.3**, b).[34-37]

In analogy to the pmHAS polymerizing enzyme, the chondroitin (C) synthase identified in the pathogen type F *Pasteurella multocida* (pmCS) forms a polysaccharide containing alternating HexNAc and hexuronic acid (HexA) connected through identical glycosidic linkages as in HA, but in which the HexNAc is the *N*-acetyl galactosamine (GalNAc) residue (Figure **2.3**, c).[38]

Finally, the testosteronan (T) is an unusual GAG chain composed of the [GlcA-α(1,4)-GlcNAc-α(1,4)] repeating unit which is synthesized by a polysaccharide synthase found in the Gram-negative

aerobic bacteria *Comamonas testosteroni* (ctTS).[39] Interestingly, while this enzyme contains the same α-GlcNAc-transferase activity as pmHS2, its GlcA-transfer active site catalyses the extension of the GAG chain with a mechanism that proceeds with retention of the anomeric configuration (Figure **2.3**, d).

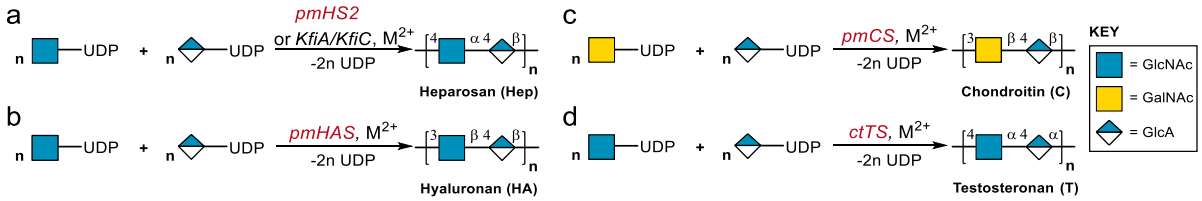


Figure 2.3. Various types of GAGs synthesized by bacterial GAG-polymerizing enzymes. These GAGs all possess the same monosaccharide units, but in different glycosidic linkages (**a**, **b** and **d**) except for the chondroitin chain (**c**) that has the same structure as HA, but with a GalNAc unit substituting for GlcNAc (pmHS2, type D *Pasteurella multocida* heparosan (Hep) synthase 2; KfiA/KfiC, *Escherichia coli* strain K5 GlcNAc-/GlcA-transferase; pmHAS, type A *Pasteurella multocida* hyaluronan (HA) synthase; pmCS, type F *Pasteurella multocida* chondroitin (C) synthase; ctTS, *Comamonas testosteroni* testosteronan (T) synthase; M^{2+}, metal cofactor; UDP, uridine diphosphate).

In this study, we aimed to obtain PG-mimics using a dual-action remodeling enzyme which together with the investigated substrate would generate a simple, two-component protein reaction mixture. Thus, for the controlled enzymatic synthesis of a Hep chain starting from a single GlcNAc residue chemically installed on a model PG (Figure **2.1**), we chose to investigate pmHS2 over pmHS1 since the former had shown to be more plastic.[40] Consequently, we envisioned that this GAG-polymerizing enzyme would be tolerant of the peculiar glycosidic *S*-linkage we chemically introduce to obtain the homogeneously primed GlcNAc-intermediate. Nevertheless, employing a system in which the two glycosyltransferase activities would be residing in separate subunits (KfiA and KfiC), and thus leading to a more complicated three-component protein reaction mixture, would also be worth investigating. The latter would facilitate explorations of different substrate:KfiA:KfiC stoichiometries should their potentially different enzymatic kinetic properties require this.

2.2 Results and discussion

2.2.1 Extraction, purification and characterization of a natural PG: decorin

Decorin is a soluble, small leucine-rich proteoglycan (SLRP) (~ 36.5 kDa) which bears a single GAG chain attached to the Ser34 residue, near the *N*-terminus.[41] This can be a CS or a DS chain, depending on the activity of the dermatan sulfate epimerases (DS-epi1 and DS-epi2) in the connective tissue it originates from.[42] Decorin adopts a curved solenoid fold composed of 12 leucine-rich repeats (LRRs) and its native conformation is complemented by six cysteine residues which form three intramolecular disulfide bridges: Cys25-Cys31, Cys29-Cys38 and Cys284-Cys317 (Figure **2.4**, a).[43]

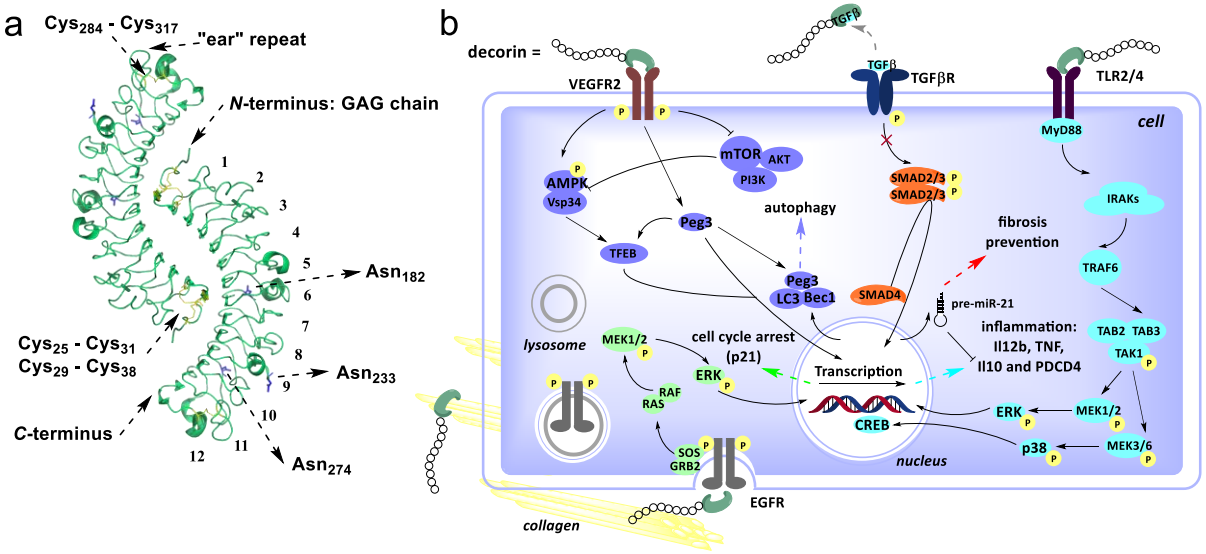


Figure 2.4. The structural features found in decorin and the diverse, cell-dependent biological effects evoked by this PG. Crystal structure of the GAG-trimmed decorin (image of 1XCD[43] created in pymol) reveals the dimer association of this PG. Each monomer consists of 12 leucine-rich repeats (LRRs), contains six Cys locked in disulfide bridges (yellow) and presents three Asn-glycosylation sites (blue) (**a**). Decorin binds TGFβ and blocks the downstream signalling cascade (orange) leading to the transcription of the profibrotic microRNA pre-miR-21. The decorin/TLR2/4 interaction activates the ERK and p38 pathways (blue) that culminate with an increased expression of the pro-inflammatory cytokines: Il12b, TNF, Il10 and PDCD4. By targeting the EGFR, decorin triggers activation of the ERK-controlled transcription of p21 (green) promoting cell cycle arrest, and EGFR internalization leading to lysosomal degradation. The decorin/VEGFR2/Peg3 signalling cascade leads to increased expression of the pro-autophagic markers Bec1 and LC3. Concomitantly, binding of decorin to VEGFR2 causes inhibition of the anti-autophagic mTOR-node (violet) (**b**) (TGFβ, transforming growth factor β; TGFβR, TGFβ receptor; SMAD 2/3/4, mothers against decapentaplegic homolog 2/3/4; TLR2/4, Toll-like receptor; MyD88, myeloid differentiation primary response 88; IRAKs, interleukin-1 receptor-associated kinases; TRAF6, TNF receptor-associated factor 6; TAB2/3, TAK1 binding proteins 2/3; TAK1, Transforming growth factor beta-activated kinase 1; MEK 1/2/3/6, mitogen-activated protein kinase kinase 1/2/3/6; ERK, extracellular signal-regulated kinase; p38, a class of mitogen-activated protein kinases; CREB, cAMP response element-binding protein; EGFR, epidermal growth factor receptor; SOS, son of Sevenless protein; GRB2, growth factor receptor-bound protein 2; RAS, rat sarcoma protein; RAF, rapidly accelerated fibrosarcoma; p21, a cyclin-dependent kinase inhibitor; VEGFR2, vascular endothelial growth factor receptor 2; AMPK, 5' AMP-activated protein kinase; Vsp34, vacuolar protein sorting 34; Peg3, paternally-expressed gene 3; mTOR, mammalian target of rapamycin; PI3K, phosphoinositide 3-kinase; AKT, protein kinase B; TFEB, transcription factor EB; Bec1, beclin 1; LC3, microtubule-associated proteins 1A/1B light chain 3B).

The latter forms a larger loop at the *C*-terminus of the protein core, the so called "ear" repeat. Moreover, decorin presents three *N*-glycosylation sites at Asn182, Asn233 and Asn274, and it is believed that as for other *N*-glycoproteins, they have a crucial role in the "quality control" check during the folding process. Indeed, data published elsewhere suggests that both the GAG chain and the *N*-linked oligosaccharides contribute to the conformation-dependent mechanism of cellular transit and secretion of this PG.[44]

Decorin is present in high concentrations in connective tissues and extracellular matrix, where it acts both as a structural component as well as a signalling mediator.[45] Thus, several studies have indicated that decorin binds with high affinity to collagen fibers thereby regulating their assembly and growth.[46-49] In addition, there is increasing evidence that this PG also regulates a plethora of other biological processes including tumorigenesis, inflammation, autophagy and angiogenesis. These roles are achieved not only by sequestering multiple growth factors, but also by acting as an endogenous ligand for the innate immunity Toll-like receptor (TLR) 2/4 and as a pan-receptor tyrosine kinase (pan-RTK) interaction partner.[41,45,50-52]

For instance, it was shown that decorin binds the transforming growth factor β (TGF-β) and by trapping this autocrine factor within the matrix, it indirectly inhibits downstream signalling pathways governed by the interaction with its designated membrane receptor (Figure **2.4**, b, orange).[53] Consequently, neutralization of the TGF-β activity represses the development of various TGF-β-dependent fibrotic conditions that may occur in response to inflammation or tissue injury.[54-56]

Research has also shown that decorin modulates inflammatory responses by engaging the TLRs (Figure **2.4**, b, blue).[57] Thus, upon interacting with TLR2 and TLR4, the PG induces expression of interleukin-12b (Il12b), tumour necrosis factor alpha (TNFα), interleukin-10 (Il10), and programmed cell death 4 (PDCD4) protein in macrophages. At the same time, *via* decorin-driven inhibition of the TGFβ signalling the translational repression of PDCD4 is also prevented, causing diminished release of the anti-inflammatory cytokine Il10.

Among the tyrosine kinase receptors that were identified as interaction partners of decorin, the endothelial growth factor receptor (EGFR) was the first one to display a decorin-binding domain (Figure **2.4**, b, green).[58] The decorin/EGFR interaction triggers a signalling cascade that causes, among the numerous events leading to cellular growth suppression, phosphorylation of the extracellular signal-regulated kinases 1/2 (ERK1/2) and induced expression of the potent cyclin-dependent kinase inhibitor p21.[59] Moreover, the persistent engagement of EGFR by decorin evokes receptor down-regulation through sustained endocytosis and intracellular degradation.[60,61] These effects cause cell cycle arrest in various EGFR-expressing tumours such as squamous, breast and prostate carcinomas.[59,62-64]

The decorin-evoked RTK-antagonism was also illustrated by studies in which this PG was able to bind to the hepatocyte growth factor receptor (HGFR), thereby releasing the glycogen synthase kinase-3β (GSK-3β) from being phosphorylated and inactivated (example not illustrated).[65,66] The latter promotes, *via* phosphorylation, the proteosomal degradation of two major oncogenic proteins, namely β-catenin and Myc.[67] By suppressing the β-catenin and Myc levels, decorin inhibits the HGFR-controlled networks conducive to tumour growth, tissue invasion, angiogenesis and metastasis.[50]

More recently published data has shown that decorin binds with high affinity to another RTK that is predominantly expressed in endothelial cells, the vascular endothelial growth factor receptor 2 (VEGFR2) (Figure **2.4**, b, violet).[68] Stimulation of VEGFR2 triggers a signalling cascade that leads to activation of the AMP-activated protein kinase (AMPK), inhibition of the mechanistic target of rapamycin (mTOR) protein and increased synthesis of the paternally-expressed gene 3 (Peg3) protein.[69] The latter is a tumour repressor which recruits beclin 1 (Bec1) and microtubule-associated proteins 1A/1B light chain 3B (LC3), two major activators of the autophagic process.[70] Autophagy is a homeostatic mechanism that cells initiate to catabolize damaged organelles, misfolded or aggregated proteins, and toxic metabolic products thereby preventing tumour initiation.

Moreover, the decorin-induced Peg3 is also capable of transcriptionally enhancing the synthesis of the autophagy markers Bec1 and LC3.[68] In parallel to the activation of the Peg3-controlled pathway, the decorin/VEGFR2 interaction causes sustained potentiation of the intracellular cascade controlled by the pro-autophagic AMPK.[70,71] In addition, the PG/RTK binding blocks the anti-autophagic signalling in the endothelial cell *via* suppression of the Akt/mTOR/PI3K pathway, and ultimately, all these three concurrent decorin-controlled events contribute to preventing tumorigenesis.[69,72]

Therefore, due to the structure, the feasible isolation from connective tissue and the promising biological applications, decorin was chosen as our model PG system on which a biocompatible thiyl-glycosylation was developed. This allowed the glycoprotein remodeling of decorin with a non-native GAG chain. In future biological investigations, the obtained PG-analogue should give better control in the study of decorin-mediated signalling pathways.

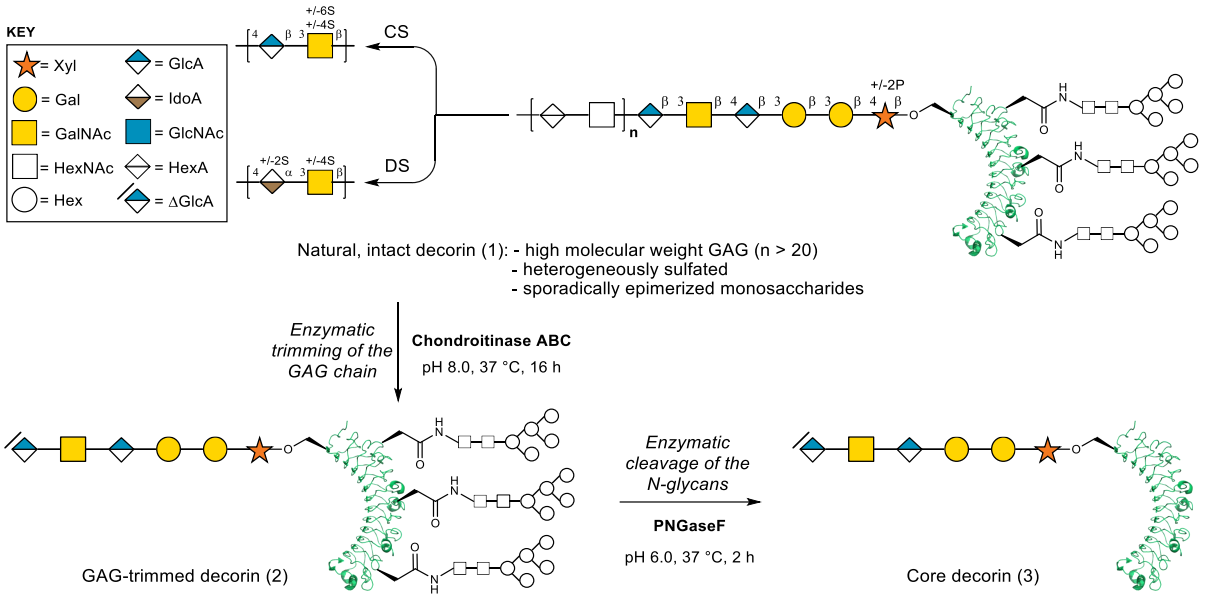


Figure 2.5. Generation of the three natural decorin variants. The extracted decorin (1) is homogenized by treatment with GAG lyases to a conserved *O*-linked hexasaccharide PG (2), and the *N*-glycans are removed upon digestion with PNGaseF to afford the core decorin (3) variant.

In this work, the natural, intact decorin-CS/DS (Figure **2.5**, 1) was extracted from bovine tendons and purified by anion-exchange chromatography, following previous reports (Figure SI 2.1).[73] This was further purified by gel filtration for structural and binding studies, or treated with commercial chondroitin ABC lyase to digest the *O*-glycan chain and generate the GAG-trimmed protein bearing a residual hexasaccharide-alkene (Figure **2.5**, 2 and figure SI 2.2). The mixture was readily purified by concavalin A (ConA) affinity chromatography followed by gel filtration, and SDS-PAGE analysis showed the expected molecular weight (MW) shift from approximately 75 kDa to around 45 kDa, confirming that trimming of the CS/DS chain had been successfully achieved (Figure **2.6**, a). Then, the three *N*-glycans were cleaved by treatment with the peptide *N*-glycosidase F (PNGaseF), and formation of the core decorin protein (Figure **2.5**, 3) was confirmed by the increased electrophoretic mobility of a 37 kDa band on the SDS-PAGE gel (Figure **2.6**, a and figure SI 2.3).

The core decorin (1) was further characterized by denaturing, intact protein LC-MS analysis, and the deconvoluted spectrum revealed that the obtained protein had a molecular mass of 37410 +/- 2 Da (Figure SI 2.3). In addition to this, we also identified a few low-abundant decorin-related species, and a comprehensive rationalization regarding their potential origin can be found in Section 2.4.2.1 (Figure SI 2.4).

Next, we sought to determine the sequence of the residual *O*-glycan attached to the protein core that had been generated upon treatment of the intact PG with the chondroitin ABC lyase. To achieve this, the core decorin was digested with trypsin and the resulted (glyco)peptides were analysed

by LC-MS/MS analysis (Figure SI 2.5). It was initially assumed that the chondroitinase would cleave at the GlcA residue situated at the non-reducing terminus of the conserved tetrasaccharide core, the common structure shared by all CS, DS and HS chains.[74] Instead, we found that the enzyme does not lyse the CS/DS chain until it reaches the innermost acidic saccharide but the digestion stalls at the preceding GlcA residue, generating the unsaturated Ser34-linked hexasaccharide: ΔGlcA-GalNAc-GlcA-Gal-Gal-Xyl (Figure SI 2.5). We were pleased to see that our glycoproteomics results were in agreement with the findings published by Zaia *et al*.[75]

Moreover, depending on the activity of sulfotransferases and kinases in the connective tissue it originates from, the residual glycan chain can present additional modifications such as sulfation of the Gal and GalNAc residues[76] or phosphorylation of the internal Xyl residue[77]. Indeed, MS/MS fragmentation of the decorin *O*-glycopeptide revealed that the unsaturated glycan was carrying either a phosphate group on the Xyl or a sulfate on the innermost Gal residue (Figure SI 2.5).

The glycoproteomics results were successfully corroborated with the rationalized intact protein LC-MS data (Figure SI 2.4) and the fully characterized, minimal ene-tagged decorin was further used for the development of the free-radical GAG ligation (Figure **2.1**).

While crystal structure of the GAG-trimmed decorin (Figure **2.5**, 2) indicated that an array of hydrophobic interactions, complemented by an extended network of hydrogen bonds and salt bridges favour the dimeric association of decorin (Figure **2.4**, a),[43] there have been a couple of contradictory studies regarding the oligomeric state of this glycoprotein in solution.

The first study claimed that both the extracted decorin (Figure **2.5**, 1) and its GAG-trimmed variant (Figure **2.5**, 2) exist as mainly dimers in solution, as determined by light scattering chromatography and solution X-ray scattering.[78]

On the contrary, a second study concluded that both species are monomers in solution since several tested crosslinker agents such as the sulfo-SMPB (sulfosuccinimidyl 4-(*N*-maleimidophenyl)butyrate), the BS3 (bis(sulfosuccinimidyl)suberate) and the glutaraldehyde failed to capture any potential dimers.[79] As bis-electrophilic compounds, these probes should have been able to trap nucleophilic amino acid chains from two different decorin molecules if associated in a dimeric complex.

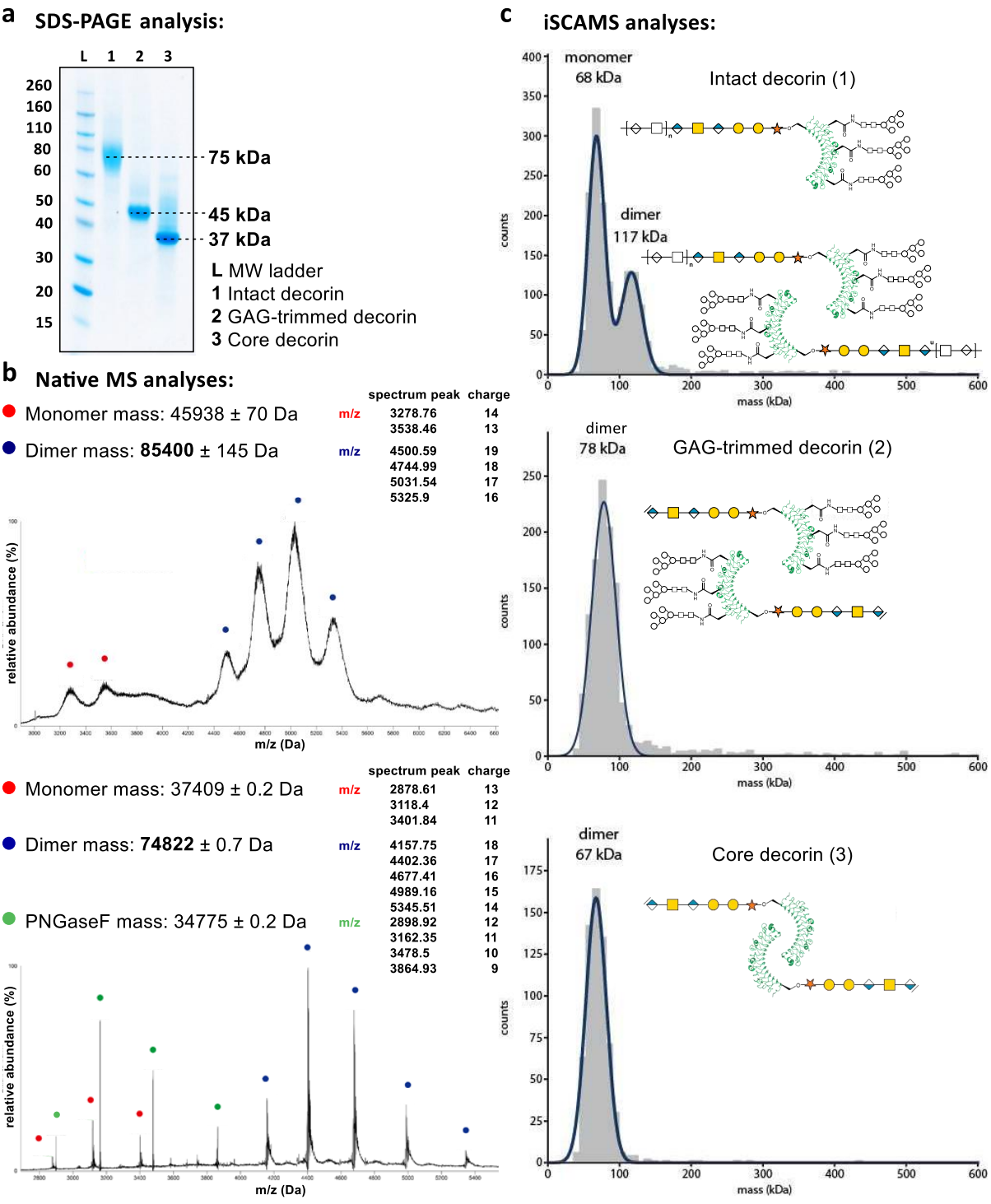


Figure 2.6. The obtained decorin variants were analysed by SDS-PAGE, native MS and iSCAMS analyses. SDS-PAGE showed that the intact (1), the GAG-trimmed (2) and the core (3) PG had a molecular mass of approximately 75, 45 and 37 kDa, respectively (**a**). The native mass spectrometry (MS) results revealed that both the GAG-trimmed (2) and the core decorin (3) were present as mainly dimers in solution (blue circles). The latter still contained PNGaseF, an impurity that originated from the incomplete removal of the enzyme upon cleavage of the *N*-glycans (green circles) (**b**). The interferometric scattering mass spectrometry (iSCAMS) data obtained from the analysis of the GAG-trimmed (2) and the core decorin (3) confirmed the native MS results, but an underestimation of the molecular masses could also be noticed. Conversely, iSCAMS analysis recorded for a solution of intact decorin (1) revealed that the PG is predominantly a monomer. Moreover, the mass error found for the lower dimer population was even more pronounced suggesting that the accuracy of the measurements might have been significantly impacted by the GAG chain (**c**).

Given the conflicting published reports, we decided to conduct our own study to determine the oligomeric state of the three decorin variants that we had obtained (Figure **2.5**). To do so, we used a different approach, one that combined analytical data obtained by native mass spectrometry (native MS) and information afforded by interferometric scattering mass spectrometry (iSCAMS) analyses, also known as mass photometry (MP).[80]

ISCAMS has recently emerged as a powerful, solution-based biophysical method that can be used to determine in real time, the mass of biomolecules.[81] This analytical technique relies on the detection of changes in the reflectivity of a glass-water/-buffer interface caused by interference between scattered and reflected light when individual proteins bind to this interface.[81] Then, the scattering contrast of the molecular binding events is extracted, and the molecular mass of the analyte is calculated based on a contrast-to-mass calibration plot.[81] The latter is built from scattering measurements of appropriate protein calibrants with known oligomeric states and corresponding masses.[81] All in all, due to its single-molecule nature, iSCAMS provides, *via* the mass distribution graphs that are being generated, reliable information on the oligomeric state composition of the analysed protein sample. In addition, the MP measurements are typically very fast (minutes), and they require minimal sample volumes (µL) and concentrations (nM).

Results obtained using the iSCAMS imaging technique were in agreement with those obtained by native MS, an already established gas-phase method used to interrogate the oligomeric state of a protein.[82] The GAG-trimmed decorin (2) and the core decorin (3) were predominantly dimeric when the two protein solutions were analysed by native MS (Figure **2.6**, b) or iSCAMS (Figure **2.6**, c) whereas the intact PG (1) was mainly present as a monomer in solution (Figure **2.6**, c), when the scattered blue light (445 nm) was measured.

The two "weighing" methods complemented each other very well because the native MS was not amenable for the analysis of intact decorin (1) due to the heterogeneity of this glycoprotein and its low positive ionization efficiency imparted by the highly sulfated GAG-chain. On the other hand, presence of monomeric GAG-trimmed (2) and core decorin (3) could only be detected by native MS and not by iSCAMS, due to the imperfect background subtraction of the latter, which has a more pronounced effect when analysing proteins smaller than 50 kDa.[81]

Despite this, some differences between the results obtained from the two analytical methods could also be observed. First, it is worth noticing that according to the iSCAMS measurement (Figure **2.6**, c), the molecular mass of the core decorin (3) dimer is 67 kDa instead of 74.8 kDa, which accounts for ~ 10 % mass error, significantly higher than the previously reported 2 % sequence mass accuracy.[81]

Lower molecular masses were also recorded when the GAG-trimmed decorin (2) dimer (78 kDa versus 85.4 kDa, the native MS dimer mass) and the intact PG (3) monomer (68 kDa instead of 75 kDa, the molecular mass revealed by SDS-PAGE analysis) were analysed by iSCAMS.

Surprisingly, the lower population of dimers present in a solution of intact decorin (1) had a recorded mass of 117 kDa, lower than the calculated mass of 136 kDa (2 x 68 kDa, the iSCAMS recorded monomer mass) and much lower than the calculated mass of 150 kDa (2 x 75 kDa, the SDS-PAGE monomer mass). The observed discrepancy could be due to several different aspects: i) only species bearing a shorter GAG chain with a total glycoprotein mass of around 58.5 kDa (117/2 kDa) could associate and form dimers; ii) the molecular mass underestimation for this complex glycoprotein varied with the length of the GAG chain attached to it; iii) both factors might have contributed to the observed results.

All in all, the combined native MS/iSCAMS strategy suggested that the monomer-dimer transition of decorin might be GAG-mediated and consequently there might be a critical minimal length for the GAG chain (around 20 kDa), below which protein dimerization is favoured.

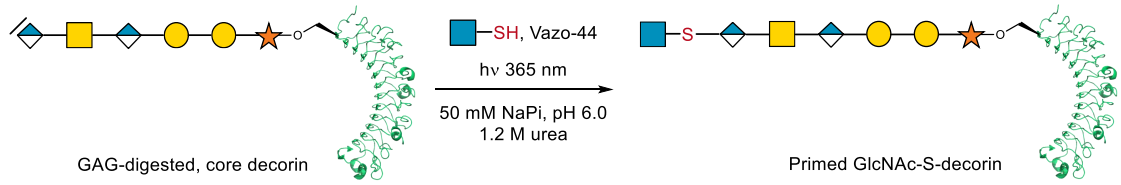
2.2.2 Development of a radical GAG ligation to obtain semi-synthetic PG-intermediates

Having obtained and characterized our desired starting material, the GAG-digested, core decorin (3), we embarked on exploring several reaction conditions for the radical thiyl-glycosylation step, using GlcNAc-SH as a model 1-thio-monosaccharide, Vazo-44 as a water-soluble radical initiator and a 125 W UV Hg lamp with maximum emission at 365 nm (Table **1**). The decision to use the core decorin as our model substrate was based on previous unsuccessful experiments performed in the group that attempted to develop benign conditions for the radical *S*-glycosylation of GAG-trimmed decorin (2) instead. The latter approach implied that, upon irradiation of the decorin variant 2 in presence of GlcNAc-SH and Vazo-44, an aliquot would be taken and treated with PNGaseF to enzymatically remove the *N*-glycans and facilitate analysis by intact protein LC-MS. Unfortunately, this strategy failed to provide good quality LC-MS spectra that unequivocally showed the molecular mass shift (+ 237), corresponding to the GlcNAc-SH addition. Although discouraging, the results suggested that this outcome could be attributed to an inefficient *N*-deglycosylation caused by residual reagents and/or by-products of the radical reaction inhibiting the enzymatic activity of the PNGaseF. We also considered the possibility that the obtained GlcNAc-S-decorin aggregated when further incubated with the enzyme for two hours, at 37 °C, particularly if unreacted reagents had not been efficiently removed.

All in all, we decided to perform the thiyl-glycosylation of decorin after the enzymatic removal of the *N*-glycans. We were pleased to see that for the first time, irradiation of core decorin in the presence of 100 mM radical initiator and 500 mM GlcNAc-SH, using a 125 W UV lamp, allowed us to monitor the formation of GlcNAc-S-decorin (Table **1**, entry **1** and figure **2.7**). Even though 70 % conversion to the desired product was obtained after only 30 minutes irradiation, a longer reaction time proved to cause significant protein degradation (Figure SI 2.10).

Therefore, we decided to replace the UV source with a 365 nm 3 W torch that would allow us to gradually generate the thiyl radicals and thus, cause less damage to the protein structure (Table **1**, entry **2** and figure **2.7**). By using a 365 nm torch that could be more safely handled we also more easily monitored the progression of the reaction by intact protein LC-MS analyses. We achieved 34 % conversion to the desired GlcNAc-S-decorin after 60 minutes irradiation, using the same reagent equivalents (Figure SI 2.11).

In order to improve the conversion while still preserving the integrity of the protein, we decided to double the concentration of reagents and perform the UV radical initiation using the same 3 W torch (Table **1**, entry **3** and figure **2.7**). Indeed, after one hour, the desired GlcNAc-S-decorin was obtained with excellent conversion (89 %) as revealed by intact protein LC-MS analysis (Figure SI 2.12).



Entry	Conc. Vazo-44 (M)	Conc. GlcNAc-SH (M)	Power 365 nm (W)	Time (min)	Reaction scale (µL)	Conversion (%)	
						- DTT	+ DTT
1	0.1	0.5	125	30	25	70	71
2	0.1	0.5	3	60	50	34	39
3	0.2	1	3	60	50	89	95
4	0.2	1	3	60	250	74	87

* Incubation time for reaction described at Entry 1 was shorter due to significant protein degradation (aggregation) observed when decorin was irradiated with a 125 W UV lamp for more than 30 minutes.

Table 1. Optimization of the thiyl-glycosylation of core decorin with GlcNAc-SH.

Then, we performed the reaction on a larger volume of protein solution (250 µL) and we observed that the conversion to GlcNAc-primed decorin was slightly decreased (Table **1**, entry **4** and figure **2.7**) due to the formation of a double GlcNAc-S-addition product (11 %) (Figure SI 2.13).

However, LC-MS analysis showed that, upon treatment of an aliquot of the reaction mixture with a disulfide reducing agent such as dithiothreitol (DTT), this side-product was no longer observed and the conversion towards desired GlcNAc-S-decorin increased to 87 % (Figure SI 2.13).

Thus, treatment with DTT of the samples analyzed by intact protein LC-MS helped us get a quick insight into the selectivity of the radical *S*-glycosylation for the ΔGlcA residue of decorin over any of the six cysteine residues present in the protein (Figure **2.8**). We therefore inferred that the formation of the side-product is most likely caused by the homolytic cleavage of one of the disulfide bridges, and quenching of the resulted Cys• radical by a GlcNAc-S• radical. The formation of this side-product resulted to be more significant when performing the thiyl-glycosylation on a 250 µL scale using the same 300 µL glass vial that was employed in all screening experiments. Nevertheless, we were able to avoid this side-reaction by allowing the release of the developing N_2 from the glass vial through a needle or a vial septum cap (Figure **2.9**).

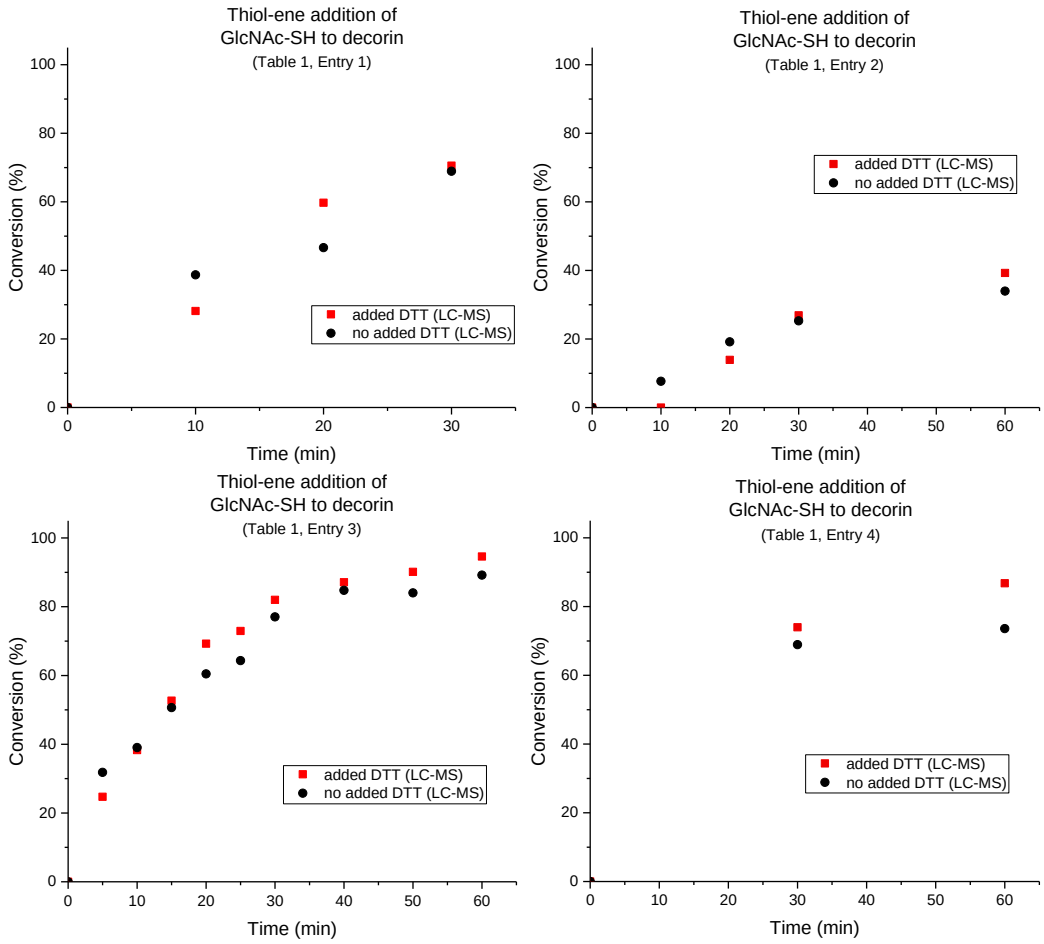


Figure 2.7. **Effect of varying UV source, GlcNAc-SH and Vazo-44 levels, and reaction volume on the *S*-glycosylation of core decorin (graphical data for the experiments in Table 1).** Graphs show conversions to GlcNAc-S-decorin as determined by LC-MS analyses of aliquots treated (red) or not treated (black) with DTT, taken in parallel for each of the single-replicate experiments.

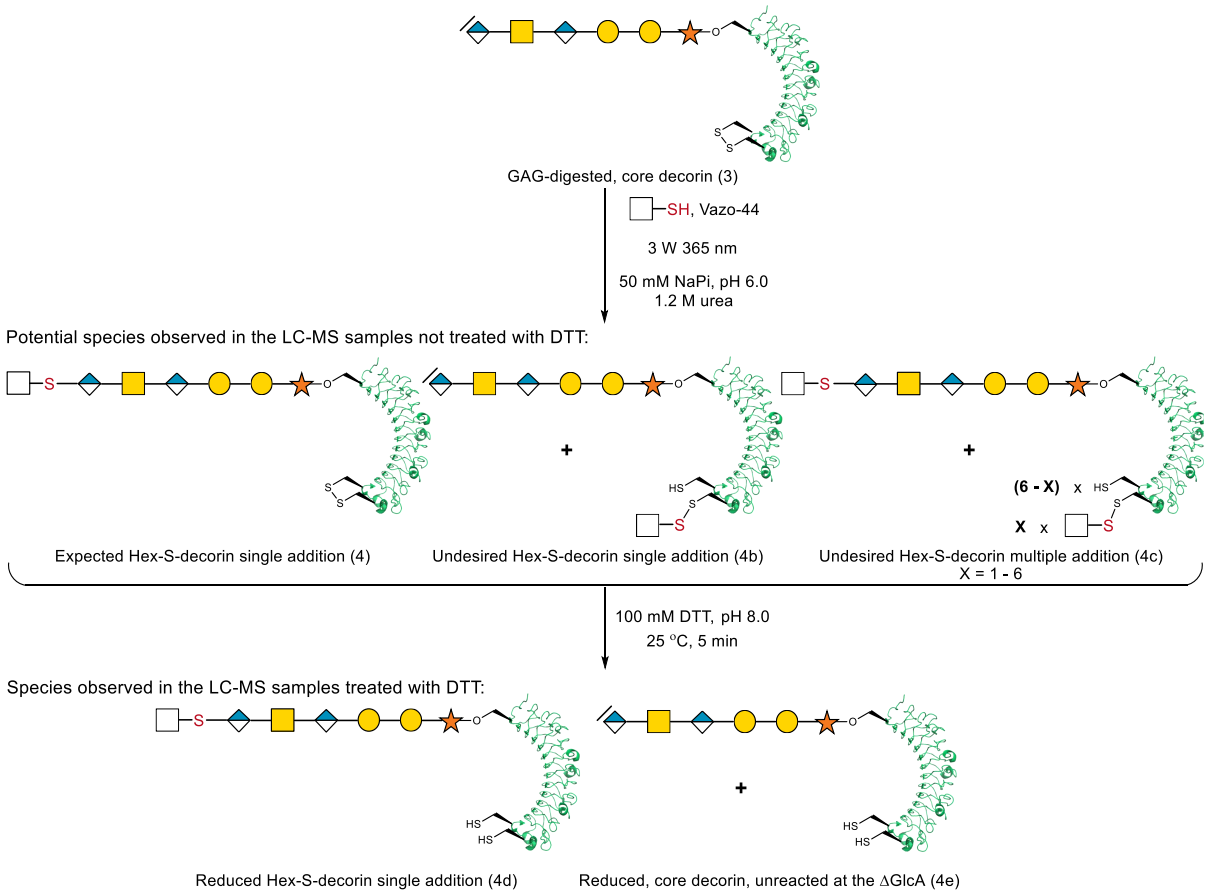


Figure 2.8. Treatment of the LC-MS samples with DTT was instrumental in understanding the selectivity of the thiyl radical addition for the ΔGlcA residue of the PG. Hex-SH can add to the unsaturated glycan of the core decorin (3) to afford the expected Hex-S-decorin (4) but it can, to a much less extent, add to the Cys residues of decorin to give the undesired Hex-S-decorin single (4b) and/or multiple (4c) addition products. However, upon treatment with the disulfide reducing agent DTT the radical addition to the ΔGlcA (4d) can be discerned from the PG species that are not *S*-glycosylated at the unsaturated glycan part (4e; X, number of additions to Cys; X = 1 – 6).

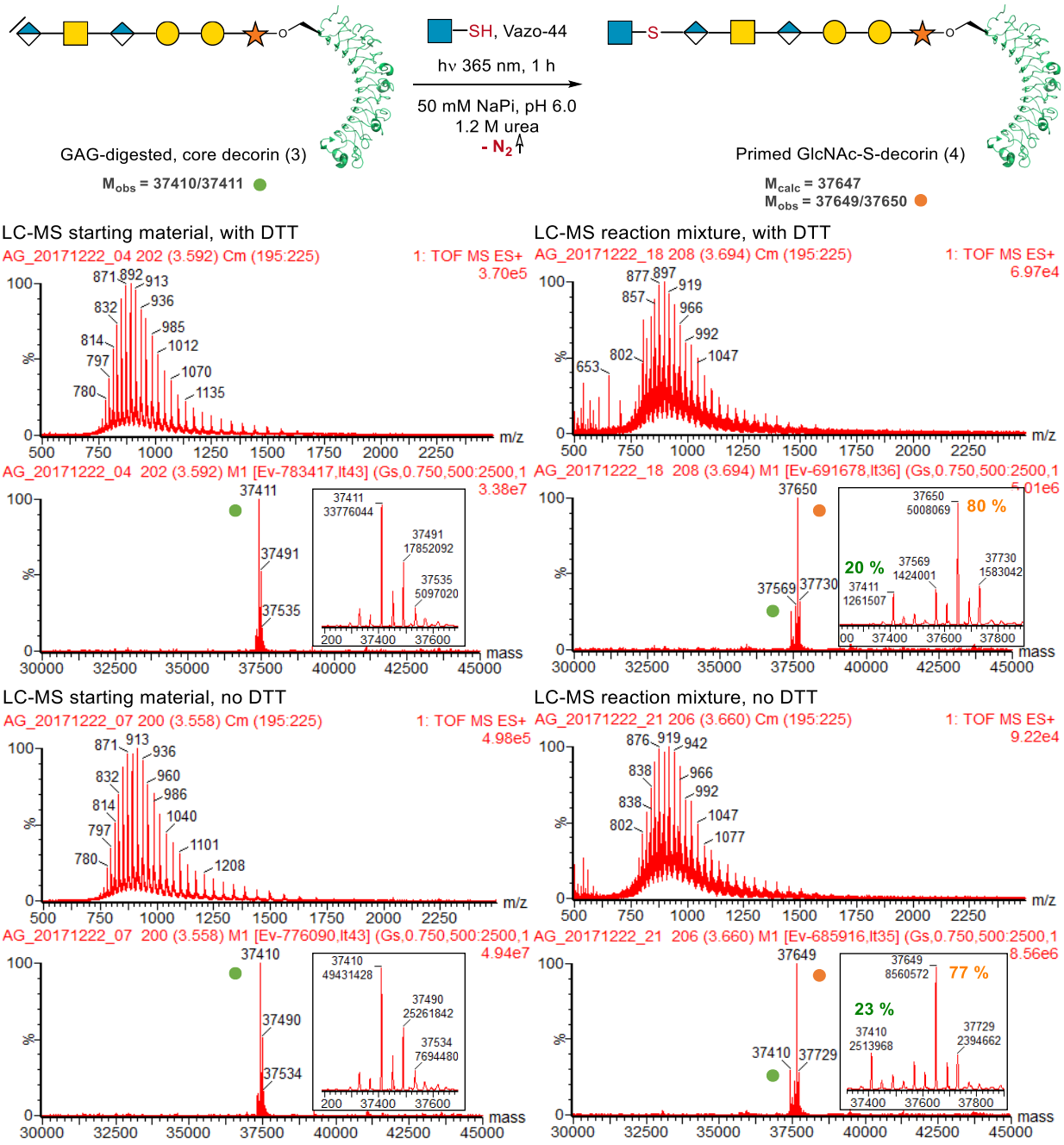


Figure 2.9. Optimized reaction conditions allowed excellent conversion of core decorin to GlcNAc-S-decorin single addition product and no undesired double addition was observed even on a 250 µL scale. Reaction was performed in triplicate and 73 ± 4 % conversions were achieved. The figure shows the LC-MS spectra that were obtained for one of the replicates (77 % conversion).

The photochemical radical initiation mechanism was supported by control reactions, in which the need of both UV light and radical initiator Vazo-44 was investigated. In fact, the reaction mixture in which the radical initiator Vazo-44 was omitted (Table **2**, control **1**, Figure SI 2.14) or that was not UV-irradiated (Table **2**, control **2**, Figure SI 2.14) showed only traces of GlcNAc-S-addition (9 % and 12 % conversion, respectively). The formation of these small amounts of GlcNAc-S-decorin was most likely caused by the nucleophilic substitution of GlcNAc-S^- at one of the disulfide bridges of the protein, and

thus generation of an S-S bond with one of the Cys residues. Although the pKa of thiols is higher (pKa 8-10)[83,84] than the pH of the thiyl-glycosylation buffer (pH 6.0, NaPi) even a low concentration of thiolate ions (~ 10^{-4}-10^{-2} M) was still in high excess with respect to the amount of protein (~ 44 µM) in the reaction mixture. Nevertheless, the observed undesired species carrying a GlcNAc-S-residue were readily abolished in the LC-MS samples treated with DTT.

All together, these control reactions proved that both the UV source and the radical initiator were necessary for the photochemical thiyl-glycosylation of decorin.

Moreover, the radical GlcNAc-S-addition attempted on decorin bearing an intact GAG chain, and thus lacking an unsaturated GlcA residue (Table **2**, control **3**, Figure SI 2.15), revealed no addition of the GlcNAc-SH priming sugar, proving once again the site-selectivity of the developed radical GAG ligation.

Control reaction	Radical initiator	UV light	Radical acceptor (ΔGlcA)	Observed GlcNAc-S-addition (%) (- DTT)	Observed GlcNAc-S-addition (%) (+ DTT)
1	No	Yes	Yes	< 10	-
2	Yes	No	Yes	12	-
3	Yes	Yes	No	-	-

Table 2. Control reactions performed in absence of radical initiator Vazo-44, without UV activation or in absence of a trimmed GAG chain proved that mechanism explained in Figure 2.2 is responsible for the GlcNAc-S-addition to decorin.

In order to further demonstrate the site-selectivity of the radical GlcNAc-S-addition to the ΔGlcA residue on decorin, we digested the GlcNAc-S-primed PG with trypsin and analysed the resulting mixture using the same glycoproteomics approach that we had previously developed for the (glyco)sequencing of core decorin (Figure SI 2.5). Gratifyingly, MS/MS fragmentation of the GlcNAc-S-glycopeptide identified fragments GlcNAc-S-GlcA, GlcNAc-S-GlcAGalNAc and HS-GlcAGalNAcGlcAGalGalXyl-peptide, thus confirming the expected site-selectivity of the developed thiyl-glycosylation (Figure **2.10**).

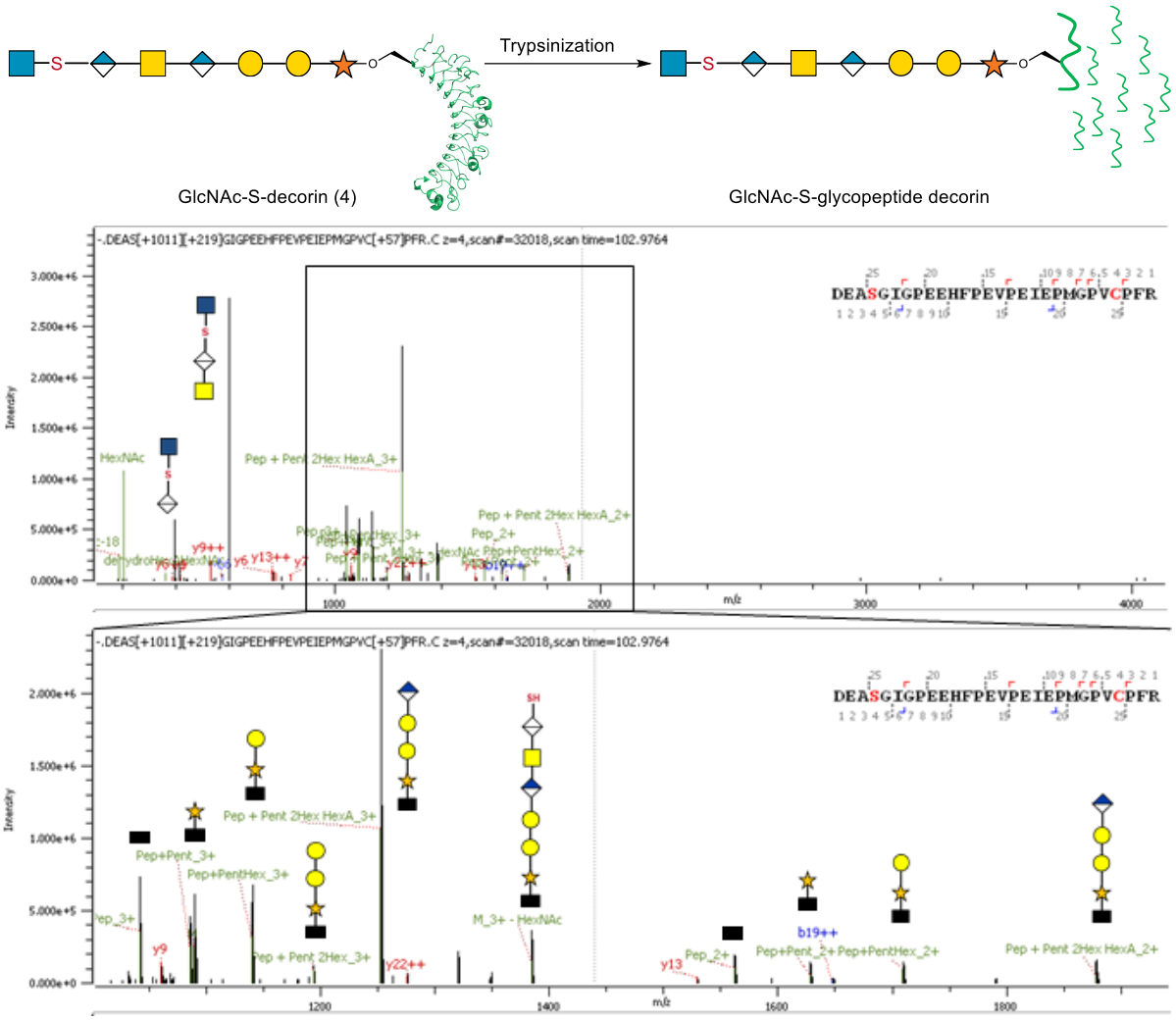
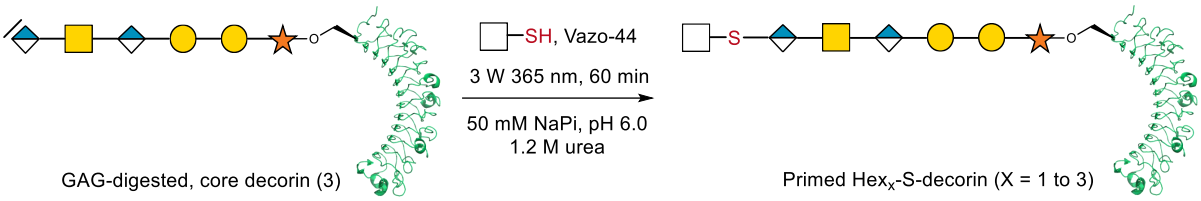


Figure 2.10. LC-MS/MS analysis of GlcNAc-S-decorin upon digestion with trypsin. The expected GlcNAc-S-glycopeptide was identified, and the MS/MS glycan fragmentation confirmed that the radical addition had taken place at the terminal ΔGlcA residue.

Extension of the reaction to other β-saccharides bearing –SH or –ethyl-SH anomeric substituents revealed successful ligation to decorin (Table **3**). Notably, the relatively low addition of GlcA-SH was paralleled by the conjugation to one of the Cys residues of the protein (Table **3**, entry **2**, Figure SI 2.16). An increased level of double addition was also detected after one hour incubation with Glc-SH (Table **3**, entry **3**, Figure SI 2.17) and even triple addition was observed for Glc-ethyl-SH (Table **3**, entry **4**, Figure SI 2.18). In these cases, one and two cysteine residues respectively, had been coupled to thiyl radicals as proven by conversion of the di- and tri-glycosylated species into the corresponding single addition product upon treatment with DTT. Interestingly, saccharides presenting a thiol attached through an ethyl-linker to the anomeric *O*-atom proceeded with higher conversion rates than the corresponding sugars in which the -SH had replaced the anomeric hydroxyl group. Lastly, a much lower conversion (24 %) was achieved after one hour of UV-irradiation of the PG with Gal-Glc-SH (Table **3**,

entry **6**, Figure SI 2.20) thus revealing the limitation of the developed GAG ligation in priming decorin with longer saccharides.



Entry	SH	% of observed species (- DTT treatment)				% of observed species (+ DTT treatment)	
		starting material	single addition	double addition	triple addition	starting material	single addition
1	SH	11	89	-	-	5	95
2	SH	28	60	12	-	59	41
3	SH	-	72	28	-	-	100
4	SH	-	53	33	14	-	100
5	SH	-	92	8	-	-	100
6	β 4 SH	73	27	-	-	76	24

Table 3. **Radical addition of different thio-saccharides to decorin proceeded with useful to high conversions and various degrees of selectivity.**

Because the GlcNAc-ethyl-SH reacted faster with the ΔGlcA-tagged decorin (Table **3**, entry **5**, Figure SI 2.19) than the GlcNAc-SH (Table **3**, entry **1**, Figure SI 2.12), to avoid formation of the double addition side-product, we decided to reduce the UV-irradiation time from 1 hour to 40 minutes. Consequently, 86 % conversion to the desired GlcNAc-ethyl-S-decorin was achieved even when the reaction was performed on a larger volume (150 µL) and no undesired side-product was detected by intact protein LC-MS analyses (Figure **2.11**). The addition of this thio-saccharide afforded a different GlcNAc-primed decorin variant that was further investigated as a promising substrate for the heparosan synthase pmHS2 and it later proved to be suitable for enzymatic remodeling (Figure **2.20**).

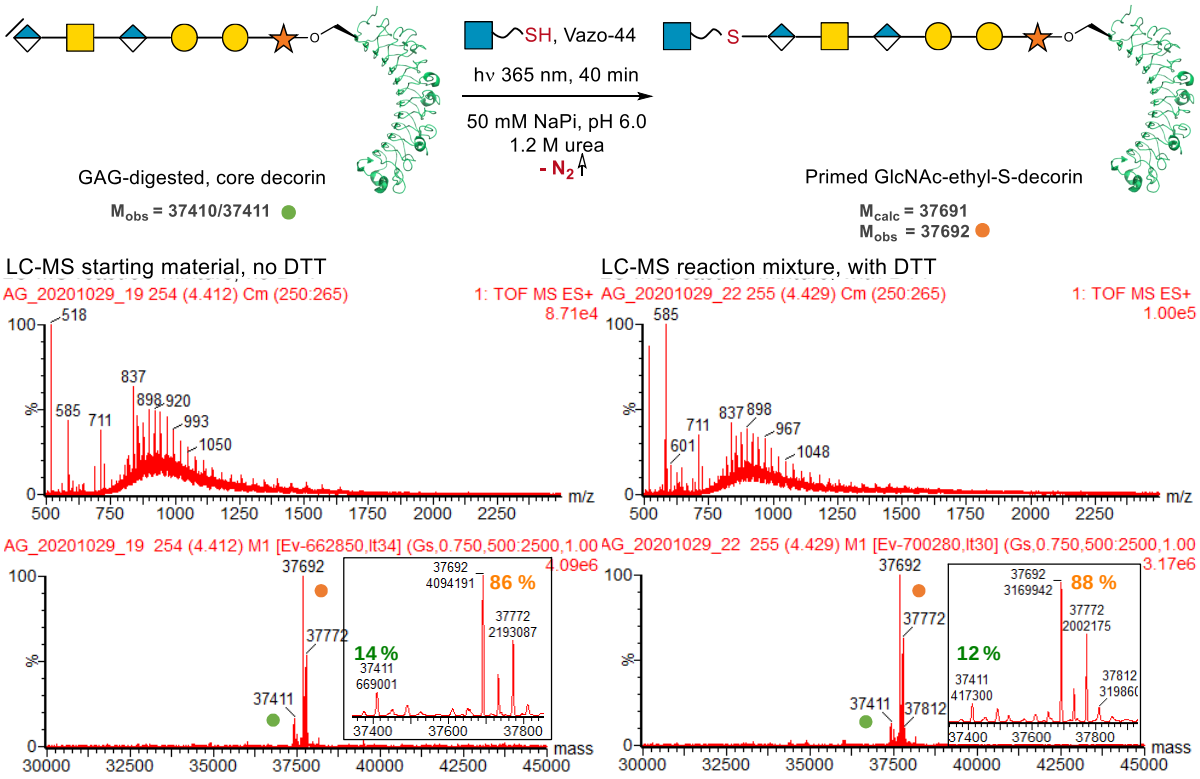


Figure 2.11. Optimized reaction conditions allowed excellent conversion of core decorin to GlcNAc-ethyl-S-decorin single addition product and no undesired double addition was observed even on a 150 µL scale. Reaction was performed in triplicate and 79 ± 7 % conversions were achieved. The figure shows the LC-MS spectra that were obtained for one of the replicates (86 % conversion).

2.2.3 Chemo-enzymatic synthesis of neoproteoglycans containing a heparosan GAG chain

Once we obtained the desired GlcNAc-S-primed PG, we embarked on exploring enzymatic reaction conditions that would install a non-native Hep chain on decorin (Figure **2.12**). To this end, the *N*-terminus His-tagged heparosan synthase pmHS2[85] was expressed in *E. coli* using a plasmid that had been previously prepared within our group to contain the DNA sequence encoding the enzyme.[86] SDS-PAGE and LC-MS analysis after purification by Ni^{2+}-affinity chromatography confirmed the successful isolation of pmHS2, and the recorded circular dichroism (CD) spectrum revealed that the enzyme had maintained a largely unaltered secondary structure (Figure SI 2.6).

Then, the enzymatic activity of pmHS2 was tested on an easily accessible substrate such as the GlcNAcylated histone H3. We envisioned that by optimizing the enzymatic extension on this model protein we would also obtain a histone-derived neoproteoglycan. Upon biological testing, the latter may reveal down-stream effects that, if similar to the ones evoked by the remodeled decorin, could be considered as being unaffected by the protein backbone and thus be attributed to the linked Hep chain (Figure **2.12**).

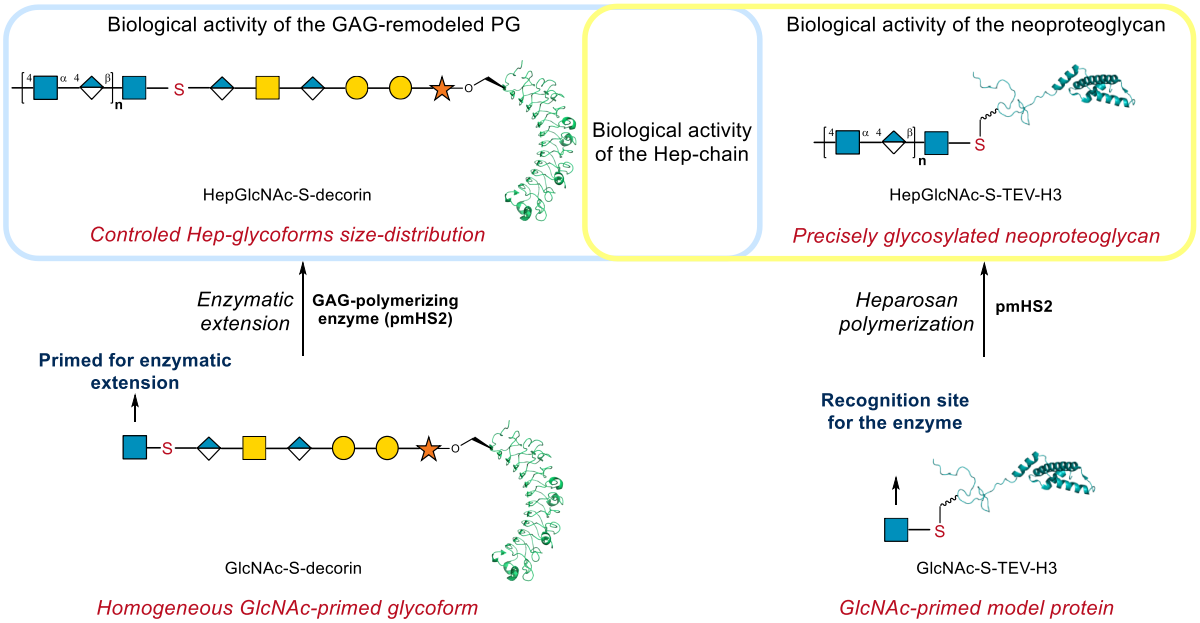


Figure 2.12. An enzymatic strategy was envisioned in which optimization of the on-protein heparosan polymerization of the GlcNAc-S-TEV-H3 histone would be easily transferable to enzymatically remodel a PG primed with the same thio-saccharide (GlcNAc-S-decorin). In this way, the PG-analogue that would eventually carry a heparosan (Hep) has the potential to exhibit a biological activity (highlighted by the blue square) different from the one observed when testing the Hep-remodeled H3 neoproteoglycan (highlighted by the yellow square). Any biological effect caused by both HepGlcNAc-S-decorin and HepGlcNAc-S-TEV-H3 may be attributed to the common structural feature: the newly installed Hep chain (intersection of the two squares).

The plasmid encoding for the histone H3 protein had been previously prepared within our group to contain a cysteine residue at the second amino acid position from the *N*-terminus, followed by the recognition cleavage site of the *Tobacco etch* virus (TEV)-protease. We hypothesized that the presence of a Cys-handle on a solvent-exposed, floppy *N*-tail qualified the Cys-TEV-H3 as a suitable enzymatic substrate for GAG-elongation. Moreover, if required, upon TEV-protease digestion the released *N*-terminal (glyco)peptide could be further analysed by LC-MS/MS.

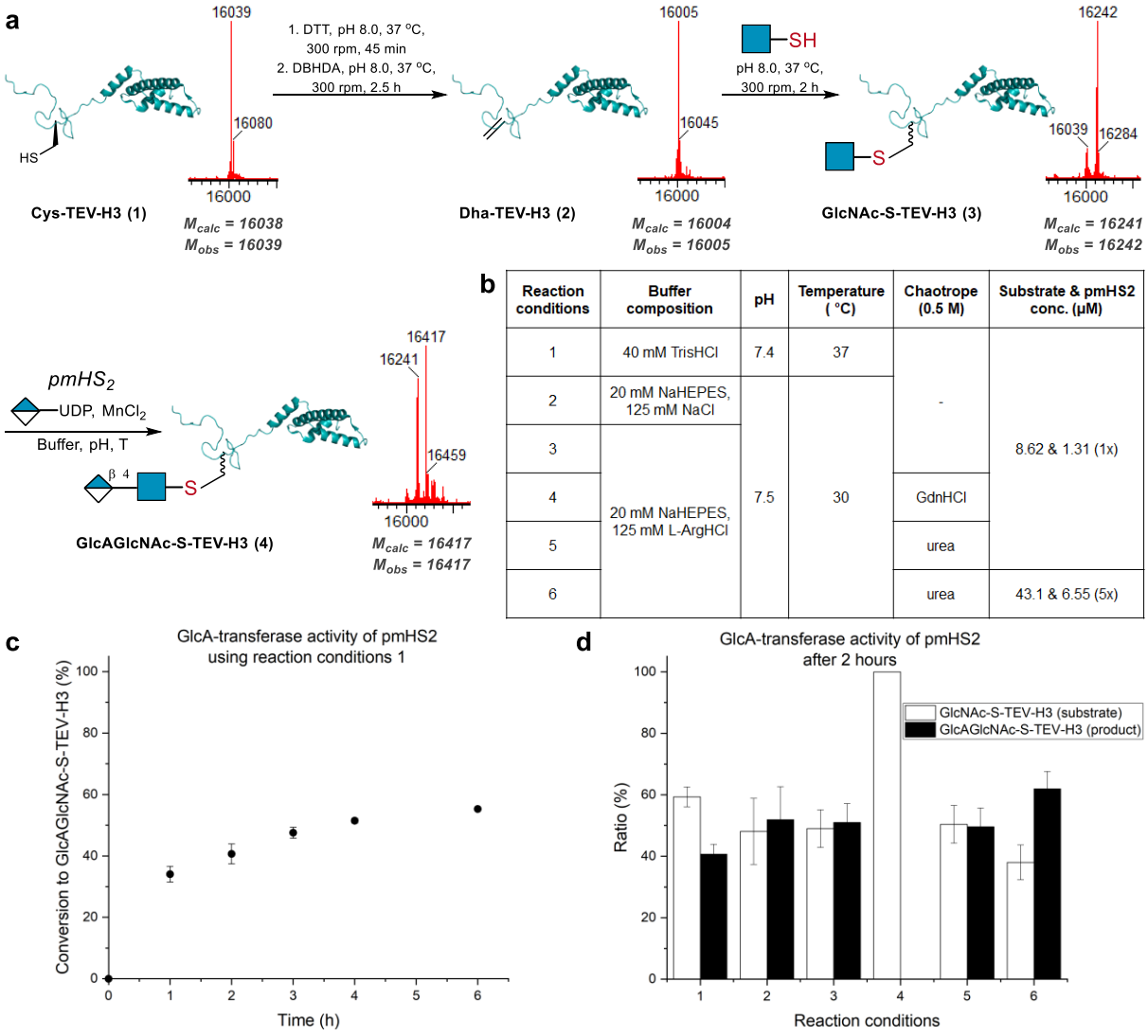
In order to access the model GlcNAc-primed protein a combined chemical mutagenesis / thia-Michael addition strategy was employed (Figure **2.13**, a and figure SI 2.22).[87] Upon reaction with the bis-alkylating reagent 2,5 dibromohexanediamide (DBHDA), Cys-TEV-H3 (1) was converted to the corresponding dehydroalanine (Dha) derivative (Dha-TEV-H3, 2). Next, thia-Michael addition of GlcNAc-SH to the unsaturated amino acid of 2 afforded GlcNAc-S-TEV-H3 (3), the desired unnatural enzymatic substrate that was further used for pmHS2 recognition and GAG-elongation.

The first enzymatic GlcA-extension of GlcNAc-S-TEV-H3 (8.62 µM) was performed in TrisHCl pH 7.4 buffer, at 37 °C, using a high concentration of UDP-GlcA (5 mM), 2.5 mM $MnCl_2$ and 0.15 equivalents enzyme (1.31 µM, Figure **2.13**, b, entry **1**), an amount that would allow the LC-separation and the reliable MS detection of both proteins. Although we were able to achieve 41 % conversion to the desired GlcAGlcNAc-S-TEV-H3 (4) intermediate after two hours, this increased to only 55 % after six hours of incubation (Figure **2.13**, c).

Since significant protein precipitation was observed in the reaction mixture, we decided to better assess the stability of the enzyme by performing a Differential Scanning Fluorimetry (DSF)/ThermoFluor Stability Assay (Figure SI 2.7 and 2.8). DSF is based on the principle that the fluorescence of a nonspecific protein-binding dye such as SYPRO Orange increases together with increasing the hydrophobicity of its environment.[88,89] Accordingly, the protein solution is heated in the presence of SYPRO Orange and upon denaturation, the dye fluoresces as it binds to the internal, hydrophobic patches of the protein core. The maximal fluorescence signal is obtained when the protein unfolds completely, and the inflection point of the obtained sigmoidal curve corresponds to the melting temperature (Tm), at which 50 % of the protein is unfolded. The DSF assay was carried out in a 96-well plate containing various buffers with a large range of pH values and different salt concentrations. The screening showed that pmHS2 is more stable at a pH ranging from 6 to 8 and the highest Tm (40.4 °C) was measured in two buffers: 20 mM NaHEPES pH 7.5 and 50 mM TrisHCl pH 8.0. Addition of NaCl (125 mM) to both buffers had a positive effect on the stability of the enzyme but most importantly, at 37 °C, the protein resulted to be unfolding in many of the screened buffers.

Therefore, in order to minimize the unfolding phenomenon, in which exposure to the aqueous environment of internal, hydrophobic groups of the protein might have contributed to the

precipitation of the enzyme, we decided to perform the enzymatic GlcA-extension in a 20 mM NaHEPES, 125 mM NaCl pH 7.5 buffer, at 30 °C (Figure **2.13**, b, entry **2**).



Reaction conditions	Buffer composition	pH	Temperature (°C)	Chaotrope (0.5 M)	Substrate & pmHS2 conc. (μM)
1	40 mM TrisHCl	7.4	37	-	8.62 & 1.31 (1x)
2	20 mM NaHEPES, 125 mM NaCl	7.5	30		
3	20 mM NaHEPES, 125 mM L-ArgHCl				
4				GdnHCl	
5				urea	
6				urea	43.1 & 6.55 (5x)

Figure 2.13. Chemical mutagenesis of Cys-TEV-H3 to Dha-TEV-H3 and subsequent thia-Michael glycosylation allow the installation of a GlcNAc-S-primer onto the protein substrate and the subsequent pmHS2-catalysed GlcA-extension. Excellent conversion of Cys-TEV-H3 (1) to the corresponding Dha-derivative (2) was achieved, and the thia-Michael addition of GlcNAc-SH to 2 afforded the desired product GlcNAc-S-TEV-H3 (3) (**a**). The enzymatic formation of GlcAGlcNAc-S-TEV-H3 (4) underwent a series of optimization experiments (**b**). The GlcA-addition performed in TrisHCl pH 7.4 buffer at 37 °C (Entry 1) revealed that after two hours, protein precipitation was causing significant slowdown of the enzymatic extension (**c**). The GlcA-elongation performed using reaction conditions identified in the DSF assay (Entry 2 and **d**) showed that the latter were causing less precipitation and the conversion to 4 was slightly increased, from 41 +/- 3 % to 52 +/- 11 %, after two hours. Similarly, incubation of the GlcNAc-S-TEV-H3 substrate with pmHS2, UDP-GlcA and $MnCl_2$ in the buffer that had shown to prevent the aggregation of decorin (Entry 3 and **d**), gave GlcAGlcNAc-S-TEV-H3 with good conversion (51 +/- 6 %). While addition of 0.5 M guanidinium hydrochloride (GdnHCl) was completely deleterious of the GlcA-transferase activity (Entry 4 and **d**), supplementation with 0.5 M urea was well tolerated by the enzyme (Entry 5 and **d**, 50 +/- 6 %). Lastly, upon concentration (5x) of the protein mixture 62 +/- 6 % conversion to the GlcAGlcNAc-S-TEV-H3 intermediate was achieved (Entry 6 and **d**). Reactions described by the entries 1, 2, 3 and 5 were performed in duplicate, entry 4 is a single-replicate experiment, and reaction 6 was performed in triplicate. Representative deconvoluted LC-MS spectra of these experiments are reported in Section 2.4.4.2, Figure SI 2.23.

Indeed, this helped us decrease the protein precipitation and enabled us to increase the conversion from 41 % to 52 % after two hours of incubation (Figure **2.13**, d).

Since our aim was to eventually apply the optimized reaction conditions to enzymatically remodel the GlcNAc-primed decorin, we needed to assess the enzymatic activity of pmHS2 in the arginine-containing buffer (Figure **2.13**, b, entry **3**) which had shown to be critical for preventing aggregation of the PG (Figure SI 2.21). Pleasingly, pmHS2 performed similarly in this buffer and 51 % conversion to the desired GlcAGlcNAc-S-TEV-H3 product was obtained after two hours (Figure **2.13**, d).

Moreover, since our PG of interest, the GAG-trimmed decorin, had shown to dimerize once the CS/DS chain was digested (Figure **2.6**) we were interested in identifying the maximum concentration of a chaotrope that would favour stabilization of the monomeric state, and would still not affect the enzymatic activity of pmHS2. Interestingly, by adding 0.5 M guanidinium hydrochloride (GdnHCl) to the reaction mixture in which the pmHS2-catalysed GlcA-extension of GlcNAc-S-TEV-H3 was evaluated (Figure **2.13**, b entry **4**), we noticed that the additive had completely abolished the GlcA-transferase activity of the enzyme (Figure **2.13**, d). On the contrary, the same amount of urea (Figure **2.13**, b entry **5**) was easily tolerated by pmHS2 and lead to 50 % conversion to the desired GlcAGlcNAc-S-TEV-H3 (Figure **2.13**, d).

In addition, it appeared that a low concentration of urea improved the solubility of the two proteins when mixed at high concentrations (43.1 µM substrate and 6.55 µM enzyme) which otherwise underwent extensive precipitation (Figure **2.13**, b entry **6**). We hypothesized that concentrating both proteins would help increase the affinity of the enzyme for the GlcNAc-S-protein substrate; indeed, by doing so, we were able to further increase the conversion (62 % after two hours) to the desired GlcAGlcNAc-S-TEV-H3 intermediate (Figure **2.13**, d).

During our research, we also investigated whether other GlcNAc-S-primed histones could be used to obtain additional neoproteoglycans bearing different protein backbones (Figure **2.14**).

Unfortunately, the histone H2B which carried an *S*-linked GlcNAc-primer at the 33rd amino acid position proved not to be a suitable substrate for the GAG-polymerizing enzyme.

Similarly, the histone H3 which lacked the *N*-terminus TEV-cleavage sequence and was GlcNAc-primed at the 10th amino acid position showed no conversion towards the GlcAGlcNAc-S-H3 product, even after six hours of incubation with pmHS2.

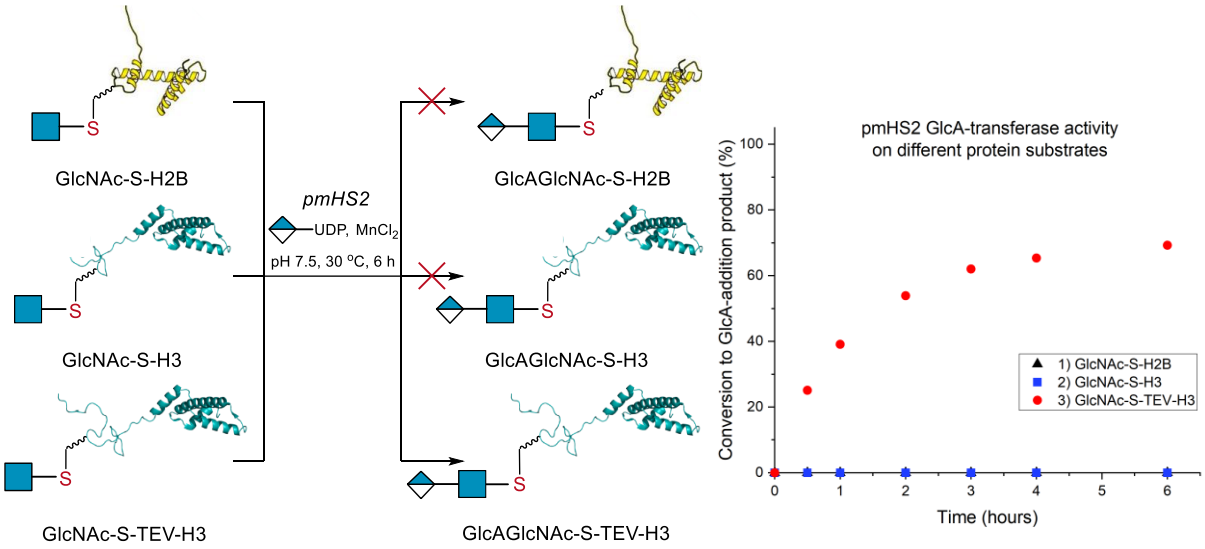
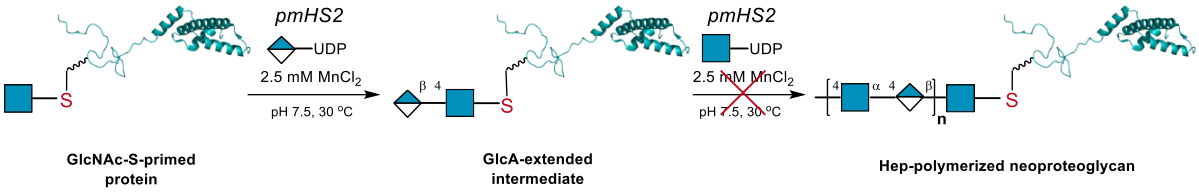
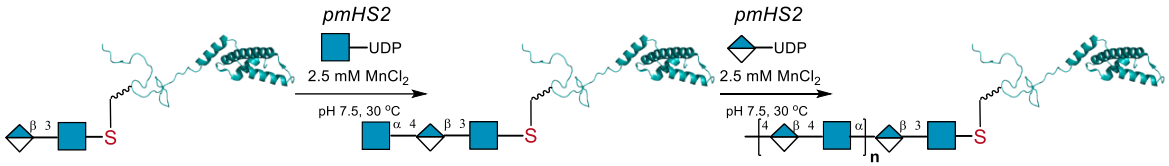


Figure 2.14. pmHS2 transferred a GlcA residue to acceptor GlcNAc-S-TEV-H3 but it was not able to extend model substrates GlcNAc-S-H2B and GlcNAc-S-H3. Graphs show conversions to GlcAGlcNAc-S-H2B (black), GlcAGlcNAc-S-H3 (blue) and GlcAGlcNAc-S-TEV-H3 (red) as determined by LC-MS analyses of aliquots taken for each of the three, single-replicate experiments that were performed in parallel. Representative deconvoluted LC-MS spectra of these experiments are reported in Section 2.4.4.3, Figure SI 2.24 and Section 2.4.4.4, Figure SI 2.26.

This experiment was instrumental in understanding that most likely, any attempted on-protein GAG-polymerization should employ a strategically chosen substrate, one that can be *S*-primed on a flexible, intrinsically disordered region and thus, able to reach the GlcA-transferase active site of the enzyme. This is a pre-requisite that we hadn't envisioned it as playing a relevant role in the remodeling of a GlcNAc-primed PG. For instance, in the specific case of decorin, we believed that the enzymatic extension would be facilitated by the linear glycan chain which, although *O*-linked to the 34th amino acid position, might be able to better present the GlcNAc-motif to the active site of pmHS2. Moreover, by protruding further away from the protein backbone the GlcNAc-S-hexasaccharide could also mitigate other phenomena that might negatively affect the remodeling process, such as any repulsive surface interactions that could occur between the enzyme and the PG-substrate, both in their native 3D conformation.



Entry	Subentry	Concentration of **GlcNAc-S-primed protein** (μM)	Concentration of ***pmHS2*** (μM)	Concentration of UDP (mM)	Concentration of UDP (mM)
1	A	8.62	1.31	5	5
	B				0.5
	C				0.1
2	A	43.1	6.55	5	5
	B				0.5
3	A	43.1	6.55	0.5	0.5



Entry	Subentry	Concentration of **GlcAGlcNAc-S-primed protein** (μM)	Concentration of ***pmHS2*** (μM)	Concentration of UDP (mM)	Concentration of UDP (mM)
4	A	43.1	6.55	0.5	0.5
	B			0.05	

Table 4. Explored reaction conditions for the on-protein Hep-polymerization revealed that depending on the starting material that was being used a different outcome could be achieved. The GlcNAc-S-primed substrate did not successfully compete with the off-protein GAG-synthesis (Entries 1 to 3) whereas a longer GlcAGlcNAc-S-primed protein (Entry 4) was enzymatically extended and based on the availability of sugar-donors, afforded neoproteoglycans with different degrees of polymerization. The table is a summary of the reaction conditions described in the figures 2.15, 2.16, 2.17 and 2.18.

After we had achieved the unprecedented pmHS2-catalysed GlcA-extension of a GlcNAc-S-primed protein substrate we embarked on exploring reaction conditions that would allow further elongation of the GlcAGlcNAc-S-TEV-H3 intermediate when both UDP-GlcA and UDP-GlcNAc were added to the reaction mixture (Table **4**). Hep-polymerization of GlcNAc-S-TEV-H3 (1) in the presence of a high concentration of UDP-GlcA (5 mM) could not be monitored by intact protein LC-MS analyses, although disappearance of the GlcA-extended intermediate (2, mol. mass 16417) was observed and higher MW species were detected by SDS-PAGE (Figure **2.15**).

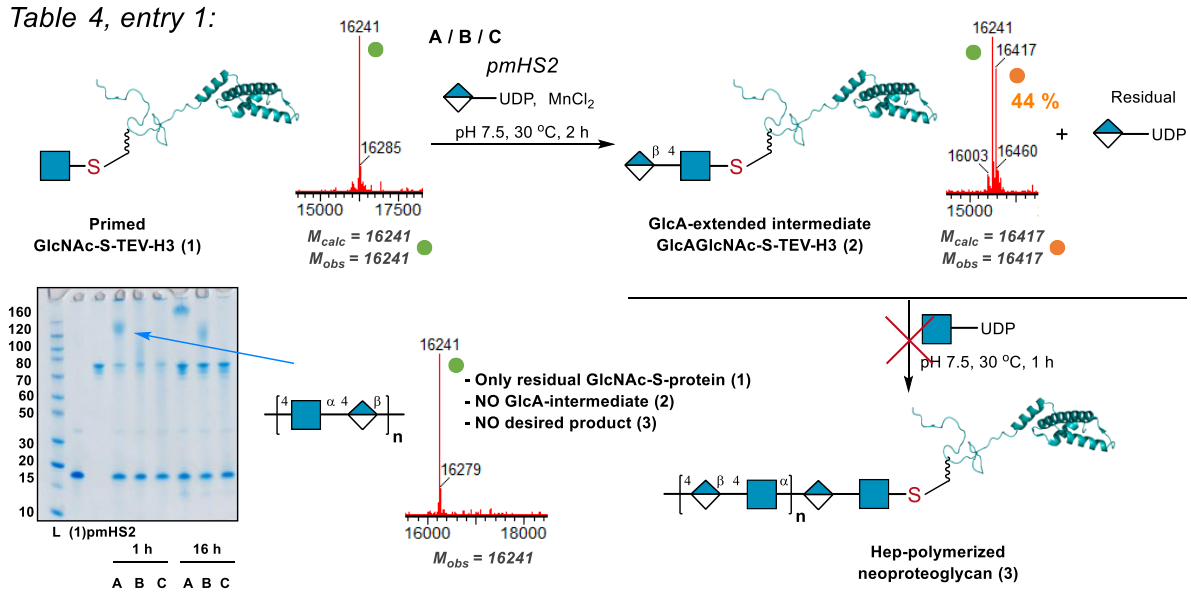


Figure 2.15. Attempted Hep-polymerization of GlcNAc-S-TEV-H3 in a diluted protein mixture, using a high concentration of UDP-GlcA and various concentrations of UDP-GlcNAc (Table 4, entry 1). Pre-incubation of the GlcNAc-S-primed protein substrate (8.62 µM) with pmHS2 (1.31 µM) and UDP-GlcA (5 mM) for 2 hours ensured that a satisfactory GlcA-addition (44 %) was achieved. However, upon addition of different concentrations of UDP-GlcNAc (A: 5 mM, B: 0.5 mM, C: 0.1 mM) the on-protein GAG-polymerization could not be monitored by intact protein LC-MS. The SDS-PAGE analysis of aliquots taken after 1 h and 16 h, for each of the three, single-replicate experiments that were performed in parallel, revealed that GAG chains had been synthesised instead.

When the reaction was performed in a concentrated solution of pmHS2 (6.55 µM) and substrate (43.1 µM) we were again unable to observe by intact protein LC-MS analyses molecular masses corresponding to protein glycoforms bearing multiple, repetitive GlcAGlcNAc disaccharides (Figure **2.16**). Moreover, western blot analysis using a primary anti-H3 antibody showed that the resulting high MW bands identified by SDS-PAGE were not attached to the H3 protein backbone but rather synthesized GAG-polymers. The ability of pmHS2 to extend a protein substrate with a GlcNAc-residue was demonstrated by isolating the GlcA-extended intermediate (2) using size-exclusion chromatography followed by incubation with UDP-GlcNAc alone, to obtain the expected GlcNAcGlcA-extended intermediate (4).

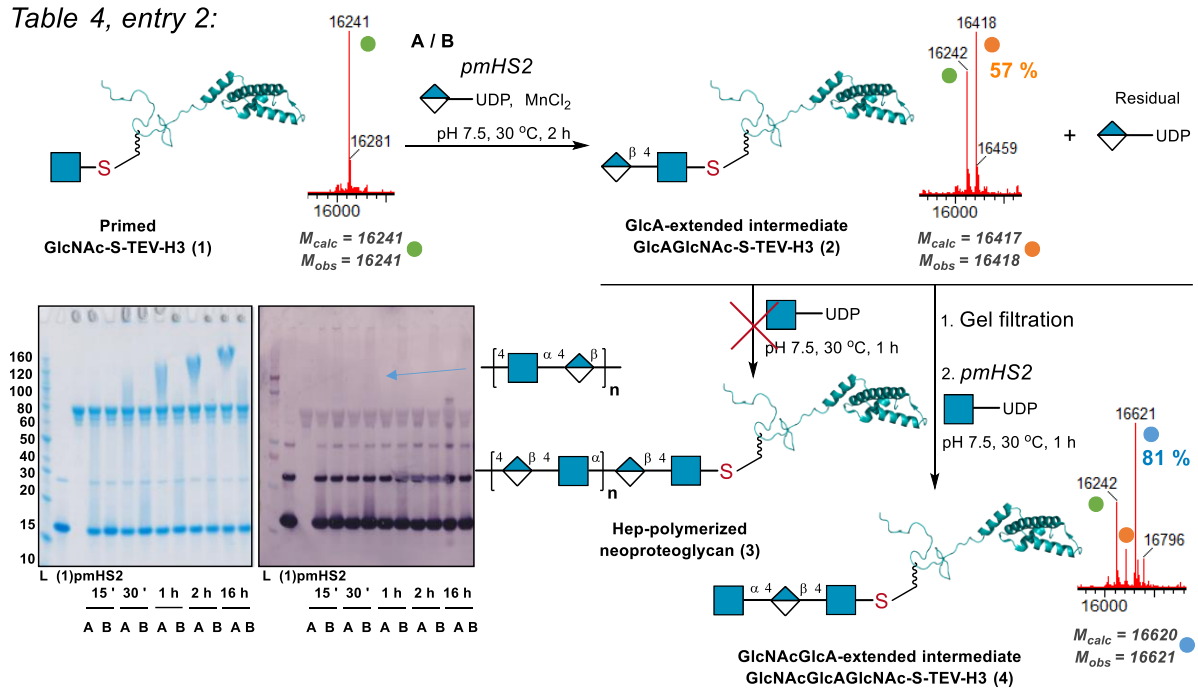


Figure 2.16. **Attempted Hep-polymerization of GlcNAc-S-TEV-H3 in a concentrated protein mixture, using a high concentration of UDP-GlcA and various concentrations of UDP-GlcNAc (Table 4, entry 2).** Pre-incubation of the GlcNAc-S-TEV-H3 (1) (43.1 µM) with pmHS2 (6.55 µM) and UDP-GlcA (5 mM) for 2 hours ensured that a satisfactory GlcA-addition (57 %) was achieved. However, upon addition of different concentrations of UDP-GlcNAc (A: 5 mM, B: 0.5 mM) the on-protein GAG-polymerization could not be monitored by intact protein LC-MS. The SDS-PAGE and western blot analyses of aliquots taken after 15 min, 30 min, 1 h, 2 h and 16 h, for each of the two, single-replicate experiments that were performed in parallel, revealed that GAG chains had been synthesised instead. To further confirm that the enzyme still possessed GlcNAc-transferase activity the GlcAGlcNAc-S-TEV-H3 (2) was purified by gel filtration prior to being incubated again with pmHS2 and UDP-GlcNAc (5 mM). Intact protein LC-MS analysis performed after 1 hour confirmed the formation of the GlcNAcGlcA-extended intermediate (4) (81 % conversion from 2).

In a third attempt, a low concentration of UDP-GlcA was added to a concentrated mixture of starting material 1 and pmHS2, and this significantly impacted the first enzymatic extension step: conversion to 2 was decreased to 41 % even though the incubation was extended with one more hour. Moreover, the subsequent addition of UDP-GlcNAc afforded only 60 % of the desired GlcNAcGlcA-extended intermediate (4), as observed by LC-MS (Figure **2.17**). This result suggested that the employed conditions still favoured the oligosaccharide extension rather than the on-protein, Hep-polymerization.

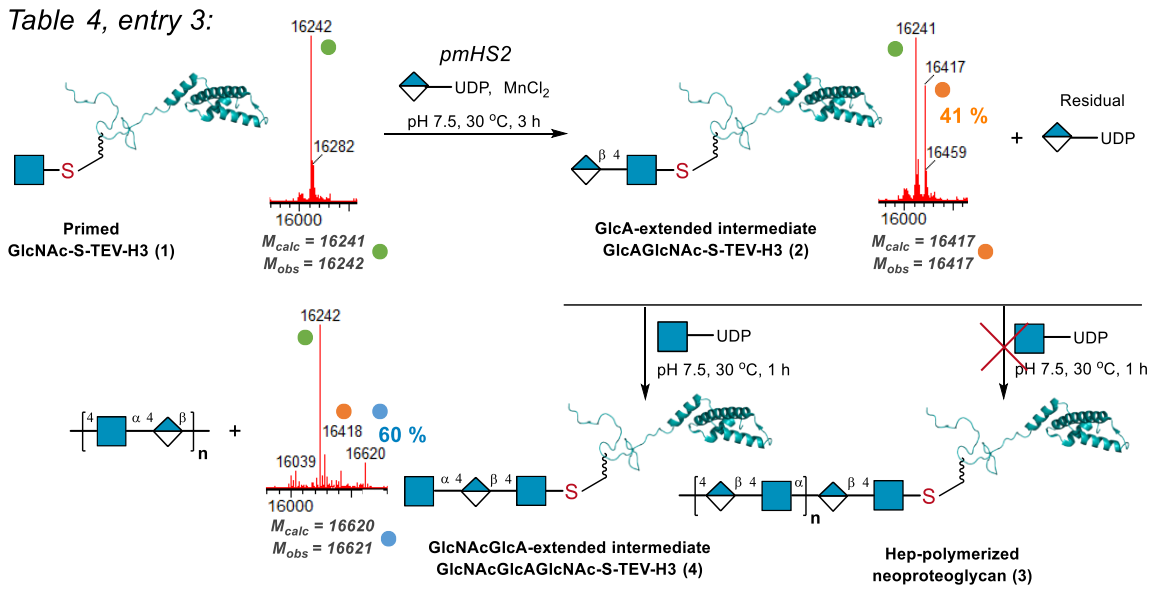


Figure 2.17. **Attempted Hep-polymerization of GlcNAc-S-TEV-H3 in a concentrated protein mixture, using low concentrations of UDP-GlcA and UDP-GlcNAc (Table 4, entry 3).** Pre-incubation of the GlcNAc-S-primed protein substrate (43.1 µM) with pmHS2 (6.55 µM) and UDP-GlcA (0.5 mM) for 3 hours ensured that a satisfactory GlcA-addition (41 %) was achieved. However, the addition of UDP-GlcNAc (0.5 mM) lead to reduced formation of the GlcNAcGlcA-extended product (60 % conversion from 2) and longer neoproteoglycans (3) were not detected by intact protein LC-MS analyses.

The latter was achieved when a GlcAGlcNAc-S-primed protein (5) obtained through the same combined chemical mutagenesis/thia-Michael addition strategy was used (Figure **2.18**). A longer primer such as the *S*-linked GlcAGlcNAc-disaccharide helped bypass the slow initiation phase, the GlcA-addition to the GlcNAc-acceptor, that required high concentrations of UDP-GlcA. Consequently, the GlcAGlcNAc-S-TEV-H3 substrate allowed the use of low concentrations of UDP-donors (0.5 mM), enabling the extension of the GlcNAcGlcAGlcNAc-protein template (6) rather than the initiation and extension of new GAG chains. Thus, this strategy afforded the desired Hep-neoproteoglycan H3 (3). Moreover, we were pleased to find that with an excess of UDP-GlcA (0.5 mM) and a limiting concentration of UDP-GlcNAc (0.05 mM, in an approximately 1:1 ratio to the protein substrate) the enzymatic extension was curtailed and a more homogeneous Hep-neoproteoglycan H3 (7) was obtained.

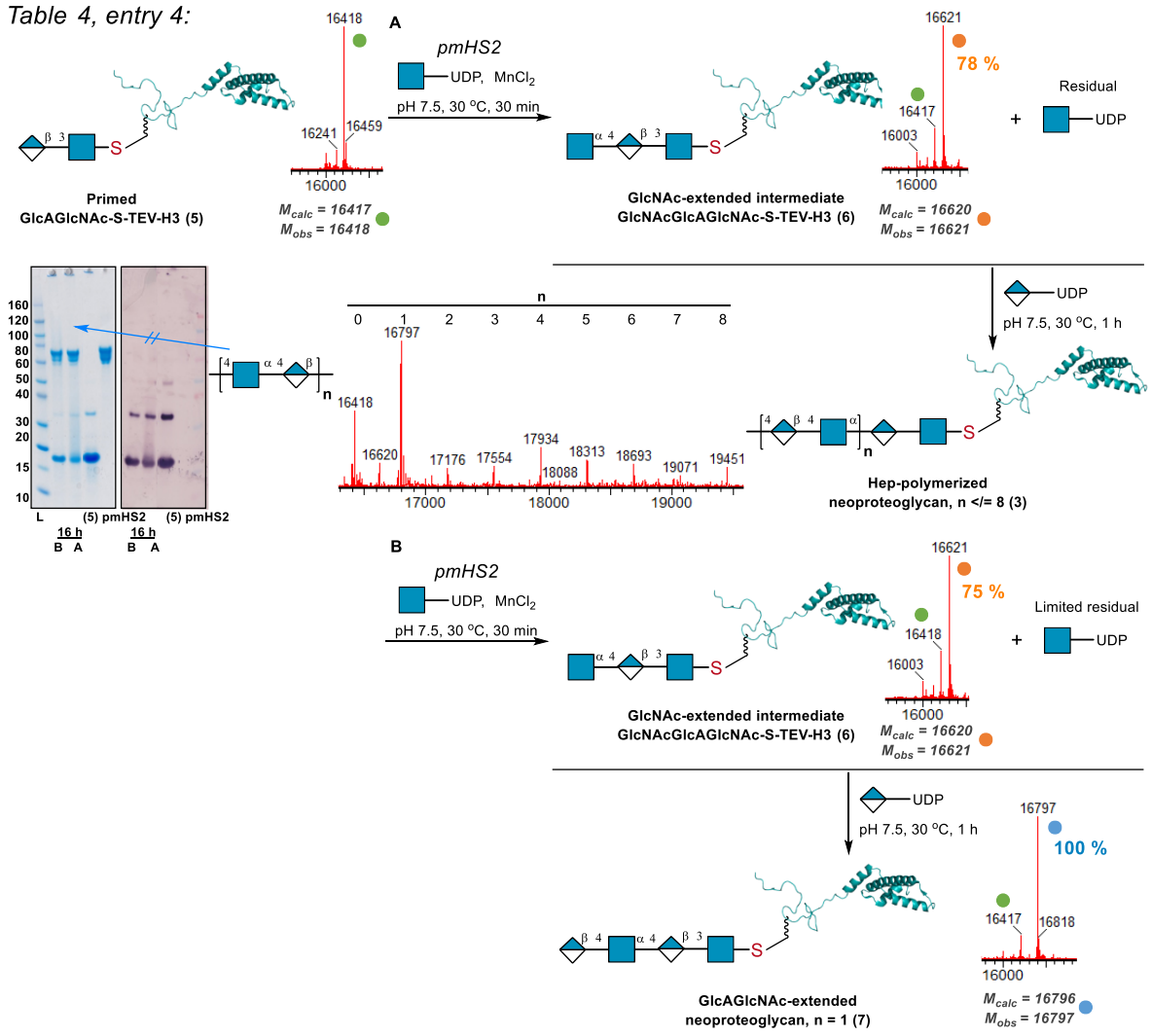


Figure 2.18. **Hep-polymerization of GlcAGlcNAc-S-TEV-H3 in a concentrated protein mixture, using low concentrations of UDP-GlcNAc and a low concentration of UDP-GlcA (Table 4, entry 4).** Pre-incubation of the GlcAGlcNAc-S-TEV-H3 (5) substrate (43.1 µM) with pmHS2 (6.55 µM) and UDP-GlcNAc (0.5 mM) for 30 min ensured that an excellent GlcA-addition (78 % conversion) was achieved. Moreover, Hep-neoproteoglycans (3) were detected by intact protein LC-MS upon further incubation with UDP-GlcA (A: 0.5 mM). The SDS-PAGE and western blot analyses of aliquots taken after 16 h, revealed that the on-protein Hep-extension had occurred and the off-protein GAG-polymerization had been significantly reduced. Pre-incubation of 5 (43.1 µM) with pmHS2 (6.55 µM) and UDP-GlcNAc (B: 0.05 mM) ensured that an excellent GlcNAc-addition (75 % conversion) was still achieved after just 30 minutes. The addition of UDP-GlcA (0.5 mM) afforded a more homogeneous, curtailed neoproteoglycan (7), the GlcAGlcNAcGlcAGlcNAc-S-TEV-H3 (100 % conversion from 6) as detected by intact protein LC-MS of an aliquot taken after 1 hour. The SDS-PAGE and western blot analyses of aliquots taken after 16 h, revealed that the off-protein GAG-polymerization had been significantly reduced.

2.2.4 Optimization of the pmHS2-catalysed enzymatic extension and development of the Hep-remodeled decorin

Having developed a protocol for the radical thiyl-glycosylation of the natural PG and explored enzymatic reaction conditions that were compatible with the on-protein GAG-polymerizing activity of pmHS2 using the histone 3 model system, we were now in a position to investigate whether the two GlcNAc-primed variants of decorin could be extended and remodeled by the enzyme into the desired non-native Hep-PG.

First, GlcNAc-S-decorin (Figure **2.19**, 1) was generated using the radical GAG ligation method, and upon removal of excess reagents and side-products, the product (8.76 µM) was incubated with pmHS2 (1.31 µM), UDP-GlcA (5 mM) and 2.5 mM $MnCl_2$, at 30 °C for two hours (Figure **2.19**, a). However, formation of desired GlcAGlcNAc-S-decorin (Figure **2.19**, 2) could not be observed by LC-MS analysis although similar reaction conditions lead to the extension of the GlcNAc-S-TEV-H3 (3) model protein and gave 54 % conversion to GlcAGlcNAc-S-TEV-H3 (4).

It was postulated that dimerization of decorin (section 2.2.1) might have reduced the accessibility of the GlcNAc-S-handle to the GlcA-transferase active site of pmHS2. Thus, we next investigated whether dissociation of the GlcNAc-S-decorin (1) dimers in the L-arginine buffer containing a high concentration of urea (6 M) would be beneficial. Purification of 1 in the denaturing 6 M urea buffer, and subsequent dilution with all components of the enzymatic reaction mixture to reach the final enzyme-compatible concentration of urea (0.5 M, Figure **2.13**) did not show the addition of a GlcA-residue to the primed decorin (Figure **2.19**, b). Moreover, the quality of the LC-MS analysis was significantly altered because the rapid change in conditions (transition from 6 M to 0.5 M urea buffer and from 4 °C to 30 °C) might have caused significant protein aggregation.

During our research, we also investigated whether other glycosyltransferases such as the well-studied bovine β-1,4-galactosyltransferase (GalT)[90] would tolerate an *S*-linked acceptor and transfer a Gal-residue from UDP-Galactose to a GlcNAc-primed protein substrate (Figure **2.19**, c and figure SI 2.28). This was readily achieved when the GlcNAc-S-TEV-H3 protein (3) was incubated with the enzyme (77 % conversion after four hours). Translation of the enzymatic reaction conditions to the GlcNAc-S-primed PG (1) did not reveal the expected Gal-extension neither after four hours nor upon longer incubation times.

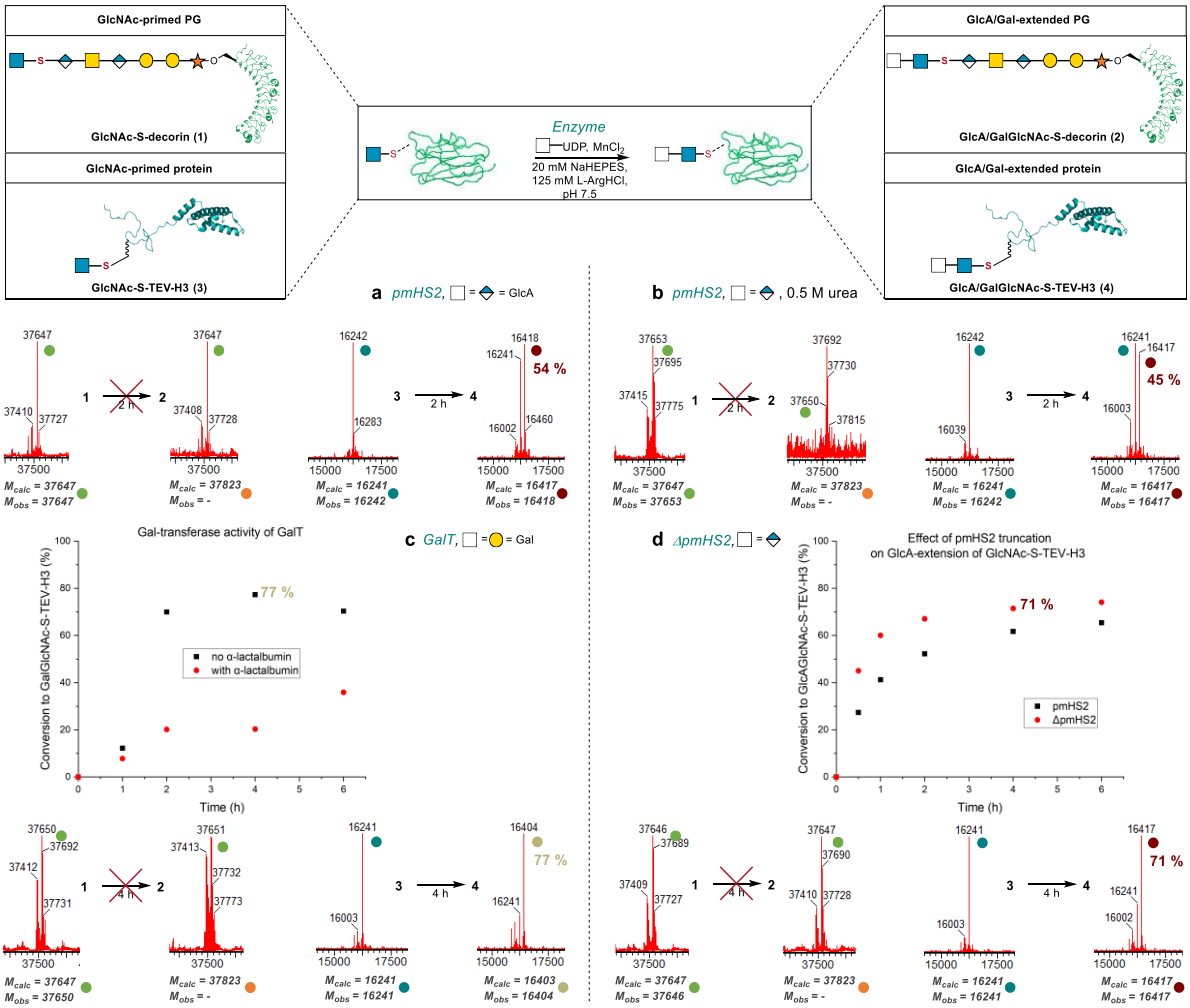


Figure 2.19. **pmHS2, ΔpmHS2 and GalT extend the primed GlcNAc-S-TEV-H3 but not the GlcNAc-S-decorin**. Attempted pmHS2-catalysed GlcA-extension of the GlcNAc-S-primed PG (1) did not afford the expected GlcAGlcNAc-S-decorin (2) although similar conditions applied for the extension of GlcNAc-S-TEV-H3 (3) gave 54 % conversion to GlcAGlcNAc-S-TEV-H3 (4) after 2 hours (**a**). Denaturation of 1 in a highly concentrated urea buffer (6 M) followed by dilution to the benign 0.5 M urea concentration (prior to adding pmHS2) did not form GlcAGlcNAc-S-decorin (2) either. Similar conditions applied for the extension of 3 gave 45 % conversion to 4 after 2 hours (**b**). The glycosyl transferase GalT was also unable to extend 1 although it successfully transferred a Gal residue to 3, revealing 77 % conversion to GalGlcNAc-S-TEV-H3 after 4 hours. The graph shows conversions to GalGlcNAc-S-TEV-H3 as determined by LC-MS analyses of aliquots taken for each of the two, single-replicate experiments (+/- α-lactalbumin)[90] that were performed in parallel (**c**). The truncated heparosan synthase ΔpmHS2 showed a slightly improved ability in extending 3 with a GlcA residue when compared to the original enzyme (71 % vs. 62 % and 74 % vs. 65 % conversion after 4 and 6 hours, respectively) but when incubated with 1, it did not afford the GlcA-extended decorin product. The graph shows conversions to GlcAGlcNAc-S-TEV-H3 as determined by LC-MS analyses of aliquots taken for each of the two, single-replicate experiments that were performed in parallel with the two, single-replicate experiments presented in Figure 2.21, b (**d**).

In our last enzymatic extension attempt with GlcNAc-S-decorin (1) as an acceptor, we explored whether a truncated version of pmHS2, lacking the first eighty *N*-terminal amino acids (ΔpmHS2),[29] would be able to extend the GlcNAc-S-primed PG (Figure **2.19**, d and figure SI 2.29). Although we had hypothesised that ΔpmHS2 might have a more accessible GlcA-transferase active site we could only observe an increased conversion (71 % vs. 62 %) of the GlcNAc-S-TEV-H3 (3) substrate to the GlcAGlcNAc-S-TEV-H3 (4) product, after four hours. The attempt to enzymatically extend GlcNAc-S-

decorin using the ΔpmHS2 variant failed to reveal any GlcA-addition even when the enzymatic reaction time was extended from four to 24 hours.

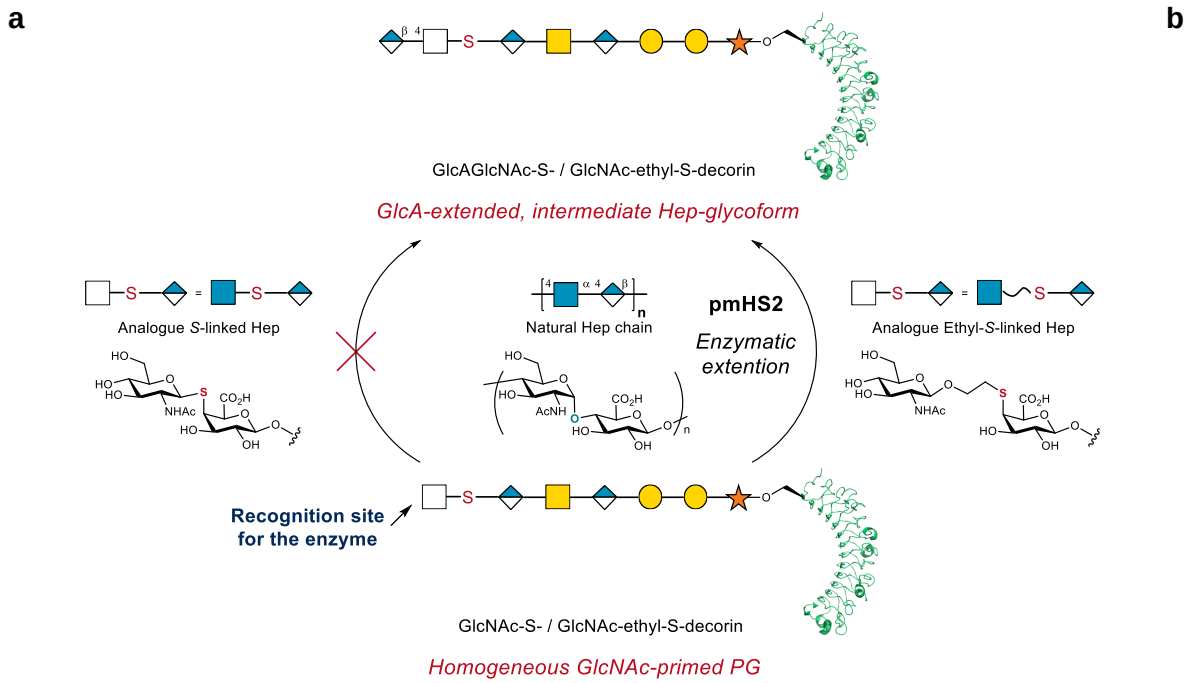


Figure 2.20. pmHS2 was unable to extend the GlcNAc-S-decorin substrate (a) but it transferred one GlcA residue to the GlcNAc-ethyl-S-decorin acceptor (b).

The work described so far demonstrates that two biocompatible "priming" strategies (chemical mutagenesis/thia-Michael addition and GAG-trimming/thiyl-glycosylation) were developed so that GlcNAc-S-proteins / -PGs be generated with high conversions. However, it became apparent that the pmHS2 enzyme did not exhibit sufficient plasticity to effect catalysis on the GlcNAc-S-decorin since the unnatural *S*-linkage, resulted from the axial thiyl addition to the ΔGlcA-tagged PG (Figure **2.2**), may have restricted the expected pmHS2 substrate, into an unproductive conformation for the GlcA-transfer to occur (Figure **2.20**, a). Therefore, we next turned our focus towards exploring whether the GlcNAc-ethyl-S-primed PG would be tolerated by the Hep-polymerizing enzyme (Figure **2.20**, b). We envisioned that insertion of a minimal ethyl-group would allow additional flexibility to the *S*-linked GlcNAc-primer and this would compensate for the restricted "galacturonic"-like conformation chemically imparted to the acidic sugar, part of the original decorin-linked hexasaccharide.

In order to assess the effect of the GlcNAc-ethyl-S-primer on the kinetics of the GlcA-addition step, we compared the activities of the two heparosan synthases (pmHS2 and its truncated variant ΔpmHS2) on both the GlcNAc-ethyl-S-decorin (1) and the GlcNAc-ethyl-S-TEV-H3 (3) substrate (Figure **2.21**). Pleasingly, the pmHS2-catalysed conversion of 3 to the GlcA-extended product GlcAGlcNAc-ethyl-S-TEV-H3 (4) reached 77 % and 80 % after four and six hours, respectively (Figure **2.21**, b, black),

which were slightly higher than those observed for the GlcNAc-S-TEV-H3 substrate (Figure **2.19**, d). As expected, a significantly increased GlcA-addition rate was observed when the GlcNAc-ethyl-S-primed histone was incubated with ΔpmHS2 (Figure **2.21**, b, red). In this case, complete conversion to the desired GlcAGlcNAc-ethyl-S-TEV-H3 (4) was observed already after four hours.[i]

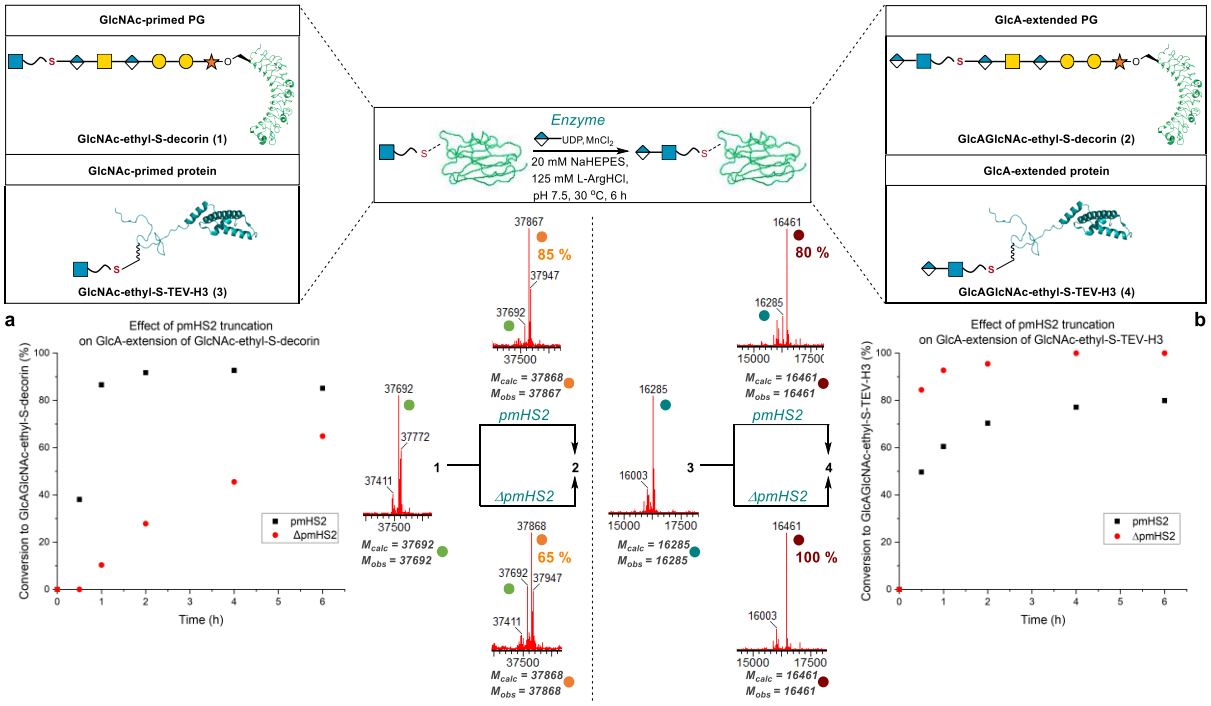


Figure 2.21. The *N*-terminus 80 amino acids truncation of the heparosan synthase has a profound effect on the Hep-remodeling of a GlcNAc-ethyl-S-primed natural PG. The truncated heparosan synthase ΔpmHS2 showed a slower GlcA-transferase activity than pmHS2 when these enzymes were incubated with the GlcNAc-ethyl-S-primed PG (1) (46 % vs. 93 % and 65 % vs. 85 % conversion after 4 and 6 hours, respectively). The graph shows conversions to GlcAGlcNAc-ethyl-S-decorin (2) as determined by LC-MS analyses of aliquots taken for each of the two, single-replicate experiments that were performed in parallel. All reactions were done in the presence of 5 mM UDP-GlcA (**a**). Conversely, ΔpmHS2 revealed an improved ability in extending the GlcNAc-ethyl-S-histone (3) with a GlcA residue, when compared to the original enzyme (100 % vs. 77 % and 100 % vs. 80 % conversion after 4 and 6 hours, respectively). The graph shows conversions to GlcAGlcNAc-ethyl-S-TEV-H3 (4) as determined by LC-MS analyses of aliquots taken for each of the two, single-replicate experiments that were performed in parallel with the two, single-replicate experiments presented in Figure 2.19, d (**b**).

Surprisingly, the outcome was different when the GlcNAc-ethy-S-decorin (1) was incubated with pmHS2 and ΔpmHS2 under the same reaction conditions (Figure **2.21**, a). The full-length enzyme was more efficient than the truncated variant in extending the primed PG; formation of the GlcA-extended decorin (2) reached 93 % and 85 % conversion after four and six hours, respectively (Figure **2.21**, a, black). Moreover, although about 20 % difference in product formation/conversion was observed when the two Hep synthases were incubated with 1 for six hours, the initial rates of pmHS2 and ΔpmHS2 were strikingly different. We therefore postulated that the reduced GlcA-addition rate

[i] It is worth mentioning that experiments herein presented (Figure **2.21**, b) were performed in parallel with the ones described in Figure **2.19**, d and a summary of the combined results is available for better visualisation (Figure SI 2.29).

observed for the ΔpmHS2 (10 % conversion after one hour) compared to the full-length pmHS2 (87 % conversion after one hour) could presumably be due to the lack of the eighty *N*-terminal amino acids. In the wild-type enzyme, even though this *N*-terminal domain may not be directly involved in the catalytic process, it may have a positive cooperative activity. Its association with the rest of the enzyme structure could potentially stabilize the protein into a more active conformation, with higher binding affinity for substrates carrying a polysaccharidic chain.[91]

We also observed a decrease in product formation for the full-length enzyme, from 93 % to 85 % after four hours (Figure **2.21**, a, black). We attributed this to the reversible glycosyltransferase activity of the enzyme. This is a characteristic that had been previously identified by Chen *et al.*[29] while investigating the ability of pmHS2 to extend isolated GAG chains.

To further demonstrate the specificity of the enzymatic GlcA-addition for the chemically inserted GlcNAc-handle, we performed control reactions in which the core decorin, bearing the non-primed ΔGlcA-tagged hexasaccharide, was incubated with UDP-GlcA and $MnCl_2$, both in the presence (Figure **2.22**, a, control **1**) and the absence of pmHS2 (Figure **2.22**, a, control **2**). Pleasingly, no GlcA-addition was observed by intact protein LC-MS analyses, suggesting that without a GlcNAc-primer, no residual saccharide on the protein core could be recognized as an enzymatic acceptor by pmHS2 (Figure SI 2.36). Moreover, the second control reaction, the one without pmHS2, confirmed that decorin does not possess any intrinsically reactive amino acids that could covalently bind GlcA upon nucleophilic attack to the UDP-GlcA (Figure SI 2.36).

Upon achieving the unprecedented pmHS2-catalysed GlcA-extension of a GlcNAc-ethyl-S-primed decorin, we embarked on exploring reaction conditions that would allow further elongation of the GlcA-intermediate.

Our main objective was to facilitate the on-protein GAG-extension and avoid the Hep synthesis occurring when both UDP-GlcA and UDP-GlcNAc are present in the reaction mixture, an issue we had previously encountered during the synthesis of the Hep-S-TEV-H3 neoproteoglycan.

Thus, we first started exploring the pre-incubation of the primed-PG (8.77 µM) with the enzyme (1.31 µM), in the presence of low concentrations of UDP-GlcA (Figure **2.22**, b). Pleasingly, a concentration of 0.5 mM UDP-GlcA would still allow the slow formation of the GlcAGlcNAc-ehtyl-S-decorin (Figure **2.22**, b, red) whereas a more pronounced reversible activity for pmHS2[29] was observed at a UDP-GlcA concentration as low as 0.05 mM (Figure **2.22**, b, black).

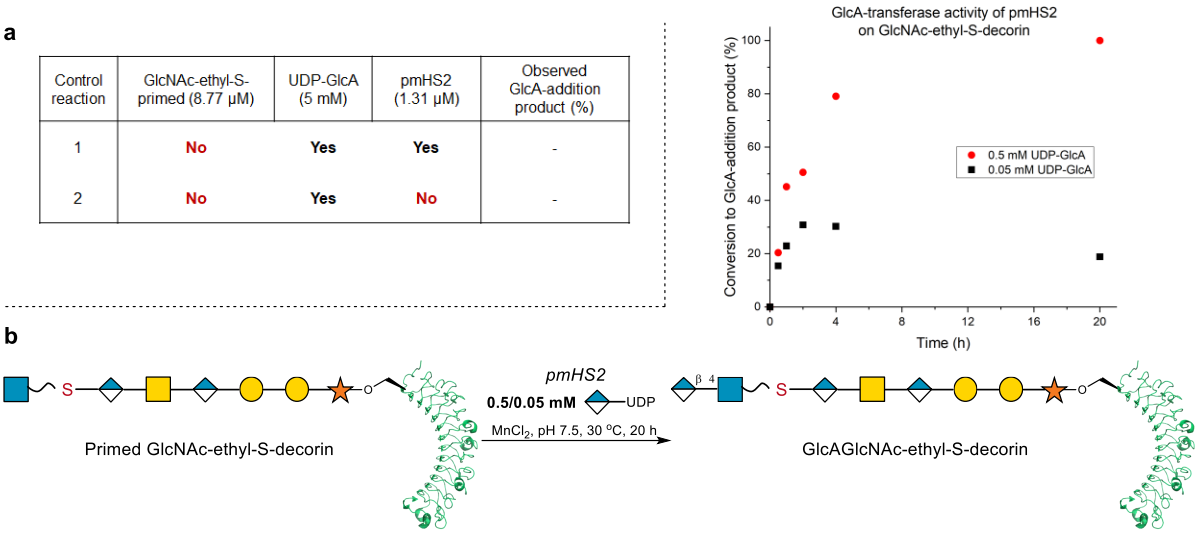


Figure 2.22. **Control enzymatic reactions and pmHS2-catalysed GlcA-extension of GlcNAc-etyl-S-decorin in the presence of low UDP-GlcA concentrations.** Control reactions were performed on the core decorin in the presence of pmHS2 (1) or without the enzyme (2) and no GlcA-addition was observed (Figure SI 2.36) by intact protein LC-MS (**a**). The installation of a GlcNAc-ethyl-S-primer onto decorin promotes the enzymatic GlcA-extension even at very low concentrations of UDP-GlcA: 100 % and 19 % obtained conversion with 0.5 mM and 0.05 mM UDP-GlcA, respectively. The graph shows conversions to GlcAGlcNAc-ethyl-S-decorin as determined by LC-MS analyses of aliquots taken for each of the two, single-replicate experiments that were performed in parallel (**b**).

Moreover, we were keen to determine whether the addition of UDP-GlcNAc (0.5 mM and 0.05 mM, respectively) after two hours pre-incubation with UDP-GlcA would result in the Hep-remodeling of the PG substrate. Not surprisingly, we encountered problems similar to the ones revealed by the Hep-S-TEV-H3 neoproteoglycan: while we could observe the disappearance of the intermediate GlcAGlcNAc-ethyl-S-decorin by LC-MS, no extended glycoforms could be detected. This suggested either the competition between the on-protein and the off-protein Hep-polymerization and the subsequent entanglement of the resulting complex reaction mixture had occurred, or that a very high size-distribution of low-abundant Hep-remodeled decorin glycoforms could not be detected by intact protein LC-MS.

Therefore, in order to promote the on-protein GAG-polymerization, we attempted the enzymatic remodeling reaction using the same PG:enzyme:UDP-GlcA ratio as described in Figure 2.22 (b, red) but we concentrated approximately three times the reaction mixture. Consequently, the resulting molarities changed from 8.77 µM: 1.31 µM: 0.5 mM to 29.23 µM: 4.37 µM: 1.67 mM.

After 14 hours incubation of GlcNAc-ethyl-S-decorin (1) with pmHS2 and UDP-GlcA, formation of the expected GlcAGlcNAc-intermediate (2) reached 81 % conversion (Figure **2.23**, a). Surprisingly though, the same amount of UDP-GlcNAc (1.7 mM) added to the reaction mixture only caused formation of the GlcNAc-extended product 3 (75 % conversion) and not further Hep-polymerization.

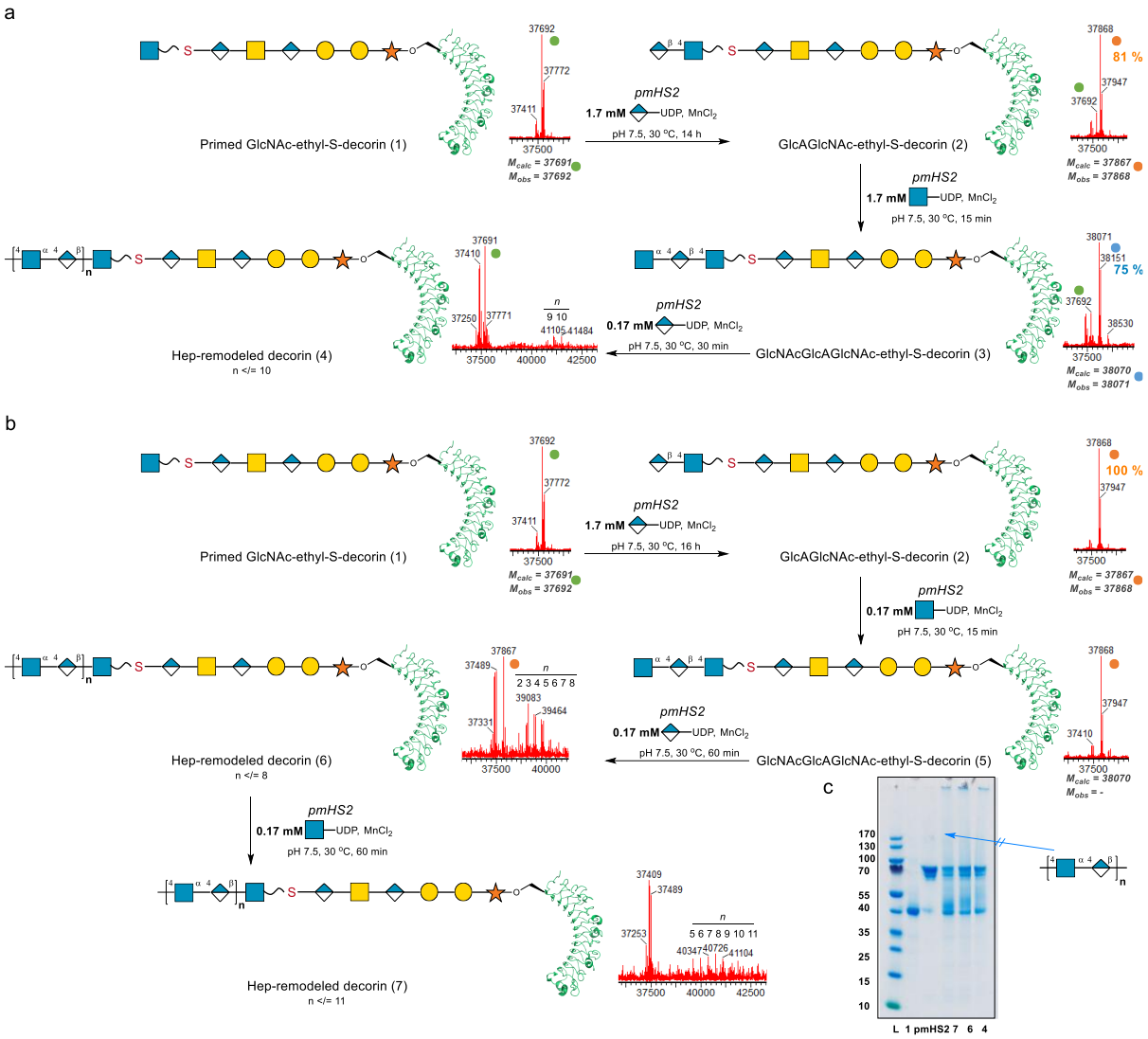


Figure 2.23. Hep-polymerization of GlcAGlcNAc-ethyl-S-decorin in a concentrated protein mixture, using low concentrations and repetitive additions of UDP-GlcA and UDP-GlcNAc. Pre-incubation of the GlcNAc-ethyl-S-decorin (1, 29.23 µM) with pmHS2 (4.37 µM) and UDP-GlcA (1.7 mM) for 14 h ensured that an excellent conversion (81 %) to GlcAGlcNAc-ethyl-S-decorin (2) was achieved. Moreover, the addition of UDP-GlcNAc (1.7 mM) led to excellent conversion (75 %) to GlcNAcGlcAGlcNAc-ethyl-S-decorin (3) in just 15 min. The Hep-remodeled decorin (4, n = 9-10) was detected by intact protein LC-MS upon further incubation with UDP-GlcA (0.17 mM) for 30 min (**a**). Pre-incubation of the the GlcNAc-ethyl-S-primed PG (1) substrate (29.23 µM) with pmHS2 (4.37 µM) and UDP-GlcA (1.7 mM) for 16 h ensured complete conversion to the GlcA-addition intermediate (2). Further incubation for 15 min with UDP-GlcNAc (0.17 mM) did not lead to the detection of GlcNAcGlcAGlcNAc-ethyl-S-decorin (5), but supplementation of the reaction mixture with additional UDP-GlcA (0.17 mM) afforded the Hep-remodeled decorin (6, n = 2-8). The latter was converted into longer PG-analogues (7) when UDP-GlcNAc (0.17 mM) was again added, and incubation was continued for 60 min (**b**). The SDS-PAGE analysis of aliquots taken after 16 hours for 4, and after 60 min for 6 and 7 revealed that the on-protein Hep-extension had occurred, and the off-protein GAG-polymerization had been significantly reduced.

This was presumably caused by the insufficient UDP-GlcA still present in the reaction mixture which, although it had been added in high excess, might have undergone chemical or enzymatic hydrolysis. Nevertheless, when more UDP-GlcA (0.17 mM) was added, this was readily used by the

GAG-polymerizing enzyme and the Hep-remodeled decorin (4) with up to n = 9 and 10 disaccharide GAG-repeating units could be detected by LC-MS (Figure **2.23**, a).

In a similar experiment, after 16 hours incubation of GlcNAc-ethyl-S-decorin (1) with pmHS2 and the first sugar-donor (1.7 mM UDP-GlcA), complete conversion to the desired GlcAGlcNAc-intermediate (2) was achieved (Figure **2.23**, b). However, upon addition of UDP-GlcNAc (0.17 mM), further extension couldn't be achieved within the first 15 minutes and the latter was detected only upon supplementing with the same amount of UDP-GlcA (0.17 mM). After 60 minutes, the Hep-remodeled decorin (6) could be observed, and the most abundant species carried 3 to 5 disaccharide repeating units (Figure **2.23**, b). The addition of a second aliquot of UDP-GlcNAc (0.17 mM) stimulated the on-protein polymerization to give longer PG-glycoforms, with up to 11 GAG-units (7). Noteworthily, in both trials the shorter glycoforms (molecular masses 37691 and 37867, respectively) appeared to be lagging behind, and this might have been an indication that the enzyme has a higher affinity for longer intermediates and consequently, it has an increased propensity to generate their corresponding longer species. Gratifyingly, the SDS-PAGE analysis of the Hep-remodeled decorin variants 4, 6 and 7 revealed the presence of higher molecular weight PG-species and no significant accumulation of extended GAG-chains (Figure **2.23**, c).

2.2.5 Decorin induces DNA transcription of the intercellular adhesion molecule 1 (ICAM-1) in human umbilical vein endothelial cells (HUVECs)

Previous studies had shown that when HUVECs were incubated with nanomolar concentrations (200 nM) of intact decorin (1) and GAG-trimmed decorin (2) for six hours, these underwent endothelial cells autophagy.[68] The autophagic process arose from the interaction of decorin with VEGFR2 and this triggered a signalling cascade that led to an increased synthesis of Bec1 and LC3, two key players in the autophagic cellular pathway (Figure **2.4**, b). Moreover, it appeared that the decorin-evoked formation of autophagosomes required the upstream expression of Peg3, since the decorin/VEGFR2/Peg3 signalling could be blocked by suppressing both VEGFR2 and Peg3 levels *via* specific, small interfering RNAs (siRNAs).

Thus, we decided to test whether HUVECs are an appropriate cellular system for the evaluation of the biological activity of our natural decorin variants. We first intended to reproduce the results reported by Iozzo R. V. *et al.*[68] so that we could later use the assay to investigate the activity of the Hep-remodeled decorin, and the extent to which the newly installed GAG chain impacts the autophagic process (Figure **2.24**). To this end, the endothelial cells were incubated for six hours with the three natural variants (1, 2 and 3) that we had prepared in two concentrations (100 and 200 nM). Then, the

cells were lysed, the mRNA was extracted and the transcriptional induction of Peg3, Bec1 and LC3 was evaluated by quantitative reverse transcription polymerase chain reaction (RT-qPCR).

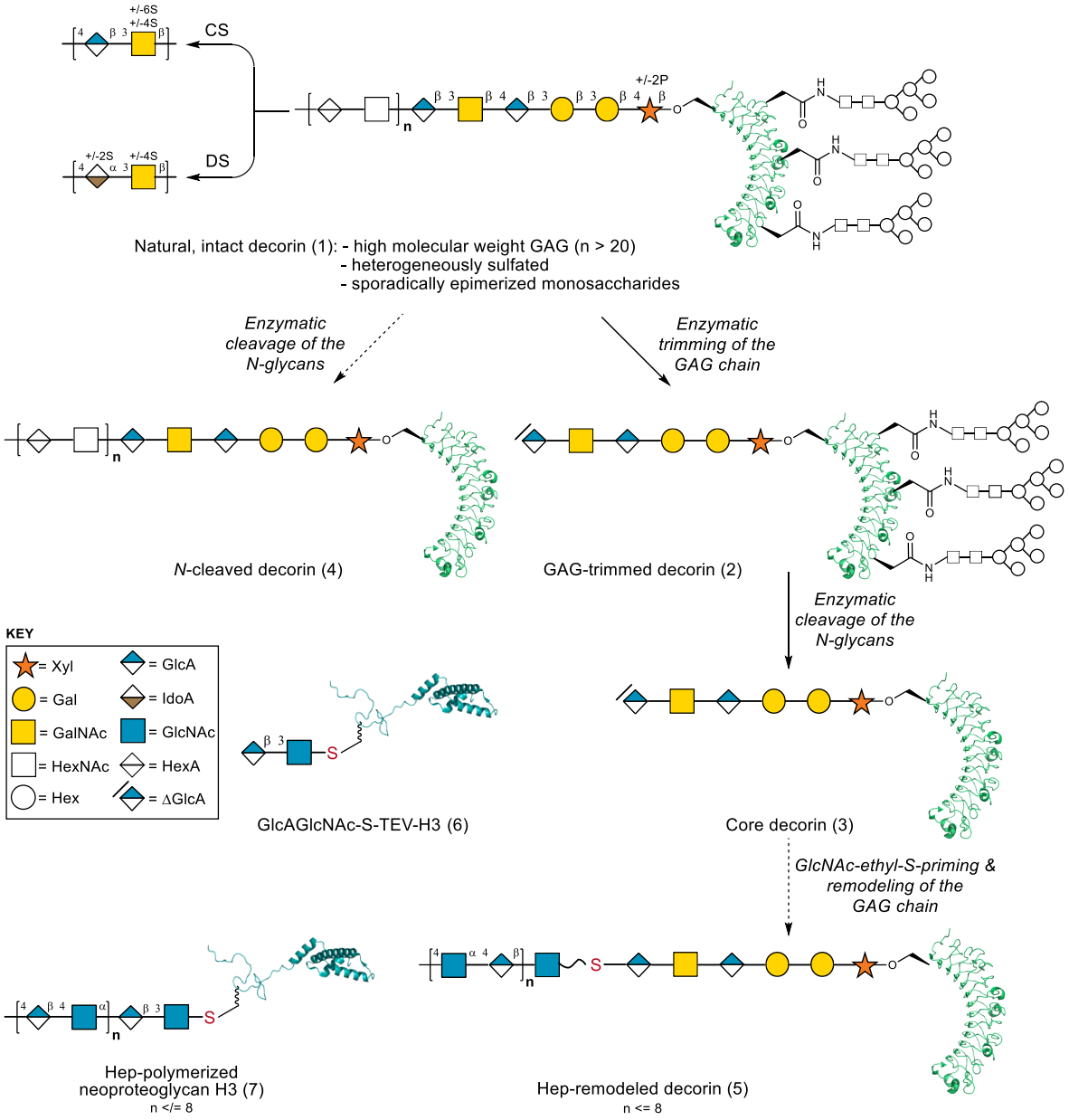


Figure 2.24. Decorin and histone H3-derived variants that were or planned to be tested using the HUVECs-based assay. Cells were incubated with intact decorin (1), GAG-trimmed PG (2) and core variant (3) at 100 nM and 200 nM, and the obtained results are presented in Figure 2.25. The assay will be optimized and further used for the evaluation of the PG resulted from the cleavage of the *N*-glycans (4), the Hep-remodeled decorin (5), the GlcAGlcNAc-S-primed H3 (6) and the corresponding Hep-containing neoproteoglycan (7).

Moreover, in the same experiment, we also investigated whether genes encoding for endothelial-specific pro-inflammatory biomolecules[92] such as the intercellular adhesion molecule 1 (ICAM-1), the vascular cell adhesion molecule 1 (VCAM-1) and the interleukin-1α (Il-1α) were upregulated as well. Unexpectedly, the positive control TNFα[92] employed in the bioassay did not promote the transcriptional activation of all the genes under investigation and thus, without additional

optimization the results that had been previously reported by Iozzo R. V. *et al.*[68] for 1 and 2 were not reproduced (Figure **2.25**).

Surprisingly, our data revealed that while the intact decorin (1) and the GAG-trimmed PG (2) upregulated the gene encoding for ICAM-1 at a concentration of 200 nM similarly to the positive control TNFα, the transcriptional activation did not occur when the cells were incubated with the decorin protein core (3) (Figure **2.25**, pink square). Moreover, since the gene transcription was not induced when the HUVECs were incubated with 1 and 2 at a lower concentration (100 nM) the transcriptional effect appeared to have been concentration dependent. In addition, while the data with regards to the upregulation of VCAM-1 remained inconclusive the transcription of Il-1α in endothelial cells resulted unaffected by the addition of the three decorin variants.

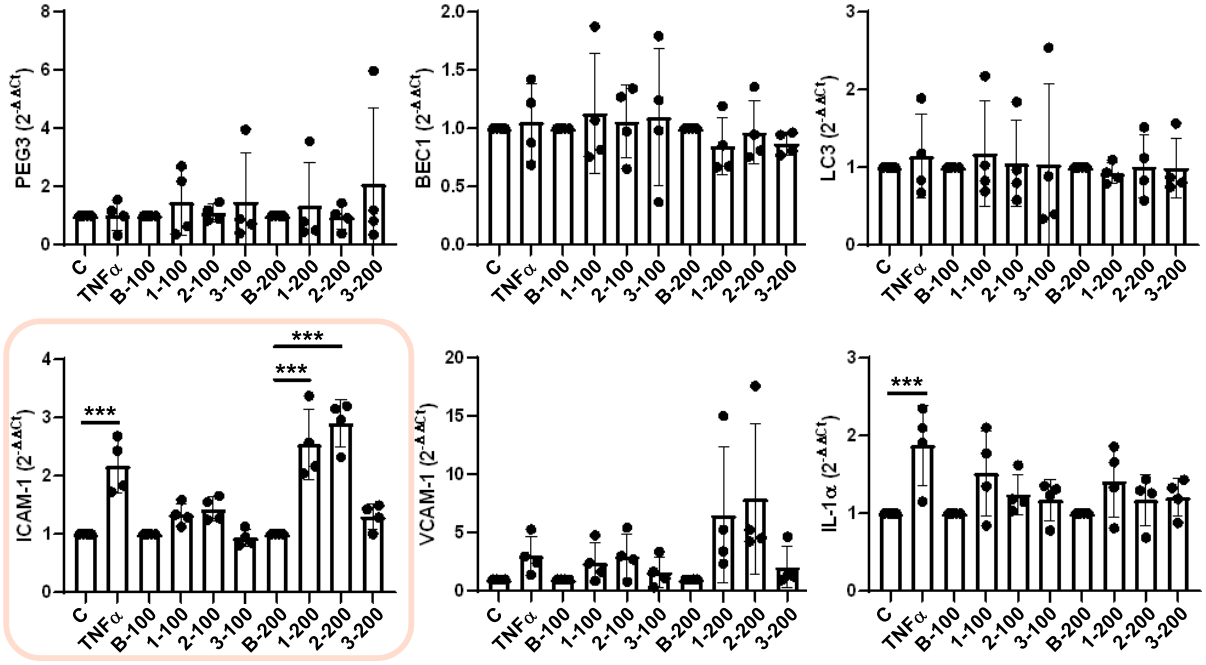


Figure 2.25. **Natural decorin (1) and its GAG-trimmed variant (2) caused upregulation of the ICAM-1 receptor whereas the core PG (3) did not elicit a similar effect.** 1 and 2 activated the transcription of the gene encoding for ICAM-1 at 200 nM (1-200, 2-200) but not when less concentrated (1-100, 2-100). Conversely, 3 could not induce the same effect at any of the two concentrations that were tested (3-100, 3-200). The images shown are representative of the fold changes that were calculated using the double ΔCt method ($2^{-\Delta\Delta CT}$) ± SEM. Data derived from two parallel trials run in quadruplicate for each gene of interest (C: negative control; B: 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer; TNFα: tumour necrosis factor α, positive control).

We are fully aware that these results will have to be confirmed by additional independent replicates able to undoubtedly prove the role of the *N*-glycans of decorin on the ICAM-1 upregulation process. However, we believe that these findings may represent the first cues that in addition to the TLR2/4-mediated pro-inflammatory role of this PG in macrophages (Figure **2.4**, b, violet),[57] decorin may also have immunomodulatory effects in endothelial cells *via* VEGFR2-activation.

In future experiments we also aim to evaluate on HUVECs the biological activity of *N*-cleaved decorin (4), Hep-remodeled decorin (5), GlcAGlcNAc-S-primed H3 (6) and Hep-polymerized neoproteoglycan H3 (7) (Figure **2.24**).

2.3 Conclusions and future perspectives

As part of our overall aim of generating GAG-remodeled proteoglycan-analogues for subsequent biological studies, in this chapter we set out to explore the viability of the thiyl glycosylation as a strategy in this regard.

The chosen PG decorin was successfully extracted from a natural source and its intact, GAG-trimmed and hexasaccharide-containing (core) variants were purified. These were characterized using multiple analytical techniques, separately or combined when more appropriate: SDS-PAGE, iSCAMS, native or denaturing protein LC-MS analysis, and glycoproteomics.

The feasibility of the free-radical *S*-ligation with respect to both product formation and selectivity for the ΔGlcA residue over any cysteine residue present in the core structure of decorin, were investigated. The successful thiyl glycosylation of the decorin core protein was demonstrated and optimised, enabling good conversions to be attained. Moreover, dialysis in the L-ArgHCl-containing buffer allowed good recovery of the GlcNAc-primed decorin that would have otherwise aggregated.

We next obtained the Hep-synthase pmHS2 and evaluated its GAG-polymerizing activity on the GlcNAc-primed (the GlcNAc-S-TEV-H3 acceptor) and on the GlcAGlcNAc-primed (the GlcAGlcNAc-S-TEV-H3 acceptor) model proteins. These were obtained *via* a combined chemical mutagenesis/thia-Michael addition strategy, and after careful optimization of the enzymatic reaction conditions the Hep-remodeled H3 neoproteoglycan was obtained.

Although it was considered that pmHS2, its truncated variant ΔpmHS2 and the more plastic GT GalT will tolerate and further polymerize the glycosidic chain of the primed GlcNAc-S-decorin substrate, no such activity was observed for any of these three enzymes.

Through the synthesis and ligation of GlcNAc-ethyl-SH to the ene-tagged decorin another minimal, structurally accurate PG-mimic was generated: the GlcNAc-ethyl-S-decorin. Enzymatic extension studies revealed that the insertion of an unnatural ethyl-group provided additional flexibility to the *S*-linked GlcNAc-primer which could then be further extended by the pmHS2. Moreover, the experiments that were performed to compare the enzymatic activity of pmHS2 and its truncated variant ΔpmHS2 offer the potential to gain new insights into the catalytic mechanism of this heparosan synthase which may be more complex than it is currently speculated.

Lastly, when we tested the natural decorin variants (intact, GAG-trimmed and core decorin) in a HUVECs biological assay we found that the first two caused increased transcription of the gene encoding for the ICAM-1 receptor whereas the same effect was not caused by the core PG.

We envisage that the developed thiyl-glycosylation/GAG-remodeling methodology will be further developed and applied to enable the study of GAG-dependent cellular mechanisms of transmembrane PGs. Thus, the evaluation of new GAG-remodeled PGs may reveal interesting aspects with respect to the ability of these cell-surface PGs to modulate in a GAG-defined manner specific signalling pathways of various cell types.

Thus, cells in various stages of their development that express transmembrane PGs would be enzymatically GAG-trimmed to generate a glycocalyx enriched in PGs that carry the residual Ser-linked hexasaccharide presenting the reactive ΔGlcA-tag. Then, the thiosugar-primer would be installed onto the terminal $\Delta_{4,5}$-unsaturation using the radical GAG ligation strategy, and the resulted "primed cell" would be subjected to PG-remodeling using GAG-polymerizing enzymes. Future enzymatic extensions of the HexNAc-S-primed cells should explore the ability of pmHS2 and of other GAG-synthases to effect 'on-cell' GAG-remodeling. These enzymes could be used either separately or in various combinations, and their plasticity (Figure SI 2.30) towards accepting non-canonical or even labelled sugar-donors may also prove informative with regards to dissecting additional, highly intricate cellular mechanisms.[85,93]

2.4 Experimental section

2.4.1 Plasmid DNA details and protein sequences

Plasmid DNA preparations were stored at -20 °C as solutions in H_2O. Amplification of plasmid DNA was carried out using QIAprep spin miniprep kit (Qiagen) according to manufacturer instructions, from overnight LB cultures of cells transformed with the plasmid of interest. Glycerol stocks of cells transformed with plasmids were prepared from overnight LB cultures, *via* dilution of the culture with sterilised 80 % (v/v) glycerol and stored at -80 °C.

The plasmids and proteins utilised in the course of this work that were generously prepared and generated by prior members of the Davis group are as follows:

pmHS2: plasmid prepared by Dr. Arghya Modak, in a pET-15b vector;

*GSSHHHHHHSSGLVPRGSH*MKGKKEMTQIQIAKNPPQHEKENELNTFQNKIDSLKTTLNKDIISQQTLLAKQDSKH
PLSASLENENKLLLKQLQLVLQEFEKIYTYNQALEAKLEKDKQTTSITDLYNEVAKSDLGLVKETNSANPLVSIIMTSHN
TAQFIEASINSLLLQTYKNIEIIIVDDDSSDNTFEIASRIANTTSKVRVFRLNSNLGTYFAKNTGILKSKGDIIFFQDSDDV
CHHERIERCVNILLANKETIAVRCAYSRLAPETQHIIKVNNMDYRLGFITLGMHRKVFQEIGFFNCTTKGSDDEFFHRI
AKYYGKEKIKNLLLPLYYNTMRENSLFTDMVEWIDNHNIIQKMSDTRQHYATLFQAMHNETASHDFKNLFQFPRIY
DALPVPQEMSKLSNPKIPVYINICSIPSRIAQLRRIIGILKNQCDHFHIYLDGYVEIPDFIKNLGNKATVVHCKDKDNSIR
DNGKFILLEELIEKNQDGYYITCDDDIIYPSDYINTMIKKLNEYDDKAVIGLHGILFPSRMTKYFSADRLVYSFYKPLEKD
KAVNVLGTGTVSFRVSLFNQFSLSDFTHSGMADIYFSLLCKKNNILQICISRPANWLTEDNRDSETLYHQYRDNDEQ
QTQLIMENGPWGYSSIYPLVKNHPKFTDLIPCLPFYFL

The *N*-terminus of the enzyme contains a polyhistidine tag which is highlighted in yellow and a thrombin-cleavable peptide which is highlighted in green. The amino acid residues that come from the vector are italicized.

ΔpmHS2: protein donated by Dr. Liang Wu;

*GSHM*FEKIYTYNQALEAKLEKDKQTTSITDLYNEVAKSDLGLVKETNSANPLVSIIMTSHNTAQFIEASINSLLLQTYK
NIEIIIVDDDSSDNTFEIASRIANTTSKVRVFRLNSNLGTYFAKNTGILKSKGDIIFFQDSDDVCHHERIERCVNILLANKE
TIAVRCAYSRLAPETQHIIKVNNMDYRLGFITLGMHRKVFQEIGFFNCTTKGSDDEFFHRIAKYYGKEKIKNLLLPLYY
NTMRENSLFTDMVEWIDNHNIIQKMSDTRQHYATLFQAMHNETASHDFKNLFQFPRIYDALPVPQEMSKLSNPKI
PVYINICSIPSRIAQLRRIIGILKNQCDHFHIYLDGYVEIPDFIKNLGNKATVVHCKDKDNSIRDNGKFILLEELIEKNQDG
YYITCDDDIIYPSDYINTMIKKLNEYDDKAVIGLHGILFPSRMTKYFSADRLVYSFYKPLEKDKAVNVLGTGTVSFRVSL

FNQFSLSDFTHSGMADIYFSLLCKKNNILQICISRPANWLTEDNRDSETLYHQYRDNDEQQTQLIMENGPWGYSSIY PLVKNHPKFTDLIPCLPFYFL

The amino acid residues that come from the vector are italicized.

Cys-TEV-H3: protein donated by Brian Josephson and Yue Yu and was generated from H3 wildtype plasmid (Xenopus laevis C110A) via T4 DNA ligation of the intended insert into the pET-3d vector;

A*C*ENLYFQGTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLV REIAQDFKTDLRFQSSAVMALQEASEAYLVALFEDTNLAAIHAKRVTIMPKDIQLARRIRGERA

The *N*-terminus contains a TEV-cleavable peptide which is highlighted in yellow and a Cys-residue which is highlighted in green.

H3-S10C: protein donated by Dr. Jitka Dadova (Xenopus laevis S10C);

ARTKQTARKCTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIRRYQKSTELLIRKLPFQRLVREIAQDF KTDLRFQSSAVMALQEASEAYLVALFEDTNLAAIHAKRVTIMPKDIQLARRIRGERA
The Ser10 was mutated to a Cys-residue, and this is highlighted in green.

H2B-S36C: protein donated by Dr. Simon Nadal as Dha-variant from the corresponding Cys-containing precursor, amino acid residue highlighted in green (Xenopus laevis S36C, mutant lacking the first three amino acids);[94]

AKSAPAPKKGSKKAVTKTQKKDGKKRRKTRKECYAIYVYKVLKQVHPDTGISSKAMSIMNSFVNDVFERIAGEASRL AHYNKRSTITSREIQTAVRLLLPGELAKHAVSEGTKAVTKYTSAK

2.4.2 Protein extraction/expression and characterisation

2.4.2.1 Decorin variants: intact, GAG-trimmed and core decorin

Isolation of intact decorin from bovine tendon cartilage:

Beef tendons were trimmed of fat and muscle. 65 g of cartilage were chopped finely, and total protein extracted overnight in 1.1 L of 50 mM NaOAc, 4 M Gdn·HCl pH 6.0 (containing 5 mM benzamidine·HCl, 10 mM EDTA, and 5 mM *N*-methylmaleimide) at 4 °C with stirring. The insoluble residue was filtered over gauze and extracted once more in 0.5 L of extraction buffer, overnight at 4 °C. The filtrates were centrifuged for 30 min at 8000 rpm, 4 °C to remove large debris, then passed down a paper filter. All procedures were performed at 4 °C. This second filtrate was further clarified by centrifugation for 20 min at 30,000 rpm, 4 °C and then subjected to vacuum filtration using a 0.45 µm cellulose acetate filter. Final filtrate was concentrated under pressure to 150 mL using an Amicon® stirred cell concentrator equipped with a 10000 MWCO membrane (Amicon PM-10®). The concentrated protein solution was dialyzed into 50 mM NaOAc, 7 M urea, pH 6.0 and purified by weak anionic exchange chromatography (HiTrap® DEAE, GE Healthcare, 5 x 5 mL connected columns, 0 to 1 M NaCl linear gradient). Collected fractions were analysed by denaturing SDS-PAGE analyses. Proteoglycan-containing fractions were confirmed by Alcian blue staining followed by destaining in a solution containing 25 % AcOH, 10 % MeOH, and 65 % ddH_2O.

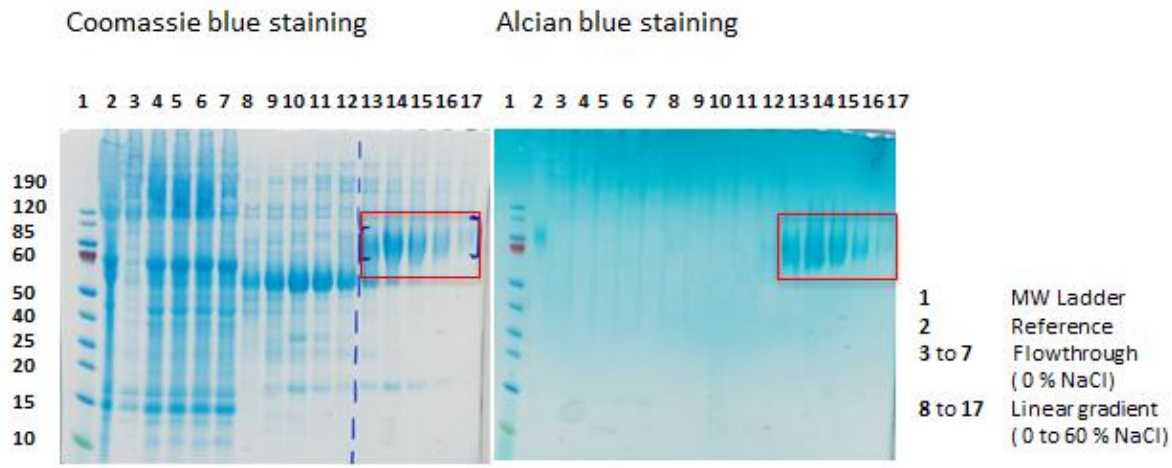


Figure SI 2.1. SDS-PAGE analyses of the intact decorin purification by ionic exchange chromatography. Fractions containing the proteoglycan are shown in red frame after staining with Coomassie blue (left) and Alcian blue (specific for sulfated polysaccharides, right).

Isolation of GAG-trimmed decorin:

The GAG chain was removed with chondroitin ABC lyase (chondroitinase) from *Proteus vulgaris* (Sigma). Intact decorin (25 mL, c = 0.5 mg/mL) was dialyzed into 100 mM TrisHCl, 50 mM NaOAc, pH 8.0 buffer and 1 unit of Chondroitinase ABC was added. The mixture was incubated at 37 °C for 16 hours and completion of the reaction was checked by SDS-PAGE gel analysis. Then, the reaction mixture was concentrated by centrifugation at 3750 rpm, 4 °C using a Vivaspin® 20, (30000 MWCO) concentrator. The obtained mixture (10 mL) was dialyzed against 50 mM NaPi, 500 mM NaCl pH 7.0 buffer and purified using lectin-binding chromatography (HiTrap® ConA 4B column, 5 mL, GE Healthcare). Trimmed decorin was eluted with 300 mM methyl-α-D-mannopyranoside and fractions containing desired glycoprotein were dialyzed against distilled water and lyophilized, yielding 12 mg of pure trimmed decorin.

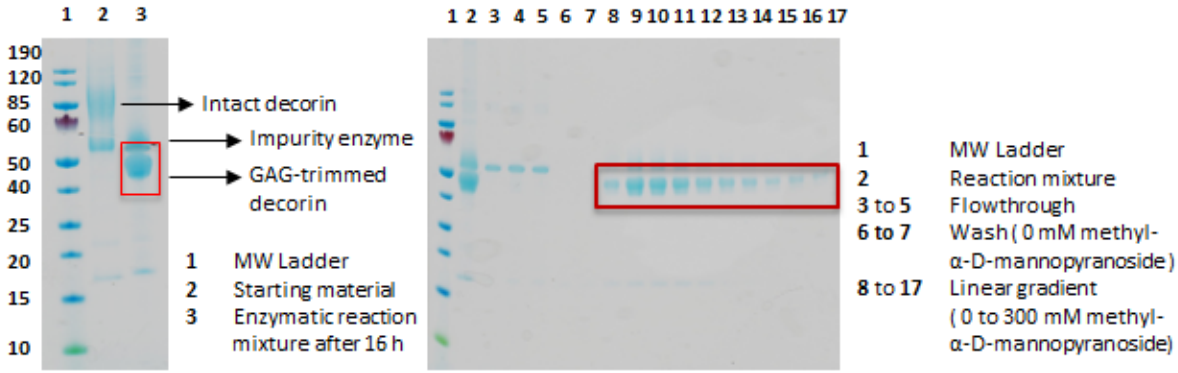


Figure SI 2.2. SDS-PAGE analyses of decorin after trimming with the GAG lyase (left) and after Concavalin A affinity purification (right). Fractions containing the GAG-trimmed decorin are shown in red frame after staining with Coomassie blue.

Isolation of core decorin:

GAG-trimmed decorin was dissolved in 50 mM NaPi pH 6.0 buffer and 6 M urea was added to reach a final 1.6 mg/mL protein concentration and a final 1.2 M urea. To digest 1 mg substrate, 1560 U PNGaseF were added from a 500000 U/mL commercial stock and reaction mixture was stirred for 2 hours at 37 °C and 300 rpm. Then, obtained core decorin was used as such for subsequent thiol-ene addition experiments or further purified by size-exclusion chromatography for characterization by native protein MS and iSCAMS analyses.

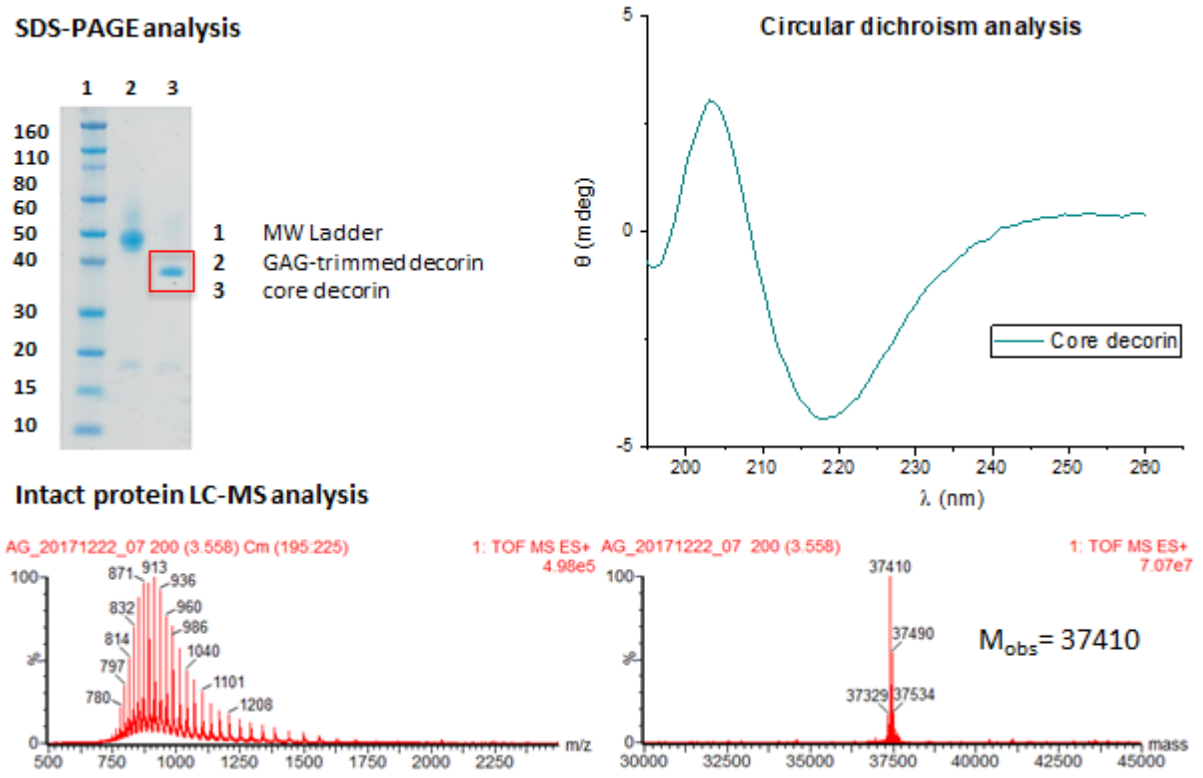
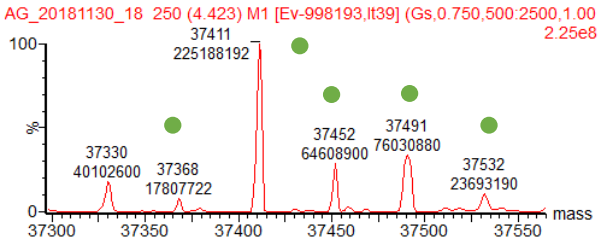


Figure SI 2.3. SDS-PAGE, circular dichroism (CD), and intact protein LC-MS analysis of decorin after digestion of the *N*-glycans by treatment with PNGaseF. SDS-PAGE analysis showed the expected molecular weight shift from approximately 45 kDa to around 37 kDa. CD analysis revealed the characteristic single minimum at 218 nm, indicative of a β-sheet protein structure. The molecular mass of the core decorin is 37410 Da as determined by denaturing, intact protein LC-MS analysis.

Molecular intact mass 37410 was thus assigned to the protein bearing the glycan ΔGlcA-Gal-GlcA-Gal-Gal-Xyl or the decarboxylated (ΔGlcA-44)-GalNAc-GlcA-Gal-Gal-Xyl whereas molecular mass 37490 was attributed to their corresponding sulfated/phosphorylated derivatives. Moreover, the species that appeared to have a molecular mass of 37452 and 37532 were assigned to the decorin forms that carry glycan ΔGlcA-GalNAc-GlcA-Gal-Gal-Xyl and the sulphated/phosphorylated version, respectively. While molecular mass 37368 was attributed to decarboxylated (ΔGlcA-44)-Gal-GlcA-Gal-Gal-Xyl, peak 37330 could not be rationalized. However, since it was not observed in all the extracted batches it might not be decorin-related, but rather a protein contaminant derived from the extraction procedures of the tendons. In addition, decarboxylation of the ΔGlcA as well as replacement of the adjacent GalNAc to a Gal residue could not be identified by MS/MS analysis of the core *O*-glycopeptide and thus -41/-44 difference in the intact protein mass might be a modification occurring at a different site of core decorin. However, the predominant glycoform identified by LC-MS/MS analysis was the one carrying the ΔGlcA-GalNAc-GlcA-Gal-Gal-Xyl hexasaccharide (37455). Therefore, we are also considering the possibility that the molecular peaks 37491 (+39) and 37532 (+39+41) could be generated during the extensive purification procedures which involve the use of buffers containing high concentrations of urea, a denaturing agent that may have caused carbamylation of Lys-residues.[95]

DEA**S**GIGPEEHFPEVPEIEPMGPVCPFRCQCHLRVVQCSDLGLEKVPKDLPPDTALLDLQNNKITEIKDGDFKNL
KNLHTLILINNKISKISPGAFAPLVKLERLYLSKNQLKELPEKMPKTLQELRVHENEITKVRKSVFNGLNQMIVVEL
GTNPLKSSGIENGAFQGMKKLSYIRIADT**N**ITTIPQGLPPSLTELHLDGNKITKVDAASLKGLNNLAKLGLSFNSIS
AVD**N**GSLANTPHLRELHLNNNKLVKVPGGLADHKYIQVVYLHNN**N**ISAIGSNDFCPPGYNTKKASYSGVSLFSN
PVQYWEIQPSTFRCVYVRAAVQLGNYK

S = *O*-glycosylation site
N = *N*-glycosylation site
Protein MW: 36467.02
=> 36461 (with *S-S* bonds)
=> 36464 (the *N*-glycosylation sites become Asp-residues upon PNGaseF treatment
Hexasaccharide molecular weight: ~ 946



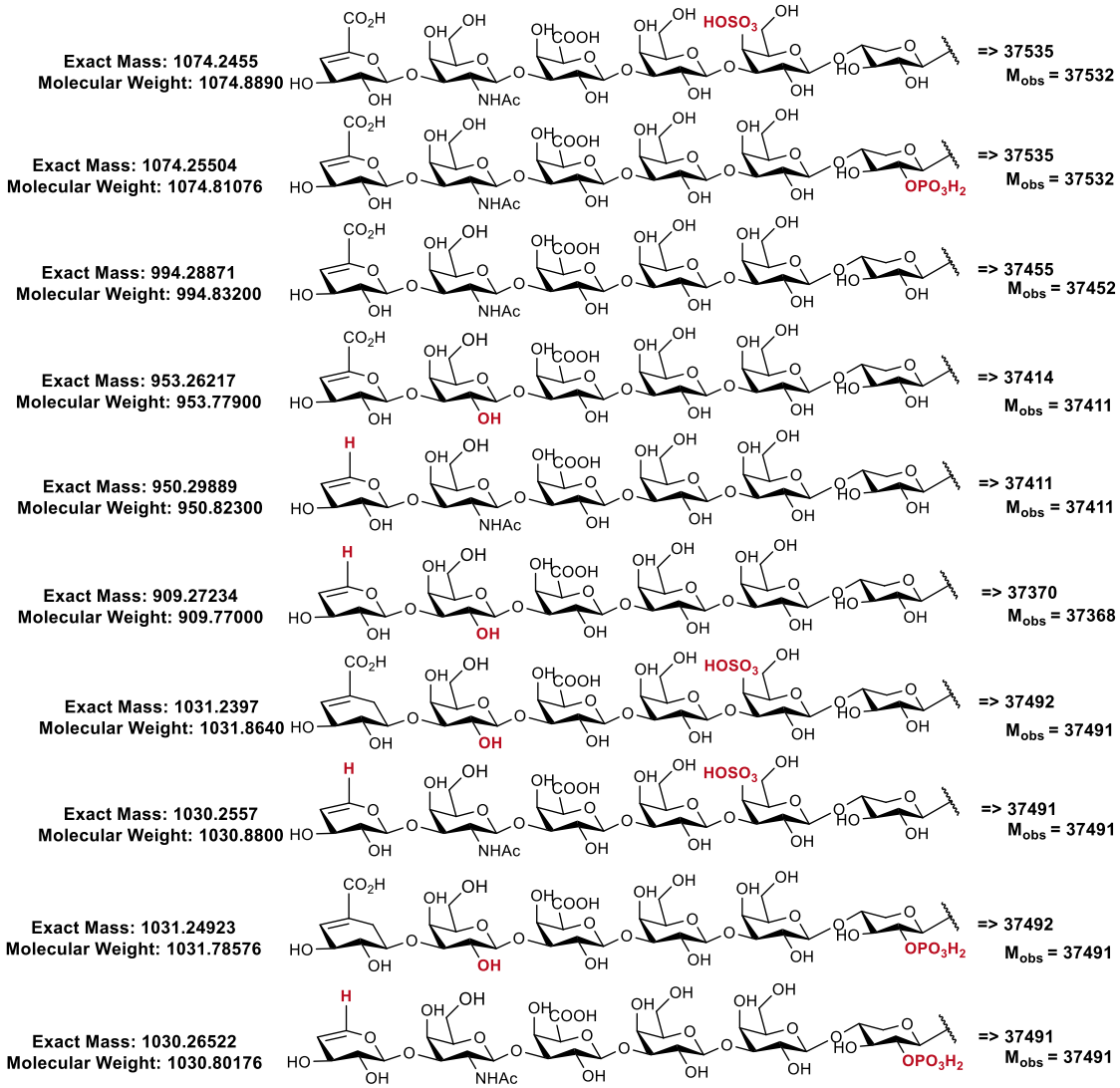


Figure SI 2.4. A rational analysis of the residual Ser-linked hexasaccharide generated by GAG-lyase digestion as suggested by molecular masses observed by intact protein LC-MS analysis.

Characterization of enzymatically digested core decorin by LC-MS/MS analyses:

Enzymatic digestion of decorin with Trypsin:

10 µg of core decorin were denatured in 8 M urea, 100 mM TEAB for 30 min at room temperature (r.t.), 400 rpm and subsequently reduced with 10 mM TCEP for 30 min at r.t., 400 rpm before being alkylated with 50 mM 2-chloroacetamide for 30 min at r.t., 400 rpm, in the dark. Samples were diluted to 0.8 M urea with 50 mM TEAB and digestion was performed with Trypsin (1:20 w/w) O/N at 37 °C, 400 rpm. Digested peptides were desalted using Oasis® HLB cartridges (10 mg) from Waters, eluting with 250 µL 50:50:0.1 MeCN/H_2O/TFA and 250 µL 80:20:0.1 MeCN/H_2O/TFA. Combined fractions were lyophilized and re-dissolved in 50 µL 90:5:5 H_2O/FA/DMSO.

Data acquisition: SCE-HCD experiment on a QExactive MS instrument.

Resulting peptides (2 µL, 1:10 dil.) were separated by nano-flow reversed-phase liquid chromatography Ultimate 3000 UHPLC system (Thermo Fisher Scientific) coupled to a Q Exactive Hybrid Quadrupole-Orbitrap mass spectrometer (Thermo Fischer Scientific). The peptides were loaded on a C18 PepMap100 precolumn (inner diameter 300 µm, 5 mm, 3 µm C18 beads; Thermo Fisher Scientific) and separated on an in-house packed analytical column (inner diameter 75 µm x 50 cm packed with ReproSil-Pur 120 C18-AQ, 1.9 µm, 120 Å, Dr. Maisch GmbH). Separation of peptides was conducted with a first step linear gradient from 15 % to 35 % of B for 30 min followed by a second step from 35 % to 55 % of B for additional 15 min, at a flow rate of 200 nL/min (A: 0.1 % formic acid, B: 0.1 % formic acid in acetonitrile). The raw data were acquired on the mass spectrometer in data-dependent mode. Automatically switching from MS to higher energy collision induced dissociation MS/MS. Full-scan spectra were acquired in the Orbitrap [scan range 350-2000 m/z, resolution 70000, automatic gain control (AGC) target 3 x 10^6, maximum injection time 50 ms]. After the MS scan, the top 10 most intense peaks were selected for HCD fragmentation in a 2 stepped-collision energy 10 %-15 % of normalised collision energy. HCD spectra were also acquired in the Orbitrap (resolution 17500; AGC target 5 x 10^4; maximum injection time 120 ms), with first fixed mass at 180 m/z. Full scan MS spectra (scan range 350-1500 m/z, resolution 120000, AGC target $1e^6$, maximum injection time 250 ms) and subsequent HCD MS/MS spectra (resolution 15000, AGC target $3e^4$, maximum injection time 500 ms) of 5 most intense peaks were acquired in the Orbitrap. HCD fragmentation was performed at 10 % and 15 % of normalized collision energy and the signal intensity threshold was kept at 500 counts.

Data analysis:

Generated raw MS/MS data files were processed using the Byonic™ software 2.9.30 version and the search was performed against a focused protein database containing the glycoprotein sequence and the identified contaminants. Using Byonic's ProteinLabel keyword, glycosylation was only allowed on the glycoprotein of interest. For all data acquired in the Orbitrap (Q Exactive), a 20 ppm mass error was set. Carbamidomethylation was set as a fixed modification while oxidation (Met), deamidation (Asn, Gln), carbamidomethylation (peptide N-terminus, Lys), phosphorylation (Ser, Thr) and oxidation (Pro) were set as variable modifications. One missed cleavage was allowed, and the data was searched with trypsin as enzyme (R, K) and semi-specific cleavage. The following database of *O*-glycans was manually curated (based on the findings of Klein *et al.*[75]) and used for the identification of the residual Ser-linked hexasaccharide: ΔGlcA-GalNAc-GlcA-Gal-Gal-Xyl.[96] Note that the ΔGlcA-residue generated by the chondroitinase digestion is defined as ΔGlcA = Glc(A) -18.0106.

Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1
Xyl(1)Hex(2)dGlcA(1) / +614,16888 @ S | rare1
Xyl(1) @ OGlycan | rare1
Xyl(1)Phospho(1) @ OGlycan | rare1
Xyl(1)Hex(1) @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(1) @ OGlycan | rare1
Xyl(1)Hex(1)Sulfo(1) @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(1)Sulfo(1) @ OGlycan | rare1
Xyl(1)Hex(2) @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2) @ OGlycan | rare1
Xyl(1)Hex(2)Sulfo(1) @ OGlycan | rare1
Xyl(1)Hex(2)Sulfo(2) @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)Sulfo(1) @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)Sulfo(2) @ OGlycan | rare1
Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1
Xyl(1)Hex(2)Sulfo(1)GlcA(1) @ OGlycan | rare1
Xyl(1)Hex(2)Sulfo(2)GlcA(1) @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(1) @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)Sulfo(2)GlcA(1) @ OGlycan | rare1
Xyl(1)Hex(2)GlcA(1)HexNAc(1) @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)GlcA(1)HexNAc(1) @ OGlycan | rare1

Xyl(1)Hex(2)Sulfo(1)GlcA(1)HexNAc(1) @ OGlycan | rare1

Xyl(1)Hex(2)Sulfo(2)GlcA(1)HexNAc(1) @ OGlycan | rare1

Xyl(1)Hex(2)Sulfo(3)GlcA(1)HexNAc(1) @ OGlycan | rare1

Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(1)HexNAc(1) @ OGlycan | rare1

Xyl(1)Phospho(1)Hex(2)Sulfo(2)GlcA(1)HexNAc(1) @ OGlycan | rare1

Xyl(1)Phospho(1)Hex(2)Sulfo(3)GlcA(1)HexNAc(1) @ OGlycan | rare1

Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1

Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1

Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1

Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1

Xyl(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1

Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1

Xyl(1)Phospho(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1

Xyl(1)Phospho(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1

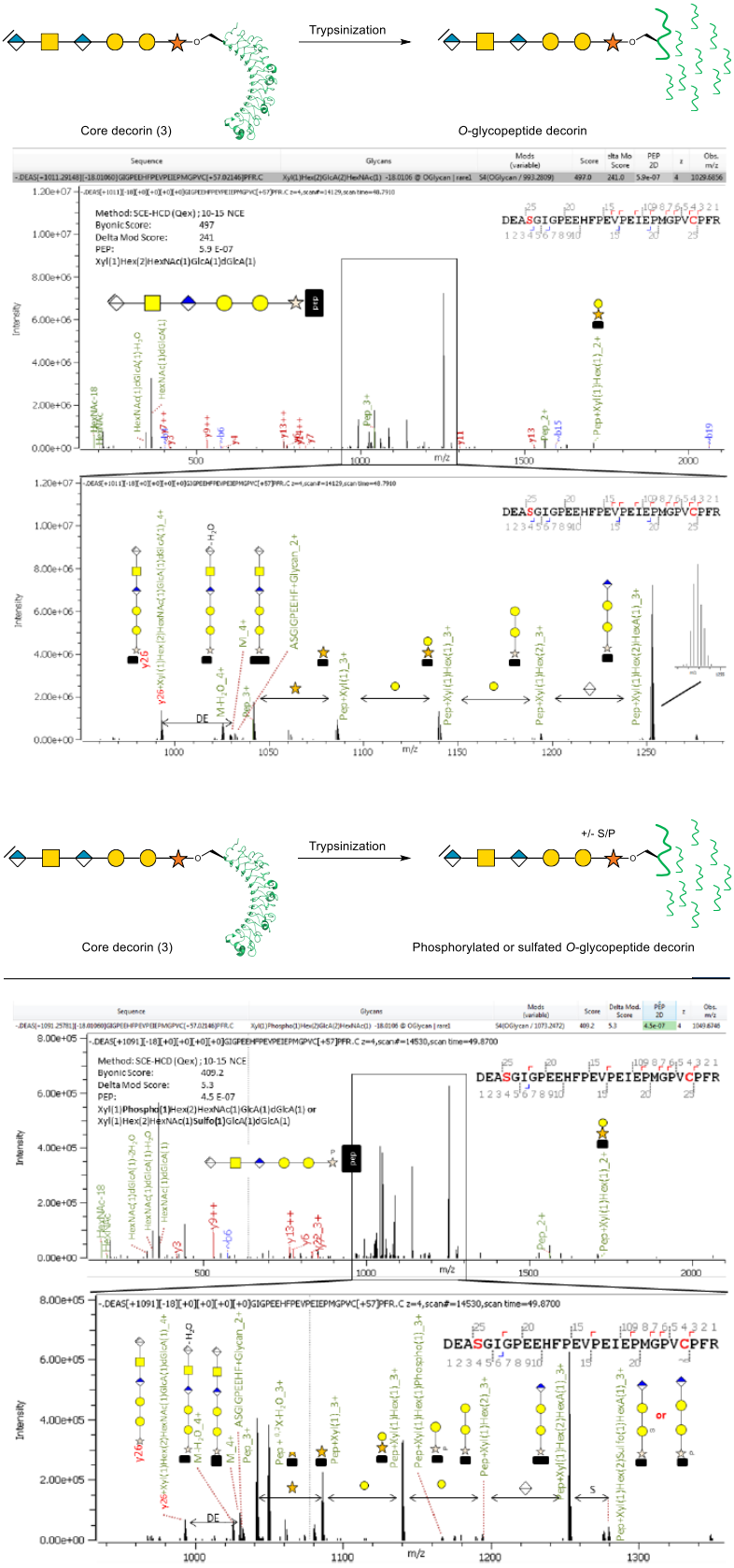


Figure SI 2.5. LC-MS/MS analysis of core decorin after digestion with trypsin. MS/MS glycan fragmentation confirmed the sequence of the residual *O*-linked hexasaccharide ΔGlcA-GalNAc-GlcA-Gal-Gal-Xyl upon GAG-trimming. Moreover, the *O*-glycopeptides attached to the Ser34 revealed phosphorylation of the Xyl residue or sulfation of the innermost Gal residue.

Characterization of intact, GAG-trimmed and core decorin by native MS and/or iSCAMS analyses:

After dialysis in 100 mM TrisHCl, 50 mM NaOAc pH 8.0 buffer, intact decorin was purified by size-exclusion chromatography using a Superdex 200 16/60 column pre-equilibrated with the same buffer and eluted protein (c = 0.2 mg/mL) was diluted with PBS buffer to 50 nM for iSCAMS analysis.
After ConA-affinity chromatography, GAG-trimmed decorin was purified by size-exclusion chromatography using a Superdex 75 10/300 column pre-equilibrated with 200 mM NH_4OAc pH 7.0 buffer and eluted protein was analysed by native MS analyses (c = 0.07 mg/mL) and diluted with PBS buffer to 50 nM for iSCAMS analysis.
After treatment with PNGaseF, core decorin was purified by size-exclusion chromatography using a Superdex 75 10/300 column pre-equilibrated with 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer and eluted protein was diluted with 200 mM NH_4OAc pH 7.0 buffer (c = 0.07 mg/mL) for native MS analysis and with PBS buffer to 50 nM for iSCAMS analysis.

iSCAMS:

Mass photometry measurements were done as described.[100] Briefly, borosilicate microscope coverslips were cleaned by sequential washes in Milli-Q H_2O, isopropanol and Milli-Q H_2O and dried under a nitrogen gas flow. Sample chambers were assembled using silicone gaskets (CultureWell™ reusable gasket, Grace Bio-Labs). Assembled Coverslips were placed on the sample stage and a single gasket was filled with 10 µL phosphate buffered saline (PBS) to find focus. Intact decorin, GAG-trimmed and core protein were diluted to 50 nM in PBS immediately prior to analysis. For each measurement, 10 µL sample was added to the 10 µL PBS in a single gasket. Data were recorded for 90 seconds. MP settings were as follows: frame averaging = 10, frame rate = 955 Hz and 4x4 pixel binning. Data was analysed using Discovery MP v2.3dev7 (Refeyn Ltd).

Native MS:

The glycoprotein sample was buffer exchanged to 200 mM ammonium acetate buffer (pH 7.0) Zeba Spin desalting column (7K MWCO, Thermo Fisher)) and loaded into a gold coated needle prepared in-house and analysed on a Q-Exactive UHMR mass spectrometer (Thermo Fisher Scientific). The typical MS setttings were spray voltage of 1.2 kV, source fragmentation of 50 V, source temperature of 100-150 degree. HCD collision energy of 0-100 V, and resolution of 17500 at m/z 200. Backing pressure was maintained at ~5*10-10 mbar. The native MS data were analyzed using Xcalibur 4.0

2.4.2.2 pmHS2

Transformation and expression in BL21(DE3) competent cells

BL21(DE3) cells were thawed on ice. A 20 µL aliquot was transferred to a pre-chilled 14 mL falcon tube. 1 µL β-mercaptoethanol was added and the cells were incubated on ice for 10 minutes. Plasmid pET-15b-pmHS2 encoding for the *N*-His_6-tagged enzyme (85 ng) was added and the tube placed on ice for 30 minutes with occasional swirling. Cells were heat shocked (42 ^oC, 45 s) then immediately placed on ice for a further 2 minutes. 200 µL of SOC media was added and cells were incubated at 37 ^oC for 1 h with shaking (200 rpm). Cells were plated on Luria-Bertani agar plates containing carbenicillin (100 µg/mL). Plates were incubated at 37 ^oC overnight.

A single colony obtained from the above transformation was used to inoculate 10 mL Luria-Bertani media, supplemented with carbenicillin (100 µg/mL) and cells were grown at 37 ^oC, overnight, with shaking (200 rpm). Resultant cells were used to inoculate 1 L Luria-Bertani media containing the same antibiotic and grown at 37 °C, 200 rpm until OD_{600} = 0.8. IPTG was added to a final concentration of 0.2 mM and the flask was incubated at 20 °C, 180 rpm, for 20 h. Cells were harvested by centrifugation (10 min, 8000 rpm, 4 °C) and the pellet (~ 4.7 g/1 L, wet) was stored at -80 °C.

Cell lysis and purification

Frozen cells (2 x 1L culture) were thawed on ice and re-suspended in 50 mL loading buffer (25 mM TrisHCl, 500 mM NaCl, 30 mM imidazole, pH 7.2). 10 mg DNAase, 50 mg lysozyme and cOmplete™ EDTA-free protease inhibitor (one tablet) and the mixture was gently shaken at 4 °C for 3 hours. Cells were then sonicated (40 % amplitude, 5 s on, 5 s off, 5 x 1 min cycles) and centrifuged at 20000 rpm and 4 °C for 10 min.

The supernatant was filtered through a 0.4 µm cellulose acetate membrane filter and loaded onto a prepacked 5 mL HisTrap™ Ni^{2+}-NTA column pre-equilibrated with loading buffer. The protein was purified at a 4 mL/min flow, using a step-wise gradient: 0 % for 7 CV, 25 % for 5 CV, 50 % for 6 CV and 100 % elution buffer (25 mM TrisHCl, 500 mM NaCl, 250 mM imidazole, pH 7.2) for 8 CV. Collected fractions were analysed by SDS-PAGE and the ones containing pure protein were combined and dialyzed against 25 mM TrisHCl, 150 mM NaCl, 10 % glycerol, pH 7.4. Desired protein (26 mL, c = 1.09 mg/mL) was obtained (~14.2 mg/1L culture), flash-frozen and stored at – 80 °C until further use.

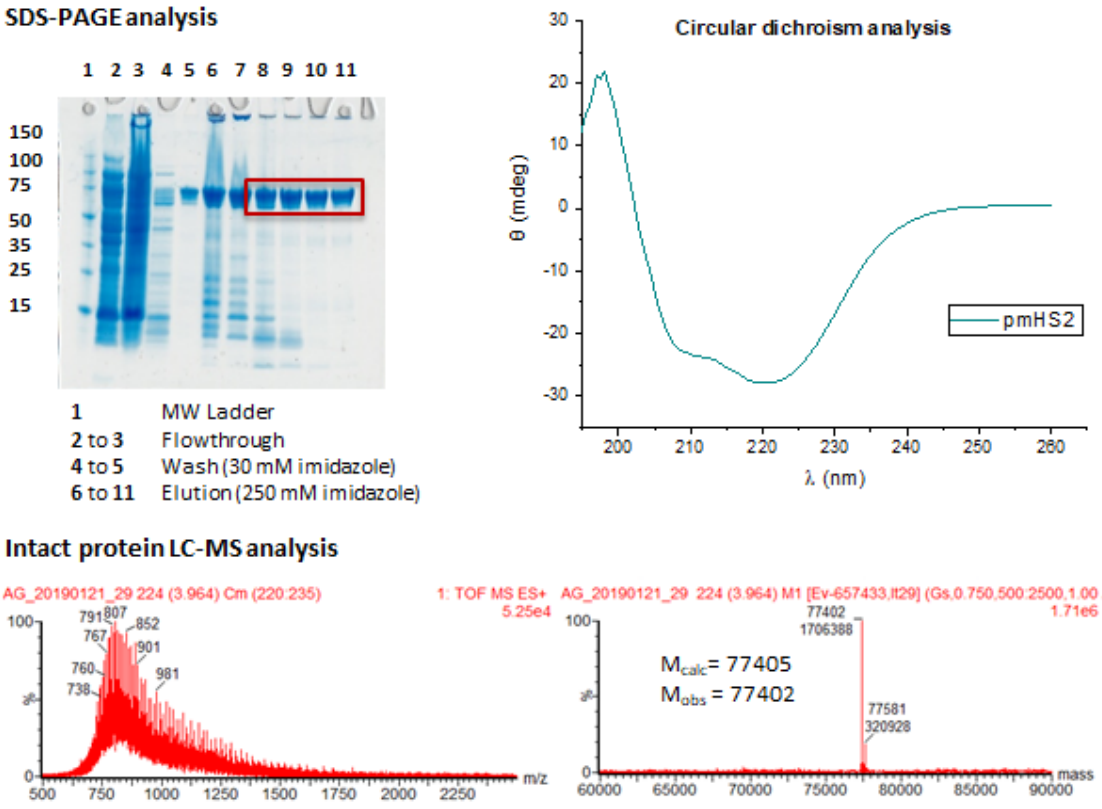


Figure SI 2.6. SDS-PAGE analysis of the pmHS2 purification by Ni^{2+}-affinity chromatography and characterization of the protein by circular dichroism (CD) analysis and intact protein LC-MS. As expected from the calculated molecular weight, the size of the purified enzyme was about 75 kDa. Fractions containing pmHS2 are shown in red frame after staining with Coomassie blue. The recorded CD spectrum showed the characteristic double minimum at 208 nm and 222 nm, indicative of an α-helical protein structure. The molecular mass of the enzyme is 77402 Da as determined by denaturing, intact protein LC-MS analysis (77581 Da peak is due to gluconoylation of *N*-terminal His_6-tagged proteins[97]).

Differential Scanning Fluorimetry (DSF) stability assay

21 µL of each buffer were transferred into a PCR-microplate. This was spun down for 30 seconds and placed on ice. To each well, 2 µL of protein solution (pmHS2 c = 1.1 mg/mL) were added. A fresh SYPRO Orange solution (Invitrogen, S6651, 5000X) was prepared by diluting 3 µL of 5000X DMSO stock in 237 µL of water. To each well, 2 µL of diluted SYPRO Orange solution were dispensed. The final working concentration of the dye was 5X. The PCR-plate was sealed with a clear adhesive seal and spun down for 30 seconds. Then, the microplate was placed in the Bio-Rad RT-PCR machine pre-equilibrated at the desired temperature (20 °C). A temperature gradient of 0.2°C/5 seconds, 20 °C to 95 °C was programmed and a pair of filters compatible with the maximum excitation and emission wavelengths of SYPRO Orange were chosen. Data analysis was performed using CFX Manager Bio-Rad software.

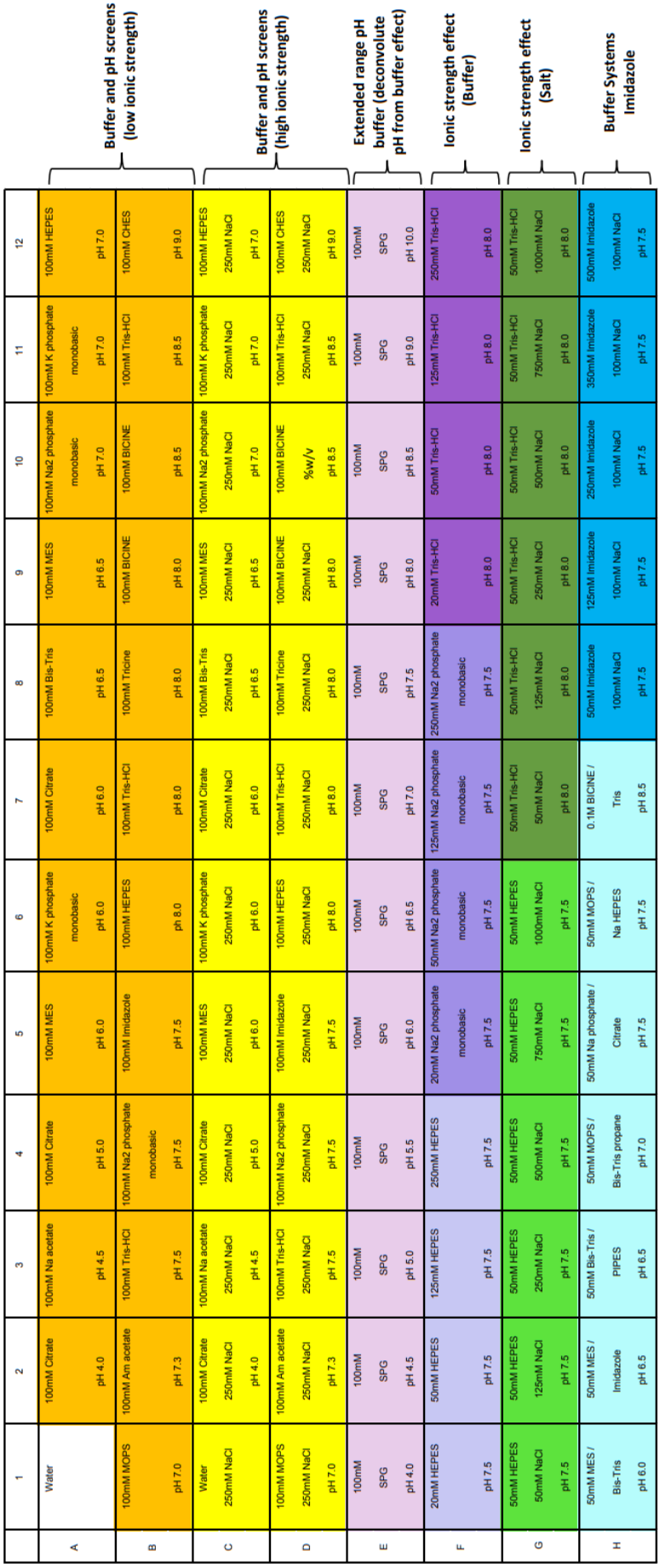


Figure SI 2.7. The 96-well plate RUBIC® buffer screen kit was used to assess the stability of pmHS2.

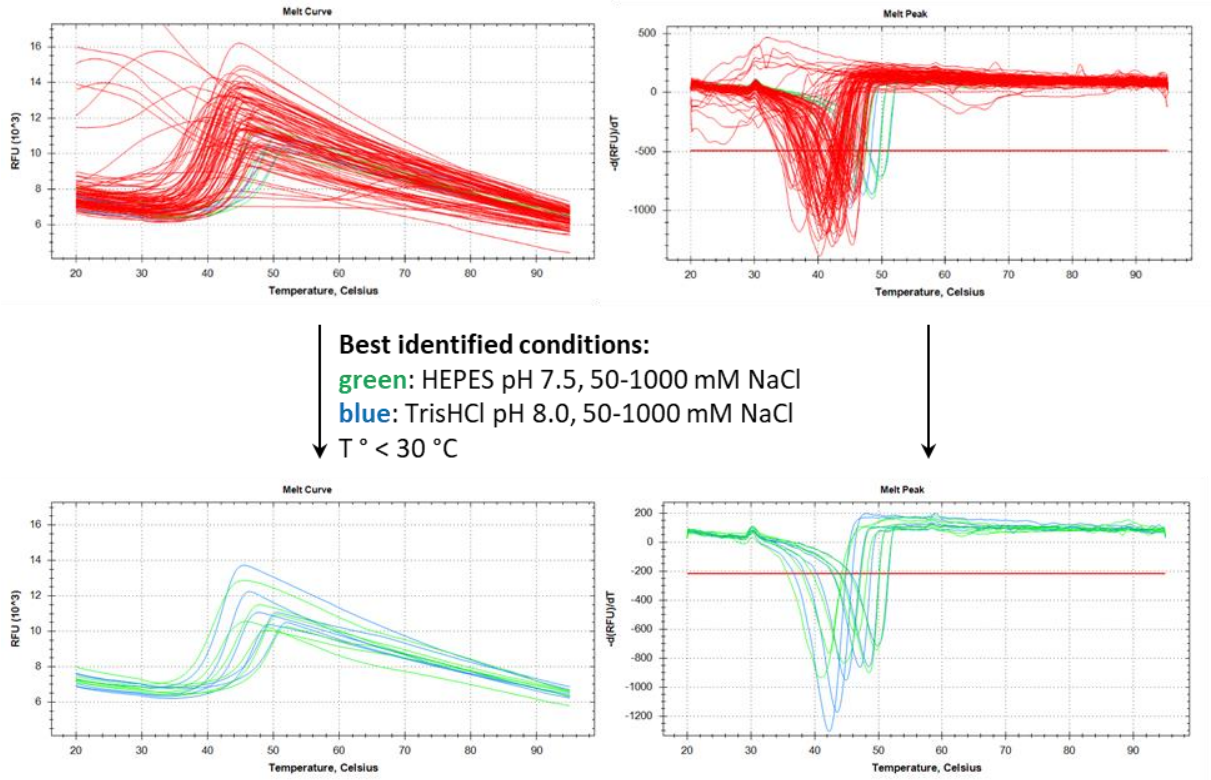


Figure SI 2.8. A typical thermal denaturation assay using the Differential Scanning Fluorimetry (DSF) technique showed that pmHS2 is most stable at temperatures lower than 30 °C, in 20 mM HEPES pH 7.5 buffer and 50 mM TrisHCl pH 8.0 buffer, both supplemented with various concentrations of NaCl.

2.4.2.3 Cys-TEV-H3

Transformation and expression in BL21(DE3)pLysS competent cells

BL21(DE3)pLysS cells were thawed on ice. A 75 µL aliquot was transferred to a pre-chilled 14 mL falcon tube. Plasmid pET-3d-Cys-TEV-H3 (75-100 ng) was added and the tube placed on ice for 30 minutes with occasional swirling. Cells were heat shocked (42 °C, 40 s) then immediately placed on ice for a further 2 minutes. 900 µL of SOC media was added, and cells incubated at 37 °C for 1 h with shaking (250 rpm). Cells were plated on Luria-Bertani agar plates containing both chloramphenicol (25 µg/mL) and ampicillin (100 µg/mL). Plates were incubated at 37 °C overnight.

A single colony obtained from the above transformation was used to inoculate 5 mL 2xTY media, supplemented with ampicillin (100 µg/mL) and chloramphenicol (25 µg/mL). Cells were grown at 37 °C overnight with shaking (200 rpm). Resultant cells were used to inoculate 500 mL 2xTY media containing the same antibiotics and grown at 37 °C until OD_{600} = 0.4-0.9. IPTG was added to a final concentration of 0.5 mM and the flask was shaken at 37 °C and 250 rpm for 2 h. Cells were harvested by centrifugation and the pellet was resuspended in lysis buffer (50 mM Tris pH 7.5, 100 mm NaCl) containing

cOmplete™ EDTA-free protease inhibitor (one tablet per 50 mL). Cells were frozen and stored at -80 °C until further use.

Cell lysis and purification

Frozen cells were thawed in a water bath and DNAase was added (approx. 1 mg per 0.5 L culture). Cells were then sonicated (40 % amplitude, 30 s on, 30 s off, 5 cycles) and centrifuged at 20000 rpm and 4 °C for 20 min. The supernatant was discarded, and the pellet was resuspended in 40 mL TW buffer (above lysis buffer + 1 % Triton-X). Sonication was repeated at 40 % amplitude for 30 s, and the suspension centrifuged at 20000 rpm and 4 °C for 10 min. The pellet was washed two more times with TW buffer and once with lysis buffer. 1 mL DMSO was added to the resulted pellet and after being minced with a spatula it was incubated at room temperature for 10 min. 15 mL unfolding buffer (7M guanidine HCl, 20 mM Tris pH 7.5, 10 mM DTT) were then added and the mixture was gently shaken for 1 hour at room temperature. It was then centrifuged at 20000 rpm and 25 °C for 10 min.

The supernatant was collected and loaded onto an HiLoad™ 26/60 Superdex 200 size-exclusion column pre-equilibrated with 1 column volume of SAU-1000 buffer (7 M urea, 20 mM NaOAc pH 5.2, 1 M NaCl, 1 mM EDTA, 10 mM DTT, 1mM benzamidine). The protein was eluted with SAU-1000 buffer and analysed by SDS-PAGE.

Fractions containing purified protein were concentrated in a Vivaspin (MWCO 10kDa) concentrator and then loaded onto a 5 mL HiTrap SP HP column equilibrated with SAU-100 buffer (7 M urea, 20 mM NaOAc pH 5.2, 0.1 M NaCl, 1 mM EDTA, 10 mM DTT, 1mM benzamidine). Protein was eluted using a linear gradient from 0 % to 70 % SAU-1000 buffer over 20 CV. Pure fractions were collected, dialyzed against deionized water containing 2 mM β-mercaptoethanol and lyophilized to obtain porous white powder (~50 mg/0.5 L).

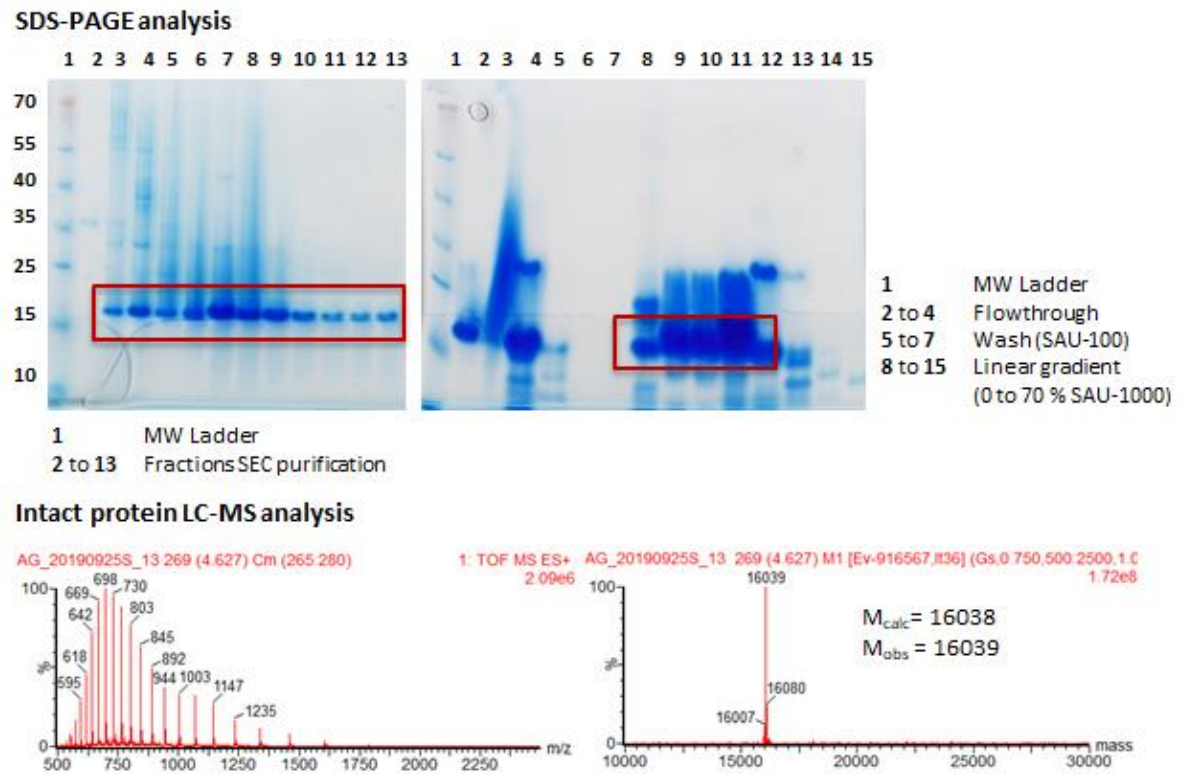
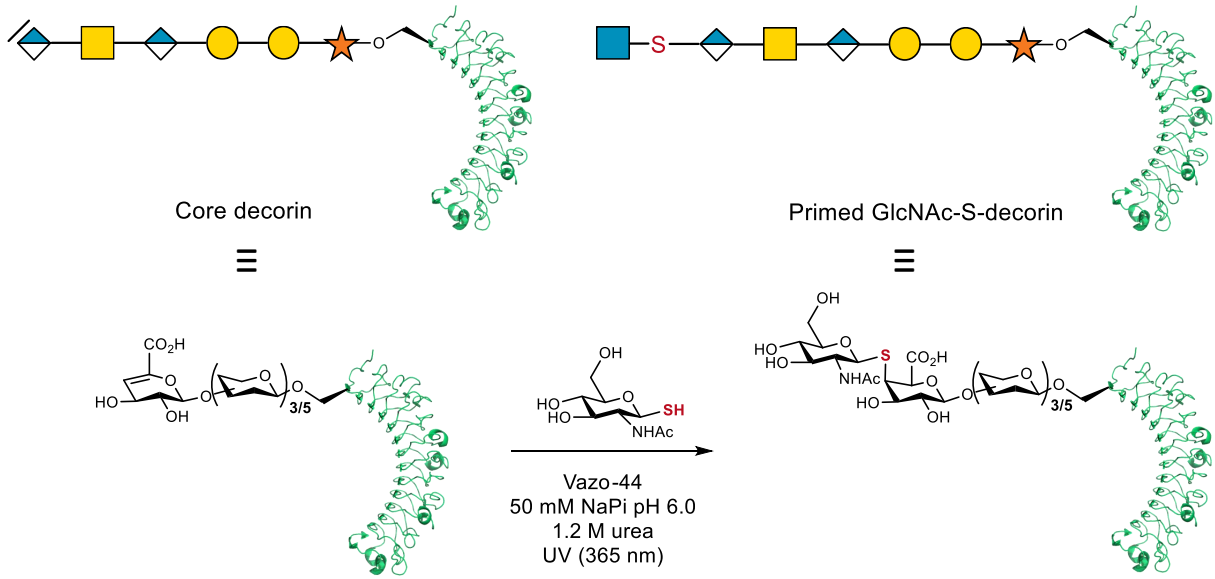


Figure SI 2.8. SDS-PAGE analyses of the Cys-TEV-H3 purification by size-exclusion chromatography (left) and cation-exchange chromatography (right), and characterization of the protein by intact protein LC-MS analysis. As expected from the calculated molecular weight, the size of the purified protein was about 15 kDa. Fractions containing Cys-TEV-H3 are shown in red frame after staining with Coomassie blue. The molecular mass of the protein is 16039 Da as determined by denaturing, intact protein LC-MS analysis (16080 Da adduct is attributable to carbamylation of Lys-residues, a known modification that occurs in buffers containing high concentrations of urea[95]).

2.4.3 Development of a free-radical hydrothiolation reaction to ligate unprotected glycans to the core decorin

2.4.3.1 Optimization of the GlcNAc-addition to the core decorin



Lyophilized GAG-trimmed decorin was dissolved in 50 mM NaPi buffer pH 6.0 and 1.2 M urea in a 1.5 mL Eppendorf tube to have a final concentration of c = 1.6 mg/mL. PNGaseF was added (1560 U/mg substrate, from a 500000 U/mL commercial stock) and reaction mixture was stirred for 2 h at 37 °C and 300 rpm. Then, the protein solution was added to the GlcNAc-SH and obtained mixture was transferred into a Chromacol® clear vial (03-FISV) containing the corresponding amount of radical initiator (Vazo-44). The mixture was irradiated with a 365 nm light source for 60 minutes and analysed by intact protein LC-MS as follows:

DTT monitoring: 2 µL sample were diluted to 64 µg/mL in 100 mM NH_4OAc pH 8.0 buffer and 10 mM (Table 1, entries 1 and 2) or 100 mM DTT (Table 1, entries 3 and 4);

No DTT monitoring: 2 µL sample were diluted to 64 µg/mL in 100 mM NH_4OAc pH 8.0 buffer.

Table 1, Entry 1: 0.1 M Vazo-44, 0.5 M GlcNAc-SH, 125 W UV lamp, 30 min

25 µL protein solution (c = 1.6 mg/mL) were added to 2.9 mg GlcNAc-SH (500 mM) and then to 0.8 mg Vazo-44 (100 mM). Reaction mixture was UV-irradiated for 30 minutes with a medium-pressure Hg-lamp (125 W) and progression was monitored by LC-MS analysis after 10, 20, and 30 minutes. However, after 30 minutes LC-MS analyses showed significant protein degradation.

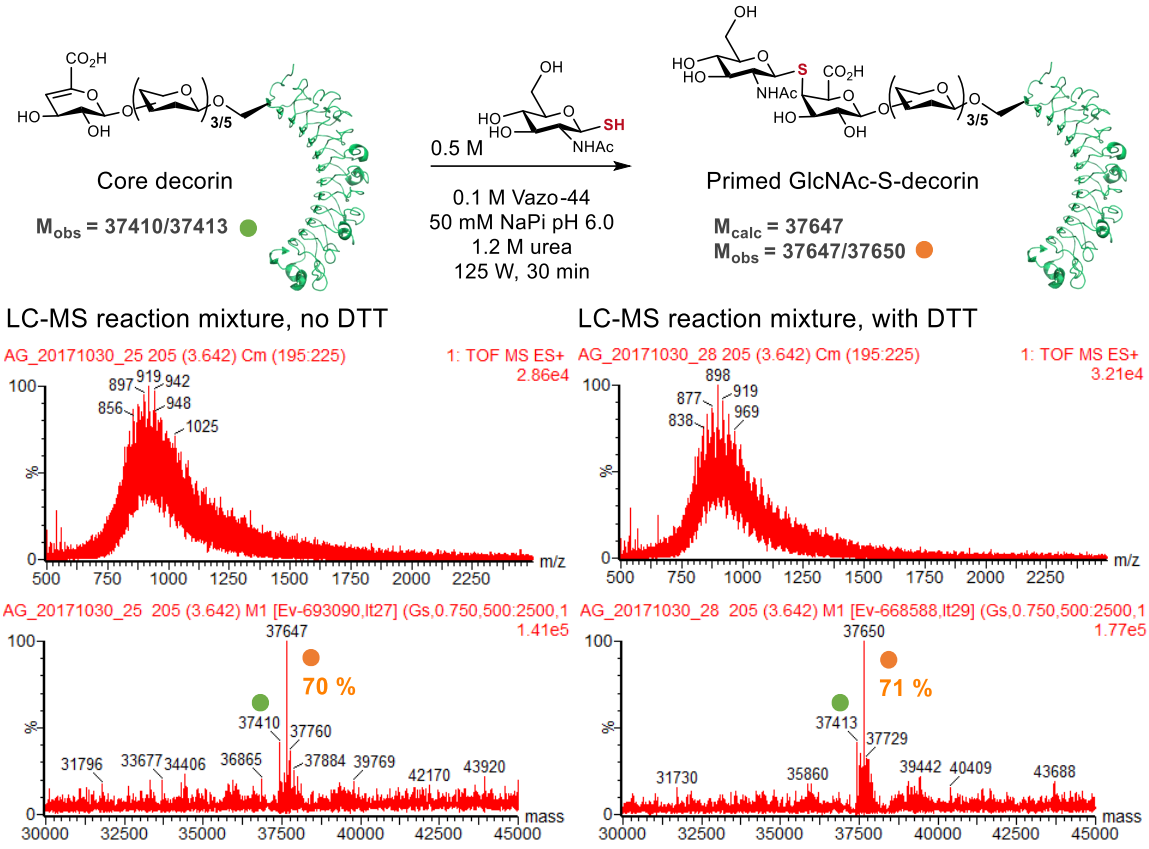


Figure SI 2.10. Intact protein LC-MS analyses of the reaction performed using 0.1 M Vazo-44, 0.5 M GlcNAc-SH, and the 125 W UV lamp for 30 minutes. Good conversion to the desired GlcNAc-S-decorin was achieved: 70 % in the LC-MS sample not treated with DTT and 71 % in the LC-MS sample treated with DTT. However, significant protein degradation could also be observed.

Table 1, Entry 2: 0.1 M Vazo-44, 0.5 M GlcNAc-SH, 3 W UV torch, 60 min

50 µL protein solution (c = 1.6 mg/mL) were added to 5.8 mg GlcNAc-SH (500 mM) and then to 1.6 mg Vazo-44 (100 mM). Reaction mixture was UV-irradiated for 1 hour with a 365 nm UV LED torch (3 W) and progression was monitored by LC-MS analysis after 10, 20, 30 and 60 minutes.

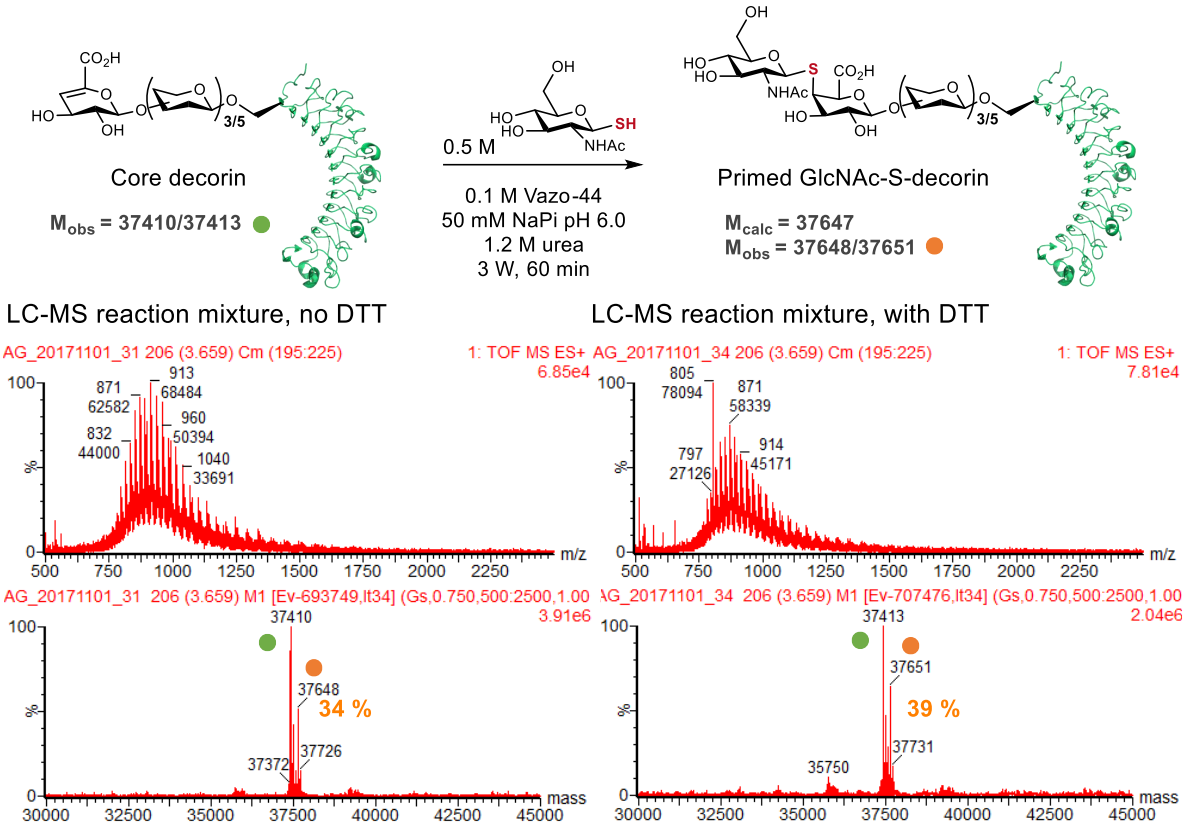


Figure SI 2.11. Intact protein LC-MS analyses of the reaction performed using 0.1 M Vazo-44, 0.5 M GlcNAc-SH, and the 3 W UV torch for 60 minutes. Moderate conversion to the desired GlcNAc-S-decorin was achieved: 34 % in the LC-MS sample not treated with DTT and 39 % in the LC-MS sample treated with DTT.

Table 1, Entry 3: 0.2 M Vazo-44, 1 M GlcNAc-SH, 3 W UV torch, 60 min, (50 µL scale)

50 µL protein solution (c = 1.6 mg/mL) were added to 11.6 mg GlcNAc-SH (1000 mM) and then to 3.2 mg Vazo-44 (200 mM). Reaction mixture was UV-irradiated for 1 hour with a 365 nm UV LED torch (3 W) and progression was monitored by LC-MS analysis after 5, 10, 15, 20, 25, 30, 40, 50 and 60 minutes.

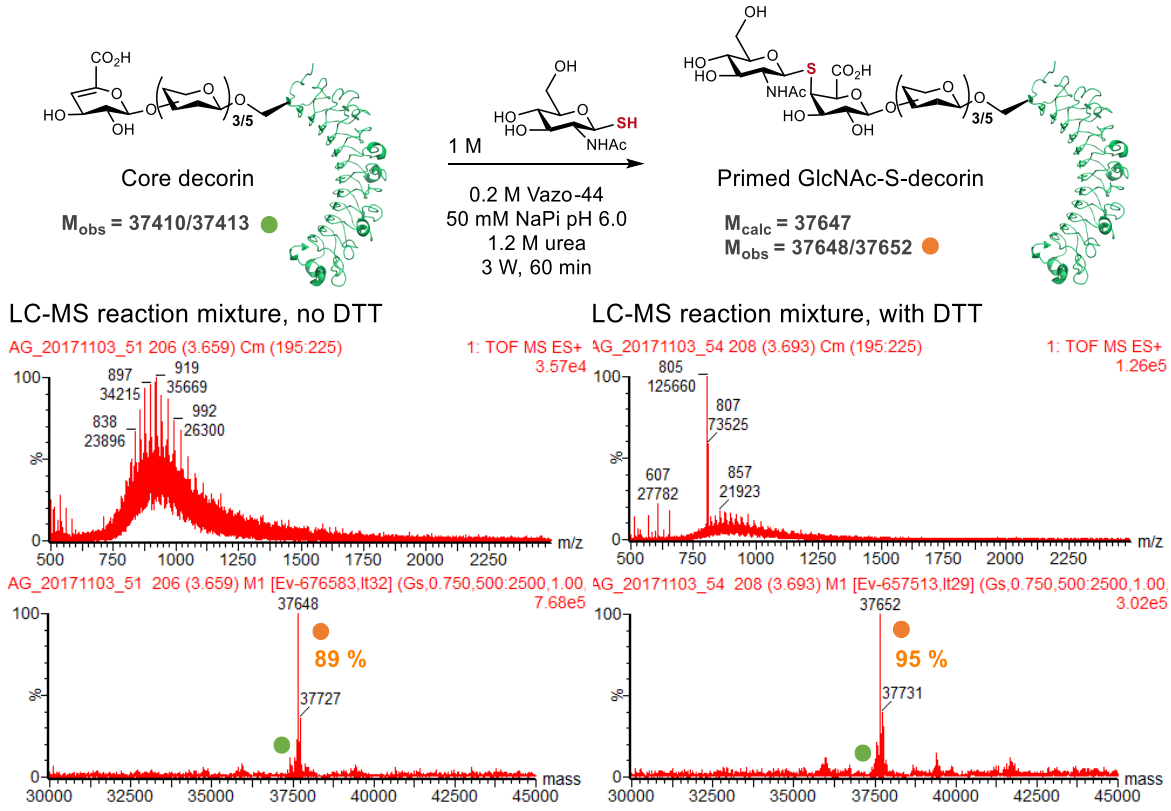


Figure SI 2.12. Intact protein LC-MS analyses of the reaction performed using 0.2 M Vazo-44, 1 M GlcNAc-SH, and the 3 W UV torch for 60 minutes. Excellent conversion to the desired GlcNAc-S-decorin was achieved: 89 % in the LC-MS sample not treated with DTT and 95 % in the LC-MS sample treated with DTT, and the protein stability was preserved.

Table 1, Entry 4: 0.2 M Vazo-44, 1 M GlcNAc-SH, 3 W UV torch, 60 min, (250 µL scale)

250 µL protein solution (c = 1.6 mg/mL) were added to 58 mg GlcNAc-SH (1000 mM) and then to 16 mg Vazo-44 (200 mM). Reaction mixture was UV-irradiated for 1 hour with a 365 nm UV LED torch (3 W) and progression was monitored by LC-MS analysis after 30 and 60 minutes.

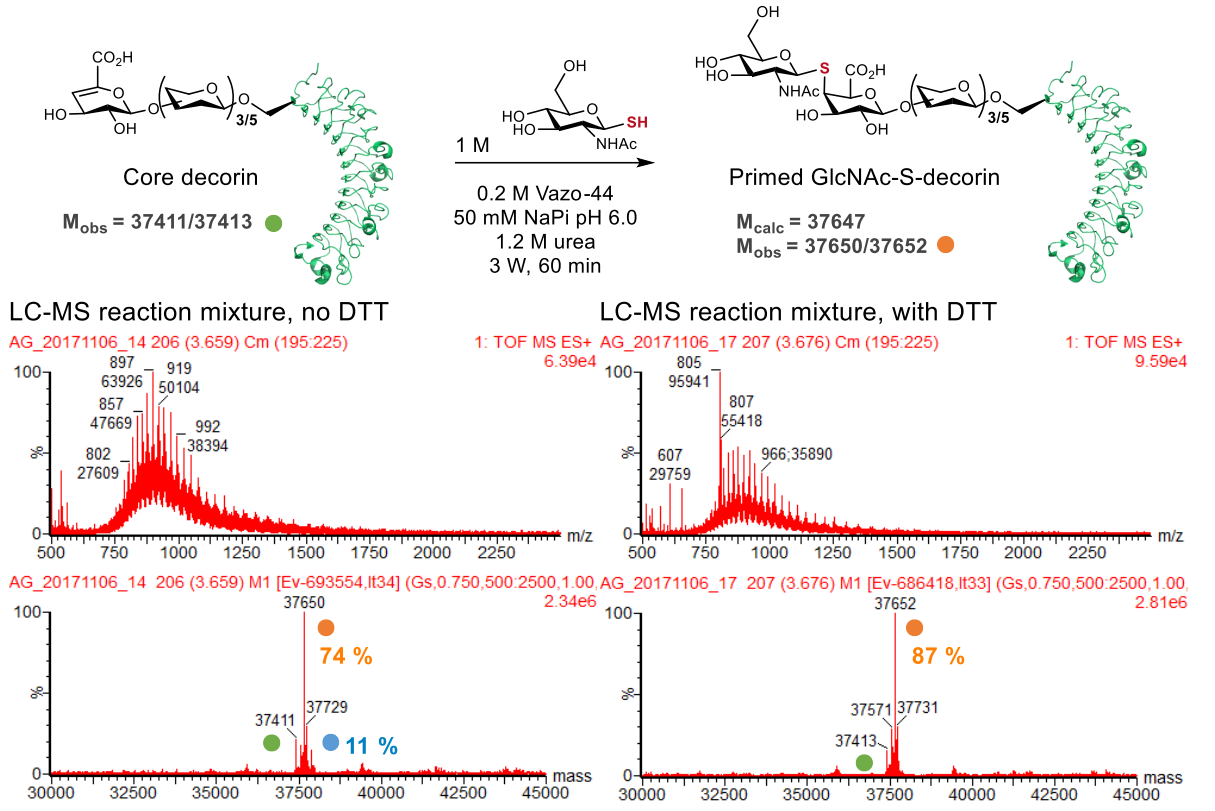


Figure SI 2.13. **Intact protein LC-MS analyses showed that excellent conversion to desired GlcNAc-S-decorin was also achieved when reaction was performed on a 250 µL scale (74 % in the samples not treated with DTT and 87 % in the samples treated with DTT).** A double GlcNAc-S-addition product could be observed (11 %) but the latter disappeared upon reduction with DTT.

2.4.3.2 Negative control reactions for the thiyl-glycosylation

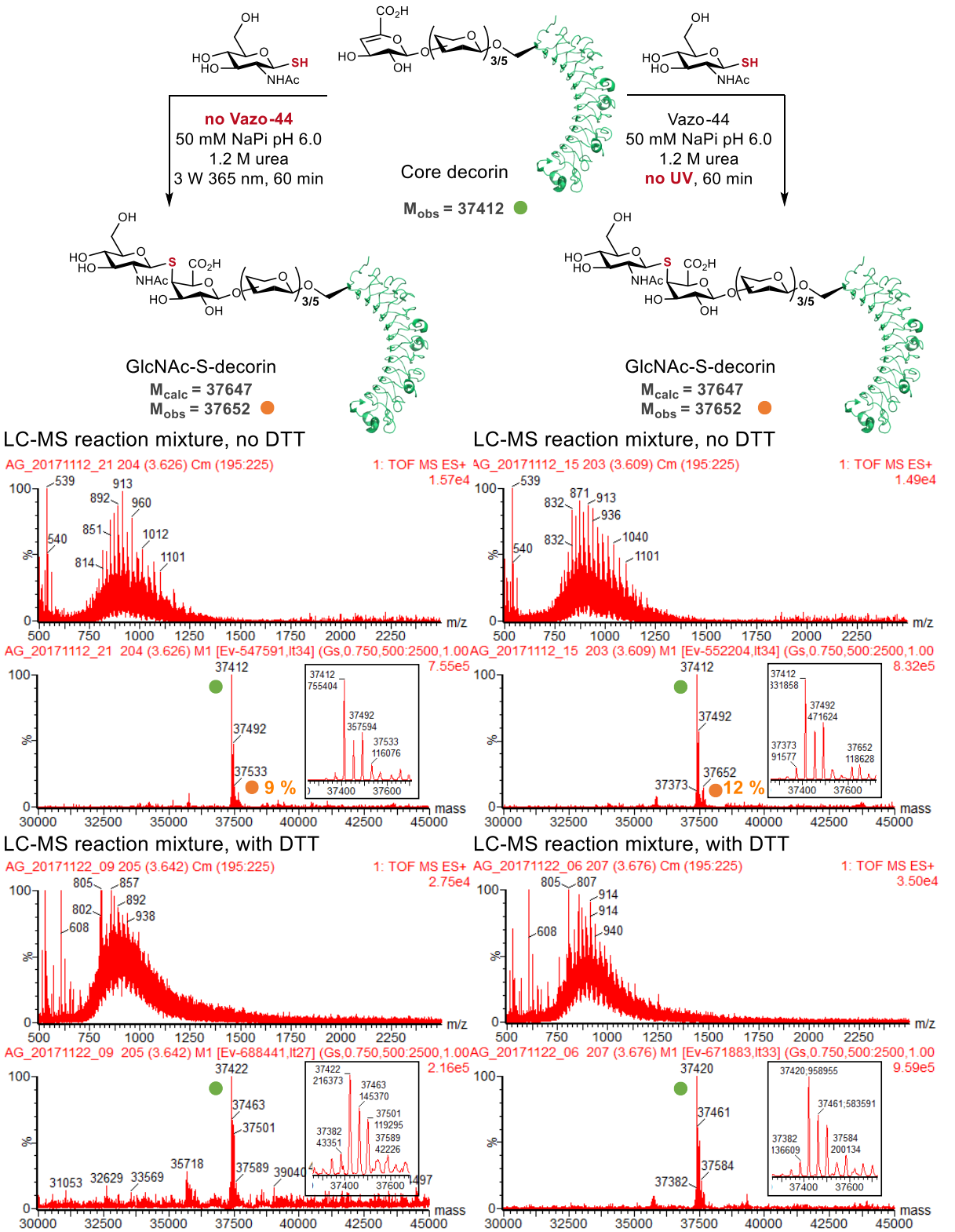


Figure SI 2.14. Intact protein LC-MS analyses did not show a significant conversion of the core decorin to the desired GlcNAc-S-decorin when no radical initiator or no UV irradiation were used. Traces of GlcNAc-SH addition were observed in the samples not treated with DTT (9 % and 12 %, respectively) and these were not detected in the samples treated with DTT.

Two negative control reactions were set up in parallel, as follows:

1: 25 µL protein solution (c = 1.6 mg/mL in 50 mM NaPi buffer pH 6.0 and 1.2 M urea) were added to 5.8 mg GlcNAc-SH (1000 mM) and the reaction mixture was irradiated for 1 hour with a 365 nm UV LED torch (3 W) and progression was monitored by LC-MS analysis after 60 minutes.

2: 25 µL protein solution (c = 1.6 mg/mL in 50 mM NaPi buffer pH 6.0 and 1.2 M urea) were added to 5.8 mg GlcNAc-SH (1000 mM) and then to 1.6 mg Vazo-44 (200 mM). The vial was covered with an aluminium foil and reaction mixture was incubated for 1 hour in the dark, at room temperature. Progression was monitored by LC-MS analysis after 60 minutes.

A third negative control reaction was set up as follows:

3: Intact decorin was buffer exchanged into 50 mM NaPi pH 6.0 buffer using Vivaspin® 500 (5 kDa MWCO) and diluted with a 6 M urea solution to have a final protein c = 1.6 mg/mL and 1.2 M final urea concentration. 200 µL protein solution were then incubated with 1 µL PNGaseF (c = 500 U/µL) for 2 hours at 37 °C and 300 rpm. Then, this was divided and treated as follows:

- 90 µL were buffer exchanged into 100 mM TrisHCl, 50 mM NaOAc pH 8.0 buffer using Vivaspin® 500 (10 kDa MWCO) and incubated with 2 µL Chondroitinase ABC (c = 1 U/100 µL) for 1.5 hours at 37 °C and 400 rpm.

- 90 µL were incubated with 5.75 mg Vazo-44 (200 mM) and GlcNAc-SH (20.9 mg) for 60 minutes, under irradiation with a 365 nm UV LED torch (3 W). It was then buffer exchanged into 100 mM TrisHCl, 50 mM NaOAc pH 8.0 buffer using Vivaspin 500 (10 kDa MWCO) and incubated with 2 µL Chondroitinase ABC (1U/100 µL conc) for 1.5 hours at 37 °C and 400 rpm. LC-MS analysis of the two reaction mixtures were identical (no GlcNAc-SH addition was observed when the GAG-chain on decorin had not been trimmed).

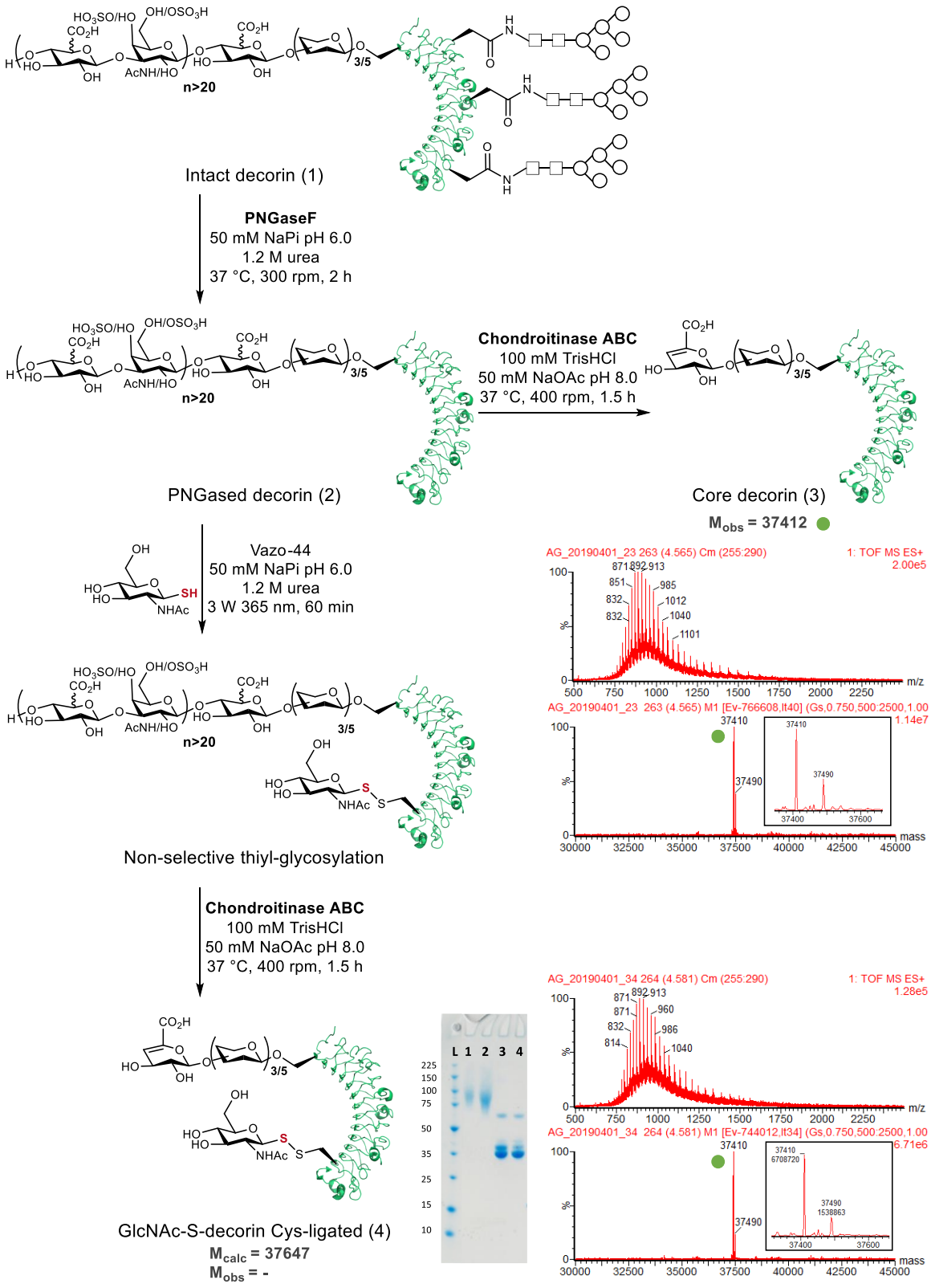
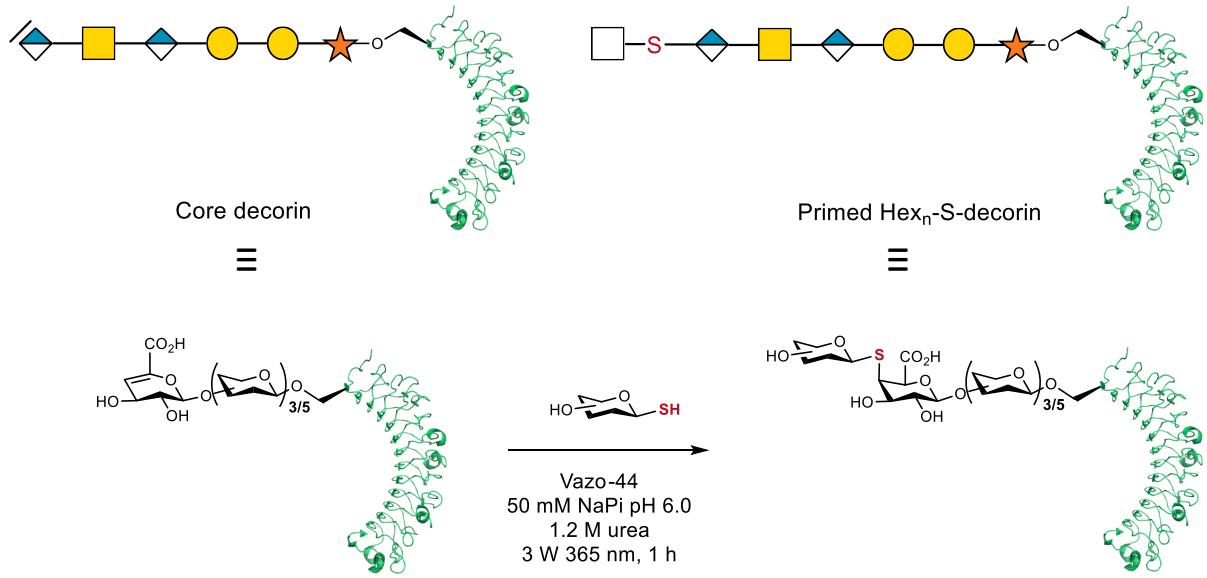


Figure SI 2.15. Intact protein LC-MS analyses showed that the thiyl-glycosylation did not occur when the GAG chain had not been trimmed, and the formation of the potential side-product GlcNAc-S-decorin (4) was not observed.

2.4.3.3 Radical hydrothiolation of decorin with various 1-thio-saccharides



Lyophilized GAG-trimmed decorin was dissolved in 50 mM NaPi buffer pH 6.0 and 1.2 M urea in a 1.5 mL Eppendorf tube to have a final concentration of c = 1.6 mg/mL. PNGaseF was added (1560 U/mg substrate, from a 500000 U/mL commercial stock) and reaction mixture was stirred for 2 h at 37 °C and 300 rpm. Then, 25/50 µL protein solution were added to each of the sugars (GlcA-SH / Glc-SH / Glc-ethyl-SH / GlcNAc-ethyl-SH / GalGlc-SH) to have a final 1000 mM sugar concentration and then each obtained mixture was transferred into a Chromacol® clear vial (03-FISV) containing the Vazo-44 (200 mM final concentration). The reaction mixtures were UV-irradiated for 1 hour with a 365 nm UV LED torch (3 W) and progression was monitored by intact protein LC-MS analysis after 60 minutes, as follows.

DTT monitoring: 2 µL sample were diluted to 64 µg/mL in 100 mM NH_4OAc pH 8.0 buffer and 100 mM DTT;

No DTT monitoring: 2 µL sample were diluted to 64 µg/mL in 100 mM NH_4OAc pH 8.0 buffer.

Table 2, Entry 2: 0.2 M Vazo-44, 1 M GlcA-SH, 3 W UV torch, 60 min

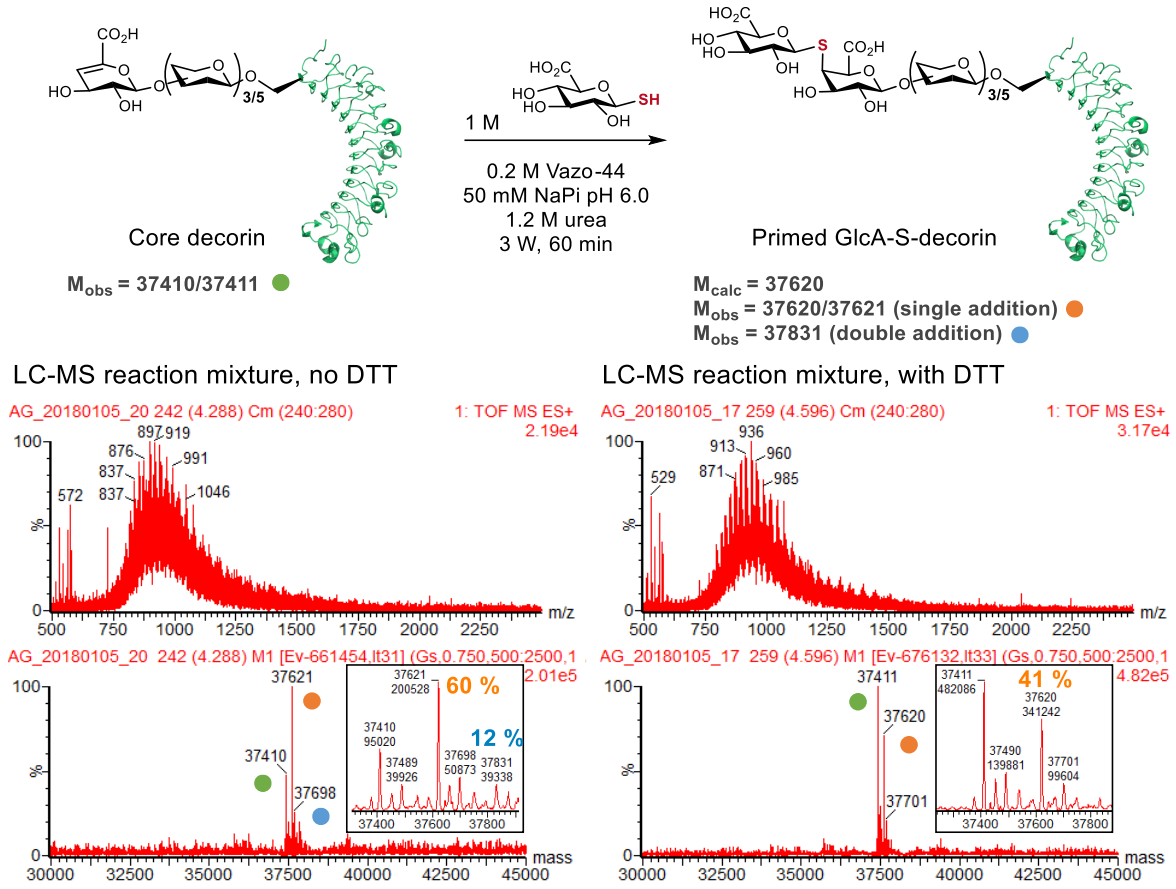


Figure SI 2.16. Intact protein LC-MS analyses of the *S*-glycosylation of decorin with GlcA-SH. Good conversion to the desired GlcA-S-decorin was achieved (60 %) but double GlcA-addition (12 %) could also be observed. However, DTT treatment of the LC-MS samples significantly decreased the conversion to the desired single GlcA-addition product (41 %).

Table 2, Entry 3: 0.2 M Vazo-44, 1 M Glc-SH, 3 W UV torch, 60 min

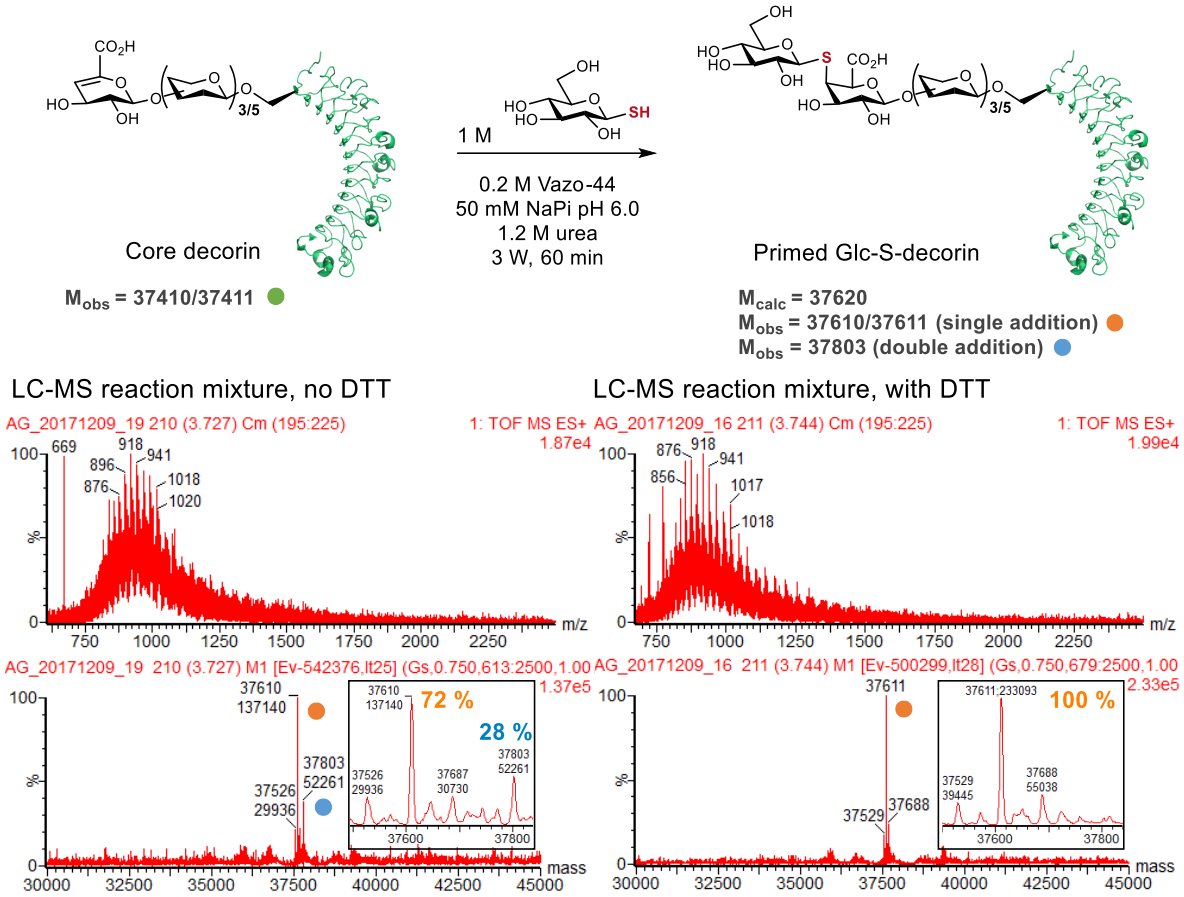


Figure SI 2.17. Intact protein LC-MS analyses of the *S*-glycosylation of decorin with Glc-SH. Good conversion to the desired Glc-S-decorin was achieved (72 %) but double Glc-addition (28 %) could also be observed. However, upon reduction of the sample with DTT the latter was completely converted into the expected single Glc-addition product (100 %).

Table 2, Entry 4: 0.2 M Vazo-44, 1 M Glc-ethyl-SH, 3 W UV torch, 60 min

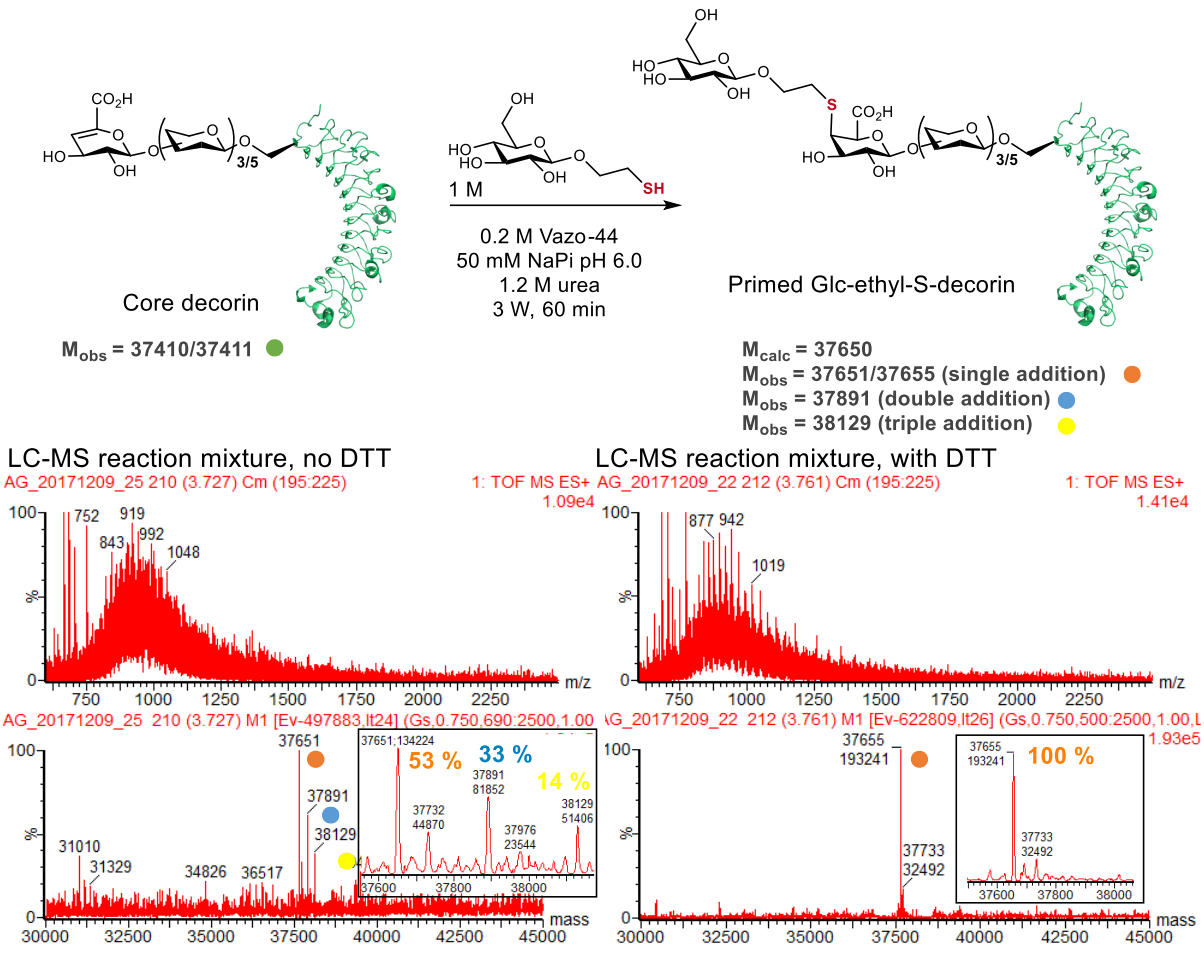


Figure SI 2.18. Intact protein LC-MS analyses of the *S*-glycosylation of decorin with Glc-ethyl-SH. Moderate conversion to the desired Glc-ethyl-S-decorin was achieved (53 %) but double Glc-ethyl-addition (33 %) and even triple Glc-ethyl-addition (14 %) could also be observed. However, upon reduction of the sample with DTT the multiple addition species were completely converted into the expected, single Glc-ethyl-addition product (100 %).

Table 2, Entry 5: 0.2 M Vazo-44, 1 M GlcNAc-ethyl-SH, 3 W UV torch, 60 min

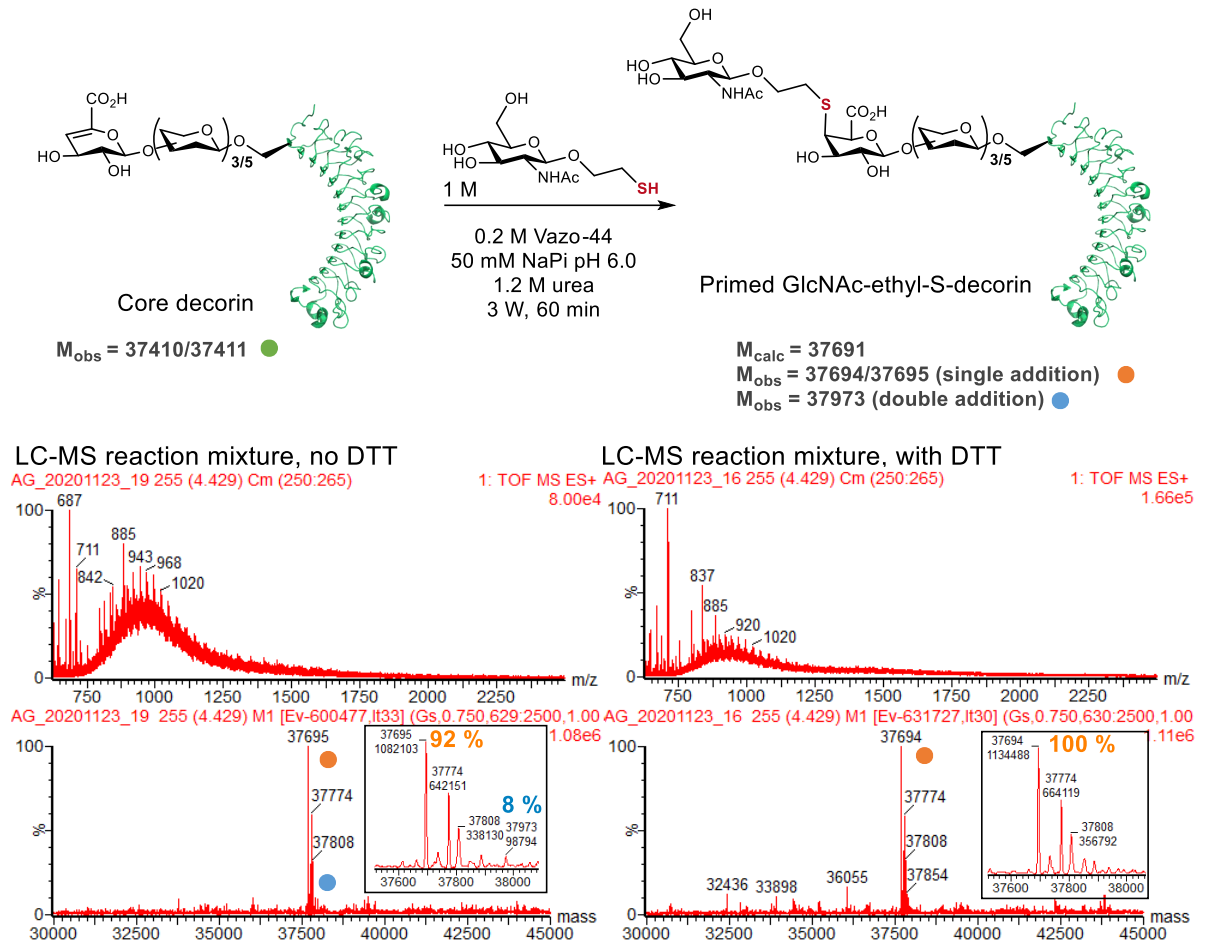
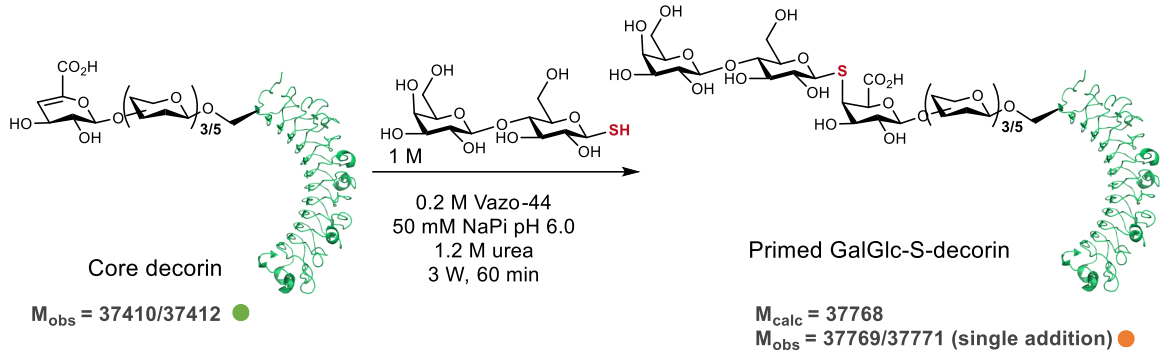


Figure SI 2.19. Intact protein LC-MS analyses of the *S*-glycosylation of decorin with GlcNAc-ethyl-SH. Excellent conversion to the desired GlcNAc-ethyl-S-decorin was achieved (92 %) but double GlcNAc-ethyl-addition (8 %) could also be observed. However, upon reduction of the sample with DTT the latter was completely converted into the expected, single GlcNAc-ethyl-addition product (100 %).

Table 2, Entry 6: 0.2 M Vazo-44, 1 M GalGlc-SH, 3 W UV torch, 60 min



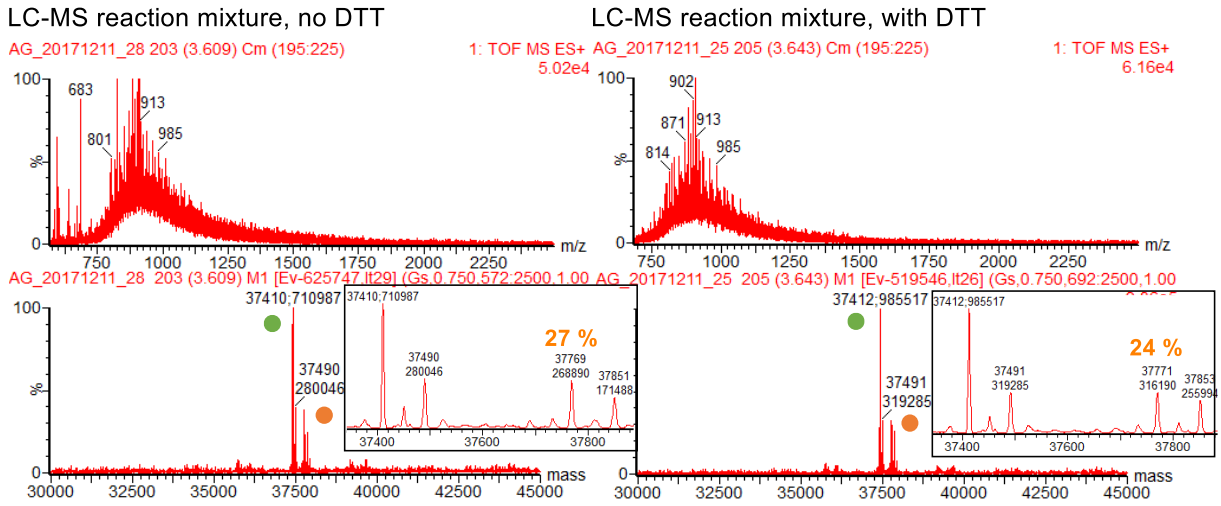


Figure SI 2.20. Intact protein LC-MS analyses of the *S*-glycosylation of decorin with GalGlc-SH. Low conversion to the desired GalGlc-S-decorin was achieved (27 %), and this was not substantially reduced upon reduction of the sample with DTT (24 %).

2.4.3.4 Characterization of enzymatically digested GlcNAc-S-decorin by LC-MS/MS analysis

Enzymatic digestion of GlcNAc-S-decorin with Trypsin:

10 µg of GlcNAc-S-decorin were denatured in 8 M urea, 100 mM TEAB for 30 min at room temperature (r.t.), 400 rpm and subsequently reduced with 10 mM TCEP for 30 min at r.t., 400 rpm before being alkylated with 50 mM 2-chloroacetamide for 30 min at r.t., 400 rpm, in the dark. Samples were diluted to 0.8 M urea with 50 mM TEAB and digestion was performed with Trypsin (1:20 w/w) O/N at 37 °C, 400 rpm. Digested peptides were desalted using Oasis® HLB cartridges (10 mg) from Waters, eluting with 250 µL 50:50:0.1 MeCN/H_2O/TFA and 250 µL 80:20:0.1 MeCN/H_2O/TFA. Combined fractions were lyophilized and re-dissolved in 50 µL 90:5:5 H_2O/FA/DMSO.

Data acquisition: SCE-HCD experiment on a QExactive MS instrument.

Resulting peptides (2 µL, 1:10 dil.) were separated by nano-flow reversed-phase liquid chromatography on an EASY-nLC 1000 UHPLC system (Proxeon) coupled to a Q Exactive Hybrid Quadrupole-Orbitrap mass spectrometer (Thermo Fischer Scientific).
The peptides were trapped on a guard column (inner diameter 75 µm, 20 mm, 3 µm Reprosil C18, 120 Å) using solvent A (0.1% formic acid in water) and separated on a C18 analytical column (inner diameter 75 µm × 15 cm, RSLC C18, 3 µm, 100 Å). Separation of peptides was conducted using a linear gradient from 3 % to 30 % of solvent B (0.1% formic acid in acetonitrile) for 120 min at a flow rate of 300 nL/min. The raw data were acquired on the mass spectrometer in data-dependent mode. Automatically switching from MS to higher energy collision induced dissociation MS/MS. Full-scan spectra were acquired in the Orbitrap [scan range 350-2000 m/z, resolution 70000, automatic gain control (AGC) target 3 x 10^6, maximum injection time 50 ms]. After the MS scan, the top 10 most intense peaks were selected for HCD fragmentation in a 2 stepped-collision energy 10 % - 15 % of normalised collision energy. HCD spectra were also acquired in the Orbitrap (resolution 17500; AGC target 5 x 10^4; maximum injection time 120 ms), with first fixed mass at 180 m/z. Full scan MS spectra (scan range 350-1500 m/z, resolution 120000, AGC target $1e^6$, maximum injection time 250 ms) and subsequent HCD MS/MS spectra (resolution 15000, AGC target $3e^4$, maximum injection time 500 ms) of 5 most intense peaks were acquired in the Orbitrap. HCD fragmentation was performed at 10 % and 15 % of normalized collision energy and the signal intensity threshold was kept at 500 counts.

Data analysis:

Generated raw MS/MS data files were processed using the Byonic™ software 3.6.0 version and the search was performed against a focused protein database containing the glycoprotein sequence and the identified contaminants. Using Byonic's ProteinLabel keyword, glycosylation was only allowed on the glycoprotein of interest. For all data acquired in the Orbitrap (Q Exactive), a 20 ppm mass error was set. Carbamidomethylation was set as a fixed modification while oxidation (Met), deamidation (Asn, Gln), carbamidomethylation (peptide N-terminus, Lys), phosphorylation (Ser, Thr) and oxidation (Pro) were set as variable modifications. One missed cleavage was allowed, and the data was searched with trypsin as enzyme (R, K) and semi-specific cleavage. The following glycans were added to the database of *O*-glycans reported in section 2.4.2.1 and the updated list was used for the identification of the Ser-linked GlcNAc-S-modified saccharide: GlcNAc-S-GlcA-GalNAc-GlcA-Gal-Gal-Xyl. Note that the *S*-glycosylation product is defined as GlcNAc-S-GlcA = Glc(A) +219.0565.

Xyl(1)Hex(2)GlcA(2)HexNAc(1) +219.0565 @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) +219.0565 @ OGlycan | rare1
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) +219.0565 @ OGlycan | rare1
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) +219.0565 @ OGlycan | rare1
Xyl(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) +219.0565 @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) +219.0565 @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) +219.0565 @ OGlycan | rare1

2.4.3.5 Refolding of GlcNAc-S-decorin in the L-arginine containing buffer.

150 µL protein solution (c = 1.6 mg/mL in 50 mM NaPi, 1.2 M urea, pH 6.0) were added to 34.8 mg GlcNAc-SH (1000 mM) and then to 9.6 mg Vazo-44 (200 mM) and the reaction mixture was irradiated for 1 hour with a 365 nm UV LED torch (3 W). Then, it was buffer exchanged into 50 mM NaOAc, 7 M urea pH 6.0 buffer and dialysed in the same buffer at room temperature, overnight (A). Then, the protein solution was divided into two equal aliquots and re-dialysed in 20 mM NaHEPES, 125 mM NaCl pH 7.5 buffer (B) or in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer (C). The stability of the two protein solutions was checked after 24 and 48 hours, by intact protein LC-MS analyses as follows:

DTT monitoring: 2 µL sample were diluted with 43 µL 100 mM NH_4OAc pH 8.0 buffer and 5 µL 1 M DTT solution to have a final concentration of 100 mM DTT;
No DTT monitoring: 2 µL sample were diluted with 48 µL 100 mM NH_4OAc pH 8.0 buffer.

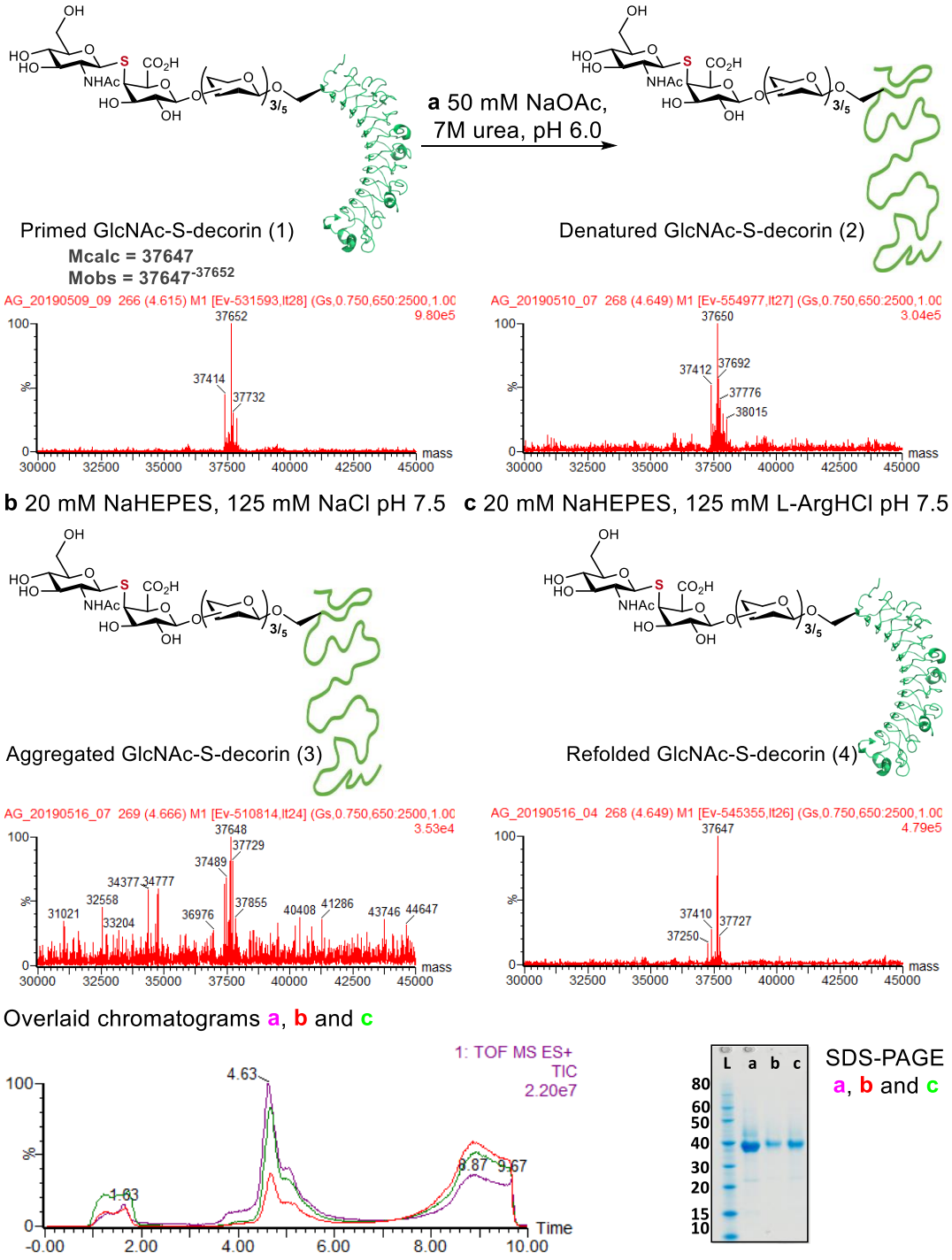


Figure SI 2.21. Denaturation of the GlcNAc-S-primed decorin in a high-concentration urea buffer, and attempted protein recovery by dialysis in the NaCl buffer or the L-ArgHCl buffer. Intact protein LC-MS and SDS-PAGE analyses showed that upon denaturation of the GlcNAc-S-decorin (1) in a high-concentration urea buffer (**a**) this might have undergone aggregation when buffer exchanged into the NaCl buffer (**b**) as indicated by the low MS intensity (deconvoluted spectrum and red trace chromatogram of 3) and the low staining intensity of the gel band (lane b). On the contrary, acceptable protein recovery was observed when the thiyl-glycosylation product (2) was re-dialysed in the L-ArgHCl buffer (**c**) as indicated by the high MS intensity (deconvoluted spectrum and green trace chromatogram of 4) and the moderate staining intensity of the gel band (lane c).

2.4.4 Chemo-enzymatic synthesis of neoproteoglycans containing a heparosan GAG chain

2.4.4.1 Formation of Dha-TEV-H3, GlcNAc-S-TEV-H3 and GlcAGlcNAc-S-TEV-H3

Dha-TEV-H3 was prepared according to a procedure developed in the group.[Ref] Lyophilized Cys-TEV-H3 (10 mg) was solubilized in 500 µL reaction buffer (RB, 3 M Gdn·HCl, 100 mM NaPi, pH 8.0) and 25 mg of DTT were added. The mixture was incubated at 37 °C, 300 rpm for 45 minutes and desalted on a PD MiniTrap G-25 equilibrated in RB. Then, 24 equiv. of DBHDA (4.5 mg dissolved in 26 µL DMSO) were added to the protein solution and the mixture was incubated at 37 °C, 300 rpm for 90 minutes. It was then filtered through a PD MiniTrap G-25 equilibrated in RB and further incubated at 37 °C, 300 rpm until completion of the reaction (60 minutes).

34 equiv. of GlcNAc-SH (2.5 mg) were added to 500 µL Dha-TEV-H3 protein solution (c = 5 mg/mL in 3 M Gdn·HCl, 100 mM NaPi, pH 8.0 buffer) and the mixture was incubated at 37 °C, 300 rpm for 2 hours. The GlcNAc-S-TEV-H3 was desalted by PD MiniTrap G-25 equilibrated in RB and stored at - 80 °C until further use.

40.6 equiv. of GlcAGlcNAc-SH (1.1 mg) were added to 150 µL Dha-TEV-H3 protein solution (c = 7 mg/mL in 3 M Gdn·HCl, 100 mM NaPi, pH 8.0 buffer) and the reaction mixture was incubated at 37 °C, 400 rpm for 1 hour. Then, 29.6 equiv. GlcAGlcNAc-SH (0.8 mg) were added and the reaction mixture was incubated again at 37 °C, 400 rpm for 1 hour. The GlcAGlcNAc-S-TEV-H3 was buffer exchanged using a SpinTrap G-25 equilibrated in 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 and used for enzymatic extension reactions catalysed by pmHS2.

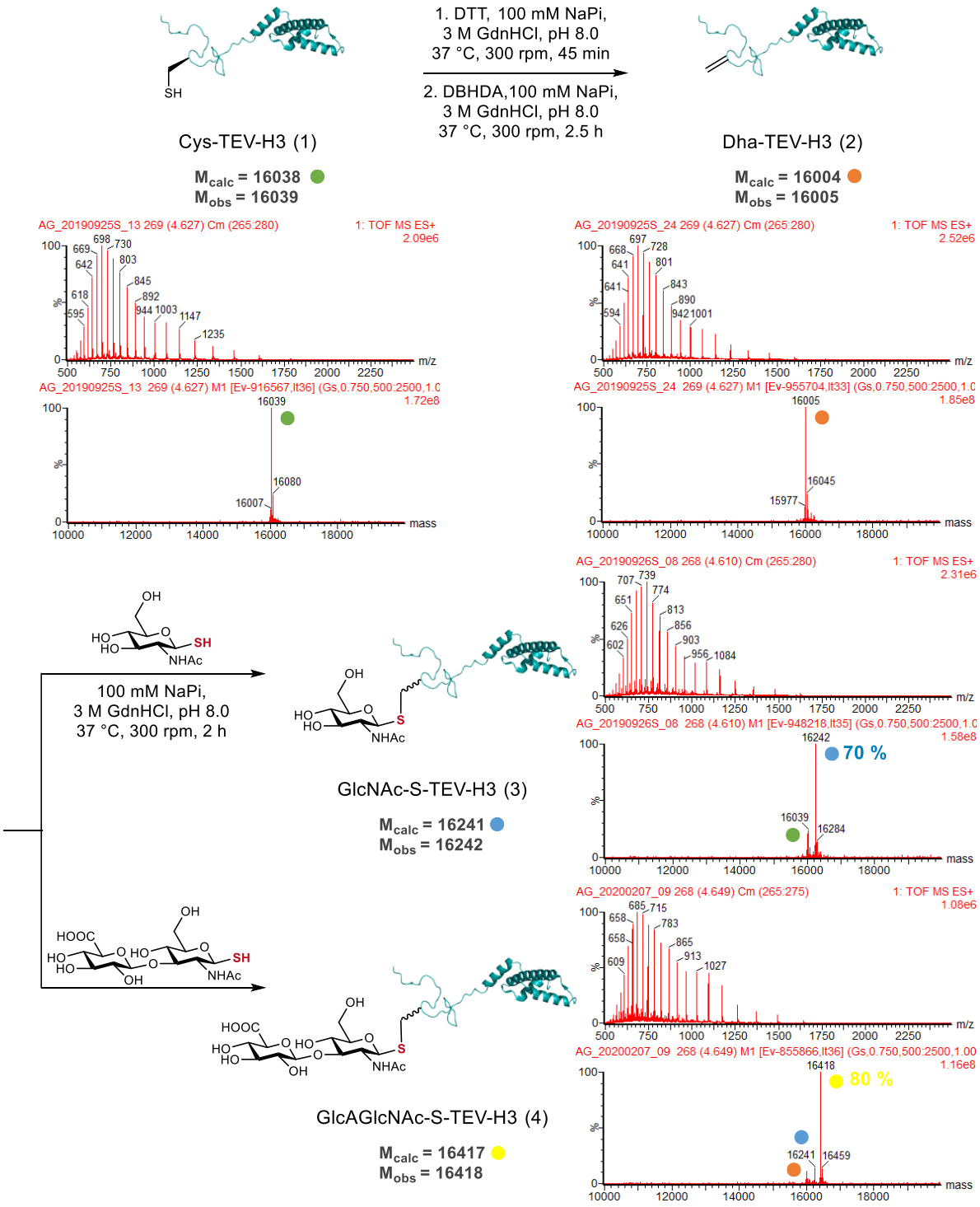


Figure SI 2.22. Intact protein LC-MS analyses of the GlcNAc-S-TEV-H3 and GlcAGlcNAc-S-TEV-H3 synthesis. Excellent conversion of Cys-TEV-H3 (1) into the corresponding Dha-derivative (2) was achieved. The thia-Michael addition of GlcNAc-SH to Dha-TEV-H3 (2) afforded the desired product GlcNAc-S-TEV-H3 (70 % conversion to 3, 14 % Dha-TEV-H3 and 16 % Cys-TEV-H3 due to incomplete reduction with DTT in the previous step). Similarly, the thia-Michael addition of GlcAGlcNAc-SH to the Dha-tagged protein (2) gave the expected product GlcAGlcNAc-S-TEV-H3 (80 % conversion to 4, 9 % Dha-TEV-H3 and 11 % GlcNAc-TEV-H3 due to the disaccharide degradation).

2.4.4.2 Optimization of the enzymatic GlcA-extension of GlcNAc-TEV-H3 by pmHS2

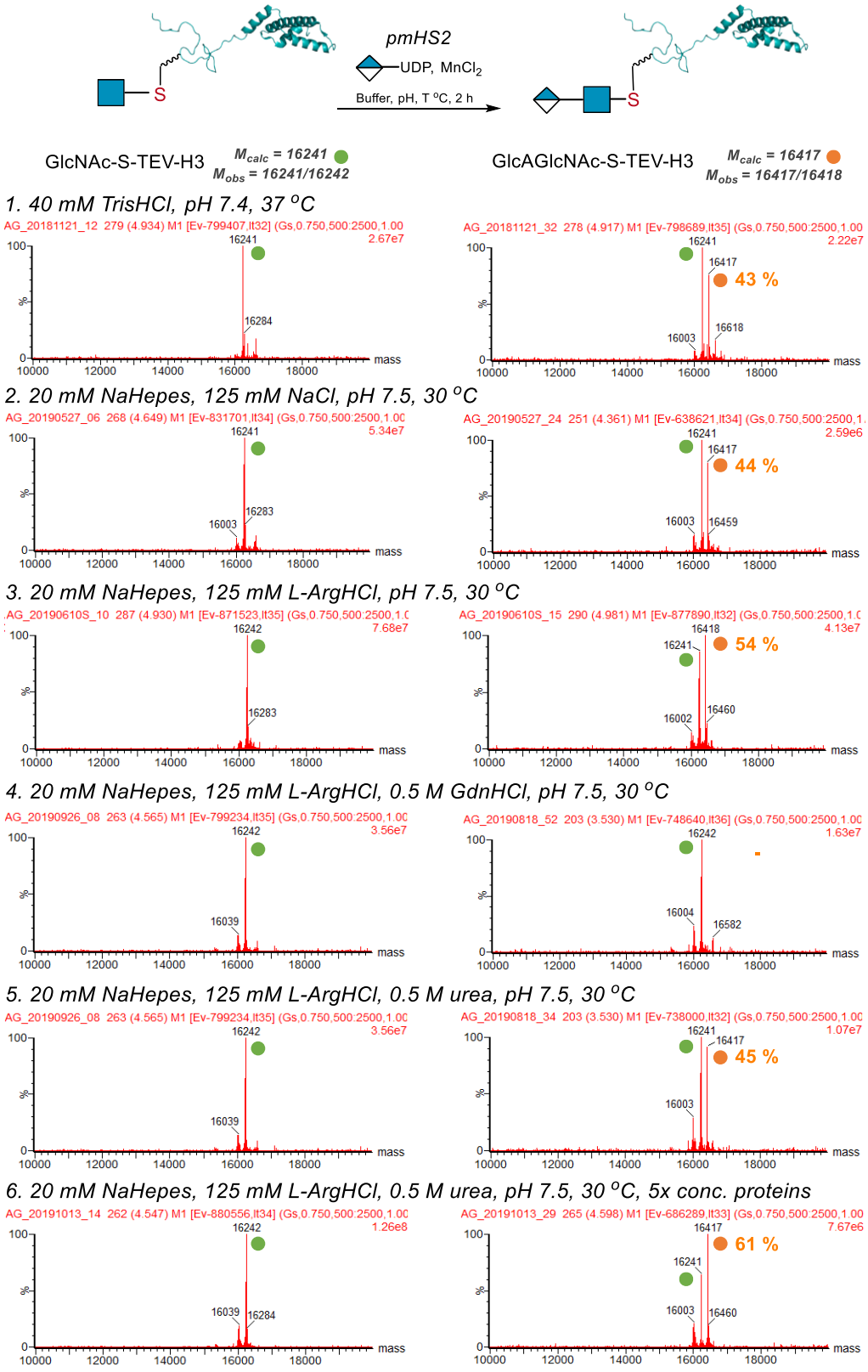


Figure SI 2.23. Representative deconvoluted spectra of intact protein LC-MS analyses that were recorded to evaluate the influence of various reaction conditions (summarized in Figure 2.13) on the pmHS2-catalysed GlcA-extension of GlcNAc-S-TEV-H3.

A typical 100 µL reaction mixture would contain 35 µL pmHS2 protein solution (c_1 = 0.29 mg/mL or c_2 = 1.45 mg/mL), 1 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 20 µL buffer, 20 µL GlcNAc-S-TEV-H3 protein solution (c_1 = 0.7 mg/mL or c_2 = 3.5 mg/mL in buffer, or buffer supplemented with 1.5 M urea), 14 µL buffer or buffer supplemented with 1.5 M urea and 10 µL UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 5 mM). The enzymatic reaction mixtures were incubated for 2 h at 30 °C.

pmHS2 enzymatic activity at different urea and Gdn·HCl concentrations

13 x 100 µL reaction mixtures were set up. In each case, to 35 µL pmHS2 protein solution (c = 0.29 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 1.31 µM final), 1 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 34 µL buffer X, 20 µL GlcNAc-S-TEV-H3 protein solution (0.7 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 8.62 µM final) and 10 µL UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 5 mM) were added. These were incubated for 2 h at 30 °C. Then, to each, 1 µL UDP-GlcNAc (50 mM solution in ddH_2O, to have a final concentration of 0.5 mM) and 9 µL ddH_2O were added and reaction mixtures were incubated for 16 h at 30 °C.

Buffer X: 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer containing 0, 0.5, 1, 1.5, 2, 4, 6 M urea/ Gdn·HCl.

GlcA-addition was monitored by LC-MS analyses after 2 h (3 µL injected after 10X dilution of a 5 µL aliquot with 0.1 % FA ddH_2O) and overall enzymatic extension was monitored by SDS-PAGE analyses after 16 h (denaturation of a 15 µL aliquot with 5 µL reducing 4xLaemmli buffer at 95 °C for 5 min).

2.4.4.3 Formation of various GlcNAc-primed proteins as potential substrates for pmHS2 (GlcNAc-S-H2B, GlcNAc-S-H3, GlcNAc-S-TEV-H3 and GlcNAc-ethyl-S-TEV-H3)

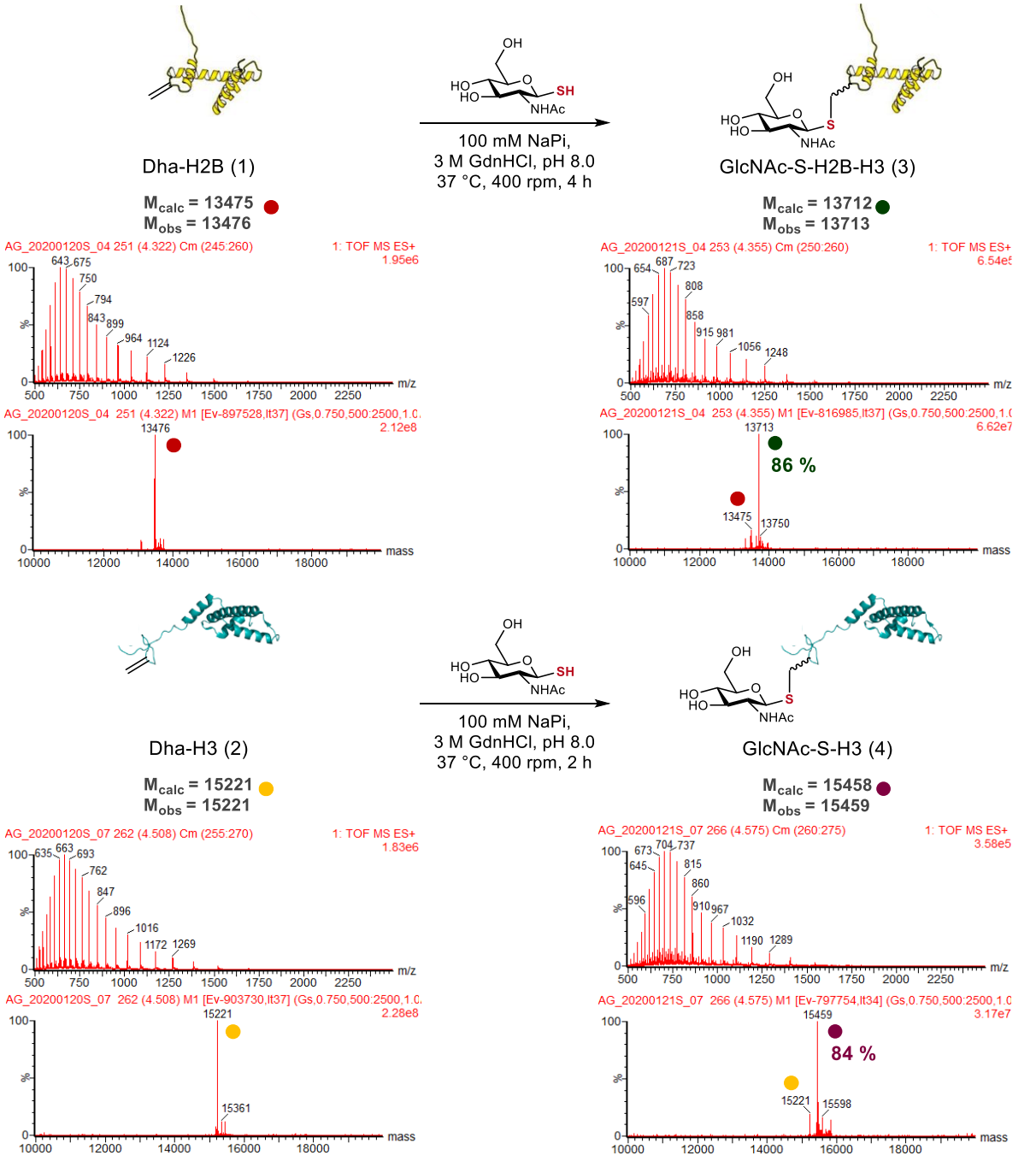


Figure SI 2.24. Intact protein LC-MS analyses of the GlcNAc-S-H2B and GlcNAc-S-H3 glycoproteins. The thia-Michael addition of GlcNAc-SH to Dha-H2B (1) and to Dha-H3 (2) generated the GlcNAc-S-H2B (3, 86 % conversion) and the GlcNAc-S-H3 (4, 84 % conversion), respectively.

To 130 µL Dha-H2B (S36Dha)/Dha-H3 (S10Dha) protein solution (c = 1.5 mg/mL in 3 M Gdn·HCl, 100 mM NaPi, pH 8.0 buffer) 0.6 mg GlcNAc-SH (175.0 and 197.6 equivalents, respectively) were added and the reaction mixtures were incubated at 37 °C and 400 rpm. After 3 hours, 0.2 mg GlcNAc-SH (58.3

and 65.9 equivalents, respectively) were added and the reaction mixtures were incubated again at 37 °C and 400 rpm for 1 hour. They were then filtered through PD SpinTrap G-25 columns equilibrated in 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 and used for enzymatic extension reactions catalysed by pmHS2.

To 500 µL Dha-TEV-H3 protein solution (c = 5 mg/mL in 3 M Gdn·HCl, 100 mM NaPi, pH 8.0 buffer) 3.0 mg GlcNAc-SH (81.0 equivalents) were added and the reaction mixture was incubated at 37 °C and 400 rpm for 2 hours. It was then filtered through PD SpinTrap G-25 columns equilibrated in 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 and used for enzymatic extension reactions catalysed by pmHS2, ΔpmHS2 and GalT.

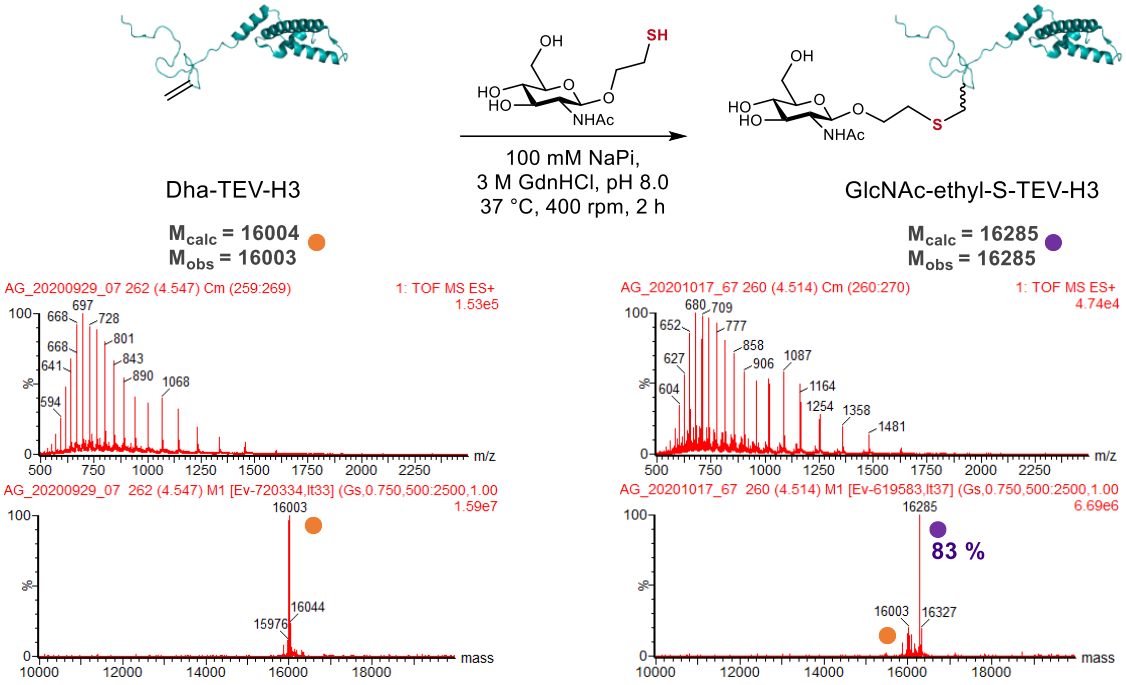


Figure SI 2.25. Intact protein LC-MS analysis of the GlcNAc-ethyl-S-TEV-H3 glycoprotein. The thia-Michael addition of GlcNAc-ethyl-SH to Dha-TEV-H3 generated the desired GlcNAc-ethyl-S-TEV-H3 (83 % conversion).

To 150 µL Dha-TEV-H3 protein solution (c = 5 mg/mL in 3 M Gdn·HCl, 100 mM NaPi, pH 8.0 buffer) 1.1 mg GlcNAc-ethyl-SH (84.0 equivalents) were added and the reaction mixture was incubated at 37 °C and 400 rpm for 2 hours. It was then filtered through PD SpinTrap G-25 columns equilibrated in 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 and used for enzymatic extension reactions catalysed by pmHS2, ΔpmHS2 and GalT. Representative LC-MS analysis are depicted in Figure SI 2.22.

2.4.4.4 pmHS2 GlcA-transferase activity on various GlcNAc-S-histone substrates (GlcNAc-S-H2B, GlcNAc-S-H3 and GlcNAc-S-TEV-H3)

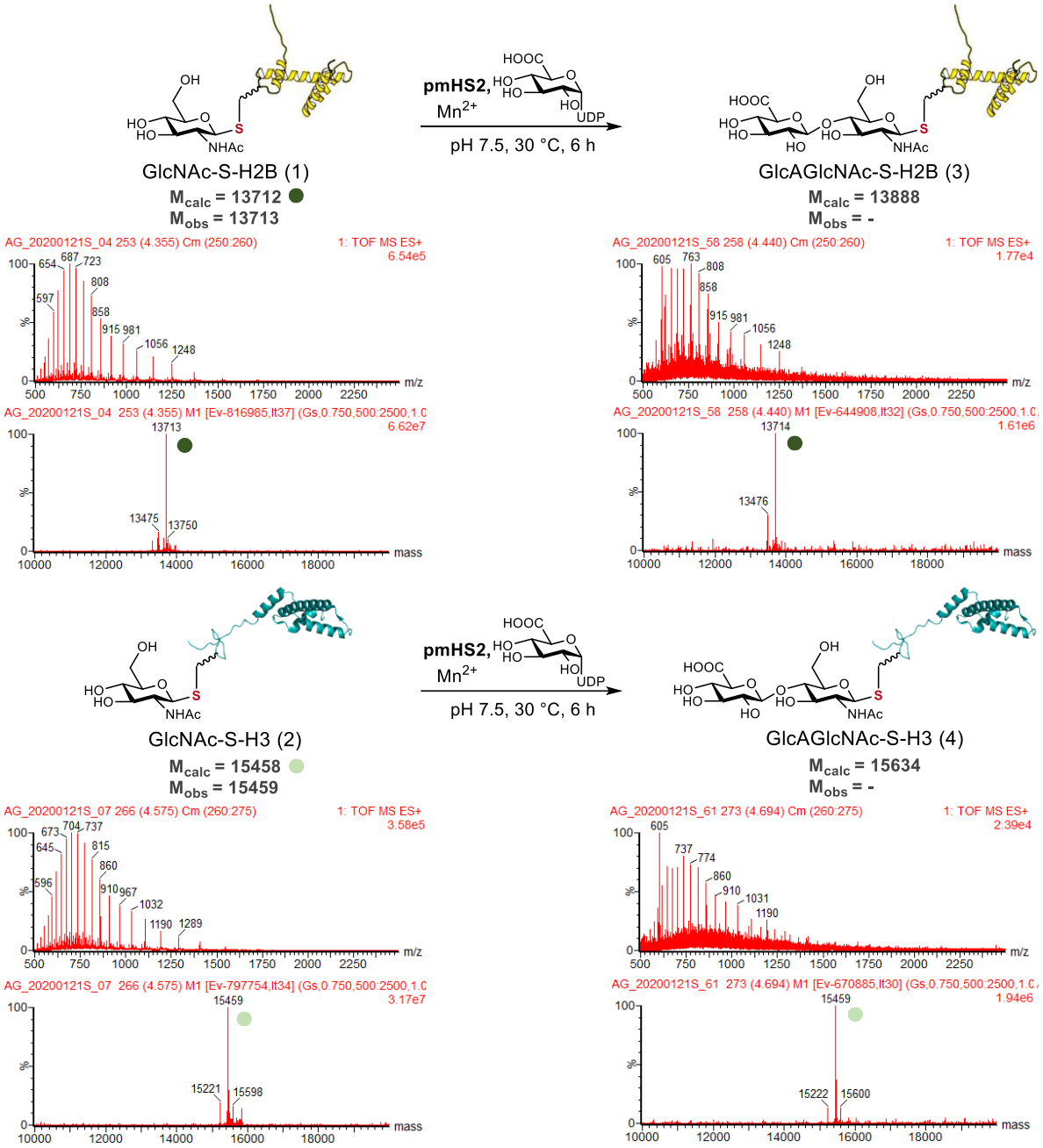


Figure SI 2.26. Intact protein LC-MS analyses showed that pmHS2 did not catalyse the GlcA-extension of GlcNAc-S-H2B (1) or GlcNAc-S-H3 (2), and GlcAGlcNAc-S-H2B (3) or GlcAGlcNAc-S-H3 (4) could not be obtained.

Three reaction mixtures were set up in parallel, as follows and incubated at 30 °C for 6 hours. To 70 µL pmHS2 protein solution (c = 0.29 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer), 2 µL $MnCl_2$ (250 mM solution in ddH_2O), 68 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 40 µL GlcNAc-S-H2B/GlcNAc-S-H3/GlcNAc-S-TEV-H3 protein solution (43.1 µM in 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea, pH 7.5 buffer) and 20 µL UDP-GlcA (50 mM solution in ddH_2O) were added. GlcA-

addition was monitored by LC-MS analyses after 0.5, 1, 2, 3, 4 and 6 h (3 µL injected after 10X dilution of a 5 µL aliquot with 0.1 % FA ddH_2O).

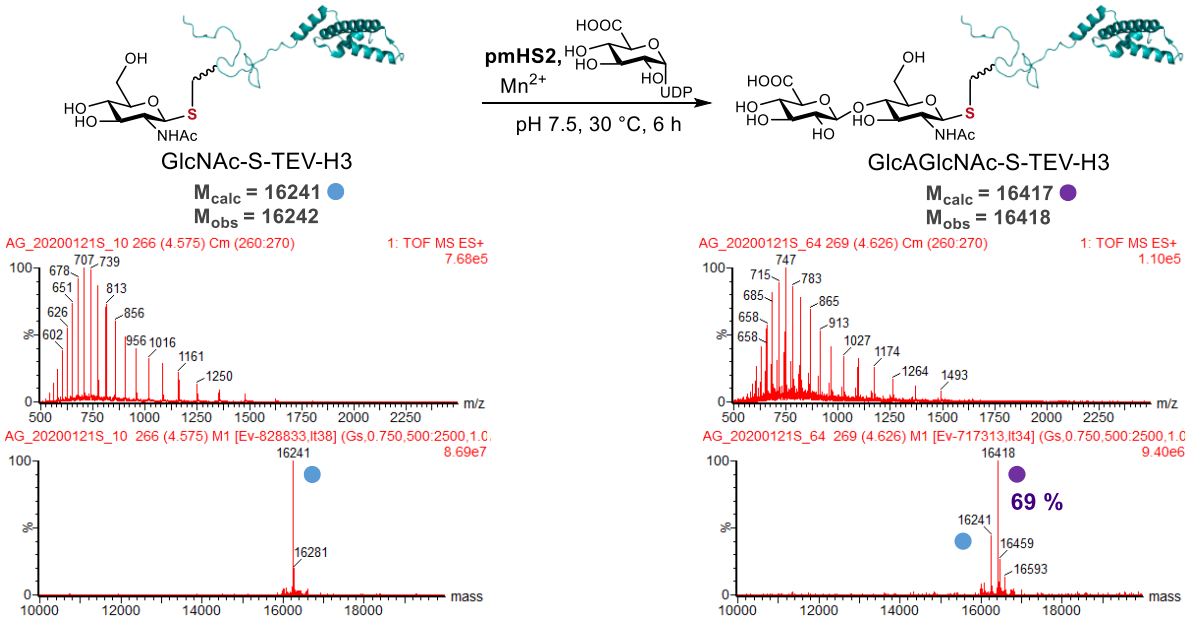


Figure SI 2.27. Intact protein LC-MS analysis showed that pmHS2 transferred one GlcA residue to acceptor GlcNAc-S-TEV-H3 and the GlcAGlcNAc-S-TEV-H3 could be obtained (69 % conversion).

2.4.4.5 Enzymatic extension of the GlcNAc-S-TEV-H3 substrate by pmHS2, formation of the GlcAGlcNAc-S-TEV-H3 intermediate and attempted on-protein Hep-polymerizations

Table 4, Entry 1:

3 x 100 µL reaction mixtures were set up. In each case, to 35 µL pmHS2 protein solution (c = 0.29 mg/mL, in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 1.31 µM final), 1 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 34 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 20 µL GlcNAc-S-TEV-H3 protein solution (0.7 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 8.62 µM final) and 10 µl UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 5 mM) were added. These were incubated for 2 h at 30 °C. Then, to each, UDP-GlcNAc (50 mM solution in ddH_2O) was added, and reaction mixtures were incubated for 16 h at 30 °C.

1. 5 mM UDP-GlcNAc: 10 µL UDP-GlcNAc (50 mM solution in ddH_2O)
2. 0.5 mM UDP-GlcNAc: 1 µL UDP-GlcNAc (50 mM solution in ddH_2O) + 9 µL ddH_2O
3. 0.1 mM UDP-GlcNAc: 0.2 µL UDP-GlcNAc (50 mM solution in ddH_2O) + 9.8 µL ddH_2O

GlcA-addition was monitored by LC-MS analyses after 2 h (3 µL injected after 10X dilution of a 5 µL aliquot with 0.1 % FA ddH_2O) and GlcNAc-addition was monitored by LC-MS analyses after 15 min, 30 min and 1 h (3 µL injected after 10X dilution of a 5 µL aliquot with 0.1 % FA ddH_2O). Overall enzymatic

extension was monitored by SDS-PAGE analyses after 1 h and 16 h (denaturation of a 15 µL aliquot with 5 µL reducing 4x Laemmli buffer at 95 °C for 5 min).

Table 4, Entry 2:

A 500 µL reaction mixture was set up as follows: to 175 µL pmHS2 protein solution (c = 1.45 mg/mL (5x0.29 mg/mL) in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 6.55 µM final), 5 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 100 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 70 µL 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 buffer, 100 µL GlcNAc-S-TEV-H3 protein solution (3.5 mg/mL (5x0.7 mg/mL) in 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 buffer, 43.1 µM final) and 50 µL UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 5 mM) were added. After incubation at 30 °C for 2 h, 2 x 200 µL reaction mixtures were set up. To each, UDP-GlcNAc (50 mM solution in ddH_2O) was added, and reaction mixtures were incubated for 16 h at 30 °C.

1. 5 mM UDP-GlcNAc: 20 µL UDP-GlcNAc (50 mM solution in ddH_2O)
2. 0.5 mM UDP-GlcNAc: 2 µL UDP-GlcNAc (50 mM solution in ddH_2O) + 18 µL ddH_2O

Overall enzymatic extension was monitored by SDS-PAGE analyses after 15 min, 30 min, 1 h, 2 h and 16 h (denaturation of a 15 µL aliquot with 5 µL reducing 4xLaemmli buffer at 95 °C for 5 min). Coomassie blue staining was performed according to the manufacturer protocol and Ab96C10 (primary anti-H3 antibody) followed by A3562 (anti-mouse IgG−alkaline phosphatase secondary antibody, BCIP/NBT substrate) were used for western blot detection.

pmHS2 GlcNAc-transferase activity on GlcAGlcNAc-S-TEV-H3, purified intermediate resulted from enzymatic extension of GlcNAc-S-TEV-H3 (Table 4, Entry 2):

A 500 µL reaction mixture was set up as follows: to 175 µL pmHS2 protein solution (c = 1.45 mg/mL (5x0.29 mg/mL) in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 6.55 µM final), 5 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 100 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 70 µL 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 buffer, 100 µL GlcNAc-S-TEV-H3 protein solution (3.5 mg/mL (5x0.7 mg/mL) in 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 buffer, 43.1 µM final) and 50 µL UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 5 mM) were added. This was incubated for 14 h at 30 °C, loaded onto a Superdex 75 10/300 analytical column and eluted with 20 mM NaHEPES, 125 mM L-ArgHCl, 6 M Gdn·HCl pH 7.5 buffer. Obtained fractions (0.5 mL each) were analysed by SDS-PAGE. The one containing desired GlcAGlcNAc-S-TEV-H3 was concentrated and buffer exchanged into 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 buffer to obtain 170 µL protein solution, c = 0.41 mg/mL. This was added to

a mixture containing 175 µL pmHS2 protein solution (c = 0.29 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 1.31 µM final), 5 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 100 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 50 µL UDP-GlcNAc (50 mM solution in ddH_2O, to have a final concentration of 5 mM) and it was incubated for 14 h at 30 °C. GlcNAc-addition was monitored by LC-MS analyses after 1, 2 and 14 h (3 µL injected after 50X dilution of a 1 µL aliquot with 0.1 % FA ddH_2O).

Table 4, Entry 3:

To 35 µL pmHS2 protein solution (c = 1.45 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 6.55 µM final), 1 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 20 µL buffer 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5, 14 µL buffer 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5, 20 µL GlcNAc-S-TEV-H3 protein solution (3.5 mg/mL, 43.1 µM in 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 buffer) and 1 µL UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 0.5 mM) were added. It was then incubated for 3 h at 30 °C. Then, 1 µL UDP-GlcNAc (50 mM solution in ddH_2O, to have a final concentration of 0.5 mM) was added and reaction mixture was incubated for 16 h at 30 °C. GlcNAc-addition was monitored by LC-MS analyses after 0.5 and 1 h (3 µL injected after 50X dilution of a 1 µL aliquot with 0.1 % FA ddH_2O).

2.4.4.6 Enzymatic extension of the GlcAGlcNAc-S-TEV-H3 substrate by pmHS2 and formation of the Hep-remodeled H3 neoproteoglycan

Table 4, Entry 4:

2 x 100 µL reaction mixtures were set up. In each case, to 35 µL pmHS2 protein solution (c = 1.45 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer), 1 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 20 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 14 µL 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 buffer and 20 µL GlcAGlcNAc-S-TEV-H3 protein solution (c = 3.5 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 buffer) were added. To each, UDP-GlcNAc (50 mM solution in ddH_2O) was added, and reaction mixtures were incubated for 1 h at 30 °C.

1. 0.5 mM UDP-GlcNAc: 1 µL UDP-GlcNAc (50 mM solution in ddH_2O)
2. 0.05 mM UDP-GlcNAc: 1 µL UDP-GlcNAc (5 mM solution in ddH_2O)

Then, to each reaction mixture, 1 µL UDP-GlcA (50 mM solution in ddH_2O to have a final concentration of 0.5 mM) was added and reaction mixtures were incubated for 16 h at 30 °C.

Enzymatic extension was monitored by LC-MS analyses after 30 and 60 min (3 µL injected after 50X dilution of a 1 µL aliquot with 0.1 % FA ddH_2O) and overall enzymatic extension was monitored by SDS-

PAGE analyses after 16 h (denaturation of a 15 µL aliquot with 5 µL reducing 4xLaemmli buffer at 95 °C for 5 min). Coomassie blue staining was performed according to the manufacturer's protocol and Ab96C10 (primary anti-H3 antibody) followed by A3562 (anti-mouse IgG−alkaline phosphatase secondary antibody, BCIP/NBT substrate) were used for western blot detection.

2.4.4.7 GalT enzymatic extension of GlcNAc-S-TEV-H3 with a Gal residue

2 x 100 µL reaction mixtures were set up. In each case, to 20 µL GlcNAc-S-TEV-H3 protein solution (0.7 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 8.62 µM final), 2 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 5 mM), 63.8 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, and 4 µL UDP-Gal (50 mM solution in ddH_2O, to have a final concentration of 2 mM) were added. To the two reaction mixtures the following reagents were added and then these were incubated for 16 h at 30 °C.

1. 3.2 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer
2. 3.2 µL α-lactalbumin protein solution (6.2 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer)

Enzymatic extension was monitored by LC-MS analyses after 1, 2, 3, 4, 6 hours (3 µL injected after 10X dilution of a 5 µL aliquot with 0.1 % FA ddH_2O).

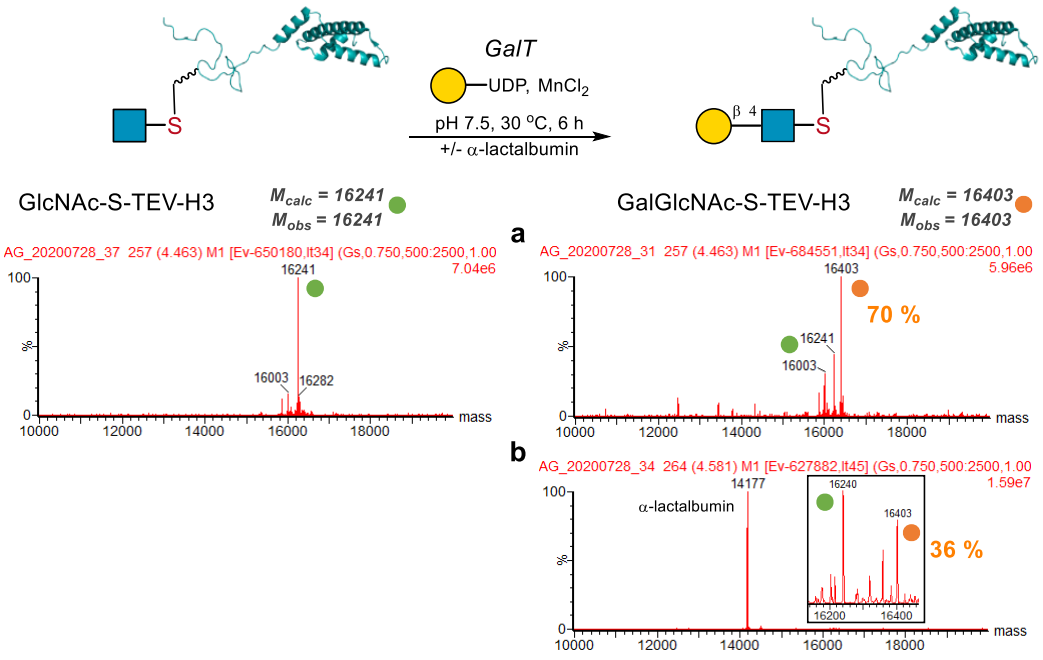


Figure SI 2.28. Intact protein LC-MS analyses showed that GalT transferred one Gal residue to acceptor GlcNAc-S-TEV-H3. The conversion to GalGlcNAc-S-TEV-H3 reached 70 % after six hours, in absence of α-lactalbumin (**a**). The addition of α-lactalbumin caused a decrease in product formation (36 %) and it also hindered the LC-MS identification of the histone species (**b**).

2.4.4.8 ΔpmHS2-catalysed extension of GlcNAc-S-TEV-H3 and GlcNAc-ethyl-S-TEV-H3 with a GlcA residue (comparison to the pmHS2 activity)

4 x 100 µL reaction mixtures were set up in parallel, as follows and incubated at 30 °C for 6 hours. In 1. and 2., to 35 µL pmHS2 protein solution (c = 0.29 mg/mL, in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 1.31 µM final), 1 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 34 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 10 µL UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 5 mM) and 20 µL GlcNAc-S-TEV-H3 protein solution (0.7 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 8.62 µM final) (1.) or 20 µL GlcNAc-ethyl-S-TEV-H3 protein solution (0.7 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 8.62 µM final) (2.) were added. In 3. and 4., to 30.1 µL ΔpmHS2 protein solution (c = 0.29 mg/mL, in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 1.31 µM final), 1 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 38.9 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 10 µL UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 5 mM) and 20 µL GlcNAc-S-TEV-H3 protein solution (0.7 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 8.62 µM final) (3.) or 20 µL GlcNAc-ethyl-S-TEV-H3 protein solution (0.7 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 8.62 µM final) (4.) were added.

Enzymatic extension was monitored by LC-MS analyses after 0.5, 1, 2, 4 and 6 hours (3 µL injected after 10X dilution of a 5 µL aliquot with 0.1 % FA ddH_2O).

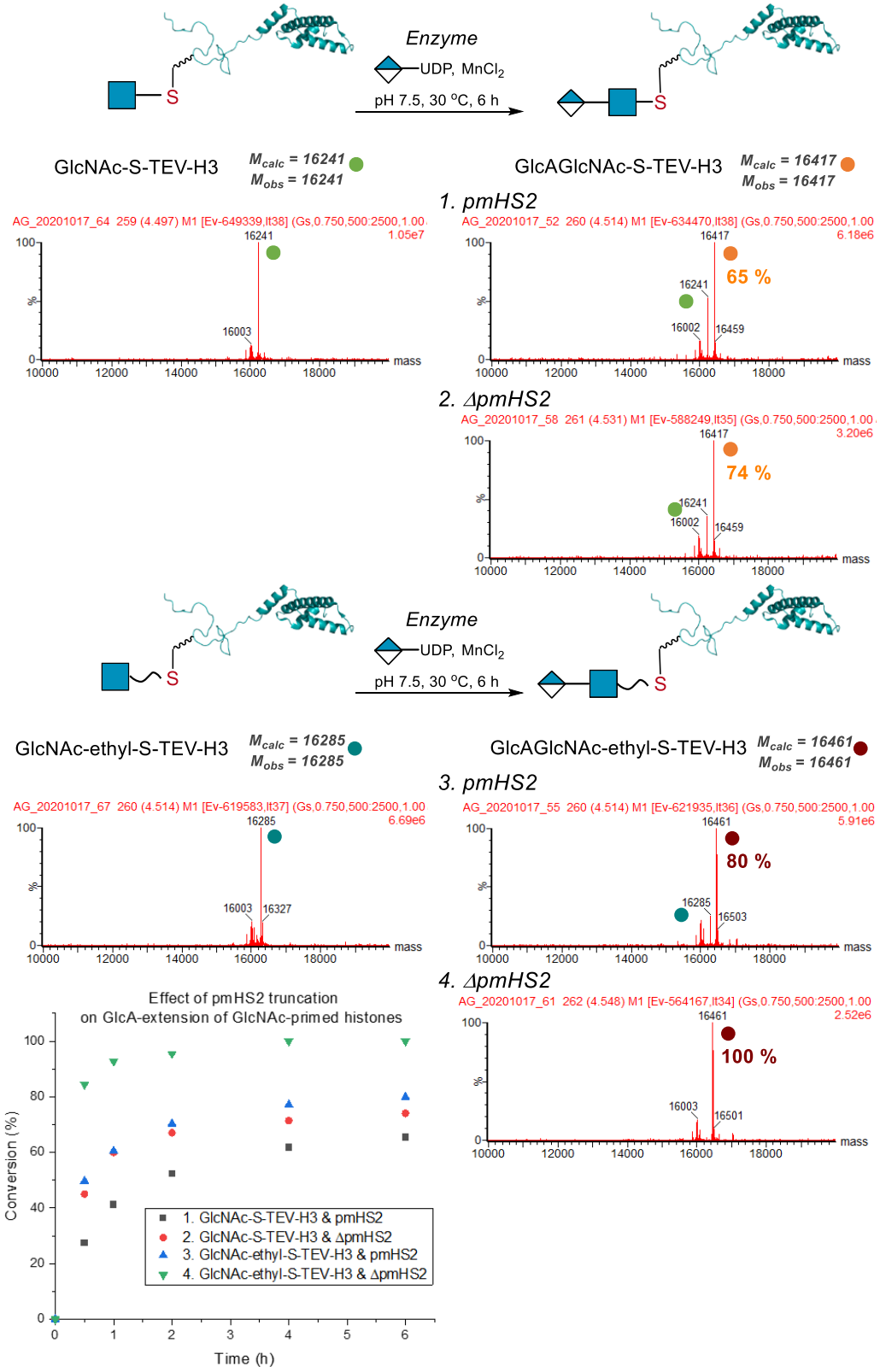


Figure SI 2.29. Intact protein LC-MS analyses showed that ΔpmHS2 had a higher activity than pmHS2 when extending both the GlcNAc-S-TEV-H3 and the GlcNAc-ethyl-S-H3. The conversion to GlcAGlcNAc-S-TEV-H3 reached 65 % when GlcNAc-S-TEV-H3 was incubated with pmHS2 for six hours. When the same starting material was subjected to ΔpmHS2-catalysed GlcA-extension an increase in product formation (74 %) could be observed. The conversion to GlcAGlcNAc-ethyl-S-TEV-H3 reached 80 % when GlcNAc-ethyl-S-TEV-H3 was incubated with pmHS2 for six hours. When the same starting material was subjected to ΔpmHS2-catalysed GlcA-extension complete product formation (100 %) could be observed.

2.4.4.9 pmHS2-catalysed extension of GlcAGlcNAc-S-TEV-H3 with a Glc residue

A 250 µL reaction mixture was set up as follows: to 87.5 µL pmHS2 protein solution (c = 1.45 mg/mL (5x0.29 mg/mL) in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 6.55 µM final), 2.5 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 50 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 35 µL 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 buffer, 50 µL GlcAGlcNAc-S-TEV-H3 protein solution (3.5 mg/mL (5x0.7 mg/mL), in 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 buffer) and 25 µL UDP-Glc (50 mM solution in ddH_2O, to have a final concentration of 5 mM) were added. The reaction mixture was incubated for 3 h at 30 °C and Glc-addition was monitored by LC-MS analyses after 1 hour (3 µL injected after 50X dilution of a 1 µL aliquot with 0.1 % FA ddH_2O).

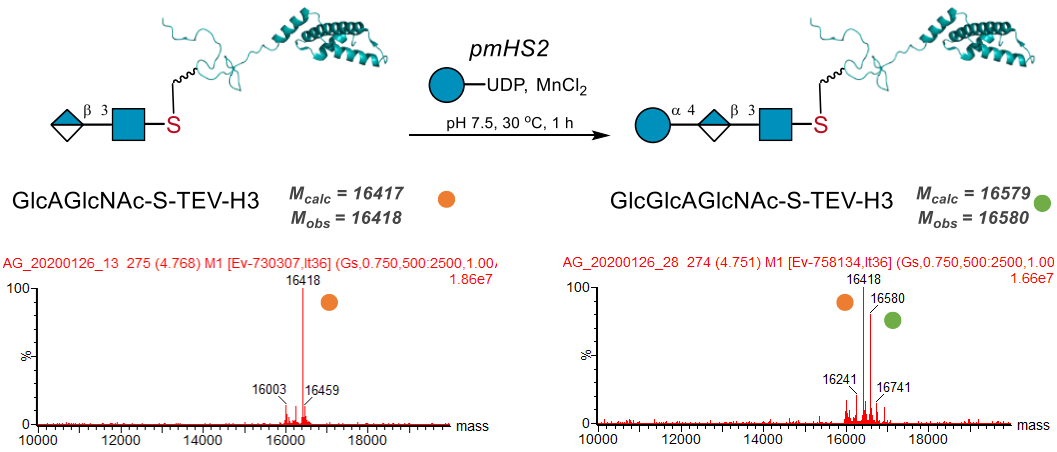


Figure SI 2.30. Intact protein LC-MS analysis showed that pmHS2 was able to catalyse the Glc-extension of GlcNAc-S-TEV-H3.

2.4.5 PmHS2, ΔpmHS2 and GalT extend the primed GlcNAc-ethyl-S-decorin but not the GlcNAc-S-variant

2.4.5.1 Enzymatic GlcA-extension attempt of GlcNAc-S-decorin by pmHS2

After dialysis of GlcNAc-S-decorin in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer:

After the thiyl-glycosylation, obtained GlcNAc-S-decorin was filtered through a PD SpinTrap G-25 column equilibrated with 50 mM NaOAc, 7 M urea pH 6.0 buffer and dialysed overnight in the same buffer. Then, the protein was refolded into 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer and added (c = 1.6 mg/mL, 20 µL) to a mixture of 35 µL pmHS2 protein solution (c = 0.29 mg/mL, in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 1.31 µM final), 1 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 34 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer and 10 µL UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 5 mM). Reaction was monitored by LC-MS analyses after 1, 2, and 4 hours (3 µL injected after 10X dilution of a 5 µL aliquot with ddH_2O) but no GlcA-addition product was identified.

After size-exclusion purification of GlcNAc-S-decorin in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer:

After the thiyl-glycosylation, obtained GlcNAc-S-decorin was filtered through PD SpinTrap G-25 columns equilibrated with 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer and then a 500 µL sample was loaded onto a Superdex 75 10/300 column equilibrated with the same buffer. After elution, fraction A3, 480 µL (c = 0.7 mg/mL) was used to set up 2x100 µL reaction mixtures in parallel, as follows. To 20 µL solution GlcNAc-S-decorin 8.5 µL 20 mM NaHEPES, 125 mM L-ArgHCl, 6 M urea pH 7.5 buffer (A) or 8.5 µL 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer (B) were added and then to both A and B 30.5 µL 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer were added, followed by 1 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 35 µL pmHS2 protein solution (c = 0.29 mg/mL, in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 1.31 µM final) and 10 µL UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 5 mM). Reaction mixtures were incubated at 30 °C for 2 hours and were monitored by LC-MS analyses after 0.5, 1, and 2 hours (3 µL injected after 10X dilution of a 5 µL aliquot with ddH_2O) but no GlcA-addition product was identified.

2.4.5.2 Enzymatic GlcA-extension attempt of GlcNAc-S-decorin by pmHS2 in presence of 0.5 M urea

After the thiyl-glycosylation, obtained GlcNAc-S-decorin was filtered through PD SpinTrap G-25 columns equilibrated with 20 mM NaHEPES, 125 mM L-ArgHCl, 6 M urea pH 7.5 buffer and then a 500 µL sample was loaded onto a Superdex 75 10/300 column equilibrated with the same buffer. After elution, fraction A4 was concentrated to obtain 200 µL (c = 1.4 mg/mL) and then 2x100 µL reaction mixtures were set up in parallel, as follows. To 8.3 µL protein solution (c = 1.4 mg/mL) 45.7 µL (A) or 44.7 µL (B) 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer were added, followed by 1 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM), 35 µL pmHS2 protein solution (c = 0.29 mg/mL, in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 1.31 µM final) and 10 µL UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 5 mM). To B, 1 µL DTT (100 mM solution in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer) was added and both reaction mixtures were incubated at 30 °C for 2 hours. Reaction was monitored by LC-MS analyses after 0.5, 1, and 2 hours (3 µL injected after 10X dilution of a 5 µL aliquot with ddH_2O) but no GlcA-addition product was identified.

2.4.5.3 Enzymatic Gal-extension attempt of GlcNAc-S-decorin by GalT

To 7 µL GalT protein solution (c = 1.4 mg/mL, in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer) 67 µL 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer were added, followed by 2 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 5 mM), 20 µL GlcNAc-S-decorin solution (c = 1.6 mg/mL, in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer) and 4 µL UDP-Gal (50 mM solution in ddH_2O, to have a final concentration of 2 mM) and reaction mixture was incubated at 30 °C for 24 hours. Reaction was monitored by LC-MS analyses after 1, 2, 4 and 24 hours (3 µL injected after 10X dilution of a 5 µL aliquot with ddH_2O) but no Gal-addition product was identified.

2.4.5.4 Enzymatic GlcA-extension attempt of GlcNAc-S-decorin by ΔpmHS2

2 x 100 µL reaction mixtures were set up in parallel, as follows and incubated at 30 °C for 24 hours. To 20 µL GlcNAc-S-decorin solution (c = 1.6 mg/mL, in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer) 5 µL 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer were added followed by 34 µL 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer (A) or 34 µL 20 mM NaHEPES, 125 mM L-ArgHCl, 1.5 M urea pH 7.5 buffer (B). Then, to each reaction mixture, 30 µL ΔpmHS2 protein solution (c = 0.29 mg/mL, in 20 mM NaHEPES, 125 mM L-ArgHCl pH 7.5 buffer, 1.31 µM final), 1 µL $MnCl_2$ (250 mM solution in ddH_2O to have a final concentration of 2.5 mM) and 10 µL UDP-GlcA (50 mM solution in ddH_2O, to have a final concentration of 5 mM) were also added. Reactions were monitored by LC-MS analyses after 1, 2, 4

and 24 hours (3 µL injected after 10X dilution of a 5 µL aliquot with ddH_2O) but no GlcA-addition product was identified.

2.4.5.5 Evaluation of the glycosyl transferase activities of pmHS2, ΔpmHS2 and GalT using the GlcNAc-ethyl-S-decorin substrate

After the thiyl-glycosylation, obtained GlcNAc-ethyl-S-decorin was filtered through a PD SpinTrap G-25 column equilibrated in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 and dialysed overnight in the same buffer. Then, three reaction mixtures were set up in parallel, as follows and incubated at 30 °C for 6 hours. In each case, to 20 µL starting material, $MnCl_2$ (250 mM solution in ddH_2O), corresponding UDP-sugar (50 mM solution in ddH_2O) and 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer were added as described in the table.

Reaction	Enzyme	$MnCl_2$ (µl)	UDP-sugar (µl)	Buffer (µl)
1	35 µl pmHS2 (0.29 mg/ml)	1	10	34
2	30.1 µl ΔpmHS2 (0.29 mg/ml)	1	10	38.9
3	7 µl GalT (1.4 mg/ml)	2	4	67

Figure SI 2.31. Enzymes investigated for the extension of GlcNAc-ethyl-S-decorin: pmHS2, ΔpmHS2 and GalT.

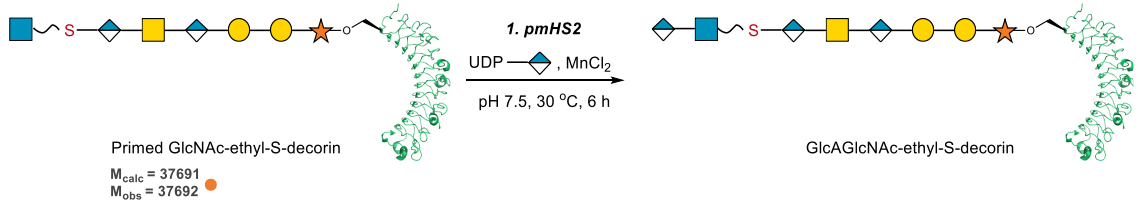


Figure SI 2.32. Intact protein LC-MS analyses of the GlcA-extension achieved by pmHS2 on the GlcNAc-ethyl-S-decorin substrate. Excellent conversion to the desired GlcAGlcNAc-ethyl-S-decorin was achieved (85 %) after six hours.

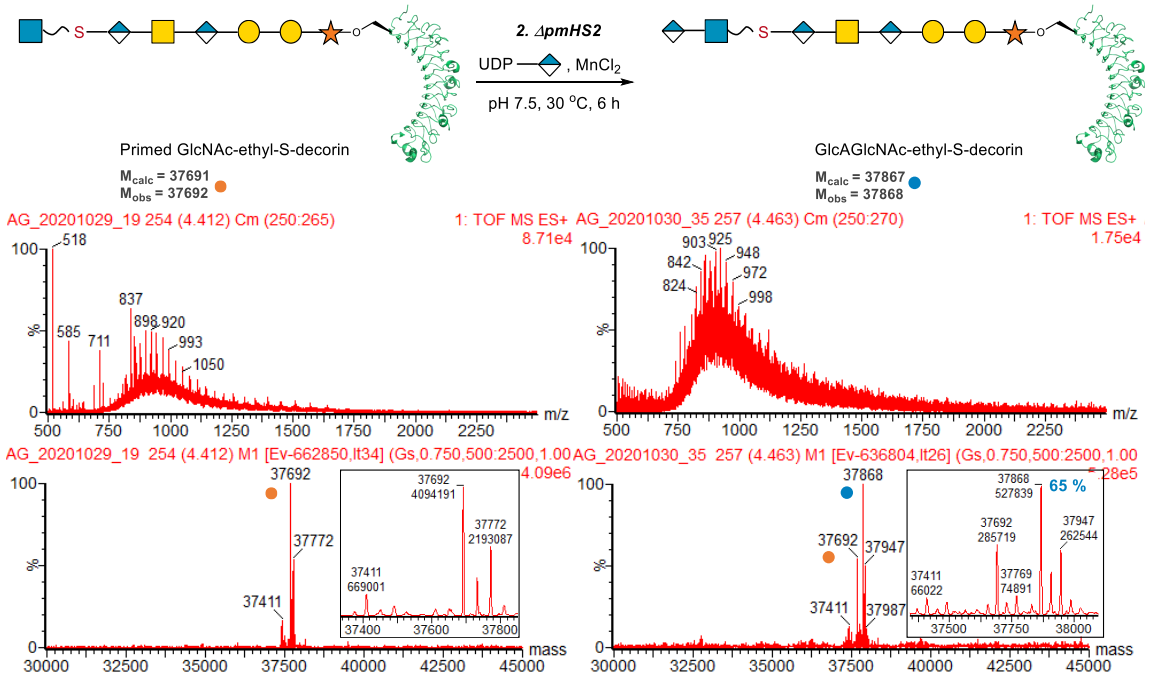


Figure SI 2.33. Intact protein LC-MS analyses of the GlcA-extension achieved by ΔpmHS2 on the GlcNAc-ethyl-S-decorin substrate. Good conversion to the desired GlcAGlcNAc-ethyl-S-decorin was achieved (65 %) after six hours.

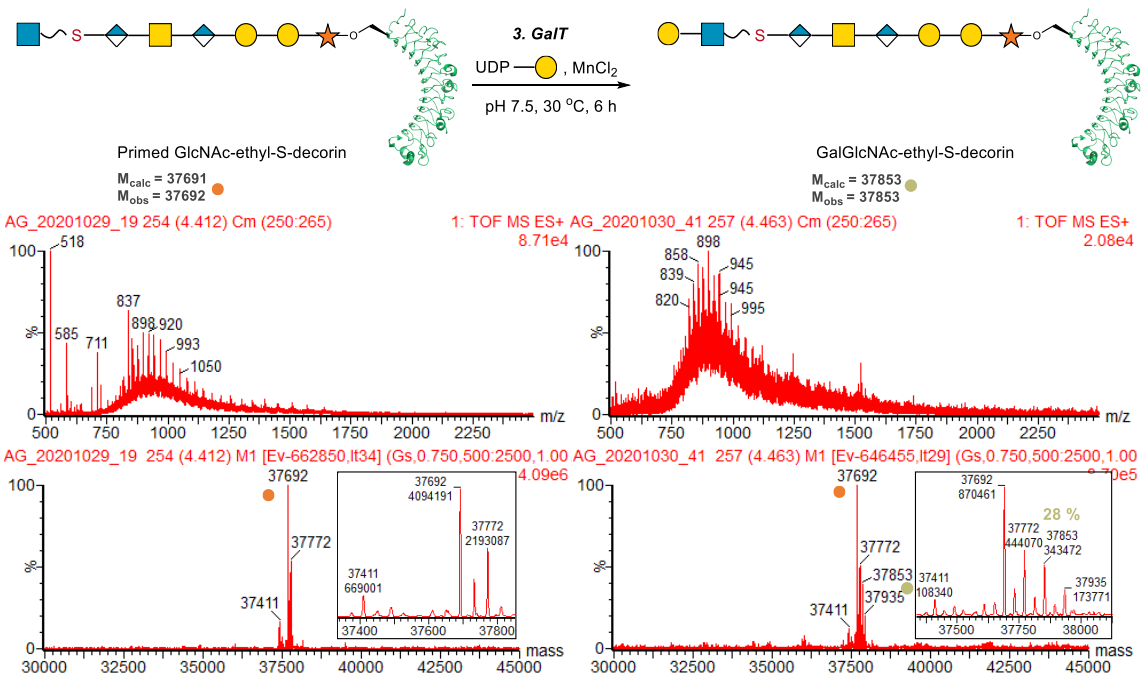


Figure SI 2.34. Intact protein LC-MS analyses of the Gal-extension achieved by GalT on the GlcNAc-ethyl-S-decorin substrate. Limited conversion to the desired GalGlcNAc-ethyl-S-decorin was achieved (28 %) after six hours.

2.4.5.6 PmHS2 GlcA-transferase activity on GlcNAc-ethyl-S-decorin at lower UDP-GlcA concentrations (0.5 mM and 0.05 mM)

Two reaction mixtures were set up in parallel, as follows and incubated at 30 °C for 20 hours. To 35 µL pmHS2 protein solution (c = 0.29 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer), 1 µL $MnCl_2$ (250 mM solution in ddH_2O), 43 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 20 µL GlcNAc-ethyl-S-decorin protein solution (pre-dialysed in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, after thiyl-glycosylation) and 1 µL UDP-GlcA (50 mM/5 mM solution in ddH_2O) were added. GlcA-addition was monitored by LC-MS analyses after 0.5, 1, 2, 3, 4 and 20 h (3 µL injected after 10X dilution of a 5 µL aliquot with ddH_2O).

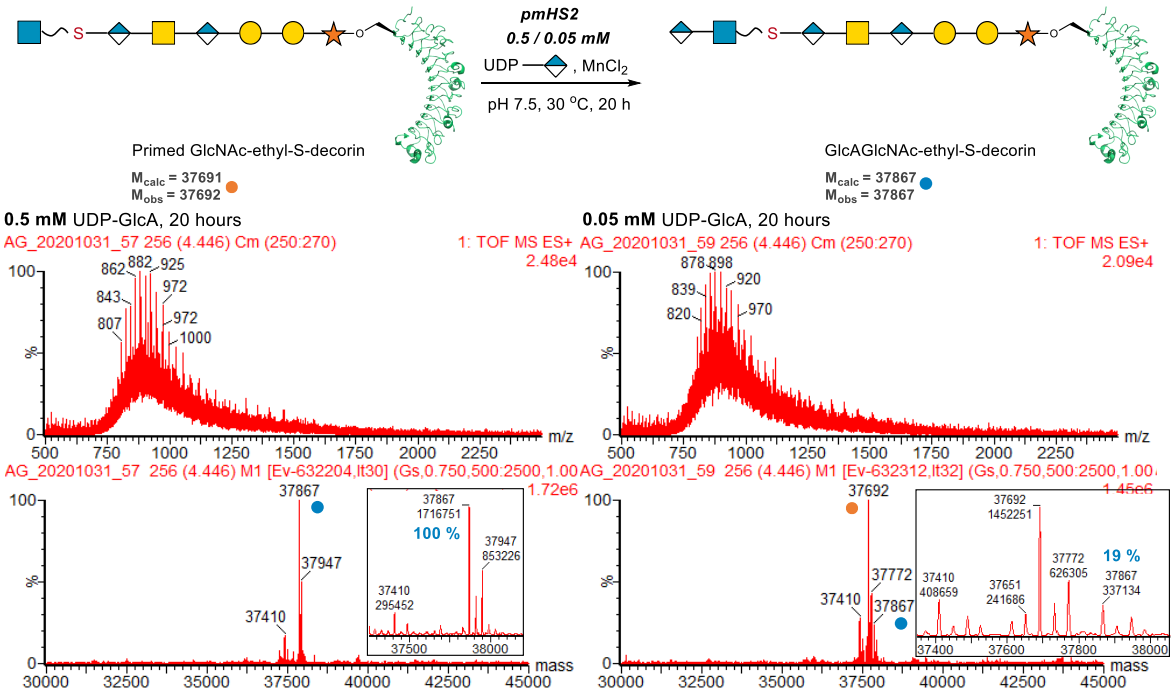


Figure SI 2.35. Intact protein LC-MS analyses of the GlcA-extension achieved by pmHS2 on the GlcNAc-ethyl-S-decorin, with 0.5 / 0.05 mM UDP-GlcA. Complete conversion to the desired GlcAGlcNAc-ethyl-S-decorin was achieved (100 %) when incubating with 0.5 mM UDP-GlcA for 20 hours, whereas a much lower concentration of sugar donor (0.05 mM) generated limited product formation (19 %).

2.4.5.7 Enzymatic extension of GlcNAc-ethyl-S-decorin with 1.7 mM UDP-GlcA, 1.7 mM UDP-GlcNAc and additional 0.17 mM UDP-GlcA

To 7 µL pmHS2 protein solution (c = 1.45 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer), 1 µL $MnCl_2$ (250 mM solution in ddH_2O), 1 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer, 20 µL GlcNAc-ethyl-S-decorin protein solution (pre-dialysed in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer) and 1 µL UDP-GlcA (50 mM solution in ddH_2O) were added and the reaction mixture was incubated for 14 hours at 30 °C. Then, 1 µL UDP-GlcNAc (50 mM solution in ddH_2O) was added, the reaction was further incubated at 30 °C and monitored by LC-MS analysis after 15 and 30 minutes. Then, 1 µL UDP-GlcA (5 mM solution in ddH_2O) was added, the reaction mixture was incubated under the same conditions and monitored by LC-MS analysis after 15 and 30 minutes and by SDS-PAGE analysis after 16 hours.

2.4.5.8 Enzymatic extension of GlcNAc-ethyl-S-decorin with 1.7 mM UDP-GlcA, 0.17 mM UDP-GlcNAc, 0.17 mM UDP-GlcA and additional 0.17 mM UDP-GlcNAc

To 52.5 µL pmHS2 protein solution (c = 1.45 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer), 7.5 µL $MnCl_2$ (250 mM solution in ddH_2O), 7.5 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5

buffer, 150 µL GlcNAc-ethyl-S-decorin protein solution (pre-dialysed in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer) and 7.5 µL UDP-GlcA (50 mM solution in ddH_2O) were added and the reaction mixture was incubated for 16 hours at 30 °C. Then, 7.5 µL UDP-GlcNAc (5 mM solution in ddH_2O) were added, and the reaction was further incubated at 30 °C for 15 minutes. Then, 7.5 µL UDP-GlcA (5 mM solution in ddH_2O) was added, and the reaction mixture was incubated under the same conditions for 60 minutes. Again, 7.5 µL UDP-GlcNAc (5 mM solution in ddH_2O) were added, and the reaction mixture was further incubated for 60 minutes. Enzymatic extension was monitored by LC-MS analysis after 16 hours from the first UDP-GlcA addition, 15 min from the first UDP-GlcNAc addition, after 30 and 60 minutes from the second UDP-GlcA addition and after 30 and 60 minutes from the second UDP-GlcNAc addition. Reaction progression was also monitored by SDS-PAGE analysis after 60 minutes from the second UDP-GlcA addition and after 60 minutes from the second UDP-GlcNAc addition.

2.4.5.9 Negative control reactions for the pmHS2 enzymatic extension of GlcNAc-ethyl-S-decorin

Two negative control reactions were set up in parallel, as follows and the reaction mixtures were incubated at 30 °C for 24 hours:

1: To 20 µL core decorin protein solution (pre-dialysed in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer), 1 µL $MnCl_2$ (250 mM solution in ddH_2O), 35 µL pmHS2 protein solution (c = 0.29 mg/mL in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer), 34 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer and 10 µL UDP-GlcA (50 mM in ddH_2O) were added.

2: To 20 µL core decorin protein solution (pre-dialysed in 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer), 1 µL $MnCl_2$ (250 mM solution in ddH_2O), 69 µL 20 mM NaHEPES, 125 mM L-ArgHCl, pH 7.5 buffer and 10 µL UDP-GlcA (50 mM in ddH_2O) were added.

GlcA-addition was checked by LC-MS analyses after 24 h (3 µL injected after 10X dilution of a 5 µL aliquot with ddH_2O).

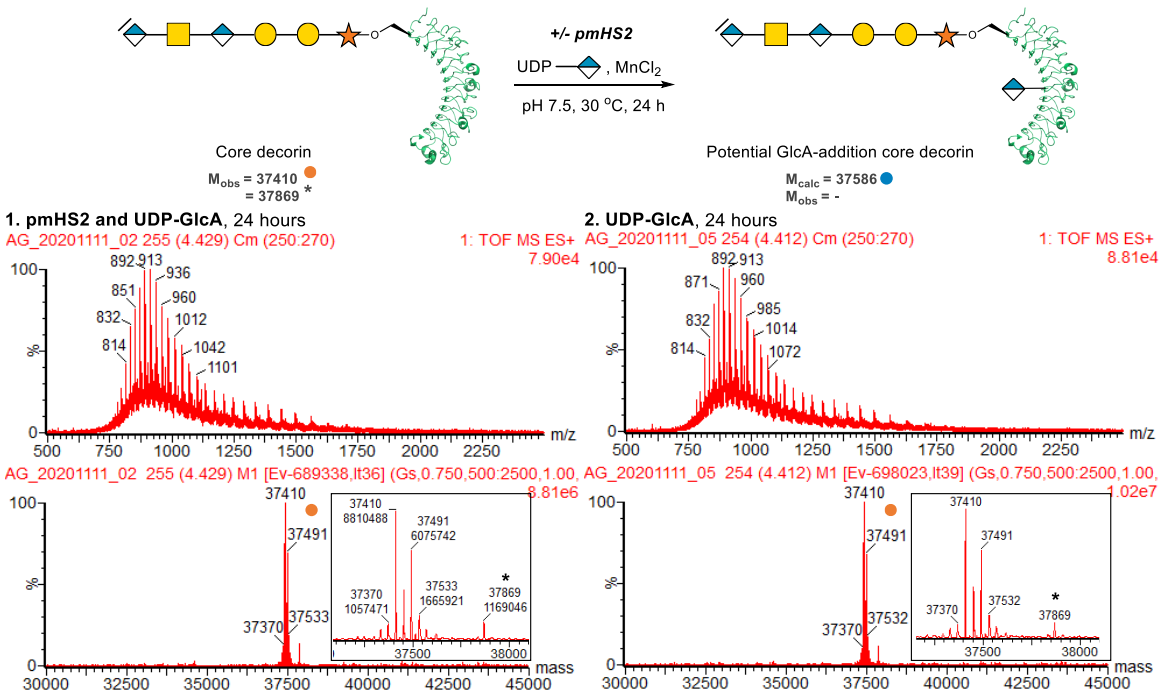
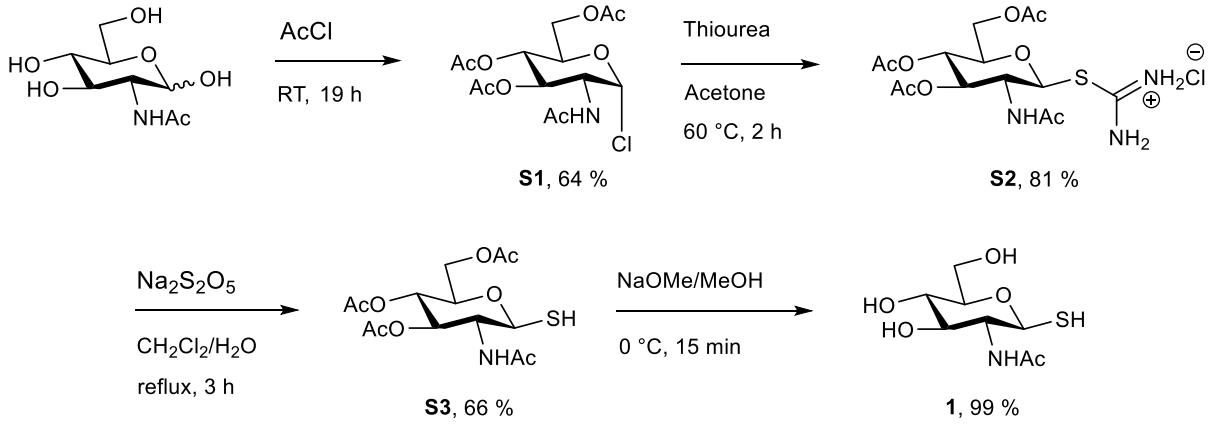


Figure SI 2.36. Intact protein LC-MS analyses did not show any GlcA-addition when a GlcNAc-primer had not been installed on decorin. Incubation of core decorin with pmHS2 and UDP-GlcA did not reveal any GlcA-extension (**1**) and the same was observed when the core PG was incubated with UDP-GlcA alone (**2**). Molecular mass 37869 (*) was an impurity present in the starting material, and it was presumably caused by the incomplete digestion of the GAG, that stalled at the octasaccharide-PG instead of the usual hexasaccharide-PG.

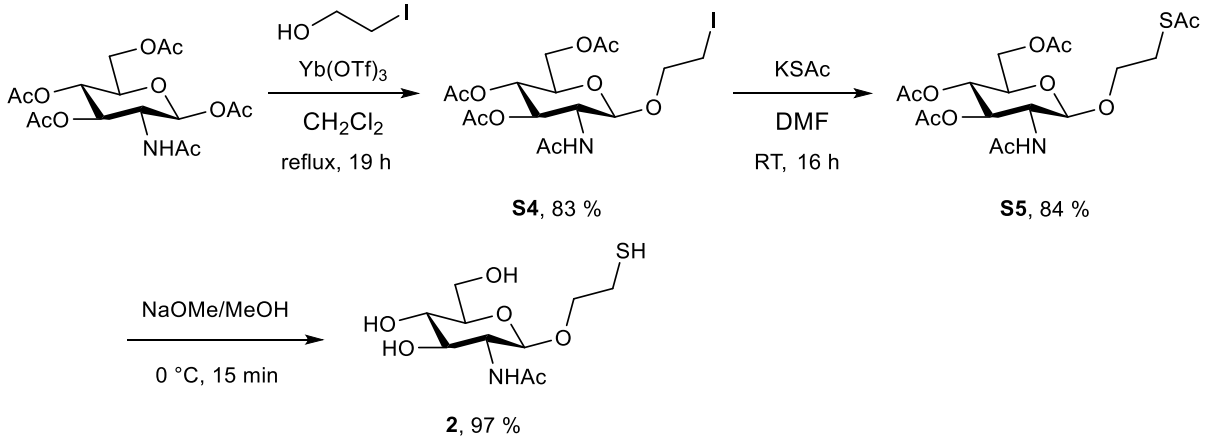
2.4.6 Quantification of the relative protein levels of Peg3, Bec1, LC3 and ICAM-1, VCAM-1, IL-1α

Expression analysis by RT-qPCR was carried out on subconfluent six-well plates seeded with ∼2 × 10^5 HUVEC cells treated for 6 h with 100 or 200 nM decorin protein core in full serum Basal Endothelial Medium (LifeCell). Following incubation, cells were lysed directly in 1 mL of TRIzol reagent (Invitrogen) to extract total RNA. Subsequently, 1 μg of total RNA was annealed with oligo (dT18–20) primers, and cDNA was synthesized with the aid of SuperScript Reverse Transcriptase II (Invitrogen). Amplicons representing target genes and the endogenous housekeeping gene, 18S ribosomal RNA, were amplified in quadruplicate reactions using the Brilliant SYBR Green Master Mix II (Agilent Technologies). All samples were run on the QuantStudio 7 Flex Real Time PCR System (ThermoFischer Scientific), and cycle number (Ct) was recorded for each reaction. Fold change determinations were made using the Comparative Ct method for expression analysis. Delta Ct (ΔCt) values represent gene expression levels to normalized 18S ribosomal RNA for each reaction. Delta Delta Ct (ΔΔCt) values then represent the experimental cDNA (for example, samples treated with 200 nM decorin protein core) minus the corresponding gene levels of the calibrator sample (control samples). Fold changes were calculated using the double ΔCt method ($2^{-\Delta\Delta CT}$) ± SEM. Data derive from two parallel trials run in quadruplicate for each gene of interest.

2.4.7 Chemical synthesis of GlcNAc-SH and GlcNAc-ethyl-SH

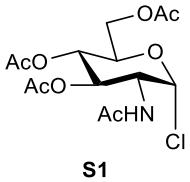


Scheme 1. Synthesis of 2-acetamido-2-deoxy-1-thio-β-D-glucose (GlcNAc-SH).



Scheme 2. Synthesis of 1-(2-thioethyl)-2-acetamido-2-deoxy-β-D-glucose (GlcNAc-ethyl-SH).

3,4,6-tri-*O*-acetyl-2-*N*-acetyl-2-deoxy-β-D-glucopyranosyl chloride (S1)



S1

N-Acetyl-D-glucosamine (7.0 g, 31.7 mmol, 1 equiv.) was added to a solution of acetyl chloride (70.2 mL, 31.1 equiv.) under N_2 atmosphere and the reaction mixture was stirred for 19 hours at room temperature. The slightly discoloured solution mixture was then diluted with $CHCl_3$ (70 mL) and poured

under strong stirring onto ice (28.0 g) and H_2O (20 mL). The layers were separated and the organic layer was transferred to a beaker containing ice (14.0 g) and saturated solution of $NaHCO_3$ (87.5 mL). The mixture was transferred into a separatory funnel and shaken until gas production ended. The layers were separated and the organic one was dried over anhydrous Na_2SO_4, filtered and concentrated *in vacuo*. The residue was then purified by flash column chromatography (petroleum ether/EtOAc 2:3) to give the desired glucopyranosyl chloride **S1** as white solid (7.4 g, 20.2 mmol, 63.8 %).

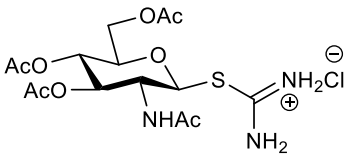
^{1}H NMR (400 MHz, $CDCl_3$): δ 6.18 (d, 1H, $J_{1,2}$= 3.7 Hz, H-1), 5.82 (d, 1H, $J_{NH,2}$ = 8.7 Hz, NH), 5.32 (dd, 1H, $J_{3,2}$ = 10.8 Hz, $J_{3,4}$ = 9.4 Hz, H-3), 5.21 (dd, 1H, $J_{4,5}$ = 10.0 Hz, $J_{4,3}$ = 9.4 Hz, H-4), 4.53 (ddd, 1H, $J_{2,3}$ = 10.7 Hz, $J_{2,NH}$ = 8.8 Hz, $J_{2,1}$ = 3.7 Hz, H-2), 4.28 (dd, 1H, $J_{6a,6b}$ = 12.4 Hz, $J_{6a,5}$ = 4.1 Hz, H-6a), 4.25 (ddd, 1H, $J_{5,4}$ = 10.0 Hz, $J_{5,6a}$ = 4.1 Hz, , $J_{5,6b}$ = 1.8 Hz, H-5), 4.13 (dd, 1H, $J_{6b,6a}$ = 12.3 Hz, $J_{6b,5}$ = 1.8 Hz, H-6b), 2.10, 2.05, 2.05 (3s, 9H, 3 x $COCH_3$), 1.98 (s, 3H, $NHCOCH_3$).

^{13}C NMR (101 MHz, $CDCl_3$): δ 171.6, 170.7, 170.2 (3 x $\underline{C}OCH_3$), 169.3 ($NH\underline{C}OCH_3$), 93.8 (C-1), 71.0 (C-5), 70.3 (C-3), 67.1 (C-4), 61.3 (C-6), 53.6 (C-2), 23.2 ($NHCO\underline{C}H_3$), 20.8, 20.7 (3 x $CO\underline{C}H_3$).

ESI^+ LRMS: calculated for $C_{14}H_{20}ClNO_8Na$ $[M+Na]^+$ 388.1, found 388.1.

Data consistent with that previously reported.[18]

2-acetamido-3,4,6-tri-O-acetyl-2-deoxy-β-D-glucopyranosyl-1-isothiouronium chloride (S2)



S2

Intermediate **S1** (7.0 g, 19.1 mmol, 1 equiv.) and thiourea (2.18 g, 28.7 mmol, 1.5 equiv.) were dissolved in acetone (60 mL) under N_2 atmosphere and the reaction mixture was stirred at 60 °C. After 2 hours, a white precipitate was formed and filtered off. The filtrate was returned to reflux and this process was repeated until the solid ceased to precipitate. The solids were combined and re-crystallized from petroleum ether/acetone 1:1 to give the desired glucopyranosyl isothiouronium chloride **S2** as white solid (6.8 g, 15.4 mmol, 80.6 %).

^{1}H NMR (400 MHz, DMSO-d_6): δ 9.32 (s, 2H, NH_2Cl), 9.13 (s, 2H, NH_2), 8.36 (d, 1H, $J_{NH,2}$ = 9.3 Hz, NH), 5.61 (d, 1H, $J_{1,2}$= 10.3 Hz, H-1), 5.09 (t, 1H, J = 9.8 Hz, H-3), 4.90 (t, 1H, J = 9.8 Hz, H-4), 4.21-4.09 (m,

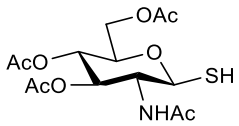
2H, $J_{6a,6b}$ = 12.0 Hz, $J_{6a,5}$ = 4.8 Hz, H-5/H-6a), 4.07-3.92 (m, 2H, $J_{6b,6a}$ = 11.9 Hz, $J_{6b,5}$ = 1.7 Hz, H-2/H-6b), 1.97, 1.94, 1.90 (3s, 9H, 3 x $COCH_3$), 1.77 (s, 3H, $NHCOCH_3$).

^{13}C NMR (101 MHz, DMSO-d_6): δ 183.9 ($\underline{C}NH_2NH_2Cl$), 170.1, 170.0, 169.7 (3 x $\underline{C}OCH_3$), 169.3 ($NH\underline{C}OCH_3$), 80.7 (C-1), 74.8 (C-5), 72.7 (C-3), 67.9 (C-4), 61.6 (C-6), 51.3 (C-2), 22.6 ($NHCO\underline{C}H_3$), 20.6, 20.5, 20.4 (3 x $CO\underline{C}H_3$).

ESI^+ LRMS: calculated for $C_{15}H_{24}N_3O_8S$ $[M\text{-}Cl]^+$ 406.1, found 406.1.

Data consistent with that previously reported.[18]

2-acetamido-3,4,6-tri-*O*-acetyl-2-deoxy-1-thio-β-D-glucopyranose (S3)



S3

Compound **S2** (2.2 g, 5.0 mmol, 1 equiv.) and sodium metabisulfite (1.6 g, 8.4 mmol, 1.7 eq.) were dissolved in CH_2Cl_2/H_2O (35 mL:15 mL). The mixture was stirred under N_2 atmosphere, at reflux, for 2 hours when TLC (100 % EtOAc) indicated formation of the product (R_f > 0.0) with complete consumption of the starting material (R_f = 0.0). Then, the mixture was allowed to reach room temperature and the layers were separated. The aqueous layer was re-extracted with CH_2Cl_2 (2 x 25 mL) and the combined organic layers were washed with H_2O (25 mL) and brine (25 mL), dried over anhydrous $MgSO_4$, filtered and concentrated under reduced pressure. The crude was purified by flash column chromatography (petroleum ether/EtOAc 3:7) to give compound **S3** as white solid (1.2 g, 3.3 mmol, 66.0 %).

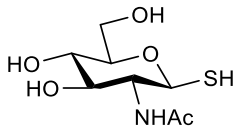
^{1}H NMR (400 MHz, $CDCl_3$): δ 5.60 (d, 1H, $J_{NH,2}$ = 9.5 Hz, NH), 5.12 (t, 1H, $J_{4,5}$ = 9.4 Hz, $J_{4,3}$ = 9.4 Hz, H-4), 5.07 (t, 1H, $J_{3,2}$ = 9.6 Hz, $J_{3,4}$ = 9.6 Hz, H-3), 4.57 (t, 1H, $J_{1,2}$ = 9.7 Hz, $J_{1,SH}$ = 9.7 Hz, H-1), 4.24 (dd, 1H, $J_{6a,6b}$ = 12.4 Hz, $J_{6a,5}$ = 4.9 Hz, H-6a), 4.15 – 4.08 (m, 2H, $J_{6b,6a}$ = 12.4 Hz, $J_{6b,5}$ = 2.3 Hz, H-2/ H-6b), 3.68 (ddd, 1H, $J_{4,5}$ = 9.7 Hz, $J_{5,6a}$ = 2.3 Hz, $J_{5,6b}$ = 4.8 Hz, H-5), 2.57 (d, 1H, $J_{SH,1}$ = 9.4 Hz, SH), 2.10, 2.04, 2.02 (3s, 9H, 3 x $OCOCH_3$), 1.98 (s, 3H, $NHCOCH_3$).

^{13}C NMR (101 MHz, $CDCl_3$): δ 171.4, 170.9, 170.5 (3 x $\underline{C}OCH_3$), 169.4 ($NH\underline{C}OCH_3$), 80.5 (C-1), 76.5 (C-5), 73.7 (C-3), 68.2 (C-4), 62.3 (C-6), 57.0 (C-2), 23.5 ($NHCO\underline{C}H_3$), 20.9, 20.8, 20.7 (3 x $OCO\underline{C}H_3$).

ESI^- LRMS: calculated for $C_{14}H_{20}NO_8S$ $[M\text{-}H]^-$ 362.1, found 362.1

Data consistent with that previously reported.[18]

2-acetamido-2-deoxy-1-thio-β-D-glucose (1)



1

A solution of sodium methoxide (25 % wt/wt solution, 1.0 mL) was added to a solution of protected sugar **S3** (1.2 g, 3.3 mmol, 1 equiv.) in dry MeOH (12 mL) at 0 °C, under N_2 atmosphere. The reaction mixture was stirred for 15 min at room temperature when TLC (100 % EtOAc) indicated formation of the product (R_f = 0.0) with complete consumption of the starting material (R_f > 0.0). Then, the pH of the reaction mixture was neutralized with activated DOWEX® 50WX8-200. The mixture was then filtered and concentrated i*n vacuo* to afford unprotected compound **1** (770 mg, 3.25 mmol, 98.5 %) as a white solid. The product was used without further purification.

^{1}H NMR (400 MHz, D_2O): δ 4.70 (d, 1H, $J_{1,2}$ = 10.2 Hz, H-1), 3.91 (dd, 1H, $J_{6a,6b}$ = 12.4 Hz, $J_{6a,5}$ = 1.6 Hz, H-6a), 3.76 (app t, 1H, $J_{2,3}$ = 9.8 Hz, $J_{2,1}$ = 10.1 Hz, H-2), 3.72 (ddd, 1H, $J_{6b,6a}$ = 12.4 Hz, $J_{6b,5}$ = 4.1 Hz, $J_{6b,4}$ = 1.6 Hz, H-6b), 3.53 (ddd, 1H, $J_{3,2}$ = 9.7 Hz, $J_{3,4}$ = 5.8 Hz, $J_{3,5}$ = 3.3 Hz, H-3), 3.50-3.46 (m, 2H, H-4, H-5), 2.06 (s, 3H, $NHCOCH_3$).

^{13}C NMR (101 MHz, D_2O): δ 174.6 ($NH\underline{C}OCH_3$), 80.2 (C-5), 79.0 (C-1), 74.9 (C-3), 69.6 (C-4), 60.7 (C-6), 57.9 (C-2), 22.2 ($NHCO\underline{C}H_3$).

IR (neat): $\tilde{\nu}$ (cm^{-1}) 3483, 3287(NH, OH), 2535 (SH), 1633, 1544 (C=O), 1422, 1373, 1315, 1096, 1050, 1003.

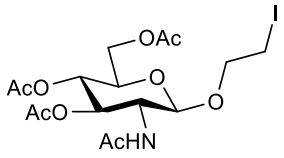
ESI^- LRMS: calculated for $C_8H_{14}NO_5S$ $[M-H]^-$ 236.1, found 236.1.

mp: 178-179 °C, lit.[98] value: 177-179 °C.

$[\alpha]_D^{25}$ = - 6.5 (c = 0.6, MeOH), lit.[99] value: $[\alpha]_D^{20}$ = -6.0 (c = 1, MeOH).

Data consistent with that previously reported.[18,98,99]

1-(2-iodoethyl)-2-acetamido-2-deoxy-β-D-glucopyranosyl 1,3,4,6-tetraacetate (S4)



S4

To a solution of 2-acetamido-2-deoxy-β-D-glucopyranose 1,3,4,6-tetraacetate (1.4 g, 3.6 mmol, 1 equiv.) and 2-iodoethanol (1.12 mL, 14.4 mmol, 4 equiv.) in dry CH_2Cl_2 (42 mL) ytterbium(III) trifluoromethanesulfonate (676 mg, 1.1 mmol, 0.3 equiv.) was added and the reaction mixture was stirred for 19 hours at reflux, under N_2 atmosphere. The next day, t.l.c. (EtOAc) showed complete conversion. Thus, the solution mixture was washed with saturated solution of $NaHCO_3$ (3 x 45 mL) and brine (25 mL), dried over anhydrous $MgSO_4$, filtered and concentrated *in vacuo*. The crude was purified by flash column chromatography (petroleum ether/EtOAc 1:9) to give compound **S4** as white solid (1.52 g, 3.0 mmol, 83.3 %).

^{1}H NMR (400 MHz, $CDCl_3$): δ 5.61 (d, 1H, $J_{NH,2}$ = 8.7 Hz, NH), 5.31 (dd, 1H, $J_{3,2}$ = 10.7 Hz, $J_{3,4}$ = 9.3 Hz, H-3), 5.06 (dd, 1H, $J_{4,5}$ = 10.0 Hz, $J_{4,3}$ = 9.2 Hz, H-4), 4.76 (d, 1H, $J_{1,2}$ = 8.4 Hz, H-1), 4.25 (dd, 1H, $J_{6a,6b}$ = 12.3 Hz, $J_{6a,5}$ = 4.8 Hz, H-6a), 4.17-4.02 (dd, 1H, $J_{6b,6a}$ = 12.2 Hz, $J_{6b,5}$ = 2.6 Hz, H-6b; ddd, 1H, J_{gem} = 11.5 Hz, $J_{A',B}$ = 6.6 Hz, $J_{A',B'}$ = 5.1 Hz, H-A'), 3.85 (ddd, 1H, $J_{2,3}$ = 10.6 Hz, $J_{2,NH}$ = 8.6 Hz, $J_{2,1}$ = 8.5 Hz, H-2), 3.78 (ddd, 1H, J_{gem} = 11.4 Hz, $J_{A,B}$ = 7.6 Hz, $J_{A,B'}$ = 6.7 Hz, H-A), 3.71 (ddd, 1H, $J_{5,4}$ = 10.0 Hz, $J_{5,6a}$ = 4.8 Hz, $J_{5,6b}$ = 2.5 Hz, H-5), 3.36-3.19 (m, 2H, H-B/H-B'), 2.09, 2.03, 2.02 (3s, 9H, 3 x $OCOCH_3$), 1.98 (s, 3H, $NHCOCH_3$).

^{13}C NMR (101 MHz, $CDCl_3$): δ 171.0, 170.8, 170.6 (3 x $\underline{C}OCH_3$), 169.5 ($NH\underline{C}OCH_3$), 100.9 (C-1), 72.3 (C-3), 72.1 (C-5), 70.3 (C-A), 68.7 (C-4), 62.2 (C-6), 54.9 (C-2), 23.7 ($NHCO\underline{C}H_3$), 20.9 ($CO\underline{C}H_3$), 20.8 (2 x $CO\underline{C}H_3$), 3.0 (C-B).

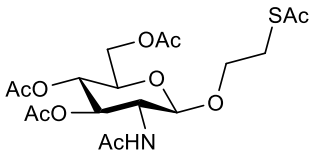
IR (neat): $\tilde{\nu}$ (cm^{-1}) 1741, 1673, 1573 (C=O), 1366, 1226, 1030.

ESI$^+$ LRMS: calculated for $C_{16}H_{24}INO_9Na$ $[M+Na]^+$ 524.0, found: 524.0.

mp: 165-166 °C.

$[\alpha]_D^{25}$ = - 9.0 (c = 0.6, $CHCl_3$).

1-(2-(thioacetate)-ethyl)-2-acetamido-2-deoxy-β-D-glucopyranosyl 1,3,4,6-tetraacetate (S5)



S5

A solution of 1-(2-iodoethyl)-2-acetamido-2-deoxy-β-D-glucopyranosyl 1,3,4,6-tetraacetate (1.5 g, 3.0 mmol, 1 equiv.) and potassium thioacetate (684 mg, 6.0 mmol, 2 equiv.) in dry DMF (10.5 mL) was stirred for 16 hours at room temperature, under N_2 atmosphere. After completion of the reaction, CH_2Cl_2 (90 mL) was added to the reaction mixture and the mixture was washed with H_2O (45 mL). The aqueous phase was re-extracted with CH_2Cl_2 (90 mL). Thereafter, the combined organic layers were washed with brine (45 mL), dried over anhydrous $MgSO_4$, filtered and concentrated *in vacuo*. The crude was purified by flash column chromatography (toluene/EtOAc 2:8) to give compound **S5** as white solid (1.13 g, 2.5 mmol, 83.8 %).

^{1}H NMR (400 MHz, $CDCl_3$): δ 5.51 (d, 1H, $J_{NH,2}$ = 8.7 Hz, NH), 5.28 (dd, 1H, $J_{3,2}$ = 10.5 Hz, $J_{3,4}$ = 9.2 Hz, H-3), 5.06 (dd, 1H, $J_{4,5}$ = 10.0 Hz, $J_{4,3}$ = 9.3 Hz, H-4), 4.70 (d, 1H, $J_{1,2}$ = 8.3 Hz, H-1), 4.25 (dd, 1H, $J_{6a,6b}$ = 12.3 Hz, $J_{6a,5}$ = 4.8 Hz, H-6a), 4.13 (dd, 1H, $J_{6b,6a}$ = 12.3 Hz, $J_{6b,5}$ = 2.5 Hz, H-6b), 3.95 (dt, 1H, , J_{gem} = 10.6 Hz, $J_{A',B}$ = 6.1 Hz, H-A'), 3.85 (ddd, 1H, $J_{2,3}$ = 10.6 Hz, $J_{2,NH}$ = 8.6 Hz, $J_{2,1}$ = 8.4 Hz, H-2), 3.70 (ddd, 1H, $J_{5,4}$ = 10.0 Hz, $J_{5,6a}$ = 4.8 Hz, $J_{5,6b}$ = 2.5 Hz, H-5), 3.64 (ddd, 1H, J_{gem} = 10.6 Hz, $J_{A,B'}$ = 7.1, $J_{A,B}$ = 6.2 Hz, H-A), 3.20-3.09 (m, 1H, H-B), 3.08-2.96 (m, 1H, H-B'), 2.33 (s, 3H, $SCOCH_3$), 2.08, 2.03, 2.02 (3s, 9H, 3 x $OCOCH_3$), 1.97 (s, 3H, $NHCOCH_3$);

^{13}C NMR (101 MHz, $CDCl_3$): δ 195.5 (S-$\underline{C}OCH_3$), 171.0, 170.8, 170.3 (3 x $\underline{C}OCH_3$), 169.5 (NH$\underline{C}OCH_3$), 100.9 (C-1), 72.4 (C-3), 72.1 (C-5), 68.7 (C-A), 68.5 (C-4), 62.2 (C-6), 54.9 (C-2), 30.7 (S-CO$\underline{C}H_3$), 29.0 (C-B), 23.5 (NHCO$\underline{C}H_3$), 20.9 (CO$\underline{C}H_3$), 20.8 (2 x CO$\underline{C}H_3$);

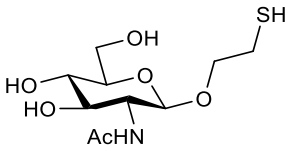
IR (neat): $\tilde{\nu}$ (cm^{-1}) 1744, 1685, 1655 (C=O), 1556, 1373, 1218, 1032.

ESI$^+$ LRMS: calculated for $C_{18}H_{27}NO_{10}SNa$ $[M+Na]^+$ 472.0, found: 472.0

mp: 144-145 °C ;

$[\alpha]_D^{25}$ = - 4.3 (c = 0.5, $CHCl_3$).

2-acetamido-2-deoxy-1-(2-thioethyl)-β-D-glucose (2)



2

A solution of sodium methoxide (25 % wt solution, 1.2 mL) was added to a solution of protected sugar **S5** (1.13 g, 2.5 mmol, 1 equiv.) in dry MeOH (10 mL) at 0 °C, under N_2 atmosphere. The reaction mixture was stirred for 15 min at room temperature when TLC (100 % EtOAc) indicated formation of the product (R_f = 0.0) with complete consumption of the starting material (R_f > 0.0). Then, the pH of the reaction mixture was neutralized with activated DOWEX® 50WX8-200. The mixture was then filtered and concentrated i*n vacuo* to afford unprotected compound **2** (680 mg, 2.42 mmol, 96.8 %) as a white solid. The product was used without further purification.

^{1}H NMR (400 MHz, D_2O): δ 4.58 (d, 1H, $J_{1,2}$ = 8.4 Hz, H-1), 4.04 (dt, 1H, J_{gem}= 10.6 Hz, $J_{A',B}$ = 5.7 Hz, H-A'), 3.95 (dd, 1H, $J_{6a,6b}$= 12.3 Hz, $J_{6a,5}$= 1.8 Hz, H-6a), 3.80-3.65 (m, 3H, H-6b, H-A, H-2), 3.59-3.51 (t, 1H, $J_{3,4}$ = 8.8 Hz, $J_{3,2}$ = 10.2 Hz, H-3), 3.49-3.43 (m, 2H, H-4, H-5), 2.74 (t, 2H, $J_{B,A}$ = 5.8 Hz, 2 H-B), 2.08 (s, 3H, $NHCOCH_3$).

^{13}C NMR (101 MHz, D_2O): δ 174.7 ($NHCOCH_3$), 101.2 (C-1), 75.9 (C-5), 73.8 (C-3), 71.6 (C-A), 70.0 (C-4), 60.8 (C-6), 55.6 (C-2), 23.5 (C-B), 22.3 ($NHCOCH_3$).

IR (neat): $\tilde{\nu}$ (cm^{-1}) 3262 (NH, OH), 1702, 1651, 1552 (C=O), 1322, 1151, 1105, 1083, 1067, 1031.

ESI$^-$ HRMS: calculated for $C_{10}H_{18}NO_6S$ $[M-H]^-$ 280.0860, found 280.0857.

mp: 155-156 °C.

$[\alpha]_D^{25}$ = - 39.4 (c = 0.1, H_2O).

2.5 References

1. Lindahl U, Couchman J, Kimata K, Esko JD. Proteoglycans and sulfated glycosaminoglycans. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

2. Karamanos NK, Piperigkou Z, Theocharis AD, et al. Proteoglycan chemical diversity drives multifunctional cell regulation and therapeutics. *Chem Rev*. 2018;118(18):9152-9232.

3. Deangelis PL, Liu J, Linhardt RJ. Chemoenzymatic synthesis of glycosaminoglycans: re-creating, re-modeling and re-designing nature's longest or most complex carbohydrate chains. *Glycobiology*. 2013;23(7):764-777.

4. Xu Y, Masuko S, Takieddin M, et al. Chemoenzymatic synthesis of homogeneous ultralow molecular weight heparins. *Science*. 2011;334(6055):498-501.

5. Huang ML, Smith RAA, Trieger GW, Godula K. Glycocalyx remodeling with proteoglycan mimetics promotes neural specification in embryonic stem cells. *J Am Chem Soc*. 2014;136(30):10565-10568.

6. Huang ML, Tota EM, Lucas TM, Godula K. Influencing early stages of neuromuscular junction formation through glycocalyx engineering. *ACS Chem Neurosci*. 2018;9(12):3086-3093.

7. Desnoyers L, Arnott D, Pennica D. WISP-1 binds to decorin and biglycan. *J Biol Chem*. 2001;276(50):47599-47607.

8. Schönherr E, Broszat M, Brandan E, Bruckner P, Kresse H. Decorin core protein fragment Leu155-Val260 interacts with TGF-beta but does not compete for decorin binding to type I collagen. *Arch Biochem Biophys*. 1998;355(2):241-248.

9. Hildebrand A, Romarís M, Rasmussen LM, et al. Interaction of the small interstitial proteoglycans biglycan, decorin and fibromodulin with transforming growth factor beta. *Biochem J*. 1994;302(2):527-534.

10. Weyers A, Linhardt RJ. Neoproteoglycans in tissue engineering. *FEBS J*. 2013;280(10):2511-2522.

11. Yamaguchi M, Takagaki K, Kojima K, et al. Novel proteoglycan glycotechnology: chemoenzymatic synthesis of chondroitin sulfate-containing molecules and its application. *Glycoconj J*. 2010;27(1):189-198.

12. Köhling S, Exner MP, Nojoumi S, Schiller J, Budisa N, Rademann J. One-pot synthesis of unprotected anomeric glycosyl thiols in water for glycan ligation reactions with highly functionalized sugars. *Angew Chem Int Ed*. 2016;55(50):15510-15514.

13. Pulsipher A, Griffin ME, Stone SE, Hsieh-Wilson LC. Long-lived engineering of glycans to direct stem cell fate. *Angew Chem Int Ed.* 2015;54(5):1466-1470.

14. Los GV, Encell LP, McDougall MG, et al. HaloTag: a novel protein labeling technology for cell imaging and protein analysis. *ACS Chem Biol*. 2008;3(6):373-382.

15. Sletten EM, Bertozzi CR. Bioorthogonal chemistry: fishing for selectivity in a sea of functionality. *Angew Chem Int Ed*. 2009;48(38):6974-6998.

16. Wright TH, Bower BJ, Chalker JM, et al. Posttranslational mutagenesis: a chemical strategy for exploring protein side-chain diversity. *Science*. 2016;354(6312):1-11.

17. Josephson B, Fehl C, Isenegger PG, et al. Light-driven post-translational installation of reactive protein side chains. *Nature*. 2020;585(7826):530-537.

18. Floyd N, Vijayakrishnan B, Koeppe JR, Davis BG. Thiyl glycosylation of olefinic proteins: *S*-linked glycoconjugate synthesis. *Angew Chem Int Ed*. 2009;48(42):7798-7802.

19. Perchyonok T. *Radical reactions in aqueous media.* Royal Society of Chemistry; 2009.

20. McSweeney L, Dénès F, Scanlan EM. Thiyl-radical reactions in carbohydrate chemistry: From thiosugars to glycoconjugate synthesis. *Eur J Org Chem*. 2016;2016(12):2080-2095.

21. Prabhakar V, Raman R, Capila I, Bosques CJ, Pojasek K, Sasisekharan R. Biochemical characterization of the chondroitinase ABC I active site. *Biochem J*. 2005;390(2):395-405.

22. Prabhakar V, Capila I, Soundararajan V, Raman R, Sasisekharan R. Recombinant expression, purification, and biochemical characterization of chondroitinase ABC II from proteus vulgaris*. *J Biol Chem*. 2009;284(2):974-982.

23. Cottier L, Descotes G. Syntheses stereoselectives des isomeres cinetiques et thermodynamiques d'alcoxy-spirocetals et de trioxa-bis-spirocetals. *Tetrahedron*. 1985;41(2):409-412.

24. Weimar T, Kreis UC, Andrews JS, Pinto BM. Conformational analysis of maltoside heteroanalogues using high-quality NOE data and molecular mechanics calculations. Flexibility as a function of the interglycosidic chalcogen atom. *Carbohydr Res*. 1999;315(3-4):222-233.

25. Little PJ, Ballinger ML, Burch ML, Osman N. Biosynthesis of natural and hyperelongated chondroitin sulfate glycosaminoglycans: new insights into an elusive process. *Open Biochem J*. 2008;2:135-142.

26. Kitagawa H, Izumikawa T, Uyama T, Sugahara K. Molecular cloning of a chondroitin polymerizing factor that cooperates with chondroitin synthase for chondroitin polymerization. *J Biol Chem*. 2003;278(26):23666-23671.

27. Izumikawa T, Koike T, Shiozawa S, Sugahara K, Tamura J, Kitagawa H. Identification of chondroitin sulfate glucuronyltransferase as chondroitin synthase-3 involved in chondroitin polymerization: chondroitin polymerization is achieved by multiple enzyme complexes consisting of chondroitin synthase family members. *J Biol Chem*. 2008;283(17):11396-11406.

28. Linhardt RJ, Dordick JS, Deangelis PL, Liu J. Enzymatic synthesis of glycosaminoglycan heparin. *Semin Thromb Hemost*. 2007;33(5):453-465.

29. Na L, Yu H, McArthur JB, Ghosh T, Asbell T, Chen X. Engineer *P. multocida* heparosan synthase 2 (PmHS2) for size-controlled synthesis of longer heparosan oligosaccharides. *ACS Catal*. 2020;10(11):6113-6118.

30. DeAngelis PL, White CL. Identification and molecular cloning of a heparosan synthase from *Pasteurella multocida* type D. *J Biol Chem*. 2002;277(9):7209-7213.

31. DeAngelis PL, White CL. Identification of a distinct, cryptic heparosan synthase from *Pasteurella multocida* types A, D, and F. *J Bacteriol*. 2004;186(24):8529-8532.

32. Kane TA, White CL, DeAngelis PL. Functional characterization of PmHS1, a *Pasteurella multocida* heparosan synthase. *J Biol Chem*. 2006;281(44):33192-33197.

33. Sugiura N, Baba Y, Kawaguchi Y, et al. Glucuronyltransferase activity of KfiC from *Escherichia coli* strain K5 requires association of KfiA KfiC AND KfiA ARE ESSENTIAL ENZYMES FOR PRODUCTION OF K5 POLYSACCHARIDE, *N*-ACETYLHEPAROSAN. *J Biol Chem*. 2010;285(3):1597-1606.

34. Kooy FK, Beeftink HH, Eppink MHM, Tramper J, Eggink G, Boeriu CG. Structural and functional evidence for two separate oligosaccharide binding sites of *Pasteurella multocida* hyaluronan synthase. *Adv Enzyme Res*. 2013;1(4):97-111.

35. Weigel PH, DeAngelis PL. Hyaluronan synthases: a decade-plus of novel glycosyltransferases*. *J Biol Chem*. 2007;282(51):36777-36781.

36. Jing W, DeAngelis PL. Analysis of the two active sites of the hyaluronan synthase and the chondroitin synthase of *Pasteurella multocida*. *Glycobiology*. 2003;13(10):661-671.

37. Jing W, DeAngelis PL. Synchronized chemoenzymatic synthesis of monodisperse hyaluronan polymers*. *J Biol Chem*. 2004;279(40):42345-42349.

38. DeAngelis PL, Padgett-McCue AJ. Identification and molecular cloning of a chondroitin synthase from *Pasteurella multocida* type F. *J Biol Chem*. 2000;275(31):24124-24129.

39. Otto NJ, Solakyildirim K, Linhardt RJ, DeAngelis PL. *Comamonas testosteronan* synthase, a bifunctional glycosyltransferase that produces a unique heparosan polysaccharide analog. *Glycobiology*. 2011;21(10):1331-1340.

40. Sismey-Ragatz AE, Green DE, Otto NJ, Rejzek M, Field RA, DeAngelis PL. Chemoenzymatic synthesis with distinct *Pasteurella* heparosan synthases. *J Biol Chem*. 2007;282(39):28321-28327.

41. Schaefer L, Iozzo RV. Biological functions of the small leucine-rich proteoglycans: from genetics to signal transduction. *J Biol Chem*. 2008;283(31):21305-21309.

42. Akatsu C, Mizumoto S, Kaneiwa T, et al. Dermatan sulfate epimerase 2 is the predominant isozyme in the formation of the chondroitin sulfate/dermatan sulfate hybrid structure in postnatal developing mouse brain. *Glycobiology*. 2011;21(5):565-574.

43. Scott PG, McEwan PA, Dodd CM, et al. Crystal structure of the dimeric protein core of decorin, the archetypal small leucine-rich repeat proteoglycan. *Proc Natl Acad Sci U S A*. 2004;101(44):15633-15638.

44. Seo N, Hocking AM, Höök M, McQuillan DJ. Decorin core protein secretion is regulated by *N*-linked oligosaccharide and glycosaminoglycan additions. *J Biol Chem*. 2005;280(52):42774-42784.

45. Neill T, Schaefer L, Iozzo RV. Decorin: a guardian from the matrix. *Am J Pathol*. 2012;181(2):380-387.

46. Danielson KG, Baribault H, Holmes DF, Graham H, Kadler KE, Iozzo RV. Targeted disruption of decorin leads to abnormal collagen fibril morphology and skin fragility. *J Cell Biol*. 1997;136(3):729-743.

47. Keene DR, San Antonio JD, Mayne R, et al. Decorin binds near the *C* terminus of type I collagen. *J Biol Chem*. 2000;275(29):21801-21804.

48. Zhang G, Ezura Y, Chervoneva I, et al. Decorin regulates assembly of collagen fibrils and acquisition of biomechanical properties during tendon development. *J Cell Biochem*. 2006;98(6):1436-1449.

49. Zhang G, Chen S, Goldoni S, et al. Genetic evidence for the coordinated regulation of collagen fibrillogenesis in the cornea by decorin and biglycan. *J Biol Chem*. 2009;284(13):8888-8897.

50. Zhang W, Ge Y, Cheng Q, Zhang Q, Fang L, Zheng J. Decorin is a pivotal effector in the extracellular matrix and tumour microenvironment. *Oncotarget*. 2018;9(4):5480-5491.

51. Karamanos NK, Piperigkou Z, Theocharis AD, et al. Proteoglycan chemical diversity drives multifunctional cell regulation and therapeutics. *Chem Rev*. 2018;118(18):9152-9232.

52. Wei J, Hu M, Huang K, Lin S, Du H. Roles of proteoglycans and glycosaminoglycans in cancer development and progression. *Int J Mol Sci*. 2020;21(17):1-28.

53. Yamaguchi Y, Mann DM, Ruoslahti E. Negative regulation of transforming growth factor-β by the proteoglycan decorin. *Nature*. 1990;346(6281):281-284.

54. Baghy K, Iozzo RV, Kovalszky I. Decorin-TGFβ axis in hepatic fibrosis and cirrhosis. *J Histochem Cytochem*. 2012;60(4):262-268.

55. Isaka Y, Brees DK, Ikegaya K, et al. Gene therapy by skeletal muscle expression of decorin prevents fibrotic disease in rat kidney. *Nat Med*. 1996;2(4):418-423.

56. Zhang Z, Li X, Liu Y, Zhang X, Li Y, Xu W. Recombinant human decorin inhibits cell proliferation and downregulates TGF-beta1 production in hypertrophic scar fibroblasts. *Burns*. 2007;33(5):634-641.

57. Merline R, Moreth K, Beckmann J, et al. Signaling by the matrix proteoglycan decorin controls inflammation and cancer through PDCD4 and MicroRNA-21. *Sci Signal*. 2011;4(199):ra75.

58. Santra M, Reed CC, Iozzo RV. Decorin binds to a narrow region of the epidermal growth factor (EGF) receptor, partially overlapping but distinct from the EGF-binding epitope. *J Biol Chem*. 2002;277(38):35671-35681.

59. Seidler DG, Goldoni S, Agnew C, et al. Decorin protein core inhibits in vivo cancer growth and metabolism by hindering epidermal growth factor receptor function and triggering apoptosis via caspase-3 activation. *J Biol Chem*. 2006;281(36):26408-26418.

60. Csordás G, Santra M, Reed CC, et al. Sustained down-regulation of the epidermal growth factor receptor by decorin. A mechanism for controlling tumor growth in vivo. *J Biol Chem*. 2000;275(42):32879-32887.

61. Zhu J, Goldoni S, Bix G, et al. Decorin evokes protracted internalization and degradation of the epidermal growth factor receptor via caveolar endocytosis. *J Biol Chem*. 2005;280(37):32468-32479.

62. Hu X, Villodre ES, Larson R, et al. Decorin-mediated suppression of tumorigenesis, invasion, and metastasis in inflammatory breast cancer. *Commun Biol*. 2021;4(1):1-14.

63. Goldoni S, Seidler DG, Heath J, et al. An antimetastatic role for decorin in breast cancer. *Am J Pathol*. 2008;173(3):844-855.

64. Hu Y, Sun H, Owens RT, et al. Decorin suppresses prostate tumor growth through inhibition of epidermal growth factor and androgen receptor pathways. *Neoplasia*. 2009;11(10):1042-1053.

65. Buraschi S, Pal N, Tyler-Rubinstein N, Owens RT, Neill T, Iozzo RV. Decorin antagonizes met receptor activity and down-regulates β-catenin and myc levels. *J Biol Chem*. 2010;285(53):42075-42085.

66. Goldoni S, Humphries A, Nyström A, et al. Decorin is a novel antagonistic ligand of the met receptor. *J Cell Biol*. 2009;185(4):743-754.

67. Horváth Z, Kovalszky I, Fullár A, et al. Decorin deficiency promotes hepatic carcinogenesis. *Matrix Biol*. 2014; 35:194-205.

68. Buraschi S, Neill T, Goyal A, et al. Decorin causes autophagy in endothelial cells *via* Peg3. *Proc Natl Acad Sci U S A*. 2013;110(28):2582.

69. Goyal A, Neill T, Owens RT, Schaefer L, Iozzo RV. Decorin activates AMPK, an energy sensor kinase, to induce autophagy in endothelial cells. *Matrix Biol*. 2014;34:46-54.

70. Buraschi S, Neill T, Iozzo RV. Decorin is a devouring proteoglycan: remodeling of intracellular catabolism via autophagy and mitophagy. *Matrix Biol*. 2019;75-76:260-270.

71. Neill T, Sharpe C, Owens RT, Iozzo RV. Decorin-evoked paternally expressed gene 3 (PEG3) is an upstream regulator of the transcription factor EB (TFEB) in endothelial cell autophagy. *J Biol Chem*. 2017;292(39):16211-16220.

72. Neill T, Schaefer L, Iozzo RV. Decorin as a multivalent therapeutic agent against cancer. *Adv Drug Deliv Rev*. 2016;97:174-185.

73. Iozzo RV. *Proteoglycan protocols.* Humana Press; 2001.

74. Lindahl U, Couchman J, Kimata K, Esko JD. Proteoglycans and sulfated glycosaminoglycans. In: Varki A, Cummings RD, Esko JD, et al, eds. *Essentials of glycobiology.* 3rd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2015.

75. Klein JA, Meng L, Zaia J. Deep sequencing of complex proteoglycans: a novel strategy for high coverage and site-specific identification of glycosaminoglycan-linked peptides. *Mol Cell Proteomics*. 2018;17(8):1578-1590.

76. Sugahara K, Ohi Y, Harada T, de Waard P, Vliegenthart JF. Structural studies on sulfated oligosaccharides derived from the carbohydrate-protein linkage region of chondroitin 6-sulfate proteoglycans of shark cartilage. I. six compounds containing 0 or 1 sulfate and/or phosphate residues. *J Biol Chem*. 1992;267(9):6027-6035.

77. Wen J, Xiao J, Rahdar M, et al. Xylose phosphorylation functions as a molecular switch to regulate proteoglycan biosynthesis. *Proc Natl Acad Sci U S A*. 2014;111(44):15723-15728.

78. Scott PG, Grossmann JG, Dodd CM, Sheehan JK, Bishop PN. Light and X-ray scattering show decorin to be a dimer in solution. *J Biol Chem*. 2003;278(20):18353-18359.

79. Goldoni S, Owens RT, McQuillan DJ, et al. Biologically active decorin is a monomer in solution. *J Biol Chem*. 2004;279(8):6606-6612.

80. Sonn-Segev A, Belacic K, Bodrug T, et al. Quantifying the heterogeneity of macromolecular machines by mass photometry. *Nat Commun*. 2020;11(1):1-10.

81. Young G, Hundt N, Cole D, et al. Quantitative mass imaging of single biological macromolecules. *Science*. 2018;360(6387):423-427.

82. Heuvel, Robert HH van den, Heck AJ. Native protein mass spectrometry: from intact oligomers to functional machineries. *Curr Opin Chem Biol*. 2004;8(5):519-526.

83. Zheng Y, Zheng W, Zhu D, Chang H. Theoretical modeling of pKa's of thiol compounds in aqueous solution. *New J Chem*. 2019;43(13):5239-5254.

84. Lian P, Johnston RC, Parks JM, Smith JC. Quantum chemical calculation of pK_as of environmentally relevant functional groups: carboxylic acids, amines, and thiols in aqueous solution. *J Phys Chem A*. 2018;122(17):4366-4374.

85. Li Y, Yu H, Thon V, et al. Donor substrate promiscuity of the *N*-acetylglucosaminyltransferase activities of *Pasteurella multocida* heparosan synthase 2 (PmHS2) and *Escherichia coli* K5 KfiA. *Appl Microbiol Biotechnol*. 2014;98(3):1127-1134.

86. Modak A. *Novel methods in glycosaminoglycan and proteoglycan synthesis.* University of Oxford; 2015.

87. Nadal S, Raj R, Mohammed S, Davis BG. Synthetic post-translational modification of histones. *Curr Opin Chem Biol*. 2018;45:35-47.

88. Niesen FH, Berglund H, Vedadi M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nat Protoc*. 2007;2(9):2212-2221.

89. Gao K, Oerlemans R, Groves MR. Theory and applications of differential scanning fluorimetry in early-stage drug discovery. *Biophys Rev*. 2020;12(1):85-104.

90. Qasba PK, Ramakrishnan B, Boeggeman E. Structure and function of β-1,4-galactosyltransferase. *Curr Drug Targets*. 2008;9(4):292-309.

91. Ardèvol A, Rovira C. Reaction mechanisms in carbohydrate-active enzymes: glycoside hydrolases and glycosyltransferases. Insights from *ab initio* quantum mechanics/molecular mechanics dynamic simulations. *J Am Chem Soc*. 2015;137(24):7528-7547.

92. Ruparelia N, Chai JT, Fisher EA, Choudhury RP. Inflammatory processes in cardiovascular disease: a route to targeted therapies. *Nat Rev Cardiol*. 2017;14(3):133-144.

93. Lane RS, St. Ange K, Zolghadr B, et al. Expanding glycosaminoglycan chemical space: towards the creation of sulfated analogs, novel polymers and chimeric constructs. *Glycobiology*. 2017;27(7):646-656.

94. Choi J, Kim H, Kim K, Lee B, Lu W, An W. Selective requirement of H2B *N*-terminal tail for p14ARF-induced chromatin silencing. *Nucleic Acids Res*. 2011;39(21):9167-9180.

95. Kollipara L, Zahedi RP. Protein carbamylation: *in vivo* modification or *in vitro* artefact? *Proteomics*. 2013;13(6):941-944.

96. Nadal S. *Studying protein interaction and glycosylation with chemical proteomics.* University of Oxford; 2020.

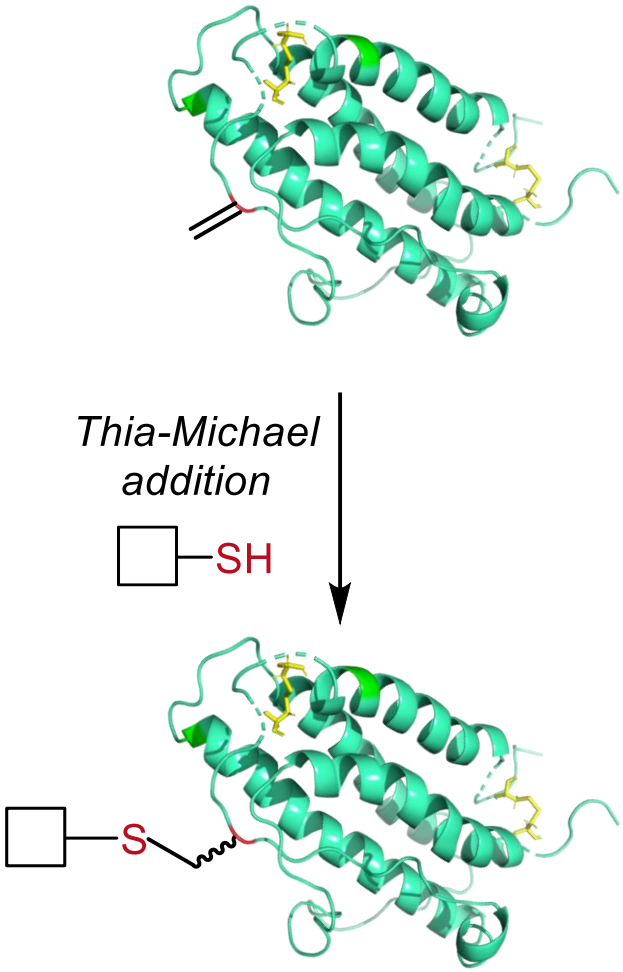
97. Geoghegan KF, Dixon HB, Rosner PJ, et al. Spontaneous α-*N*-6-phosphogluconoylation of a "His tag" in *Escherichia coli*: the cause of extra mass of 258 or 178 da in fusion proteins. *Anal Biochem*. 1999;267(1):169-184.

98. Gamblin DP, Garnier P, Kasteren Sv, Oldham NJ, Fairbanks AJ, Davis BG. Glyco-SeS: selenenylsulfide-mediated protein glycoconjugation—a new strategy in post-translational modification. *Angew Chem Int Ed*. 2004;43(7):828-833.

99. Alexander SR, Lim D, Amso Z, Brimble MA, Fairbanks AJ. Protecting group free synthesis of glycosyl thiols from reducing sugars in water; application to the production of *N*-glycan glycoconjugates. *Org Biomol Chem*. 2017;15(10):2152-2156.

100. Soltermann F, Foley EDB, Pagnoni V, et al. Quantifying protein–protein interactions by molecular counting with mass photometry. *Angew Chem Int Ed*. 2020;59(27):10774-10779.

Chapter 3



Towards a single glycoform of erythropoietin through site-selective chemical glycosylation

Contents

3.1 Introduction 168

3.1.1 Natural heterogeneity impedes the structure-activity relationship study of *N*-glycoproteins..... 168

3.1.2 A combined chemical mutagenesis/thia-Michael addition approach to access defined EPO glycoform-mimics 172

3.2 Results and discussion..... 177

3.2.1 Introduction of dehydroalanine to the MBP-fused EPO (MBP-EPO-1)..... 177

3.2.2 Introduction of dehydroalanine to the MBP-fused EPO which was engineered for reduced aggregation and disulfide shuffling (MBP-EPO-2)..... 189

3.2.3 Introduction of dehydroalanine to the MBP-fused EPO which was engineered for reduced aggregation (MBP-EPO-3) 198

3.3 Conclusions and future perspectives..... 202

3.4 Experimental section 205

3.4.1 Plasmid DNA details and protein sequences 205

3.4.2 Protein expression and characterisation 206

3.4.2.1 Protein expression of different MBP-EPO variants:..... 206

3.4.2.2 Optimization of lysis conditions for cells overexpressing the MBP-EPO-1 variant..... 207

3.4.2.3 TEV-cleavage of MBP-EPO-1 and attempted purification of EPO-1..... 211

3.4.3 Formation of Dha-MBP-EPO-1 and GlcNAc-S-MBP-EPO-1..... 213

3.4.3.1 Implementation of a DTT/TEV-cleavage method for the LC-MS detection of Dha-MBP-EPO-1 and GlcNAc-S-MBP-EPO-1 213

3.4.3.2 Formation of Dha-MBP-EPO-1 using 300 equivalents MDBP with (A) or without (B) pre-treatment with 0.5 mM DTT (DTT/TEV-cleavage LC-MS method) 214

3.4.3.3 Formation of Dha-MBP-EPO-1 using 900 equivalents MDBP after pre-treatment with 0.5 mM DTT (DTT/TEV-cleavage LC-MS method) 214

3.4.3.4 Thia-Michael addition of GlcNAc-SH to Dha-MBP-EPO-1 (DTT/TEV-cleavage LC-MS method) 214

3.4.4 Formation of Dha-MBP-EPO-2 215

3.4.4.1 Attempted Dha-reaction in 50 mM TrisHCl, pH 8.0 buffer 215

3.4.4.2 Screening of different bis-alkylating reagents (300 equiv.) in different buffers (A and B), on ice 215

3.4.4.3 Screening of different bis-alkylating reagents (900 equiv.) in buffer B (100 mM NaPi, 3 M GdnHCl pH 8.0), on ice and at 25 °C 216

3.4.4.4 Screening of different equivalents (900/1800) of three bis-alkylating reagents (MDBP, DBHDA, DBDGla) in buffer B (100 mM NaPi, 3 M GdnHCl pH 8.0), at 37 °C 218

3.4.4.5 Screening of different equivalents (300/600) of the bis-alkylating reagent MDBP in buffer B (100 mM NaPi, 3 M GdnHCl pH 8.0), at 37 °C 218

3.4.4.6 Screening of different organic solvents for their ability to affect the stability of the MBP-EPO-2 protein 221

3.4.4.7 Quantification of free Cys residues in the MBP-EPO-2 protein 221

3.4.4.8 Identification of the pre-requisite needed for monitoring formation of the Dha-MBP-EPO-2 (3 M GdnHCl buffer) 222

3.4.4.9 Thia-Michael addition of β-mercaptoethanol to Dha-MBP-EPO-2 223

3.4.5 Formation of Dha-MBP-EPO-3 and GlcNAc-S-MBP-EPO-3 224

3.4.5.1 Formation of Dha-MBP-EPO-3 using 300 equivalents MDBP and subsequent thia-Michael addition of GlcNAc-SH or β-mercaptoethanol 224

3.4.6 LC-MS/MS analysis of enzymatically digested GlcNAc-S-MBP-EPO-1, βME-S-MBP-EPO-2 and βME-S-MBP-EPO-3 225

3.4.7 Molecular Dynamics (MD) of EPO-1 227

3.5 References 228

3.1 Introduction

3.1.1 Natural heterogeneity impedes the structure-activity relationship study of *N*-glycoproteins

Erythropoietin (EPO) is a glycoprotein hormone produced mainly in kidneys[1-6] and to a lesser extent in liver, spleen, bone marrow, lungs and brain,[7-9] in response to a hypoxia-induced stimulus (Figure **3.1**).[10] Under physiological concentrations of oxygen, the levels of hypoxia-inducible transcription factor alpha (HIFα) are maintained constant by the iron- and oxygen-dependent prolyl hydroxylase domain (PHD) proteins.[11] These enzymes are able to direct the constitutively expressed HIFα on a continuous path of ubiquitination and ultimately proteosomal degradation, thereby restricting the accumulation of this protein in the cytosol.[12] In contrast, when PHDs are inhibited by hypoxic conditions, HIFα accumulates and translocates into the nucleus, where it forms a stable heterodimer with a constitutive nuclear HIFβ subunit.[13] The resulting HIFα-HIFβ complex is then able to bind hypoxia-response elements (HREs), thereby activating the transcription of the EPO gene[14] and of other genes which encode proteins involved in iron transport,[15-18] glucose metabolism[19,20] and angiogenesis,[21-23] processes that ultimately facilitate oxygen delivery to the affected tissue. EPO secreted into the blood stream is then able to reach erythroid progenitor cells in the bone marrow and to stimulate their differentiation into red blood cells (RBCs).[24] The erythropoietic function of EPO is mediated by the preformed membrane homodimer receptor (EPO-R) to which it binds with high affinity.[25] Upon binding of EPO to EPO-R, a conformation change in EPO-R is induced.[26,27] The latter triggers the essential partner Janus Kinase 2 (JAK2) to phosphorylate itself, the cytoplasmatic tail of EPO-R and other proteins such as the signal transducer and activator of transcription 5 (STAT5), thereby simultaneously activating a series of signalling pathways involved in erythropoiesis.[26,27]

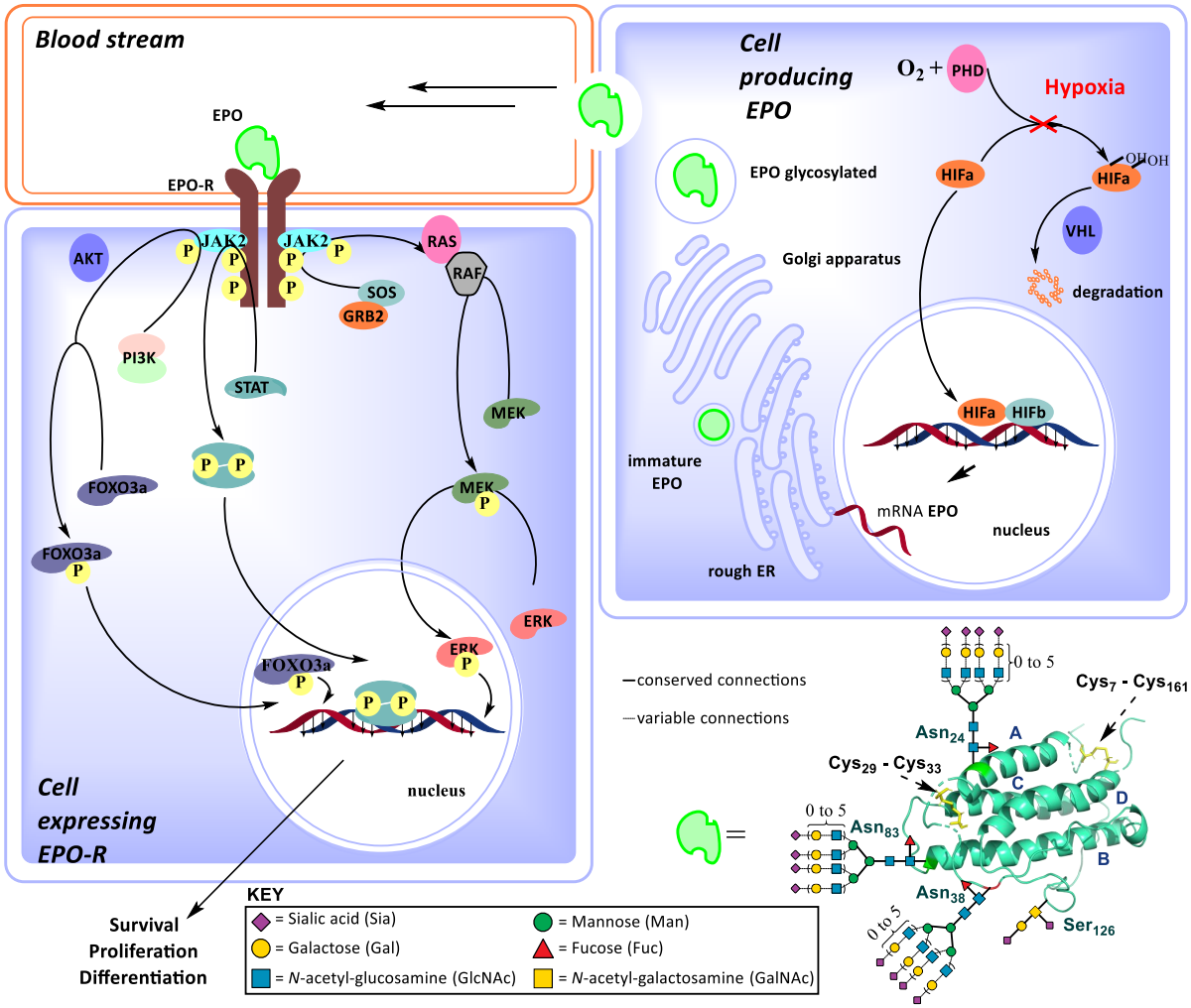


Figure 3.1. EPO gene transcription is activated by the HIF-pathway (right) and upon release in the blood, circulating EPO binds to its receptor, altering the downstream cascade of effectors responsible with numerous biological events (left). At low oxygen concentrations, HIFα escapes PHD-controlled proteosomal degradation and after association with HIFβ transcription of the EPO gene is induced. The resulting mRNA is translated into the corresponding EPO protein which is then folded and subjected to glycosylation, first in the endoplasmic reticulum (ER) and then in the Golgi apparatus. The vesicular secretion of the heterogeneously glycosylated EPO ensures that this glycohormone reaches its target cells which express the transmembrane erythropoietin receptor (EPO-R). Activation of the EPO-R triggers a signal transduction cascade that leads to anti-apoptotic, proliferation and maturation effects in hematopoietic or selected non-hematopoietic cells (HIF, hypoxia-inducible transcription factor; PHD, prolyl hydroxylase domain; VHL, von Hippel–Lindau protein; P, phosphorylation; JAK2, Janus Kinase 2; RAS, rat sarcoma protein; RAF, rapidly accelerated fibrosarcoma; MEK, mitogen-activated protein kinase kinase; ERK, extracellular-regulated kinase; SOS, son of Sevenless protein; GRB2, growth factor receptor-bound protein 2; STAT, signal transducer and activator of transcription 5; PI3K, phosphoinositide 3-kinase; AKT, protein kinase B; FOXO3a, forkhead box class O 3a).

Endogenous human EPO consists of 166 amino acids,[28,29] four of which are cysteine (Cys) residues that form two disulfide bridges: Cys7-Cys161 and Cys29-Cys33 (Figure **3.1**)[30]. The former has shown to be essential for maintaining the biological activity of this four-helical (A-D) cytokine since it secures the native protein folding in which the *N*-terminus (Cys7 helix A) is tightly linked to the distant *C*-terminus (Cys161 helix D). Conversely, the Cys29-Cys33 disulfide bridge that forms a small loop

between helix A and helix B can be safely abolished and after site-directed mutagenesis of these two residues to tyrosine (Tyr) the resulting cytokine still maintains its native function.[31]

In addition, approximately ~ 40 % (w/w) of EPO consists of glycans that are unevenly distributed onto four sites of glycosylation: one serine residue (Ser126) that allows for mainly homogeneous *O*-glycosylation, and three asparagine residues (Asn24, Asn38, Asn83) that carry the *N*-linked, differently configured, branched oligosaccharides.[30,32] Previous studies have shown that while the *O*-glycan does not influence the *in vitro* and *in vivo* biological activities of EPO, the more complex *N*-glycans are essential for the *in vivo* hematopoietic activity of this cytokine.[33] Moreover, the composition of these *N*-linked oligosaccharides which can have various antennary structures (bi-, tri-, and tetra-antennary), different numbers of *N*-acetyllactosamine (LacNAc) repeats, and various degrees of sialylation, has a significant effect on the biological activity of this glycohormone.[34]

For instance, a positive correlation was established between the *in vivo* activity of EPO and the ratio of tetra-antennary to bi-antennary *N*-linked glycans, with the highly branched EPO glycoforms displaying a better specific activity in a polycythaemic animal model.[35] In addition, sialylation of the tetra-antennary *N*-glycoforms reduces the hepatic plasma clearance of EPO thereby increasing its bioavailability. However, the same terminal saccharide does not significantly impact the plasma levels of the smaller bi-antennary species because these are extensively removed from the systemic circulation through renal excretion.[36] Lastly, EPO glycoforms bearing LacNAc repeats have been identified[37] and studies have shown that these display a higher propensity to be cleared from circulation by the galactose binding protein of hepatocytes.[38]

Due to its crucial hematopoietic role, EPO-based therapies made available to patients suffering from severe anaemia rely on administering a complex mixture of non-separable EPO glycoforms with variable glycan chains at each of the three *N*-glycosylation sites.[39] However, since these are produced by recombinant DNA technology (rEPO) in eukaryotic cell lines equipped with a slightly different set of glycosyltransferases, the resulting biosynthetic EPO preparations ('epoetins') have glycosylation profiles that differ from the ones found in the human serum EPO.[40,41] For instance, the latter does not contain fully sialylated tetra-antennary glycans whereas these are present in high ratios in the commercial epoetins (alfa, 19 %; beta, 46 %; omega, 21 %).[42] Another significant difference is represented by the high content in neutral (20 %) and di-acidic (48 %) glycans present in human EPO that cannot be found in comparable amounts in the rEPOs.[42]

In addition, large scale production of epoetins has also encountered reproducibility issues.[34,43] This has led to EPO preparations characterized by different glycosylation profiles which were not only cell line dependent, but which also varied from batch to batch when using the same eukaryotic host.

This type of issue may occur because the smallest change in culture conditions has the potential to greatly impact the metabolism of the producing organism.[44]

EPO therapies have been highly efficacious in treating anaemia caused by chronic renal failure, antineoplastic chemotherapy or anti-viral medication. However, over the last 20 years, it has become increasingly concerning that patients treated with 'epoetins' may present a significantly higher risk of developing venous thromboembolism.[45-47] In contrast, there is also mounting evidence that EPO may be associated with other important, beneficial biological processes such as tissue protection/repair,[48,49] angiogenesis[50] and metabolic regulation[51].

All things considered, this glycohormone appears to be involved in a variety of complex cellular pathways but its rigorous, structure-dependent, systematic study is hindered by its natural heterogeneity. Consequently, the established hematopoietic role and recently emerging non-hematopoietic functions of EPO were not correlated to any specific glycan features of this cytokine because of the difficulty in accessing homogeneous EPO glycoforms. Therefore, there is still a high need to develop rapid synthetic methodologies that give access to homogeneously glycosylated EPO, presenting a predetermined glycan, and thus exhibiting a structure-dependent biological activity.

3.1.2 A combined chemical mutagenesis/thia-Michael addition approach to access defined EPO glycoform-mimics

The fundamental goal of this project was the generation of a homogeneous EPO glycoform bearing a human, core-fucosylated, tetra-antennary glycan at one of the three *N*-glycosylation sites. This would allow us to establish more precisely the role that these highly branched, heavily glycosylated EPO species have in a biological context, and obtain a therapeutic *N*-glycoprotein with uniform, structure-related, predictable effects. Thus, the generation of such a glycoform or a closely resembling glycoform-mimic would contribute to further elucidate the SAR of this glycohormone while establishing a novel, flexible and convergent strategy, easily transferable to other heterogeneously *N*-glycosylated proteins that have important biological applications.

With this in mind, we chose to obtain a site-specifically glycosylated EPO using the so called "tag-and-modify" strategy (Figure **3.2**, a).[52-54] This would involve the initial installation of a uniquely reactive amino acid residue "tag" at a specific *N*-glycosylation site in the protein. Then, this would be chemo-selectively glycosylated ("modified") with the orthogonal glycan reaction partner, appropriately functionalized at its reducing end to match the reactivity of the "tag". Thus, the approach would generate the desired EPO glycoform-mimic in a convergent fashion, and the mimicry of the resulting glycosidic bond would ultimately depend on the developed chemical strategy (Figure **3.2**, b).

In this project, our aim was to choose a "tag-and-modify" approach that would allow the construction of a glycosidic linkage that closely resembles the substituted amide of the natural EPO glycoform. This would cause minimum disturbance of the protein 3D structure, and thereby the biological activity would not be significantly impacted.

Since Flitsch and co-workers[55] had already explored the nucleophilic substitution of a small 1-iodoacetamido-β-monosaccharide by Cys-tagged EPO proteins (Figure **3.2**, b, 1 and Chapter 1, figure **1.11**) we decided to focus on other site-selective glycosylation methods that could employ a complex 1-thio-saccharide. A highly convergent and flexible strategy had been previously developed in the Davis group to synthesize an extensively branched EPO glycan, functionalized at the reducing end with a thiol group (Figure **3.2**, b, data not published). Thus, we believed that other conjugation chemistries, either Cys-based or facilitated by the insertion of an unnatural amino acid at the *N*-glycosylation site intended for modification, may generate similar human-like glycosylation linkers carrying in turn a complex, tetra-antennary EPO glycan.

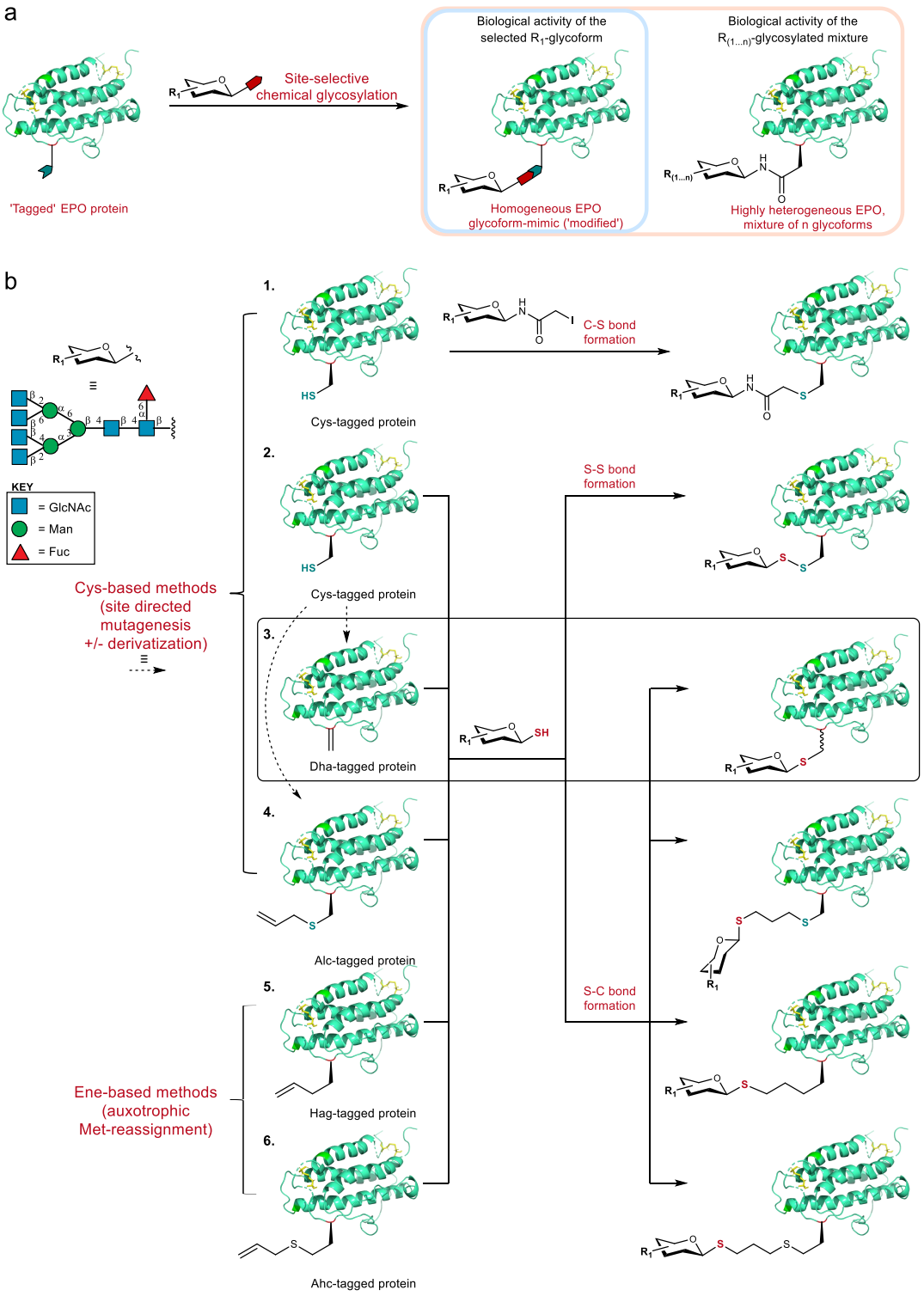


Figure 3.2. A "tag-and-modify" strategy was envisioned for the generation of a single *N*-glycoform-mimic of EPO. A heterogeneous mixture of a natural *N*-glycoprotein is expressed devoid of glycans in a bacterial host, and it contains a reactive 'tag' at the originally glycosylated amino acid position. A biocompatible, ligation strategy is then used to install the desired saccharide. The closely resembling semi-synthetic EPO-analogue has the potential to display a specific biological activity (highlighted by the blue square), that could not be otherwise dissected from the myriad of effects displayed by the natural, heterogeneous EPO mixture (highlighted by the pink square, **a**). A few ligation strategies would generate relatively flexible EPO *N*-glycoform-mimics, but the one we chose (3) appeared to be most advantageous (**b**). A more detailed description of the tags that have been explored in the past for the site-selective chemical glycosylation of proteins is made in chapter 1.

We envisioned that an alternative "tag-and-modify" strategy could either exploit the nucleophilicity of the anomeric thiolate (S^-) group formed at slightly basic pH and thus use it in traditional two-electron chemistry (Figure **3.2**, b, 2-4), or the homolytic reactivity of the thiol (S-H) bond that makes it amenable to single-electron conjugation reactions (Figure **3.2**, b, 4-6).

At first, we considered that a viable method could be the construction of a disulfide-linkage with a Cys-tagged protein scaffold *via* a selenenylsulfide-protein/-glycan intermediate, but the resulting glycoform while often stable *in vivo* may be labile under reducing conditions (Figure **3.2**, b, 2).[56]

Next, our attention was directed towards the use of a dehydroalanine (Dha) amino acid that can be derived from the same proteinogenic Cys residue inserted by site-directed mutagenesis, at one of the solvent accessible *N*-glycosylation sites of the protein (Figure **3.2**, b, 3).[57] The nucleophilic tetra-antennary 1-thiolate-glycan would react with the electrophilic, thia-Michael acceptor Dha-tag, and the resultant thioether-linked glycoconjugate would be analogous to the natural EPO glycoform. Noteworthily, the thia-Michael addition to the α,β-unsaturated amide of the Dha may generate two diastereomeric glycoforms. Davis *et al*.[58] had previously demonstrated that the addition to Dha of a small, non-chiral thiophosphate group generated a mixture of C_α-epimers (L:D), and the moderate selectivity (1:2) towards the D configuration might had been imparted by the chiral protein substrate. Consequently, we believed that it would be particularly interesting to understand what the influence on the C_α-configuration might be when a sterically demanding, chiral molecule such as our 1-thio-decasaccharide would be added to the Dha-tagged EPO protein.

Moreover, since there is a rather limited understanding about the general, biological role of *S*-glycoproteins in mammals, it would be extremely valuable to learn more about the *in vivo* stability of the *S*-linked EPO glycoform. Currently, there is increasing evidence that Cys-linked glycosylation is a post-translational modification (PTM) of proteins which occurs, albeit rarely, in humans as well.[59,60] In addition, one of these studies also showed that *O*-GlcNActransferase (OGT), the enzyme responsible with the addition of a GlcNAc-residue to serine and threonine residues on proteins, is able to catalyse the *S*-GlcNAcylation of cysteine residues as well.[60] On the contrary, *O*-GlcNAcase (OGA), the enzyme that removes the Ser/Thr-linked GlcNAc-modification, is not able to also catalyse the hydrolysis of the corresponding Cys-glycoside.[61]

Finally, the free-radical glycosylation of an ene-tagged EPO protein (Figure **3.2**, b, 4-6) represents another viable method for the site-selective synthesis of *S*-linked glycoprotein-mimics. However, the resulting *S*-glycoconjugates would contain considerably longer linkers than the natural *N*-glycoforms. Moreover, since the insertion of the ene-tagged unnatural amino acids homoallylglycine (Hag, Figure **3.2**, b, 5)[62] and allylhomocysteine (Ahc, Figure **3.2**, b, 6)[63] is made using

the Met-auxotrophic method (involving global replacement of methionine), the developed thiyl-glycosylation strategy could not be applied to dissect the activity of any chosen glycoprotein. For instance, in the case of *N*-glycosylated proteins that contain a high number of Met residues, the extensive mutagenesis required to replace these with other amino acids (e.g. isoleucine) in order to ensure the desired site-selectivity, may perturb the structure and thus the function of the resulted *S*-linked glycoform.

In conclusion, we felt that the combined chemical mutagenesis/thia-Michael addition strategy would be a straightforward route towards formation of a homogeneous EPO glycoform-mimic in which the linker would be structurally similar to the one naturally occurring at the Asn-glycosylation site (Figure **3.3**, 1). Firstly, a Cys residue is introduced by site-directed mutagenesis at one of the solvent accessible, *N*-glycosylation Asns (Figure **3.3**, 2). Then, by treating with a bis-alkylating reagent such as the commercially available methyl 2,5-dibromopentanoate (MDBP) the Cys-mutant is converted into the corresponding Dha-derivative (Figure **3.3**, 5). The reaction mechanism involves a double alkylation, first intermolecular to give the mono-alkylated intermediate (Figure **3.3**, 3), and then intramolecular to afford a sulfonium-derivative (Figure **3.3**, 4).[58]

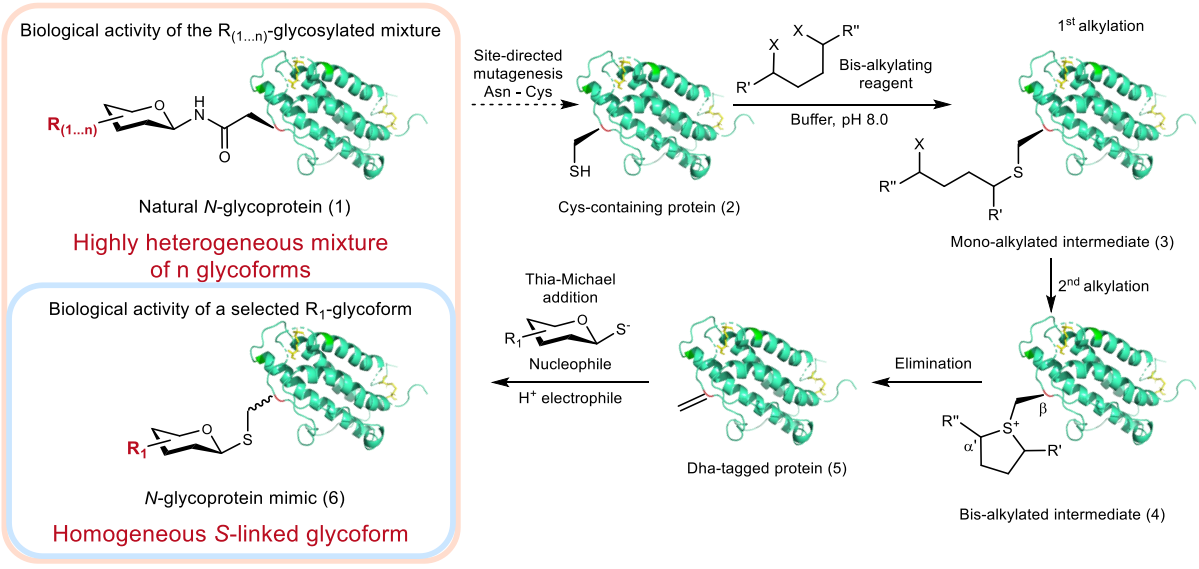


Figure 3.3. A combined chemical mutagenesis ("tag")/thia-Michael glycosylation ("modify") strategy was designed to generate selected *N*-glycoform-mimics for SAR studies. The Asn-glycosylation site of interest in the natural EPO (1) is substituted by a Cys residue through site-directed mutagenesis. Conversion of the Cys-containing protein (2) into the Dha-tagged analogue (5) is achieved *via* a double alkylation reaction (2 to 3, 3 to 4) followed by elimination of the resulting sulfonium intermediate (4). The chemo-selective thia-Michael reaction with the 1-thio-saccharide affords the desired *S*-linked glycoform (5). In this way, 5 has the potential to display a specific biological activity (highlighted by the blue square), that could not be otherwise dissected from the multitude of effects displayed by the natural, heterogeneously glycosylated EPO (highlighted by the pink square; bis-alkylating reagent, e.g. MDBP: R' = COOMe, R'' = H, X = Br; $R_{(1..n)}$: differently glycosylated species; dashed arrow: biological manipulation; plain arrow: chemical manipulation).

The latter undergoes a base catalysed α'-β elimination to generate the desired Dha-tagged protein (Figure **3.3**, 5).[58] Finally, the homogeneously glycosylated EPO (Figure **3.3**, 6) was envisioned to be obtained through conjugate addition of the 1-thio-oligosaccharide to the unsaturated amino acid residue of the Dha-tagged protein.

3.2 Results and discussion

3.2.1 Introduction of dehydroalanine to the MBP-fused EPO (MBP-EPO-1)

Preliminary work had been carried out in the Davis group to design a bacterial system appropriate for the expression of non-glycosylated EPO (Figure **3.4**, 1). Since it was known that EPO devoid of any glycan is prone to extensive aggregation into the inclusion bodies of *E. coli* cells[55] a pETM-41 vector had been engineered so that the inserted EPO gene be expressed as a fused protein construct with improved solubility properties (Figure **3.4**, 2). Therefore, the resulted plasmid encoded a polyhistidine (His6) tag which facilitates purification by Ni^{2+}-affinity chromatography, a Maltose Binding Protein (MBP) tag that promotes soluble overexpression in the cytoplasm[64] and the desired EPO-1 sequence, Cys-mutated at the Asn38 glycosylation site, and separated from the MBP by the recognition cleavage site of the *Tobacco etch* virus (TEV)-protease. Thus, upon formation of the *S*-linked MBP-EPO-1 glycoconjugate (Figure **3.4**, 3), the TEV-protease could be used to cleave the His6-MBP and afford the desired *S*-linked EPO-1 glycoform-mimic (Figure **3.4**, 4).

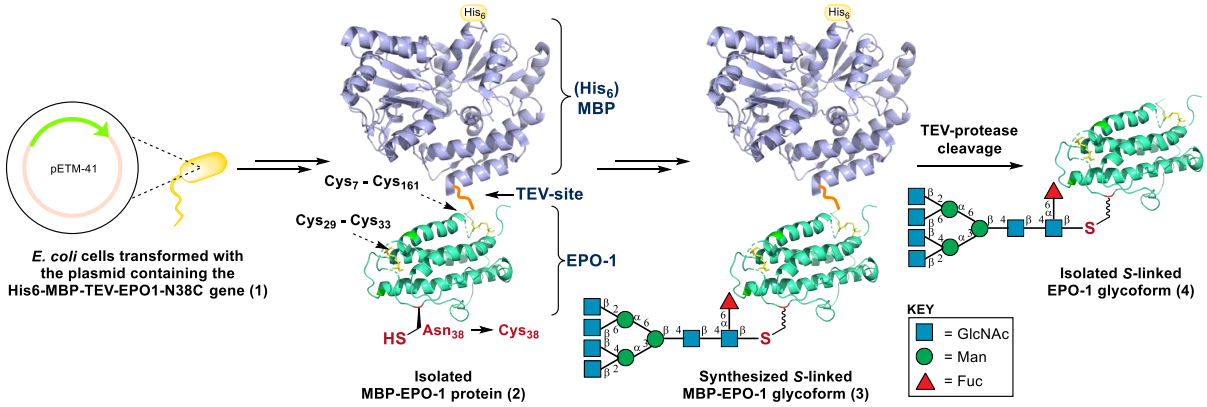
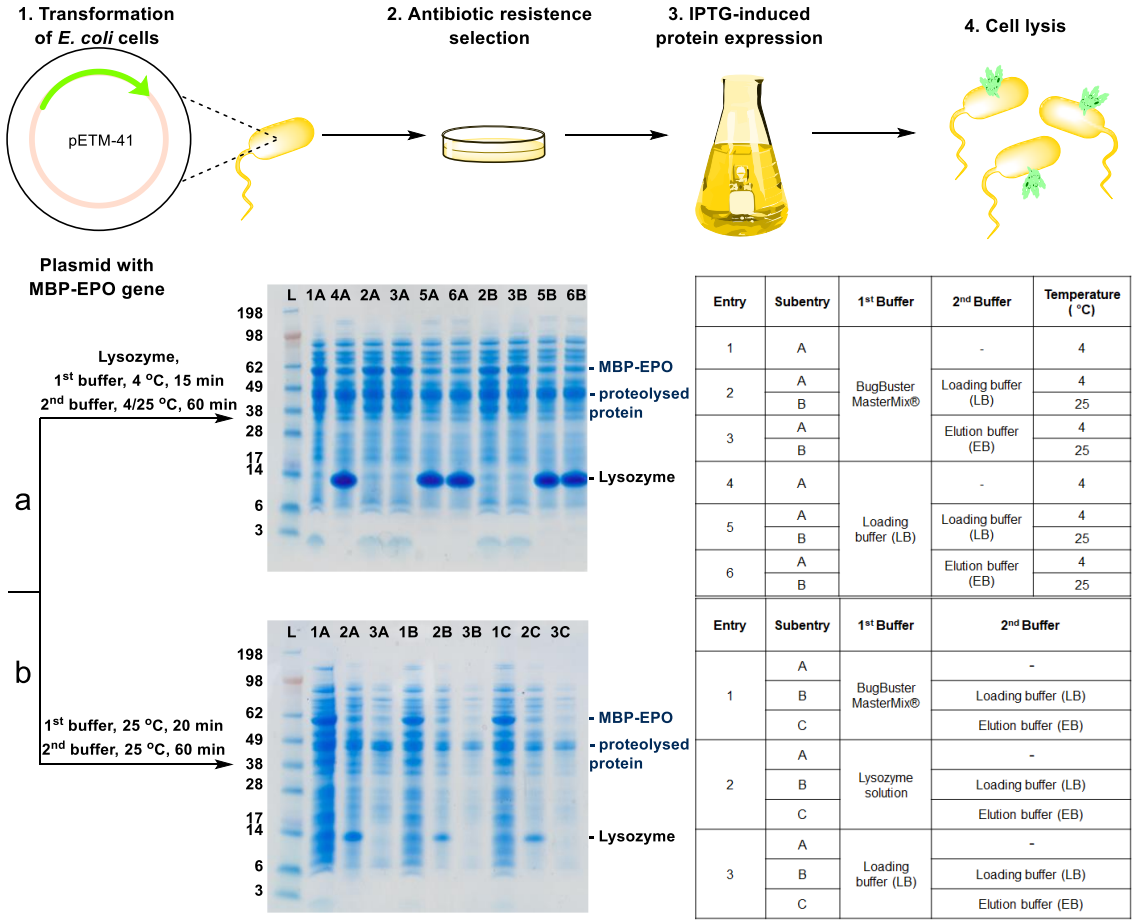


Figure 3.4. A general semi-synthetic strategy was envisioned to generate a selected *N*-glycoform-mimic of EPO-1. The competent *E. coli* strain transformed with the plasmid encoding for the MBP-EPO-1 construct (1) expresses the non-glycosylated protein containing a cysteine at the Asn38-glycosylation site (2). The latter is used to synthesize the *S*-linked MBP-EPO-1 glycoform (3) and upon TEV-protease cleavage, the homogeneous *S*-linked EPO-1 glycoform (4) is obtained (⟹ His6-MBP-TEV-EPO1-N38C gene).

To express the MBP-EPO-1 protein, BL21(DE3) competent cells, an economically advantageous biological strain that gives high yields of non-post-translationally modified protein, were transformed with the engineered plasmid (Figure **3.5**, 1). pETM-41 contained a kanamycin resistance gene and thus, selection of the transformants relied on the viability of the cells in the presence of this antibiotic (Figure **3.5**, 2). Isopropyl β-D-thiogalactoside (IPTG) was added to the recombinant bacterial culture to induce transcription of the His6-MBP-TEV-EPO1-N38C gene that had

been placed under the control of the T7 *lac* promoter (Figure **3.5**, 3). During initial large-scale cell lysis and purification procedures, MBP-EPO-1 underwent extensive proteolytic degradation and consequently, the intact protein could not be isolated and characterized. Therefore, in order to identify the optimal cell lysis conditions that preserved the integrity of the MBP-EPO-1 construct we first proceeded with testing different combinations of detergent (BugBuster MasterMix®), lysis buffers (with/without imidazole), lysis temperatures and other additives (lysozyme), in small-scale reactions (Figure **3.5**, 4 a and b).



Entry	Subentry	1st Buffer	2nd Buffer	Temperature (°C)
1	A	BugBuster MasterMix®	-	4
2	A		Loading buffer (LB)	4
	B			25
3	A		Elution buffer (EB)	4
	B			25
4	A	Loading buffer (LB)	-	4
5	A		Loading buffer (LB)	4
	B			25
6	A		Elution buffer (EB)	4
	B			25

Entry	Subentry	1st Buffer	2nd Buffer
1	A	BugBuster MasterMix®	-
	B		Loading buffer (LB)
	C		Elution buffer (EB)
2	A	Lysozyme solution	-
	B		Loading buffer (LB)
	C		Elution buffer (EB)
3	A	Loading buffer (LB)	-
	B		Loading buffer (LB)
	C		Elution buffer (EB)

Figure 3.5. *E. coli* protein expression and optimization of the cell lysis conditions allowed for the MBP-EPO-1 to be obtained. The first screening identified the combination of lysozyme and BugBuster MasterMix® as being efficient in promoting lysis of the cells (**a**, conditions 1A, 2A, 2B, 3A, 3B). The second screening revealed that the same outcome can be achieved when lysozyme is omitted from the lysis mixture (**b**, conditions 1A, 1B, 1C; LB: 50 mM TrisHCl, 150 mM NaCl, 5 % glycerol, pH 8.0; EB: 50 mM TrisHCl, 150 mM NaCl, 5 % glycerol, 500 mM imidazole, pH 8.0).

In the first screening, cells were treated with lysozyme and then either with BugBuster MasterMix® or loading buffer (LB: 50 mM TrisHCl, 150 mM NaCl, 5 % glycerol, pH 8.0) at 4 °C for 15 minutes. Then, each of the mixtures was further incubated with LB or an Ni^{2+}-affinity chromatography elution buffer (EB: 50 mM TrisHCl, 150 mM NaCl, 5 % glycerol, 500 mM imidazole, pH 8.0), at either 4 °C or 25 °C, for one hour. Since the best results were obtained when the cells were first incubated with

lysozyme and then with BugBuster MasterMix® (Figure **3.5**, a, conditions **1A**, **2A**, **2B**, **3A, 3B**) we next investigated whether the enzyme that hydrolyses the peptidoglycan of the cell wall was necessary in the mixture.[65]

Thus, a second screening was performed in which the cells were separately treated with BugBuster MasterMix®, lysozyme and LB for 20 minutes, at 25 °C. Then, to each of these, LB or EB was added, and the mixtures were further incubated for one hour at 25 °C. Again, BugBuster MasterMix® resulted to be instrumental in preserving the integrity of MBP-EPO-1 and the need of lysozyme was excluded (Figure **3.5**, b, conditions **1A**, **1B**, **1C**). Therefore, the large-scale lysis of the cells was performed using the commercial mixture of detergents further supplemented with protease inhibitors, and after optimization of the Ni^{2+}-affinity chromatography gradient (Figures SI 3.1 and 3.2), MBP-EPO-1 with acceptable purity was obtained (Figure **3.6**, a).

Intact protein LC-MS analysis of the dialyzed MBP-EPO-1 barely allowed the identification of the expected molecular mass (M_{calc} = 62768), but treatment of the LC-MS sample with a disulfide reducing agent such as dithiothreitol (DTT) clearly improved the molecular ion series that corresponded to our protein construct (Figure **3.6**, a, M_{calc} = 62772). Moreover, non-reducing SDS-PAGE analysis revealed bands indicative of high molecular weight (MW) species (lane **-**) which converged to form the expected MBP-EPO-1 band (~ 63 kDa) under reducing conditions (lane **+**). The LC-MS analyses were an indication that either the protein construct might have formed an intermolecular disulfide bridge at the mutated Cys38 residue (R = MBP-EPO-1) and the resulted dimer might be out of the set mass range of the instrument, or the MBP-EPO-1 (R = H) had poor ionization properties.

Conversely, the SDS-PAGE showed a different situation, one in which the natural four cysteine residues of EPO might have not formed the native intramolecular Cys7-Cys161 and Cys29-Cys33 disulfide bonds when expressed by the BL21(DE3) cells. Consequently, their subsequent intermolecular oxidation might have generated large disulfide-interconnected protein oligomers, observed as high MW species in the gel.

Therefore, the expression of MBP-EPO-1 was performed in a different *E. coli* bacterial strain, the SHuffleT7® competent cells. These cells constitutively express the disulfide bond isomerase DsbC, an enzyme that corrects mis-oxidized proteins into their native folded state, in the cytoplasm.[66] Nevertheless, no further improvement was observed when analysing the MBP-EPO-1 protein produced by this strain, neither by intact protein LC-MS nor by SDS-PAGE analysis (Figure **3.6**, b).

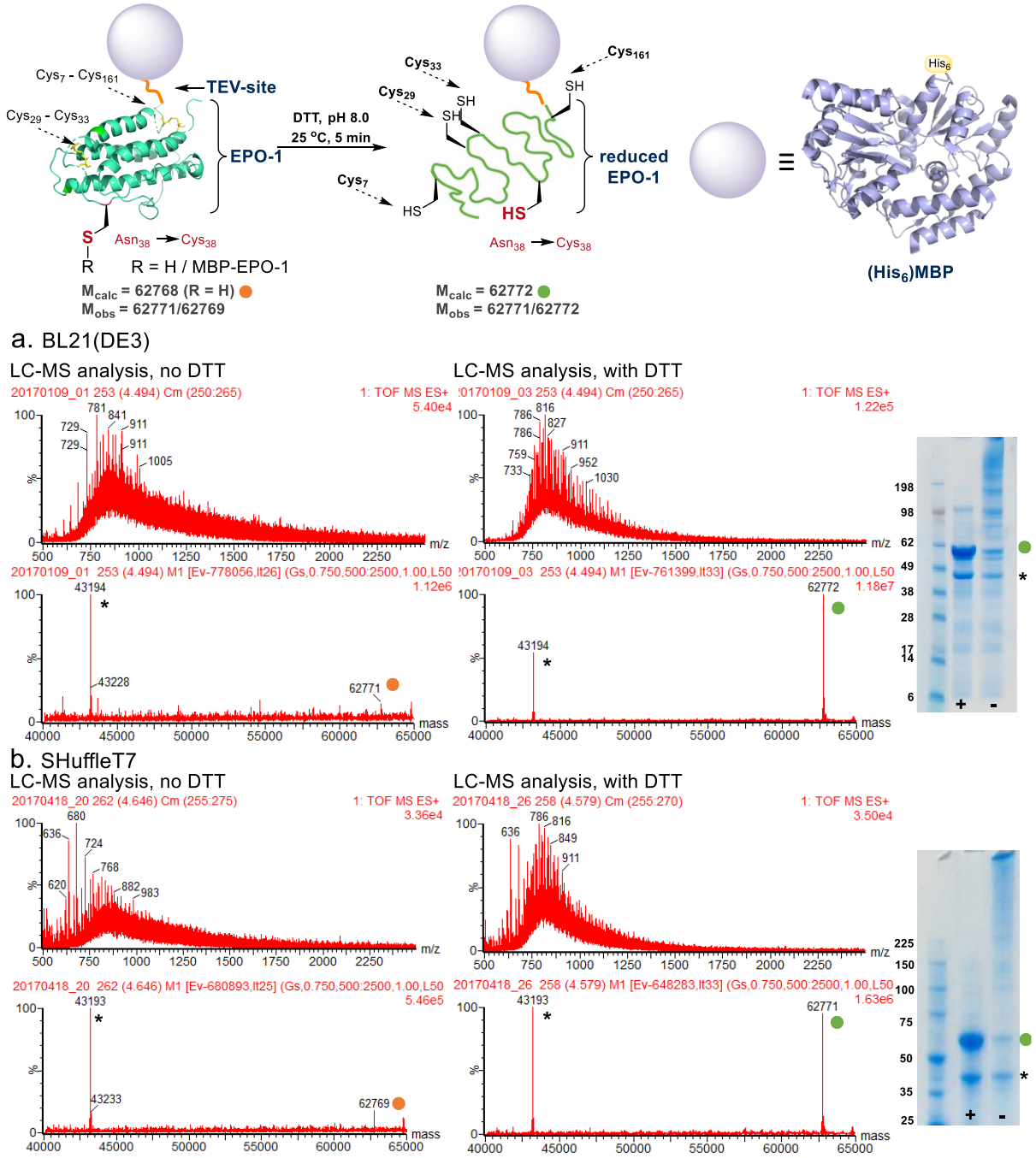


Figure 3.6. Characterization of MBP-EPO-1 by intact protein LC-MS and SDS-PAGE analyses. The expected molecular mass of the protein construct (M_{calc} = 62768/62772) expressed in BL21(DE3) (**a**) or SHuffleT7 (**b**) can be identified by LC-MS upon treatment with 10 mM DTT (right spectra) and by reducing (+) SDS-PAGE analyses (* = proteolysed MBP-EPO-1 impurity).

Then, the TEV-protease cleavage of MBP-EPO-1 was attempted, but the purification procedures that were performed to obtain pure EPO-1 were unsuccessful (Figure **3.7**).

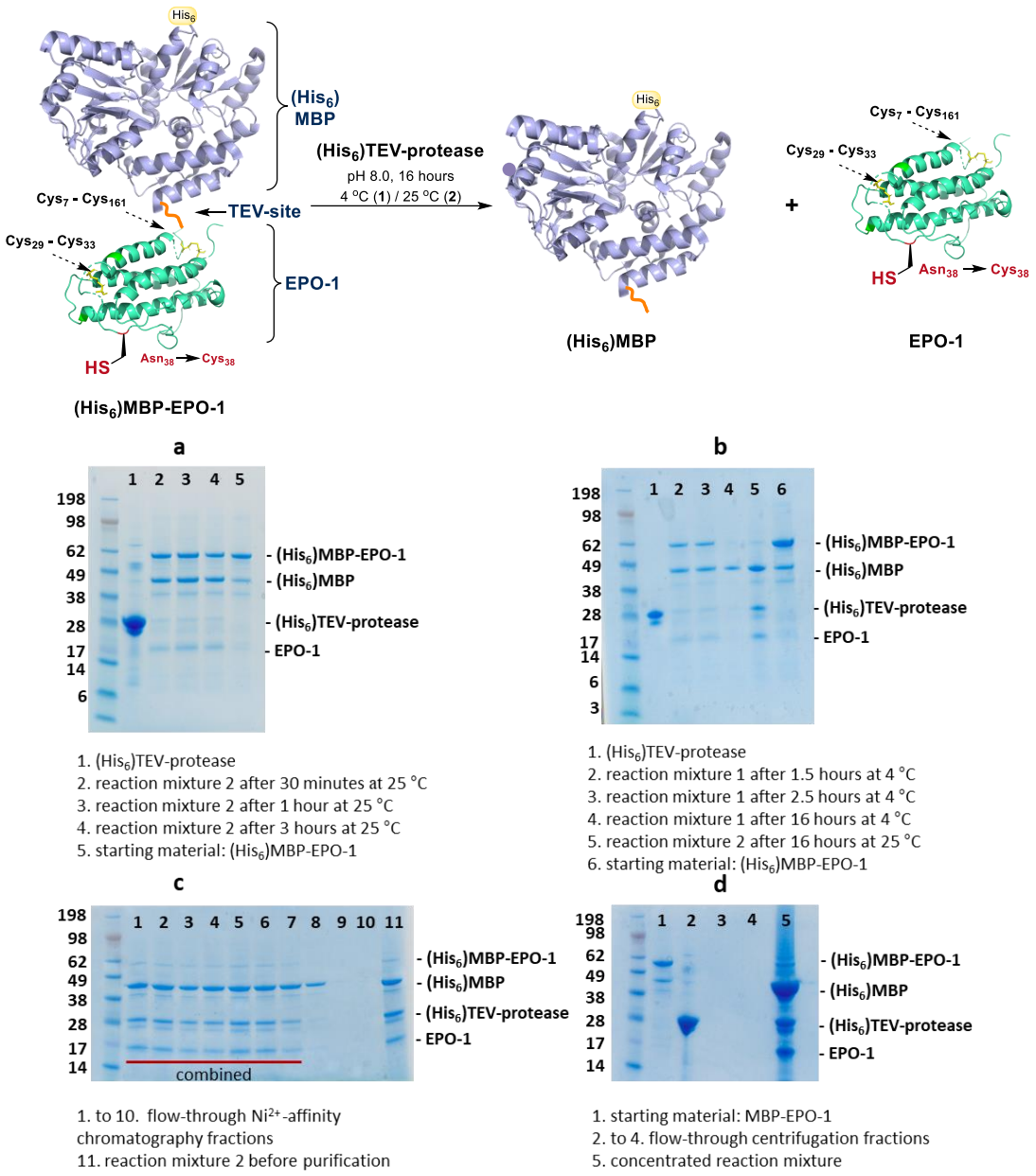


Figure 3.7. TEV-protease cleavage of MBP-EPO-1 and subsequent purification procedures did not afford pure EPO-1. Reducing SDS-PAGE analyses were performed to monitor both the progression of the TEV-cleavage of MBP-EPO-1 at 4 °C (1) and 25 °C (2), and the stability of the reaction mixtures (**a** and **b**). As revealed by the SDS-PAGE gels, the attempted purification by Ni^{2+}-affinity chromatography (**c**) and the following stepwise concentration of the combined fractions (**d**) did not afford pure EPO-1.

Two equal aliquots of MBP-EPO-1 were treated in parallel with the TEV-protease, and the resulted mixtures were incubated for 16 hours at 4 °C (1) and 25 °C (2), respectively. Since after 3 hours slow progression was observed in the reaction performed at 25 °C (Figure **3.7**, a), more enzyme was added to the reaction mixtures and incubation was continued for up to 16 hours.

Reducing SDS-PAGE gel analysis showed that the protein incubated at 4 °C had precipitated (Figure **3.7**, b, lane **4**), and thus, only the reaction mixture incubated at 25 °C (Figure **3.7**, b, lane **5**) was next subjected to Ni^{2+}-affinity chromatography. It was believed that the chosen purification

procedure will be able to afford the pure EPO-1 since after TEV-cleavage, it is the only protein in the mixture not bearing a His6-tag. However, SDS-PAGE analysis showed that this was not achieved. As expected, EPO-1 was not retained on the Ni^{2+}-affinity column, but the same occurred with the enzyme and the His6-MBP, and all three proteins were eluted with the flow-through (Figure **3.7**, c).

Despite this, the collected fractions were combined and stepwise concentrated by centrifugation. Interestingly, when analysing the flow-through fractions by SDS-PAGE, we could notice that although EPO-1 is rather small (~ 18 kDa), it did not pass a filter membrane with a 30 kDa MW cut-off (MWCO) whereas the TEV-protease (~ 28 kDa) was present in the flow-through, after the first centrifugation cycle (Figure **3.7**, d, lanes **2**-**4**). We speculated that this might have been caused either by the peculiar physico-chemical properties of EPO-1 that favoured binding to the filter membrane, or by its improper folding that caused protein aggregation or oligomerization.

Nevertheless, EPO-1 was separated from the other components of the reaction mixture by SDS-PAGE, and upon excision of the corresponding gel band (Figure **3.7**, d) it was digested with trypsin and submitted for LC-MS/MS analysis. These were able to confirm that the protein cleaved by the TEV-protease had the expected amino acid sequence, and that the Asn38 had been mutated to a cysteine residue (Cys38).

Given the results obtained when attempting to isolate EPO-1, we decided to proceed with experiments aimed at converting the MBP-EPO-1 construct (Figure **3.8**, 1) into the Dha-MBP-EPO-1 (3). First, the starting material was incubated with the disulfide reducing agent (50 equiv. DTT ~ 0.5 mM) for one hour at 4 °C to reduce any potential dimers, and the LC-MS analysis of an aliquot further treated with 10 mM DTT confirmed that the expected molecular mass (M_{calc} = 62772) could be detected (M_{obs} = 62775, Figure **3.8**, 1, a). Unexpectedly, upon incubating with the bis-alkylating reagent MDBP for 1.5 hours at 25 °C, the LC-MS analysis of an aliquot further treated with 10 mM DTT did not reveal any presence of either the expected mono-alkylated MBP-EPO-1 (M_{calc} = 62965, Figure **3.8**, 2, a) or the unreacted MBP-EPO-1 (M_{calc} = 62772, Figure **3.8**, 1, a). The results appeared even more surprising when the SDS-PAGE analysis confirmed that the integrity of the protein construct had not been substantially affected over the time course of the reaction.

Having obtained this unexpected outcome, we next envisioned to monitor the Dha reaction using a combined approach: an aliquot of the reaction mixture would be both treated with 10 mM DTT and cleaved by the TEV-protease before the LC-MS analysis (Figure **3.8**, b). Despite the low solubility of the non-glycosylated EPO at the optimum pH of the Dha reaction (pH = 8.0; calculated protein pI = 8.7[67]) and its propensity to undergo extensive precipitation (Figure **3.7**), the TEV-protease cleavage of the reduced LC-MS samples at pH 7.0 allowed us to monitor the Dha formation (Figure **3.8**, b). Thus, after 1.5 hours incubation with 300 equivalents of MDBP at 25 °C the newly optimized

LC-MS method (Figure SI 3.5) revealed that 21 % conversion to the desired mono-alkylated intermediate was achieved (Figure **3.8**, 2, b). Then, after 3.5 hours the DTT/TEV-cleavage method showed that 36 % conversion to the desired Dha-MBP-EPO-1 had been reached (Figure **3.8**, 3, b). The result was further confirmed by the almost quantitative conjugate addition of a simple model sugar, the 1-thio-*N*-acetylglucosamine (GlcNAc-SH). This led to the formation of GlcNAc-S-MBP-EPO-1 (Figure **3.8**, 4), as revealed by the DTT/TEV-cleavage method: 35 % of EPO-1 appeared to have been converted to GlcNAc-S-EPO-1.

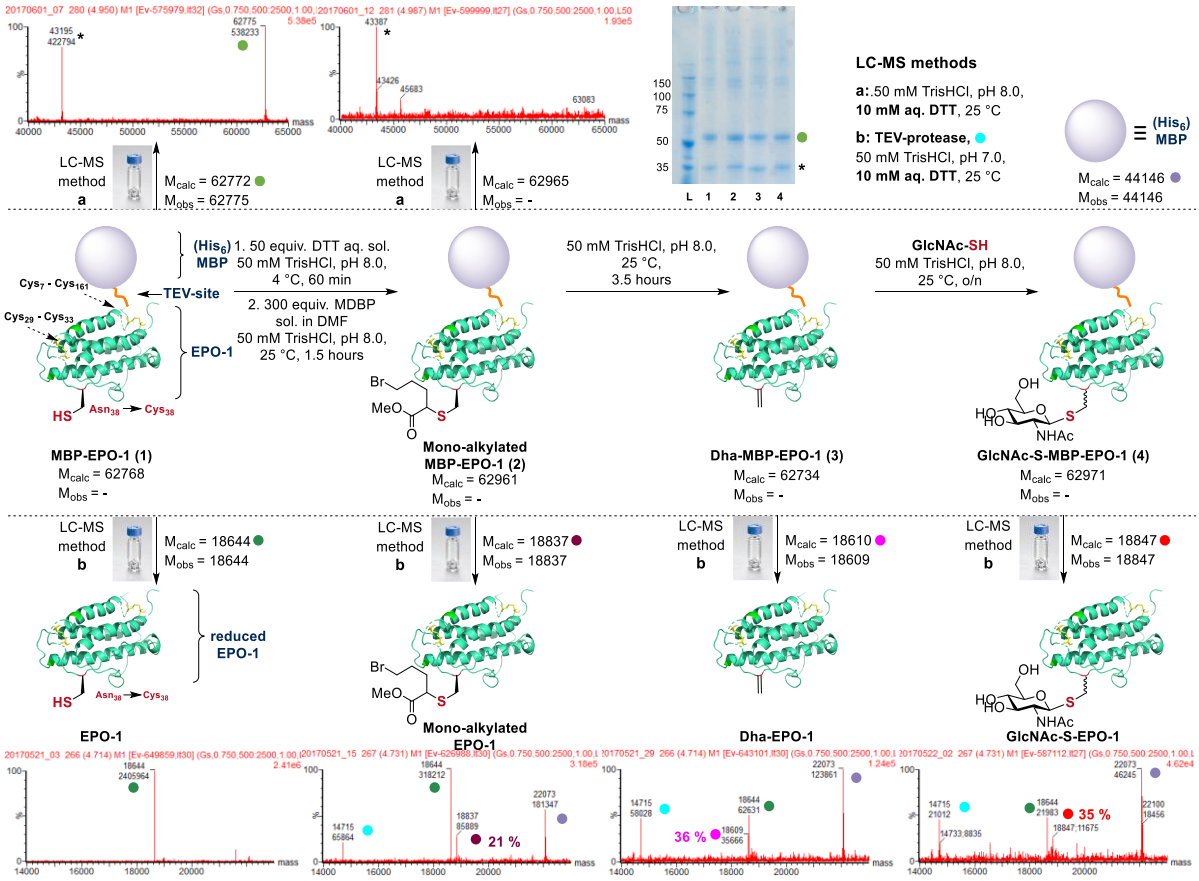
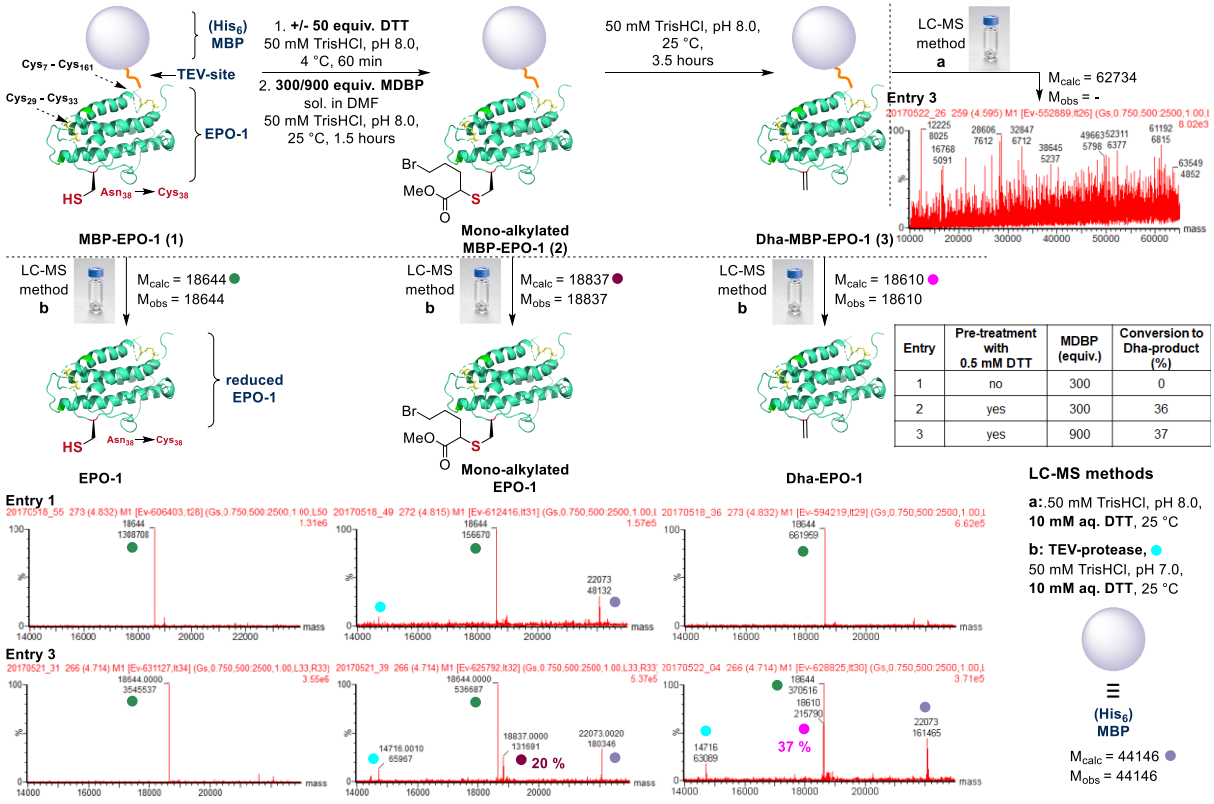


Figure 3.8. Formation of Dha-MBP-EPO-1 and GlcNAc-S-MBP-EPO-1. The chemical mutagenesis of MBP-EPO-1 (1) could not be monitored by treatment with DTT (10 mM) (**a**) but by using the combined DTT/TEV-cleavage method (**b**). The latter revealed that the mono-alkylated MBP-EPO-1 (2) was formed after 1.5 hours from the MDBP addition (21 % conversion of EPO-1 to mono-alkylated EPO-1) and the desired Dha-MBP-EPO-1 (2) was generated upon further incubation (36 % conversion of EPO-1 to Dha-EPO-1). The following thia-Michael addition of GlcNAc-SH to Dha-MBP-EPO-1 revealed that GlcNAc-S-MBP-EPO-1 (4) was formed almost quantitatively from 3 (35 % conversion of EPO-1 to GlcNAc-S- EPO-1). SDS-PAGE analyses of the reaction mixture were performed, and these revealed that the protein did not incur significant alterations at any of the four time points that led to the conversion of MBP-EPO-1 (1) to the desired GlcNAc-S-MBP-EPO-1 (4) (* = proteolysed MBP-EPO-1-derived impurity).

Interestingly, when we attempted the reaction without performing the initial reduction step of the starting material (50 equiv. DTT ~ 0.5 mM) no desired product could be observed in the reaction mixture that was analysed using the same DTT/TEV-cleavage method (Figure **3.9**, entry **1**). This was an indication that the starting material had dimerized through an intermolecular disulfide bridge and that the reducing treatment prior to the MDBP addition was required for the Dha-forming reaction to occur.

Moreover, an increase in the number of equivalents of MDBP from 300 (Figure **3.8** and Figure **3.9**, entry **2**) to 900 (Figure **3.9**, entry **3**) did not substantially improve the conversion of MBP-EPO-1 to the corresponding Dha-derivative (36 % vs. 37 %). The observed small difference between the two conversion yields corroborated with the almost complete conversion of Dha-MBP-EPO-1 to GlcNAc-S-MBP-EPO-1 (Figure **3.8**) prompted us to believe that the limiting factor had been the initial, intermolecular disulfide reduction step with DTT.



Entry	Pre-treatment with 0.5 mM DTT	MDBP (equiv.)	Conversion to Dha-product (%)
1	no	300	0
2	yes	300	36
3	yes	900	37

Figure 3.9. Additional insights into the Cys-Dha chemical mutagenesis of MBP-EPO-1. Bypassing the initial reducing step of the starting material with 50 equiv. DTT (0.5 mM) did not allow the formation of Dha-MBP-EPO-1 upon addition of the MDBP reagent (Entry 1). The reducing treatment of MBP-EPO-1 with 50 equiv. DTT at 4 °C for 60 minutes, followed by the addition of 300 equivalents MDBP afforded 36 % conversion to the desired Dha-product (Entry 2 and Figure 3.8). The same, initial reducing step of the starting material followed by the addition of 900 equivalents MDBP afforded 37 % conversion to the desired Dha-product (Entry 3). For the reaction described by the conditions summarized at entry 3 the results of the LC-MS analyses performed using the two methods, DTT treatment (**a**) and DTT/TEV-cleavage (**b**) are depicted for visual comparison.

Consequently, to improve the overall yield, we were tempted to explore different reaction conditions such as higher concentrations of reducing agent, a longer incubation time or a temperature increase during the initial reduction step with DTT.

However, to identify the site of the modification, LC-MS/MS analysis of the enzymatically digested GlcNAc-S-MBP-EPO-1 was first performed (Figure **3.10**).

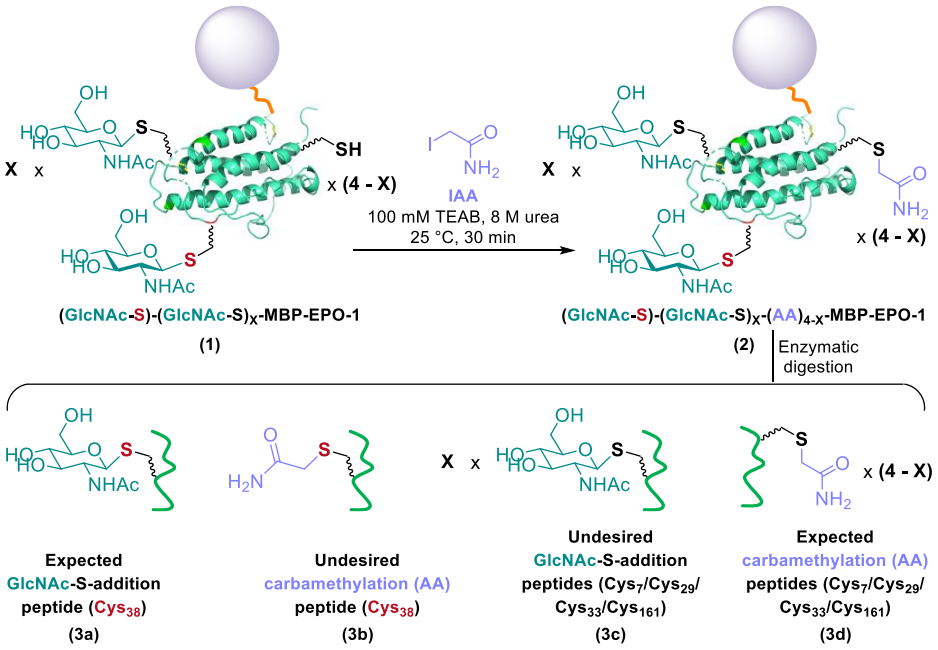


Figure 3.10. LC-MS/MS analysis of the *S*-glycosylated MBP-EPO-1 product (Figure 3.8) was instrumental in understanding the selectivity of the chemical mutagenesis/thia-Michael addition for the Cys38 residue of EPO-1. Dha-formation and subsequent GlcNAc-S-addition can occur at the mutated cysteine (**Cys38**) and/or at any of the native Cys residues (**Cys7/Cys29/Cys33/Cys161**) if the latter are not engaged in non-reactive disulfide bonds (1). Upon reduction and treatment with 2-iodoacetamide (IAA) an intermediate is obtained in which the previously unreacted cysteines are carbamethylated (2). Then, after enzymatic digestion, the GlcNAc-modification can be identified by LC-MS/MS at the expected Cys38 (3a) and/or at the undesired native cysteines (3c). The Cys residues found as carbamethylated – non-native (3b) or native (3d) indicate that prior to the proteomics analysis these had not undergone the GlcNAc-modification (X, number of modifications to native Cys; X = 0 – 4).

Upon reduction of any unmodified Cys residue of the *S*-glycosylated MBP-EPO-1 protein (Figure **3.10**, 1), the subsequent alkylation with 2-iodoacetamide (IAA) affords the carbamethylated (AA) *S*-glycosylated MBP-EPO-1 construct (2) that can be further subjected to enzymatic digestion. Then, to obtain information regarding the selectivity of the chemical mutagenesis/thia-Michael addition strategy for the unnatural Cys38 over any of the other four native residues, the generated (glyco)peptides are analysed by LC-MS/MS analysis. Ideally, the resulted glycoproteomics data should have indicated that the desired GlcNAc-modification had occurred at the expected Cys38-handle (3a). Moreover, a good conversion of the MBP-EPO-1 starting material to the corresponding Cys38-glycosylated derivative would ensure that the mutated residue would not substantially react with the

IAA to give the corresponding carbamethylated peptide (3b). Furthermore, if the native Cys residues would be found as GlcNAc-modified (3c) this would be an indication that they were not engaged in disulfide bridges in the first place and thus, they underwent Dha-conversion/thia-Michael glycosylation. Conversely, if involved in intra-/intermolecular disulfide bonds these residues would be chemically inert during the Cys-directed chemical manipulations and upon being subjected to the enzymatic digestion protocol (reduction and IAA alkylation) they would appear as carbamethylated (3d).

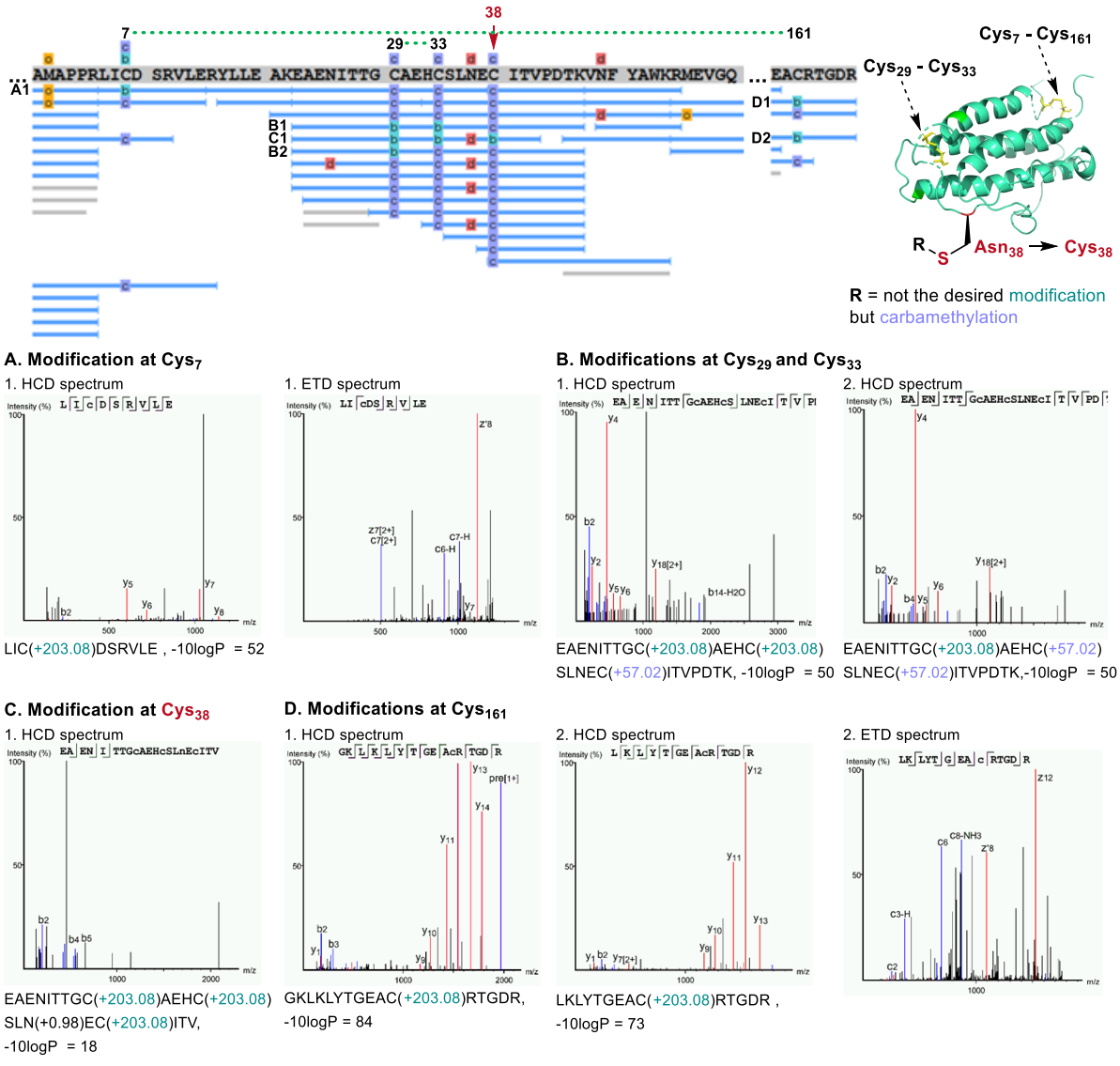


Figure 3.11. LC-MS/MS analysis of GlcNAc-S-MBP-EPO-1; amino acid sequence that covers the region of interest (EPO-1 part). The recorded higher-energy collisional dissociation (HCD) and electron-transfer dissociation (ETD) spectra showed that the desired GlcNAc-modification (cyan) was most reliably identified (-10logP score > 20[68]) on all the native cysteines (peptides A1, B1 and B2, D1 and D2) and mainly on the Cys161 residue of EPO (D1 and D2). The cysteine residue intended to be modified (Cys38) did not present the expected GlcNAc-addition (cyan) as revealed by the single, unreliably identified peptide C1 (-10logP score < 20). Moreover, the data indicated that numerous peptides containing the Cys38 had been carbamethylated (blue) during the enzymatic digestion procedure.

Several high-quality mass spectra reliably identified the GlcNAc-addition at all the cysteines that form intramolecular disulfide (S-S) bridges in the native, biologically active EPO (Figure **3.11**, A1, B1 and B2, D1 and D2). In addition, the desired modification was not found at the expected Cys38 residue, as revealed by the HCD-spectrum of the single GlcNAc-modified peptide that was identified with a rather low significance score (-10logP < 20) (Figure **3.11**, C1).[68] Instead, the proteomics analysis showed that all the other peptides containing the Cys38 had been carbamethylated during the enzymatic digestion, as a result of the Cys38 side-chain that, by virtue of not being modified, was able to effect a nucleophilic substitution on the employed iodoacetamide reagent.

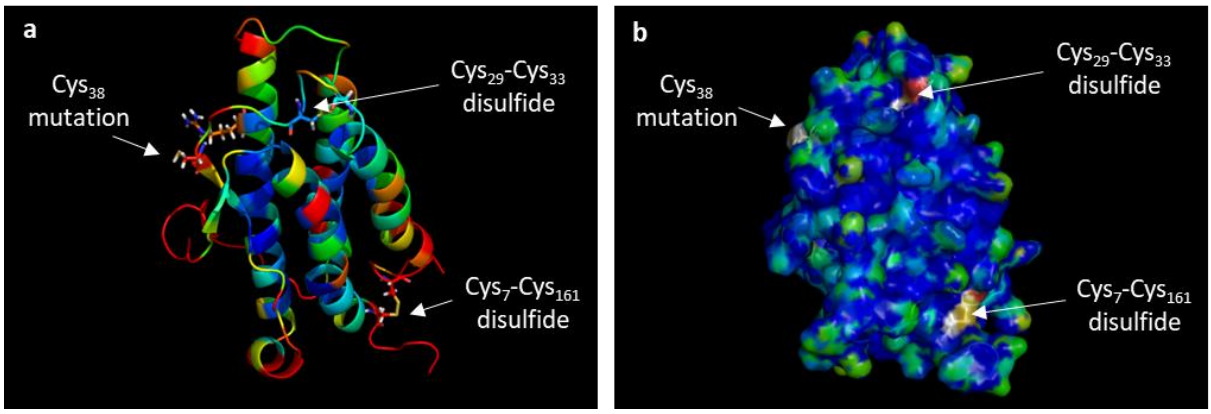


Figure 3.12. Molecular Dynamics (MD) of EPO-1. The results revealed that mutated Cys38 and disulfide Cys7-Cys161 are situated in highly flexible regions (red) in contrast to the disulfide Cys29-Cys33 which appeared to be rather rigid (blue) (**a**). In addition, Cys38 and disulfide Cys7-Cys161 have also high solvent-accessible surface area (**b**). Visualisation of the results was performed with PyMOL and colouring of the protein structure 1BUY[69] was done based on the obtained root-mean-square fluctuation (RMSF, **a**) and B-factor (**b**) values, respectively, using the RGB spectrum.[70]

Molecular dynamics (MD) simulations of EPO-1 were performed to evaluate the flexibility and accessibility of the five cysteine residues: the four natural ones involved in the disulfide bridges (Cys29-Cys33, Cys7-Cys161) and the fifth (Cys38) that was inserted by site-directed mutagenesis (Figure **3.12**). The results helped gain a better understanding of the previously obtained experimental data. For instance, the root-mean-square fluctuation (RMSF) of the residues showed that the mutated Cys is very flexible, but so is the region containing the Cys7-Cys161 disulfide (Figure **3.12**, a red).

Interestingly, these two cysteines (Cys7 and Cys161) are also the residues that were undoubtedly, chemically modified as determined by the previous LC-MS/MS analysis (Figure **3.11**). It was thus sensible to suspect that when overexpressed in a bacterial host as an MBP-construct, folding of the EPO-1 mutant into the state of lowest potential energy might favour shuffling or incorrect intra- and intermolecular S-S bonds formation. Moreover, the high flexibility of this area explains why ablation of this disulfide (Cys7-Cys161) through mutations to other residues other than cysteines causes immediate unfolding of the protein and therefore loss of its biological activity.[30]

Finally, the obtained solvent-accessible surface area revealed that in addition to the inserted Cys38-handle, the two native disulfide bridges are also not buried into the protein's core, especially the Cys7-Cys161 (Figure **3.12**, b). This makes them susceptible even to very low concentrations of DTT (50 equiv. ~ 0.5 mM), the reagent that had been used to reduce any potential S-S-linked dimers of the starting material (Figure **3.8** and Figure **3.9**).

3.2.2 Introduction of dehydroalanine to the MBP-fused EPO which was engineered for reduced aggregation and disulfide shuffling (MBP-EPO-2)

Having obtained the non-selective Dha-formation on the MBP-EPO-1 protein construct, we next decided to use the valuable insights we gained in the process and completely restructure our approach already from the initial phase, the heterologous expression of our target protein. Accordingly, our efforts were directed towards obtaining the site-selectively modified EPO by starting from a simpler system, one in which two of the four natural cysteine residues (Cys29 and Cys33) were mutated to tyrosines (Figure **3.13**, a).

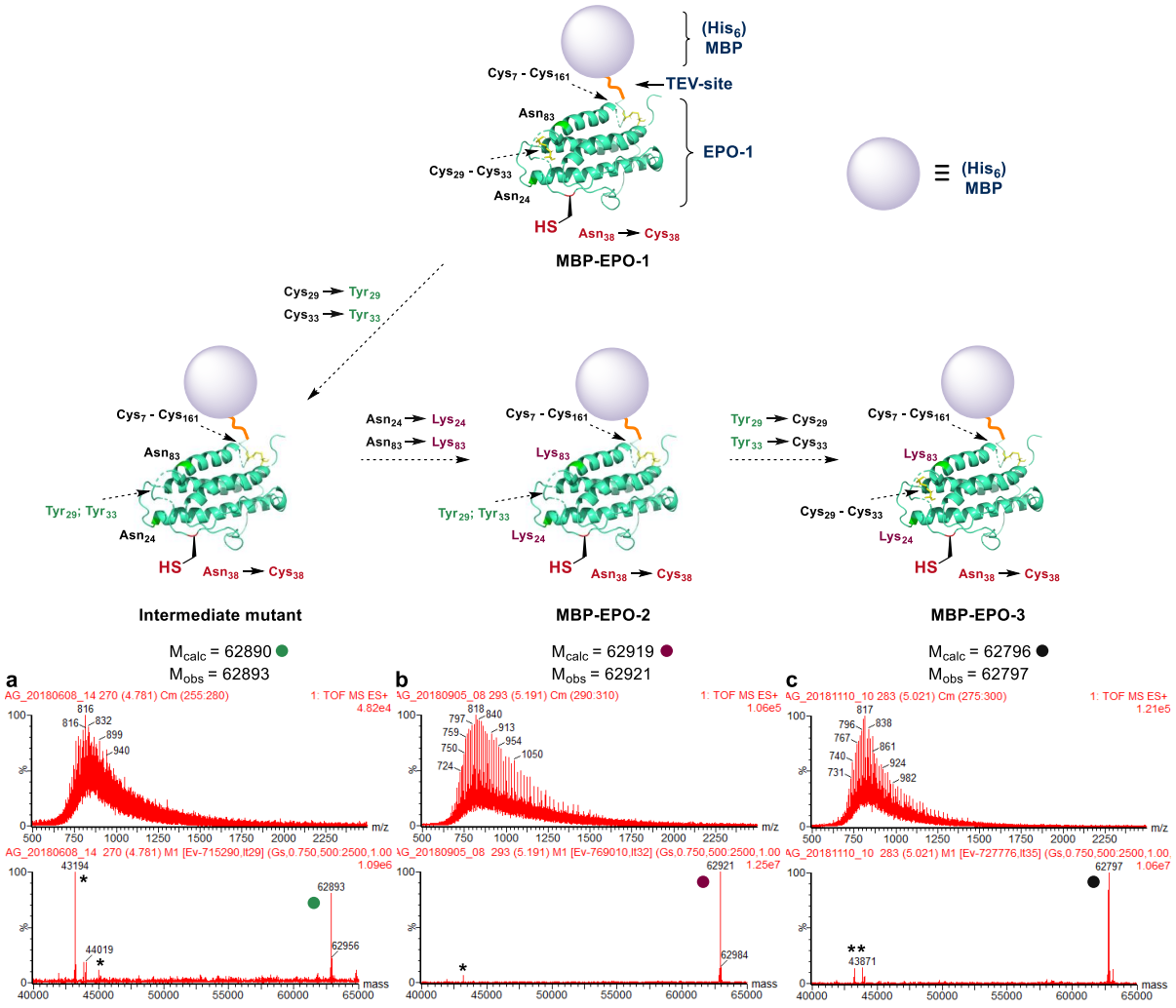


Figure 3.13. Various MBP-EPO mutants with different structural and physico-chemical properties were obtained which did not require reducing treatment with DTT (10 mM) for LC-MS characterization. First, the Cys29 and Cys33 residues forming a disulfide bridge were replaced with tyrosines, but the resulted intermediate mutant afforded a deconvoluted LC-MS spectrum of rather poor quality in absence of DTT (**a**, M_{calc} = 62890). Conversely, the additional mutation of Asn24 and Asn83 (the other two *N*-glycosylation sites of native EPO) to lysines significantly improved the molecular ion series of the obtained MBP-EPO-2 construct (**b**, M_{calc} = 62919). Reintroduction of the Cys29 and Cys33 residues into the latter gave a protein variant (MBP-EPO-3) that still had acceptable ionization properties (**c**, M_{calc} = 62796; * = proteolysed impurities).

Thus, it was envisioned that a π-π stacking interaction between the aromatic side chains of these residues[71] could benignly replace the native but dispensable[31] disulfide bridge and favour the site-selective Dha-formation at the Cys38.

Moreover, a previously published study had reported that mutation of the *N*-glycosylation sites of EPO to lysine (Lys) residues reduced the aggregation of the protein.[72] Therefore, in addition to the Asn38 glycosylation site which we had already mutated to a cysteine intended for further derivatization, the other two *N*-glycosylation sites (Asn24 and Asn83) were mutated to Lys residues (Figure **3.13**, b).

Finally, a last mutant was obtained which featured the Lys24 and the Lys83 residues instead of the two native Asn-glycosylation sites, and all of the four Cys residues that are naturally locked into the Cys29-Cys33 and Cys7-Cys161 disulfide bridges (Figure **3.13**, c).

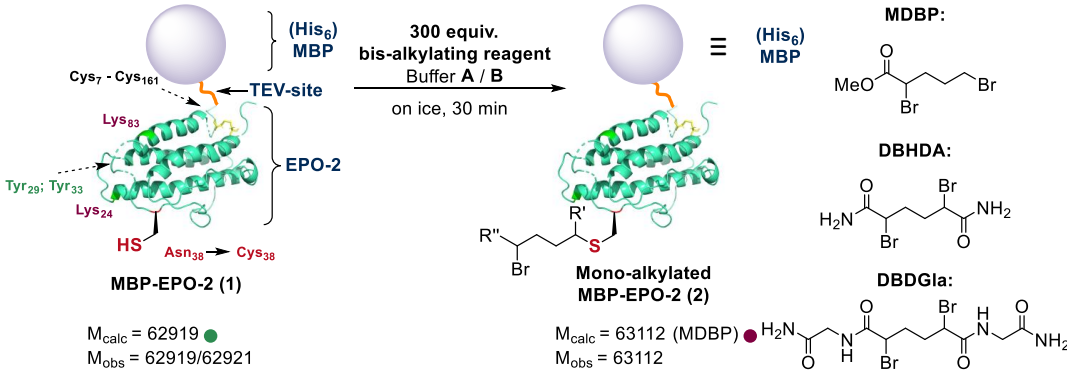
SHuffleT7® cells were transformed with the newly obtained plasmids, and upon lysis of the cells using the previously optimized conditions, purification of the EPO variants was performed at neutral pH instead of pH 8.0. Considering a pK value of 8.6 for the side-chain ionizable group of cysteine[73] we reasoned that the lower pH would prevent the intra- and intermolecular thiolate-mediated disulfide exchange reactions. In addition, to minimize the dimerization at the Cys38 chemical handle and thus avoid the reducing pre-treatment with 0.5 mM DTT required for the Dha reaction (Figure **3.9**), the slow dialysis that was necessary to remove the imidazole after the Ni^{2+}-affinity chromatography, was replaced by the faster MBP-affinity purification.

Non-reducing SDS-PAGE analysis of MBP-EPO-2 showed less oligomerization compared to previously obtained results (Figure SI 3.3) and intact protein LC-MS analysis of this construct (not treated with DTT) afforded the expected molecular mass (M_{calc} = 62919, Figure **3.13**, b). Mutation of the two Asn residues at the *N*-linked glycosylation sites to Lys changed the calculated isoelectric point (pI)[67] of the protein from 8.7 (MBP-EPO-1) to 8.8 (intermediate mutant), 9.2 (MBP-EPO-2) and 9.0 (MBP-EPO-3). This had presumably resulted in an increase in the net positive charge of the protein at neutral pH and had decreased its likelihood of undergoing aggregation. Moreover, the strategy not only led to higher repulsive electrostatic interactions in solution which decreased aggregation, but it also improved the efficiency of the positive ionization during the intact protein LC-MS analyses of aliquots that were not treated with 10 mM DTT.

Attempted conversion of the MBP-EPO-2 variant into the corresponding Dha-MBP-EPO-2 using previously employed reaction conditions (Figure **3.9**, entry **3**) could still not be monitored *via* DTT treatment of the analysed LC-MS samples (Figure SI 3.6). We again attributed the obtained results to the extensive aggregation phenomena that the protein underwent in the TrisHCl buffer, upon

addition of the bis-alkylating reagent MDBP. Therefore, we next embarked on exploring different Dha-forming reaction conditions.

In the first screening, in addition to MDBP, two bis-alkylating reagents with different physico-chemical properties and reactivities were tested in parallel: the more hydrophilic and symmetrical reagent 2,5-dibromohexadiamide (DBHDA) and its bulkier analogue, the *N*1,*N*6-bis(2-amino-2-oxoethyl)-2,5-dibromohexanediamide (DBDGla) (Figure **3.14**).[74] Moreover, two new buffers were also explored for their potential to suppress protein aggregation and improve the stability of MBP-EPO-2 during the Cys-Dha conversion. Thus, the protein construct was incubated with each of the three reagents on ice, either in a TrisHCl buffer with increased ionic strength and supplemented with maltose, the natural ligand of MBP[75] (Figure **3.14**, buffer A), or in the buffer used in Chapter 2, a NaPi buffer containing the protein denaturant guanidinium hydrochloride[76] (Figure **3.14**, buffer B).



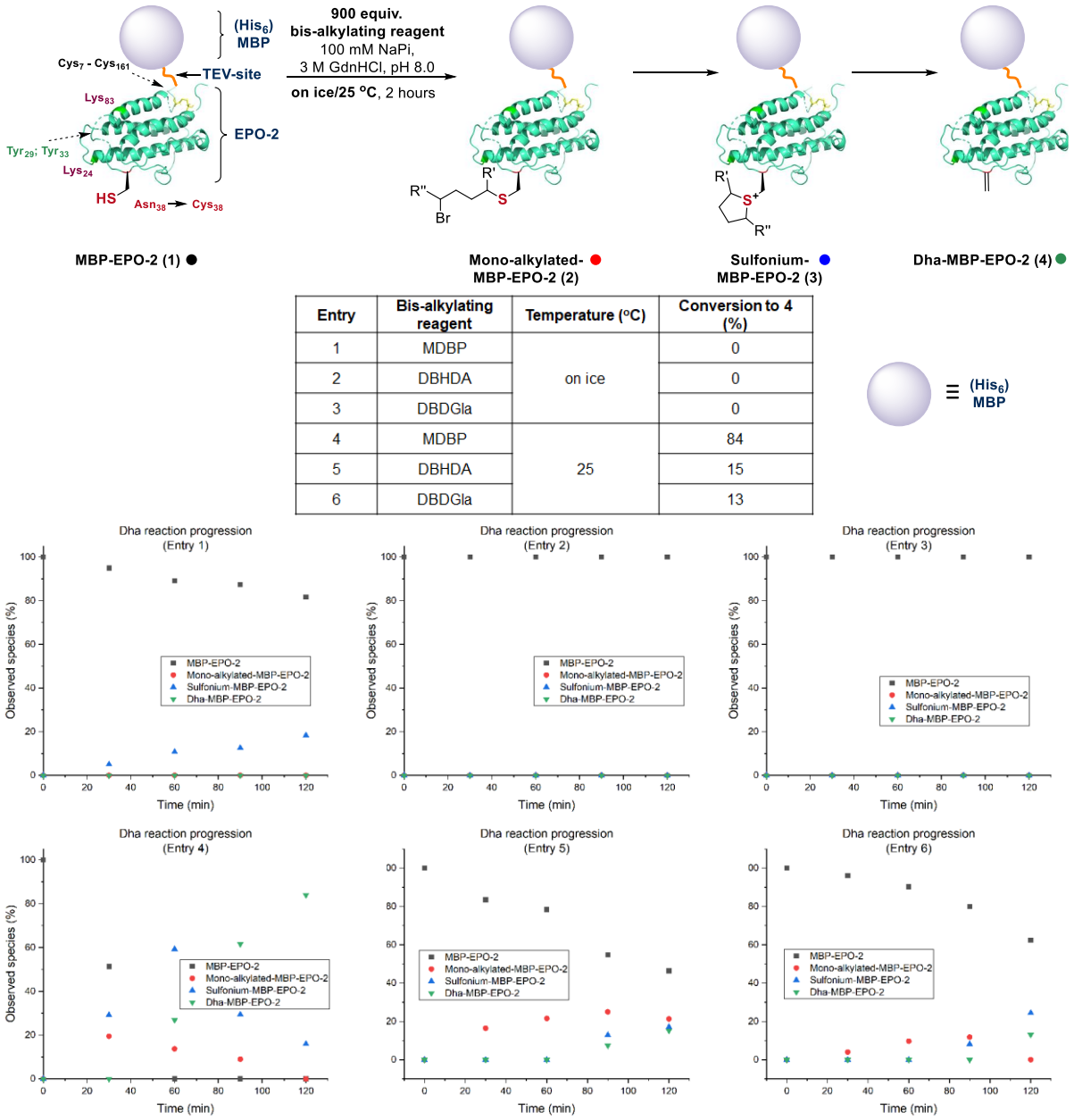
Entry	Bis-alkylating reagent	Buffer	Conversion to 2 (%)
1	MDBP	**A**: 50 mM TrisHCl, 150 mM NaCl, 100 mM maltose, pH 8.0	45 (degradation)
2	DBHDA		0
3	DBDGla		0
4	MDBP	**B**: 100 mM NaPi, 3 M GdnHCl, pH 8.0	0
5	DBHDA		0
6	DBDGla		0

Figure 3.14. Initially attempted conditions for the formation of Dha-MBP-EPO-2. 45 % conversion to the mono-alkylated intermediate was observed (M_{calc} = 63112) when MBP-EPO-2 was incubated with 300 equiv. MDBP on ice, in a maltose-containing buffer (A). However, the employed reaction conditions caused protein instability (entry 1; * = proteolysed impurity). All the other reactions described in the table, employing the bis-alkylating reagents DBHDA and DBDGla in buffer A (entries 2 and 3) or in a guanidinium hydrochloride (GdnHCl)-containing buffer (B: entries 4-6) did not reveal progression towards the formation of the mono-alkylated intermediate. Nevertheless, the integrity of the starting material remained intact in the reaction mixtures that employed buffer B (e. g. entry 4).

Upon incubation of the protein with 300 equivalents MDBP LC-MS analysis performed after just 30 minutes revealed that 45 % of the starting material had been converted to the expected mono-alkylated intermediate (Figure **3.14**, entry **1**).

In addition, the appearance of a protein cleavage side-product (M_{obs} = 43868) and the significant decrease of the MS signal intensity were an indication that the desired Cys-Dha conversion had been accompanied by significant protein degradation and aggregation. Conversely, all the other reaction conditions that were explored did not lead to the formation of the mono-alkylated intermediate, but the protein stability did not appear to have been affected (Figure **3.14**, entry **4** as representative example).

Therefore, in a second attempt, higher amounts of MDBP, DBHDA and DBDGla (900 equivalents) were screened in the chaotropic buffer (buffer B), both on ice and at 25 °C, for two hours (Figure **3.15**). Interestingly, when DBHDA and DBDGla were used at 0 °C, the reaction did not proceed towards the formation of the mono-alkylated MBP-EPO-2 (2, Figure **3.15**, entry **2** and **3**), whereas limited conversion, less than 20 % to the sulfonium-MBP-EPO-2 (3) was observed when MDBP was employed (Figure **3.15**, entry **1** and Figure SI 3.7). The mono-alkylation reaction rate was greatly increased when any of the three reagents was used at a higher temperature (25 °C) but again, conversion to the final Dha-MBP-EPO-2 (4) product was significantly higher (84 %) when the bis-alkylating reagent MDBP was employed (Figure **3.15**, entry **4**, and Figure SI 3.7). In addition, when the protein was incubated with DBHDA and DBDGla at 25 °C significant amounts of starting material (1) were still present in the reaction mixture after two hours and only limited conversions to 4 could be achieved: 15 % and 13 % of Dha-MBP-EPO-2, respectively (Figure **3.15**, entries **5** and **6,** and Figure SI 3.7).

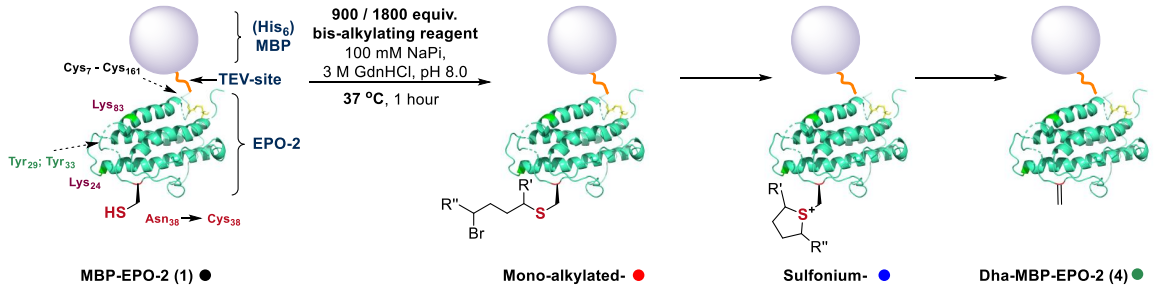


Entry	Bis-alkylating reagent	Temperature (°C)	Conversion to 4 (%)
1	MDBP	on ice	0
2	DBHDA		0
3	DBDGla		0
4	MDBP	25	84
5	DBHDA		15
6	DBDGla		13

Figure 3.15. Formation of Dha-MBP-EPO-2 using 900 equivalents bis-alkylating reagent on ice or at 25 °C. Limited conversion of MBP-EPO-2 (1) towards the sulfonium-MBP-EPO-2 intermediate (3) was observed (18 %) when the starting material was incubated with MDBP, on ice (entry 1). Incubation of 1 with DBHDA or DBDGla on ice did not proceed to give the mono-alkylated intermediate (2), the sulfonium-derivative (3) or the Dha-MBP-EPO-2 (4) product (entries 2 and 3). Excellent conversion to 4 was achieved (84 %) when MBP-EPO-2 was incubated with MDBP at 25 °C, and the employed reaction conditions preserved the stability of the protein (entry 4). Limited conversions of 1 to 4 were reached when MBP-EPO-2 was incubated at 25 °C with DBHDA (15 %, entry 5) or DBDGla (13 %, entry 6), and the predominant specie observed in these two reaction mixtures was still the starting material: 47 % and 63 %, respectively. Graphs show ratios (%) of observed species at different time points of the Cys-Dha transformation, as determined by LC-MS analyses of aliquots taken in parallel for each of the six single-replicate experiments.

To understand better the limitations of the chemical mutagenesis of MBP-EPO-2 to Dha-MBP-EPO-2, a third screening was performed in which the starting material was incubated with 900 or 1800 equivalents of MDBP, DBHDA or DBDGla, at an even higher temperature: 37 °C (Figure **3.16**). Pleasingly, when 900 equivalents of MDBP were added, full conversion to the desired Dha-MBP-EPO-

2 was achieved in just one hour (Figure **3.16**, entry **1**), and unsurprisingly, a similar outcome was obtained with 1800 equivalents of MDBP (Figure **3.16**, entry **4** and Figure SI 3.8).



Entry	Bis-alkylating reagent	Equiv.	Conversion to 4 (%)
1	MDBP	900	100
2	DBHDA		26
3	DBDGla		21
4	MDBP	1800	100 (degradation)
5	DBHDA		62
6	DBDGla		50

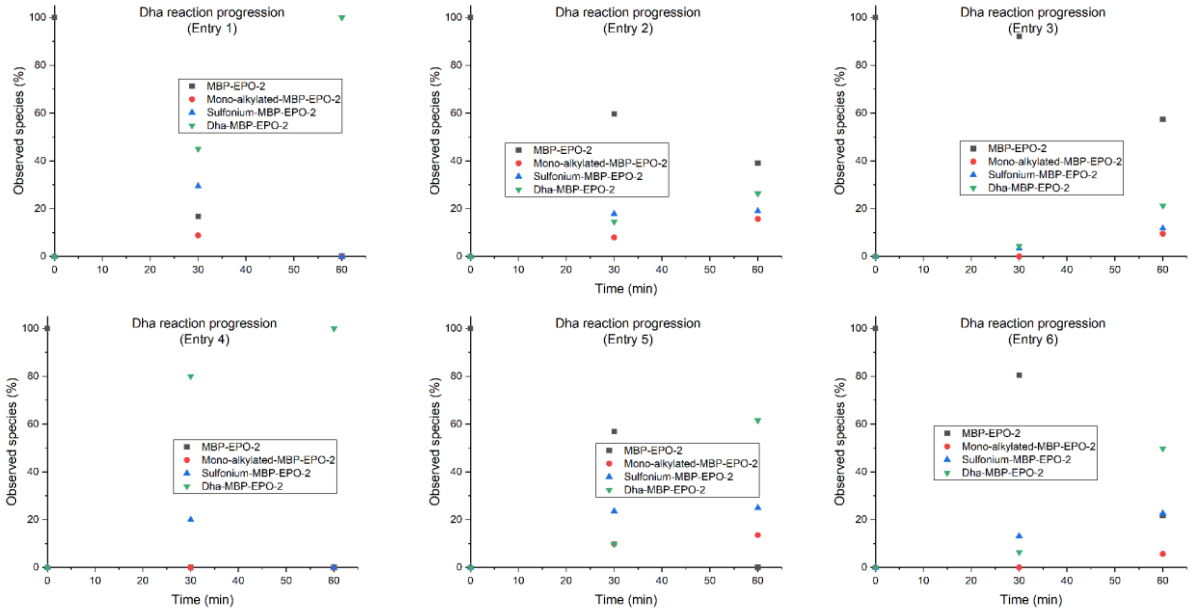


Figure 3.16. Formation of Dha-MBP-EPO-2 using 900 or 1800 equivalents bis-alkylating reagent at 37 °C. Excellent conversion of MBP-EPO-2 (1) towards the Dha-MBP-EPO-2 (4) product was observed (100 %) when the starting material was incubated for one hour with 900 equiv. (entry 1) or 1800 equiv. (entry 4) MDBP, at 37 °C. Incubation of 1 with DBHDA or DBDGla gave satisfactory amounts of the mono-alkylated intermediate (2), the sulfonium-derivative (3) or the Dha-MBP-EPO-2 (4) product (entries 2 and 3). Moderate conversions of 1 to 4 were reached when MBP-EPO-2 was incubated at 37 °C with DBHDA (62 %, entry 5) or DBDGla (50 %, entry 6). Graphs show ratios (%) of observed species at different time points of the Cys-Dha transformation, as determined by LC-MS analyses of aliquots taken in parallel for each of the six single-replicate experiments.

On the other hand, upon incubation with DBHDA (Figure **3.16**, entry **2** and **5**) or DBDGla (Figure **3.16**, entry **3** and **6**), the Dha-formation did not reach completion when 1800 equivalents were used. In fact, even when employing such a high excess of bis-alkylating reagent DBHDA or DBDGla, the two reaction mixtures revealed that the Dha-MBP-EPO-2 had been generated only in moderate yields: 62 % (Figure **3.16**, entry **5**) and 50 % (Figure **3.16**, entry **6**), respectively.

Thus, it was inferred from the latter results that when MBP-EPO-2 was incubated with MDBP at 37 °C, a lower amount of bis-alkylating reagent (300 or 600 equivalents MDBP) could have been employed.

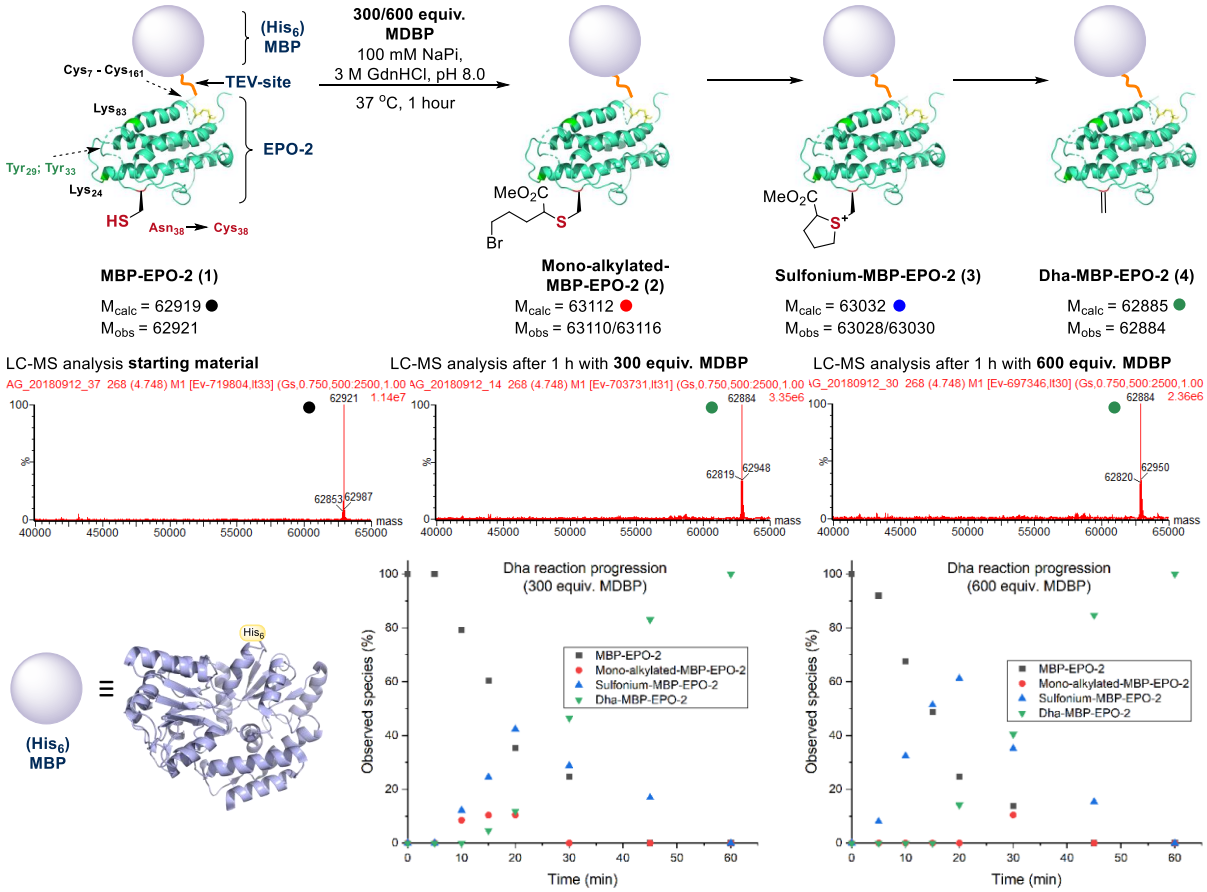


Figure 3.17. Formation of Dha-MBP-EPO-2 using 300 or 600 equivalents MDBP at 37 °C. MDBP was able to fully convert the MBP-EPO-2 to the desired Dha-derivative in just one hour at 37 °C, and a higher number of equivalents (600) only increased the rate at which the sulfonium-intermediate was formed but not the overall reaction progression. Graphs show ratios (%) of observed species at different time points of the Cys-Dha transformation, as determined by LC-MS analyses of aliquots taken in parallel for each of the two single-replicate experiments.

Indeed, the last screening demonstrated that the quantitative conversion to Dha-MBP-EPO-2 could also be achieved even when using only 300 equivalents of MDBP (Figure **3.17**). Moreover, the addition of 600 equivalents of this bis-alkylating reagent revealed that the rate at which the sulfonium-MBP-EPO-2 was formed had been slightly increased, but the overall rate of the Dha reaction was unaltered. The latter could presumably be due to the α'-β elimination, the rate-limiting step of the Cys-Dha transformation,[58] which unsurprisingly, was not influenced by the addition of more MDBP.

To identify the site of the Dha-modification, upon thia-Michael addition of β-mercaptoethanol (βME), enzymatic digestion and LC-MS/MS analysis were performed. For this EPO variant and mutant MBP-EPO-3, the βME was used instead of the GlcNAc-SH probe because it helped avoid the use of an

ETD fragmentation method. This MS/MS technique is crucial for the identification of labile protein modifications (e.g. glycosylation)[77] but it is often more difficult to implement.

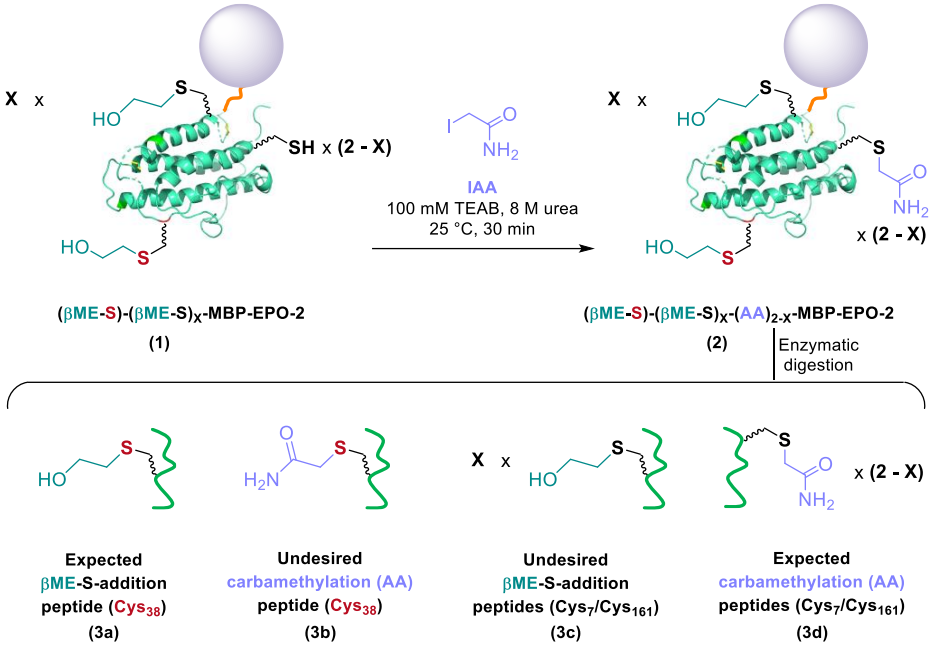


Figure 3.18. LC-MS/MS analysis of the βME-S-MBP-EPO-2 product was instrumental in understanding the selectivity of the chemical mutagenesis/thia-Michael addition for the Cys38 residue of EPO-2. Dha-formation and subsequent βME-S-addition can occur at the mutated cysteine (**Cys38**) and/or at any of the native Cys residues (**Cys7**/**Cys161**) if the latter are not engaged in a non-reactive disulfide bond (1). Upon reduction and treatment with 2-iodoacetamide (IAA) an intermediate is obtained in which the previously unreacted cysteines are carbamethylated (2). Then, after enzymatic digestion, the βME-modification can be identified by LC-MS/MS at the expected Cys38 (3a) and/or at the undesired native cysteines (3c). The Cys residues found as carbamethylated – non-native (3b) or native (3d) indicate that prior to the proteomics analysis these had not undergone the βME-modification (X, number of modifications to native Cys; X = 0 – 2).

Several high-quality mass spectra reliably identified the βME-addition at the two natural cysteines that form the Cys7-Cys161 intramolecular S-S bridge in the native EPO (Figure **3.19**, A and C). However, in this mutant, which lacked the natural Cys29 and Cys33 residues, the expected modification was found at the Cys38 residue. This was revealed by the HCD spectrum of the fragmented, modified peptide which was characterized by the highest significance score (-10logP = 78) (Figure **3.19**, B).

The data was corroborated with results obtained from a test reaction aimed at quantifying the free Cys residues of MBP-EPO-2 using the Ellman's reagent (Figure SI 3.11). Unsurprisingly, the generated data showed that this variant participated to the disulfide-exchange reaction with mainly one cysteine. Thus, we believe that the latter is the Cys38 that had been converted to Dha and identified *via* proteomics analysis.

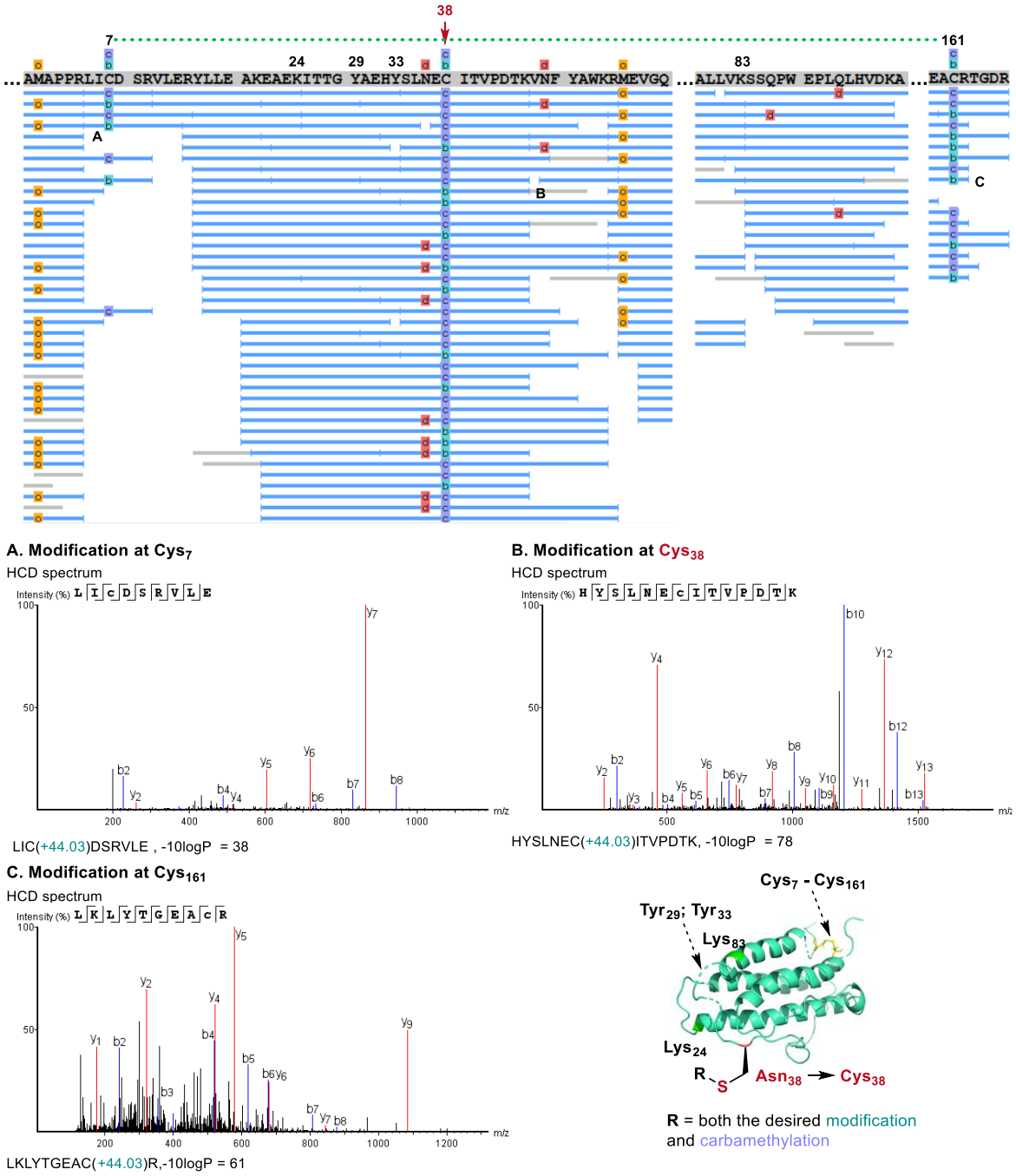


Figure 3.19. LC-MS/MS analysis of βME-S-MBP-EPO-2; amino acid sequence that covers the region of interest (EPO-2 part). The recorded HCD spectra showed that the desired βME-modification (cyan) was most reliably identified (-10logP score = 78)[68] on the expected cysteine residue (Cys38, peptide B) but also on the two native residues (Cys7, peptide A) and Cys161, peptide C) The latter had been also mostly carbamethylated (blue) during the enzymatic digestion procedure.

3.2.3 Introduction of dehydroalanine to the MBP-fused EPO which was engineered for reduced aggregation (MBP-EPO-3)

Pleased with the results obtained in section 3.2.2, we decided to apply the optimized conditions found for the Dha-formation on mutant MBP-EPO-2 to the protein construct in which the native Cys29 and Cys33 residues were reintroduced. Again, non-reducing SDS-PAGE analysis performed during the purification procedures showed less oligomerization compared to MBP-EPO-1 (Figure SI 3.4) and intact protein LC-MS analysis revealed the desired molecular mass upon MS deconvolution without addition of DTT to the analysed sample (M_{calc} = 62796, Figure **3.20**, 1).

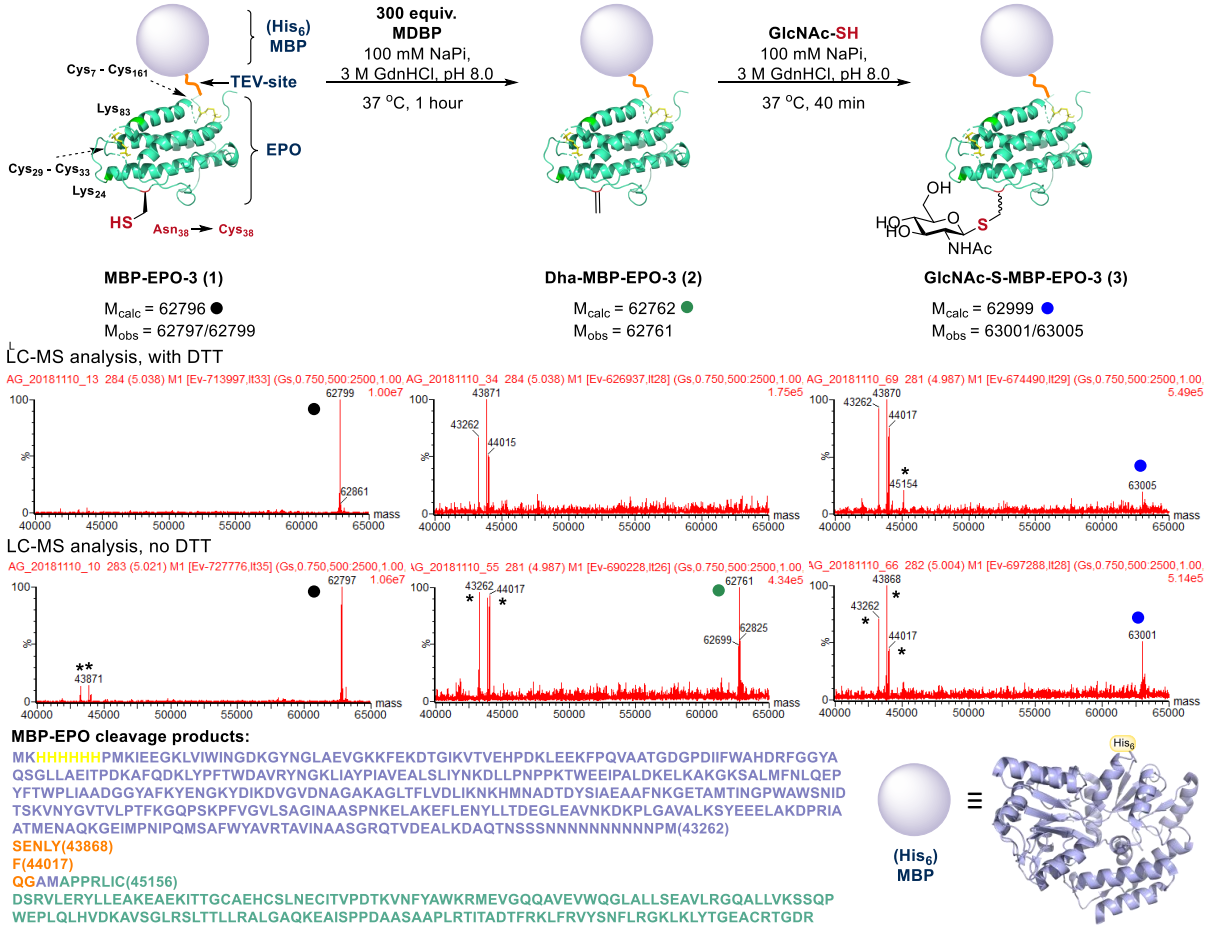


Figure 3.20. Formation of Dha-MBP-EPO-3 using 300 equivalents MDBP at 37 °C, and subsequent thia-Michael addition of GlcNAc-SH. The previously optimized reaction conditions for the Cys-Dha conversion were applied to the MBP-EPO-3 mutant (1) and these afforded the desired Dha-MBP-EPO-3 (2). The LC-MS identification was possible (M_{obs} = 62761) only when the samples were not treated with DTT. Generation of the desired Dha-derivative was accompanied by decreased protein stability and formation of proteolysed species (M_{obs} : 43262, 43868, 44017 and 45154) that corresponded to various construct fragments: MBP (43262), MBP-SENLY (43868), MBP-SENLYF (44017) and MBP-EPO(until C7: 45156). Nevertheless, upon incubation with GlcNAc-SH quantitative conversion to GlcNAc-S-MBP-EPO (3) was obtained, and the LC-MS identification was straightforward (M_{obs} = 63001) in the samples not treated with DTT.

The new mutant thus obtained was incubated with MDBP using the reaction conditions previously optimized for MBP-EPO-2 (Figure **3.17**). LC-MS analyses performed after one hour showed mostly presence of proteolysed products (M_{obs} : 43262, 43869, 44017 and 45154) and in the LC-MS samples treated with DTT, the desired product Dha-MBP-EPO-3 could not be identified (Figure **3.20,** 2). Presumably, this outcome had been caused again by the extensive aggregation of the protein which resulted in the observed significant decrease of its ionization properties. Interestingly though, addition of DTT to the analysed LC-MS samples again improved ionization of the starting material compared to the cleavage products present in the mixture, but not of the Dha-MBP-EPO-2 derivative or the subsequently obtained GlcNAc-S-MBP-EPO-2 (Figure **3.20,** 3).

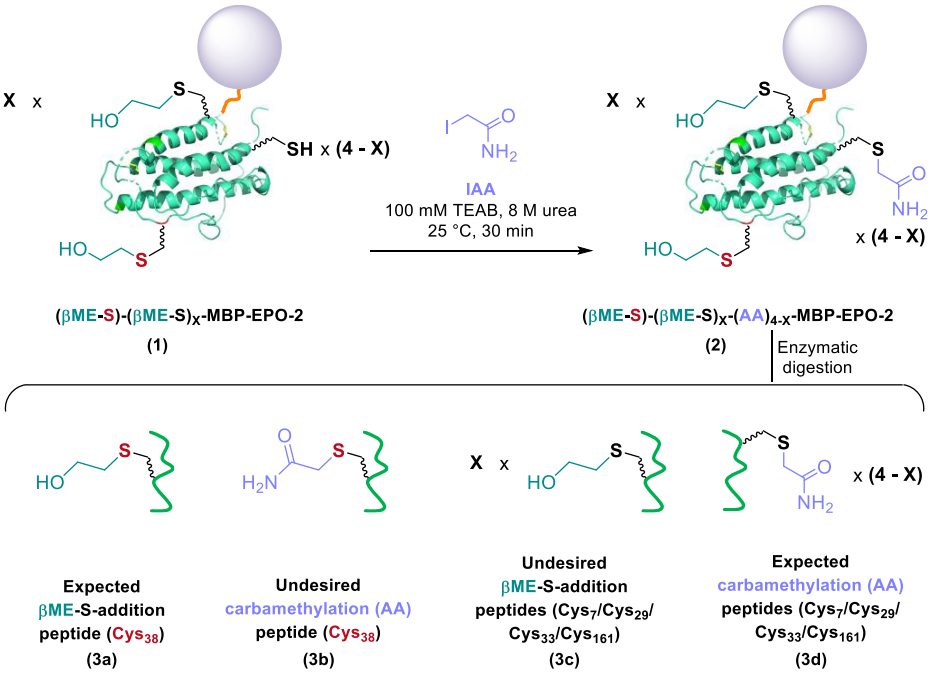


Figure 3.21. LC-MS/MS analysis of the βME-S-MBP-EPO-3 product was instrumental in understanding the selectivity of the chemical mutagenesis/thia-Michael addition for the Cys38 residue of EPO-3. Dha-formation and subsequent βME-S-addition can occur at the mutated cysteine (**Cys38**) and/or at any of the native Cys residues (**Cys7**/**Cys29**/**Cys33**/**Cys161**) if the latter are not engaged in non-reactive disulfide bonds (1). Upon reduction and treatment with 2-iodoacetamide (IAA) an intermediate is obtained in which the previously unreacted cysteines are carbamethylated (2). Then, after enzymatic digestion, the βME-modification can be identified by LC-MS/MS at the expected Cys38 (3a) and/or at the undesired native cysteines (3c). The Cys residues found as carbamethylated – non-native (3b) or native (3d) indicate that prior to the proteomics analysis these had not undergone the βME-modification (X, number of modifications to native Cys; X = 0 – 4).

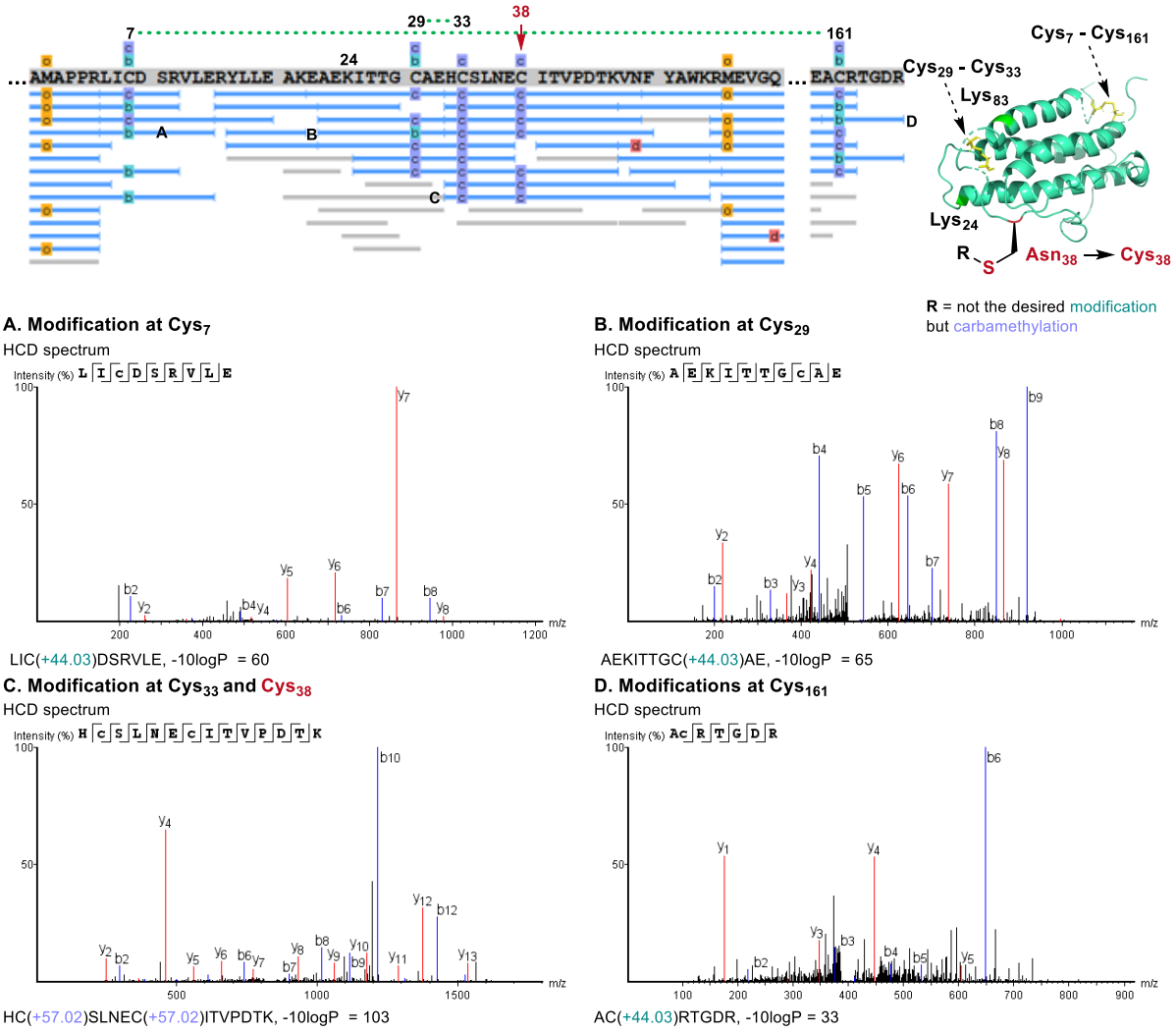


Figure 3.22. LC-MS/MS analysis of βME-S-MBP-EPO-3; amino acid sequence that covers the region of interest (EPO-3 part). The recorded HCD spectra showed that the desired βME-modification (cyan) was most reliably identified on three of the native cysteine residues: Cys7 (peptide A), Cys29 (peptide B) and Cys161 (peptide D) but not on the native residue Cys33. Moreover, the data indicated that numerous peptides contained the native Cys33 and the mutated Cys38 that had been carbamethylated (blue) during the enzymatic digestion procedure and, among these, peptide C was identified with a very high confidence score (-10logP score = 103).[68]

To identify the site of the Dha-modification, upon thia-Michael addition of β-mercaptoethanol (βME) and enzymatic digestion LC-MS/MS analysis were performed. Surprisingly, the proteomics results obtained for βME-MBP-EPO-3 revealed that the desired modification had taken place on three of the four native cysteine residues (Cys7, Cys29 and Cys161, Figure **3.22**, A, B and D), but the βME-addition was not identified at the desired Cys38 mutated residue nor at the native Cys33. Moreover, the peptide carrying the βME-modified Cys29 that should have been involved in forming one of the intramolecular S-S bridges, was characterized by one of the highest significance scores (-10logP = 65) (Figure **3.22**, B). In addition, the inserted Cys38 only appeared to have been carbamethylated during the enzymatic digestion protocol (-10logP = 103) (Figure **3.22**, C). This was an indication that before digestion, the mutated cysteine had not been modified to a Dha residue.

We believe that when overexpressed in a bacterial host as an MBP-construct, folding of the MBP-EPO-3 mutant into the state of lowest energy has favoured incorrect intramolecular S-S bonds formation (shuffling of the disulfides). We speculate that the native Cys33 residue was not crosslinked to its natural partner, the Cys29, but with the adjacent cysteine introduced at the Asn38-glycosylation site. Thus, this mismatch left exposed the Cys29 which was converted by the MDBP bis-alkylating reagent to a Dha residue.

3.3 Conclusions and future perspectives

This chapter has described the implementation of the Cys-Dha chemical mutagenesis/thia-Michael addition strategy which had been attempted to conjugate glycans to an aggregation prone protein that contains multiple native Cys residues (Cys29-Cys33 and Cys7-Cys161): erythropoietin (EPO).

Soluble expression and isolation of the MBP-construct of EPO in which one of the three *N*-glycosylation sites (Asn38) had been mutated to a Cys (MBP-EPO-1) was successfully achieved. Conversely, the purification procedures that were attempted upon cleavage of this construct with the designated enzyme TEV-protease did not afford pure EPO-1. This was attributed to the poor solubility properties of non-glycosylated EPO-1 which had readily precipitated upon cleavage from the MBP solubilizing tag.

Characterization of MBP-EPO-1 by intact protein LC-MS analysis was only possible upon treatment of the analysed sample with 10 mM disulfide reducing agent dithiothreitol (DTT). However, the attempted conversion of MBP-EPO-1 to its Dha-derivative in 50 mM TrisHCl pH 8.0 buffer and using MDBP as a bis-alkylating reagent could not be monitored using the same reducing DTT protocol.

We initially speculated that the addition of this disulfide reducing agent to the sample analysed by LC-MS was able to increase the ionization of MBP-EPO-1 *via* unfolding of the protein and exposure of additional amino acid residues that could be protonated. The fact that the same effect could not be obtained during the Cys-Dha conversion suggested that many different mechanisms might have been at play. For instance, when fused to the EPO-1 protein, the MBP tag did favour the expression of the construct in the cytoplasm. However, as found in the case of other problematic proteins such as the MBP-fused HPV E6 oncoprotein,[78] soluble "micelle-like" aggregates[79] might have been formed that gave good quality MS spectra when disrupted by DTT but unresolved MS spectra upon addition of the hydrophobic MDBP solution.

For the latter, a new sample preparation protocol was developed that involved DTT reduction and cleavage by the TEV-protease of small aliquots taken from the reaction mixture at different time points. The DTT/TEV-cleavage method generated a three-component protein mixture which could be separated using an optimized LC-MS gradient. Consequently, formation of Dha-MBP-EPO-1 and subsequent thia-Michael addition of GlcNAc-SH to the generated Dha residue were monitored on EPO-1, part of the original MBP-EPO-1 construct. We proved that after dialysis, the Cys-Dha conversion occurred in moderate yield (36 %) only if the starting material had undergone a preceding treatment with DTT (50 equiv. ~ 0.5 mM) aimed at reducing any dimers that might have been formed *via* an

intermolecular (Cys38-Cys38) disulfide bridge. Nevertheless, the LC-MS/MS analysis of the enzymatically digested GlcNAc-S-MBP-EPO-1 revealed that even if dimerization had occurred, the pre-treatment with DTT had non-selectively reduced the intramolecular disulfide bonds of EPO-1 (Cys7-Cys161 and Cys29-Cys33) rather than the intended intermolecular disulfide bridge. Then, the data obtained experimentally was corroborated with results generated by the Molecular dynamics (MD) simulations of EPO-1 which revealed that the two native disulfide bridges are not buried into the core protein structure (the Cys7-Cys161 in particular), but rather susceptible to the reducing action of DTT that had been employed to generate MBP-EPO-1 monomers.

Consequently, these valuable insights directed us towards exploring new reaction conditions for the formation of Dha-MBP-EPO starting from a mutant that lacked the non-critical Cys29-Cys33 disulfide bond and in which the other two *N*-glycosylation sites (Asn24 and Asn83) were mutated to Lys residues. Thus, we were able to obtain the MBP-EPO-2 variant that showed improved ionization properties and consequently, it did not require treatment with 10 mM DTT of the LC-MS analysed sample. Nevertheless, the attempted conversion of MBP-EPO-2 to its Dha-derivative in the TrisHCl buffer and using MDBP as a bis-alkylating reagent could not be monitored again by LC-MS without using the reducing DTT protocol. The latter was an indication that the protein reaction mixture was still undergoing aggregation but optimisation of the Dha reaction conditions (different buffers, temperatures, and bis-alkylating reagents) resulted in the successful formation of the desired Dha-MBP-EPO-2. By employing the MDBP bis-alkylating reagent at 37 ^{o}C, in a pH 8.0 buffer supplemented with 3M GdnHCl the Dha-MBP-EPO-2 was generated in almost quantitative yield. Thia-Michael addition of β-mercaptoethanol to the Dha-tagged protein ensured the complete conversion of Dha-MBP-EPO-2 to the corresponding β-ME-derivative (βME-MBP-EPO-2). LC-MS/MS analysis of the enzymatically digested βME-MBP-EPO-2, the site-selective Dha-addition product, revealed that the desired Cys-Dha conversion had most likely taken place at the expected Cys38 residue and presumably to a less extent at the two native cysteines forming the Cys7-Cys161 disulfide bridge.

Finally, reintroduction of the native cysteines forming the intramolecular Cys29-Cys33 disulfide bond of EPO generated the MBP-EPO-3 variant that displayed acceptable ionization properties, and thus it did not require treatment with 10 mM DTT for its LC-MS characterization. By employing the reaction conditions that had been previously optimised for the MBP-EPO-2 mutant the Dha-MBP-EPO-3 was successfully generated with an excellent yield. Moreover, side-products could also be observed as the Dha reaction progressed, and after careful analysis these were attributed to MBP-containing fragments of various length that had been cleaved off from the MBP-EPO-3 construct. Nevertheless, upon thia-Michael addition of β-ME the LC-MS/MS analysis of the enzymatically digested βME-MBP-EPO-3 revealed that the desired Cys-Dha conversion had mainly taken place at the

native Cys29 residue and presumably to a less extent at the two native cysteines forming the Cys7-Cys161 disulfide bridge. In addition, the data strongly indicated that insertion of the unnatural Cys38 residue might have favoured the formation of the shifted Cys33-Cys38 disulfide bond instead of the native Cys29-Cys33.

Future work in this area should primarily focus on generating a homogeneous EPO glycoform *via* the Cys-Dha chemical mutagenesis/thia-Michael addition strategy by investigating the potential of the other two *N*-glycosylated asparagines (Asn24 and Asn83) to serve as glycoconjugation sites. While we do anticipate that a similar outcome for the Dha reaction may be achieved when expressing the Cys24-containing protein in a bacterial host, we believe that explorations at the Cys-mutated Asn83 amino acid position would be worth further consideration.

3.4 Experimental section

3.4.1 Plasmid DNA details and protein sequences

The plasmids utilised in the course of this work were as follows:

MBP-EPO-1 (C7-N24-C29-C33-C38-N83-C161): donated by Dr. Maola Khan, previously prepared in a pETM-41 vector;

*MKHHHHHHPMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWAH DRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKELKAK GKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAF NKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAV NKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQTVDEALKDAQTN SSSNNNNNNNNNNPMSENLYFQGAM*APPRLICDSRVLERYLLEAKEAENITTGCAEHCSLNECITVPDTKVNFYA WKRMEVGQQAVEVWQGLALLSEAVLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPDA ASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR; pI = 8.7 (EPO-1)

The *N*-terminus of the protein construct contains a polyhistidine tag which is highlighted in yellow and a TEV-cleavable peptide which is highlighted in green. The amino acid residues that contain the MBP protein tag and come from the vector are italicized. For convenience, this is abbreviated as **MBP** in the following protein sequences.

Intermediate mutant (C7-N24-Y29-Y33-C38-N83-C161): generated from MBP-EPO-1 through single site-directed mutagenesis

MBPAPPRLICDSRVLERYLLEAKEAENITTGYAEHYSLNECITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTITADTFRKLFRVYSNFLRG KLKLYTGEACRTGDR; pI = 8.8 (intermediate EPO)

MBP-EPO-2 (C7-K24-Y29-Y33-C38-K83-C161): generated from the intermediate mutant through double, site-directed mutagenesis

MBPAPPRLICDSRVLERYLLEAKEAEKITTGYAEHYSLNECITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA VLRGQALLVKSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTITADTFRKLFRVYSNFLRG KLKLYTGEACRTGDR; pI = 9.2 (EPO-2)

MBP-EPO-3 (C7-K24-C29-C33-C38-K83-C161): generated from MBP-EPO-2 through single site-directed mutagenesis

MBPAPPRLI**C**DSRVLERYLLEAKEAE**K**ITTG**C**AEH**C**SLNE**C**ITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA VLRGQALLV**K**SSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTITADTFRKLFRVYSNFLRG KLKLYTGEA**C**RTGDR; pI = 9.0 (EPO-3)

3.4.2 Protein expression and characterisation

3.4.2.1 Protein expression of different MBP-EPO variants:

Transformation and expression of MBP-EPO variants in BL21(DE3) competent cells

To 50 µL BL21(DE3) competent cells 2 µL β-mercaptoethanol were added and the cells were incubated on ice for 10 minutes. The plasmid encoding the MBP-EPO variant (5 µL, 10 ng/µL) was added followed by incubation on ice for 30 minutes. Then the tube was heated in a water bath at 42 °C for 45 seconds and immediately placed on ice for 2 minutes. Luria-Bertani medium (1 mL, preheated to 37 °C) was added and the mixture was incubated at 37 °C and 250 rpm for 1 hour. Transformed cells were plated onto Luria-Bertani agar plates supplemented with kanamycin (50 µg/mL) and these were incubated at 37 °C overnight. Cell colonies were picked and inoculated into Luria-Bertani medium (20 mL) supplemented with kanamycin (50 µg/mL) and shaken for 16 hours at 37 °C and 225 rpm.

Starter cultures were transferred into 800 mL Luria-Bertani medium supplemented with kanamycin (50 µg/mL) and incubated at 37 °C and 200 rpm until OD_{600} ~ 0.4-0.6. Media was then cooled down to room temperature, IPTG (0.5 mM) was added to induce protein expression and incubation was continued for 16 hours at 20 ^{0}C and 180 rpm. Then, cells were harvested by centrifugation (10 min, 8000 rpm, 4 ^{0}C) and the obtained cell pellet was stored at – 80 ^{0}C.

Transformation and expression of MBP-EPO variants in SHuffleT7 competent cells

To 50 µL SHuffleT7 competent cells plasmid encoding the MBP-EPO variant (2 µL, 37 ng/µL) and the cells were incubated on ice for 30 minutes. Then the tube was heated in a water bath at 42 °C for 30 seconds and immediately placed on ice for 5 minutes. SOC medium (950 µL, preheated to 25 °C) was added and the mixture was incubated at 30 °C and 250 rpm for 1 hour. Transformed cells were plated onto Luria-Bertani agar plates supplemented with kanamycin (40 µg/mL) preheated to 30 °C, and these were incubated at 37 °C overnight. Cell colonies were picked and inoculated into Luria-Bertani medium (10 mL) supplemented with kanamycin (40 µg/mL) and shaken for 16 hours at 37 °C and 225 rpm.

Starter cultures were transferred into 800 mL Luria-Bertani medium supplemented with kanamycin (40 µg/mL) and incubated at 30 °C and 200 rpm until OD_{600} ~ 0.4-0.6. Media was then

cooled down to room temperature, IPTG (1 mM) was added to induce protein expression and incubation was continued for 21 hours at 16 ^{0}C and 180 rpm. Then, cells were harvested by centrifugation (10 min, 8000 rpm, 4 ^{0}C) and the obtained cell pellet was stored at – 80 ^{0}C.

3.4.2.2 Optimization of lysis conditions for cells overexpressing the MBP-EPO-1 variant

First screening of lysis conditions:

The wet cell pellet (~ 100 mg) was suspended in 200 µL ddH_2O water and 200 µL aqueous solution of Lysozyme (0.05 mg/mL) was added and the resulting suspension was divided into two equal aliquots A and B (150 µL each). Then, these were treated as follows:

A: 500 µL BugBuster MasterMix™ were added and the mixture was incubated for 15 minutes at 4 °C and after centrifugation (5 min, 15 000 g, 4 °C) the supernatant was collected and treated as follows:

1. 100 µL control 1, 1 hour at 4 °C;
2. 100 µL treated with 100 µL loading buffer for 1 hour at 4 °C
3. 100 µL treated with 100 µL loading buffer for 1 hour at 25 °C
4. 100 µL treated with 100 µL elution buffer for 1 hour at 4 °C
5. 100 µL treated with 100 µL elution buffer for 1 hour at 25 °C

B: 500 µL Loading buffer were added and the mixture was incubated for 15 minutes at 4 °C and after centrifugation (5 min, 15 000 g, 4 °C) the supernatant was collected and treated as follows:

1. 100 µL control 2, 1 hour at 4 °C;
2. 100 µL treated with 100 µL loading buffer for 1 hour at 4 °C
3. 100 µL treated with 100 µL loading buffer for 1 hour at 25 °C
4. 100 µL treated with 100 µL elution buffer for 1 hour at 4 °C
5. 100 µL treated with 100 µL elution buffer for 1 hour at 25 °C

Loading buffer (LB): 50 mM TrisHCl, 150 mM NaCl, 5 % glycerol, pH 8.0

Elution buffer (EB): 50 mM TrisHCl, 150 mM NaCl, 5 % glycerol, 500 mM imidazole, pH 8.0

Second screening of lysis conditions:

The wet cell pellet (~ 100 mg) was suspended in 300 µL ddH_2O water and the resulting suspension was divided into three equal aliquots A, B and C (100 µL each). Then, these were treated as follows:

A: 500 µL BugBuster MasterMix were added and the mixture was vortexed for 5 seconds, incubated for 20 minutes at 25 °C and after centrifugation (5 min, 15 000 g, 4 °C) the supernatant was collected and treated as follows:

1. 100 µL control 1, 1 hour at 25 °C;
2. 100 µL treated with 100 µL loading buffer for 1 hour at 25 °C
3. 100 µL treated with 100 µL elution buffer for 1 hour at 25 °C

B: 500 µL solution of Lysozyme (0.05 mg/mL) were added and the mixture was vortexed for 5 seconds, incubated for 20 minutes at 25 °C and after centrifugation (15 000 g, 5 min, 4 °C) the supernatant was collected and treated as follows:

1. 100 µL control 2, 1 hour at 25 °C;
2. 100 µL treated with 100 µL loading buffer for 1 hour at 25 °C
3. 100 µL treated with 100 µL elution buffer for 1 hour at 25 °C

C: 500 µL loading buffer were added and the mixture was vortexed for 5 seconds, incubated for 20 minutes at 25 °C and after centrifugation (15 000 g, 5 min, 4 °C) the supernatant was collected and treated as follows:

1. 100 µL control 3, 1 hour at 25 °C;
2. 100 µL treated with 100 µL loading buffer for 1 hour at 25 °C
3. 100 µL treated with 100 µL elution buffer for 1 hour at 25 °C

Loading buffer (LB): 50 mM TrisHCl, 150 mM NaCl, 5 % glycerol, pH 8.0

Elution buffer (EB): 50 mM TrisHCl, 150 mM NaCl, 5 % glycerol, 500 mM imidazole, pH 8.0

Optimized cell lysis procedure:

To 35 mL cold solution of BugBuster MasterMix™, 4 tablets cOmplete EDTAfree™ cocktail of protease inhibitors were added at 4 °C and the mixture was gently shaken until complete dissolution. It was then poured onto the cell pellet previously thawed on ice, and cold 200 µL Pepstatin solution in methanol (1 mg/mL) were added. The obtained suspension was homogenized by gently shaking and additional 200 µL Pepstatin solution in MeOH (1 mg/mL) were added. It was immediately sonicated on ice for 4 cycles (40 % amplitude, 2 s on, 2 s off, 4 cycles: each cycle lasting for 30 seconds and followed by a 30 seconds pause). Again, 200 µL Pepstatin solution in MeOH (1 mg/mL) were added and the mixture was centrifuged for 5 minutes, at 29000 rpm and 4 °C. After supplementing again with

200 µL Pepstatin solution in MeOH (1 mg/mL) and after adding DNase (1 mg), the obtained mixture was quickly loaded, without prior filtration, onto an AKTA™ Fast Protein Liquid Chromatography (FPLC) system and then subjected to Ni^{2+}-affinity chromatography.

Optimization of the protein purification:

A: The lysate was loaded (~ 35 mL, flow = 4 mL/min) onto a prepacked 5 mL HisTrap™ HP Ni^{2+}-NTA column pre-equilibrated with loading buffer and the desired protein was purified using a linear gradient, from 0 to 100 % EB in LB, in 20 column volumes (CV) (flow = 3 mL/min) monitoring by UV absorption (λ = 280 nm). Collected fractions were analysed by SDS-PAGE gel and it was concluded that the desired protein was eluted when the linear gradient reached ~ 35 % EB in LB.

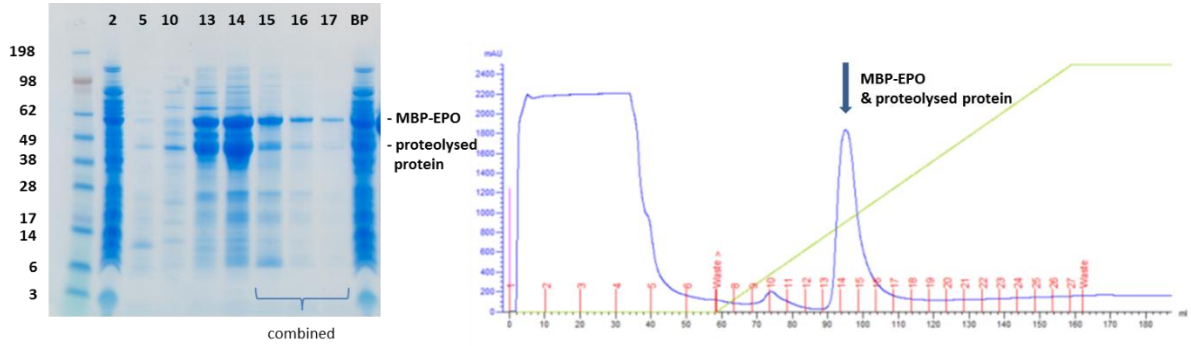


Figure SI 3.1. Reducing SDS-PAGE gel of collected fractions and chromatogram of the Ni^{2+}-NTA purification of MBP-EPO-1 using a linear gradient. The full-length construct MBP-EPO-1 and the proteolysed protein coeluted when the Ni^{2+}-affinity purification was performed using a linear gradient.

B: The lysate was loaded (~35 mL, flow = 4 mL/min) onto a prepacked 5 mL HisTrap™ HP Ni^{2+}-NTA column pre-equilibrated with loading buffer, and the desired protein was purified using a stepwise gradient (flow = 3 mL/min) monitoring by UV absorption (λ = 280 nm):

- 100 % LB1 for 7 CV;
- isocratic 20 % EB1 in LB1 for 6 CV;
- isocratic 35 % EB1 in LB1 for 6 CV;
- isocratic 100 % EB1 for 6 CV;

Collected fractions were analysed by SDS-PAGE gel and the ones containing mainly the desired protein were combined and dialysed in 50 mM TrisHCl, pH 8.0 (3 x 4 L, 1 h/1 h/12 h). Thus, it was concluded that B was a better method for obtaining the MBP-EPO-1 variant in moderate yields (15 mL protein solution, c = 0.63 mg/mL) and with acceptable purity. This was aliquoted, flash-frozen in liquid N_2 and stored at – 80 °C.

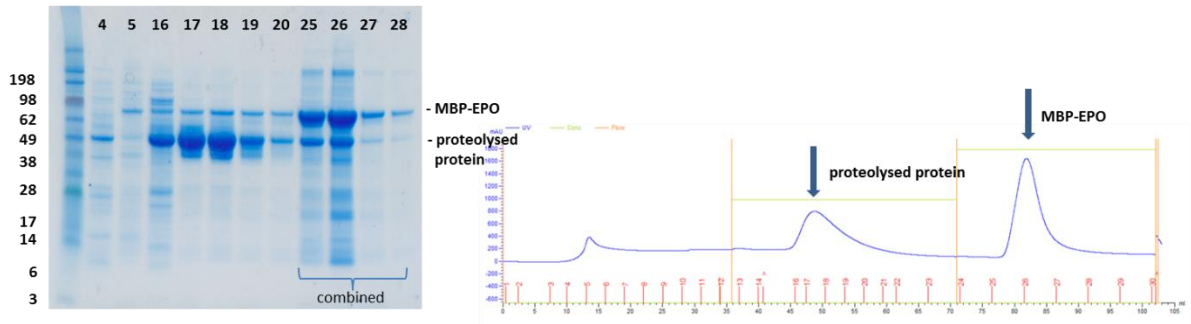


Figure SI 3.2. Reducing SDS-PAGE gel of collected fractions and chromatogram of the Ni^{2+}-NTA purification of MBP-EPO-1 using a stepwise gradient. Most of the proteolysed protein eluted earlier (20 %) than the full-length construct MBP-EPO-1 (35 %) when the Ni^{2+}-affinity purification was performed using a stepwise gradient.

C: The lysate was loaded (~35 mL, flow = 4 mL/min) onto a prepacked 5 mL HisTrap™ HP Ni^{2+}-NTA column pre-equilibrated with loading buffer, and the desired protein was purified using a stepwise gradient (flow = 3 mL/min) monitoring by UV absorption (λ = 280 nm):

- isocratic 100 % LB2 for 5 CV;
- isocratic 20 % EB2 in LB2 for 5 CV;
- isocratic 40 % EB2 in LB2 for 5 CV;
- isocratic 60 % EB2 in LB2 for 5 CV;
- isocratic 100 % EB2 in LB2 for 5 CV.

Loading buffer 2 (LB2): 50 mM NaPi, 150 mM NaCl, 0.5 mM Na-EDTA, pH 7.0

Elution buffer 2 (EB2): 50 mM NaPi, 150 mM NaCl, 0.5 mM Na-EDTA, 500 mM imidazole, pH 7.0

Collected fractions were analysed by SDS-PAGE gel and the ones containing mainly the desired protein were combined. Then, 3 mL protein solution were subjected to MBP-affinity purification. Thus, protein was loaded onto ~ 1 mL MBP-affinity resin pre-equilibrated with LB2 and manually purified as follows:

- isocratic 100 % LB3 for 5 CV;
- isocratic 100 % EB3 for 10 CV;

Collected fractions were analysed by SDS-PAGE gel and the ones containing mainly the desired protein were combined to obtain MBP-EPO variants in moderate yield (~ 6 mL MBP-EPO-2 protein solution, c = 0.7 mg/mL; ~ 4 mL MBP-EPO-3 protein solution, c = 0.8 mg/mL) and with acceptable purity. This was aliquoted, flash-frozen in liquid N_2 and stored at – 80 °C.

Loading buffer 3 (LB3): 50 mM Tris, 150 mM NaCl, 5% glycerol, pH 7.0

Elution buffer 3 (EB3): 50 mM Tris, 150 mM NaCl, 5% glycerol, 100 mM Maltose, pH 7.0

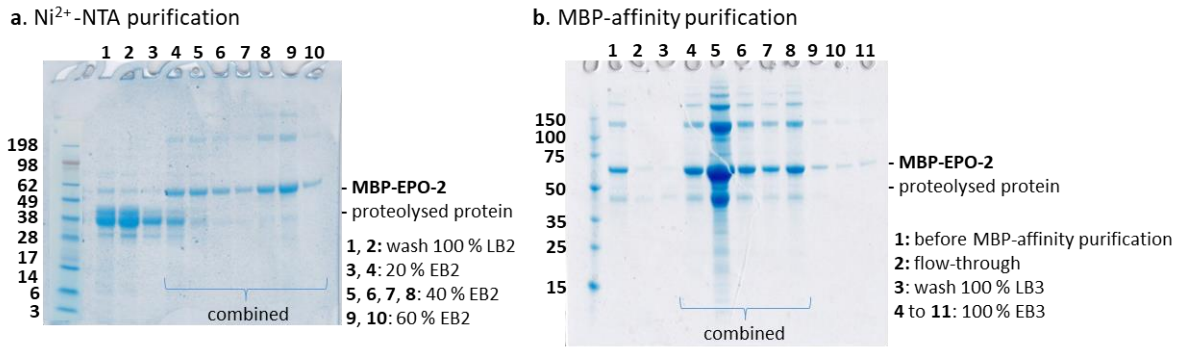


Figure SI 3.3. Non-reducing SDS-PAGE analyses of collected fractions for the Ni^{2+}-NTA and the MBP-affinity purification of MBP-EPO-2. Most of the proteolysed protein eluted earlier (20 %) than the full-length construct MBP-EPO-2 (40 % and 60 %) when the Ni^{2+}-affinity purification was performed using a stepwise gradient (**a**). The MBP-affinity purification allowed the fast removal of imidazole, and generated MBP-EPO-2 that had already started to aggregate/oligomerize but to a less degree than the original MBP-EPO-1 (Figure 3.6) (**b**).

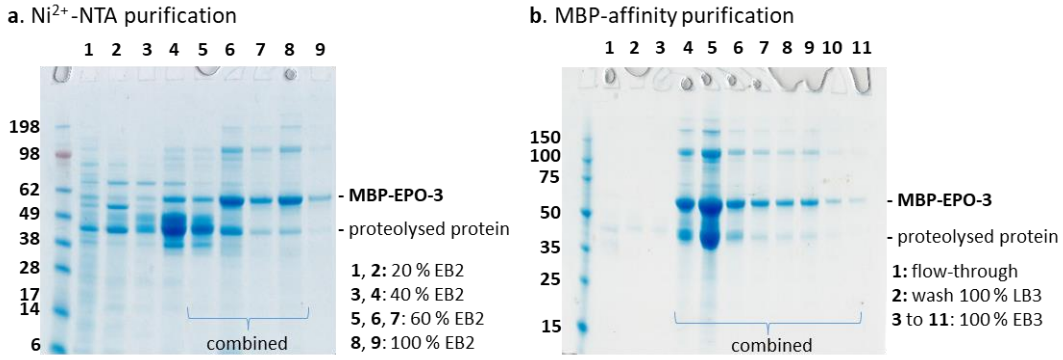


Figure SI 3.4. Non-reducing SDS-PAGE analyses of collected fractions for the Ni^{2+}-NTA and the MBP-affinity purification of MBP-EPO-3. Most of the proteolysed protein eluted earlier (20 % and 40 %) than the full-length construct MBP-EPO-3 (60 % and 100 %) when the Ni^{2+}-affinity purification was performed using a stepwise gradient (**a**). The MBP-affinity purification allowed the fast removal of imidazole and generated MBP-EPO-3 that had already started to aggregate/oligomerize but to a less degree than the original MBP-EPO-1 (Figure 3.6) (**b**).

3.4.2.3 TEV-cleavage of MBP-EPO-1 and attempted purification of EPO-1

Two equal aliquots of MBP-EPO-1 (7 mL, c = 0.23 mg/mL in 25 mM TrisHCl, pH 8.0) were treated in parallel with a TEV-protease solution (31 µL, c = 1.15 mg/mL in 25 mM TrisHCl, 10 % glycerol, pH 8.0), in a 1:45 protease-to-target protein ratio (w/w), and the mixtures were incubated for 16 hours at 4 °C (1) and 25 °C (2), respectively. After 3 hours, since slow progression was observed in the reaction performed at 25 °C, more enzyme, up to 1:2 protease-to-target protein ratio (w/w) was added to both reaction mixtures and incubation was continued for up to 16 hours. Then, since reducing SDS-PAGE gel analysis showed that the protein incubated at 4 °C had precipitated, only the reaction mixture incubated at 25 °C was subjected to Ni^{2+}-affinity chromatography. Thus, 6.5 mL reaction mixture obtained above was slowly loaded onto a 5 mL HisTrap™ Ni^{2+}-NTA column (flow = 1 mL/min) and eluted with loading buffer (25 mM TrisHCl, pH 8.0). Reducing SDS-PAGE analysis showed that also

the impurities were eluted with the flow-through and thus, the combined collected fractions (3 mL) were concentrated by centrifugation (3 times, 5 min each, 5000 g, 4 °C) in a 5 mL Vivaspin™ vial (30000 MWCO, PES membrane). EPO-1 was digested with trypsin, upon excision of the corresponding band from the polyacrylamide gel that was run to separate the reaction mixture.

3.4.3 Formation of Dha-MBP-EPO-1 and GlcNAc-S-MBP-EPO-1

3.4.3.1 Implementation of a DTT/TEV-cleavage method for the LC-MS detection of Dha-MBP-EPO-1 and GlcNAc-S-MBP-EPO-1

A stock solution of TEV-protease (c = 3.8 mg/mL) was diluted with 50 mM TrisHCl pH 7.0 buffer to have a final concentration of 0.038 mg/mL and to this, 1M DTT solution in ddH_2O was added to have a final concentration of 0.5 mM. Then, in a LC-MS vial, to 5 µL reaction mixture 45 µL of the prepared TEV-protease solution were added, followed by 0.5 µL 1M DTT solution in ddH_2O, to have a final DTT concentration of ~ 10 mM. After incubation at 25 °C for 5 minutes, the sample was analysed by denaturing, intact protein LC-MS.

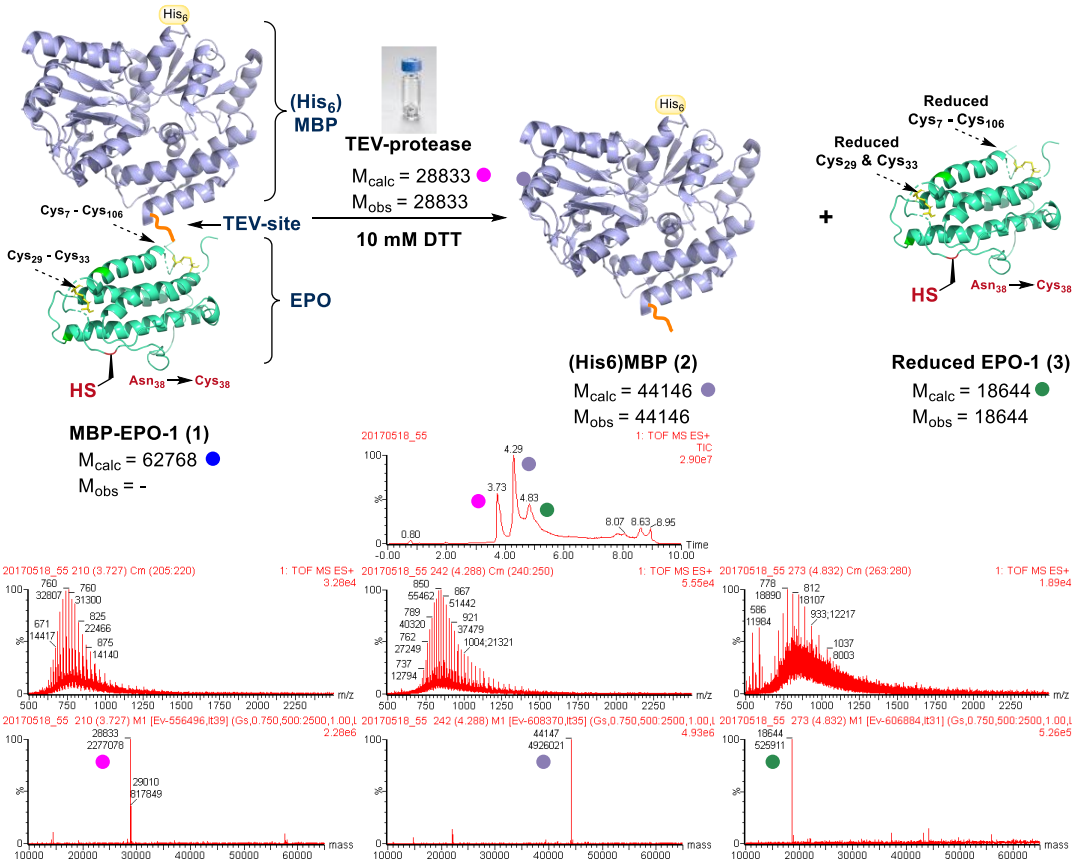


Figure SI 3.5. LC-MS analysis of MBP-EPO-1 performed using a DTT/TEV-cleavage method that was applied to a small LC-MS aliquot of the protein solution. An optimized LC gradient allowed separation of the three resulted proteins: TEV-protease (M_{obs} = 28833), (His6)MBP (M_{obs} = 44147) and the reduced EPO-1 (M_{obs} = 18644).

3.4.3.2 Formation of Dha-MBP-EPO-1 using 300 equivalents MDBP with (A) or without (B) pre-treatment with 0.5 mM DTT (DTT/TEV-cleavage LC-MS method)

A: To 140 µL of MBP-EPO-1 protein solution (c = 0.6 mg/mL) in 50 mM TrisHCl pH 8.0 buffer 50 equivalents DTT (0.7 µL of a 100 mM solution in ddH_2O to have a final concentration of 0.5 mM) were added and the reaction mixture was incubated on ice for 60 minutes. It was then filtered through a PD SpinMiniTrap G-25™ and treated with 300 equivalents MDBP (4.02 µL of a 100 mM emulsion obtained from diluting 10x a 1 M stock solution of MDBP in DMF with 50 mM TrisHCl pH 8.0 buffer). The reaction mixture was incubated at 25 °C for 5 hours monitoring the progression by intact protein LC-MS analysis, using the DTT/TEV-cleavage method. Then, the reaction mixture was filtered through a PD Spin MiniTrap G-25™ column and used for the thia-Michael addition step by incubating with GlcNAc-SH.

B: To 140 µL of MBP-EPO-1 protein solution (c = 0.6 mg/mL) in 50 mM TrisHCl pH 8.0 buffer 300 equivalents MDBP (4.02 µL of a 100 mM emulsion obtained from diluting 10x a 1 M stock solution of MDBP in DMF, with 50 mM TrisHCl pH 8.0 buffer) were added. The reaction mixture was incubated at 25 °C for 5 hours monitoring the progression by intact protein LC-MS analysis, using the DTT/TEV-cleavage method.

3.4.3.3 Formation of Dha-MBP-EPO-1 using 900 equivalents MDBP after pre-treatment with 0.5 mM DTT (DTT/TEV-cleavage LC-MS method)

To 140 µL of MBP-EPO-1 protein solution (c = 0.6 mg/mL) in 50 mM TrisHCl pH 8.0 buffer 50 equivalents DTT (0.7 µL of a 100 mM solution in ddH_2O to have a final concentration of 0.5 mM) were added and the reaction mixture was incubated on ice for 60 minutes. It was then filtered through a PD SpinMiniTrap G-25™ and treated with 900 equivalents MDBP (12.06 µL of a 100 mM emulsion obtained from diluting 10x a 1 M stock solution of MDBP in DMF with 50 mM TrisHCl pH 8.0 buffer). The reaction mixture was incubated at 25 °C overnight, monitoring the progression by intact protein LC-MS analysis, using the DTT/TEV-cleavage method.

3.4.3.4 Thia-Michael addition of GlcNAc-SH to Dha-MBP-EPO-1 (DTT/TEV-cleavage LC-MS method)

To the Dha-MBP-EPO-1 obtained above (section 3.4.3.2), 1.24 mg GlcNAc-SH were added (11000 equiv.) and the reaction mixture was incubated at 25 °C overnight. The progression of the reaction was evaluated by intact protein LC-MS analysis, using the DTT/TEV-cleavage method.

3.4.4 Formation of Dha-MBP-EPO-2

3.4.4.1 Attempted Dha-reaction in 50 mM TrisHCl, pH 8.0 buffer

To 140 µl of MBP-EPO-2 protein solution (c = 0.6 mg/ml in 50 mM TrisHCl pH 8.0 buffer) 300 equivalents MDBP (4.02 µl of a 100 mM emulsion obtained from diluting 10x a 1 M stock solution of MDBP in DMF with 50 mM TrisHCl pH 8.0 buffer) were added and the reaction mixture was incubated at 25 °C. After 30 minutes already, progression of the reaction could not be monitored by intact protein LC-MS analysis.

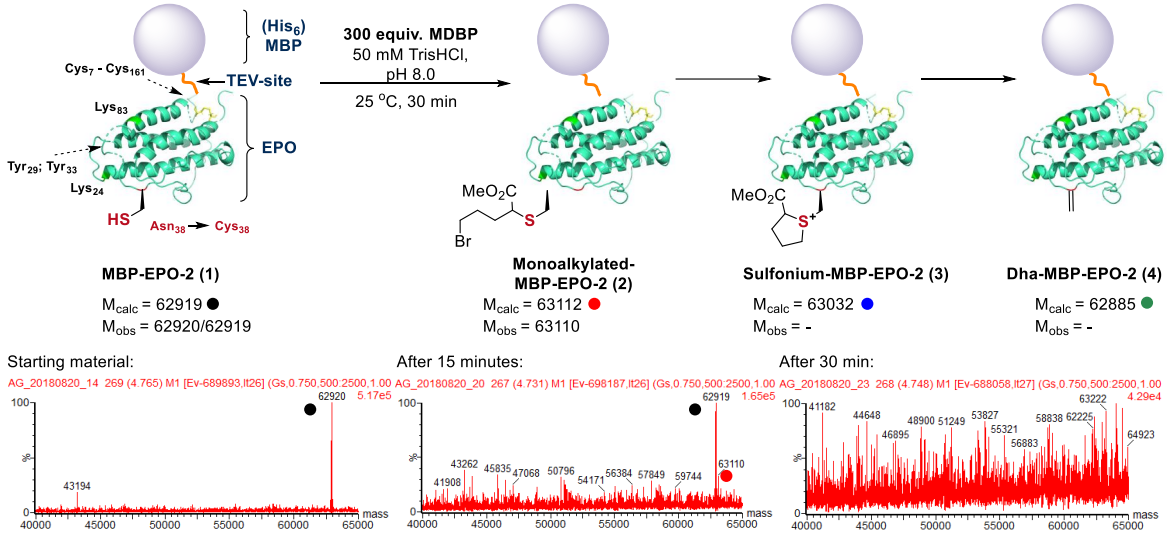


Figure SI 3.6. Formation of Dha-MBP-EPO-2 could still not be followed by intact protein LC-MS analysis (without DTT treatment) most likely because of the protein aggregation that had rapidly occurred in the TrisHCl buffer.

3.4.4.2 Screening of different bis-alkylating reagents (300 equiv.) in different buffers (A and B), on ice

Two aliquots of protein solution (160 µL each) were buffer-exchanged into two different buffers **A** and **B** using two PD Spin MiniTraps G-25 columns and then six reactions were set up in parallel as follows: to 50 µL protein solution (c = 0.6 mg/mL in buffer A: 50 mM TrisHCl, 150 mM NaCl, 100 mM maltose pH 8.0 or buffer B: 100 mM NaPi, 3 M GdnHCl pH 8.0) 0.286 µL reagent 1 (MDBP, 0.5 M solution in DMF), 2 (DBHDA, 0.5 M solution in DMSO), or 3 (DBDGla, 0.5 M solution in DMSO) were added and reaction mixtures were incubated on ice for 30 minutes. Reaction progression was monitored by LC-MS analyses after 30 minutes (5 µL injected after 50X dilution of a 1 µL aliquot with 0.1 % FA ddH_2O).

3.4.4.3 Screening of different bis-alkylating reagents (900 equiv.) in buffer B (100 mM NaPi, 3 M GdnHCl pH 8.0), on ice and at 25 °C

Two aliquots of protein solution (160 µL each) were buffer-exchanged into 100 mM NaPi, 3 M GdnHCl pH 8.0 buffer using two PD Spin MiniTraps G-25 columns and then six reactions were set up in parallel as follows: to 50 µL protein solution (c = 0.6 mg/mL) 0.858 µL reagent 1 (MDBP, 0.5 M solution in DMF), 2 (DBHDA, 0.5 M solution in DMSO), or 3 (DBDGla, 0.5 M solution in DMSO) were added and reaction mixtures were incubated on ice or at 25 °C for two hours. Reaction progression was monitored by LC-MS analyses after 30, 60, 90 and 120 minutes (5 µL injected after 50X dilution of a 1 µL aliquot with 0.1 % FA ddH2O).

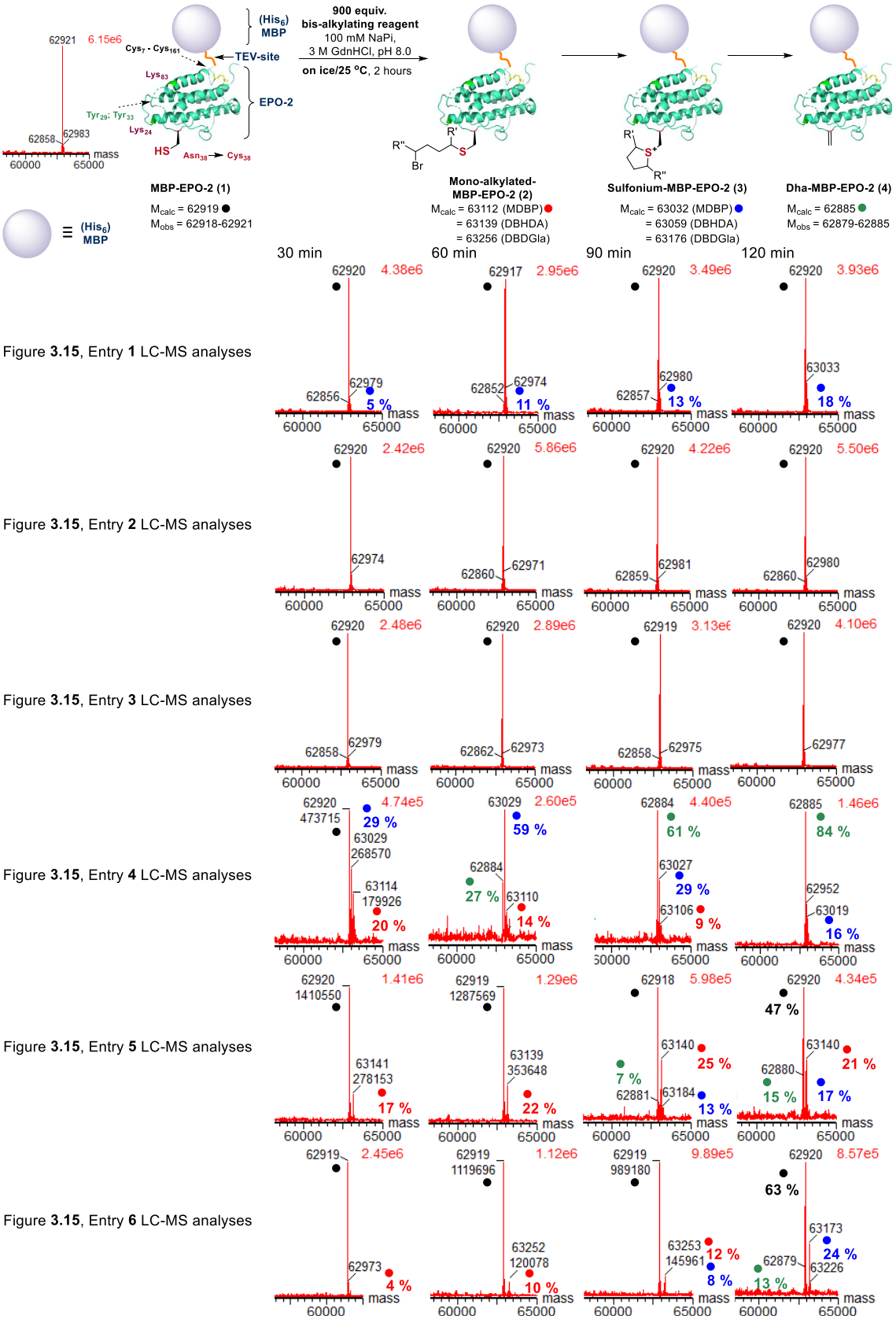


Figure **3.15**, Entry **1** LC-MS analyses

Figure **3.15**, Entry **2** LC-MS analyses

Figure **3.15**, Entry **3** LC-MS analyses

Figure **3.15**, Entry **4** LC-MS analyses

Figure **3.15**, Entry **5** LC-MS analyses

Figure **3.15**, Entry **6** LC-MS analyses

Figure SI 3.7. **Intact protein LC-MS analyses of the screening performed to generate the Dha-MBP-EPO-2 using 900 equivalents MDBP, DBHDA and DBDGla, on ice or at 25 °C.** Excellent conversion to the desired product was achieved (84 %) when MDBP was incubated with the starting material at 25 °C for two hours (Figure 3.15, entry 4).

3.4.4.4 Screening of different equivalents (900/1800) of three bis-alkylating reagents (MDBP, DBHDA, DBDGla) in buffer B (100 mM NaPi, 3 M GdnHCl pH 8.0), at 37 °C

Two aliquots of protein solution (160 µL each) were buffer-exchanged into 100 mM NaPi, 3 M GdnHCl pH 8.0 buffer using two PD Spin MiniTraps G-25 columns and then six reactions were set up in parallel as follows: to 50 µL protein solution (c = 0.6 mg/mL) 0.858 µL or 1.716 µL reagent 1 (MDBP, 0.5 M solution in DMF), 2 (DBHDA, 0.5 M solution in DMSO), or 3 (DBDGla, 0.5 M solution in DMSO) were added and reaction mixtures were incubated at 37 °C for one hour. Reaction progression was monitored by LC-MS analyses after 30 and 60 minutes (5 µL injected after 50X dilution of a 1 µL aliquot with 0.1 % FA ddH_2O).

3.4.4.5 Screening of different equivalents (300/600) of the bis-alkylating reagent MDBP in buffer B (100 mM NaPi, 3 M GdnHCl pH 8.0), at 37 °C

160 µL protein solution were buffer-exchanged into 100 mM NaPi, 3 M GdnHCl pH 8.0 buffer using a PD Spin MiniTraps G-25 column and then two reactions were set up in parallel as follows: to 50 µL protein solution (c = 0.6 mg/mL) 300 or 600 equiv. MDBP reagent (0.5 M solution in DMF) were added and the reaction mixtures were incubated at 37 °C for 60 minutes. Reaction progression was monitored by LC-MS analyses after 5, 10, 15, 20, 30, 45 and 60 minutes (5 µL injected after 50X dilution of a 1 µL aliquot with 0.1 % FA ddH_2O).

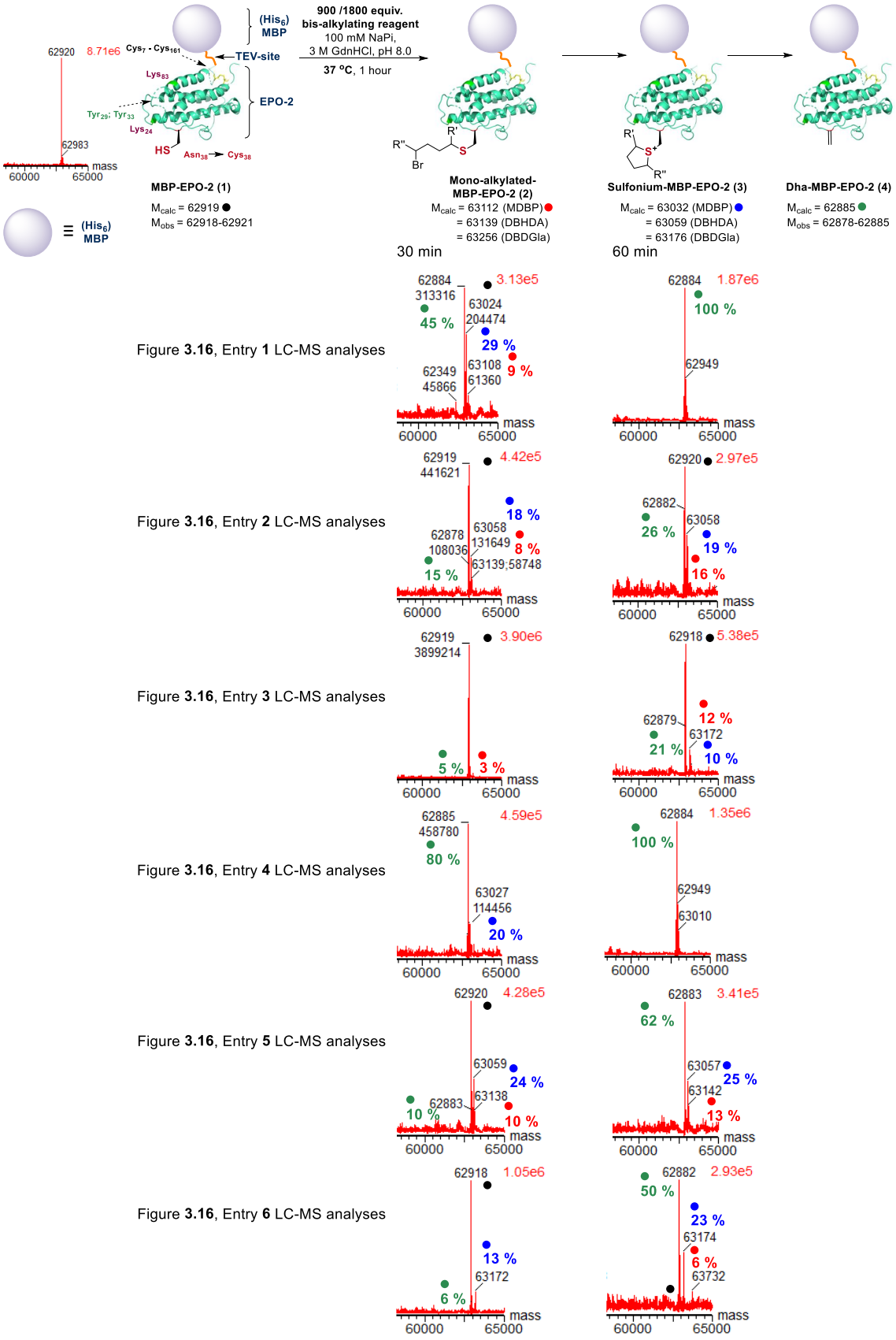


Figure SI 3.8. Intact protein LC-MS analyses of the screening performed to generate the Dha-MBP-EPO-2 using 900/1800 equivalents MDBP, DBHDA and DBDGla at 37 °C. Excellent conversion to the desired product was achieved (100 %) when the starting material was incubated with either 900 or 1800 equiv. MDBP but with such high excess of reagents the quality of the spectra appeared slightly negatively affected.

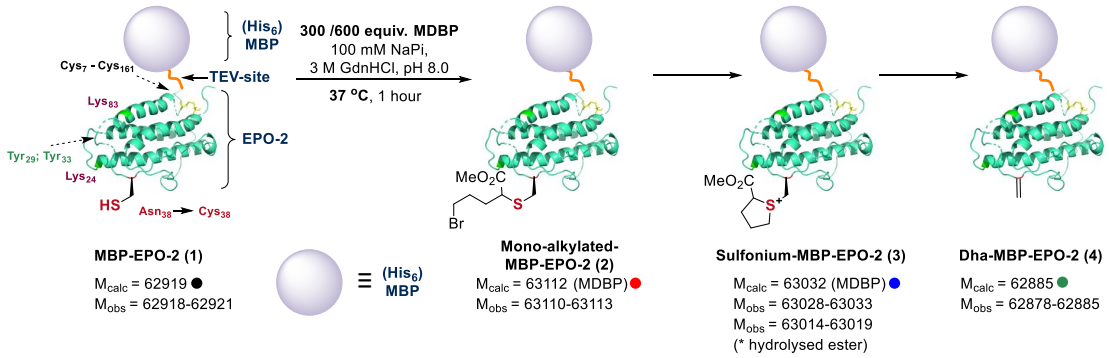


Figure **3.17**, 300 equiv. MDBP LC-MS analyses

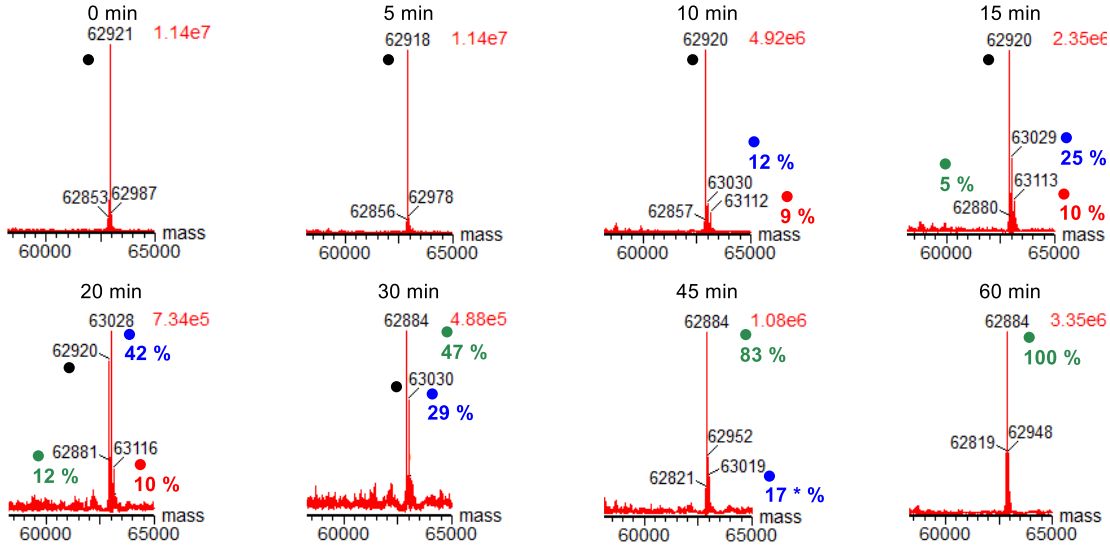


Figure **3.17**, 600 equiv. MDBP LC-MS analyses

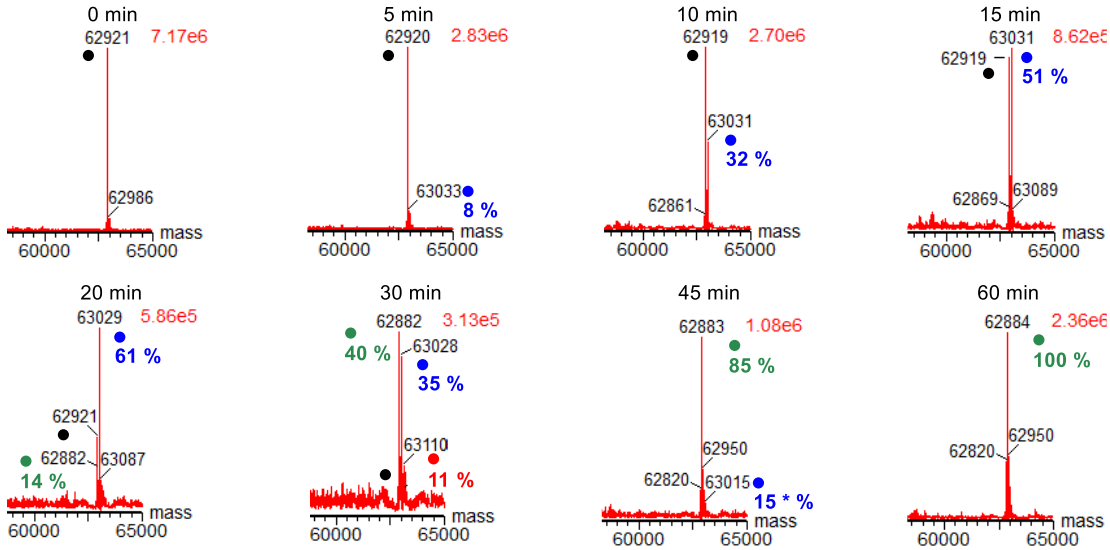


Figure SI 3.9. **Intact protein LC-MS analyses of the screening performed to generate the Dha-MBP-EPO-2 using 300/600 equivalents MDBP at 37 °C.** Excellent conversion to the desired product was achieved (100 %) when the starting material was incubated with either 300 or 600 equiv. Nevertheless, with a higher concentration of bis-alkylating reagent the MBP-EPO-2 was more rapidly converted to the sulfonium-MBP-EPO-2 (3): after 20 minutes of incubation with 300 and 600 equivalents MDBP 3 was generated with a conversion of 42 % and 61 %, respectively. The molecular masses 62820 and 62950 observed in the obtained Dha-MBP-EPO-2 are impurities that were also present in the starting material (molecular masses 62853 and 62987), and which had also undergone the Cys-Dha transformation.

3.4.4.6 Screening of different organic solvents for their ability to affect the stability of the MBP-EPO-2 protein

Eight reaction mixtures were set up in parallel, as follows and incubated at 37 °C for 60 minutes. To 25 µL protein solution (c = 0.6 mg/mL in 100 mM NaPi, 3M GdnHCl pH 8.0 buffer) 0.15 µL solvent (DMF, DMSO, MeCN, MeOH, iPrOH, THF, TFE, ddH_2O) were added and protein stability was monitored by LC-MS analyses after 60 minutes (5 µL injected after 50X dilution of a 1 µL aliquot with 0.1 % FA ddH_2O).

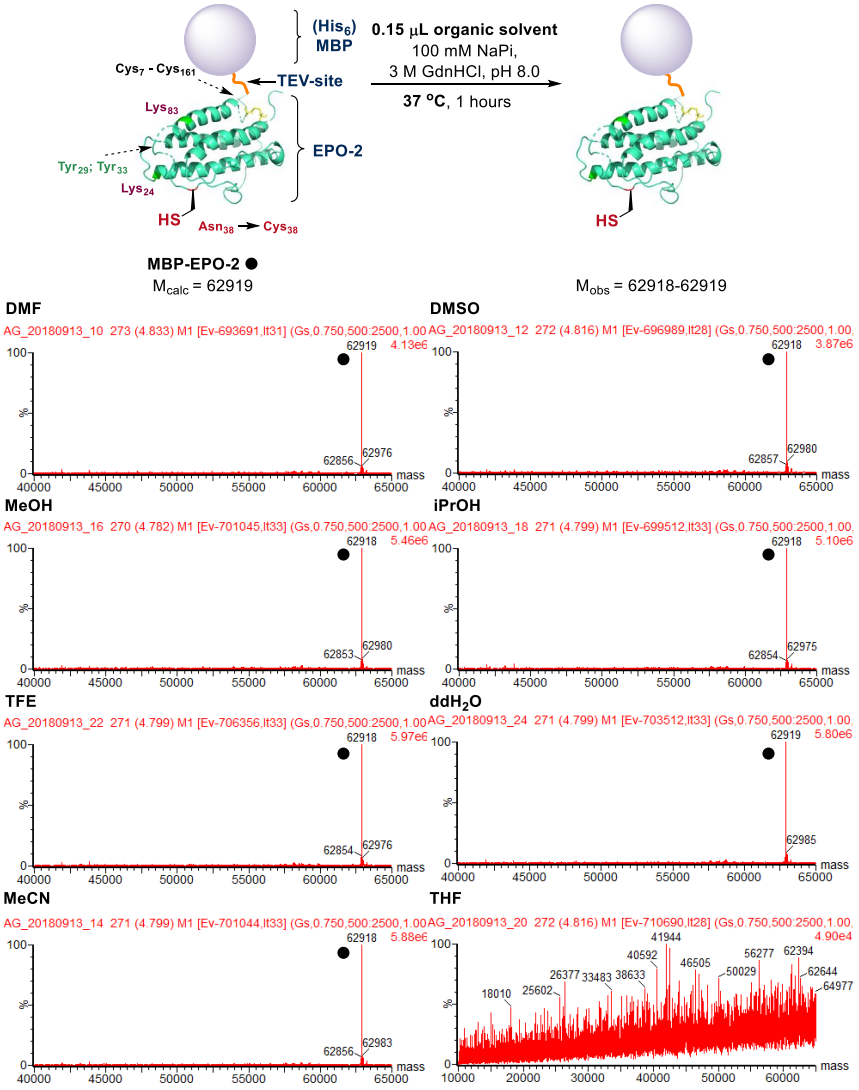


Figure SI 3.10. **Intact protein LC-MS analyses of the screening performed to assess the stability of MBP-EPO-2 when incubated with various organic solvents at 37 °C for 1 hour.** Excellent protein stability was observed when MBP-EPO-2 was incubated with DMF, DMSO, MeOH, iPrOH, TFE, ddH_2O or MeCN but the same outcome was not achieved when THF was added. In this experiment, the added amount of organic solvent was identical to the one used to dilute 300 equiv. bis-alkylating reagent for its addition to the solution of starting material.

3.4.4.7 Quantification of free Cys residues in the MBP-EPO-2 protein

To 50 µL protein solution (c = 0.6 mg/mL in 100 mM NaPi, 3M GdnHCl pH 8.0 buffer), 0.286 µL Ellman's reagent (0.5 M solution in DMF) were added and the reaction mixture was incubated 37 °C for 60

minutes. Reaction progression was monitored by LC-MS analysis (5 µL injected after 50X dilution of a 1 µL aliquot with 0.1 % FA ddH_2O).

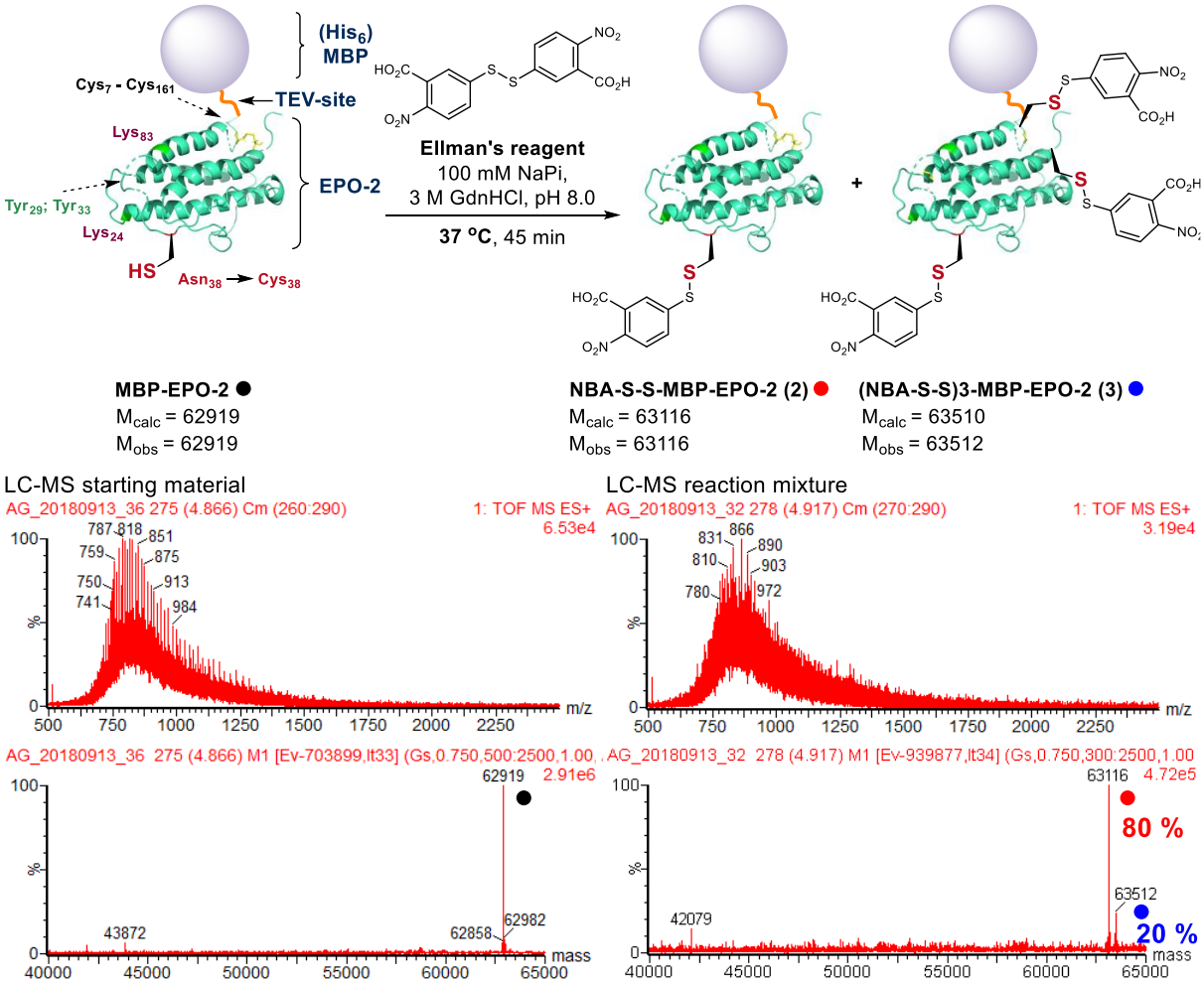


Figure SI 3.11. Intact protein LC-MS analyses of the reaction performed to relatively quantify the free Cys residues of MBP-EPO-2. Excellent conversion of MBP-EPO-2 (1) to the desired nitrobenzoic acid (NBA) disulfide single addition product (2) was achieved (80 %) when the starting material was incubated with the 5,5'-dithio-bis-(2-nitrobenzoic acid, Ellman's reagent) at 37 °C for 45 min. A much lower conversion (20 %) to the NBA disulfide triple addition product (3) could also be observed in the analysed sample.

3.4.4.8 Identification of the pre-requisite needed for monitoring formation of the Dha-MBP-EPO-2 (3 M GdnHCl buffer)

Four aliquots of protein solution (160 µL each) were buffer-exchanged into 100 mM NaPi, 3 M GdnHCl pH 8.0 buffer using four PD Spin MiniTraps G-25 columns and then four reactions were set up in parallel as follows: to 150 µL protein solution (c = 0.6 mg/mL in buffer A: 100 mM NaPi, 3 M GdnHCl pH 8.0 / B: 100 mM NaPi, pH 8.0 / C: 10 mM NaPi, pH 8.0 / D: 10 mM NaPi, 100 mM maltose, pH 8.0) 2.574 µL MDBP (0.5 M solution in DMF) were added and reaction mixtures were incubated at 25 °C for one hour. Reaction progression was monitored by LC-MS analyses after 60 minutes (5 µL injected after 50X dilution of a 1 µL aliquot with 0.1 % FA ddH2O).

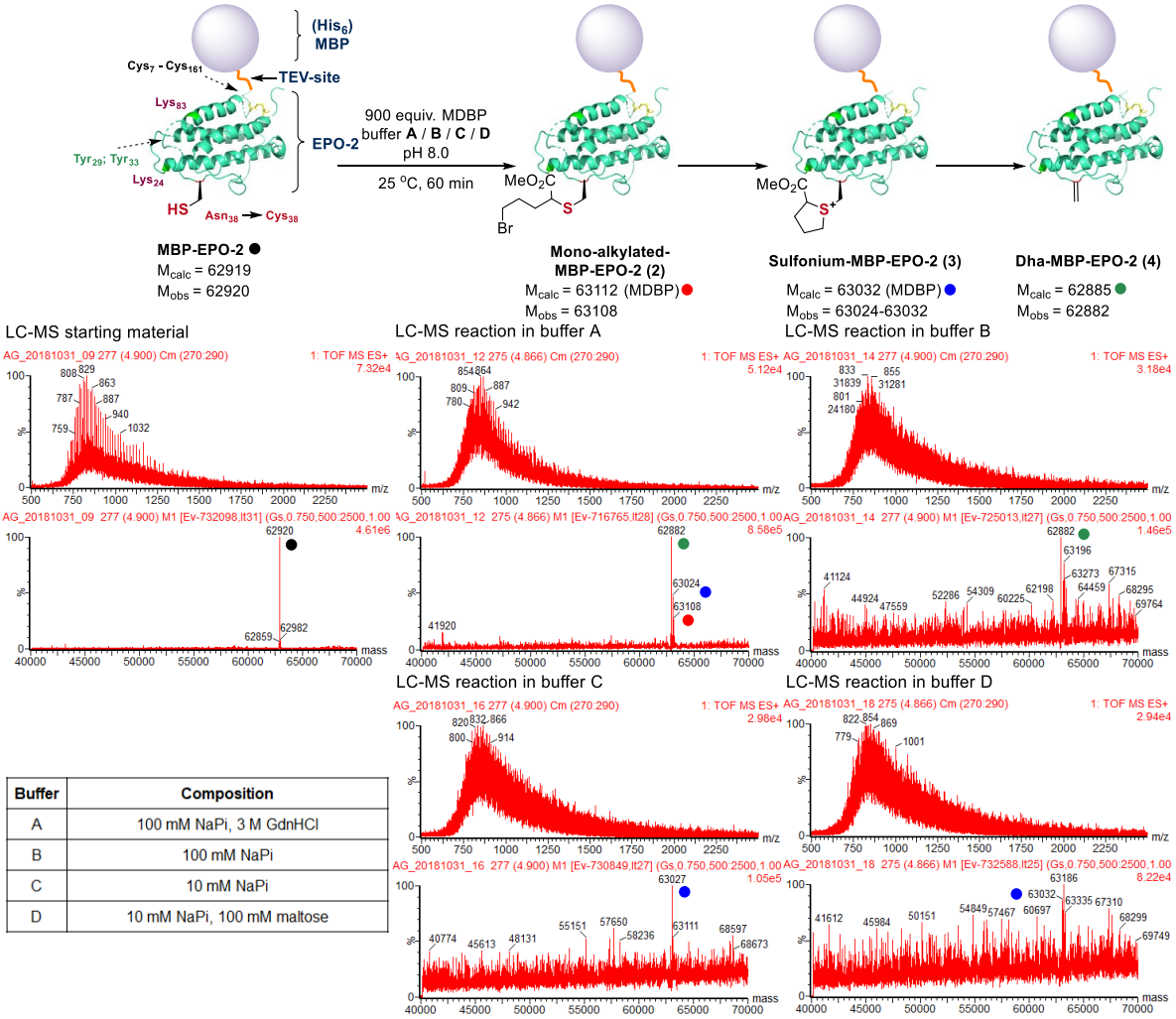


Figure SI 3.12. **Intact protein LC-MS analyses of the reactions performed in different buffers revealed that GdnHCl was crucial for the formation of Dha-MBP-EPO-2.** Excellent conversion of MBP-EPO-2 (1) to the desired Dha-derivative (4) was achieved in buffer A but since the reaction was performed with 900 equiv. MDBP at 25 °C, low amounts of residual mono-alkylated intermediate (2) and sulfonium-MBP-EPO-2 (3) could also be observed. The reactions performed in buffer B, C or D did not afford good quality MS spectra and thus, the conversions could not be calculated.

3.4.4.9 Thia-Michael addition of β-mercaptoethanol to Dha-MBP-EPO-2

To 180 µL protein solution (c = 0.6 mg/mL in 100 mM NaPi, 3M GdnHCl pH 8.0 buffer) 300 equiv. MDBP reagent (0.5 M solution in DMF, 1.03 µL) were added and the reaction mixture was incubated at 37 °C for 45 minutes. Then, 2 µL β-mercaptoethanol were added and the reaction mixture was further incubated at room temperature for 60 minutes. Progression of the two consecutive reactions was monitored by LC-MS analyses after 45 and 60 minutes, respectively (3 µL injected after 100X dilution of a 1 µL aliquot with 0.1 % FA ddH2O). A 20 µL sample was digested and analysed by LC-MS/MS analysis as well.

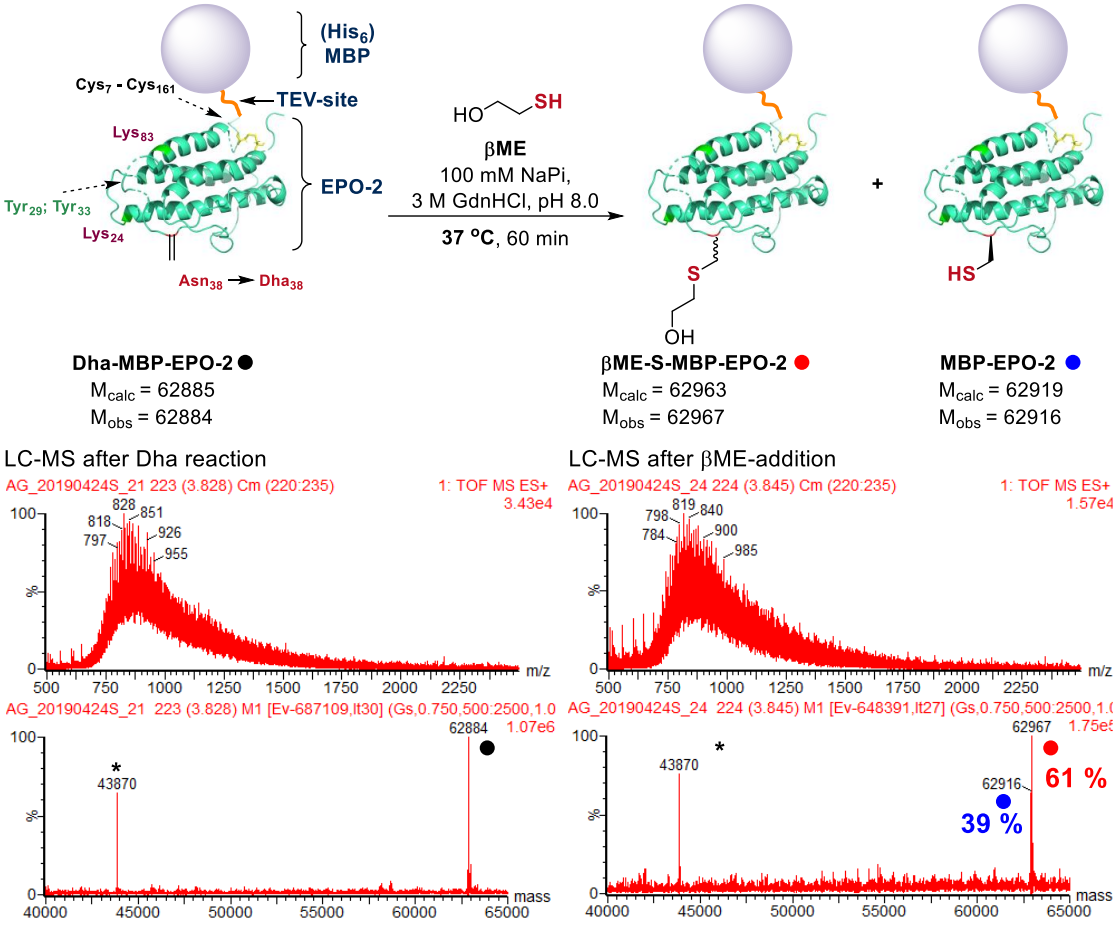


Figure SI 3.13. Intact protein LC-MS analyses of the reaction performed to generate βME-MBP-EPO-2. Moderate conversion of Dha-MBP-EPO-2 to the desired βME-derivative was achieved (61 %) when the starting material was incubated with βME at 37 °C, for one hour. 39 % of residual MBP-EPO-2 could also be observed in the reaction mixture, and this had presumably originated from the potential dimers that might have been cleaved by the addition of βME (* = proteolysed impurity).

3.4.5 Formation of Dha-MBP-EPO-3 and GlcNAc-S-MBP-EPO-3

3.4.5.1 Formation of Dha-MBP-EPO-3 using 300 equivalents MDBP and subsequent thia-Michael addition of GlcNAc-SH or β-mercaptoethanol

To 100 µL protein solution (c = 0.6 mg/mL in 100 mM NaPi, 3M GdnHCl pH 8.0 buffer) 300 equiv. MDBP reagent (0.5 M solution in DMF, 0.572 µL) were added and the reaction mixture was incubated at 37 °C for 60 minutes. Then, 75 µL reaction mixture were transferred into an Eppendorf containing 2.2 mg GlcNAc-SH and it was further incubated at 37 °C for 40 minutes. Progression of the two consecutive reactions was monitored by LC-MS analyses after 60 and 40 minutes, respectively (3 µL injected after 100X dilution of a 1 µL aliquot with 0.1 % FA ddH_2O or with 100 mM NH_4OAc buffer, pH 8.0 + 10 mM DTT).

To 180 µL protein solution (c = 0.6 mg/mL in 100 mM NaPi, 3M GdnHCl pH 8.0 buffer) 300 equiv. MDBP reagent (0.5 M solution in DMF, 1.03 µl) were added and the reaction mixture was incubated at 37 °C for 45 minutes. Then, 2 µL β-mercaptoethanol were added and the reaction mixture was further incubated at room temperature for 60 minutes. Progression of the two consecutive reactions was monitored by LC-MS analyses after 45 and 60 minutes, respectively (3 µL injected after 100X dilution of a 1 µL aliquot with 0.1 % FA ddH_2O) and despite not being able to identify the expected products by intact protein LC-MS analyses, a 20 µL sample was digested and analysed by LC-MS/MS analysis as well.

3.4.6 LC-MS/MS analysis of enzymatically digested GlcNAc-S-MBP-EPO-1, βME-S-MBP-EPO-2 and βME-S-MBP-EPO-3

GluC and trypsin combined digest of modified MBP-EPO proteins:

10 µg of modified MBP-EPO were denatured in 8 M urea in 100 mM TEAB for 30 min at room temperature (r.t.) and subsequently reduced with 10 mM TCEP for 30 minutes, at r.t. before being alkylated with 50 mM 2-iodoacetamide for 30 min, at r.t., in the dark. Samples were diluted to 0.8 M urea with 50 mM TEAB and digestion was performed with GluC (4 µg/10 µg sample) over night at 37 °C. Samples were further digested with trypsin (1 µg/40 µg sample) at 37 °C, for 4 hours. Digested peptides were enriched and purified using a C18 column (Oasis® HLB 10mg) and freeze-dried.

LC-MS/MS data acquisition for GlcNAc-S-MBP-EPO-1:

Dried peptides were resuspended in a 5 % DMSO, 5 % formic acid aqueous solution. Then, they were analysed on an EASY-nLC 1000 UHPLC system (Proxeon) connected to an Orbitrap Elite mass spectrometer (Thermo Fischer Scientific) possessing an EASY-Spray nano-electrospray ion source (Thermo Fischer Scientific). The peptides were trapped on an in-house packed guard column (75 µm i.d. x 20 mm, reprosil C18, 3µm, 120 Å) using solvent A (0.1% Formic Acid, 5 % DMSO in water) at a pressure of 500 bar. The peptides were separated on an EASY-spray Acclaim PepMap® analytical column (75 µm i.d. × 500 mm, RSLC C18, 2 µm, 100 Å) using a linear gradient (length: 58 minutes, 8 % to 30 % solvent B (0.1% formic acid, 5 % DMSO in acetonitrile), flow rate: 200 nL/min). The separated peptides were electrosprayed directly into the mass spectrometer operating in a data-dependent mode using an HCD/ETD based method. Full scan MS spectra (scan range 350-1500 m/z, resolution 30000, AGC target $1e^6$, maximum injection time 250 ms) and subsequent HCD MS/MS spectra

(resolution 15000, AGC target $3e^4$, maximum injection time 500 ms) of 5 most intense peaks were acquired in the Orbitrap. HCD fragmentation was performed at 30 % of normalized collision energy and the signal intensity threshold was kept at 500 counts. If peaks at any of m/z 138.0550, 168.0655 or 204.0867 were detected among top 20 most intense peak of a HCD spectrum, an ETD MS/MS spectrum of the precursor was triggered. ETD spectra (AGC cation target $5e^4$, AGC anion target $2e^5$, cation maximum injection time 100 ms) were acquired in the Iontrap.

Processing of the data for GlcNAc-S-MBP-EPO-1:

The raw MS file was searched against the respective protein sequence. Trypsin with [D|P] with a maximum number of 3 missed cleavages was selected as the protease, allowing for diverging cleavage at one end. Carbamidomethylation (Cysteine), Oxidation (Methionine), Carbamylation (*N*-terminus), Deamidation (Asparagine, Glutamine) and the custom thio-GlcNAcylation (+203.079) were set as variable modifications. Precursor mass tolerance was set as 50 ppm. Fragment mass tolerances for HCD and ETD were set to 0.2 Da. All spectra were manually validated.

LC-MS/MS data acquisition for βME-S-MBP-EPO-2 and βME-S-MBP-EPO-3:

Dried peptides were redissolved in 0.1 % formic acid, 2% acetonitrile (100 µL) and analysed on a NanoAcquity-UPLC system (Waters) coupled to an Orbitrap Elite mass spectrometer (Thermo Fischer Scientific) possessing an EASY-Spray nano-electrospray ion source (Thermo Fischer Scientific). Peptides (2 µL injection in 0.05% TFA) were separated on an EASY-spray Acclaim PepMap analytical column (75 µm i.d. × 15 mm, RSLC C18, 3 µm, 100 Å) using a linear gradient (length: 60 min; 0–3 min 1 % B, 3–33 min 35 % B, 33–38 min 99 % B, 38–41 min 99 % B, 41–44 min 1 % B, 44–47 min 99 % B, 47–50 min 99% B, 50–60 min 1% B; solvent A (0.1% formic acid), B (0.1% formic acid in acetonitrile), flow rate: 300 nL/min). Data-dependent MS acquisition full-scan MS spectra were collected (scan range 350–1,500 m/z, resolution 120,000, AGC target $1e^6$, maximum injection time 100 ms). After the MS scans, the ten most intense peaks were selected for HCD fragmentation at 35 % of normalized collision energy. HCD spectra were acquired in the Iontrap (AGC target $1e^3$, maximum injection time 50 ms).

Processing of the data for βME-S-MBP-EPO-2 and βME-S-MBP-EPO-3:
The raw MS file was searched against the respective protein sequence. Trypsin with [D|P] with a maximum number of 3 missed cleavages was selected as the protease, allowing for diverging cleavage at one end. Carbamidomethylation (Cysteine), Oxidation (Methionine), Carbamylation (*N*-terminus),

Deamidation (Asparagine, Glutamine) and the custom thio-βME (+44.03) were set as variable modifications. Precursor mass tolerance was set as 15 ppm and fragment mass tolerance for HCD was set to 0.8 Da. All spectra were manually validated.

3.4.7 Molecular Dynamics (MD) of EPO-1

The structure of EPO (PDB code 1BUY) was downloaded, and the following residues mutated: K24N, K38C, K83N. All simulations were carried out with the GROMACS 4.6 package[80] and the AMBER99SB*-ILDN force field.[81-83] The structure was added to a triclinic box with periodic boundary conditions with the edge of the box being 12 Å from the nearest protein atom. The box was populated with SPC/E waters.[84] The system was neutralized by adding Cl^- and Na^+ ions to a final concentration of 150 mM. Energy minimization using the steepest descent method was followed by equilibration with positional restraints on heavy atoms for 100 ps. Trajectories were collected at 300 K and 1 bar with standard NPT ensemble settings. The temperature was kept at 300 K with velocity rescaling (interval 0.1 ps)[85] and the pressure was kept at 1 bar using the Parrinello-Rahman coupling scheme (interval 2 ps).[86] Electrostatic interactions were calculated at every step with the particle-mesh Ewald method[87] and short-range repulsive and attractive dispersion interactions were simultaneously described by a Lennard-Jones potential, which was cut off at 1.0 nm. The SETTLE algorithm was used to constrain bonds and angles of water molecules.[88] LINCS was used for all other bonds.[89,90] Additionally, virtual interaction sites (VSITE) replaced the fastest angular degrees of freedom involving hydrogen atoms, allowing a time step of 4 fs. Visualisation of the results was performed with PyMol[70] and VMD[91].

3.5 References

1. Bachmann S, Le Hir M, Eckardt KU. Co-localization of erythropoietin mRNA and ecto-5'-nucleotidase immunoreactivity in peritubular cells of rat renal cortex indicates that fibroblasts produce erythropoietin. *J Histochem Cytochem*. 1993;41(3):335-341.

2. Lacombe C, Da Silva JL, Bruneval P, et al. Peritubular cells are the site of erythropoietin synthesis in the murine hypoxic kidney. *J Clin Invest*. 1988;81(2):620-623.

3. Zeisberg M, Kalluri R. Physiology of the renal interstitium. *Clin J Am Soc Nephrol*. 2015;10(10):1831-1840.

4. Fisher JW, Koury S, Ducey T, Mendel S. Erythropoietin production by interstitial cells of hypoxic monkey kidneys. *Br J Haematol*. 1996;95(1):27-32.

5. Shanks JH, Hill CM, Lappin TRJ, Maxwell AP. Localization of erythropoietin gene expression in proximal renal tubular cells detected by digoxigenin-labelled oligonucleotide probes. *J Pathol*. 1996;179(3):283-287.

6. Pan X, Suzuki N, Hirano I, Yamazaki S, Minegishi N, Yamamoto M. Isolation and characterization of renal erythropoietin-producing cells from genetically produced anemia mice. *PLoS One*. 2011;6(10):e25839.

7. Göpfert T, Gess B, Eckardt KU, Kurtz A. Hypoxia signalling in the control of erythropoietin gene expression in rat hepatocytes. *J Cell Physiol*. 1996;168(2):354-361.

8. Jelkmann W. Regulation of erythropoietin production. *J Physiol*. 2011;589(6):1251-1258.

9. Maxwell PH, Ferguson DJP, Nicholls LG, et al. Sites of erythropoietin production. *Kidney Int*. 1997;51(2):393-401.

10. Semenza GL, Nejfelt MK, Chi SM, Antonarakis SE. Hypoxia-inducible nuclear factors bind to an enhancer element located 3' to the human erythropoietin gene. *Proc Natl Acad Sci U S A*. 1991;88(13):5680-5684.

11. Berra E, Benizri E, Ginouvès A, Volmat V, Roux D, Pouysségur J. HIF prolyl-hydroxylase 2 is the key oxygen sensor setting low steady-state levels of HIF-1α in normoxia. *EMBO J*. 2003;22(16):4082-4090.

12. Schofield CJ, Ratcliffe PJ. Oxygen sensing by HIF hydroxylases. *Nat Rev Mol Cell Biol*. 2004;5(5):343-354.

13. Haase VH. Regulation of erythropoiesis by hypoxia-inducible factors. *Blood Rev*. 2013;27(1):41-53.

14. Semenza GL, Wang GL. A nuclear factor induced by hypoxia *via de novo* protein synthesis binds to the human erythropoietin gene enhancer at a site required for transcriptional activation. *Mol Cell Biol*. 1992;12(12):5447-5454.

15. Peyssonnaux C, Nizet V, Johnson RS. Role of the hypoxia inducible factors HIF in iron metabolism. *Cell Cycle*. 2008;7(1):28-32.

16. Rolfs A, Kvietikova I, Gassmann M, Wenger RH. Oxygen-regulated transferrin expression is mediated by hypoxia-inducible factor-1. *J Biol Chem*. 1997;272(32):20055-20062.

17. Tacchini L, Bianchi L, Bernelli-Zazzera A, Cairo G. Transferrin receptor induction by hypoxia. HIF-1-mediated transcriptional activation and cell-specific post-transcriptional regulation. *J Biol Chem*. 1999;274(34):24142-24146.

18. Lok CN, Ponka P. Identification of a hypoxia response element in the transferrin receptor gene. *J Biol Chem*. 1999;274(34):24147-24152.

19. Ebert BL, Firth JD, Ratcliffe PJ. Hypoxia and mitochondrial inhibitors regulate expression of glucose transporter-1 *via* distinct cis-acting sequences. *J Biol Chem*. 1995;270(49):29083-29089.

20. Semenza GL, Roth PH, Fang HM, Wang GL. Transcriptional regulation of genes encoding glycolytic enzymes by hypoxia-inducible factor 1. *J Biol Chem*. 1994;269(38):23757-23763.

21. Andrew P. Levy, Nina S. Levy, Scott Wegner, Mark A. Goldberg. Transcriptional regulation of the rat vascular endothelial growth factor gene by hypoxia. *J Biol Chem*. 1995;270(22):13333-13340.

22. Liu Y, Cox SR, Morita T, Kourembanas S. Hypoxia regulates vascular endothelial growth factor gene expression in endothelial cells. Identification of a 5' enhancer. *Circ Res*. 1995;77(3):638-643.

23. Forsythe JA, Jiang BH, Iyer NV, et al. Activation of vascular endothelial growth factor gene transcription by hypoxia-inducible factor 1. *Mol Cell Biol*. 1996;16(9):4604-4613.

24. Grover A, Mancini E, Moore S, et al. Erythropoietin guides multipotent hematopoietic progenitor cells toward an erythroid fate. *J Exp Med*. 2014;211(2):181-188.

25. Crystallographic evidence for preformed dimers of erythropoietin receptor before ligand activation. *Science*. 1999;283(5404):987-990.

26. Moraga I, Wernig G, Wilmes S, et al. Tuning cytokine receptor signaling by re-orienting dimer geometry with surrogate ligands. *Cell*. 2015;160(6):1196-1208.

27. Witthuhn BA, Quelle FW, Silvennoinen O, et al. JAK2 associates with the erythropoietin receptor and is tyrosine phosphorylated and activated following stimulation with erythropoietin. *Cell*. 1993;74(2):227-236.

28. Shoemaker C, Shimizu T, Jacobs K, et al. Isolation and characterization of genomic and cDNA clones of human erythropoietin. *Nature*. 1985;313(6005):806-810.

29. Lin FK, Suggs S, Lin CH, et al. Cloning and expression of the human erythropoietin gene. *PNAS*. 1985;82(22):7580-7584.

30. Lai PH, Everett R, Wang FF, Arakawa T, Goldwasser E. Structural characterization of human erythropoietin. *J Biol Chem*. 1986;261(7):3116-3121.

31. Boissel JP, Lee WR, Presnell SR, Cohen FE, Bunn HF. Erythropoietin structure-function relationships. Mutant proteins that test a model of tertiary structure. *J Biol Chem*. 1993;268(21):15983-15993.

32. Wang FF, Kung CK, Goldwasser E. Some chemical properties of human erythropoietin*. *Endocrinology*. 1985;116(6):2286-2292.

33. Wasley LC, Timony G, Murtha P, et al. The importance of *N*- and *O*-linked oligosaccharides for the biosynthesis and *in vitro* and *in vivo* biologic activities of erythropoietin. *Blood*. 1991;77(12):2624-2632.

34. Yuen C, Storring PL, Tiplady RJ, et al. Relationships between the *N*-glycan structures and biological activities of recombinant human erythropoietins produced using different culture conditions and purification procedures. *Br J Haematol*. 2003;121(3):511-526.

35. Takeuchi M, Inoue N, Strickland TW, et al. Relationship between sugar chain structure and biological activity of recombinant human erythropoietin produced in chinese hamster ovary cells. *Proc Natl Acad Sci U S A*. 1989;86(20):7819-7822.

36. Misaizu T, Matsuki S, Strickland TW, Takeuchi M, Kobata A, Takasaki S. Role of antennary structure of *N*-linked sugar chains in renal handling of recombinant human erythropoietin. *Blood*. 1995;86(11):4097-4104.

37. Bones J, McLoughlin N, Hilliard M, Wynne K, Karger BL, Rudd PM. 2D-LC analysis of BRP 3 erythropoietin *N*-glycosylation using anion exchange fractionation and hydrophilic interaction UPLC reveals long poly-*N*-acetyl lactosamine extensions. *Anal Chem*. 2011;83(11):4154-4162.

38. Fukuda MN, Sasaki H, Lopez L, Fukuda M. Survival of recombinant erythropoietin in the circulation: the role of carbohydrates. *Blood*. 1989;73(1):84-89.

39. Deicher R, Hörl WH. Differentiating factors between erythropoiesis-stimulating agents: a guide to selection for anaemia of chronic kidney disease. *Drugs*. 2004;64(5):499-509.

40. Lalonde M, Durocher Y. Therapeutic glycoprotein production in mammalian cells. *J Biotechnol*. 2017;251:128-140.

41. Goh JB, Ng SK. Impact of host cell line choice on glycan profile. *Crit Rev Biotechnol*. 2018;38(6):851-867.

42. Skibeli V, Nissen-Lie G, Torjesen P. Sugar profiling proves that human serum erythropoietin differs from recombinant human erythropoietin. *Blood*. 2001;98(13):3626-3634.

43. Cowper B, Li X, Yu L, Zhou Y, Fan WH, Rao CM. Comprehensive glycan analysis of twelve recombinant human erythropoietin preparations from manufacturers in China and Japan. *J Pharm Biomed*. 2018;153:214-220.

44. Hossler P, Khattak SF, Li ZJ. Optimal and consistent protein glycosylation in mammalian cell culture. *Glycobiology*. 2009;19(9):936-949.

45. Bohlius J, Wilson J, Seidenfeld J, et al. Recombinant human erythropoietins and cancer patients: updated meta-analysis of 57 studies including 9353 patients. *J Natl Cancer Inst*. 2006;98(10):708-714.

46. Phrommintikul A, Haas SJ, Elsik M, Krum H. Mortality and target haemoglobin concentrations in anaemic patients with chronic kidney disease treated with erythropoietin: a meta-analysis. *Lancet*. 2007;369(9559):381-388.

47. Bennett CL, Silver SM, Djulbegovic B, et al. Venous thromboembolism and mortality associated with recombinant erythropoietin and darbepoetin administration for the treatment of cancer-associated anemia. *JAMA*. 2008;299(8):914-924.

48. Nachbauer W, Boesch S, Reindl M, et al. Skeletal muscle involvement in Friedreich ataxia and potential effects of recombinant human erythropoietin administration on muscle regeneration and neovascularization. *J Neuropathol Exp Neurol*. 2012;71(8):708-715.

49. Calvillo L, Latini R, Kajstura J, et al. Recombinant human erythropoietin protects the myocardium from ischemia-reperfusion injury and promotes beneficial remodeling. *Proc Natl Acad Sci U S A*. 2003;100(8):4802-4806.

50. Westenbrink BD, Lipsic E, van der Meer P, et al. Erythropoietin improves cardiac function through endothelial progenitor cell and vascular endothelial growth factor mediated neovascularization. *Eur Heart J*. 2007;28(16):2018-2027.

51. Teng R, Gavrilova O, Suzuki N, et al. Disrupted erythropoietin signalling promotes obesity and alters hypothalamus proopiomelanocortin production. *Nat Commun*. 2011;2(1):520.

52. Bernardes GJL, Chalker JM, Errey JC, Davis BG. Facile conversion of cysteine and alkyl cysteines to dehydroalanine on protein surfaces: versatile and switchable access to functionalized proteins. *J Am Chem Soc*. 2008;130(15):5052-5053.

53. Raj R, Lercher L, Mohammed S, Davis BG. Synthetic nucleosomes reveal that GlcNAcylation modulates direct interaction with the FACT complex. *Angew Chem Int Ed*. 2016;55(31):8918-8922.

54. Lercher L, Raj R, Patel NA, et al. Generation of a synthetic GlcNAcylated nucleosome reveals regulation of stability by H2A-Thr101 GlcNAcylation. *Nat Commun*. 2015;6(1):7978.

55. Macmillan D, Bill RM, Sage KA, Fern D, Flitsch SL. Selective *in vitro* glycosylation of recombinant proteins: semi-synthesis of novel homogeneous glycoforms of human erythropoietin. *Chem Biol*. 2001;8(2):133-145.

56. Gamblin DP, Garnier P, Kasteren Sv, Oldham NJ, Fairbanks AJ, Davis BG. Glyco-SeS: selenenylsulfide-mediated protein glycoconjugation - a new strategy in post-translational modification. *Angew Chem Int Ed*. 2004;43(7):828-833.

57. Dadová J, Galan SR, Davis BG. Synthesis of modified proteins *via* functionalization of dehydroalanine. *Curr Opin Chem Biol*. 2018;46:71-81.

58. Galan SRG, Wickens JR, Dadova J, et al. Post-translational site-selective protein backbone α-deuteration. *Nat Chem Biol*. 2018;14(10):955-963.

59. Xiao H, Wu R. Global and site-specific analysis revealing unexpected and extensive protein *S*-GlcNAcylation in human cells. *Anal Chem*. 2017;89(6):3656-3663.

60. Maynard JC, Burlingame AL, Medzihradszky KF. Cysteine *S*-linked *N*-acetylglucosamine (*S*-GlcNAcylation), a new post-translational modification in mammals. *Mol Cell Proteomics*. 2016;15(11):3405-3411.

61. De Leon CA, Levine PM, Craven TW, Pratt MR. The sulfur-linked analogue of *O*-GlcNAc (*S*-GlcNAc) is an enzymatically stable and reasonable structural surrogate for *O*-GlcNAc at the peptide and protein levels. *Biochemistry*. 2017;56(27):3507-3517.

62. Floyd N, Vijayakrishnan B, Koeppe JR, Davis BG. Thiyl glycosylation of olefinic proteins: *S*-linked glycoconjugate synthesis. *Angew Chem Int Ed*. 2009;48(42):7798-7802.

63. Bhushan B, Lin YA, Bak M, et al. Genetic incorporation of olefin cross-metathesis reaction tags for protein modification. *J Am Chem Soc*. 2018;140(44):14599-14603.

64. Jeong TH, Son YJ, Ryu HB, et al. Soluble expression and partial purification of recombinant human erythropoietin from *E. coli*. *Protein Expr Purif*. 2014;95:211-218.

65. Salazar O, Asenjo JA. Enzymatic lysis of microbial cells. *Biotechnol Lett*. 2007;29(7):985-994.

66. Lobstein J, Emrich CA, Jeans C, Faulkner M, Riggs P, Berkmen M. SHuffle, a novel *Escherichia coli* protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microb Cell Factories*. 2012;11(1):56.

67. Walker MJ. *The proteomics protocols handbook.* Humana Press; 2005.

68. Zhang J, Xin L, Shan B, et al. PEAKS DB: *de novo* sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics*. 2012;11(4):M111.010587.

69. Cheetham JC, Smith DM, Aoki KH, et al. NMR structure of human erythropoietin and a comparison with its receptor bound conformation. *Nat Struct Biol*. 1998;5(10):861-866.

70. The PyMOL molecular graphics system, version 1.8 Schrödinger, LLC.

71. McGaughey GB, Gagné M, Rappé AK. π-stacking interactions alive and well in proteins. *J Biol Chem*. 1998;273(25):15458-15463.

72. Narhi LO, Arakawa T, Aoki K, et al. Asn to Lys mutations at three sites which are *N*-glycosylated in the mammalian protein decrease the aggregation of *Escherichia coli*-derived erythropoietin. *Protein Eng Des Sel*. 2001;14(2):135-140.

73. Grimsley GR, Scholtz JM, Pace CN. A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Sci*. 2009;18(1):247-251.

74. Galan S. *Site-selective chemical phosphorylation and activation of kinases.* University of Oxford; 2016.

75. Shilton BH, Shuman HA, Mowbray SL. Crystal structures and solution conformations of a dominant-negative mutant of *Escherichia coli* maltose-binding protein. *J Mol Biol*. 1996;264(2):364-376.

76. Meuzelaar H, Panman MR, Woutersen S. Guanidinium-induced denaturation by breaking of salt bridges. *Angew Chem Int Ed*. 2015;127(50):15470-15474.

77. Saba J, Dutta S, Hemenway E, Viner R. Increasing the productivity of glycopeptides analysis by using higher-energy collision dissociation-accurate mass-product-dependent electron transfer dissociation. *Int J Proteomics*. 2012;2012:560391.

78. Nominé Y, Ristriani T, Laurent C, Lefèvre J, Weiss É, Travé G. Formation of soluble inclusion bodies by HPV E6 oncoprotein fused to maltose-binding protein. *Protein Expr Purif*. 2001;23(1):22-32.

79. Raran-Kurussi S, Waugh DS. The ability to enhance the solubility of its fusion partners is an intrinsic property of maltose-binding protein but their folding is either spontaneous or chaperone-mediated. *PloS One*. 2012;7(11):e49589.

80. Pronk S, Páll S, Schulz R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*. 2013;29(7):845-854.

81. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple AMBER force fields and development of improved protein backbone parameters. *Proteins*. 2006;65(3):712-725.

82. Best RB, Hummer G. Optimized molecular dynamics force fields applied to the Helix−Coil transition of polypeptides. *J Phys Chem B*. 2009;113(26):9004-9015.

83. Lindorff-Larsen K, Piana S, Palmo K, et al. Improved side-chain torsion potentials for the amber ff99SB protein force field. *Proteins*. 2010;78(8):1950-1958.

84. Berendsen HJC, Grigera JR, Straatsma TP. The missing term in effective pair potentials. *J Phys Chem*. 1987;91(24):6269-6271.

85. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. *J Chem Phys*. 2007;126(1):014101.

86. Parrinello M, Rahman A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J Appl Phys*. 1981;52(12):7182-7190.

87. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh ewald method. *J Chem Phys*. 1995;103(19):8577-8593.

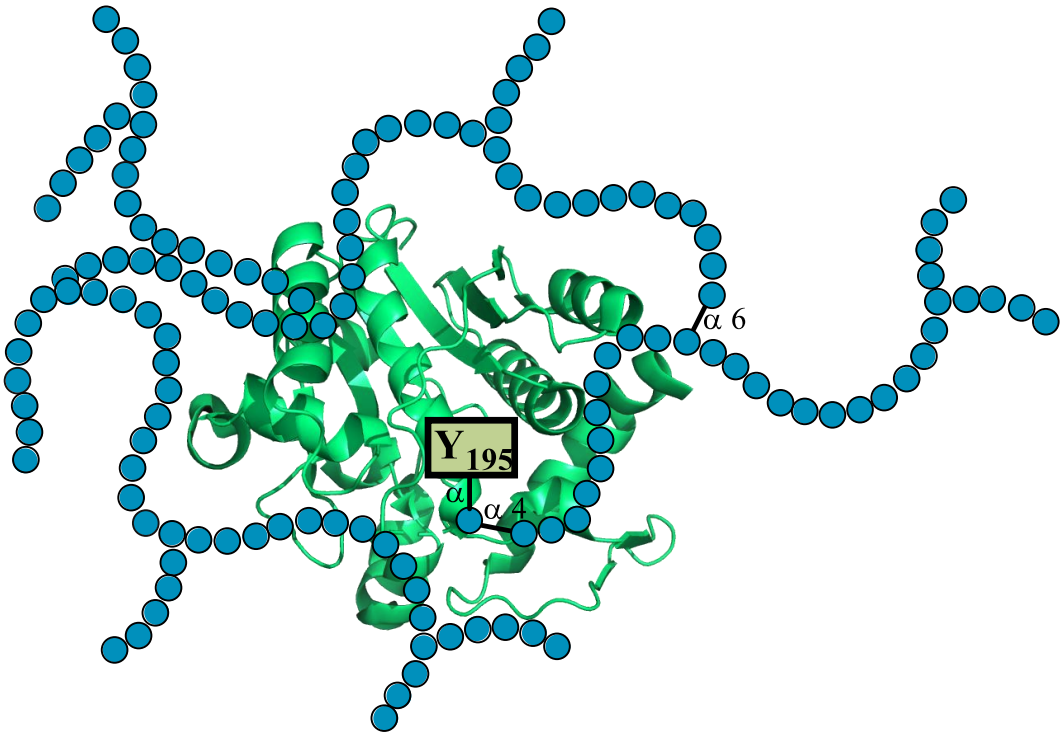
88. Miyamoto S, Kollman PA. Settle: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem*. 1992;13:952-962.

89. Hess B. P-LINCS: a parallel linear constraint solver for molecular simulation. *J Chem Theory Comput*. 2008;4(1):116-122.

90. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput*. 2008;4(3):435-447.

91. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *J Mol Graph*. 1996;14(1):33-38.

Chapter 4



Insights into the early stages of the glycogen biogenesis

Contents

4.1 Introduction 238

4.1.1 Glycogen: a readily accessible reserve for energy-depleting conditions..... 238

4.2 Results and discussion..... 242

4.2.1 Enzyme-compatible Suzuki-Miyaura cross-coupling (SMC) enabled mechanistic insights into the biocatalytic activity of GYG 242

4.2.2 ISCAMS analyses revealed dissociation of the autoglucosylated GYG-wt-nGlc 246

4.2.3 Intact protein LC-MS analyses revealed hydrolysis of the autoglucosylated GYG-wt-nGlc and the GYS-extended GYG-wt-(m+n)Glc polyglucosan chains in absence of UDP-Glucose 250

4.3 Conclusions and future perspectives..... 256

4.4 Experimental section 259

4.4.1 Plasmid DNA details and protein sequences 259

4.4.2 Protein expression and characterisation 260

4.4.2.1 GYG-wt-0Glc..... 260

4.4.2.2 GYG-Y195F 262

4.4.3 Autoglucosylation and glucosylation assays..... 265

4.4.3.1 Autoglucosylation of GYG-wt-0Glc in the presence of GYG-Y195F 265

4.4.3.2 Autoglucosylation of GYG-wt-0Glc: LC-MS and iSCAMS analyses of GYG-wt-0Glc and GYG-wt-nGlc (n = 9-11 avg. Glc residues) 267

4.4.3.3 Glucosylation of GYG-wt-nGlc (n = 9-11 avg. Glc residues) with 10 mM UDP-Glucose 268

4.4.4 Hydrolysis assays 273

4.4.4.1 Hydrolytic activity observed in the autoglucosylated GYG-wt-nGlc (n = 9-11 avg. Glc residues) 273

4.4.4.2 Hydrolytic activity observed in the GYS-extended GYG-wt-(n+m)Glc (n+m = 19-21 avg. Glc residues) 276

4.5 Supplementary text and figures 281

4.5.1 Additional figures 281

4.5.2 Note on the use of GYS as a complex with GYG-wt 285

4.6 References 287

4.1 Introduction

4.1.1 Glycogen: a readily accessible reserve for energy-depleting conditions

Glycogen is the essential glucose storage molecule in a wide variety of organisms, and its formation (glycogenesis) involves a trio of enzymes that catalyse in a concerted manner, the different biosynthetic reactions (Figure **4.1**, full pink square).[1]

The initial stage of the *de novo* glycogen biogenesis is under the control of glycogenin (GYG),[2,3] a protein that in humans, exists in two isoforms: GYG1 is predominantly found in skeletal muscle and heart,[4] whereas GYG2 is expressed mainly in liver.[5,6] Due to its ability to autoglucosylate its tyrosine (Tyr) 195 residue, glycogenin (herein called GYG-wt-0Glc, 1)[i] is a peculiar glycosyltransferase (GT) that is able to convert itself into a *O*-linked maltosaccharide-containing glycoprotein.[7] As a member of the GT-8 family of retaining GTs, this dimeric, non-processive enzyme exhibits the classic GT-A fold[8] and contains the DXD motif that binds the Mn^{2+} cofactor required for catalysis.[9-11]

Upon priming with the first glucose (Glc) residue, GYG further catalyses its linear Glc-extension.[12] The formation of α(1,4) glycosidic linkages through a stepwise, front-face, 'S_Ni-like' reaction mechanism generates GYG-glycoforms (GYG-wt-nGlc, 2) containing no more than 20 Glc residues and around 9-11 average Glc.[13,14] The unmatched self-glucosylating capacity of this protein, for which its extending glycan chain plays an important catalytic role, makes GYG unique among all the biosynthetic enzymes. GYG is not a catalyst in the strict sense of the term, since while being simultaneously an enzyme and a substrate, it is transformed with each Glc-addition step. Consequently, at different lengths of the nascent Tyr-linked maltosaccharide chain, GYG adopts distinct modes of autoglucosylation.[11,12,14]

GYG-wt-nGlc is a substrate for the key regulatory GT in glycogen biogenesis, the glycogen synthase (GYS).[15] This enzyme further catalyses the α(1,4)-glucosylation of the polyglucosan GYG-chain and similarly to GYG, it is encoded by two genes: the GYS1 expressed mainly in the skeletal and muscle heart, and the liver-specific GYS2.[16,17] Mammalian GYS,[i] a member of the GT-3 family of retaining GTs, is inhibited by protein phosphorylation and is activated upon binding to glucose 6-phosphate (G6P).[18-20] Crystallography studies revealed that GYS may possess a tetrameric structure with each monomer presenting a GT-B fold and, unlike GYG, GYS does not require the metal cofactor for catalysis.[18,21]

[i] GYG and GYS are referring to GYG1 and GYS1 respectively, unless otherwise mentioned.

The last, but not least important enzyme involved in the biosynthesis of glycogen is the glycogen branching enzyme (GBE).[22] By virtue of its unique capability of transferring an α(1,4)-linked polyglucosan chain containing from 8 to 14 sugar units from the outer, non-reducing end of a GYS-extended maltosaccharide-GYG into the α(1,6) position of the same or of a nearby chain, GBE generates the branched architecture of the GYG-bound glycogen (3).

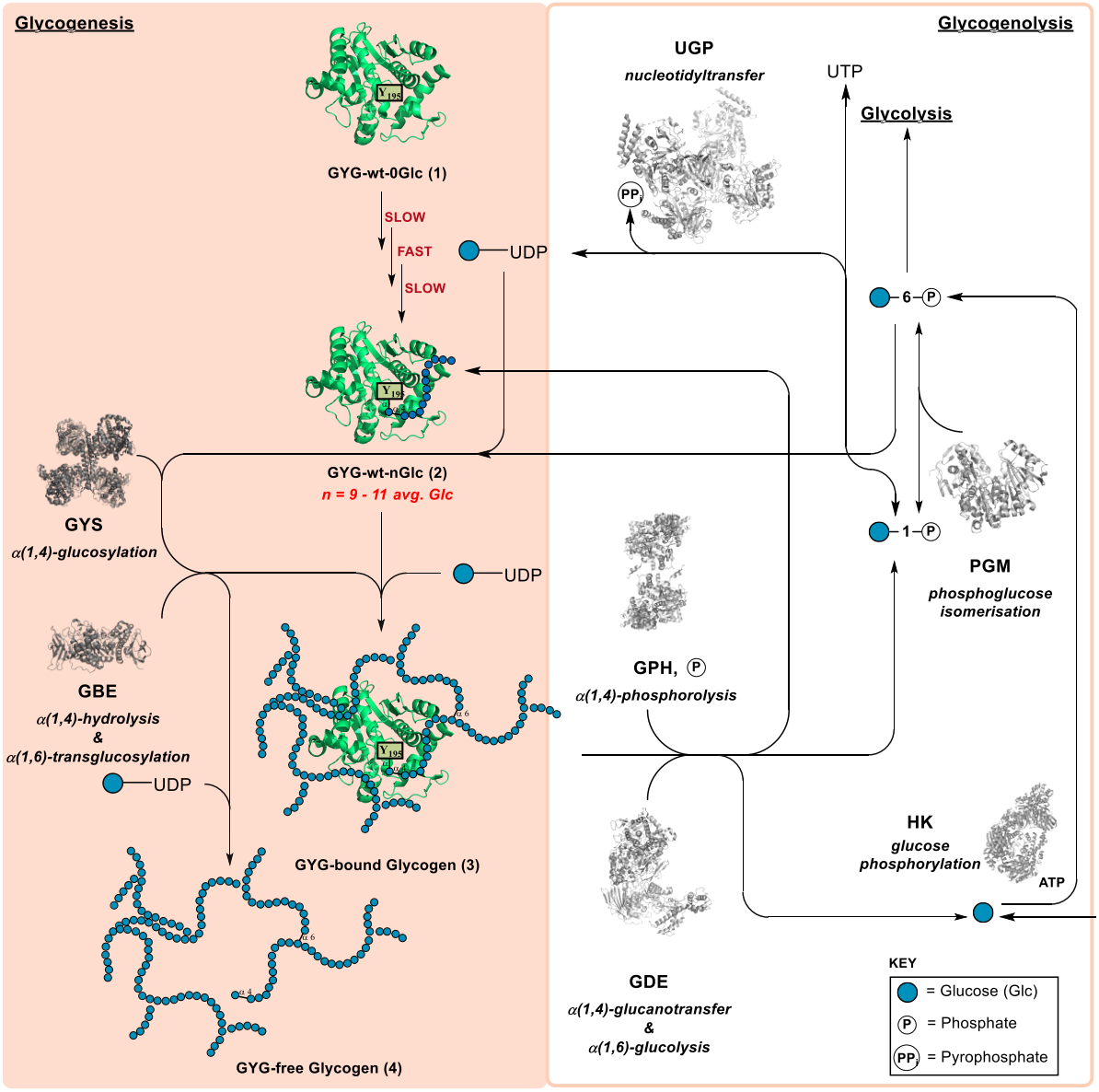


Figure 4.1 The glycogenesis and glycogenolysis processes in mammalians and the associated glycogen metabolic enzymes. In times of low energy demand GYG, GYS and GBE catalyse the glycogenesis process, whereas in times of energy depletion GPH and GDE catalyse the glycogenolytic pathway. In addition, HK, PGM and UGP are the three enzymes that are involved in maintaining Glc, G6P and UDP-Glc at their homeostatic levels, based on the energetic requirements of the cell (GYG, glycogenin; GYS, glycogen synthase; GBE, glycogen branching enzyme; GPH, glycogen phosphorylase; GDE, glycogen debranching enzyme; HK, hexokinase; PGM, phosphoglucomutase; UGP, UDP-Glc pyrophosphorylase; UDP, uridine diphosphate; UTP, uridine triphosphate; ATP, adenosine triphosphate).

The single isoform GBE belongs to the family of glycosyl hydrolases (GH) 13, also known as the α-amylase family)[23] and has a monomeric functional structure.[22] This is composed of four regions: a short *N*-terminal helical segment, a carbohydrate-binding module 48 (CBM48), a canonical (β/α) barrel core containing the presumed Asp357-Glu412-Asp481 catalytic triad, and the *C*-terminal amylase-like barrel domain.

The coordinated activity of GYS and GBE is responsible for the spherical, compact nature of a nascent glycogen granule. This not only preserves the osmotic neutrality of the cell but also displays many accessible, non-reducing termini that in times of increased energy demand can be immediately hydrolysed by the glycogenolytic enzymes.

The currently established theory argues that in times of nutritional deprivation, the glycogen breakdown process (glycogenolysis) is accomplished in the cytosol (Figure **4.1**, empty pink square) by the coordinated action of two catabolic enzymes: the glycogen phosphorylase (GPH) and the glycogen debranching enzyme (GDE).[24]

A second glycogenolytic pathway exists, the lysosomal degradation carried out by the acid α-glucosidase, but this is a secondary process which affects less than 10 % of the total glycogen existing in a mammalian's body.[25,26] In addition, since many aspects regarding this pathway remain elusive to this day,[27] the topic is beyond the scope of this thesis and thus it will not be addressed in detail herein.

GPH is a dimeric allosteric enzyme that releases glucose 1-phosphate (G1P) by cleaving the α(1,4)-linkages of the numerous glycosidic termini of the particle, up to the fourth Glc residue from a branching point. GPH has two independent catalytic sites and it exists in three isoforms in the human body: muscle, liver and brain.[28,29] The three isoenzymes catalyse the same phosphorolytic reaction by using pyridoxal phosphate (PLP) as a cofactor,[30] but they differ in the extent to which are regulated by the allosteric modulator adenosine monophosphate (AMP) or by the serine (Ser)-phosphorylation.[31,32] For instance, the muscular isoform is strongly and cooperatively activated by Ser-phosphorylation and binding of the AMP, the brain enzyme mainly responds to a non-cooperative AMP activating event, whereas the liver isoform is regulated *via* Ser-phosphorylation in response to hypo- and hyperglycemic hormonal changes.[33]

On the other hand, GDE is the 4-α-glucanotransferase amylo-1,6-glucosidase that removes the branches of a glycogen particle. Although the crystal structure of the human GDE has not been reported so far, a study that focused on the analysis of the *Candida glabrata* homolog instead, revealed that GDE is a large, elongated monomer that is composed of four regions.[34,35] The *N*-terminal domain has GT activity and transfers a maltotriosyl group from the α(1,6) tetrasaccharidic glycogen branch (resulted from the GPH activity) to the non-reducing end of a proximal branch, forming an α(1,4)-glycosidic linkage between them. On the other hand, the *C*-terminal region has an amylo-α(1,6)-

glucosidase activity similar to that of GH15 family members and therefore is able to cleave the residual Glc of the branch resulted from the preceding GT step.[34]

The glucose-1-phosphate (G1P) released by GPH is isomerized into G6P by the phosphoglucomutase (PGM) enzyme, and the resulted product supplies the glycolytic pathway that generates the energy required by the cell.[15] PGM-1 is the most abundant among the many PGM isozymes found in humans and is a monomeric, metal-dependent enzyme.[36] Depending on the energetic state of the cell, PGM1 catalyses the reversible interconversion between G1P and G6P *via* a phosphoryl transfer mechanism between one of the two substrates and the phosphoserine residue present in the active site.[37]

The *de novo* G6P synthesis is operated by one of the two hexokinase isozymes found in the human skeletal muscle (HK1/HK2), from the Glc that enters the cell and the mitochondrial adenosine triphosphate (ATP).[38,39] If not immediately used to produce energy in the glycolysis pathway the G6P can be directed towards glycogenesis *via* conversion into G1P and then into UDP-Glc.[24] UDP-glucose pyrophosphorylase (UGP) is the octameric enzyme that forms UDP-Glc and pyrophosphate (PP_i) from uridine triphosphate (UTP) and G1P in a Mg^{2+}-dependent, reversible manner.[40,41]

All things considered, the glycogen metabolism is a highly dynamic process, and it involves the multiple orchestrating partners described so far. Moreover, the process is made even more complex by many other proteins which by acting in an indirect manner on the glycogenesis and glycogenolysis pathways, have modulatory effects on the overall energetic status of the cell.[1,15]

Since the fundamental knowledge of the glycogen biogenesis is still surprisingly limited, we decided to undertake additional scientific endeavours in this area and further investigate the activity of the three biosynthetic key players: GYG, GYS and GBE.

4.2 Results and discussion

4.2.1 Enzyme-compatible Suzuki-Miyaura cross-coupling (SMC) enabled mechanistic insights into the biocatalytic activity of GYG

The collaboration between the groups of Davis, Yue and Rovira[42] found that by accessing catalytically active mimics of GYG-glucosylation states through a benign Pd-mediated $C(sp^2)$–$C(sp^2)$ ligation, an unusual triphasic mechanism could be delineated for the catalytic domain of human GYG (Figure **4.2** and Figure **1.18**). Accordingly, by replacing the catalytic Tyr195 of the natural protein (Figure **4.2**, 1) with the tagging amino acid *p*-iodophenylalanine (*p*IPhe) a suitable starting material for the SMC could be obtained (Figure **4.2**, 2). Then, after meticulous optimization of the cross-coupling reaction conditions, three glycoform-mimics of different GYG glucosylation states were generated (Figure **4.2**, 3).[43] Most importantly, upon incubation with UDP-Glc and $MnCl_2$ the obtained structurally accurate mimics GYG-Glc, GYG-Glc_2 and GYG-Glc_6 displayed an unaltered biocatalytic activity and gave the autoglucosylated GYG-Glc_n (4, n ~ 9 avg. Glc).[43] However, these three mimics of natural GYG glycoforms had different glucosylation kinetics which appeared to be dependent on the length of the GYG-ligated sugar chains. Hence, the cross-coupled GYG-Glc, GYG-Glc_2 and GYG-Glc_6 revealed that the GYG autoglucosylation is a complex process characterized by a very rapid intermediate phase that leads to up to 4/5 Glc residues attached ($K_{app} > 0.126\ s^{-1}$), flanked by a slow initiation phase in which the enzyme extends with up to 2 Glc residues ($K_{app} = 0.016\ s^{-1}$) and by a second slower phase in which the stepwise glucosylation from 6 Glc ($K_{app} = 0.003\ s^{-1}$) is eventually curtailed.

In order to validate the proposed triphasic mechanistic model we further used the autoglucosylation apparent rate constants of the three GYG glycoform-mimics to reconstruct a kinetic profile that was in good agreement with the autoglucosylation of GYG-wt-0Glc. The wild-type GYG in its non-glucosylated form (1), when incubated with UDP-Glc and $MnCl_2$ for 15 min using the same reaction conditions affords GYG-wt-nGlc (5, n ~ 9 avg. Glc), the natural heterogeneous mixture of self-glucosylated glycoforms.[12]

To this end, the His-tagged GYG-wt0Glc protein was expressed into the *E. coli* M226 strain lacking UDP-glucose pyrophophorylase activity, according to a protocol previously developed and optimised by the Yue group.[11] SDS-PAGE and LC-MS analyses after purification by Ni^{2+}-affinity chromatography confirmed successful isolation of the His_6-tagged GYG-wt-0Glc (Figure **4.3**, a and figure SI 4.1). Then, upon enzymatic cleavage of the polyhistidine tag and the required purification steps, the almost sugar-free GYG-wt-0Glc protein was obtained (Figure **4.3**, b and figure SI 4.1).

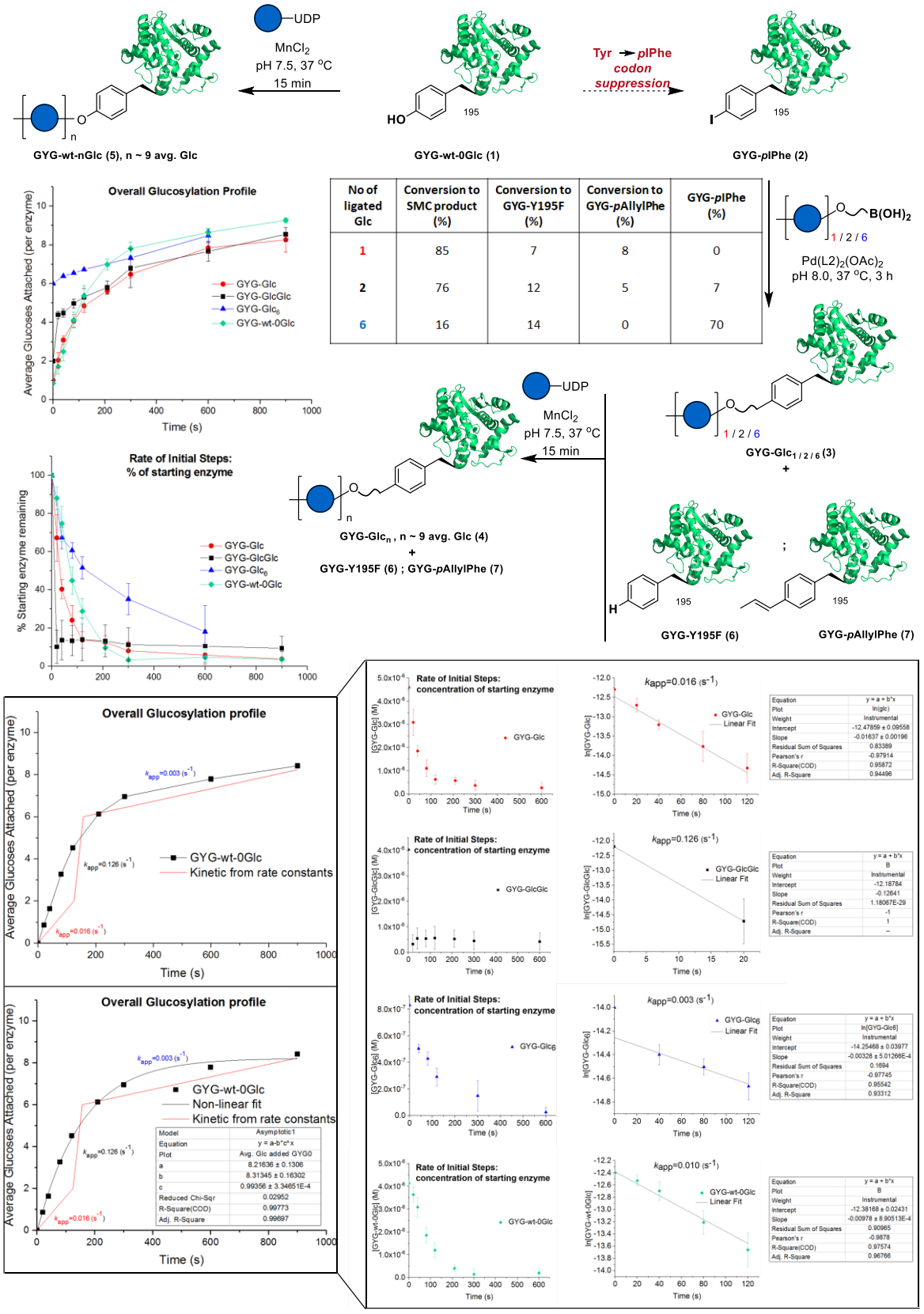


Figure 4.2. Autoglucosylation kinetic analyses and comparison of GYG-wt-0Glc with the three GYG-glycomimetics obtained *via* a protein-compatible SMC which also generated low amounts of enzymatically inert side-products (GYG-Y195F and/or GYG-*p*AllylPhe).[42,43] Data for GYG-wt-0Glc are shown as mean ± s.d. of the three independent replicate kinetic assays that complemented studies previously performed in the Davis group (⌒ = allyl linker).

Interestingly, protein LC-MS analysis revealed that although more than 60 % of the protein was in its non-glucosylated form, small amounts of GYG carrying up to 4 Glc residues could still be detected. Despite the residual, low level of GYG glucosylation, the name GYG-wt-0Glc assigned to the protein expressed by the M226 cells defective in UDP-Glc production was considered to be pertinent. Moreover, although the ratios between these GYG-species may slightly vary from one batch of cells to another (± 2 %), among the low abundant, partially glucosylated species, the glycoform carrying one Glc residue is always the predominant one (~ 16 % vs. less than 10 % for all the others). This finding represented an additional proof that the first two autoglucosylation steps (GYG-wt-0Glc -> GYG-wt-1Glc -> GYG-wt-2Glc) were much slower than the two subsequent Glc-extensions (GYG-wt-2Glc -> GYG-wt-3Glc -> GYG-wt-4Glc). This had ultimately preserved the protein in its non-glucosylated form (GYG-wt-0Glc) and had caused an accumulation of the GYG-wt-1Glc glycoform.

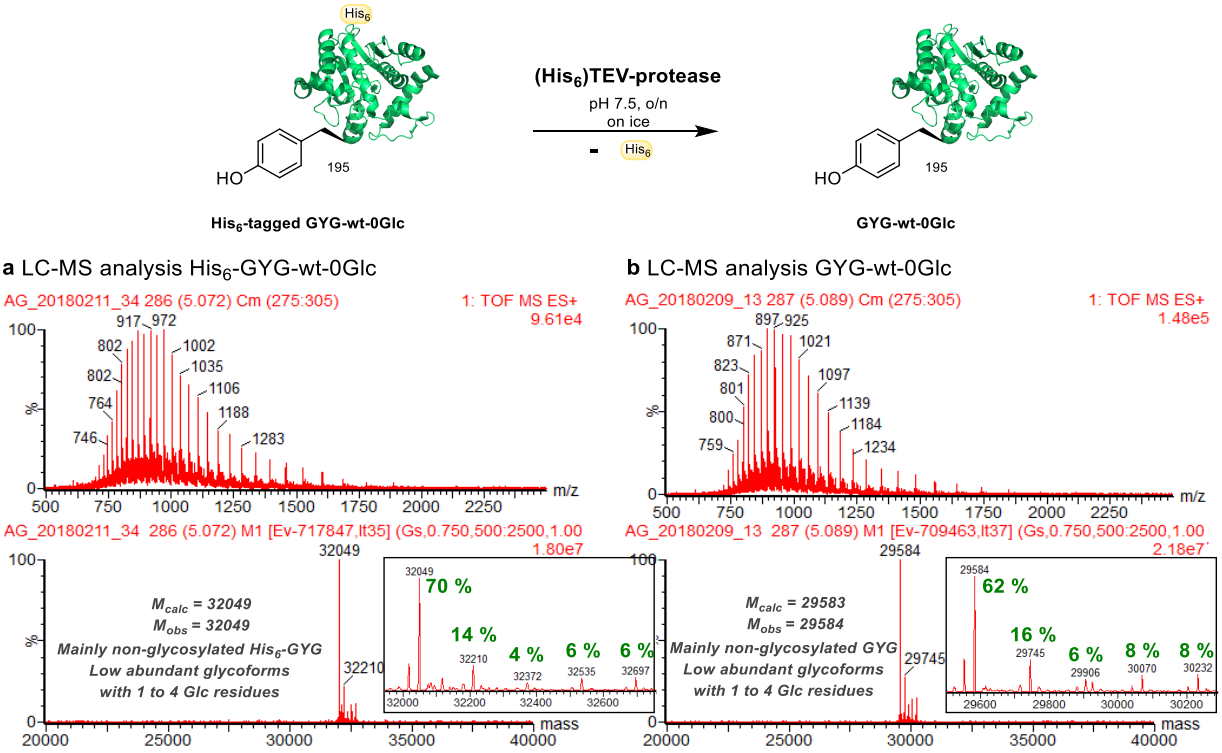


Figure 4.3. Characterization of His_6-GYG-wt-0Glc and GYG-wt-0Glc by intact protein LC-MS analyses. The molecular mass of the enzyme before (M_{calc} = 32049) and after cleavage of the His_6-tag (M_{calc} = 29583) was readily identified by denaturing, intact protein LC-MS analyses. In addition, small amounts of the protein carrying 1 Glc, 2 Glc, 3 Glc or 4 Glc could be observed both for the His_6-GYG-wt-0Glc (M_{obs}: 32210, 32372, 32535, 32697) and the GYG-wt-0Glc (M_{obs}: 29745, 29906, 30070, 30232) protein.

As expected, upon incubation of GYG-wt-0Glc with UDP-Glc and $MnCl_2$ for 15 minutes the protein glucosylated itself to generate glycoforms carrying an average number of approximately 9 Glc residues (Figure **4.2**, 5). Its kinetic profile overlayed reasonably well onto the one generated from the Glc-extension of the three synthetic GYG-intermediates performed under the same reaction conditions, and the calculated apparent rate constant was $K_{app} = 0.010\ s^{-1}$.

In addition, Davis and co-workers[42] had found that the Suzuki–Miyaura cross-coupling (SMC) of GYG-*p*IPhe to the three glycosylated boronic acids proceeded with high to useful reaction conversions to afford the desired GYG-glycomimetics (Figure **4.2**, Table). The non-quantitative yield observed for GYG-Glc (85 %) arose not from the partial consumption of the starting material, but it was due to the formation of GYG-Y195F (7 %) and GYG-*p*AllylPhe (8 %) (Figure **4.2**, Table, entry **1**). The former was plausibly considered a side-product that had been generated by the well-known de-iodination pathway of SMC (Figure **4.2**, 6),[44-46] and the latter was attributed to a GYG-variant that had resulted upon "reductive substitution" of the carbohydrate moiety (Figure **4.2**, 7).[42] While less documented than the de-iodination side-reaction, this alternative SMC side-pathway, likely involving a Pd-catalysed hydrogenolysis of the allylic glycoside,[47-49] led to 5 % GYG-*p*AllylPhe during formation of the GYG-Glc_2 product (76 %) and it was not observed when GYG-*p*IPhe was reacted with the maltohexasaccharide boronic acid (Figure **4.2**, Table, entry **2** and **3**). Lastly, despite numerous optimization attempts, the level of the cross-coupled product GYG-Glc_6 stalled at 16 % and the main product that resulted from the SMC of the maltohexaose boronic acid was the unreacted *p*IPhe-tagged protein (70 %) (Figure **4.2**, Table, entry **3**).[43]

Given the surprisingly fast autoglucosylation kinetics of GYG-Glc_2 we were prone to consider that, this might have been an artefact caused by the residual, most abundant SMC side-product GYG-Y195F (12 %) which had not been removed prior to the autoglucosylation assay of GYG-Glc_2. Since GYG-Y195F lacks the native acceptor capacity (*p*OH) and thus it cannot undergo Glc-extension,[11] we next embarked on testing whether this was able to catalyse, in an intermolecular manner, the glucosylation of an extendable GYG-variant, the GYG-wt0Glc.

To accomplish this goal, the His-tagged GYG-Y195F variant was expressed in *E. coli* BL21(DE3) cells and after purification by Ni^{2+}-affinity chromatography the successful isolation of the protein was confirmed by SDS-PAGE and LC-MS analyses. Then, upon enzymatic cleavage of the polyhistidine tag and subsequent purification steps, the desired GYG-Y195F protein was obtained (Figure SI 4.2).

We then proceeded with performing an autoglucosylation assay in which prior to the incubation with UDP-Glc and $MnCl_2$, the GYG-wt0Glc protein solution was spiked with the GYG-Y195F variant in a 1:1 ratio (Figure **4.4**). Pleasantly, although the latter was added in a much higher ratio (50 %) than the one present in the autoglucosylation reaction mixture of GYG-Glc_2 (12 %), the data not

only confirmed that this side-product was unable of self-glucosylation but, by slightly suppressing the GYG-wt0Glc glucosylation, it helped us discard the hypothesis by which GYG-Y195F would be able to catalyse the elongation of the sugar-chain on an extendable GYG-variant.

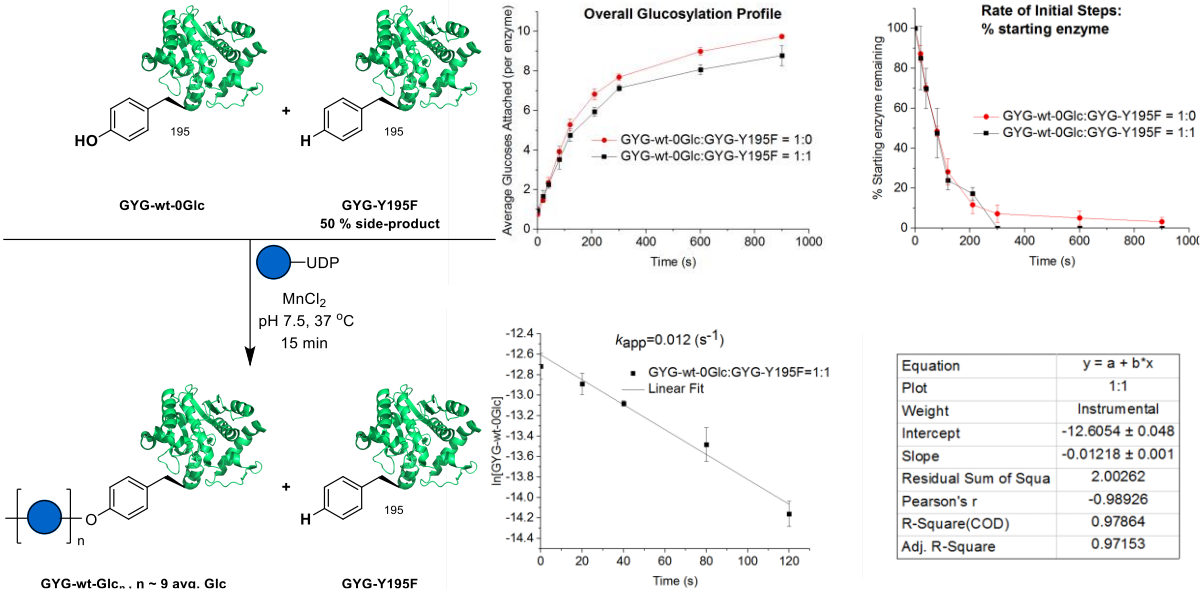


Figure 4.4. The autoglucosylation of GYG-wt-0Glc was slightly inhibited rather than enhanced when incubated in a 1:1 mixture with the GYG-Y195F. The findings helped overrule the hypothesis that the unprecedented fast autoglucosylation observed for GYG-Glc_2 might have been caused by the SMC side-product GYG-Y195F. Data are shown as mean ± s.d. for three independent replicate kinetic assays.

4.2.2 ISCAMS analyses revealed dissociation of the autoglucosylated GYG-wt-nGlc

The glycogenesis is a dynamic process that mainly involves the orchestrating partners GYG, GYS and GBE (Figure **4.1**). Nevertheless, our knowledge of the glycogen biogenesis has been limited by the lack of solution-based imaging techniques that are able to identify in real time, the interactions that occur between GYG, GYS and/or GBE. Therefore, we believed that by applying the recently emerging iSCAMS technology[50] to the study of the glycogenesis process we would be able to detect both the homogeneous associations that occur between monomers of the same protein (e. g. GYG, GYS) and the heterogeneous assembly of subunits belonging to different macromolecules that may presumably generate various biosynthetic complexes (e. g. GYG-GYS, GYG-GBE, GYG-GYS-GBE). Not only would this be instrumental in understanding the enzymatic stoichiometry that leads to the formation of a glycogen particle but upon changing the reaction conditions (e. g. concentration of UDP-Glc, co-factors or allosteric modulators), this strategy could have the potential to offer additional insights into the glycogen biogenesis, a process that is dependent on the energetic status of the cell.

However, before proceeding with the analysis of complex biosynthetic mixtures of GYG, GYS and GBE we first decided to use iSCAMS to detect the oligomeric state of the catalytic domain of human GYG in solution, before (GYG-wt-0Glc) and after autoglucosylation (GYG-wt-nGlc).

Hurley *et al.*[10] were the first to crystalize the full-length wild-type (partially glucosylated) rabbit muscle GYG as well as the seleno-methionine-substituted variant which aided with the phasing of the MAD crystallography method.[51] Firstly, the authors were able to obtain dimeric crystal structures of the apo-GYG and of its complex with UDP–Glc and Mn^{2+}, and upon crystallographic analyses they proposed a pertinent enzymatic mechanism for the retention of the anomeric configuration that occurs in the transferred Glc residues. Secondly, Hurley and co-workers[10] also speculated that within the dimeric structure, the GYG-autoglucosylation might take place in an intermolecular manner until a critical polyglucosan minimal length allows the shift to an intramolecular mechanism. Thirdly, they were able to determine that due to enzymatic proteolysis the protein expressed in *E. coli* lacks the sixty *C*-terminal amino acids (out of the 332 residues), and that nevertheless, the absence of this fragment does not impact the activity of the catalytic domain of GYG.[52] Lastly, Hurley *et al*.[10] disclosed that in addition to the dimer crystal structure that GYG presented, the analytical ultracentrifugation data and the gel filtration analysis also indicated that the catalytic domain of GYG (262 residues) exists as a dimer in solution.

Yue and co-workers[11] were the first to crystallize the dimeric catalytic domain of the human muscle GYG. In addition to achieving this goal, from the recorded gel filtration chromatograms they determined that this enzyme is a dimer also in solution.[11] Moreover, upon analysing the three crystallized GYG variants (partially glucosylated GYG-wt, non-glucosylated GYG-wt-0Glc and mutant GYG-Y195F) the authors were able to demonstrate that after the initial autoglucosylation, the growing maltosaccharide chain of each of the two subunits forming the GYG dimer is oriented to be extended with up to 4 Glc residues in an intramolecular fashion.[11] In contrast, the study also showed that a longer polyglucosan GYG chain containing at least 6 Glc ensures that the acceptor GYG-linked maltosaccharide reaches the GT active site of the adjacent subunit to be further elongated in an intermolecular manner.[11]

This work was later complemented by additional findings which were generated within the collaboration framework established by the groups of Davis, Yue and Rovira.[42] Accordingly, the MD simulations of 'UDP-Glc + GYG-wt-Glc_n' complexes (n = 0–5) suggested that the first GYG-catalysed glucosylation steps (n = 0/1) are preferentially inter-monomeric.[42] Moreover, the computational study confirmed that this is due to a peculiar loop within the GYG structure that impedes the protein from adopting the active conformation responsible for the intra-monomeric Glc-extension.[42] On the contrary, the study demonstrated that longer sugar chains (n = 2/3) can circumvent the blocking loop,

and the subsequent intra-monomeric autoglucosylation steps generate GYG-linked polyglucosans containing up to 4/5 Glc residues.[42] The latter are long enough to reach the active site of the adjacent monomer and the study concluded that these are further glucosylated *via* an inter-monomeric mechanism.[42]

Curtino *et al.*[53] described for the first time the existence of rabbit muscle GYG variants (catalytic domains) as monomers, in dilute protein solutions. Gel filtration studies carried out by loading 17–20 μM protein solutions showed that the partially glucosylated GYG-wt as well as the non-glucosylated GYG-wt-0Glc were eluted mainly as dimers.[53] Conversely, when loaded at a concentration of 0.5 μM, the elution profiles of both GYG variants were shifted to that of a GYG monomer form.[53] A residual population of dimers was noticeable in the eluted GYG-wt-0Glc (0.5 μM) whereas the chromatogram of the partially glucosylated GYG-wt (0.5 μM) revealed presence of only monomers. However, the study neither acknowledged the presence of residual dimerized GYG-wt-0Glc nor it elaborated on the factors that might have led to the obtained results.

All in all, we believed that most likely the predominant GYG dimer forms that were identified in the studies described above might have been generated by the high enzyme concentrations that were required for crystal formation (210–450 μM), analytical ultracentrifugation or gel-filtration (17–20 μM). Thus, the obtained results might not have been accurate representations of the actual oligomerization state of GYG that occurs in humans, at biologically relevant *in vivo* protein concentrations.

Therefore, we decided to investigate the oligomerization state of the human muscle GYG catalytic domain at nanomolar concentrations, before and after autoglucosylation. To this end, in addition to the intact protein LC-MS spectra (Figure **4.5**, a) that were recorded for GYG, before and after incubation with UDP-Glc and $MnCl_2$ for 60 minutes, iSCAMS analyses (Figure **4.5**, b) were also performed for each of the two variants: the non-glucosylated GYG-wt-0Glc (Figure **4.5**, 1) and the resulted autoglucosylated GYG-wt-nGlc (Figure **4.5**, 2). The two analytical methods were used in parallel because the intact protein LC-MS by virtue of its dissociating nature cannot be used to investigate the oligomerization state of a protein. However, due to its high MS resolution and accuracy intact protein LC-MS is amenable for the precise identification of heterogeneously glucosylated GYG-glycoforms.[54,55] Conversely, iSCAMS, a biophysical method that only relies on the scattered light to measure the mass of (glyco)proteins and related complexes without causing the dissociation of the latter,[50] appeared to be suitable for the identification of the GYG oligomerization state. In addition, it did not seem that iSCAMS will deliver the same analytical performance as the LC-MS method (2 % sequence mass accuracy, up to 19-kilodalton resolution, and 1-kilodalton precision)[56] that allows the identification of the mass increase caused by the autoglucosylation of GYG.

Indeed, the approximately 1.7 kDa mass difference that was generated by the GYG autoglucosylation process was determined by intact protein LC-MS (Figure **4.5**, a) and could not be identified by performing iSCAMS recordings of GYG-wt-0Glc (1) and GYG-wt-nGlc (2).

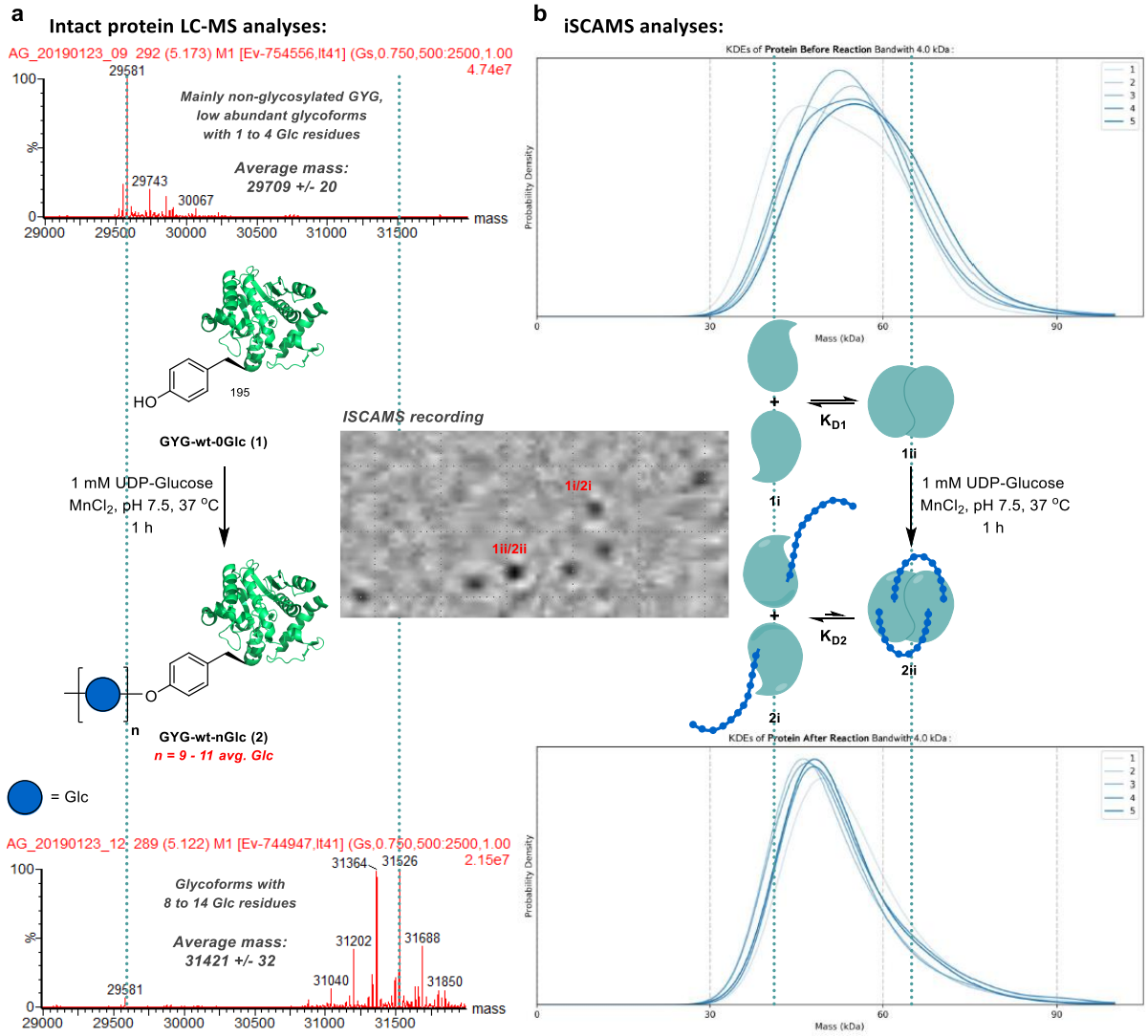


Figure 4.5. GYG was analysed by intact protein LC-MS and iSCAMS analyses before (GYG-wt-0Glc) and after (GYG-wt-nGlc) autoglucosylation. Intact protein LC-MS data showed that upon incubation with UDP-Glc and $MnCl_2$ for 60 min GYG extends its Tyr195 residue to form a short (n = 9-11 Glc residues) enzyme-bound maltosaccharide (**a**). iSCAMS results obtained from the analysis of the non-glucosylated protein (1) and its self-extended version (2) revealed that while the former was present both as a monomer (1i) and a dimer (1ii), the latter existed as mainly a monomer (2i) and had presumably resulted from the dissociation of the GYG-wt-nGlc dimer intermediate (2ii) that might have a higher dissociation constant than 1ii ($K_{D2} > K_{D1}$) (**b**).

Surprisingly, the obtained iSCAMS data revealed that equal populations of monomers (1i) and dimers (1ii) were present in the GYG-wt-0Glc (1) sample. The same analytical method applied to determine the oligomerization state of GYG-wt-nGlc (2) showed that after autoglucosylation, the enzyme existed mainly as a monomer (2i) and a significantly much lower population of dimers (2ii) could still be noticed in the analysed protein solution. This suggested that at the end of the third autoglucosylation inter-monomeric phase the GYG dimer may first dissociate before being further

processed by the downstream enzymes GYS and GBE, and that the extended polyglucosan chain may have increased the dissociation constant (K_D) of the dimer, thus favouring its disassembly into separate autoglucosylated monomers.

4.2.3 Intact protein LC-MS analyses revealed hydrolysis of the autoglucosylated GYG-wt-nGlc and the GYS-extended GYG-wt-(m+n)Glc polyglucosan chains in absence of UDP-Glucose

Before proceeding with further analysing by iSCAMS the associations of autoglucosylated GYG-wt-nGlc with GYS and/or GBE in various biosynthetic complexes, we first embarked on exploring reaction conditions that would allow the glycogen biosynthetic enzymes GYS and GBE to further glucosylate and transglucosylate respectively, the GYG-liked maltosaccharide chain.

To this end, we performed a set of experiments in which GYG-wt-0Glc (Figure **4.6**, 1) was first incubated with 1 mM UDP-Glc and $MnCl_2$ for one hour (Figure **4.6**, A) and then, the resulted autoglucosylated GYG-wt-nGlc (Figure **4.6**, 2) was further incubated with 10 mM UDP-Glc, with or without enzymes GYS and GBE and in presence or absence of 3 mM G6P, the allosteric activator of GYS (Figures **4.6** and **4.7**).[18] The progression of the reactions was monitored by intact protein LC-MS and SDS-PAGE analyses and the obtained results revealed important aspects about the fate of the GYG-wt-nGlc substrate when subjected to all of these different enzymatic conditions.

Firstly, when 2 was incubated with GYS in the presence of 3 mM G6P (Figure **4.6**, B) long GYG-wt-(n+m)Glc species bearing up to 36 Glc residues could be detected by LC-MS analysis (Figure **4.6**, 3a, n+m = 19-21 avg. Glc) already in the first four hours, and the obtained size distribution of GYG-glycoforms remained largely unaltered when the reaction time was extended to 16 hours. The GYS-catalysed extension of GYG-wt-nGlc was also confirmed by the SDS-PAGE analysis in which the GYG-wt-(n+m)Glc mixture of glycoforms migrated as slightly higher MW species than the starting material GYG-wt-nGlc (Figure **4.6**).

Secondly, when GYG-wt-nGlc was incubated with 10 mM UDP-Glc and 3 mM G6P in absence of GYS (Figure **4.6**, C) medium GYG-wt-(n+m)Glc species bearing up to 23 Glc residues could be detected by LC-MS analysis (Figure **4.6**, 3b, n+m = 15.8 avg. Glc) after four hours, but the obtained GYG-glycoforms underwent hydrolysis to approximately 14 avg. Glc residues in the next 12 hours.

A similar outcome was obtained for the reaction mixture in which 2 was incubated with 10 mM UDP-Glc in absence of GYS and G6P (Figure **4.6**, D). Again, medium GYG-wt-(n+m)Glc glycoforms could be detected by LC-MS analysis (Figure **4.6**, 3c, n+m = 15.8 avg. Glc) after four hours, and the

hydrolysis of the polyglucosan chains generated GYG-wt-(n+m)Glc species that carried approximately 13 avg. Glc residues when the assay was terminated.

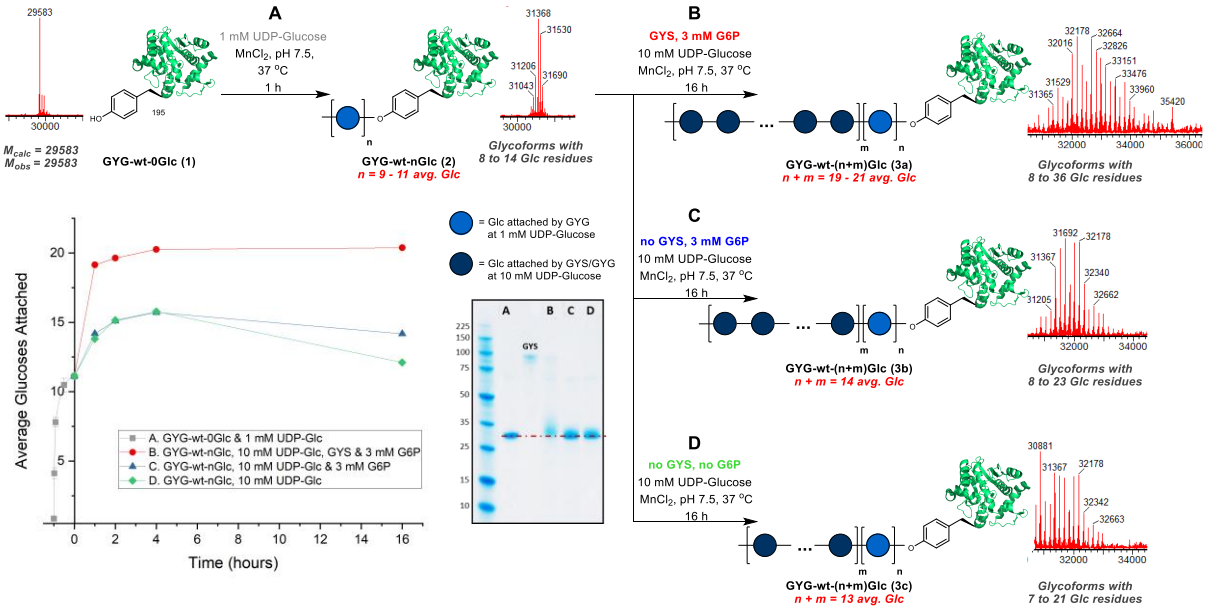


Figure 4.6. GYS further extends autoglucosylated GYG-wt-nGlc to generate long GYG-linked polyglucosan chains that are stable glucose-reservoirs. At 1 mM UDP-Glc concentration, GYG-wt-0Glc (**1**) catalyses a rapid, multi-step, autoglucosylation reaction (**A**) and the GYG-wt-(n)Glc (n = 9-11 avg. Glc residues, **2**) is formed. Subsequent addition of GYS and its allosteric activator glucose-6-phosphate (G6P) (**B**) allows for GYG-wt-(n+m)Glc species (n + m = 19-21 avg. Glc residues, **3a**) to be obtained after 4 hours. This mixture of extensively glucosylated GYG-glycoforms presents a wider size distribution that remains stable for up to 16 hours. In contrast, further autoglucosylation of GYG-wt-nGlc (GYS not added) at high UDP-glucose concentrations, with and without G6P (**C** and **D**, respectively), generates shorter GYG-glycoforms (n + m = 15.8 avg. Glc residues) after 4 hours. These intermediates slowly hydrolyse at longer incubation times, to generate slightly deglucosylated species carrying up to 14 (**3b**) and 13 (**3c**) average Glc residues.

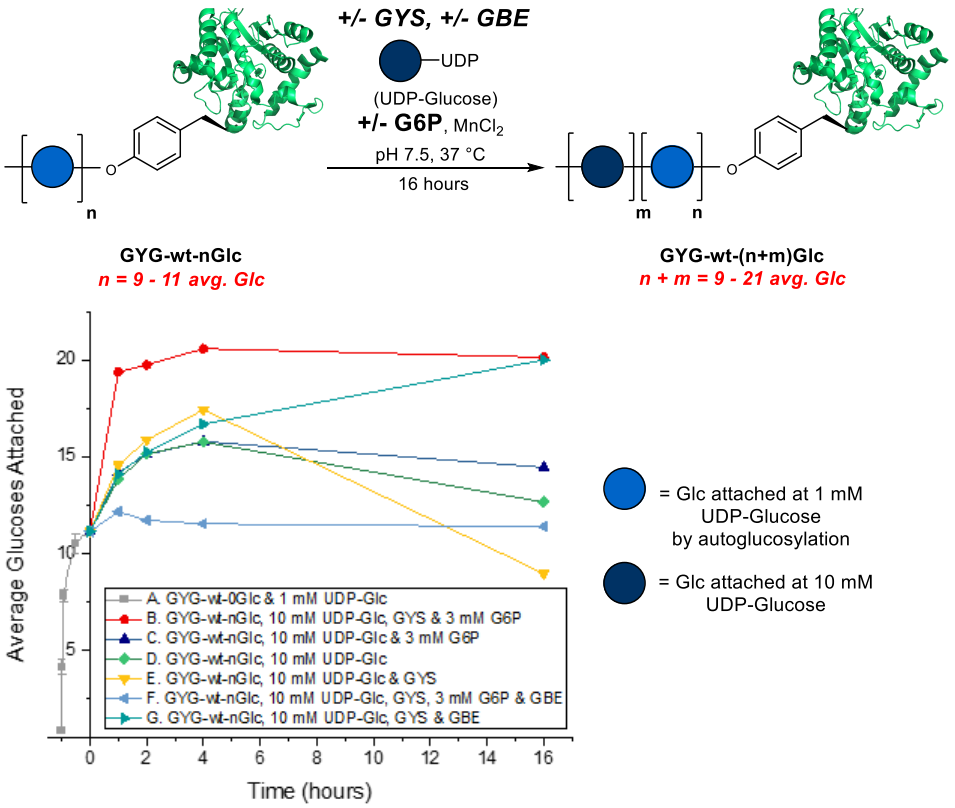
These reactions (Figure 4.6) were part of a more comprehensive screening of enzymatic conditions in which the activity of GBE was investigated as well (Figure 4.7, E to I).

LC-MS analyses of the reaction mixture in which GYG-wt-nGlc was incubated with 10 mM UDP-Glc and GYS but in which the enzymatic allosteric activator G6P was omitted (Figure **4.7**, E) revealed that after 4 hours, the Glc-extension afforded medium GYG-wt-(n+m)Glc glycoforms (n + m = 17 avg. Glc residues). Surprisingly, analyses performed after 16 hours showed that the initially observed glucosylation phase was followed by a hydrolytic phase during which the GYG-linked polyglucosan chains underwent a drastic decrease of their length to afford trimmed GYG-wt-nGlc (n = 9 avg. Glc residues) glycoforms (Figure SI 4.8).

Unexpected results were also obtained for the reaction mixture that contained GYG-wt-nGlc, 10 mM UDP-Glc, GYS, G6P and GBE (Figure **4.7**, F). It appeared that the employed enzymatic conditions did not significantly alter the glucosylation profile of the starting material and at the end of the

incubation time the recorded LC-MS showed that GYG-wt-nGlc carried approximately 11 avg. Glc residues (Figure SI 4.8).

Lastly, when GYG-wt-nGlc was incubated with 10 mM UDP-Glc, GYS and GBE but in absence of G6P (Figure **4.7**, G) medium GYG-wt-(n+m)Glc species (n+m = 17 avg. Glc) bearing up to 22 Glc residues could be detected by LC-MS analysis in the first four hours. However, within the next 12 hours these intermediates were gradually Glc-extended and afforded the GYG-wt-(n+m)Glc (n+m = 20 avg. Glc) mixture of glycoforms that had a more narrow size distribution that the enzymatic product obtained in reaction B (Figure SI 4.8). Moreover, low levels of a second population of even longer glycoforms GYG-wt-(n+m)Glc (n+m = 30-32 avg. Glc) could also be identified in the LC-MS analysis performed after 16 hours. The obtained result might have been caused by the enzymatic activity of GBE which was able to transfer α(1,4)-linked maltosaccharides containing from 8 to 14 sugar units from the outer, non-reducing end of the polyglucosan-GYG into the α(1,6) position of the same or of an adjacent chain. We believe that by this means, GBE was able to create additional non-reducing termini in the Glc-extended GYG that could be further α(1,4)-glucosylated by any of the other two enzymes (GYG or GYS) present in the reaction mixture.



Entry	GYG-wt-nGlc (0.1 mg/ml)	GYG-wt-nGlc (0.2 mg/ml)	GYS (0.1 mg/ml)	GBE (0.1 mg/ml)	G6P (3 mM)	GYS (0.4 mg/ml)
A*	-	+	-	-	-	-
B	+	-	+	-	+	-
C	+	-	-	-	+	-
D	+	-	-	-	-	-
E	+	-	+	-	-	-
F	+	-	+	+	+	-
G	+	-	+	+	-	-
H**	-	-	-	+	+	+
I**	-	-	-	+	-	+

Figure 4.7. Summary of the reaction conditions employed to glucosylate GYG. Marked with "*" is the autoglucosylation reaction that generated GYG-wt-nGlc (n = 9-11 avg. Glc residues) from GYG-wt-0Glc (Figure **4.6**, A). GYG-wt-nGlc was the starting material that was used in the reactions B to G. Marked with "**" are the reactions in which GYG-wt-nGlc (n = 9-11 avg. Glc residues) was not added (H and I), and because the LC-MS could not detect the existence of any GYG variant in these reaction mixtures the construction of a GYG glucosylation profile was not possible. See *Note on the use of GYS as a complex with GYG-wt for the glucosylation of GYG-wt-nGlc (n = 9-11 avg. Glc residues)* for additional information (Section 4.5.2).

From data gathered here we first speculated that GYG might be a reversible glycosyltransferase. With this hypothesis in mind, we started to investigate whether GYG was able to catalyse the sugar excision from its own glycan chain to an acceptor uridine diphosphate (UDP) molecule, and whether by doing this, its enzymatic activity could be influenced by the concentration of UDP available in the reaction mixture. Thus, autoglucosylated GYG-wt-nGlc (n =9-11 avg Glc residues) was purified by gel filtration to remove excess reagents (UDP-Glc, $MnCl_2$) and the UDP formed during the autoglucosylation of GYG (Figure SI 4.9).

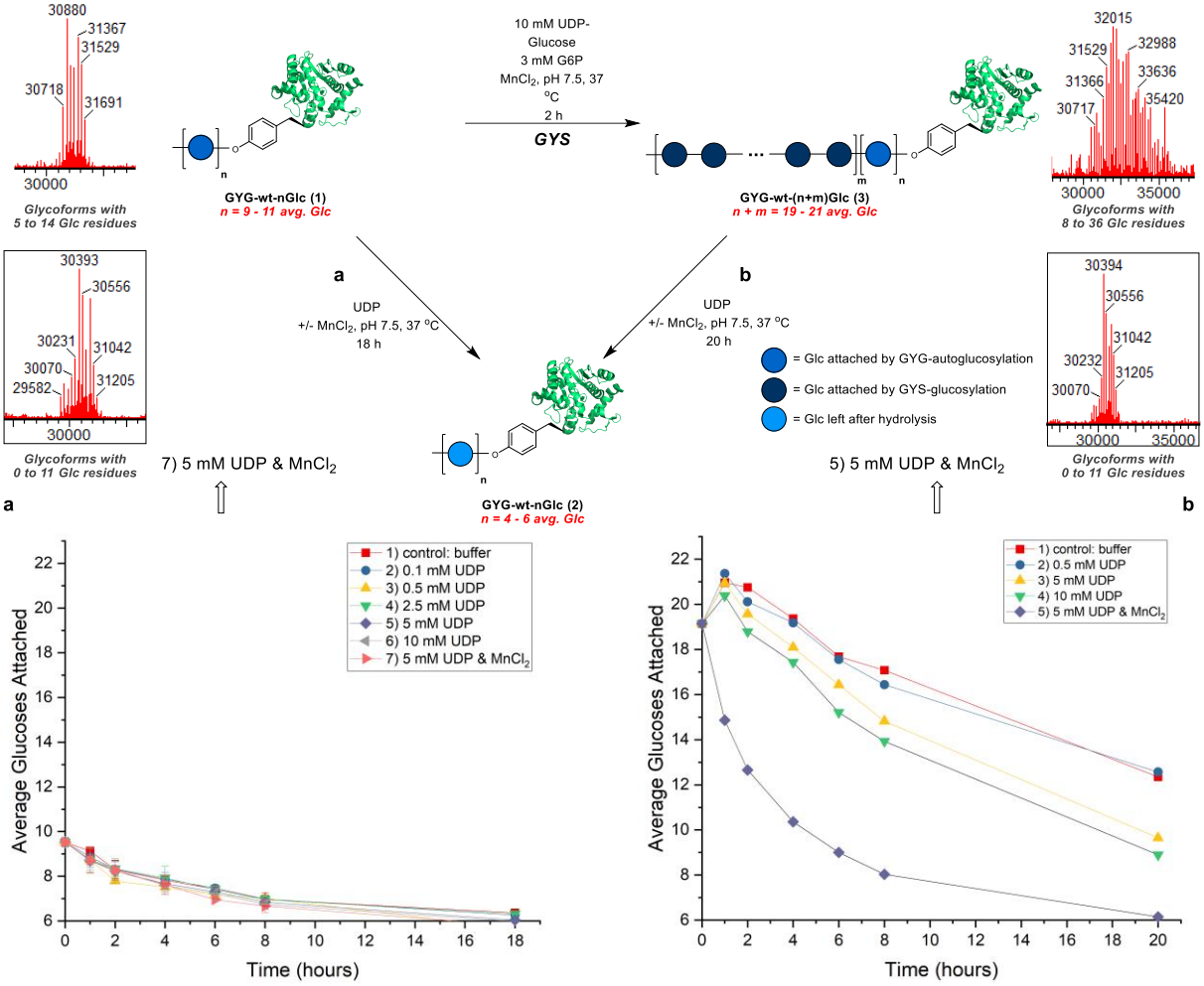


Figure 4.8. The hydrolytic activity observed for GYG dropped dramatically as the length of its polyglucosidic chain decreased. GYG-wt-nGlc (n = 9-11 avg. Glc residues) was incubated in 25 mM NaHEPES pH 7.5 buffer with increasing concentrations of UDP, with or without $MnCl_2$. The experiment showed that the decrease of the sugar chain length to about 4-6 avg. Glc residues did not depend on the presence of UDP or of UDP and $MnCl_2$. Graph shows the avg. number of Glc residue attached to GYG as determined by LC-MS analyses of aliquots taken for each of the seven experiments that were performed in parallel. This screening of enzymatic conditions was performed in duplicate and representative deconvoluted LC-MS spectra are reported in Section 4.4.4.1 (**a**). Also the GYG-wt-(n+m)Glc (n = 19-21 avg. Glc residues) underwent hydrolysis of its polyglucosidic chain in the presence of UDP. However, when this substrate was incubated with both UDP and $MnCl_2$ trimming of the sugar chain occurred in two distinct phases: fast during the first 6 hours and slow upon a longer incubation time. Nevertheless, curtailed glycoforms of GYG-wt-nGlc (n = 4-6 avg. Glc residues) were again obtained. Graph shows the avg. number of Glc residue attached to GYG as determined by LC-MS analyses of aliquots taken for each of the five single-replicate experiments that were performed in parallel. Deconvoluted LC-MS spectra of this single-replicate screening of enzymatic conditions are reported in Section 4.4.4.2 (**b**).

Upon incubation in the NaHEPES pH 7.5 buffer, with various concentrations of UDP or with UDP and $MnCl_2$, GYG-wt-nGlc (n = 9-11 avg. Glc residues, 1) underwent slow hydrolysis of its polyglucosidic chain and curtailed glycoforms of GYG-wt-nGlc (n = 4-6 avg. Glc residues, 2) were generated in a rather UDP-independent manner (Figure **4.8**, a).

Then, a similar set of experiments was performed in which the purified GYS-extended GYG-wt-(n+m)Glc (n+m ~ 19-21 avg. Glc residues, 3) was incubated in the NaHEPES pH 7.5 buffer, with various concentrations of UDP or with UDP and $MnCl_2$ (Figure **4.8**, b). Interestingly, the reaction mixtures containing 3 in the NaHEPES pH 7.5 buffer, with or without UDP, showed that the polyglucosan chain of the substrate was being slowly hydrolysed and the reaction rate appeared only slightly influenced by the concentration of UDP present in the reaction mixture. Conversely, upon incubation of GYG-wt-(n+m)Glc with 5 mM UDP and $MnCl_2$ an initial, very rapid hydrolytic phase afforded GYG-glycoforms carrying only 9 average Glc residues after 6 hours. Then, the fast trimming of extensively glucosylated GYG-species was followed by a second hydrolytic phase during which the intermediate GYG-glycoforms followed a slower hydrolytic profile, similarly to what observed for 1, to give again minimal glycoforms of GYG-wt-nGlc (n= 4-6 avg. Glc residues, 2).

Lastly, the LC-MS analysis of the purified starting material GYG-wt-(n+m)Glc (n+m ~ 19-21 avg. Glc residues) revealed presence of oxonium ions corresponding to cleaved oligosaccharides maltohexaose, maltoheptaose and maltooctaose (Figure SI 4.13). The latter could not be detected after 20 hours of incubation with 5 mM UDP and $MnCl_2$ (Figure SI 4.14). Further investigations are needed in order to fully elucidate whether the observed molecular masses were due to the MS fragmentation of the polyglucosan chain or caused by the enzymatic hydrolytic activity present in the solution of GYG-wt-(n+m)Glc.

4.3 Conclusions and future perspectives

Firstly, in this chapter we were able to demonstrate that the kinetic data previously obtained in the group from the autoglucosylation of catalytically active mimics of GYG-glucosylation states (GYG-Glc, GYG-Glc_2, GYG-Glc_6) reconstructed an overall glucosylation profile that aligned well with the average level of Glc incorporation into the GYG-wt-0Glc, the wild-type GYG in its non-glucosylated form.

In addition, we demonstrated that the SMC-derived side-product GYG-Y195F could not catalyse in an intermolecular manner the glucosylation of GYG-Glc_2. Since autoglucosylation of GYG-wt-0Glc was not enhanced by the deliberate addition of GYG-Y195F we confirmed by analogy, that the fast Glc-extending activity observed for the GYG-Glc_2 cross-coupled mimic could not have resulted from the enzymatic action of residual GYG-Y195F.

Next, we used iSCAMS analyses to investigate the oligomerization state of GYG in solution, before (GYG-wt-0Glc) and after autoglucosylation (GYG-wt-nGlc, n = 9-11 avg. Glc residues). Obtained results indicated that monomers and dimers of the non-glucosylated protein were in equilibrium at the low protein concentration at which the GYG-wt-0Glc solution was analysed. Conversely, under the same conditions, iSCAMS recordings of the GYG-wt-nGlc solution showed that the autoglucosylated protein existed predominantly as a monomer.

Then, with the intention to carry out additional iSCAMS experiments and interrogate the biosynthetic associations that may occur between extensively glucosylated GYG (GYG-wt-(n+m)Glc (n+m ~ 19-21 avg. Glc residues) and the glycogen biosynthetic enzymes GYS and GBE, we serendipitously found that both GYG-wt-(n+m)Glc and GYG-wt-nGlc underwent hydrolysis of their polyglucosan chains upon UDP-Glc removal. Although the underlying mechanisms are yet to be elucidated the kinetics of the observed hydrolytic process appeared to have been influenced by the length of the GYG-ligated sugar chains (Figure **4.9**).

In order to exclude the hypothesis by which a contaminant α(1,4)-glucosidase derived from the expression of GYG in *E. coli* might have been responsible for the observed hydrolytic activity, we digested and submitted the protein for in-depth shotgun proteomics analysis. The results confirmed that one peptide belonging to the maltodextrin glucosidase (malZ) protein sequence could be identified (Appendix 5.5). MalZ is a bacterial α(1,4)-glucosidase able to hydrolyse maltotriose, an inducer of the maltodextrin system in bacteria, to the noninducer maltose metabolite.[57] It was

previously demonstrated that this enzyme is able to hydrolyse maltodextrins ranging from maltotriose to maltoheptaose, but the data reported herein represents the first evidence that this enzyme hydrolyses polyglucosan chains containing up to 36 Glc residues and that in doing so, maltohexaose, maltoheptaose and maltooctaose oligosaccharides may first be produced (Figure SI 4.13).

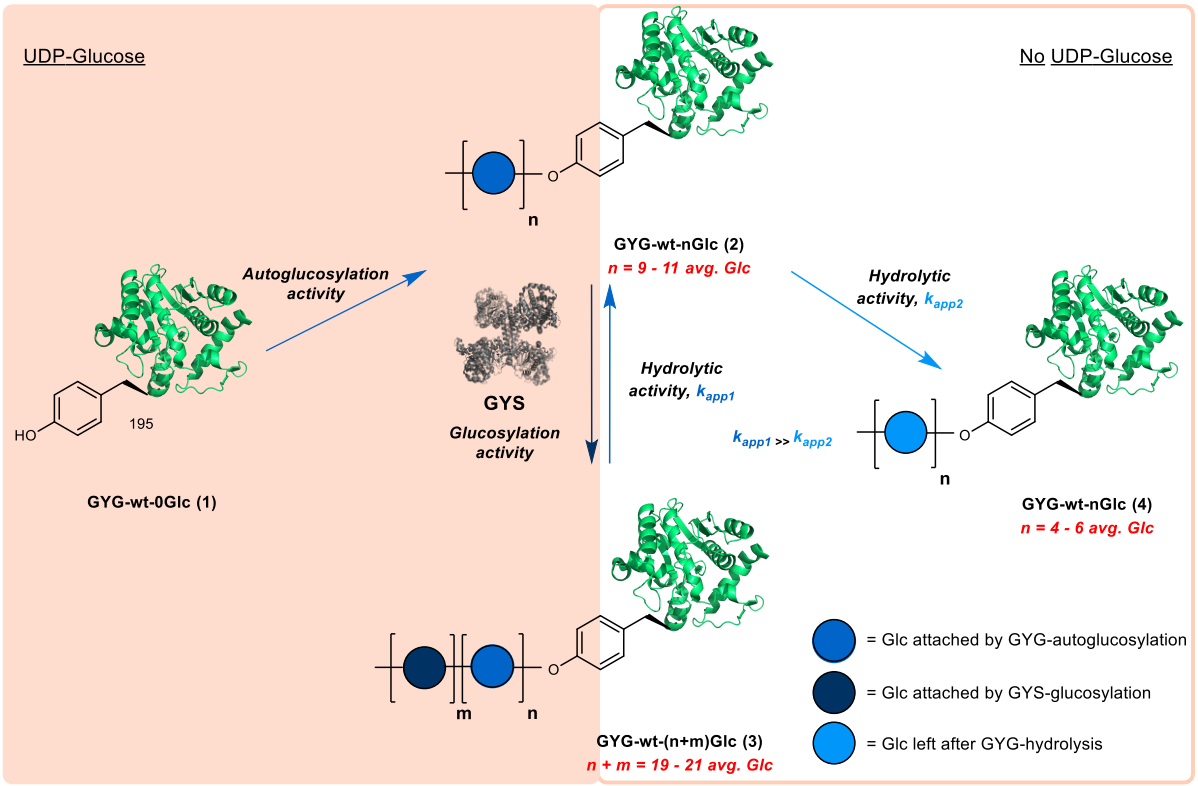


Figure 4.9. Glycogenin displays functional plasticity during the early stages of glycogen biosynthesis. In presence of UDP-Glucose the glycosyltransferase enzyme glycogenin GYG-wt-0Glc (**1**) autoglucosylates its Tyr195 residue to form a short enzyme-bound maltooligosaccharide: GYG-wt-nGlc (n = 9-11 avg. Glc residues, **2**). This undergoes further enzymatic extension by glycogen synthase (GYS) to form GYG-wt-(n+m) (n + m = 19-21 avg. Glc residues, **3**), a mixture of glycoforms characterized by a high degree of Glc-polymerization. In absence of UDP-Glucose, both glucosylation states (**2** and **3**) displayed hydrolytic activity and generated trimmed species GYG-wt-nGlc (n = 4-6 avg. Glc residues) (**4**). The observed deglucosylation is faster for longer glycan chains (k_{app1}>>k_{app2}) and both hydrolytic reactions stall when the GYG-linked polysaccharide reaches a minimal length n (n = 4-6 avg. Glc residues).

We envision that it will be possible to prove the hydrolytic activity of GYG, GYS or malZ using an activity-based probe such as the one reported by Withers *et al.*[58], the trinitrophenyl-2,2-difluoroglucoside. This activated fluorinated glucoside should allow the inactivation of the glucosidase responsible for the hydolysis of the GYG-linked sugar chain. Moreover, the resultant covalent modification of the protein could be monitored both by ^{19}F-NMR and intact protein LC-MS analyses. Finally, upon enzymatic digestion of the covalently inhibited protein LC-MS/MS analyses could further elucidate the amino acid residue involved in the hydrolytic mechanism of the acting glucosidase. If

future investigations will find that the hydrolytic activity is attributable to GYG this may lead us to rethink our current understating about the human glycogenesis and glycogenolysis processes.

Although GYG is considered as exclusively essential in glycogenesis - it primes itself with Glc for the subsequent polyglucosan self-catalysed extension - new data has emerged to indicate that there may be unknown aspects about the GYG-governed processes that are worth further investigation. Perhaps the most striking findings were revealed by Guinovart *et al.*[59] who found that GYG1 knocked out (KO) mice were still able to convert Glc into GYG-free glycogen (Figure **4.1**, 4). Surprisingly, although 4 appeared to be excessively accumulated in the muscles of GYG1 KO mice it could not ensure the same level of endurance to exercise that was reached by the control group of animals.

Moreover, studies published by Vissing and co-workers[60] revealed that glycogen formation occurs in the muscles of GYG1-defficient humans as well. However, unlike mice, humans have the liver-specific isoform GYG2 which appeared to have been upregulated in these patients. This comes to suggest that GYG1 and GYG2 may coexist in the muscle in special circumstances, and when the former is affected by an inborn error the latter may take over the initiation of glycogen biosynthesis.

Despite these differences, both studies found that the muscular glycogenesis process can occur in absence of GYG1 and counterintuitively, the resulted glycogen particles appear larger than the normal, GYG1-linked particles.

All in all, the data indicates that GYG may not just influence the glycogen biosynthesis, but it may, *via* mechanisms that are still to be elucidated, affect the consumption of the glycogen particles as well as the energy expenditure.

4.4 Experimental section

4.4.1 Plasmid DNA details and protein sequences

The plasmids utilised in the course of this work were as follows:

GYG-wt-0Glc: donated by the group of Prof. Wyatt Yue, previously prepared by Chaikuad A., Froese D. S. and Berridge G.[11] in a pNIC28-Bsa4 vector;

*M*HHHHHH*SSGVDLGTENLYFQ*SMTDQAFVTLTTNDAYAKGALVLGSSLKQHRTTRRLVVLATPQVSDSMRKVLE
TVFDEVIMVDVLDSGDSAHLTLMKRPELGVTLTKLHCWSLTQYSKCVFMDADTLVLANIDDLFDREELSAAPDPG
WPDCFNSGVFVYQPSVETYNQLLHLASEQGSFDGGDQGILNTFFSSWATTDIRKHLPFIYNLSSISIYSYLPAFKVFG
ASAKVVHFLGRVKPWNYTYDPKTKSVKSEAHDPNMTHPEFLILWWNIFTTNVLPLLQ

The *N*-terminus of the protein construct contains a polyhistidine tag which is highlighted in yellow and a TEV-cleavable peptide which is highlighted in green. The peptide containing the first 21st amino acid residues that are removed upon TEV-protease cleavage are italicized. For the protein sequence, numbers refer to the amino acid sequence without this cleavable peptide. The autoglucosylation site, the 195 residue is in red.

GYS: donated by Dr Thomas McCorvie (Prof. Wyatt Yue group) as a complex with GYG since the latter is required for optimal glucotransferase activity;

SMPLNRTLSMSSLPGLEDWEDEFDLENAVLFEVAWEVANKVGGIYTVLQTKAKVTGDEWGDNYFLVGPYTEQGV
RTQVELLEAPTPALKRTLDSMNSKGCKVYFGRWLIEGGPLVVLLDVGASAWALERWKGELWDTCNIGVPWYDRE
ANDAVLFGFLTTWFLGEFLAQSEEKPHVVAHFHEWLAGVGLCLCRARRLPVATIFTTHATLLGRYLCAGAVDFYNN
LENFNVDKEAGERQIYHRYCMERAAAHCAHVFTTVSQITAIEAQHLLKRKPDIVTPNGLNVKKFSAMHEFQNLHA
QSKARIQEFVRGHFYGHLDFNLDKTLYFFIAGRYEFSNKGADVFLEALARLNYLLRVNGSEQTVVAFFIMPARTNNF
NVETLKGQAVRKQLWDTANTVKEKFGRKLYESLLVGSLPDMNKMLDKEDFTMMKRAIFATQRQSFPPVCTHNM
LDDSSDPILTTIRRIGLFNSSADRVKVIFHPEFLSSTSPLLPVDYEEFVRGCHLGVFPSYYEPWGYTPAECTVMGIPSIST
NLSGFGCFMEEHIADPSAYGIYILDRRFRSLDDSCSQLTSFLYSFCQQSRRQRIIQRNRTERLSDLLDWKYLGRYYMS
ARHMALSKAFPEHFTYEPNEADAAQGYRYPRPASVPPSPSLSRHSSPHQSEDEEDPRNGPLEEDGERYDEDEEAAK
DRRNIRAPEWPRRASCTSSTSGSKRNSVDTATSSSLSTPSEPLSPTSSLGEERN

GYG (from GYS-complex):

SMTDQAFVTLTTNDAYAKGALVLGSSLKQHRTTRRLVVLATPQVSDSMRKVLETVFDEVIMVDVLDSGDSAHLTL MKRPELGVTLTKLHCWSLTQYSKCVFMDADTLVLANIDDLFDREELSAAPDPGWPDCFNSGVFVYQPSVETYNQL LHLASEQGSFDGGDQGILNTFFSSWATTDIRKHLPFIYNLSSISIYSYLPAFKVFGASAKVVHFLGRVKPWNYTYDPKT KSVKSEAHDPNMTHPEFLILWWNIFTTNVLPLLQQFGLVKDTCSYVNVEDVSGAISHLSLGEIPAMAQPFVSSEERK ERWEQGQADYMGADSFDNIKRKLDTYLQ

The part of the protein that is identical to the sequence of GYG-wt-0Glc is underlined. The *C*-terminal peptide of GYG that binds to GYS is in red.

GBE: donated by Dr Thomas McCorvie (Prof. Wyatt Yue group).

MAAPMTPAARPEDYEAALNAALADVPELARLLEIDPYLKPYAVDFQRRYKQFSQILKNIGENEGGIDKFSRGYESFG VHRCADGGLYCKEWAPGAEGVFLTGDFNGWNPFSYPYKKLDYGKWELYIPPKQNKSVLVPHGSKLKVVITSKSGEI LYRISPWAKYVVREGDNVNYDWIHWDPEHSYEFKHSRPKKPRSLRIYESHVGISSHEGKVASYKHFTCNVLPRIKGL GYNCIQLMAIMEHAYYASFGYQITSFFAASSRYGSPEELQELVDTAHSMGIIVLLDVVHSHASKNSADGLNMFDGT DSCYFHSGPRGTHDLWDSRLFAYSSWEVLRFLLSNIRWWLEEYRFDGFRFDGVTSMLYHHHGVGQGFSGDYSEY FGLQVDEDALTYLMLANHLVHTLCPDSITIAEDVSGMPALCSPISQGGGGFDYRLAMAIPDKWIQLLKEFKDEDW NMGDIVYTLTNRRYLEKCIAYAESHDQALVGDKSLAFWLMDAEMYTNMSVLTPFTPVIDRGIQLHKMIRLITHGLG GEGYLNFMGNEFGHPEWLDFPRKGNNESYHYARRQFHLTDDDLLRYKFLNNFDRDMNRLEERYGWLAAPQAYV SEKHEGNKIIAFERAGLLFIFNFHPSKSYTDYRVGTALPGKFKIVLDSDAAEYGGHQRLDHSTDFFSEAFEHNGRPYSL LVYIPSRVALILQNVDLPNAENLYFQSHHHHHHHHHHDYKDDDDK

The *C*-terminus of the protein construct contains a polyhistidine tag which is highlighted in yellow and a TEV-cleavable peptide which is highlighted in green.

4.4.2 Protein expression and characterisation

4.4.2.1 GYG-wt-0Glc

Expression in M226 competent cells

GYG-wt-0Glc expression was carried out based on protocols from a previous report.[11]

M226 cells containing GYG-wt plasmid and lacking UDP-glucose pyrophosphorylase activity were plated on TB-agar plates containing both kanamycin (50 μg/mL) and streptomycin (100 μg/mL) and these were incubated at 37 °C, overnight. A single colony was used to inoculate 10 mL TB media supplemented with kanamycin (50 μg/mL) and streptomycin (100 μg/mL) and the cells were grown at

37 °C, overnight, with shaking (200 rpm). Resultant cells were used to inoculate 1 L of same TB media and grown (37 °C, 180 rpm) to a final OD_{600} of 0.6. IPTG was added to a final concentration of 0.1 mM and the cells were incubated at 18 °C, overnight, with shaking (180 rpm). The next day, the cells were harvested through centrifugation (7000 x g, 7 minutes, 4 °C).

Cell lysis and purification

Cell pellet was suspended in lysis buffer (approx. 15 mL per L culture) with addition of cOmplete™ EDTA-free protease inhibitor cocktail. Lysozyme (approx. 5 mg per L culture) and DNaseI (approx. 1 mg per L culture) were added and the suspension was gently shaken at 4 °C for 60 minutes. Cells were then sonicated (40 % amplitude, 5 s on, 5 s off, 30 sec pause intervals, 3 cycles) and cell debris removed by centrifugation (29000 rpm, 4 °C, 20 minutes) and filtration (0.2 μm filter).

Lysis Buffer: 20 mM imidazole, 50 mM Na-HEPES, 0.5 M NaCl, 0.5 mM TCEP, 5 % glycerol, pH 7.5

Purification of His_6-tagged GYG-wt-0Glc

The clarified lysate was loaded onto a gravity flow Ni^{2+}-affinity column. The flowthrough was collected, and the column washed with lysis buffer (ca. 40 mL/mL resin). Protein was purified using a step gradient of imidazole (40 mM – 60 mM – 250 mM, protein elutes at 250 mM). Fractions were analysed by SDS-PAGE. For isolation of the His-tagged protein, fractions containing the desired protein were combined and dialysed against dialysis buffer (3 x 2L, 10000 MWCO, 4 °C, 2 h, 2 h then overnight). Protein was concentrated using a Vivaspin® 20 (10000 MWCO) concentrator to a concentration c > 1 mg/mL and used in the TEV-protease cleavage experiment.

Dialysis Buffer: 50 mM Na-HEPES, 0.5 M NaCl, 0.5 mM TCEP, 5 % glycerol, pH 7.5

Purification of non-His_6-tagged GYG-wt-0Glc

Following purification by Ni^{2+}-affinity chromatography, 10 mL His-tagged GYG-wt-0Glc obtained as described above were incubated with TEV-protease overnight, on ice. A 9:1 w/w ratio of substrate:TEV-protease was used. Then, the sample was passed over Ni^{2+}-affinity resin. The flowthrough was collected, and the resin was washed successively with dialysis buffer (4 mL/mL Ni^{2+}-affinity resin), lysis buffer (4 mL/mL Ni^{2+}-affinity resin), and then with lysis buffer containing increasing concentrations of imidazole (60 mM – 80 mM – 250 mM, 2 mL/mL Ni^{2+}-affinity resin). Fractions containing the non-His_6-tagged GYG-wt-0Glc were pooled and concentrated to a volume of 5 mL using a Vivaspin® 20, (10000 MWCO) concentrator, and then purified by gel filtration chromatography (S200 16/600 column), eluting with gel filtration buffer (50 mM Na-HEPES, 0.5 M NaCl, 0.5 mM TCEP, 5 %

glycerol, pH 7.5). Fractions containing GYG-wt-0Glc were pooled, concentrated (c > 1 mg/mL), flash frozen and stored at – 80 °C.

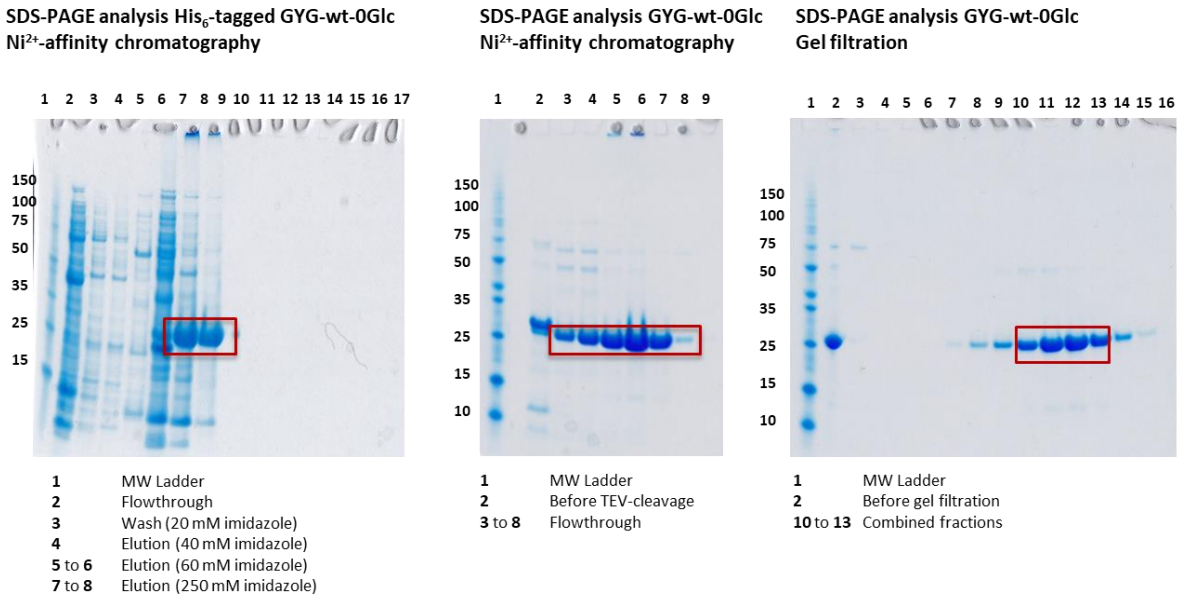


Figure SI 4.1. SDS-PAGE analyses of the GYG-wt-0Glc purification by Ni^{2+}-affinity chromatography and gel filtration. As expected from the calculated molecular weight, the size of the purified His_6-tagged GYG-wt-0Glc variant was about 32 kDa. Fractions containing the desired protein are shown in red frame after staining with Coomassie blue. Upon cleavage of the His_6-tag the size of the purified non-His_6-tagged GYG-wt-0Glc variant decreased to about 29.5 kDa.

4.4.2.2 GYG-Y195F

Transformation and expression in BL21(DE3) competent cells

BL21(DE3) cells were thawed on ice. A 50 µL aliquot was transferred to a pre-chilled 14 mL falcon tube. Plasmid GYG-Y195F (1.0 µL, ca. 80 ng/µL) was added, and the tube was placed on ice for 30 minutes, with occasional swirling. Cells were heat shocked (42 °C, 30 s), then immediately placed on ice for a further 2 minutes. 250 µL of SOC media was added, and cells were incubated at 37 °C, for 1 h, with shaking (200 rpm). Cells were plated on LB-agar plates containing kanamycin (50 µg/mL); 20 µL of the transformation mixture is typically ample to obtain many colonies. Plates were incubated at 37 °C overnight.

A single colony obtained from the above transformation was used to inoculate 10 mL LB media, supplemented with kanamycin (50 µg/mL). Cells were grown at 37 °C, overnight, with shaking (200 rpm). Resultant cells were used to inoculate 1 L of LB media. Cells were grown (37 °C, 180 rpm) to a final OD_{600} of 0.6-0.7. For each culture, IPTG was added to a final concentration of 0.5 mM. Cells were

cooled to 30 °C and incubated with shaking at 180 rpm for 2 h, then cooled to 18 °C and incubated overnight. After this time, cells were harvested through centrifugation (7000 x g, 7 minutes, 4 °C).

Cell lysis and purification

Cell pellets were resuspended in lysis buffer (approx. 15 mL per L culture) with addition of cOmplete™ EDTA-free protease inhibitor cocktail. Lysozyme (approx. 5 mg per L culture) and DNaseI (approx. 0.5 mg per L culture) were added and the suspension was gently shaken at 4 °C, for 30 minutes. Cells were then sonicated (35 % amplitude, 5 s on, 10 s off, 12 cycles) and cell debris was removed by centrifugation (50000 x g, 4 °C, 20 minutes).

Lysis Buffer: 20 mM imidazole, 50 mM Na-HEPES, 0.5 M NaCl, 0.5 mM TCEP, 5 % glycerol, pH 7.5

Purification of His_6-tagged GYG-Y195F

The clarified lysate was loaded onto a gravity flow Ni^{2+}-affinity column. The flowthrough was collected, and the column washed with lysis buffer (ca. 40 mL per mL resin). Protein was purified using a step gradient of imidazole (20 mM – 40 mM – 60 mM – 250 mM imidazole). Fractions were analysed by SDS-PAGE and the ones containing the desired protein were combined and dialysed against dialysis buffer (2 x 2L, 10000 MWCO, 4 °C, 2 h then overnight). Protein was concentrated using a Vivaspin® 20 (10000 MWCO) concentrator to a concentration c > 1 mg/mL, flash frozen and stored at - 80 °C.

Dialysis Buffer: 10 mM Hepes pH 7.5, 500 mM NaCl, 1 mM DTT, 1 mM Na-EDTA, 5 % glycerol

Purification of non-His_6-tagged GYG-Y195F

Following purification by Ni^{2+}-affinity chromatography as described above, fractions containing GYG were combined and concentrated to a volume of 5 mL, then purified by gel filtration chromatography (S200 16/600 column), eluting with gel filtration buffer. Fractions containing the protein were pooled and concentrated to 9 mg/mL. For His_6-tag cleavage, protein was incubated with TEV-protease overnight, on ice (9:1 w/w ratio of substrate:TEV-protease was used). The sample was passed over Ni^{2+}-affinity resin. The flowthrough was collected, and resin was washed successively with gel filtration buffer (4 mL/mL Ni^{2+}-affinity resin), lysis buffer (4 mL/mL Ni^{2+}-affinity resin), and then with lysis buffer containing increasing concentrations of imidazole (60 mM – 80 mM – 250 mM, 2 mL/mL Ni^{2+}-affinity resin each). Fractions containing non-His_6-tagged GYG-Y195F were pooled, flash frozen and stored at - 80 °C.

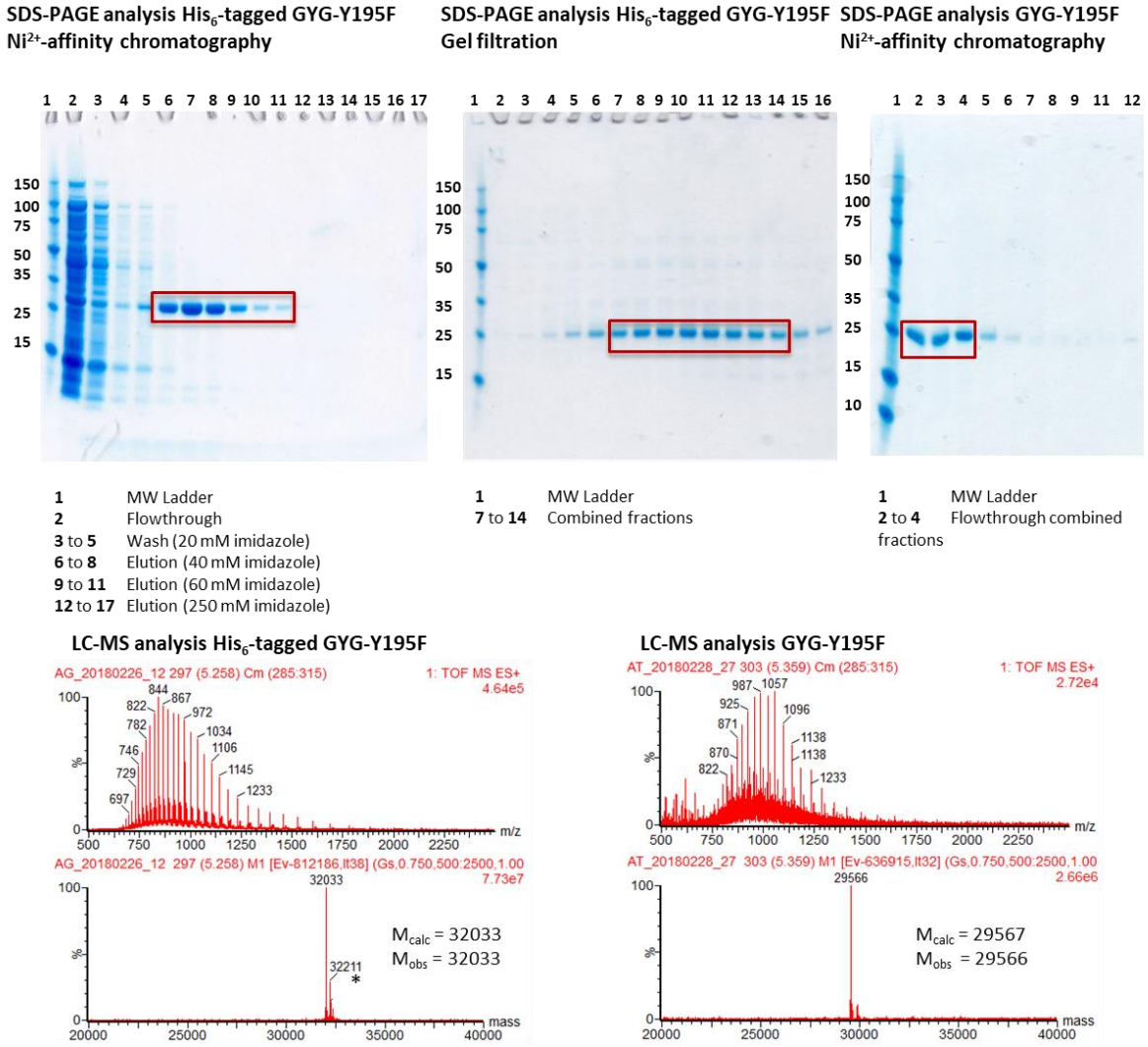


Figure SI 4.2. SDS-PAGE analyses of the GYG-Y195F purification by Ni^{2+}-affinity and gel filtration, and characterization of the His_6-tagged and non-His_6-tagged protein by intact protein LC-MS. As expected from the calculated MW, the size of the purified His_6-tagged GYG-Y195F mutant was about 32 kDa. Fractions containing the desired protein are shown in red frame after staining with Coomassie blue. The molecular mass of the enzyme before cleavage of the His_6-tag is 32033 Da as determined by intact protein LC-MS analysis. Note that the +178 Da peak is attributable to the well-known *N*-terminal gluconoylation, a commonly observed feature during expression of His_6-tagged fusion proteins.[61] The molecular mass of the non-His_6-tagged GYG is 29567 Da as determined by intact protein LC-MS analysis.

4.4.3 Autoglucosylation and glucosylation assays

4.4.3.1 Autoglucosylation of GYG-wt-0Glc in the presence of GYG-Y195F

Proteins were buffer exchanged into pH 7.5 25 mM Na-HEPES by PD SpinTrap G-25® and diluted using the same buffer to a concentration of 0.2 mg/mL. Then, 2 reaction mixtures were prepared as follows and incubated for 1 hour at 37 °C with shaking (500 rpm).

a. GYG-wt-0Glc:GYG-Y195F 1:1
    - 32 µL GYG-wt-0Glc solution (c = 0.2 mg/mL in 25 mM Na-HEPES pH 7.5 buffer)
    - 32 µL GYG-Y195F solution (c = 0.2 mg/mL in 25 mM Na-HEPES pH 7.5 buffer)
    - 8 µL $MnCl_2$ (10 mM solution in ddH_2O)
    - 8 µL UDP-glucose (10 mM solution in ddH_2O)

b. GYG-wt-0Glc:GYG-Y195F 1:0
    - 32 µL GYG-wt-0Glc solution (c = 0.2 mg/mL in 25 mM Na-HEPES pH 7.5 buffer)
    - 32 µL 25 mM Na-HEPES pH 7.5 buffer
    - 8 µL $MnCl_2$ (10 mM solution in ddH_2O)
    - 8 µL UDP-glucose (10 mM solution in ddH_2O)

After the desired time points (20 s, 40 s, 80 s, 120 s, 210 s, 300 s, 600 s, 900 s, 1800 s, 2700 s, 3600 s), aliquots were removed and quenched through the addition of Na-EDTA to a final concentration of 20 mM, and immediately placed on ice.

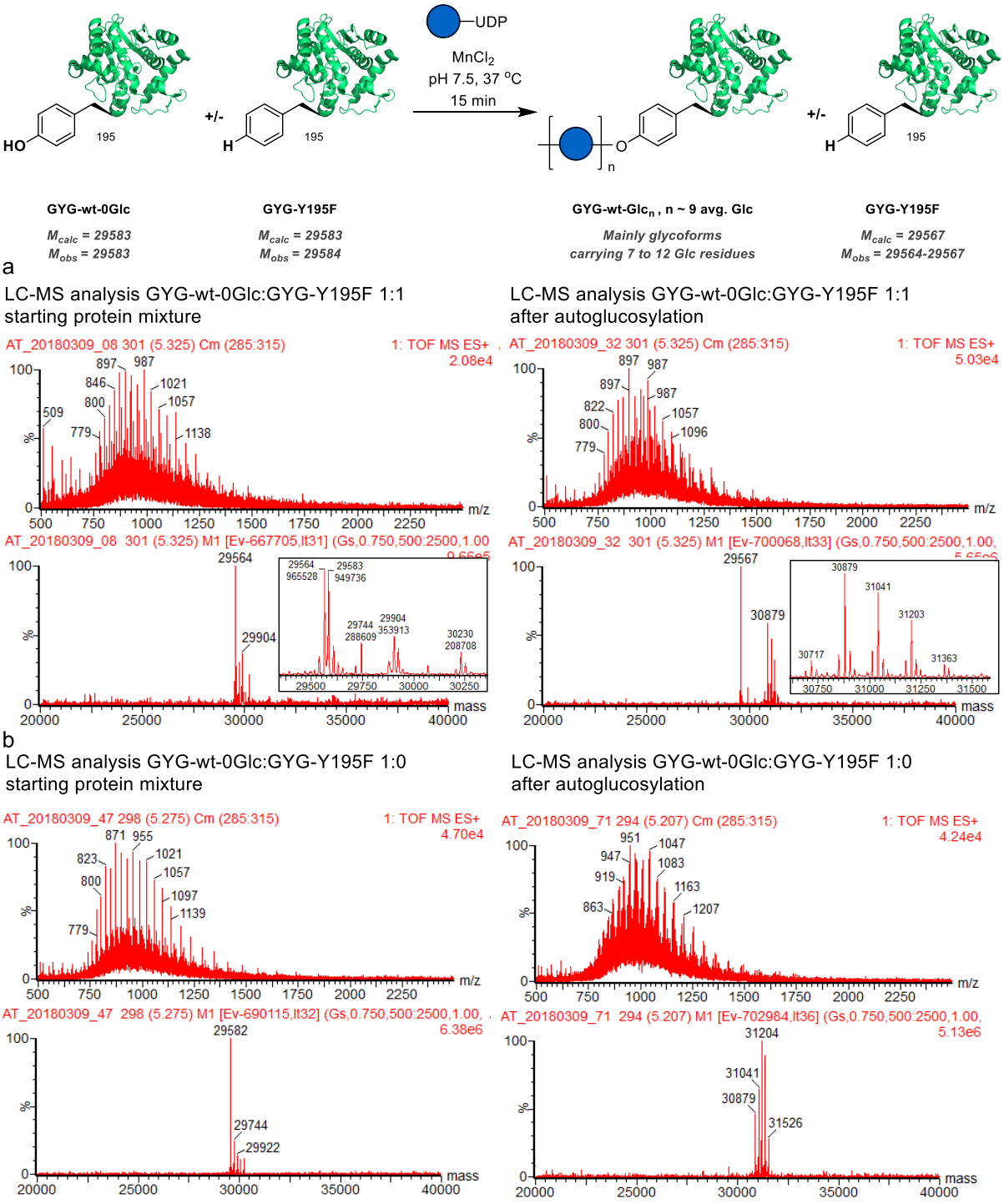


Figure SI 4.3. Autoglucosylation of GYG-wt-0Glc in the presence/absence of GYG-Y195F. LC-MS analyses showed that the addition of GYG-Y195F to the reaction mixture of GYG-wt-0Glc, UDP-Glc and $MnCl_2$ did not cause an increase in the glucosylation level of GYG-wt-nGlc but on the contrary, it slightly suppressed the autoglucosylation process.

4.4.3.2 Autoglucosylation of GYG-wt-0Glc: LC-MS and iSCAMS analyses of GYG-wt-0Glc and GYG-wt-nGlc (n = 9-11 avg. Glc residues)

Protein was buffer exchanged into 25 mM Na-HEPES pH 7.5 buffer by PD SpinTrap G-25® and diluted using the same buffer to a concentration of 0.2 mg/mL. Then, 3 reaction mixtures were prepared as follows and incubated for 1 hour at 37 °C with shaking (500 rpm).

c. Blank:
    - 80 µL buffer 25 mM Na-HEPES, pH 7.5
    - 10 µL $MnCl_2$ (10 mM solution in ddH_2O)
    - 10 µL ddH_2O
d. Protein mixture:
    - 80 µL GYG-wt-0Glc solution (c = 0.2 mg/mL in 25 mM Na-HEPES pH 7.5 buffer)
    - 10 µL $MnCl_2$ (10 mM solution in ddH_2O)
    - 10 µL ddH_2O
e. Reaction mixture:
    - 80 µL GYG-wt-0Glc solution (c = 0.2 mg/mL in 25 mM Na-HEPES pH 7.5 buffer)
    - 10 µL $MnCl_2$ (10 mM solution in ddH_2O)
    - 10 µL UDP-Glucose (10 mM solution in ddH_2O)

After 1 hour, these were filtered through PD SpinTrap G-25® columns into 25 mM Na-HEPES pH 7.5 buffer to remove small molecules reagents (using 80 µL buffer to top-up). Successful autoglucosylation was confirmed by LC-MS analysis.

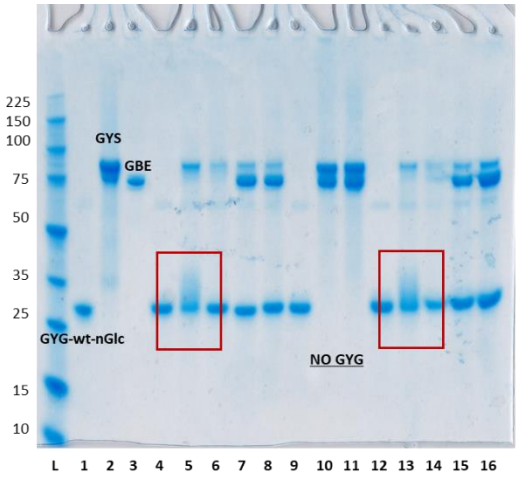
iSCAMS:

Mass photometry measurements were done as described.[50] Briefly, borosilicate microscope coverslips were cleaned by sequential washes in Milli-Q H_2O, isopropanol and Milli-Q H_2O and dried under a nitrogen gas flow. Sample chambers were assembled using silicone gaskets (CultureWell™ reusable gasket, Grace Bio-Labs). Assembled Coverslips were placed on the sample stage and a single gasket was filled with 10 µL phosphate buffered saline (PBS) to find focus. Then, the 3 samples were diluted 1:100 with 25 mM Na-HEPES pH 7.5 buffer (27 nM final protein concentration) prior to analysis. For each measurement, 5 µL of the 10 µL of Blank in the well were replaced with 5 µL Protein/Reaction mixture followed by gentle pipette mixing. Five measurements were performed for each of the 3 mixtures described above. Data were recorded for 90 seconds. MP settings were as follows: frame averaging = 10, frame rate = 955 Hz and 4x4 pixel binning. Data was analysed using Discovery MP v2.3dev7 (Refeyn Ltd).

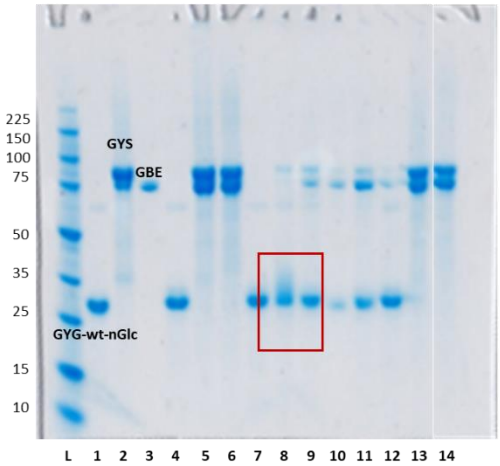
4.4.3.3 Glucosylation of GYG-wt-nGlc (n = 9-11 avg. Glc residues) with 10 mM UDP-Glucose

Protein GYG-wt-0Glc was buffer exchanged into 25 mM Na-HEPES pH 7.5 buffer by PD SpinTrap G-25® and diluted using the same buffer to a concentration of 0.2 mg/mL. Then, to 480 µL protein solution 60 µL $MnCl_2$ (10 mM solution in ddH_2O) and 60 µL UDP-Glucose (10 mM solution in ddH_2O) were added and the reaction mixture was incubated for 1 hour at 37 °C with shaking (500 rpm) to obtain GYG-wt-nGlc (n = 9-11 avg. Glc residues) (Figure **4.6** and Figure **4.7**, A). Proteins GYS and GBE were buffer-exchanged into 25 mM Na-HEPES pH 7.5 buffer by PD SpinTrap G-25 and diluted using the same buffer to a concentration of 0.6 mg/mL. Then, eight reaction mixtures (120 µL scale) were prepared as follows and incubated at 37 °C with shaking (500 rpm) for 16 hours. Glucosylation was monitored by LC-MS (Figure **4.7**, B to G) and SDS-PAGE analyses (Figure **4.7**, B to I) after 1, 2, 4 and 16 hours.

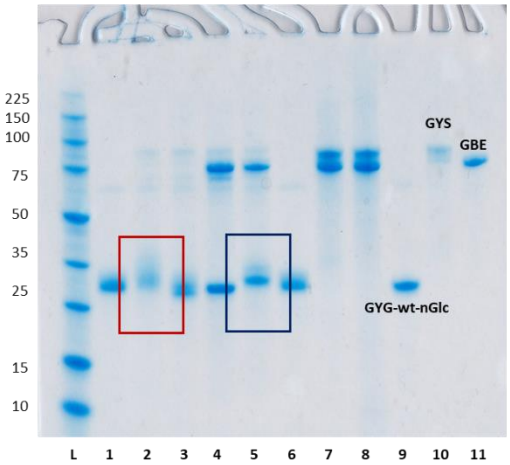
- 60 µl GYG-wt-nGlc + 24.4 µL buffer 25 mM Na-HEPES, pH 7.5 + 12 µL UDP-Glucose (100 mM solution in buffer) + 3.6 µL G6P (100 mM solution in buffer) + 20 µL GYS (0.6 mg/mL) (Figure **4.6** and Figure **4.7**, B)
- 60 µL GYG-wt-nGlc + 44.4 µL buffer 25 mM Na-HEPES, pH 7.5 + 12 µL UDP-Glucose (100 mM solution in buffer) + 3.6 µL G6P (100 mM solution in buffer) (Figure **4.6** and Figure **4.7**, C)
- 60 µL GYG-wt-nGlc + 48 µL buffer 25 mM Na-HEPES, pH 7.5 + 12 µL UDP-Glucose (100 mM solution in buffer) (Figure **4.6** and Figure **4.7**, D)
- 60 µL GYG-wt-nGlc + 28 µL buffer 25 mM Na-HEPES, pH 7.5 + 12 µL UDP-Glucose + 20 µL GYS (0.6 mg/mL) (Figure **4.7**, E)
- 60 µL GYG-wt-nGlc + 4.4 µL buffer 25 mM Na-HEPES, pH 7.5 + 12 µL UDP-Glucose (100 mM solution in buffer) + 3.6 µL G6P (100 mM solution in buffer) + 20 µL GYS (0.6 mg/mL) + 20 µL GBE (0.6 mg/mL) (Figure **4.7**, F)
- 60 µL GYG-wt-nGlc + 8 µL buffer 25 mM Na-HEPES, pH 7.5 + 12 µL UDP-Glucose (100 mM solution in buffer) + 20 µL GYS (0.6 mg/mL) + 20 µL GBE (0.6 mg/mL) (Figure **4.7**, G)
- 80 µL GYS (0.6 mg/mL) + 20 µL GBE (0.6 mg/mL) + 12 µL UDP-Glucose (100 mM solution in buffer) + 3.6 µL G6P (100 mM solution in buffer) + 4.4 µL buffer 25 mM Na-HEPES, pH 7.5 (Figure **4.7**, H)
- 80 µL GYS (0.6 mg/mL) + 20 µL GBE (0.6 mg/mL) + 12 µL UDP-Glucose (100 mM solution in buffer) + 4.4 µL buffer 25 mM Na-HEPES, pH 7.5 (Figure **4.7**, I)



1. **A** (GYG-wt-nGlc, **2**)
2. GYS (0.4 mg/ml)
3. GBE (0.1 mg/ml)
4. **C** - 1 hour
5. **B** – 1 hour
6. E – 1 hour
7. F – 1 hour
8. G – 1 hour
9. **D** – 1 hour
10. H – 1 hour
11. I – 1 hour
12. **C** – 2 hours
13. **B** – 2 hours
14. E – 2 hours
15. F – 2 hours
16. G – 2 hours



1. **A** (GYG-wt-nGlc, **2**)
2. GYS (0.4 mg/ml)
3. GBE (0.1 mg/ml)
4. **D** - 2 hours
5. H – 2 hours
6. I – 2 hours
7. **C** – 4 hours
8. **B** – 4 hours
9. E – 4 hours
10. F – 4 hours
11. G – 4 hours
12. **D** – 4 hours
13. H – 4 hours
14. I – 4 hours



1. **C** – 16 hours
2. **B** – 16 hours
3. E – 16 hours
4. F – 16 hours
5. G – 16 hours
6. **D** – 16 hours
7. H – 16 hours
8. I – 16 hours
9. **A** (GYG-wt-nGlc, **2**)
10. GYS (0.1 mg/ml)
11. GBE (0.1 mg/ml)

Figure SI 4.4. SDS-PAGE analyses of the GYG-wt-nGlc (n = 9-11 avg. Glc residues) glucosylation assay in the presence of 10 mM UDP-Glucose. Noteworthy, GYG-wt-nGlc (n = 9-11 avg. Glc residues) underwent extensive glucosylation upon incubation with GYS and G6P and the newly formed GYG-wt-(n+m)Glc (n+m = 19-21 avg. Glc residues) glycoforms migrated as higher MW species than GYG-wt-nGlc (red square, Figure 4.6/4.7 and Figure SI 4.5). Large GYG-wt-(n+m)Glc glycoforms were also generated when GYG-wt-nGlc was incubated with both GYS and GBE, in absence of G6P but in this case, the obtained glycoform distribution appeared as being composed of two different sub-populations (blue square, Figure 4.7, G and Figure SI 4.8).

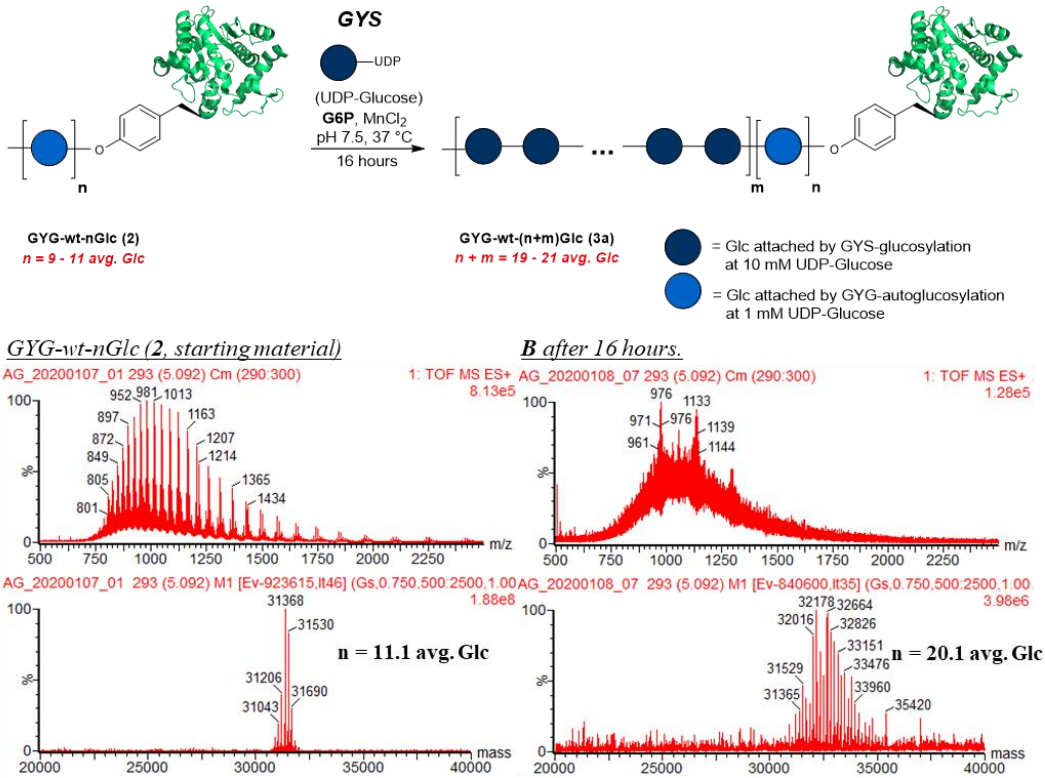


Figure SI 4.5. Glucosylation of GYG-wt-nGlc (n = 9-11 avg. Glc residues) by GYS, using 10 mM UDP-Glucose and 3 mM G6P (Figure 4.6, B). LC-MS analyses showed that the allosterically activated GYS is able to further glucosylate GYG-wt-nGlc (n = 9-11 avg. Glc residues) to give GYG-wt-(n+m)Glc (n+m = 19-21 avg. Glc residues), a heterogeneous mixture of GYG-species that has a much wider glycoform distribution than the starting material.

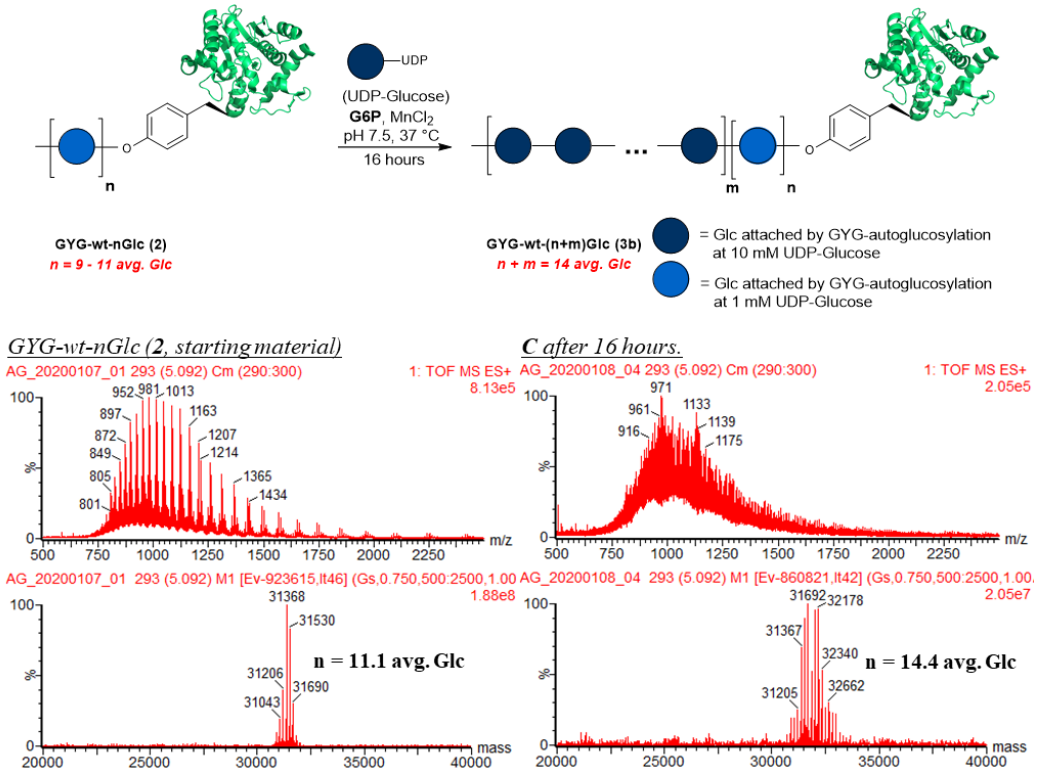


Figure SI 4.6. Autoglucosylation of GYG-wt-nGlc (n = 9-11 avg. Glc residues) using 10 mM UDP-Glucose and 3 mM G6P, in the absence of GYS (Figure 4.6, C). At higher concentrations of UDP-Glc, GYG continues to autoglucosylate its polyglucosidic chain and after 16 hours GYG-wt-nGlc (n = 14.4 avg. Glc residues) is obtained.

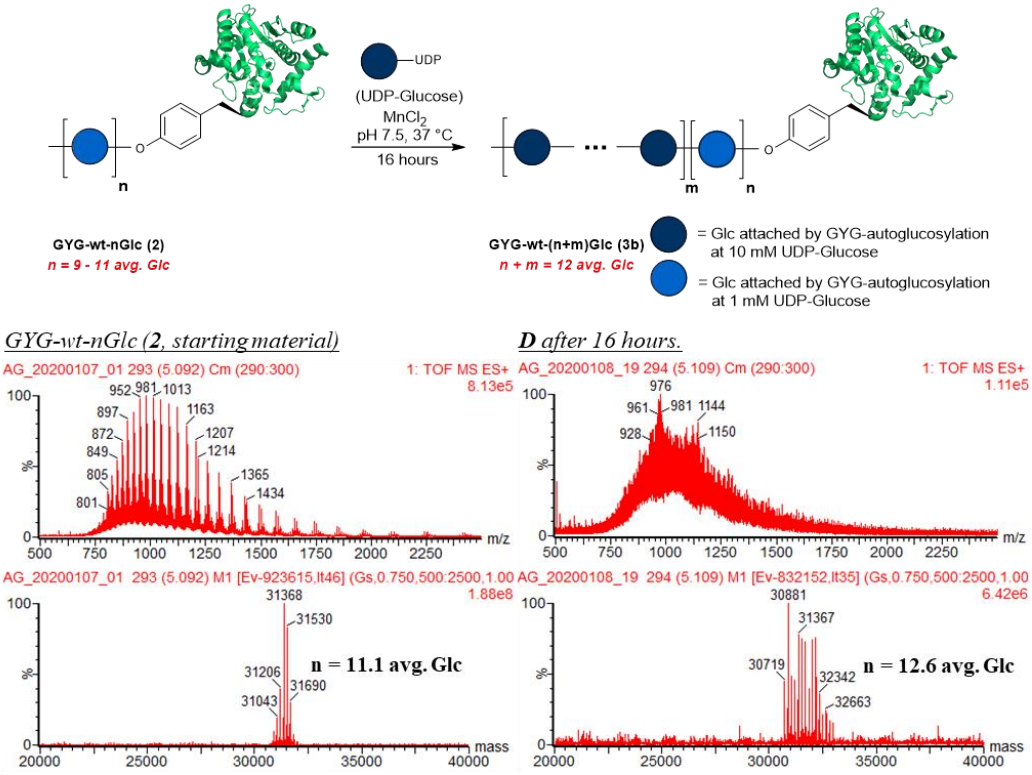


Figure SI 4.7. Autoglucosylation of GYG-wt-nGlc (n = 9-11 avg. Glc residues) using 10 mM UDP-Glucose, in the absence of both GYS and G6P (Figure 4.6, D). At higher concentrations of UDP-Glc, GYG continues to autoglucosylate its polyglucosidic chain for the first 4 hours but after a longer incubation time (16 hours) shorter GYG-wt-(n+m)Glc (n + m = 12.6 avg. Glc residues) glycoforms are obtained.

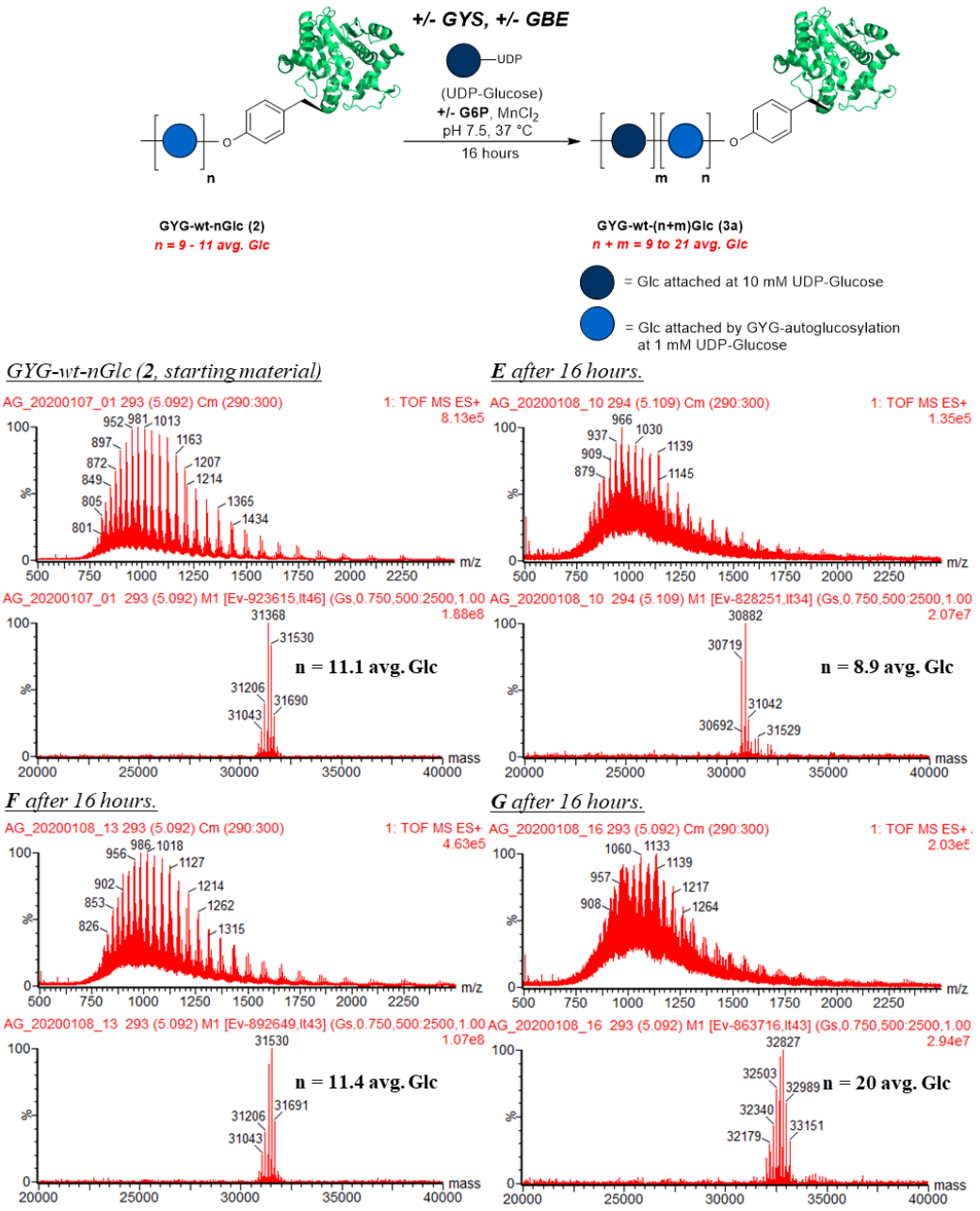


Figure SI 4.8. Glucosylation of GYG-wt-nGlc (n = 9-11 avg. Glc residues) using 10 mM UDP-Glucose and reaction conditions described in Figure 4.7 entries E, F and G. At higher concentrations of UDP-Glc, GYS is not able to further glucosylate GYG-wt-nGlc (n = 9-11 avg. Glc residues) in absence of its allosteric activator G6P. Instead, autoglucosylation of GYG-wt-nGlc is favoured by the high UDP-Glc concentration during the first 4 hours and generates GYG-wt-(m+n)Glc (n+m = 17 avg. Glc residues), but after this initial phase extensive hydrolysis of the GYG glucosidic chain can be observed and much shorter GYG-wt-nGlc (n = 8.9 avg. Glc residues) glycoforms are being formed (**E**). Surprisingly though, addition of GBE to a reaction mixture containing 10 mM UDP-Glc and GYS activated by G6P ensures that the glucosylation degree of GYG-wt-nGlc remains constant over time (**F**). Moreover, the addition of GBE to a reaction mixture containing GYS, GBE and 10 mM UDP-Glc in absence of G6P leads to the formation of GYG-wt-(m+n)Glc (n+m = 20 avg. Glc residues) a mixture of glycoforms that appeared composed of two different sub-populations (**G**). (Note the low abundance of the longer GYG-species and see the Additional figure **SI 4.16** for a better visualization of these glycoforms).

4.4.4 Hydrolysis assays

4.4.4.1 Hydrolytic activity observed in the autoglucosylated GYG-wt-nGlc (n = 9-11 avg. Glc residues)

GYG-wt-0Glc was buffer exchanged into 25 mM Na-HEPES pH 7.5 buffer by PD SpinTrap G-25® and diluted using the same buffer to a concentration of 0.2 mg/mL. Then, 3 x 300 µL reaction mixtures were set up. In each case, 240 µL protein solution were incubated with 30 µL $MnCl_2$ (10 mM solution in ddH_2O, to have a final concentration of 1 mM) and 30 µL UDP-Glucose (10 mM solution in ddH_2O, to have a final concentration of 1 mM). These were incubated for 1h at 37 °C with shaking (500 rpm) and then combined and concentrated to less than 500 µL using a Vivaspin 500 concentrator (10 kDa MWCO). The protein solution was loaded onto a Superdex 75 10/300 analytical column and eluted with 25 mM Na-HEPES pH 7.5 buffer (0.3 mL/min flow). Obtained fractions (0.5 mL each) were analysed by SDS-PAGE (Figure SI 4.9). The ones containing desired GYG-wt-nGlc (n = 9-11 avg. Glc residues) were combined and concentrated using a Vivaspin 500® concentrator (10 kDa MWCO) to obtain 185 µL protein solution, c=0.27 mg/mL. It was then diluted to a concentration of 0.2 mg/mL with 25 mM Na-HEPES pH 7.5 buffer, seven reaction mixtures were set up as follows and incubated at 37 °C with shaking (500 rpm) for 18 hours. Hydrolysis was monitored by LC-MS analyses after 1, 2, 4, 6, 8 and 18 hours (Figure SI 4.10). This experiment was performed in duplicate and the representative LC-MS spectra presented below were generated from one of the replicates.

1. Control (buffer):
    - 30 µL GYG-wt-nGlc
    - 30 µL 25 mM Na-HEPES, pH 7.5 buffer
2. 0.1 mM UDP:
    - 30 µL GYG-wt-nGlc
    - 0.3 µL UDP (20 mM in 25 mM Na-HEPES, pH 7.5 buffer)
    - 29.7 µL 25 mM Na-HEPES, pH 7.5 buffer
3. 0.5 mM UDP:
    - 30 µL GYG-wt-nGlc
    - 1.5 µL UDP (20 mM in 25 mM Na-HEPES, pH 7.5 buffer)
    - 28.5 µL 25 mM Na-HEPES, pH 7.5 buffer
4. 2.5 mM UDP:
    - 30 µL GYG-wt-nGlc
    - 7.5 µL UDP (20 mM in 25 mM Na-HEPES, pH 7.5 buffer)

- 22.5 µL 25 mM Na-HEPES, pH 7.5 buffer

5. 5 mM UDP:
    - 30 µL GYG-wt-nGlc
    - 15 µL UDP (20 mM in 25 mM Na-HEPES, pH 7.5 buffer)
    - 15 µL 25 mM Na-HEPES, pH 7.5 buffer
6. 10 mM UDP:
    - 30 µL GYG-wt-nGlc
    - 30 µL UDP (20 mM in 25 mM Na-HEPES, pH 7.5 buffer)
7. 5 mM UDP & $MnCl_2$:
    - 30 µL GYG-wt-nGlc
    - 15 µL UDP (20 mM in 25 mM Na-HEPES, pH 7.5 buffer)
    - 1 µL $MnCl_2$ (100 mM in 25 mM Na-HEPES, pH 7.5 buffer)
    - 14 µL 25 mM Na-HEPES, pH 7.5 buffer

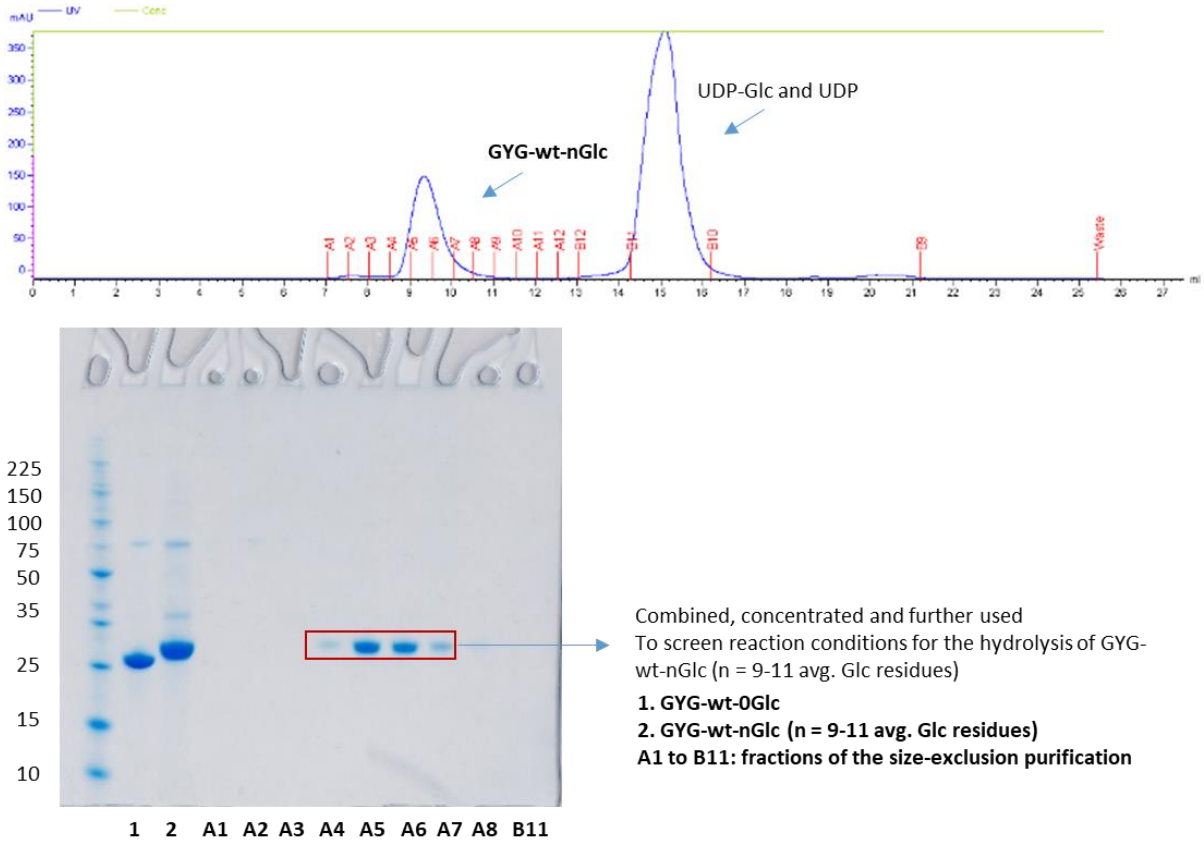


Figure SI 4.9. Purification of GYG-wt-nGlc (n = 9-11 avg. Glc residues): chromatogram and SDS-PAGE analysis.

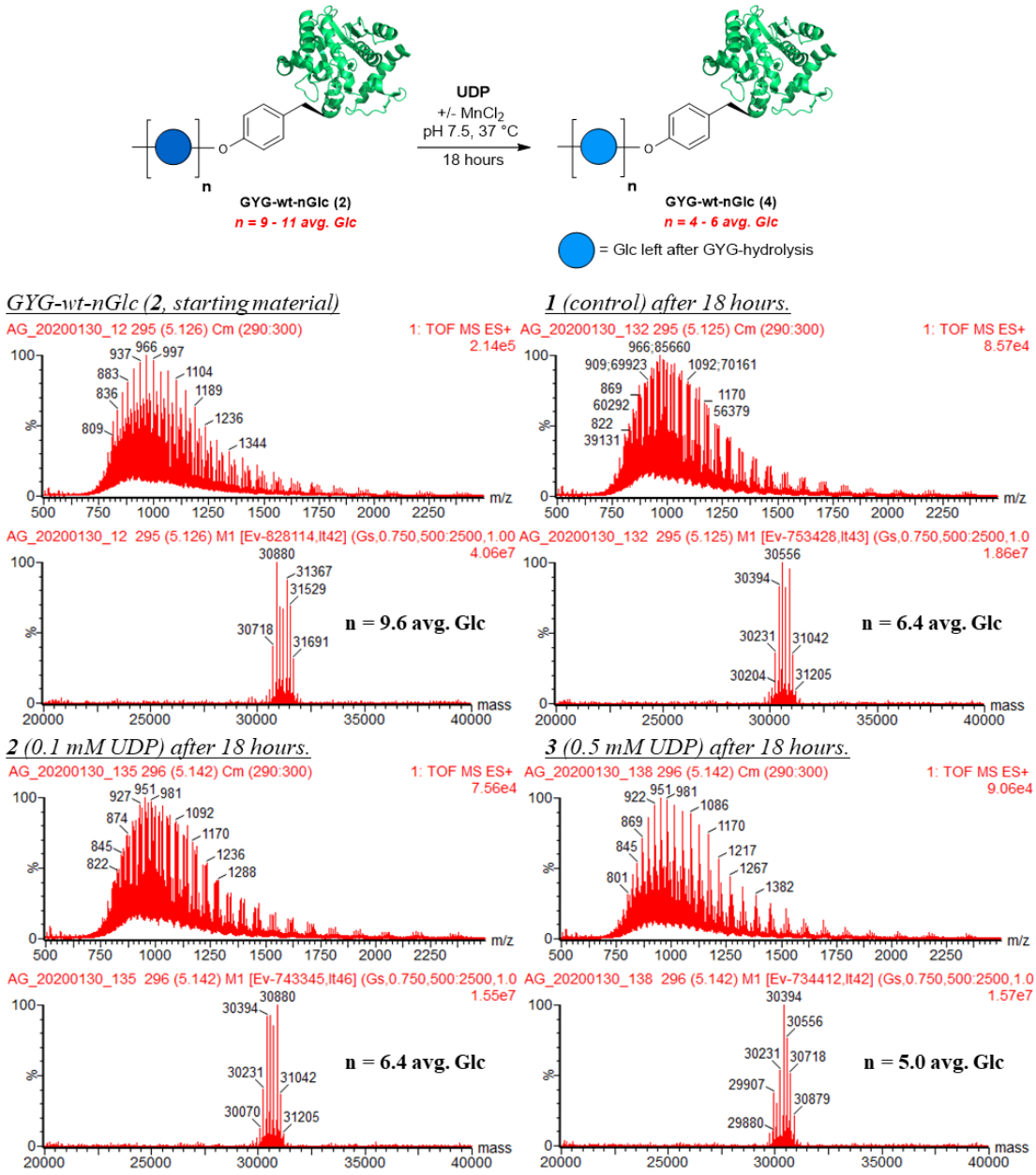


Figure SI 4.10. Hydrolysis of GYG-wt-nGlc (n = 9-11 avg. Glc residues) to GYG-wt-nGlc (n = 4-6 avg. Glc residues). LC-MS analyses showed that in absence of UDP-Glucose the polyglucosidic chain of GYG-wt-nGlc (n = 9-11 avg. Glc residues) underwent hydrolysis. This generated shorter GYG-wt-nGlc (n = 4-6 avg. Glc residues) species in all the reaction mixtures, either supplemented with buffer (1), 0.1 mM UDP (2) or 0.5 mM UDP (3).

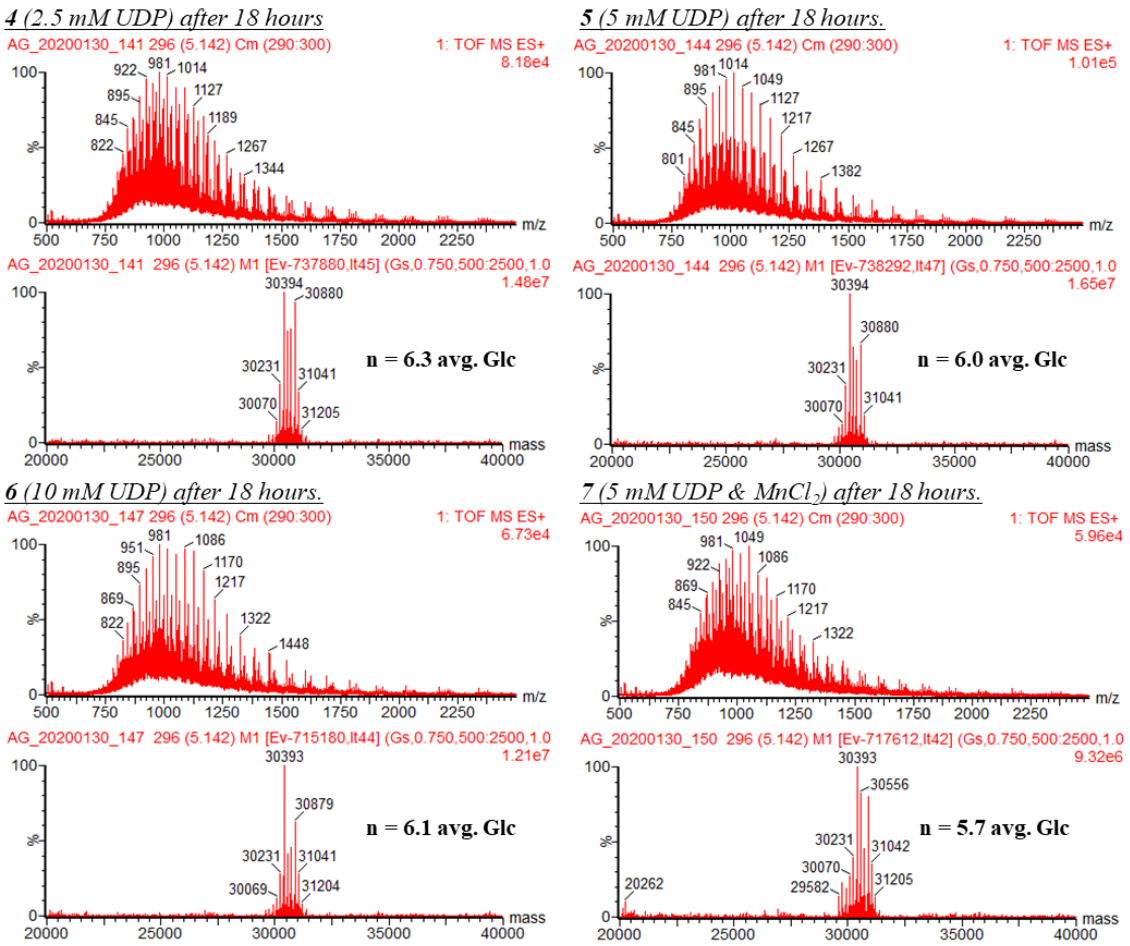


Figure SI 4.11. Hydrolysis of GYG-wt-nGlc (n = 9-11 avg. Glc residues) to GYG-wt-nGlc (n = 4-6 avg. Glc residues). LC-MS analyses showed that in absence of UDP-Glucose the polyglucosidic chain of GYG-wt-nGlc (n = 9-11 avg. Glc residues) underwent hydrolysis. This generated shorter GYG-wt-nGlc (n = 4-6 avg. Glc residues) species in all the reaction mixtures, either supplemented with 2.5 mM (4), 5 mM (5), 10 mM UDP (6) or 5 mM UDP and $MnCl_2$ (7).

4.4.4.2 Hydrolytic activity observed in the GYS-extended GYG-wt-(n+m)Glc (n+m = 19-21 avg. Glc residues)

GYG-wt-0Glc was buffer exchanged into 25 mM Na-HEPES pH 7.5 buffer by PD SpinTrap G-25® and diluted using the same buffer to a concentration of 0.2 mg/mL. Then, 3 x 300 µL reaction mixtures were set up. In each case, 240 µL protein solution were incubated with 30 µL $MnCl_2$ (10 mM solution in ddH_2O, to have a final concentration of 1 mM) and 30 µL UDP-Glucose (10 mM solution in ddH_2O, to have a final concentration of 1 mM). These were incubated for 1 hour at 37 °C with shaking (500 rpm). Then, they were combined and 750 µL obtained GYG-wt-nGlc (n = 9-11 avg. Glc residues) protein solution was incubated with 150 µL UDP-Glucose (100 mM solution in 25 mM Na-HEPES, pH 7.5 buffer to have a final concentration of 10 mM), 250 µL GYS (0.6 mg/mL in 25 mM Na-HEPES, pH 7.5 buffer),

45 µL glucose-6-phosphate (100 mM solution in 25 mM Na-HEPES, pH 7.5 buffer to have a final concentration of 3 mM) and 305 µL 25 mM Na-HEPES, pH 7.5 buffer. The reaction mixture was incubated for 2 hours at 37 °C with shaking (500 rpm) and then concentrated to less than 500 µL using a Vivaspin 500® concentrator (10 kDa MWCO). The protein solution was loaded onto a Superdex 75 10/300 analytical column and eluted with 25 mM Na-HEPES pH 7.5 buffer (0.3 mL/min flow). Obtained fractions (0.5 mL each) were analyzed by SDS-PAGE (Figure SI 4.12). The ones containing desired GYG-wt-(n+m)Glc (n+m = 19-21 avg. Glc residues) were combined and concentrated using a Vivaspin 500® concentrator (10 kDa MWCO) to obtain 85 µL protein solution, c=0.36 mg/mL. It was then diluted to a concentration of 0.2 mg/mL with 25 mM Na-HEPES pH 7.5 buffer, five reaction mixtures were set up as follows and incubated at 37 °C with shaking (500 rpm) for 20 hours. Hydrolysis was monitored by LC-MS analyses after 1, 2, 4, 6, 8 and 20 hours.

1. Control (buffer):
    - 30 µL GYG-wt-(n+m)Glc
    - 30 µL 25 mM Na-HEPES, pH 7.5 buffer
2. 0.5 mM UDP:
    - 30 µL GYG-wt-(n+m)Glc
    - 1.5 µL UDP (20 mM in 25 mM Na-HEPES, pH 7.5 buffer)
    - 28.5 µL 25 mM Na-HEPES, pH 7.5 buffer
3. 5 mM UDP:
    - 30 µL GYG-wt-(n+m)Glc
    - 15 µL UDP (20 mM in 25 mM Na-HEPES, pH 7.5 buffer)
    - 15 µL 25 mM Na-HEPES, pH 7.5 buffer
4. 10 mM UDP:
    - 30 µL GYG-wt-(n+m)Glc
    - 30 µL UDP (20 mM in 25 mM Na-HEPES, pH 7.5 buffer)
5. 5 mM UDP & $MnCl_2$:
    - 30 µL GYG-wt-(n+m)Glc
    - 15 µL UDP (20 mM in 25 mM Na-HEPES, pH 7.5 buffer)
    - 1 µL $MnCl_2$ (100 mM in 25 mM Na-HEPES, pH 7.5 buffer)

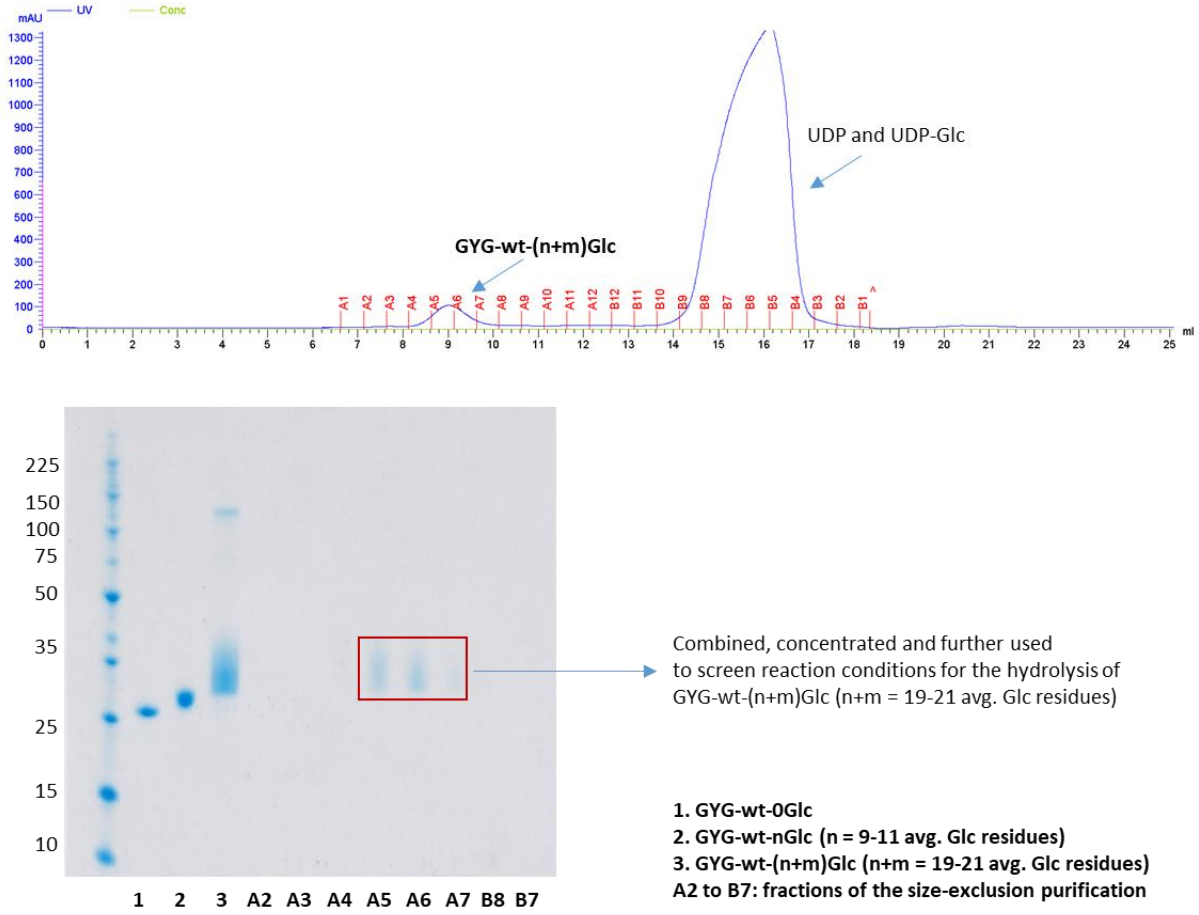


Figure SI 4.12. Purification of GYG-wt-(n+m)Glc (n+m = 19-21 avg. Glc residues): chromatogram and SDS-PAGE analysis.

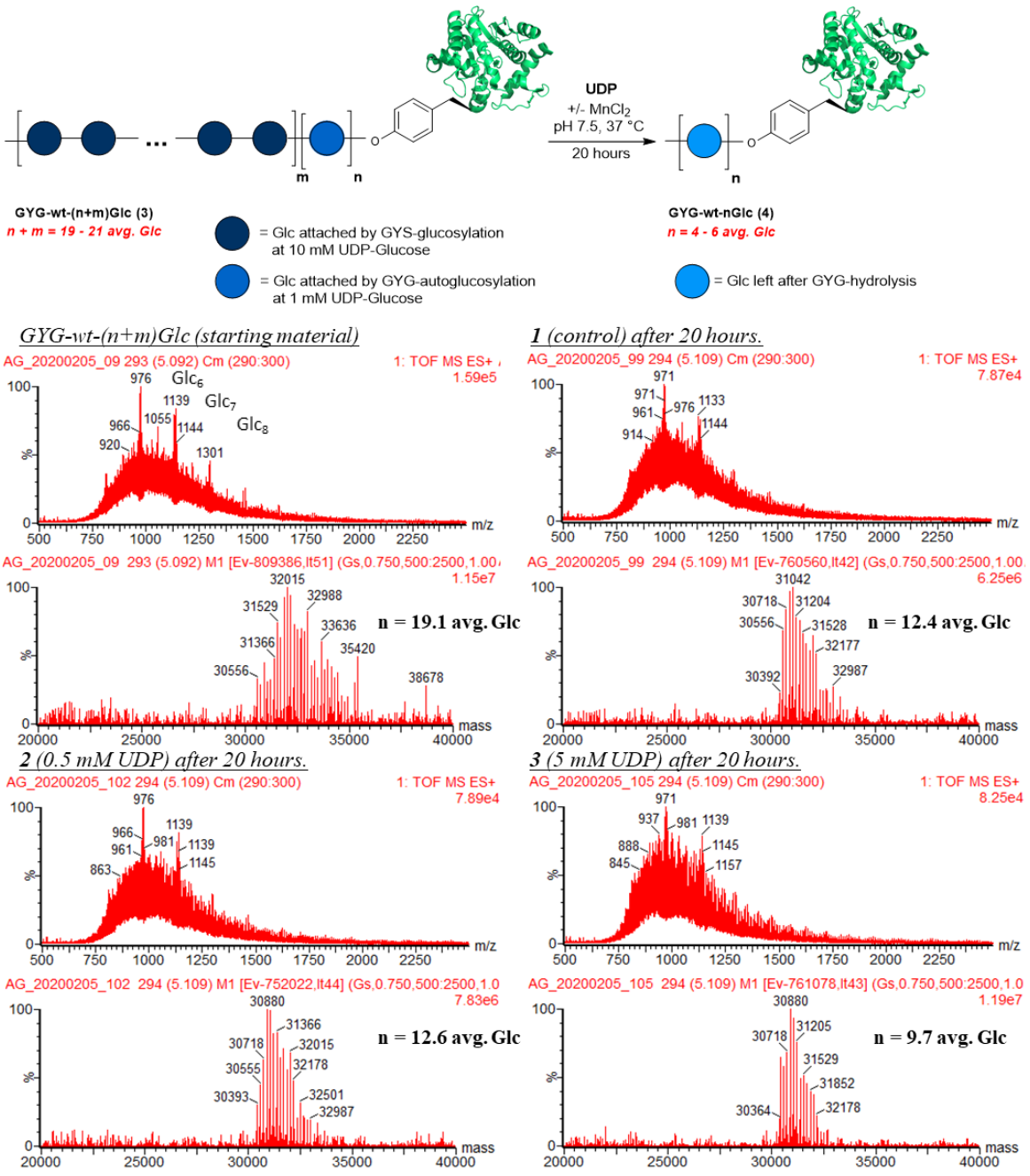


Figure SI 4.13. Hydrolysis of GYG-wt-(n+m)Glc (n = 19-21 avg. Glc residues) to GYG-wt-nGlc (n = 4-6 avg. Glc residues). In absence of UDP-Glucose, GYG-wt-(n+m)Glc (n = 19-21 avg. Glc residues) underwent hydrolysis of its polyglucosidic chain when incubated alone (1), with 0.5 mM UDP (2) or 5 mM UDP (3). This was highly accelerated if both UDP and $MnCl_2$ were added and consequently, trimmed glycoforms GYG-wt-nGlc (n = 4-6 avg. Glc residues) were formed. Molecular masses corresponding to oxonium ions of the oligosaccharides Glc_6, Glc_7 and Glc_8 were readily identified in the GYG-wt-(n+m)Glc starting material.

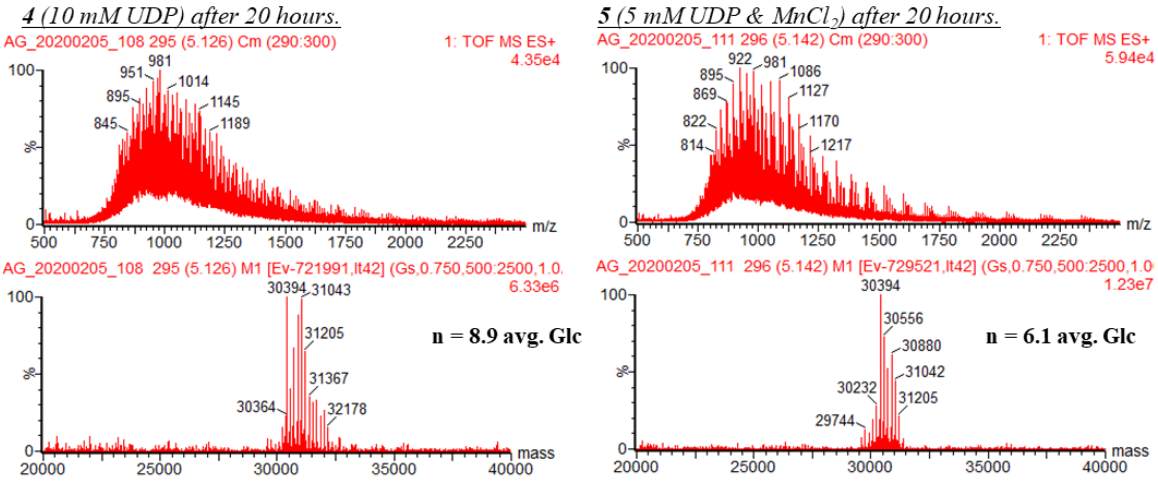
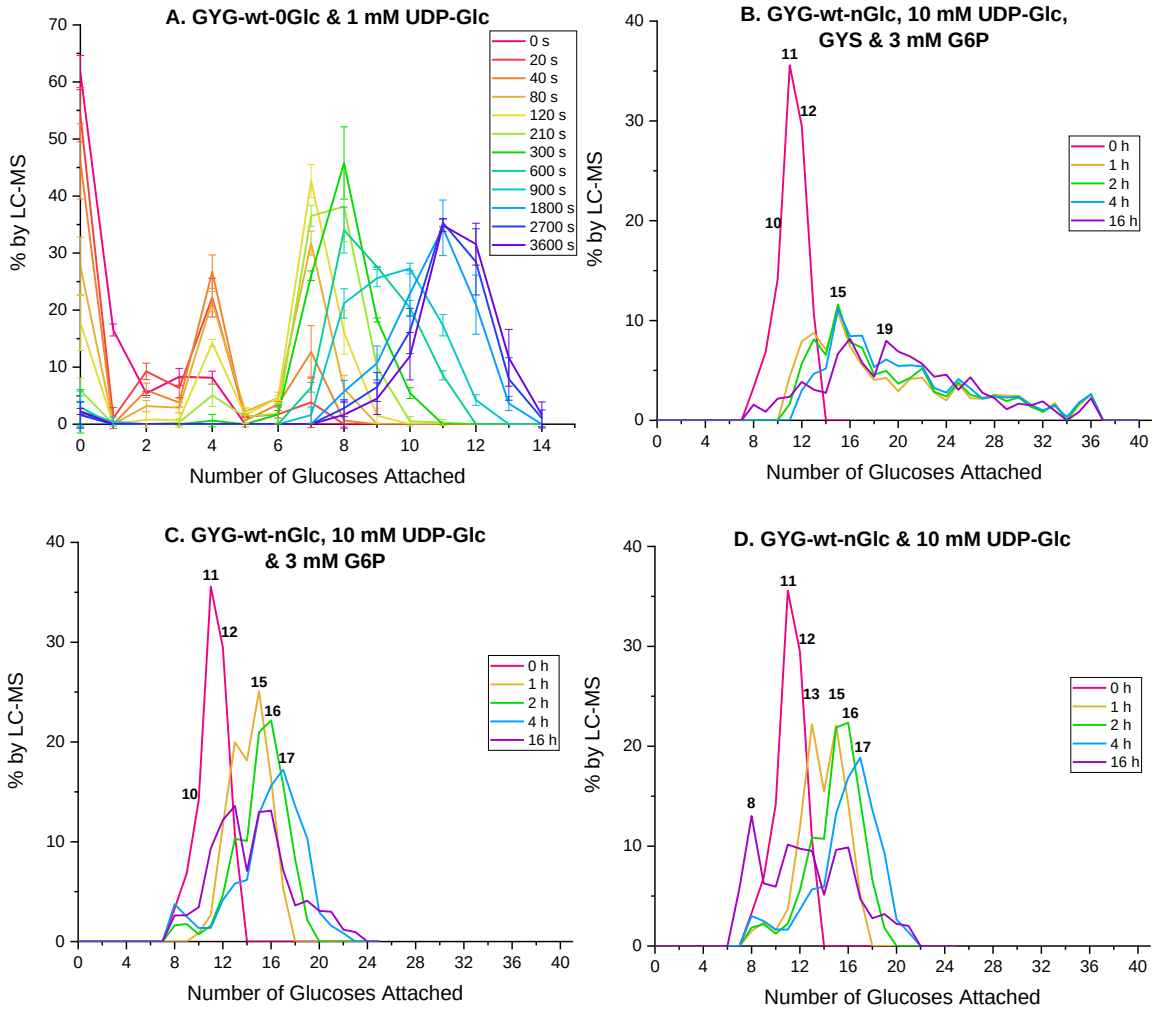


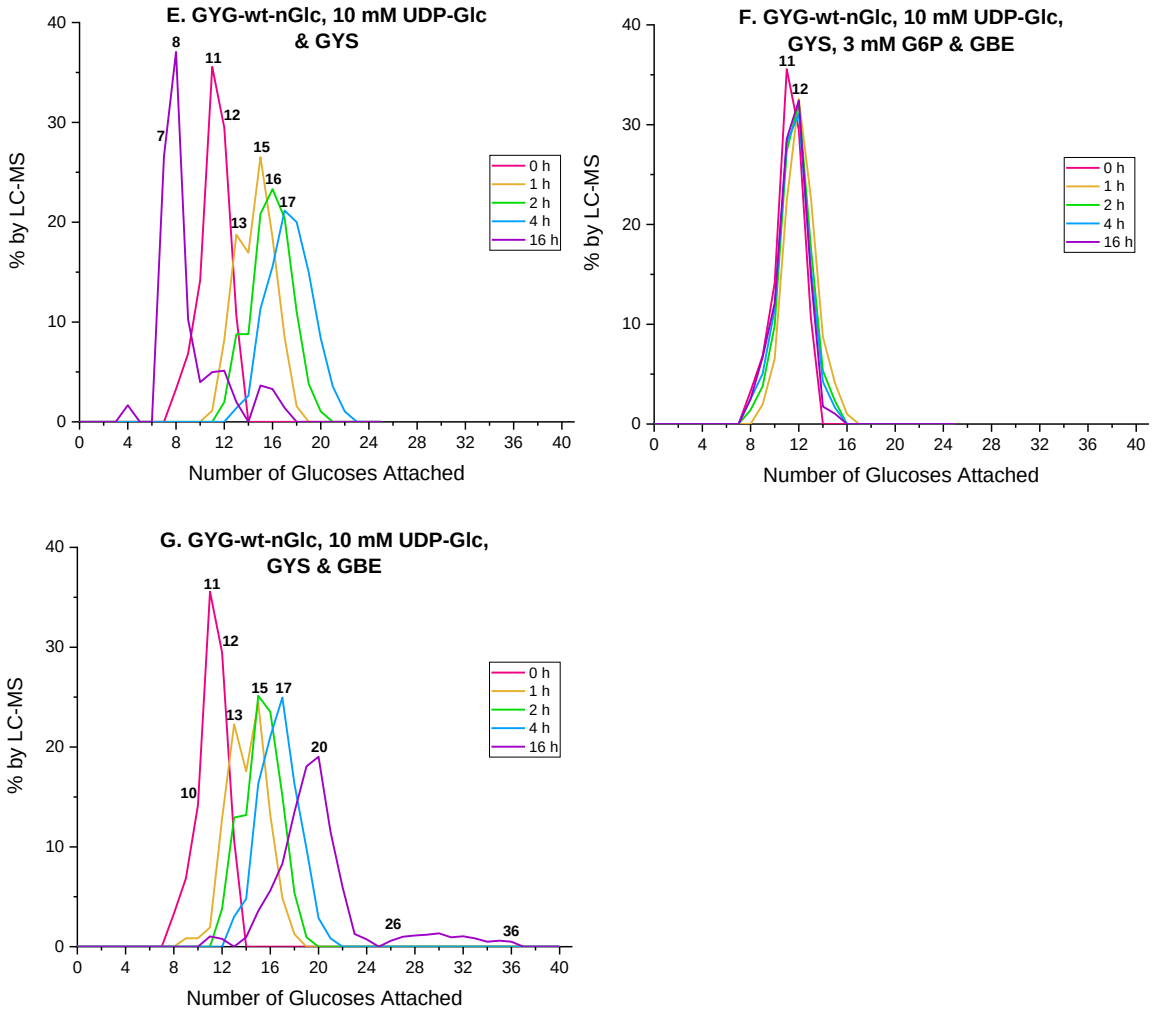
Figure SI 4.14. Hydrolysis of GYG-wt-(n+m)Glc (n = 19-21 avg. Glc residues) to GYG-wt-nGlc (n = 4-6 avg. Glc residues). In absence of UDP-Glucose, GYG-wt-(n+m)Glc (n = 19-21 avg. Glc residues) underwent hydrolysis of its polyglucosidic chain when incubated with 10 mM UDP (4) but trimming was highly accelerated when both UDP and $MnCl_2$ (5) were added and consequently, GYG-wt-nGlc (n = 4-6 avg. Glc residues) glycoforms were formed.

4.5 Supplementary text and figures

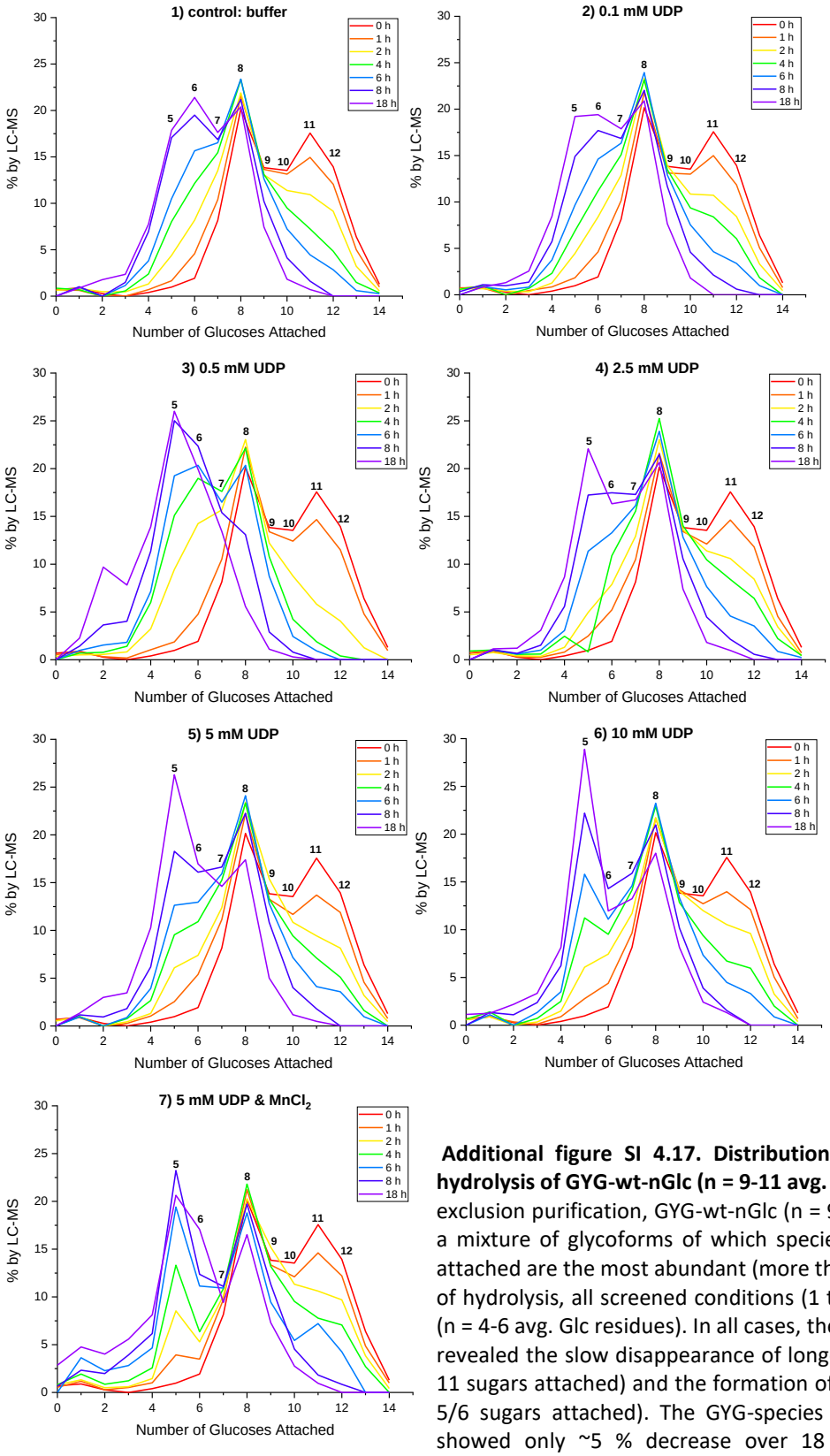
4.5.1 Additional figures



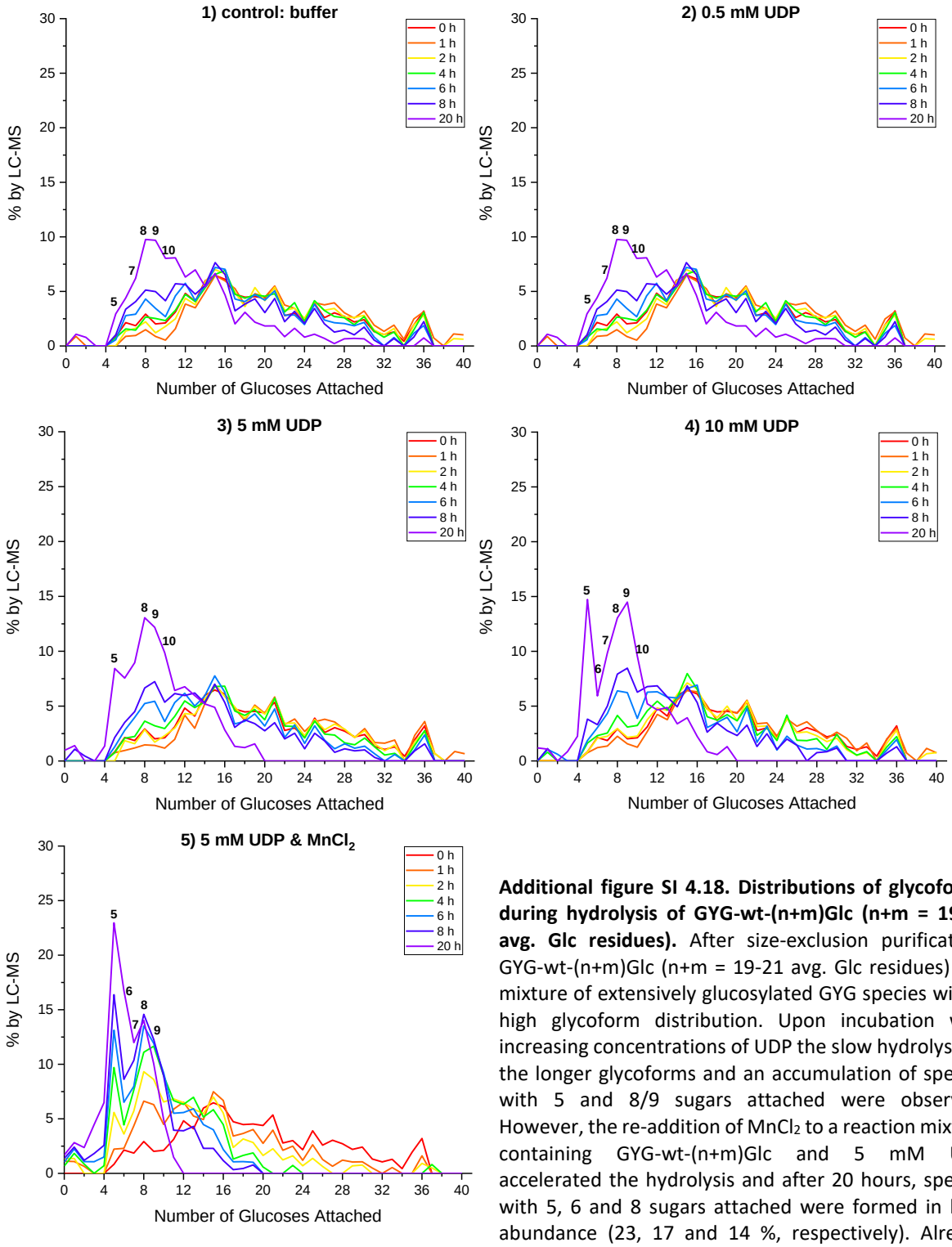
Additional figure SI 4.15. Distributions of glycoforms during autoglucosylation of GYG-wt-0Glc with 1 mM UDP-Glucose and during glucosylation of GYG-wt-nGlc (n = 9-11 avg. Glc residues) with 10 mM UDP-Glucose (Figure 4.6). After autoglucosylation (**A**), GYG-wt-nGlc (n = 9-11 avg. Glc residues) appeared as a mixture of glycoforms among which the ones with 11 and 12 sugars attached are the most abundant (more than 30 % after 1 hour). When this was further incubated with UDP-Glc, GYS and G6P a mixture with a high glycoform distribution of extensively glucosylated GYG-species (up to 36 Glc residues attached) was obtained (**B**). On the contrary, in absence of GYS (**C**) or in presence of only UDP-Glc (**D**), GYG was able to further extend its glucosydic chain but glycoforms with no more than 24 Glc residues were formed.



Additional figure SI 4.16. Distributions of glycoforms during glucosylation of GYG-wt-nGlc (n = 9-11 avg. Glc residues) with 10 mM UDP-Glucose (Figure 4.7, entries E to G). The reaction mixture containing the GYG-wt-nGlc starting material and GYS without its allosteric activator G6P showed a similar glucosylation pattern to **C** and **D** in the first 4 hours of incubation and gave mostly GYG-species carrying 17 Glc residues. However, these underwent hydrolysis to GYG-wt-7/8Glc glycoforms after a longer incubation time (**E**). Unexpectedly, when GYG-wt-nGlc was incubated with both GYS and GBE in the presence of G6P the glucosylation extent did not change significantly over the 16 hours (**F**). On the contrary, when the same reaction was set up in the absence of G6P, low abundant, highly glucosylated GYG-species (that carried from 26 to 36 Glc residues) were formed, in addition to the highly abundant population of glycoforms that contained predominantly 18 to 22 Glc residues (**G**).



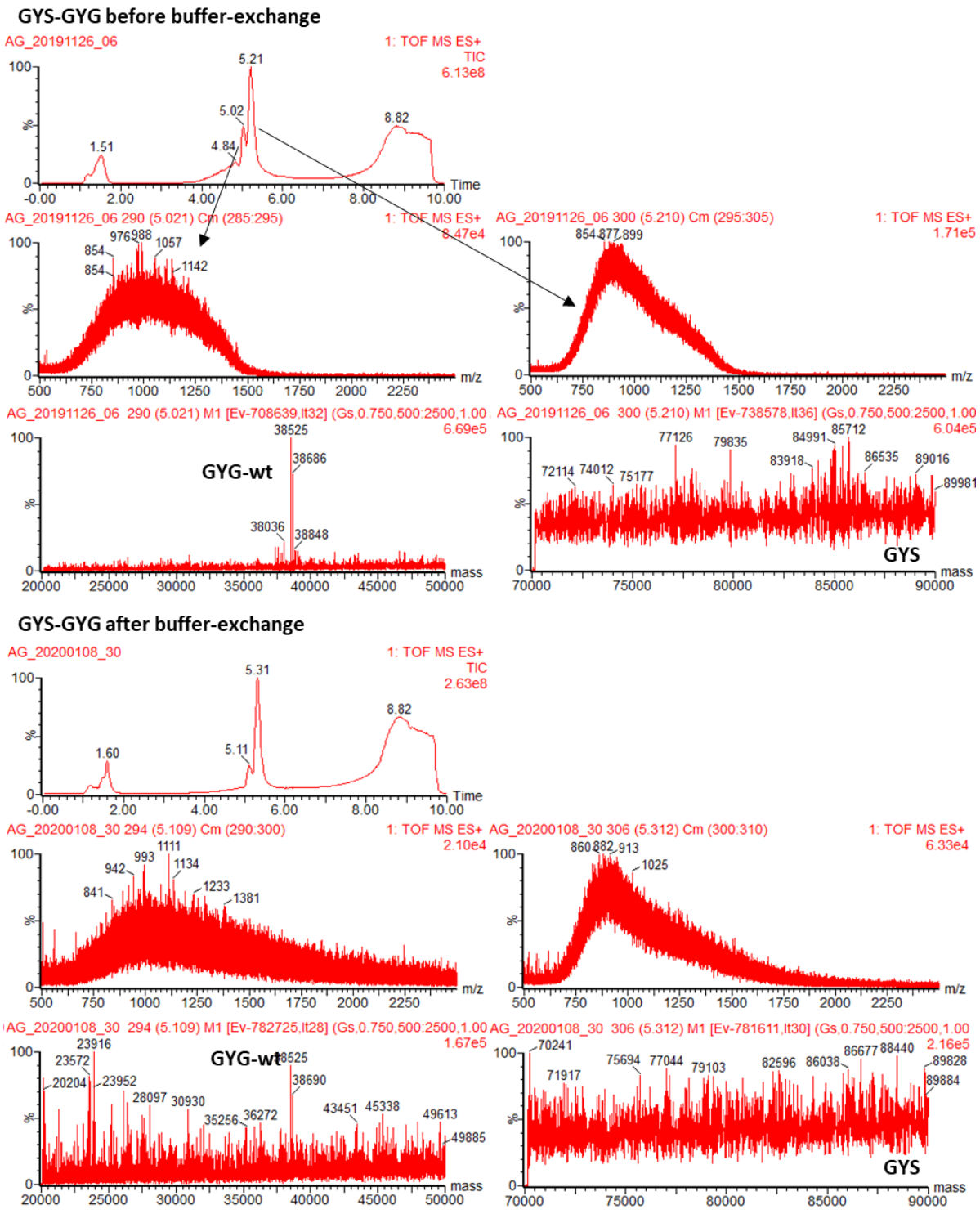
Additional figure SI 4.17. Distributions of glycoforms during hydrolysis of GYG-wt-nGlc (n = 9-11 avg. Glc residues). After size-exclusion purification, GYG-wt-nGlc (n = 9-11 avg. Glc residues) is a mixture of glycoforms of which species with 8 and 11 sugars attached are the most abundant (more than 15 %). After 18 hours of hydrolysis, all screened conditions (1 to 7) led to GYG-wt-nGlc (n = 4-6 avg. Glc residues). In all cases, the glycoform distributions revealed the slow disappearance of longer glycoforms (e. g. with 11 sugars attached) and the formation of shorter ones (e. g. with 5/6 sugars attached). The GYG-species with 8 sugars attached showed only ~5 % decrease over 18 hours, in all screened conditions. Note that these kinetic profiles and the data shown in figures SI 4.10 and 4.11 are generated from the same LC-MS samples; proportions of glycoforms are determined from their relative peak intensities (see Appendix).



Additional figure SI 4.18. Distributions of glycoforms during hydrolysis of GYG-wt-(n+m)Glc (n+m = 19-21 avg. Glc residues). After size-exclusion purification, GYG-wt-(n+m)Glc (n+m = 19-21 avg. Glc residues) is a mixture of extensively glucosylated GYG species with a high glycoform distribution. Upon incubation with increasing concentrations of UDP the slow hydrolysis of the longer glycoforms and an accumulation of species with 5 and 8/9 sugars attached were observed. However, the re-addition of $MnCl_2$ to a reaction mixture containing GYG-wt-(n+m)Glc and 5 mM UDP accelerated the hydrolysis and after 20 hours, species with 5, 6 and 8 sugars attached were formed in high abundance (23, 17 and 14 %, respectively). Already after 6 hours, no species with more than 20 glucose residues attached could be observed. Note that these kinetic profiles and data shown in figures SI 4.13 and 4.14 are generated from the same LC-MS samples; proportions of glycoforms are determined from their relative peak intensities (see Appendix).

4.5.2 Note on the use of GYS as a complex with GYG-wt

The use of GYS as a complex with the GYG-wt protein (partially glucosylated full-length protein) had the potential to cause misassignment of the molecular masses during the deconvolution of the LC-MS spectra due to a possible overlap between the molecular masses corresponding to glycoforms of the complexed GYG-wt and the ones generated by the GYS-catalysed glucosylation of GYG-wt-nGlc (n = 9-11 avg. Glc residues). Fortunately, although LC-MS analysis of the protein complex initially revealed presence of the glucosylated GYG-wt, the latter was lost during the buffer-exchange step required prior to the glucosylation assay. Thus, the subsequent LC-MS (Additional figure SI 4.19) and SDS-PAGE analysis (Figure SI 4.4) revealed that the protein solution which was next used to glucosylate GYG-wt-nGlc (n = 9-11 avg. Glc residues) contained only non-complexed GYS. Moreover, the reactions described in Figure **4.7** entries 7 and 8 tested whether increasing the concentration of the GYG-wt complexed GYS (4x) while omitting the addition of GYG-wt-nGlc would have facilitated the identification of any GYG-wt glycoforms, either by LC-MS or SDS-PAGE analysis. This was not the case, only GYS was identified, and so we are confident that all the analyses following the glucosylation of GYG-wt-nGlc are a true reflection of what occurred on this substrate and not on the added GYG-wt, part of the initial GYS-complex.



Additional figure SI 4.19. While GYG-wt, part of the GYS expression complex could not be confidently detected by LC-MS analysis after buffer-exchange into 25 mM NaHEPES pH 7.5 buffer, GYS did not have good ionization properties, neither in the storage buffer nor in the 25 mM NaHEPES pH 7.5 buffer. Nevertheless, while GYS was always detectable by SDS-PAGE the opposite was valid for GYG-wt (Figure SI 4.4). Therefore, the enzymatic glucosylation activity observed on GYG-wt-0Glc upon incubation with the complex was always attributed to GYS.

4.6 References

1. Roach PJ, Depaoli-Roach AA, Hurley TD, Tagliabracci VS. Glycogen and its metabolism: some new developments and old themes. *Biochem J*. 2012;441(3):763-787.

2. Lomako J, Lomako WM, Whelan WJ. A self-glucosylating protein is the primer for rabbit muscle glycogen biosynthesis. *FASEB J*. 1988;2(15):3097-3103.

3. Smythe C, Cohen P. The discovery of glycogenin and the priming mechanism for glycogen biogenesis. *Eur J Biochem*. 1991;200(3):625-631.

4. Barbetti F, Rocchi M, Bossolasco M, et al. The human skeletal muscle glycogenin gene: cDNA, tissue expression and chromosomal localization. *Biochem Biophys Res Commun*. 1996;220(1):72-77.

5. Mu J, Skurat AV, Roach PJ. Glycogenin-2, a novel self-glucosylating protein involved in liver glycogen biosynthesis. *J Biol Chem*. 1997;272(44):27589-27597.

6. Mu J, Roach PJ. Characterization of human glycogenin-2, a self-glucosylating initiator of liver glycogen metabolism. *J Biol Chem*. 1998;273(52):34850-34856.

7. Alonso MD, Lomako J, Lomako WM, Whelan WJ. Tyrosine-194 of glycogenin undergoes autocatalytic glucosylation but is not essential for catalytic function and activity. *FEBS Lett*. 1994;342(1):38-42.

8. Liu J, Mushegian A. Three monophyletic superfamilies account for the majority of the known glycosyltransferases. *Protein Sci*. 2003;12(7):1418-1431.

9. Hurley TD, Stout S, Miner E, Zhou J, Roach PJ. Requirements for catalysis in mammalian glycogenin. *J Biol Chem*. 2005;280(25):23892-23899.

10. Gibbons BJ, Roach PJ, Hurley TD. Crystal structure of the autocatalytic initiator of glycogen biosynthesis, glycogenin. *J Mol Biol*. 2002;319(2):463-477.

11. Chaikuad A, Froese DS, Berridge G, von Delft F, Oppermann U, Yue WW. Conformational plasticity of glycogenin and its maltosaccharide substrate during glycogen biogenesis. *Proc Natl Acad Sci U S A.* 2011;108(52):21028–21033.

12. Hurley TD, Walls C, Bennett JR, Roach PJ, Wang M. Direct detection of glycogenin reaction products during glycogen initiation. *Biochem Biophys Res Commun*. 2006;348(2):374-378.

13. Ardèvol A, Rovira C. Reaction mechanisms in carbohydrate-active enzymes: glycoside hydrolases and glycosyltransferases. insights from *ab initio* quantum mechanics/molecular mechanics dynamic simulations. *J Am Chem Soc*. 2015;137(24):7528-7547.

14. Spicer CD, Davis BG. Palladium-mediated site-selective Suzuki–Miyaura protein modification at genetically encoded aryl halides. *Chem Commun*. 2011;47(6):1698-1700.

15. Adeva-Andany MM, González-Lucán M, Donapetry-García C, Fernández-Fernández C, Ameneiros-Rodríguez E. Glycogen metabolism in humans. *BBA Clin*. 2016;5:85-100.

16. Browner MF, Nakano K, Bang AG, Fletterick RJ. Human muscle glycogen synthase cDNA sequence: a negatively charged protein with an asymmetric charge distribution. *Proc Natl Acad Sci U S A*. 1989;86(5):1443-1447.

17. Nuttall FQ, Gannon MC, Bai G, Lee EY. Primary structure of human liver glycogen synthase deduced by cDNA cloning. *Arch Biochem Biophys*. 1994;311(2):443-449.

18. Baskaran S, Roach PJ, DePaoli-Roach AA, Hurley TD. Structural basis for glucose-6-phosphate activation of glycogen synthase. *Proc Natl Acad Sci U S A*. 2010;107(41):17563-17568.

19. Skurat AV, Wang Y, Roach PJ. Rabbit skeletal muscle glycogen synthase expressed in COS cells. Identification of regulatory phosphorylation sites. *J Biol Chem*. 1994;269(41):25534-25542.

20. Skurat AV, Roach PJ. Phosphorylation of sites 3a and 3b (Ser640 and Ser644) in the control of rabbit muscle glycogen synthase. *J Biol Chem*. 1995;270(21):12491-12497.

21. Alzari PM, Shepard W, Guerin ME, Ugalde JE, Ugalde RA, Buschiazzo A. Crystal structure of glycogen synthase: homologous enzymes catalyze glycogen synthesis and degradation. *EMBO J*. 2004;23(16):3196-3205.

22. Froese DS, Michaeli A, McCorvie TJ, et al. Structural basis of glycogen branching enzyme deficiency and pharmacologic rescue by rational peptide design. *Hum Mol Genet*. 2015;24(20):5667-5676.

23. MacGregor EA, Janecek S, Svensson B. Relationship of sequence and structure to specificity in the α-amylase family of enzymes. *Biochim Biophys Acta*. 2001;1546(1):1-20.

24. Curtino JA, Aon MA. From the seminal discovery of proteoglycogen and glycogenin to emerging knowledge and research on glycogen biology. *Biochem J*. 2019;476(21):3109-3124.

25. Geddes R, Harvey JD, Wills PR. The molecular size and shape of liver glycogen. *Biochem J*. 1977;163(2):201-209.

26. Geddes R, Jeyarathan P, Taylor JA. Molecular and metabolic aspects of lysosomal glycogen. *Carbohydr Res*. 1992;227:339-349.

27. Prats C, Graham TE, Shearer J. The dynamic life of the glycogen granule. *J Biol Chem*. 2018;293(19):7089-7098.

28. Johnson LN, Madsen NB, Mosley J, Wilson KS. The crystal structure of phosphorylase b at 6 Å resolution. *J Mol Biol*. 1974;90(4):703-717.

29. Newgard CB, Hwang PK, Fletterick RJ. The family of glycogen phosphorylases: structure and function. *Crit Rev Biochem Mol Biol*. 1989;24(1):69-99.

30. Johnson GF, Tu JI, Bartlett ML, Graves DJ. Physical-chemical studies on the pyridoxal phosphate binding site in sodium borohydride-reduced and native phosphorylase. *J Biol Chem*. 1970;245(21):5560-5568.

31. Sprang SR, Withers SG, Goldsmith EJ, Fletterick RJ, Madsen NB. Structural basis for the activation of glycogen phosphorylase b by adenosine monophosphate. *Science*. 1991;254(5036):1367-1371.

32. Sprang SR, Acharya KR, Goldsmith EJ, et al. Structural changes in glycogen phosphorylase induced by phosphorylation. *Nature*. 1988;336(6196):215-221.

33. Mathieu C, Dupret J, Lima FR. The structure of brain glycogen phosphorylase—from allosteric regulation mechanisms to clinical perspectives. *FEBS J*. 2017;284(4):546-554.

34. Zhai L, Feng L, Xia L, Yin H, Xiang S. Crystal structure of glycogen debranching enzyme and insights into its catalysis and disease-causing mutations. *Nat Commun*. 2016;7(1):11229.

35. Zmasek CM, Godzik A. Phylogenomic analysis of glycogen branching and debranching enzymatic duo. *BMC Evol Biol*. 2014;14:183.

36. Muenks AG, Stiers KM, Beamer LJ. Sequence-structure relationships, expression profiles, and disease-associated mutations in the paralogs of phosphoglucomutase 1. *PLoS One*. 2017;12(8):e0183563.

37. Stiers KM, Kain BN, Graham AC, Beamer LJ. Induced structural disorder as a molecular mechanism for enzyme dysfunction in phosphoglucomutase 1 deficiency. *J Mol Biol*. 2016;428(8):1493-1505.

38. Nawaz M, Ferreira J, Nedyalkova L, et al. The catalytic inactivation of the *N*-half of human hexokinase 2 and structural and biochemical characterization of its mitochondrial conformation. *Biosci Rep*. 2018;38(1):BSR20171666.

39. Ritov VB, Kelley DE. Hexokinase isozyme distribution in human skeletal muscle. *Diabetes*. 2001;50(6):1253-1262.

40. Yu Q, Zheng X. The crystal structure of human UDP-glucose pyrophosphorylase reveals a latch effect that influences enzymatic activity. *Biochem J*. 2012;442(2):283-291.

41. Führing JI, Cramer JT, Schneider J, Baruch P, Gerardy-Schahn R, Fedorov R. A quaternary mechanism enables the complex biological functions of octameric human UDP-glucose pyrophosphorylase, a key enzyme in cell metabolism. *Sci Rep*. 2015;5(1):9618.

42. Bilyard MK, Bailey HJ, Raich L, et al. Palladium-mediated enzyme activation suggests multiphase initiation of glycogenesis. *Nature*. 2018;563(7730):235-240.

43. Bilyard M. *Chemical probing of glycogenesis.* University of Oxford; 2018.

44. Dumas A, Spicer CD, Gao Z, et al. Self-liganded Suzuki–Miyaura coupling for site-selective protein PEGylation. *Angew Chem Int Ed*. 2013;52(14):3916-3921.

45. Ahmadi Z, McIndoe JS. A mechanistic investigation of hydrodehalogenation using ESI-MS. *Chem Commun*. 2013;49(98):11488-11490.

46. Lennox AJJ, Lloyd-Jones GC. The slow-release strategy in Suzuki–Miyaura coupling. *Isr J Chem*. 2010;50(5-6):664-674.

47. Wu L, Chen Y, Pei J, Tang M, Wang S, Liu Z. Tetrahedron. *Chin J Chem*. 2019;37(8):834-842.

48. Welch JT. Tetrahedron report number 221. *Tetrahedron*. 1987;43(14):3123-3197.

49. Tsuji J, Yamakawa T. A convenient method for the preparation of 1-olefins by the palladium catalyzed hydrogenolysis of allylic acetates and allylic phenyl ethers with ammonium formate. *Tetrahedron Lett*. 1979;20(7):613-616.

50. Soltermann F, Foley EDB, Pagnoni V, et al. Quantifying protein–protein interactions by molecular counting with mass photometry. *Angew Chem Int Ed*. 2020;59(27):10774-10779.

51. Hendrickson WA, Ogata CM. [28] Phase determination from multiwavelength anomalous diffraction measurements. *Methods Enzymol*. 1997;276:494-523.

52. Alonso MD, Lomako J, Lomako WM, Whelan WJ. Catalytic activities of glycogenin additional to autocatalytic self-glucosylation. *J Biol Chem*. 1995;270(25):15315-15319.

53. Bazán S, Issoglio FM, Carrizo ME, Curtino JA. The intramolecular autoglucosylation of monomeric glycogenin. *Biochem Biophys Res Commun.* 2008;371(2):328-332.

54. Donnelly DP, Rawlins CM, DeHart CJ, et al. Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nat Methods*. 2019;16(7):587-594.

55. Mann M, Kelleher NL. Precision proteomics: the case for high resolution and high mass accuracy. *Proc Natl Acad Sci U S A*. 2008;105(47):18132-18138.

56. Young G, Hundt N, Cole D, et al. Quantitative mass imaging of single biological macromolecules. *Science*. 2018;360(6387):423-427.

57. Dippel R, Boos W. The maltodextrin system of *Escherichia coli*: metabolism and transport. *J Bacteriol*. 2005;187(24):8322-8331.

58. Zhang R, McCarter JD, Braun C, Yeung W, Brayer GD, Withers SG. Synthesis and testing of 2-deoxy-2,2-dihaloglycosides as mechanism-based inhibitors of α-glycosidases. *J Org Chem*. 2008;73(8):3070-3077.

59. Testoni G, Duran J, García-Rocha M, et al. Lack of glycogenin causes glycogen accumulation and muscle function impairment. *Cell Metab*. 2017;26(1):256-266.

60. Krag TO, Ruiz-Ruiz C, Vissing J. Glycogen synthesis in glycogenin 1–deficient patients: a role for glycogenin 2 in muscle. *J Clin Endocrinol Metab*. 2017;102(8):2690-2700.

61. Geoghegan KF, Dixon HB, Rosner PJ, et al. Spontaneous α-*N*-6-phosphogluconoylation of a "His tag" in *Escherichia coli*: the cause of extra mass of 258 or 178 da in fusion proteins. *Anal Biochem*. 1999;267(1):169-184.

Appendix

Contents

1. General procedures for small molecule chemistry 293

2. General procedures for biological work and protein modification 294

3. Denaturing intact protein LC-MS 295

4. NMR spectra 298

5. Results of protein and peptide identifications obtained from the shotgun proteomics analysis of GYG-wt-0Glc 305

1. General procedures for small molecule chemistry

Chemicals and solvents were purchased from Sigma-Aldrich UK, Carbosynth, Acros UK, Alfa Aesar UK or Fisher UK and were used as received unless otherwise stated. All non-aqueous reactions were carried out in oven-dried or flame-dried glassware under an atmosphere of dry nitrogen or argon, as indicated. Room temperature (r.t) refers to ambient temperature. Temperatures of 0 ^{o}C were maintained using an ice-water batch. Pet Ether refers to the fraction of petroleum ether boiling in the range 40-60 ^{o}C. Deionized water was used in all chemical reactions.

Chromatography

Chromatographic separations were done by flash chromatography on Biotage®Isolera™ Systems or manually using Merck Geduran® Si 60 40-63 µm silica gel. Analytical thin-layer chromatography was carried out using Merck aluminium-backed sheets coated with silica gel 60 F_{254}. Compounds were visualised with either of the following stains: potassium permanganate; sulfuric acid; ceric ammonium molybdate (CAM). Flash column chromatography was performed using Geduran® silica 60 40-63 µm.

Mass spectrometry (small molecules)

Low-resolution mass spectra (LRMS) were recorded on an Agilent 6120 Quadrupole spectrometer using electrospray ionisation (ESI) in either positive and/or negative mode as indicated. High-resolution mass spectra (HRMS) were recorded on a Bruker MicroTOF ESI mass spectrometer using a lock-spray source.

Nuclear Magnetic Resonance spectroscopy (NMR spectroscopy)

^{1}H-, ^{13}C-NMR spectra were recorded on a Bruker AVIII HD 400 nanobay or on a Bruker AVII 500. Deuterated solvents were used as the lock. Spectra were analysed using MestReNova software. Chemical shifts are quoted in parts per million (ppm) to the nearest 0.01 ppm and are referenced to the centre of the residual non-deuterated solvent peak. Coupling constants (*J*) are reported in Hertz (Hz) to the nearest 0.1 Hz. Data are reported as follows: chemical shift, multiplicity (br = broad; s = singlet, d = doublet; t = triplet; q = quartet; qn = quintet; m = multiplet; or as a combination of these), coupling constant(s), integration and assignment. Spectra were assigned on the basis of a combination of the following: unambiguous chemical shift; 2D experiments (COSY, HSQC, HMBC); analogy to fully interpreted spectra of related compounds.

IR spectroscopy

Infrared (IR) spectra were recorded on a Bruker Tensor 27 Fourier Transform spectrophotometer equipped with a diamond ATR probe on a Thermo Nexus 670 equipped with a silicon probe. Samples were loaded directly onto the probe.

Melting points

Melting points were measured using a Leica Galen III microscope equipped with a heating stage and a Testo 720 thermocouple probe. Heating was conducted up to 300 °C at a heating rate of 5 °C/minute.

Optical rotations

Optical rotation values were recorded using a Schmidt Haensch Unipol polarimeter at 25 °C using a cell with a path length of 10 cm and are reported in units of. Samples were dissolved in the solvents chloroform, methanol, or water depending on their solubility. The polarimeter was blanked with the corresponding solvent before use. Values represent mean averages from at least 8 successive readings.

2. General procedures for biological work and protein modification

All biological equipment and prepared biological solutions used in the handling of bacterial cultures were sterilized prior to use by autoclaving for 20 minutes at 121 °C in either a Falcon LTE or a Touchclave®-R tabletop autoclave. Biological work was carried out in a Kendro HERAsafe KS-12 Class II Laminar Flow Cabinet. Water was deionized and passed through a 0.22 µm Millipore filter before use. For growth of bacterial cultures, the following incubators were used: New Brunswick Scientific model G25, or New Brunswick Scientific Excella E25. Centrifugation of bacterial cultures and of cell lysates was carried out in a Beckmann Coulter Avanti-J25 centrifuge. JA10 and JA16.250 rotors were used for volumes of 200-500 ml, and a JA25.50 rotor was used for volumes of 25-50 ml. UV/Vis spectroscopy was used to determine the optical density of cell cultures. Therefore, 1 mL of the cell culture was transferred to a 1 cm x 1 cm plastic cuvette and the absorption measured at 600 nm.

SDS-PAGE

Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was carried out using pre-cast Invitrogen Novex Bis-Tris gels (4-12 %). Samples were prepared by mixing with 5x SDS loading dye in a 4:1 sample:dye ratio and were then heated at 90 ^{o} C for 5 minutes prior to gel loading. Perfect Protein^{TM} markers were used as a reference. Gels were run at a voltage of 200 V for 30-40 minutes or 130 V for 80 min, using 1x MES buffer. Proteins were visualised through staining with InstantBlue® (Expedeon Ltd.) for ca. 60 minutes. Staining solution was then decanted and the gel washed with water (1 h then overnight) prior to scanning. Samples containing intact decorin were analysed by SDS-PAGE by staining with Alcian Blue followed by destaining with AcOH:MeOH:H_2O 25:10:65.

Protein modification reactions

All protein modification reactions were carried out in 1.5 or 0.5 mL Eppendorf® tubes. Water refers to MilliQ water. Protein concentrations were measured using the Pierce BCA Protein Assay Kit® according to the manufacturer's protocol with Bovine Serum Albumin (BSA) as a standard, or by using a NanoDrop® spectrophotometer which measures the A_{280} of the protein according to its MW (molecular weight) and E (molar extinction coefficient).

Circular dichroism

Measurements were made at room temperature on a Chirascan® CD-spectrophotometer (Applied Photophysics), using a 0.1 cm quartz cuvette. A background spectrum was recorded using 0.2 ml of appropriate buffer. Protein samples (0.2 mL) were run at concentrations of 0.15-0.25 mg/ml. Scans were run using 1.0 s per time point and 0.5 nm increments, scanning from 185 nm to 260 nm. 3 repeats were carried out. For each protein sample measurement, the background spectrum was subtracted, then the runs averaged (mean) and the resultant data smoothed using the literature standard S.G algorithm (3 times).

3. Denaturing intact protein LC-MS

Data Acquisition:

Liquid-chromatography mass spectrometry (LC-MS) was performed using the methods detailed below. Solvents A and B are water and acetonitrile respectively, both containing 0.1 % formic acid.

Waters Xevo G2-S or Xevo G2-XS QToF MS coupled to a Waters Acquity UPLC using a Thermo Proswift (250 x 4.6 mm x 5 µm) column or a Waters 2.1 x 50 mm BEH C18 1.7 µm column. Electrospray source parameters were as follows: capillary voltage 3.0 kV and cone voltage 20 V. Nitrogen was used as the nebulizer and desolvation gas at a total flow of 700 L/hr. The protein mixture was diluted to 0.01-0.06

mg/mL with buffer pH 8.0 100 mM NH_4OAc, +/- 10-100 mM DTT or water/formic acid (100:0.1) and 1-10 µL were injected. Spectra were calibrated using an internal lock-spray. The gradient program is shown below; a flow rate of 0.3 mL/min was used throughout.

Time (min)	% Solvent A	% Solvent B
0.0	95	5
1.0	95	5
7.0	5	95
8.0	5	95
8.1	95	5
10.0	95	5

Data Processing and Analysis:

Raw MS data containing multiply charged ion series were deconvoluted using the MaxEnt1 algorithm of MassLynx version 4.1 (Waters). The deconvolution was set up as follows: output mass ranges 10000:30000 for histone species, 10000:65000 for decorin/EPO species and 20000:40000 for GYG species, resolution 1 Da/channel; damage model, uniform Gaussian width at half height 0.75 Da, minimum intensity ratios left 50 % and right 50 % and completion options, iterate to convergence.
The average number of glucoses attached per GYG enzyme after autoglucosylation, GYS-catalyzed glucosylation or hydrolysis assays was determined *via* LC-MS analysis. Analysis of the spectra obtained to determine the glucosylation extent, was carried out according to the method of Hurley *et al.* (Biochem Biophys Res Commun. 2006;348(2):374-378) and the following parameters were calculated:

1. **Proportions of each GYG-glycoform present**

   The proportion of each GYG-glycoform – as a fraction of the total glycoforms present – is obtained from the intensity of its corresponding peak in LC-MS spectrum, relative to the combined intensities of all such MS peaks:

$$proportion\ of\ glycoform = \frac{MS\ peak\ intensity\ of\ glycoform}{Total\ MS\ peak\ intensities\ of\ all\ glycoforms}$$

The above is used to construct the graphs illustrating abundance of different glycoforms before and after glucosylation or hydrolysis assay (proportions are there expressed as percentages), for example in Additional figure **SI 4.15** etc.

2. Average number of glucoses attached (per enzyme)

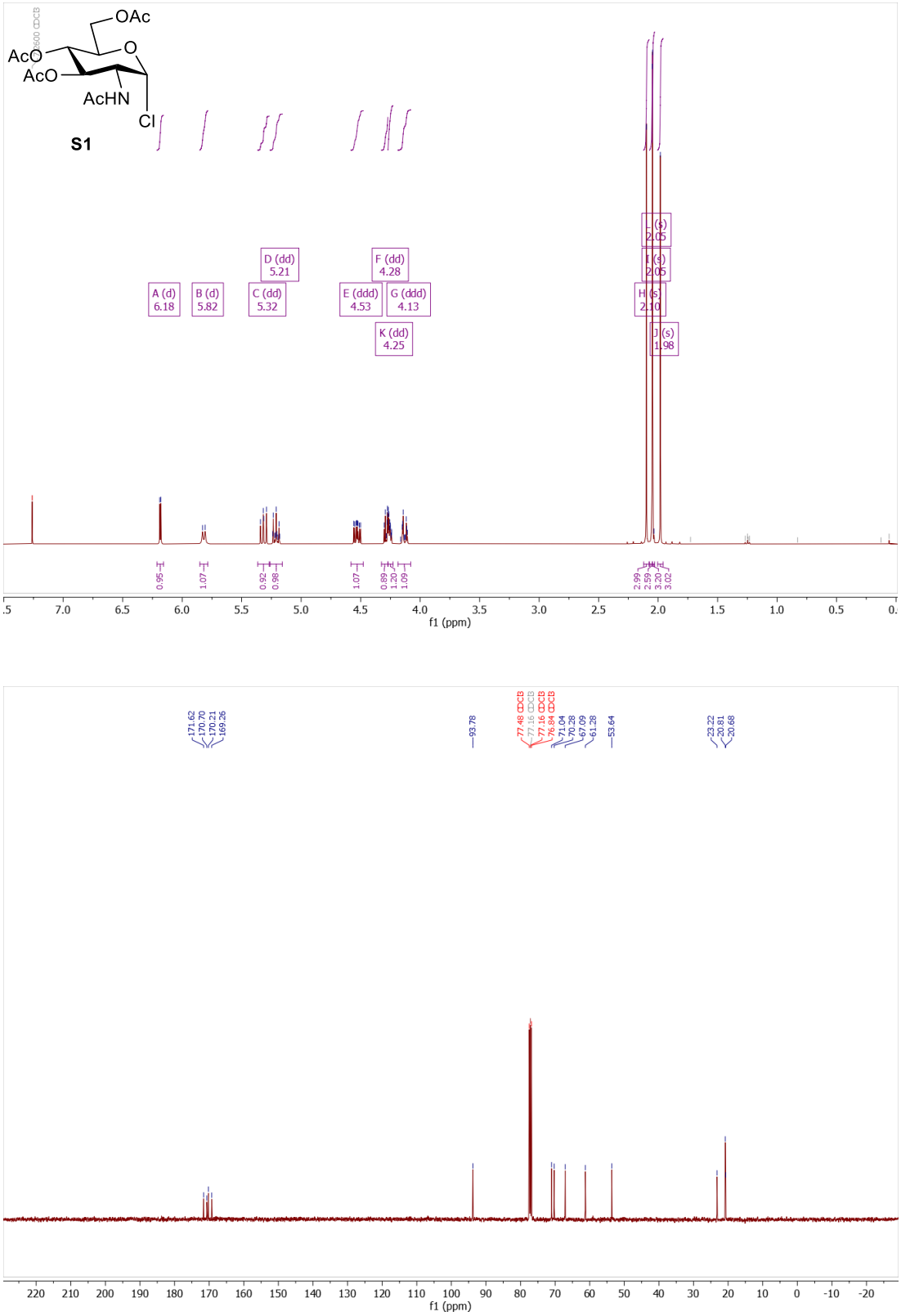
For each GYG-glycoform, the proportion calculated in 1) above (as a value between 0 and 1, i.e. 1 = 100 % of the mixture) is multiplied by the number of glucoses associated with it. Summing the values thus obtained across all glycoforms yields the average number of glucose residues attached. This may be represented by the formula below, in which "i" refers to the number of glucose residues associated with a given glycoform:

$$Average\ no.of\ glucoses\ attached = \sum_{i=1}^{n} i \times (proportion\ of\ glycoform)_i$$

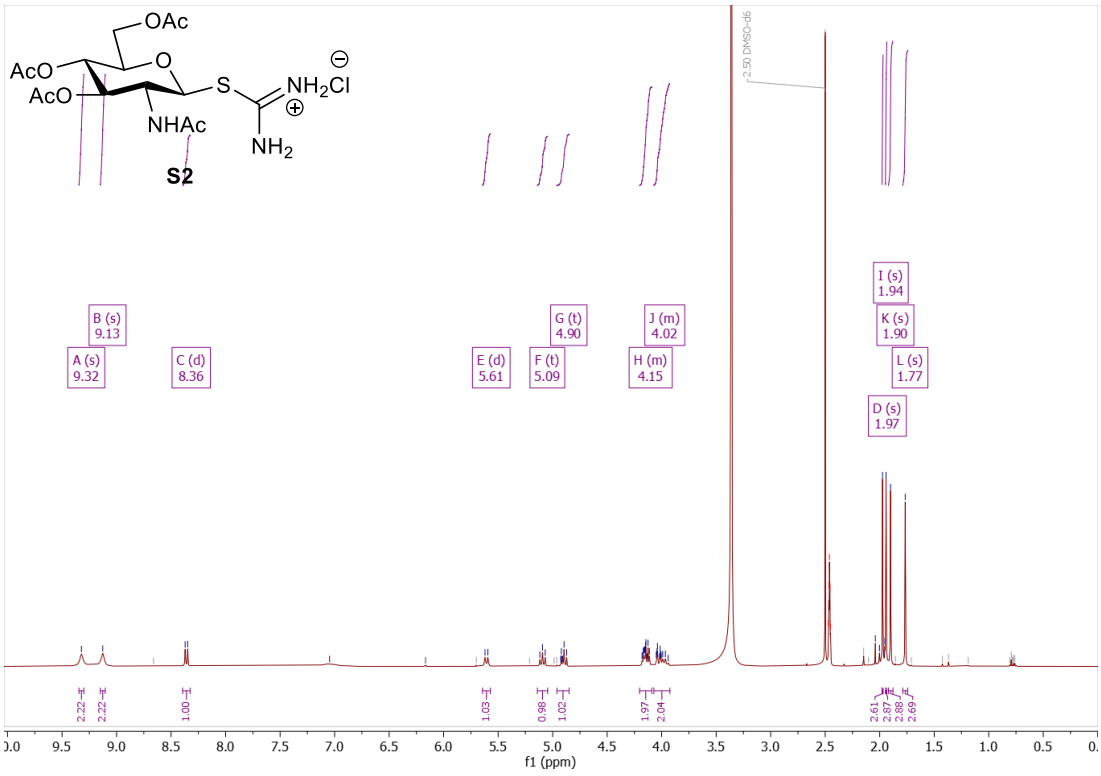
This is effectively a weighted average, in which the "weights" are the number of glucose residues (i).

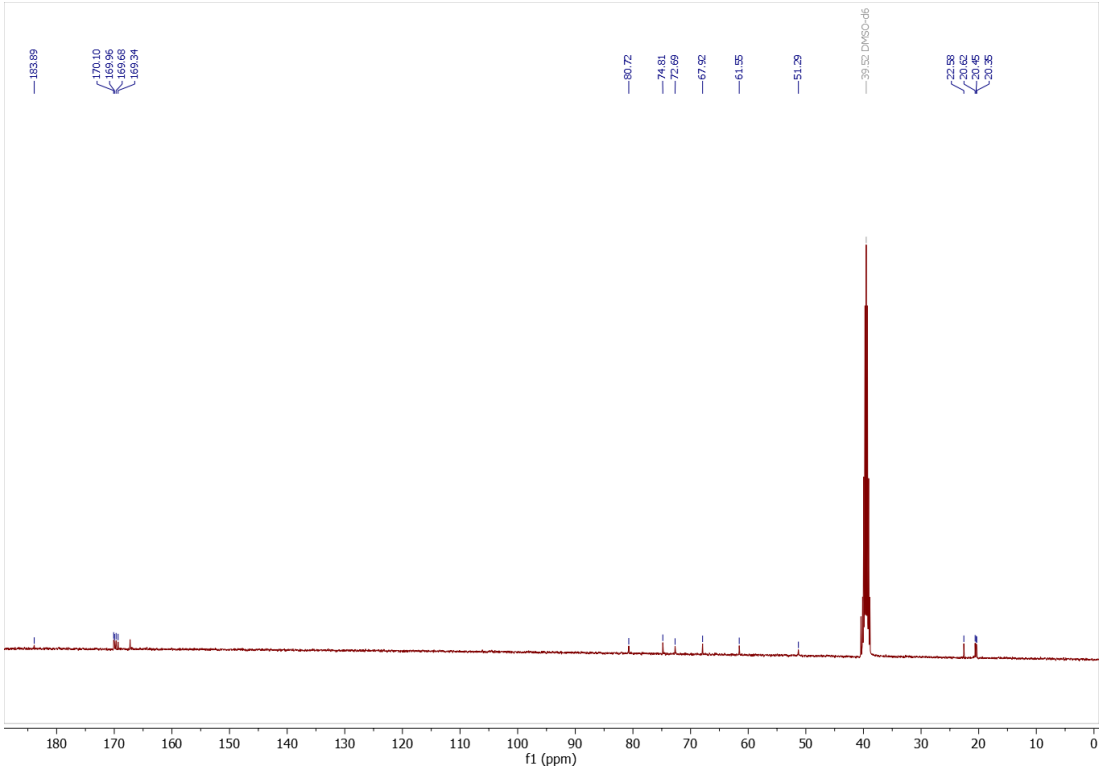
The above is used to construct the graphs illustrating the average number of glucose residues attached to GYG during the glucosylation and hydrolysis assays, for example in Figure **4.8** etc.

4. NMR spectra

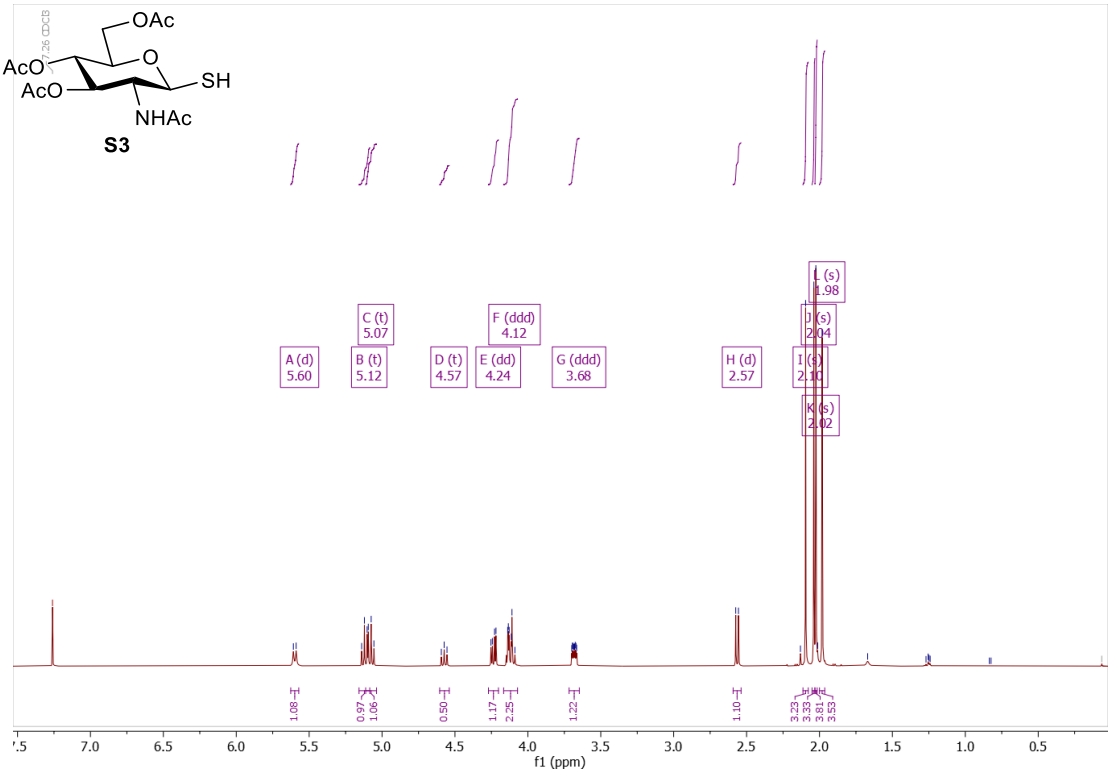


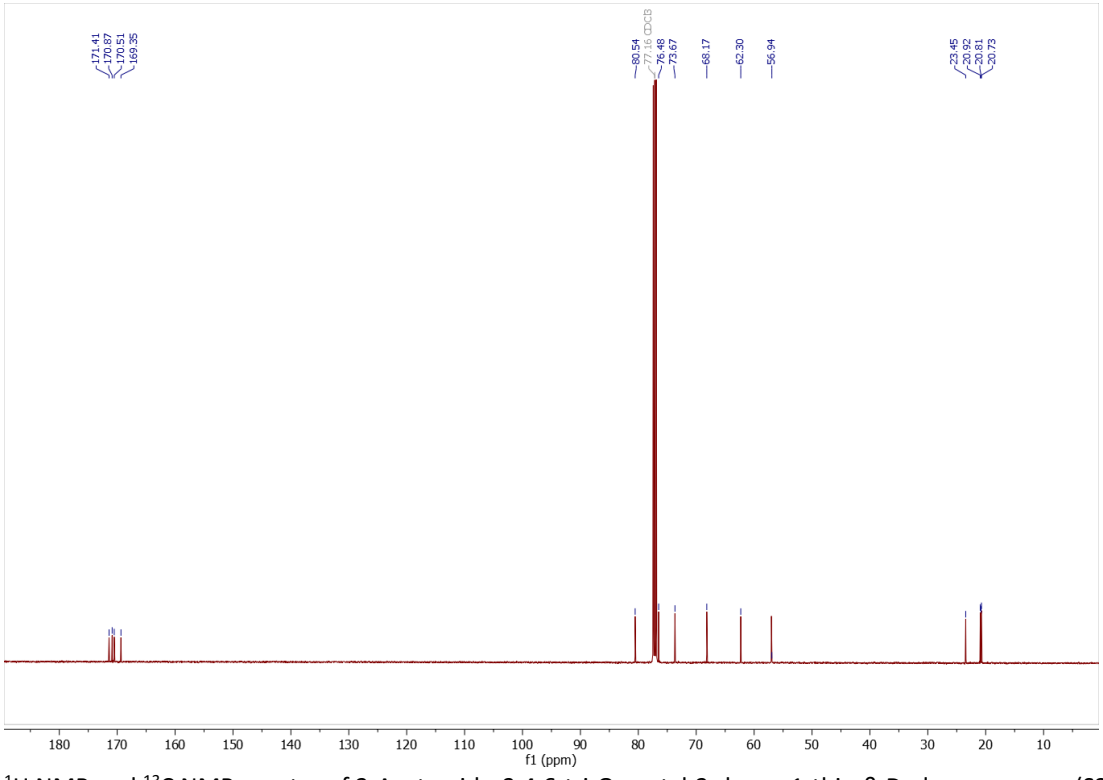
^{1}H NMR and ^{13}C NMR spectra of 3,4,6-Tri-O-acetyl-2-N-acetyl-2-deoxy-β-D-glucopyranosyl chloride (S1)



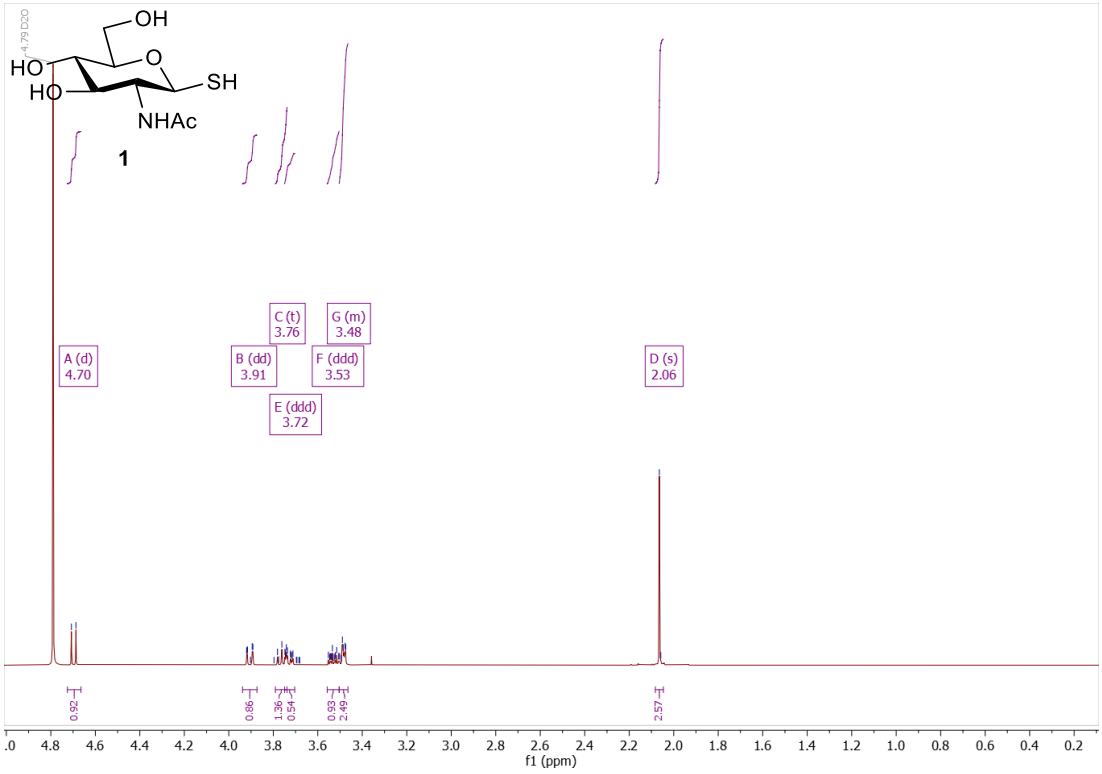


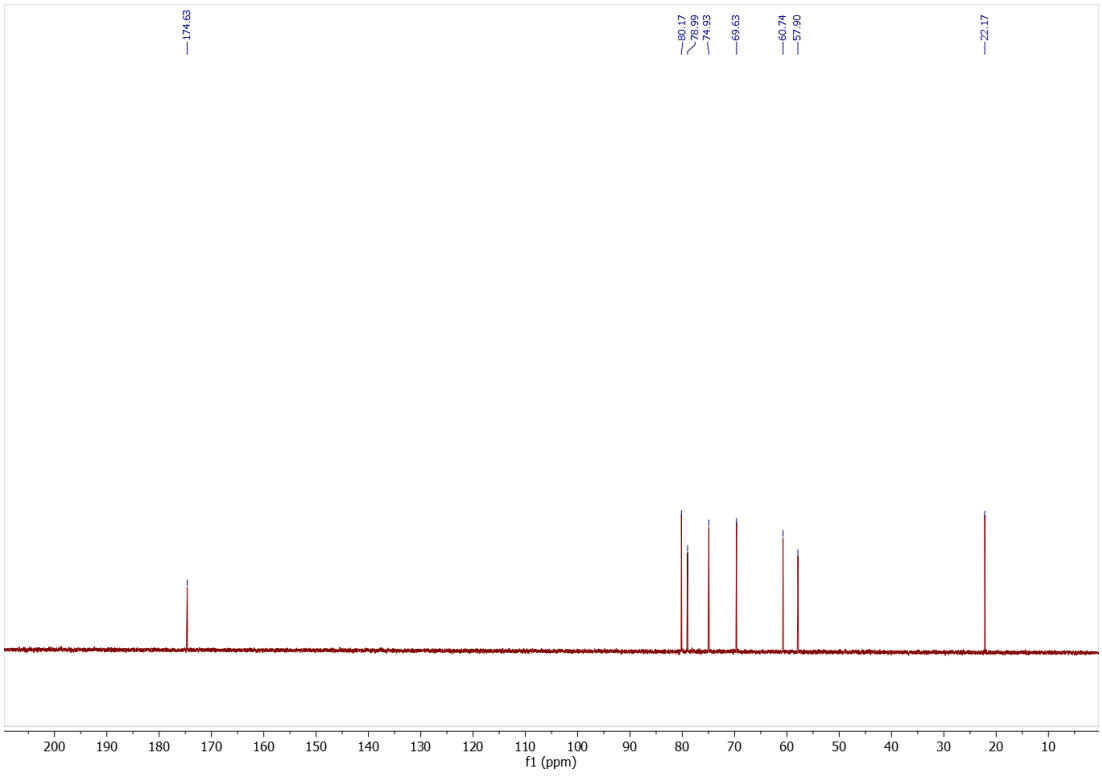
^{1}H NMR and ^{13}C NMR spectra of 2-Acetamido-3,4,6-tri-O-acetyl-2-deoxy-β-D-glucopyranosyl-1-isothiouronium chloride (S2)



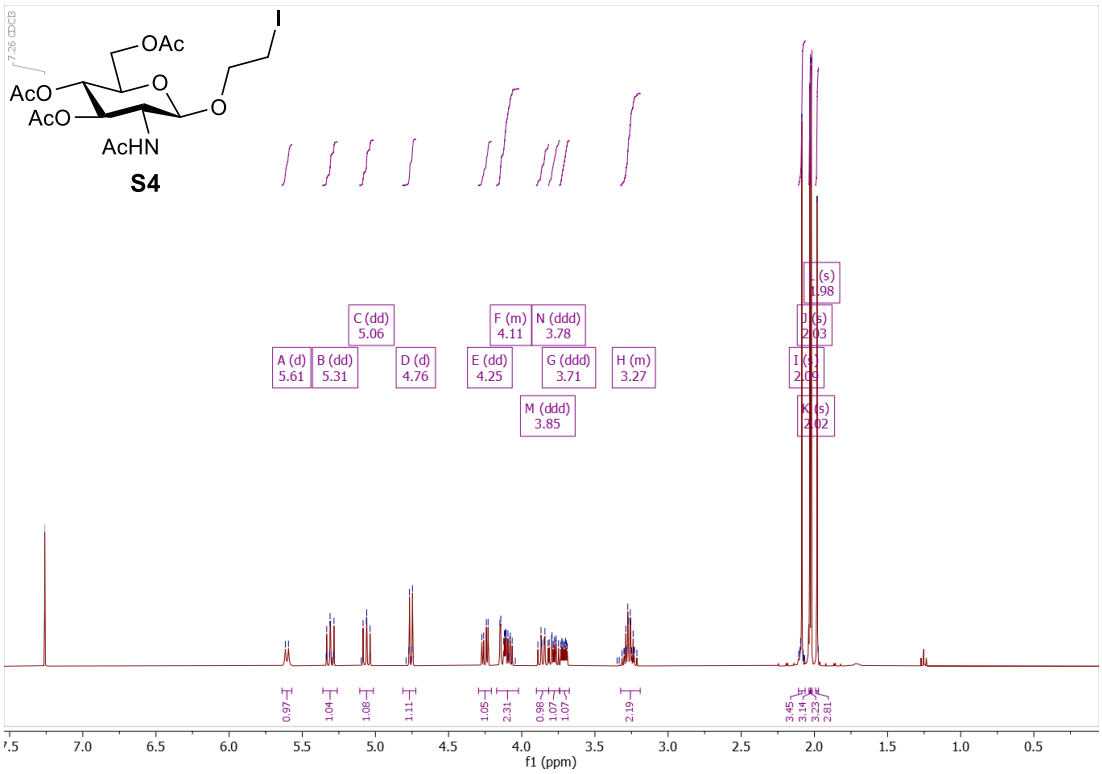


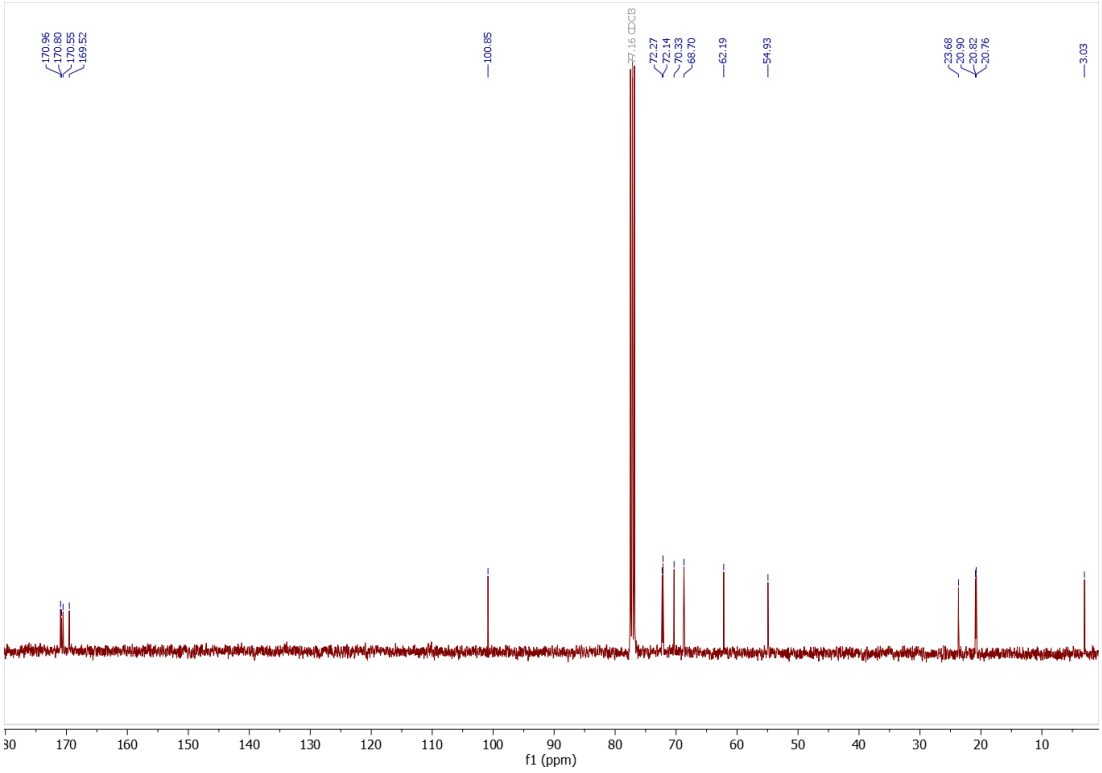
^{1}H NMR and ^{13}C NMR spectra of 2-Acetamido-3,4,6-tri-O-acetyl-2-deoxy-1-thio-β-D-glucopyranose (S3)



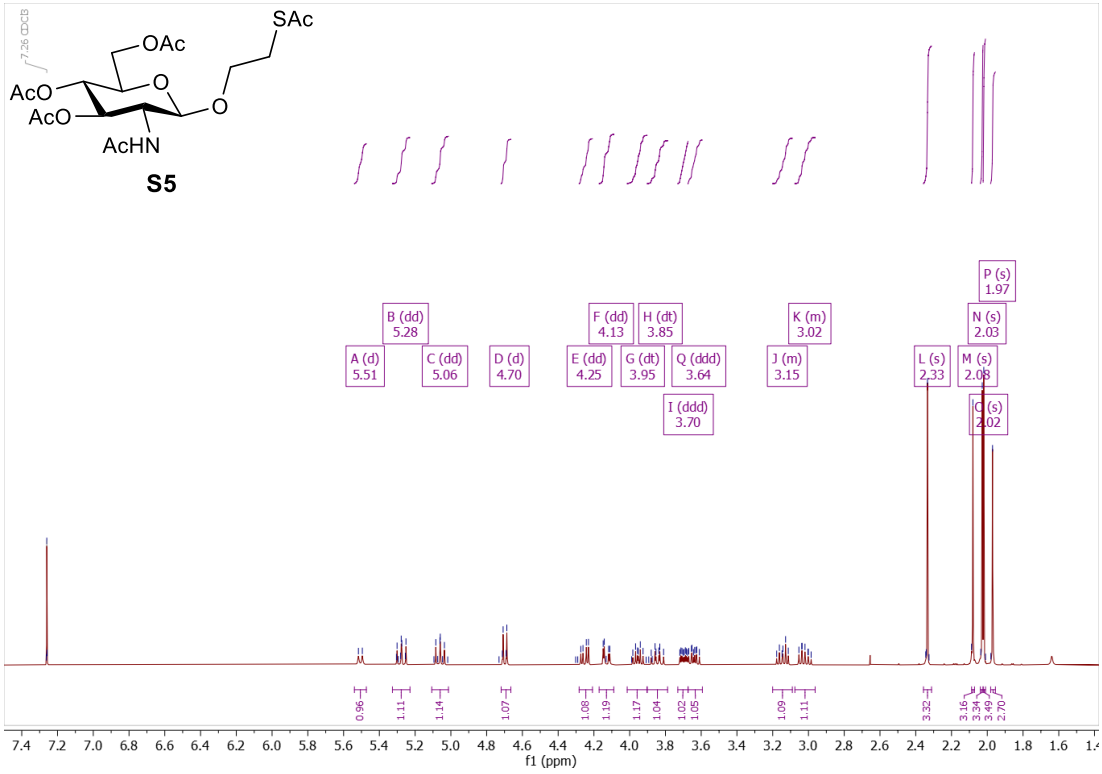


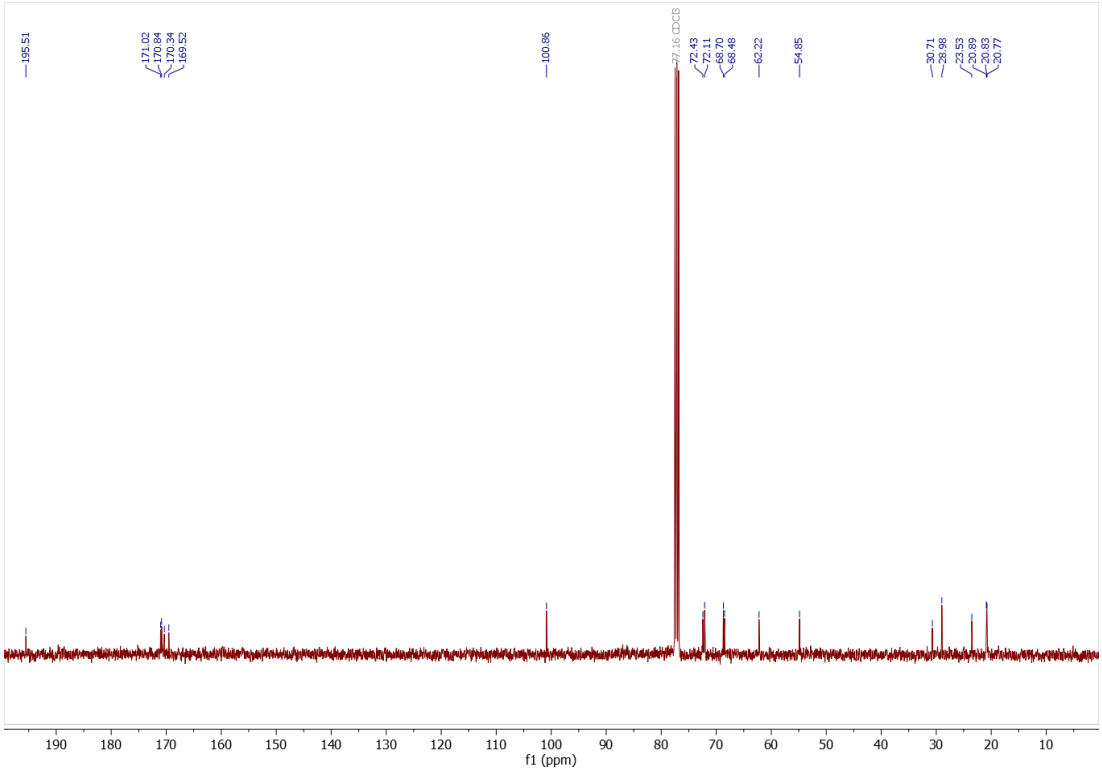
^{1}H NMR and ^{13}C NMR spectra of 2-Acetamido-2-deoxy-1-thio-β-D-glucose (1)



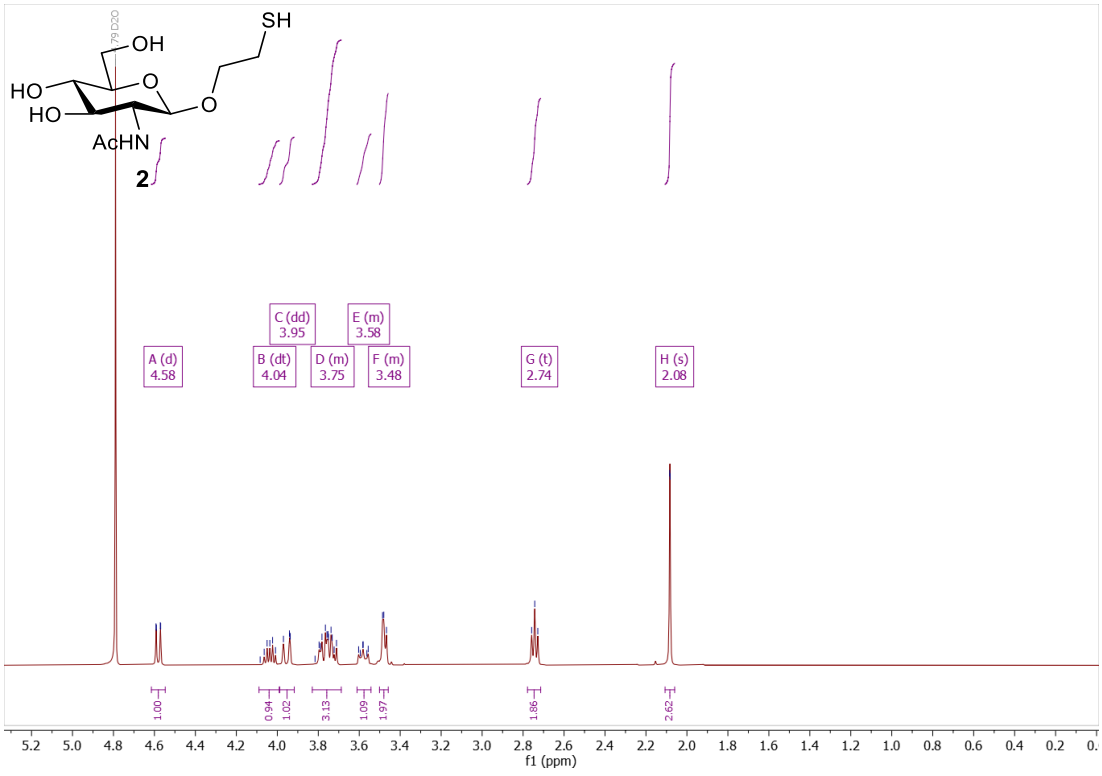


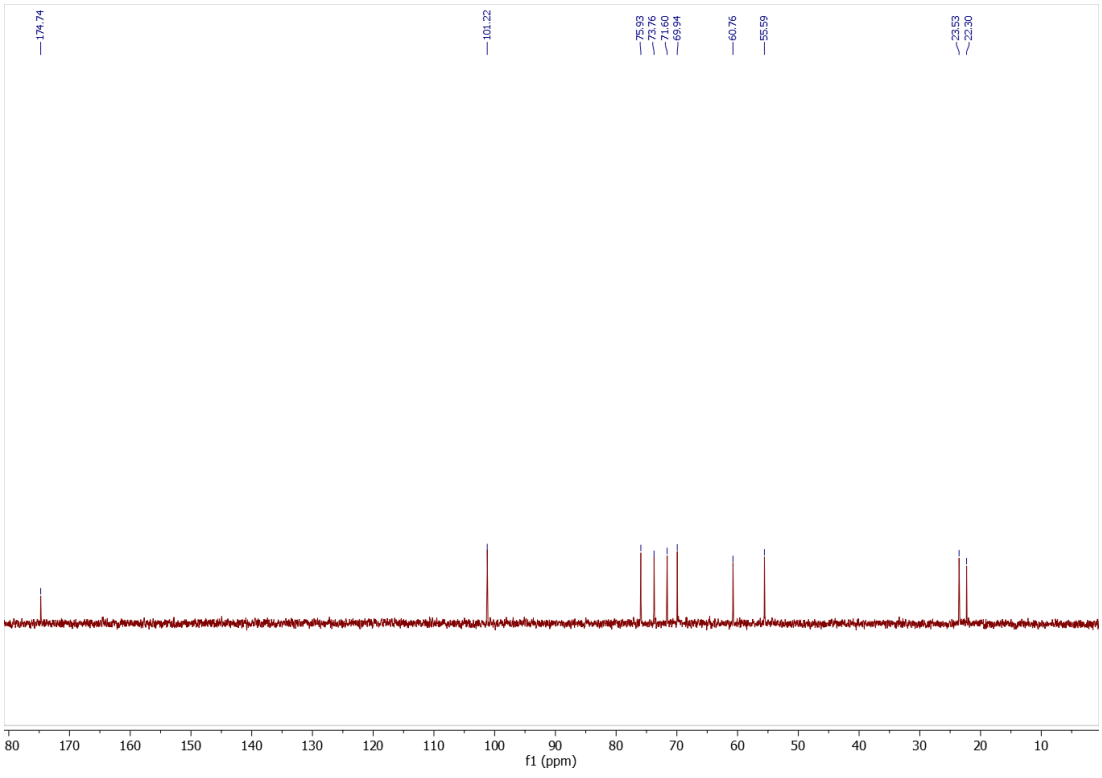
^{1}H NMR and ^{13}C NMR spectra of 1-(2-(thioacetate)-ethyl)-2-acetamido-2-deoxy-β-D-glucopyranosyl 1,3,4,6-tetraacetate (S5)





^{1}H NMR and ^{13}C NMR spectra of 1-(2-(thioacetate)-ethyl)-2-acetamido-2-deoxy-β-D-glucopyranosyl 1,3,4,6-tetraacetate (S5)





^{1}H NMR and ^{13}C NMR spectra of 2-Acetamido-2-deoxy-1-(2-thioethyl)-β-D-glucose (2)

5. Results of protein and peptide identifications obtained from the shotgun proteomics analysis of GYG-wt-0Glc

Protein names	Gene names	Peptides	Sequence coverage [%]	Mol. weight [kDa]	Q-value	Score	Intensity	MS/MS count	Protein IDs
		25	74.1	29.582	0	323.31	12.03294094	1140	GYG
		12	22.9	83.785	0	259.59	9.382125152	15	GYS
Ribonucleoside-diphosphate reductase 1 subunit alpha	nrdA	8	18.4	85.774	0	92.905	8.488409689	8	P00452;P00452-2
Beta-galactosidase	lacZ	3	3	116.48	0	45.169	7.792888742	3	P00722
Aspartate--ammonia ligase	asnA	2	8.5	36.65	0	30.3	7.934468151	2	P00963
30S ribosomal protein S7	rpsG	2	15.1	20.019	0	29.332	8.328379603	3	P02359
Chromosomal replication initiator protein DnaA	dnaA	2	3.9	52.55	0	26.018	7.200440076	2	P03004
DNA helicase II	uvrD	6	12.6	81.989	0	102.48	8.387140569	5	P03018
4-hydroxy-tetrahydrodipicolinate reductase	dapB	2	11.7	28.756	0	23.724	7.481241918	2	P04036
GMP synthase [glutamine-hydrolyzing]	guaA	1	2.1	58.679	0.005128	5.4364	7.001257631	0	P04079
D-lactate dehydrogenase	dld	1	1.4	64.612	0	6.4582	6.939504277	1	P06149
Pyruvate dehydrogenase [ubiquinone];Alpha-peptide	poxB	1	1.7	62.011	0	7.0505	7.350596865	1	P07003
tRNA pseudouridine synthase A	truA	3	15.9	30.399	0	84.205	8.003934206	3	P07649
L-threonine 3-dehydrogenase	tdh	1	2.3	37.239	0	5.5529	6.900858505	1	P07913
Bifunctional protein FolC;Folylpolyglutamate synthase;	folC	4	14.5	45.405	0	58.477	7.854548936	3	P08192
Phenylalanine--tRNA ligase alpha subunit	pheS	1	2.4	36.831	0	8.0278	7.129496543	1	P08312
Type I restriction enzyme EcoKI M protein	hsdM	4	12.3	59.306	0	43.384	8.034387784	4	P08957
Valine--pyruvate aminotransferase	avtA	2	6.7	46.711	0	27.055	7.305458855	2	P09053
Mannitol-1-phosphate 5-dehydrogenase	mtlD	2	9.7	41.139	0	10.53	7.428248214	2	P09424
Acetate kinase	ackA	6	27.2	43.29	0	102.35	8.471482407	6	P0A6A3
Cysteine desulfurase IscS	iscS	10	28.2	45.089	0	136.66	8.988732818	11	P0A6B7
ATP synthase epsilon chain	atpC	2	36.7	15.068	0	22.357	8.1703204	2	P0A6E6
Glycerol kinase	glpK	3	7.6	56.23	0	40.388	8.002382075	3	P0A6F3
60 kDa chaperonin	groL	6	21.9	57.328	0	72.31	8.655253719	8	P0A6F5
ATP-dependent Clp protease ATP-binding subunit ClpX	clpX	5	20	46.355	0	42.421	8.244252374	5	P0A6H1
Pantothenate kinase	coaA	1	2.8	36.359	0	5.9491	7.297782389	1	P0A6I3
Phosphopentomutase	deoB	1	3.4	44.369	0	24.718	7.324981703	1	P0A6K6
Elongation factor G	fusA	1	2.1	77.58	0	7.4216	6.847480115	0	P0A6M8
Glycerol-3-phosphate dehydrogenase [NAD(P)+]	gpsA	1	5.6	36.361	0	10.183	7.18335527	1	P0A6S7
GTP cyclohydrolase 1	folE	4	23.9	24.83	0	55.449	8.354242934	5	P0A6T5
Glucokinase	glk	2	8.4	34.723	0	13.389	8.185088643	2	P0A6V8
33 kDa chaperonin	hslO	2	14.4	32.534	0	35.641	8.219270086	2	P0A6Y5
Chaperone protein DnaK	dnaK	3	5.3	69.114	0	41.914	7.750909334	3	P0A6Y8
Chaperone protein HtpG	htpG	7	15.2	71.422	0	122.41	8.637049208	8	P0A6Z3
Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase	lpxA	2	10.7	28.08	0	47.274	9.067814511	3	P0A722
ATP-dependent 6-phosphofructokinase isozyme 1	pfkA	1	2.8	34.842	0	8.8727	7.252416091	1	P0A796
CTP synthase	pyrG	1	2.6	60.373	0	9.6182	7.109274738	1	P0A7E5
Recombination protein RecR	recR	1	4	21.963	0	10.102	7.38936103	1	P0A7H6
Peptide chain release factor 3	prfC	4	10.4	59.573	0	38.571	8.225335132	3	P0A7I4
3,4-dihydroxy-2-butanone 4-phosphate synthase	ribB	2	13.8	23.353	0	23.215	8.093666782	3	P0A7J0
50S ribosomal protein L10	rplJ	2	17.6	17.711	0	27.187	7.707502046	2	P0A7J3
30S ribosomal protein S10	rpsJ	2	30.1	11.735	0	33.261	8.299223923	4	P0A7R5
30S ribosomal protein S11	rpsK	2	31	13.845	0	48.187	8.521517272	3	P0A7R9
30S ribosomal protein S13	rpsM	4	38.1	13.099	0	100.06	8.861361136	8	P0A7S9
30S ribosomal protein S2	rpsB	5	35.3	26.743	0	105.22	9.068297205	8	P0A7V0
30S ribosomal protein S3	rpsC	8	40.3	25.983	0	190.63	9.957927982	17	P0A7V3
30S ribosomal protein S4	rpsD	5	34	23.469	0	103.86	8.870005717	6	P0A7V8

Protein names	Gene names	Peptides	Sequence coverage [%]	Mol. weight [kDa]	Q-value	Score	Intensity	MS/MS count	Protein IDs
30S ribosomal protein S5	rpsE	1	11.4	17.603	0	46.392	8.138870857	1	P0A7W1
Ribonuclease 3	rnc	1	4	25.55	0	14.244	7.559451754	1	P0A7Y0
Trigger factor	tig	1	2.3	48.192	0.01	5.3081	7.095273988	0	P0A850
Triosephosphate isomerase	tpiA	4	33.3	26.972	0	55.255	8.238221794	4	P0A858
Tryptophan synthase beta chain	trpB	7	30.7	42.983	0	62.849	8.778961178	7	P0A879
RNase adapter protein RapZ	rapZ	11	57.7	32.492	0	170.44	8.915674151	14	P0A894
tRNA (guanine-N(7)-)-methyltransferase	trmB	2	18.8	27.307	0	28.173	8.037864556	3	P0A8I5
ATP-dependent RNA helicase RhlB	rhlB	13	41.6	47.125	0	309.88	9.413785768	26	P0A8J8
Threonine--tRNA ligase	thrS	2	3.6	74.013	0	18.719	7.37542583	1	P0A8M3
Lysine--tRNA ligase	lysS	4	10.9	57.603	0	40.664	7.913273296	4	P0A8N3
Lysine--tRNA ligase, heat inducible	lysU	2	5	57.826	0	9.8442	7.032860416	1	P0A8N5
UPF0301 protein YqgE	yqgE	1	8.6	20.686	0	26.68	7.466511738	1	P0A8W5
Ferric uptake regulation protein	fur	1	8.1	16.795	0	21.944	7.113909944	1	P0A9A9
Glyceraldehyde-3-phosphate dehydrogenase A	gapA	3	14.8	35.532	0	43.096	8.249736305	3	P0A9B2
HTH-type transcriptional regulator CysB	cysB	1	3.1	36.15	0	11.713	7.037426498	1	P0A9F3
P-protein;Chorismate mutase;Prephenate dehydratase	pheA	4	9.1	43.111	0	51.179	8.194431083	3	P0A9J8
Phosphate-specific transport system accessory protein PhoU	phoU	4	20.7	27.417	0	45.372	7.927231736	3	P0A9K7
FKBP-type peptidyl-prolyl cis-trans isomerase SlyD	slyD	1	5.6	20.853	0	21.609	7.483373306	1	P0A9K9
Dihydrolipoyl dehydrogenase	lpdA	4	16	50.688	0	57.726	8.109544801	4	P0A9P0
Thioredoxin reductase	trxB	2	9.3	34.623	0	18.429	7.444762819	2	P0A9P4
Lactaldehyde reductase	fucO	1	3.9	40.513	0.010101	5.3295	7.668087586	1	P0A9S1
Uncharacterized ABC transporter ATP-binding protein YbiT	ybiT	1	1.3	59.857	0	6.1986	7.463639203	1	P0A9U3
Lipopolysaccharide export system ATP-binding protein LptB	lptB	8	37.8	26.8	0	94.093	8.575106985	10	P0A9V1
Rod shape-determining protein MreB	mreB	9	35.2	36.952	0	133.99	8.878291949	11	P0A9X4
50S ribosomal protein L13	rplM	2	14.1	16.018	0	24.162	8.242118854	2	P0AA10
Ribosomal small subunit pseudouridine synthase A	rsuA	3	31.2	25.865	0	42.933	7.8460772	3	P0AA43
UTP--glucose-1-phosphate uridylyltransferase	galF	1	4	32.829	0	9.0313	7.144792078	1	P0AAB6
Fructose-bisphosphate aldolase class 2	fbaA	1	4.7	39.147	0	6.1425	7.105305763	1	P0AB71
Phospho-2-dehydro-3-deoxyheptonate aldolase, Phe-sensitive	aroG	10	50	38.009	0	186.37	9.419030029	15	P0AB91
ATP synthase gamma chain	atpG	4	11.5	31.577	0	27.676	8.086359831	4	P0ABA6
Poly(A) polymerase I	pcnB	1	1.7	53.87	0	6.9016	NaN	1	P0ABF1
Lipid II flippase FtsW	ftsW	1	1.9	45.987	0.00995	5.3006	6.891470552	0	P0ABG4
Cell division protein FtsA	ftsA	3	19.8	45.329	0	52.533	8.215637563	3	P0ABH0
ATP-dependent Clp protease ATP-binding subunit ClpA	clpA	3	3.4	84.206	0	28.262	7.824028156	1	P0ABH9
D-erythro-7,8-dihydroneopterin triphosphate epimerase	folX	1	8.3	14.082	0	7.2286	7.245685364	0	P0AC19
Fumarate hydratase class I, aerobic	fumA	1	2	60.298	0	7.1361	7.324549948	1	P0AC33
Succinate dehydrogenase flavoprotein subunit	sdhA	6	17.3	64.421	0	129.52	8.635523957	7	P0AC41
GDP-mannose 4,6-dehydratase	gmd	22	70.5	42.046	0	323.31	10.31228317	52	P0AC88
Iron-sulfur cluster insertion protein ErpA	erpA	2	16.7	12.1	0	13.09	7.589010779	1	P0ACC3
cAMP-activated global transcriptional regulator CRP	crp	3	14.8	23.64	0	44.485	9.163578765	5	P0ACJ8
Glycerol-3-phosphate regulon repressor	glpR	4	18.7	28.048	0	72.88	8.353666233	5	P0ACL0
Exu regulon transcriptional regulator	exuR	2	15.5	29.835	0	40.856	7.812539034	2	P0ACL2
HTH-type transcriptional regulator GntR	gntR	1	2.4	36.422	0	6.2094	8.947566049	1	P0ACP5
HTH-type transcriptional repressor PurR	purR	1	3.2	38.174	0	21.652	7.720374329	1	P0ACP7
Hydrogen peroxide-inducible genes activator	oxyR	3	13.8	34.276	0	24.771	7.782615932	2	P0ACQ4
Transcriptional repressor MprA	mprA	1	8.5	20.563	0	6.6079	7.161008437	1	P0ACR9
Lipid A biosynthesis palmitoleoyltransferase	lpxP	1	2.3	35.493	0.005102	5.4065	7.012837225	0	P0ACV2
Uncharacterized protein YecJ	yecJ	3	54.2	9.1225	0	39.992	8.254040258	4	P0AD10

Protein names	Gene names	Peptides	Sequence coverage [%]	Mol. weight [kDa]	Q-value	Score	Intensity	MS/MS count	Protein IDs
Enterobactin synthase component B;Isochorismatase;Aryl carrier protein	entB	1	5.6	32.554	0	15.228	7.674153574	1	P0ADI4
Uncharacterized protein YiiD	yiiD	1	5.5	37.094	0	20.343	7.499013129	1	P0ADQ2
Ribosomal RNA large subunit methyltransferase M	rlmM	1	2.2	41.905	0.005076	5.3531	7.099058789	1	P0ADR6
LOG family protein YgdH	ygdH	1	2.4	50.971	0	8.6882	7.077876404	1	P0ADR8
Uncharacterized protein YhbW	yhbW	5	31.3	37.129	0	91.694	8.560910443	6	P0ADV5
Magnesium and cobalt efflux protein CorC	corC	4	25.3	33.298	0	77.055	8.36599312	4	P0AE78
DNA-cytosine methyltransferase	dcm	2	5.9	53.464	0	12.25	7.13219563	1	P0AED9
tRNA-2-methylthio-N(6)-dimethylallyladenosine synthase	miaB	4	12	53.662	0	53.511	8.36574998	3	P0AEI1
Ribosomal protein S12 methylthiotransferase RimO	rimO	3	10.2	49.581	0	38.585	8.50609896	3	P0AEI4
Inosine-guanosine kinase	gsk	1	3.5	48.448	0	18.607	7.058919261	1	P0AEW6
Maltose-binding periplasmic protein	malE	1	2.8	43.387	0.005155	5.4529	7.184975191	1	P0AEX9
Molybdenum cofactor biosynthesis protein B	moaB	1	8.2	18.665	0	10.069	7.450895449	1	P0AEZ9
Protein mrp	mrp	1	4.6	39.938	0	14.717	8.25190577	2	P0AF08
N-acetylglucosamine repressor	nagC	1	2.5	44.541	0	6.2445	7.196590654	1	P0AF20
Cell division protein ZapB	zapB	1	25.9	9.6347	0	15.189	7.385141348	1	P0AF36
2-oxoglutarate dehydrogenase E1 component	sucA	5	7.3	105.06	0	80.22	8.471775621	5	P0AFG3
Pyruvate dehydrogenase E1 component	aceE	1	1.5	99.667	0	18.66	7.407254901	1	P0AFG8
Pyridoxine/pyridoxamine 5-phosphate oxidase	pdxH	6	38.5	25.545	0	45.587	8.168880342	3	P0AFI7
Negative modulator of initiation of replication	seqA	1	9.4	20.315	0	15.321	7.206960729	1	P0AFY8
Stringent starvation protein B	sspB	3	21.2	18.262	0	55.296	9.375279436	6	P0AFZ3
Riboflavin biosynthesis protein RibF;Riboflavin kinase;FMN adenylyltransf	ribF	9	50.2	34.734	0	151.57	8.733413957	9	P0AG40
50S ribosomal protein L17	rplQ	1	11.8	14.364	0	7.0089	7.568741322	1	P0AG44
30S ribosomal protein S14	rpsN	1	18.8	11.58	0	27.54	8.993467093	4	P0AG59
Protein-export protein SecB	secB	3	31.6	17.277	0	88.96	7.613936912	4	P0AG86
tRNA (guanosine(18)-2-O)-methyltransferase	trmH	1	5.2	25.343	0	5.495	7.302287627	1	P0AGJ2
HTH-type transcriptional regulator IscR	iscR	1	13.6	17.336	0	11.184	7.22605831	1	P0AGK8
Protein sirB1	sirB1	1	4.5	30.841	0	7.0588	6.97337727	1	P0AGM5
Small heat shock protein IbpA	ibpA	1	6.6	15.774	0	10.501	6.868544522	1	P0C054
D-tagatose-1,6-bisphosphate aldolase subunit GatZ	gatZ	3	10.5	47.108	0	20.171	7.842864999	2	P0C8J8;P0C8K0
Elongation factor Tu 2;Elongation factor Tu 1	tufB;tufA	16	61.9	43.313	0	290.27	10.05296313	36	P0CE48;P0CE47
UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase	murF	1	2.4	47.447	0	9.5862	7.413048169	1	P11880
Selenocysteine-specific elongation factor	selB	8	21.2	68.867	0	120.93	8.700824818	9	P14081
Helicase IV	helD	2	2.2	77.975	0	19.534	6.913692021	1	P15038
L-serine dehydratase 1	sdaA	3	6.6	48.906	0	19.315	7.125090713	3	P16095
Arabinose 5-phosphate isomerase GutQ	gutQ	4	32.7	34.031	0	35.773	7.930143848	4	P17115
Glutamine--fructose-6-phosphate aminotransferase [isomerizing]	glmS	3	7.2	66.894	0	28.305	7.902650041	3	P17169
Oxygen-dependent choline dehydrogenase	betA	4	12.8	61.877	0	68.029	8.031973689	3	P17444
UDP-N-acetylmuramate--L-alanine ligase	murC	4	15.3	53.625	0	86.97	7.978235314	4	P17952
Ubiquinone biosynthesis O-methyltransferase	ubiG	1	7.9	26.555	0	10.647	6.812425336	1	P17993
Sulfate adenylyltransferase subunit 2	cysD	2	11.6	35.188	0	24.861	7.713734083	2	P21156
Ribonuclease E	rne	4	3.8	118.2	0	36.395	7.808332531	2	P21513
Maltodextrin glucosidase	malZ	1	1.5	69.04	0	5.9383	6.950014161	1	P21517
Pyruvate kinase II	pykA	10	33.1	51.357	0	209.75	9.253870852	15	P21599
UDP-3-O-(3-hydroxymyristoyl)glucosamine N-acyltransferase	lpxD	1	7.6	36.038	0	7.6783	7.793545938	1	P21645
UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2,6-diaminopimelate ligas	murE	2	6.1	53.343	0	23.911	7.718244191	2	P22188
tRNA/tmRNA (uracil-C(5))-methyltransferase	trmA	1	8.2	41.966	0	21.441	7.861731501	1	P23003
Transcriptional regulatory protein PhoP	phoP	1	6.3	25.535	0.01005	5.3108	6.445635334	0	P23836
Sulfate adenylyltransferase subunit 1	cysN	4	15.4	52.558	0	54.288	7.717279295	3	P23845

Protein names	Gene names	Peptides	Sequence coverage [%]	Mol. weight [kDa]	Q-value	Score	Intensity	MS/MS count	Protein IDs
Mannose-1-phosphate guanylyltransferase	manC	3	13.2	53.016	0	34.144	7.948261327	3	P24174
Biotin carboxylase	accC	4	12	49.32	0	40.918	8.27510395	4	P24182
DNA repair protein RadA	radA	4	13.3	49.472	0	44.71	8.190695792	3	P24554
Riboflavin biosynthesis protein RibD;	ribD	1	3.3	40.338	0	16.566	7.461273368	1	P25539
Guanosine-5-triphosphate,3-diphosphate pyrophosphatase	gppA	1	3.8	54.871	0	6.8354	6.75882109	1	P25552
ATP-dependent RNA helicase RhlE	rhlE	10	24.7	49.989	0	140.79	9.136117843	11	P25888
Soluble pyridine nucleotide transhydrogenase	sthA	8	38.4	51.56	0	235.38	9.089304452	10	P27306
Trehalose-6-phosphate hydrolase	treC	4	10.3	63.837	0	44.818	8.642820516	5	P28904
L-serine dehydratase 2	sdaB	1	2	48.752	0	6.7703	7.309672743	1	P30744
Methionine import ATP-binding protein MetN	metN	4	21	37.788	0	33.24	7.932575223	3	P30750
Inorganic triphosphatase	ygiF	2	4.6	48.388	0	11.756	7.848872362	1	P30871
Molecular chaperone Hsp31 and glyoxalase 3	hchA	2	11	31.19	0	21.479	7.826249782	2	P31658
GDP-L-fucose synthase	fcl	3	19	36.141	0	45.743	8.036987998	3	P32055
Formate dehydrogenase-O major subunit	fdoG	1	1.3	112.55	0	7.2822	7.37821615	1	P32176
NADH pyrophosphatase	nudC	3	16.3	29.689	0	48.144	8.025919999	4	P32664
Uncharacterized HTH-type transcriptional regulator YijO	yijO	1	3.9	32.149	0.005181	5.4766	6.636858844	0	P32677
NADH-quinone oxidoreductase subunit C/D	nuoC	1	2.5	68.235	0	5.7317	7.142827295	1	P33599
NADH-quinone oxidoreductase subunit G	nuoG	5	7.9	100.3	0	37.438	8.338755217	3	P33602
Alkyl hydroperoxide reductase subunit F	ahpF	4	12.5	56.176	0	43.734	8.289031235	5	P35340
HTH-type transcriptional regulator TreR	treR	1	5.1	34.53	0	10.407	7.08228259	1	P36673
Aconitate hydratase B	acnB	3	4.3	93.497	0	20.188	7.705359022	2	P36683
Formyltetrahydrofolate deformylase	purU	8	42.9	31.934	0	110.89	9.271679006	14	P37051
dTDP-4-dehydrorhamnose 3,5-epimerase	rfbC	1	4.9	21.27	0	8.3265	6.907282343	1	P37745
UDP-N-acetylmuramate--L-alanyl-gamma-D-glutamyl-meso-2,6-diaminoh	mpl	2	9	49.874	0	10.831	7.477642096	1	P37773
Universal stress protein G	uspG	1	11.3	15.935	0	9.7734	6.987755617	1	P39177
Ribosomal RNA large subunit methyltransferase G	rlmG	7	36.2	42.331	0	137.34	9.381205361	13	P42596
Arabinose 5-phosphate isomerase KdsD	kdsD	6	32.6	35.196	0	103.84	8.783510536	8	P45395
RNA chaperone ProQ	proQ	2	12.9	25.892	0	19.043	7.906582831	2	P45577
50S ribosomal protein L2	rplB	2	10.3	29.86	0	11.474	8.2511757	2	P60422
50S ribosomal protein L3	rplC	1	9.6	22.243	0	10.234	7.214366922	1	P60438
Lipoyl synthase	lipA	2	14	36.071	0	12.607	7.761115595	1	P60716
50S ribosomal protein L22	rplV	2	18.2	12.226	0	20.783	8.042851043	2	P61175
Carbonic anhydrase 2	can	4	23.2	25.096	0	53.553	8.83773991	4	P61517
50S ribosomal protein L5	rplE	2	11.2	20.301	0	17.706	7.412174844	2	P62399
Chaperone protein ClpB	clpB	3	6.3	79.869	0	32.438	7.468716472	3	P63284-2;P63284
Ribosomal RNA small subunit methyltransferase I	rsmI	4	21	31.348	0	63.454	8.219217657	6	P67087
HTH-type transcriptional repressor NemR	nemR	2	10.6	22.275	0	11.953	7.572581201	1	P67430
Long-chain-fatty-acid--CoA ligase	fadD	3	8.9	62.331	0	38.4	7.871333541	3	P69451
Probable diguanylate cyclase YeaP	yeaP	1	4.7	38.545	0	13.772	6.835341728	1	P76245
tRNA (cmo5U34)-methyltransferase	cmoA	7	41.7	27.776	0	73.89	8.376631697	6	P76290
D-cysteine desulfhydrase	dcyD	3	12.2	35.153	0	41.406	8.465472081	3	P76316
Uncharacterized protease YegQ	yegQ	5	16.3	51.193	0	54.605	8.369531256	5	P76403
1-deoxy-D-xylulose-5-phosphate synthase	dxs	1	1.5	67.616	0	6.7601	6.771065426	1	P77488
Succinylornithine transaminase	astC	7	33.3	43.665	0	119.06	8.33611937	6	P77581