Studying protein interaction and glycosylation with chemical proteomics

A thesis submitted in partial fulfilment for the degree of

Doctor of Philosophy

Simon Nadal

Department of Chemistry

University of Oxford

Trinity Term 2020

The Queen's College

Abstract

Protein glycosylation is present in all kingdoms of life, from small virus particles to complex living organisms. It is the most abundant and diverse form of post-translational modification, ranging from the attachment of a single monosaccharide to entire sugar chains onto cellular proteins. This modification affects the recognition and interaction of proteins with their interaction partners, altering their role in health and diseases. Despite its importance, the study of protein glycosylation has lacked behind other biomolecules and post-translational modifications, partly because of absence of available tools and methods. This doctoral thesis aims at expanding existing analytical approaches and developing new technologies to study the different types of protein glycosylation using tandem mass spectrometry.

The second chapter demonstrates *de novo* determination of the sequence and protein *N*-glycosylation of the anti-γH2AX monoclonal antibody, enabling structural modelling of the antigen recognition and laying the groundwork for further development of this diagnostic reagent for DNA damage in cancers. Chapter 3 focusses on improving analysis of protein *O*-glycosylation using tandem mass spectrometry. The use of low-energy stepped collision energy (SCE) methods is described for improved characterisation of the glycan part and site-localisation of protein *O*-glycosylation. SCE-HCD methods such as 10-20% of normalised collision energy (NCE) show the formation of valuable fragment ions enabling a more accurate identification of *O*-glycopeptides.

The *O*-GlcNAc modification is a particular form of *O*-glycosylation, consisting in the attachment of a single *N*-acetylglucosamine sugar to proteins. In chapter 4, the role of *O*-GlcNAc as an epigenetic mark on H2B-S36 is investigated. Homogeneously GlcNAcylated histones and nucleosomes were synthesised and used in pull-down experiments to enrich putative interaction partners for the H2B-S36-GlcNAc modification. Mass spectrometry-based quantitative interaction proteomics revealed a direct interaction between H2B-S36-*O*-GlcNAcylated nucleosomes and the TCF20 transcription factor, suggesting a possible signalling role for the *O*-GlcNAc modification. Lastly, chapter 5 shows that synthetic histones with reactive side chains can be used to covalently trap interaction partners in a native context, identifying their binding site by crosslinking mass spectrometry with amino-acid resolution inside the interaction.

Acknowledgements

First, I would like to thank Prof. Ben Davis for accepting me in his lab, for mentorship and supervision as well as for showing me that the big questions in science are the most rewarding ones. Also, for teaching me never to judge too early and to always keep an open mind.

My deep gratitude goes to Prof. Shabaz Mohammed for his patience, for guiding me when I was loosing focus. Also, for always being able to point me in the right direction no matter which abstruse mass spectrometry question or task I had. Shabaz' expertise in fragmentations has truly been an asset to this thesis I cannot emphasise enough. Last but not least, thank you for keeping the proteomics instruments and infrastructure running along the clock, while always keeping high standards in MS data quality.

A heartfelt thank-you (!!!) to Svenja Hester for incredible support and advise on sample handling and preparation (and jewellery), in which her knowledge is unmatched, and for teaching me how to repair nanoLC's. Thank you for being a good friend − take it from Mannie (he does have the haircut).

My gratitude goes to Dr. Philip Charles for immense knowledge in bioinformatic data analysis, always provided with a big smile. And for being one of the persons working tirelessly in the background to keep IT infrastructure running.

A special thanks goes to Tobi Arisa and Marielle Tamigney for being good and understanding friends.

My gratitude goes to numerous collaborators on fruitful collaborations: Brian Josephson, Alexandra Gafitescu, Dr. Abirami Lakshminarayanan, Dr. Marielle Tamigney, Dr. Charlie Fehl, Yue-Jin Oh, Dr. Ritu Raj, Dr. Alberto Fernandez-Tejada; some of the results are shown in this thesis. Equally, I thank Dounia Krouch for doing her Master's thesis with me, her hard work, her optimism and joyful participation. Thanks to all members of the BGD group, past and present, particularly Seb, Yibo, Nan, Sarah, Jitka, Artem, Amy, Matt, Sadra for lab fun and passionate discussions. Equally, members of the SM group past & present: Monika, Caroline, Sabrina, Deniz, Kasia, Chen and Wael.

Many have supported this thesis on technical subjects. A special thank you to Dr. Sanjay Sanghani, Protein Metrics Inc, for being the best sales rep ever − and turning a "no" into a "yes", which meant a lot for this thesis. Equally, thanks to Dr. Marshall Bern, Protein Metrics Inc, for implementing proteoglycan support into Byonic. I also thank the mass spectrometry service of the Department of Chemistry for support, instrumentation and their incredible patience − particularly Dr. Elisabete Pires for using her network to get access to the PeaksAB software. In addition, I want to acknowledge IT and the Chemistry workshops, particularly Ron, Jude and John, for showing me many tricks in repairing instruments.

I thank my examiners, Prof. Albert Heck and Prof. Justin Benesch, for taking the time to evaluate this thesis and for insightful discussions and feedback.

I want to express my deep gratitude to the FR goup at the TDI for adopting and encouraging me while finishing writing and to Prof. Fraydoon Rastinejad for his patience.

Finally, I thank my family for steady support and acceptance during the last years, and my fiancée Yuliya for bearing with me through highs and lows. Lastly, Harry thanks countless tourists for a small contribution towards this thesis.

Declaration

I declare the work presented in this thesis was carried out at the Chemistry Research Laboratory and Department of Biochemistry, University of Oxford, UK. All the work is my own, except where otherwise stated, and has not been submitted for any other degree at this or any other university or academic institution.

Oxford, November 6$^{\text{th}}$, 2020

This DPhil thesis was carried out under the supervision of:

1. Prof. Benjamin G. Davis
2. Prof. Shabaz Mohammed

and examined by

1. Prof. Albert J. R. Heck, University of Utrecht
2. Prof. Justin L. P. Benesch, University of Oxford

À Pedro Nadal
und meiner Familie

Short contents

Abstract ii

Acknowledgements iii

Short contents vii

Detailed contents viii

1. Introduction 1

I. Tandem mass spectrometry methods to study glycosylated proteins 106

2. From tube to 3D model: *de novo* sequencing, glycosylation analysis and modelling of the anti-γH2AX murine antibody 107

3. Energy-resolved fragmentation of *O*-linked glycopeptides 215

II. Chemical proteomics for the identification of interaction partners of histone PTMs and their interaction site 315

4. Synthetic nucleosomes reveal H2B-Ser36 interaction partners 316

5. Site-specific proximity-enabled protein crosslinking allows identification of protein-protein interaction site 398

6. General appendix 454

Detailed contents

Abstract ii

Acknowledgements iii

Short contents vii

Detailed contents viii

1. Introduction 1

1.1. General introduction: protein glycosylation 2

1.1.1. Protein glycosylation is a highly variable and diverse post-translational modification 2

1.1.2. Conserved structural elements in protein glycosylation 6

1.1.2.1. Available monosaccharides and glycosyltransferases limit the combinatorial diversity 6

1.1.2.2. Protein glycosylation has conserved linkages and cores . . 6

1.1.2.3. Protein glycosylation has conserved epitopes 16

1.1.3. The importance of protein glycosylation in modern medicine . . . 18

1.2. Objective 20

1.2.1. Discoveries in glycobiology are limited by lack of analytical tools . 20

1.2.2. Aim of this thesis: MS/MS and chemical proteomics methods to study protein glycosylation and PTM-dependent protein-protein interaction 21

1.3. Technical introduction: proteomics and glycoproteomics 23

1.3.1. Proteomics, glycomics and glycoproteomics 23

1.3.2. Sample preparation and processing 25

1.3.2.1. Isolation of proteomes and sub-proteomes from biological samples 25

1.3.2.2. Sample preparation prior to MS analysis 25

1.3.2.3. Fractionation, enrichment at peptide level 30

1.3.3. Nano liquid chromatography (nLC) 34

1.3.4. Mass spectrometry 36

1.3.4.1. Ionisation techniques 36

1.3.4.2. Mass analysers and detectors . . . 37
1.3.4.3. Tandem mass spectrometry & peptide fragmentation . . 41
1.3.4.4. Data acquisition methods . . . 48
1.3.5. Data analysis: peptide identification and quantitation . . . 52
1.3.5.1. Preprocessing . . . 52
1.3.5.2. Database searching . . . 52
1.3.5.3. De-novo sequencing . . . 56
1.3.5.4. Quantitative proteomics . . . 57
1.3.5.5. Crosslinking tandem mass spectrometry . . . 60
1.3.6. Proteomics applications and their respective post-processing methods . . . 65
1.3.6.1. Proteome profiling . . . 65
1.3.6.2. Spacial proteomics . . . 66
1.3.6.3. Functional analysis . . . 66
1.3.6.4. Protein-protein interaction partners . . . 67
1.3.6.5. PTM analysis . . . 68
1.3.6.6. Structural proteomics . . . 70
1.3.6.7. Chemical proteomics . . . 71
1.3.6.8. Multi-omics data integration . . . 73
1.4. References . . . 74

I. Tandem mass spectrometry methods to study glycosylated proteins 106

2. From tube to 3D model: *de novo* sequencing, glycosylation analysis and modelling of the anti-γH2AX murine antibody 107
2.1. Introduction . . . 110
2.1.1. γH2AX is a mark for DNA double-strand break . . . 110
2.1.2. Structure of monoclonal IgG antibodies . . . 112
2.1.3. *De novo* sequencing of monoclonal antibodies . . . 114
2.1.3.1. RNA sequencing of the hybridoma cell line . . . 114
2.1.3.2. *De novo* sequencing of purified antibodies by LC-MS/MS 115
2.1.4. Glycosylation analysis of monoclonal antibodies . . . 117
2.1.5. Antibody molecular modelling . . . 118
2.2. Objective: Project description and analytical strategy . . . 120
2.3. Materials & methods . . . 122
2.3.1. General methods . . . 122
2.3.2. Intact denaturing protein LC-MS . . . 122

2.3.3. *De novo* antibody sequencing by MS/MS − Experimental 123
2.3.3.1. Sample preparation for *de novo* sequencing 123
2.3.3.2. Standard data acquisition − Q Exactive 124
2.3.4. *De novo* antibody sequencing by MS/MS − data analysis 125
2.3.4.1. Short description of the PEAKS software and its key modules 125
2.3.4.2. Description of the *de novo* sequencing methodology by MS/MS using the ALPS assembler 126
2.3.4.3. *De novo* sequencing and sequence validation using PeaksAB 2.0 128
2.3.5. Glycosylation analysis 129
2.3.5.1. Sample preparation for glycosylation analysis 129
2.3.5.2. Data acquisition for glycosylation analysis – Orbitrap Elite 129
2.3.5.3. Data analysis for glycosylation analysis 131
2.3.6. RNA sequencing 131
2.3.7. Antibody Molecular Modelling 131
2.3.7.1. Homology model of the variable region 131
2.3.7.2. Data mining of the RCSB PDB 132
2.3.7.3. Docking of the peptide antigen in the F_V antibody receptor 132
2.3.7.4. Homology model for the entire anti-γH2Ax monoclonal antibody 133
2.4. Results 135
2.4.1. Assessing *de novo* sequencing tools on the Herceptin test system . 135
2.4.1.1. Both ALPS and PeaksAB successfully assemble Herceptin's light chain 136
2.4.1.2. PeaksAB outperforms ALPS for the Herceptin heavy chain assembly 140
2.4.2. *De novo* sequencing of the anti-γH2AX mAb light chain 143
2.4.2.1. Manual correction of the ALPS assembly allows successful sequencing 143
2.4.2.2. PeaksAB has substantial difficulties with the CDR-L and CL regions 148
2.4.3. Manual sequencing of the anti γ-H2Ax heavy chain 150
2.4.3.1. Generation of a consensus template 150
2.4.3.2. Correction of the consensus template in the variable domain 152
2.4.3.3. Investigating sequence coverage gaps in the constant domain 158

2.4.3.4. Comparison with the RNAseq and intact protein MS results 163
2.4.3.5. PeaksAB fails on most of the anti-γH2Ax heavy chain variable domain . . . 163
2.4.3.6. PTM analysis goes hand in hand with *de novo* sequencing 166
2.4.4. Lessons from the *de novo* sequencing and discussion of its future . 168
2.4.5. Glycosylation analysis . . . 171
2.4.6. Molecular modelling of the antibody structure . . . 175
2.4.6.1. Homology modelling and antigen-antibody docking . . . 175
2.4.6.2. Bridging the gap: Combining structural and germline frequency data to explain antibody binding specificity . . 180
2.4.6.3. Molecular modelling of the full anti-γH2AX mAb . . . 184
2.5. Outlook . . . 187
2.6. Conclusion . . . 188
2.7. Author contributions . . . 190
2.8. Acknowledgements . . . 190
2.9. Conflict of interest . . . 191
2.10. Confidentiality declaration . . . 191
2.11. Appendix . . . 192
2.11.1. Copy of code and scripts used in this chapter . . . 192
2.11.2. Advice for manual *de novo* sequencing by Bioinformatic Solutions 193
2.11.3. RNA sequencing report from Absolute Antibody . . . 194
2.11.4. Antibody Homology Modelling – Supplementary information . . . 197
2.11.5. Full Haddock docking results − short peptide . . . 198
2.11.6. Full Haddock docking results − longer peptide . . . 202
2.12. References . . . 205

3. Energy-resolved fragmentation of *O*-linked glycopeptides 215
3.1. Abstract . . . 216
3.2. Introduction: Glycopeptide fragmentation and analysis . . . 217
3.2.1. Fragmentation mechanisms of glycopeptides . . . 217
3.2.1.1. Nomenclature and general aspects . . . 217
3.2.1.2. CID-based fragmentation mechanisms . . . 219
3.2.2. ETD-based fragmentation mechanisms . . . 224
3.2.3. The complementarity paradigm of glycopeptide fragmentation . . 228
3.2.4. Hybrid fragmentation methods . . . 228

3.2.5. Energy-resolved fragmentation of glycopeptides 230
3.2.5.1. Fragmentation energetics differ between peptides and glycopeptides 230
3.2.5.2. Activation and dissociation channels are energetically linked to structural features in *N*-glycopeptides 231
3.2.5.3. Stepped collision energy methods for fragmentation of *N*-glycopeptides 233
3.2.5.4. Difference in optimal dissociation methods for *N*- and *O*-glycopeptides are related to their distinct structures . 235
3.2.6. Data analysis and interpretation of glycopeptide MS/MS spectra . 237
3.3. Objective 239
3.4. Materials & methods 240
3.4.1. Sample preparation 240
3.4.2. Tandem mass spectrometry methods 241
3.4.2.1. Preparation of nLC columns for electrospray ionisation . 241
3.4.2.2. Data acquisition for glycosylated samples 242
3.4.3. Data analysis 245
3.4.3.1. Search parameters 245
3.4.3.2. Identification of novel *O*-glycopeptides and database curation 246
3.4.3.3. Rationale and flaws for the selected Byonic scoring metrics 246
3.5. Results 249
3.5.1. A case study: Identifying the linker *O*-glycopeptides of heparinase-trimmed decorin 249
3.5.2. Investigating the generality of low-NCE energy-resolved fragmentation of *O*-glycopeptides 262
3.5.2.1. Single NCE runs pinpoint the collision energy range for glycan fragmentation of *O*-glycopeptides between 10 and 20 % 262
3.5.2.2. Low-energy SCE-HCD methods are advantageous for *O*-glycopeptide glycan sequencing 266
3.5.2.3. Low-energy HCD methods affect the fragmentation and scoring of other components 272
3.5.3. ETD and EThcD methods remain essential for glycan site localisation 275
3.5.3.1. Implementing EThcD on an Orbitrap Elite instrument for glycopeptide analysis 275

3.5.3.2. Analysis of the model glycopeptides by ETD and EThcD 277
3.6. Conclusion and outlook . . . 280
3.7. Author contributions . . . 281
3.8. References . . . 282
3.9. Appendix . . . 290
3.9.1. Supplementary discussion . . . 290
3.9.1.1. Supplementary discussion 1: Collision energy values in the literature and their conversion to Thermo's NCE . . 290
3.9.2. Supplementary figures, tables & lists . . . 292
3.9.2.1. List of Byonic keywords . . . 292
3.9.2.2. Data for the decorin case study . . . 293
3.10. Supplementary references . . . 314

II. Chemical proteomics for the identification of interaction partners of histone PTMs and their interaction site 315

4. Synthetic nucleosomes reveal H2B-Ser36 interaction partners 316
4.1. Abstract . . . 317
4.2. Introduction . . . 318
4.2.1. Synthetic posttranlational modification of histones . . . 318
4.2.1.1. Background and motivation . . . 318
4.2.1.2. Searching for histone modifications using tandem mass spectrometry . . . 320
4.2.1.3. The chemical biology toolkit . . . 321
4.2.1.4. Application of chemical biology tools for the study of chromatin biology . . . 325
4.2.1.5. Probing interaction partners and downstream spreading of modifications . . . 326
4.2.2. *O*-GlcNAc is part of the histone code − the case of H2B-Ser36 . . 328
4.2.3. Aim of the project . . . 331
4.3. Results & Discussion . . . 334
4.3.1. Synthesis of post-translationally modified histone H2B-Ser36 variants . . . 334
4.3.2. Tandem mass spectrometry of modified proteins . . . 339
4.3.3. Reconstitution of histone octamers with H2B-Ser36 variants . . . 341
4.3.4. Generation of Widom 601 nucleosome-positioning DNA . . . 343
4.3.4.1. Generation of 145bp DNA and 186 bp biotinylated DNA 343

4.3.4.2. DNA degradation and subsequent purification 344
4.3.5. Nucleosome assembly 346
4.3.6. Interaction proteomics with synthetic nucleosomes 349
4.4. Outlook 355
4.5. Author contributions 359
4.6. Experimental 360
4.6.1. Materials 360
4.6.1.1. Chemicals and enzymes 360
4.6.1.2. Plasmids 360
4.6.1.3. Primers 361
4.6.2. Expression and purification of histone proteins 362
4.6.3. Bacterial transformation 364
4.6.4. Plasmid amplification 365
4.6.5. Expression and purification of histone proteins 365
4.6.6. Post-translational chemical mutagenesis 367
4.6.6.1. Synthesis of H2B-Dha36 protein 367
4.6.6.2. Synthesis of H2B-CysGlcNAc36 368
4.6.6.3. Attempts for the synthesis of thiophosphorylated histone H2B-pCys36 369
4.6.6.4. Synthesis of H2B-Pfa36 371
4.6.7. LC-MS/MS of modified histone proteins 373
4.6.8. Histone octamer assembly 374
4.6.9. Large scale DNA preparation for nucleosome reconstitution 376
4.6.9.1. Preparation of the 145 bp "601" DNA 376
4.6.9.2. Preparation of the 186 bp biotinylated "601" DNA 378
4.6.9.3. Southern Blot with anti-biotin detection 380
4.6.9.4. Purification of DNA fragments 381
4.6.10. Nucleosome reconstitution 382
4.6.11. Preparation of nuclear cell extract 383
4.6.12. Interaction proteomics experiment with synthetic nucleosomes . . 384
4.7. References 387

5. Site-specific proximity-enabled protein crosslinking allows identification of protein-protein interaction site 398

5.1. Introduction: Chemical crosslinking using unnatural amino acids 399

5.2. Results . . . 404
5.2.1. Generation of synthetic alkylator protein eH3 containing Bromohomonorleucine unnatural amino acid . . . 404
5.2.2. Protein-protein crosslink with the unnatural amino acid bromohomonorleucine . . . 404
5.2.3. Identified crosslinks show nucleophilicty of cysteines at Zn-binding site . . . 407
5.2.3.1. Crosslinking tandem mass spectrometry for low abundance xLink . . . 407
5.2.3.2. Parameters for xLink MS/MS data analysis . . . 408
5.2.3.3. Identification of cysteines at Zn-binding site . . . 410
5.2.3.4. Identification of inter-histone crosslinking sites show crosslinking reactivity beyond cysteine . . . 412
5.3. Conclusion & outlook . . . 414
5.4. Author contributions . . . 417
5.5. Experimental . . . 418
5.5.1. General Experimental Procedures . . . 418
5.5.2. Histone protein expression, chemical modification and crosslinking 419
5.5.3. Tandem mass spectrometry . . . 429
5.6. References . . . 433
5.7. Appendix: MS/MS spectra . . . 436

6. General appendix **454**
6.1. Abbreviations . . . 454
6.2. List of Figures . . . 463
6.3. List of Tables . . . 469
6.4. SNFG monosaccharide symbols & codes . . . 470

Evolution has failed to generate a living cell devoid of surface glycosylation.

— Ajit Varki

1

Introduction

Contents

1.1. General introduction: protein glycosylation . . . 2
1.1.1. Protein glycosylation is a highly variable and diverse post-translational modification . . . 2
1.1.2. Conserved structural elements in protein glycosylation . . . 6
1.1.2.1. Available monosaccharides and glycosyltransferases limit the combinatorial diversity . . . 6
1.1.2.2. Protein glycosylation has conserved linkages and cores . . . 6
1.1.2.3. Protein glycosylation has conserved epitopes . . . 16
1.1.3. The importance of protein glycosylation in modern medicine . . . 18
1.2. Objective . . . 20
1.2.1. Discoveries in glycobiology are limited by lack of analytical tools . . . 20
1.2.2. Aim of this thesis: MS/MS and chemical proteomics methods to study protein glycosylation and PTM-dependent protein-protein interaction . . . 21
1.3. Technical introduction: proteomics and glycoproteomics . . . 23
1.3.1. Proteomics, glycomics and glycoproteomics . . . 23
1.3.2. Sample preparation and processing . . . 25
1.3.2.1. Isolation of proteomes and sub-proteomes from biological samples 25
1.3.2.2. Sample preparation prior to MS analysis . . . 25
1.3.2.3. Fractionation, enrichment at peptide level . . . 30
1.3.3. Nano liquid chromatography (nLC) . . . 34
1.3.4. Mass spectrometry . . . 36
1.3.4.1. Ionisation techniques . . . 36
1.3.4.2. Mass analysers and detectors . . . 37
1.3.4.3. Tandem mass spectrometry & peptide fragmentation . . . 41
1.3.4.4. Data acquisition methods . . . 48

1.3.5. Data analysis: peptide identification and quantitation 52
1.3.5.1. Preprocessing 52
1.3.5.2. Database searching 52
1.3.5.3. De-novo sequencing 56
1.3.5.4. Quantitative proteomics 57
1.3.5.5. Crosslinking tandem mass spectrometry 60
1.3.6. Proteomics applications and their respective post-processing methods . . . 65
1.3.6.1. Proteome profiling 65
1.3.6.2. Spacial proteomics 66
1.3.6.3. Functional analysis 66
1.3.6.4. Protein-protein interaction partners 67
1.3.6.5. PTM analysis 68
1.3.6.6. Structural proteomics 70
1.3.6.7. Chemical proteomics 71
1.3.6.8. Multi-omics data integration 73
1.4. References 74

1.1. General introduction: protein glycosylation

1.1.1. Protein glycosylation is a highly variable and diverse post-translational modification

The importance of carbohydrates in biological and medical sciences has been established in the late 19th and early 20th century, yet it is an often overlooked class of biomolecules. From virus and single-cell organisms to complex animals and plants, carbohydrates are prominent for the many biochemical processes in which they play an essential role. Their dazzling complexity, lack of templates and the absence of simple methods for analysis, amplification or synthesis has prevented the broader study of these biomolecules.[1,2]

Protein glycosylation is only an aspect of the wider biological roles of carbohydrates, it is a structurally diverse post-translational modification (PTM) that consists in the covalent attachment of a carbohydrate unit to a protein. Without any known exceptions, all cells and numerous macromolecules carry glycans − the carbohydrate part − on their surface, which can consist of a single sugar unit (monosaccharide) or longer sugar chains (oligosaccharides). It is estimated that half of all proteins are glycosylated.[1,3]

Protein glycosylation plays an essential role in every aspect of cellular organisation, as well as tissues and organs.[4] Important biochemical functions such as exogenous and endogenous cell-cell recognition[5], signalling processes[6,7], immune reactions[8,9] as

well as cellular protein quality control[10] depend on this highly diverse and heterogeneous post-translational modification[11]. Noteworthy, cancer cells have altered surface glycosylation[12], which is starting to emerge as potential diagnostic markers[13,14] and is linked to drug resistance.[15] Protein glycosylation has further been shown central in pathogen invasion, such as the *Helicobacter pylori* recognition of gastric sialoglycans, or toxin interaction such as ricin toxin binding to terminal β-linked galactose residues.[4]

From a physico-chemical point of view, protein glycosylation alters the stability of proteins by influencing their macromolecular assembly or protecting against proteolytic degradation. Moreover, the hydrophilic glycosylation increases protein solubility and is used to help protein folding and its quality control, protein transport and trafficking by preventing possible precipitation and aggregation of misfolded proteins. In contrast to polypeptide chains that fold into a compact protein tertiary structure, glycans "stretch out" in solution and have significant freedom of motion in aqueous solutions. Through discrete structural changes glycans can be used to fine-tuning protein activity or alter protein-protein interaction (PPI).[16]

On the cellular surface, glycoproteins and glycolipids determine the physical properties of the glycocalix − the sugar coating outside cells − including its hydration, lubrication, elasticity and protection against external threats. Long carbohydrate polymer chains (glucosaminoglycans, GAG) spanning the extracellular matrix shape tissue properties by regulating water content, influencing cell migration and organising the extracellular space. This is of particular importance in epithelial surfaces such as the gut or the lungs, where the thick glycocalyx is essential for organ protection from pathogens or exogenous chemicals.[4]

In truth, protein glycosylation is a multitude of structurally and functionally very different carbohydrate-based post-translational modifications, which each require tailored methods for their analysis. Figure 1.1 provides a more detailed overview over common classes of animal glycans, those biological functions will be briefly outlined in the following.

In the cytosol, the site-specific attachment of an *O*-GlcNAc modification on serine/-threonine has the role of a site-specific signalling post-translational modification, as will be detailed further in chapter 4. In the nucleus, *O*-GlcNAc acts as an epigenetic modification on histone tail and core, affecting gene expression and recruiting of interaction partners (see chapter 4). In both cases, *O*-GlcNAc serves regulatory functions at the single cell level.

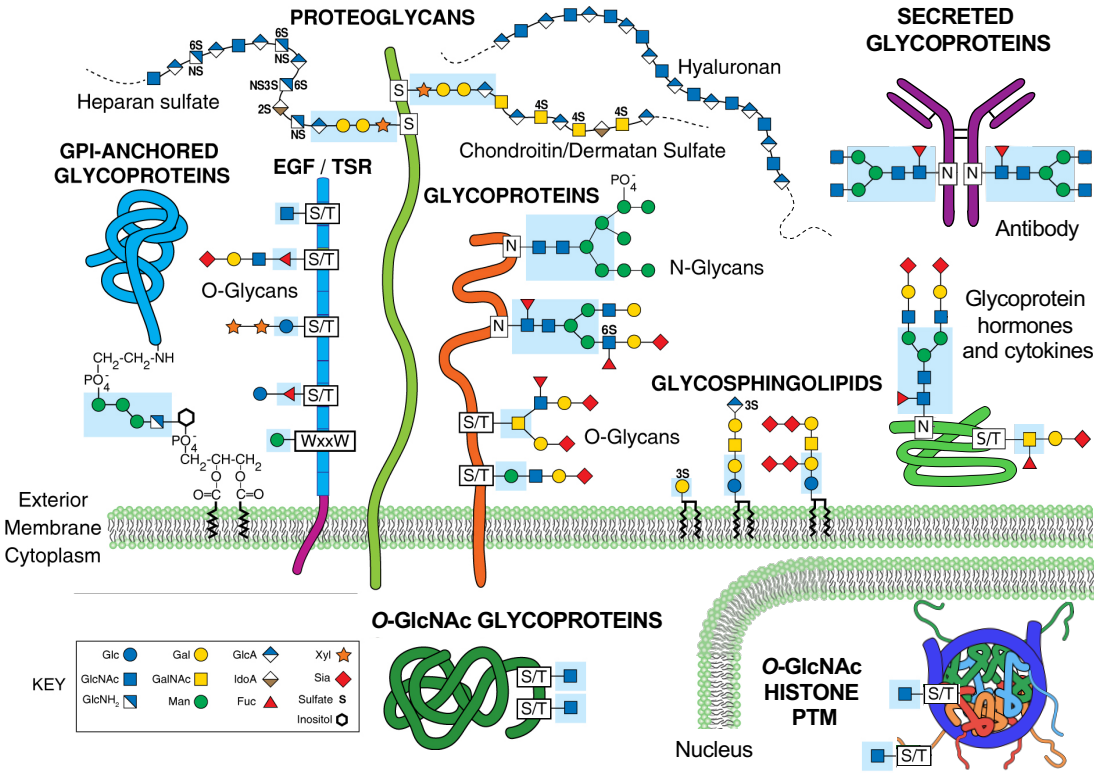


Figure 1.1. Common classes of animal glycans. Proposed glycoRNA's have been omitted as their role and structure remains to be validated.[17] Structures highlighted in teal indicate the conserved core that determines glycosylation type. Extracellular glycans are formed in the ER and Golgi secretory system. Glycan drawings follow SNFG nomenclature (see appendix 6.4). Modified and updated from[18], with permission from ©CSH Press.

In contrast, more complex glycans are required at the multicellular (organismic) level. On the cell surface, the organisation of glycolipids and glycoproteins that make up the glycocalyx has been compared in a famous analogy by Ajit Varki to the canopy of a tropical forest[19] (Figure 1.1): the glycan class (sphingolipids, GPI-anchored-, *N*-linked or *O*-linked glycoproteins – see Figure 1.1) corresponds to the trees and bushes of the forest. The branching pattern of the sugars relates to branching in trees; terminal modifications such as sialic acids, fucose, phosphorylation and sulfation etc. to the leaves and flowers visible on top of the canopy. The spacial organisation on the cell surface in "saccharide patches" or "glycosynapse" reflects the microorganisation in the forest and allows for multivalent recognition and subsequent cellular reaction, such as activation of a signalling pathway or triggering of endocytosis.

Targeted cell-cell recognition processes and several specific signalling pathways rely on glycoprotein notches sticking out of cells. These long transmembrane proteins consist

of repeating protein domains of the types EGF-like repeat or TSRs (see section 1.1.2.2). These domains are highly and site-specifically glycosylated, the exact type of glycosylation will affect how the signal is taken over into the cell and influence cellular activation. Prominent examples are the Notch signalling pathway or the recognition of the PSGL-1 glycan structure by selectins, their understanding has been instrumental to the discovery of these very precise communication routes over the cell membrane.[1]

The biological role of proteoglycans and glucosaminoglycans is better explained with another famous analogy of a sea-floor kelp bed, as suggested by Pascal Gagneux.[20] The long polymeric carbohydrate chains (glucosaminoglycans, GAG), anchored to a long protein chain (GAG + protein anchor = proteoglycan) organise the extracellular matrix further away from the cell. Proteoglycans are categorised by their repeating disaccharide unit into Heparan sulfates, Chondroitin/Dermatan sulfates, Keratan sulfates (-3Gal(β1-4)GlcNAcβ1- ; ○-■) and Hyaluronan. These acidic polymers are dynamic, diverse in their structure with multiple branching patterns; they retain electrolytes, control the elastic and hydration properties around the cells while at the same time allowing migrating cells to move across, to situate their location and guiding them to access the cell surface glycans for more specific recognition. Moreover, they can act as micro-compartment or diffusion barriers around the cell surface, allowing to keep secreted proteins in an enclosed space and releasing those when necessary, such as components of the innate immunity against a pathogen.[21]

Numerous glycoproteins in the extracellular space are not attached by any means to the cell membrane. Released glycoproteins such as antibodies, glycoprotein hormones and cytokines or serum proteins are very abundant in body fluids or tissues. Notorious examples for glycoprotein hormones are pancreatic FSH and LH, as well as erythropoietin (EPO).[1] Equally, the occurrence of free oligosaccharides related to *N*-glycans (syalyl FNGs) in animal sera was long supposed and has been recently identified.[22] Oligosaccharides in milk are another example of secreted glycans.[1]

In summary, carbohydrates and glycoproteins mediate a vast range of functions essential for organism survival. At first, the complexity of protein glycosylation appears dazzling, however there are common patterns / rules and structures that allow to categorise and guide through the heterogeneity.

1.1.2. Conserved structural elements in protein glycosylation

1.1.2.1. Available monosaccharides and glycosyltransferases limit the combinatorial diversity

The monosaccharide diversity across all kingdoms of life used to build up glycans is vast. A huge combinational diversity of complex glycans and polymers is theoretically accessible via their linkages and branching − not mentioning additional sulfation, phosphorylation etc. Yet, eukaryotes use only a very limited subset of glycan structures, the glycosylation in Bacteria and Archae is far more diverse.[23]

In humans only 9 monosaccharides build up the entire glycome[1] − symbiotic microbiota taken apart: glucose (Glc, ●), *N*-acetylglucosamine (GlcNAc, ■), glucuronic acid (GlcA, ◆), galactose (Gal, ○), *N*-acetylgalactosamine (GalNAc, □), mannose (Man, ●), xylose (Xyl, ★), fucose (Fuc, ▲) and *N*-acetylneuraminic acid (Neu5Ac, ◆) which is often confounded with the sialic acid class (Sia, ◆ − please refer to Appendix 6.4 for full list of monosaccharide names and codes, including SNFG symbols).

In the following, classes of glycans, general rules for the biosynthesis of glycans (resulting in their conserved cores), as well as common structural features are described, showing that the huge theoretical possibilities has been narrowing down in evolution to very distinct and organism-specific carbohydrate structures. This is a consequence of their biosynthesis: glycans are not synthetised following a template, but instead by an array of sequentially and competitively acting glycosyltransferases (GTase) in the ER and Golgi apparatus, and are subsequently trafficked to the cell surface via vesicles. Knock-outs of these glycosyltransferases are often embryonically lethal, their combined activity gives rise to the conserved patterns. Expression levels of GTases are tissue and cell-type specific, explaining the distribution of several of the later mentioned structures across the human body. Discussions of the evolutionary development of glycans and its implications in pathogen/host co-evolution are beyond the scope of this thesis and the reader is kindly referred to excellent accounts in the literature.[4]

1.1.2.2. Protein glycosylation has conserved linkages and cores

Figure 1.2 shows a widely adopted classification of protein glycosylation into 6 types by the linkage chemistry to the protein, named after the key functional atom connecting the two entities. This distinction in glycan linkage classes is key as it affects the physicochemical properties of glycopeptides (see chapter 3).

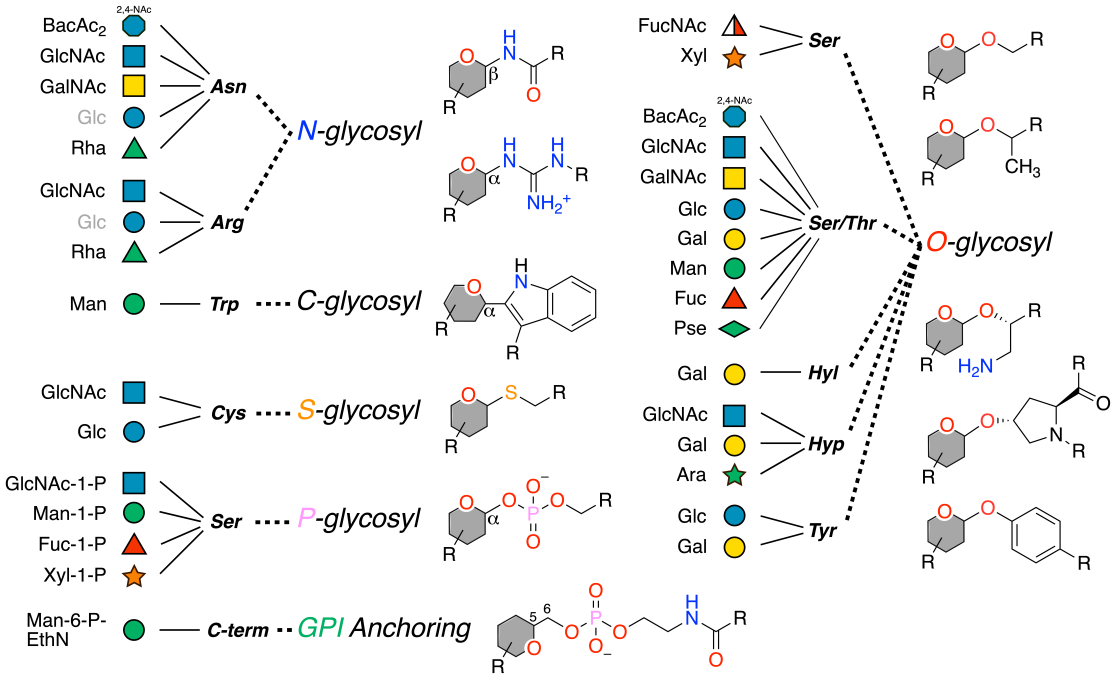


Figure 1.2. Glycan-protein linkages reported in nature. Diagrammatic representation of six distinct types of protein glycosylation that have been identified in nature, classified by their protein linkage chemistry. The molecular structure of the sugar-protein bond is indicated, together with the linkage geometry (α or β) for cases where it is clearly determined by the stereochemistry of attached functional groups. The reducing sugars (directly linked to protein) identified across all kingdoms are listed, grey monosaccharide names indicate poor experimental evidence. (Original data from R. Spiro[11]; for updated references see text.)

The chemical logic behind this classification is immediately apparent, nonetheless this system is also pertinent from a biological point of view. Indeed glycans are build sequentially, either on a carrier lipid or directly on the target protein. The amino acid on which the first sugar is loaded determines what the subsequent glycosyltransferase enzymes sees before it loads the next monosaccharide. Therefore, the biosynthesis of each class is individual, i.e. the linkage class dictates which structures can / cannot be formed – analytical workflows take advantage of this knowledge.

This chapter will cover in the following mostly animal glycans and only highlight specific features relevant to a generalised understanding. As a side note, two other carbohydrate-based modifications exist: ADP-ribosylation and unspecific protein glycation. These are strictly speaking not counted as protein glycosylation and not further discussed in this thesis.

N-glycoproteins

Asn *N*-linked protein glycosylation is in the most well-known form a branched glycan attached to the amide nitrogen of asparagine, typically in the consensus motif Asn-Xxx-Ser/Thr where Xxx represents any amino except proline − exceptions where *N*-glycans are located on nonconsensus sequons (NXN, NXQ, NXC and reverse motifs) exist.[24] This glycosylation is ubiquitous amongst eukaryotic organisms, as well as found in some Archae, and is divided into three types of classes: oligomannose, complex and hybrid (see Figure 1.3), whose prevalence differ between organisms. The conserved chitobiose core Man(α1-6)[Man(α1-3)]Man(β1-4)GlcNAc(β1-4)GlcNAc(β1)Asn is a result of the biosynthesis of *N*-glycans in eukaryotes in the endoplasmatic reticulum (ER) and the Golgi. Asn glycosylation occurs co-translationally: upon recognition of the signal sequence by the protein SRP, the ribosome docks on the SEC61 translocon in the ER membrane. As the nascent polypeptide chain emerges, the oligosaccharyl transferase (OST) adds a core glycan ($GlcNAc_2Man_9Glc_3$) *en bloc* from the dolichol-linked precursor. Glycan processing is coupled to the ER protein folding quality control mechanism, terminal glucose residues are removed by the calnexin/calreticulin system if the protein is normally folded to yield a $GlcNAc_2Man_9$ protein, which can then be further processed in the Golgi to the final glycan structures. Misfolded proteins undergo further mannose trimming to Man_7 and are removed via the ER associated degradation (ERAD) pathway.[25,26]

The variety of Asn *N*-glycosylation in Archaea and Bacteria taxa is very diverse, yet their assembly of *N*-glycoproteins follows a functionally similar pathway to Eukarya: the *N*-glycan is assembled on a carrier lipid phosphate − structurally similar variations of the dolichol lipid carrier. The glycan is assembled on the cytoplasmic site, translocated to the external site of the membrane by a flippase and transferred onto the target protein. Linkages on Asn of $BacNAc_2$, GalNAc, Glc and Rha have been identified, and are extended into complex glycans.[27–29]

Arg Arginine *N*-GlcNAcylation is particularly prominent since the recent discovery that enteropathogenic *E. coli* strains GlcNAcylate arginine in TRADD death domains to block host death receptor signalling, especially TNF signalling.[30–32] Furthermore, Arg rhamnosylation is a regulatory mechanism in bacterial protein expression.[33,34]

Two examples in the literature suggest that Asn glycosylation this might not be the only form of *N*-glycosylation in higher organisms (i.e. fungi, animals and plants). In animals, the extracellular laminin protein has been reported to be modified by a

single Glc(β) on Asn, localised in the kidney glomeruli.[35] In plants, amylogenin which is believed to be a self-glycosylating protein from sweet corn, has been reported to contain a Glc-β-Arg linkage to the guanidino group.[36] Although both these examples for linkages have found their way into glycobiology textbooks (such as the *Essentials*[1]), it is worth mentioning that the evidence presented beyond its initial discovery in the 90s remains sparse and would not suffice by today's standards, MS/MS proof being at oxonium ion level or large fragments, respectively.

Linkage stereochemistry Interestingly, all well-characterised *N*-glycosylations of **Asn** reported at the time of writing were **β**-linked at their reducing terminus, suggesting a strong influence of the amide group on the linkage geometry. Supposedly, this is due to the exo-anomeric effect, the amide group being stereoelectronically more favourable in the equatorial position as a result of sterics, higher dipole moment (in water) and orbital overlap that permits expression of an $\pi_{nb,\underline{N}CO} \rightarrow \sigma^*_{C\text{-}O}$ stabilizing orbital interaction.[2,37–39] On the contrary, the well-characterised *N*-glycosylations of **Arg** appear to all be **α**-linked at their reducing end.[32,34] Alkylation of arginine has been shown to have only a minor effect on its pK_a, as such protonation / positive charge is expected to remain on the guanidinium at physiological pH.[40] The α-configuration is consistent with observations of the endo-anomeric effect, where the antiperiplanar $n_O \rightarrow \sigma^*_{C\text{-}N}$ donation is favoured by the electron-deficient guanidinium moiety.[37,38]

C-glycoproteins

C-mannosylation features a unique C-C bond between Man C-1 and the C-2 atom of the indole moiety of tryptophan, the linkage is in the α-configuration. *C*-mannosylation is widespread in mammals and follows the WxxW consensus sequence, where the first Trp is modified. Dolichol-P-mannose produced from GDP-Man is the biosynthetic donor for *C*-mannosylation, which occurs in the ER. It is thought to assist in the protein folding process and play a role in extracellular signalling and recognition, including of TSRs.[41,42] Several complement proteins and interleukins contain *C*-Man, it has been further suggested as a marker for renal dysfunction and for ovarian cancer.[43] Interestingly, the Ebola virulence factor soluble glycoprotein sGP is *C*-mannosylated, although this modification does not seem essential for protein function.[44]

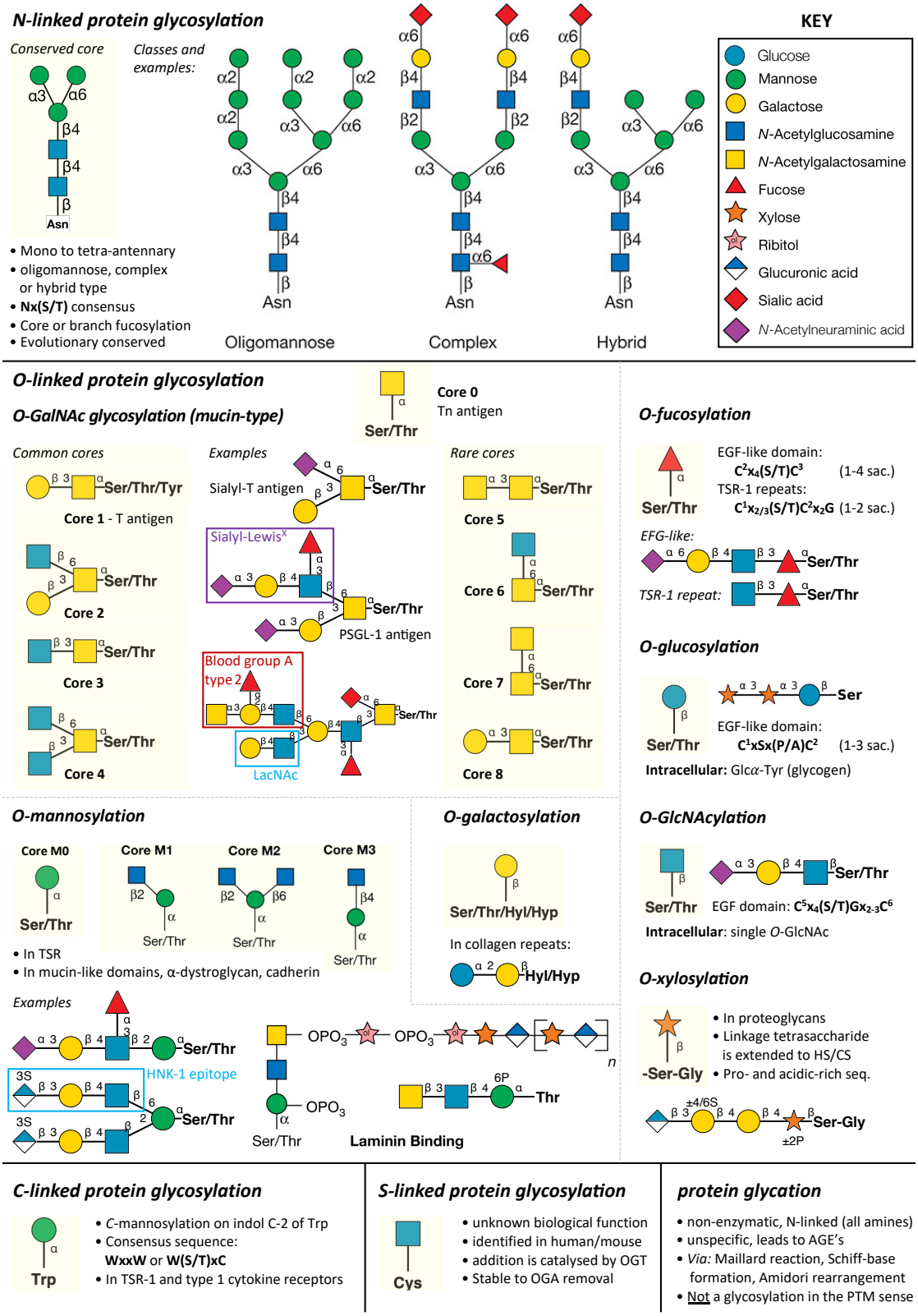


Figure 1.3. Mammalian protein-glycan cores with examples. Core structures are highlighted in pale yellow and examples of characterised glycans given.

O-glycoproteins

Figure 1.2 highlights the immense structural variety in *O*-glycosylation found in the living world. Indeed, *O*-glycosylation has been identified in all three kingdoms of life.[45,46] It can be linked through Ser /Thr, Tyr, Hyp as well as Hyl. The next section focusses on mammals, with glycosylation mostly on Ser/Thr unless otherwise stated.

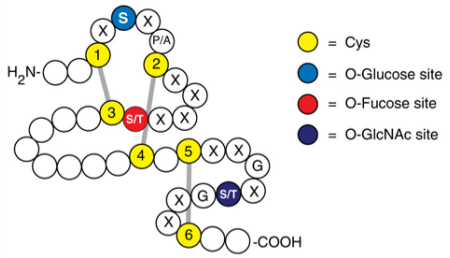
Linkage stereochemistry Given that the anomeric effect in aqueous solution is not as strong as in apolar solvents[2,47], *O*-linked protein glycosylation can be found in both α and β linkage geometries. As such, the stereochemistry of the monosaccharide is determined by the glycosyltransferase, classified in retaining and inverting glycosyltransferases.[47] As a result, in humans the linkage geometry is defined: GalNAc-α, Fuc-α, Man-α, Glc-α (on Tyr-Collagen), Glc-β (extracellular), Gal-β, GlcNAc-β, Xyl-β.[11]

Mucins / *O*-GalNAc *O*-GalNAc-α-linked glycosylation of Ser/Thr (and Tyr[48]) is often referred to as mucin-type. It is important to mention that this type of *O*-glycosylation also takes place on proteins that have no mucin domains. First described in the 1860s, mucins are highly *O*-GalNAc glycosylated polymers typically found in slimy surfaces on membranes of many epithelial surfaces, such as the gastrointestinal and respiratory tracts where they form the mucus layer and shield cells against pathogen infections or physical & chemical damage. Soluble mucins, such as small monomeric mucins or the gel-forming mucins are also secreted by the cells in mucus glands to the ECM. The typical *O*-GalNAc glycans found on mucins are short (di- to tetrasaccharides) and often sialylated.[49,50]

All *O*-GalNAc glycans are built from the sugars GalNAc, Gal, GlcNAc, Fuc and Sia, whereas Man, Glc or Xyl residues are not represented.[1] Gal and GlcNAc may be modified by sulfation and Sia by *O*-acetylation. Their length varies from 1-20 residues and their composition is highly heterogeneous, the longer forms can include blood groups and other glycan epitopes[50] (see Figure 1.3). 8 conserved cores have been identified in animals (Figure 1.3), of which core 1 and core 2 are the most frequent *O*-GalNAc backbones, followed by core 3 and core 4; core 5-8 are rare. Only cores 1-4 and core 6 *O*-GalNAc glycans have been shown as extended. Typical extensions are Sia (α3 or α6) or Fuc (α2 or α3).[51] The biosynthesis proceeds sequentially through galactosyltransferases in the Golgi. Although it is known that *O*-GalNAc glycans tend to be shorter /

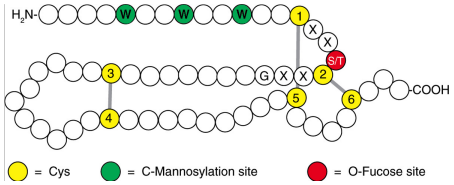
truncated in cancer (mucins are anti-adhesive), the rules of assembly are currently not understood.[14]

EGF-like repeats Epidermal Growth Factor (EGF)-like repeats are small protein motifs (30-40 amino acids) that are defined by six conserved cysteine residues, which form three disulfide bonds. This structural motif is found in hundreds of cell-surface and secreted proteins in animals and can be modified site-specifically by *O*-fucosylation ▲, *O*-glucosylation ● and *O*-GlcNAcylation ■. The position of each glycosylation is defined in respect to the disulfide bridges and will regulate ligand recognition and signal transduction such as the Notch signalling pathway (discrimination Serrate/Jagged).[42,52]



Scheme of an EGF-like repeat.[1]

TSRs Thrombospondin type 1 repeats (TSRs) is another commonly *O*-glycoslated small protein domain (50-60 amino acids), also with six conserved cysteines forming three disulfide bonds. As for EGF-like repeats, they are also found in a subset of cell-surface and secreted proteins in animals and appear to function in protein–protein interactions. They can be modified by *O*-fucosylation ▲ and *O*-mannosylation ●.[42]



Scheme of a TSR.[1]

***O*-fucose ▲** EGF-like repeats can be modified by *O*-fucosylation immediately before Cys^3, with 1-4 saccharides from the Siaα2-3/6Galβ1-4GlcNAcβ1-3Fucα-O-Ser/Thr (◆-●-■-▲-S/T) sequence. The regulation of Notch signalling by *O*-fucosylation is one of the clearest examples to date of how a signalling pathway can be regulated by altering the glycosylation of a receptor: single Fuc on Notch results in the receiving cell favouring signals from Serrate/Jagged ligand, the extended tetrasaccharide leads to favouring the Delta signalling ligand. Similarly, *O*-fucose is found on TSRs, although the biological function of this modification remains unclear.[42,52]

***O*-glucose ●** Found on EGF-like repeats in the $C^1xSx(P/A)C^2$ consensus, it typically exists as the trisaccharide Xylα1–3Xylα1–3Glcβ-O-Ser (★-★-●-S/T), although

mono- and disaccharide forms are also seen.[42] The biological function is poorly explored, there are indications for similar signalling effects as *O*-fucosylation and hints for links with the Notch pathway.[42] Initial findings indicate *O*-Glc could have a role in intracellular signalling as well.[53]

_O_-GlcNAc ■ This modification is particularly prominent as intracellular single sugar PTM in nucleus and cytosol – reviewed in more detail in chapter 4. An extracellular variant exists as well found on EGF-like repeats that can be further extended to Siaα2/3-6Galβ1-4GlcNAc (◆-○-■-S/T), although its identification so far has been rather rare. The gene encoding the EGF-specific *O*-GlcNAc transferase, EOGT, is distinct from OGT, the GlcNAc-transferase that modifies nuclear and cytoplasmic proteins. Mutations in human EOGT result in skin and skeletal problems in AOS (Adams–Oliver syndrome type 4), a disorder with links to the Notch pathway.[42,54]

_O_-mannose ● From a phylogenetic perspective, *O*-mannosylation is another major class of *O*-glycosylation found in many eukaryotes.[55] In humans, *O*-Man is particularly prevalent in the brain, accounting for ∼30% of glycosylations. As for other types, the biosynthesis is initiated in the ER but in contrary to other *O*-glycosylations, the initial donor is a dolichol phosphate lipid – the same precursor as for *C*-glycosylations. Three different cores, M1-M3 are build on this initial monosaccharide. One particularity of *O*-Man glycans is the use of glucuronic acid as a terminal sugar alongside the usual Fuc and Sia for example, leading to formation of the HNK-1 epitope in neurons. Defects in *O*-mannosylation of the protein α-dystroglycan, the major extracellular glycoprotein component of the dystrophin–glycoprotein complex, leads to loss of anchoring of the sarcoplasm with the extracellular laminin that binds to the glycan and results in multiple forms of congenital muscular dystrophy (CMD), known as the secondary dystroglycanopathies.[42,55,56]

GAG / _O_-xylose ★ The linkage tetrassacharide ◇-○-○-★-SG that anchors two classes of long glucosaminoglycan chains, chondroitin sulfate/dermatan sulfate and heparan sulfate/heparin, to anchor proteins (such as Decorin, see chapter 3) is *O*-linked through xylose to a Ser-Gly conserved consensus sequence, typically in proline and acidic amino acid rich sequence. The linkage tetrasaccharide lies at a bifurcation in the biosynthetic pathway. Two types of reactions occur: β4-addition of GalNAc, which initiates chondroitin sulfate assembly, or α4-addition of GlcNAc, which initiates heparan sulfate assembly.

Collagen / *O*-Gal ○ Proteins with collagen domains can be modified with the disaccharide Glcα1-2Galβ-, assembled on hydroxylysine (Hyl) or hydroxyproline (Hyp) residues. This modification is not restricted to extracellular collagen, and occurs in other proteins that have collagen-like modules such as adiponectin and pulmonary surfactant proteins. It is thought that this glycosylation might have a regulatory function in the assembly of collagen fibres.[42]

S-glycoproteins

S-glycosylation is a recently identified PTM, its biological role is unclear yet its structural properties are often compared to the *O*-glycosylated analogues, *S*-GlcNAc is used as a structural surrogate for *O*-GlcNAc in many chemical biology studies[57] including this thesis (chapter 4). Early accounts of naturally occurring *N*-terminal cysteine glycosylation in human urine or erythrocyte membrane are almost fifty years old and retrospectively doubted as the peptide sequences are not found in the human proteome (P02728, P02729).[58,59] However, in the last ten years *S*-glycosylation has been characterised on bacteriocins glycocin F[60,61] and lantibiotic sublancin[61,62], as well as native *S*-GlcNAcylation in mammals[63].

The transfer is, as for *O*-GlcNAc, catalysed by *O*-GlcNAc transferase[63] (OGT, see chapter 4 for details). Thioglycosides seems to remain stable to enzymatic hydrolysis by most *O*-GlcNAcases (OGA)[57], as OGA uses acid catalysis to facilitate cleavage of *O*-glycosides but not *S*-glycosides.[64]* Enzymes that hydrolyse *S*-glycosides, such as mammalian thioglycosidase[65], must be able to effectively stabilize a transition state bearing a sulfur atom in the exocyclic position.[64] As such, the question of removal as well as biological significance of *S*-glycosylation as a PTM remains open. This is an important question for the chemical biology community as per-acetylated GlcNAz (azido-GlcNAc) and derivates, used in metabolic labelling studies to investigate *O*-glycosylation, are shown to increase the levels of *S*-GlcNAcylation, raising doubts on findings obtained with this methodology (see discussion in 1.3.2.3).

P-glycoproteins

*Note that Vocadlo and co-workers showed hOGA activity on pNP-S-GlcNAc as a substrate[64] (not peptide), which is thought by Pratt and co-workers to be confined to the more activated leaving group.[57]

Attachment of a sugar to Ser/Thr by means of a phosphodiester bond has been coined phosphoglycosylation. To date, this modification has been identified on cysteine proteases of the slime mold *Dictyostelium discoideum*, an amoeba (multicellular eukaryote), as well as phosphoglycans of Leishmania and Trypanosoma species, both euglenae (unicellular eukaryotes). Identified glycans range from a single GlcNAcα1-PO_4-Ser in *D. discoideum* to complex branched glycans on proteophosphoglycans of Leishmania and Trypanosoma. As expected, the identified linkage geometry is exclusively α due to the anomeric effect, analogous to the geometry of UDP-sugars.[66–68]

GPI-linked glycoproteins

GPI-anchored glycoproteins are a small but diverse group, found almost ubiquitously amongst eukaryotes. They float in the extracellular space and are tethered to the membrane by a glycolipid anchor. The minimal core structure of the GPI anchor consists of a glycerophospholipid attached to inositol at C-1 position, then a linear tetrasaccharide consisting of GlcNAc followed by three mannoses and phosphate-linked ethanolamine that connects with an amide bond to the protein C-terminus (Protein-CONHEt-OPO_3-6Manα1-2Manα1-6Manα1-4GlcNH_2α1-6myo-Ins1-PO_4-L, see Figure 1.1). This minimal core structure can be extended by additional glycans on Man, fatty acid esters on Ins-C2 and C-3 and phosphoethanolamine groups on Man. The biosynthesis occurs in the ER, it is transferred *en bloc* from the GPI transamidase to the target protein. Although there is no consensus sequences, identified patterns and likely sequences have led to the development of prediction tools.[69]

GPI anchors are very prominent in parasites of different classes, such as *Trypanosoma*, *Leishmania*, *Plasmodium*, where they act as virulence factors for host invasion. In addition, many mammalian proteins including cell-surface hydrolases, receptors, adhesion molecules and complement defence molecules (such as CD59) are GPI-anchored.[69]

In summary, understanding conserved glycosylation linkages and cores, their origin from biosynthesis and their intrinsic chemical properties has a profound meaning for analytical abilities: the ultimate aim in glycoproteomic analytics is to know which glycan occupies which site of the protein, including monosaccharide linkages and connections. This predefined knowledge helps in analysing data − which might be incomplete − by providing a framework for likely glycan structures and site localisation. In the following section, the focus will be on the extension and endings of glycosylation cores to complete glycans.

1.1.2.3. Protein glycosylation has conserved epitopes

Despite having different cores, many *N*-glycans, *O*-GalNAc and *O*-Man glycans as well as glycolipids share common structural motives in the extension away from the core as well as the endings of their glycans, i.e. common epitopes such as Lewis^x motif or human blood group determinants. In the previously mentioned analogy by A. Varki to a tropical rainforest, extensions and epitopes would correspond to the branches and leave stems. These structures can be further modified by sialylation or poly-sialylation, sulfation, phosphorylation, antennae fucosylation and more − especially the terminal sugars. In our analogy, this would correspond to the leaves and flowers, an integral part of the mature glycan. Whereas cores are conserved amongst mammals, each species has its own extension and terminal epitopes, with variants being different amongst individual sub-populations of each species (see Figure 1.4).[70]

Typical extensions in humans are the *N*-acetyl-lactosamine unit (LN or LacNAc) in which Gal is β1-4 linked to GlcNAc, and the neo-LacNAc with the β1-3 linkage. These can be α2-3 or α2-6 sialylated to form sialyl-LacNAc, a very common terminal motif in vertebrates. Note that in humans, sialylation would be with Neu5Ac, as humans do not have the Neu5Gc sugar present for example in the Great Apes. Several repeating LacNAc units can form complex epitopes such as the i and I antigens on erythrocytes. Similarly, the LacDiNAc extension consisting of GalNAcβ1-4GlcNAc is often found in humans. A particularity of some *N*-glycans, such as of the pituitary hormones lutropin (LH) and thyrotropin (TSH) is to have a terminal sulfated GalNAc instead of the more usual sialic acid.[70]

Often, the human ABO blood groups are found build on a LacNAc unit, they are expressed on membrane glycoproteins and glycolipids on the surface of erythrocytes and many epithelial and endothelial tissues. Variants such as directly on O-GalNAc core 1 have been identified, among others. It is astonishing that the difference of a single stereocenter makes the difference between A and B blood groups, yet the consequence of a blood transfusion mix-up would be lethal. The Lewis blood group is another epitope family, Sialyl-Lewis^x is crucial for cell-cell interaction and immune reactions, as well as egg fertilisation by sperm.[70]

More examples of cell and tissue-specific epitopes are shown in Figure 1.4. Often, these epitopes are not only expressed on cell surface glycoproteins, but also on proteoglycans. It appears that the further away from the cell, the more important epitopes become.

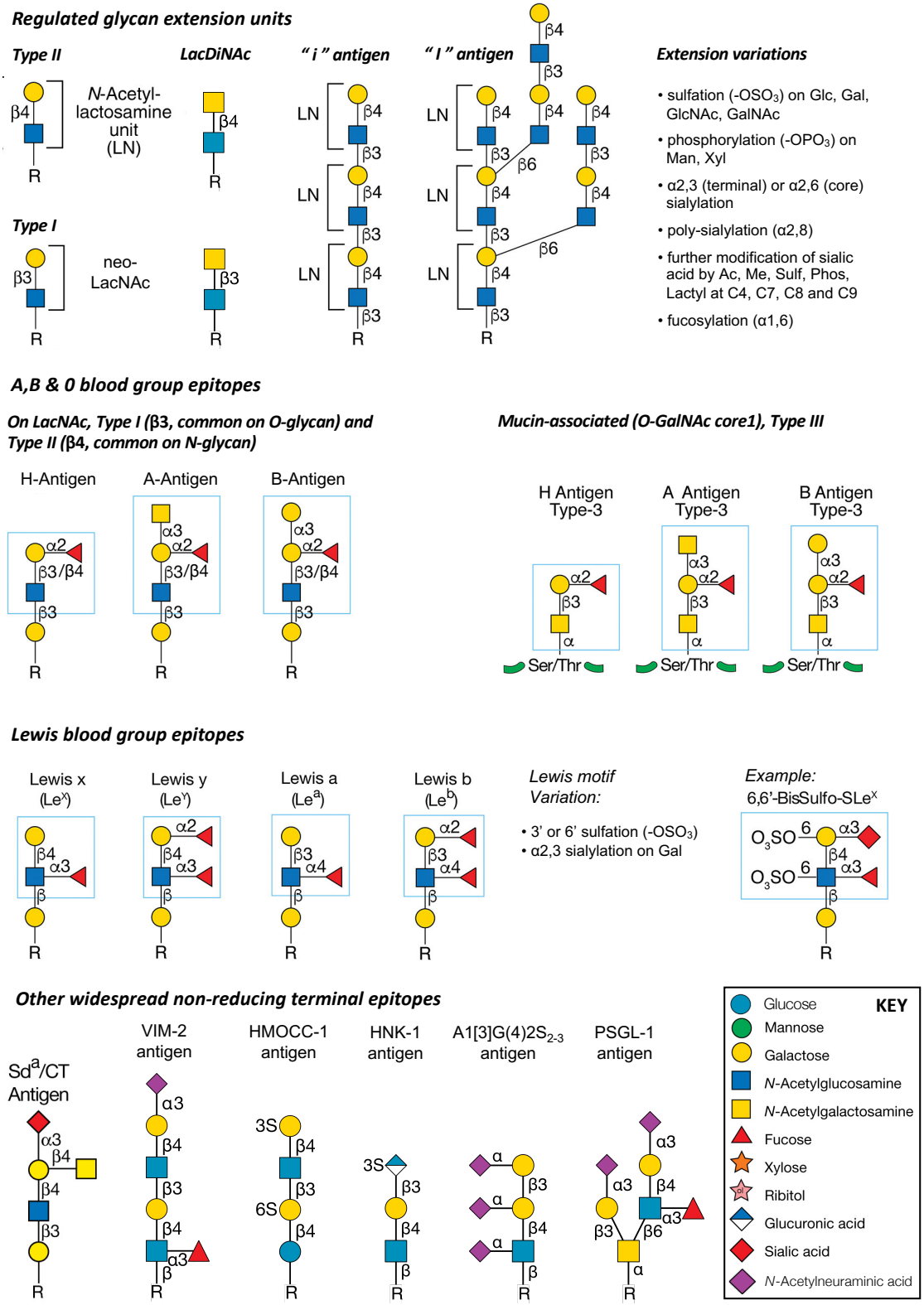


Figure 1.4. Common extensions, epitopes and antigens on human glycans.

1.1.3. The importance of protein glycosylation in modern medicine

Beyond the well-known importance of blood groups for transfusion and serum cross-compatibilities for ***transplantation*** (and xenotransplantation), human glycobiology presents many opportunities for small molecule therapeutics, glycopeptides as well as for biologics and vaccines. Indeed, many disease processes that affect humans and other animals involve glycans. Chronic disease pathologies such as cancer,[12] autoimmunity[8] or genetically inherited (congenital) disorders of glycosylation (CDG)[71] as well as disease pathologies resulting from infections of pathogens or virus[71] have been investigated for the therapeutic exploitation of glycosylation, for the development of drugs[71] or diagnostic markers.[72]

Small molecule therapeutics derived from carbohydrates that take advantage of the functional understanding of carbohydrate-protein interactions are known as ***glycomimetics***. Although their design and development has to overcome properties intrinsic to carbohydrates, such as low affinity, serum half-life and poor oral bioavailability as well as complex chemical synthesis,[73] the development of famous drugs such as zanamivir (Relenza) or oseltamivir (Tamiflu) for treatment of influenza infection has shown the potential of this therapeutics class.[74,75]

Conceptually, one approach for the development of small molecule drugs has been to mimic the geometry of carbohydrate recognition structures. This is exemplarily the case for developments of selectin inhibitors for acute myeloid leukaemia (AML) by GlycoMimetic Inc, in which the active pharmaceutical ingredient (API) mimics the biologically active conformation of the carbohydrate ligand PSGL-1 and thus takes advantage of the native glycan-protein interactions.[75]

Another strategy is to inhibit catalytic enzyme activity of human carbohydrate processing enzymes. This has been particularly prominent as an antiviral strategy, where inhibition of the ER glycosyltransferases prevents virus glycoprotein folding and assembly and thus circumvents the issue of high mutation rates of virus allowing them to bypass direct inhibition.[71] Similarly, some CDGs resulting in metabolic disorders can be treated by ***substrate reduction therapy***, as is the case for type I galactosaemia. In this approach, GALK, an enzyme upstream in the biosynthetic Leloir pathway of the dysfunctional mutated enzyme GALT is inhibited by a monosaccharide analogue, leading to a less severe pathology.[76,77] These examples illustrate that the field of glycobiology offers multiple possibilities for treating untapped diseases with small molecules – although much remains to be explored.

Beyond small molecules, many therapeutic proteins are glycoproteins, a noteworthy class being ***monoclonal antibodies***. The glycoform not only affects the therapeutic efficiency of the drug,[78] but has led in some extreme cases to severe adverse effects in patients such as with the approved colorectal cancer drug cetuximab.[79] As a consequence, methods for characterising the glycosylation of therapeutic antibodies as well as the rationale of their effects have been increasingly in the focus of the manufacturing biopharmaceutical industry, which will be further highlighted in chapter 2.[80] Similarly, ***hormone drugs*** such as the anaemia drug erythropoietin (EPO) can be glycoproteins. Analysing the glycoforms of EPO to produce optimally glycosylated protein for maximum efficacy has been an ongoing effort in synthetic and analytical glycoscience.[81]

Glycopeptide and glycoprotein ***vaccines*** are another very active area of medical research in the glycosciences, equally for the API discovery as for the formulation. The influenza and meningitis shots are famous glycoprotein vaccine.[82] In more recent years, cancer vaccines based on MUC1-derived glycopeptide containing the cancer-associated Tn epitope have been evaluated in a phase III trial, however the vaccine did not impart in a significant survival benefit presumably due to too high individual variation.[71] Glycopoteins from cancer serum have been established as ***biomarkers*** correlated with specific tumours.[72,83] Whereas most FDA-approved markers have their diagnostic value based on protein quantity,[84] the use of glycosylation itself has recently emerged: the altered glycosylation of α-fetoprotein (AFP) resulting in increased fucosylation (AFP-F3) in hepatocellular carcinoma has been approved by the FDA as clinical biomarker.

Taken together, our understanding of protein glycosylation in health and disease has a profound impact on appreciating diseases, identifying plausible biomarkers and developing novel therapeutic strategies. As an underlining example, the study of the underlining pathology of some congenital diseases of glycosylation has enabled their treatment by ***substrate replacement therapy*** (mannose for Man-6-P-isomerase CDG, fucose for GDP-Fuc transporter deficiency). Similarly, ***enzyme replacement therapy*** (ERT) which consists in replacing the deficient enzyme by intravenous infusion with recombinant analogues has proven efficient for lysosomal storage diseases such as Gaucher (glucocerebrosidase deficiency) or Fabry disease (α-galactosidase deficiency). The progress in understanding diseases is thus seamlessly coupled to investigative technologies available to researchers.[71,85]

1.2. Objective

1.2.1. Discoveries in glycobiology are limited by lack of analytical tools

> For the great potential of glycobiology to be realized, efforts need to be directed toward (...) advancing technologies that will allow ease of experimentation. (Ajit Varki and Gerald W Hart, 2017[85])

Whereas the study of carbohydrates was in phase with investigation of important biomolecules such as proteins, DNA and RNA in the beginning of the 20$^{\text{th}}$ century, the postulation of the central dogma of molecular biology and the following technological revolution in molecular biology in the second half of the 20$^{\text{th}}$ century has led the glycosciences out of focus. Further, the knowledge on carbohydrates has lacked behind for several reasons: protein glycosylation does not follow a template that one can study, carbohydrates cannot easily be manipulated in cells and despite much progress, methods for carbohydrate analysis in a biological context remain challenging. Paradoxally, the setback of the Human Genome Project − set-up as the scientific community expected that knowing the human genome would explain our biology − led to an increased attention to what lies beyond genes. The resulting appreciation of post-translational modifications has ignited a renaissance of a broader interest into glycobiology.[86]

Nonetheless, the study of protein glycosylation is still hampered by the lack of available tools and methods. Understanding a very specific protein glycosylation requires isolation of the glycoprotein in its native state, the analysis of present glycans (typically using tandem mass spectrometry or lectins recognition) and structural studies to identify linkage geometries and connectivities (often by NMR or X-ray crystallography). Each of these steps requires extensive resources, expert knowledge and specialised equipment.

Glycomics has enabled the global, high-throughput analysis of released glycans with mass spectrometry (see 1.3.1) and thus contributed to a wider understanding of cell surface glycosylation. However, it does not provide any information about the location of glycan attachment, limiting the discovery of glycoproteins and the understanding of the meaning of their glycosylations. This being said, the release of glycans in itself can be an issue: whereas the enzyme PNGase F allows reliable removal of native *N*-glycans for subsequent analysis, β-elimination of *O*-glycans is an acidic hydrolysis step in which information is inevitably lost. The impact on the field of *O*-glycomics by the recent discovery of StcE[87] by the Bertozzi lab or the OpeRATOR glycosidase[88] that enables the release of intact mucin-like *O*-glycans exemplary illustrates the need for further tools in the glycosciences.

The field of glycoproteomics bears the promise of popularising the glycosciences through linking both fields of proteomics and glycomics, enabling to pinpoint the localisation of each identified glycan on its carrier protein. Yet, major technical difficulties are still to be overcome (see 1.3). Ultimately, a thorough understanding of protein glycosylation in a cellular context will require the use of multiple methods and platforms together. In this line, this thesis attempts at contributing new or combining existing methods that allow to study aspects of protein glycosylation using a (glyco)proteomic readout.

1.2.2. Aim of this thesis: MS/MS and chemical proteomics methods to study protein glycosylation and PTM-dependent protein-protein interaction

This DPhil project aims at making a contribution in developing tandem mass spectrometry methods for investigation of protein glycosylation, both as a PTM on proteins (chapters 2-4) as well as in the context of protein-protein interactions (chapter 5,6).

Chapter 2 attempts at fully characterising a glycoprotein without a genetic template, namely the anti-γH2AX monoclonal antibody, from scratch. The aim is to explore if i) *de novo* sequencing approaches can be combined with PTM and glycosylation analysis to obtain the composition of the biomolecule and ii) this information can translated in a structural model of the full protein and its peptide ligand recognition, to understand the molecular details of the mode of action. iii) This analytical project also intends assessment of various available tools in tandem mass spectrometry and modelling for use in an antibody characterisation project.

Chapter 3 asks the question, whether a fine-tuning of the fragmentation energies allows to generate tandem mass spectra of *O*-glycopeptides that contain more information and can be used to improve MS/MS analysis of *O*-glycoproteins.

Whereas the first three chapters explore the use of analytical methods based on tandem mass spectrometry for protein glycosylation, the following two chapters aim at asking precise questions in biology using synthetic mimics of post-translational modifications, including glycosylation. This is investigated in the context of histone post-translational modifications, which includes the *O*-GlcNAc modification.

Chapter 4 asks the question: what is the biological role of the *O*-GlcNAc modification on histone H2B at serine 36 in the context of epigenetic signalling? For this, a chemical proteomics method based on structural surrogates is used to search for interaction partners.

Chapter 5 aims at looking deeper into the molecular details of protein-protein interaction. It investigates whether reactive amino acids, incorporated into histone proteins, can be used to form of covalent crosslinks in the recognition site of a protein-protein interaction. It also assess specificity of these crosslinking reactions for a given interaction.

All these approaches heavily rely on tandem mass spectrometry and include method development or tailored data acquisitions. As such, a detailed introduction of proteomics and glycoproteomics is given in the following section to cover the technical background required for the subsequent chapters.

1.3. Technical introduction: proteomics and glycoproteomics

1.3.1. Proteomics, glycomics and glycoproteomics

Mass spectrometry of biological macromolecules has been conceptually separated into two major tandem mass spectrometry workflows, termed "top-down" and "bottom-up".[89] The ***top-down*** approach uses intact protein mixtures (but not necessarily the native 3D conformation), typically in samples of limited complexity.[90] It enables the characterisation of exiting proteoforms,[91] degradation products[92] as well as mapping to some extend the micro- and macro-heterogeneity of post-translational modifications of a (glyco-)protein.[93] It is however in practice limited to single protein/complexes, requiring isolation of target proteins and extensive separation.[94] From a technical viewpoint, fragmentation efficiency is limited[95] and the approach is not yet capable of disentangling complex proteomes.[94] Yet, it's progressive merging with the field of native mass spectrometry is promising for the study of PTM in structural biology.[96] Although both native mass spectrometry and "top down" tandem mass spectrometry are intrinsically limited by the sample complexity, their use for the in-detail study and understanding of purified protein complexes such as drug targets is very powerful.[97]

The ***bottom-up*** approaches[98,99] rely on the enzymatic digestion of proteins or glycoproteins using either sequence-specific proteases such as trypsin[100,101] to generate proteolytic peptides, or glycosidases such as PNGase F to release glycans.[102,103] This digestion simplifies the analyte, the resulting mixture is more homogenous in its physicochemical properties and analysis is facilitated.

In bottom-up ***proteomics*** (also termed "shotgun proteomics"),[98,99] proteolytic peptides are subject to further processing such as pre-fractionation,[104,105] sub-proteome enrichment[106,107] or chemical/enzymatic derivatisation before being loaded onto a liquid chromatography (LC) system coupled to a mass spectrometer (MS). The mass spectrometer records the mass-to-charge ratio of the peptide and subsequently fragments it (MS/MS) to determine its sequence. The recorded tandem mass spectrum is compared to a theoretical spectrum, obtained by an *in-silico* digest of a protein database of the relevant organism, to determine the peptide sequence and in extension also the protein identity.[108,109] Proteomics gives comprehensive identification of present proteins and their abundance levels.[89]

Conceptually, most ***glycomics*** workflows are also "bottom-up" in that glycans are released from proteins before analysis – this is not to be confused with the recently

introduced term "bottom-up glycomics" which refers to breaking down glycans to modified monosaccharides with linkage information.[110] Released glycans are then derivatised and analysed by LC-MS/MS, information about the peptide attachment site is lost. Glycomics gives information about which glycans are present on proteins, without any site-localisation.[111–113]

Another approach coined ***N-glycosite mapping*** (or *de*-glycoproteomics) takes advantage of the recognition and cleavage of *N*-glycans from the protein by PNGase F, while combining this with an enrichment step or a labelling step. This allows the identification of sites of glycosylation with very good proteome coverage, although without the attached glycan.[114,115]

The aim to obtain glycan composition and peptide site of attachment at the same time has led to the the emerging field of ***glycoproteomics***. In this approach, proteolytic glycopeptides are enriched and analysed by LC-MS/MS. Although this provides the most comprehensive information, glycoproteomics experiments are technically challenging.[114,115]

The following technical introduction aims at providing background information in all aspects of proteomics and glycoproteomics relevant to this thesis, from sample preparation to LC-MS/MS data acquisition, data analysis and post-processing methods (Figure 1.5).

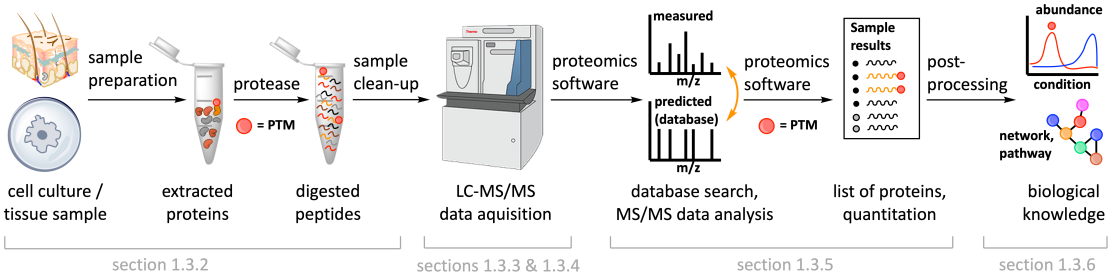


Figure 1.5. Simplified overview of a proteomics workflow. In the case of a glycoproteomics workflow, the red PTM label would be a glycosylation. The different parts of the workflow are discussed in more detail in the indicated sections.

1.3.2. Sample preparation and processing

1.3.2.1. Isolation of proteomes and sub-proteomes from biological samples

Any proteomics workflow starts with the extraction of proteins from the biological sample. Physical disruption, reagents such as chaotropic salts or detergents are possible. The method of choice depends on the physico-chemical characteristics of the proteins of interest, their cellular localisation and whether preservation of the native protein structure is required. Protocols differ regarding the source of the sample and the desired experiment, best practice methods for sample generation and avoiding the introduction of experimental bias have been reviewed elsewhere.[116,117]

Extracted proteins may be subjected to further enrichment at the protein level to obtain sub-proteomes. As an example, these can consist of the enrichment of protein classes, e.g. kinases using affinity matrices[118,119], of an enriched subset of proteins targeted by a compound class[120], of interaction partners binding to a selected protein, obtained by immunoprecipitation or tandem affinity purification.[121] Sub-proteomes can also be prepared as proteins performing particular tasks in the cell, such as DNA-binding proteins[122], or simply fractions of the proteome based on size or electrochemical properties. In all cases, the aim is to better take advantage of the limited dynamic range of the MS instrument and identify less abundant protein or track changes between samples that could be lost if sample complexity is too high.

1.3.2.2. Sample preparation prior to MS analysis

Most protocols for sample processing follow conceptually the same steps, variants have emerged to facilitate the processing or improve yields for samples that contain high salt, detergents or other challenging characteristics. Typically, proteins are denatured, disulfide bonds reduced and cysteines alkylated to render these chemically inert. A protease is then used to digest proteins into peptides which are desalted and transferred into an appropriated buffer for LC-MS.

Proteolytic digestion A step that comes in multiple flavours is the protease digestion, many of which are listed in Table 1.1. Trypsin has established itself as the standard tool: it cleaves the amide bond N-terminally of the basic amino acids Lys and Arg with high efficiency, allowing a well-distributed charge distribution (2^+ to 4^+) and a peptide size typically favourable for MS/MS (m/z in the range $700-2000$).[126–128] However, in

Table 1.1. Characteristics of selected proteases relevant to this thesis.

Protease	**Family**	**Cleavage site**[a]	**Missed cleavages**	**Comment**
ArgC	cysteine protease	R' (K')[b]	0-2 (0-3)[c]	Carbamylation (*N-term*) and amidation (*C*-term) observed[po]
AspN	zinc metalloprotease	'D	0-3 (0-4)[c]	Used in denaturant before trypsin 80% activity in 4.5M urea[p] 100% activity in 1M Gd.HCl [p] 16% activity in 3M GD.HCl [p] Carbamylation (*N-term*) observed
Chymotrypsin	serine protease	F/Y/L/W' (M')	0-4	Orthogonal to trypsin
Elastase	serine protease	A/V/I/L/ T/S/G'	0-9	Used in difficult cases[123] Used with trypsin for xL-MS[124]
GluC	serine protease	E' (D')[d]	0-3 (0-4)[c]	Cleavage specificity is pH dependent Carbamylation (*N-term*) observed
LysC	serine protease	K'	0-2	Used in denaturant before trypsin 86% activity in 4M urea[r] 12% activity in 1M Gd.HCl [r]
Trypsin	serine protease	R/K'	0-2	Proximity of D, E, pS, T leads to missed cleavages[125]

[a] Cleavage sites are indicated by a mark ['']. Unless otherwise stated, cleavage before proline is not possible.
[b] ArgC can also cleave at K residues, but with much less efficiency.
[c] Expected number of missed cleavages in database searches with relaxed specificity.
[d] At pH=8, Asp residues are also cleaved but at much lower rate than Glu residues.
Data from:[125] Giansanti *et al.*; [p] Promega technical documentation, [r] Roche technical documentation, and [po] personal observation.

some cases when serine protease inhibitors are present or the protein sequence of interest does not present many Lys/Arg, alternative proteases can be considered.[125,129,130] Combinations of trypsin with LysC, a protease that retains activity in high urea, or AspN, which retains activity in guanidinium hydrochloride solutions, have allowed to improve the digestion efficiency. The combination of trypsin with elastase has been highlighted for preparation of crosslinked samples.[124] Recently, commercial kits with thermostable trypsin have been developed for accelerated digestion at higher temperatures, taking advantage of simultaneous protein denaturation (RapiDigest™, SmartDigest™ etc.) as well as proteases immobilised to a solid support to reduce contaminants.

Glycosylation-dependent proteases In the last three years, enzymes from commensal bacteriae have been described that catalyse peptide bond hydrolysis dependent on the glycosylation context and have subsequently been applied to glycoproteomics projects. Flavastacin from *Flavobacterium menigosepticum* cleaves *C*-terminally

of complex-type *N*-glycosylated Asn and has been combined with trypsin for *N*-glycoproteomics.[131]

A major breakthrough for the analysis of mucin-type *O*-glycosylation (Figure 1.6) has been the description of the use of StcE, a bacterial mucinase from *E. coli*, by the Bertozzi group.[87] This zinc metalloprotease recognises the sequon S/T^{MUC}-X-S/T, which it cleaves C-terminally after the X.[87] Similarly, the OpeRATOR® protease from *Akkermansia muciniphila* commercialised by Genovis Inc. cleaves N-terminal of asialo core 1 mucin-type glycosylated S/T (and is therefore combined with sialidases). Both proteases offer the unique advantage of being able to digest densely glycosylated domains, such as mucin domains, which would not be accessible to workhorse proteases such as trypsin.[132] These discoveries have led to new glycoproteomics sample preparation workflows such as the EXoO method for site-specific extraction of *O*-linked glycopeptides (see below).[133,134]

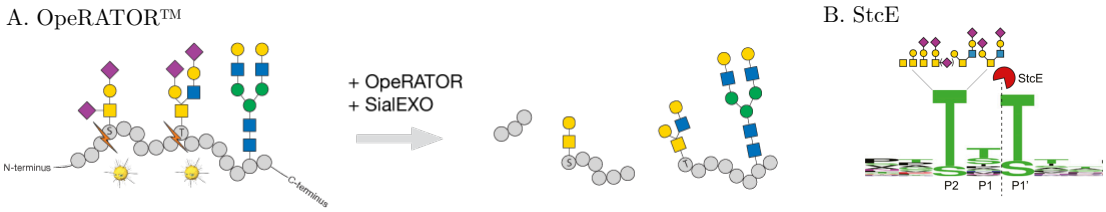


Figure 1.6. Recently discovered *O*-glycosylation-dependent proteases. A. The OpeRATOR® commercialised by Genovis Inc. cleaves C-terminally of de-sialylated core 1 mucin-type *O*-glycans. Adapted from[135]. Multiply glycosylated peptides remain common and the use of electron-based methods is required for analysis.[136] **B.** StcE cleaves at the sequon S/T^{MUC}-X-S/T *C*-terminally of X, where MUC is a mucin-type *O*-glycan. Adapted from[87].

Deglycosylation If needed, glycans can be removed either from intact proteins or from digested glycopeptides, depending on application. Glycans can interfere with regular proteomics experiments by impairing on ionisation or data analysis. Equally, deglycosylation can be combined with immobilisation (using hydrazide chemistry[137]) or labelling steps for glycosite mapping[138] (*de-glycoproteomics*). Generally speaking, enzymatic and chemical methods for deglycosylation exist, the latter being rather blunt (thermal β-elimination in mild alkali media − BMAD) and leading often to partial sample degradation or loss of information.

Several enzymes for removal of *N-glycans* exist. *PNGase F* is a notorious amidase for efficient and reliable removal of most *N*-glycans and has become the dominant workhorse in many laboratories. However, α1-3 core fucosylated glycans (but not α1-6 Fuc) as well as peptides carrying their N-glycan on the *C*- or *N*-terminal Asn (as from flavastacin

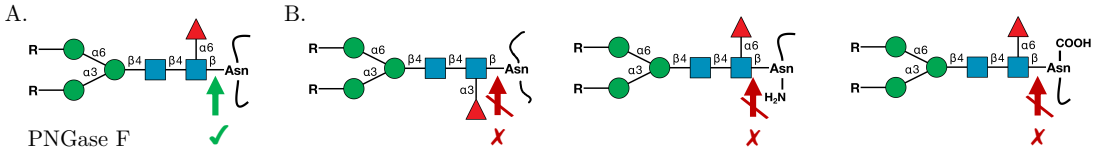


Figure 1.7. PNGase F resistant *N*-glycopeptides. A. PNGase F cleaves all *N*-glycans without fucosylation and those with α1-6 core fucosylation. **B.** PNGase F does not cleave *N*-glycans with α1-3 core fucosylation or those glycopeptides carrying the glycan on the peptide *N*- or *C*-terminus.

digest) are so-called "PNGaseF resistant" glycopeptides[139] – the latter can be bypassed by *N*-terminal succinylation[140] (Figure 1.7).

Beyond PNGase F, *endoglycosidases* hydrolyse glycosidic bonds inside complex glycans. This feature has been exploited to use endoglycosidases such as EndoS/D/H[141,142] to cleave different types of high mannose/complex *N*-glycans leaving the terminal *O*-GlcNAc on the peptide, thus allowing to identify the site of attachment or even enzymatic remodelling[143]. However, endoglycosidases typically have a high specificity for their substrates limiting the types of bonds that can be hydrolysed.[102] Similarly, *exoglycosidases* such as neuraminidases, fucosidases, glucosidases or galactosidases can be used to remove specific monosaccharides at the ends or branching point of glycans, but often recognise only a specific linkage geometry.[144] Although readily available tools for deglycosylation, the fact that each of these enzymes is having very distinct glycan structure/monosaccharide linkages they can recognise and hydrolyse has channelled their use towards specialised applications rather than a promiscuous use.[144,145] Therefore, efforts to constitute *enzyme cocktails* aiming at a broader applicability have emerged[146], including commercial sets (i.e. NEB protein deglycosylation mix II), but their use remains rather restrained.

Encouraging discoveries have been recently made in enzymes that are able to remove *O-glycans*. Whereas formerly limited to the previously mentioned heated β-elimination, the commercialisation of NEB's *O*-glycosidase as well as Genovis' OglyZOR® able to hydrolyse core 1 and core 3 disaccharides holds to promise of further useful tools to emerge.

Clean-up & desalting Technologies for desalting and clean-up of samples from a broad range of buffers and biological sources that would interfere with mass spectrometry have emerged. Salts, detergents and other ion-suppressing contaminants lead to adduct formation, increased chemical noise and consequently to lower peak intensities and lower

protein sequence coverage.[147] Whereas historically, protein precipitation using organic solvents – typically acetone – was performed, new solid phase extraction techniques have emerged that improve sample recovery. C18 and SCX materials come in various formats, such as cartridges, StageTips[148] and commercial ZipTips™, and have been used both for *in-solution* or *in gel* digests. Although widely applied, desalting short or highly hydrophilic *N*-glycopeptides with C18 (or Oasis HLB) has been suggested to have partial sample loss[149,150] and carbohydrate-based HILIC[150] and PGC[151] materials have been proposed for such cases.

Solid support-assisted sample preparation Protocols that combine both digestion and clean-up have been developed, aiming at improving reproducibility, sample loss, speed and detergent compatibility. The in-StageTips (iST) method allows to perform all steps in a single StageTip vessel.[152] Solid-phase enhanced sample preparation (SP3) beads surface-functionalised with carboxylic acids allow extraction of proteins from solutions with high amounts of detergents and chaotropes, yet loose efficiency in high salt ($>$ 1M).[153,154] Filter-aided sample preparation (FASP) allows washing of detergents and high salt using a molecular weight cut-off (MWCO) filter, before digestion.[155,156] However, FASP has been reported to drop performance in the low µg range and reproducibility issues are noted.[157] For more niche applications, the gel-aided sample preparation (GASP) variant thereof has been proposed for scalability and application to intact cells.[158] More recently, the S-Trap method that uses a porous material as filter has been suggested as a broadly applicable, although costly, sample preparation method.[159] Its efficiency has been reported superior to FASP.[160]

1.3.2.3. Fractionation, enrichment at peptide level

Fractionation Fractionation may be required to achieve comprehensive proteome coverage considering that the mass spectrometer lacks the required scan speed to analyse all individual peptides in complex mixtures.[89,161] Offline fractionation separates peptides based on their physico-chemical properties. As most MS workflows will employ C18 reverse-phase at a later stage as part of their LC-MS setup,[162,163] orthogonal techniques are used, such as strong cation/anion exchange chromatography (SCX, SAX),[164,165] high-pH reverse phase chromatography (HpH),[166] hydrophobic interaction chromatography (HILIC),[167] isoelectric focussing[168] or mixed-phase chromatography.[105]

PTM-specific enrichment Additionally, peptides carrying specific post-translational modifications (PTMs) can be enriched to overcome sub-stoichiometry and thus enabling a higher number of identifications by MS.[169,170] For example, phosphopeptides are readily captured using titanium dioxide IMAC matrices.[171] A multitude of PTMs exist,[121] this section will focus solely on the enrichment of glycopeptides.

Importance of glycopeptide enrichment This has been a major field of research in glycoproteomics, since the success of any large-scale glycoproteomics experiment depends on the quality of the enrichment, as non-glycosylated peptides would otherwise dominate the data acquisition. The difficulties encountered in glycosylation analysis are multiple: for the peptide same concentration, glycopeptides exhibit significantly *lower signal intensities* (10 - 50% compared to unglycosylated peptides with identical sequence)[172]. Similarly, *overlapping retention times* of highly glycosylated peptides can hamper detection.[51] In addition, glycoproteins can have considerable *micro-* and *macroheterogeneity*, i.e. different glycoforms for the same protein, lowering the abundance of a single species and thus impairing on detection. Macroheterogeneity refers to the occupancy of different glycosites on a protein, i.e. the number of glycosites on a protein that carry a glycan. Microheterogeneity describes the glycan distribution at a single site on a given protein. These properties differ between *N*- and *O*-glycopeptides.[51] *N*-glycosites are believed to be highly or fully occupied, whereas the occupancy of *O*-glycosites spans the range between a few percentages to fully occupied. On the other hand, microheterogeneity is much higher for *N*-glycans than for *O*-glycans. As such, in either case enrichment is necessary for samples of higher complexity.

Methods for glycopeptide enrichment have been thoroughly reviewed in several publications,[173–179] the following aims at briefly highlighting subjectively selected, currently

applied methods and promising new discoveries.

Lectins Lectin resins are one of the earliest applied methods, these glycan-binding proteins recognise and bind with high affinity to their specific glycan epitopes (Figure 1.8.a). Concavalin A (ConA) and Wheat Germ Agglutinin (WGA) are commonly used, however they only enrich a specific subset of glycopeptides. ConA lectins recognise oligomannosyl motifs in *N*-linked glycans and WGA binds to sialic acids and D-GlcNAc. Several combinations of lectins for enrichments, additional derivatisation with boronic acids as well as immobilisation techniques have been tried to overcome limitations in the efficiency, leakage and selectivity bias of lectin enrichments. Whereas their use as staining or imaging reagents has been widely adopted, their use for enrichment before mass spectrometry is less frequent. To date, Jacalin is the only lectin that has been used for an enrichment geared towards mucin-type *O*-glycopeptides, however it's high cross-reactivity and nonspecific binding impairs on performance.[173,174]

ZIC-HILIC Zwitterionic HILIC (ZIC-HILIC) has been extensively used to enrich glycopeptides based on their increased hydrophilicity[180, Fig.9] and is currently the most widely applied method (Figure 1.8.b). Its performance in *N*-glycoproteomics has been recently evaluated and branded as the most universal enrichment method.[181] However, it cannot distinguish glycopeptides from many hydrophilic non-glycopeptides, thus requiring further off-line fractionation for more complex samples.[173,174] Its performance on *O*-glycopeptides has neither been well studied, as such the quality of the enrichment of *O*-glycopeptides remains unclear.

Benzoboroxole-functionalised dendrimer A promising new discovery for the enrichment of glycopeptides comes from a modification to boronic acid resins. Monovalent boronic acids have been used frequently for the capture of the cis-diol motive in the literature, and thus for glycopeptides, but the methods suffer from weak interaction and a very high background capture. Also, a bias towards high mannose *N*-glycans has been reported.[181] A new version using a benzoboroxole dendrimer[182] has shown to be effective by using synergistic, weak, reversible covalent capture of the *cis*-diol motive enabling as well the enrichment of *O*-glycopeptides including single *O*-GlcNAc-glycopeptides (Figure 1.8.c).[182,183]†

†A possible interpretation is that the reduction of the available coordination number from 3 in boronic acids to 2 in the benzoboroxole might have a key influence on the selectivity. The question whether the

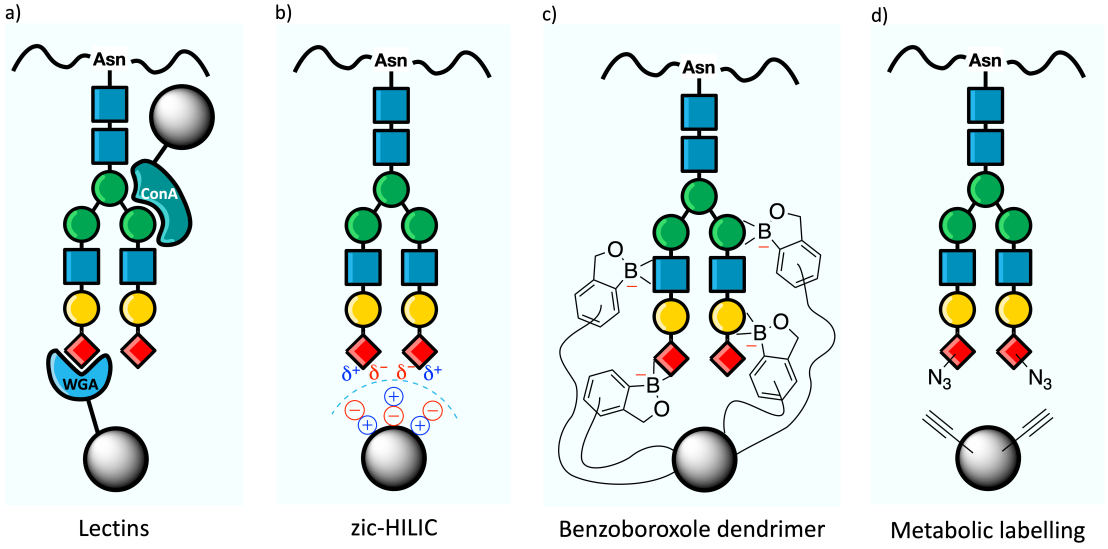


Figure 1.8. Selected *N*-glycopeptide enrichment principles. a. Immobilised lectins recognise specific motives on *N*-glycan structures and bind with weak K_D. **b.** Zwitterionic hydrophilic interaction liquid chromatography (zic-HILIC) takes advantage of hydrophilc / polar interactions that occur with a zwitterionic (amphoteric) phase which contains two oppositely charged groups in close proximity at stoichiometric ratio.[186] The weak, amphoteric hydrophilic and electrostatic interactions allow retention of the highly polar and partially charged *N*-glycopeptides. **c.** Benzoboroxoles bind 1,2- and 1,3-*cis*-diols with low affinity and specificity. Their coupling onto a dendrimeric polymer allows to use the reversible covalent interactions synergistically for enrichment. This technique has emerged only very recently[182] and its selection here is subjective. **d.** Metabolic labelling takes advantage of the limited permeability of the cellular machinery to unnatural monosaccharides, added to the cell culture media. Once incorporated into cellular glycans, they can be covalently enriched by copper-catalysed azide-alkyne cycloaddition.

Partial peptidolysis Another approach towards increasing the relative amount of glycopeptides in a sample has been the exploitation of glycosylation as a protection from proteolytic digest. *Unspecific proteolysis* with enzymes such as pronase has been used to digest *N*- and *O*-glycopeptides until only the glycan and a few amino acid-long peptide remain. These remnants can be further enriched using PGC (which does not enrich tryptic glycopeptides efficiently[182]). Although a destructive approach, it has been successfully applied for the characterisation samples with limited complexity.[187] In particular, limitations are dictated by the identification of the short peptide part at the data analysis stage, requiring searches with unspecific cleavage.

enrichment could be enhanced by electron-withdrawing groups on the aromatic ring by conjugation with the empty p-orbital on the boron to strengthen *cis*-diol interaction of benzoboroxoles has not yet been assessed in the context of glycopeptide enrichment. This effect is seen on the free phenylboronic acid and synthetic publications suggest this could be the case.[184,185]

Nonetheless, the use of *glycosylation-dependent enzymes* has proven an interesting emerging technology for *O*-glycopeptide enrichment by partial peptidolysis. In the site-specific extraction of *O*-linked glycopeptides (EXoO), glycoproteins are digested first with trypsin, followed by lysine guanidination[138] and immobilisation on a solid support by conjugation with the *N*-terminus of the peptide. *O*-linked core 1 glycopeptides are then selectively released using the OpeRATOR protease (*vide supra*), yielding *N*-ragged tryptic peptides.[133,134]

Metabolic labelling Metabolic labelling exploits the enzymatic promiscuity of the biosynthetic machinery to incorporate unnatural monosaccharides with bioorthogonal reactive tags into cellular glycans. These can subsequently be enriched or visualised by high yielding bioorthogonal reactions (Figure 1.8.d). Elegant proteomic readout for metabolic labelling strategies, such as IsoTaG, have been developed.[188–190] However, there are several drawbacks using metabolic labelling: the metabolic incorporation efficiency is low, unnatural sugars can perturb cellular systems or even be toxic and epimerases convert one monosaccharides into another.[191] Importantly, a long overlooked issue is the artificial, non-enzymatic cysteine *S*-glycosylation induced by the use of per-*O*-acetyated unnatural monosaccharides[192] and sugars with azides at C-2[193] leading to false identifications of presumed glycosites.[192,193] Mechanistic investigation showed that in alkaline protein microenvironments, per-*O*-acetylated monosaccharides undergo base-promoted β-elimination to form thiol-reactive α,β-unsaturated aldehydes, which then react with cysteine residues via Michael addition.[194] A new generation of better characterised unnatural monosaccharides such as 1,3-Pr2GalNAz[195] or 1,6-Pr2GalNAz[194] has been suggested to bypass these limitations.

In summary, no method currently exists for the broad, efficient, unbiased and non-disruptive enrichment of glycopeptides, as opposed to for example titanium dioxide matrices for phosphoproteomics,[171] which has been one of the limiting factors in glycoproteomics.[114,115]

1.3.3. Nano liquid chromatography (nLC)

The most widespread implementation for LC-MS is the online coupling of a reverse phase high performance liquid chromatography (RP-HPLC) to the electrospray ion source (described in 1.3.4.1), taking advantage that the buffers applied for separation are salt-free, volatile and therefore compatible with mass spectrometry.[162] In RP-HPLC, peptides are retained based on their interactions with the hydrophobic stationary phase and the mobile phase, the partition between both phases relates to the total hydrophilicity of its amino acids.[196] The separation of the peptide analytes occurs by applying linear gradients of increasing organic solvent content (typically acetonitrile) according to their hydrophobicity in the presence of small amounts of acid (i.e. 0.1 % formic acid), which serves to ensure protonation of acidic groups, increases retention and sharpens peaks. The combination of the high peak capacity of RP-HPLC with the low flow rates required for electrospray ionisation (see 1.3.4.1) enhances the accessibility of analytes to the mass spectrometer by providing time to sequence individual analytes, widening the dynamic range.[197,198]

Although often overlooked, nLC settings can have a major impact on the performance of a given method. Electrospray is a concentration-dependent ionisation source, i.e. its signal is inversely proportional to the flow rate. Nano-flow ESI has been described to enhance the detection of carbohydrates and glycopeptides when compared with micro-flow ESI, due to their hydrophilicity and lack of surface activity.[172,199] On the other hand, micro-flow LC has been shown to strongly extend the dynamic range of TMT experiments and thus impact on the quality of the quantification.[200]

The addition of "neutral" sugar units, for example an extra antenna on an *N*-glycan, does not alter the retention times in C18-RP significantly: different *N*-glycoforms of the same acidity often co-elute.[51] In contrast, *O*-glycopeptides differ in their RP behaviour: acidic and neutral elute at different times, multiple *O*-glycosylation lead to RT shifts > 5 min, facilitating separation.[51] Therefore, in *N*-glycoproteomics, two additional online nLC chemistries are occasionally used: HILIC[201] and porous graphitized carbon (PGC)[202,203], the choice of which depends on the glycopeptide type − i.e. the digestion procedure. Both are particularly advantageous in cases where separation of highly similar glycopeptides with identical peptide part and small changes in glycan structure is required (HILIC), or in protocols leading to very hydrophilic analytes such as digestion with pronase (PGC). Combinations are possible: specialist laboratories occasionally use an integrated C18-PGC-LC setup upon pronase digestion, where the sample is first loaded onto a C18 precolumn and the flow through, containing hydrophilic glycopeptides

that are not retained, is directed online towards a PGC precolumn.[187] Glycopeptides are retained on either stationary phase and elution is performed sequentially over the respective analytical columns into a combined acquisition run.[187]

1.3.4. Mass spectrometry

Mass spectrometer are built up by (at least) three components, found in all instruments: ion source, mass analyser and detector. The ion source transfers the analyte into the gas phase. High vacuum technology ensures high quality of measurement and ion optics allow the transfer of charged ions into the mass spectrometer, where further ion optics steer the flux of charged analytes across the compartments with different pressures. Mass analysers manipulate the trajectory of ions: they can influence velocity, momentum, resonance and energy of ions and by this filter, deflect or separate ions. These properties can be related back to the m/z ratio (mass-to-charge ratio). Finally, the detector records the current induced by the ions. Electronics or data processing subsequently converts it to the m/z domain. The instrument delivers a mass spectrum plotting ion abundance against the m/z ratio.[89,108]

1.3.4.1. Ionisation techniques

In biological mass spectrometry − as opposed to the analysis of small molecules or the use of mass spectrometry in the physical sciences − so-called "soft" ionisation methods are used: matrix assisted laser desorption ionisation (MALDI − not discussed further) and electrospray ionisation (ESI).

ESI allows continuous ionisation from a liquid phase.[204] The liquid running through the narrow capillary needle, the emitter, is sprayed into the gas phase due to the repulsive electric potential (several thousand volt) between the emitter and the mass spectrometer of same polarity as the analyte. This results in the formation of a Taylor Cone: droplet dispersion occurs when the electrostatic force exceeds the surface tension of the analyte solution.[205] Solvent evaporates from these airborne, highly charged droplets, which subsequently collapse when the Rayleigh limit is exceeded − that is the surface tension is smaller than the repulsion of charges in the droplet, which subsequently undergoes droplet fission (Coulomb fission). This process carries on until the desolvated analyte ions remain. The exact mechanism of the ionisation process is not yet fully understood, two main theories namely the ion evaporation theory[206,207] and the charged residue model[205,208] explain the final production of the gas-phase ion. Evidence suggests that the first is more applicable to peptides and small molecules, while the second to larger analytes such as proteins.[209,210] The efficiency and thus signal intensity of ESI is dependent on droplet size, it can be improved by factors such as low flow rates[205] or LC additives[211,212] that modulate the surface tension of the liquid. Nanoflow-ESI,

in contrast to microflow-ESI, is highly efficient in generating multiply charged ions[213] and is today's standard ionisation technique. For analytes that form lower than ideal charge states (i.e. some glycopeptides that are difficult to ionise), supercharging additives such as 0.1 % *m*-NBA can be used to increase charge density and thus ESI ionisation efficiency.[214,215]

Mass spectrometers for proteomics almost exclusively operate in positive mode, where quasimolecular ions are created by the addition of one proton, i.e. $[M + H^+]$, or several i.e. $[M + nH^+]^{n+}$. Positive mode data acquisition is employed in conjunction with acidic LC solvent (0.1 % FA), which encourages protonation and enhances separation by peak sharpening through neutralising negative charges such as carboxylic acids and by ion pairing. As a consequence, tryptic peptides have predictable charge states. They are typically doubly protonated: at the *N*-terminus (H^+ also coordinated by the peptide backbone) and at the *C*-terminus carrying a basic Lys or Arg residue. Peptides may carry additional charges if they have additional basic residues due to a missed cleavage, if the sequence contains residues such as His, or if they are large in size (>15 AA).[216,217]

The negative mode is occasionally used for the analysis of carbohydrates[218], released derivatised glycans[219] and (occasionally) of glycopeptides.[220] The negative mode provides improved ionisation efficiency for some *O*-glycopeptides and for sialic acid carrying glycopeptides, in addition to diagnostic ions from the $[M - H]^-$ ions resulting from hydroxyl groups.[220] Although this has been exploited in glycomics, glycopeptide analysis workflows are closely related to well-established proteomics methods. In fact, the *N*- and *O*-glycopeptide analysis is almost exclusively performed in positive ionisation mode.[114,115]

1.3.4.2. Mass analysers and detectors

Mass analysers separate analytes by their m/z in space and time. They are described in terms of resolution, mass accuracy, sensitivity and speed.[221,222] The resolution relates to the minimum difference between two distinct m/z peaks that can be distinguished. [‡] The mass accuracy is obtained by comparing measured to expected value of calibrants, the difference normalised by the mass is reported in ppm. The acquisition time is the time needed from the entry of the analytes into the unit until the spectrum is recorded, the acquisition speed describes how many spectra are generated in a given time.[224,225] Figure 1.9 shows important mass analysers relevant to this thesis, discussed further in the text.

[‡]It is commonly given as peak resolution defined as $(m/m_p)_{50\%}$ where m_p is the FWHM.[223]

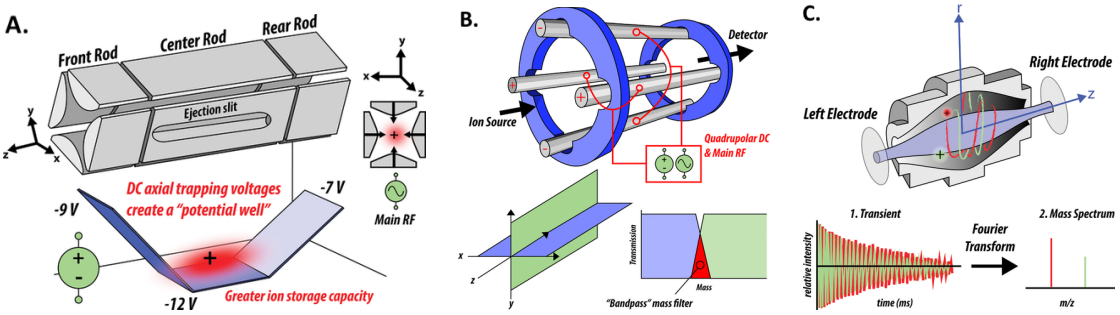


Figure 1.9. Overview of mass analysers relevant to this thesis. A. Linear ion trap. **B.** Quadrupole mass filter. **C.** Orbitrap mass analyser. Details are discussed in the text. Image reproduced from[226] with permission from ©Wiley-VCH.

High resolution mass spectrometers (resolving power $(m/m_p)_{50\%} > 10000$[223])are able to distinguish peptide isotopic pattern. The analyte can contain the heavier isotope of a chemical element, such as ^{13}C, ^{15}N or ^{2}H. These isotope peaks differ by one neutron, the peak intensity of the isotope results from the elemental composition weighted by the natural abundance of each isotope.[224,225] Consequently, isotope pattern can be used to calculate the charge state of the analyte in the MS^1 scan (or MS^2 for large glycopeptide fragments).[227] *N*-glycopeptides and large *O*-glycopeptides have a different isotopic pattern than regular peptides due to their large carbon content.

The operation mode of mass analysers to provide m/z separation can be based on momentum, velocity, selective stable trajectories or resonance frequencies and are distinguished in scanning/non-scanning as well as trapping/non-trapping types.[224] The following section provides the background for the discussions in chapter 3.

Linear ion trap (IT) Linear ion traps (IT) consist of four rod-shaped electrodes which are able to trap ions in space and scan for their m/z (Figure 1.9.A).[228] The ions are confined radially between the rods through a fast oscillating alternating current (AC) at radio frequency (main RF). Additionally, the ions are trapped axially by applying a valley-shaped DC potential: the front rod and rear rod of the ion trap have a higher potential than the axial center part, steering the ions to its lower potential. The trapped ions are in a secular motion (forced motion of an ion in a field), with larger molecules moving slower than smaller ones. A buffer gas, often helium, is used to focus/trap the ions more efficiently through collisional cooling. The main RF amplitude controls which ions, defined by their m/z, will have a stable trajectory in the trap. Ejection slits are carved on two opposite center rods (exit rods), which allows ions that do not have a stable trajectory and cannot be contained in the ion trap to be ejected.[228] After trapping of the

ionised analyte mixture, an additional RF is applied and progressively increased. When the frequency of this additional RF matches the resonance frequency of a given ion (i.e. a given m/z), it starts resonating back and forth until the ion packet is ejected through the ejection slit (resonant ejection). Scans can be performed by ramping the main RF, leading to sequential ejection of all all m/z ion populations.[228] The ions ejected from the trap through the exit rods hit a detector, typically an electron multiplier.[229] This detector amplifies the signal: upon impact, it generates a cascade of secondary electrons which are eventually measured by an anode.[230] Calibration of the ion trap to external standards allows to calculate the m/z ratio of the ion packet from the main RF and exit rod AC frequency. Ion traps have high sensitivity and high overall scan speed, with the trade-off of low resolution and low mass accuracy.[231] The resolution (and to a limited extend the mass accuracy) of the ion trap can be increased by slowing the main RF ramp of each scan, at the cost of lower scan speed.[89]

Quadrupole Quadrupole mass analysers consist of four rods. A quadrupolar RF + DC potential is applied to its rods by applying the same sign of voltage to opposite rods, i.e. the opposite rods are either both positive or both negative, creating a quadrupolar and slightly fluctuating electric field. This confines the ions radially: when the ions enter the quadrupole they adopt a corkscrew-like motion.[232] At a given RF/DC voltage, only ions of a specific m/z have stable trajectories that allows them to pass through the quadrupole. Ions with unstable trajectories collide with the rods. As such, a quadrupole is an efficient mass filter able to isolate single ion populations from a continuous stream of ions.[232] However, in contrast to the ion trap, the quadrupole is not able to store the isolated ions. Quadrupoles exhibit fast scan speed in a narrow scan range, which has been used in applications for scanning, or for isolating individual ions (see SRM).[89] The acquisition of large scan ranges and thus MS^1 spectra is slow.[89] On their own, they are often coupled with electron multipliers as detectors, located at the rear end and continuously recording the ion current.[89] In high-end mass spectrometers, quadrupoles are mostly employed as mass filters in combination with a downstream TOF or Orbitrap mass analyser.[224,225]

Time of Flight (TOF) The principle of TOF mass analysers relies on the recoding the time of flight t for analytes for a given distance d. Ions are collected, i.e. focussed and bundled, and accelerated to a velocity v by a pulsed potential V that is converted into kinetic energy E_{kin}:

$$E_{kin} = \frac{1}{2}mv^2 = Vze \tag{1.1}$$

Ions with the same energy but different m/z will take a different time t to pass through the defined distance d, small ions will reach the detector (typically an electron multiplier) faster than large ions. Accurate measurement of time allows to obtain the m/z value:

$$m/z = 2eV \left(\frac{t}{d}\right)^2 \tag{1.2}$$

TOF mass analyser exhibit fast scan speed while providing high mass accuracy, however the resolution suffers from the need to initially distribute the kinetic energy evenly among all ions. Improvements in resolution have come from the development of reflectors, that allow to elongate the drift distance while symmetrically restoring the kinetic energy of ions of same m/z.[224,225]

Orbitrap Orbitrap mass analysers are Fourier transform mass analysers (as are ion cyclotron analysers).[233] It is made up of a coaxial central electrode and a barrel-shaped outer electrode. Ion packets are collected in the upstream C-trap (bent quadrupole with one exit rod) before being injected tangentially at an offset from the center of the orbitrap electrode. The bent shape of the orbitrap subsequently squeezes the ions towards the central electrode, leading to an orbiting motion of the ions around the central electrode. A harmonic oscillation of the ions in the axial dimension follows, determined by their m/z. This oscillating movement of the ion packets induces an image current in the left and right electrodes of the Orbitrap analyser, recorded as signal intensity over time. The measured current is converted from the time domain to the frequency domain using the Fourier transformation. The frequency of axial oscillation ω_z is converted to an m/z ratio by calibration of an instrument constant k:[233]

$$\omega_z = \sqrt{\frac{k}{m/z}} \tag{1.3}$$

Orbitrap mass analysers combine high mass accuracy with high resolution which is important for MS^1,[234] and for MS^2 in *de novo* sequencing (chapter 2). However, their speed is much lower than for other detector types and the sensitivity is reduced due to the amount of ions being required to generate a signal.[221,234] They cannot capture a constant stream of ions and are therefore used in conjunction with a C-trap that stores and collects ions.[231] Orbitrap-based mass spectrometer are to date the most common instrument platform in bottom-up proteomics.[89]

Other types of mass analysers are not used in this thesis and reviewed elsewhere.[224,225]

1.3.4.3. Tandem mass spectrometry & peptide fragmentation

The composition of a peptide cannot be obtained from its intact mass. Fragmentation of peptides along the amide bond into their amino acid building blocks allows to obtain sequence information. Tandem MS involves three steps: Recording the intact mass of all peptides eluting simultaneously in a first scan, the MS^1 scan; isolating a single ion population (= precursor ion); fragmenting the precursor ion into its building blocks by a combination of physical force and chemical reaction in the gas phase in a MS^2 scan. Three fundamental mechanistic types are used in peptide fragmentation, which are sub-divided in methods that differ in their technical implementation:

- Protonation-mediated fragmentation such as collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD, formerly higher energy C-trap dissociation[235]). Both are classified as low-energy CID methods (kinetic energy < 1 keV). HCD is not to be confused with high energy CID (HECID) which occurs at 1 keV - 20 keV and is not routinely used in proteomics.[236] CID (also termed "ion-trap CID") is typically performed in an ion trap mass analyser, HCD (also termed "beam-type CID") is performed in a quadrupole mass analyser, often termed "collision cell".

- Electron-based mechanism such as electron-capture dissociation (ECD) and electron-transfer dissociation (ETD). ECD needs an ultra-high vacuum and is usually implemented on FT-ICR instruments, not routinely used in proteomics.

- Photon-mediated fragmentation, categorised by the energy of the photon: ultraviolet photodissociation (UVPD) and infrared multiphoton dissociation (IRMPD). The latter is also usually implemented in FT-ICR instruments and therefore not routinely used in proteomics. Neither of these methods has been used in this thesis.

Modern tandem mass spectrometers combine mass analysers and use their modular build to parallelise ion collection, filtering and scanning or to perform advanced scan types combining several modules. For example, to record a tandem MS scan one analyser can act as a mass filter isolating the precursor ion population (such as a quadrupole or an IT), while the second mass analyser is used to record the MS^2 scan after fragmentation.[89]

Fragment ions The fragment ions resulting from peptide fragmentation are termed as defined by the nomenclature of Roepstorff, Fohlman and Biemann.[237–239] Peptide fragmentation occurs along the amide bond, typically generating ions derived from the N-terminus (a, b and c-ions) and the C-terminus (x, y and z-ions),[243] as shown in

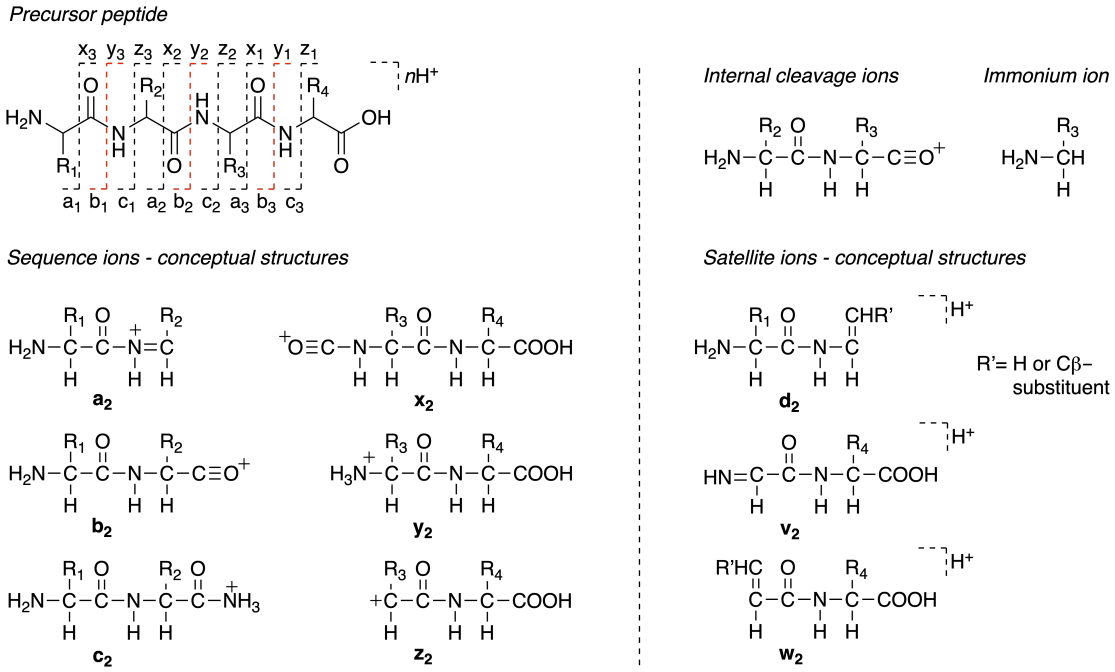


Figure 1.10. Peptide fragment ion nomenclature. Numbering follows nomenclature by Roepstorff, Fohlman and Biemann.[237–239] The exact site of bond breakage and the localisation of the retained charge determines whether sequence ions are termed *a*, *b* and *c*-ions (*N*-terminal) or *x*, *y* and *z*-ions (*C*-terminal). For *a*, *b* and *y*-ions, neutral loss of water (ion^0) or ammonia (ion^*) is observed depending on amino acid composition (see text). Internal cleavage ions and immonium ions are formed by two subsequent internal cleavages. Satellite ions are formed by side chain cleavages observed at higher (typically keV) energies[239,240] or occasionally in MS^3 strategies[241,242], and allow differentiation of isobaric residues Leu and Ile.[239,241,242] R' denotes the remaining substituent, if any, after side chain cleavages − that is either H or the secondary Cβ substituent of Ile and Thr. $w^\bullet$-ions are observed in ETD as a result radical side chain cleavages (disulfides etc.). The structures shown are conceptual structures, it is not the connectivity of the ions formed in the gas phase. Moreover, higher charged fragments (i.e. 2^+) can be formed if the precursor ions carries enough charges.

Figure 1.10. Noteworthy, *a*-ions can result from loss of CO from *b*-ions.[244] In addition, other fragment ions can be observed. For fragments containing RKNQ, neutral losses of ammonia (-17 Da) can occur, denotated as $a^\star$, $b^\star$ or $y^\star$. Similarly, for fragments containing STED neutral loss of water (-18 Da) is denotated as a^0, b^0 or y^0. Internal fragments resulting from double backbone cleavage (prevalent at Pro and His),[245] as well as single amino acid immonium ions[246,247] or satellite ions (*d*, *w* and *v*) resulting from side-chain cleavages[239,240] are also possible − depending on the used energy. The individual fragment ions are formed with different probabilities, which leads to a non-uniform but reproducible distribution of fragment ion intensities in the MS^2 spectrum.[248] The processes leading to fragmentation are a system of competing gas phase rearrangements.[243] Which

fragments are observed is the results of statistical chemical kinetics, the likeliness of a fragmentation pathway is dependent of parameters such as amino acid sequence, the applied fragmentation technique, the energy and reaction time used for fragmentation.[243] Many of these parameters can be controlled by the operator, and are typically normalised on analyte m/z (see NCE).

Ion trap CID Collision induced dissociation (CID) is a low-energy fragmentation technique that uses excitation and subsequent collision of the analyte molecules with an inert gas.[249–251] The implementation in ion traps is very common. After the initial MS^1 scan, all arriving ion populations are collected and stored in the high pressure chamber of the ion trap, followed by ejection of ions except the desired precursor m/z. Activation is achieved by bringing the isolated ion population to resonance and undergoing multiple collisions with the inert gas, typically helium. The kinetic energy is converted into molecular vibrational energy distributed over the entire peptide (ergodic), which accumulates until the molecule fragments by one of its most accessible fragmentation pathways, usually the amide bond in the peptide backbone. CID employs energies between 1-100 eV on the millisecond timescale.[249–251]

In low-energy CID, the disintegration of activated ions into fragments follows charge-directed dissociation pathways, which are described mechanistically by the *mobile proton* model.[255–257] Protons that are transferred to the analyte during ionisation in positive mode ESI are sequestered to the most basic residues, that is the side chains of Lys, Arg, His and the N-terminus. Upon excitation in the ion trap (i.e. "slow heating"), the proton can populate energetically less favourable sites on the amide bonds of the peptide backbone. The following disintegration has been described as a competing set of reactions in the *peptide fragmentation pathway framework*.[243] In the IT-CID-dominating b_x-y_z fragmentation pathway, protonation of the amide nitrogen of the backbone leads to a weakening of the peptide bond and increases electrophilicity of the amide carbonyl, leading to nucleophilic addition of the N-terminal neighbouring amide oxygen (see discussion in Figure 1.11). This results in the formation of a 5-membered oxazolone derivate and elimination of the linear amine. Depending on basicity, the proton can be retained on either dissociation product, either the cyclic b-ion or the linear y-ion.[258,259] Alternative peptide fragmentation pathways are possible and have been described for the cleavage of the most N-terminal amide bond,[260] cleavage due to side chain nucleophilic attacks (e.g. the histindine effect[255]) and charge-directed neutral losses from amino acid side chains.[261–263] Further fragmentation of the generated b- and y-ions can lead to the formation of the aforementioned internal ions,[264] immonium ions[265] and a-ions

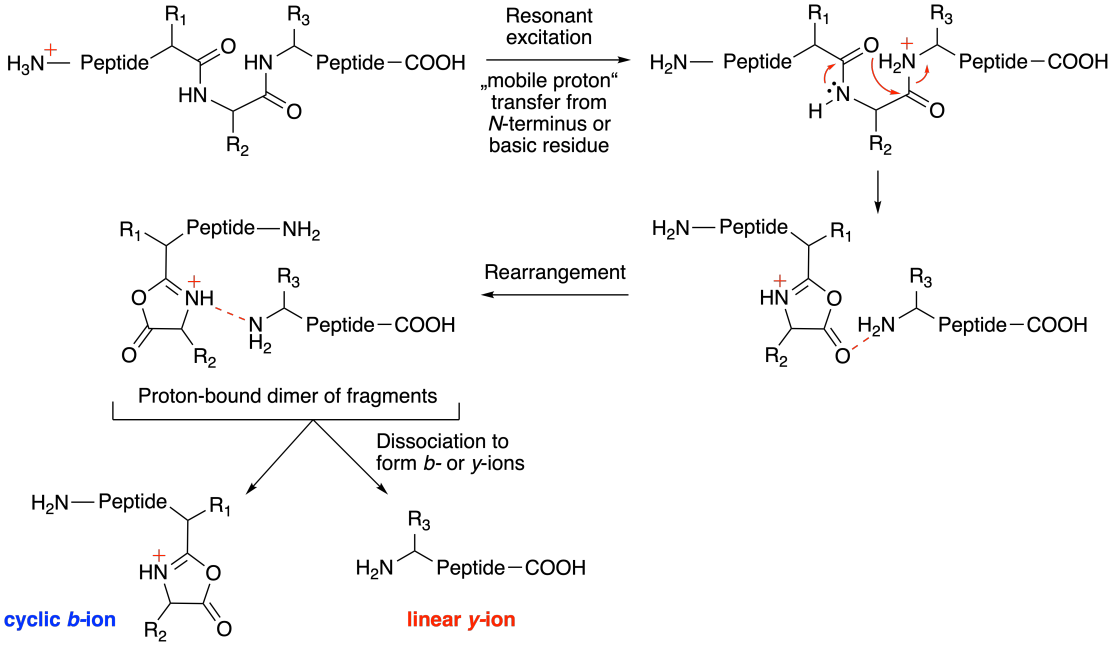


Figure 1.11. CID fragmentation mechanism. Fragmentation scheme of a protonated peptide following Paisz *et al.* via 5-membered oxazolone.[243] Although amide *O*-protonation is themodynamically more favourable than *N*-protonation, and indeed described in chemistry textbooks,[252] it has been suggested that in the context of resonance-activation, protonated peptides reach the energy levels of *N*-protonation (which further weakens the amide bond) as an intermediate before fragmentation.[253,254] The higher-energy, *N*-protonated isomers represent only a small fraction of the ion population, but they are energetically poised for fragmentation.[253]

through loss of CO from *b*-ions.[244] The observation of these ions is dependent of the applied energy.[266]

Ion-trap CID spectra exhibit a good coverage of the peptide sequence in *b*- and *y*-ions, neutral losses are frequent. An issue of CID spectra is the "low mass cutoff",[267] commonly referred to as the "1/3 rule". The mechanism for CID does not allow for trapping of product ions below 28 % of the precursor ion mass due to failed stabilisation of these small ions by the applied RF current. This leads to the loss of immonium ions, y_1 and b_1 ions as well as reporter ions in TMT / iTRAQ.

HCD / beam-type CID Higher energy collisional dissociation (HCD, formerly know as higher energy C-trap dissociation[235]) is a low-energy CID method that differs from the previously described ion-trap CID in its implementation. Nevertheless, the underlying principles remain the same. In HCD, the precursor ion is isolated using an ion trap (Orbitrap Elite) or a quadrupole (Q Exactive) and accelerated into a dedicated collision

cell (quadrupolar or octopolar) that is filled with an inert gas.[235] The acceleration is driven by a DC voltage offset between the ion optics and the collision cell. The impact with the inert gas, typically nitrogen, provides the energy required to render the previously described competing peptide fragmentation pathways accessible. In comparison to ion-trap CID, HCD applies higher energy (20-100 eV) for a shorter amount of time (<1 ms), which implicates that bond cleavage can occur prior to full energy redistribution. Therefore the spectra exhibit different characteristics: typically, comprehensive y-type ion series dominate HCD spectra,[89] large *b*-ions are depleted with increasing collision energy and tend to be less represented.[89,268] Internal fragments and satellite ions are prominent, as well as immonium ions of His, Lys and Tyr due to fragmentation of larger product ions.[235,269] Neutral losses of ammonia have been commonly reported, the neutral loss of phosphoric acids from phosphopeptides is present but slightly less pronounced as in CID.[270,271] The efficiency of HCD on tryptic peptides, independent of charge state, together with the small activation time required have led to HCD becoming the standard fragmentation method for bottom-up proteomics.[272]

Normalised collision energy (NCE) To adjust fragmentation energies at comparable levels for distinct precursor peptides, the normalised collision energy was defined. It relates the absolute collision energy in a linear relationship to the *m/z* of the isolated precursor, as an experimentally accessible proxy for the energy distribution across the degrees of freedom of the analyte.[273] Thermo Scientific implemented the NCE concept to enhance reproducibility of its MS^2 spectra across its range of instruments. The NCE relation to the actual collision energy is define for Q Exactive instruments as:[274]

$$\begin{matrix}\text{Absolute}\\ \text{energy (eV)}\end{matrix} = \begin{matrix}\text{settling}\\ \text{NCE}\end{matrix} \times \frac{\text{Isolation center}}{500\text{m/z}} \times \begin{matrix}\text{charge}\\ \text{factor}\end{matrix} \tag{1.4}$$

The peptide *m/z* is defined as $[\mathrm{M} + n\mathrm{H}]^{n+}$, with the following charge factor conversion: $2^+/0.9$; $3^+/0.85$; $4^+/0.8$; $\geq 5^+/0.75$. Note that the charge factor are lower for glycopeptides than peptides.[275, Fig. 2]

Electron-transfer dissociation (ETD) ETD relies on electron transfer from a ionised carrier molecule to the peptide backbone, which induces fragmentations via a radical mechanism.[276,277] The radical anion, typically fluoranthene,[278] is generated in the ETD source. It is then conducted through ion optics to the ion trap where the radical reacts with the isolated precursor for a set time. This results in cleavage of the $\mathrm{N}-\mathrm{C}_\alpha$ bond in the peptide backbone, generating *c* and *z* ions while acid-labile post-translational

modifications remain intact.[279] Disulfide bonds S−S are also cleaved in ETD to w and u ions.[280]

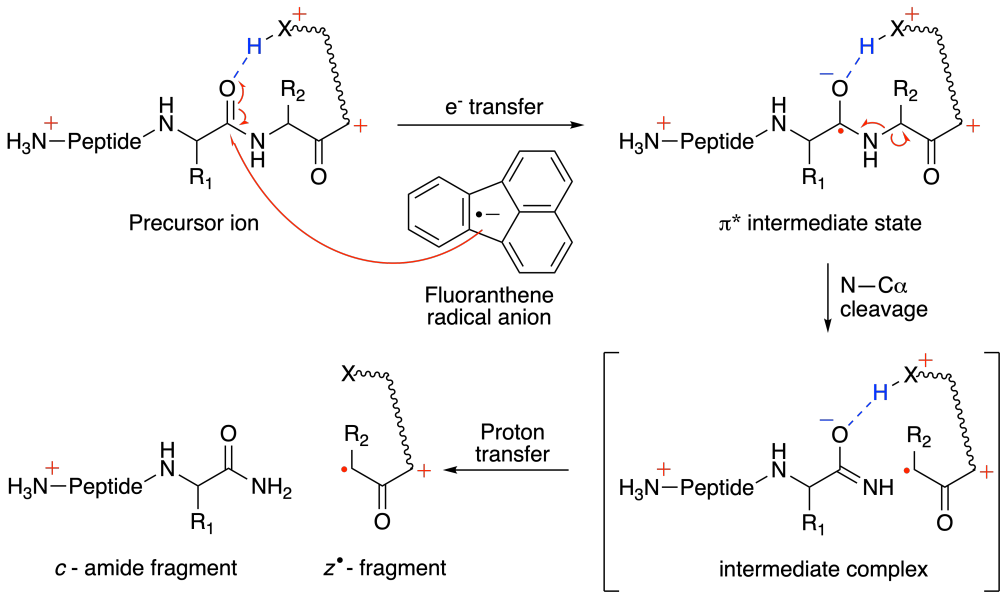


Figure 1.12. ETD fragmentation mechanism. Fragmentation scheme of a multiple protonated peptide by electron-transfer from fluoranthene radical anion leading to formation of c- and z-ions following the Utah branch of the Utah-Washington mechanism.[281,282]

The exact fragmentation mechanism of ETD has been an ongoing debate, with two mechanisms being suggested in the literature: the Cornell[283] and the Utah-Washington mechanism[284,285] – as well as variants of both. Recent evidence obtained by IR analysis of the fragment ions against computational predictions[286,287] and synthetic peptides[288] seems to favour the Utah branch of the Utah-Washington model[281,282], which is depicted in Figure 1.12. This charge-remote mechanism suggests a Marcus-type electron transfer from the low electron-affinity fluoranthene radical anion (possibly via a Rydberg orbital at a charged site[289]) to the peptide LUMO, the π* antibonding orbital of the amide bond, leading to aminoketyl radical formation. Subsequently, this results in cleavage of the N−C_α bond C-terminal to the carbonyl group and proton transfer, yielding c'- and $z^\bullet$-product ions.[281] It has been suggested that the c' and $z^{+\bullet}$ ions are held together in close proximity via non-covalent interactions, as a long-lived intermediate radical species.[290] This can result in hydrogen atom rearrangement, leading to the formation of c-1 (i.e. $c^\bullet = c' - H^\bullet$) and z+1 ions (i.e. $z' = z^\bullet + H^\bullet$), prior to separation of the two fragment ions.[291] The extend of formation of these (potentially unwanted[292]) side products increases with the lifetime of the charge-reduced precursors[293] and was found to be higher for doubly charged precursors than for higher charged ones.[294] This

highlights the importance of precursor charge and reaction time for ETD efficiency: in addition to the aforementioned side products, some precursor ions capture the electron and form a long-lived[295], charge-reduced species $[M + nH]^{(n-1)+\bullet}$ that does not dissociate (an "ETnoD" product).[291] ETD fragmentation efficiency is thus poor for doubly and triply charged peptides, or peptides tending towards high m/z, due to the low charge density.[296] Strategies to overcome these limitations have been to increase the energy of the precursor $[M + nH]^{n+}$ before or during electron transfer reaction[297] as well as additional collisional activation of the charge-reduced precursor $[M+nH]^{(n-1)+\bullet}$ (ETcaD)[293] and calibration routines.[298]

The efficiency of collision-induced methods (CID/HCD) and ETD has been reviewed in numerous reports.[272,296,299] In summary, ETD performance is less efficient and slower, reducing scan speed of the instrument. Therefore ETD is not routinely applied in bottom-up proteomics. However, in cases of higher charged peptides[300,301], labile post-translational modifications[302,303] or intact proteins[304] ETD has clear advantages. As such, complementary use has been recommended leading to the development of data-dependent decision-trees[272] and product-dependent triggering[305] as a pragmatic implementation.

Combination of fragmentation events – EThcD Modern mass spectrometers allow the triggering of subsequent fragmentation events on the same ion population, a useful feature if the peptide sequence coverage is limited. Methods that combine CID/HCD with ETD fragmentation have been developed to overcome limitations of the latter (2^+ charged peptides etc.) and termed ETciD[291] and EThcD.[306] Peptides are fragmented with ETD first and highly abundant fragments further subjected to CID or HCD fragmentation, generating a comprehensive sequence coverage of the peptide by yielding four distinct ion series. The identification of precursors with low charge-density is thus improved, which is in turn beneficial for the correct site-localisation of PTMs.[307] However, combining scan types comes at the cost of decreased scan speed, rendering these approaches impractical for the analysis of complex samples.[89]

Fragmentation of glycopeptides – key points The fragmentation of glycopeptides is thoroughly reviewed in chapter 3, as such only key points are highlighted here. The glycan part of glycopeptides predominantly fragments in CID/HCD, leading to lower peptide backbone fragmentation. However, glycans rarely fragment in ETD whereas the peptide backbone does. This observation has been coined the *complementarity*

paradigm of glycopeptide fragmentation.[114,115] Fragmentation methods are often combined in glycoproteomics, as will be detailed in chapter 3. Glycans fragment down to diagnostic low mass *oxonium ions*, characteristic for each monosaccharide. These can be used to inform about the presence of glycans,[308] triggering events for product-dependent data acquisition strategies[305] and more.

1.3.4.4. Data acquisition methods

Instrumentation Figure 1.13 shows a schematic representation of the instruments used for tandem mass spectrometry in this thesis. The operation of scans is defined for each instrument.

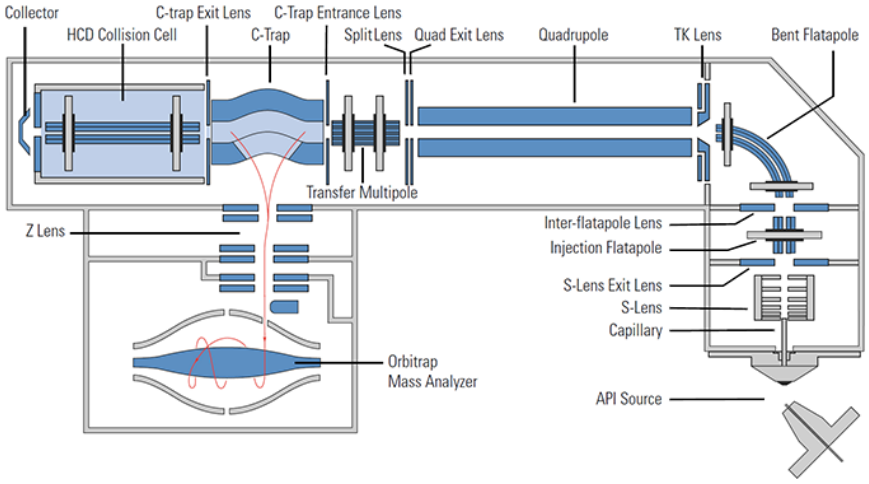
The Q Exactive is one of the workhorses of proteomics.[309] In MS^1 scans, ions are acquired by passing all ions through the ion optics into the orbitrap mass analyser. MS^2 scans are performed using HCD fragmentation. The precursor ion is selected by filtering the entering ion population through the quadrupole, fragmented in the HCD collision cell and analysed in the orbitrap.

The Orbitrap Elite is a very versatile instrument that allows the definition of complex scan modes.[310] Particularly, two mass analysers are used: the upfront linear ion trap (LIT) and the orbitrap. In a typical MS^1 scan, ions are guided to the orbitrap, taking advantage of the high resolution and mass accuracy. The operation of the MS^2 scans depends on the fragmentation method: CID and ETD fragmentations are performed in the LIT, and the spectra typically acquired there, making use of the high speed of the analyser. HCD scans are performed by fragmentation in the HCD collision cell and spectrum acquisition in the orbitrap. The Orbitrap Elite allows to change these combinations of fragmentations and data acquisition, described in the following exemplarily on the EThcD fragmentation.

Implementation of mixed fragmentation schemes – EThcD on the Orbitrap Elite EThcD is not implemented by default on the Orbitrap Elite – as opposed to the sucessors such as the Lumos. Nevertheless, scans can be defined by the user using the modular build of the instrument.[§] The process of EThcD in the Elite is described in Figure 1.13. After ❶ precursor isolation upon MS^1 in the High Pressure Cell (part of the ion trap) and ❷ anion formation in the ETD source, ETD fragmentation ❸ is conducted by

[§]Technical note: in this setup on the Elite, EThcD is hard-coded as a HCD scan with first mass 151. Only with this setting is a MS^3 step in the HCD cell without precursor isolation defined. Please refer to section 3.4.2.2 for more details.

a.



b.

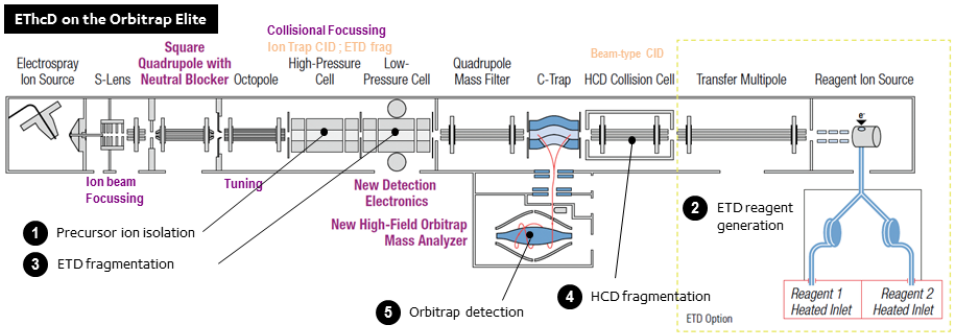


Figure 1.13. Ion routing scheme in a Q Exactive and Orbitrap Elite. a. The Q Exactive mass spectrometer combines a quadrupole mass filter with an HCD collision cell and on Orbitrap mass analyser.[309] **b.** The Orbitrap Elite combines a linear ion trap, an octopole mass filter, an HCD collision cell and an Orbitrap mass analyser. The ETD source generates the fluranthene anion reagent. The numbers indicate the steps in the process of EThcD described in the text.[310] Images modified from ©Thermo Fisher Scientific.

electron transfer from the fluoranthene anion in the Low Pressure Cell (part of the ion trap). Fragment ions are transferred – without detection – to the HCD collision cell ❹ for further fragmentation before ❺ transfer to the Orbitrap for fourier-transformed detection.[306]

Data dependent acquisition (DDA) DDA is the classical operation mode for proteomics. The top N most abundant peptide ion species in the MS^1 spectrum are individually isolated and fragmented for MS^2.[311] To increase the dynamic range, precursors selected for MS^2 are excluded from repeated fragmentation for a set period preventing the instrument from repeatedly fragmenting specific peptides (dynamic exclusion). Important

parameters considered for DDA are scan speed, sensitivity, time for collection of the ion species ("ion injection time") and the total number of ions collected.[312] DDA doesn't require prior knowledge of the sample used and is the most common method of operation in proteomics. However, two main disadvantages are to be disclaimed: poor longitudinal reproducibility due to stochastic picking and a bias towards identification of peptides with high abundancy. Improvements in DDA techniques to identify low abundant peptides − that are often biologically interesting − have focussed on using segmented MS^1 windows[313,314] or *on-the-fly* analysis.[315–317]

All data acquired in this thesis is DDA, as such the two other main techniques for data acquision are not further introduced, namely targeted data acquisition (SRM,[318] MRM,[83] PRM[319] etc.) and data-independent acquisition (DIA[320], SWATH[321], GlycoDIA[322]).

Optimisation of data acquisition The need for optimisation of the acquisition method for sample and experiment types has been highlighted numerous times. In quantitative proteomics, the use of MS^3 strategies has been highlighted to improve quantitative accuracy in TMT experiments.[323] Similarly, crosslinking MS benefits from ignoring lower charge states ($< 3^+$) (see chapter 5) and MS^3 strategies have also been developed for MS-cleavable crosslinkers.[324]

New emerging techniques for glycopeptides: ion mobility, cold-ion IR The potential for ion mobility to discriminate isobaric and epimeric glycans at MS^1 level[325] and MS^2 level[326] and glycopeptides[325] has illustrated its potential as complementary analytical technique for glycoproteomics. Ion mobility separates ions by their size, shape and charge, resulting in different drift times for glycans/glycopeptides which is directly related to the rotationally averaged collision cross section (CCS).[327] The implementation and use of ion mobility in combination with other fragmentation techniques has been explored,[328] this would allow complementary information such as distinguishing sialic acid linkage,[329] glycosidic bond stereochemistry,[326] and glycan stereoisomers[325] − which could potentially be automatised with the help of specialised databases (GlycoMob[330]).

Tunable benchtop lasers for infrared multiple photon dissociation (IRMPD) and related gas-phase techniques have empowered the growing area of gas-phase IR spectroscopy-MS of carbohydrates.[331] In principle, ion clouds of a selected m/z are held in a mass analyser and irradiated with an IR laser.[331] The light absorption is indirectly measured (via fragmentation yield → action spectroscopy) and a gas-phase vibrational spectrum

constructed.[332] The comparison of the experimental spectrum with the theoretically predicted (by DFT) allows differentiation of isomeric glycans. Cold-ion (also: cryogenic) gas-phase IR spectra have allowed to improve IR resolution, allowing identification of analytes differing in a single stereocenter.[333] Although these applications not yet translated into commercial mass spectrometers, their application appears promising.[333]

The recent description of MS-separation, soft landing and subsequent imaging of synthetic glycans is another exciting future perspective for glycopeptides.[334]

1.3.5. Data analysis: peptide identification and quantitation

Post acquisition, the generated data has to be analysed with the aim of generating a list of confident peptides and by inference protein identifications. The used computational analysis workflow depends on the available resources, broadly speaking strategies can be classified into database searching,[109,335] *de novo* sequencing[336,337] and spectral library-based approaches.[321,338] The latter was not used in this thesis and will not be further discussed.

1.3.5.1. Preprocessing

Generated raw files containing the spectra are stored scan-by-scan as peak list, that is paired *m/z* and intensity information. Additionally, metadata including acquisition and machine parameters are included. Typically, spectra are pre-processed before submitting to database search or *de novo* sequencing to remove impurities and reduce complexity.[339,340] These steps can include noise removal (baseline correction), normalisation, peak picking, converting from profile to centroid peaks, deconvolution of charge states and removal of isotopic peaks. The resulting spectra exhibit a higher signal-to-noise level and are less complex, which allows more precise database searching and facilitates *de novo* sequencing.[341]

Charge deconvolution is particularly important for glycopeptides, that typically exhibit a different, wider isotopic pattern as unmodified peptides due to their increased carbon content − for larger glycopeptides with substantial glycan part, the second peak is often the highest. For deconvolution, the *m/z* spacing of observed isotopes is used to calculate the charge state of the fragment (for ^{13}C: $\frac{1\text{Da}}{\text{charge}}$). Fragment ion intensities of different charge states are accumulated at the nominal mass or the lowest charge state, depending on algorithm implementation. In glycoproteomics, this is handled differently across software: Byonic[342] and pGlyco2[343] do not deconvolute MS^2 spectra while others such as GlycReSoft do.[344]

1.3.5.2. Database searching

DDA acquisition combined with a database search is currently the standard workflow in bottom-up proteomics. At its core, an *in silico* digest of a protein sequence or protein sequence database is performed, which is matched against an experimental spectra under consideration of instrument type, fragmentation method and mass accuracy among

other: a peptide spectrum match (PSM). Depending on the quality of the match, a score is given to each candidate peptide sequence. The peptide sequence with best match to the experimental spectrum (highest score) is termed rank 1 PSM and reported. The scoring distribution of peptides is further analysed by statistical means to help distinguish between true and false identifications (details see Figure 1.14). The entries in the protein database used for searching of the data will determine which peptides and by parsimony which proteins (or gene-based protein groups) are identified.[345]

Several search engines such as Mascot,[335] Sequest,[109] Andromeda[346] (integrated in MaxQuant[347]) or MSFragger[348] among others exist, their scoring model varies but the outlined fundamental features are shared. In the following, common features of database search engines and new developments are highlighted.

***In silico* spectrum prediction** Fundamentally, algorithms for *in silico* prediction have remained similar to the initially published version in the 1990s[109] to the late 2010s. Prediction of fragment ions have no particular information in intensity except for a few empirical rules (i.e. proline effect). Recent years have seen the emerging of deep learning technologies such as Prosit[349] or DeepMass:Prism[350] that have been trained on large spectral libraries such as the ProteomeTools[197] or the PRIDE repository, respectively, to accurately predict fragment ion intensity and retention times with very high overlap with experimental data. The application of these machine learning-based prediction tools to generate predicted spectral libraries is poised to have a strong impact of the quality of peptide identification, PSM scoring and have so far been implemented in Proteome Discoverer, Mascot, Skyline, Spectronaut for Prosit and MaxQuant for DeepMass:Prism. Note that although an extended version includes the prediction of some PTMs[351], the prediction of glycosylated peptides by deep learning is not yet possible.

Search parameters & search space Multiple parameters are specified for database searches: digestion enzyme specificity, allowed missed cleavages, fixed and variable amino acid modifications, fragmentation type and error tolerance among others. These parameters directly affect the search space, that is the combinatorial possibilities of theoretical spectra against an experimental spectra that can be matched. As an example, variable modifications such as oxidation are also used to search for post-translational modifications, they are mass increments that can occur, but don't necessarily have to (as opposed to fixed modifications such as cysteine alkylation / carbamidomethylation). Each additional modification vastly increases the search space as permutated versions of each peptide must be generated. This effect is emphasised the larger the protein database used.[352]

The increase in search space is of particular importance in glycoproteomics projects. The high number of possible glycosylations (glycoforms) multiplied by the number of possible digested peptides leads to a *combinatorial explosion in glycopeptide search space.*[342] Whereas *N*-glycosylation has a conserved sequon, the vast combinatorial possibilities render the identification of *O*-glycopeptides very difficult. Specialised search engines such as Byonic have implemented *fine modification control* features allowing to limit searches to specified protein entries or the creation of focussed databases.[342]

Open searches Database searches mandate the existence of an estimate of the presumed sample content, such as the availability of genomic or reference proteome data. For non-sequenced organisms, organisms with partial proteome information or samples from a mixture of organisms such as found in gut samples, very large database are employed which interferes with the statistic error control. In addition, proteins contained in a sample may have unexpected modifications, sequence mutations and differentially expressed isoforms or truncations. Open searches were introduced to overcome these limitations. This approach uses high mass tolerance in MS^1 to identify peptides not present in the database or peptides carrying modifications while requiring high mass accuracy of fragment ions to identify the amino acid sequence.[348,353] Open search approaches are commonly used in crosslinking mass spectrometry (see 1.3.5.5).

Probabilistic scoring and error control Database searches match a large number of spectra to candidate peptide sequences, resulting in PSM of variable quality and confidence. Erroneous matching can be a result of incomplete fragmentation, low signal-to-noise in the spectrum, unknown composition, restrictive search space or by random chance. This leads to false-positive error, i.e. when an incorrect sequence is matched to a spectrum, and false-negative errors, through absence of the peptide sequence from the search space. Particularly false positive PSM present a challenge in bottom-up proteomics as their false identification interferes with the interpretation of the dataset.[354,355] Hence, mechanisms have been created to estimate and control the false discovery rate (FDR), commonly defined in proteomics as the percentage of false identifications of the total identifications, often arbitrary set at 1 %.[355] There are several methods to relate the FDR to the scoring method of the scoring model of the search algorithm, in which typically a -10logP score is used, i.e. the negative logarithm of the determined probability that the computed PSM is an incorrect assignment.[346] One of these FDR estimation methods is the target-decoy approach.

Target-decoy approach As it is not possible to know with certainty which PSM are false, one option is to artificially generate wrong PSM. The target decoy approach[356] is used to extend the target sequence database with a database of the same size composed of shuffled or reversed protein sequences (decoys). Matches to the decoy database and false matches are assumed to follow a similar distribution. In order to be able to estimate the score cutoff for a desired FDR rate, the global distribution of all PSM scores for both target and decoy database is compiled (Figure 1.14). The threshold for the PSM score is defined such that a user-defined number of decoy-identifications is retained (typically 1 %). This instrument allows to globally control the number of false-positive identifications and can generally be used to assess the quality of the data.[356] However, in many cases the user is interested in a specific protein or complex, it then becomes relevant to know the probability of the identification to be wrong. The target-decoy approach allows to calculate the posterior error probability (PEP, or 'local FDR') by dividing, at the given score of the identification in question, the number of incorrect PSM by the numbers of all PSM.[357] Recently, FDR methods for glycopeptides based on the target-decoy method have been implemented.[343,358]

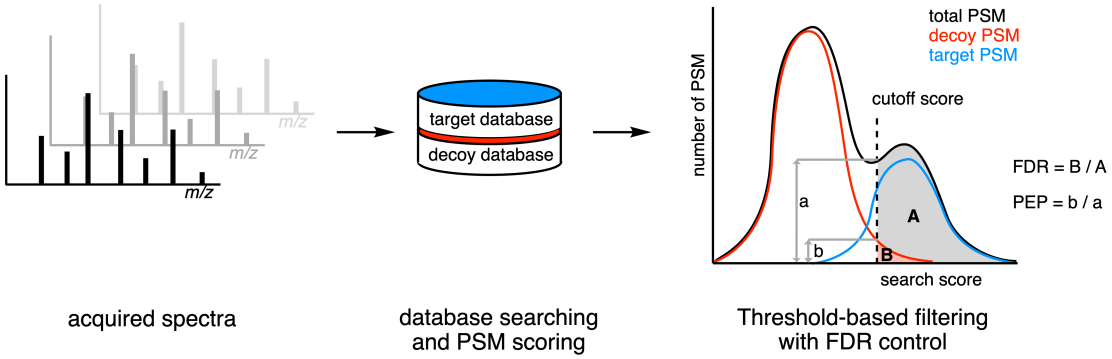


Figure 1.14. The target-decoy approach. Spectra are scored against a concatenated database containing the target species database and a database of reverse or shuffled sequences, the decoys. The score cutoff for retained identifications is chosen globally in relation to the desired global FDR. Only a given percentage of wrong (decoy) matches (the FDR) is retained (typically 1 %). The score cutoff is set at a value so that the ratio of the area under the curve B / A equals the desired global FDR. The PEP or local FDR (at any given score) is defined by b / a, which are the number of decoy/total PSM at that particular score.[357]

Localisation of a modification When analysing site-specific modifications such as post-translational modifications, the information of the exact amino acid carrying the modification is central for the biological question. False localisation (FL ; FLR, false localisation rate) of a modification on a peptide is another source of error in database searches.[359] If

only one amino acid in the peptide sequence can be modified, the MS^1 and parts of the fragment ion series will be sufficient for identification. *N*-glycoproteomics as an example requires the (rare) N-X-S/T sequon, thus the glycan mass can be derived from the MS^1 and the localisation is evident. However, peptides with multiple acceptor sites may result in ambiguous identifications. For example, in *O*-glycoproteomics, modifications often occur in Ser/Thr rich sequences, yet the exact site of modification can have a strong effect on the downstream biological signalling. In the best case scenario, fragment ions flanking the modified amino acid, so-called *site-determining ions*,[360] pinpoint the exact site of modification. In practice, they might be low abundant or absent. Search engines and post-processing tools employ additional validation and probability tools to address this issue of assigning the correct positional isomer. This can be a metric assessing the likelihood that a site-determining peak is random noise, such a probability-based localisers such as A-Score,[360] ptmRS[361] or PTM Score[362] in Andromeda/MaxQuant. Alternatively, a calculated search engine delta score of the PSMs for different positional isomers is also encountered, such as Mascot delta score[363] or Byonic delta score.[342] Reported PSM can be further classified or filtered by localisation likelihood[362] or score cuts.[364] In the case of complex samples, all these strategies are limited to providing means the filter the dataset,[359,365] assessing the true FLR rate in large samples without knowing the sample composition is difficult.[89]

1.3.5.3. De-novo sequencing

De-novo sequencing attempts to obtain sequence information directly from the ion series in the MS^2 spectrum and the precursor mass. By identifying the ion series type and comparing mass shifts between peaks with catalogued mass tables, the sequence can be iteratively build up. Manual *de novo* sequencing is time consuming, automated peptide sequencing software tools have been written to automatise the process: Novor,[366] PepNovo+,[367] pNovo 3[368] and PEAKS.[369] These tools bring pre-processing functionalities and local confidence scoring models that allow to assess different suggested sequencing results. Nevertheless, incomplete fragmentations, complex and noisy MS^2 render the process error-prone, and requires stringent − often manual − error control. Recently, deep learning[370] as well as post-processing FDR error control[371] have made their apparition in the field (These tools have not yet been combined with extensive spectral libraries except for Novor[366]). The assembly of protein sequences is a separate process, generally speaking a De Bruijn assembler is used that attempts to assemble the protein sequence, or protein contigs.[372] *De novo* sequencing methods, particularly of monoclonal antibodies, are further highlighted in chapter 2.

1.3.5.4. Quantitative proteomics

Mass spectrometry of peptides is inherently not quantitative in absolute terms, a result from the different physico-chemical properties of peptides affecting ionisation efficiency and detection.[373,374] Upon assignment of confident peptide identifications, quantitative proteomics aims at calculating changes in peptide/protein levels across samples.[375] Several strategies have been developed to achieve this goal (Figure 1.15), the underlying assumption is that the signal amount obtained by mass spectrometry is proportional to the abundance of the analyte, enabling relative quantitation of many proteins in parallel.[376] In this thesis, only label-free quantitation (LFQ) is used (see chapter 4), therefore this section introduces general possibilities for quantitative approaches using MS data and focusses on LFQ.

Historically, spectral counting has been used in DDA approaches (Figure 1.15.A). It relies on the assumption that higher abundant proteins will generate more abundant peptides, which will be picked more often for MS^2 fragmentation yielding more spectra. However, modern mass spectrometer operate with dynamic exclusion, a data acquisition strategy that avoids re-sequencing the same peptide to achieve deeper sampling.[376] Today, quantitation in DDA approaches is more frequently performed by calculating the area under the curve through summing of the peak intensity over time (extracted ion chromatogram, XIC) for features of interest (Figure 1.15.B). This can be performed for precursor ions for a given peptide at the MS^1 level, or by summing up fragment ion intensities at MS^2 level (i.e. in targeted proteomics) or by integrating both MS^1 and MS^2 (often in DIA).[376]

Proteomics workflow include multiple handling steps that can potentially introduce a bias into the data. The labelling of samples before combining these early in the workflow allow to minimise selective losses as these are processed and analysed together (Figure 1.15.D). Metabolic labelling strategies with stable isotope labelling by amino acids in cell culture (SILAC) or by ^{15}N media allows the earliest combining, its readout is typically at MS^1 level. Digested peptides can be labelled chemically by dimethyl labelling (also MS^1 readout) or using isobaric reagents such as iTRAQ or TMT,[377] before being combined. In this case, the peptides from different sample conditions, all carrying the isotopomeric tag, co-elute and are fragmented together. Upon fragmentation, of each isobaric tag releases a unique reporter ion in the low m/z range that allows to distinguish the sample. The intensity ratio of the reporter ions in the MS^2 can be used for quantitation (Figure 1.15.C). In the cases of overlapping precursor ion series in the MS^1, co-isolation of another precursor peptide can lead to a distortion of the ratio leading to *ratio compression*.[323] Tailored MS^3 strategies have been developed to counteract this effect.[323,378]

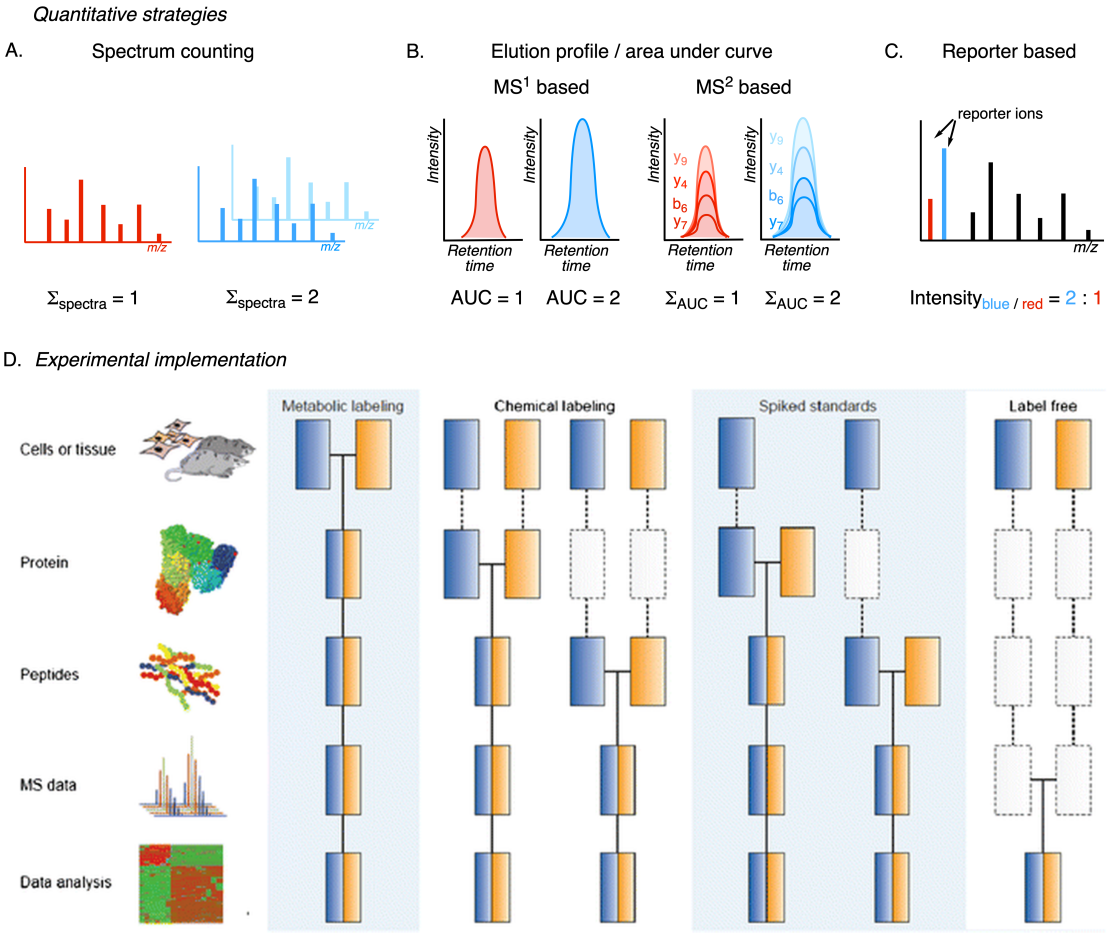


Figure 1.15. Quantification methods used in proteomics. A. Spectral counting sums the number of acquired spectra as a proxy for abundance for a given protein. Blue and Red represent two experimental conditions. **B.** Extracted ion chromatograms (XIC) allow integration of a peak of a given m/z over its elution time. The area under the curve of the elution profile allows intensity-based quantitation of the precursor ions (MS^1 based) or fragment ions (MS^2 based). **C.** Reporter ion-based quantification methods such as TMT rely on the co-isolation of precursors carrying distinct isobaric tags, which generate diagnostic reporter ion for each of the pooled samples after fragmentation and can therefore be used for accurate relative quantification. **D.** Experimental workflows implementing the different quantitative strategies. Blue and orange represent two experimental conditions. Horizontal lines indicate pooling of the samples together, dashed lines highlights points at which experimental variation and thus quantitation errors can occur. Image adapted from[89] (A-C); reprinted from[374] (D) with permission from ©Springer Nature.

Quantitative information extracted from bottom-up proteomics is on the peptide level and needs to be rolled up to the protein level. Building the sum if all peptide intensities, or a sub-selection thereof is one possible option. Algorithms have been implemented in several software suites for quantitative proteomics such as MaxQuant,[379] PEAKS Q, Progenesis QIP among others.

Label-free quantitation (LFQ) Label-free quantitation methods combine samples at the data analysis stage, they have therefore the advantage of not needing labelling. Two main methods have emerged for LFQ, spectral counting (SC) and extracted ion chromatogram (XIC).[376] As described above, spectral counting has many drawbacks but can provide a quantitative readout if applied correctly.[380] Normalised spectral abundance factor (NSAF) and variants thereof have been developed to account for protein size among other factors.[376]

Most label-free methods are intensity-based. Conceptually, in label-free experiments the detected peptide peaks (XIC) of two or more LC-MS based runs are compared, aligned and the area under the curve of the peptide and its isotopic peaks (peptide feature area) integrated. Several normalisation and processing steps are performed by the LFQ processing algorithms, such as MaxLFQ[381]: peak identification, noise reduction, peptide retention time alignment ('delayed normalisation'), peak intensity calculation, match between runs, intensity normalisation based on total ion count before the maximal peptide ratio extraction. The protein intensity profile is then constructed from its peptide signals. If no peptide ratio can be found in a given sample, because no or only one peptide is identified in the sample, the algorithm such as MaxLFQ cannot produce an LFQ intensity for the protein.[381] This leads to the issue of missing values, which can be overcome at the data analysis stage, by imputation or by replacing missing values with a constant.[382]

As such, LFQ methods are highly dependent on the quality and reproducibility of the nLC separation and the acquisition technique: the number of high-resolution (and thus time-consuming) MS^1 scans strongly affects the quality of the quantitation, while MS^2 scans are necessary for peptide identification. New generation of instruments such as the Tribrid-type instruments allow parallel acquisition of MS^1 in the orbitrap while collecting MS^2 spectra in the quicker ion trap. A very interesting study by Hoehenwarter and colleagues has assessed the performance of several leading LFQ software techniques, with MaxLFQ and Progenesis QIP ranking among the best available, while at the same time describing pitfalls of the methods.[383]

Whereas LFQ experiments typically yield *relative* changes of protein abundance amongst samples, it is also possible to obtain *absolute* values for protein concentration in the cell using the 'proteomics ruler'.[384] The MS signal of histone proteins is proportional to the amount of DNA in the sample which in turn is dictated by the number of cells.

1.3.5.5. Crosslinking tandem mass spectrometry

Crosslinking (xLink) converts non-covalent interactions between proteins, into covalent bonds. The resulting crosslinks formed are a direct result of the 3D geometry of the interaction and provides data on the native state of proteins, with a bottom-up readout.[385] This section provides the technical background to chapter 5 and as such focuses primarily on the data analysis of crosslinked samples, additional aspects for sample preparation and data acquisition have been included for readability.

Crosslinking reagents and sample preparation The crosslinking reagent reacts with two amino acids from the protein complex, provided both are within distance on either side of the crosslinker for the reaction to happen. Although formaldehyde is still used, most crosslinking mass spectrometry applications use bifunctional linkers such as DSS and BS3 or their MS-cleavable analogues DSSO and DSBU.[388] Heavy (deuterated) and light variants of these crosslinkers have been designed to enable quantitative crosslinking mass spectrometry.[388,389] Similarly, heterobifunctional crosslinkers are commercially available, such as variants ending with a UV-photoactivable moiety for photo-induced proximity labelling.[390]

Customised linkers have been designed and synthesised for distinct applications. Examples includes linker with functional groups enabling enrichment of crosslinked peptides, such as a phosphate group for IMAC enrichment (PhoX[391]), copper-catalysed[392] or strain-promoted[393] azide-alkyne cycloaddition. Interesting for the glycosciences is the TRICEPS trifunctional linker that allows crosslinking protein with oxidised glycans.[394] This thesis uses a synthetic protein as a crosslinking agent (see chapter 5).

Following on the observation that crosslinks in complex samples are mainly observed between abundant proteins (top 100 - 500),[395] the kinetics of the formation of crosslinking have been studied. Increasing the concentration of crosslinker was shown to allow identification of low-abundant proteins, in a pseudo first order kinetics model.[396]

From the point of view of sample preparation for mass spectrometry, XL-MS has benefited from improved digestion protocols (elastase/trypsin[124]) as well as SCX enrichment for crosslinked peptides. Their development goes hand in hand, as SCX is most effective when protonation sites are available at both the *N*- and *C*-termini, which can be achieved by use of successive enzymatic digests.[397]

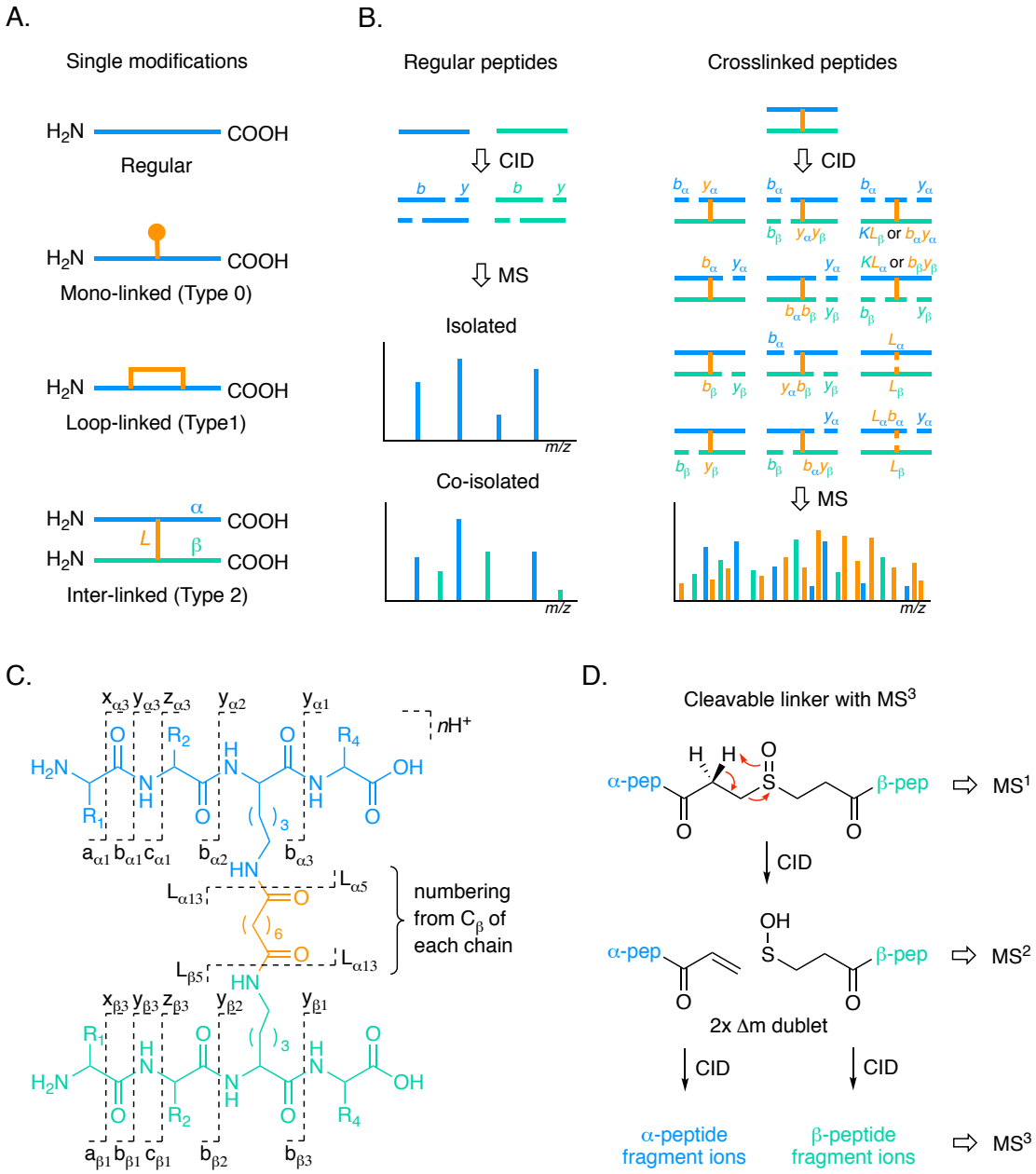


Figure 1.16. Fragment ions in crosslinking mass spectrometry. A. Crosslinking products. Digestion of chemically crosslinked proteins generates mono-, loop- and interlinked peptides besides regular peptides unmodified by the linker.[386, S1] Although rare, multiple modifications are also possible. These are named after the modifications occurring, i.e. a inter-liked peptide with a single additional mono modification would be Type 2,0.[387] **B.** Crosslinked peptides have more complex fragmentation patterns than regular peptides, leading to more complex spectra. Fragment ion nomenclature follows Schilling *et al.*[387] **C.** Molecular detail of the fragmentation nomenclature. Note that peptide strands are numbered by residues, linkers by atoms.[387] **D.** The MS^3 strategy commonly used with cleavable linkers. The sulfoxide group fragments at lower CID/HCD energy than the amide bond. In the MS^3 strategy, this is used to first separate the 2 strands by their own mass, and sequence these individually, rendering the MS^3 spectrum easier to interpret.[385]

Fragmentation techniques Due to their generally low abundance, crosslinked peptides have been difficult to detect among the many linear peptides, leading to poor signal patterns. As a consequence, elution peaks contain few data points, isotope envelopes are incomplete and signal to noise is low.[398] Typically, collision-induced methods are used for fragmentation, with both MS^1 and MS^2 recorded in the orbitrap.[399] However, HCD or CID are fragmenting preferably one strand leading to the formation of product ions from majorly one of the two constituent peptides.[400] As such, alterning CID-ETD fragmentations were used to improve fragmentation and confidence in peptide identification.[401] The efficiency of fragmentation techniques was subsequently assessed, HCD was found to be the method of choice for most cases due to fastest acquisition and high efficiency. Nevertheless, EThcD gives the best sequence coverage for peptides with high charge density, confirming DDDT methods for high-complexity samples or enriched crosslinked peptides can improve performance.[402] Combining fragmentation methods can improve cross-link peptide identifications.[403]

An alternative approach to bypass difficulties in identification has been the use of cleavable crosslinkers together with the use of tailored MS^2-MS^3 acquisition strategies (Figure 1.16.D).[385, Fig.4] After full MS^1 spectrum, a precursor is selected for MS^2. Several groups such as sulfoxides in crosslinkers fragment in CID/HCD at a distinct, lower energy than the amide bond. This results in a MS^2 spectrum that contains the reporter ion and the precursors for each branch of the crosslinked peptide. Each branch is then selected for an additional HCD fragmentation, the MS^3 spectra gives the sequence of each peptide.[404] A stepped-energy fragmentation approach has been suggested to improve the spectra.[405] At the stage of data analysis, cleavable crosslinker require far less computation time than non-cleavable.[395]

Data analysis Data analysis in crosslinking tandem mass spectrometry is rendered difficult by three key obstacle well-known to the literature: the increased complexity of fragmentations, the n-square problem and the statistic evaluation of credibility/false discovery.

Crosslinked peptides can fragment from either end of both strands connected at the branching point, leading to 4 directions of fragmentations with respective b^{1+}, b^{2+} and y^{1+}, y^{2+} ions with their neutral losses and the less common y^{3+}, a^{1+}, a^{2+} and rendering the spectra very complex (Figure 1.16).[387] Additionally yb^{1+} and ya^{1+} cross-chain fragments (including the linker) are commonly observed (but not necessarily recognised by the software). In the case of lysine-specific crosslinkers (BS3 etc.), additional K-type fragment that correspond to the single α or β chain + linker + iminium ion of K are

observed, named $K^L\alpha/K^L\beta$. Cleavable linkers produce additional L-type fragments $L\alpha$ and $L\beta$.[386, SI p.62f]

The second problem refers to the quadratic expansion of search space as each of the n proteins in the database can potentially crosslink with n other proteins as well as itself, leading to $(n^2 + n)/2$ possibilities that have to be assessed in exhaustive search strategies.[399] As such, including post-translational modifications or additional missed cleavages has a more dramatic effect on the expansion of the search space than in regular bottom-up proteomics database searches. Crosslinking engines designed for proteome-scale identifications, that is most modern engines such as pLink 2[406] (used in this thesis), Xi[407] or Kojak[408], solve this problem using open search approaches (1.3.5.2) which consider one cross-linked peptide pair as two single peptides, each bearing a modification of large mass yet unknown composition. Peptide pairs are the associated using their precursor ion masses.

The third problem of statistical validation comes from the need to avoid false-positive peptide identifications. As an example, non-covalent adducts of co-eluting peptides can be mis-identified.[409] Commonly in bottom-up proteomics, this is dealt with by incorporating a global false-discovery rate (FDR) which is related to peptide-spectrum match (PSM) score (see Figure 1.17). Defining a FDR for crosslinked peptides is however much more challenging then for linear peptides.[410] In XL-MS, three additional level exist: PSMs aggregate to peptide pairs, which then aggregate to linked residue pairs and ultimately in turn in protein pairs.[410] Peptide identification is further hampered in that the number of peptides that represent a particular interaction are non-stoichiometric, rendering quantifications based on spectral counts or AUC's particularly vulnerable. Fortunately, the relationship of local FDR rates and their plausible settings can be benchmarked using crosslinked dataset on protein complexes where the structure has been solved, by analysing PSM, peptide- or residue pairs on the basis of distance restraints.[401,410] pLink 2,[406] XiFDR,[410] XlinkX[401] and others make the choice of using a target-decoy approach (TDA) adapted for crosslinking datasets to dertermine the global FDR. The global FDR is calculated by database construction on peptide pair level: target-target, target-decoy and decoy-decoy, the latter two serving as mis-identification for FDR calculation.[411] FDR calculations by the different search engines were recently benchmarked against a synthetic peptide library.[412]

Quantitative protocols for crosslinking mass spectrometry[395,413] including LFQ[414] − not used in this thesis − follow similar principles as highlighted in section 1.3.5.4 and are excellently reviewed by Chen and Rappsilber.[398]

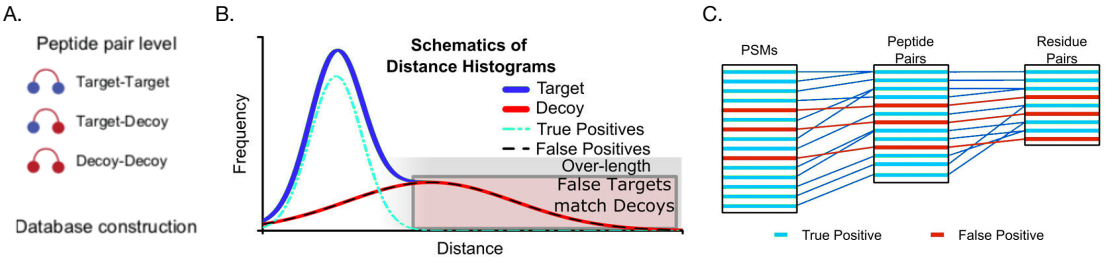


Figure 1.17. False discovery rate in crosslinking mass spectrometry. A. The target-decoy approach adapted to crosslinking mass spectrometry. For peptide pairs, the decoys can consist of target-decoy and decoy-decoy pairs. **B.** Schematic distance-histogram showing the use of distance metrics as a validation tool for FDR. The graph illustrates the overlap of false positive and decoys and resulting overlap of overlength crosslinks with decoy crosslinks. **C.** FDR error propagation in crosslinking mass spectrometry, in form of wrong identifications, from PSMs to peptide pairs to residue pairs. True positives tend to cluster, for example several correctly identified PSMs can support the same unique peptide pair. Similarly, several correctly identified peptide pairs can in turn support one residue pair. This is not the case for false identifications, which are random and do not cluster to the same extend. Modified from[410] with permission from ©ACS.

Data visualisation and validation The use of crosslinking mass spectrometry in structural proteomics research projects is described in section 1.3.6.6, here only necessary data visualisation for validation of the results is highlighted. Most software for crosslink analysis such as Xi (i.e. xiVIEW), pLink 2, XlinkX 2 (a node in PD), MeroX and Kojak facilitate the manual validation and investigation of the crosslinked spectrum.

Validation of the data beyond FDR often takes known structures into account, and plotting the distance distribution for the identified crosslinks and setting an upper limit.[399, Fig.2C] Although possible with many computational chemistry software (such as PyMol etc.), several specialised software applications facilitate this analysis, such as xiVIEW or Xlink analyser (in UCSF Chimera). In addition, inclusion of protein dynamics has been shown to greatly improve the interpretation of identified crosslinks, using tools such a DynamXL.[415]

In addition to annotated spectra, 3D structure view and distograms, visualisation solutions include network and residue-level views of cross-linked peptides.[385, Fig.3] Further downstream analysis and post-processing, such as PPI networks, is described in section 1.3.6.6.

1.3.6. Proteomics applications and their respective post-processing methods

At its core, proteomics workflows up to database searches yield a list of proteins and their intensities. Whereas the data acquisition follows fundamentally common principles (and same instrumentation), different fields and applications in proteomic research differ vastly in their sample preparation method and their post-processing, that is the subsequent analysis of the generated proteomics data. Post-processing is inherently project-dependent, multiple possibilities of varying bioinformatic complexity exist that can combine additional experimental information or archived data from databases, up to 'multi-omics' approaches. Ultimately, the aim is to translate the obtained extensive, high quality proteomics data into biological, biochemical or chemical knowledge.[416,417]

The aim of the following section is to provide a very brief summary, common outcomes, opinions and practices sorted by different sub-fields of proteomics relevant to this thesis. The following discussion only applies to a setting of 'discovery proteomics', not to clinical proteomics or targeted proteomics. An important point to stress is that experimental design and post-processing are linked. Opinions on appropriate controls, adequate sample preparation, tailored data acquisition, quality controls and data analysis evolve together with their respective fields, which eventually reach a level of maturity enabling standardisation of optimised procedures. Interdisciplinary projects that combine elements from different sub-fields should aim at respecting their established or evolving conventions; the following attempts to provide a summary relevant to this thesis.

1.3.6.1. Proteome profiling

Protein ID In its simplest case, a list of protein with their intensity is used to know which proteins are identified in the sample and potentially which isoform is expressed, by comparing sequence coverage with the genomic sequence. This information allows for individual follow-up with classic biochemical techniques such as Western Blot etc.

Global proteome Global proteome profiling tries to identify, and possibly quantify (relative or absolute), a maximum of proteins in the samples analysed. This allows to identify up- and down-regulated proteins in different sample conditions (differential expression, such as tumor vs healthy) or at different timepoints (temporal dynamics) by statistical significance analysis and plots,[99,418] using statistical packages such as Perseus,[419] MSstats[420] or the R statistical computing language.[421] The proteomic data is then interpreted by inferring from protein function to observed phenotypes.[422]

Additionally, it allows to identify over-expressed proteins that could have analytical value as biomarkers.[423]

1.3.6.2. Spacial proteomics

Spacial proteomics aims at identifying the subcellular localisation of proteins. Mislocalisation of proteins is frequently associated with cellular dysfunction and disease, spacial proteomics aims at integrating the exact cellular localisation of a given protein in our biological understanding.[424] MS-based organelle profiling uses separation techniques for organelles such as cellular nuclei (i.e. sucrose gradient centrifugation), however these techniques unavoidably include constituents from other organelles as contaminants (see chapter 4). Data analysis and post-processing in spacial proteomics therefore focusses on identifying the contaminants using abundance distribution profile analysis over the fractions followed by a cluster analysis − also referred to as protein correlation profiling (PCP). Other types of analysis have emerged including the TMT-based localisation of organelle protein by isotope tagging (LOPIT) or dynamic organellar maps versus an internal reference.[424] Additionally, retrospective meta-analysis of subcellular proteins with R has been described in the MetaMass project, based on the use of standardised lists of subcellular markers.[425]

Of interest, the development of chemical proteomics methods (see below) has also enabled the emerging of additional, chemical proteomics-based techniques for subcellular proteome analysis, such as organelle-specific ABPP, proximity labelling and organelle-localisable reactive small molecules.[426] Conceptually, control experiments and data analysis should remain the same.

1.3.6.3. Functional analysis

Gene ontology An initial step for the interpretation of proteomics protein lists is to connect the protein identifier with its associated gene ontology (GO) terms. GO is a bioinformatics initiative to develop a classification for eukaryotic genes or proteins into a functional category. These categories are organised hierarchically, starting from three original domains: 'biological process', 'molecular function', 'cellular component'.[427]

Once identified proteins have been linked to their GO terms, enrichment analysis can be performed. This procedure identifies GO terms and in extension biological pathways that are most likely related to the proteomic data using p-values. In its simplest form, singular enrichment analysis (SEA) compares the abundance of GO terms in the dataset

with the natural abundance in a reference dataset, selected manually or implemented in the tool, such as Panther, DAVID, GOrilla or the MS-specialised aGO tool. More sophisticated algorithms allow for gene set enrichment (GSEA, also tool name) or modular enrichment analysis (MEA).[428,429]

Several post-processing tools allow for GO enrichment analysis, yet the information obtained remains in many cases on a higher / phenomenological level and is typically not broken down into single pathways or a molecular understanding. It is however very useful as a 'first-pass' technique (see chapter 4).

Pathway analysis Biological pathways are commonly defined as series of cellular chemical reactions that together cause a biological effect. This includes both proteins involved in the chemical reaction (as enzyme or signalling partner) and those that have regulatory influence. Proteins have been combined in pathway databases such as KEGG, Reactome, BioCarta or the proprietary Ingenuity Pathway Analysis Knowledge Base. This allow to interpret or identify types of biological processes within the proteomics dataset that can be particularly present.[427]

Pathway analysis in its simplest method is the analysis of protein lists for abundances that represent a particular pathway; more complex versions build biological network models. Modern, feature rich software such as the commercial Ingenuity Pathway Analysis, or open source Cytoscape (modules such as BioNSi), Pathview, Metascape and to some extend STRING allow to additionally annotate protein families, localisation, perform statistical testing of the network and aid in the interpretation of proteomics data by simulating biological systems[427]

1.3.6.4. Protein-protein interaction partners

Transient or stable complexes are at the core of biochemical processes, such as signalling, molecular assembly and function inhibition. Protein-protein interactions can be analysed in the complex of molecular networks with bioinformatic tools such as Cytoscape, STRING and EnrichNet; these tools access specialised interaction databases which include data from experimental sources and literature mining: IntAct (protein interactions), InterPro (Protein families, domains), Corum (protein complex information)[427], Human Reference Interactome (HuRi, by yeast-two-hybrid) and BioPlex (AE-MS).[430]

Despite these bioinformatic resources, protein-protein interactions are often investigated with tailored experiments,[121] such as affinity-enrichment mass spectrometry (AP-

MS), proximity-based labellings (BioID-MS with BirA[431]; APEX-MS[431] ; photoredox carbene MicroMap[432]) or, less frequently, by crosslinking mass spectrometry (see 1.3.6.6). In these cases, the data analysis is typically performed only on the acquired data with statistical packages such as Perseus, using functionalities such as hierarchical clustering, statistical significance test and volcano plots among other[433] – see chapter 4 for example. Additionally, several scoring algorithm have been widely adopted for the identification of high confidence interactors, such as SAINTexpress[434] and MiST.[435]

More recently, meta-analysis has become possible to combine obtained experimental data by AP-MS or BioID with previously deposited knowledge. This enabled deeper analysis of protein-protein interactions/associations leading to a map of human protein complexes[436] and a database of protein associations (Tabloid Proteome).[437] An exciting development is the inclusion of 3D protein complex and protein structure data in these interaction networks.[438]

A remarkable example of the impact of protein-protein interaction approaches can have is demonstrated by the recent collaborative paper on repurposing existing drugs for the treatment of Covid-19. In this collaborative effort between groups as UCSF, Mount Sinai and the Institut Pasteur, AP-MS was performed for all proteins of the SARS-CoV-2 virus to identify interaction partners that are already targeted by existing drugs in the clinic. This lead to the identification of *in vivo* active compounds and the modulation of Sigma1/Sigma2 receptors as a promising therapeutic strategy.[439]

1.3.6.5. PTM analysis

The challenges in PTM analysis reside predominantly in the correct sample preparation (PTM enrichment), data acquisition (MS-labile modifications) and database search (see page 64). In the case of single proteins or samples with limited complexity, the focus is typically on site-identification and quantitation of micro- and macro-heterogeneity (see chapter 2). The challenges of the quantitation of post-translational modifications has been highlighted in section 1.3.5.4.

Large-scale proteomic PTM analysis can currently be conducted for phosphorylation, acetylation, ubiquitinylation and very recently, to some extend, for *N*-glycoproteomics (using ZIC-HILIC with SCE-HCD fragmentation methods).[440–442] Additionally, specialised workflows for large-scale analysis of histone PTM have allowed to identify their multiple PTMs simultaneously.[443,444] Post-processing tools tend to be predominantly used for analysis of numbers of identifications and tracking site-specific changes cross-comparing conditions or for statistical validation of the identifications using statistics

packages such as Perseus. Network analysis tools and known databases of PTM (UniProt, PhosphoSitePlus, BioGRID,[445] PTMsigDB[446]) can be included in the data analysis. An inclusive approach is PTM signature enrichment analysis that allows PTM-site-centric pathway analysis while looking at signalling pathway level to identify site-specific modifications that reflect the activity level of a given biological pathway.[446] Glycosylation has to the best of knowledge not yet been combined in post-processing with network analysis, although databases such as UniCarbKB, CAZy or ProCarbDB contain readily accessible data.[447]

Despite the difficulties, the central role PTMs have in signalling and in altering protein activity makes the development of improved methods for their future study indispensable. Quantitation of PTMs bears the promise to highlight which pathways are active or turned on, as opposed to the information about which pathway are overexpressed.[448] Additionally, tracking the formation of protein-protein complexes as a function of PTM is another promising avenue.[422] A holistic understanding of site-specific PTM-based signalling, intricated with its structural recognition, has great potential for accelerating the discovery of new therapeutic approaches and biomarkers.[449] As an example from the field of epigenetics, bromodomains are a notorious class of drug targets that recognise lysine acetylations and whose structural study has enabled the development of new drug candidates.[450]

Chemically modified proteins One major drawback has long been the limited ability to recombinantly generate post-translationally modified proteins. Synthetic chemical biology has developed strategies to generate synthetic proteins with close or identic PTM mimicks (see section 4.2.1), which can be used for the study of PTMs, particularly with the focus of finding interaction partners (see chapter 4) or the structural study of their recognition (see Chapter 5).

The MS/MS analysis of single purified chemically modified proteins carrying unnatural modifications, for purposes of quality control (proof of *site-specificity*), brings its own set of challenges. The purpose of data analysis and post-processing is to assess the quality of the MS/MS method regarding possible side reactions on the modifications (acid labile, radical labile, photoreactive, rearrangements) and the quality of the identification.¶

¶Personal observation. This has not yet been reviewed concisely in the literature.

1.3.6.6. Structural proteomics

Structural proteomics aims at combining information about all proteins together with their 3D information and interactions, to gain an understanding of how structural assembly and biological function are integrated.[451,452]

Protein assemblies and their structural models In recent years, cross-linking mass spectrometry (XL-MS) have become a powerful structural tool for mapping PPIs and thus elucidating architectures of large protein complexes. The advantages are speed, versatility, need for little materials and the possibility to study of heterogeneous and dynamic protein complexes.[397] In its simplest case, post-processing of XL-MS data allows modelling of a single protein, by establishing single residue crosslinking maps. This allows to define distance constraints for the 3D geometry of the protein, which can be used for modelling applications such as structure prediction or homology modelling.[399] Additionally, it allows to track conformational changes, for example inducted by PTMs.[385]

The use of XL-MS to elucidate the conformation of unknown protein complexes with native conformations and topologies has led to widespread interest[385], identifying key residues in the protein-protein interaction (see chapter 5). Several flexible protein-protein docking algorithms such as HADDOCK or Rosetta allow the incorporation of experimental crosslinking data to guide the modelling, generating experimentally plausible structural models of the protein-protein interaction.[453] A particularly powerful approach is the combination of identified crosslinks with data from other structural biology methods, such as electron microscopy (EM),[454–456] X-ray or NMR, which allows to reconstruct macromolecular assemblies.

Hydrogen-deuterium exchange (HDX) is another MS/MS technology used for structural proteomics. The exchange of deuterated protein backbone hydrogens for ^{1}H gives information about the solvent accessibility of a site[457] and allows to identify binding interface of protein-protein and protein-ligand interactions.[458]

Proteome-wide applications The improvement of MS/MS methods over the last decade has allowed to study the conformational properties of proteins and protein-ligand complexes on a proteome-wide scale.[451] XL-MS methods are to date able to deliver large-scale or systems-level interactome data that contain topological information from isolated organelles, cell lysates, virus particles, intact bacterial and mammalian cultured cells and tissue samples.[459] Although still limited and technically challenging,

their use complements AP-MS by also being able to capture transient and dynamic interactions.[401]

Beyond XL-MS, other techniques have been used for structural proteomics.[451,460] This includes native state approaches such as limited proteolysis methods[461], denaturant-based approaches that probe protein stability (CETSA/TPP – see below) or covalent labelling methods with limited range.[452]

Structural glycoproteomics Initial crosslinking approaches have been developed for structural glycoproteomics allowing to study glycan-protein interactions, based either on crosslinking the protein with oxidised carbohydrate structures[394] or by photocrosslinkers included in carbohydrates by metabolic labelling[462] (see paragraph on chemical glycoproteomics in 1.3.6.7).

A recent example is the site-specific analysis of the Sars-Cov-2 glycan shield. By combining *N*-glycoproteomics with electron microscopy and molecular modelling, the Crispin group was able to generate a structural model of the virus capsid that includes the glycans at its surface, mapped together with their respective abundance. This led to the identification of a physiologically rarely present oligomannose glycan that is sandwiched between two protein domains and could potentially be a promising epitope for antibody targeting.[463]

1.3.6.7. Chemical proteomics

Chemical proteomics is the use of small molecules, synthetic peptides and synthetic proteins in proteomics experiments to study and quantify their binding to other cellular proteins. Equally, the concept of chemical proteomics also includes the use of chemical tools to advance proteomic research, from small molecule crosslinkers to chemical biology methods for the enrichment of specific peptides/proteins.[464,465] In this thesis, chemical proteomics projects with synthetic proteins are presented (chapter 4, chapter 5).

Small molecule chemical proteomics Most chemical proteomics follow tailored protocols in which a ligand (such as a small molecule) is incubated with the cells or *in situ* with cellular lysate. Efforts in preclinical drug discovery such as small molecule target identification, target validation and target engagement projects use a range of methods from the following categories:[466]

- compound-dependent global proteome profiling[467,468] (including FITExP[469,470]).
- activity-based protein profiling (ABPP)[120,471]. ABPP was also done with carbohydrates for probing glycosidases[472,473], or using photoreactive moieties.[474,475]
- compound-centric chemoproteomics (CCCP, an affinity purification).[467,468] This was tested for carbohydrates with a DNJ-matrice.[476]
- Stability shift: thermal proteome profiling (TPP/CETSA)[477] and other denaturant-based approaches (PP, SPROX).[451]
- limited proteolysis methods (LiP[478], DARTS[479,480]).
- covalent labelling methods such as hydrogen-deuterium exchange (HDX-MS) and hydroxyl-radical footprinting (HFR).
- compound-modulated protein interaction profiling.[467]

The assay description for each method is beyond the scope of this thesis and has been highlighted elsewhere.[451,467] Many approaches have been combined with methods for quantitative proteomics, allowing multiplexing.[481] All chemical proteomics experiments rely on appropriate control experiments for statistical validation at post-processing stage, depending on the subject study this could be (sorted by similarity to the active compound): an inactive enantiomer‖ (ABPP with natural products), an inactive structurally close analogue,[482] background reactivity controls (photochemical labelling[483,484], covalent warheads), DMSO-controls (CCCP, TPP). Post-processing mainly uses Perseus and R/python-based scripts, which allow the calculation of K_D^{app} or EC_{50} for compound affinity[119] and the elaboration of advanced metrics such as CATDS for compound selectivity.[485] Several databases catalogue ligand affinity values for different proteins, such as DrugBank,[486] ChEMBL,[487] BioGRID,[445] PubChem,[488] BindingDB,[489] chemicalprobes.org[482], probes-drug.org[490] among others.

Chemical glycoproteomics Chemical glycoproteomics mainly refers to the use of chemical tools to enable glycoproteomics projects.[491] A large focus has been the of enrichment tools for glycopeptides based on metabolic labelling, such as the IsoTaG method[188], or engineered enzymatic extension with unnatural monosaccharides (see 1.3.2.3). For functional studies, a trifunctional chemoproteomics probe called TRICEPS was designed to crosslink periodate-oxidised cell surface glycans with its trifluoroacetylated hydrazine moiety and their interacting proteins by *N*-hydroxysuccinimide, while having a biotin

‖ R.B. Woodward "the only model system worth using is the enantiomer", London 1968, on the synthesis of vitamin B12.

affinity tag.[394] Similarly, photoactivable probes for the photocrosslinking of glycans that can be incorporated by metabolic labelling have been developed.[462] In all cases, the complex data analysis requires tailored code that is able to analyse the complex fragmentation patterns generated. The limitations of these method have thus far prevented large-scale use.[491]

1.3.6.8. Multi-omics data integration

The central dogma of molecular biology is not quantitative.[492] This means there is no clear quantitative correlation between datasets in genomics, transcriptomics and proteomics.[492,493] Yet, the qualitative pattern contain valuable information and multi-omics approach have a great potential to approach a biological question from different angles.[494] Multi-omics studies have been facilitated by the practice of uploading raw data into public data repositories such as PRIDE, part of the ProteomeXchange network, or OmicsDI[495]. The use of publicly available data has allowed re-analysis and combining datasets.[496,497]

In proteogenomics, proteomics information derived from mass spectrometry is used to complement genomics or transcriptomics (RNAseq) data generated using next generation sequencing. This allows to identify splice variants, single nucleotide polymorphisms (SNP) or single amino acid variants (SAAVs).[352] An example study in colorectal cancer used a proteogenomic approach to highlight the importance of the 20q amplification, a previously controversially discussed concept.[498] However despite the promises, bioinformatic data analysis in the field faces significant challenges regarding analysis, including the need to overcome "data hoarding" and false discoveries resulting from inaccurate filtering thresholds and incorrect identifications of homologous peptides.[352] Going even further, Mun *et al.* combined mRNA, protein, phosphorylation and *N*-glycosylation data on a set of 80 tissues from young gastric cancer patients. Their analysis identified four subtypes and their associated distinct cellular pathways. This data allows to now categorise patients in positive or negative prognosis.[499]

In summary, proteomics approaches have a tremendous impact as discovery tools for the life sciences. Tandem mass spectrometry is a highly versatile technology which allows acquisition methods tailored to the research subject. There are numerous opportunities for the development of new MS/MS techniques, particularly for the glycosciences, and the combination of MS/MS with chemical biology allows to ask very precise biological questions (synthetic precision) with a powerful readout.

1.4. References

1. *Essentials of Glycobiology* 3rd (eds Varki, A. *et al.*) (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).
2. Vranken, D. V. & Weiss, G. A. *Introduction to Bioorganic Chemistry and Chemical Biology* 1 edition. 504 pp. (Garland Science, New York, 2012).
3. Krištić, J. & Lauc, G. in *High-Throughput Glycomics and Glycoproteomics: Methods and Protocols* (eds Lauc, G. & Wuhrer, M.) 1–12 (Springer, New York, NY, 2017).
4. Varki, A. Biological Roles of Glycans. *Glycobiology* **27,** 3–49 (2017).
5. Cummings, R. D., Schnaar, R. L., Esko, J. D., Drickamer, K. & Taylor, M. E. in *Essentials of Glycobiology* (eds Varki, A. *et al.*) 3rd (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).
6. Jiang, X., Dias, J. A. & He, X. Structural Biology of Glycoprotein Hormones and Their Receptors: Insights to Signaling. *Molecular and Cellular Endocrinology* **382,** 424–451 (2014).
7. Parker, R. B. & Kohler, J. J. Regulation of Intracellular Signaling by Extracellular Glycan Remodeling. *ACS chemical biology* **5,** 35–46 (2010).
8. Kolarich, D., Lepenies, B. & Seeberger, P. H. Glycomics, Glycoproteomics and the Immune System. *Current Opinion in Chemical Biology. Bioinorganic Chemistry • Biocatalysis and Biotransformation • Omics* **16,** 214–220 (2012).
9. Johnson, J. L., Jones, M. B., Ryan, S. O. & Cobb, B. A. The Regulatory Power of Glycans and Their Binding Partners in Immunity. *Trends in Immunology* **34,** 290–298 (2013).
10. Lederkremer, G. Z. Glycoprotein Folding, Quality Control and ER-Associated Degradation. *Current Opinion in Structural Biology. Carbohydradtes and Glycoconjugates / Biophysical Methods* **19,** 515–523 (2009).
11. Spiro, R. G. Protein Glycosylation: Nature, Distribution, Enzymatic Formation, and Disease Implications of Glycopeptide Bonds. *Glycobiology* **12,** 43R–56R (2002).
12. Pinho, S. S. & Reis, C. A. Glycosylation in Cancer: Mechanisms and Clinical Implications. *Nature Reviews Cancer* **15,** 540–555 (9 2015).
13. Fuster, M. M. & Esko, J. D. The Sweet and Sour of Cancer: Glycans as Novel Therapeutic Targets. *Nature Reviews Cancer* **5,** 526–542 (7 2005).
14. Pearce, O. M. T. Cancer Glycan Epitopes: Biosynthesis, Structure and Function. *Glycobiology* **28,** 670–696 (2018).
15. Very, N., Lefebvre, T. & Yazidi-Belkoura, I. E. Drug Resistance Related to Aberrant Glycosylation in Colorectal Cancer. *Oncotarget* **9,** 1380–1402 (1 2017).
16. Rudd, P. M. & Dwek, R. A. Glycosylation: Heterogeneity and the 3D Structure of Proteins. *Critical Reviews in Biochemistry and Molecular Biology* **32,** 1–100 (1997).
17. Flynn, R. A., Smith, B. A. H., Johnson, A. G., Pedram, K., George, B. M., Malaker, S. A., Majzoub, K., Carette, J. E. & Bertozzi, C. R. Mammalian Y RNAs Are Modified at Discrete Guanosine Residues with N-Glycans. *bioRxiv,* 787614 (2019).
18. Varki, A. & Kornfeld, S. in *Essentials of Glycobiology* (eds Varki, A. *et al.*) 3rd (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

19. Cohen, M. M. & Varki, A. The Sialome–Far More than the Sum of Its Parts. *Omics : a journal of integrative biology* (2010).

20. Kavli Frontiers of Science. *Glycobiology - Pascal Gagneux, University of California, San Diego* `https://vimeo.com/31120368` (2020).

21. Lindahl, U., Couchman, J., Kimata, K. & Esko, J. D. in *Essentials of Glycobiology* (eds Varki, A. *et al.*) 3rd (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

22. Seino, J., Fujihira, H., Nakakita, S.-i., Masahara-Negishi, Y., Miyoshi, E., Hirabayashi, J. & Suzuki, T. Occurrence of Free Sialyl Oligosaccharides Related to N-Glycans (Sialyl Free N-Glycans) in Animal Sera. *Glycobiology* **26,** 1072–1085 (2016).

23. Gagneux, P., Aebi, M. & Varki, A. in *Essentials of Glycobiology* (eds Varki, A. *et al.*) 3rd (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

24. Asperger, A., Marx, K., Albers, C., Molin, L. & Pinato, O. Low Abundant N-Linked Glycosylation in Hen Egg White Lysozyme Is Localized at Nonconsensus Sites. *Journal of Proteome Research* **14,** 2633–2641 (2015).

25. Chavan, M. & Lennarz, W. The Molecular Basis of Coupling of Translocation and N-Glycosylation. *Trends in Biochemical Sciences* **31,** 17–20 (2006).

26. Stanley, P., Taniguchi, N. & Aebi, M. in *Essentials of Glycobiology* (eds Varki, A. *et al.*) 3rd (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

27. Jarrell, K. F., Ding, Y., Meyer, B. H., Albers, S.-V., Kaminski, L. & Eichler, J. N-Linked Glycosylation in Archaea: A Structural, Functional, and Genetic Analysis. *Microbiology and Molecular Biology Reviews* **78,** 304–341 (2014).

28. Yurist-Doutsch, S., Chaban, B., VanDyke, D. J., Jarrell, K. F. & Eichler, J. Sweet to the Extreme: Protein Glycosylation in Archaea. *Molecular Microbiology* **68,** 1079–1084 (2008).

29. Abu-Qarn, M., Eichler, J. & Sharon, N. Not Just for Eukarya Anymore: Protein Glycosylation in Bacteria and Archaea. *Current Opinion in Structural Biology. Carbohydrates and Glycoconjugates / Biophysical Methods* **18,** 544–550 (2008).

30. Li, S. *et al.* Pathogen Blocks Host Death Receptor Signalling by Arginine GlcNAcylation of Death Domains. *Nature* **501,** 242–246 (7466 2013).

31. Pearson, J. S. *et al.* A Type III Effector Antagonizes Death Receptor Signalling during Bacterial Gut Infection. *Nature* **501,** 247–251 (7466 2013).

32. Park, J. B. *et al.* Structural Basis for Arginine Glycosylation of Host Substrates by Bacterial Effector Proteins. *Nature Communications* **9,** 1–15 (1 2018).

33. Lassak, J. *et al.* Arginine-Rhamnosylation as New Strategy to Activate Translation Elongation Factor P. *Nature Chemical Biology* **11,** 266–270 (4 2015).

34. Li, X. *et al.* Resolving the α-Glycosidic Linkage of Arginine-Rhamnosylated Translation Elongation Factor P Triggers Generation of the First ArgRha Specific Antibody. *Chemical Science* **7,** 6995–7001 (2016).

35. Schreiner, R., Schnabel, E. & Wieland, F. Novel N-Glycosylation in Eukaryotes: Laminin Contains the Linkage Unit Beta-Glucosylasparagine. *The Journal of Cell Biology* **124,** 1071–1081 (1994).

36. Singh, D. G., Lomako, J., Lomako, W. M., Whelan, W. J., Meyer, H. E., Serwe, M. & Metzger, J. W. β-Glucosylarginine: A New Glucose-Protein Bond in a Self-Glucosylating Protein from Sweet Corn. *FEBS Letters* **376,** 61–64 (1995).

37. Randell, K. D., Johnston, B. D., Green, D. F. & Pinto, B. M. Is There a Generalized Reverse Anomeric Effect? Substituent and Solvent Effects on the Configurational Equilibria of Neutral and Protonated N-Arylglucopyranosylamines and N-Aryl-5-Thioglucopyranosylamines. *The Journal of Organic Chemistry* **65,** 220–226 (2000).

38. *The Anomeric Effect and Associated Stereoelectronic Effects* (ed Thatcher, G. R. J.) 299 pp. (American Chemical Society, Washington, DC, 1993).

39. Fleming, I. *Molecular Orbitals and Organic Chemical Reactions - Student Edition* 1. 378 pp. (Wiley, Chichester, West Sussex, U.K, 2009).

40. Evich, M., Stroeva, E., Zheng, Y. G. & Germann, M. W. Effect of Methylation on the Side-chain pK a Value of Arginine. *Protein Science : A Publication of the Protein Society* **25,** 479–486 (2016).

41. Furmanek, A. & Hofsteenge, J. Protein C-Mannosylation: Facts and Questions. *Acta Biochimica Polonica* **47,** 781–789 (3 2000).

42. Haltiwanger, R. S., Wells, L., Freeze, H. H. & Stanley, P. in *Essentials of Glycobiology* (eds Varki, A. *et al.*) 3rd (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

43. Iwahashi, N., Inai, Y., Minakata, S., Sakurai, S., Manabe, S., Ito, Y., Ino, K. & Ihara, Y. C-Mannosyl Tryptophan Increases in the Plasma of Patients with Ovarian Cancer. *Oncology Letters* **19,** 908–916 (2020).

44. Falzarano, D., Krokhin, O., Van Domselaar, G., Wolf, K., Seebach, J., Schnittler, H.-J. & Feldmann, H. Ebola sGP–the First Viral Glycoprotein Shown to Be C-Mannosylated. *Virology* **368,** 83–90 (2007).

45. Schäffer, C. & Messner, P. Emerging Facets of Prokaryotic Glycosylation. *FEMS Microbiology Reviews* **41,** 49–91 (2017).

46. Joshi, H. J., Narimatsu, Y., Schjoldager, K. T., Tytgat, H. L. P., Aebi, M., Clausen, H. & Halim, A. SnapShot: O-Glycosylation Pathways across Kingdoms. *Cell* **172,** 632–632.e2 (2018).

47. Davis, B. G. & Fairbanks, A. J. *Carbohydrate Chemistry: 99* 104 pp. (Oxford University Press, USA, Oxford ; New York, 2002).

48. Halim, A., Brinkmalm, G., Rüetschi, U., Westman-Brinkmalm, A., Portelius, E., Zetterberg, H., Blennow, K., Larson, G. & Nilsson, J. Site-Specific Characterization of Threonine, Serine, and Tyrosine Glycosylations of Amyloid Precursor Protein/Amyloid β-Peptides in Human Cerebrospinal Fluid. *Proceedings of the National Academy of Sciences* **108,** 11848–11853 (2011).

49. Hollingsworth, M. A. & Swanson, B. J. Mucins in Cancer: Protection and Control of the Cell Surface. *Nature Reviews Cancer* **4,** 45–60 (1 2004).

50. Brockhausen, I. & Stanley, P. in *Essentials of Glycobiology* (eds Varki, A. *et al.*) 3rd (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

51. Darula, Z. & Medzihradszky, K. F. Analysis of Mammalian O-Glycopeptides—We Have Made a Good Start, but There Is a Long Way to Go. *Molecular & Cellular Proteomics* **17,** 2–17 (2018).

52. Harvey, B. M. & Haltiwanger, R. S. in *Molecular Mechanisms of Notch Signaling* (eds Borggrefe, T. & Giaimo, B. D.) 59–78 (Springer International Publishing, Cham, 2018).

53. Darabedian, N., Gao, J., Chuh, K. N., Woo, C. M. & Pratt, M. R. The Metabolic Chemical Reporter 6-Azido-6-Deoxy-Glucose Further Reveals the Substrate Promiscuity of O-GlcNAc Transferase and Catalyzes the Discovery of Intracellular Protein Modification by O-Glucose. *Journal of the American Chemical Society* **140,** 7092–7100 (2018).

54. Ogawa, M. & Okajima, T. Structure and Function of Extracellular O-GlcNAc. *Current Opinion in Structural Biology. Sequences and Topology · Carbohydrates* **56,** 72–77 (2019).

55. Sheikh, M. O., Halmo, S. M. & Wells, L. Recent Advancements in Understanding Mammalian O-Mannosylation. *Glycobiology* **27,** 806–819 (2017).

56. Stalnaker, S. H., Stuart, R. & Wells, L. Mammalian O-Mannosylation: Unsolved Questions of Structure/Function. *Current Opinion in Structural Biology. Carbohydrates and Glycoconjugates/-Biophysical Methods* **21,** 603–609 (2011).

57. De Leon, C. A., Levine, P. M., Craven, T. W. & Pratt, M. R. 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* **56,** 3507–3517 (2017).

58. Lote, C. J. & Weiss, J. B. Identification of Digalactosylcysteine in a Glycopeptide Isolated from Urine by a New Preparative Technique. *FEBS Letters* **16,** 81–85 (1971).

59. Weiss, J. B., Lote, C. J. & Bobinski, H. New Low Molecular Weight Glycopeptide Containing Triglucosylcysteine in Human Erythrocyte Membrane. *Nature New Biology* **234,** 25–26 (44 1971).

60. Venugopal, H., Edwards, P. J. B., Schwalbe, M., Claridge, J. K., Libich, D. S., Stepper, J., Loo, T., Patchett, M. L., Norris, G. E. & Pascal, S. M. Structural, Dynamic, and Chemical Characterization of a Novel S-Glycosylated Bacteriocin. *Biochemistry* **50,** 2748–2755 (2011).

61. Stepper, J., Shastri, S., Loo, T. S., Preston, J. C., Novak, P., Man, P., Moore, C. H., Havlíček, V., Patchett, M. L. & Norris, G. E. Cysteine S-Glycosylation, a New Post-Translational Modification Found in Glycopeptide Bacteriocins. *FEBS letters* **585,** 645–650 (2011).

62. Oman, T. J., Boettcher, J. M., Wang, H., Okalibe, X. N. & van der Donk, W. A. Sublancin Is Not a Lantibiotic but an S-Linked Glycopeptide. *Nature Chemical Biology* **7,** 78–80 (2 2011).

63. Maynard, J. C., Burlingame, A. L. & Medzihradszky, K. F. Cysteine S-Linked N-Acetylglucosamine (S-GlcNAcylation), A New Post-Translational Modification in Mammals. *Molecular & Cellular Proteomics* **15,** 3405–3411 (2016).

64. Macauley, M. S., Stubbs, K. A. & Vocadlo, D. J. O-GlcNAcase Catalyzes Cleavage of Thioglycosides without General Acid Catalysis. *Journal of the American Chemical Society* **127,** 17202–17203 (2005).

65. Goodman, I., Fouts, J. R., Bresnick, E., Menegas, R. & Hitchings, G. H. A Mammalian Thioglycosidase. *Science* **130,** 450–451 (1959).

66. Haynes, P. A. Phosphoglycosylation: A New Structural Class of Glycosylation? *Glycobiology* **8,** 1–5 (1998).

67. Ilg, T. Proteophosphoglycans of Leishmania. *Parasitology Today* **16,** 489–497 (2000).

68. Capul, A. A., Barron, T., Dobson, D. E., Turco, S. J. & Beverley, S. M. Two Functionally Divergent UDP-Gal Nucleotide Sugar Transporters Participate in Phosphoglycan Synthesis in Leishmania Major. *Journal of Biological Chemistry* **282,** 14006–14017 (2007).

69. Ferguson, M. A. J., Hart, G. W. & Kinoshita, T. in *Essentials of Glycobiology* (eds Varki, A. *et al.*) 3rd (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

70. Stanley, P. & Cummings, R. D. in *Essentials of Glycobiology* (eds Varki, A. *et al.*) 3rd (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

71. Dalziel, M., Crispin, M., Scanlan, C. N., Zitzmann, N. & Dwek, R. A. Emerging Principles for the Therapeutic Exploitation of Glycosylation. *Science* **343** (2014).

72. Almeida, A. & Kolarich, D. The Promise of Protein Glycosylation for Personalised Medicine. *Biochimica et Biophysica Acta (BBA) - General Subjects. Glycans in Personalised Medicine* **1860,** 1583–1595 (2016).

73. Ernst, B. & Magnani, J. L. From Carbohydrate Leads to Glycomimetic Drugs. *Nature Reviews Drug Discovery* **8,** 661–677 (2009).

74. Magnani, J. L. & Ernst, B. Glycomimetic Drugs–a New Source of Therapeutic Opportunities. *Discovery Medicine* **8,** 247–252 (2009).

75. Magnani, J. *Glycobiology: An Untapped Source of Novel Drugs* The Biomedical & Life Sciences Collection. `https://hstalks.com/t/2435/glycobiology-an-untapped-source-of-novel-drugs/` (2020).

76. Odejinmi, S. I., Rascon, R. G., Tang, M., Vankayalapati, H. & Lai, K. Structure–Activity Analysis and Cell-Based Optimization of Human Galactokinase Inhibitors. *ACS Medicinal Chemistry Letters* **2,** 667–672 (2011).

77. Liu, L. *et al.* Structure Activity Relationships of Human Galactokinase Inhibitors. *Bioorganic & Medicinal Chemistry Letters* **25,** 721–727 (2015).

78. Upton, R., Migas, L. G., Pacholarz, K. J., Beniston, R. G., Estdale, S., Firth, D. & Barran, P. E. Hybrid Mass Spectrometry Methods Reveal Lot-to-Lot Differences and Delineate the Effects of Glycosylation on the Tertiary Structure of Herceptin®. *Chemical Science* **10,** 2811–2820 (2019).

79. Chung, C. H. *et al.* Cetuximab-Induced Anaphylaxis and IgE Specific for Galactose-α-1,3-Galactose. *New England Journal of Medicine* **358,** 1109–1117 (2008).

80. Spearman, M., Dionne, B. & Butler, M. in *Antibody Expression and Production* (ed Al-Rubeai, M.) 251–292 (Springer Netherlands, Dordrecht, 2011).

81. Hossler, P., Khattak, S. F. & Li, Z. J. Optimal and Consistent Protein Glycosylation in Mammalian Cell Culture. *Glycobiology* **19,** 936–949 (2009).

82. Adamo, R., Nilo, A., Castagner, B., Boutureira, O., Berti, F. & Bernardes, G. J. L. Synthetically Defined Glycoprotein Vaccines: Current Status and Future Directions. *Chemical Science* **4,** 2995–3008 (2013).

83. Zhang, B., Whiteaker, J. R., Hoofnagle, A. N., Baird, G. S., Rodland, K. D. & Paulovich, A. G. Clinical Potential of Mass Spectrometry-Based Proteogenomics. *Nature Reviews Clinical Oncology* **16,** 256–268 (4 2019).

84. Peracaula, R., Barrabés, S., Sarrats, A., Rudd, P. M. & de Llorens, R. Altered Glycosylation in Tumours Focused to Cancer Diagnosis. *Disease Markers* **25,** 207 (2008).

85. Hart, G. W. & Varki, A. in *Essentials of Glycobiology* (eds Varki, A. *et al.*) 3rd (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

86. Agre, P. *et al.* Training the next Generation of Biomedical Investigators in Glycosciences. *The Journal of Clinical Investigation* **126,** 405–408 (2016).

87. Malaker, S. A. *et al.* The Mucin-Selective Protease StcE Enables Molecular and Functional Analysis of Human Cancer-Associated Mucins. *Proceedings of the National Academy of Sciences* **116,** 7278–7287 (2019).

88. Yang, S., Onigman, P., Wu, W. W., Sjogren, J., Nyhlen, H., Shen, R.-F. & Cipollo, J. Deciphering Protein O-Glycosylation: Solid-Phase Chemoenzymatic Cleavage and Enrichment. *Analytical Chemistry* **90,** 8261–8269 (2018).

89. Zolg, D. P. *Building ProteomeTools Based on a Complete Synthetic Human Proteome* (Technical University of Munich, Freising, 2019). 182 pp.

90. Kelleher, N. L. Peer Reviewed: Top-Down Proteomics. *Analytical Chemistry* **76,** 196 A–203 A (2004).

91. Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annual Review of Analytical Chemistry* **9,** 499–519 (2016).

92. Tholey, A. & Becker, A. Top-down Proteomics for the Analysis of Proteolytic Events - Methods, Applications and Perspectives. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. Proteolysis as a Regulatory Event in Pathophysiology* **1864,** 2191–2199 (11, Part B 2017).

93. Yang, Y., Liu, F., Franc, V., Halim, L. A., Schellekens, H. & Heck, A. J. R. Hybrid Mass Spectrometry Approaches in Glycoprotein Analysis and Their Usage in Scoring Biosimilarity. *Nature Communications* **7,** 1–10 (1 2016).

94. Catherman, A. D., Skinner, O. S. & Kelleher, N. L. Top Down Proteomics: Facts and Perspectives. *Biochemical and Biophysical Research Communications. Advances in OMICs-Based Disciplines* **445,** 683–693 (2014).

95. Breuker, K. Raw Protein from the Top Down. *Nature Chemistry* **10,** 114–116 (2 2018).

96. Li, H., Nguyen, H. H., Loo, R. R. O., Campuzano, I. D. G. & Loo, J. A. An Integrated Native Mass Spectrometry and Top-down Proteomics Method That Connects Sequence to Structure and Function of Macromolecular Complexes. *Nature Chemistry* **10,** 139–148 (2 2018).

97. Liko, I., Allison, T. M., Yen, H.-Y., Hopper, J. S. T. & Robinson, C. V. Using Native Mass Spectrometry to Inform Drug Discovery. *Drug Target Review* **4,** 44–47 (2017).

98. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chemical Reviews* **113,** 2343–2394 (2013).

99. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-Generation Proteomics: Towards an Integrative View of Proteome Dynamics. *Nature Reviews Genetics* **14,** 35–48 (1 2013).

100. Olsen, J. V., Ong, S.-E. & Mann, M. Trypsin Cleaves Exclusively C-Terminal to Arginine and Lysine Residues. *Molecular & Cellular Proteomics* **3,** 608–614 (2004).

101. Tsiatsiani, L. & Heck, A. J. R. Proteomics beyond Trypsin. *The FEBS Journal* **282,** 2612–2626 (2015).

102. Maley, F., Trimble, R. B., Tarentino, A. L. & Plummer, T. H. Characterization of Glycoproteins and Their Associated Oligosaccharides through the Use of Endoglycosidases. *Analytical Biochemistry* **180,** 195–204 (1989).

103. Tretter, V., Altmann, F. & März, L. Peptide-N4-(N-Acetyl-β-Glucosaminyl)Asparagine Amidase F Cannot Release Glycans with Fucose Attached A1 → 3 to the Asparagine-Linked N-Acetylglucosamine Residue. *European Journal of Biochemistry* **199,** 647–652 (1991).

104. Ritorto, M. S., Cook, K., Tyagi, K., Pedrioli, P. G. A. & Trost, M. Hydrophilic Strong Anion Exchange (hSAX) Chromatography for Highly Orthogonal Peptide Separation of Complex Proteomes. *Journal of Proteome Research* **12,** 2449–2457 (2013).

105. Yu, P., Petzoldt, S., Wilhelm, M., Zolg, D. P., Zheng, R., Sun, X., Liu, X., Schneider, G., Huhmer, A. & Kuster, B. Trimodal Mixed Mode Chromatography That Enables Efficient Offline Two-Dimensional Peptide Fractionation for Proteome Analysis. *Analytical Chemistry* **89,** 8884–8891 (2017).

106. Ruprecht, B., Koch, H., Medard, G., Mundt, M., Kuster, B. & Lemeer, S. Comprehensive and Reproducible Phosphopeptide Enrichment Using Iron Immobilized Metal Ion Affinity Chromatography (Fe-IMAC) Columns. *Molecular & Cellular Proteomics* **14,** 205–215 (2015).

107. Bantscheff, M. *et al.* Quantitative Chemical Proteomics Reveals Mechanisms of Action of Clinical ABL Kinase Inhibitors. *Nature Biotechnology* **25,** 1035–1044 (2007).

108. Steen, H. & Mann, M. The Abc's (and Xyz's) of Peptide Sequencing. *Nature Reviews Molecular Cell Biology* **5,** 699–711 (9 2004).

109. Eng, J. K., McCormack, A. L. & Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry* **5,** 976–989 (1994).

110. Zaare, S., Aguilar, J. S., Hu, Y., Ferdosi, S. & Borges, C. R. Glycan Node Analysis: A Bottom-up Approach to Glycomics. *Journal of Visualized Experiments: JoVE* (2016).

111. Rudd, P., Karlsson, N. G., Khoo, K.-H. & Packer, N. H. in *Essentials of Glycobiology* (eds Varki, A. *et al.*) 3rd (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

112. Lu, H., Zhang, Y. & Yang, P. Advancements in Mass Spectrometry-Based Glycoproteomics and Glycomics. *National Science Review* **3,** 345–364 (2016).

113. Zhu, Z. & Desaire, H. Carbohydrates on Proteins: Site-Specific Glycosylation Analysis by Mass Spectrometry. *Annual Review of Analytical Chemistry* **8,** 463–483 (2015).

114. Thaysen-Andersen, M. & Packer, N. H. Advances in LC–MS/MS-Based Glycoproteomics: Getting Closer to System-Wide Site-Specific Mapping of the N- and O-Glycoproteome. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1844,** 1437–1452 (2014).

115. Thaysen-Andersen, M., Packer, N. H. & Schulz, B. L. Maturing Glycoproteomics Technologies Provide Unique Structural Insights into the N-Glycoproteome and Its Regulation in Health and Disease. *Molecular & Cellular Proteomics* **15,** 1773–1790 (2016).

116. Bodzon-Kulakowska, A., Bierczynska-Krzysik, A., Dylag, T., Drabik, A., Suder, P., Noga, M., Jarzebinska, J. & Silberring, J. Methods for Samples Preparation in Proteomic Research. *Journal of Chromatography B. Analytical Tools for Proteomics* **849,** 1–31 (2007).

117. Feist, P. & Hummon, A. B. Proteomic Challenges: Sample Preparation Techniques for Microgram-Quantity Protein Analysis from Biological Samples. *International Journal of Molecular Sciences* **16,** 3537–3563 (2 2015).

118. Bantscheff, M. *et al.* Quantitative Chemical Proteomics Reveals Mechanisms of Action of Clinical ABL Kinase Inhibitors. *Nature Biotechnology* **25,** 1035–1044 (9 2007).

119. Médard, G. *et al.* Optimized Chemical Proteomics Assay for Kinase Inhibitor Profiling. *Journal of Proteome Research* **14,** 1574–1586 (2015).

120. Niphakis, M. J. & Cravatt, B. F. Enzyme Inhibitor Discovery by Activity-Based Protein Profiling. *Annual Review of Biochemistry* **83,** 341–377 (2014).

121. Yakubu, R. R., Nieves, E. & Weiss, L. M. The Methods Employed in Mass Spectrometric Analysis of Posttranslational Modifications (PTMs) and Protein–Protein Interactions (PPIs). *Advances in experimental medicine and biology* **1140,** 169–198 (2019).

122. Makowski, M. M., Gräwe, C., Foster, B. M., Nguyen, N. V., Bartke, T. & Vermeulen, M. Global Profiling of Protein–DNA and Protein–Nucleosome Binding Affinities Using Quantitative Mass Spectrometry. *Nature Communications* **9,** 1–10 (1 2018).

123. Rietschel, B., Arrey, T. N., Meyer, B., Bornemann, S., Schuerken, M., Karas, M. & Poetsch, A. Elastase Digests: New Ammunition for Shotgun Membrane Proteomics. *Molecular & Cellular Proteomics* (2008).

124. Dau, T., Gupta, K., Berger, I. & Rappsilber, J. Sequential Digestion with Trypsin and Elastase in Cross-Linking/Mass Spectrometry. *bioRxiv* (2018).

125. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six Alternative Proteases for Mass Spectrometry–Based Proteomics beyond Trypsin. *Nature Protocols* **11,** 993–1006 (2016).

126. Vandermarliere, E., Mueller, M. & Martens, L. Getting Intimate with Trypsin, the Leading Protease in Proteomics. *Mass Spectrometry Reviews* **32,** 453–465 (2013).

127. Hustoft, H. K., Malerod, H., Wilson, S. R., Reubsaet, L., Lundanes, E. & Greibrokk, T. A Critical Review of Trypsin Digestion for LC-MS Based Proteomics. *Integrative Proteomics* (2012).

128. Burkhart, J. M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A. & Zahedi, R. P. Systematic and Quantitative Comparison of Digest Efficiency and Specificity Reveals the Impact of Trypsin Quality on MS-Based Proteomics. *Journal of Proteomics* **75,** 1454–1462 (2012).

129. Tsiatsiani, L. & Heck, A. J. R. Proteomics beyond Trypsin. *The FEBS Journal* **282,** 2612–2626 (2015).

130. Zhang, X. Less Is More: Membrane Protein Digestion Beyond Urea–Trypsin Solution for Next-Level Proteomics. *Molecular & Cellular Proteomics* **14,** 2441–2453 (2015).

131. Pralow, A., Hoffmann, M., Nguyen-Khuong, T., Rapp, E. & Reichl, U. Improvement of the Glycoproteomic Toolbox with the Discovery of a Unique C-Terminal Cleavage Specificity of Flavastacin for N-Glycosylated Asparagine. *Scientific Reports* **7,** 11419 (2017).

132. Kesimer, M. & Sheehan, J. K. in *Mucins: Methods and Protocols* (eds McGuckin, M. A. & Thornton, D. J.) 67–79 (Humana Press, Totowa, NJ, 2012).

133. Yang, W., Ao, M., Hu, Y., Li, Q. K. & Zhang, H. Mapping the O-Glycoproteome Using Site-Specific Extraction of O-Linked Glycopeptides (EXoO). *Molecular Systems Biology* **14,** e8486 (2018).

134. Ma, C. *et al.* Comprehensive N- and O-Glycosylation Mapping of Human Coagulation Factor V. *Journal of Thrombosis and Haemostasis* **n/a.**

135. Genovis Inc. *Analysis of O-Linked Glycosylation - Genovis - SmartEnzymes* Genovis. https://www.genovis.com/applications/analysis-of-o-linked-glycosylation/ (2020).

136. Riley, N. M., Malaker, S. A. & Bertozzi, C. R. Electron-Based Dissociation Is Needed for O-Glycopeptides Derived from OpeRATOR Proteolysis. *Analytical Chemistry* **92,** 14878–14884 (2020).

137. Zhang, H., Li, X.-j., Martin, D. B. & Aebersold, R. Identification and Quantification of N-Linked Glycoproteins Using Hydrazide Chemistry, Stable Isotope Labeling and Mass Spectrometry. *Nature Biotechnology* **21,** 660–666 (2003).

138. Sun, S. *et al.* Comprehensive Analysis of Protein Glycosylation by Solid-Phase Extraction of N-Linked Glycans and Glycosite-Containing Peptides. *Nature Biotechnology* **34,** 84 (2015).

139. Stadlmann, J., Hoi, D. M., Taubenschmid, J., Mechtler, K. & Penninger, J. M. Analysis of PNGase F-Resistant N-Glycopeptides Using SugarQb for Proteome Discoverer 2.1 Reveals Cryptic Substrate Specificities. *PROTEOMICS* **18,** 1700436 (2018).

140. Weng, Y., Sui, Z., Jiang, H., Shan, Y., Chen, L., Zhang, S., Zhang, L. & Zhang, Y. Releasing N-Glycan from Peptide N-Terminus by N-Terminal Succinylation Assisted Enzymatic Deglycosylation. *Scientific Reports* **5,** 9770 (1 2015).

141. Tarentino, A. L. & Maley, F. Purification and Properties of an Endo-β-N-Acetylglucosaminidase from Streptomyces Griseus. *Journal of Biological Chemistry* **249,** 811–817 (1974).

142. Hsieh, P., Rosner, M. R. & Robbins, P. W. Selective Cleavage by Endo-Beta-N-Acetylglucosaminidase H at Individual Glycosylation Sites of Sindbis Virion Envelope Glycoproteins. *Journal of Biological Chemistry* **258,** 2555–2561 (1983).

143. Parsons, T. B. *et al.* Optimal Synthetic Glycosylation of a Therapeutic Antibody. *Angewandte Chemie International Edition* **55,** 2361–2367 (2016).

144. New England BioLabs Inc. *Glycoproteomics Technical Guide* 56 pp. (2018).

145. New England Biolabs Inc. *Glycoproteomics Understanding Protein Modifications* 20 pp. (2017).

146. Hägglund, P., Matthiesen, R., Elortza, F., Højrup, P., Roepstorff, P., Jensen, O. N. & Bunkenborg, J. An Enzymatic Deglycosylation Scheme Enabling Identification of Core Fucosylated N-Glycans and O-Glycosylation Site Mapping of Human Plasma Proteins. *Journal of Proteome Research* **6,** 3021–3031 (2007).

147. Donnelly, D. P. *et al.* Best Practices and Benchmarks for Intact Protein Analysis for Top-down Mass Spectrometry. *Nature Methods* **16,** 587–594 (7 2019).

148. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for Micro-Purification, Enrichment, Pre-Fractionation and Storage of Peptides for Proteomics Using StageTips. *Nature Protocols* **2,** 1896–1906 (8 2007).

149. Berven, F. S., Ahmad, R., Clauser, K. R. & Carr, S. A. Optimizing Performance of Glycopeptide Capture for Plasma Proteomics. *Journal of Proteome Research* **9,** 1706–1715 (2010).

150. Li, J., Li, X., Guo, Z., Yu, L., Zou, L. & Liang, X. Click Maltose as an Alternative to Reverse Phase Material for Desalting Glycopeptides. *Analyst* **136,** 4075–4082 (2011).

151. Chin, E. T. & Papac, D. I. The Use of a Porous Graphitic Carbon Column for Desalting Hydrophilic Peptides Prior to Matrix-Assisted Laser Desorption/ Ionization Time-of-Flight Mass Spectrometry. *Analytical Biochemistry* **273,** 179–185 (1999).

152. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, Encapsulated Proteomic-Sample Processing Applied to Copy-Number Estimation in Eukaryotic Cells. *Nature Methods* **11,** 319–324 (3 2014).

153. Hughes, C. S., Moggridge, S., Müller, T., Sorensen, P. H., Morin, G. B. & Krijgsveld, J. Single-Pot, Solid-Phase-Enhanced Sample Preparation for Proteomics Experiments. *Nature Protocols* **14,** 68–85 (1 2019).

154. Waas, M., Pereckas, M., Jones Lipinski, R. A., Ashwood, C. & Gundry, R. L. SP2: Rapid and Automatable Contaminant Removal from Peptide Samples for Proteomic Analyses. *Journal of Proteome Research* **18,** 1644–1656 (2019).

155. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal Sample Preparation Method for Proteome Analysis. *Nature Methods* **6,** 359–362 (5 2009).

156. Wiśniewski, J. R. Filter Aided Sample Preparation – A Tutorial. *Analytica Chimica Acta* **1090,** 23–30 (2019).

157. Sielaff, M., Kuharev, J., Bohn, T., Hahlbrock, J., Bopp, T., Tenzer, S. & Distler, U. Evaluation of FASP, SP3, and iST Protocols for Proteomic Sample Preparation in the Low Microgram Range. *Journal of Proteome Research* **16,** 4060–4072 (2017).

158. Fischer, R. & Kessler, B. M. Gel-Aided Sample Preparation (GASP)—A Simplified Method for Gel-Assisted Proteomic Sample Generation from Protein Extracts and Intact Cells. *PROTEOMICS* **15,** 1224–1229 (2015).

159. Zougman, A., Selby, P. J. & Banks, R. E. Suspension Trapping (STrap) Sample Preparation Method for Bottom-up Proteomics Analysis. *PROTEOMICS* **14,** 1006–1000 (2014).

160. Ludwig, K. R., Schroll, M. M. & Hummon, A. B. Comparison of In-Solution, FASP, and S-Trap Based Digestion Methods for Bottom-Up Proteomic Studies. *Journal of Proteome Research* **17,** 2480–2490 (2018).

161. Kuljanin, M., Dieters-Castator, D. Z., Hess, D. A., Postovit, L.-M. & Lajoie, G. A. Comparison of Sample Preparation Techniques for Large-Scale Proteomics. *PROTEOMICS* **17,** 1600337 (2017).

162. Ducret, A., Oostveen, I. V., Eng, J. K., Yates, J. R. & Aebersold, R. High Throughput Protein Characterization by Automated Reverse-Phase Chromatography/Electrospray Tandem Mass Spectrometry. *Protein Science* **7,** 706–719 (1998).

163. Meiring, H. D., van der Heeft, E., ten Hove, G. J. & de Jong, A. P. J. M. Nanoscale LC–MS(n): Technical Design and Applications to Peptide and Protein Analysis. *Journal of Separation Science* **25,** 557–568 (2002).

164. Ritorto, M. S., Cook, K., Tyagi, K., Pedrioli, P. G. A. & Trost, M. Hydrophilic Strong Anion Exchange (hSAX) Chromatography for Highly Orthogonal Peptide Separation of Complex Proteomes. *Journal of Proteome Research* **12,** 2449–2457 (2013).

165. Ye, M., Zou, H., Liu, Z. & Ni, J. Separation of Peptides by Strong Cation-Exchange Capillary Electrochromatography. *Journal of Chromatography A* **869,** 385–394 (2000).

166. Yang, F., Shen, Y., Camp, D. G. & Smith, R. D. High-pH Reversed-Phase Chromatography with Fraction Concatenation for 2D Proteomic Analysis. *Expert Review of Proteomics* **9,** 129–134 (2012).

167. Boersema, P. J., Mohammed, S. & Heck, A. J. R. Hydrophilic Interaction Liquid Chromatography (HILIC) in Proteomics. *Analytical and Bioanalytical Chemistry* **391,** 151–159 (2008).

168. Krijgsveld, J., Gauci, S., Dormeyer, W. & Heck, A. J. R. In-Gel Isoelectric Focusing of Peptides as a Tool for Improved Protein Identification. *Journal of Proteome Research* **5,** 1721–1730 (2006).

169. Černý, M., Skalák, J., Cerna, H. & Brzobohatý, B. Advances in Purification and Separation of Posttranslationally Modified Proteins. *Journal of Proteomics. Special Issue: Posttranslational Protein Modifications in Biology and Medicine* **92,** 2–27 (2013).

170. Zhang, Y., Zhang, C., Jiang, H., Yang, P. & Lu, H. Fishing the PTM Proteome with Chemical Approaches Using Functional Solid Phases. *Chemical Society Reviews* **44,** 8260–8287 (2015).

171. Humphrey, S. J., Karayel, O., James, D. E. & Mann, M. High-Throughput and High-Sensitivity Phosphoproteomics with the EasyPhos Platform. *Nature Protocols* **13,** 1897–1916 (9 2018).

172. Stavenhagen, K., Hinneburg, H., Thaysen-Andersen, M., Hartmann, L., Silva, D. V., Fuchser, J., Kaspar, S., Rapp, E., Seeberger, P. H. & Kolarich, D. Quantitative Mapping of Glycoprotein Micro-Heterogeneity and Macro-Heterogeneity: An Evaluation of Mass Spectrometry Signal Strengths Using Synthetic Peptides and Glycopeptides. *Journal of Mass Spectrometry* **48,** 627–639 (2013).

173. Ongay, S., Boichenko, A., Govorukhina, N. & Bischoff, R. Glycopeptide Enrichment and Separation for Protein Glycosylation Analysis. *Journal of Separation Science* **35,** 2341–2372 (2012).

174. Yamamoto, S., Kinoshita, M. & Suzuki, S. Current Landscape of Protein Glycosylation Analysis and Recent Progress toward a Novel Paradigm of Glycoscience Research. *Journal of Pharmaceutical and Biomedical Analysis. Review Issue 2016* **130,** 273–300 (2016).

175. Tissot, B., North, S. J., Ceroni, A., Pang, P.-C., Panico, M., Rosati, F., Capone, A., Haslam, S. M., Dell, A. & Morris, H. R. Glycoproteomics: Past, Present and Future. *FEBS Letters* **583,** 1728–1735 (2009).

176. Yang, Y., Franc, V. & Heck, A. J. R. Glycoproteomics: A Balance between High-Throughput and In-Depth Analysis. *Trends in Biotechnology* **35,** 598–609 (2017).

177. Yu, A., Zhao, J., Peng, W., Banazadeh, A., Williamson, S. D., Goli, M., Huang, Y. & Mechref, Y. Advances in Mass Spectrometry-Based Glycoproteomics. *ELECTROPHORESIS* **0** (2018).

178. Shajahan, A., Heiss, C., Ishihara, M. & Azadi, P. Glycomic and Glycoproteomic Analysis of Glycoproteins—a Tutorial. *Analytical and Bioanalytical Chemistry* **409,** 4483–4505 (2017).

179. Riley, N. M., Bertozzi, C. R. & Pitteri, S. J. A Pragmatic Guide to Enrichment Strategies for Mass Spectrometry-Based Glycoproteomics. *Molecular & Cellular Proteomics* (2020).

180. Alley, W. R., Mann, B. F. & Novotny, M. V. High-Sensitivity Analytical Approaches for the Structural Characterization of Glycoproteins. *Chemical Reviews* **113,** 2668–2732 (2013).

181. Xue, Y., Xie, J., Fang, P., Yao, J., Yan, G., Shen, H. & Yang, P. Study on Behaviors and Performances of Universal N -Glycopeptide Enrichment Methods. *Analyst* **143,** 1870–1880 (2018).

182. Xiao, H., Chen, W., Smeekens, J. M. & Wu, R. An Enrichment Method Based on Synergistic and Reversible Covalent Interactions for Large-Scale Analysis of Glycoproteins. *Nature Communications* **9,** 1692 (2018).

183. Rowe, L., Khoury, G. E. & Lowe, C. R. A Benzoboroxole-Based Affinity Ligand for Glycoprotein Purification at Physiological pH. *Journal of Molecular Recognition* **29,** 232–238 (2016).

184. Alam, M. A., Arora, K., Gurrapu, S., Jonnalagadda, S. K., Nelson, G. L., Kiprof, P., Jonnalagadda, S. C. & Mereddy, V. R. Synthesis and Evaluation of Functionalized Benzoboroxoles as Potential Anti-Tuberculosis Agents. *Tetrahedron* **72,** 3795–3801 (2016).

185. Mereddy, G. R., Chakradhar, A., Rutkoski, R. M. & Jonnalagadda, S. C. Benzoboroxoles: Synthesis and Applications in Medicinal Chemistry. *Journal of Organometallic Chemistry. Organometallic Chemistry of Boranes and Carboranes (to Be Dedicated to Professor Narayan Hosmane)* **865,** 12–22 (2018).

186. Di Palma, S., Hennrich, M. L., Heck, A. J. R. & Mohammed, S. Recent Advances in Peptide Separation by Multidimensional Liquid Chromatography for Proteome Analysis. *Journal of Proteomics* **75,** 3791–3813 (2012).

187. Stavenhagen, K., Hinneburg, H., Kolarich, D. & Wuhrer, M. in *High-Throughput Glycomics and Glycoproteomics: Methods and Protocols* (eds Lauc, G. & Wuhrer, M.) 109–119 (Springer New York, New York, NY, 2017).

188. Woo, C. M., Iavarone, A. T., Spiciarich, D. R., Palaniappan, K. K. & Bertozzi, C. R. Isotope-Targeted Glycoproteomics (IsoTaG): A Mass-Independent Platform for Intact N- and O-Glycopeptide Discovery and Analysis. *Nature Methods* **12,** 561–567 (2015).

189. Woo, C. M., Felix, A., Byrd, W. E., Zuegel, D. K., Ishihara, M., Azadi, P., Iavarone, A. T., Pitteri, S. J. & Bertozzi, C. R. Development of IsoTaG, a Chemical Glycoproteomics Technique for Profiling Intact N- and O-Glycopeptides from Whole Cell Proteomes. *Journal of Proteome Research* **16,** 1706–1718 (2017).

190. Woo, C. M., Lund, P. J., Huang, A. C., Davis, M. M., Bertozzi, C. R. & Pitteri, S. Mapping and Quantification of over 2,000 O-Linked Glycopeptides in Activated Human T Cells with Isotope-Targeted Glycoproteomics (IsoTaG). *Molecular & Cellular Proteomics,* mcp.RA117.000261 (2018).

191. Almaraz, R. T., Aich, U., Khanna, H. S., Tan, E., Bhattacharya, R., Shah, S. & Yarema, K. J. Metabolic Oligosaccharide Engineering with N-Acyl Functionalized ManNAc Analogs: Cytotoxicity, Metabolic Flux, and Glycan-Display Considerations. *Biotechnology and Bioengineering* **109,** 992–1006 (2012).

192. Qin, W. *et al.* Artificial Cysteine S-Glycosylation Induced by Per-O-Acetylated Unnatural Monosaccharides during Metabolic Glycan Labeling. *Angewandte Chemie International Edition* **57,** 1817–1820 (2018).

193. Darabedian, N., Yang, B., Ding, R., Cutolo, G., Zaro, B. W., Woo, C. M. & Pratt, M. R. O-Acetylated Chemical Reporters of Glycosylation Can Display Metabolism-Dependent Background Labeling of Proteins but Are Generally Reliable Tools for the Identification of Glycoproteins. *Frontiers in Chemistry* **8** (2020).

194. Qin, K., Zhang, H., Zhao, Z. & Chen, X. Protein S-Glyco-Modification through an Elimination–Addition Mechanism. *Journal of the American Chemical Society* (2020).

195. Hao, Y., Fan, X., Shi, Y., Zhang, C., Sun, D.-e., Qin, K., Qin, W., Zhou, W. & Chen, X. Next-Generation Unnatural Monosaccharides Reveal That ESRRB O-GlcNAcylation Regulates Pluripotency of Mouse Embryonic Stem Cells. *Nature Communications* **10,** 4065 (2019).

196. Simm, S., Einloft, J., Mirus, O. & Schleiff, E. 50 Years of Amino Acid Hydrophobicity Scales: Revisiting the Capacity for Peptide Classification. *Biological Research* **49,** 31 (2016).

197. Zolg, D. P. *et al.* Building ProteomeTools Based on a Complete Synthetic Human Proteome. *Nature Methods* **14,** 259–262 (2017).

198. Hennrich, M. *Expansion of the Toolbox to Decipher the (Phospho)Proteome* (Utrecht University, 2012).

199. Bahr, U., Pfenninger, A., Karas, M. & Stahl, B. High-Sensitivity Analysis of Neutral Underivatized Oligosaccharides by Nanoelectrospray Mass Spectrometry. *Analytical Chemistry* **69,** 4530–4535 (1997).

200. Bian, Y. *et al.* Robust, Reproducible and Quantitative Analysis of Thousands of Proteomes by Micro-Flow LC–MS/MS. *Nature Communications* **11,** 1–12 (1 2020).

201. Zauner, G., Koeleman, C. A. M., Deelder, A. M. & Wuhrer, M. Protein Glycosylation Analysis by HILIC-LC-MS of Proteinase K-Generated N- and O-Glycopeptides. *Journal of Separation Science* **33,** 903–910 (2010).

202. Alley, W. R., Mechref, Y. & Novotny, M. V. Use of Activated Graphitized Carbon Chips for Liquid Chromatography/Mass Spectrometric and Tandem Mass Spectrometric Analysis of Tryptic Glycopeptides. *Rapid Communications in Mass Spectrometry* **23,** 495–505 (2009).

203. Nwosu, C. C., Huang, J., Aldredge, D. L., Strum, J. S., Hua, S., Seipert, R. R. & Lebrilla, C. B. In-Gel Nonspecific Proteolysis for Elucidating Glycoproteins: A Method for Targeted Protein-Specific Glycosylation Analysis in Complex Protein Mixtures. *Analytical Chemistry* **85,** 956–963 (2013).

204. Wilm, M. Principles of Electrospray Ionization. *Molecular & Cellular Proteomics* **10** (2011).

205. Wilm, M. S. & Mann, M. Electrospray and Taylor-Cone Theory, Dole's Beam of Macromolecules at Last? *International Journal of Mass Spectrometry and Ion Processes* **136,** 167–180 (1994).

206. Iribarne, J. V. & Thomson, B. A. On the Evaporation of Small Ions from Charged Droplets. *The Journal of Chemical Physics* **64,** 2287–2294 (1976).

207. Thomson, B. A. & Iribarne, J. V. Field Induced Ion Evaporation from Liquid Surfaces at Atmospheric Pressure. *The Journal of Chemical Physics* **71,** 4451–4463 (1979).

208. Dole, M., Mack, L. L., Hines, R. L., Mobley, R. C., Ferguson, L. D. & Alice, M. B. Molecular Beams of Macroions. *The Journal of Chemical Physics* **49,** 2240–2249 (1968).

209. Nguyen, S. & Fenn, J. B. Gas-Phase Ions of Solute Species from Charged Droplets of Solutions. *Proceedings of the National Academy of Sciences* **104,** 1111–1117 (2007).

210. Fernandez de la Mora, J. Electrospray Ionization of Large Multiply Charged Species Proceeds via Dole's Charged Residue Mechanism. *Analytica Chimica Acta* **406,** 93–104 (2000).

211. Hahne, H., Pachl, F., Ruprecht, B., Maier, S. K., Klaeger, S., Helm, D., Médard, G., Wilm, M., Lemeer, S. & Kuster, B. DMSO Enhances Electrospray Response, Boosting Sensitivity of Proteomic Experiments. *Nature Methods* **10,** 989–991 (10 2013).

212. Yu, P., Hahne, H., Wilhelm, M. & Kuster, B. Ethylene Glycol Improves Electrospray Ionization Efficiency in Bottom-up Proteomics. *Analytical and Bioanalytical Chemistry* **409,** 1049–1057 (2017).

213. Wilm, M. & Mann, M. Analytical Properties of the Nanoelectrospray Ion Source. *Analytical Chemistry* **68,** 1–8 (1996).

214. Lin, C.-w., Haeuptle, M. A. & Aebi, M. Supercharging Reagent for Enhanced Liquid Chromatographic Separation and Charging of Sialylated and High Molecular Weight Glycopeptides for NanoHPLC-ESI-MS/MS Analysis. *Analytical Chemistry* **88,** 8484–8494 (2016).

215. Kjeldsen, F., Giessing, A. M. B., Ingrell, C. R. & Jensen, O. N. Peptide Sequencing and Characterization of Post-Translational Modifications by Enhanced Ion-Charging and Liquid Chromatography Electron-Transfer Dissociation Tandem Mass Spectrometry. *Analytical Chemistry* **79,** 9243–9252 (2007).

216. Nair, H. & Wysocki, V. H. Are Peptides without Basic Residues Protonated Primarily at the Amino Terminus? *International Journal of Mass Spectrometry and Ion Processes. Polyatomic Ion-Surface Interactions* **174,** 95–100 (1998).

217. Schnier, P. D., Gross, D. S. & Williams, E. R. On the Maximum Charge State and Proton Transfer Reactivity of Peptide and Protein Ions Formed by Electrospray Ionization. *Journal of the American Society for Mass Spectrometry* **6,** 1086–1097 (1995).

218. Harvey, D. J. Fragmentation of Negative Ions from Carbohydrates: Part 3. Fragmentation of Hybrid and Complex N-Linked Glycans. *Journal of the American Society for Mass Spectrometry* **16,** 647–659 (2005).

219. Harvey, D. J. in *Analysis of Protein Post-Translational Modifications by Mass Spectrometry* (eds Griffiths, J. R. & Unwin, R. D.) 89–159 (John Wiley & Sons, Inc., 2016).

220. Nwosu, C. C., Strum, J. S., An, H. J. & Lebrilla, C. B. Enhanced Detection and Identification of Glycopeptides in Negative Ion Mode Mass Spectrometry. *Analytical Chemistry* **82,** 9654–9662 (2010).

221. Mann, M. & Kelleher, N. L. Precision Proteomics: The Case for High Resolution and High Mass Accuracy. *Proceedings of the National Academy of Sciences* **105,** 18132–18138 (2008).

222. Zubarev, R. A., Håkansson, P. & Sundqvist, B. Accuracy Requirements for Peptide Characterization by Monoisotopic Molecular Mass Measurements. *Analytical Chemistry* **68,** 4060–4063 (1996).

223. Xian, F., Hendrickson, C. L. & Marshall, A. G. High Resolution Mass Spectrometry. *Analytical Chemistry* **84,** 708–719 (2012).

224. De Hoffmann, E. & Stroobant, V. *Mass Spectrometry: Principles and Applications* 3 edition. 502 pp. (Wiley-Interscience, 2013).

225. Gross, J. H. *Mass Spectrometry: A Textbook* 3rd ed. (Springer International Publishing, 2017).

226. Savaryn, J. P., Toby, T. K. & Kelleher, N. L. A Researcher's Guide to Mass Spectrometry-Based Proteomics. *PROTEOMICS* **16,** 2435–2443 (2016).

227. Tabb, D. L., Shah, M. B., Strader, M. B., Connelly, H. M., Hettich, R. L. & Hurst, G. B. Determination of Peptide and Protein Ion Charge States by Fourier Transformation of Isotope-Resolved Mass Spectra. *Journal of the American Society for Mass Spectrometry* **17,** 903–915 (2006).

228. Douglas, D. J., Frank, A. J. & Mao, D. Linear Ion Traps in Mass Spectrometry. *Mass Spectrometry Reviews* **24,** 1–29 (2005).

229. Allen, J. S. The Detection of Single Positive Ions, Electrons and Photons by a Secondary Electron Multiplier. *Physical Review* **55,** 966–971 (1939).

230. Allen, J. S. An Improved Electron Multiplier Particle Counter. *Review of Scientific Instruments* **18,** 739–749 (1947).

231. Olsen, J. V. *et al.* A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed*. *Molecular & cellular proteomics : MCP* (2009).

232. Miller, P. E. & Denton, M. B. The Quadrupole Mass Filter: Basic Operating Concepts. *Journal of Chemical Education* **63,** 617 (1986).

233. Zubarev, R. A. & Makarov, A. Orbitrap Mass Spectrometry. *Analytical Chemistry* **85,** 5288–5296 (2013).

234. Olsen, J. V., de Godoy, L. M. F., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S. & Mann, M. Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-Trap. *Molecular & Cellular Proteomics* **4,** 2010–2021 (2005).

235. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S. & Mann, M. Higher-Energy C-Trap Dissociation for Peptide Modification Analysis. *Nature Methods* **4,** 709–712 (2007).

236. Murray, K. K., Boyd, R. K., Eberlin, M. N., Langley, G. J., Li, L. & Naito, Y. Definitions of Terms Relating to Mass Spectrometry (IUPAC Recommendations 2013). *Pure and Applied Chemistry* **85,** 1515–1609 (2013).

237. Roepstorff, P. & Fohlman, J. Letter to the Editors. *Biomedical Mass Spectrometry* **11,** 601–601 (1984).

238. Biemann, K. Contributions of Mass Spectrometry to Peptide and Protein Structure. *Biomedical & Environmental Mass Spectrometry* **16,** 99–111 (1988).

239. Johnson, R. S., Martin, S. A., Biemann, K., Stults, J. T. & Watson, J. T. Novel Fragmentation Process of Peptides by Collision-Induced Decomposition in a Tandem Mass Spectrometer: Differentiation of Leucine and Isoleucine. *Analytical Chemistry* **59,** 2621–2625 (1987).

240. Johnson, R. S., Martin, S. A. & Biemann, K. Collision-Induced Fragmentation of (M + H)+ Ions of Peptides. Side Chain Specific Sequence Ions. *International Journal of Mass Spectrometry and Ion Processes* **86,** 137–154 (1988).

241. Xiao, Y., Vecchi, M. M. & Wen, D. Distinguishing between Leucine and Isoleucine by Integrated LC–MS Analysis Using an Orbitrap Fusion Mass Spectrometer. *Analytical Chemistry* **88,** 10757–10766 (2016).

242. Zhokhov, S. S., Kovalyov, S. V., Samgina, T. Y. & Lebedev, A. T. An EThcD-Based Method for Discrimination of Leucine and Isoleucine Residues in Tryptic Peptides. *Journal of The American Society for Mass Spectrometry* **28,** 1600–1611 (2017).

243. Paizs, B. & Suhai, S. Fragmentation Pathways of Protonated Peptides. *Mass Spectrometry Reviews* **24,** 508–548 (2005).

244. Paizs, B., Szlávik, Z., Lendvay, G., Vékey, K. & Suhai, S. Formation of A2+ Ions of Protonated Peptides. An Ab Initio Study. *Rapid Communications in Mass Spectrometry* **14,** 746–755 (2000).

245. Papayannopoulos, I. A. The Interpretation of Collision-Induced Dissociation Tandem Mass Spectra of Peptides. *Mass Spectrometry Reviews* **14,** 49–73 (1995).

246. Falick, A. M., Hines, W. M., Medzihradszky, K. F., Baldwin, M. A. & Gibson, B. W. Low-Mass Ions Produced from Peptides by High-Energy Collision-Induced Dissociation in Tandem Mass Spectrometry. *Journal of the American Society for Mass Spectrometry* **4,** 882–893 (1993).

247. Steen, H., Küster, B., Fernandez, M., Pandey, A. & Mann, M. Detection of Tyrosine Phosphorylated Peptides by Precursor Ion Scanning Quadrupole TOF Mass Spectrometry in Positive Ion Mode. *Analytical Chemistry* **73,** 1440–1448 (2001).

248. Toprak, U. H., Gillet, L. C., Maiolica, A., Navarro, P., Leitner, A. & Aebersold, R. Conserved Peptide Fragmentation as a Benchmarking Tool for Mass Spectrometers and a Discriminating Feature for Targeted Proteomics. *Molecular & Cellular Proteomics* **13,** 2056–2071 (2014).

249. Cooks, R. G. Special Feature: Historical. Collision-Induced Dissociation: Readings and Commentary. *Journal of Mass Spectrometry* **30,** 1215–1221 (1995).

250. Mitchell Wells, J. & McLuckey, S. A. in *Methods in Enzymology* 148–185 (Academic Press, 2005).

251. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S. & Hauer, C. R. Protein Sequencing by Tandem Mass Spectrometry. *Proceedings of the National Academy of Sciences* **83,** 6233–6237 (1986).

252. Morgan, J., Greenberg, A. & Liebman, J. F. Paradigms and Paradoxes: O- and N-Protonated Amides, Stabilization Energy, and Resonance Energy. *Structural Chemistry* **23,** 197–199 (2012).

253. Haeffner, F., Merle, J. K. & Irikura, K. K. N-Protonated Isomers as Gateways to Peptide Ion Fragmentation. *Journal of the American Society for Mass Spectrometry* **22,** 2222–2231 (2011).

254. Bythell, B. J., Suhai, S., Somogyi, Á. & Paizs, B. Proton-Driven Amide Bond-Cleavage Pathways of Gas-Phase Peptide Ions Lacking Mobile Protons. *Journal of the American Chemical Society* **131,** 14057–14065 (2009).

255. Wysocki, V. H., Tsaprailis, G., Smith, L. L. & Breci, L. A. Mobile and Localized Protons: A Framework for Understanding Peptide Dissociation. *Journal of Mass Spectrometry* **35,** 1399–1406 (2000).

256. Dongré, A. R., Jones, J. L., Somogyi, Á. & Wysocki, V. H. Influence of Peptide Composition, Gas-Phase Basicity, and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model. *Journal of the American Chemical Society* **118,** 8365–8374 (1996).

257. Jones, J. L., Dongre, A. R., Somogyi, A. & Wysocki, V. H. Sequence Dependence of Peptide Fragmentation Efficiency Curves Determined by Electrospray Ionization/Surface-Induced Dissociation Mass Spectrometry. *Journal of the American Chemical Society* **116,** 8368–8369 (1994).

258. Yalcin, T., Khouw, C., Csizmadia, I. G., Peterson, M. R. & Harrison, A. G. Why Are B Ions Stable Species in Peptide Spectra? *Journal of the American Society for Mass Spectrometry* **6,** 1165–1174 (1995).

259. Shen, Y., Nemati, R., Wang, L. & Yao, X. Determining Linear Free Energy Relationships in Peptide Fragmentation Using Derivatization and Targeted Mass Spectrometry. *Analytical Chemistry* **90,** 1587–1594 (2018).

260. Paizs, B. & Suhai, S. Theoretical Study of the Main Fragmentation Pathways for Protonated Glycylglycine. *Rapid Communications in Mass Spectrometry* **15,** 651–663 (2001).

261. Ballard, K. D. & Gaskell, S. J. Dehydration of Peptide [M + H]+ Ions in the Gas Phase. *Journal of the American Society for Mass Spectrometry* **4,** 477–481 (1993).

262. Harrison, A. G. Fragmentation Reactions of Protonated Peptides Containing Glutamine or Glutamic Acid. *Journal of Mass Spectrometry* **38,** 174–187 (2003).

263. Csonka, I. P., Paizs, B., Lendvay, G. & Suhai, S. Proton Mobility and Main Fragmentation Pathways of Protonated Lysylglycine. *Rapid Communications in Mass Spectrometry* **15,** 1457–1472 (2001).

264. Ballard, K. D. & Gaskell, S. J. Sequential Mass Spectrometry Applied to the Study of the Formation of "Internal" Fragment Ions of Protonated Peptides. *International Journal of Mass Spectrometry and Ion Processes* **111,** 173–189 (1991).

265. Ambihapathy, K., Yalcin, T., Leung, H.-W. & Harrison, A. G. Pathways to Immonium Ions in the Fragmentation of Protonated Peptides. *Journal of Mass Spectrometry* **32,** 209–215 (1997).

266. Yalcin, T., Csizmadia, I. G., Peterson, M. R. & Harrison, A. G. The Structure and Fragmentation of Bn (N>=3) Ions in Peptide Spectra. *Journal of the American Society for Mass Spectrometry* **7,** 233–242 (1996).

267. Yang, Y.-H., Lee, K., Jang, K.-S., Kim, Y.-G., Park, S.-H., Lee, C.-S. & Kim, B.-G. Low Mass Cutoff Evasion with Qz Value Optimization in Ion Trap. *Analytical Biochemistry* **387,** 133–135 (2009).

268. Diedrich, J. K., Pinto, A. F. M. & Yates, J. R. Energy Dependence of HCD on Peptide Fragmentation: Stepped Collisional Energy Finds the Sweet Spot. *Journal of the American Society for Mass Spectrometry* **24,** 1690–1699 (2013).

269. Dieterich, D. C., Link, A. J., Graumann, J., Tirrell, D. A. & Schuman, E. M. Selective Identification of Newly Synthesized Proteins in Mammalian Cells Using Bioorthogonal Noncanonical Amino Acid Tagging (BONCAT). *Proceedings of the National Academy of Sciences* **103,** 9482–9487 (2006).

270. Nagaraj, N., D'Souza, R. C. J., Cox, J., Olsen, J. V. & Mann, M. Feasibility of Large-Scale Phosphoproteomics with Higher Energy Collisional Dissociation Fragmentation. *Journal of Proteome Research* **9,** 6786–6794 (2010).

271. Nagaraj, N., D'Souza, R. C. J., Cox, J., Olsen, J. V. & Mann, M. Correction to Feasibility of Large-Scale Phosphoproteomics with Higher Energy Collisional Dissociation Fragmentation. *Journal of Proteome Research* **11,** 3506–3508 (2012).

272. Frese, C. K., Altelaar, A. F. M., Hennrich, M. L., Nolting, D., Zeller, M., Griep-Raming, J., Heck, A. J. R. & Mohammed, S. Improved Peptide Identification by Targeted Fragmentation Using CID, HCD and ETD on an LTQ-Orbitrap Velos. *Journal of Proteome Research* **10,** 2377–2388 (2011).

273. Scientific, T. F. PSB 104 - Normalized Collision Energy Technology, 2.

274. Orsburn, B. *News in Proteomics Research: Normalized Collision Energy Calculation for Q Exactive* News in Proteomics Research. `http://proteomicsnews.blogspot.com/2014/06/normalized-collision-energy-calculation.html` (2020).

275. Hinneburg, H., Stavenhagen, K., Schweiger-Hufnagel, U., Pengelley, S., Jabs, W., Seeberger, P. H., Silva, D. V., Wuhrer, M. & Kolarich, D. The Art of Destruction: Optimizing Collision Energies in Quadrupole-Time of Flight (Q-TOF) Instruments for Glycopeptide-Based Glycoproteomics. *Journal of The American Society for Mass Spectrometry* **27,** 507–519 (2016).

276. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and Protein Sequence Analysis by Electron Transfer Dissociation Mass Spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 9528–9533 (2004).

277. Riley, N. M. & Coon, J. J. The Role of Electron Transfer Dissociation in Modern Proteomics. *Analytical Chemistry* **90,** 40–64 (2018).

278. Martens, J., Berden, G. & Oomens, J. Structures of Fluoranthene Reagent Anions Used in Electron Transfer Dissociation and Proton Transfer Reaction Tandem Mass Spectrometry. *Analytical Chemistry* **88,** 6126–6129 (2016).

279. Zhurov, K. O., Fornelli, L., Wodrich, M. D., Laskay, Ü. A. & Tsybin, Y. O. Principles of Electron Capture and Transfer Dissociation Mass Spectrometry Applied to Peptide and Protein Structure Analysis. *Chemical Society Reviews* **42,** 5014–5030 (2013).

280. Kempkes, L.J.M. *Reaction Mechanisms of Collision and Electron Induced Peptide Dissociation Revealed by Ion Spectroscopy* (Radboud University Nijmegen, 2019).

281. Tureček, F. & Julian, R. R. Peptide Radicals and Cation Radicals in the Gas Phase. *Chemical Reviews* **113,** 6691–6733 (2013).

282. Anusiewicz, I., Skurski, P. & Simons, J. Refinements to the Utah–Washington Mechanism of Electron Capture Dissociation. *The Journal of Physical Chemistry B* **118,** 7892–7901 (2014).

283. Zubarev, R. A., Kelleher, N. L. & McLafferty, F. W. Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *Journal of the American Chemical Society* **120,** 3265–3266 (1998).

284. Sawicka, A., Skurski, P., Hudgins, R. R. & Simons, J. Model Calculations Relevant to Disulfide Bond Cleavage via Electron Capture Influenced by Positively Charged Groups. *The Journal of Physical Chemistry B* **107,** 13505–13511 (2003).

285. Syrstad, E. A. & Tureček, F. Toward a General Mechanism of Electron Capture Dissociation. *Journal of the American Society for Mass Spectrometry* **16,** 208–224 (2005).

286. Martens, J., Grzetic, J., Berden, G. & Oomens, J. Structural Identification of Electron Transfer Dissociation Products in Mass Spectrometry Using Infrared Ion Spectroscopy. *Nature Communications* **7,** 1–7 (1 2016).

287. Kempkes, L. J. M., Martens, J., Berden, G. & Oomens, J. Spectroscopic Characterization of an Extensive Set of C-Type Peptide Fragment Ions Formed by Electron Transfer Dissociation Suggests Exclusive Formation of Amide Isomers. *The Journal of Physical Chemistry Letters* **9,** 6404–6411 (2018).

288. Asakawa, D., Yamashita, A., Kawai, S., Takeuchi, T. & Wada, Y. N–Cα Bond Cleavage of Zinc-Polyhistidine Complexes in Electron Transfer Dissociation Mediated by Zwitterion Formation: Experimental Evidence and Theoretical Analysis of the Utah–Washington Model. *The Journal of Physical Chemistry B* **120,** 891–901 (2016).

289. Świerszcz, I., Skurski, P. & Simons, J. Dipole and Coulomb Forces in Electron Capture Dissociation and Electron Transfer Dissociation Mass Spectroscopy. *The Journal of Physical Chemistry A* **116,** 1828–1837 (2012).

290. O'Connor, P. B., Lin, C., Cournoyer, J. J., Pittman, J. L., Belyayev, M. & Budnik, B. A. Long-Lived Electron Capture Dissociation Product Ions Experience Radical Migration via Hydrogen Abstraction. *Journal of the American Society for Mass Spectrometry* **17,** 576–585 (2006).

291. Swaney, D. L., McAlister, G. C., Wirtala, M., Schwartz, J. C., Syka, J. E. P. & Coon, J. J. Supplemental Activation Method for High-Efficiency Electron-Transfer Dissociation of Doubly Protonated Peptide Precursors. *Analytical Chemistry* **79,** 477–485 (2007).

292. Frese, C. *Development and Application of Novel Electron Transfer Dissociation-Based Technologies for Proteomics* (Utrecht University, 2013).

293. Ledvina, A. R., McAlister, G. C., Gardner, M. W., Smith, S. I., Madsen, J. A., Schwartz, J. C., Stafford, G. C., Syka, J. E. P., Brodbelt, J. S. & Coon, J. J. Infrared Photoactivation Reduces Peptide Folding and Hydrogen-Atom Migration Following ETD Tandem Mass Spectrometry. *Angewandte Chemie International Edition* **48,** 8526–8528 (2009).

294. Sun, R.-X. *et al.* Improved Peptide Identification for Proteomic Analysis Based on Comprehensive Characterization of Electron Transfer Dissociation Spectra. *Journal of Proteome Research* **9,** 6354–6367 (2010).

295. Hamidane, H. B., Chiappe, D., Hartmer, R., Vorobyev, A., Moniatte, M. & Tsybin, Y. O. Electron Capture and Transfer Dissociation: Peptide Structure Analysis at Different Ion Internal Energy Levels. *Journal of the American Society for Mass Spectrometry* **20,** 567–575 (2009).

296. Good, D. M., Wirtala, M., McAlister, G. C. & Coon, J. J. Performance Characteristics of Electron Transfer Dissociation Mass Spectrometry. *Molecular & cellular proteomics : MCP* (2007).

297. Pitteri, S. J., Chrisman, P. A. & McLuckey, S. A. Electron-Transfer Ion/Ion Reactions of Doubly Protonated Peptides: Effect of Elevated Bath Gas Temperature. *Analytical chemistry* (2005).

298. Rose, C. M., Rush, M. J. P., Riley, N. M., Merrill, A. E., Kwiecien, N. W., Holden, D. D., Mullen, C., Westphall, M. S. & Coon, J. J. A Calibration Routine for Efficient ETD in Large-Scale Proteomics. *Journal of The American Society for Mass Spectrometry* **26,** 1848–1857 (2015).

299. Molina, H., Matthiesen, R., Kandasamy, K. & Pandey, A. Comprehensive Comparison of Collision Induced Dissociation and Electron Transfer Dissociation. *Analytical Chemistry* **80,** 4825–4835 (2008).

300. Van den Toorn, H. W. P., Mohammed, S., Gouw, J. W., van Breukelen, B. & Heck, A. J. R. Targeted SCX Based Peptide Fractionation for Optimal Sequencing by Collision Induced, and Electron Transfer Dissociation. *Journal of Proteomics and Bioinformatics,* 379–388 (2008).

301. Mikesh, L. M., Ueberheide, B., Chi, A., Coon, J. J., Syka, J. E. P., Shabanowitz, J. & Hunt, D. F. The Utility of ETD Mass Spectrometry in Proteomic Analysis. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. Posttranslational Modifications in Proteomics* **1764,** 1811–1822 (2006).

302. Wiesner, J., Premsler, T. & Sickmann, A. Application of Electron Transfer Dissociation (ETD) for the Analysis of Posttranslational Modifications. *PROTEOMICS* **8,** 4466–4483 (2008).

303. Kim, M.-S. & Pandey, A. Electron Transfer Dissociation Mass Spectrometry in Proteomics. *PROTEOMICS* **12,** 530–542 (2012).

304. Fornelli, L., Damoc, E., Thomas, P. M., Kelleher, N. L., Aizikov, K., Denisov, E., Makarov, A. & Tsybin, Y. O. Analysis of Intact Monoclonal Antibody IgG1 by Electron Transfer Dissociation Orbitrap FTMS. *Molecular & Cellular Proteomics* **11,** 1758–1767 (2012).

305. 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. *International Journal of Proteomics* **2012,** e560391 (2012).

306. Frese, C. K., Altelaar, A. F. M., van den Toorn, H., Nolting, D., Griep-Raming, J., Heck, A. J. R. & Mohammed, S. Toward Full Peptide Sequence Coverage by Dual Fragmentation Combining Electron-Transfer and Higher-Energy Collision Dissociation Tandem Mass Spectrometry. *Analytical Chemistry* **84,** 9668–9673 (2012).

307. Frese, C. K., Zhou, H., Taus, T., Altelaar, A. F. M., Mechtler, K., Heck, A. J. R. & Mohammed, S. Unambiguous Phosphosite Localization Using Electron-Transfer/Higher-Energy Collision Dissociation (EThcD). *Journal of Proteome Research* **12,** 1520–1525 (2013).

308. Yu, J., Schorlemer, M., Gomez Toledo, A., Pett, C., Sihlbom, C., Larson, G., Westerlind, U. & Nilsson, J. Distinctive MS/MS Fragmentation Pathways of Glycopeptide-Generated Oxonium Ions Provide Evidence of the Glycan Structure. *Chemistry – A European Journal* **22,** 1114–1124 (2016).

309. Michalski, A., Damoc, E., Hauschild, J.-P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M. & Horning, S. Mass Spectrometry-Based Proteomics Using Q Exactive, a High-Performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Molecular & Cellular Proteomics* (2011).

310. Michalski, A. *et al.* Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes. *Molecular & Cellular Proteomics* **11** (2012).

311. Stahl, D. C., Swiderek, K. M., Davis, M. T. & Lee, T. D. Data-Controlled Automation of Liquid Chromatography/Tandem Mass Spectrometry Analysis of Peptide Mixtures. *Journal of the American Society for Mass Spectrometry* **7,** 532–540 (1996).

312. Davis, S., Charles, P. D., He, L., Mowlds, P., Kessler, B. M. & Fischer, R. Expanding Proteome Coverage with CHarge Ordered Parallel Ion aNalysis (CHOPIN) Combined with Broad Specificity Proteolysis. *Journal of Proteome Research* **16,** 1288–1299 (2017).

313. Vincent, C. E., Potts, G. K., Ulbrich, A., Westphall, M. S., Atwood, J. A., Coon, J. J. & Weatherly, D. B. Segmentation of Precursor Mass Range Using "Tiling" Approach Increases Peptide Identifications for MS1-Based Label-Free Quantification. *Analytical Chemistry* **85,** 2825–2832 (2013).

314. Meier, F., Geyer, P. E., Winter, S. V., Cox, J. & Mann, M. BoxCar Acquisition Method Enables Single-Shot Proteomics at a Depth of 10,000 Proteins in 100 Minutes. *Nature Methods* **15,** 440–448 (6 2018).

315. Erickson, B. K., Mintseris, J., Schweppe, D. K., Navarrete-Perea, J., Erickson, A. R., Nusinow, D. P., Paulo, J. A. & Gygi, S. P. Active Instrument Engagement Combined with a Real-Time Database Search for Improved Performance of Sample Multiplexing Workflows. *Journal of Proteome Research* **18,** 1299–1306 (2019).

316. Schweppe, D. K., Eng, J. K., Yu, Q., Bailey, D., Rad, R., Navarrete-Perea, J., Huttlin, E. L., Erickson, B. K., Paulo, J. A. & Gygi, S. P. Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics. *Journal of Proteome Research* **19,** 2026–2034 (2020).

317. Wichmann, C., Meier, F., Winter, S. V., Brunner, A.-D., Cox, J. & Mann, M. MaxQuant.Live Enables Global Targeting of More Than 25,000 Peptides. *Molecular & Cellular Proteomics* **18,** 982–994 (2019).

318. Picotti, P. & Aebersold, R. Selected Reaction Monitoring–Based Proteomics: Workflows, Potential, Pitfalls and Future Directions. *Nature Methods* **9,** 555–566 (6 2012).

319. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. *Molecular & Cellular Proteomics* (2012).

320. Bruderer, R., Bernhardt, O. M., Gandhi, T., Xuan, Y., Sondermann, J., Schmidt, M., Gomez-Varela, D. & Reiter, L. Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Molecular & Cellular Proteomics* **16,** 2296–2309 (2017).

321. Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R. & Aebersold, R. Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics* **11** (2012).

322. Ye, Z., Mao, Y., Clausen, H. & Vakhrushev, S. Y. Glyco-DIA: A Method for Quantitative O -Glycoproteomics with in Silico-Boosted Glycopeptide Libraries. *Nature Methods* **16,** 902–910 (9 2019).

323. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 Eliminates Ratio Distortion in Isobaric Multiplexed Quantitative Proteomics. *Nature Methods* **8,** 937–940 (11 2011).

324. Kao, A., Chiu, C.-l., Vellucci, D., Yang, Y., Patel, V. R., Guan, S., Randall, A., Baldi, P., Rychnovsky, S. D. & Huang, L. Development of a Novel Cross-Linking Strategy for Fast and Accurate Identification of Cross-Linked Peptides of Protein Complexes. *Molecular & Cellular Proteomics* **10** (2011).

325. Both, P. *et al.* Discrimination of Epimeric Glycans and Glycopeptides Using IM-MS and Its Potential for Carbohydrate Sequencing. *Nature Chemistry* **6,** 65–74 (2014).

326. Gray, C. J. *et al.* Bottom-Up Elucidation of Glycosidic Bond Stereochemistry. *Analytical Chemistry* **89,** 4540–4549 (2017).

327. Struwe, W. B. & Harvey, D. J. in *Ion Mobility-Mass Spectrometry : Methods and Protocols* (eds Paglia, G. & Astarita, G.) 203–219 (Springer US, New York, NY, 2020).

328. Kolli, V., Schumacher, K. N. & Dodds, E. D. Ion Mobility-Resolved Collision-Induced Dissociation and Electron Transfer Dissociation of N-Glycopeptides: Gathering Orthogonal Connectivity Information from a Single Mass-Selected Precursor Ion Population. *Analyst* **142,** 4691–4702 (2017).

329. Hinneburg, H., Hofmann, J., B. Struwe, W., Thader, A., Altmann, F., Silva, D. V., H. Seeberger, P., Pagel, K. & Kolarich, D. Distinguishing N -Acetylneuraminic Acid Linkage Isomers on Glycopeptides by Ion Mobility-Mass Spectrometry. *Chemical Communications* **52,** 4381–4384 (2016).

330. Struwe, W. B., Pagel, K., Benesch, J. L. P., Harvey, D. J. & Campbell, M. P. GlycoMob: An Ion Mobility-Mass Spectrometry Collision Cross Section Database for Glycomics. *Glycoconjugate Journal* **33,** 399–404 (2016).

331. Gray, C. J., Compagnon, I. & Flitsch, S. L. Mass Spectrometry Hybridized with Gas-Phase InfraRed Spectroscopy for Glycan Sequencing. *Current Opinion in Structural Biology* **62,** 121–131 (2020).

332. Mucha, E., Stuckmann, A., Marianski, M., Struwe, W. B., Meijer, G. & Pagel, K. In-Depth Structural Analysis of Glycans in the Gas Phase. *Chemical Science* **10,** 1272–1284 (2019).

333. Mucha, E., Marianski, M., Xu, F.-F., Thomas, D. A., Meijer, G., von Helden, G., Seeberger, P. H. & Pagel, K. Unravelling the Structure of Glycosyl Cations via Cold-Ion Infrared Spectroscopy. *Nature Communications* **9,** 1–5 (1 2018).

334. Wu, X. *et al.* Imaging Single Glycans. *Nature* **582,** 375–378 (7812 2020).

335. Perkins, D. N., Pappin, D. J. C., Creasy, D. M. & Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *ELECTROPHORESIS* **20,** 3551–3567 (1999).

336. Dančík, V., Addona, T. A., Clauser, K. R., Vath, J. E. & Pevzner, P. A. De Novo Peptide Sequencing via Tandem Mass Spectrometry. *Journal of Computational Biology* **6,** 327–342 (1999).

337. Seidler, J., Zinn, N., Boehm, M. E. & Lehmann, W. D. De Novo Sequencing of Peptides by MS/MS. *PROTEOMICS* **10,** 634–649 (2010).

338. Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S. & MacCoss, M. J. Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. *Analytical Chemistry* **78,** 5678–5684 (2006).

339. Listgarten, J. & Emili, A. Statistical and Computational Methods for Comparative Proteomic Profiling Using Liquid Chromatography-Tandem Mass Spectrometry. *Molecular & Cellular Proteomics* **4,** 419–434 (2005).

340. Renard, B. Y., Kirchner, M., Monigatti, F., Ivanov, A. R., Rappsilber, J., Winter, D., Steen, J. A. J., Hamprecht, F. A. & Steen, H. When Less Can Yield More – Computational Preprocessing of MS/MS Spectra for Peptide Identification. *PROTEOMICS* **9,** 4978–4984 (2009).

341. Gentzel, M., Köcher, T., Ponnusamy, S. & Wilm, M. Preprocessing of Tandem Mass Spectrometric Data to Support Automatic Protein Identification. *PROTEOMICS* **3,** 1597–1610 (2003).

342. Bern, M., Kil, Y. J. & Becker, C. Byonic: Advanced Peptide and Protein Identification Software. *Current Protocols in Bioinformatics* **40,** 13.20.1–13.20.14 (2012).

343. Liu, M.-Q. *et al.* pGlyco 2.0 Enables Precision N-Glycoproteomics with Comprehensive Quality Control and One-Step Mass Spectrometry for Intact Glycopeptide Identification. *Nature Communications* **8,** 1–14 (2017).

344. Klein, J., Carvalho, L. & Zaia, J. Application of Network Smoothing to Glycan LC-MS Profiling. *Bioinformatics* **34,** 3511–3518 (2018).

345. Nesvizhskii, A. I. & Aebersold, R. Interpretation of Shotgun Proteomic Data: The Protein Inference Problem. *Molecular & Cellular Proteomics* **4,** 1419–1440 (2005).

346. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V. & Mann, M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research* **10,** 1794–1805 (2011).

347. Cox, J. & Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nature Biotechnology* **26,** 1367–1372 (12 2008).

348. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics. *Nature Methods* **14,** 513–520 (5 2017).

349. Gessulat, S. *et al.* Prosit: Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning. *Nature Methods* **16,** 509–518 (6 2019).

350. Tiwary, S., Levy, R., Gutenbrunner, P., Soto, F. S., Palaniappan, K. K., Deming, L., Berndl, M., Brant, A., Cimermancic, P. & Cox, J. High-Quality MS/MS Spectrum Prediction for Data-Dependent and Data-Independent Acquisition Data Analysis. *Nature Methods* **16,** 519–525 (6 2019).

351. Zolg, D. P., Wilhelm, M., Schmidt, T., Medard, G., Zerweck, J., Knaute, T., Wenschuh, H., Reimer, U., Schnatbaum, K. & Kuster, B. ProteomeTools: Systematic Characterization of 21 Post-Translational Protein Modifications by LC-MS/MS Using Synthetic Peptides. *Molecular & Cellular Proteomics,* mcp.TIR118.000783 (2018).

352. Nesvizhskii, A. I. Proteogenomics: Concepts, Applications and Computational Strategies. *Nature Methods* **11,** 1114–1125 (11 2014).

353. Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L. & Gygi, S. P. A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides. *Nature Biotechnology* **33,** 743–749 (7 2015).

354. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and Validation of Proteomic Data Generated by Tandem Mass Spectrometry. *Nature Methods* **4,** 787–797 (10 2007).

355. Nesvizhskii, A. I. A Survey of Computational Methods and Error Rate Estimation Procedures for Peptide and Protein Identification in Shotgun Proteomics. *Journal of Proteomics. Model Organism Proteomics* **73,** 2092–2123 (2010).

356. Elias, J. E. & Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nature Methods* **4,** 207–214 (3 2007).

357. Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *Journal of Proteome Research* **7,** 40–44 (2008).

358. Lakbub, J. C., Su, X., Zhu, Z., Patabandige, M. W., Hua, D., Go, E. P. & Desaire, H. Two New Tools for Glycopeptide Analysis Researchers: A Glycopeptide Decoy Generator and a Large Data Set of Assigned CID Spectra of Glycopeptides. *Journal of Proteome Research* **16,** 3002–3008 (2017).

359. Chalkley, R. J. & Clauser, K. R. Modification Site Localization Scoring: Strategies and Performance. *Molecular & Cellular Proteomics* **11,** 3–14 (2012).

360. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A Probability-Based Approach for High-Throughput Protein Phosphorylation Analysis and Site Localization. *Nature Biotechnology* **24,** 1285–1292 (10 2006).

361. Taus, T., Köcher, T., Pichler, P., Paschke, C., Schmidt, A., Henrich, C. & Mechtler, K. Universal and Confident Phosphorylation Site Localization Using phosphoRS. *Journal of Proteome Research* **10,** 5354–5362 (2011).

362. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P. & Mann, M. Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell* **127,** 635–648 (2006).

363. Savitski, M. M., Lemeer, S., Boesche, M., Lang, M., Mathieson, T., Bantscheff, M. & Kuster, B. Confident Phosphorylation Site Localization Using the Mascot Delta Score. *Molecular & Cellular Proteomics* **10** (2011).

364. Wiese, H., Kuhlmann, K., Wiese, S., Stoepel, N. S., Pawlas, M., Meyer, H. E., Stephan, C., Eisenacher, M., Drepper, F. & Warscheid, B. Comparison of Alternative MS/MS and Bioinformatics Approaches for Confident Phosphorylation Site Localization. *Journal of Proteome Research* **13,** 1128–1137 (2014).

365. Lee, D. C. H., Jones, A. R. & Hubbard, S. J. Computational Phosphoproteomics: From Identification to Localization. *PROTEOMICS* **15,** 950–963 (2015).

366. Ma, B. Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of the American Society for Mass Spectrometry* **26,** 1885–1894 (2015).

367. Frank, A. & Pevzner, P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Analytical Chemistry* **77,** 964–973 (2005).

368. Yang, H., Chi, H., Zeng, W.-F., Zhou, W.-J. & He, S.-M. pNovo 3: Precise de Novo Peptide Sequencing Using a Learning-to-Rank Framework. *Bioinformatics* **35,** i183–i190 (2019).

369. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A. & Lajoie, G. PEAKS: Powerful Software for Peptide de Novo Sequencing by Tandem Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **17,** 2337–2342 (2003).

370. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De Novo Peptide Sequencing by Deep Learning. *Proceedings of the National Academy of Sciences* **114,** 8247–8252 (2017).

371. Miller, S. E., Rizzo, A. I. & Waldbauer, J. R. Postnovo: Postprocessing Enables Accurate and FDR-Controlled de Novo Peptide Sequencing. *Journal of Proteome Research* **17,** 3671–3680 (2018).

372. Tran, N. H., Rahman, M. Z., He, L., Xin, L., Shan, B. & Li, M. Complete De Novo Assembly of Monoclonal Antibody Sequences. *Scientific Reports* **6,** 31730 (2016).

373. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative Mass Spectrometry in Proteomics: A Critical Review. *Analytical and Bioanalytical Chemistry* **389,** 1017–1031 (2007).

374. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative Mass Spectrometry in Proteomics: Critical Review Update from 2007 to the Present. *Analytical and Bioanalytical Chemistry* **404,** 939–965 (2012).

375. Ong, S.-E. & Mann, M. Mass Spectrometry–Based Proteomics Turns Quantitative. *Nature Chemical Biology* **1,** 252–262 (5 2005).

376. Ankney, J. A., Muneer, A. & Chen, X. Relative and Absolute Quantitation in Mass Spectrometry–Based Proteomics. *Annual Review of Analytical Chemistry* **11,** 49–77 (2016).

377. Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T. & Hamon, C. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Analytical Chemistry* **75,** 1895–1904 (2003).

378. McAlister, G. C., Nusinow, D. P., Jedrychowski, M. P., Wühr, M., Huttlin, E. L., Erickson, B. K., Rad, R., Haas, W. & Gygi, S. P. MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes. *Analytical Chemistry* **86,** 7150–7158 (2014).

379. Tyanova, S., Temu, T. & Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nature Protocols* **11,** 2301–2319 (12 2016).

380. Bubis, J. A., Levitsky, L. I., Ivanov, M. V., Tarasova, I. A. & Gorshkov, M. V. Comparative Evaluation of Label-Free Quantification Methods for Shotgun Proteomics. *Rapid Communications in Mass Spectrometry* **31,** 606–612 (2017).

381. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N. & Mann, M. Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics* **13,** 2513–2526 (2014).

382. Keilhauer, E. C., Hein, M. Y. & Mann, M. Accurate Protein Complex Retrieval by Affinity Enrichment Mass Spectrometry (AE-MS) Rather than Affinity Purification Mass Spectrometry (AP-MS). *Molecular & Cellular Proteomics* **14,** 120–135 (2015).

383. Al Shweiki, M. R., Mönchgesang, S., Majovsky, P., Thieme, D., Trutschel, D. & Hoehenwarter, W. Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. *Journal of Proteome Research* **16,** 1410–1424 (2017).

384. Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A "Proteomic Ruler" for Protein Copy Number and Concentration Estimation without Spike-in Standards. *Molecular & Cellular Proteomics* **13,** 3497–3506 (2014).

385. O'Reilly, F. J. & Rappsilber, J. Cross-Linking Mass Spectrometry: Methods and Applications in Structural, Molecular and Systems Biology. *Nature Structural & Molecular Biology* **25,** 1000–1008 (2018).

386. Yang, B. *et al.* Identification of Cross-Linked Peptides from Complex Samples. *Nature Methods* **9,** 904–906 (2012).

387. Schilling, B., Row, R. H., Gibson, B. W., Guo, X. & Young, M. M. MS2Assign, Automated Assignment and Nomenclature of Tandem Mass Spectra of Chemically Crosslinked Peptides. *Journal of the American Society for Mass Spectrometry* **14,** 834–850 (2003).

388. *Thermo Scientific - Crosslinking Technical Handbook* `http://tools.thermofisher.com/content/sfs/brochures/1602163-Crosslinking-Reagents-Handbook.pdf` (2019).

389. Fischer, L., Chen, Z. A. & Rappsilber, J. Quantitative Cross-Linking/Mass Spectrometry Using Isotope-Labelled Cross-Linkers. *Journal of Proteomics. Special Issue: New Horizons and Applications for Proteomics [EuPA 2012]* **88,** 120–128 (2013).

390. Preston, G. W. & Wilson, A. J. Photo-Induced Covalent Cross-Linking for the Analysis of Biomolecular Interactions. *Chemical Society Reviews* **42,** 3289–3301 (2013).

391. Steigenberger, B., Pieters, R. J., Heck, A. J. R. & Scheltema, R. A. PhoX: An IMAC-Enrichable Cross-Linking Reagent. *ACS Central Science* **5,** 1514–1522 (2019).

392. Stadlmeier, M., Runtsch, L. S., Streshnev, F., Wühr, M. & Carell, T. A Click-Chemistry-Based Enrichable Crosslinker for Structural and Protein Interaction Analysis by Mass Spectrometry. *ChemBioChem* **21,** 103–107 (2020).

393. Matzinger, M., Kandioller, W., Doppler, P., Heiss, E. H. & Mechtler, K. Fast and Highly Efficient Affinity Enrichment of Azide-A-DSBSO Cross-Linked Peptides. *Journal of Proteome Research* **19,** 2071–2079 (2020).

394. Frei, A. P. *et al.* Direct Identification of Ligand-Receptor Interactions on Living Cells and Tissues. *Nature Biotechnology* **30,** 997–1001 (10 2012).

395. Klykov, O., Steigenberger, B., Pektaş, S., Fasci, D., Heck, A. J. R. & Scheltema, R. A. Efficient and Robust Proteome-Wide Approaches for Cross-Linking Mass Spectrometry. *Nature Protocols* **13,** 2964–2990 (12 2018).

396. Fürsch, J., Kammer, K.-M., Kreft, S. G., Beck, M. & Stengel, F. Proteome-Wide Structural Probing of Low-Abundant Protein Interactions by Cross-Linking Mass Spectrometry. *Analytical Chemistry* **92,** 4016–4022 (2020).

397. Holding, A. N. XL-MS: Protein Cross-Linking Coupled with Mass Spectrometry. *Methods. Mass Spectrometry-Based Structural Biology* **89,** 54–63 (2015).

398. Chen, Z. A. & Rappsilber, J. Protein Dynamics in Solution by Quantitative Crosslinking/Mass Spectrometry. *Trends in Biochemical Sciences* **43,** 908–920 (2018).

399. Rappsilber, J. The Beginning of a Beautiful Friendship: Cross-Linking/Mass Spectrometry and Modelling of Proteins and Multi-Protein Complexes. *Journal of Structural Biology. Combining Computational Modeling with Sparse and Low-Resolution Data* **173,** 530–540 (2011).

400. Trnka, M. J., Baker, P. R., Robinson, P. J. J., Burlingame, A. L. & Chalkley, R. J. Matching Cross-Linked Peptide Spectra: Only as Good as the Worse Identification. *Molecular & Cellular Proteomics* **13,** 420–434 (2014).

401. Liu, F., Rijkers, D. T. S., Post, H. & Heck, A. J. R. Proteome-Wide Profiling of Protein Assemblies by Cross-Linking Mass Spectrometry. *Nature Methods* **12,** 1179–1184 (12 2015).

402. Kolbowski, L., Mendes, M. L. & Rappsilber, J. Optimizing the Parameters Governing the Fragmentation of Cross-Linked Peptides in a Tribrid Mass Spectrometer. *Analytical Chemistry* **89,** 5311–5318 (2017).

403. Zhao, B., Reilly, C. P., Davis, C., Matouschek, A. & Reilly, J. P. Use of Multiple Ion Fragmentation Methods to Identify Protein Cross-Links and Facilitate Comparison of Data Interpretation Algorithms. *Journal of Proteome Research* **19,** 2758–2771 (2020).

404. Sinz, A. Divide and Conquer: Cleavable Cross-Linkers to Study Protein Conformation and Protein–Protein Interactions. *Analytical and Bioanalytical Chemistry* **409,** 33–44 (2017).

405. Stieger, C. E., Doppler, P. & Mechtler, K. Optimized Fragmentation Improves the Identification of Peptides Cross-Linked by MS-Cleavable Reagents. *Journal of Proteome Research* **18,** 1363–1370 (2019).

406. Chen, Z.-L. *et al.* A High-Speed Search Engine pLink 2 with Systematic Evaluation for Proteome-Scale Identification of Cross-Linked Peptides. *Nature Communications* **10,** 3404 (2019).

407. Mendes, M. L. *et al. An Integrated Workflow for Crosslinking Mass Spectrometry* preprint (Biochemistry, 2018).

408. Hoopmann, M. R., Zelter, A., Johnson, R. S., Riffle, M., MacCoss, M. J., Davis, T. N. & Moritz, R. L. Kojak: Efficient Analysis of Chemically Cross-Linked Protein Complexes. *Journal of Proteome Research* **14,** 2190–2198 (2015).

409. Giese, S. H., Belsom, A., Sinn, L., Fischer, L. & Rappsilber, J. Noncovalently Associated Peptides Observed during Liquid Chromatography-Mass Spectrometry and Their Effect on Cross-Link Analyses. *Analytical Chemistry* **91,** 2678–2685 (2019).

410. Fischer, L. & Rappsilber, J. Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. *Analytical Chemistry* **89,** 3829–3833 (2017).

411. Walzthoeni, T., Claassen, M., Leitner, A., Herzog, F., Bohn, S., Förster, F., Beck, M. & Aebersold, R. False Discovery Rate Estimation for Cross-Linked Peptides Identified by Mass Spectrometry. *Nature Methods* **9,** 901–903 (9 2012).

412. Beveridge, R., Stadlmann, J., Penninger, J. M. & Mechtler, K. A Synthetic Peptide Library for Benchmarking Crosslinking-Mass Spectrometry Search Engines for Proteins and Protein Complexes. *Nature Communications* **11,** 1–9 (1 2020).

413. Chen, Z. A. & Rappsilber, J. Quantitative Cross-Linking/Mass Spectrometry to Elucidate Structural Changes in Proteins and Their Complexes. *Nature Protocols* **14,** 171–201 (1 2019).

414. Müller, F., Fischer, L., Chen, Z. A., Auchynnikava, T. & Rappsilber, J. On the Reproducibility of Label-Free Quantitative Cross-Linking/Mass Spectrometry. *Journal of the American Society for Mass Spectrometry* **29,** 405–412 (2018).

415. Degiacomi, M. T., Schmidt, C., Baldwin, A. J. & Benesch, J. L. P. Accommodating Protein Dynamics in the Modeling of Chemical Crosslinks. *Structure* **25,** 1751–1757.e5 (2017).

416. Larance, M. & Lamond, A. I. Multidimensional Proteomics for Cell Biology. *Nature Reviews Molecular Cell Biology* **16,** 269–280 (5 2015).

417. Aebersold, R. & Mann, M. Mass-Spectrometric Exploration of Proteome Structure and Function. *Nature* **537,** 347–355 (7620 2016).

418. Sinitcyn, P., Rudolph, J. D. & Cox, J. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annual Review of Biomedical Data Science* **1,** 207–234 (2018).

419. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M. & Cox, J. The Perseus Computational Platform for Comprehensive Analysis of (Prote)Omics Data. *Nature Methods* **13,** 731–740 (9 2016).

420. Choi, M., Chang, C.-Y., Clough, T., Broudy, D., Killeen, T., MacLean, B. & Vitek, O. MSstats: An R Package for Statistical Analysis of Quantitative Mass Spectrometry-Based Proteomic Experiments. *Bioinformatics* **30,** 2524–2526 (2014).

421. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2020).

422. Bludau, I. & Aebersold, R. Proteomic and Interactomic Insights into the Molecular Basis of Cell Functional Diversity. *Nature Reviews Molecular Cell Biology,* 1–14 (2020).

423. Schubert, O. T., Röst, H. L., Collins, B. C., Rosenberger, G. & Aebersold, R. Quantitative Proteomics: Challenges and Opportunities in Basic and Applied Research. *Nature Protocols* **12,** 1289–1294 (7 2017).

424. Lundberg, E. & Borner, G. H. H. Spatial Proteomics: A Powerful Discovery Tool for Cell Biology. *Nature Reviews Molecular Cell Biology* **20,** 285–302 (5 2019).

425. Lund-Johansen, F. *et al.* MetaMass, a Tool for Meta-Analysis of Subcellular Proteomics Data. *Nature Methods* **13,** 837–840 (10 2016).

426. Zhu, H., Tamura, T. & Hamachi, I. Chemical Proteomics for Subcellular Proteome Analysis. *Current Opinion in Chemical Biology. Omics* **48,** 1–7 (2019).

427. Schmidt, A., Forne, I. & Imhof, A. Bioinformatic Analysis of Proteomics Data. *BMC Systems Biology* **8,** S3 (2014).

428. Tipney, H. & Hunter, L. An Introduction to Effective Use of Enrichment Analysis Software. *Human Genomics* **4,** 202 (2010).

429. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Research* **37,** 1–13 (2009).

430. Huttlin, E. L. *et al.* Architecture of the Human Interactome Defines Protein Communities and Disease Networks. *Nature* **545,** 505–509 (7655 2017).

431. Trinkle-Mulcahy, L. Recent Advances in Proximity-Based Labeling Methods for Interactome Mapping. *F1000Research* **8,** 135 (2019).

432. Geri, J. B. *et al.* Microenvironment Mapping via Dexter Energy Transfer on Immune Cells. *Science* **367,** 1091–1097 (2020).

433. Keilhauer, E. C., Hein, M. Y. & Mann, M. Accurate Protein Complex Retrieval by Affinity Enrichment Mass Spectrometry (AE-MS) Rather than Affinity Purification Mass Spectrometry (AP-MS). *Molecular & Cellular Proteomics* **14,** 120–135 (2015).

434. Teo, G., Liu, G., Zhang, J., Nesvizhskii, A. I., Gingras, A.-C. & Choi, H. SAINTexpress: Improvements and Additional Features in Significance Analysis of INTeractome Software. *Journal of Proteomics. Special Issue: Can Proteomics Fill the Gap Between Genomics and Phenotypes?* **100,** 37–43 (2014).

435. Jäger, S. *et al.* Global Landscape of HIV–Human Protein Complexes. *Nature* **481,** 365–370 (7381 2012).

436. Drew, K., Lee, C., Huizar, R. L., Tu, F., Borgeson, B., McWhite, C. D., Ma, Y., Wallingford, J. B. & Marcotte, E. M. Integration of over 9,000 Mass Spectrometry Experiments Builds a Global Map of Human Protein Complexes. *Molecular Systems Biology* **13,** 932 (2017).

437. Gupta, S., Turan, D., Tavernier, J. & Martens, L. The Online Tabloid Proteome: An Annotated Database of Protein Associations. *Nucleic Acids Research* **46,** D581–D585 (2018).

438. Postic, G. *et al.* Probing Protein Interaction Networks by Combining MS-Based Proteomics and Structural Data Integration. *Journal of Proteome Research* (2020).

439. Gordon, D. E. *et al.* A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug Repurposing. *Nature,* 1–13 (2020).

440. Shu, Q. *et al.* Large-Scale Identification of N-Linked Intact Glycopeptides in Human Serum Using HILIC Enrichment and Spectral Library Search. *Molecular & Cellular Proteomics* **19,** 672–689 (2020).

441. Fang, P. *et al.* Multilayered N-Glycoproteome Profiling Reveals Highly Heterogeneous and Dysregulated Protein N-Glycosylation Related to Alzheimer's Disease. *Analytical Chemistry* **92,** 867–874 (2020).

442. Aguilar, H. A., Iliuk, A. B., Chen, I.-H. & Tao, W. A. Sequential Phosphoproteomics and N-Glycoproteomics of Plasma-Derived Extracellular Vesicles. *Nature Protocols* **15,** 161–180 (1 2020).

443. Huang, H., Lin, S., Garcia, B. A. & Zhao, Y. Quantitative Proteomic Analysis of Histone Modifications. *Chemical Reviews* **115,** 2376–2418 (2015).

444. Schwämmle, V., Aspalter, C.-M., Sidoli, S. & Jensen, O. N. Large Scale Analysis of Co-Existing Post-Translational Modifications in Histone Tails Reveals Global Fine Structure of Cross-Talk. *Molecular & Cellular Proteomics* **13,** 1855–1865 (2014).

445. Oughtred, R. *et al.* The BioGRID Interaction Database: 2019 Update. *Nucleic Acids Research* **47,** D529–D541 (2019).

446. Krug, K. *et al.* A Curated Resource for Phosphosite-Specific Signature Analysis. *Molecular & Cellular Proteomics* **18,** 576–593 (2019).

447. Abrahams, J. L., Taherzadeh, G., Jarvas, G., Guttman, A., Zhou, Y. & Campbell, M. P. Recent Advances in Glycoinformatic Platforms for Glycomics and Glycoproteomics. *Current Opinion in Structural Biology* **62,** 56–69 (2020).

448. Olsen, J. V. & Mann, M. Status of Large-Scale Analysis of Post-Translational Modifications by Mass Spectrometry. *Molecular & Cellular Proteomics* **12,** 3444–3452 (2013).

449. Pascovici, D., Wu, J. X., McKay, M. J., Joseph, C., Noor, Z., Kamath, K., Wu, Y., Ranganathan, S., Gupta, V. & Mirzaei, M. Clinically Relevant Post-Translational Modification Analyses—Maturing Workflows and Bioinformatics Tools. *International Journal of Molecular Sciences* **20,** 16 (1 2019).

450. Weigt, D., Hopf, C. & Médard, G. Studying Epigenetic Complexes and Their Inhibitors with the Proteomics Toolbox. *Clinical Epigenetics* (2016).

451. Kaur, U., Meng, H., Lui, F., Ma, R., Ogburn, R. N., Johnson, J. H. R., Fitzgerald, M. C. & Jones, L. M. Proteome-Wide Structural Biology: An Emerging Field for the Structural Analysis of Proteins on the Proteomic Scale. *Journal of Proteome Research* **17,** 3614–3627 (2018).

452. Leitner, A. Cross-Linking and Other Structural Proteomics Techniques: How Chemistry Is Enabling Mass Spectrometry Applications in Structural Biology. *Chemical Science* **7,** 4792–4803 (2016).

453. Politis, A., Stengel, F., Hall, Z., Hernández, H., Leitner, A., Walzthoeni, T., Robinson, C. V. & Aebersold, R. A Mass Spectrometry–Based Hybrid Method for Structural Modeling of Protein Complexes. *Nature Methods* **11,** 403–406 (4 2014).

454. McCafferty, C. L., Verbeke, E. J., Marcotte, E. M. & Taylor, D. W. Structural Biology in the Multi-Omics Era. *Journal of Chemical Information and Modeling* (2020).

455. Cianfrocco, M. A. & Kellogg, E. H. What Could Go Wrong? A Practical Guide to Single-Particle Cryo-EM: From Biochemistry to Atomic Models. *Journal of Chemical Information and Modeling* (2020).

456. Ho, C.-M., Li, X., Lai, M., Terwilliger, T. C., Beck, J. R., Wohlschlegel, J., Goldberg, D. E., Fitzpatrick, A. W. P. & Zhou, Z. H. Bottom-up Structural Proteomics: cryoEM of Protein Complexes Enriched from the Cellular Milieu. *Nature Methods* **17,** 79–85 (1 2020).

457. Masson, G. R. *et al.* Recommendations for Performing, Interpreting and Reporting Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS) Experiments. *Nature Methods* **16,** 595–602 (7 2019).

458. Bouyssié, D., Lesne, J., Locard-Paulet, M., Albigot, R., Burlet-Schiltz, O. & Marcoux, J. HDX-Viewer: Interactive 3D Visualization of Hydrogen–Deuterium Exchange Data. *Bioinformatics* **35,** 5331–5333 (2019).

459. Chavez, J. D. & Bruce, J. E. Chemical Cross-Linking with Mass Spectrometry: A Tool for Systems Structural Biology. *Current Opinion in Chemical Biology. Omics* **48,** 8–18 (2019).

460. De Souza, N. & Picotti, P. Mass Spectrometry Analysis of the Structural Proteome. *Current Opinion in Structural Biology. Folding and Binding · Proteins* **60,** 57–65 (2020).

461. Schopper, S., Kahraman, A., Leuenberger, P., Feng, Y., Piazza, I., Müller, O., Boersema, P. J. & Picotti, P. Measuring Protein Structural Changes on a Proteome-Wide Scale Using Limited Proteolysis-Coupled Mass Spectrometry. *Nature Protocols* **12,** 2391–2410 (11 2017).

462. Wu, H. & Kohler, J. Photocrosslinking Probes for Capture of Carbohydrate Interactions. *Current Opinion in Chemical Biology. Mechanistic Biology • Chemical Biophysics* **53,** 173–182 (2019).

463. Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S. & Crispin, M. Site-Specific Glycan Analysis of the SARS-CoV-2 Spike. *Science* (2020).

464. *Mass Spectrometry-Based Chemical Proteomics* 1st ed. (eds Tao, W. A. & Zhang, Y.) 448 pp. (Wiley, Hoboken, NJ, 2019).

465. Parker, C. G. & Pratt, M. R. Click Chemistry in Proteomic Investigations. *Cell* **180,** 605–632 (2020).

466. Drewes, G. & Knapp, S. Chemoproteomics and Chemical Probes for Target Discovery. *Trends in Biotechnology* (2018).

467. Schirle, M., Bantscheff, M. & Kuster, B. Mass Spectrometry-Based Proteomics in Preclinical Drug Discovery. *Chemistry & Biology* **19,** 72–84 (2012).

468. Schirle, M., Bantscheff, M. & Kuster, B. in *Proteomics for Biological Discovery* 247–263 (John Wiley & Sons, Ltd, 2019).

469. Chernobrovkin, A., Marin-Vicente, C., Visa, N. & Zubarev, R. A. Functional Identification of Target by Expression Proteomics (FITExP) Reveals Protein Targets and Highlights Mechanisms of Action of Small Molecule Drugs. *Scientific Reports* **5,** 1–9 (1 2015).

470. Gaetani, M. & Zubarev, R. A. in *Mass Spectrometry-Based Chemical Proteomics* (eds Tao, W. A. & Zhang, Y.) 1st ed., 257–266 (Wiley, 2019).

471. Wright, M. H. & Sieber, S. A. Chemical Proteomics Approaches for Identifying the Cellular Targets of Natural Products. *Natural Product Reports* **33,** 681–708 (2016).

472. Kallemeijn, W. W. *et al.* Novel Activity-Based Probes for Broad-Spectrum Profiling of Retaining β-Exoglucosidases In Situ and In Vivo. *Angewandte Chemie International Edition* **51,** 12529–12533 (2012).

473. Wu, L., Armstrong, Z., Schröder, S. P., de Boer, C., Artola, M., Aerts, J. M., Overkleeft, H. S. & Davies, G. J. An Overview of Activity-Based Probes for Glycosidases. *Current Opinion in Chemical Biology. Mechanistic Biology • Chemical Biophysics* **53,** 25–36 (2019).

474. Yu, S.-H., Wands, A. M. & Kohler, J. J. Photoaffinity Probes for Studying Carbohydrate Biology. *Journal of Carbohydrate Chemistry* **31,** 325–352 (2012).

475. Sakurai, K., Ozawa, S., Yamada, R., Yasui, T. & Mizuno, S. Comparison of the Reactivity of Carbohydrate Photoaffinity Probes with Different Photoreactive Groups. *ChemBioChem* **15,** 1399–1403 (2014).

476. Cruz, I. N., Barry, C. S., Kramer, H. B., Chuang, C. C., Lloyd, S., van der Spoel, A. C., Platt, F. M., Yang, M. & Davis, B. G. Glycomimetic Affinity-Enrichment Proteomics Identifies Partners for a Clinically-Utilized Iminosugar. *Chemical Science* **4,** 3442–3446 (2013).

477. Mateus, A., Kurzawa, N., Becher, I., Sridharan, S., Helm, D., Stein, F., Typas, A. & Savitski, M. M. Thermal Proteome Profiling for Interrogating Protein Interactions. *Molecular Systems Biology* **16,** e9232 (2020).

478. Piazza, I., Kochanowski, K., Cappelletti, V., Fuhrer, T., Noor, E., Sauer, U. & Picotti, P. A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication. *Cell* **172,** 358–372.e23 (2018).

479. Lomenick, B. *et al.* Target Identification Using Drug Affinity Responsive Target Stability (DARTS). *Proceedings of the National Academy of Sciences* **106,** 21984–21989 (2009).

480. Pai, M. Y., Lomenick, B., Hwang, H., Schiestl, R., McBride, W., Loo, J. A. & Huang, J. in *Chemical Biology: Methods and Protocols* (eds Hempel, J. E., Williams, C. H. & Hong, C. C.) 287–298 (Springer, New York, NY, 2015).

481. Bantscheff, M. *et al.* Chemoproteomics Profiling of HDAC Inhibitors Reveals Selective Targeting of HDAC Complexes. *Nature Biotechnology* **29,** 255–265 (2011).

482. Arrowsmith, C. H. *et al.* The Promise and Peril of Chemical Probes. *Nature Chemical Biology* **11,** 536–541 (2015).

483. Kleiner, P., Heydenreuter, W., Stahl, M., Korotkov, V. S. & Sieber, S. A. A Whole Proteome Inventory of Background Photocrosslinker Binding. *Angewandte Chemie International Edition* **56,** 1396–1401 (2017).

484. Kawamura, A., Hindi, S., Mihai, D. M., James, L. & Aminova, O. Binding Is Not Enough; Flexibility Is Needed for Photocrosslinking of Lck Kinase by Benzophenone Photoligands. *Bioorganic & medicinal chemistry* **16,** 8824–8829 (2008).

485. Klaeger, S. *et al.* The Target Landscape of Clinical Kinase Drugs. *Science* **358,** eaan4368 (2017).

486. Wishart, D. S. *et al.* DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Research* **46,** D1074–D1082 (2018).

487. Gaulton, A. *et al.* The ChEMBL Database in 2017. *Nucleic Acids Research* **45,** D945–D954 (2017).

488. Kim, S. *et al.* PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Research* **47,** D1102–D1109 (2019).

489. Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L. & Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Research* **44,** D1045–D1053 (2016).

490. Skuta, C., Popr, M., Muller, T., Jindrich, J., Kahle, M., Sedlak, D., Svozil, D. & Bartunek, P. Probes & Drugs Portal: An Interactive, Open Data Resource for Chemical Biology. *Nature Methods* **14,** 759–760 (8 2017).

491. Palaniappan, K. K. & Bertozzi, C. R. Chemical Glycoproteomics. *Chemical Reviews* (2016).

492. Liu, Y. *et al.* Multi-Omic Measurements of Heterogeneity in HeLa Cells across Laboratories. *Nature Biotechnology* **37,** 314–322 (3 2019).

493. Wang, D. *et al.* A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues. *Molecular Systems Biology* **15,** e8503 (2019).

494. Misra, B. B., Langefeld, C., Olivier, M. & Cox, L. A. Integrated Omics: Tools, Advances and Future Approaches. *Journal of Molecular Endocrinology* **62,** R21–R45 (2019).

495. Perez-Riverol, Y. *et al.* Discovering and Linking Public Omics Data Sets Using the Omics Discovery Index. *Nature Biotechnology* **35,** 406–409 (5 2017).

496. Martens, L. & Vizcaíno, J. A. A Golden Age for Working with Public Proteomics Data. *Trends in Biochemical Sciences* **42,** 333–341 (2017).

497. Vaudel, M., Verheggen, K., Csordas, A., Ræder, H., Berven, F. S., Martens, L., Vizcaíno, J. A. & Barsnes, H. Exploring the Potential of Public Proteomics Data. *PROTEOMICS* **16,** 214–225 (2016).

498. Zhang, B. *et al.* Proteogenomic Characterization of Human Colon and Rectal Cancer. *Nature* **513,** 382–387 (7518 2014).

499. Mun, D.-G. *et al.* Proteogenomic Characterization of Human Early-Onset Gastric Cancer. *Cancer Cell* **35,** 111–124.e10 (2019).

Part I.

Tandem mass spectrometry methods to study glycosylated proteins

Sólo el que ensaya lo absurdo es capaz de conquistar lo imposible.
In order to attain the impossible, one must attempt the absurd.

— Miguel de Unamuno, commentary on *Vida de D. Quijote y Sancho*

2

From tube to 3D model: *de novo* sequencing, glycosylation analysis and modelling of the anti-γH2AX murine antibody

Contents

2.1. Introduction . . . 110
2.1.1. γH2AX is a mark for DNA double-strand break . . . 110
2.1.2. Structure of monoclonal IgG antibodies . . . 112
2.1.3. *De novo* sequencing of monoclonal antibodies . . . 114
2.1.3.1. RNA sequencing of the hybridoma cell line . . . 114
2.1.3.2. *De novo* sequencing of purified antibodies by LC-MS/MS . . . 115
2.1.4. Glycosylation analysis of monoclonal antibodies . . . 117
2.1.5. Antibody molecular modelling . . . 118
2.2. Objective: Project description and analytical strategy . . . 120
2.3. Materials & methods . . . 122
2.3.1. General methods . . . 122
2.3.2. Intact denaturing protein LC-MS . . . 122
2.3.3. *De novo* antibody sequencing by MS/MS − Experimental . . . 123
2.3.3.1. Sample preparation for *de novo* sequencing . . . 123
2.3.3.2. Standard data acquisition − Q Exactive . . . 124
2.3.4. *De novo* antibody sequencing by MS/MS − data analysis . . . 125
2.3.4.1. Short description of the PEAKS software and its key modules . . . 125
2.3.4.2. Description of the *de novo* sequencing methodology by MS/MS using the ALPS assembler . . . 126

2.3.4.3. *De novo* sequencing and sequence validation using PeaksAB 2.0 . 128
2.3.5. Glycosylation analysis . . . 129
2.3.5.1. Sample preparation for glycosylation analysis . . . 129
2.3.5.2. Data acquisition for glycosylation analysis – Orbitrap Elite . . . 129
2.3.5.3. Data analysis for glycosylation analysis . . . 131
2.3.6. RNA sequencing . . . 131
2.3.7. Antibody Molecular Modelling . . . 131
2.3.7.1. Homology model of the variable region . . . 131
2.3.7.2. Data mining of the RCSB PDB . . . 132
2.3.7.3. Docking of the peptide antigen in the F_V antibody receptor . . . 132
2.3.7.4. Homology model for the entire anti-γH2Ax monoclonal antibody 133
2.4. Results . . . 135
2.4.1. Assessing *de novo* sequencing tools on the Herceptin test system . . . 135
2.4.1.1. Both ALPS and PeaksAB successfully assemble Herceptin's light chain . . . 136
2.4.1.2. PeaksAB outperforms ALPS for the Herceptin heavy chain assembly . . . 140
2.4.2. *De novo* sequencing of the anti-γH2AX mAb light chain . . . 143
2.4.2.1. Manual correction of the ALPS assembly allows successful sequencing . . . 143
2.4.2.2. PeaksAB has substantial difficulties with the CDR-L and CL regions . . . 148
2.4.3. Manual sequencing of the anti γ-H2Ax heavy chain . . . 150
2.4.3.1. Generation of a consensus template . . . 150
2.4.3.2. Correction of the consensus template in the variable domain . . 152
2.4.3.3. Investigating sequence coverage gaps in the constant domain . . 158
2.4.3.4. Comparison with the RNAseq and intact protein MS results . . 163
2.4.3.5. PeaksAB fails on most of the anti-γH2Ax heavy chain variable domain . . . 163
2.4.3.6. PTM analysis goes hand in hand with *de novo* sequencing . . . 166
2.4.4. Lessons from the *de novo* sequencing and discussion of its future . . . 168
2.4.5. Glycosylation analysis . . . 171
2.4.6. Molecular modelling of the antibody structure . . . 175
2.4.6.1. Homology modelling and antigen-antibody docking . . . 175
2.4.6.2. Bridging the gap: Combining structural and germline frequency data to explain antibody binding specificity . . . 180
2.4.6.3. Molecular modelling of the full anti-γH2AX mAb . . . 184
2.5. Outlook . . . 187
2.6. Conclusion . . . 188
2.7. Author contributions . . . 190
2.8. Acknowledgements . . . 190
2.9. Conflict of interest . . . 191
2.10. Confidentiality declaration . . . 191
2.11. Appendix . . . 192
2.11.1. Copy of code and scripts used in this chapter . . . 192
2.11.2. Advice for manual *de novo* sequencing by Bioinformatic Solutions . . . 193
2.11.3. RNA sequencing report from Absolute Antibody . . . 194
2.11.4. Antibody Homology Modelling – Supplementary information . . . 197
2.11.5. Full Haddock docking results − short peptide . . . 198
2.11.6. Full Haddock docking results − longer peptide . . . 202

2.12. References 205

Abstract

Monoclonal antibodies represent a ∼90 bn US$ market in diagnostics and biopharmaceutical research. Yet their development is hampered by technical difficulties of identifying the correct sequence, analysing their varying post-translational modifications − in particular the *N*-glycosylation pattern − and relating this information to properties of the antibody: from its stability and its immunogenicity up to the molecular understanding of antigen recognition. This is the case for the widely used anti-γH2AX murine monoclonal antibody. It recognises phosphorylation of H2AX at Ser 139 specifically, a well-known mark for DNA double-strand break (DSB) occurring in several cancers types. Its potential for diagnostics or targeted delivery is actively explored, yet the lack of sequence, glycosylation and structural information has prevented its fine tuning or humanisation and in extension hampered the use of this promising DSB biomarker to generate a diagnostic tool. In this case study, we illustrate how the sequence and a 3D homology model could be obtained from 20 µg of the commercial reagent. Using *de novo* tandem mass spectrometry methods, we were able to confidently sequence the antibody and further provide evidence for the glycosylation form G0F as well as other post-translational modifications. This information was used for homology modelling by combining well established software tools with structural data from the RCSB PDB database to generate a 3D model of the entire antibody. Based on this structure, we postulate a mechanism and rationale for the specific recognition of phosphorylated H2AX S139 by the antibody. This information was passed on to our collaborators for humanisation and the development of antibody fragments as preclinical candidates, enabling the exploitation of γH2AX as diagnostic marker. Along with our efforts, we highlight current possibilities, pitfalls and limitations in antibody protein research for academic laboratories and mass spectrometry facilities using both Herceptin and the anti-γH2AX antibody as examples. Conversely, the used state of the art experimental methodology and software tools are critically assessed on this real-world case study for antibody development.

Please note that parts of this chapter relating to the sequence of the murine anti-γH2AX antibody are confidential, see the confidentiality declaration in section 2.10.

2.1. Introduction

Monoclonal antibodies (mAbs) are glycoproteins used as biologics − pharmaceuticals from biotechnological origin. As of 6th January 2020, 89 antibodies have been clinically approved as biopharmaceuticals.[500] The vast majority is either of human origin or humanised. The humanisation process requires knowledge of sequence: loops that contribute to the highly specific antigen recognition are kept while other parts of the sequence that could lead to an immune reaction are changed to the human counterparts.[501] Knowing the amino acid sequence of the antibody is therefore a prerequisite for humanisation and thus allows to accelerate preclinical development.

Beyond the protein sequence, post-translational modifications such as *N*-glycosylation have a profound effect on the antibody structure and function. The need to understand the structure and function of the glycan portion of mAbs has increased in importance: first, any biologic destined for therapeutic use is required to insure a consistent structure with minimal changes between batches. Environmental factors in large-scale production can perturb the glycosylation process, leading to batch-to-batch variations. Secondly, different glycoforms for the same antibody have been reported to impact on the antigen recognition process. In fact, the removal of fucose from the glycan structures of mAbs improved their efficiency for antibody-dependent cellular cytotoxicity (ADCC) by nearly two orders of magnitude.[80]

Antibodies remain some of the best protein-protein interaction tools available for research, diagnostics and therapeutics. This project was initially started to track changes in glycosylation on a non-sequenced antibody against phosphorylated histone H2AX. This commercial reagent detects double-strand breaks relevant for tracking oncogenesis. The project evolved into a sequencing and modelling study to allow for humanisation of the antibody to enable its potential clinical use. The humanisation and preclinical development were conducted by Dr. Abirami Lakshminarayanan in the Department of Oncology, Churchill Hospital Oxford, and are not discussed in detail in this thesis.

2.1.1. γH2AX is a mark for DNA double-strand break

H2AX is a non-canonical variant of histone H2A required for checkpoint-mediated cell cycle arrest and DNA repair following double-stranded DNA breaks (DSB), it represents approximately 10% of the total H2A proteins.[502] Within minutes following DNA damage, H2AX is phosphorylated at Ser139 at sites in proximity of DNA damage by

PI3K-likes kinases including ATM, ATR and DNA-PK.[503–505] Depending of the source of the DNA damage, different kinases are responsible for H2AX S139 phosphorylation: DNA-PK upon cell death receptor activation, c-Jun N-terminal Kinase (JNK1) in response to UV-A irradiation, and p38 MAPK in response to serum starvation.[506–508]

Depending on cellular context, flagging out DNA double-stranded breaks by H2AX S139ph can lead to the activation of different pathways. In the DSB repair response, this phosphorylation mechanism is supposed to be in antiplay with *O*-GlcNAc glycosylation at Ser139[509] and controls the regulation of downstream DNA-damage response proteins such as MDC1, BRCA1, MRN, NBS1, RAD50, MRE11 and 53BP1 among others.[510,511]

Additionally, DNA-PK mediated phosphorylation of H2AX S139 has been reported to increase in response to apoptotic signals, H2AX being required for controlled DNA fragmentation. In this scenario, the interplay with H2AX Tyr142 phosphorylation is central. Upon DNA damage, and concurrent with phosphorylation of Ser139, Tyr142 is dephosphorylated at sites of DNA damage by recruiting EYA1 and EYA3 phosphatases. While phosphorylation at Ser139 facilitates the recruitment of DNA repair proteins and apoptotic proteins to sites of DNA damage, phosphorylation at Tyr142 determines which set of proteins are recruited. Phosphorylation of H2AX at Tyr142 inhibits the recruitment of DNA repair proteins and promotes binding of pro-apoptotic factors such as JNK1. Thus, it appears that the balance of H2AX Tyr142 phosphorylation and dephosphorylation provides a switch mechanism to determine cell fate after DNA damage.[512,513]

Being a convergence point of pathways relating to DNA damage − particularly in DSB, it is not surprising that phosphorylation of H2AX at Ser139 has been observed in many cancers. It is thus thought to be a potential hallmark for diagnostics. With the availability of selective monoclonal antibodies of murine and rabbit origin against phosphorylated H2AX Ser139ph, research in this field has expanded and the reagent is now routinely used in flow cytometry, immunohistochemistry and immunofluorescence.[504–506]

These studies have illustrated the potential of anti-γH2AX antibodies as promising diagnostic reagent for several cancers and senescence. Yet, improvement of these antibodies is lacking information about their sequence and glycosylation, preventing fine tuning or humanisation of these promising biomarkers. By carrying out *de novo* sequencing using a set of tandem mass spectrometry techniques, we were able to obtain the sequence of the broadly used anti-γH2AX antibody JBW301, determine that it has the glycosylation form G0F and suggest a 3D model for antigen recognition. This work enabled further preclinical development for a diagnostic biomarker for DNA double-strand break.

2.1.2. Structure of monoclonal IgG antibodies

Antibodies are glycoproteins of the immunoglobulin superfamily that bind specific antigens; the elucidation of their structural properties by Rodney R. Porter and Gerald M. Edelman was awarded the Nobel Prize in 1972. The isotype of the antibody is determined by the conserved part, 5 different types are known in placental mammals, IgA, IgD, IgE, IgG and IgM. Four subclasses of IgG (IgG1, IgG2, IgG3 and IgG4) are defined by differences in the Fc polypeptide, disulphide bonds, valency and antigen type.[501]

The typical Y-shaped structure of the IgG monomer is shown in Figure 2.1. It consists of two heavy and two light chains, each chain is composed of two structural domains for the light chain, or four for the heavy chain, called immunoglobulin domains and forming the characteristic immunoglobulin fold. Conceptually, the IgG monomer is divided in two parts: a shorter variable region (V_L and V_H1) and a larger constant part (C_L or C_H1-3). Historically, the division along the papain cleavage at the hinge in two fragments was frequently used: the F_{ab} (fragment, antigen-binding) and F_C (fragment, crystallisable) region. The antigen binds in the variable domain (F_V) at the tip of the Y-shaped molecule. The antigen binding site is also referred to as *hypervariable region*, it consists of six loops of β-strands − 3 on each light and heavy chain − the *complementarity determining regions* (CDRs). These CDRs and their 3D structure are unique for each antibody and are a result of both random mutations, alterations and the germline rearrangement.[515]

The vast diversity of the antibodies and their corresponding antigen binding sites (or *paratopes*) is a result of the genetic encoding of the variants of segments of the variable domain. These variants, V, D and J can be assembled during the maturation of the antibody producing plasma cells in a combinatorial fashion (see Figure 2.1); additional random mutations, base excisions and additions bring the potential range of combinatorial possibilities to an estimate of 10^{13}, with a diversity of the clone-defining CDR-H3 of 10^7 alone.[501]

This combinatorial diversity is further expanded by post-translational modifications, which fine-tune the immunogenicity.[515–517] Among those, *N*-glycosylation at Asn297 in the C_H2 domain has a major effect on the conformational dynamics and gives rise to allosteric structural effects which extend beyond the F_C domain into the F_{ab} region.[78] Furthermore, point mutations in the C_H2 domain[518] and C_H3 domain[519] have been reported to strongly affect the stability of the antibody.

Understanding these structural features is fundamental for this study; sequencing strategies of the variable domain rely on conserved sequence snippets from the V, D and

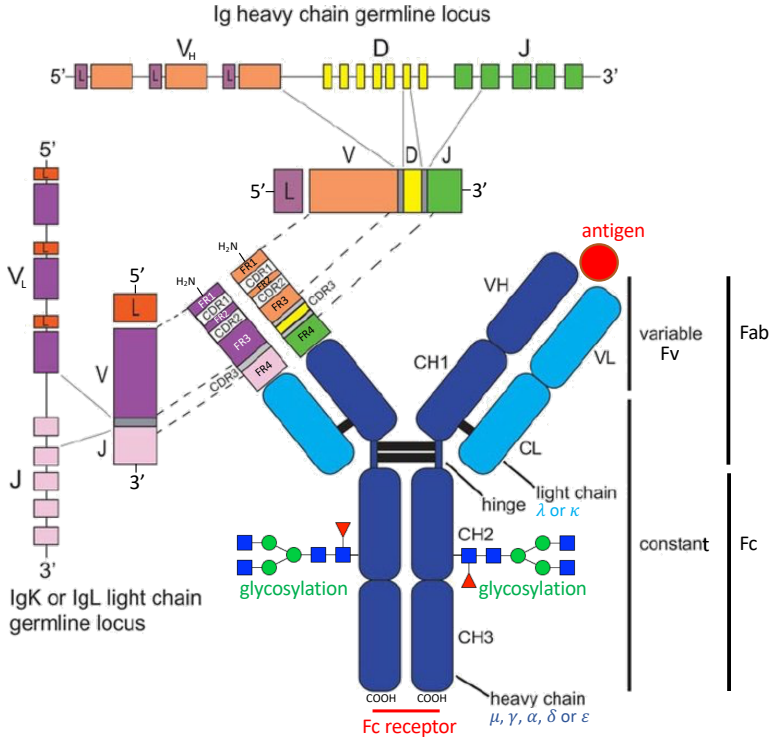


Figure 2.1. Structure of monoclonal IgG antibodies and their genetic encoding. The germline is the non-rearranged gene locus of the Variable (V), Diversity (D) and Joining (J) gene segments that encode for the variable region and constitute the antibody diversity. The configuration of the heavy chain locus is shown on the top, showing the tandem arrays of V, D and J segments that include multiple variants of each. Similarly, the light chain is encoded in a κ or λ locus, consisting of V and J segments. One variant of each segment is chosen in a stepwise rearrangement of the germline DNA while maturing to the antibody-producing plasma cells. Additional variability is introduced by somatic mutations, exonuclease digestion and randomised addition of non-template bases, particularly in the VDJ junctional region that will constitute the CDR3 loop, the region of the antibody heavy chain that often has the greatest impact on antigen recognition. The constant regions of both heavy and light chain are encoded by downstream exons, which are joined to the F_V part by mRNA splicing. The structural assembly of the four chains is further anchored by intra- and interchain disulfide bonds (the latter is highlighted in black). Finally, the immunogenic property and structural conformation is fine-tuned by post-translational *N*-glycosylation of the C_H2 domain (as well as occasionally *O*-glycosylation in the C_H1 domain) and editing of the *C*-terminus by carboxypeptidases, influencing Fc receptor recognition. Modified from[514] ©ASM Press.

J germline sequences to improve and occasionally guide *de novo* sequencing. Without an accurate prediction of the CDR loops, particularly CDR-H3, no reliable modelling project can be conducted. Finally, the key effect of glycosylation on determining the antibody 3D structure has been suspected for long in the literature, with supporting experimental data emerging only very recently.[78]

2.1.3. *De novo* sequencing of monoclonal antibodies

Protein sequencing remains a challenge in analytical biochemistry, a large part of deposited protein sequences has been transposed from genomic data. As explained previously, genomic information for antibodies is limited due to their creation through germline rearrangement and somatic mutations / editing. Obtaining sequence information for novel proteins and particularly for monoclonal antibodies with their highly variable CDRs is non-trivial. Two sequencing techniques are used in this project for elucidation of the sequence of the anti-γH2Ax murine monoclonal antibody: RNA sequencing of the antibody-producing hybridoma cell line (performed by Absolute Antibody Ltd) and *de novo* MS/MS of the purified antibody.

2.1.3.1. RNA sequencing of the hybridoma cell line

Hybridoma sequencing relies on Next Generation Sequencing (NGS) technologies for whole transcriptome RNA shotgun sequencing to obtain the transcribed antibody sequence. As this technology was executed externally and only used for sequence validation purpose in this chapter, only an overview is given below (Figure 2.2).

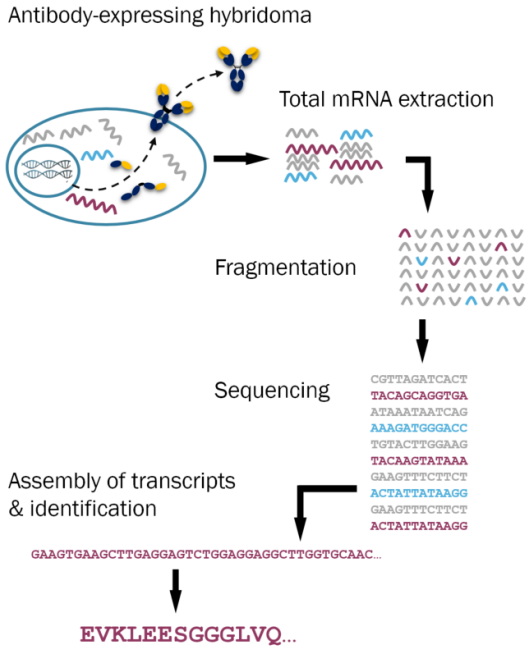


Figure 2.2. Hybridoma sequencing by RNAseq. Adapted from[520] ©Absolute Antibody Ltd.

Briefly, the RNA is isolated from the hybridoma cells and the mRNA enriched to avoid the presence of ribosomal rRNA that would mask signal from the protein-encoding RNA. The mRNA is then reverse transcribed to the more stable cDNA using random primers with hexamer barcodes to generate a cDNA library. This library is subsequently fragmented to obtain sequences of the appropriate length for the sequencing machine. cDNA fragments are purified and immobilised on a solid support for NGS (Illumina HiSeq) using a paired-ends approach: both ends of the dsDNA fragment are sequenced to generate high-quality and alignable data. The sequencing reads are subsequently assembled into contigs, from which antibody transcripts are identified based on homo-

logy using data mining tools. In subsequent data analysis step, sequences for the variable and heavy chain are assembled and transcript variants quantified.[521,522]

RNA-seq using NGS technologies has many advantages over the formerly widespread V-region PCR and 5' RACE, the latter leading to misidentification of aberrant light chains, primer bias towards particular sequences, non-amplification of variable regions due to primer mismatches or low abundance. In addition, amplification of cDNA allows to perform sequencing on low quantities of material.[523]

However, RNA sequencing technologies remain inherently restricted by the information available at transcriptome level which can differ from the actual form of the protein. The expression of additional chains might mask the biologically active form of the antibody, the most abundant mRNA transcripts found in a hybridoma cell line do not necessarily encode the antibody chains providing the correct specificity.[524] Moreover, sequence information obtained at the transcriptome level does not contain additional information about post-translational modifications (PTM) occurring spontaneously such as pyroglutamate formation or enzymatically such as C-terminal lysine clipping, phosphorylation or glycosylation. Yet, these modifications are quintessential for the physiological activity of the antibody, influencing physicochemical properties[515] or the recognition by the complement system[525].

2.1.3.2. *De novo* sequencing of purified antibodies by LC-MS/MS

Over the past decade, tremendous progress in liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) of digested proteins and proteomes (proteomics) has been achieved. Most identifications rely heavily on sequence databases, generated in a large part through translation of genomic sequencing data. This identification of peptides is based on matching to *in silico* spectra of databases using scoring models. *De novo* protein sequencing extracts the peptide sequence directly from the ion series contained in the MS/MS spectra independently of external sequence information, it is thus a powerful technology to identify peptides and proteins that are not inscribed in the catalogue of the cellular genome.

Antibodies are constantly created anew and thus not directly encoded in genomic databases. Although there are conserved sequences for the constant domain or parts of the variable chains, the presence of hypervariable elements and particularly the CDRs generated by random mutations prevents the use of database for direct sequence identification. Strong academic and commercial interest in the *de novo* sequencing of monoclonal antibodies has driven technological advances in this field.

A landmark study by researchers at Genentech combined Edman degradation − limited to 30 amino acids − and analysis by mass spectrometry of overlapping complementary proteolytic and chemical digestions to fully sequence both heavy and light chain of a monoclonal antibody against the OX40 ligand for which a cDNA sequence could not be obtained.[526] With this experience in hand, Genentech continued to develop automation of data analysis using contig algorithms and their assembly using antibody templates[527] and a De Bruijn assembler which could recently be extended to polyclonal antibodies in the mSPS software[528]. Other early reports of the use of MS/MS for antibody sequencing[529,530], particularly the CDRs[531], have emphasised the potential of this technology.

Conversely, improvements of speed, sensitivity, selectivity and fragmentation techniques in modern mass spectrometers have contributed to overcoming major hurdles. Noteworthy, Resemann *et al.* were able to sequence a (shorter) cameloid antibody by top-down MALDI-TOF/TOF.[532] Further studies have followed that emphasise the potential of top-down approaches in this field − a comprehensive review is beyond the scope of this thesis. In addition, the possibility to discriminate between isobaric Ile and Leu amino acids by *w*-ions[239] using MS^3 acquisition strategies based on EThcD fragmentation[241,242] has allowed to move away from specialised digestion protocols[533] or simple comparison with germline databases. Dual HCD/EThcD fragmentation schemes have the advantage of increasing sequence coverage, as has been recently exemplified through the sequencing of the anti-FLAG antibody and the correction of the deposited crystal structure by Snijder and co-workers.[534]

Recent progress in antibody *de novo* sequencing has been mainly driven by progress in data analysis technologies for bottom-up MS/MS. Building on the commercial proteomics software PEAKS™, Tran *et al.* developed a weighted De Bruijn assembler named ALPS to assemble identified peptides from distinct enzyme digestions into long protein contigs or the full protein sequence, available via the publisher.[372] This tool eventually matured into the commercial PeaksAB™ software. Using a different approach based on Byonic's wildcard searches, the Supernovo™algorithm was commercialised − diverging from the commonly used fragment ion sequencing approach.[535] Noteworthy, data analysis methods for diverging sample preparation techniques have emerged. The DiPS/pTA method, also building on PEAKS, provides an alternative De Bruijn assembler accounting for differences in sample preparation when using microwave-assisted amino acid hydrolysis.[536] The pNovoM tools allows to handle digestions with mirror proteases such as Trypsin/Ac-LysargiNase.[537]

The latest innovation in data analysis technologies has been the use of deep learn-

ing for improved interpretation of peptide fragmentation pattern, particularly of fragment signal intensity, such as in DeepNovo[370,538,539] and pNovo3[368]. Additionally, efforts have been ongoing for a more robust statistical validation of identified *de novo* peptides.[371,540,541] The real-world impact of these innovations on the *de novo* sequencing of antibodies, although at times assessed in the publications, remains to be seen. In this regard, the present study provides an application study critically assessing selected state of the art software tools.

Altogether, these technological improvements have allowed antibody *de novo* sequencing by MS/MS to mature from a last resort option when no cDNA could be obtained (loss of hybridoma,[529] isolation from complex mixture,[530] reverse engineering,[542] intellectual property or other reasons) to a mature − yet still challenging − technique for antibody sequencing.

2.1.4. Glycosylation analysis of monoclonal antibodies

Glycosylation patterns are very diverse across the classes of immunoglobulins and can strongly influence their activity and antigen binding: whereas glycosylations do not affect the type of antigen recognised, they influence the quality and strength of the interaction and in consequence affect the recognition process.[80,543]

IgG has the simplest glycosylation pattern with one *N*-linked structure at Asn297 on the Fc region within the C_H2 domain, maintained in all four subclasses of IgG (IgG1, IgG2, IgG3 and IgG4).[80,544] *O*-glycosylation of IgG has been reported on the Fab domain on the light chain[545] and the hinge domain[546,547], observed as well in Fc-fusion proteins.[548]

The glycosylation of monoclonal antibodies affects their 3D structure and conformation,[543] as has been demonstrated on Herceptin[78], influencing protein stability and solubility. The "glycosylation status" of an antibody effects on the cellular transport and clearance; it influences ADCC effector function such as Fc receptor interactions: non-glycosylated IgG has a reduced affinity for the FcγIIIR receptor.[80,543] Crystallographic studies have shown the α1-6 glycan arm of IgG binds non-covalently with hydrophobic amino acids of the C_H2 domain of Fc, whereas the α1-3 branches interact and occupy the space between the C_H2 domains, creating a horseshoe shape of the Fc region with the glycan sequestered in between.[80,543]

Elucidating the glycosylation of a therapeutic, diagnostic or molecular biology reagent antibody has therefore been a long going effort, with a range of technologies in

use.[544,549] Among these, mass spectrometry stands out by its direct molecular readout, analysis speed, sensitivity, specificity and high-throughput.[550,551] Variants have been developed such as LC-MS/MS of digested glycopeptides[552], LC-MS/MS of released and labelled glycans[552], native protein mass spectrometry[553–556] and top-down fragmentation methods.[557,558] Recently, a comparative interlaboratory study has shown a substantial difference in antibody glycosylation analysis results across laboratories, illustrating that the field of mAb glycosylation analysis has not yet reached maturity or harmonised protocols.[559]

Whereas each of these methods has its advantages, glycoproteomic approaches allow both site localisation and glycan composition[560] Our approach in this study is similar to Kolarich *et al.*[552] using LC-ESI-MS/MS and combining the CID and ETD fragmentation techniques for glycan and site identification.

2.1.5. Antibody molecular modelling

The computational modelling of antibodies has long been in the focus of pharmaceutical and academic software development due to its financial and technological prospects. Crystallisation of intact antibodies is particularly challenging due to their macromolecular flexibility. Whereas the F_C part is well-known to crystallise, its conserved nature makes is less interesting for biopharmaceutical projects. The Fab region is more challenging to crystallise, yet remains to this day the gold standard for characterisation of antibody-antigen interaction. Since the crystallisation studies require large amount of protein, costly to produce as monoclonal antibodies are usually obtained from mammalian hybridoma cells, advances in computational modelling technology have carried the promise of savings in time and funds.[561,562]

Antibody homology modelling allows to obtain a 3D structure for the F_V region. It requires however specialised protocols that differ from conventional homology modelling. Only the region outside the antigen binding loops (the framework regions - FR) is expected to exhibit commonality in the structural folding pattern, and the regions of interest (the antigen binding CDR loops) are expected, by their very nature, to be diverse and undergo major conformational changes. Antibody modelling takes advantage of the conserved domain structure and focusses on prediction of structures of the loops. Fortunately, 5 out of the 6 loops have a set of canonical structures adopted, facilitating their prediction. * The H3 loop however is highly individual and a full prediction from

*Three canonical structure types have been identified for H1, five types for H2, four for L1, one for L2,

scratch is needed.[567] Several specialised commercial software packages (MOE, BIOVIA, Schrödinger etc.) and academic web-tools have recently emerged for homology modelling of the F_V region.[568,569]

Computational prediction of the antigen binding follows the same mantra: identification of the binding site is straightforward as it is conserved amongst different antibodies - again relating to their conserved F_V domain structure. However, when docking the antigen into the epitope binding site to obtain the exact coordinates, the immense flexibility and conformational possibilities of the loops needs to be accounted for. This requires broadly sampling docking protocols with antibody-special scoring functions that allow for flexibility in the docking process[570], or in the case of molecular dynamics (MD), simulation times that allow for conformational changes in loop structure to happen (10 - 100 ns range) or enhanced sampling methods.[571] The application of MD simulation to antibody-antigen recognition is currently one of the most promising avenues in this field.[571,572]

and five for L3.[563–566]

2.2. Objective: Project description and analytical strategy

The anti-γH2Ax antibody is at the core of a large oncology project in Oxford centred around a cooperation of groups in the Department of Oncology (Vallis and Cornelissen), Chemistry (Davis) and EMD Millipore and coordinated by Dr. Abirami Lakshminarayanan. The part of the study presented here aims at identifying both the sequence of the murine antibody and its *N*-glycosylation, required for humanisation of the protein sequence and enzymatic re-glycosylation, respectively. Additionally, a 3D model of the highly specific molecular recognition of Ser139 phosphorylated H2AX by the murine antibody is developed for understanding the antigen binding and the effects of mutations surrounding the binding site, as well as the effect of glycosylation on the overall 3D structure (Figure 2.3). Ultimately, this work aims at enabling humanisation.

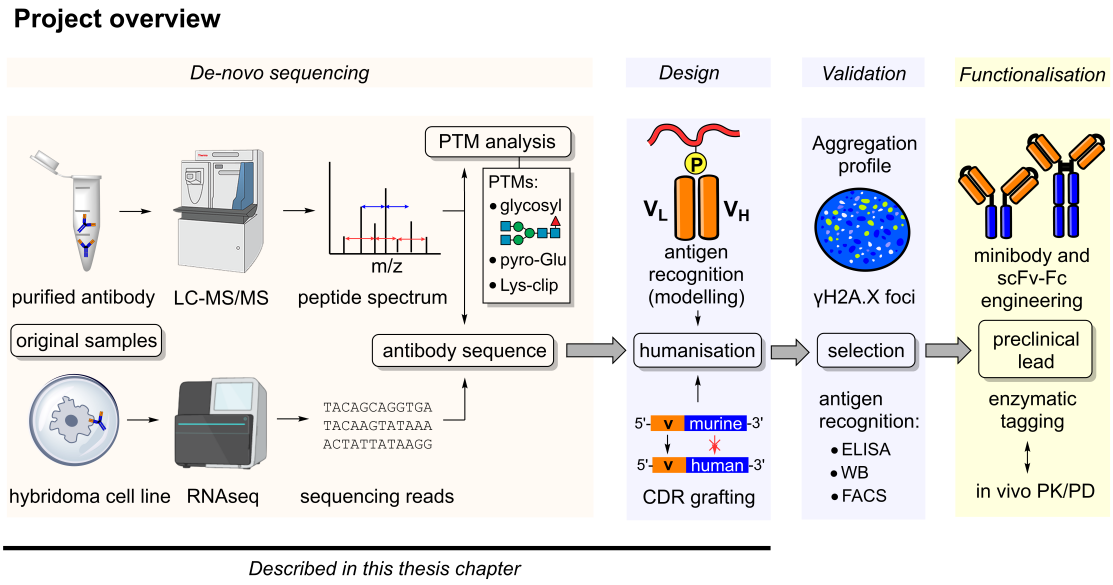


Figure 2.3. Project overview. The sequence of the murine monoclonal anti-γH2AX antibody is obtained independently by two orthogonal techniques: RNAseq starting from the hybridoma cell line provided by EMD Millipore, performed by Absolute Antibody Ltd, and tandem mass spectrometry (MS/MS) using a commercially available sample of the purified antibody protein. MS/MS analysis also reveals post-translational modifications (PTM) of the murine mAb, including *N*-glycosylation. Subsequently, computational modelling enables to generate a 3D model of the antigen recognition and identify key residues for the interaction. With sequence information in hand, humanised constructs of the antibody can be designed based on the CDR grafting approach and subsequently tested for expression, aggregation and antigen recognition. Following on these results, clinically relevant antibody fragments were designed, experimentally validated and their post-translational functionalisation evaluated. CDR grafting, experimental validation and antibody fragment functionalisation were conducted by Absolute Antibody Ltd or by Dr. Abirami Lakshminarayanan and are only briefly mentioned in this thesis.

Overall, the chosen approach focusses on determination of protein sequence, enabling the next step of glycosylation analysis. Similarly, the modelling of the F_V domain is enabled by the protein sequence information − which, together with deposited structural data, allows building a model of the antigen recognition. In a final step, all information can be combined using a close template of an entire antibody structure from the PDB, correcting differences in sequence / glycosylation and grafting the F_V region onto it.

More precisely, the analytical strategy for the *de novo* sequencing of the murine anti-γH2Ax antibody relies on bottom-up tandem mass spectrometry to identify the sequence and denaturing intact mass spectrometry of the deglycosylated antibody as well as complementary RNAseq to cross-validate the result. The RNAseq was performed by Absolute Antibody Ltd as part of a larger commercial package including antibody humanisation in a contract with the Department of Oncology groups. This validated protein sequence is a prerequisite for subsequent glycosylation analysis, using again a combination of tandem mass spectrometry for glycopeptide identification and intact glycoprotein mass spectrometry to obtain the glycan mass as cross-validation.

Building on the obtained sequence by *de novo* MS/MS, a 3D homology model of the F_V region is generated using specialised algorithms for antibody modelling. This model is further used in a fully flexible docking procedure that allows both for small conformational changes in the backbone (loop openings). A structurally well-studied γH2AX C-terminal phosphopeptide containing S139ph is docked into the defined antigen binding site. Manual validation and comparison with known phosphoserine-peptide co-crystal structure enable the suggestion of a 3D model of the antigen recognition. The information gathered in the different parts of this sequencing and modelling endeavour is then translated into a comprehensive 3D model of the full antibody structure that reflects experimental evidence.

The sequence and structural information obtained in this work does not only provide a foundation on which construct design and further experimental validation of our collaborator is built on − highlighted in the outlook 2.5. This study also allows to critically evaluate current state of the art software tools for *de novo* sequencing by MS/MS and for antibody modelling, for which we use both data from the anti-γH2AX mAb as well as for Herceptin, an approved therapeutic widely used in method development for antibody analytics. To our knowledge, this is the first reported tour-de-force study that works its way from *de novo* antibody sequencing to the *in-vivo* evaluation of pre-clinical candidates, including a full antibody structure that comprises both a 3D model for antigen binding and the *N*-glycosylation.

2.3. Materials & methods

2.3.1. General methods

The anti-phospho-histone H2AX (Ser139), clone JBW301 (Cat 05-636) from EMD Millipore was used for this study. Herceptin was kindly provided by Dr. Abirami Lakshminanrayanan, Department of Oncology, University of Oxford.

2.3.2. Intact denaturing protein LC-MS

2 µg of antibody were diluted with DTT (20 mM solution) to achieve a concentration of 0.02-0.1 mg/mL. Intact protein mass spectrometry was performed on a Waters Xevo G2-S Q-TOF operating in V-mode coupled to Water Acquity UPLC. Separation was achieved using a Thermo Proswift (250 mm × 4.6 mm × 5 µm) column with water + 0.1 % formic acid (solvent A) and acetonitrile + 0.1 % formic acid (solvent B) as the eluent system over a 10-minute linear gradient. Nitrogen was used as the desolvation gas (600 L/h) for positive electrospray ionization. Voltages used were capillary: 3000 V, cone: 160 V. Lock-spray analysis ensured continual calibration against a leucine enkephalin standard solution. Raw spectra containing multiple charged ion series were deconvoluted using MassLynx (Waters) and its maximum entropy (MaxEnt1) deconvolution algorithm (Resolution: 1.00 Da/channel, Width at half height: ion series/protein dependent, Minimum intensity ratios: 33 % left and right). Raw data was not subjected to substraction, smoothing or centring prior to deconvolution. Spectra were kindly provided by Dr. Abirami Lakshminarayanan.

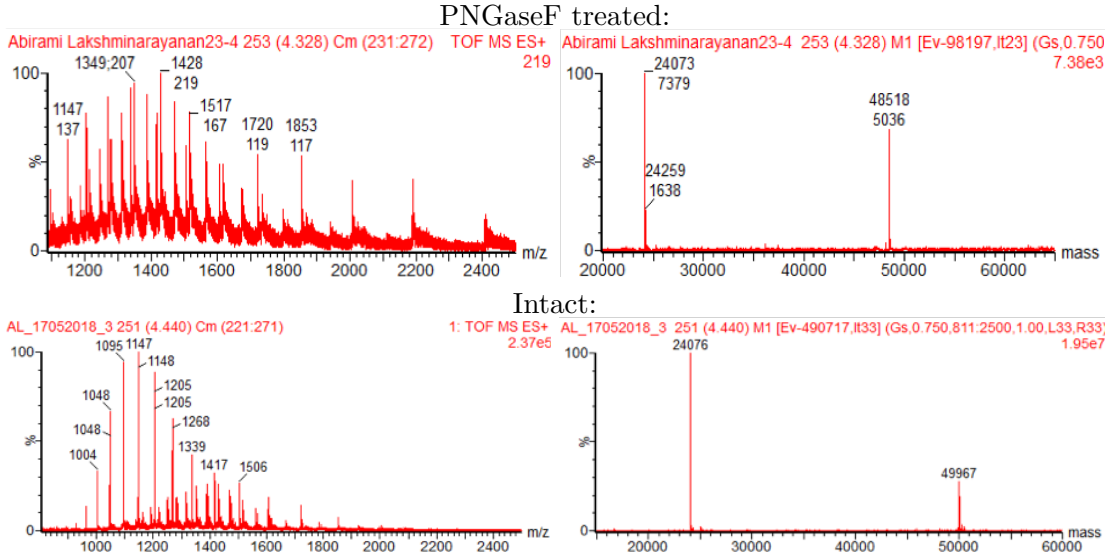


Figure 2.4. Reducing denaturing intact protein LC-MS of the anti-γH2AX mAb. Top. PNGaseF deglycosylated anti-γH2AX monoclonal antibody. **Bottom.** Non-treated antibody. Data kindly provided by Dr. Abirami Lakshminarayanan.

2.3.3. *De novo* antibody sequencing by MS/MS – Experimental

2.3.3.1. Sample preparation for *de novo* sequencing

Both antibodies, Herceptin or anti-γH2Ax were deglycosylated with 50 U PNGase F per µg protein in 100 mM TEAB for 2 h at 37 °C. The heavy and the light chain of each antibody were separated by SDS-PAGE. 1 µg of antibody was reduced and denatured 10 min at 95 °C in gel loading buffer. Samples were run in a Novex 4-12 % Bis-Tris precast gel with MOPS buffer following the manufacturer's instructions. Great care was taken to exclude any source of dust or keratins. Gels were stained in a closed keratin-free box using freshly prepared and filtered solutions. Gel bands were excised on a clean bench using a small sterile scalpel, wearing defuzzled lab coat and face shield protection. Bands were excised, cut to 2-3 small pieces and proceeded for in-gel digestion. Gel pieces were covered in 200 µL 50 % acetonitrile 50 % TEAB, incubated 10 min at r.t. and subsequently 20 min at 65 °C until fully destained. The supernatant was discarded. 200 µL of 10 mM TCEP in 100 mM TEAB were added and the sample reduced for 30 min at r.t. The supernatant was discarded and the gel pieces washed 2x with 200 µL acetonitrile for 10 min at r.t. Samples were alkylated by addition of 200 µL 50 mM chloroacetamide in 100 mM TEAB and incubation for 30 min at r.t. in the dark. The supernatant was discarded and the gel pieces washed 2× with 200 µL acetonitrile for 10 min at r.t. Digestions were performed as follows:

Trypsin	LysC	Chymotrypsin	GluC
1:10 w/w	1:10 w/w	1:10 w/w, 10 mM $CaCl_2$	1:10 w/w
AspN	Elastase	AspN/Chymo	AspN/Chymo/trypsin
1:100 w/w	1:10 w/w	1:200/1:20 w/w	1:200/1:20/1:20 w/w

50 µL of 25 mM ammonium bicarbonate were added. Digestion steps were performed for 4h at 37 °C. After digestion, the supernatant was taken out and the bands rinsed with 15 µL 10 % formic acid per tube (15 min soaking), which was added to the previously extracted supernatant and used for analysis.

2.3.3.2. Standard data acquisition – Q Exactive

Resulting peptides 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 (75 µm inner diameter × 50cm packed with ReproSil-Pur 120 C18-AQ, 1.9 µm, 120 Å, Dr. Maisch GmbH). Separation of cross-linked 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 3×10^6, maximum injection time 50 ms]. After the MS scan, the top 10 most intense peaks were selected for HCD fragmentation at 30 % of normalised collision energy. HCD spectra were also acquired in the Orbitrap (resolution 17500; AGC target 5×10^4; maximum injection time 120 ms), with first fixed mass at 180 *m/z*.

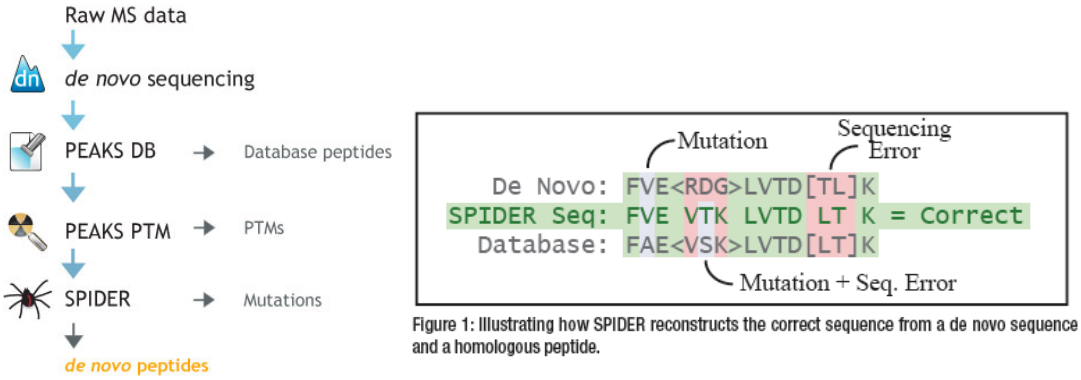


Figure 2.5. Left. Flowchart of Peaks 8.5 software package and its core modules. **Right.** Illustration of the PEAKS SPIDER algorithm allowing to identify mutated peptides. ©Bioinformatics Solutions Inc.

2.3.4. *De novo* antibody sequencing by MS/MS – data analysis

2.3.4.1. Short description of the PEAKS software and its key modules

Unlike other bottom-up proteomics software packages, the approach of PEAKS is built on its *de novo* sequencing capacities (Figure 2.5.A.). *De novo* sequencing derives the peptide sequence directly from the MS^2 spectrum by matching differences between fragment ion peaks to amino acid masses. PEAKS' *de novo* algorithm assigns an amino acid sequence and a confidence percentage for each amino acid to each spectrum.[369] These *de novo* identified peptides are subsequently searched and matched against a protein database for high confidence peptide identifications using the PEAKS DB module. As with other database search software packages, the advantages are an improvement of accuracy and robust statistical validation, allowing sequence assignment with higher confidence.[573]

Peptides identified by *de novo* sequencing but not matched to the database are subsequently searched for common post-translational modifications (PTMs) beyond the set variable modifications. Mass matching of sequence motifs of a single or few amino acids from the closest identified database peptide plus a given modification to the mass obtained by *de novo* sequencing allows to identify additional PTM that might have been missed otherwise. However, the statistical validation of this method is not robust and as a general guideline, only very convincing peptide-spectrum matches (PSM) or PTMs with a high number of distinct PSM should be trusted. Moreover, determination of the exact site of modification needs the presence of site-determining fragment ions.[574]

In addition, PEAKS' SPIDER module allows for mutation search, again by mass match-

ing (Figure 2.5.B.). When similarity is found, the algorithm tries to use both *de novo* sequencing errors and homology peptide mutations to find a consensus sequence with the minimal presence of errors possible and attempts to assign a PSM.[575] Similarly as with PEAKS PTM searches, the results have to be assessed very critically.

2.3.4.2. Description of the *de novo* sequencing methodology by MS/MS using the ALPS assembler

In 2016, Tran *et al.* described an automated antibody *de novo* sequencing method based on output of the different PEAKS modules named ALPS.[372] This method was used in this work for sequencing of the anti-γH2Ax light chain and is briefly described.

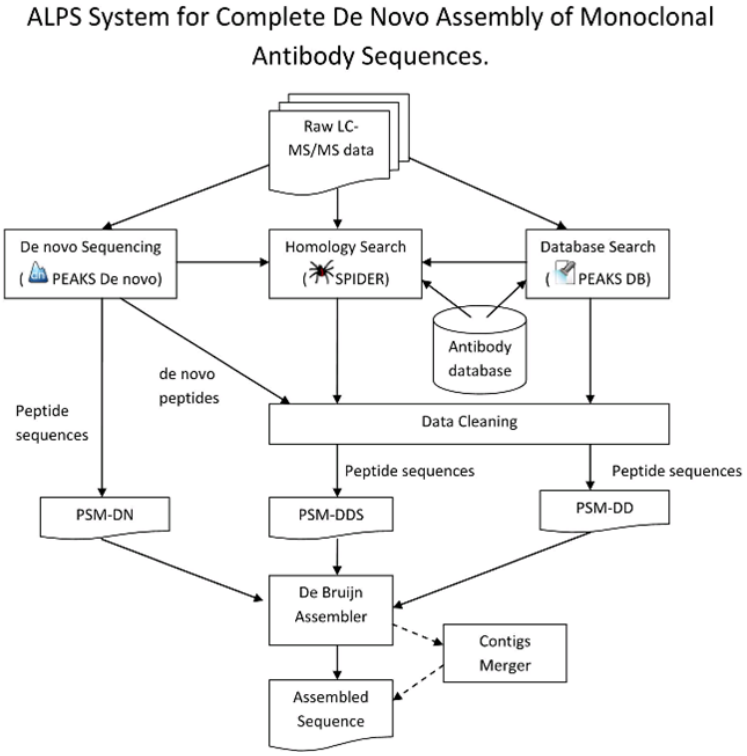


Figure 2.6. ALPS algorithm as described by Tran *et al.*[372] ©Nature Publishing Group.

The raw data were imported into PEAKS Studio 8.5, preprocessed (precursor mass correction, MS/MS de-isotoping and deconvolution, peptide feature detection) and searched together (all modules including *de novo*, DB and SPIDER) against the antibody database published by Tran *et al.*[372] and the MaxQuant contaminants database. A precursor error tolerance of 10 ppm and a fragment error tolerance of 0.03 Da was set. Searches

were performed with carbamidomethylation (Cys) as a fixed modification and oxidation (Met), deamidation (Asn/Gln) and carbamylation (Lys/Peptide N-term) were set. For the heavy chain, pyroglutamate (Glu) was additionally included (see section 2.4.3). 5 variable modifications were allowed with non-specific cleavage at both ends. A 1 % false discovery rate (FDR) was applied to any database related results and an average local confidence (ALC) score of 50.

Subsequently, the antibody sequence was followed using the ALPS method[372], schematised in Figure 2.6. From the results of each PEAKS module, three lists of peptides were generated: PSM-DN, PSM-DD and PSM-DDS.

The PSM-DN was obtained by exporting the 'de novo peptides.csv' from the PEAKS De-Novo module and removing contaminants using an in-house Perl script that performs a BLAST search of the peptide sequences against a contaminant database (courtesy of Dr. Philip Charles, Target Discovery Institute, University of Oxford).
The PSM-DD list was obtained from combining the previous PSM-DN with the 'DB search psm.csv' list from the Peaks DB module. Note that identified contaminants, annotated in PEAKS, were filtered out, as well as peptides without an area value. An ALC of 99 was artificially set for database identified peptides (code listed in appendix 2.11.1).
The PSM-DDS was obtained from combining the previous PSM-DD with the 'DB search psm.csv' list from the PEAKS SPIDER module. Note that identified contaminants were filtered out, as well as peptides without an area value. An ALC of 99 was artificially set for SPIDER identified peptides.

All three lists were individually processed with the ALPS De Bruijn contig assembler[372]. This assembler constructs a De Bruijn graph with a weighted scoring model that accounts for the peptide intensity (feature area) and positional confidence score for each amino acid.

In all cases, the De Bruijn assembler produced a few contigs rather than a full-length assembly. All contigs were inspected for the presence of contaminant sequences by BLAST searches against the NCBI database. Contigs were manually assembled by template alignment in an iterative process. The closest template from the database search was used and validated by a IgBlast search. Contigs were merged using the positional confidence scores and constant domains inferred from the template. For variable domain, priority was given to the *de novo* results. Known disagreements such as typical *de novo* sequencing errors (I/L, AG/Q, PK/KP, N/GG)[576,577] were inspected and corrected using the template in the constant region and the *de novo* peptides in the variable region.

In iterative search rounds, insertion/deletion/mutation reported by SPIDER were corrected, as well as residues with no or low peptide coverage. Details of this process are described in the respective sections.

The final sequence was validated by a search against the entire dataset, and an IgBLAST search against the germline database. Further validation was performed by intact denaturing protein LC-MS and RNAseq of the hybridoma.

2.3.4.3. *De novo* sequencing and sequence validation using PeaksAB 2.0

At the end of the *de novo* sequencing study, our method is compared to the PeaksAB 2.0 software, a commercial solution for antibody *de novo* sequencing. Its methodology is therefore presented. PeaksAB is built on PEAKS' functionalities and has evolved from the ALPS sequence assembly algorithm[372]. It consists of 2 modules: the sequencing module designed to obtain the antibody sequence directly from LC-MS/MS runs with different enzymatic digests and the sequence validation module designed with biopharmaceutical quality control in mind which focuses on sequence confidence, PTMs etc.

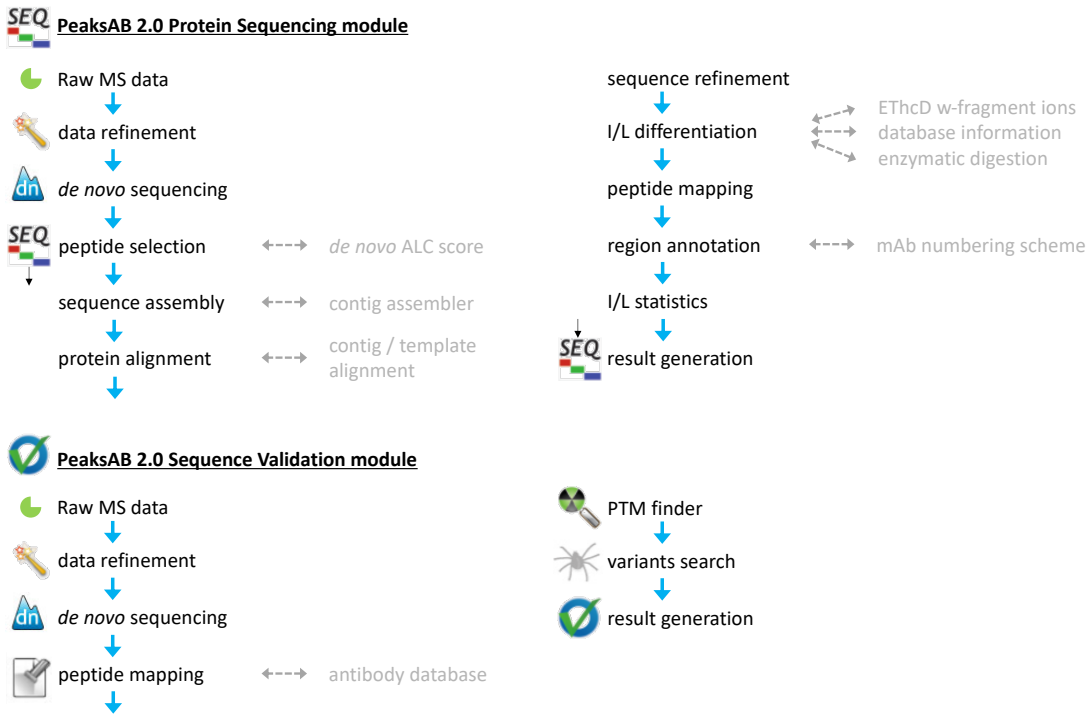


Figure 2.7. Flowchart of PeaksAB 2.0 *de novo* sequencing algorithm.

PeaksAB 2.0 *de novo* sequencing module The sequencing module is derived from PEAKS' *de novo* module. RAW data is refined and *de novo* sequencing performed on MS^2 spectra as described in 2.3.4.1. Subsequently, steps unique to antibody sequencing have been included: selection of peptide on ALC score (typically >50), sequence or contig assembly, protein alignment and sequence refinement steps. (Figure 2.7, top)

PeaksAB 2.0 *de novo* sequence validation module The sequence validation modules takes functionalities from PEAKS DB, PTM and SPIDER modules described previously which have been tailored to antibodies, i.e. antibody databases and PTM (including glycosylations), common mutations. Its structure is very similar to a full run through all of PEAKS modules (Figure 2.7, bottom).

2.3.5. Glycosylation analysis

2.3.5.1. Sample preparation for glycosylation analysis

10 µg of anti γH2Ax (S139) murine antibody from Merck JBW301 was denatured in 8 M urea in 100 mM TEAB in a volume of 100 µL for 30 min at r.t. Sample was subsequently reduced with 10 mM TCEP for 30 min at r.t. and alkylation with 50 mM 2-iodoacetamide for 30 min at r.t. in the dark. Samples were diluted to 6 M urea with 50 mM TEAB and digestion was performed with either LysC, trypsin or elastase (1:20 w/w) O/N at 37 °C, followed by trypsin (1:40 w/w) for 4 h at 37 °C. Samples were dried for 4 h in a speed-vac before being resuspended in 5 % FA 5 % DMSO immediately before data acquisition.

2.3.5.2. Data acquisition for glycosylation analysis – Orbitrap Elite

Resulting peptides were analysed on an EASY-nLC 1000 UHPLC system (Proxeon) connected to a Orbitrap Elite mass spectrometer (Thermo Fischer Scientific) possessing an EASY-Spray nanoelectrospray ion source (Thermo Fischer Scientific). The peptides were trapped on a guard column (75 µm i.d. x 20 mm, reprosil C18, 3 µm, 120 Å) using solvent A (0.1 % formic acid in water) at a pressure of 500 bar. The peptides were separated on a C18 analytical column (75 µm i.d. × 750 mm, RSLC C18, 2 µm, 100 Å) using a linear gradient (length: 58 minutes, 8 % to 30 % solvent B (0.1 % formic acid in acetonitrile, flow rate: 200 nL/min). The separated peptides were electrosprayed directly into the mass spectrometer operating in a data-dependent mode.

CID experiments

Full scan MS spectra (scan range 350-1500 m/z, resolution 120000, AGC target $1e^6$, maximum injection time 250 ms) were acquired in the Orbitrap and subsequent CID MS/MS spectra (resolution 15000, AGC target $3e^4$, maximum injection time 500 ms) of 5 most intense peaks were acquired in the ion trap. CID fragmentation was performed at 30 % of normalized collision energy and the signal intensity threshold was kept at 500 counts.

ETD-DDDT experiments

Full scan MS spectra (scan range 350-1500 m/z, resolution 120000, AGC target $1e^6$, maximum injection time 250 ms) were acquired in the Orbitrap. CID and ETD spectra were acquired in Ion Trap (resolution 7500, AGC cation target $3e^4$, AGC Anion target $2e^5$, maximum injection time 100 ms). After the MS scans, the 5 most intense peaks were selected for fragmentation based on data-dependent decision tree (DDDT). The signal threshold for fragmentation of parent ion was set to 500 ion counts. For CID fragmentation, the normalized collision energy and default charge state was set to 35 % and 2 respectively. For ETD fragmentation, reaction time was set to 50 ms and supplemental activation was enabled. Charge state screening was enabled; parent ion with unassigned charge state and charge state 1 was rejected. Dynamic exclusion was enabled (exclusion size list 75, exclusion duration 5 s). The setting for DDDT is as such: ETD fragmentation was followed for charge states 3, 4 and 5 with m/z less than 750. For charge states greater than 5, ETD fragmentation was always performed. In all other cases, CID activation was performed.

HCD-pd-ETD experiments

Full scan MS spectra (scan range 350-1500 m/z, resolution 30000, AGC target 1e6, maximum injection time 250 ms) and subsequent HCD MS/MS spectra (resolution 15000, AGC target 3e4, 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. This method uses HexNAc oxonium triggers: if peaks at any of m/z 138.0546, 168.0563 or 204.0867 (50 ppm error) were detected among top 20 most intense peak of a HCD spectrum, an ETD MS/MS spectrum of the precursor was triggered.[305,578]

2.3.5.3. Data analysis for glycosylation analysis

The raw data was processed using Byonic™ software (Protein Metrics, version 2.3). The raw MS file was searched against the respective protein sequence as well as list of common contaminants. Trypsin or elastase (semi-specific) with a maximum number of 3 missed cleavages was selected as the protease. Carbamidomethylation (cysteine) was set as fixed modification. Oxidation (methionine), carbamylation (*N*-terminus) and deamination (asparagine) and the Bionyc™ mammalian *N*-glycan database were set as variable modifications. Precursor mass tolerance was set as 15 ppm. Fragment mass tolerances for CID and ETD were set to 0.6 Da. All spectra were manually validated.

2.3.6. RNA sequencing

The RNA sequencing was performed by Absolute Antibody Ltd. as described in 2.1.3.1. The result is attached in appendix 2.11.3.

2.3.7. Antibody Molecular Modelling

2.3.7.1. Homology model of the variable region

abYmod Homology model of the F_V fragment was generated from the MS/MS & RNAseq consensus sequence using the abYmod 1.2 beta version (`http://abymod.abysis.org`), developed by Andrew C. R. Martin and co-workers at UCL, released in September 2019.
Framework templates 1t660 (2.3 Å resolution, R-factor 0.20 and R-free 0.25) and 1zea0 (1.8 Å resolution, R-factor 0.20 and R-free 0.25) were selected. For the loop templates, 1c1e0 (1.9 Å resolution, B-factor 28.6, RSSQSLV**H**SNGNTYLH) was selected for L1, 1c1e0 (1.9 Å resolution, B-factor 27.6, KVSNFS) again for L2, 3qcu0 (2.0 Å resolution, B-factor 24.9, SQSTH**F**P**F**T); 4n8c0 (1.6 Å resolution, B-factor 25.4, GY**T**FTDY**A**MH) for H1, 3mo10 (1.8 Å resolution, B-factor 31.0, WINTETGEPTYADDFKG) for H2 and 5l1k0 (1.6 Å resolution, B-factor 27.1, **YD**G**ELD**F) for H3. The output structure of abYmod was used without further modification, pyro-Glu being not recognised by most computational chemistry software.

ABodyBuilder Oxford's own antibody builder was used via the web app: `http://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/abodybuilder/`. Briefly, ABodyBuilder renumbers the submitted sequences for heavy and light chain and chooses the best templates from the curated SAbDab database. The VH-VL orientation is predicted from the angles the template domains placed in the predicted pose, before CDR loop modelling. Side chains are generated in a subsequent step, the final model is renumbered and annotated.[569]

Rosetta Antibody The Rosetta Antibody homology modelling protocol 3 was accessed via the ROSIE interface (`https://rosie.graylab.jhu.edu`). Briefly, canonical templates are selected and assembled in a first stage. Templates are chosen using BLAST from antibody crystal structures in the PDB. Template searches are independent for each CDR (L1, L2, L3, H1, H2, and H3), identified by a knowledge-based system, and for each framework region (FRL and FRH), respectively. They are assembled using a Rosetta protocol, resulting in a crude structure. All CDRs other than H3 are modelled using templates in the PDB whereas CDR-H3 is re-modelled *de novo* using a kinematic loop modelling algorithm in a Rosetta protocol. In the second stage, CDR-H3 is remodelled *de novo* and VL/VH orientation is optimized using a Rosetta protocol. Paratope side chains and loop backbones are refined simultaneously. The server returns templates information, one crude assembled model, 10 low-energy refined models as PDB files and the score file.[568]

2.3.7.2. Data mining of the RCSB PDB

For identification of phosphorylated H2AX peptides deposited as complex with their receptor, the UniProt-PDB link was used and all entries manually inspected.

To identify all murine IgG1 κ co-crystal structure with a peptide antigen, the automatically updated SACS database[579] from Andrew CR Martin's group at UCL was downloaded as XML file, filtered for crystal structures of murine IgG1 κ antibodies in complex with peptidic ligands that are not lysozyme. The 17 resulting entries were manually inspected.

2.3.7.3. Docking of the peptide antigen in the F_V antibody receptor

CABS-Dock is a specialised protein-peptide docking application allowing for full flexibility of the peptide and small fluctuations of the receptor backbone.[580] The docking was

performed on the abYmod homology model using the webserver: `http://biocomp.chem.uw.edu.pl/CABSdock`. The CDR loops were defined as flexible, the framework regions as 'unlikely to bind'. Unfortunately, the server does not support phosphoserine, docking was tried with the ASQEY and AEQEY peptides, as well as longer N-terminal variants. Docking poses were manually investigated, no plausible docking conformation was found.

HADDOCK is a flexible docking platform for integrative modelling that allows incorporation of experimental data in form of constraints.[581] The docking procedure is executed in three steps: 1) rigid body sampling, 2) simulated annealing of 200 rigid based conformations and 3) molecular dynamics refinement in explicit solvent (50 ps, TIP3P water). The 'Easy' interface of the HADDOCK webserver was used: `https://milou.science.uu.nl/services/HADDOCK2.2/haddockserver-easy.html`. The pdb file was renumbered and prepared according to the demands of the webserver. The protein/peptide protocol was selected, the CDR loops defined as active residues and all other ones as passive. All analysis was done with open source PyMol v 2.3, the detailed results are listed in appendix 2.11.5 and 2.11.6.

2.3.7.4. Homology model for the entire anti-γH2Ax monoclonal antibody

To our knowledge, the PDB structure 1IGY is the only crystallised murine IgG1 antibody.[582]† The HADDOCK result cluster2_1 derived from the abYmod homology model and containing the ApSQEY peptide in the presumed active conformation was aligned in PyMol to the F_V domain of 1IGY. The F_V of 1IGY was removed and the cluster2_1 F_V merged to the constant part of the 1IGY F_{ab}.

After sequence comparison, mutations were added. No mutations were made in the light chain. The following mutations were made in the heavy chain: Pro188 → Thr, Arg189 → Trp, Thr241 → Val, Leu243 → Thr, Val246 → Leu, Asn272 → Asp, Val293 → Ser, Ala296 → Glu, Lys333 → Arg, Arg335 → Lys, Ser374 → Trp, Asp375 → Asn, Ala378 → Pro, Pro379 → Ala. In addition, the amino acids SPG were added at the C-terminus. The K was omitted as it is trimmed, according to the MS/MS data.

†Data mining of the RCSB PDB database identified only the following crystallised full antibody constructs: 1IGT, 1IGY, 1HZH, 3BQU, 1FSK, 5DK3, 4QOU, 4QOV, 4QOW

Finally, the glucose in the G1F glycan was removed to obtain the correct G0F glycan. The resulting structure was then energy-minimised by steepest descent using the Amber12 forcefield [‡] , keeping the antigen binding site fixed, as this site has been previously optimised with a higher computational method.

[‡] *Technical note on carbohydrate forcefields:* Despite specialised carbohydrate forcefields having been developed such as Amber99sb/Glycam06, these are generally speaking not implemented in common molecular modelling software.[583–587]

2.4. Results

Please note that parts of the Results section relating to the sequence of the murine anti-γH2AX antibody are confidential, see the confidentiality declaration in section 2.10.

2.4.1. Assessing *de novo* sequencing tools on the Herceptin test system

The monoclonal antibody Trastuzumab, often referred to by his brand name Herceptin, is used in the treatment of cancers that have high expression levels of human epidermal growth factor 2 (HER2), such as early and advanced breast cancer and stomach cancer. In research, it is used as model for monoclonal antibodies (mAb) in several methodological studies. Here, we used Herceptin as a model antibody due to its IgG subfamily and ready commercial availability to assess the capacity of two accessible software tools for automated *de novo* assembly of complete monoclonal antibodies from tandem mass spectrometry data: the ALSP De Bruijn assembler and PeaksAB. To our knowledge, specialised software tools for mAb *de novo* sequencing by MS/MS have not yet been compared in the literature.

In a field dominated by commercial antibody sequencing services and proprietary software, the ALPS assembler source code was released as part of the original *Scientific Reports* publication[372], and as an academic lab we were able to obtain a 2-week trial version for PeaksAB™ thanks to the efforts of Elisabete Pires, CRL Mass Spectrometry Research Facility, Oxford. Further efforts to obtain access to other specialised software, such as Supernovo™ were not fruitful. Many academic *de novo* sequencing software developers, such as Novor[366] or pNovo 3[368], both reasonably accessible through DeNovoGUI[588], do not include features tailored to antibodies (see discussion in 2.4.4 explaining the relevance). Some developers understandably keep the necessary extensions for antibodies proprietary to offer an in-house commercial antibody sequencing service (ie. RapidNovor's RemAb® and WILD®). The meta-SPS platform has been only used for antibody sequencing in a highly customised environment.[528] It has not escaped our notice that both tools available for our evaluation, ALPS and PeaksAB, are both based on the PEAKS software package and have similar algorithms. Nonetheless, results differ – as will be described in the following.

Ultimately, our goal is to evaluate the available tools on a known target to enable us to define our own *de novo* sequencing strategy for the anti-γH2AX antibody.

2.4.1.1. Both ALPS and PeaksAB successfully assemble Herceptin's light chain

De novo sequencing requires a complete sequence coverage of the antibody by the digested peptides. We separated the heavy and light chain of the deglycosylated antibody by SDS-PAGE and performed eight separate enzymatic digests to provide full sequence coverage of the antibody (namely: trypsin, chymotrypsin, LysC, AspN, GluC and elastase, AspN+chymotrypsin and AspN+chymotrypsin+trypsin). This diversity of digestions provided peptides overlapping in their amino acid sequence which can be assembled to the full protein sequence.

The separately acquired datasets were combined and analysed jointly. Figure 2.8 shows a PEAKS DB search of the dataset against the known sequence of the Herceptin light chain. The search results in 100 % sequence coverage, including many overlapping peptides. Thus the data set can be seen as comprehensive and constitutes an appropriate test system for the following evaluation of *de novo* sequencing software tools.

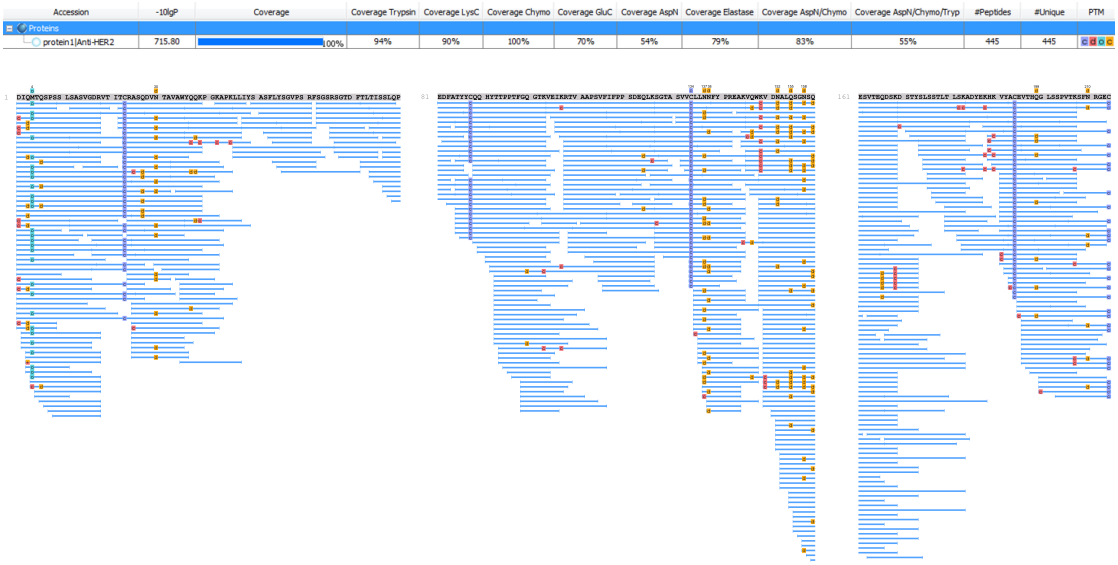


Figure 2.8. De novo sequencing of Herceptin light chain – PSM-DD alignment.

First, we used the workflow described by Tran *et al.* to assemble the Herceptin light chain sequence with the ALPS assembler (see 2.3.4.2 for details), using as input either the PSM-DD list (no PTM/mutation search) or the PSM-DDS list (includes PTM/mutation search) to understand the influence the peptide list has on the assembly result. Conceptually, de-novo spectrum sequencing differs substantially from spectral matching to a database (see 1.3.5). PTM and mutation searches allow to bridge the gap between the two algorithms and thus provide additional identifications.

Initially, we started with the PSM-DD list which, as mentioned, includes the *de novo* PSM and the database PSM, but lacks PSM identified by PEAKS' PTM or SPIDER mutation searches. The alignments of the 2 best contigs suggested by the ALPS assembler (seq1 and seq0) to the Herceptin light chain shows that the assembly can be made from the two longest contigs (Figure 2.9).

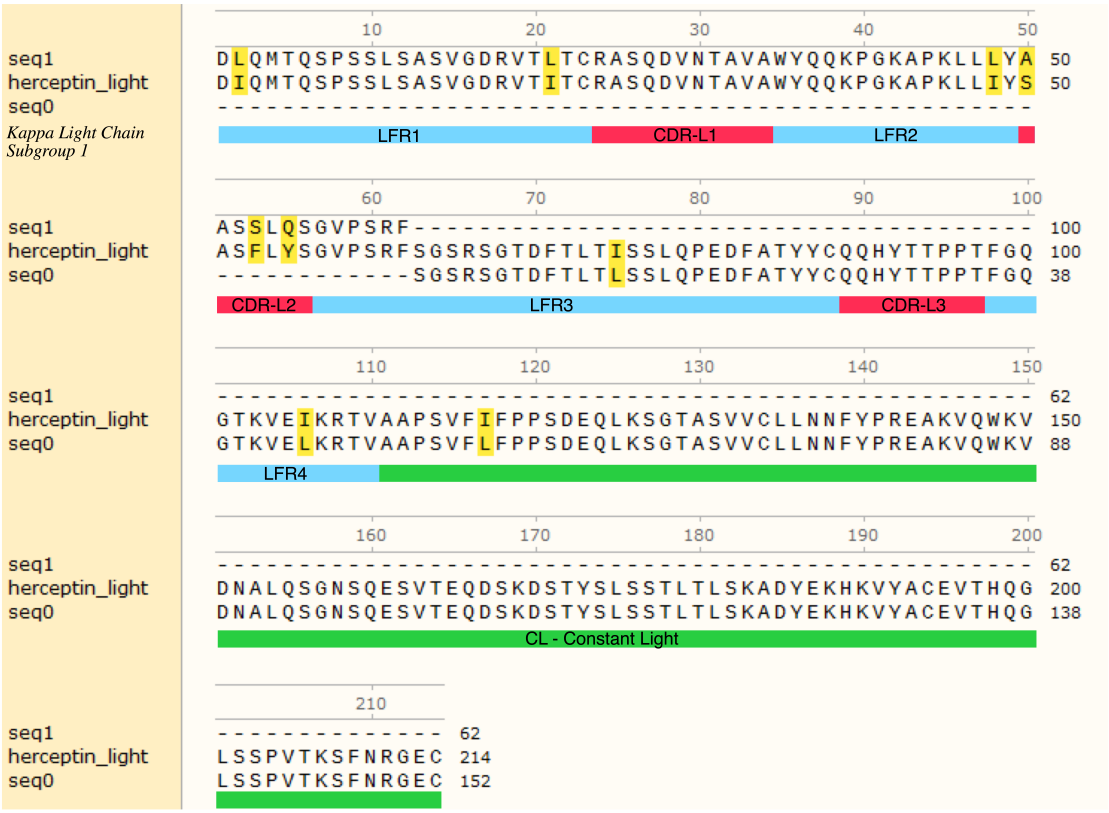


Figure 2.9. De novo sequencing of Herceptin light chain – PSM-DD alignment. All annotations in this thesis follow the Kabat numbering scheme. Annotations were counter-checked with the abYsis 3.4.1 software[589] (http://www.abysis.org). LFR - light chain framework, CDR - complementarity determining region, CL - light chain constant region. Discrepancies in sequence are highlighted in yellow.

This assembly finds the right sequence with the 2 longest contigs, however errors other than Leu $\longleftrightarrow$ Ile mismatch can be seen in the CDR-2 region. This indicates the weakness of the ALPS system when using only the PSM-DD list: a major input comes from peptides identified by database search on the antibody database, which makes correct sequencing of highly variable CDR domains challenging. This can be solved by manual validation of the sequence, identifying the gap in the PEAKS DB (Figure 2.10.A) and using *de novo* peptides (in grey, example in Figure 2.10.B) or PEAKS' SPIDER mutation search (Figure 2.10.C) to identify the correct sequence.

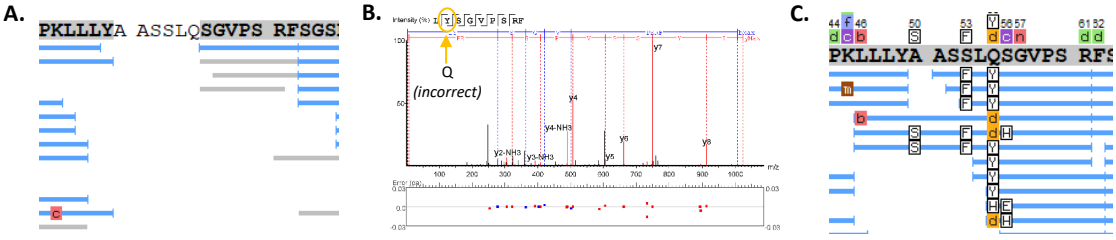


Figure 2.10. ***De novo*** **sequencing of Herceptin light chain – PSM-DD alignment.** Sequence errors result in loss of peptide coverage; these can be manually corrected.

When using the PSM-DDS list (contains *de novo*, DB and SPIDER PSM) directly for the assembly with the ALPS De Bruijn assembler, the correct contig is generated in the first attempt (Figure 2.11). Mismatches in CDR-L2 are not produced, because the SPIDER peptides add additional emphasis on the corrected peptides as the ion count intensity (or AUC, area under the curve) is accounted for each PSM in the assembly. However, Leu/Ile mismatches are frequent, PEAKS assigns by default Leu. This can be manually curated by querying germline frequency data for the production organism, for example using abYsis[589].

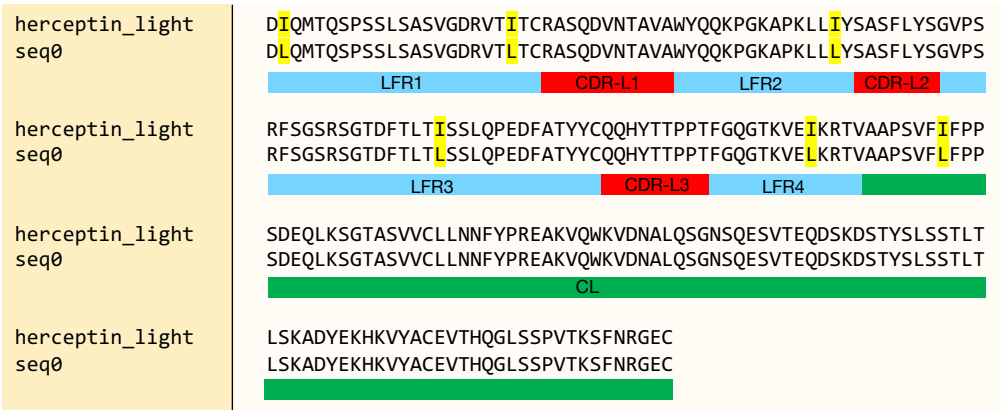


Figure 2.11. ***De novo*** **sequencing of Herceptin light chain – PSM-DDS alignment.** Inclusion of PTM / mutation search results improves contig assembly.

Taken together, the Herceptin light chain test system shows that we can successfully use the ALPS system for sequencing. The inclusion of identifications by PTM/mutation searches (PSM-DDS peptide list) proved critical for correct CDR determination. Nonetheless, manual analysis using PEAKS functionalities allowed to obtain the correct sequence even from the PSM-DD list by a 'fill the gap'-approach: using template alignment, high-scoring *de novo* peptides or SPIDER-mutated peptides suggested by PEAKS, a self-consistent sequence with the dataset was generated and found to be correct. This methodological validation of the manual sequencing approach is important for the following work on the anti-γH2AX mAb, as similar approaches will be used.

We also evaluated PeaksAB v2.0 on the Herceptin light chain. Without surprise, PeaksAB was able to sequence and assemble the Herceptin light chain – showing only Ile/Leu mismatch and an additional N-terminal Gly (with 0 % confidence – see Figure 2.12). Noteworthy, the Herceptin sequence is deposited in most antibody databases. Nonetheless, this shows that antibody sequencing by MS/MS has reached a considerable level of automation; the more recent PeaksAB outperforming the PEAKS+ALPS approach in the Leu/Ile differentiation. PeaksAB uses EThcD fragmentation (via *w*-ions, not used here), enzyme specificity and/or homology database analysis for this task, taking advantage of the fact that chymotrypsin (and pepsin) prefers to cleave C-terminal of leucine over isoleucine.[590]

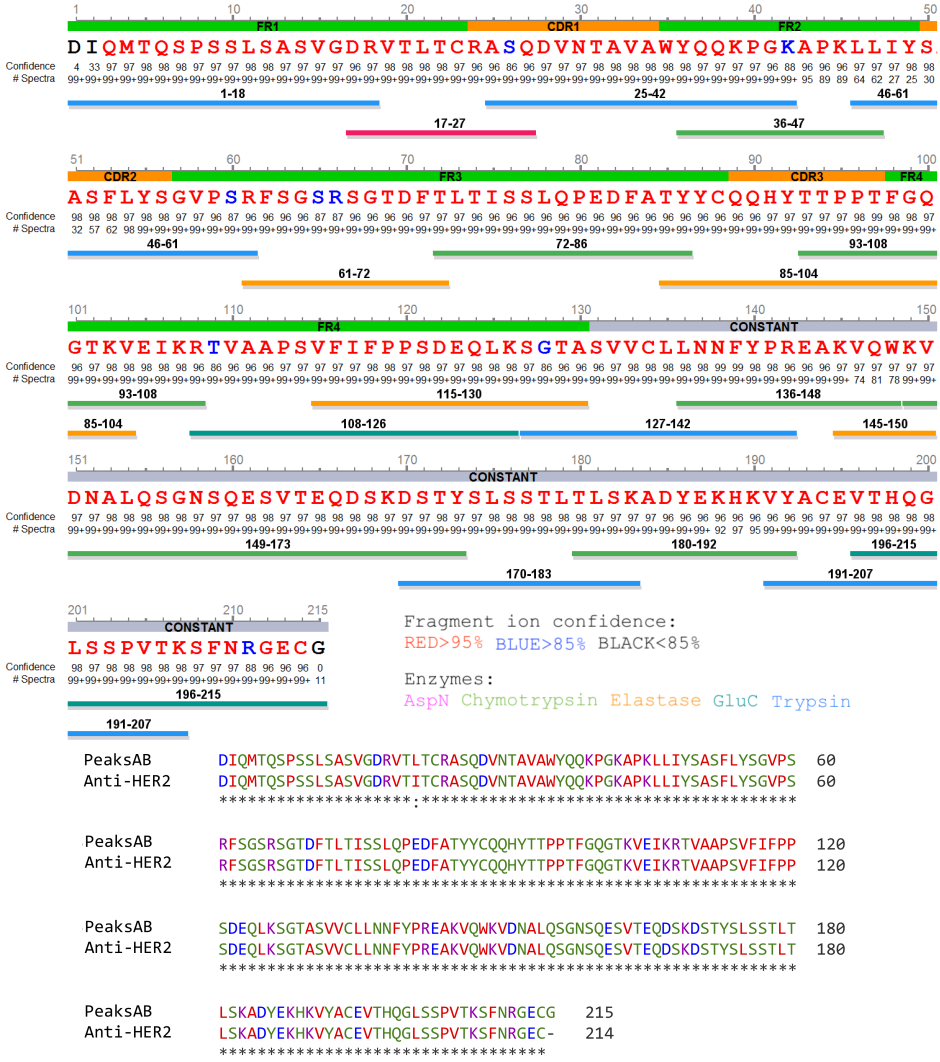


Figure 2.12. ***De novo*** **sequencing of Herceptin light chain by PeaksAB. Top.** Assembled result for the light chain. The report by PeaksAB indicates local confidence and number of spectra. Overlapping peptides with highest scores from the multi-enzyme digest are indicated. **Bottom.** Alignment of the suggested *de novo* sequence by PeaksAB to the deposited Herceptin light chain sequence by Clustal Ω.

2.4.1.2. PeaksAB outperforms ALPS for the Herceptin heavy chain assembly

The previous assembly of the light chain highlights the importance of the inclusion of PTM / mutations and that manual inspection enables local sequence corrections. The heavy chain of Herceptin is a more complex test system. The corresponding gel band was similarly sequenced by MS/MS and the assembly attempted. The database search against the known sequence in Figure 2.13 illustrates the depth in peptide sequence coverage of the acquired dataset, emphasising that for this particular dataset, *de novo* sequencing is mainly a data analysis problem.

Figure 2.13. Sequence coverage of the Herceptin heavy chain in the combined dataset of all 8 digestions.

As shown in Figure 2.14, the protein sequence can be assembled from the PSM-DDS contigs using a template to fill the gaps. The heavy chain constant region can be build from overlapping contigs seq0+2+3, the variable domain in principle from seq1+7+5+4 – with errors in the CDR's (Figure 2.14).

This alignment highlights key difficulties in antibody *de novo* sequencing. Apart from the well-known Leu/Ile mismatch due to their homologous mass, mismatches mainly occur in the variable chain, as it holds higher variability (including in the sequences

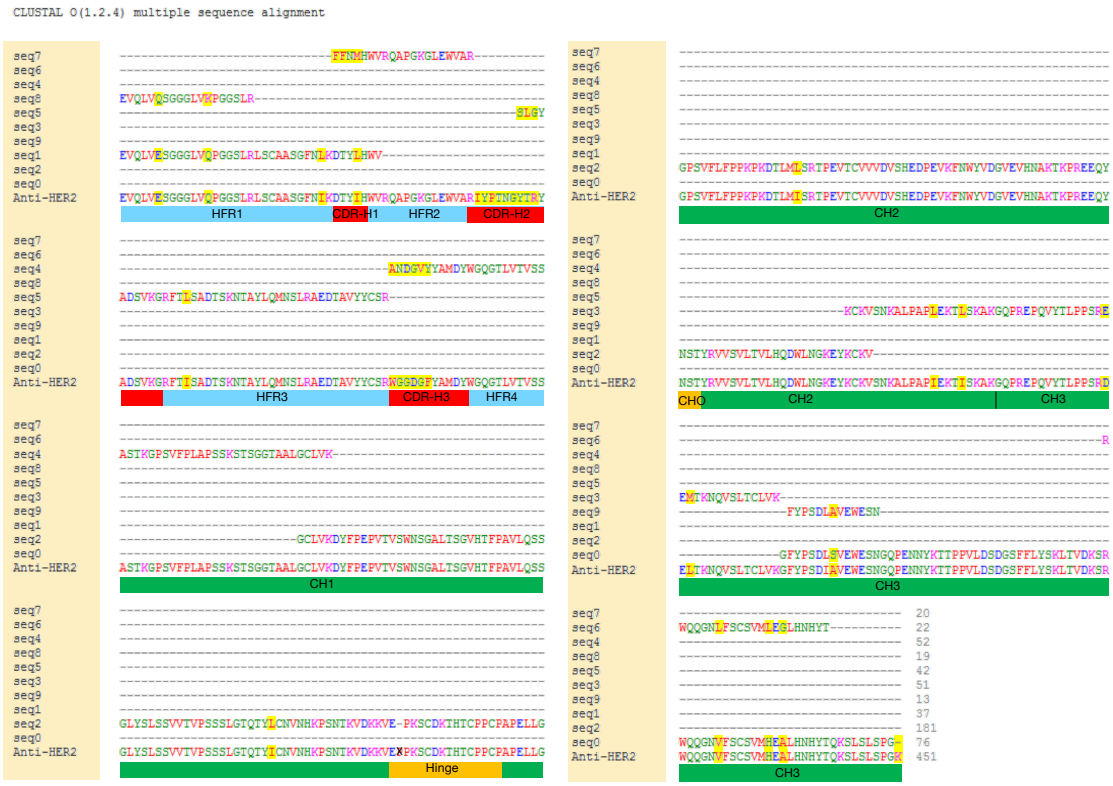


Figure 2.14. Contig alignment to the Herceptin heavy chain. Mismatches between contigs and the Herceptin sequence are highlighted in yellow, sequence domains are highlighted. CHO = *N*-glycosylation site. Kabat numbering.

deposited in the used antibody database). The algorithm performed particularly poorly in the highly individual CDR-H2 and CDR-H3 domains, these being either not assembled or having large chunks incorrectly sequenced. Manual validation and additional analysis of gaps is required to assemble the full Herceptin sequence, illustrating the limitation of the ALPS automatised approach.
This reality check on one of the best established De Bruijn assembler illustrates that more work is still needed to improve current algorithms for *de novo* sequencing.

Similarly as for the light chain, PeaksAB on the other hand was able to sequence the Herceptin heavy chain with minor errors and a notable discrepancy in the CDR-H3, which would require manual correction using the build-in tools of the software (Figure 2.15). Impressively, CDR-H1 and CDR-H2 were correctly sequenced and only few I/L mismatches are found. Deamidated Asn / Asp mismatches such as on the N-S-T glycosylation site (labelled CHO) are noted more frequently.

Nonetheless, this case highlights drawbacks of PeaksAB. Some sites with errors can

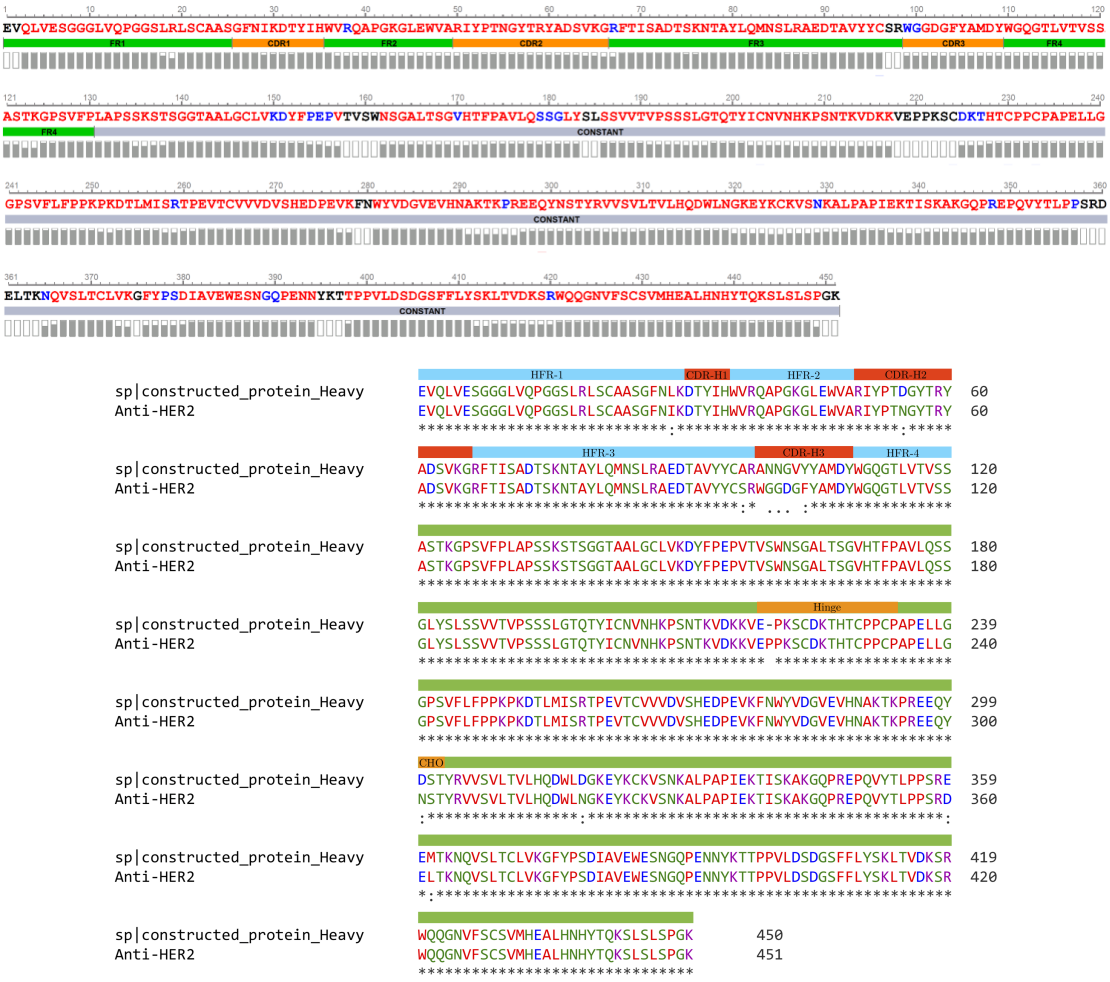


Figure 2.15. ***De novo*** **sequencing of Herceptin heavy chain by PeaksAB. Top.** Probability annotation of the assembled sequence. Red > 95 %, blue > 85 % and black < 85 %. **Bottom.** Alignment of the suggested *de novo* sequence by PeaksAB to the deposited Herceptin sequence by Clustal Ω. Kabat numbering / CDR annotation.

be spotted by inspection of the sequence coverage view with the local confidence, such as in the Hinge domain or the M/L error in the C_H3. Yet, these errors would be easily missed without a close manual inspection. However other error sites, including CDR-H3, have high confidence scores, misleading the user to believe they would have been correctly assigned. This underlines the importance of cross-validating the sequence obtained by MS/MS with another analytical technique, such as MW determination by intact protein LC-MS using the deglycosylated antibody. In this case, the sequence suggested by PeaksAB would be 49270 Da, which does not match the observed 49142 Da (calculated from deposited sequence: 49153 Da).

2.4.2. *De novo* sequencing of the anti-γH2AX mAb light chain

2.4.2.1. Manual correction of the ALPS assembly allows successful sequencing

Having assessed the performance of the available tools on a well-studied antibody in the previous sections, we take this experience forward for the *de novo* sequencing of the murine anti-γH2AX antibody – the goal of this project. Sequencing the anti-γH2AX antibody additionally provides a real-world test case for *de novo* sequencing by MS/MS. Using the ALPS assembler[372], contigs were found and assembled into a first manual draft sequence using the highest scoring template from the database. To assess if the gaps in contig coverage can be filled up from the template, the template sequence coverage was analysed and the variable domain checked against an IgBLAST search. As shown in Figure 2.17, the gap in contig coverage, highlighted in grey, matches the known CDR-L1

Figure 2.16. Alignment of the contig to the highest score antibody database (highest -10logP and sequence coverage). Left: PSM-DN. The template matches to seq1+gap+seq4+seq6+seq0. **Right:** PSM-DDS. The template matches to seq1+gap+seq6+seq3+seq0. The start of the template highlighted in blue is a known signal peptide, which is in the genomic sequence but cannot be found in the expressed antibody.

(IMGT numbering) variable domain part. Therefore, the gaps can be filled in from the highest scoring template (high LogP). Building on this, a manual draft sequence was assembled, taking known I/L mismatches and other known flaws in *de novo* sequencing into consideration by comparison with similar deposited sequences.

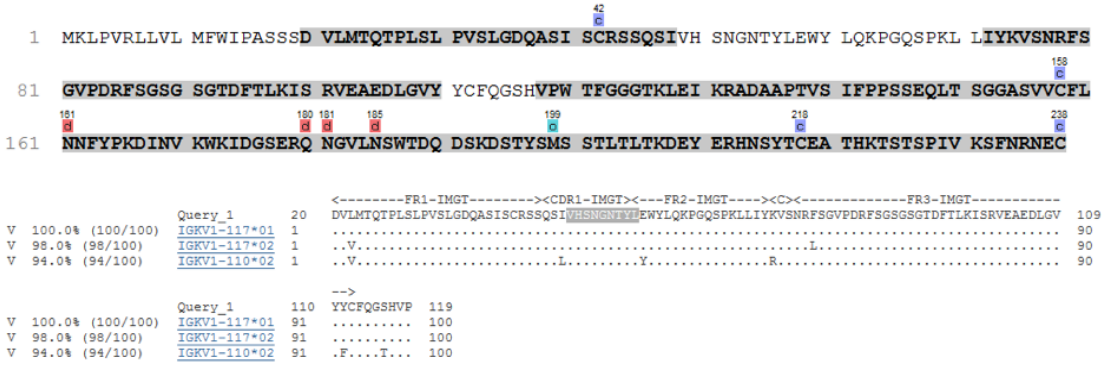
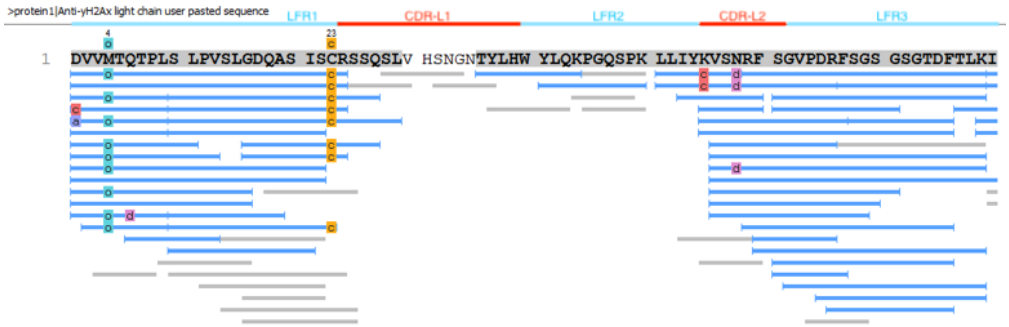


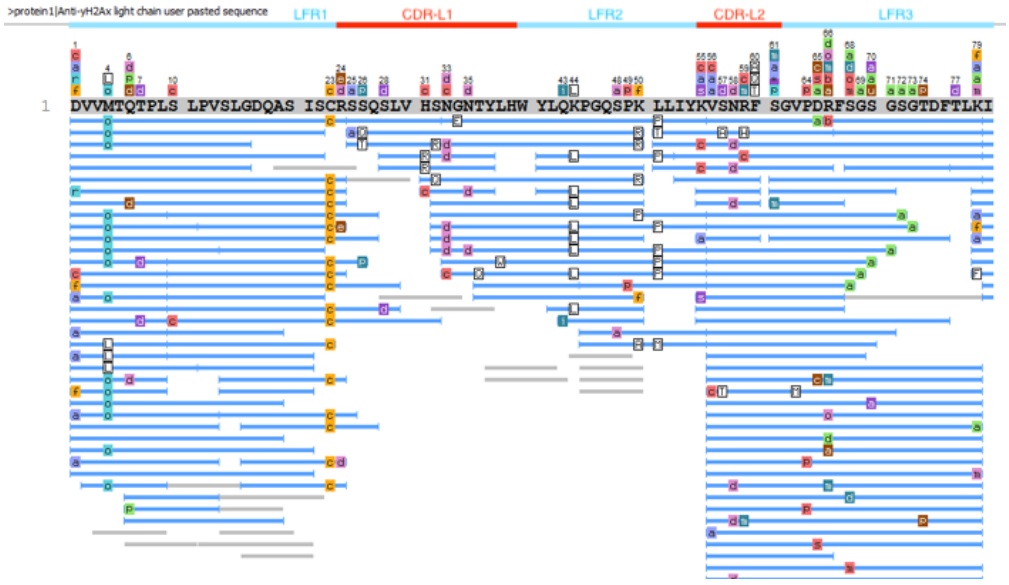
Figure 2.17. Validation and inspection of the high-scoring template (high LogP) for anti-γH2Ax light chain contig assembly. Top: Sequence coverage in PEAKS DB search (1% FDR). Signal peptide is not covered. **Bottom:** IgBLAST of the template suggests IgG1 κ light chain variant 1. The part taken over from the template for manual assembly is highlighted in grey. Note the IMGT numbering scheme used by IgBLAST differs from the Kabat scheme used in all annotations in this thesis chapter.

Once a draft sequence was assembled, it was searched again in PEAKS. The search resulted in 97 % coverage, with two gaps in the sequence − suggesting sites of error or mismatches to be corrected. A successful *de novo* sequencing requires 100 % sequence coverage and the sequence has to be self-consistent with the dataset. In the following, the correction of the draft sequence using a '*filling the gap*' manual sequencing approach is detailed, working our way from the *N*- to the *C*-terminus, while illustrating how to take advantage of different features of PEAKS − a readily available software in academic labs.

First, the VHSNGN gap in the **CDR-L1** was targeted:



The number of *de novo* peptides (grey) found in the gap already indicated that the sequence must be only slightly different at this place. As this corresponds to the variable domain of the antibody, template alignment most likely was misleading. To identify potential point mutations – i.e. single amino acid errors, a SPIDER mutation search was conducted:



The SPIDER search indicated that H31 should be R31 (all peptides following C-terminally are tryptic), K44→L44/I44 and L51→P51, which was confirmed by a subsequent validation search in PEAKS DB yielding peptide coverage with novel additional PSM. This observation emphasises once again the importance of mutation searches, enabling speedy and straightforward progress. For Ile/Leu-44 differentiation, abYsis frequency information for *mus musculus* was consulted, revealing 0 known sequences with a leucine at this position. In this case, the differentiation of the isobaric amino acids, indistinguishable by mass spectrometry without further fragmentation, can be solved using the germline data.

Next, the second gap at HVPY in the **CDR-L3** was investigated. Both a *de novo* peptide and a peptide identified in the PEAKS SPIDER search suggested a V→A mutation (Figure 2.18). This illustrates a simple case, were different sources of evidence converge.

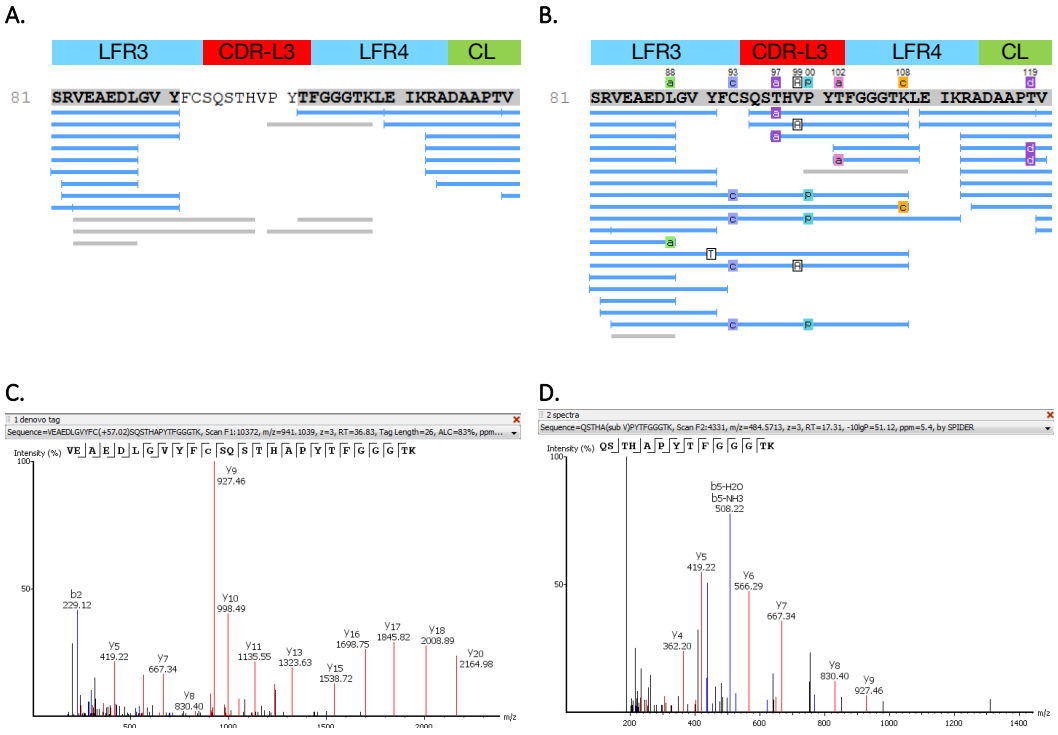


Figure 2.18. Solving the CDR-3 region of anti-γH2Ax light chain. A. The CDR-L3 gap, no matching peptides are identified. **B.** The SPIDER search suggest a mutation Val→Ala. This is confirmed by a *de novo* peptide (**C.**) and a SPIDER peptide (**D.**).

The final sequence leads to 100% sequence coverage (Figure 2.19.A). In summary, the described '*filling the gap*' procedure enabled to manually curate errors in an iterative approach and thus complete the sequencing task. The identified subtype of the anti-γH2AX antibody is IgG1 κ light chain subgroup II, the IgBLAST search highlights the similarity to deposited sequences (Figure 2.19.B).

Whereas the original goal of this project was to identify the antibody protein sequence and glycosylation by MS/MS, the opportunity of obtaining confirmation of the assembled sequence by RNAseq presented itself approximatively one year after the MS/MS was started. At that time, the oncology partner group started a commercial collaboration with Absolute Antibody Ltd. on the humanisation of the anti-γH2AX mAb. The RNAseq was part of the paid-for service portfolio. As can be seen in Figure 2.19.C, the RNAseq sequence is identical with the MS/MS sequence. The sequence has a calculated molecular weight of 24079 Da, which matches the experimentally observed 24075 by denaturing intact protein LC-MS (Figure 2.19.D)

In conclusion, both techniques have yielded the same sequencing result that fits the molecular weight determined by intact protein LC-MS. This suggests that our approach to antibody *de novo* sequencing is valid and yields the correct result.

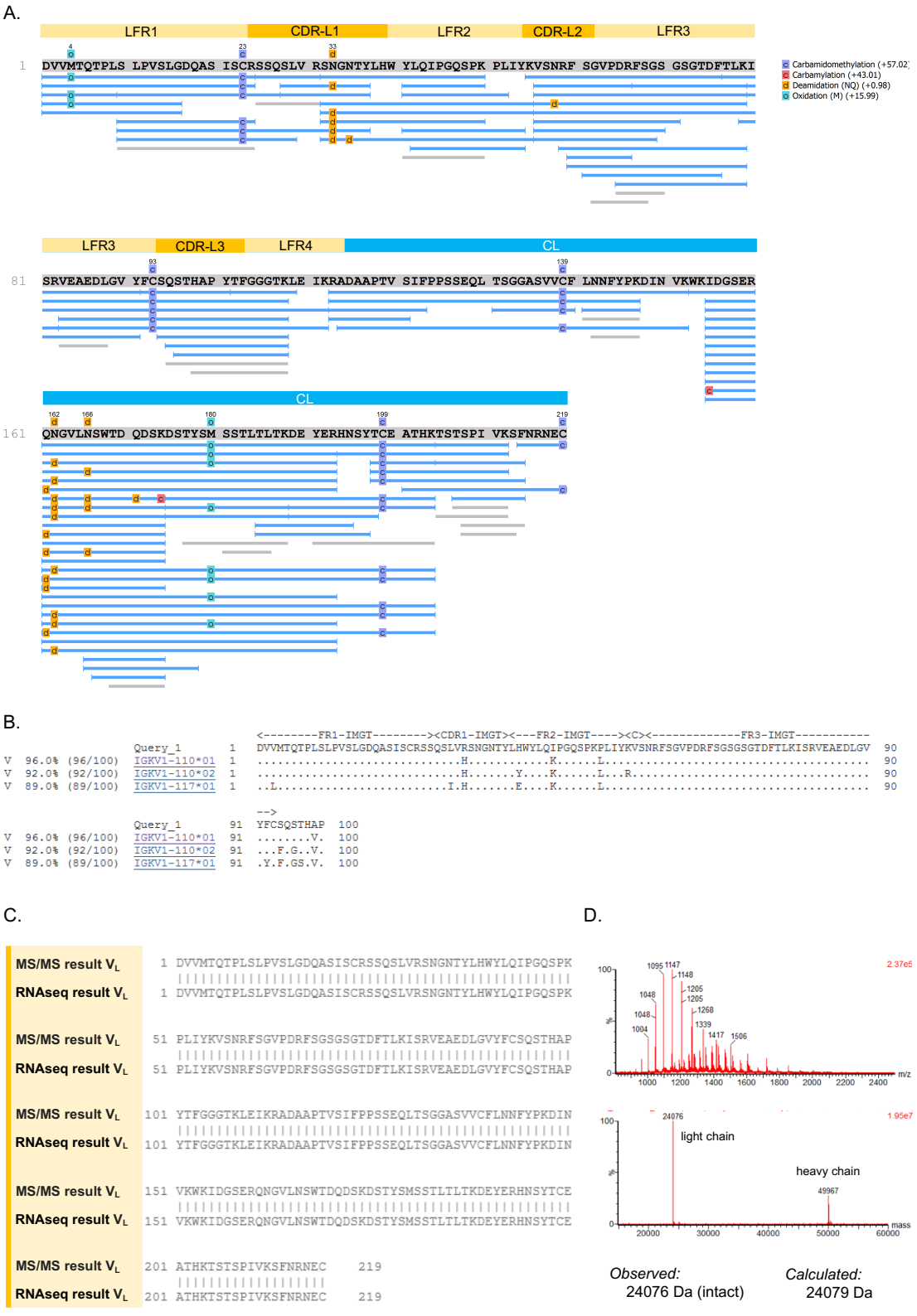


Figure 2.19. ***De novo*** **sequencing summary of the anti-γH2AX light chain.**

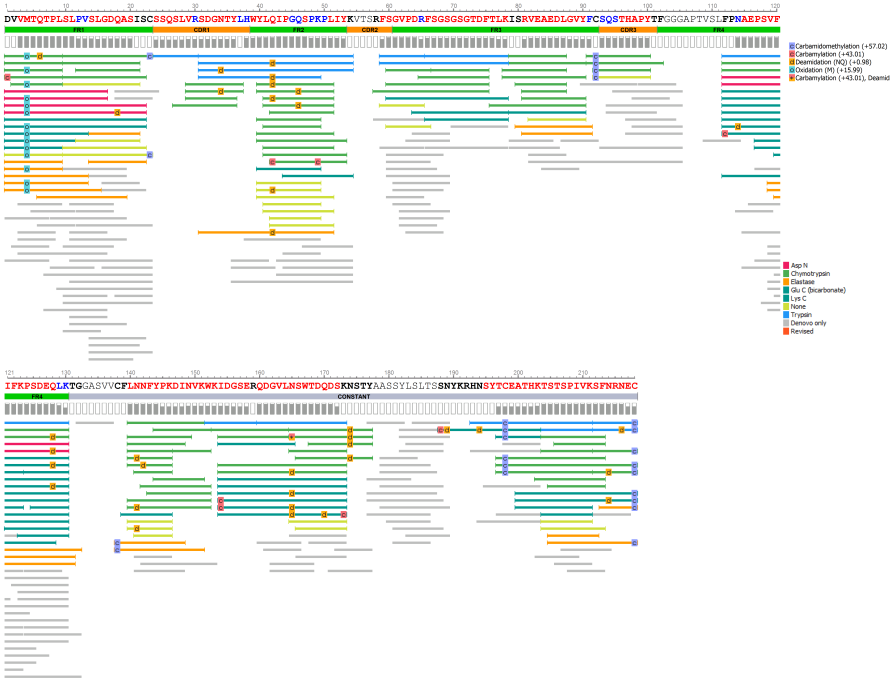
Figure 2.19. (***continued***) **A.** Sequence coverage by MS/MS showing the identified peptides at 1% FDR (100% sequence coverage). The sequence is annotated using the Kabat numbering scheme. **B.** IgBLAST search of the final sequence. **C.** Alignment of the sequence obtained by *de novo* MS/MS sequencing with the RNAseq result shows identical sequences obtained independently by the two methods. **D.** Molecular weight determination of the intact murine antibody by denaturing intact protein LC-MS, showing the ion series (top) and the deconvoluted spectrum (bottom). The mass obtained by MW determination fits the calculated mass from the sequence.

2.4.2.2. PeaksAB has substantial difficulties with the CDR-L and CL regions

In the previous section, the light chain could be successfully sequenced using the ALPS assembler followed by manual curation. For comparison, the commercial PeaksAB was tested on the anti-γH2Ax mAb 8-digest dataset to assess its ability to sequence the light chain in an unsupervised manner. Figure 2.20 shows that the software was not able to accurately sequence the light chain and would require further manual investigation to obtain the correct sequence. Noteworthy, difficulties are occurring in the CDR2 and HFR4 as expected, but also in the constant chain, supposedly more straightforward to sequence. This is particularly surprising as the ALPS contig assembler, from which code PeaksAB was further developed, was able to fully assemble the heavy chain from the dataset.

In order to facilitate manual *de novo* sequencing, PeaksAB has features that help perform the same tasks on identifying clues and spectra enabling to correct the missing parts. Grey lines in Figure 2.20.A show *de novo* peptides which would help the sequencing. Particularly, the possibility to modify the sequencing result in real-time is very valuable. However, the approach described in the previous section correcting the ALPS assembly using a template has the advantage of quicker obtaining the constant domain and allowing to settle I/L and D/N mismatches using germline frequency information. It also makes better use of the explicit PTM and mutation searches (SPIDER) which allow to quickly pinpoint single-residue errors.

A.



B.

```
CLUSTAL O(1.2.4) multiple sequence alignment

Consensus                       DVVMTQTPLSLPVSLGDQASISCRSSQSLVRSNGNTYLHWYLQIPGQSPKPLIYKVSNRF  60
sp|constructed_protein_Light    DVVMTQTPLSLPVSLGDQASISCS-SQSLVRSDGNTYLHWYLQIPGQSPKPLIYKVTSRF  59
                                **********************  *******:**********************:.**

Consensus                       SGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCSQSTHAPYTFGGGTKLEIKRADAAPTV  120
sp|constructed_protein_Light    SGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCSQSTHAPYTFGGGAPTVSLFPNAEPSV  119
                                ************************************************:       :* *:*

Consensus                       SIFPPSSEQLTSGGASVVCFLNNFYPKDINVKWKIDGSERQNGVLNSWTDQDSKDSTYSM  180
sp|constructed_protein_Light    FIFKPSDEQLKTGGASVVCFLNNFYPKDINVKWKIDGSERQDGVLNSWTDQDSKNSTYAA  179
                                 ** **.***.:*****************************:***********:***:

Consensus                       SSTLTLTKDEYERHNSYTCEATHKTSTSPIVKSFNRNEC       219
sp|constructed_protein_Light    SSYLSLTSSNYKRHNSYTCEATHKTSTSPIVKSFNRNEC       218
                                ** *:**..:*:**************************
```

Figure 2.20. ***De novo*** **sequencing of the anti-γH2Ax mAb light chain by PeaksAB. A.** Sequence coverage along with suggested sequence and average local confidence score (ALC) for each amino acid (grey bars). Peptides from each digest are highlighted in differing colours and *de novo* sequenced peptides not matched to the sequence highlighted in grey. Sequence annotation follows the IMGT scheme, amino acids are coloured by confidence (red >95%, blue >85%, black <85%; non-bold suggests gap). **B.** Alignment of the suggested sequence by PeaksAB to the consensus sequence of MS/MS and RNAseq (see 2.4.2).

2.4.3. Manual sequencing of the anti γ-H2Ax heavy chain

The combined digestion dataset was searched against the antibody database. An initial contig assembly with the previously described ALPS assembler was unsuccessful as large parts of the variable region could not be identified. At the time this project was conducted, PeaksAB was not available − looking over the fact that PeaksAB would not have provided much help either (see 2.4.3.5). Considering that no suitable template for assembly could be obtained directly from the database − all high-scoring antibodies having sequence coverages of <80% and many gaps in the variable domains − a different methodological approach was adopted than previously described for the light chain or for Herceptin.

It was thus decided to fall back on a fully manual sequencing strategy. This manual sequencing was performed in four steps: 1) Generation of a consensus template, 2) Local correction of the consensus template in the F_V domain, 3) Investigation of sequence coverage gaps in the constant domain, 4) comparison and validation of the manually assembled sequence with orthogonal techniques (RNAseq; MW determination). Although conceptually not innovative,[576,577] this approach has not been explicitly described in the literature, except for a document published on the website of Bioinformatics Solution Inc. which is reprinted in appendix 2.11.2.

2.4.3.1. Generation of a consensus template

Three templates were retained for assembly of a consensus template, using the highest scoring (-10logP) database hit (id_1725) as template for the heavy chain. Additionally, two hits with high sequence coverage in the variable domain were retained (id_6062 and id_2255) to allow assembly of a consensus template that covers the variable part of the sequence. The first underlying assumption in this approach is that the heavy chain frameworks in the variable domains (HFR) have conserved elements and therefore HFR and potentially the highly variable CDR domains could be solved from comparing discrepancies in the templates directly with the MS/MS data to suggest a corrected sequence for each site. The second underlying assumption is that the 8-digest dataset contains all evidence required to solve the sequence (i.e. that the data contains 100 % sequence coverage) and that a correct sequence has to be self-consistent with the underlying data.

Figure 2.21. Alignment of the three templates sequences from the antibody

Figure 2.21. (*continued*) **database used for construction of the consensus template for the anti-γH2Ax antibody heavy chain.** Alignment was performed with Clustal ω (1.2.4). id_1725 was used as template for the constant domain, id_6062 and id_2255 for the variable domain. Discrepancies in the sequences are highlighted in yellow, with sites numbered according to the Kabat numbering scheme. The CMITDFFPEDITVEWQ sequence is highlighted as MS/MS coverage of this region is lacking, as discussed later in this chapter.

Figure 2.21 shows the sequence coverage of the three templates and their sequence alignment and annotation, discrepancies within the templates are highlighted in yellow and subsequently analysed using the MS/MS data to elaborate a consensus template using the experimental evidence presented in Figure 2.22. Each corrected sequence was confirmed in further downstream analysis by new additional overlapping PSM identified in database searches with the corrected sequence.

Importantly for the choice of the two variable domain templates, both templates selected show pyroglutamate formation on the N-terminal Gln (Q) as identified initially in PEAKS SPIDER searches (Figure 2.22.A), a prevalent modification for Q at the protein *N*-terminus. Observation of this spontaneous lactam cyclisation was taken as a stark hint for glutamine in position H1 of the HFR1.

2.4.3.2. Correction of the consensus template in the variable domain

Having a consensus template in hand, the next efforts focussed on correcting the template in the F_V domain by working our way from gap to gap through the sequence. Along with elucidating the light chain composition, the different manual *de novo* sequencing strategies applied are described, highlighting their strengths and pitfalls.

On the ***N*-terminus**, the same SPIDER peptides identifying pyro-Glu further provide information on the lack of coverage in the K.ASGY sequence, with 2 peptides suggesting a mutation from serine to alanine in position H25 (KASGY → KAAGY). The corrected sequence was validated by subsequent identification of several new LysC and tryptic peptides PSM, adding further evidence for the corrected sequence.

This approach of '*filling the gap*' by taking up a hint/indication, correcting the sequence accordingly and subsequently validating it by a new database search could be applied on several occasions. The **CDR-H1** domain for example has no sequence coverage in both templates selected implying it is incorrect. However, a *de novo* peptide is identified for id_6062. This peptide suggests the DYSMH sequence for the CDR-H1 domain and a T → I or L mutation at position H28 (Figure 2.22.B). A database search

against the corrected sequence found a high number of different overlapping PSM, confirming the suggested sequence. The I/L mismatch was solved comparing to the germline database, in which I is listed in this combination. Similarly, the template discrepancies W ↔ R in the **HFR2** and T ↔ A in the **CDR-H2** could be solved by inspecting the sequence coverage, id_2255 leading to identification of chymotryptic peptides containing the latter CDR-H2 sequence (Figure 2.22.C).

Solving the **HFR3** sequence provides an example of a more complex case, as it requires simultaneous inspection of three sites: H69; H72 & 73; H79 & H81 & H82A. As will be explained in the following, only parts of the HFR3 could be determined by the '*fill the gap*' manual sequencing strategy, requiring the introduction of a new approach. Figure 2.22.D shows a database-identified peptide confirming Phe at H69 as in template id_6062. Similarly, the sequence of id_6062 at positions H79/81/82A is confirmed by a high number of PSM covering the site (Figure 2.22.E). These observations settle both ends of the HFR3 region, while the middle part remains unknown.

Unfortunately, no further hints or indications could be found to solve the missing part, highlighting the limitations of the '*fill the gap*' approach. Therefore, a new manual sequencing strategy was used, termed '*conserved sequon search*' or '*tag-based de novo sequencing*'.[591] Concretely, site H72/H73 could be elucidated by searching for the "SASTA" sequence motif in the PEAKS *de novo* peptides dataset, which was obtained in the first processing step of the raw data. The "SASTA" sequence motif being conserved in the two templates, it was assumed that it would likely be present in the anti-γH2Ax antibody. Figure 2.22.F shows a *de novo* PSM that contains this sequence motif, thus providing a possible sequence for the gap. Its *C*-terminal sequence matches very well the consensus sequence for HFR3 and has high local ALC scores (>90 %), which supports the identification suggesting therefore a high likelihood for this peptide to indeed be the correct sequence. However, the *N*-terminal end doesn't quite match. As a matter of fact, the yellow box in Figure 2.22.F indicates the local ALC for each residue and shows poor scores for the N-terminal amino acids. Indeed, the low MW area of the spectrum has major peaks unassigned; moreover LS is isobaric to TV. Upon closer inspection, the manually assigned SL-E-T *b*-ion series (highlighted in Figure 2.22.F) matches the conserved SL part of the previously established HFR3 templates. This sequence was therefore retained and subsequently confirmed in a database search, yielding a high number of PSM in this region. In summary, the example of the HFR3 gap shows that the combination of two different manual sequencing strategies, the '*fill the gap*' and '*conserved sequon search*' search approaches can be a powerful approach to solve complex cases.

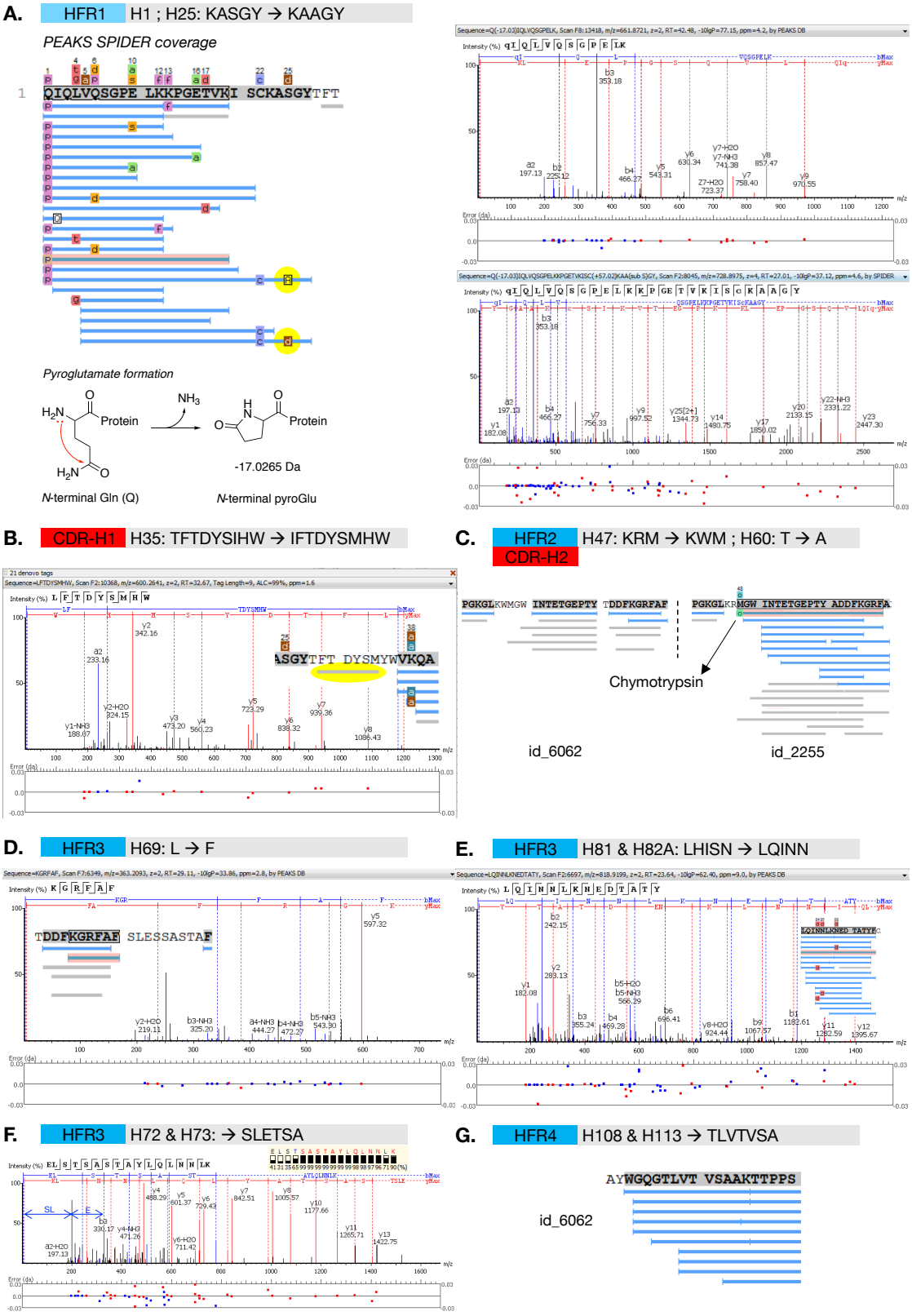


Figure 2.22. (*Continued on the next page.*)

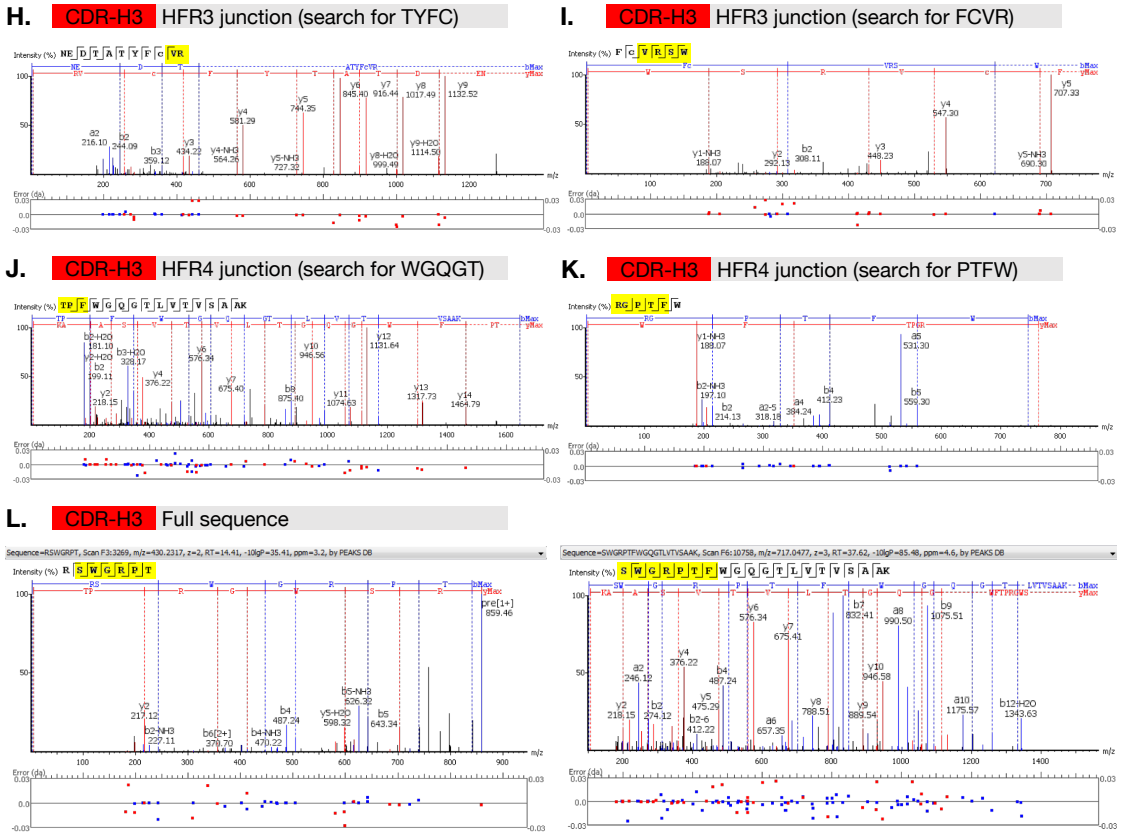


Figure 2.22. Site-by-site inspection of the variable domain peptide-spectrum matches to identify the correct sequence for assembly of the consensus template. Templates (see Figure 2.21.B) are compared with MS/MS data, each site is referred with the Kabat number. **A.** *N*-terminus and HFR-1. Pyroglutamate formation was identified on many PSM by PEAKS SPIDER module (violet p) following the described mechanism, the top spectra shows a high scoring PSM for pyro-Glu. Additionally, the SPIDER identification suggests an S → A mutation at position H25. **B.** A *de novo* identified peptide suggests the sequence for the CDR-H1. **C.** Both templates for the variable domain differ in the HFR2 at position H47: R ↔ M. The identification of chymotryptic peptides by PEAKS DB C-terminally located to that site suggest W, as chymotrypsin cleavage site. The sequence coverage in the CDR-H2 region of id_2255 suggest alanine at position H60. **D-F.** Inspection of the HFR3 region. **D.** A PSM identified by PEAKS DB suggests Phe at position H69. **E.** A high number of database PSM validate the id_6062 template sequence. **F.** A *de novo* peptide elucidates the central part of HFR3 with 2 possible annotations. **G.** A high number of database PSM for the id_6062 template validates its sequence. **H-L.** Manual *de novo* sequencing of the CDR-H3 region, starting with identifying its *N*-terminus (H, I), continuing with the C-terminus (J, K) before validation of the full CDR-H3 in a subsequent PEAKS DB search of the suggested sequence. **H.** Search for TYFC in *de novo* peptides data identifies a PSM suggesting VR as following residues. **I.** Subsequent search for FCVR in the same dataset confirms the FCVR sequence with a new PSM and suggests SW as following residues. **J.** Search of WGQGT (part of HFR4) suggests TP-F at the *C*-terminus of the CDR-H3. **K.** Search of PTFW identifies a PSM suggesting the RGPTFW sequence for the *C*-terminus of the CDR-H3. **L.** Two PSM identified in a subsequent PEAKS DB search of the suggested sequence confirm the CDR-H3 sequence.

In some cases, *de novo* sequencing can be very straightforward: the **HFR4** region appeared to be identical to the sequence of id_6062 region (Figure 2.22.G). However, there are also cases were solving the sequence can be difficult, as will be illustrated with the CDR-H3 region.

Solving the **CDR-H3** region appeared particularly challenging with both templates having no overlap, different length and high variability, thus emphasising that a substantial chunk of the amino acid chain is unknown. As such, neither the previous '*filling the gap*' nor the '*conserved sequon search*' strategy are applicable. A third manual sequencing approach was chosen, termed '*incremental expansion*': starting from the *C*- and *N*-terminus, searches of 4 or 5 amino acid sequence motifs (the 'search-sequon window' or 'sequon window') typically consisting of the last adjacent conserved amino acids are conducted against the PEAKS *de novo* dataset. If the search is unsuccessful, the 'sequon window' is moved by a small increment (here 2 amino acids) into the known sequence. In the case that the search is successful, a 'sequon window' in the newly identified peptide sequence is chosen. In this procedure, the 'search sequon window' is incrementally moved into the gap, while expanding evidence for the previously unkwnown sequence.

The CDR-H3 example illustrates this explanation. Starting from the *N*-terminus of the CDR-H3 gap, first the "FCAR" sequence was tried which corresponds to the end of the available HFR3 sequence, yielding no matches. As this search sequon was unsuccessful, the 'search sequon window' was shifted up towards the N-terminus by 2 amino acid positions. Subsequently, the "TYFC" motif was searched, resulting in the identification of the NEDTATYFCVR peptide (Figure 2.22.H). This suggests an A → V mutation in the *C*-terminal end of the HFR3 domain. This corrected sequence sets the base for the following '*incremental expansion*'. Next, the "FCVR" motif − corresponding to the end of the known sequence − was searched, leading to identification of the FCVRSW peptide (Figure 2.22.I), thus elucidating the first two amino acids of the CDR. Following searches with "TRSW" or other combinations did not result in further matches, as such the same approach was tried from the *C*-terminus. A search for "WGWQT" identified TPF or PTF as *C*-terminal amino acids of the CDR-H3 gap (Figure 2.22.J). A subsequent search for "PTFW" leads to the finding of additional RG or GR amino acids and the *b*-ion series clarifies the order *N*-P-T-*C* (Figure 2.22.K). As such, the proposed sequences VRSWRGPTFW and VRSWGRPTFW, obtained by '*incremental expansion*' from both ends, were assembled in the consensus template and searched against the dataset using PEAKS DB. This led to the identification of several high-scoring PSM for the CDR-H3, two of which shown in Figure 2.22.L, validating the latter of the two sequences.

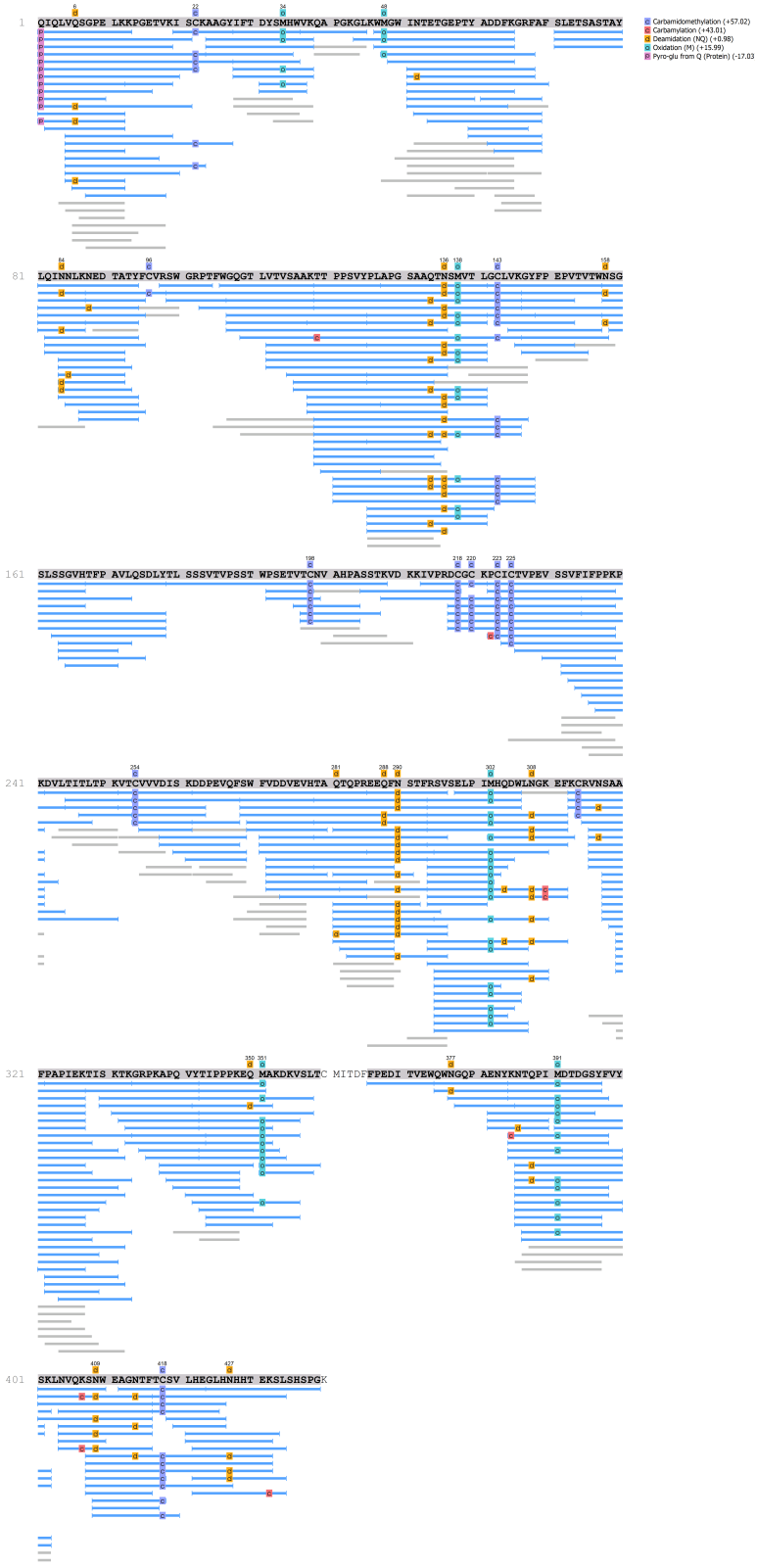


Figure 2.23. Sequence coverage of the consensus template. Note that this search was performed with unspecific cleavage at both end of the peptide.

In summary, our approach of creating a consensus template for the variable domain using two high-scoring templates and correcting these using the MS/MS data allowed us to sequence the variable domain manually. As part of this endeavour, different manual *de novo* sequencing strategies were introduced: '*filling the gap*', '*conserved sequon search*' and '*incremental expansion*'. In common to all of them is they take advantage of modern software tools, expanding the repertoire of *de novo* sequencing strategies described in the literature.[576,577,592] The entire consensus template with the conserved domain derived from template id_1725 was subsequently searched against the entire dataset using PEAKS DB, Figure 2.23 shows the sequence coverage. Noteworthy, whereas the variable domain shows full sequence coverage, two parts of the heavy chain lack sequence coverage and therefore need further investigation: the CMITDF gap at position H390 and the C-terminal K position H478.

2.4.3.3. Investigating sequence coverage gaps in the constant domain

As stated at the beginning of this section, a fundamental assumption underlying all work presented here was made, i.e. that the dataset of 8 digests contains all the data required to solve the sequence (i.e. an implicit 100 % sequence coverage). We will only later see that this assumption is wrong. As such, I initially assumed the sequence − although in the conserved domain − could be incorrect and set out to find alternative sequences contained in the dataset following the strategies described in this chapter.

The search was rendered difficult by common contaminants having overlapping sequences with this domain, hampering identification of *de novo* peptides for the extension of *N*-terminal or *C*-terminal side of the gap (Figure 2.24.A):

Contaminant peptide sequence	Contaminant
MVTSFYPDYIAVEWQR	BSA
YPDYIAVWEQR	BSA
VSLAGACGVGGYGSR	Keratin, type II cytoskeletal 5
VATVSLPR	Trypsin
VSLQDKTGF	Chymotrypsin

Initially, no peptide was found for the entire gap until Q at position H405, only when searching with unspecific cleavage at both ends did peptide FPEDITVEW appear (Figure 2.24.B). As such, the identification of further peptides was investigated by searching for sequence variants following the conservation of heavy chain residues amongst species

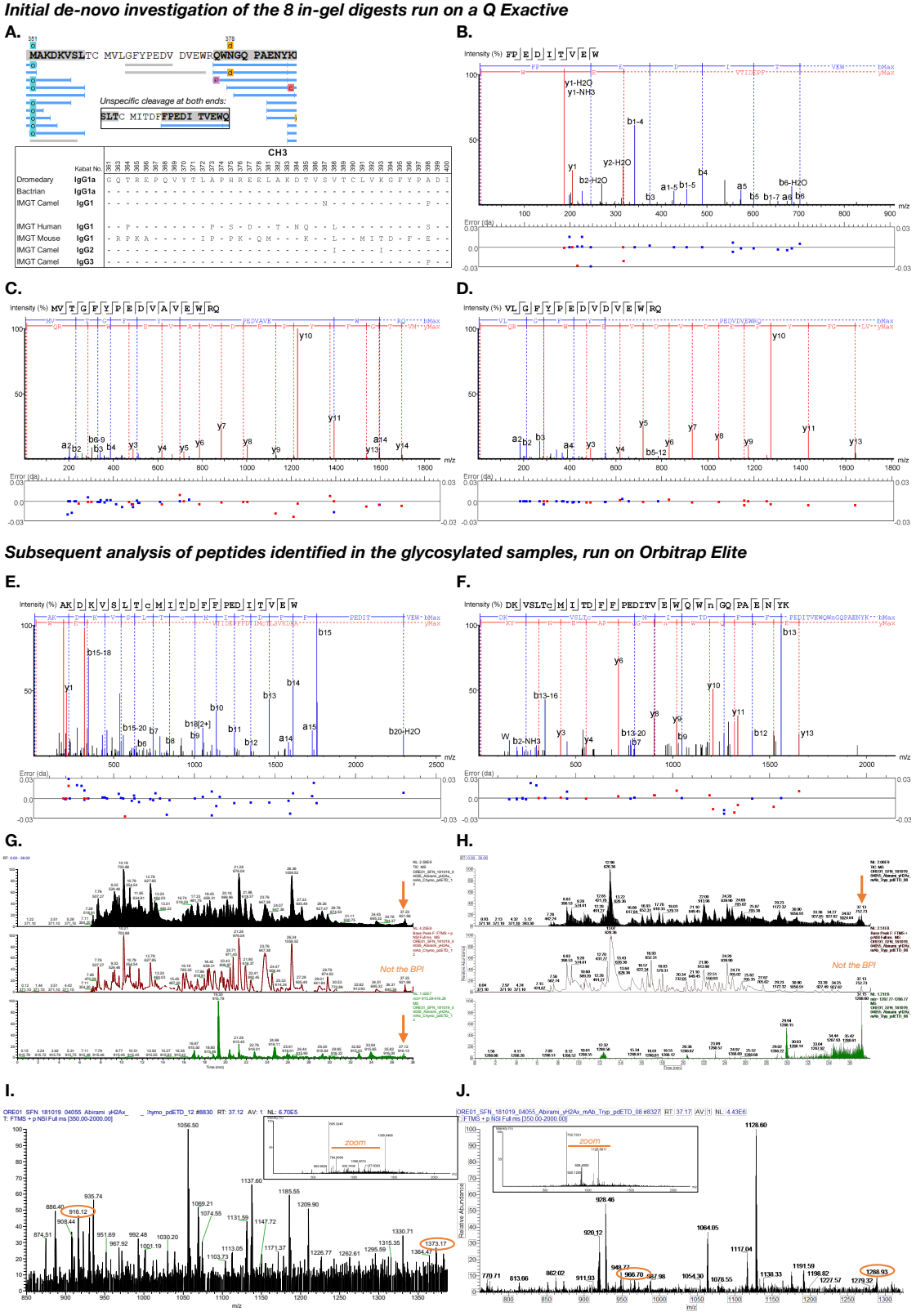


Figure 2.24. Investigating the CMITDF sequence coverage gap in the heavy

Figure 2.24. (***continued***) **chain. A.** The mouse IgG1 is unique in the H390 region among mammalian species, several residues are highly conserved amongst species. Identical residues are displayed as "−". Adapted from Griffin *et al.*[593]. **B.** MS^2 spectrum of peptide FPEDITVEW identified in the AspN.Chymo.Trypsin sample. This peptide was only identified in a database search when unspecific cleavage at both ends was allowed (-10lgP = 27.66 ; RT = 48.15 / 50 min end ; ppm = 0.4). **C.** MS^2 spectrum of *de novo* identified peptide MVTGFYPEDVAVEWRQ in the trypsin sample (-10lgP = 62.54 ; RT = 40.98 / 50 min end; ppm = 4.2). **D.** MS^2 spectrum of *de novo* identified peptide VLGFYPEDVDVEWRQ in the trypsin sample (-10lgP = 64.95 ; RT = 40.71 / 50 min end ; ppm = 1.4). **E.** MS^2 spectrum of database identified peptide AKDKVSLTCMITDFFPEDITVEW in the chymotryptic in-solution digest of the entire glycosylated antibody (-10lgP = 55.86 ; RT = 37.12 / 38min end ; ppm = 2.2). **F.** MS^2 spectrum of database identified peptide DKVSLTCMITDFFPEDITVEWQWNGQPAENYK in the tryptic in-solution digest of the entire glycosylated antibody (-10lgP = 67.45 ; RT = 37.18 / 38 min end ; ppm = 7.5). **G.** Chromatogram for the chymotryptic digest of the entire glycosylated antibody with TIC, BPI and XIC for the above peptide, elution is indicated by an orange arrow. **I, J.** MS^1 spectrum of the precursor ion, zoomed in the region of interest. Peptides of different charge states corresponding to the identified sequence are highlighted in orange.

(Figure 2.24.A), assuming changes in conserved residues would be highly unlikely. Two *de novo* peptides were identified following this strategy, shown in Figure 2.24.C and Figure 2.24.D. However, neither of these two peptides yielded any coverage of the *N*-terminal TCMIT gap, nor did any further *de novo* or SPIDER peptides provide any indication. As such, the meaning of the two identified peptides remains questionable, it is not excluded that these could come from contaminants.

Despite high investment in time, this gap could not be identified in the 8-digest dataset run on the Q Exactive and the *de novo* approach was abandoned. Only when performing glycosylation analysis at a later stage in time, with samples being run on an Orbitrap Elite (HCD fragmentation), a peptide coverage of the entire stretch of the sequence was identified (Figure 2.24.E-J). The data was processed in exactly the same way as before and the chymotryptic peptide presented in Figure 2.24.E and the tryptic peptide in Figure 2.24.F provide high scoring identifications, thus confirming that the conserved murine IgG1 sequence is indeed the correct one.

This identification of the relevant peptides in the in-solution digest of the intact glycosylated antibody on the Orbitrap Elite appeared surprising at first and therefore the elution profile of the peptides was analysed. An important detail is that these samples were run on a recently changed column. Indeed, both hydrophobic peptides elute at the very end of the run, a part that can be lost if the system is not entirely clean (Figure 2.24.E and H). Additionally, these peptides are masked by others in the MS^1 spectrum (Figure 2.24.I and J), with the chymotryptic peptide (left) being quite literally the 10^{th} most intense peak in this top10 run.

Retrospectively, this finding is not surprising accounting for the hydrophobicity of the peptide sequence. A multitude of hydrophobicity scales[196] and retention time prediction methods[594] have been developed, such as indexed retention times (iRT)[595,596]. As an example, the chymotryptic peptide AKDKVSLTCMITDFFPEDITVEW has a calculated hydrophobicity index (HI)[597] of 25.75 (in acetonitrile % units) according to the Sequence Specific Retention Calculator version Q (SSRCalcQ),[598,599] which corresponds to an iRT value of 142.42. These values mean that the peptide is at the hydrophobic extreme end of the respective scales, at which peptides are often not measured or elute in the end of the run.[597, Fig.4C] Similarly, the tryptic peptide DKVSLTCMITDFFPED ITVEWQWNGQPAENYK has a HI value of 23.97 and an iRT value of 129.51. Taking into account the high degree of identity of antibody sequences in the conserved domain, knowledge of this peptide when analysing murine IgG1 by mass spectrometry is important for *de novo* sequencing as it is likely to often be missing − as in the case of our Q Exactive data.

A similar observation can be made for the *C*-terminus, missing the *C*-terminal lysine. Whereas this is a well-documented observation for antibodies expressed in mammalian cells referred to as *C*-terminal lysine clipping by exopeptidases[516,600], important for complement activation[525], we could identify the SLSHSPGK peptide in the glycosylated sample suggesting trace presence of C-terminal lysine (Figure 2.25.A,D). Its elution at the beginning of the run might explain lack of identification in the 8-digest dataset (Figure 2.25.B). Its calculated hydrophilicity index (HI) is of -0.92 and iRT of -50.97, placing this peptide in the top 0.1 % most hydrophilic of database-registered peptides.[597, Fig.4C] The other identified peptide for the C-terminus (Figure 2.25.A,C), NHHTEKSLSHSPKG, is equally highly polar and has a calculated HI of -1.07 and an iRT of -52.05. It is therefore not surprising that both peptides elute at the very beginning of the run (Figure 2.25.E,F) and remarkable that these peptides are retained by the PepMap C18 on-line precolumns. Noteworthy, the C-terminal glycine resulting from clipping is presumably prone to amidation[601], yet this modification was not identified.

Taken together, this data shows that the underlying assumption in this work of all data necessary for *de novo* sequencing being contained in the 8-digest dataset is not entirely correct. Highly hydrophilic and hydrophobic peptides can be missed due to low intensity and/or co-elution with contaminants/other peptides in the LC system. This highlights the importance of chromatographic separation and sequencing speed for *de novo* sequencing projects, along with the < 0.03 Da accuracy achieved by Orbitrap detectors. Additionally, it emphasises the importance of high laboratory standards: the low amount of contaminants along with system performance proved decisive in this study.

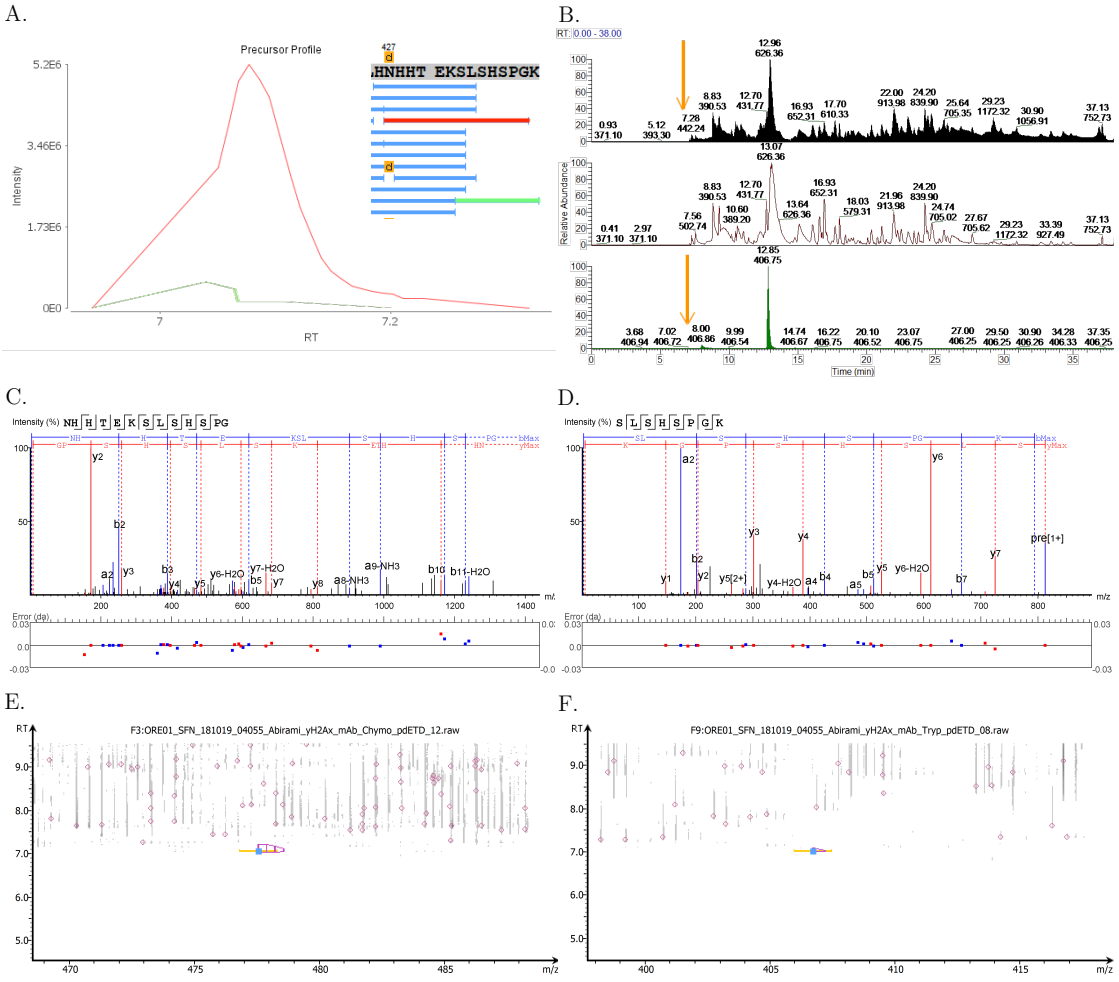


Figure 2.25. Investigation of the C-terminus of the anti-γH2Ax heavy chain. *C*-terminal lysine clipping by exopeptidases[516,600] is a commonly observed post-translational modification on antibodies expressed in mammalian cells. In the case of the anti-γH2AX heavy chain, two peptides corresponding to the C-terminus are identified, one with and one without the C-terminal lysine. Both peptides confirm the correct sequence of the C-terminus, which can vary amongst IgG1 heavy chains. **A.** MS^1 intensity profile (extracted ion chromatogram of the precursor peptide) for both identified peptides covering the C-terminus. Red: NHHTEKSLSHSPGK, green: SLSHSPGK. The sequence coverage at the C-terminus is shown, highlighting the two peptides. The green SLSHSPGK peptide which conserves the C-terminal K has only trace abundance. Nonetheless, this trace presence of the uncleaved peptide confirms the C-terminal K in the genetic sequence. **B.** Both peptides covering the C-terminus elute very early in the C18-nLC chromatogram. The elution time is indicated by an orange arrow. Data for SLSHSPGK, tryptic digest of the intact glycosylated antibody. **C.** MS^2 spectrum confirming the identity of the red chymotryptic NHHTEKSLSHSPKG peptide (-10logP = 33.55 ; RT = 7.06 ; ppm= 0.8). **D.** MS^2 spectrum confirming the identity of the green tryptic SLSHSPGK peptide (-10logP = 60.57 ; RT = 7.02 ; ppm = 1.7). **E, F.** Retention time vs m/z plots highlighting the early elution and highly polar nature of both peptides.

In summary, the sequence of the heavy chain of the anti-γH2Ax murine monoclonal antibody could be solved by MS/MS in its entirety. The final sequence coverage is shown in Figure 2.26.A. Whereas our initial contig assembly with the ALPS system was unsuccessful, a fully manual approach based on high-scoring templates identified in the antibody database using the MS/MS data allowed elucidation of the antibody sequence. This is a further reality check stating that automatisation of antibody *de novo* sequencing has yet to reach its full potential.

2.4.3.4. Comparison with the RNAseq and intact protein MS results

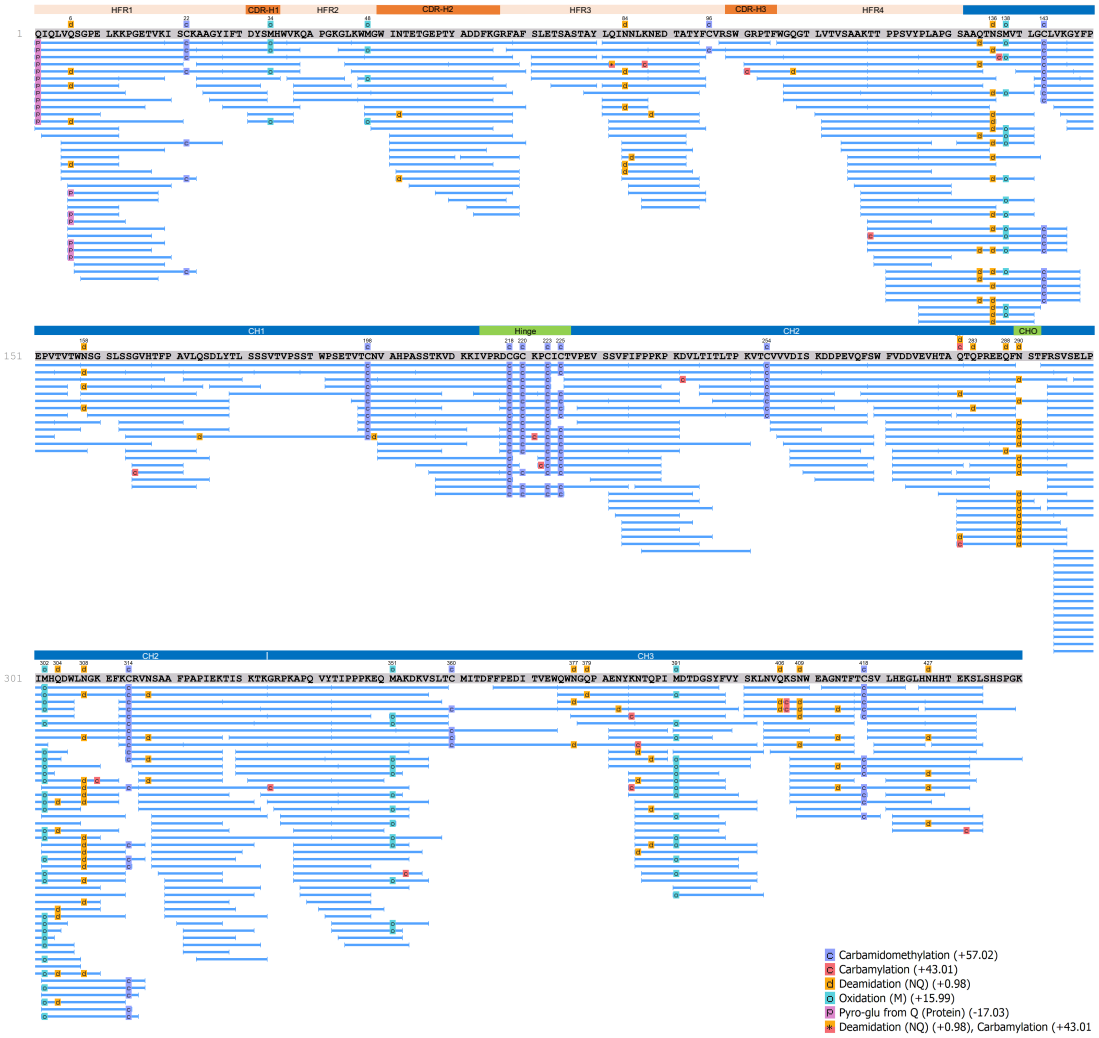
The sequence obtained by manual sequencing is identical to the one obtained by RNAseq, emphasising excellent convergence between both methods.

The sequence was calculated to have a MW of 48674.19 Da. Accounting for *N*-terminal lysine clipping (-128.10 Da[602]) and pyroglutamate formation (-17.03 Da), the resulting mass of 48529.06 is slightly higher than the measured 48518 PNGaseF deglycosylated heavy chain by 11 Da, which is related to the mass error on the Waters Xevo system at the time of use (Figure 2.26.B,C). Nevertheless, the obtained sequence matches the intact mass analysis.

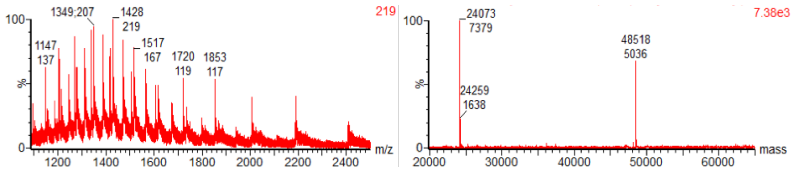
2.4.3.5. PeaksAB fails on most of the anti-γH2Ax heavy chain variable domain

Whereas PeaksAB was able to assemble a protein sequence, it contains several chunks in the HFR2, CDR-H2, HFR3, CDR-H3, HFR4 and constant region which appear to be incorrect − in addition to minor mismatches and inserts (Figure 2.27). The suggested sequence immediately after the initial search undeniably needs further manual elaboration to a large extent, similar to the previously described procedure in 2.4.3. PeaksAB sequencing module does not appear to be superior to our previously described manual approach, independently of the inclusion of the glycosylation analysis runs on the Orbitrap Elite (annotated PeaksAB_gly) that contain missing peptides. However, the interface of PeaksAB greatly facilitates the task of manual analysis, providing ALC scores directly for each amino acid as well as an enzyme-coloured peptide coverage which allows to quickly spot "gaps" indicating the requirement of manual validation. Additionally, PeaksAB build-in tools allow easy alterations in sequence with real-time preview to facilitate manual analysis, removing the need of setting up a new database search with the corrected sequence for each alteration.

A.



B.



C.

Observed
48518 Da (PNGase F)

Calculated
48674.19 Da (sequence)
– 128.10 Da (C-term K clip)
– 17.03 Da (pyro-Glu)

= 48529.06 Da

Figure 2.26. Summary of *de novo* sequencing of the anti-γH2AX heavy chain. A. Sequence coverage by MS/MS showing the identified peptides at 1% FDR (100% sequence coverage), using the 8-digest dataset acquired on a Q Exactive and the dataset acquired on the Orbitrap Elite. The sequence is annotated for the different domains using the Kabat numbering scheme. This sequence is identical to the sequence obtained independently by RNAseq. The peptide identifying the C-terminal lysine is the low abundant peptide discussed. **B.** Molecular weight determination of the PNGase F deglycosylated murine antibody by denaturing intact protein LC-MS, showing the ion series (left) and the deconvoluted spectrum (right). **C.** Calculation explaining the match between the observed mass and the mass calculated under consideration of the dominant PTMs: N-terminal pyroglutamate formation and C-terminal lysine clipping.

```
CLUSTAL O(1.2.4) multiple sequence alignment


Consensus       QIQLVQSGPELKKPGETVKISCKAAGYIFT-DYSMHWVKQAPGKGLKWMGWINTETG---   56
PeaksAB         QIQLVQSGPELKKPGETVKLSCKAAGYTLSSDNSVGWVRQAPGKALEWVGGISSSYGGSA   60
PeaksAB_gly     QIQLVQSGPELKKPGETVKLSCKAAGYSLS-DNSVGWVRQAPGKALEWVGGISSG-GSTG   58
                ********************:******* :: * *: **:*****.*:*:* *.:  *

Consensus       -------EPTYADDFKGRFAFSLETSASTAYLQINNLKNEDTATYFCVRS----WGRPTF   105
PeaksAB         YYASWAKSPTYADDFKGRFAFSIETSASTAYLQLNNLKNEDTATYYCVLANNGVYYAMNF   120
PeaksAB_gly     YINTETGEPTYADDFKGRFAFSIETSASTAYLQLNNLKNEDTATYFCVRANNGVYWNVDA   118
                       .**************:**********:**********:** :    :

Consensus       WGQGTLVTVSAAKTTPPSVYPLAPGSAAQTNSMVTLGCLVKGYFPEPVTVTWNSGSLSSG   165
PeaksAB         WGQGTLVTVSAAKTTPPSVYPLAPGSAAQTDSMVTLGCLVKGYFPEPVTVTWNSGSLSSG   180
PeaksAB_gly     WGQGLLVTVSSASTTAPKVYPLAPCSAAQTDSMVTLGCLVKGYFPEPVTVTWNSGSLSSG   178
                **** *****:*.** *.****** *****:*****************************

Consensus       VHTFPAVLQS-DLYTLSSSVTVPSSTWPSETVTCNVAHPASSTKVDKKIVPRDCGCKPCI   224
PeaksAB         VHTFPAVLQS-DLYTLSSSVTVPSSTWPSETVTCNVAHPASSTKVDKKIVPRDCGCKPCI   239
PeaksAB_gly     VHTFPAVLQSSGLYTLSSSVTVPSSTWPSETVTCNVAHPASSTKVDKKIVPRDCGCKPCI   238
                ********** .**********************************************

Consensus       CTVPEVSSVFIFPPKPKDVLTITLTPKVTCVVVDISKDDPEVQFSWFVDDVEVHTAQTQP   284
PeaksAB         CTVPEVSSVFIFPPKPKNVLTITLTPKVTCVVVDISKDDPEVQFSWFVDDVEVHTAQTQP   299
PeaksAB_gly     CTVPEVSSVFIFPPKPKDVLTITLTPKVTCVVVDISKDDPEVQFSWFVDDVEVHTAQTQP   298
                ****************:*******************************************

Consensus       REEQFNSTFRSVSELPIMHQDWLNGKEFKCRVNSAAFPAPIEKTISKTKGRPKAPQVYTI   344
PeaksAB         REEQFDSTFRSVSELPIMHQDWLDGKEFKCRVNSAAFPAPIEKTISKTKGRPKAPQVYTI   359
PeaksAB_gly     REEQFDSTFRSVSELPIMHQDWLDGKEFKCRVNSAAFPAPIEKTISKTKGRPKAPQVYVI   358
                *****:****************:*********************************.*

Consensus       PPPKEQMAKDKVSLTCMITDFFPEDITVEWQWNGQPAENYKNTQPIMDTDGSYFVYSKLN   404
PeaksAB         PPPKQQMAKDKVSLTCMITNFFPQNITVQWQWNGQPAENYKNTQPIMDTDGSYFVYSKLN   419
PeaksAB_gly     APPQEELSKSKVSLTCMITNFFPQNITVQWQWNGQPAENYKNTQPIMDTDGSYFVYSKLN   418
                 **:::::*.*********:***::***:*******************************

Consensus       VQKSNWEAGNTFTCSVLHEGLHNHHTEKSLSHSPGK-- 440
PeaksAB         VQWSNWQW-DTYTCVVMHEALHNHYTQKSLSHSPGKAS 456
PeaksAB_gly     VQWSN-QW-DTYTCVVMHEALHNHYTQKSTSHSPGK-- 452
                ** ** :  :*:** *:**.****:*:** ******
```

Figure 2.27. Alignment of the suggested sequence by PeaksAB to the deposited anti-γH2Ax mAb heavy chain consensus sequence by Clustal Ω.

2.4.3.6. PTM analysis goes hand in hand with *de novo* sequencing

Antibody protein sequencing allows to identify post-translational modifications, which can have a strong influence on the antibody's activity.[516,517] Common PTM such as deamidation or oxidation are typically monitored for quality control (QC) purposes, yet their role is thought to be minor.[516,517]

PEAKS X as well as PeaksAB include a PTM viewer which allows MS^1 quantification of PTMs using the XIC of modified vs unmodified peptide. [§] Its application to the N-terminal pyroglutamate is exemplary illustrated in Figure 2.28. The biological role of pyroglutamate, if any, is unclear.[516,517] It is historically notorious for hampering Edman sequencing, requiring specialised protocols to overcome the *N*-terminal blockage by this PTM (using pyroglutamate aminopeptidase, PGAP[603]).

The importance of PTM as an intrinsic part of the *de novo* sequencing process has been highlighted in the previous section, including a discussion on *C*-terminal lysine clipping (Figure 2.25). *N*-glycosylation analysis however is a separate process, sequencing tasks being usually performed on the PNGaseF deglycosylated antibody. With the validated sequence in hand, identification and glycoproteomic sequencing of the glycan on anti-γH2Ax was attempted next.

[§]Whereas this approach is not novel, the design of the tool allows for convenient comparision of both XIC intensity and retention time for each peptide pair and quantification. The potential application of such a tool together with a robust LFQ normalisation as for example introduced in the MaxLFQ algorithm could greatly impact on the field of protein modification chemistry. Whereas conversion for reactions in protein modification chemistry is usually calculated from deconvolution of denaturing intact protein mass spectrometry ion series, several cases have been encountered during this DPhil where this was not possible due to difficulties in ionisation, sample complexity or competing side reactions (not discussed in this thesis). A LFQ tool building on the Peaks PTM profile tool could potentially add to the current methodology and help protein chemist to troubleshoot and design better reactions.

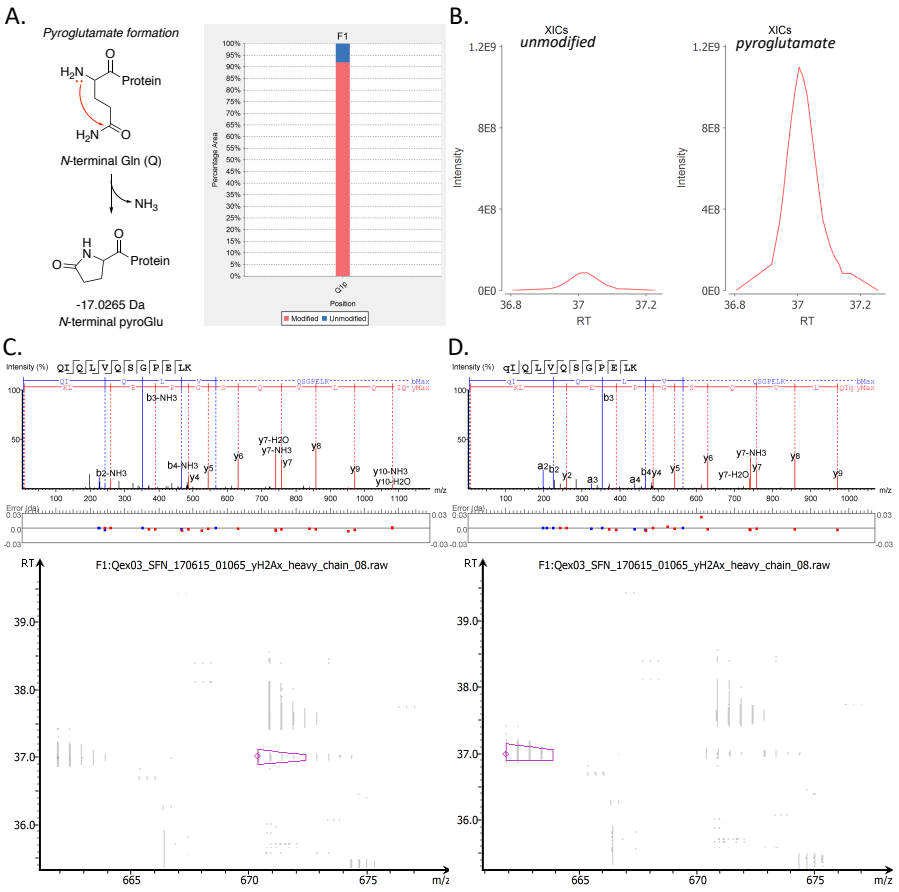


Figure 2.28. *N*-terminal pyroglutamate formation on the anti-γH2Ax heavy chain. **A.** Mechanism of pyroglutamate formation and quantification by integration of peptide MS^1 signal (AUC). **B.** Extracted ion chromatogram (XIC) of unmodified (left) and modified (pyroGlu, right) peptide. **C.** MS^2 spectra of unmodified peptide and 2D chromatogram. **D.** MS^2 spectra of modified (pyroGlu) peptide and 2D chromatogram.

2.4.4. Lessons from the *de novo* sequencing and discussion of its future

Back in 2015, the widely assumed paradigm, expressed for example in the mAbs journal was that:

> Modern mass spectrometry-based sequencing will rarely, if ever, provide a single unambiguous sequence for the variable domains. (Baca and co-workers[542])

Since then, tremendous progress in data analysis techniques, particularly in *de novo* sequencing software and sequence assembly from MS/MS data, has contributed to overcome limitations in *de novo* sequencing of monoclonal antibodies by LC-MS/MS. The work presented in this thesis provides a real-world application and illustrates that it has indeed become a reality, but remains nonetheless a far from accessible technology.

Indeed, the literature on antibody *de novo* sequencing by MS/MS rarely contains case studies[526,529–531,542] in which state-of-the-art software is used to elucidate the sequence of an expressed antibody – RNAseq of hybridoma cells is more established in this context.[521,522] Recent publications on antibody *de novo* sequencing by MS/MS are mainly of methodological nature,[372,527,528,542] introducing novel data analysis procedures[370,538,557] – often repeating the sequencing of known antibodies or antibodies of commonly studied sub-types and thus well-covered in existing databases.[372]

As such, it is difficult to obtain data on the use and application of the different sequencing software in "real world" projects. Many facilities offering *de novo* antibody sequencing by MS/MS use proprietary in-house software and expertise. It appears that the commercial market leaders in *de novo* antibody sequencing software by MS/MS, Novor[366]+REmAb, PeaksAB and SuperNovo[535], are very popular among biopharmaceutical manufacturer that can afford the high license fees.¶

The order of sections in this results part is sorted by increasing complexity of the *de novo* sequencing task, to highlight the increasing difficulties, or failure, of automated tools. The underlying conclusion is that in cases that are not straightforward, two of the best tools available still require manual analysis. The difficulty lies less in the *de novo* sequencing of spectra (alternatives such as Novor[366] or DeepNovo[538] are available) than in the assembly of the sequence from the identified spectra. This does not come as a surprise: all commercial antibody protein sequencing services such as from RapiNovor,

¶personal communication by Sanjay Sanghani, Protein Metrics Inc.

Bioinformatics Solutions Inc, Absolute Antibody, Alphalyse, Creative Biolabs and many more promise manual validation. For this reason, LC-MS/MS will likely remain the 2nd choice technology after RNA-seq. However, in many cases, i.e. where the hybridoma is not available, MS/MS has established itself as a powerful alternative.

A potential improvement could thus come from increasing the performance of sequence assembler. Important differences exist between the *de novo* sequencing of proteins from other families vs. antibodies due to genomic encoding, high mutations, variability and structure. The ALPS De Bruijn assembler[372] for example does not include any antibody-specific approaches, such as domain definitions or template-based approaches. PeaksAB has elements thereof and therefore performed better. Yet an antibody-centric approach was key for the success of the manual *de novo* sequencing presented in this work: frequency data and deposited germline sequences, accessible via abYsis[589] and IgBLAST, respectively, enabled corrections in the CDRs and FR regions. Templates allowed to fill in gaps in conserved domains and parts of the FR regions.

Similarly, the importance of PTMs beyond "decorations" might be very relevant to the sequencing process itself. The identification of pyroglutamate settled the N-terminus of the anti-γH2AX heavy chain. Known deamidation sites, annotated in abYsis, facilitate N/D and Q/E discrimination.

So far, this discussion has been following the trend of the field to evermore sophisticated data analysis methods. Yet, one conclusion from our sequencing effort is that the project' success is determined by the quality of the acquired data, and particularly the chromatographic retention and separation of all peptides. We show that peptides on the C-terminus and a hydrophobic patch in the conserved domain were missing on Q Exactive datasets, highly used in the laboratory, but could be identified on an Orbitrap Elite equipped with a new column thus taking better into account the hydrophobic range of the peptides. These details impact on whether 100 % sequence coverage can be achieved. In principle, this holds true for any proteomics experiment; the absolute threshold is less forgiving in *de novo* sequencing.

Whereas the field of antibody *de novo* sequencing by MS/MS has obvious links to commercial interests − as a service technology or for intellectual property − neglecting its impact on the academic world would be a fallacy. Multiple reports[604,605] and indeed an international appeal[606] have alerted the scientific community of serious issues concerning the reliability of antibodies used as laboratory reagents. This is widely regarded[606] as one of the main causes of the reproducibility crisis[607,608] in modern science. Estimates suggest an annual cost of US$ 350 million to biomedical research

in the U.S.A. only.[606] Many voices therefore ask for an increase of recombinant antibodies, offering higher purity, reproducibility and animal welfare.[606] This will require a tremendous amount in sequencing. Idealists might wish for resellers of antibody reagents to perform this sequencing and recombinant expression, or provide access to the hybridoma (if any) for external RNA-seq. Yet many will doubt all companies will be willing to stem this effort without financial incentive, let alone placing the sequences in the public domain. Absolute Antibody has been a notable exception in their pioneering effort in recombinant antibody expression. Considering *de novo* antibody sequencing by LC-MS/MS requires only a few µg of the commercial protein sample, without substantial barriers in intellectual property, the popularity of this technique is likely to increase amongst academic labs with its accessibility.

2.4.5. Glycosylation analysis

Several methods have been developed for antibody glycosylation analysis, such as LC-MS/MS of digested glycopeptides[552], LC-MS/MS of released and labelled glycans,[609] native MS[553–556] and top-down MS/MS.[557,558] These techniques all have their advantages, no consensus and indeed a notable variation in the results exists.[559]

We performed both intact denaturing protein LC-MS as well as a bottom-up tandem mass spectrometry to obtain site-localisation of the glycan, glycan identity and glycan connectivity.[552] Three different enzymatic digests, namely trypsin, chymotrypsin and elastase as well as three different acquisition methods were tried on the Orbitrap Elite:

- Resonance-activated (ion-trap) CID, to identify the glycan and obtain a fragmentation pattern indicating the connectivity of the saccharides. We reasoned that a tryptic peptide at the site of interest would have three acidic amino acids and only one Arg leading to an expected charge state of 2^+, in which case resonance activated CID has been reported previously to be a sensible choice.[115]

- A CID-ETD data-dependent decision tree (DDT) previously described.[272] This approach triggers an ETD scan whenever reasonable from a peptide fragmentation perspective.

- The HCD-pd-ETD method.[305,610] This approach takes advantage of the distinct preferential fragmentation of the glycan and the peptide chain following the complementarity paradigm of glycoproteomics, which suggests glycan fragmentation in a collision-induced type fragmentation experiment and backbone cleavage leaving the glycan intact in an ETD fragmentation experiment.[115]

It should be mentioned that using the option "show all N-glycopeptides" in Byonic was decisive. It prevents the removal of glycoforms from the results list, which particularly affected low-energy CID results. In addition, it was noted that Byonic has difficulties with product-dependent methods, typically only showing one of the two scans for the same glycopeptide precursor ion (mostly HCD). Identification of the second scan, mostly ETD needs explicit search or annotation directly from the raw file − the scan number is known and the precursor identified.

Chymotryptic digests did not reliably identify glycopeptides, presumably due to the F.NSTF.R sequence. Several spectra contain oxonium ions, yet peptides are difficult to identify.

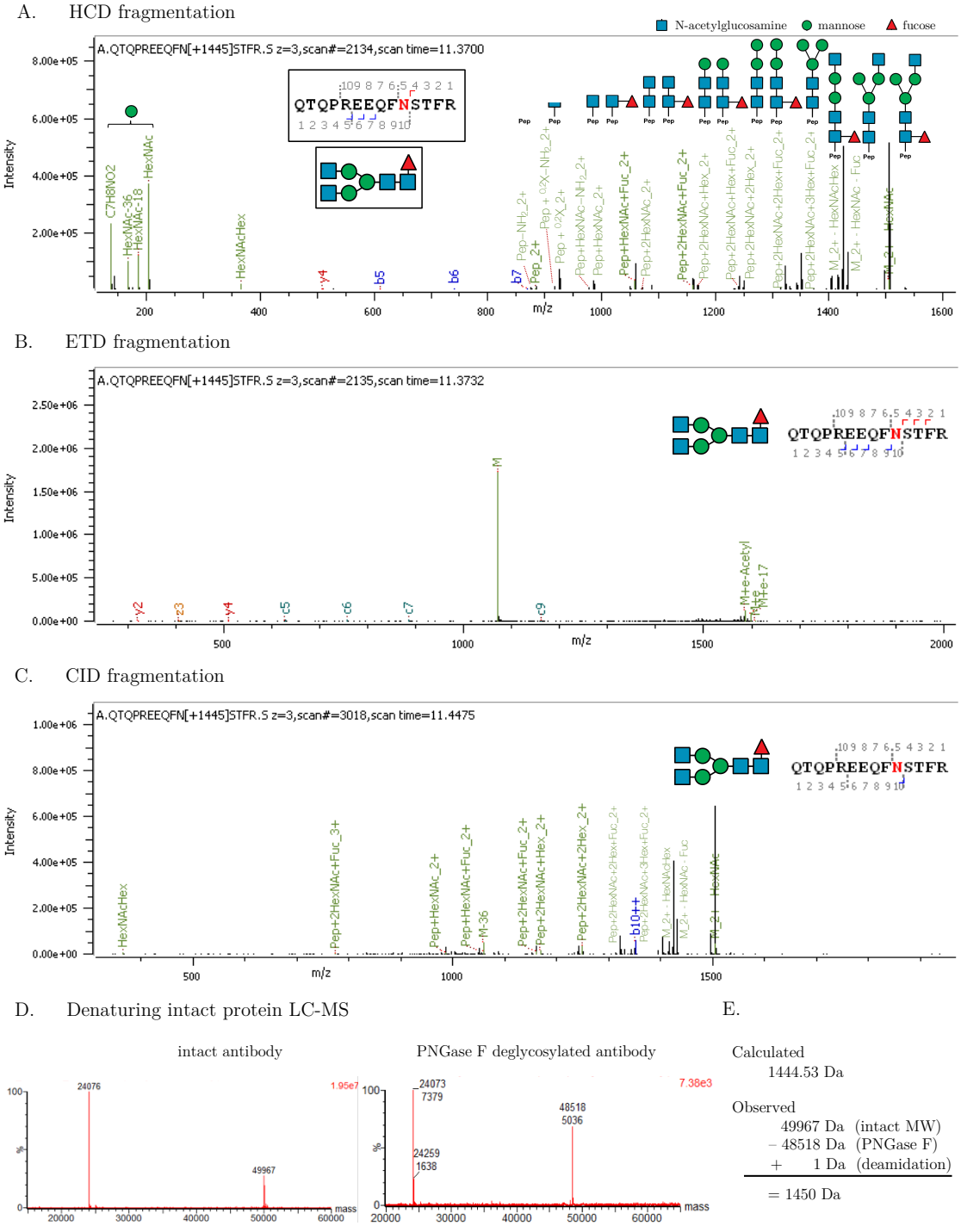


Figure 2.29. Glycosylation analysis of the murine anti-γH2Ax monoclonal antibody. A & B. Elastase digest analysed with the HCD-pd-ETD method on an Orbitrap Elite reveals G0F glycosylation (Oxford name FA2G0). **A.** HCD fragmentation shows glycan connectivity and core fucosylation (PEP 0.015, Byonic score: 198). The presence of the oxonium ions triggers the subsequent ETD scan. **B.** ETD fragmentation reveals site-localisation of G0F glycan at Asn314, identified in elastase digest run on an Orbitrap

Figure 2.29. (***continued***) Elite using HCD-pd-ETD (PEP 0.015, Byonic score: 56). **C.** CID fragmentation shows a comparable fragmentation pattern to CID (PEP 0.013, Byonic score 87). **D.** Molecular weight determination by intact protein LC-MS. Both peaks in each spectrum correspond to either heavy or light chain. **E.** Calculation showing the G0F glycan form matches the observed mass difference.

Elastase digestions identified the G0F (Oxford name FA2G0) glycan using HCD-pd-ETD. Figure 2.29.A shows the HCD fragmentation pattern for the glycosylation, suggesting core fucosylation. The stereochemical identity of the monosaccharides is inferred from the literature and known biosynthesis (see 1.1.2.2). To confirm site localisation of the glycan − expected at the NST site (Asn314) − ETD fragmentation was performed on an Orbitrap Elite. As expected, the glycan was left intact and is identified in Figure 2.29.B as a +1445 Da modification. This suggests a glycan composition of Hex(3)HexNAc(4)Fuc(1) of calculated MW of 1444.53 Da. This glycoform was also identified in both CID and DDT datasets. However, as can be seen in Figure 2.29.C, the spectrum is more noisy, missing the low mass end (due to 1/3rd rule) and has lower fragment intensities for shorter glycan fragments.

Protein LC-MS gave a mass of 49967 Da for heavy chain with glycan, which was lowered to 48518 Da upon treatment with PNGase F (Note that there is a 6 Da mass error on the light chain). Accounting for the deamidation at Asn to Asp (0.98 Da) by hydrolysis of the glycosidic bond by PNGaseF, intact protein LC-MS suggests a glycan mass of 1450 Da (Figure 2.29.D,E). This matches the mass and glycan identified by MS/MS on elastase digests and confirms that the G0F is indeed the major glycoform present.

Noteworthy, this identification was not found in all digests. The tryptic digests yielded the identification of the FA1G0 glycan by both HCD-pd-ETD and CID (Figure 2.30). The G0F glycopeptide was not identified in the tryptic samples. This could either be a minor glycoform or a hydrolysis product, the trypsin samples being run after elastase digestions while kept in 5 % DMSO 5% formic acid at 4 °C. This glycoform was also identified in the elastase digest analysed by DDT, making it more likely to be a minor isoform.

In summary, by combining intact protein LC-MS and glycopeptide MS/MS, the *N*-glycans on the murine anti-γH2AX could be identified.

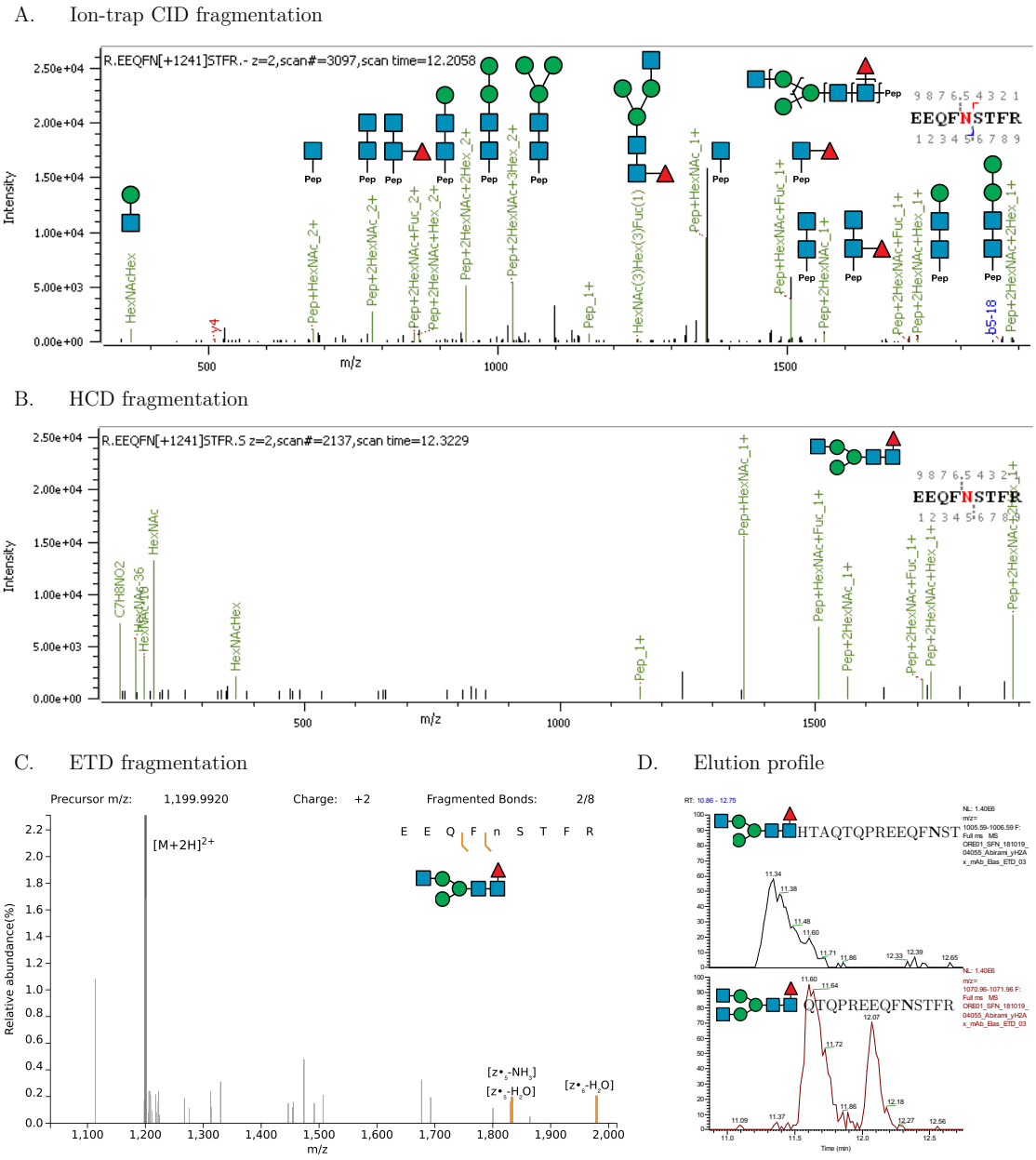


Figure 2.30. Analysis of the second glycopeptide of the murine anti-γH2Ax monoclonal antibody. Tryptic and elastase digests analysed with HCD-pd-ETD or CID method on an Orbitrap Elite reveal additional FA1G0 glycopeptide. This could either represent a hydrolysis product with loss of terminal HexNAc or an alternative, minor glycoform. **A.** Ion-trap CID MS2 spectrum (Orbitrap Elite) identifies FA1G0 glycan on the tryptic peptide EEQFNSTFR (PEP: 1.3e-05, Byonic score: 228). **B.** HCD fragmentation in the HCD-pd-ETD method. (PEP: 0.032, Byonic score 148.6) **C.** The ETD fragmentation from the HCD-pd-ETD method, triggered by the oxonium ions shown in B, could not be annotated by Byonic. Instead, the Interactive Peptide Spectrum Annotator (IPSA)[611] was used for the spectrum following the HCD scan in the RAW file. The fragmentation efficiency of ETD on this peptide is poor. **D.** Elution profile of the two glycopeptides identified in the elastase DDDT run, where both glycoforms were detected.

2.4.6. Molecular modelling of the antibody structure

Having performed *de novo* antibody sequencing and glycosylation analysis, the next steps are focussed on generating a 3D model of the antigen recognition site as well as a full 3D model of the entire antibody.

2.4.6.1. Homology modelling and antigen-antibody docking

With the protein sequence in hand, the first step was to create a homology model of the variable F_V region. The Ig fold of the domains being conserved and the CDR loops except H3 forming a very limited set canonical structures, specialised algorithms have been developed for structure prediction of the variable domain. These focus essentially on template-based modelling of the CDR loops and particularly the prediction of the hypervariable CDR-H3 loop, essential for correct binding predictions.[567] For this, we compared the result of three different tools: abYmod, ABodyBuilder[569] and Rosetta Antibody[568]. Although this list is not exhaustive and other tools have been developed, these are freely accessible academic tools. The rationale for these choices is that Rosetta Antibody is a well-known tool in the community, ABodyBuilder is developed in Oxford and abYmod is the newest tool created by Andrew Martin's group at UCL − a long-standing reference in the field of antibody modelling.

All three tools generated highly similar homology models, only slight divergences in the H3 loop could be noted (see appendix 2.11.4). The abYmod model being the closest structure to the average of the three methods, this structure was retained. Moreover, the abYmod interface is the most feature-intense, it offers the choice to carefully select which template to use, and gave a confident prediction value for the CDRs. As a matter of fact, most of the automatically suggested templates by abYmod were dismissed and replaced by template sequences that are closer from their amino acid chemical structure or displaying highest resolution. The resulting antibody F_V structure is shown in Figure 2.31.A and the CDR's highlighted.

The PDB was then mined for structures of murine IgG1κ antibodies in complex with peptide, to identify the broader binding site location and experimentally observed binding poses of the peptides (Figure 2.31.B). Residues susceptible to participate in phosphoserine recognition[612] − the experimentally confirmed key driver for specificity and binding affinity of the antibody[610] − are highlighted in Figure 2.31.C. Two potential sites of coordination emerge from this view: in proximity of Ser91 / Tyr96 or in proximity of His35, Ser33 and Ser95.

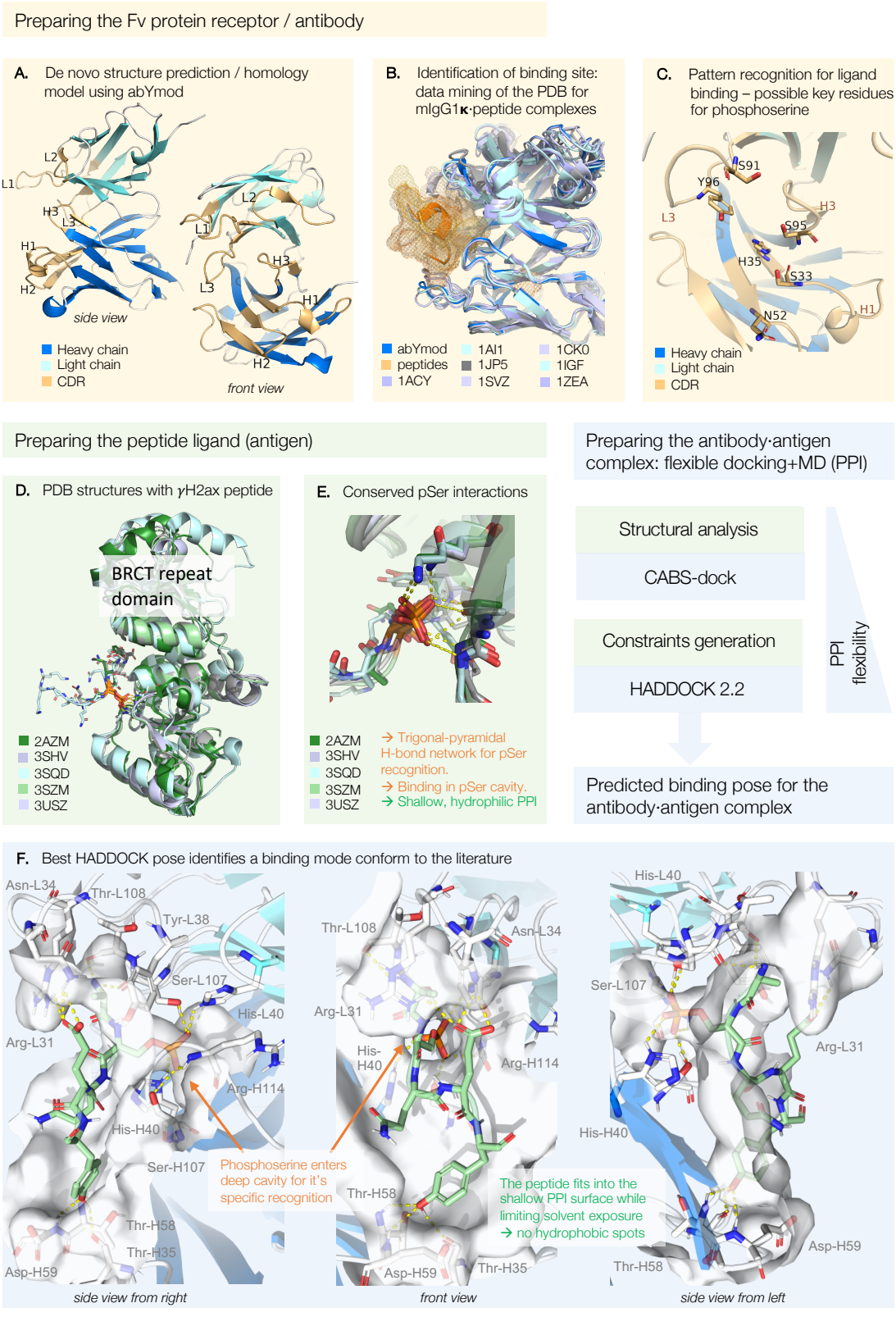


Figure 2.31. Homology modelling followed by antibody docking suggests a phosphoserine-specific antibody-antigen recognition. Top. Homology model genera-

Figure 2.31. (***continued***) -tion and preliminary analysis, highlighted in light orange. **A.** Front view and side view of the F_V homology model generated using the abYmod tool. CDRs are highlighted in gold. **B.** Identification of peptide binding sites for murine IgG1κ-type antibodies based on structures deposited in the PDB. All peptides bind in a groove between the CDR loops − as expected for peptidic antigens. **C.** Amino acid side chains with promiscuity for phosphoserine coordination. The visualisation of these side chains allows to spot potential binding sites of phosphoserine in the antigen-binding region. **Center.** Preparation of the experimentally studied *C*-terminal γH2AX peptide pSQEY, highlighted in light green. **D.** Known PDB structures with γH2AX peptides bound, all co-crystallised with the BRCT repeat domain. **E.** The recognition of phosphoserine has a typical pattern of H-bonds in a specialised basic cavity for phosphate binding. The coordination of the negatively phosphate inside the protein structure is a major enthalpic driver for the binding. **Bottom.** Modelling of the complex and binding pose prediction, highlighted in light blue. **F.** Best-ranked binding pose for the HADDOCK docking, highlighting the key interactions of the peptide (green) and its positional features. Polar contacts are shown in yellow. Image generated using PyMol. IMGT numbering.

With the structure of the F_v protein in hand, we aimed at investigating the antigen recognition by the anti-γH2Ax mAb. Experimental studies are mainly based on the pSQEY peptide, therefore we also used this peptide as the antigen in this computational study. Indeed, several structures have been deposited in the PDB containing this γH2Ax *C*-terminal peptide or variants thereof in complex with the BRCT recognition domain. The peptide binds in a shallow hydrophilic groove at the protein surface, which bears a deeper cavity specific for recognition of the phosphate group (Figure 2.31.D). Phosphoserine is recognised in a typical trigonal-pyramidal H-bond network (Figure 2.31.E).

Having prepared both the antibody receptor and the antigen peptidic ligand, the next endeavour was to identify the binding pose of the ligand epitope in the receptor paratope. Again, two different approaches were chosen: i) site-directed protein-peptide docking with CABS-dock − a docking protocol specialised on peptide-protein docking; and ii) protein-peptide docking with MD refinement in HADDOCK. The latter is able to incorporate experimental data into the docking process in form of docking restraints and use these to guide the docking process alongside traditional energetics and shape complementarity. Of note, HADDOCK support phosphoserine, but neither do CABS-dock, nor the widespread SnugDock (Flexible Rosetta Docking) which was therefore not used.

In summary, no plausible docking poses were generated with CABS-dock. In HADDOCK, only the best-ranked cluster (with multiple poses) showed a plausible docking conformation (appendix 2.11.5). As shown in Figure 2.31.F and Figure 2.32, the peptide binds in a shallow grove at the surface and negatively charged phosphoserine enters a deep, positively charged cavity in which it is recognised by the previously described

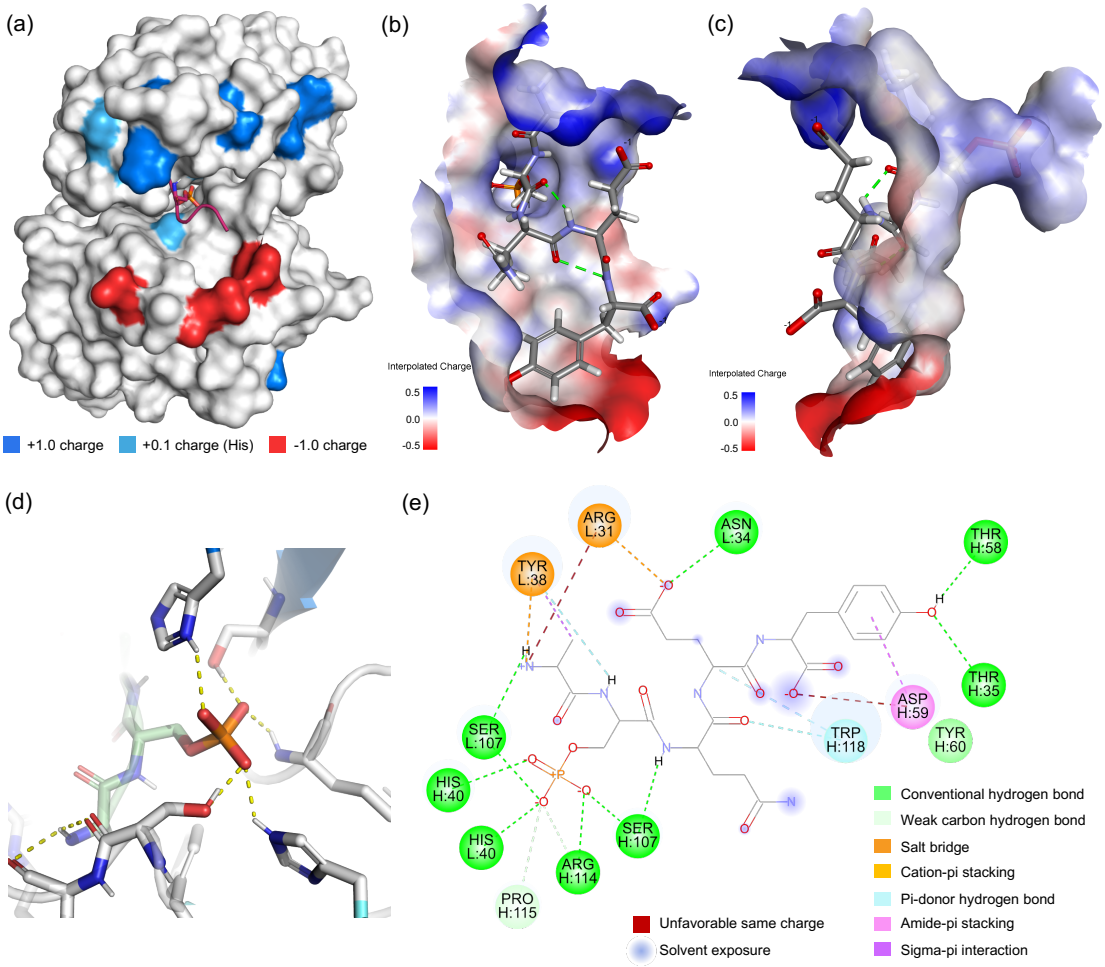


Figure 2.32. Analysis of the antibody-antigen interactions as predicted by ligand docking. Results obtained by flexible docking (using HADDOCK) of the ApSQEY peptide into the homology model of the murine antibody (abYmod model). This phosphopeptide corresponds to the C-terminal end of γH2AX with phosphorylation at S139. **A.** Protein surface depiction highlighting the spatial separation of charged residues around the active site, coloured by their formal charge. Positively charged residues are located on the side that comprises the phosphate binding pocket. Negatively charged residues are located on the opposite side, below the pocket, where the C-terminal end of the peptide is accommodated. Image generated using PyMol. **B.** A detailed front view of the electrostatics in the binding site, highlighting the positively charged character of the phosphate binding pocket. Image generated with Biovia Discovery Studio visualizer. **C.** The side view emphasizes the specificity of the recognition of the phosphate through a deep, positively charged cavity. Image generated with Biovia Discovery Studio visualizer. **D.** Structural details of the phosphoserine recognition inside the basic pocket. The H-bond network fixes the phosphoserine at this position, anchoring the peptide via this amino acid. **E.** 2D ligand interaction analysis schematising the contacts in the protein-ligand complex, coloured by interaction type. Image generated with Biovia Discovery Studio visualiser. IMGT numbering.

typical H-bond network (Figure 2.31.E, Figure 2.32).

The structural recognition of phosphoserine has been thoroughly studied in 14-3-3, WW, FHA and other phosphopeptide binding domains; molecular details underlying the specific recognition of phosphoserine peptides over non-modified ones − as is the case for the anti-γH2Ax mAb − are well understood.[613] In our antigen binding model, the negatively charged phosphoserine group binds in a positively charged pocket (Figure 2.32B,C). Each of the residues in this basic pocket interacts directly with the peptide phosphate group, forming five hydrogen bonds to all three of the phosphate oxygens. This strong network of interactions forms the critical anchor that orients the remainder of the peptide.[613]

This is not surprising, the recognition of the γH2AX C-terminal pSQEY peptide by MDC1 and MCHP1 has been studied structurally.[610] These studies emphasise the recognition of Glu, Tyr and the C-terminus of the pSQEY sequence.[614] The mentioned salt bridge from glutamate of the antigen can be found in our structure with the arginine and lysine in the L1 loop. Together with the phosphoserine recognition, this is a major enthalpic contribution for the peptide binding and provides a strong case for this binding pose to be correct or close to correct.

Whereas HADDOCK does include a MD refinement step of the final structure with explicit solvent (50 ps), a more extensive MD simulation starting from the proposed docking pose might be necessary to further validate the structure, including explicit solvent (correct entropy of PPI) and either longer simulation times allowing for conformational changes (100 ns) or using extensive sampling methods such as REMD or Monte-Carlo.[571] Similarly to HADDOCK, a clustering analysis followed by MM-GBSA[615] or MM-PBSA[616] estimation of ligand-binding affinity would be a practicable high-end molecular modelling confirmation to add further certainty.

A possible practicable experimental validation of the antigen-antibody recognition could potentially be attempted in a future study by crosslinking MS of the antigen with the antibody[617] and / or hydrogen-deuterium exchange MS[78,600]. The gold standard for structural validation would be co-crystallisation of the papain-cleaved Fab fragment with the antigen, although this particular experiment would be limited by the low amount of materials accessible.

2.4.6.2. Bridging the gap: Combining structural and germline frequency data to explain antibody binding specificity

Antibodies against γH2AX have emerged in 1999,[505] the clone JWB301 sequenced in this study was developed in the early 2000 by Upstate Biotechnology Inc,[618] later acquired by MilliporeSigma (Merck KGaA) and has been used in more than 2300 publications. Its use for radioimmunoconjugates that target γH2AX *in vivo* has highlighted the potential of this antibody in preclinical and potentially clinical setting.[619] This section aims at obtaining a molecular understanding of the underlying principles in antigen recognition for the anti-γH2AX mAb using a structural genomics approach by combining germline frequency information with structural data from the modelling.

In a landmark study in 1995, Lara-Ochoa *et al* demonstrated that two distinct frequency distributions exist for amino acids in monoclonal antibodies.[620,621] The frequency distribution of structure-determining amino acids, i.e. at those positions that determine the Ig fold or are determinants for the CDR canonical structures, follows an inverse power law. In contrast, the frequency distribution at sites responsible for antigen specificity and thus recognition follows an exponential law.[620,621] Therefore in reverse conclusion, the frequency distribution of all amino acids at a given position is indicative of the role of the amino acid that is present at this position. This frequency distribution builds on the standardised numbering schemes and is catalogued in antibody databases, easily accessible by tools such as abYsis.[589]

When considering specificity-determining amino acids (i.e. those with exponential distribution), a second effect has to be taken into account. The antibody maturation process involves selection against a particular antigen in iterative rounds, dismissing antibodies with poor selectivity or cross reactivity.[501] Which antibodies enter this process is a result of germline rearrangement and somatic mutations, amongst other. Therefore, there is no even distribution of frequencies at a given site, some amino acids will be present more often in certain positions before entering the selection process.[622] Therefore, if unusual or very rare amino acids remain at the end of the selection process, it might be because they have an important function in antigen recognition, meaning that specificity primes over natural occurrence.[623]

Combining these two effects together with structural data enables additional interpretation which interaction is likely to be important, what its function could be and which residues could contribute to antigen specificity. This structural genomics analysis is relevant for later humanisation, enabling to pinpoint which amino acids are structural determinants and which ones are critical for antigen recognition, expanding the

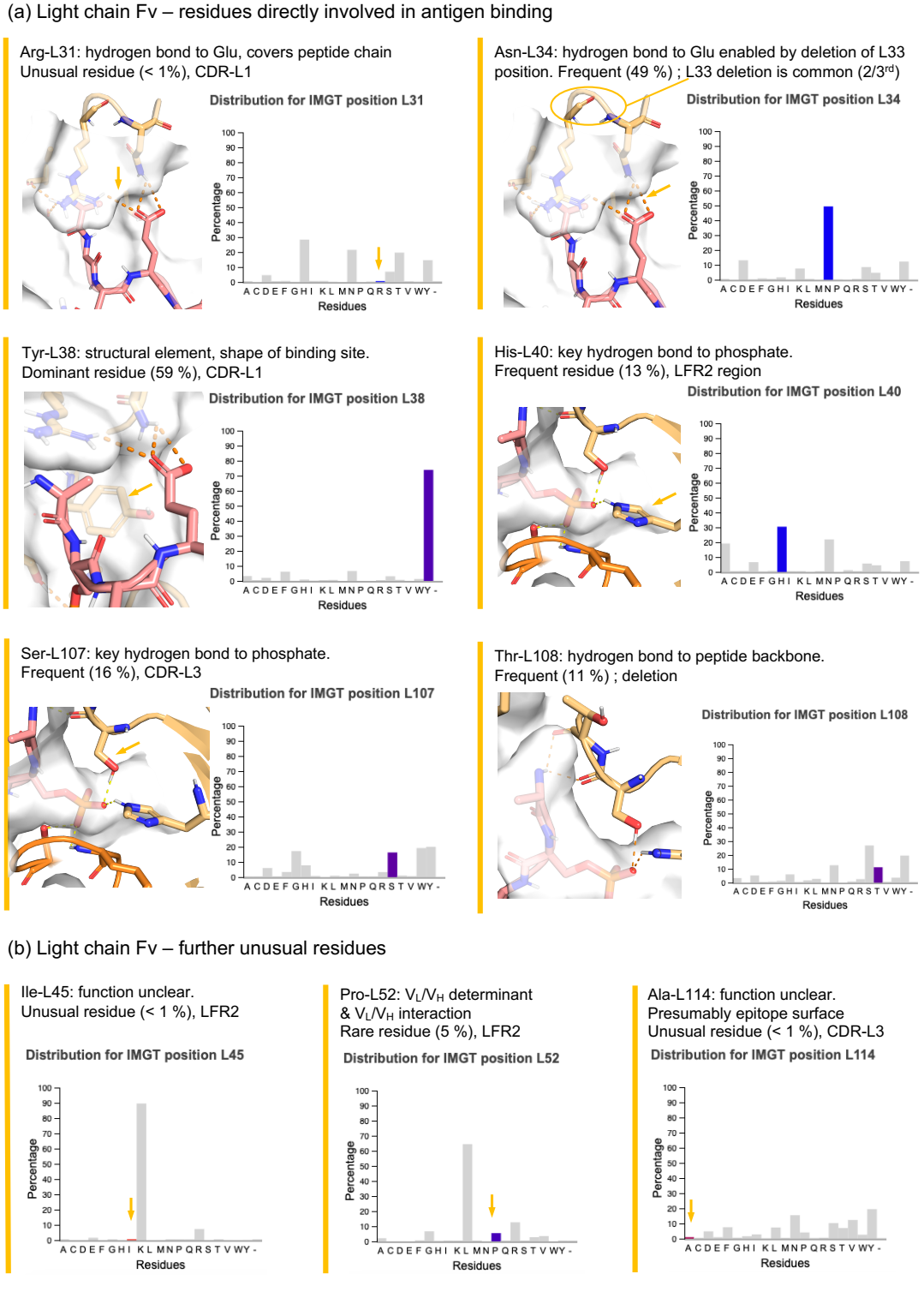


Figure 2.33. Structural genomic analysis of the murine anti-γH2AX mAb light chain. Integrating the structural recognition of the antigen with germline frequency data, which indicates how often a residue is found at a given position, allows to pinpoint which interactions are likely to be unique for the murine anti-γH2AX antibody. In this figure,

Figure 2.33. (***continued***) the frequency of occurrence of residues directly involved in antigen binding across known murine antibodies is indicated together with their key interaction for antigen recognition. **(a)** For each key interaction of light chain residues, a structural representation − as predicted by the docking pose − is indicated together with its natural occurance in *mus musculus*. The frequency of occurrence of the residues directly involved in antigen binding across known murine antibodies was obtained through abYsis.[589] **(b)** Unusual and rare residues of the VL domain that are not directly involved in antigen binding and might contribute to the specificity by other means. Another particularity of this antibody is the very short CDR-L2 loop, consisting of only 2 amino acids (K,V). IMGT numbering.

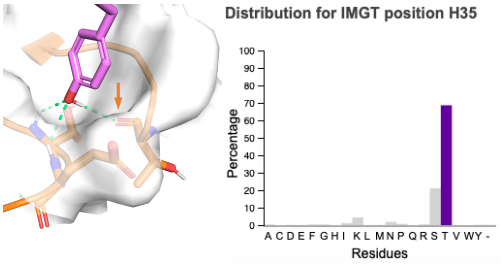
decision-making for humanisation beyond simple sequence analysis. Structural determinants can potentially be changed when transposing to another organism provided the overall structure is kept intact.

Figure 2.33 shows the analysis for the light chain. As an example for a structural determinant, the L38 position exemplarily shows an inverse power law distribution of frequency and Tyr is the dominant residue. The docking structure shows that Tyr affects the shape of the binding site decisively. It does however not contribute to any antigen-specific interaction. His-L40, Ser-L107 and Thr-L108 all show exponential distributions, indicating that these positions are contributing towards antigen-specificity. Indeed, all three are key residues for coordinating the phosphate group. Arg-L31 is another example of an exponential distribution, however this is a rare residue and therefore likely to be important for the recognition of the antigen. The modelling highlights its key role in forming a hydrogen bond or salt bridge to the glutamate of the antigen, further stabilising the peptide in a conformation where the phosphate can bind inside the pocket.

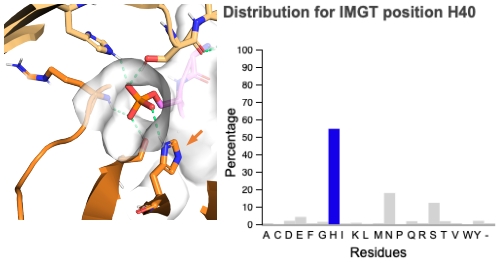
The heavy chain F_V stands out by a high number of unusual or rare residues. This is particularly the case for the position Thr-H58: a structural residue (Pro) is replaced by Thr to enable the correct interaction of Thr-H58 and Glu-H59 to the antigen tyrosine hydroxyl. Similarly as for the light chain, the interactions with the phosphate involve residues with exponential distribution, highlighting their importance for specificity. His-L40, which is a structural residue nonetheless making a key interaction to the phosphate, illustrates that both structural and frequency data together enable the better comprehension of the antibody function.

(a) Heavy chain Fv – residues directly involved in antigen binding

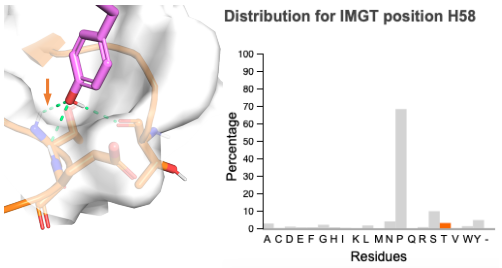
Thr-L35: hydrogen bond to Tyr
Dominant residue (69 %), CDR-H1



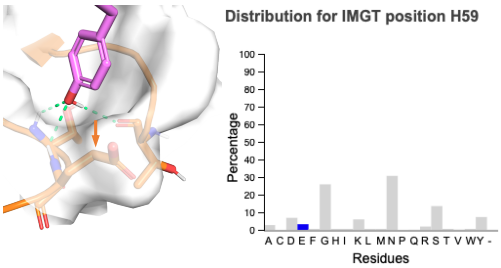
His-H40: hydrogen bond to phosphate
Dominant residue (69 %), CDR-H1



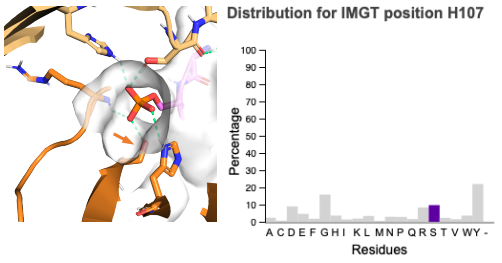
Thr-H58: hydrogen bond to Tyr
Rare residue (3 %), CDR-H2



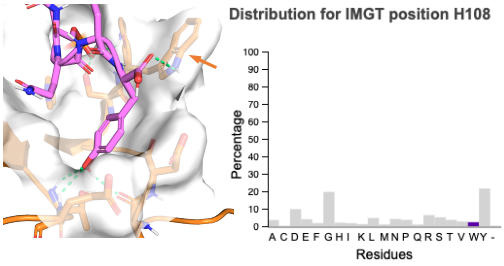
Glu-H59: hydrogen bond to Tyr
Rare residue (3 %), CDR-H2



Ser-H107: hydrogen bond to phosphate
Occasional residue (10 %), CDR-H3

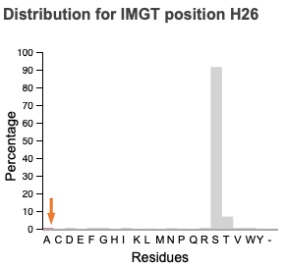


Trp-H108: binding site topology at C-terminus
Rare residue (2 %), CDR-H3

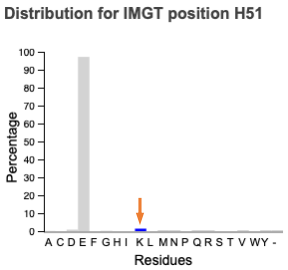


(b) Light chain Fv – further unusual residues

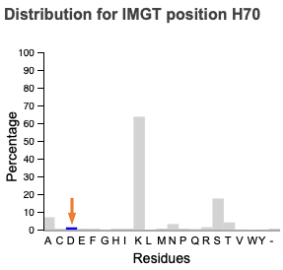
Ala-H26: Immediately before CDR1
Unusual residue (< 1 %), HFR1



Lys-H51: change of charge at V_L/V_H
Unusual residue (1 %), HFR2



Asp-H70: V_H/V_L determinant
Unusual residue (< 1 %), HFR3



(c) Light chain Fv – frequency distribution in the CDR's

A	G	Y	I	F	T	D	Y	S
H26	H27	H28	H29	H30	H35	H36	H37	H38
<1%	96%	66%	2%	81%	69%	20%	83%	1%
CDR-H1								

W	I	N	T	E	T	G	E	P	T
H55	H56	H57	H58	H59	H62	H63	H64	H65	H66
6%	94%	18%	3%	3%	4%	74%	4%	2%	4%
CDR-H2									

V	R	S	W	G	R	P	T	F
H105	H106	H107	H108	H109	H114	H115	H116	H117
4%	80%	10%	2%	13%	2%	<1%	1%	1%
CDR-H3						CDR-H3		

Figure 2.34. Structural genomic analysis of the murine anti-γH2AX mAb heavy chain. Integrating the structural recognition of the antigen with germline frequency data,

Figure 2.34. (***continued***) which indicates how often a residue is found at a given position, allows to pinpoint which interactions are likely to be unique for the murine anti-γH2AX antibody. In this figure, the frequency of occurrence of residues directly involved in antigen binding across known murine antibodies is indicated together with their key interaction for antigen recognition. **(a)** For each key interaction of light chain residues, a structural representation − as predicted by the docking pose − is indicated together with its natural occurance in *mus musculus*, obtained through abYsis[589] **(b)** Unusual and rare residues of the VL domain that are not directly involved in antigen binding and might contribute to the specificity by other means. **(c)** The CDRs of the heavy chain are characterised by a high presence of rare, unusual or low-abundance residues. Although His-L40 and His-H40 which coordinate the phosphate in the active site are frequent residues, most other residues that provide H-bonds or salt bridges to the γH2AX C-terminal peptide are rare or unusual residues. Particularly the CDR-H2 and the CDR-H3, decisive for antigen recognition, have a high number of rare residues. IMGT numbering.

An interesting observation is that the murine anti-γH2AX mAb has unusual residues in the framework regions, those frequency distributions indicate a structural function. The role of these residues is unclear but provides a hint that the framework regions have a strong influence on this antibody and could potentially explain aggregations identified after humanisation (data not shown). CDR and FR regions are not the only parts of the antibody influencing antigen binding, further remote parts such as the constant domains and *N*-glycosylation can also influence binding.[624,625]

2.4.6.3. Molecular modelling of the full anti-γH2AX mAb

Having studied the antigen recognition in detail, the next step was to investigate how remote sites could influence antigen binding.[624,625]

From a modelling perspective, having structural information upon the F_V and antigen-recognition in hand, the next aim was to develop a model for the full antibody including glycosylation. The PDB was searched for murine IgG1κ structures that could serve as a template, and indeed the 1IGY structure was identified, with the corresponding G1F glycan.[582]. Figure 2.35.A and B. show the sequence alignment to our antibody consensus sequence. The light chain constant region is identical, however there are discrepancies in the constant chain which were mutated to the amino acid sequence of the consensus sequence (exhaustive list see 2.3.7.4).

In order to create the full structural model for the anti-γH2AX mAb, the previously generated F_V model containing the antigen was aligned to the 1IGY structure, taking advantage of the conserved Ig fold (Figure 2.35.D). The glycan was edited from G1F to the identified G0F form (by removal of yellow galactose from Figure 2.35.E) and the ends

of the anti-γH2AX F_V connected to the C_L and C_H of the mutated 1IGY model, again by taking advantage of the very close conformation of the amino acids in both structures (Figure 2.35.E). In a last step, the resulting comprehensive model was energy-minimised using the Amber12 forcefield to generate the final model shown in Figure 2.35.G.

This model shows direct interactions between the hinge polypeptides, the glycosylated loop of one C_{H2} domain and the oligosaccharide. As speculated in the original 1IGY paper[582], these could strongly influence the conformation of the antibody in solution, possibly explaining why attempts at re-glycosylation were unsuccessful as compared to the glycosylated Herceptin structure.[78].

As for the generated pose for the F_V region, molecular dynamics simulations of the entire antibody could be an additional high-end computational validation, to identify accessible conformations in solution. Yet, this would be a challenging simulation to set-up: the accurate handling of the glycan part would require the use of a combined specialised forcefield system for both protein and carbohydrate[584] compatible with the simulation engine – the GROMACS + Amber99sb-ILDN + GLYCAM06 combination[626] has been reportedly used successfully.[627] It goes without saying that the computational costs might be quite extensive.

Taken together, this study suggests for the first time a sequence, glycosylation and structural model for the anti-γH2Ax murine antibody that has been used in research for over a decade, by using a combination of homology modelling, structural analysis, flexible peptide docking and molecular dynamics. Our result are cross-validated across three methodologies for sequencing and F_V modelling. The structure for the antibody-antigen recognition appears plausible in light of the current literature, but could potentially benefit from additional computational and/or experimental validation. These results provide a groundwork for humanisation of this murine antibody, remodelling of its glycan and hopefully will promote its potential as a promising diagnostic marker for double-strand breaks and associated cancers.

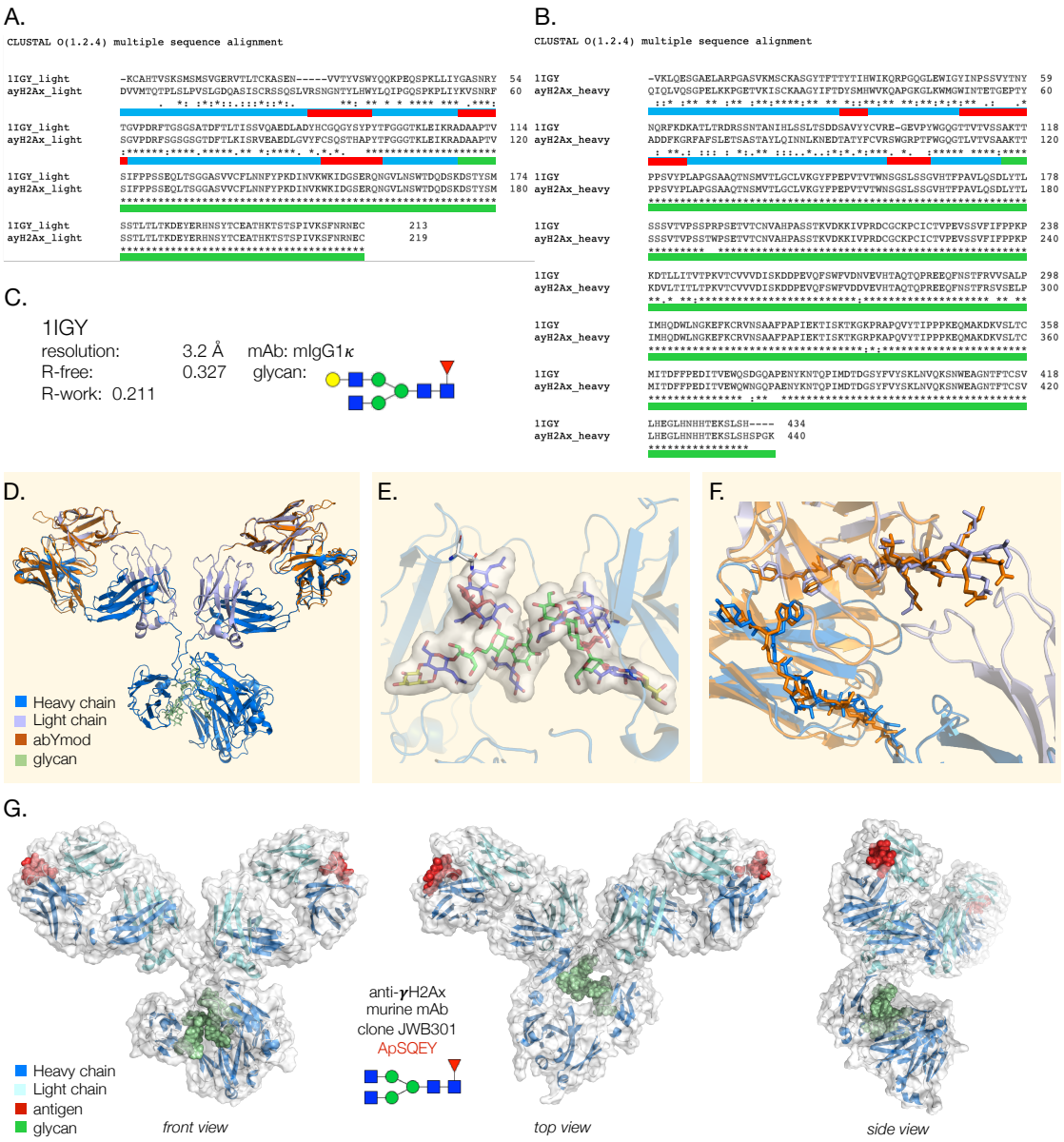
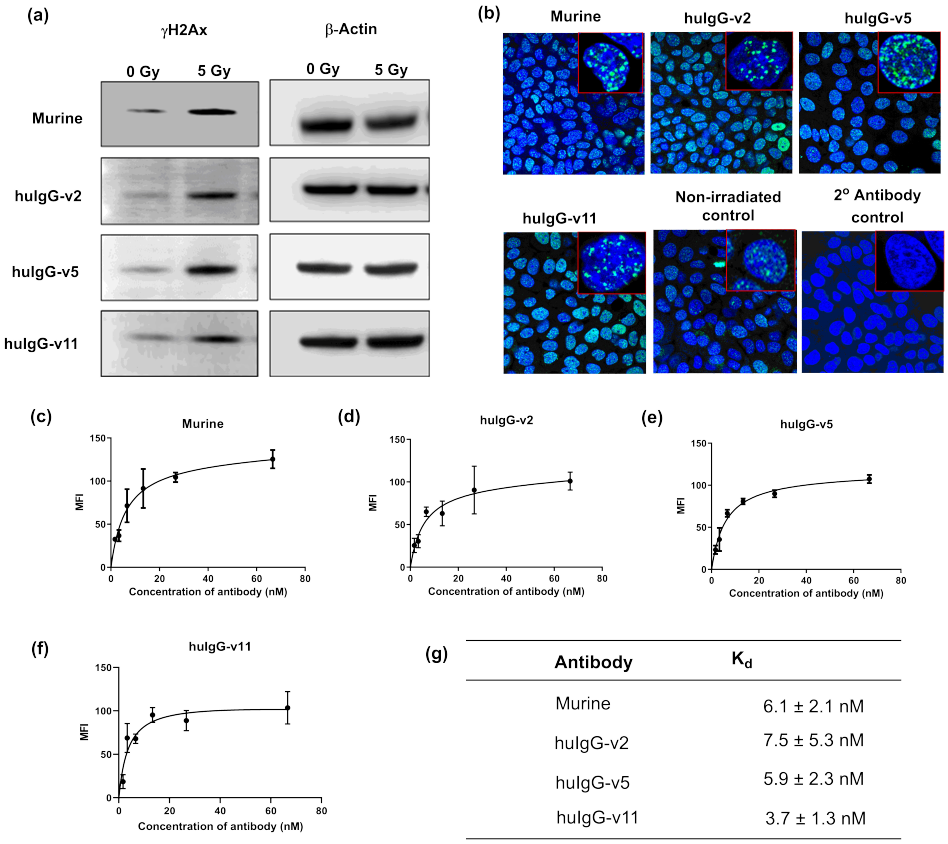


Figure 2.35. Building the full antibody model of the murine anti-γH2AX. A. Alignment of the light chain of the 1IGY template to the anti-γH2AX sequence. **B.** Alignment of the heavy chain of the 1IGY template to the anti-γH2AX sequence shows discrepancies in the constant region which were mutated in the 1IGY structure to the anti-γH2AX sequence. **C.** Background information on the selected antibody template and its glycosylation. **D.** Structural alignment of the homology model for the F_V region to the 1IGY template taking advantage of the conserved IgG fold. **E.** Detailed structure of the 1IGY glycosylation (G1F *N*-glycan) and its interactions with the C_H2 region. **F.** The structural alignment allows for facile merging of the F_V and constant parts, as the position of the amino acid side chains is structurally very similar. This allows the creation of a starting structure for the anti-γH2AX mAb that does not include stark torsional strains. **G.** The final energy-minimised 3D model of the anti-γH2AX mAb, that includes all information generated in this study.

2.5. Outlook

The *de novo* sequencing and computational modelling presented in this chapter enabled sequence and structure-guided humanisation of the anti-γH2AX antibody. In particular, sequencing of the CDRs allowed for humanisation by CDR grafting. Furthermore, computational modelling enabled the identification of key residues in the framework regions. Humanisation, antibody expression and purification work was conducted at Absolute Antibody Ltd and testing of the resulting constructs in γH2AX foci assay in the Department of Oncology. Humanised IgG1 as well as minibodies and scFv-Fc antibody fragments were also generated and shown to be preclinically viable.



Antibody	K_d
Murine	6.1 ± 2.1 nM
huIgG-v2	7.5 ± 5.3 nM
huIgG-v5	5.9 ± 2.3 nM
huIgG-v11	3.7 ± 1.3 nM

Figure 2.36. Results of the humanisation of the anti-γH2AX antibody. A. Western blot of irradiated and non-irradiated SQ20B lysates. The antibody binds to γH2AX, formed upon irradiation. **B.** Irradiation of SQ20B cells leads to DNA damage and results in the formation of foci, where γH2AX is highly present. These foci can be visualised by the anti-γH2AX antibody. **C-F** Plots of mean fluorescence intensity (MFI) v/s concentration of antibody. **G.** Table of the dissociation constants (K_D) of the antibodies determined from the non-linear fit of the plots. Data kindly provided by Dr. Abirami Lakshminarayanan, reproduced with permission.

2.6. Conclusion

The increasing cost of biopharmaceutical innovation in drug discovery and the very limited success rate in the discovery of new treatments is contributing to the current public health crisis. Truly novel treatments for ever increasing complex pathologies such as cancer, dementia, antibiotic resistance or rare diseases pain to emerge from a scientific community entrenched in its traditional disciplines and the pharmaceutical or biotech industry that prefers to keep its results under secrecy [Chas Bountra, 2020[628]].

In no field has this been more true than the development of monoclonal antibodies as biopharmaceutically active molecules to target diseases. Whereas pharmaceutical corporations have been able to build ever more performant pipelines for screening and developing antibodies, the success of these approaches did not correlate with the amount of investments, producing medicines often too similar to targets that are already known or, should a novel cure indeed been found, often too expensive for many patients to afford [Chas Bountra, 2020[628]].

Perhaps it is time for the community to add an investigative approach to its portfolio that is more target-focussed, aiming at a detailed molecular understanding of the mode of action of a single but effective antibody and building upon this knowledge instead of relying on large screening campaigns. This study proposes to do exactly this: the anti-γH2Ax murine antibody has been a workhorse for the double-strand break research community for more than a decade and diagnostic as well as therapeutic strategies have been built on this asset. We managed to perform state-of-the-art antibody *de novo* sequencing using widely available proteomics software and free software tools, combined with in-house scripts and manual analysis, to obtain missing information on protein and glycosylation sequence of this antibody. This was prerequisite to build a 3D molecular model of the antigen recognition and the full antibody using, again, well-established tools from computational chemistry. Our model suggests that the glycosylation might strongly affect the antibody dynamics and conformation, and gives structural details how the phosphoserine is recognised, anchoring the antigen into the antibody paratope.

Several lessons have been learned along the way. Firstly, manual analysis remains the Achilles heel of the *de novo* sequencing endeavour. Whereas modern software tools and methodological papers tend to claim ease of use and indeed facilitate the deed enormously, the sequencing of an antibody remains a daunting task requiring manual analysis and experience in MS/MS and a profound knowledge of antibody structure and function. Whereas most academic mass spectrometry facilities should have the

technical capabilities and knowledge to perform this work, the time constraints are not practicable yet − much effort has still to go into software that really is able to speed up the analysis of real-world case studies − PeaksAB 2.0 assessed here is a promising step in the right direction, facilitating manual analysis but the underlying methodology might need further improvement. Unfortunately, despite an increased sophistication in search engines and *de novo* sequencing algorithms, manual analysis is still required. The lack of case studies in the literature, difficult to publish, stresses this perceived discrepancy further.

Similarly, for glycosylation analysis, better guidance is required to make antibody glycosylation analysis a truly broadly applicable technology. Much of the literature is dominated by laboratories explaining a specific implementation or pipeline, yet labs that need to perform occasional routine analysis will not find the guidance on enzymatic digestion and acquisition protocols needed, resulting − as in our case − in a trial-and-error approach that fits into a rationale at a later stage.

Our sequencing study further highlights the need for methodological cross-validation to assess the correctness and validity of the result, in our case we used RNAseq and denaturing intact mass spectrometry. MS/MS, although less established than RNAseq, undoubtedly offers the advantage to operate on the protein level and such offer the possibility of identification of post-translational modifications, that can affect the antibody behaviour. Additionally, the independence from hybridoma cells allows to resuscitate lost antibodies provided a very small amount of protein is present.

In recent years, computational chemistry tools have reached a level of maturity that allows for scientists from neighbouring disciplines to use these for their projects. This is particularly true for antibody homology modelling of the F_V region, where tools developed by academic labs have been made available publicly via web apps and interfaces. Although research is still ongoing on improving the models, particularly for prediction of the H3 loop, the tools have arguably reached a quality of output that allows benchside scientists to use these for their projects. Further computational studies down the line, such as antigen docking, are less prevalent and limited by the lack of inclusion of standard PTMs such as phosphorylation in some of the most well-established web tools. This clearly is an area where the community could easily improve its resources. The question of validation of these results, both computationally and experimentally, remains an open one. The relevance, though, is high: knowing the epitope as well as the exact sequence and structure of the F_V regions is important for legal fights, as highlighted in the *Noelle v. Lederman*[629] and *Centocor v. Abbott*[630] cases.

At its core, we demonstrate in this study that the current technologies allow us to obtain a molecular understanding of this reagent starting from scratch, providing guidance for further development, be it humanisation, fine-tuning of its properties or mode of action. However, this remains a "tour-de-force" endeavour unique to this day in the field. The technological demands of such a project require the antibody community to improve its knowledge dissemination and extend skills transfer. In our opinion, web-based tools as well as user-friendly software cannot be overstated as key drivers to guide the researcher and drive knowledge in mAb development.

2.7. Author contributions

Simon Nadal designed and performed all MS/MS experiments, data analysis, sequencing analysis, glycosylation analysis, structural analysis and molecular modelling. Dr. Abirami Lakshminarayanan performed intact protein LC-MS and is conducting a separate project on humanisation of the anti-γH2Ax monoclonal antibody in the research groups of Prof. Katherine Vallis with the co-supervision of Prof. Bart Cornelissen in the Department of Oncology. The humanisation project is based on the sequence obtained in this work and by RNAseq. The RNAseq hybridoma sequencing was performed by Absolute Antibody Ltd as part of a commercial package using a hybridoma provided by EMD / Merck Millipore.

2.8. Acknowledgements

I thank Dr. Philip Charles for providing the Perl script for filtering off contaminants from a peptide list. I also thank Elisabete Pires who obtained the trial version of Peaks AB from Bioinformatics Solutions Inc, with the generous support from Ms Katherine Tran. The efforts of Dr. Sanjay Sanghani to obtain us a trial version of SuperNovo™, although fruitless, are also gratefully acknowledged. I also thank Svenja Hester for experimental support in mass spectrometry and Aziz Khan for helpful discussion on antibody PTMs. Finally, the European e-Infrastructure projects WeNMR, WestLife and EOSC-hub are acknowledged for enabling the use of the HADDOCK web portal.

2.9. Conflict of interest

The authors declare no conflict of interest on the murine anti-γH2Ax monoclonal antibody and its sequence or 3D structure. A patent filing to the University of Oxford is ongoing for humanised versions of the murine antibody on which Simon Nadal is listed as a contributor, Ben Davis and Shabaz Mohammed as inventors. This patent might afford authors royalties were it to be licensed.

2.10. Confidentiality declaration

The public disclosure of the sequence of the murine anti-γH2AX monoclonal antibody antibody is subject to an MTA between the University of Oxford with EMD Milipore, enclosing both heavy and light chain sequences. Any disclosure of sequence information requires 30 day prior notice and EMD reserves itself the right to withhold the publication of sequence information. This chapter has been placed under Dispensation of Consultation until October 2021, which can be renewed upon the demand of the authors.

2.11. Appendix

2.11.1. Copy of code and scripts used in this chapter

MS Excel regular expressions

SPIDER peptide list clean-up was performed using the following regular expressions in MS Excel: `(sub ?)`, `(del ?)`, `(ins)`.

VBA code

The following module was defined to set an ALC of 99 for database and SPIDER peptides.

```
Function ExtractAlpha(rC As Range) As String

With CreateObject("VBSCRIPT.REGEXP")
  .Pattern = "[^a-z]"
  .Global = True
  .IgnoreCase = True
  ExtractAlpha = .Replace(rC.Value, "")
End With
End Function
```

In the Excel sheet, the following function was used:

```
=REPT("99 ",LEN(ExtractAlpha(B2)))
```

2.11.2. Advice for manual *de novo* sequencing by Bioinformatic Solutions

Whole Protein *de novo* Sequencing from MS/MS

Lian Yang[1], Baozhen Shan[1], Mingjie Xie[1], Bin Ma[2]
[1] Bioinformatics Solutions Inc, Waterloo, ON
[2] University of Waterloo, Waterloo, ON

Overview
We present a semi-automatic workflow for characterizing antibody primary structure utilizing protein contigs assembled from *de novo* peptide sequences.

Introduction
De novo peptide sequencing is a practical technique for characterizing uncharted proteins. Among all proteins, antibodies are the most common subject to be sequenced. In general, it is required to achieve complete sequence coverage with multiple enzymatic digestion and produce evidence for every residue, specifically for the hypervariable CDR regions. While automated *de novo* peptide sequencing is routinely conducted, it is still puzzling and tedious to infer the original protein sequence from those short peptides.

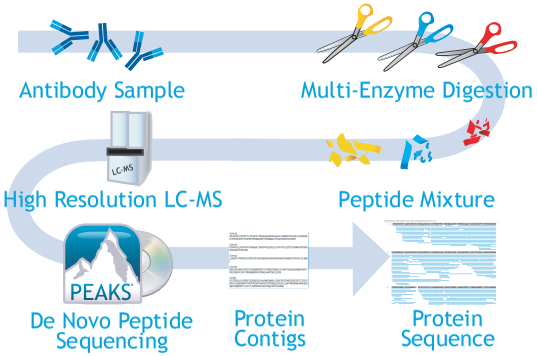


Figure 1. Workflow of Monoclonal Antibody Characterization

In the world of genomics, DNA sequencing is achieved by assembling short reads using software. In principle, one can do the same in proteomics by assembling the short *de novo* peptides into "protein contigs", and it will significantly reduce the workload in manual analysis.

Method
1. Assembling of protein contigs
A program is developed using a greedy algorithm for assembling protein contigs. It first detects the overlapping *de novo* peptides and iteratively extends the contigs using the consensus residues at both direction.
In the extending process, contigs are merged if their sequences collide with each other. Protein contigs are extended until no more residues can be confidently added.

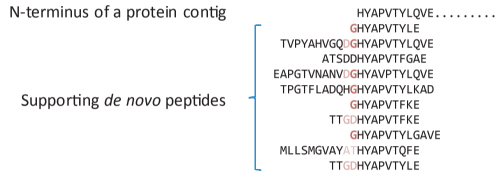


Figure 2. Extending a Contig at N-Terminus

2. Workflow for antibody sequencing
1. Assemble an antibody template
- Blast protein contigs in NCBI nr database
- Select a protein hit corresponding to the constant region
- Select the closest protein hit corresponding to the variable region.

2. Draft a sequence using the contigs and template
- Map the contigs to the antibody template
- Solve disagreements between contigs and the template
  - Correct typical *de novo* sequencing errors (I/L, AG/Q, PK/KP, variable PTMs)
  - In principle, trust contigs in variable region and the template in constant region
- Solve the residues not covered by contigs, if any, using *de novo* peptides

3. Refine the sequence using PEAKS' SPIDER homology search
- Perform SPIDER homology search on the draft sequence
- Examine insertion/deletion/mutation reported by SPIDER
- Examine residues with low peptide coverage
- Examine residues at the protein n-terminus
- Compare the sequence mass with protein intact mass, if available
- Make corrections and repeat step 3 until the sequences converge

Result
1. Experiment
The antibody, I5154 (human IgG1/kappa), was purchased from Sigma-Aldrich to evaluate the proposed workflow.
The sample was reduced with DTT, alkylated with iodoacetamide. Glycans were removed and heavy/light chains were separated. Each chain was digested with six enzymes AspN, chymotrypsin, GluC, LysC, pepsin and trypsin. MS/MS spectra was acquired using LTQ-Orbitrap at high resolution with HCD fragmentation.

The mass spec data for the heavy and light chains were separately analyzed. *De novo* peptide sequencing was performed using PEAKS. Protein contigs were assembled from *de novo* peptides. The majority of assembled contigs had a length of 60~120 residues. Antibody sequences were manually constructed using the described workflow.
For both the heavy and light chains, sequencing results were fully covered by MS/MS spectra. 98% of amino acid residues were supported by fragment ions in at least 5 MS/MS spectra.

Figure 3. Coverage of a Light Chain Sequencing Result

2. Blind trials
The proposed workflow was also tested in a series of blind trials. Our collaborators provide antibody mass spec data to us for in-house analysis, while not disclosing the true sequences.
The sequencing results were correct at a minimal of 96% of residues, meanwhile all the wrong residues were previously identified with low confidence during the sequencing analysis.

Conclusion
Protein sequencing can be efficiently conducted by utilizing protein contigs assembled from *de novo* peptides.
The described semi-automated workflow provides a reliable solution for antibody sequencing with MS/MS of multiple enzymatic digestion.
The proposed workflow has been routinely used for data analysis in the **CHAMPS** - Antibody Sequencing Service, provided by Bioinformatics Solution Inc.
(http://www.bioinfor.com/peaks/products/champs.html)

2.11.3. RNA sequencing report from Absolute Antibody

absolute antibody

ANTIBODY SEQUENCING REPORT

Client: University of Oxford
Date: 22-Feb-18
Prepared by: Stephen Anderson

MATERIALS AND METHODS

HYBRIDOMA SEQUENCING
Sequencing is performed by whole transcriptome shotgun sequencing (RNA-Seq). This is a proprietary approach to antibody sequencing developed exclusively at Absolute Antibody.

If required hybridomas are cultured in IMDM medium containing 10% FBS and incubated at 37°C in a 5% CO_2 environment. Total RNA is extracted from cells and a barcoded cDNA library generated through RT-PCR using a random hexamer. Next Generation Sequencing is performed on an Illumina HiSeq sequencer. Contigs are assembled using a propriatary approach and data is mined for antibody sequences idenfitying all viable antibody sequences (i.e. those not containing stop codons). Variable heavy and variable light domains are identified separately and relative abundance of each identified gene is reported in transcripts per million (TPM). The species and isotype of the identified antibody genes are confirmed. Sequences are compared with known aberrant (i.e. non-functional) antibody genes that are present in many hybridomas and these genes are removed from analysis when necessary.

DATA INTERPRETATION
All identified antibody genes are shown on the "Sequence data" worksheets. One worksheet reports protein sequences and the other worksheet reports DNA sequences. Genes have been separated into heavy chains and light chains. For each chain the variable domain is reported along with the signal peptide and constant domain where it has been possible to identify these regions. If the entire constant domain is identified the integrity will be reported as complete. Where it has not been possible to identify the full constant domain sequence the integrity will be reported as partial. This does not impact the validity of the variable domain sequence but it simply a reflection of the fact the data has become truncated in the constant domain region.

The heavy chain and light chain type will be identified based on the sequence of the constant domain (e.g. mouse IgG1, rat kappa etc). If only a short fragment of constant domain has been identified it may not be possible to determine the species and isotype of antibody.

For each heavy and light chain two measurements of abundance are reported. TPM (transcripts per million) is an abudance measurement from next generation sequencing which scores the relative abundance of each gene in the whole transciptome pool. The % column is a more relevant number which converts the TPM into a percentage antibody gene abundance. This enables simple identification of high and low abundance antibody genes. The data is automatically filtered to only report antibody genes with an abundance of greater than 1%.

CDR IDENTIFICATION
The complementarity determining regions (CDRs) have been automatically identified on the "CDR identification" worksheet using Excel formulas working to the Kabat definition for CDRs. CDR identification is only performed for the primary VH and VL sequences. The CDR identification scheme can be changed from Kabat to IMGT or Chothia if desired.

If you wish to analyse the CDRs for other sequences please use the "Manual CDR tool". Copy and paste your VH sequence into column C and oyur VL sequence into column G. Select your CDR scheme and the CDRs will be automatically identified.

Although we have done our best to design formulas to auto-identify CDRs and have tested this thoroughly, we take no responsibility for the accuracy of our Excel based system. We have based our system on highly conserved antibody motifs but if your sequence contains a particular unusual sequence or extra short or long CDRs then the system will fail. This may result in Excel either mis-identifying the CDR regions or reporting an error. In these situations CDRs can manually be identified using the guidance provided on the following website: http://www.bioinf.org.uk/abs/#cdrid

FREQUENTLY ASKED QUESTIONS

Why does my antibody contain more than one heavy chain and one light chain?
For most hybridomas only one VH and one VL sequence are identified. These are reported as the primary sequences. It is well known that certain myelomas used in the hybridoma fusion process contain aberrant antibody genes, in particular variable light genes. In the case that these known sequences are found they are automatically removed and not reported here. Occasionally hybridomas contain additional heavy and or light genes. These will be reported as secondary and tertiary sequences ranked by decreasing abundance. In the case of hybridomas containing multiple genes it would be expected that the most abundance genes are most likely to be the ones representing the functional antibody. However, this is not always be the case and it is impossible to know for certain which sequence represents the correct one without going through the process of making all possible antibodies and testing activity.

Why do you use Next Generation Sequencing for hybridima sequencing?
When we first started sequencing antibodies from hybridomas we used the traditional methods of V-region PCR and 5' RACE. By the time we had sequenced approximately 300 antibodies by these methods we had become increasingly frustrated by their limitations. These include mis-identification of aberrant light chains, primer bias, non-amplification of variable regions due to primer mis-matches or low abundance, low throughput and the need to redesign primers for every new isotypoe and species to be sequenced. We felt that by switching to an NGS based approached and literally sequence every transcript in the hybridoma (antibody and everything else) would be the most thorough approach. Due to the data intensive nature of the process this does mean our turn around times are not fast as our competitors who still use traditional sequencing approaches but we felt that being able to provide customers with more detailed and accurate information was more important than turnaround time.

Why is my antibody missing an entire heavy or light chain?
On rare occasions it is not possible to identify a heavy chain and/or light chain. We are continually improving our process to try and minimise these occurences. The absence of a chain could be real and indicate that the hybridoma is no longer secreting the antibody. This can occur for older hybridomas that have been in long term storage and have not been retested for expression/functionality before sending to us for sequencing. Alternatively the lack of a chain may represent an mRNA quality problem or low abundance of this particular transcript. One of our scientists will be in touch to discuss the next stages of your project and how we will go about trying to find the "missing" chain.

Why have you been unable to idenfity the full antibody sequence?
Although we make our best efforts to provide a full intact antibody sequence (i.e. signal peptide, variable domain and constant domains) at no extra cost this is not always possible. The length of the constant domains in particular can make it challenging to get a full read all the way to the end. However, we will always aim to provide you with the variable domain sequence as a minimum.

Why does the percentage value not add up to 100?
As described previously hybridomas can contain multiple heavy and light chains. NGS data provides a wealth of information that you wouldn't necessarily see via other techniques and often highlights the fact that hybridomas have multiple heavy and light chain genes expressed at very low abundance. These are what would be classed as background noise and so we filter out any heavy and light chains with a percentage value below 1%. This can then result in the remaining percentage scores of the heavy and light chains not adding up to 100.

Why have the CDRs of my antibody not been identified correctly?
Although we have done our best to design formulas to auto-identify CDRs and have tested this thoroughly, we take no responsibility for the accuracy of our Excel based system. We have based our system on highly conserved antibody motifs but if your sequence contains a particular unusual sequence or extra short or long CDRs then the system will fail. This may result in Excel either mis-identifying the CDR regions or reporting an error. In these situations CDRs can manually be identified using the guidance provided on the following website: http://www.bioinf.org.uk/abs/#cdrid

CONTACT US
For further information and support please contact us.

Email: support@absoluteantibody.com
Phone: +44 (0)1865 920810

Clone name	AbAb ID	No of HCs	No of LCs
OxJBW301	cAb1740	1	1

Primary heavy chain protein sequence						
HC type	**Integrity**	**Sig. pep.**	**VH**	**Constant**	**TPM**	**%**
Mouse IgG1	Complete	ATGGCTTGG GTGTGGACCT TGCTATTCCT GATGGCAGC TGCCCAAAGT ATCCAAGCG	CAGATCCAGTTGGTGCAG TCTGGACCTGAGCTGAAG AAGCCTGGAGAGACAGT CAAGATCTCCTGCAAGGC TGCTGGTTATATCTTCAC AGACTATTCAATGCACTG GGTGAAGCAGGCTCCAG GAAAGGGTTTAAAGTGG ATGGGCTGGATAAACACT GAGACTGGTGAACCAAC ATATGCAGATGACTTCAA GGGACGGTTTGCCTTCTC TTTGGAAACCTCTGCCAG CACTGCCTATTTGCAGAT CAACAACCTCAAAAATGA GGACACGGCTACATATTT CTGTGTTAGATCGTGGGG TAGACCGACTTTCTGGGG CCAAGGGACTCTGGTCAC TGTCTCTGCA	GCCAAAACGACACCCCCATCTGTCTATCCACTGGCCCCTGGATCT GCTGCCCAAACTAACTCCATGGTGACCCTGGGATGCCTGGTCAAG GGCTATTTCCCTGAGCCAGTGACAGTGACCTGGAACTCTGGATCC CTGTCCAGCGGTGTGCACACCTTCCCAGCTGTCCTGCAGTCTGAC CTCTACACTCTGAGCAGCTCAGTGACTGTCCCCTCCAGCACCTGG CCCAGCGAGACCGTCACCTGCAACGTTGCCCACCCGGCCAGCAG CACCAAGGTGGACAAGAAAATTGTGCCCAGGGATTGTGGTTGTA AGCCTTGCATATGTACAGTCCCAGAAGTATCATCTGTCTTCATCTT CCCCCCAAAGCCCAAGGATGTGCTCACCATTACTCTGACTCCTAA GGTCACGTGTGTTGTGGTAGACATCAGCAAGGATGATCCCGAGG TCCAGTTCAGCTGGTTTGTAGATGATGTGGAGGTGCACACAGCTC AGACGCAACCCCGGGAGGAGCAGTTCAACAGCACTTTCCGCTCA GTCAGTGAACTTCCCATCATGCACCAGGACTGGCTCAATGGCAAG GAGTTCAAATGCAGGGTCAACAGTGCAGCTTTCCCTGCCCCCATC GAGAAAACCATCTCCAAAACCAAAGGCAGACCGAAGGCTCCACA GGTGTACACCATTCCACCTCCCAAGGAGCAGATGGCCAAGGATA AAGTCAGTCTGACCTGCATGATAACAGACTTCTTCCCTGAAGACA TTACTGTGGAGTGGCAGTGGAATGGGCAGCCAGCGGAGAACTA CAAGAACACTCAGCCCATCATGGACACAGATGGCTCTTACTTCGT CTACAGCAAGCTCAATGTGCAGAAGAGCAACTGGGAGGCAGGA AATACTTTCACCTGCTCTGTGTTACATGAGGGCCTGCACAACCAC CATACTGAGAAGAGCCTCTCCCACTCTCCTGGTAAA	247425.1	36.05

Primary light chain protein sequence						
LC type	**Integrity**	**Sig. pep.**	**VL**	**Constant**	**TPM**	**%**
Mouse kappa	Complete	ATGAAGTTGC CTGTTAGGCT GTTGGTGCTG ATGTTCTGGA TTCCTGCCTC CAGCAGT	GATGTTGTGATGACCCAA ACTCCACTCTCCCTGCCTG TCAGTCTTGGAGATCAAG CCTCCATCTCTTGCAGATC TAGTCAGAGCCTTGTACG CAGTAATGGAAACACCTA TTTACATTGGTACCTGCA GATTCCAGGCCAGTCTCC AAAGCCCCTGATTTACAA AGTCTCCAACCGATTTTCT GGGGTCCCAGACAGGTT CAGTGGCAGTGGATCAG GGACAGATTTCACACTCA AGATCAGCAGAGTGGAG GCTGAGGATCTGGGAGT TTATTTCTGCTCTCAAAGT ACACATGCTCCGTATACG TTCGGAGGGGGGACCAA GCTGGAAATAAAA	CGGGCTGATGCTGCACCAACTGTATCCATCTTCCCACCATCCAGT GAGCAGTTAACATCTGGAGGTGCCTCAGTCGTGTGCTTCTTGAAC AACTTCTACCCCAAAGACATCAATGTCAAGTGGAAGATTGATGGC AGTGAACGACAAAATGGCGTCCTGAACAGTTGGACTGATCAGGA CAGCAAAGACAGCACCTACAGCATGAGCAGCACCCTCACGTTGA CCAAGGACGAGTATGAACGACATAACAGCTATACCTGTGAGGCC ACTCACAAGACATCAACTTCACCCATTGTCAAGAGCTTCAACAGG AATGAGTGT	438789.5	63.94

Clone name	AbAb ID	No of HCs	No of LCs
OxJBW301	cAb1740	1	1

Primary heavy chain protein sequence						
HC type	Integrity	Sig. pep.	VH	Constant	TPM	%
Mouse IgG1	Complete	MAWVWTLLFLMAAAQSIQA	QIQLVQSGPELKKPGETVKISCKAAGYIFTDYSMHWVKQAPGKGLKWMGWINTETGEPTYADDFKGRFAFSLETSASTAYLQINNLKNEDTATYFCVRSWGRPTFWGQGTLVTVSA	AKTTPPSVYPLAPGSAAQTNSMVTLGCLVKGYFPEPVTVTWNSGSLSSGVHTFPAVLQSDLYTLSSSVTVPSSTWPSETVTCNVAHPASSTKVDKKIVPRDCGCKPCICTVPEVSSVFIFPPKPKDVLTITLTPKVTCVVVDISKDDPEVQFSWFVDDVEVHTAQTQPREEQFNSTFRSVSELPIMHQDWLNGKEFKCRVNSAAFPAPIEKTISKTKGRPKAPQVYTIPPPKEQMAKDKVSLTCMITDFFPEDITVEWQWNGQPAENYKNTQPIMDTDGSYFVYSKLNVQKSNWEAGNTFTCSVLHEGLHNHHTEKSLSHSPGK	247425.1	36.05

Primary light chain protein sequence						
LC type	Integrity	Sig. pep.	VL	Constant	TPM	%
Mouse kappa	Complete	MKLPVRLLVLMFWIPASSS	DVVMTQTPLSLPVSLGDQASISCRSSQSLVRSNGNTYLHWYLQIPGQSPKPLIYKVSNRFSGVPDRFSGSGSGTDFTLKISRVEAEDLGVYFCSQSTHAPYTFGGGTKLEIK	RADAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKDINVKWKIDGSERQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYERHNSYTCEATHKTSTSPIVKSFNRNEC	438789.5	63.94

2.11.4. Antibody Homology Modelling – Supplementary information

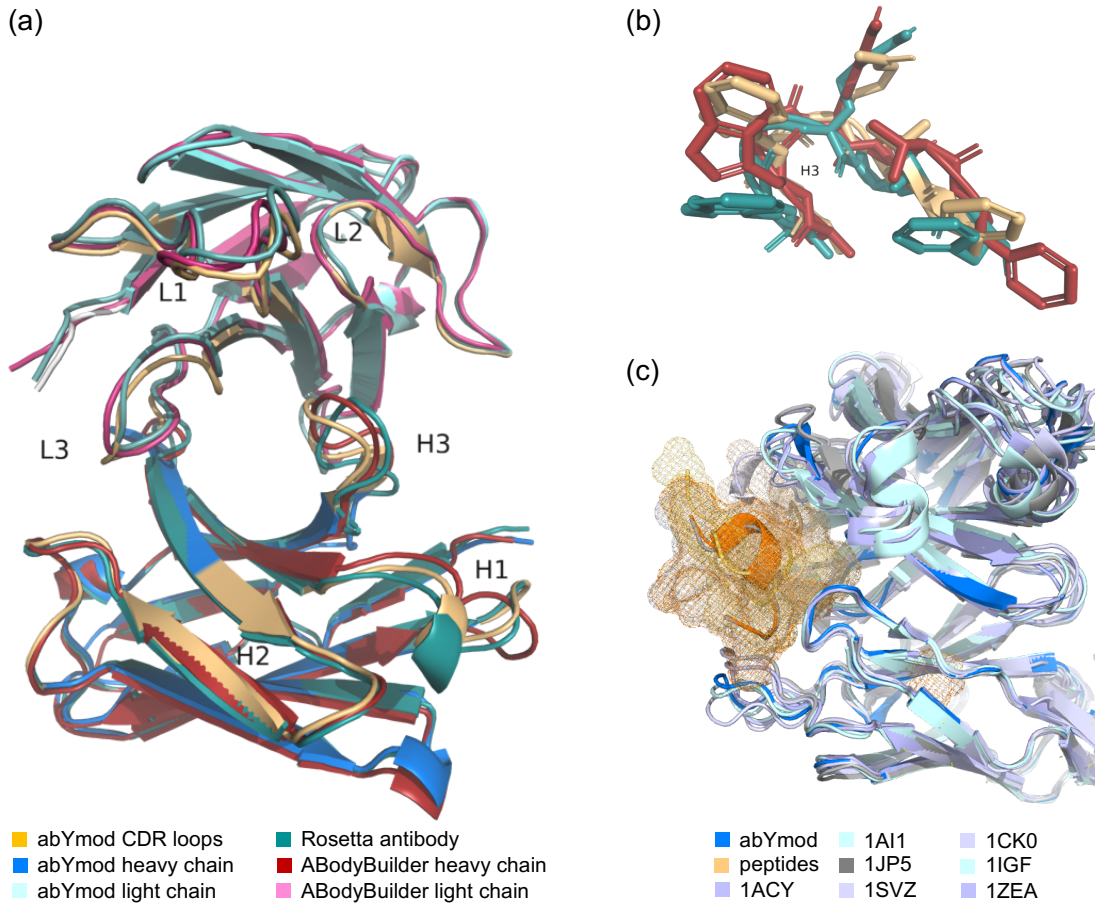


Figure 2.37. Homology modelling of the murine F_V. Antibody modelling takes advantage of the conserved domain structure. Specialised tools for antibody modelling have been developed, focussing on the prediction of CDR loops.[570,631] **A.** Structural analysis of the 3 generated homology models using the abYmod,[632] ABodyBuilder[569] and Rosetta antibody[568] tools. Overall, the result of all three modelling platforms yield structurally very similar results. One major noticeable difference is in the H1 loop, where ABodyBuilder yields a different pose. **B.** The second major difference is in the H3 loop, notoriously difficult to predict.[567] The abYmod has in all cases an overlap with either one of the two other predicted models and was used in subsequent modelling tasks. **C.** The binding site of peptides to mIgG1-κ F_V crystal structures deposited in the PDB is conserved and suggests this will be the case for the murine anti-γH2AX mAb. This structural information was subsequently used to identify residues likely to form the paratope and to generate restraints for the flexible docking with HADDOCK.

2.11.5. Full Haddock docking results – short peptide

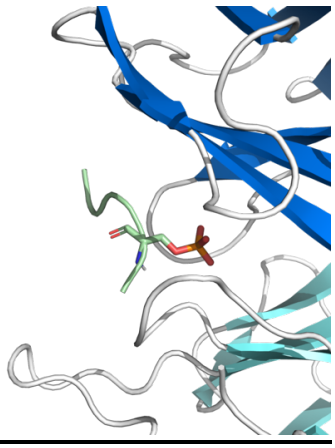
HADDOCK Docking yH2Ax shorter peptide: ApSQEY

Summary

HADDOCK clustered **171** structures in **10** cluster(s), which represents **85.5 %** of the water-refined models HADDOCK generated. Note that currently the maximum number of models considered for clustering is 200.
The statistics of the top 10 clusters are shown below. The top cluster is the most reliable according to HADDOCK. Its Z-score indicates how many standard deviations from the average this cluster is located in terms of score (the more negative the better). The view of the best-ranked pose for each cluster is depicted. The modelled peptide is shown in light green, phosphoserine is shown as sticks.

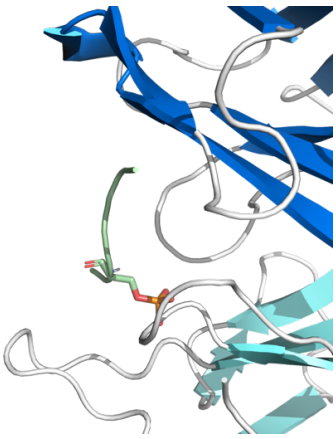
Cluster 2

HADDOCK score	-92.8 +/- 5.6
Cluster size	35
RMSD from the overall lowest-energy structure	1.5 +/- 0.2
Van der Waals energy	-46.4 +/- 4.5
Electrostatic energy	-312.6 +/- 36.3
Desolvation energy	-6.5 +/- 4.3
Restraints violation energy	226.2 +/- 23.08
Buried Surface Area	1026.0 +/- 22.5
Z-Score	-1.7



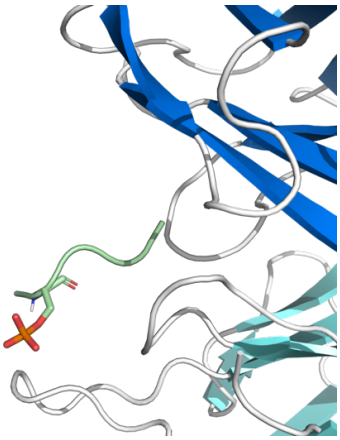
Cluster 3

HADDOCK score	-87.1 +/- 8.9
Cluster size	22
RMSD from the overall lowest-energy structure	0.9 +/- 0.5
Van der Waals energy	-39.1 +/- 2.2
Electrostatic energy	-310.0 +/- 47.4
Desolvation energy	-13.8 +/- 4.3
Restraints violation energy	279.2 +/- 49.72
Buried Surface Area	1010.2 +/- 31.1
Z-Score	-1.3



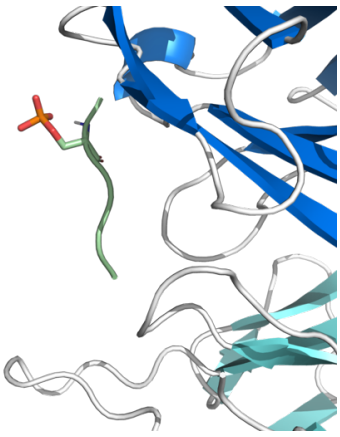
Cluster 1

HADDOCK score	-81.1 +/- 9.4
Cluster size	44
RMSD from the overall lowest-energy structure	1.3 +/- 0.1
Van der Waals energy	-28.1 +/- 4.6
Electrostatic energy	-319.1 +/- 38.9
Desolvation energy	-18.8 +/- 6.4
Restraints violation energy	295.8 +/- 114.98
Buried Surface Area	888.8 +/- 47.3
Z-Score	-1.0



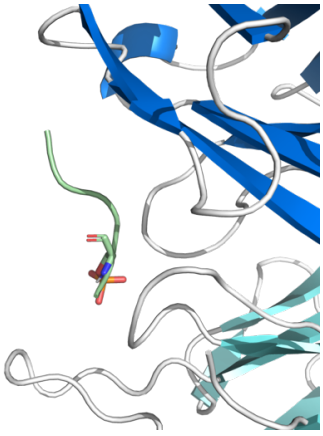
Cluster 6

HADDOCK score	-66.4 +/- 3.9
Cluster size	16
RMSD from the overall lowest-energy structure	1.7 +/- 0.1
Van der Waals energy	-39.4 +/- 3.9
Electrostatic energy	-181.1 +/- 18.1
Desolvation energy	-12.7 +/- 4.4
Restraints violation energy	219.0 +/- 20.04
Buried Surface Area	1022.9 +/- 29.9
Z-Score	-0.1



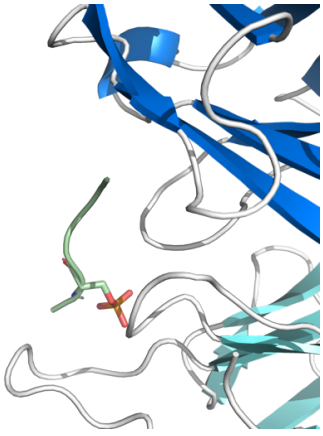
Cluster 5

HADDOCK score	-64.6 +/- 2.5
Cluster size	16
RMSD from the overall lowest-energy structure	1.4 +/- 0.2
Van der Waals energy	-34.0 +/- 7.2
Electrostatic energy	-271.4 +/- 64.7
Desolvation energy	-8.0 +/- 7.9
Restraints violation energy	317.3 +/- 91.76
Buried Surface Area	1014.1 +/- 41.9
Z-Score	-0.0



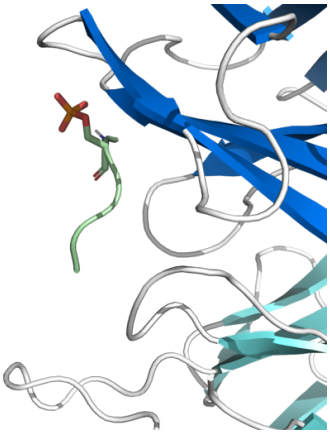
Cluster 4

HADDOCK score	-59.2 +/- 11.6
Cluster size	17
RMSD from the overall lowest-energy structure	0.9 +/- 0.2
Van der Waals energy	-30.7 +/- 9.0
Electrostatic energy	-283.1 +/- 54.4
Desolvation energy	-4.1 +/- 4.8
Restraints violation energy	322.0 +/- 74.21
Buried Surface Area	967.2 +/- 55.3
Z-Score	0.3



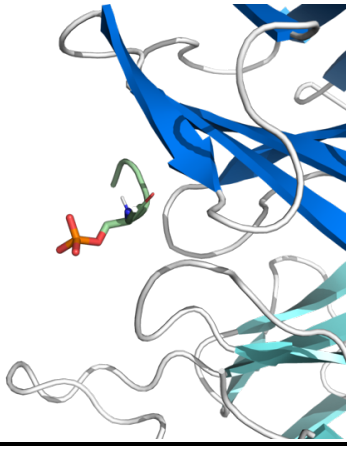
Cluster 10

HADDOCK score	-57.5 +/- 11.2
Cluster size	5
RMSD from the overall lowest-energy structure	1.7 +/- 0.1
Van der Waals energy	-35.8 +/- 7.3
Electrostatic energy	-159.0 +/- 41.9
Desolvation energy	-12.5 +/- 5.1
Restraints violation energy	226.2 +/- 37.12
Buried Surface Area	939.3 +/- 58.9
Z-Score	0.4



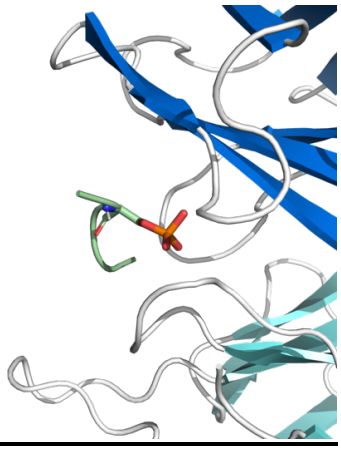
Cluster 7

HADDOCK score	-50.2 +/- 7.5
Cluster size	6
RMSD from the overall lowest-energy structure	2.0 +/- 0.2
Van der Waals energy	-38.7 +/- 7.8
Electrostatic energy	-134.4 +/- 22.3
Desolvation energy	-7.9 +/- 8.1
Restraints violation energy	233.3 +/- 65.92
Buried Surface Area	954.7 +/- 64.0
Z-Score	0.8



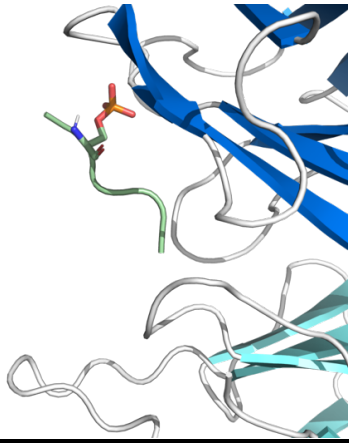
Cluster 8

HADDOCK score	-45.3 +/- 5.0
Cluster size	5
RMSD from the overall lowest-energy structure	1.9 +/- 0.1
Van der Waals energy	-30.5 +/- 4.5
Electrostatic energy	-167.9 +/- 14.3
Desolvation energy	-9.4 +/- 4.4
Restraints violation energy	281.4 +/- 36.01
Buried Surface Area	911.4 +/- 69.8
Z-Score	1.1



Cluster 9

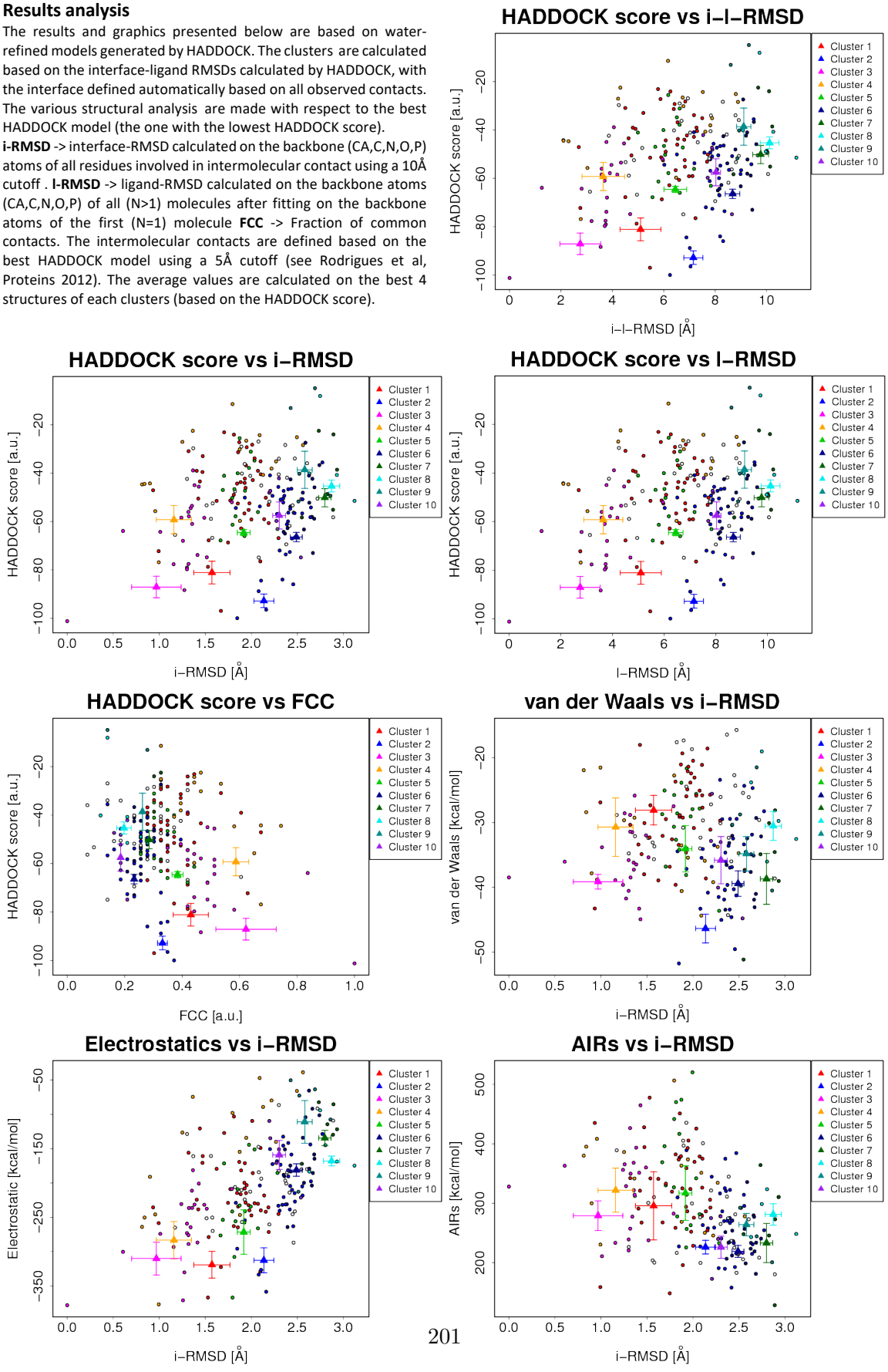
HADDOCK score	-38.6 +/- 15.3
Cluster size	5
RMSD from the overall lowest-energy structure	1.8 +/- 0.2
Van der Waals energy	-34.8 +/- 5.3
Electrostatic energy	-110.8 +/- 62.2
Desolvation energy	-8.0 +/- 5.5
Restraints violation energy	264.1 +/- 37.60
Buried Surface Area	954.8 +/- 32.6
Z-Score	1.5



Results analysis

The results and graphics presented below are based on water-refined models generated by HADDOCK. The clusters are calculated based on the interface-ligand RMSDs calculated by HADDOCK, with the interface defined automatically based on all observed contacts. The various structural analysis are made with respect to the best HADDOCK model (the one with the lowest HADDOCK score).

i-RMSD -> interface-RMSD calculated on the backbone (CA,C,N,O,P) atoms of all residues involved in intermolecular contact using a 10Å cutoff . **l-RMSD** -> ligand-RMSD calculated on the backbone atoms (CA,C,N,O,P) of all (N>1) molecules after fitting on the backbone atoms of the first (N=1) molecule **FCC** -> Fraction of common contacts. The intermolecular contacts are defined based on the best HADDOCK model using a 5Å cutoff (see Rodrigues et al, Proteins 2012). The average values are calculated on the best 4 structures of each clusters (based on the HADDOCK score).



2.11.6. Full Haddock docking results – longer peptide

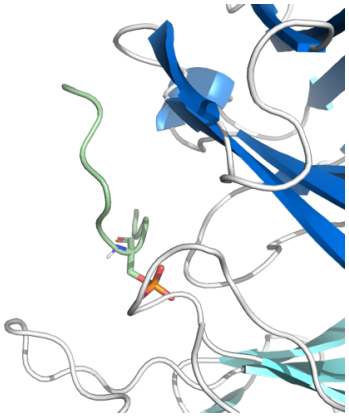
HADDOCK Docking yH2Ax Longer peptide : KKATQApSQEY

Summary

HADDOCK clustered 168 structures in 7 cluster(s), which represents 84.0 % of the water-refined models HADDOCK generated. Note that currently the maximum number of models considered for clustering is 200.
The statistics of the top 10 clusters are shown below. The top cluster is the most reliable according to HADDOCK. Its Z-score indicates how many standard deviations from the average this cluster is located in terms of score (the more negative the better). The view of the best-ranked pose for each cluster is depicted. The modelled peptide is shown in light green, phosphoserine is shown as sticks.

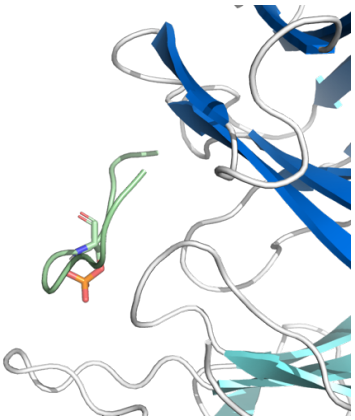
Cluster 2

HADDOCK score	-125.7 +/- 9.8
Cluster size	39
RMSD from the overall lowest-energy structure	0.6 +/- 0.4
Van der Waals energy	-48.9 +/- 4.2
Electrostatic energy	-468.0 +/- 85.9
Desolvation energy	-8.7 +/- 4.5
Restraints violation energy	254.9 +/- 67.01
Buried Surface Area	1465.3 +/- 63.4
Z-Score	-1.9



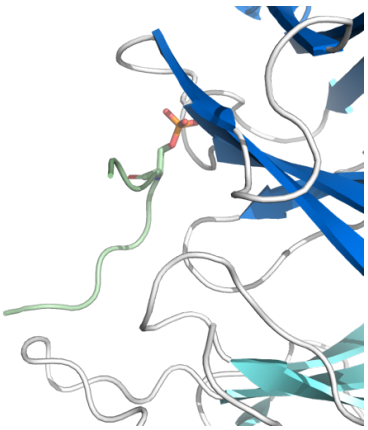
Cluster 1

HADDOCK score	-107.5 +/- 1.1
Cluster size	41
RMSD from the overall lowest-energy structure	2.2 +/- 0.1
Van der Waals energy	-55.7 +/- 7.9
Electrostatic energy	-357.6 +/- 26.5
Desolvation energy	-1.1 +/- 7.6
Restraints violation energy	208.6 +/- 13.04
Buried Surface Area	1526.2 +/- 39.5
Z-Score	-0.9



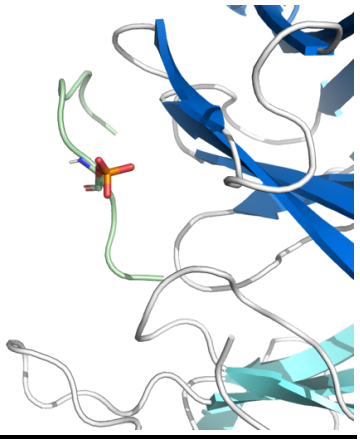
Cluster 6

HADDOCK score	-91.5 +/- 3.0
Cluster size	12
RMSD from the overall lowest-energy structure	2.9 +/- 0.1
Van der Waals energy	-53.7 +/- 7.2
Electrostatic energy	-196.1 +/- 39.2
Desolvation energy	-17.8 +/- 2.8
Restraints violation energy	191.8 +/- 42.31
Buried Surface Area	1373.2 +/- 29.5
Z-Score	0.0



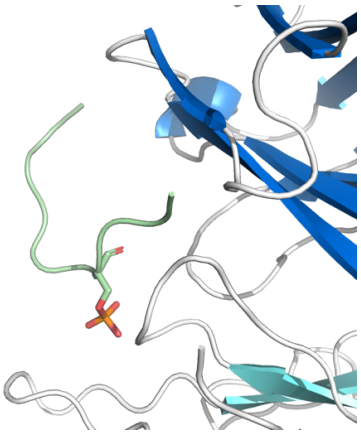
Cluster 3

HADDOCK score	-91.2 +/- 1.5
Cluster size	35
RMSD from the overall lowest-energy structure	2.0 +/- 0.2
Van der Waals energy	-42.6 +/- 0.9
Electrostatic energy	-327.5 +/- 31.9
Desolvation energy	-6.8 +/- 6.8
Restraints violation energy	237.5 +/- 37.25
Buried Surface Area	1365.4 +/- 42.8
Z-Score	0.0



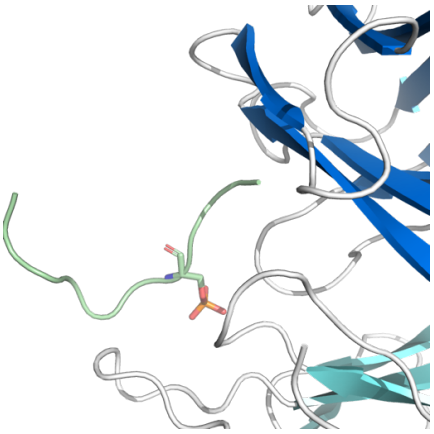
Cluster 7

HADDOCK score	-81.3 +/- 1.0
Cluster size	5
RMSD from the overall lowest-energy structure	2.0 +/- 0.2
Van der Waals energy	-32.6 +/- 2.6
Electrostatic energy	-347.8 +/- 33.4
Desolvation energy	3.1 +/- 7.0
Restraints violation energy	178.2 +/- 64.08
Buried Surface Area	1138.2 +/- 53.5
Z-Score	0.6



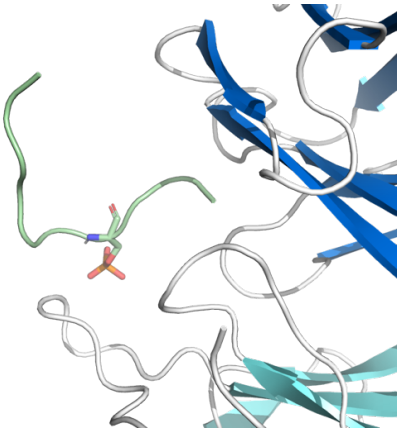
Cluster 4

HADDOCK score	-78.8 +/- 3.6
Cluster size	22
RMSD from the overall lowest-energy structure	2.5 +/- 0.2
Van der Waals energy	-40.2 +/- 9.0
Electrostatic energy	-290.7 +/- 46.4
Desolvation energy	-8.2 +/- 2.9
Restraints violation energy	276.7 +/- 47.68
Buried Surface Area	1141.6 +/- 91.3
Z-Score	0.7



Cluster 5

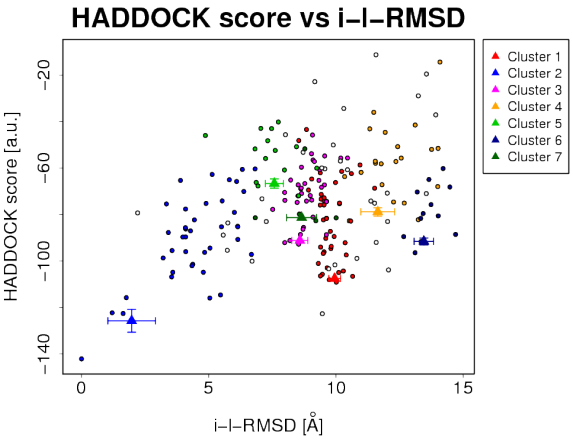
HADDOCK score	-66.6 +/- 4.0
Cluster size	14
RMSD from the overall lowest-energy structure	1.9 +/- 0.2
Van der Waals energy	-33.3 +/- 11.0
Electrostatic energy	-307.7 +/- 50.5
Desolvation energy	-9.5 +/- 5.1
Restraints violation energy	377.5 +/- 19.72
Buried Surface Area	1080.6 +/- 112.5
Z-Score	1.4

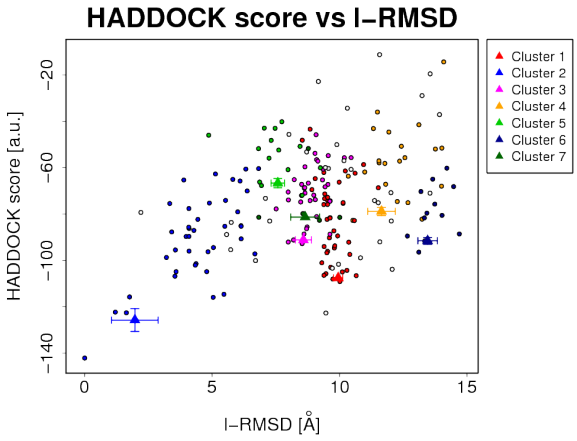


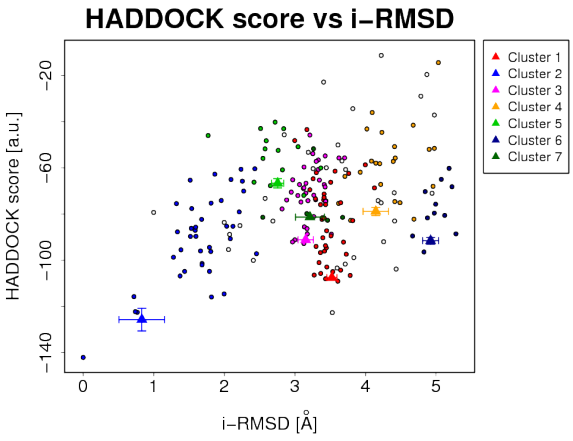
Results analysis

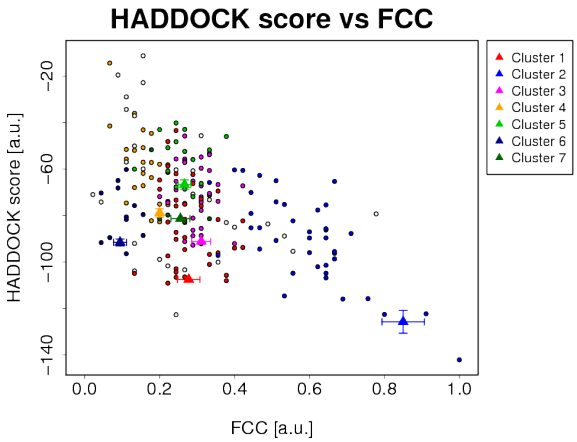
The results and graphics presented below are based on water-refined models generated by HADDOCK. The clusters are calculated based on the interface-ligand RMSDs calculated by HADDOCK, with the interface defined automatically based on all observed contacts. The various structural analysis are made with respect to the best HADDOCK model (the one with the lowest HADDOCK score).

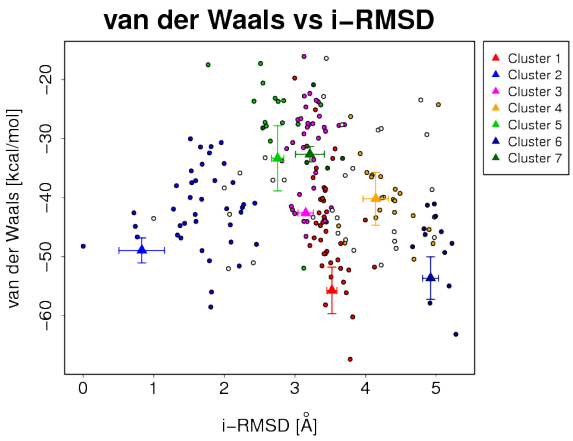
i-RMSD -> interface-RMSD calculated on the backbone (CA,C,N,O,P) atoms of all residues involved in intermolecular contact using a 10Å cutoff . **l-RMSD** -> ligand-RMSD calculated on the backbone atoms (CA,C,N,O,P) of all (N>1) molecules after fitting on the backbone atoms of the first (N=1) molecule **FCC** -> Fraction of common contacts. The intermolecular contacts are defined based on the best HADDOCK model using a 5Å cutoff (see Rodrigues et al, Proteins 2012). The average values are calculated on the best 4 structures of each clusters (based on the HADDOCK score).

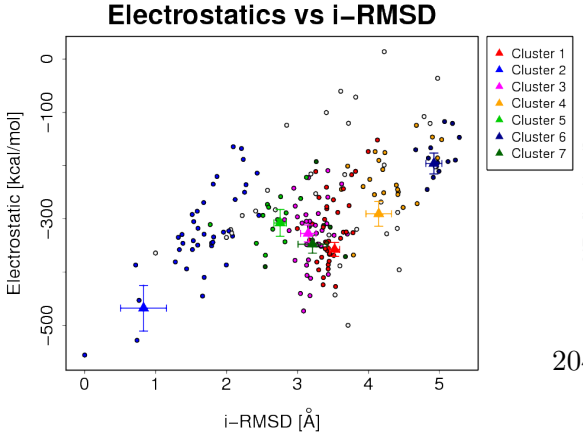


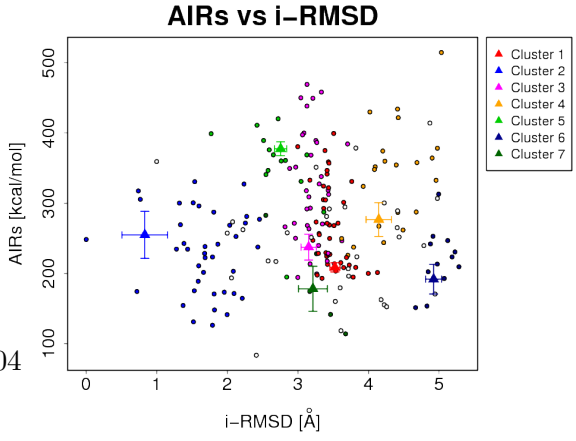












2.12. References

78. Upton, R., Migas, L. G., Pacholarz, K. J., Beniston, R. G., Estdale, S., Firth, D. & Barran, P. E. Hybrid Mass Spectrometry Methods Reveal Lot-to-Lot Differences and Delineate the Effects of Glycosylation on the Tertiary Structure of Herceptin®. *Chemical Science* **10,** 2811–2820 (2019).

80. Spearman, M., Dionne, B. & Butler, M. in *Antibody Expression and Production* (ed Al-Rubeai, M.) 251–292 (Springer Netherlands, Dordrecht, 2011).

115. Thaysen-Andersen, M., Packer, N. H. & Schulz, B. L. Maturing Glycoproteomics Technologies Provide Unique Structural Insights into the N-Glycoproteome and Its Regulation in Health and Disease. *Molecular & Cellular Proteomics* **15,** 1773–1790 (2016).

196. Simm, S., Einloft, J., Mirus, O. & Schleiff, E. 50 Years of Amino Acid Hydrophobicity Scales: Revisiting the Capacity for Peptide Classification. *Biological Research* **49,** 31 (2016).

239. Johnson, R. S., Martin, S. A., Biemann, K., Stults, J. T. & Watson, J. T. Novel Fragmentation Process of Peptides by Collision-Induced Decomposition in a Tandem Mass Spectrometer: Differentiation of Leucine and Isoleucine. *Analytical Chemistry* **59,** 2621–2625 (1987).

241. Xiao, Y., Vecchi, M. M. & Wen, D. Distinguishing between Leucine and Isoleucine by Integrated LC–MS Analysis Using an Orbitrap Fusion Mass Spectrometer. *Analytical Chemistry* **88,** 10757–10766 (2016).

242. Zhokhov, S. S., Kovalyov, S. V., Samgina, T. Y. & Lebedev, A. T. An EThcD-Based Method for Discrimination of Leucine and Isoleucine Residues in Tryptic Peptides. *Journal of The American Society for Mass Spectrometry* **28,** 1600–1611 (2017).

272. Frese, C. K., Altelaar, A. F. M., Hennrich, M. L., Nolting, D., Zeller, M., Griep-Raming, J., Heck, A. J. R. & Mohammed, S. Improved Peptide Identification by Targeted Fragmentation Using CID, HCD and ETD on an LTQ-Orbitrap Velos. *Journal of Proteome Research* **10,** 2377–2388 (2011).

305. 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. *International Journal of Proteomics* **2012,** e560391 (2012).

366. Ma, B. Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of the American Society for Mass Spectrometry* **26,** 1885–1894 (2015).

368. Yang, H., Chi, H., Zeng, W.-F., Zhou, W.-J. & He, S.-M. pNovo 3: Precise de Novo Peptide Sequencing Using a Learning-to-Rank Framework. *Bioinformatics* **35,** i183–i190 (2019).

369. Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A. & Lajoie, G. PEAKS: Powerful Software for Peptide de Novo Sequencing by Tandem Mass Spectrometry. *Rapid Communications in Mass Spectrometry* **17,** 2337–2342 (2003).

370. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De Novo Peptide Sequencing by Deep Learning. *Proceedings of the National Academy of Sciences* **114,** 8247–8252 (2017).

371. Miller, S. E., Rizzo, A. I. & Waldbauer, J. R. Postnovo: Postprocessing Enables Accurate and FDR-Controlled de Novo Peptide Sequencing. *Journal of Proteome Research* **17,** 3671–3680 (2018).

372. Tran, N. H., Rahman, M. Z., He, L., Xin, L., Shan, B. & Li, M. Complete De Novo Assembly of Monoclonal Antibody Sequences. *Scientific Reports* **6,** 31730 (2016).

500. Reichert, J. *Antibody Therapeutics Approved or in Regulatory Review in the EU or US* The Antibody Society. https://www.antibodysociety.org/resources/approved-antibodies/ (2020).

501. Murphy, K. & Weaver, C. *Janeway's Immunobiology* 9 edition. 924 pp. (Garland Science, New York, NY, 2016).

502. Yuan, J., Adamski, R. & Chen, J. Focus on Histone Variant H2AX: To Be or Not to Be. *FEBS Letters* **584,** 3717–3724 (2010).

503. Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. DNA Double-Stranded Breaks Induce Histone H2AX Phosphorylation on Serine 139. *Journal of Biological Chemistry* **273,** 5858–5868 (1998).

504. Burma, S., Chen, B. P., Murphy, M., Kurimasa, A. & Chen, D. J. ATM Phosphorylates Histone H2AX in Response to DNA Double-Strand Breaks. *Journal of Biological Chemistry* **276,** 42462–42467 (2001).

505. Rogakou, E. P., Boon, C., Redon, C. & Bonner, W. M. Megabase Chromatin Domains Involved in DNA Double-Strand Breaks in Vivo. *The Journal of Cell Biology* **146,** 905–916 (1999).

506. Mukherjee, B., Kessinger, C., Kobayashi, J., Chen, B. P. C., Chen, D. J., Chatterjee, A. & Burma, S. DNA-PK Phosphorylates Histone H2AX during Apoptotic DNA Fragmentation in Mammalian Cells. *DNA Repair* **5,** 575–590 (2006).

507. Solier, S., Sordet, O., Kohn, K. W. & Pommier, Y. Death Receptor-Induced Activation of the Chk2- and Histone H2AX-Associated DNA Damage Response Pathways. *Molecular and Cellular Biology* **29,** 68–82 (2009).

508. Lu, C., Shi, Y., Wang, Z., Song, Z., Zhu, M., Cai, Q. & Chen, T. Serum Starvation Induces H2AX Phosphorylation to Regulate Apoptosis via P38 MAPK Pathway. *FEBS Letters* **582,** 2703–2708 (2008).

509. Chen, Q. & Yu, X. OGT Restrains the Expansion of DNA Damage Signaling. *Nucleic Acids Research* **44,** 9266–9278 (2016).

510. Stucki, M., Clapperton, J. A., Mohammad, D., Yaffe, M. B., Smerdon, S. J. & Jackson, S. P. MDC1 Directly Binds Phosphorylated Histone H2AX to Regulate Cellular Responses to DNA Double-Strand Breaks. *Cell* **123,** 1213–1226 (2005).

511. Blackford, A. N. & Jackson, S. P. ATM, ATR, and DNA-PK: The Trinity at the Heart of the DNA Damage Response. *Molecular Cell* **66,** 801–817 (2017).

512. Cook, P. J., Ju, B. G., Telese, F., Wang, X., Glass, C. K. & Rosenfeld, M. G. Tyrosine Dephosphorylation of H2AX Modulates Apoptosis and Survival Decisions. *Nature* **458,** 591–596 (2009).

513. Xiao, A. *et al.* WSTF Regulates the H2A.X DNA Damage Response via a Novel Tyrosine Kinase Activity. *Nature* **457,** 57–62 (2009).

514. Boyd, S. D. & Joshi, S. A. High-Throughput DNA Sequencing Analysis of Antibody Repertoires. *Antibodies for Infectious Diseases,* 345–362 (2015).

515. Wang, W., Singh, S., Zeng, D. L., King, K. & Nema, S. Antibody Structure, Instability, and Formulation. *Journal of Pharmaceutical Sciences* **96,** 1–26 (2007).

516. Jefferis, R. *Posttranslational Modifications and the Immunogenicity of Biotherapeutics* Journal of Immunology Research. https://www.hindawi.com/journals/jir/2016/5358272/ (2019).

517. Liu, H., Ponniah, G., Zhang, H.-M., Nowak, C., Neill, A., Gonzalez-Lopez, N., Patel, R., Cheng, G., Kita, A. Z. & Andrien, B. In Vitro and in Vivo Modifications of Recombinant and Human IgG Antibodies. *mAbs* **6,** 1145–1154 (2014).

518. Weber, B., Brandl, M. J., Pulido Cendales, M. D., Berner, C., Pradhan, T., Feind, G. M., Zacharias, M., Reif, B. & Buchner, J. A Single Residue Switch Reveals Principles of Antibody Domain Integrity. *Journal of Biological Chemistry* **293,** 17107–17118 (2018).

519. Dall'Acqua, W., Simon, A. L., Mulkerrin, M. G. & Carter, P. Contribution of Domain Interface Residues to the Stability of Antibody CH3 Domain Homodimers. *Biochemistry* **37,** 9266–9273 (1998).

520. *Guaranteed Hybridoma Sequencing* Absolute Antibody. https://absoluteantibody.com/custom-services/antibody-sequencing/hybridoma-sequencing/ (2020).

521. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).

522. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A Revolutionary Tool for Transcriptomics. *Nature Reviews Genetics* **10,** 57–63 (2009).

523. Kuniyoshi, Y., Maehara, K., Iwasaki, T., Hayashi, M., Semba, Y., Fujita, M., Sato, Y., Kimura, H., Harada, A. & Ohkawa, Y. Identification of Immunoglobulin Gene Sequences from a Small Read Number of mRNA-Seq Using Hybridomas. *PLOS ONE* **11,** e0165473 (2016).

524. Bradbury, A. R. M. *et al.* When Monoclonal Antibodies Are Not Monospecific: Hybridomas Frequently Express Additional Functional Variable Regions. *mAbs* **10,** 539–546 (2018).

525. Van den Bremer, E. T. *et al.* Human IgG Is Produced in a Pro-Form That Requires Clipping of C-Terminal Lysines for Maximal Complement Activation. *mAbs* **7,** 672–680 (2015).

526. Pham, V., Henzel, W. J., Arnott, D., Hymowitz, S., Sandoval, W. N., Truong, B.-T., Lowman, H. & Lill, J. R. De Novo Proteomic Sequencing of a Monoclonal Antibody Raised against OX40 Ligand. *Analytical Biochemistry* **352,** 77–86 (2006).

527. Bandeira, N., Pham, V., Pevzner, P., Arnott, D. & Lill, J. R. Automated *de Novo* Protein Sequencing of Monoclonal Antibodies. *Nature Biotechnology* **26,** 1336–1338 (2008).

528. Guthals, A., Gan, Y., Murray, L., Chen, Y., Stinson, J., Nakamura, G., Lill, J. R., Sandoval, W. & Bandeira, N. De Novo MS/MS Sequencing of Native Human Antibodies. *Journal of Proteome Research* **16,** 45–54 (2017).

529. Castellana, N. E. *et al.* Resurrection of a Clinical Antibody: Template Proteogenomic de Novo Proteomic Sequencing and Reverse Engineering of an Anti-Lymphotoxin-α Antibody. *PROTEOMICS* **11,** 395–405 (2011).

530. Bogdanoff, W. A., Morgenstern, D., Bern, M., Ueberheide, B. M., Sanchez-Fauquier, A. & DuBois, R. M. De Novo Sequencing and Resurrection of a Human Astrovirus-Neutralizing Antibody. *ACS Infectious Diseases* **2,** 313–321 (2016).

531. Dekker, L. J. M., Zeneyedpour, L., Brouwer, E., van Duijn, M. M., Sillevis Smitt, P. A. E. & Luider, T. M. An Antibody-Based Biomarker Discovery Method by Mass Spectrometry Sequencing of Complementarity Determining Regions. *Analytical and Bioanalytical Chemistry* **399,** 1081–1091 (2011).

532. Resemann, A., Wunderlich, D., Rothbauer, U., Warscheid, B., Leonhardt, H., Fuchser, J., Kuhlmann, K. & Suckau, D. Top-Down de Novo Protein Sequencing of a 13.6 kDa Camelid Single Heavy Chain Antibody by Matrix-Assisted Laser Desorption Ionization-Time-of-Flight/Time-of-Flight Mass Spectrometry. *Analytical Chemistry* **82,** 3283–3292 (2010).

533. Poston, C. N., Higgs, R. E., You, J., Gelfanova, V., Hale, J. E., Knierman, M. D., Siegel, R. & Gutierrez, J. A. A Quantitative Tool to Distinguish Isobaric Leucine and Isoleucine Residues for Mass Spectrometry-Based De Novo Monoclonal Antibody Sequencing. *Journal of The American Society for Mass Spectrometry* **25,** 1228–1236 (2014).

534. Peng, W., Pronker, M. F. & Snijder, J. Mass Spectrometry-Based de Novo Sequencing of the Anti-FLAG-M2 Antibody Using Multiple Proteases and a Dual Fragmentation Scheme. *bioRxiv,* 2021.01.07.425675 (2021).

535. Sen, K. I., Tang, W. H., Nayak, S., Kil, Y. J., Bern, M., Ozoglu, B., Ueberheide, B., Davis, D. & Becker, C. Automated Antibody De Novo Sequencing and Its Utility in Biopharmaceutical Discovery. *Journal of The American Society for Mass Spectrometry* **28,** 803–810 (2017).

536. Savidor, A., Barzilay, R., Elinger, D., Yarden, Y., Lindzen, M., Gabashvili, A., Tal, O. A. & Levin, Y. Database-Independent Protein Sequencing (DiPS) Enables Full-Length de Novo Protein and Antibody Sequence Determination. *Molecular & Cellular Proteomics* **16,** 1151–1161 (2017).

537. Yang, H. *et al.* Precision De Novo Peptide Sequencing Using Mirror Proteases of Ac-LysargiNase and Trypsin for Large-Scale Proteomics. *Molecular & Cellular Proteomics* **18,** 773–785 (2019).

538. Tran, N. H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A. & Li, M. Deep Learning Enables de Novo Peptide Sequencing from Data-Independent-Acquisition Mass Spectrometry. *Nature Methods* **16,** 63–66 (2019).

539. Qiao, R., Tran, N. H., Xin, L., Shan, B., Li, M. & Ghodsi, A. *DeepNovoV2: Better de Novo Peptide Sequencing with Deep Learning* `http://arxiv.org/abs/1904.08514` (2019).

540. Yang, H., Chi, H., Zhou, W.-J., Zeng, W.-F., Liu, C., Wang, R.-M., Wang, Z.-W., Niu, X.-N., Chen, Z.-L. & He, S.-M. pSite: Amino Acid Confidence Evaluation for Quality Control of De Novo Peptide Sequencing and Modification Site Localization. *Journal of Proteome Research* **17,** 119–128 (2018).

541. Avtonomov, D. M., Kong, A. & Nesvizhskii, A. I. DeltaMass: Automated Detection and Visualization of Mass Shifts in Proteomic Open-Search Results. *Journal of Proteome Research* **18,** 715–720 (2019).

542. Rickert, K. W., Grinberg, L., Woods, R. M., Wilson, S., Bowen, M. A. & Baca, M. Combining Phage Display with de Novo Protein Sequencing for Reverse Engineering of Monoclonal Antibodies. *mAbs* **8,** 501–512 (2016).

543. Kiyoshi, M., Tsumoto, K., Ishii-Watabe, A. & Caaveiro, J. M. M. Glycosylation of IgG-Fc: A Molecular Perspective. *International Immunology* **29,** 311–317 (2017).

544. De Haan, N., Falck, D. & Wuhrer, M. Monitoring of Immunoglobulin N- and O-Glycosylation in Health and Disease. *Glycobiology* **30,** 226–240 (2020).

545. Valliere-Douglass, J. F., Brady, L. J., Farnsworth, C., Pace, D., Balland, A., Wallace, A., Wang, W., Treuheit, M. J. & Yan, B. O-Fucosylation of an Antibody Light Chain: Characterization of a Modification Occurring on an IgG1 Molecule. *Glycobiology* **19,** 144–152 (2009).

546. Bongers, J. *et al.* Characterization of Glycosylation Sites for a Recombinant IgG1 Monoclonal Antibody and a CTLA4-Ig Fusion Protein by Liquid Chromatography–Mass Spectrometry Peptide Mapping. *Journal of Chromatography A* **1218,** 8140–8149 (2011).

547. Plomp, R. *et al.* Hinge-Region O-Glycosylation of Human Immunoglobulin G3 (IgG3). *Molecular & Cellular Proteomics : MCP* **14,** 1373–1384 (2015).

548. Stavenhagen, K., Gahoual, R., Vega, E. D., Palmese, A., Ederveen, A. L. H., Cutillo, F., Palinsky, W., Bierau, H. & Wuhrer, M. Site-Specific N- and O-Glycosylation Analysis of Atacicept. *mAbs* **11,** 1053–1063 (2019).

549. Huhn, C., Selman, M. H. J., Ruhaak, L. R., Deelder, A. M. & Wuhrer, M. IgG Glycosylation Analysis. *PROTEOMICS* **9,** 882–913 (2009).

550. Zhang, L., Luo, S. & Zhang, B. Glycan Analysis of Therapeutic Glycoproteins. *mAbs* **8,** 205–215 (2016).

551. Aich, U., Lakbub, J. & Liu, A. State-of-the-Art Technologies for Rapid and High-Throughput Sample Preparation and Analysis of N-Glycans from Antibodies. *ELECTROPHORESIS* **37,** 1468–1488 (2016).

552. Kolarich, D., Jensen, P. H., Altmann, F. & Packer, N. H. Determination of Site-Specific Glycan Heterogeneity on Glycoproteins. *Nature Protocols* **7,** 1285–1298 (2012).

553. Rosati, S., Yang, Y., Barendregt, A. & Heck, A. J. R. Detailed Mass Analysis of Structural Heterogeneity in Monoclonal Antibodies Using Native Mass Spectrometry. *Nature Protocols* **9,** 967–976 (2014).

554. J. Thompson, N., Rosati, S., J. Rose, R. & R. Heck, A. J. The Impact of Mass Spectrometry on the Study of Intact Antibodies: From Post-Translational Modifications to Structural Analysis. *Chemical Communications* **49,** 538–548 (2013).

555. Thompson, N. J., Hendriks, L. J., de Kruif, J., Throsby, M. & Heck, A. J. Complex Mixtures of Antibodies Generated from a Single Production Qualitatively and Quantitatively Evaluated by Native Orbitrap Mass Spectrometry. *mAbs* **6,** 197–203 (2014).

556. Wohlschlager, T., Scheffler, K., Forstenlehner, I. C., Skala, W., Senn, S., Damoc, E., Holzmann, J. & Huber, C. G. Native Mass Spectrometry Combined with Enzymatic Dissection Unravels Glycoform Heterogeneity of Biopharmaceuticals. *Nature Communications* **9,** 1–9 (2018).

557. Tran, B. Q. *et al.* Comprehensive Glycosylation Profiling of IgG and IgG-Fusion Proteins by Top-down MS with Multiple Fragmentation Techniques. *Journal of Proteomics. Towards Deciphering Proteomes via the Proteoform, Protein Speciation, Moonlighting and Protein Code Concepts* **134,** 93–101 (2016).

558. Fornelli, L., Ayoub, D., Aizikov, K., Liu, X., Damoc, E., Pevzner, P. A., Makarov, A., Beck, A. & Tsybin, Y. O. Top-down Analysis of Immunoglobulin G Isotypes 1 and 2 with Electron Transfer Dissociation on a High-Field Orbitrap Mass Spectrometer. *Journal of Proteomics* **159,** 67–76 (2017).

559. Leoz, M. L. A. D. *et al.* NIST Interlaboratory Study on Glycosylation Analysis of Monoclonal Antibodies: Comparison of Results from Diverse Analytical Methods. *Molecular & Cellular Proteomics* (2019).

560. Zauner, G., Selman, M. H. J., Bondt, A., Rombouts, Y., Blank, D., Deelder, A. M. & Wuhrer, M. Glycoproteomic Analysis of Antibodies. *Molecular & Cellular Proteomics* **12,** 856–865 (2013).

561. Harris, U. J., Larson, S. B. & Mcpherson, A. in *Advances in Immunology* 191–208 (Elsevier, 1999).

562. Padlan, E. A. in *Advances in Protein Chemistry* (eds Richards, F. M., Eisenberg, D. E. & Kim, P. S.) 57–133 (Academic Press, 1996).

563. Chothia, C. & Lesk, A. M. Canonical Structures for the Hypervariable Regions of Immunoglobulins. *Journal of Molecular Biology* **196,** 901–917 (1987).

564. Chothia, C. *et al.* Conformations of Immunoglobulin Hypervariable Regions. *Nature* **342,** 877–883 (6252 1989).

565. Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. Structural Repertoire of the Human VH Segments. *Journal of Molecular Biology* **227,** 799–817 (1992).

566. Tomlinson, I. M., Cox, J. P., Gherardi, E., Lesk, A. M. & Chothia, C. The Structural Repertoire of the Human V Kappa Domain. *The EMBO Journal* **14,** 4628–4638 (1995).

567. Marks, C. & Deane, C. M. Antibody H3 Structure Prediction. *Computational and Structural Biotechnology Journal* **15,** 222–231 (2017).

568. Weitzner, B. D., Jeliazkov, J. R., Lyskov, S., Marze, N., Kuroda, D., Frick, R., Adolf-Bryfogle, J., Biswas, N., Dunbrack, R. L. & Gray, J. J. Modeling and Docking of Antibody Structures with Rosetta. *Nature Protocols* **12,** 401–416 (2017).

569. Leem, J., Dunbar, J., Georges, G., Shi, J. & Deane, C. M. ABodyBuilder: Automated Antibody Structure Prediction with Data–Driven Accuracy Estimation. *mAbs* **8,** 1259–1268 (2016).

570. Raybould, M. I. J., Wong, W. K. & Deane, C. M. Antibody–Antigen Complex Modelling in the Era of Immunoglobulin Repertoire Sequencing. *Molecular Systems Design & Engineering* **4,** 679–688 (2019).

571. Yamashita, T. Toward Rational Antibody Design: Recent Advancements in Molecular Dynamics Simulations. *International Immunology* **30,** 133–140 (2018).

572. Margreitter, C., Mayrhofer, P., Kunert, R. & Oostenbrink, C. Antibody Humanization by Molecular Dynamics Simulations—in-Silico Guided Selection of Critical Backmutations. *Journal of Molecular Recognition* **29,** 266–275 (2016).

573. Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A. & Ma, B. PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Molecular & Cellular Proteomics* **11,** M111.010587 (2012).

574. Han, X., He, L., Xin, L., Shan, B. & Ma, B. PeaksPTM: Mass Spectrometry-Based Identification of Peptides with Unspecified Modifications. *Journal of Proteome Research* **10,** 2930–2936 (2011).

575. Han, Y., Ma, B. & Zhang, K. Spider: Software for Protein Identification from Sequence Tags with de Novo Sequencing Error. *Journal of Bioinformatics and Computational Biology* **03,** 697–716 (2005).

576. Ma, B. & Johnson, R. De Novo Sequencing and Homology Searching. *Molecular & Cellular Proteomics* **11** (2012).

577. Medzihradszky, K. F. & Chalkley, R. J. Lessons in de Novo Peptide Sequencing by Tandem Mass Spectrometry. *Mass Spectrometry Reviews* **34,** 43–63 (2015).

578. Singh, C., Zampronio, C. G., Creese, A. J. & Cooper, H. J. Higher Energy Collision Dissociation (HCD) Product Ion-Triggered Electron Transfer Dissociation (ETD) Mass Spectrometry for the Analysis of N-Linked Glycoproteins. *Journal of Proteome Research* **11,** 4517–4525 (2012).

579. Allcorn, L. C. & Martin, A. C. R. SACS—Self-Maintaining Database of Antibody Crystal Structure Information. *Bioinformatics* **18,** 175–181 (2002).

580. Kurcinski, M., Badaczewska-Dawid, A., Kolinski, M., Kolinski, A. & Kmiecik, S. Flexible Docking of Peptides to Proteins Using CABS-Dock. *Protein Science* **29,** 211–222 (2020).

581. Van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., Melquiond, A. S. J., van Dijk, M., de Vries, S. J. & Bonvin, A. M. J. J. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology. Computation Resources for Molecular Biology* **428,** 720–725 (2016).

582. Harris, L. J., Skaletsky, E. & McPherson, A. Crystallographic Structure of an Intact IgG1 Monoclonal antibody11Edited by I. A. Wilson. *Journal of Molecular Biology* **275,** 861–872 (1998).

583. Fadda, E. & Woods, R. J. Molecular Simulations of Carbohydrates and Protein–Carbohydrate Interactions: Motivation, Issues and Prospects. *Drug Discovery Today* **15,** 596–609 (2010).

584. Foley, B. L., Tessier, M. B. & Woods, R. J. Carbohydrate Force Fields. *WIREs Computational Molecular Science* **2,** 652–697 (2012).

585. Grant, O. C. & Woods, R. J. Recent Advances in Employing Molecular Modelling to Determine the Specificity of Glycan-Binding Proteins. *Current Opinion in Structural Biology* **28,** 47–55 (2014).

586. Nivedha, A. K., Thieker, D. F., Makeneni, S., Hu, H. & Woods, R. J. Vina-Carb: Improving Glycosidic Angles during Carbohydrate Docking. *Journal of Chemical Theory and Computation* **12,** 892–901 (2016).

587. Woods, R. J. Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. *Chemical Reviews* **118,** 8005–8024 (2018).

588. Muth, T., Weilnböck, L., Rapp, E., Huber, C. G., Martens, L., Vaudel, M. & Barsnes, H. DeNovoGUI: An Open Source Graphical User Interface for de Novo Sequencing of Tandem Mass Spectra. *Journal of Proteome Research* **13,** 1143–1146 (2014).

589. Swindells, M. B., Porter, C. T., Couch, M., Hurst, J., Abhinandan, K. R., Nielsen, J. H., Macindoe, G., Hetherington, J. & Martin, A. C. R. abYsis: Integrated Antibody Sequence and Structure—Management, Analysis, and Prediction. *Journal of Molecular Biology. Computation Resources for Molecular Biology* **429,** 356–364 (2017).

590. Bioinformatics Solutions Inc. *PEAKS AB Version 2.0 User Guide* 2018.

591. Mann, M. & Wilm, M. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Analytical Chemistry* **66,** 4390–4399 (1994).

592. Muth, T., Hartkopf, F., Vaudel, M. & Renard, B. Y. A Potential Golden Age to Come – Current Tools, Recent Use Cases, and Future Avenues for de Novo Sequencing in Proteomics. *PROTEOMICS* **0,** 1700150 (ja 2018).

593. Griffin, L. M., Snowden, J. R., Lawson, A. D., Wernery, U., Kinne, J. & Baker, T. S. Analysis of Heavy and Light Chain Sequences of Conventional Camelid Antibodies from Camelus Dromedarius and Camelus Bactrianus Species. *Journal of Immunological Methods* **405,** 35–46 (2014).

594. Moruz, L. & Käll, L. Peptide Retention Time Prediction. *Mass Spectrometry Reviews* **36,** 615–623 (2017).

595. Escher, C., Reiter, L., MacLean, B., Ossola, R., Herzog, F., Chilton, J., MacCoss, M. J. & Rinner, O. Using iRT, a Normalized Retention Time for More Targeted Measurement of Peptides. *PROTEOMICS* **12,** 1111–1121 (2012).

596. Bruderer, R., Bernhardt, O. M., Gandhi, T. & Reiter, L. High-Precision iRT Prediction in the Targeted Analysis of Data-Independent Acquisition and Its Impact on Identification and Quantitation. *PROTEOMICS* **16,** 2246–2256 (2016).

597. Grigoryan, M., Shamshurin, D., Spicer, V. & Krokhin, O. V. Unifying Expression Scale for Peptide Hydrophobicity in Proteomic Reversed Phase High-Pressure Liquid Chromatography Experiments. *Analytical Chemistry* **85,** 10878–10886 (2013).

598. Krokhin, O. V. Sequence-Specific Retention Calculator. Algorithm for Peptide Retention Prediction in Ion-Pair RP-HPLC: Application to 300- and 100-Å Pore Size C18 Sorbents. *Analytical Chemistry* **78,** 7785–7795 (2006).

599. Spicer, V., Yamchuk, A., Cortens, J., Sousa, S., Ens, W., Standing, K. G., Wilkins, J. A. & Krokhin, O. V. Sequence-Specific Retention Calculator. A Family of Peptide Retention Time Prediction Algorithms in Reversed-Phase HPLC: Applicability to Various Chromatographic Conditions and Columns. *Analytical Chemistry* **79,** 8762–8768 (2007).

600. Tang, L., Sundaram, S., Zhang, J., Carlson, P., Matathia, A., Parekh, B., Zhou, Q. & Hsieh, M.-C. Conformational Characterization of the Charge Variants of a Human IgG1 Monoclonal Antibody Using H/D Exchange Mass Spectrometry. *mAbs* **5,** 114–125 (2013).

601. Tsubaki, M., Terashima, I., Kamata, K. & Koga, A. C-Terminal Modification of Monoclonal Antibody Drugs: Amidated Species as a General Product-Related Substance. *International Journal of Biological Macromolecules* **52,** 139–147 (2013).

602. Füssl, F., Trappe, A., Cook, K., Scheffler, K., Fitzgerald, O. & Bones, J. Comprehensive Characterisation of the Heterogeneity of Adalimumab via Charge Variant Analysis Hyphenated On-Line to Native High Resolution Orbitrap Mass Spectrometry. *mAbs* **11,** 116–128 (2018).

603. Mozdzanowski, J. in *Protein Sequencing Protocols* (ed Smith, B. J.) 365–369 (Humana Press, Totowa, NJ, 2003).

604. Bordeaux, J., Welsh, A. W., Agarwal, S., Killiam, E., Baquero, M. T., Hanna, J. A., Anagnostou, V. K. & Rimm, D. L. Antibody Validation. *BioTechniques* **48,** 197–209 (2010).

605. Berglund, L. *et al.* A Genecentric Human Protein Atlas for Expression Profiles Based on Antibodies. *Molecular & Cellular Proteomics* **7,** 2019–2027 (2008).

606. Bradbury, A. & Plückthun, A. Reproducibility: Standardize Antibodies Used in Research. *Nature News* **518,** 27 (2015).

607. Begley, C. G. & Ellis, L. M. Raise Standards for Preclinical Cancer Research. *Nature* **483,** 531–533 (7391 2012).

608. Collins, F. S. & Tabak, L. A. Policy: NIH Plans to Enhance Reproducibility. *Nature News* **505,** 612 (2014).

609. Wang, T. *et al.* LC/MS-Based Intact IgG and Released Glycan Analysis for Bioprocessing Applications. *Biotechnology Journal* **13,** 1700185 (2018).

610. Singh, N. *et al.* Dual Recognition of Phosphoserine and Phosphotyrosine in Histone Variant H2A.X by DNA Damage Response Protein MCPH1. *Proceedings of the National Academy of Sciences* **109,** 14381–14386 (2012).

611. Brademan, D. R., Riley, N. M., Kwiecien, N. W. & Coon, J. J. Interactive Peptide Spectral Annotator: A Versatile Web-Based Tool for Proteomic Applications. *Molecular & Cellular Proteomics* (2019).

612. Joughin, B. A., Tidor, B. & Yaffe, M. B. A Computational Method for the Analysis and Prediction of Protein:Phosphopeptide-Binding Sites. *Protein Science* **14,** 131–139 (2005).

613. Yaffe, M. B. & Smerdon, S. J. PhosphoSerine/Threonine Binding Domains: You Can't pSERious? *Structure* **9,** R33–R38 (2001).

614. Campbell, S. J., Edwards, R. A. & Glover, J. N. M. Comparison of the Structures and Peptide Binding Specificities of the BRCT Domains of MDC1 and BRCA1. *Structure* **18,** 167–176 (2010).

615. Lees, W. D., Stejskal, L., Moss, D. S. & Shepherd, A. J. Investigating Substitutions in Antibody–Antigen Complexes Using Molecular Dynamics: A Case Study with Broad-Spectrum, Influenza A Antibodies. *Frontiers in Immunology* **8** (2017).

616. Kumari, R., Kumar, R. & Lynn, A. G_mmpbsa—A GROMACS Tool for High-Throughput MM-PBSA Calculations. *Journal of Chemical Information and Modeling* **54,** 1951–1962 (2014).

617. Xu, J., Tack, D., Hughes, R. A., Ellington, A. D. & Gray, J. J. Structure-Based Non-Canonical Amino Acid Design to Covalently Crosslink an Antibody–Antigen Complex. *Journal of Structural Biology* **185,** 215–222 (2014).

618. Xu, X. & Stern, D. F. NFBD1/KIAA0170 Is a Chromatin-Associated Protein Involved in DNA Damage Signaling Pathways. *Journal of Biological Chemistry* **278,** 8795–8803 (2003).

619. Cornelissen, B., Kersemans, V., Darbar, S., Thompson, J., Shah, K., Sleeth, K., Hill, M. A. & Vallis, K. A. Imaging DNA Damage In Vivo Using γH2AX-Targeted Immunoconjugates. *Cancer Research* **71,** 4539–4549 (2011).

620. Lara-Ochoa, F., Vargas-Madrazo, E. & Almagro, J. C. Distributions of the Use Frequencies of Amino Acids in the Hypervariable Regions of Immunoglobulins. *Journal of Molecular Evolution* **41,** 98–103 (1995).

621. Lara-Ochoa, F., Almagro, J. C., Vargas-Madrazo, E. & Conrad, M. Antibody-Antigen Recognition: A Canonical Structure Paradigm. *Journal of Molecular Evolution* **43,** 678–684 (1996).

622. DeKosky, B. J. *et al.* Large-Scale Sequence and Structural Comparisons of Human Naive and Antigen-Experienced Antibody Repertoires. *Proceedings of the National Academy of Sciences* **113,** E2636–E2645 (2016).

623. Akbar, R. *et al.* A Compact Vocabulary of Paratope-Epitope Interactions Enables Predictability of Antibody-Antigen Binding. *bioRxiv,* 759498 (2020).

624. Chiu, M. L., Goulet, D. R., Teplyakov, A. & Gilliland, G. L. Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies* **8,** 55 (4 2019).

625. Sela-Culang, I., Kunik, V. & Ofran, Y. The Structural Basis of Antibody-Antigen Recognition. *Frontiers in Immunology* **4** (2013).

626. Bernardi, A., Faller, R., Reith, D. & Kirschner, K. N. ACPYPE Update for Nonuniform 1–4 Scale Factors: Conversion of the GLYCAM06 Force Field from AMBER to GROMACS. *SoftwareX* **10,** 100241 (2019).

627. Bumba, L., Laaf, D., Spiwok, V., Elling, L., Křen, V. & Bojarová, P. Poly-N-Acetyllactosamine Neo-Glycoproteins as Nanomolar Ligands of Human Galectin-3: Binding Kinetics and Modeling. *International Journal of Molecular Sciences* **19,** 372 (2018).

628. Bountra, C. *We Are Not Good at Translating Lab Science Into New Medicines for Patients | University of Oxford Podcasts - Audio and Video Lectures* `https://podcasts.ox.ac.uk/we-are-not-good-translating-lab-science-new-medicines-patients` (2020).

629. Crouch, D. & Rantanen, J. *Centocor v. Abbott: Written Description and Antibody Claims* Patently-O. `https://patentlyo.com/patent/2011/02/centocor-v-abbott.html` (2020).

630. Zuhn, D. *Centocor Ortho Biotech, Inc. v. Abbott Laboratories (Fed. Cir. 2011)* Patent Docs. `https://www.patentdocs.org/2011/02/centocor-ortho-biotech-inc-v-abbott-laboratories-fed-cir-2011.html` (2020).

631. Norman, R. A., Ambrosetti, F., Bonvin, A. M. J. J., Colwell, L. J., Kelm, S., Kumar, S. & Krawczyk, K. Computational Approaches to Therapeutic Antibody Design: Established Methods and Emerging Trends. *Briefings in Bioinformatics* (2019).

632. *abYmod v 1.20, Prof. Andrew C. R. Martin, UCL* `http://abymod.abysis.org/` (2020).

Durch Heftigkeit ersetzt der Irrende, was ihm an Wahrheit und Kräften fehlt.
The astray replaces through vehemence what he lacks in truth and force.

— Johann Wolfgang von Goethe, *Torquato Tasso*

3

Energy-resolved fragmentation of *O*-linked glycopeptides

Contents

3.1. Abstract . . . 216
3.2. Introduction: Glycopeptide fragmentation and analysis . . . 217
3.2.1. Fragmentation mechanisms of glycopeptides . . . 217
3.2.1.1. Nomenclature and general aspects . . . 217
3.2.1.2. CID-based fragmentation mechanisms . . . 219
3.2.2. ETD-based fragmentation mechanisms . . . 224
3.2.3. The complementarity paradigm of glycopeptide fragmentation . . . 228
3.2.4. Hybrid fragmentation methods . . . 228
3.2.5. Energy-resolved fragmentation of glycopeptides . . . 230
3.2.5.1. Fragmentation energetics differ between peptides and glycopeptides 230
3.2.5.2. Activation and dissociation channels are energetically linked to structural features in *N*-glycopeptides . . . 231
3.2.5.3. Stepped collision energy methods for fragmentation of *N*-glycopeptides . . . 233
3.2.5.4. Difference in optimal dissociation methods for *N*- and *O*-glycopeptides are related to their distinct structures . . . 235
3.2.6. Data analysis and interpretation of glycopeptide MS/MS spectra . . . 237
3.3. Objective . . . 239
3.4. Materials & methods . . . 240
3.4.1. Sample preparation . . . 240
3.4.2. Tandem mass spectrometry methods . . . 241
3.4.2.1. Preparation of nLC columns for electrospray ionisation . . . 241
3.4.2.2. Data acquisition for glycosylated samples . . . 242

3.4.3. Data analysis . . . 245
3.4.3.1. Search parameters . . . 245
3.4.3.2. Identification of novel *O*-glycopeptides and database curation . . 246
3.4.3.3. Rationale and flaws for the selected Byonic scoring metrics . . . 246
3.5. Results . . . 249
3.5.1. A case study: Identifying the linker *O*-glycopeptides of heparinase-trimmed decorin . . . 249
3.5.2. Investigating the generality of low-NCE energy-resolved fragmentation of *O*-glycopeptides . . . 262
3.5.2.1. Single NCE runs pinpoint the collision energy range for glycan fragmentation of *O*-glycopeptides between 10 and 20 % . . . 262
3.5.2.2. Low-energy SCE-HCD methods are advantageous for *O*-glycopeptide glycan sequencing . . . 266
3.5.2.3. Low-energy HCD methods affect the fragmentation and scoring of other components . . . 272
3.5.3. ETD and EThcD methods remain essential for glycan site localisation . . 275
3.5.3.1. Implementing EThcD on an Orbitrap Elite instrument for glycopeptide analysis . . . 275
3.5.3.2. Analysis of the model glycopeptides by ETD and EThcD . . . 277
3.6. Conclusion and outlook . . . 280
3.7. Author contributions . . . 281
3.8. References . . . 282
3.9. Appendix . . . 290
3.9.1. Supplementary discussion . . . 290
3.9.1.1. Supplementary discussion 1: Collision energy values in the literature and their conversion to Thermo's NCE . . . 290
3.9.2. Supplementary figures, tables & lists . . . 292
3.9.2.1. List of Byonic keywords . . . 292
3.9.2.2. Data for the decorin case study . . . 293
3.10. Supplementary references . . . 314

3.1. Abstract

Protein *O*-glycosylation is notoriously difficult to study, efforts are ongoing to develop tandem mass spectrometry techniques able to comprehensively characterise the glycan chain and its site of attachment on the protein. Several fragmentation methods have been deployed to interrogate *O*-glycopeptides, including collisional activation methods (CID, HCD) and electron-transfer methods (ETD, EThcD). This chapter aims at obtaining a more elaborate understanding of the energetic accessibility of different fragmentation pathways and how these can be used in conventional proteomics equipment. The data acquired on purified *O*-glycoproteins shows that the fragmentation behaviour and the relevant energy levels differ for *N*- and *O*-linked glycopeptides. *O*-glycopeptides show extensive glycan fragmentation at lower energies than *N*-glycopeptides. I propose to ex-

ploit these findings in low-energy stepped-HCD methods and obtain encouraging results with 10-20%, 10-20-30% and 10-15-20% NCE settings.

3.2. Introduction: Glycopeptide fragmentation and analysis

The following introduction complements the technical introduction on tandem mass spectrometry of peptides and glycopeptides given in chapter 1.3 specifically on the topics of glycopeptide fragmentation and their analysis. Aspects relating to sample preparation, nano-LC chromatography and glycobiology in general are described in chapter 1.

3.2.1. Fragmentation mechanisms of glycopeptides

3.2.1.1. Nomenclature and general aspects

The exact fragmentation behaviour of glycoconjugates depends on the nature of the analyte: released native glycans, glycans derivatised at their reducing terminus, derivatised glycans (such as permethylation) or glycopeptides share common fragmentation principles, but can differ in the observed fragment ions.[219,633] The presence of salt adducts (such as Na^+ or Li^+) further affects which fragment ions are observed, by means of influencing the fragmentation mechanism.[634] The following discussion focusses primarily on glycopeptides of the $[M+nH^+]^{n+}$ type obtained by positive mode electrospray ionisation (ESI). ESI is advantageous for sialylated and sulfated glycans compared to MALDI because of their lower tendency to fragment during ionisation (both by in-source and post-source decay).[172,218,552] The reader is kindly referred to the literature for the fragmentation behaviour of other types of glycoconjugates and fragmentations by other ionisation sources (EI, FAB, MALDI or negative mode ESI), which are beyond the scope of this thesis.[180,218–220,633–636]

Before starting the discussion on fragmentation of glycopeptides, it is worth mentioning that the ionisation efficiency of glycopeptides in positive mode nano-flow ESI is of 40-60 % of their non-glycosylated counterparts[172] (see 1.3.4.1 for comments on ionisation methods). Mass-spectrometric detection of glycopeptides in the presence of peptides is difficult because of ion suppression effects,[219,637] weaker ionisation[172,638], glycosylation micro- and macroheterogeneity[639] limiting precursor ion selection in data-dependent acquisition methods[552]. However, the relative signal intensity of different glycopeptides in a given sample is unaffected by the presence of other peptides, enabling the use of MS^1 signals for label-free quantitation.[172,640]

Glycopeptides can be fragmented along their peptide or glycan part, providing sequence information in the MS^2 spectra. The nomenclature of Roepstorff, Fohlman and Biemann[237–239] for peptide fragments (see 1.3.4.3) can be expanded by using the Domon

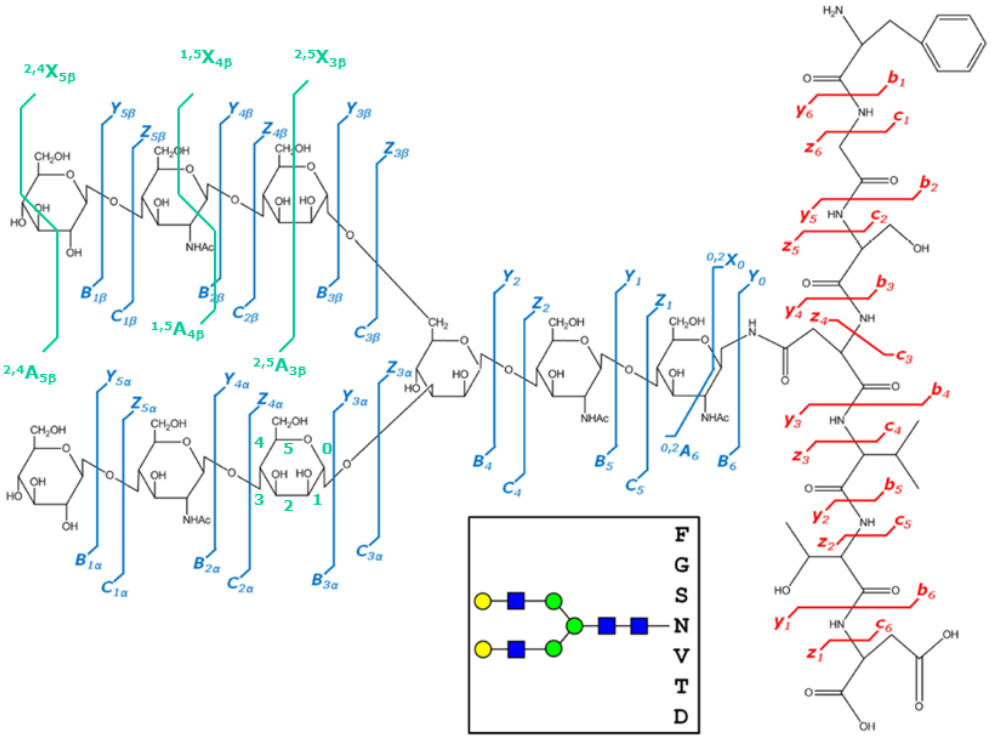


Figure 3.1. Glycopeptide fragment ion nomenclature following the Roepstorff-Fohlman system for peptides and the Domon-Costello system for glycans. Fragment ion assignment is demonstrated for an *N*-linked glycopeptide example having the composition FGSNVTD + $GlcNAc_4Man_3Gal_2$. Glycan fragments are coloured in blue / cyan, peptide fragments in red. Monosaccharides are numbered from the non-reducing end for *A*, *B* and *C* ions and from the reducing end (i.e. peptide-linked end) for *X*, *Y* and *Z* ions, numbering the glycosidic bond in subscript. Fragments from branched glycans are distinguished by Greek letters following the subscript, with α representing the largest chain. Cross-ring fragment ions (*A*, *X*) are preceded by two superscript numbers showing the broken bonds, starting with 0 for the $O-C_1$ bond. Note that for negative mode fragmentation of glycans, additional *D*, *E* and *F* ions are defined[219] (not shown here). Figure adapted from[642], with permission from ©John Wiley and Sons.

and Costello nomenclature for glycans[641] (Figure 3.1). The glycan part of glycopeptides undergoes two main types of fragmentations. *Glycosidic bond cleavage* occurs along the glycosidic linkage between two adjacent monosaccharides, it provides sequence information including branching information.[634] *Crossring cleavage* involves a bond breaking inside the sugar ring and can provide valuable information concerning linkage and branching.[634] Glycan fragments further break down to diagnostic *oxonium ions* with a unique, low m/z (Table 3.1).

$[M+nH^+]^{n+}$ type glycopeptide precursor ions form mainly B/Y glycosidic fragments with few, if any, A/X crossring cleavage ions.[219] The fragmentation of glycopeptides is different with each fragmentation techniques, which will be outlined in the following for

CID- and ETD-type fragmentations. For ion/photon interactions such as in UVPD or IRMPD, the reader is kindly referred to the literature.[642]

3.2.1.2. CID-based fragmentation mechanisms

Collision-induced dissociation (CID) fragmentations are classified as ion/neutral interactions[642] since the energy for the dissociation is provided through collision of the precursor ion with a gaseous target, thus converting translational energy to internal energy (see 1.3.4.3). Mechanistically, CID fragmentation is a proton-mediated fragmentation (Figure 1.11). The proton, located on a basic residue, is excited by collisional activation and becomes "mobile", enabling it to transfer to remote sites on the glycopeptides. The glycosidic bond is the most acid-labile bond in glycopeptides, it was suggested that the fragmentation of glycopeptides in CID follows an A_1-type cleavage (unlike for alkali metal adducts, which fragment via β-cleavage).[636,643] With peptides forming horseshoes in the gas phase,[644] it is assumed that glycans of the glycopeptide contribute in the gas-phase hydrogen-bonding network[332,645] thus yielding facile fragmentation of the glycosidic bond[646]. Notably, an S-shaped peptide backbone has been suggested in the case of *N*-glycosylation.[647]

Resulting fragments can further fragment down to oxonium ions as illustrated in Figure 3.2. These diagnostic oxonium ions irrevocably indicate the presence of a glycopeptide. They can further be used to assess the presence of a given monosaccharide or even discriminate different types of glycosidic linkages or isobaric monosaccharides of different stereochemistry (Table 3.1 and references therein). Indeed, “memory” of anomeric configuration is retained following gas-phase glycosidic bond fragmentation.[326]

Equally, *B*-ions can be indicative of diagnostic structural motives in glycans, which are particularly relevant for longer glycan.[648] As with oxonium ions, diagnostic *B* ions can be used to distinguish between glycopeptide structural isomers (Table 3.2 and references therein). For example, the presence of *B*-ions can be indicative of antenna fucosylation (α1-3 linkage) versus core fucosylation (α1-6 linkage).[649]

The peptide backbone can equally fragment in CID to b/y ions as has been described in 1.3.4.3. However, the glycosidic bond is more labile in CID and mostly fragments first − depending on the applied collision energy and the actual implementation (ion trap vs beam-type). The different fragmentation pathways of glycopeptides have an energetical hierarchy that is described in further details in 3.2.5. In addition, effects such as the charge state, peptide size and proton mobility influence the observation of glycopeptide backbone fragmentation in CID.[642]

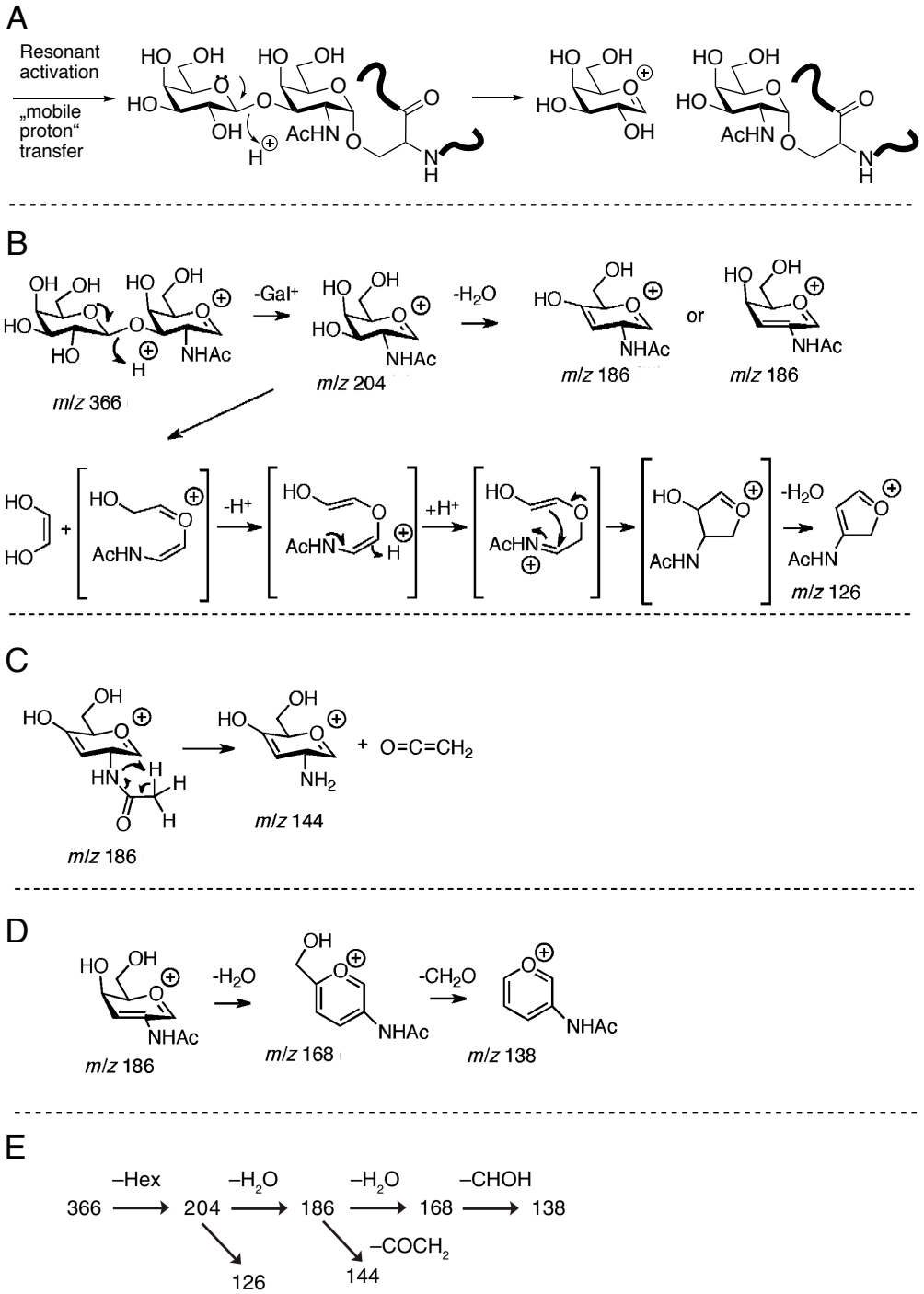
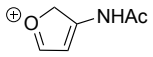
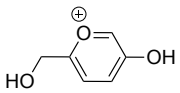
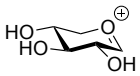
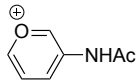
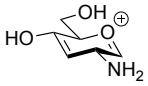
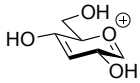
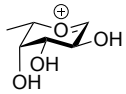
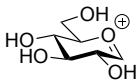
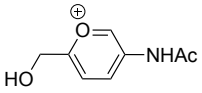
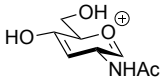
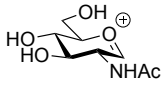
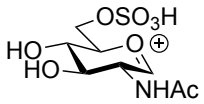


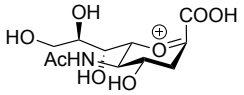
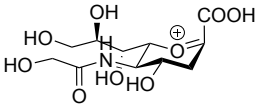
Figure 3.2. Glycopeptide fragmentation in collision-induced dissociation down to the different oxonium ions. A. Fragmentation of a core 1 *O*-glycopeptide via A_1 type mechanism.[636,643] An oxonium ion (or *B* ion) is formed together with the charge-reduced complementary Y-ion. **B.** Proposed pathway by Westerlind, Nilsson and co-workers[308] for the loss of Gal from the $[GalGalNAc]^+$ oxonium ion and further dehydration to the 186 ion or retro-Diels-Alder to the 126 ion. **C.** Cyclic elimination of the acetyl side chain. **D.** Decomposition pathways for the formation of 168 and 138 ions. **E.** Summary diagram highlighting the fragmentations sequences. Figure adapted from[308], partially reprinted with permission from ©Wiley VCH.

Table 3.1. Commonly observed oxonium ions in collision-induced fragmentations. Commonly observed oxonium ions collected from the literature and personal observation.[308,638,650] Only oxonium ions larger than 125 amu are shown. The structure of Hex- and HexNAc oxonium ions is depicted according to Westerlind, Nilsson and co-workers.[308] Note that several stereochemistries are possible, oxonium ions structures shown here are of Glc series.

Fragment ***m/z***	**Structure** [a]	**Common name or composition**	**Glycosylation type**	**Complementary Y-ion**
Oxonium ions				
126.05		$[HexNAc - C_2H_6O_3]^+$ $C_6H_8NO_2^+$	all [c]	$[M-204.09]^{z-1}$
127.06		$[Hex - 2\,H_2O]^+$ $C_7H_8NO_2^+$	all	$[M-163.06]^{z-1}$
133.05		$[Xyl]^+$ $C_5H_9O_4^+$	proteoglycan	- [b]
138.06		$[HexNAc - CH_6O_3]^+$ $C_7H_8NO_2^+$	all [c, d]	$[M-204.09]^{z-1}$
144.07		$[HexNAc - C_2H_4O_2]^+$ $C_6H_{10}NO_3^+$	all [c, d]	$[M-204.09]^{z-1}$
145.05		$[Hex - H_2O]^+$ $C_6H_7O_3^+$	all	$[M-163.06]^{z-1}$
147.08		$[Fuc]^+$ $C_6H_{11}O_4^+$	all	- [b]
163.06		$[Hex]^+$ $C_6H_{11}O_5^+$	all [a]	$[M-163.06]^{z-1}$
168.07		$[HexNAc - 2\,H_2O]^+$ $C_8H_{10}NO_3^+$	all [c]	$[M-204.09]^{z-1}$
186.08		$[HexNAc - H_2O]^+$ $C_8H_{12}NO4^+$	all	$[M-204.09]^{z-1}$
204.09		$[HexNAc]^+$ $C_8H_{14}NO_5^+$	all [a]	$[M-204.09]^{z-1}$
274.09	-	$[NeuAc - H_2O]^+$ $C_{11}H_{16}NO_7^+$	all sialo [e]	$[M-292.08]^{z-1}$
284.26		$[Sulfo\text{-}HexNAc]^+$ $C_8H_{14}NO_8S^+$	sulfated glycans [a]	- [b]
290.05	-	$[NeuGc - H_2O]^+$ $C_{11}H_{16}NO_8^+$	all sialo	$[M-308.11]^{z-1}$

(Table continued on next page.)

Table 3.1. (***continued***)

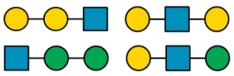
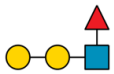
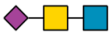
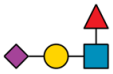
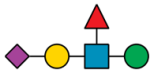
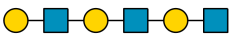
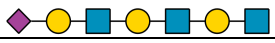
Fragment ***m/z***	**Structure** [a]	**Common name or composition**	**Glycosylation type**	**Complementary Y-ion**
Oxonium ions (continued)				
292.08		$[NeuAc]^+$ $C_{11}H_{18}NO_8^+$	all sialo [e]	$[M\text{-}292.08]^{z\text{-}1}$
308.11		$[NeuGc]^+$ $C_{11}H_{18}NO_9^+$	all sialo	$[M\text{-}308.11]^{z\text{-}1}$
316.11	- [f]	$[O\text{-AcNeuAc} - H_2O]^+$	*O*-Ac on sialo	$[M\text{-}316.11]^{z\text{-}1}$
332.10	- [f]	$[O\text{-AcNeuGc} - H_2O]^+$	*O*-Ac on sialo	$[M\text{-}332.10]^{z\text{-}1}$
334.11	- [f]	$[O\text{-AcNeuAc}]^+$	*O*-Ac on sialo	$[M\text{-}334.11]^{z\text{-}1}$
350.11	- [f]	$[O\text{-AcNeuGc}]^+$	*O*-Ac on sialo	$[M\text{-}350.11]^{z\text{-}1}$

[a] Other stereochemistry at C-3 or C-4 is possible in annotated cases (Gal or Man), here the Glc series are depicted. [b] Rarely in a non-reducing terminal position or on its own, i.e. complementary *Y*-ion results from additional fragmentation. [c] GlcNAcβ1-O- and GalNAcα1-O-substituted peptide can be distinguished by their oxonium ions (m/z138 +m/z168) to (m/z126 +m/z144) relative abundance ratios, referred to as the GlcNAc/GalNAc ratio. A ratio $2.0 - 2.5$ is indicative of GlcNAcβ-O- glycopeptides, a ratio of $0.4 - 0.6$ is indicative for GalNAcα1-O-substitution.[651]
[d] The 138 / 144 ratio enables distinguishing *N*- and *O*-glycosylation.[652, Fig.S20] [e] Sialic acid oxonium ions enable discrimination of α2,3- vs α2,6-linkages.[653] [f] Several positions can be *O*-acetylated.[650]

With reference to the energy-dependency of fragmentation pathways (see 3.2.5), it is worth highlighting the differences in fragmentation of glycopeptides in ion-trap CID and beam-type CID (also named higher energy CID, HCD) resulting from the different implementations and in fact, ion activation mechanisms (see 1.3.4.3 for details). Ion-trap CID uses a "slow heating" process which results predominantly in the cleavage of glycosidic linkages, leading to extensive B/Y and C/Z fragmentation, but little to no information of the peptide backbone.[114] Note that *N*-glycopeptides having very short peptide tags often (but not always) undergo peptide backbone cleavage via low-energy vibrational activation − depending on charge state (1^+ vs 2^+, influencing number of mobile protons) and position of the base (proximal of glycan vs distant, *C*-terminally vs interior).[642] Higher-energy ion-trap CID and beam-type CID (= HCD), the latter using accelerated ion collision, equally lead to substantial B/Y ion series and highly abundant oxonium ions (in HCD) while producing $\sim b/ \sim y$ peptide backbone fragments with loss of glycan, which can result (depending on energy and glycopeptide size) in lower glycan sequence coverage but provides information on the peptide.

Using the different implementations (ion-trap CID vs HCD) as a method to differentiate glycan and peptide fragments has been suggested,[114] yet a more subtle vision of the energetics and mechanistic processes involved in fragmentation is gaining acceptance.

Table 3.2. Commonly observed *B*-ions in collision-induced fragmentations. Commonly *B*-ions are collected from the literature and personal observations. Compiled from[638,648,649].

B*-ion *m/z	**Structure**	**Common name or composition**	**Glycosylation type**	**Complementary Y-ion**
B-ions				
350.14		Fuco-GalNAc	all	$[M-350.14]^{z\cdot1}$
366.14		HexNAcHex	all [g]	$[M-366.14]^{z\cdot1}$
407.16		LacdiNAc	all	$[M-407.16]^{z\cdot1}$
495.18		Sialo-GalNAc	all sialo	$[M-495.18]^{z\cdot1}$
528.19		Galili HexNAcHex(2)	all [g]	$[M-528.19]^{z\cdot1}$
553.22		Fuco-LacdiNAc	all	$[M-553.22]^{z\cdot1}$
657.23		HexNAcHexNeuAc	all	$[M-657.23]^{z\cdot1}$
674.25		Fuco-Galili HexNAc Hex(2)Fuc	all / *N*-glyco [f]	$[M-674.25]^{z\cdot1}$
698.26		Sialo-LacdiNAc	all	$[M-698.26]^{z\cdot1}$
699.28		Difuco-LacdiNAc	all	$[M-699.28]^{z\cdot1}$
803.29		HexNAcHexFucSia	all / *N*-glyco [f]	$[M-803.29]^{z\cdot1}$
965.35		HexNAcHex(2)Fuc-Sia	*N*-glyco [f]	$[M-965.35]^{z\cdot1}$
1055.37		PolyLacNAc	*N*-glyco	$[M-1055.37]^{z\cdot1}$
1096.41		PolyLacNAc	all	$[M-1096.41]^{z\cdot1}$
1387.49		Sialo-polyLacNAc	all	$[M-1387.49]^{z\cdot1}$

[f] The presence of these *B*-ions can be used to distinguish antennae fucosylation in *N*-glycopeptides.[649]
[g] A ratio of $I_{528}/I_{366} > 0.3$ was suggested as indicative for α-Gal motif.[654]

Complementing an energetic-centred understanding, observations suggest a preferential peptide backbone cleavage in the absence of mobile protons, which is then influenced by factors such as charge state, charge carrier and overall composition.[642]

Nonetheless, a major disadvantage of all ion-trap CID implementations is the loss of diagnostic oxonium ions and B-ions due to the 1/3rd rule.[267] Together with its lower mass accuracy than Orbitrap-detected HCD, its higher dependency on signal-to-noise and the induction of mostly single bond cleavages of glycopeptides, additionally preventing formation of important oxonium ions,[115] the use of ion-trap CID for glycopeptide analysis has limitations. Therefore, whereas earlier studies in glycopeptide fragmentation would have widely used ion-trap CID, in recent years more and more groups are moving to stepped HCD-based methods and/or hybrid fragmentation methods (see 3.2.4).

3.2.2. ETD-based fragmentation mechanisms

Both ETD and ECD lead to peptide fragmentation following a radical, $1e^-$ mechanism (explained in section 1.3.4.3). The two methods differ in their implementation, ECD uses free electron transfer (1-10 eV[279]) and is implemented in less common FT-ICR type instruments, while ETD uses an electron carrier molecule,[276,277] typically fluoranthene[278], and is implemented in widely available ion-trap instruments. As a result, the energy imparted into the isolated precursor ion by the free electron capture in ECD is greater than during electron transfer in ETD, which can result in higher vibrational excitation states being reached in ECD.[279] Additionally, the difference in vacuum leads to faster collisional cooling in ETD, reducing the likelihood of more energetic fragmentation routes.[279] The following discussion focusses primarily on ETD, used in this thesis.

Both techniques produce c' and $z^\bullet$ peptide backbone ions resulting from the scission of the N$-$C$_\alpha$ bond, which are the dominant fragment ions.[277] Occasionally, $c^\bullet$ and z' ions are observed if the $[c+z]^{(n-1)+\bullet}$ complex has a relatively long lifetime before separation[279] (see 1.3.4.3 for details). Generally speaking, little loss of glycan or glycan fragmentation are observed[638] – the exact nature of glycan fragments observed in ECD/ETD will be outlined in the following. Thus, ETD provides c'+glycan ions which, together with the $z^\bullet$ ions, allows to identify the site of attachment of the glycan. ETD is highly suited for the sequencing of the peptide backbone of glycopeptides and enables site-localisation even for densely populated *O*-glycan clusters on a single peptide.[655]

Despite of its promising fragmentation chemistry, ETD of glycopeptides suffers from several drawbacks. (i) The fragmentation efficiency is low ($\sim$ 20% for glycopeptides[638]) because the major reaction pathway observed is charge reduction (i.e. to $[\mathrm{M}+n\mathrm{H}^+]^{(n-1)+\bullet}$, ETnoD) instead of the desired fragmentation reaction.[638] This can be slightly remediated by the use of post-ETD supplemental collisional activation[277,656] in ion-trap collision cells (ETcaD[291]), which can increase fragmentation of charge-reduced precursor ions generated by ETD.[657] (ii) In addition, glycopeptides frequently feature precursor ion of low charge density ($m/z >\sim 850$) which do not undergo efficient ETD fragmentation.[658]

Although glycosylations are generally considered not to participate in ETD fragmentation, there are noteworthy exceptions in which glycan fragmentation has been observed or in which the glycosylation influences which peptide backbone fragments are formed. While loss of sialic acid is common, other cases can mostly be traced back to features in

the peptide sequence, and are instructive for further understanding the fragmentation mechanism of glycopeptides in ETD.

Glycan fragmentation to B/Y ions has been observed in cases where peptide backbone N$-$C$_\alpha$ cleavage is repressed while supplemental activation is added. Peptides with a high number of acidic residues (glutamate) exceeding the number of basic sites[659] or long peptides[660] have been shown to inhibit backbone fragmentation by stabilising the charge-reduced precursor ion $[\mathrm{M}+n\mathrm{H}^+]^{(n-1)+\bullet}$. Similarly, fragmentation efficiency in ETD is lower for highly sialylated *N*- and O-glycopeptides.[655,661] * B/Y fragmentation without c/z ions in ECD has been described by Hongo *et al.* for a cysteine-rich and Glu-rich triantennae sialylated complex glycopeptide,[663] among other examples[642, Fig.9]. Presumably, cysteine can act as a "radical trap"[663] leading to the formation of an even-electron species − which is vibrationally excited if supplemental activation is provided − and ultimately leads to ergodic fragmentations (i.e. with energy redistribution) as described in the CID mechanism.

N-glycopeptides have two amide bonds which can potentially accept an electron in their π* antibonding orbital: the amide group of HexNAc and the amide bond to Asn. The first is occasionally observed,[664] its mechanism is depicted in Figure 3.3.A. The loss of the entire Asn side chain (with glycosylation) has also been reported (Figure 3.3.B).[665] Interestingly, the glycosylation is immediately adjacent to a *C*-terminal Arg. This positively charged sites might guide the electron transfer to the nearby Asn$-$CO$-$N$-$HexNAc as has been suggested in theoretical calculations for the ETD mechanism.[289]

Even when they do not participate directly in the fragmentation, glycosylations can affect which fragments are formed. Using synthetic *N*-glycopeptides carrying only the paucimannose core in ETD, higher peptide backbone fragmentation was noted when the glycan is located at the *N*-terminus of the peptide.[661] This localised effect of glycan position does not come as a surprise as ETD is a non-ergodic process. Further, the extend of $\mathrm{H}^\bullet$ rearrangement determining $c'/z^\bullet$ vs $c^\bullet/z'$ formation is dependent on the residues adjacent to the N$-$C$_\alpha$ bond.[666] Threonine GlcNAcylation fosters $\mathrm{H}^\bullet$ transfer in ETD for the formation of immediately *C*-terminally adjacent $z^\bullet$ fragments.[667]

*Note that a side reaction of Tris reacting with Sialic acid has been recently described, which might imply that some fragments could have been overlooked or precursor ion m/z misidentified.[662]

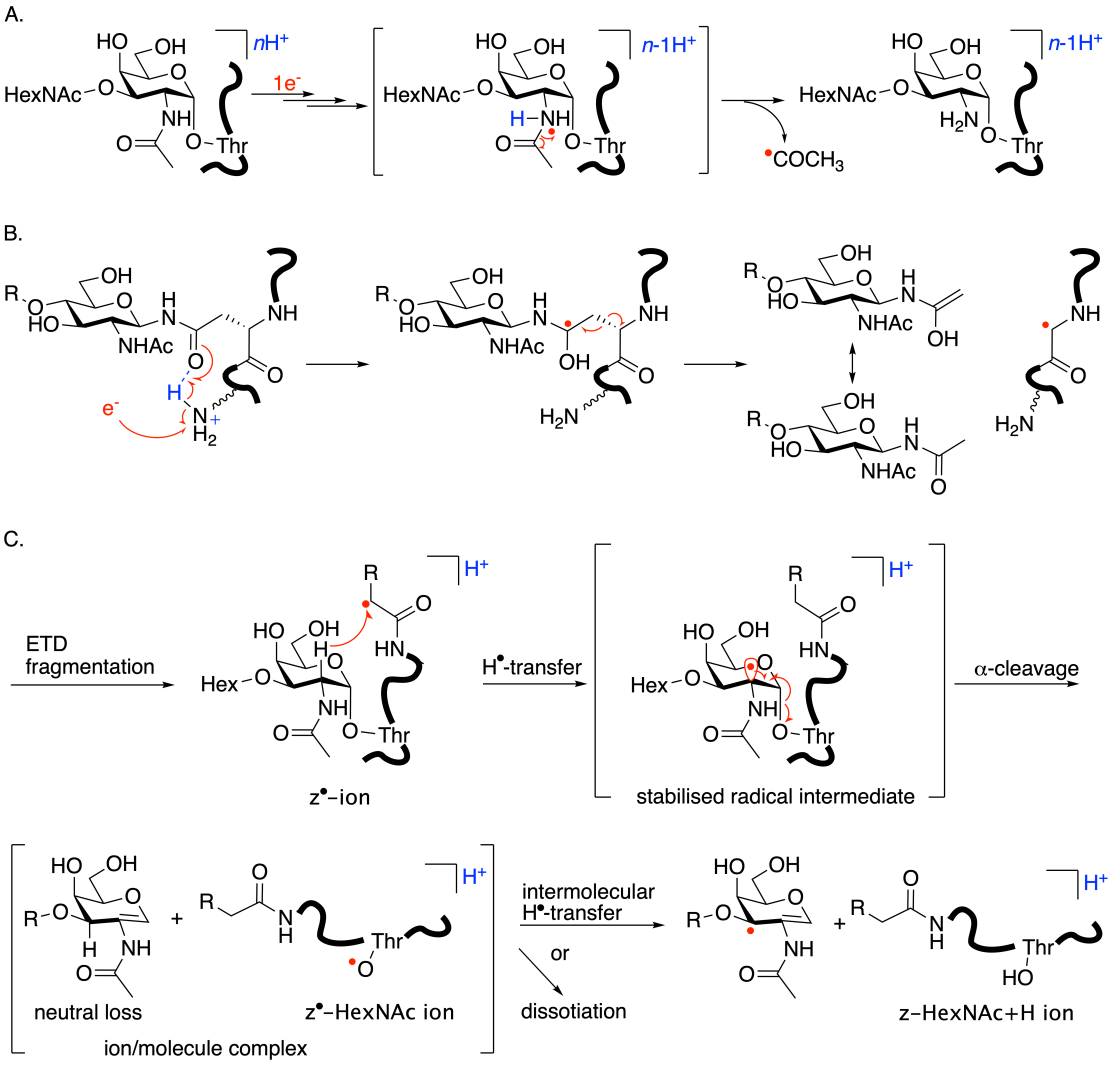


Figure 3.3. Glycopeptide-specific fragmentation pathways observed in ETD. A. Proposed mechanism for the radical site-induced loss of $^{\cdot}COCH_3$ from the charge-reduced radical cationic species $[M+2H]^{+\bullet}$, observed in *N*- and *O*-glycopeptides. Addition of a hydrogen at the nitrogen of the *N*-acetyl group leads to an *N*-centered hypervalent radical which subsequently fragments by cleaving of an acetyl radical. Note that several steps are likely to precede the depicted fragmentation, as the attack of the hydrogen or electron attachment is more likely to happen at the carbonyl oxygen. Adapted from[664]. **B.** Radical elimination of the entire Asn side chain in proximity of a basic amino acid in *N*-glycopeptides. Electron transfer to the polyprotonated glycopeptide, most likely at one of the protonation sites, leads to formation of the hypervalent glycopeptide cation ($R-NH_3^{\cdot}$). Subsequent transfer of the radical hydrogen $H^{\cdot}$ to the Asn amide carbonyl leads to a fragmentation mechanism analogue to $c'/z^{\bullet}$ formation. The resulting enol in the neutral loss can tautomerise. Adapted from[665]. **C.** Formation of $z^{\bullet}$-HexNAcHex fragment ions in *O*-glycopeptides by secondary fragmentation of odd-electron $z^{\bullet}$-type radical cation. $H^{\cdot}$ transfer from the glycan to the radical site of the α-carbon gives rise to a stabilised radical on the C^2 of the sugar ring. Antiperiplanar cleavage of the α-anomeric bond leads to the oxygen-centered $z^{\bullet}$-HexNAcHex radical cation ($= Z_0 z^{\bullet}$), which can abstract a hydrogen from the departing disaccharide to form the z'-HexNAcHex ion ($= Z_0 z'$). Adapted from[664]

Several studies on the characterisation of O-glycopeptides were performed with FT-ICR type instruments, which are less frequently found in proteomics laboratories. Although the fragmentation mechanisms are comparable between ETD and ECD, these differ in the accessible energy levels from ETD[279] (depending on instrumentation[277]). Nonetheless, findings in ECD instruments indicate which (energetically higher) fragmentation pathways are possible in ETD − this does not mean that these fragments are necessarily observed.

Glycans themselves can participate in electron-based fragmentations. Secondary fragmentation of *O*-glycopeptides from $z^{\bullet}$ in ECD[664] and ETD[667] are not uncommon (see Figure 3.3.C). Similarly, loss of sialic acid $[M+nH^+\text{-Sia}]^{n+}$ is observed with supplemental activation, but very limited when not applied.[667]

In ECD/ETD with activated ions, radical site-induced elimination of most of the glycan down to one monosaccharide (Z_1, Z_1-1H$^{\bullet}$) or of the entire glycan (Y_0-2H$^{\bullet}$, Y_0+2H$^{\bullet}$, Z_0) from the charge-reduced precursor ion were also observed.[664,668] These fragmentation mechanisms were initially characterised for *O*-glycopeptide without basic amino acid.[664,668] These proceed via hydrogen transfer processes from the glycan to the peptide and *vice versa*, described in detail by Mormann *et al.*[664] Noteworthy, the loss of $^{\bullet}O$-GalNAc (219 Da) was confirmed in ETD from charge-reduced molecular cation $[M+2H^+]^{+\bullet}$ to form Z_0 ions (as well as of $z^{\bullet}$ to $Z_0z^{\bullet}$), at a site remote from the basic amino acid.[667, Fig.3] Although these fragmentation pathways are energetically higher, they seem to be observable in ETD. This discussion might be misunderstood as of merely academic interest, yet these fragmentation pathways provide an opportunity to distinguish β1,3-linked from β1,6-linked monosaccharides (i.e. core 3 vs core 6 *O*-glycopeptides), which is otherwise only achievable by ion mobility![664, Sch.6]

Taken together, glycopeptides of $[M+nH^+]^{n+}$ type − the most commonly used analytes in glycoproteomics − rarely fragment in ETD beyond loss of sialic acid or loss of the HexNAc *N*-acetyl group. They can occasionally produce glycosidic bond cleavages in ETD framentation, depending on the peptide sequence which influences both charge density and localised fragmentation behaviour. This contrasts to doubly charged glycopeptide ions in the presence of Mg^{2+} or Ca^{2+}, obtained using other MS methods, which generates cross-ring fragmentation of *X*-type in ETD.[219] Current development using AI-ETD have illustrated its promise for *N*-glycoproteomics.[656]

3.2.3. The complementarity paradigm of glycopeptide fragmentation

The observation that the glycan part of glycopeptides fragments preferably in collision-induced methods, whereas the peptide part preferentially fragments in electron-transfer reactions has led to the formulation of the *complementarity paradigm* of glycopeptide fragmentation.[642] The orthogonal information provided through both fragmentation techniques is beneficial for glycoproteomics projects.

Although a simplification − many exceptions to this paradigm have been described in the previous sections − this conceptual approach is useful as it provides a broadly applicable framework for glycopeptide analysis. The cases where exceptions are relevant is determined mostly by the peptide sequence, in the form of charge state vs proton mobility for CID and the presence of cysteine in the case of electron-dependent mechanisms. Also, the presence of sialic acid residues can lead to additional fragmentations.

3.2.4. Hybrid fragmentation methods

As a consequence of the *complementarity paradigm*, several hybrid fragmentation methods have been applied to or developed for the analysis of glycopeptides.[669]

EThcD Electron-transfer/higher-energy collision dissociation (EThcD) fragmentation (see 1.3.4.3) provides a rich series of c/z ions pinpointing the site of glycan attachment, while also providing insight into the glycan composition via B/Y ion, including diagnostic oxonium and B-ions.[669] Although EThcD can be used to generate high-quality spectra for N-glycopeptides,[670] complementing single-NCE HCD methods,[670] its application to O-glycopeptide analysis has been a game changer[51,652] − allowing site-localisation of multiply glycosylated peptides.[652,671] In an application study, EThcD was applied to identify the presentation of O-GlcNAcylated peptides as antigens in the HLA Class I antigen presentation system for T-cell activation.[672] EThcD was chosen to enable confident modification site localisation as well as glycopeptide validation via generated oxonium ions.[672] Recently, improvement of identification was reported by extending the mass range to 4000 m/z to identify high m/z fragments (which is not required for HCD).[673] Nonetheless, EThcD suffers from the drawback of being the slowest fragmentation method, requiring the long reaction time of the ETD reaction as well as subsequent spectrum acquisition in the orbitrap. In addition, whereas glycan fragments from the non-reducing end are highly present in EThcD, core fragments (reducing terminus) are often missing.[670]

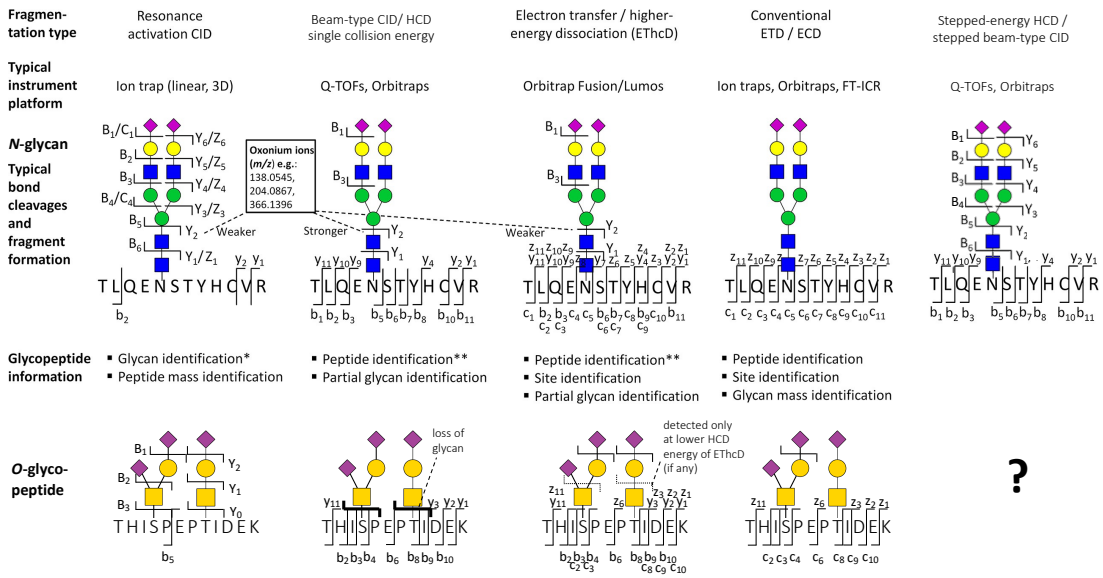


Figure 3.4. The main dissociation methods in glycoproteomics. The resulting bond cleavages and fragment ions are indicated. * *N*-glycan identification is usually restricted to monosaccharide composition and partial or complete topology. ** *b*/*y* ions in HCD and EThcD usually loose the conjugated glycan and are not indicative of the glycosylation site. Data for *N*-glycopeptides from Thaysen-Andersen[115], for *O*-glycopeptides from Darula[51] and Riley[652]. *O*-glycopeptides have been annotated for observed fragment ions on a fictive peptide sequence. Updated and modified from[115], with permission from ©ASBMB.

Product-ion triggered fragmentation strategies The use of complementary methods of fragmentation for maximal sequence coverage of both peptide and glycan part of glycopeptides can be performed either by alternating the fragmentations methods ("-alt"), or by product-ion ("-pd-") triggered approaches. Product-dependent strategies trigger an event in the mass spectrometer if a given ion is identified in the MS^2 spectrum (or another type − depending on what is desired). In the case of glycopeptides, oxonium ions provide the perfect trigger, their mass is invariable and does not overlap with any other commonly observed ions − presenting the signature of a glycopeptide. The first product-dependent strategies have used the HexNAc oxonium ions (204, 186, 168 and 138) in HCD-pd-ETD scheme, triggering the slower ETD only when required, thus allowing an increased sampling.[305,578,674] Different schemes have built on these methods, such as HCD-pd-EThcD&CID[675], used for glycosylation changes on specific proteins (insulin-mediated GLUT4 trafficking) in insulin resistance[675], or complex decision trees.[676] Product-dependent methods may not be necessary in samples that have been enriched for *N*-glycopeptides, where the majority of precursors are glycopeptides.[652] However, their use for samples with little enrichment efficiency, where many non-glycosylated peptides are present, has provided a pipeline in which the respective advantages of different

fragmentation methods can be combined.

Energy-stepping in collision-induced dissociation (sce-HCD) Different CID/HCD fragmentation energies give rise to different types of fragmentations.[275,646,677] Therefore, workflows have been developed that make use of collision-energy stepping, that is the subsequent application of different collision energies on the same group of ions in a single spectrum.[669] This enables to gather glycan and peptide backbone information. Stepped-HCD methods have been optimised[343,678] and are widely used to date for *N*-glycoproteomics,[343] as site-localisation of *N*-glycan is facilitated by the conserved N-X-S/T synthon.

Combined fragmentation strategies The emergence of sce-HCD, EThcD and other fragmentation methods such IRMPD or UVPD combined with the already established techniques for product-dependent triggering or data-dependent decision trees has enabled the development of complex hybrid acquisition sequences. The promises for biological knowledge generation are starting to fulfil.[656,672,675,679–681]

3.2.5. Energy-resolved fragmentation of glycopeptides

3.2.5.1. Fragmentation energetics differ between peptides and glycopeptides

The previous sections have introduced a mechanistic view of the fragmentation of glycopeptides in CID and ETD, emphasising the role of the implementation in the instrument for the resulting outcome, i.e. the formation of distinct fragments. Focussing on beam-type CID (Q-TOF or Thermo HCD) mechanisms, this section augments the previous description by introducing a hierarchical view of which fragmentation pathways are accessible at a given collision energy level.

Influence of charge state In beam-type CID, the acceleration of the ions for fragmentation is dictated by the voltage difference ΔU. The resulting collision energy which the ions take up upon collision, and re-distribute across the molecule, is dependent on both charge state and voltage: $\mathrm{CE}(eV) = z \cdot e \cdot \Delta U(V)$, where e is the elementary charge $1.602210^{-19}C$. Similarly as for peptides, glycopeptides with higher charge states fragment at lower collision energies (and lower voltages) than lower charge states.[275,682] This is because charge repulsion is increased by the number of charges, lowering the

activation energy required for peptide / glycopeptide fragment ions to dissociate, as fragmentation mainly proceeds via charge separation processes.[682] Neutral losses (such as of sialic acid) are not affected by charge repulsion, therefore the activation energy is similar for all charge states.[682] One key finding in the literature is that the influence of the charge state on the glycan fragmentation, albeit observed, seems to be less pronounced than for peptides.[275, Fig.1]

Linear relationship between *m/z* and CE is different for glycopeptides This finding in turn suggests that, although glycopeptide fragmentation energies are also dependent on *m/z*, their dependency is slightly different from peptides. Hinneburg *et al.* investigated this relationship, reproduced in Figure 3.5. Two main differences can be observed from this graph: *N*-glycopeptides fragment at lower collision energies (which is well-known, factor 2.5 to 1.8 depending on *m/z*), and more importantly, the slope of the trendline is different. This has concrete consequences for users of Thermo instruments using normalised collision energies (NCE - see 1.3.4.3). Normalised collision energies enable to relate peptide *m/z* to the collision energy, and are calibrated across instruments to yield highly reproducible spectra. However, they are defined using the peptide slope and peptide-specific correction factors for the charge state, which leads in the case of glycopeptides of high *m/z* to the addition of much more energy than actually desired. As will be demonstrated later in this chapter, this is particularly detrimental for energy-resolved fragmentation of *O*-glycopeptides with long peptide sequences.

The observed differences in fragmentation energetics of peptides and glycopeptides can be related to distinct structural features fragmenting first, as will be described in the following.

3.2.5.2. Activation and dissociation channels are energetically linked to structural features in *N*-glycopeptides

Aside from small neutral losses, fragmentation of glycopeptides commences with a charge separation cleavage of a glycosidic bond.[682] The stability of a glycopeptide precursor ion (survival rate) and thus the collision energy at which it fragments is dependent on peptide sequence,[683] charge state[683] and proton mobility[684] (which influences only glycan fragmentation, but not peptide fragmentation[684]).

Slight changes in *N*-glycan size or composition do not affect fragmentation behaviour, the CE needed for similar glycan fragmentations is comparable – although this does not

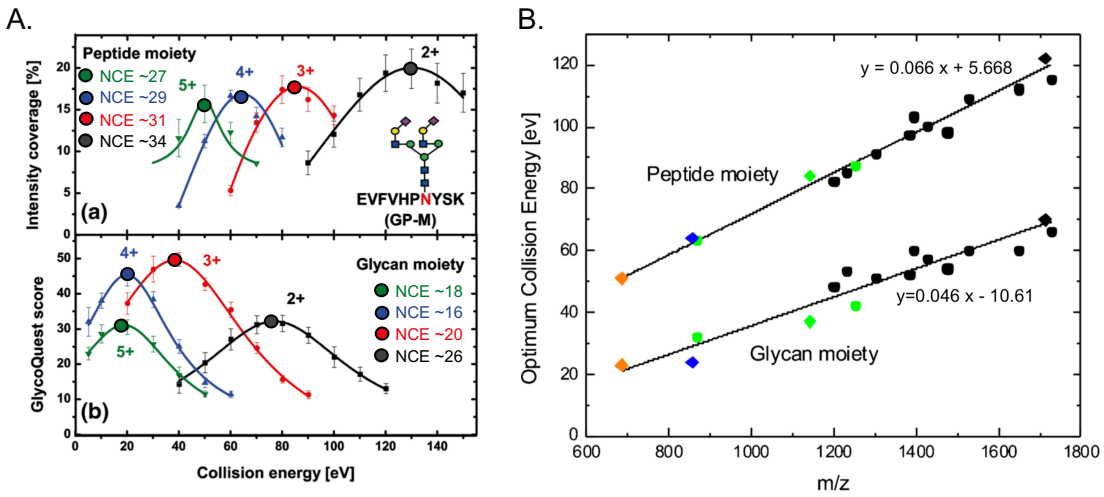


Figure 3.5. Peptides and glycopeptides differ in their fragmentation energetics. **A.** Collision energy optimum in relation to the charge state. For the peptide part, the intensity coverage is used as readout, for the glycan part the GlycoQuest score. The respective maximum were converted to approximative NCE for comparison with Thermo instruments, see 3.8.1.1 for conversion methodology. **B.** The linear relationship between precursor m/z and optimal collision energy. This linear relationship underlies the normalised collision energies on Thermo instruments.[273] Peptides and glycopeptides have different parameters. Dataset consisting of synthetic (diamond) and natural glycopeptides, coloured by charge state (orange - 5^+, blue - 4^+, green - 3^+, black - 2^+) Reproduced from[275], with permission from ©ASMS.

apply to charged NeuAc or NeuGc type of glycans, which undergo neutral loss at lower energy.[683,684] Equally, the fragmentation energy needed for glycan fragmentation is independent of its position on the peptide.[275, Fig.1d] N-glycopeptides have strikingly similar energy-resolved HCD behaviours from a qualitative perspective.[685] Generally, $Y_{n>1}$ carbohydrate cleavage is observed as the lowest-energy, first generation dissociation pathway. With increasing collision energy, production of the Y_1 glycan fragment is observed, which in turn gave rise to the Y_0 product along with b and y ions.[685] Based on the energy-resolved CID precursor and product ion abundance curves, the peptide b and y ions appear to be tertiary fragments which, accordingly, have much higher appearance energy thresholds as compared to the glycan-related scissions (Figure 3.6).[685] The ${}^{0,2}X_0$ is a cross-ring fragment diagnostic for N-glycans, its formation is at similar or higher energy levels as b and y ions.[686] These observations are in line with the expectation that amide bonds are more resistant in proton-mediated fragmentations than acetals such as the glycosidic bond − trends also observed for acidic hydrolysis of the functional groups in the condensed phase. Cross-ring fragmentations are equally expected to be energetically higher.

A particularity of the gas-phase mechanism of collision induced fragmentation are mobile protons.[255–257] The vibrational energy redistribution leads to activation of a proton, which can be separated from the charge-carrying base and wander along the peptide ultimately leading to fragmentation (see Figure 1.11). Therefore, three cases are usually considered which differ in the number of carge-carrying protons (n_H) and number of bases (n_B): $n_H > n_B$; $n_H = n_B$ and $n_H < n_B$. These cases can be used to explain quantitative differences in energy needed for fragmentation of glycopeptides − note that the qualitative hierarchy remains the same.[684] Precursor ions characterised by lower proton mobility require greater vibrational energy deposition per vibrational mode to enable unimolecular dissociation processes.[684] This means that less vibrational energy is required for peptide with high proton mobility ($n_H > n_B$) and increasingly more for $n_H = n_B$ and $n_H < n_B$.[684]

Interestingly, the position of the glycan on the peptide affects the collision energy needed for optimal peptide fragmentation[275, Fig.1c], which is probably due to glycan interfering sterically in proton mobility as peptides with the glycan remote from the base fragmenting best at lower CE. Except from *N*-glycan position, peptide fragmentation is generally speaking independent of proton mobility.[684]

These mechanistic observations have direct consequences for large-scale *N*-glycoproteomics: the collision energy does not need major adjustments for different types of *N*-glycans. The combination of low energy beam-type CID for glycan fragmentation and higher energy for peptide fragmentation has been shown by multiple studies to be beneficial.[275,677,682,685] Low collision energy further allows the identification of diagnostic *B*-ions.[648] With the introduction of NCEs on Thermo instruments,[273,687] collision energy can be normalised for *m/z*. Although this is standardised (for peptides), there is a slight dependency for glycans on the charge state (particularly at NCE < 20%[678] − limited energy emphasises role of charge-repulsion) and the peptide sequence will still affect proton mobility. This has nonetheless enabled widespread use of stepped-collision energy methods for *N*-glycopeptides.

3.2.5.3. Stepped collision energy methods for fragmentation of *N*-glycopeptides

The idea of stepped collision energy methods is to divide the total ion activation time into parts with distinct energy levels. For example a 30 msec fragmentation can be divided in 3 × 10 msec steps of 20%, 30% and 40% normalised collision energy (NCE) each − a method termed SCE-HCD-20-30-40.[687]

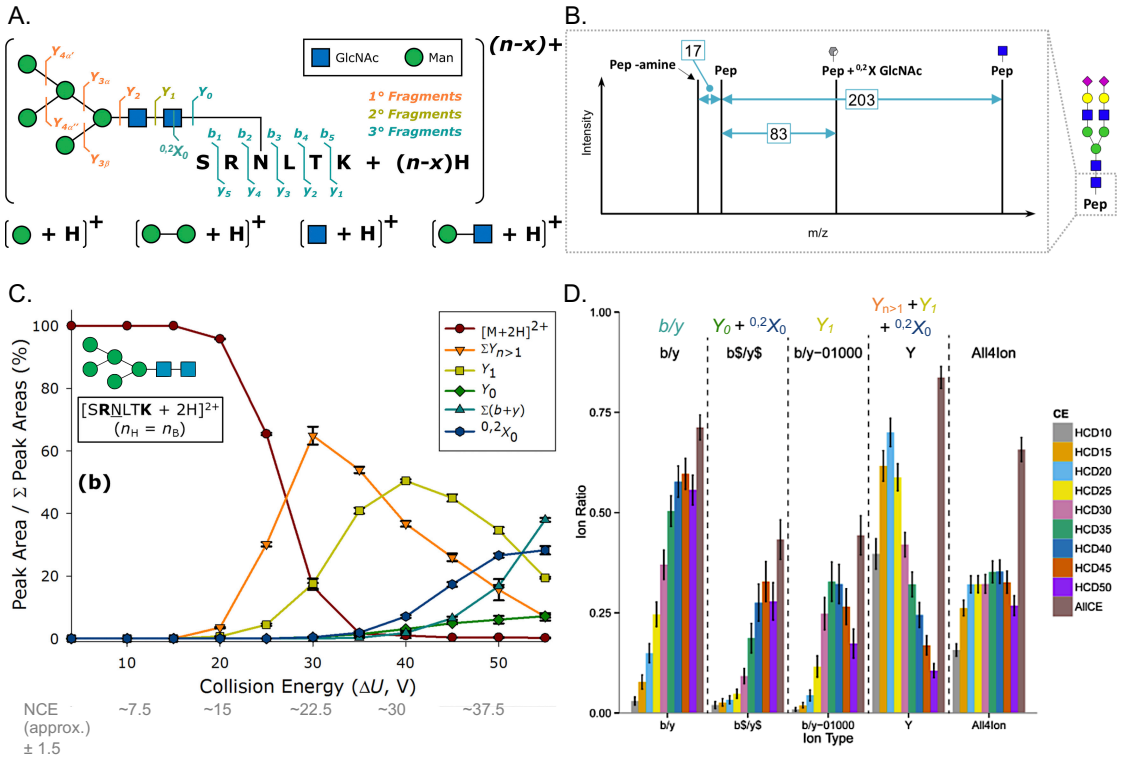


Figure 3.6. Energy-resolved fragmentation of *N*-glycopeptides. A. The three main energy levels in *N*-glycopeptide fragmentation, highlighted by the colour of fragment annotation. Reproduced from[684] with permission from ©Elsevier. **B.** Fragments obtained at the glycan attachment site, yielding characteristic fragments in SCE-HCD. Modified from[686] with permission from Wiley-VCH. **C.** Energy-resolved formation of different glycopeptide fragments. This diagram holds true qualitatively for all *N*-glycopeptides, but the collision energy values change depending on proton mobility. Collision energy values have been converted to Thermo NCE's as described in 3.8.1.1. Reproduced from[684] with permission from ©Elsevier. **D.** NCE-resolved fragmentation of *N*-glycopeptides, relevant for the implementation of SCE-HCD methods on Thermo instruments. Reproduced from[343, Note1-2a] under Creative Commons CC BY.

Stepped collision energy methods have been used in combination with ion-trap CID or beam-type HCD for small molecules,[687] TMT-labelled peptides[688] and glycopeptides.[275] In particular, the fragmentation behaviour of *N*-glycopeptides has been extensively studied and stepped-energy methods underlie many modern *N*-glycoproteomics applications (in combination with ZIC-HILIC enrichment).[343,440–442]

In a landmark study,[343, SI] the pGlyco team studied the stepped energy settings for a wide range of *N*-glycopeptides on Thermo instruments, relating the different energetical fragmentation levels described previously to NCE settings.[343, Note1] As such, maximal formation of $Y_{n>1}$ ions was observed at a NCE of 15-25 %, peaking at 20 % (Figure 3.6.D).

Maximal b/y ion formation was observed in the 35% - 50% range, peaking at 40 % (Figure 3.6.D). The authors performed extensive work assessing different NCE combinations and found the 3-step SCE-HCD with NCEs of 20-30-40 to be one of the most effective ways of glycopeptide sequencing (among 30±15 and 35±15). Follow-up studies of other groups have confirmed this approach (and setting[652]) for *N*-glycoproteomics,[678] one suggesting slightly different settings of 20-30-30.

In a sense, low-energy HCD (beam-type CID) used in sceHCD combines the advantages of low-energy ion-trap CID used in the older literature while bypassing the limitations (1/3rd rule, mass accuracy) at the cost of acquisition speed. The latter can be remediated by adjusting gradient time and loading. Indeed, several studies[668] use pure MS^1 acquisition for quantitative LFQ glycoproteomics.

3.2.5.4. Difference in optimal dissociation methods for *N*- and *O*-glycopeptides are related to their distinct structures

The guiding principle of this introduction is to present the mechanisms underlying glycopeptide fragmentation in a energy-dependent perspective, i.e. which fragmentation pathway is accessible at which energy. This knowledge can then be used to tailor appropriate acquisition methods. My ambition was to place the data acquired in this DPhil on *O*-glycopeptides in this framework, as my findings will show that the energy dependency is different for *N*- and *O*-linked glycopeptides. However, in the course of finishing this DPhil and writing the dissertation, two key publications appeared that come to similar conclusions[689] or provide data which can be interpreted with this understanding.[652] This data corroborates my findings and therefore is presented in this introduction together with other reports in the literature, which retrospectively point to the same conclusions.

O-glycopeptides show qualitatively similar trends to *N*-glycopeptides in the order of their fragmentation. The loss of sialic acid can be observed at very low energy or without applying collision energy. At low energies, $Y_{n>1}$ ions are formed, followed by Y_0 ions, corresponding to glycan loss. At higher energy, fragments from both peptide and glycosidic bonds ($Y_0 + a, b, y$) become the abundant ions (Figure 3.7.A).[689] Importantly, it is possible to generate b/y ions carrying the intact glycan at low energy, taking advantage that *O*-glycosylation is often located in Pro-rich stretches: the Xxx-Pro bond is weak (*Proline effect*) and can therefore be fragmented at lower CID energy values (yellow squares in Figure 3.7.A), facilitating site-localisation of *O*-glycopeptides in CID.[690]

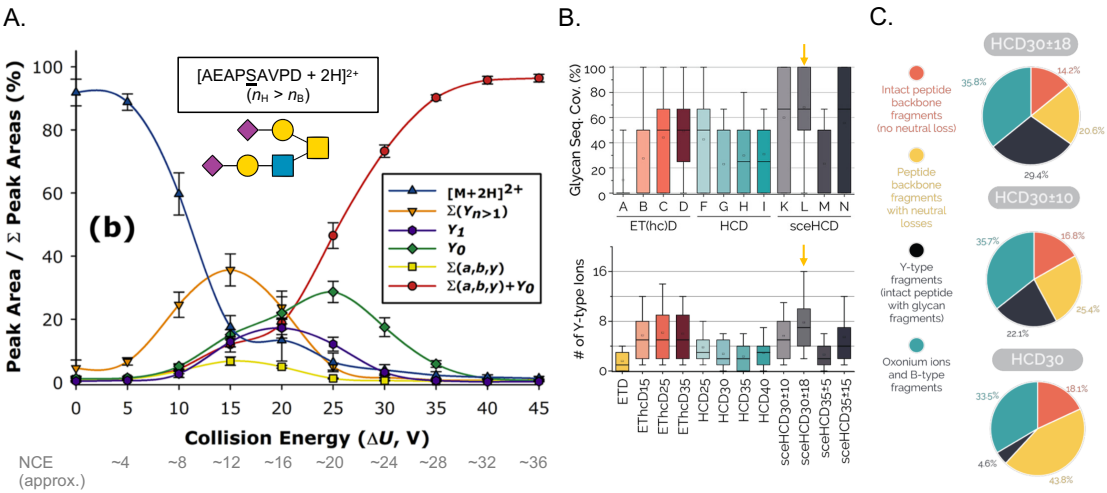


Figure 3.7. Energy-resolved fragmentation of *O*-glycopeptides. All data presented in this figure was published during the time of writing this DPhil thesis. **A.** Energy-resolved formation of different *O*-glycopeptide fragments. Note that yellow squares indicate peptide fragments with intact glycan attached. Collision energy values have been converted to Thermo NCE's as described in 3.8.1.1. Reproduced from[689] with permission from ©ACS. **B.** *O*-glycan sequence coverage and number of identified *Y* ions benefit from SCE-HCD starting at low NCE, here 12-30-48. Reproduced from[652] under ACS AuthorsChoice license. ©ACS. **C.** The SCE-HCD method 12-30-48 has the highest proportion of *Y* fragments, important for *O*-glycan sequencing. Reproduced from[652] under ACS AuthorsChoice license. ©ACS.

However, there are key differences between *O*- and *N*-glycopeptides. First, the collision energy at which 50% precursor ion survival is detected is substantially lower for *O*-glycopeptides. Second, settings of 50 % ion survival for *O*-glycopeptides for the same charge state are very close to each other, despite high differences in mass.[689] *N*-glycopeptides have an amide bond as attachment between glycan and peptide, more stable towards CID than the glycosidic linkage. Therefore, Y_0 and Y_1 formation emerges at similar energies, not far apart from other glycosidic linkages (except sialic acid). As a consequence, the energy settings for *O*-glycopeptides are more delicate to set, small difference in collison energy can lead to radically different structural features being observed.[689]

Pioneering examples have shown that 2-step collision energy methods (Q-TOF instruments) can be beneficial for *O*-glycopeptides, using a low energy for glycan fragmentation and a higher energy for peptide backbone fragmentation.[275,668] The study of Riley *et al.*, published at the time of writing of this chapter, suggests that SCE-HCD methods are sufficient for most *N*-glycoproteomic applications but are ill-suited for *O*-glycopeptide characterisation.[652] While the reported data in the NCE range of 20-40 clearly supports this claim, this chapter aims at making a case that SCE-HCD methods at lower NCE

(10-15) can indeed contribute to elucidate structural features in O-glycopeptides that would be missed by other techniques. Indeed, this hypothesis is supported by data in the Riley *et al.* study[652] for SCE-HCD 12-30-48, showing that SCE-HCD including a first step at very low energy has the highest *O*-glycan sequence coverage, highest number and proportion of *Y*-ions (Figure 3.7.B,C).

While stepped HCD methods will not be able to solve the prominent issue of site-localisation of *O*-glycopeptides − unlike for *N*-glycopeptides there is no conserved synthon, they might be helpful identifying connectivities. ETD-based methods (including EThcD) remain indispensable for the site-specific localisation of glycans on *O*-glycopeptides,[652] their use in conjunction with stepped HCD methods for *O*-glycopeptides could however be a promising avenue.

3.2.6. Data analysis and interpretation of glycopeptide MS/MS spectra

Glycopeptide spectra are particularly rich in information, including glycan fragmentation (along the glycosidic bond & cross-ring), peptide backbone fragmentation, fragments with either one or both (+ potential modifications such as methionine oxidation), B-ions and oxonium ions. Isotope effects lead to a signal decrease and *N*-glycopeptides have similar retention times on classical C18 materials[172] (but not *O*-glycopeptides[51]). Micro- and macro-heterogeneity[639] of the glycans add to the technical challenges. As such, probably one of the most difficult part in glycoproteomics (once glycopeptide enrichment and fragmentation have been optimised) is the data analysis. A detailed introduction is given in 1.3.5, therefore only key points relevant to this chapter are highlighted here.

Available tools and methodologies Neither established glycomics software (SimGlycan) nor proteomics software (MaxQuant, PEAKS etc.) are able to cope with glycoproteomics data. MaxQuant and PEAKS allow for entering glycosylations as a modification on peptides, but this does neither account for their fragmentations nor for all other specific features of glycopeptides. Although many academic groups have attempted to write their own specialist code, two broadly applicable softwares have emerged for glycoproteomics and glycopeptide identification: Byonic[342] and pGlyco2[343]. On glycopeptide level Byonic does mass matching, pGlyco2 database search. pGlyco2 is very promising from a methodology and performance perspective and has made great progress on its *N*-glycopeptide FDR, Byonic allows for more customisation and flexibility. Recently, the MSFragger-Glyco software has been released, its value is yet to be assessed.

In the authors' opinion, Byonic still is the superior software for *O*-glycopeptides. Using its modification fine control markUp language, Byonic remains more customisable and is therefore able to deal with *C*- and *S*-glycosylations. Whereas Byonic does a great job in identifying glycosylation sites, it fails at assigning glycan fragmentation in the MS2 spectra if the database has custom defined glycans not included in the list of Byonic keywords[†].

Similarly, in most cases Byonic performs well in picking up spectral features, however its performance at scoring is limited. Byonic's choices in the weighing of spectral features have been highlighted,[652] as the software weighs particularly high intensity matching fragments favourably (even $\sim b/ \sim y$ ions with glycan loss) without distinguishing their type and informative content.[342] A case has been made for altering search algorithms to more prominently score peptide backbone fragments that retain an intact glycan or the $Y_n \geq 1$ fragments. Similarly, *B*-type ions can be very informative of structural components inside the glycan itself.[648]

Combinatorial search explosion With peptide fragmentations, glycan fragmentations, modifications and cross-ring fragmentations, the number of combinatorial possibilities is very high. This is a problem particularly for *O*-glycosylation.[51] Whereas *N*-glycosylation is easy to locate thanks to the conserved synthon, *O*-glycans appear often in 'patches' and the assignments of fragmentations is difficult. The strength of Byonic's modification fine control comes to hand with the problem of *O*-glycopeptide combinatorial search explosion, allowing for protein specific definitions that allows to look for glycosylations in a complex database, thus limiting false PSM and keeping a user control on the FDR.

[†]Personal communication by Marshall Bern, Protein Metrics Inc. See appendix 3.8.2.1 for keywords

3.3. Objective

This chapter aims at obtaining further insights into the energy landscape of the competing *O*-glycopeptide fragmentation pathways, in particular for the following points:

- The energy-dependency of all fragment ion types across a range of *O*-glycopeptides
- How this energetic resolution can best be exploited in stepped-CE (SCE- HCD) methods
- Which normalised collision settings can be broadly applied across a range of *O*-glycopeptides
- If site localisation can be obtained using SCE-HCD methods or if such an approach still requires supporting ETD data.

3.4. Materials & methods

3.4.1. Sample preparation

Preparation of chondroitinase-trimmed decorin Trimmed decorin was kindly prepared and donated by Alexandra Gafitescu. Briefly, beef tendons were trimmed for fat and muscle, 65 g of cartilage were chopped finely and total protein extracted O/N (1.1 L 50 mM NaOAc pH 6.0, 4 M Gd.HCl, 5 mM benzamidine, 10 mM EDTA, 5 mM *N*-methylmaleimide, stirring at 4 °C). The insoluble residue was filtered over gauze and extracted once more (0.5 L). The filtrate was centrifuged (30 min, 8000 rpm, 4 °C) and filtered (paper, 4 °C), clarified (20 min, 30,000 rpm, 4 °C) and vacuum filtrated (0.45 µm, cellulose acetate). The filtrate was concentrated (10000 MWCO, Amicon PM-10), dialysed (50 mM NaOAc pH 6.0, 7 M urea, 100 mM NaCl) and purified by WAX (HiTrap DEAE, 5 × 5 mL connected columns, GE). Collected fractions were analysed by SDS-PAGE and proteoglycan fractions identified by Alcian Blue staining. Concentration of intact decorin was measured by BCA assay. The GAG chains were removed with chondroitin ABC lyase (= chondroitinase) from *Proteus vulgaris* (Sigma, 1.75 u/18 mg in 100 mM Tris, 50 mM NaOAc pH 8.0, 37 °C, 16h, completion by SDS-PAGE). The reaction mixture was concentrated (Vivaspin 30,000 MWCO, 4 °C), dialysed (50 mM sodium phosphate, 500 mM NaCl pH 7.0) and purified by lectin chromatography (HiTrap ConA 4B 5 mL column, GE) eluting with 300 mM methyl-α-D-mannosyranoside. Fractions containing desired trimmed decorin were dialysed against MiliQ water and lyophilised.

LysC + trypsin digest 40 µg of human IgG (P01859, Sigma Aldrich) 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 min at r.t. before being alkylated with 50 mM 2-iodoacetamide for 30 min at r.t. in the dark. Samples were diluted to 6 M urea with 50 mM TEAB and digestion was performed with LysC (Promega, 1 µg/40 µg sample) overnight at 37 °C. Samples were further diluted to 1-2 M urea with 50 mM TEAB. IgG samples were further digested with trypsin (Promega, 1 µg/40 µg sample), for 4 h at r.t. For full glycosylation analysis, samples were dried in a speed-vac overnight and resuspended in 5% DMSO 5% formic acid. Samples were not desalted to avoid loss of the short IgG glycopeptide.

LysC digest 40 µg of bovine RNAse B (P61823, Sigma Aldrich) 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 min at r.t. before being alkylated with 50 mM 2-iodoacetamide for 30 min at r.t. in the dark. Samples were diluted to 6 M urea with 50 mM TEAB and digestion was performed with LysC (Promega, 1 µg/40 µg sample) overnight at 37 °C. Samples were further diluted to 1-2 M urea with 50 mM TEAB. RNAse B samples were further digested with LysC (Promega, 1 µg/40 µg sample), for 4 h at r.t. For full glycosylation analysis, samples were dried in a speed-vac overnight and resuspended in 5% DMSO 5% formic acid. Samples were not desalted to avoid loss of the short RNase B glycopeptide.

Tryptic digest 10 µg each of PNGase F-deglycosylated, chondroitinase-trimmed bovine decorin (P21793, provided by Alexandra Gafitescu, see above), intact bovine fetuin (P12763, Sigma Aldrich F3004), intact bovine Asialofetuin (P12763, Sigma Aldrich A4781) were processed, digested and run individually. Samples 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 min at r.t. before being alkylated with 50 mM 2-iodoacetamide for 30 min at r.t. in the dark. Samples were diluted to 1-2 M urea with 50 mM TEAB and digestion was performed with trypsin (Promega, 1 µg/20 µg sample) overnight at 37 °C. Samples were subsequently desalted by C18 (Oasis HLB), dried in a speed-vac overnight and resuspended in 5% DMSO 5% formic acid.

3.4.2. Tandem mass spectrometry methods

3.4.2.1. Preparation of nLC columns for electrospray ionisation

Needle pulling A 75 µm i.d. fused silica capillaries roll (TSP075375, Polimicro Technologies) was cut in capillaries of appropriate lengths (2 × desired length of column + ∼2 cm) using a ceramic tile cutter (Chromres). Using a micro torch, a 2 cm part of the plastic sheath in the middle of the cappilary was burned and wiped clean with a delicate wipe soaked in MeOH. This exposes the quartz glass. The fused silica is loaded into a warmed up (10-15 min) laser puller (Sutter Instruments Co, model P-2000) with the exposed quartz glass immobilised in the laser accessible region. The laser puller was used to pull the needles under precise heating (Screw setting: 8-5 ; Program: HEAT = 250, FIL = 0, VEL = 15, DEL = 128, PULL = 0). The pulled needles were inspected under the microscope for cone shape.

Column packing For the opening of the tip, a 2 mL flat-bottom glass vial containing MiliQ water was placed in the packing bomb. The capillary with the needle was mounted with its open end dipped into the vial, leaving 1-2 mm space above the glass vial bottom, and fixed with the top nut. The pressure bomb system was pressurised to 10 bar. The needle tip was delicately heated with the micro torch, ensuring the resulting water jet is continuous and straight. Once a reliable steady stream is reached, the pressure is released and the pressure bomb dismantled. 50-100 mg of sorbent (Reprosil-Pur 120 C18-AQ, 1.9 µm, Dr. Maisch GmbH) is suspended in 1 mL of a 2:1 pentane/acetone mixture in a 2 mL flat-bottom glass vial containing a small magnet stirrer bar, stirring at 400-500 rpm. The pressure bomb is then assembled and the capillary with open needle tip inserted by its open capillary end, leaving 3-4 mm space above the vial bottom. Immediately upon assembly of the pressure bomb, the pressure was applied (10 bar). The ejection of liquid through the tip was verified and the progress monitored by shining the through the capillary with a desktop lamp. Once the entire column is packed, the pressure connection is closed and the packet left for slow depressurising O/N.

Similar packing method have been published.[691,692]

3.4.2.2. Data acquisition for glycosylated samples

Q Exactive

Resulting peptides 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 x 5 mm, 3 µm C18 beads; Thermo Fisher Scientific) and separated on an in-house packed analytical column (75 µm inner diameter x 50cm packed with ReproSil-Pur 120 C18-AQ, 1.9 µm, 120 Å, Dr. Maisch GmbH) using a linear gradient (length: 28 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 positive mode, automatically switching from full 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 3×10^6, maximum injection time 50 ms]. After the MS scan, the top 10 most intense peaks were selected for HCD fragmentation at the indicated normalised collision energy (standard:

30 %) or stepped collision energy. HCD spectra were also acquired in the Orbitrap (resolution 17500; AGC target 5×10^4; maximum injection time 120 ms), with first fixed mass at 180 m/z.

Orbitrap Elite

General LC method Resulting peptides were analysed on an EASY-nLC 1000 UHPLC system (Proxeon) connected to a Orbitrap Elite mass spectrometer (Thermo Fischer Scientific) possessing an EASY-Spray nanoelectrospray ion source (Thermo Fischer Scientific). The peptides were trapped on a 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 a C18 analytical column (75 µm i.d. × 750 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: 250 nL/min). For supercharging experiments, 0.1 % 3-nitrobenzyl alcohol was added to the mobile phase. The separated peptides were electrosprayed directly into the mass spectrometer operating in a data-dependent mode.

HCD experiments Full scan MS spectra (scan range 350-1500 m/z, resolution 120000, AGC target 1e6, maximum injection time 250 ms) and subsequent HCD MS/MS spectra (resolution 15000, AGC target 3e4, maximum injection time 500 ms) of 5 most intense peaks were acquired in the Orbitrap. HCD fragmentation was performed at 30 % of normalised collision energy and the signal intensity threshold was kept at 500 counts.

ETD-DDDT experiments Full scan MS spectra (scan range 350-1500 m/z, resolution 120000, AGC target 1e6, maximum injection time 250 ms) were acquired in the Orbitrap. CID and ETD spectra were acquired in Ion Trap (resolution 7500, AGC cation target 3e4, AGC Anion target 2e5, maximum injection time 100 ms). After the MS scans, the 5 most intense peaks were selected for fragmentation based on data-dependent decision tree (DDDT). The signal threshold for fragmentation of parent ion was set to 500 ion counts. For CID fragmentation, the normalised collision energy and default charge state was set to 35% and 2 respectively. For ETD fragmentation, reaction time was set to 50 ms and supplemental activation was enabled. Charge state screening was enabled; parent ion with unassigned charge state and charge state 1 was rejected. Dynamic exclusion was enabled (exclusion size list 75, exclusion duration 5 s). The setting for DDDT

is as such: ETD fragmentation was followed for charge states 3, 4 and 5 with m/z less than 750. For charge states greater than 5, ETD fragmentation was always performed. In all other cases, CID activation was performed.[272]

HCD-alt-ETD Full scan MS spectra (scan range 350-1500 m/z, resolution 120000, AGC target 1e6, maximum injection time 250 ms) were acquired in the Orbitrap. ETD spectra were acquired in Ion Trap (resolution 7500, AGC cation target 3e4, AGC Anion target 2e5, maximum injection time 100 ms). HCD spectra were acquired in the Orbitrap (resolution 15000, AGC target 3*e*4, maximum injection time 500 ms) and the described NCE used (typically NCE of 35). After the MS scans, the 3 most intense peaks were selected for fragmentation. Each peak was sequentially selected first for HCD and then for ETD fragmentation. For ETD fragmentation, reaction time was set to 50 ms and supplemental activation was enabled. Charge state screening was enabled; parent ion with unassigned charge state and charge state 1 was rejected. Dynamic exclusion was enabled (exclusion size list 75, exclusion duration 15 s).

HCD-pd-ETD experiments Full scan MS spectra (scan range 350-1500 m/z, resolution 30000, AGC target 1e6, maximum injection time 250 ms) and subsequent HCD MS/MS spectra (resolution 15000, AGC target 3e4, maximum injection time 500 ms, first mass m/z 120.00 for decorin [Xyl trigger]) 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. **HexNAc oxonium triggers** (*N*- and *O*-GlcNAc / *O*-GalNAc glycans). If peaks at any of m/z 138.0546, 168.0563 or 204.0867 (50 ppm error) were detected among top 20 most intense peak of a HCD spectrum, an ETD MS/MS spectrum of the precursor was triggered.[305,578] **Xyl oxonium triggers** (*O*-Xyl / proteoglycan core). If peaks at any of m/z 133.0500 [Xyl-H_2O]$^+$, 163.0601 [Hex-H_2O]$^+$ or 273.0444 [dHexA-H_2O]$^+$ (50 ppm error) 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 5e4, AGC anion target 2e5, cation maximum injection time 100 ms) were acquired in the Iontrap.

EThcD experiments The conceptual implementation of EThcD experiments on an Orbitrap Elite is described in Figure 1.13 and section 1.3.4.4. In this setup, EThcD is hard-coded as a HCD scan triggered if first mass is set as 151. Only with this setting is a MS^3 step in the HCD cell without precursor isolation defined. Instead, the

instrument will do an ETD event using the precursor mass of the HCD scan. The necessary ftmx2Qitmanipulate.ICB and toggle.ICB files were provided by Dr. Martin Zeller, Thermo Fisher Scientific. Full scan MS spectra (scan range 350-1500 m/z, resolution 120000, AGC target 1e6, maximum injection time 250 ms) were acquired in the Orbitrap. After the MS scans, the 5 most intense peaks were selected for fragmentation based on data-dependent decision tree (DDDT). The signal threshold for fragmentation of parent ion was set to 500 ion counts. EThcD fragment ions were generated in the IonTrap (GC cation target 3e4, AGC Anion target 2e5, maximum injection time 100 ms), the maximum injection time was also the reaction time, and in the HCD collision cell (AGC target 3e4, maximum injection time 500 ms, signal intensity threshold at 500 counts.) with collision energies defined for each experiment (typically NCE 35). The dynamic exclusion was set to 10.0 s to account for the slower acquisition speed.

HCD-alt-EThcD experiments Full scan MS spectra (scan range 350-1500 m/z, resolution 120000, AGC target 1e6, maximum injection time 250 ms) and subsequent HCD MS/MS spectra (resolution 15000, AGC target 3e4, maximum injection time 500 ms) were acquired in the Orbitrap. EThcD spectra were acquired as described previously. After the MS scan, the top 3 most intense peaks in the full MS spectrum were each selected for an HCD scan (fragmentation in the HCD collision cell, AGC target 3e4, maximum injection time 500 ms, signal intensity threshold at 500 counts) immediately followed by an EThcD scan (see above).

Note that HCD-pd-EThcD methods are not possible with this implementation of EThcD on the Orbitrap Elite, as the instrument doe not allow for HCD to be parent for a CHD MS3 scan event, which would be the case as EThcD is hard-coded as an HCD with 151 first mass.

3.4.3. Data analysis

3.4.3.1. Search parameters

The contaminants in each sample were first identified by Mascot or a PEAKS X search against the glycoprotein and the MaxQuant contaminant database. Each glycoprotein was then searched with Byonic version 2.9.30 against a focussed protein database containing the glycoprotein sequence and the identified contaminants, except for decorin

which was searched with version 3.6.0. [‡] Using Byonic's ProteinLabel keyword, glycosylation was only allowed on the glycoprotein of interest.

For all data acquired in the Orbitrap (Q Exactive and Orbitrap Elite), a 20 ppm mass error was set. For spectra acquired in the ion trap (Orbitrap Elite), a mass error of 0.6 Da was set. Carbamidomethylation was set as a fixed modification while oxidation (Met), deamidation (Asn, Gln), carbamidomethylation (peptide *N*-terminus, Lys) and phosphorylation (Ser, Thr) were set as variable modifications. On decorin, proline oxidation was included. If not specified otherwise, 1 missed cleavage was allowed and the data was searched with trypsin as enzyme (R, K) and semi-specific cleavage. The Byonic database of mammalian *N*-glycans was used. *O*-glycopeptides were initially identified using a compiled *O*-glycan database in searches with limited parameters (see 3.4.3.2) and a focussed database created, which was then used in the main search. For decorin, a *O*-glycan database was manually curated based on the findings of Klein *et al.*[693], reported in appendix 3.8.2.2.

3.4.3.2. Identification of novel *O*-glycopeptides and database curation

A comprehensive database of *O*-glycopeptides was generated in two steps. First, for each glycoprotein, known *O*-glycans were collected from the literature (i.e. for bovine fetuin[275,670,678], bovine decorin[693]). Additionally, the GlyGen database (glygen.org, accessed 29th of July 2020) was mined for mammalian O-glycans (including human, bovine) and an annotated database of known O-glycans created. Note that the GlyGen database contains almost exclusively mucin-type *O*-glycans. The dataset was searched against this extensive *O*-glycan database as common1, allowing only 1 *O*-glycan per peptide to limit search space. A focussed database of plausible *O*-glycan was then created, searched again with the previously optimised search parameters and the identified *O*-glycans manually validated. *O*-glycans that did not pass 2nd validation were dismissed from the focussed database.[694]

3.4.3.3. Rationale and flaws for the selected Byonic scoring metrics

As stated before, my aim is to evaluate fragmentation methods for *O*-glycoproteomics. Different scoring metrics were chosen to assess performance of each fragmentation method

[‡]Marshall Bern implemented support for dHexA from this version of Byonic on, which was tested on decorin data presented in this thesis as well as data from the Joseph Zaia lab at BU. The decorin data was processed with a trail version of Byonic v 3.6.

assessed in this study. We used Protein Log Score, Intensity, sequence coverage and total PSM as useful metrics to assess the performance of each method on regular, non-glycosylated peptide identifications. For the different types of glycosylated peptides the Byonic Score, Byonic Delta Modifications Score, Posterior Error Probability Score and manual analysis observations were used assess the performance and quality of PSMs.

The choice for these metrics and scores, which are all PSM related, builds on the fundamental assumption of this study that only well fragmented and correctly annotated spectra contain the necessary information for correct assignment of glycopeptides in mass matching searches − such as Byonic does. In this context, a well fragmented spectrum shows *presence* of a maximum of fragment ions from peptide / glycan with adequate intensity, enabling high sequence coverage on both peptide / glycan. Thus, the quality of fragmentation is controlled via the instrument acquisition method. Similarly, a correct annotation implies the *identification* by the software of peptide fragment ions with and without attached glycan, Y-type glycopeptide fragment ions and B-type glycan and oxonium ions. The quality of the annotation is given by the software version (ie. Byonic v 3.6 for proteoglycan) and can only be validated manually.

While this study attempts to assess the suitability for different fragmentation techniques and energies for *O*-glycoproteomics, a fundamental flaw in this approach needs to be kept in mind: the correctness of the assignment by the software influences the PSM scoring. Although Byonic is arguably the most advanced software for *O*-glycoproteomics, some fragment ions such as those with fragmentation on both peptide and glycan (i.e. b_nY_n and y_nY_n-type fragments observed for example on decorin) are not identified by the software. These fragments are nonetheless very relevant to assess glycan site localisation. To conclude, even though all datasets for a given sample are processed with the same software / parameters, the quality of annotation and identification will affect the scoring metrics to the disadvantage of methods that provide extensive *O*-glycopeptide fragmentation.

Taken together, it is therefore difficult to assess the presence of false positives. Assignments based on mass difference between MS1 and backbone sequencing ID are very prone to error – the caveat of regular HCD data acquisition on QEx for glycopeptides that yield many identifications but do not give a 'secure' identification. The comparison of metrics is required to conclude if the quality of the assignment is good and the assessment of a fragmentation technique cannot be done alone on the number of PSMs or any other metric alone. Although Byonic metrics are inherently geared towards peptides, the combination of the following metrics has been chosen to describe the quality of glycopeptide assignment:

Number of glycopeptide PSM Number of peptide-spectrum matches for glycopeptides, i.e. the number of identifications. Whereas generally speaking, more identifications are desired, this does not necessary mean that all these identifications are correct.

Posterior Error Probability (PEP) A PEP of 1.0 E^{-1} means 10% chance that the peptide is assigned wrongly. As such, identifications with a PEP in the range of E^{-1} are generally to be dismissed, E^{-2} to E^{-3} as highly sceptical (i.e. 1% to 0.1% chance that the PSM is wrong, that is a non-negligible error probability), then the higher the better. The PEP range allows to assess the general reliability of PSM with a certain fragmentation technique, it is dependent on the type of glycosylation and thus directly on the chemistry of the fragmentation process. Byonic's PEP includes 10 features, counting the Byonic Score, Delta, Precursor Mass Error, digestion specificity etc. and arguably provides the best metric to assess performance.[695] Frequently, its logarithm ($\log_{10}$) is used, the |LogProb| score.

Peptide Score Byonic score for the PSM, a 'raw' indicator of correctness. Byonic scores reflect the absolute quality of the PSM, not the relative quality compared to other candidate peptides. Ranging from 0 – 1000, a score > 300 is considered as ‚good', > 400 as ‚very good', >500 as ‚almost sure to be correct'.[695] Thresholds of >200 are often used in the literature by groups reporting their methods for FDR.[676] Generally, this score represents matching of peaks, presence of fragments etc. It works usually well unless the spectra have poor signal-to-noise ratio.

Delta Mod Score The drop in Byonic score from the top-scoring peptide to the next peptide different in any way, including placement of modifications. For glycopeptides, the DeltaMod score gives typically an indication of whether modifications are confidently localised, i.e. an indication on the reliability of site localisation/specificity at the named site of modification: <20 is uncertain, >40 is good, >100 very good.

3.5. Results

Our investigation of energy-resolved fragmentation of glycopeptides originally resulted from efforts to setup a MS/MS method to monitor artificial glycosylation of the decorin proteoglycan. Surprised by the findings in this case study, I started a broader investigation into the the use of low-NCE energy-resolved fragmentation of *O*-glycopeptides for their analysis.

3.5.1. A case study: Identifying the linker *O*-glycopeptides of heparinase-trimmed decorin

Decorin is a glycoprotein part of cell surface proteoglycans. The Ben Davis group attempts at engineering proteoglycans with controlled glycosylation states using decorin, the smallest proteoglycan, as a model protein. For this, decorin is extracted from natural sources and naturally occurring glucosaminoglycan chains are trimmed down using heparinase enzyme. This cleavage results in a double bond on 4,5 position on the terminal saccharide of the trimmed glycan, which can potentially be targeted by tailored chemistries to enable glycan extension. Synthetic proteoglycans bear high promise for application in fundamental research, medical research and cosmetics.

These efforts in synthetic chemical biology call for the parallel development of an analytical platform able to monitor changes in glycosylation as well as investigating possible pitfalls (side reactions). The sequencing of the glycan part with B/Y fragments to show connectivity is a central requirement.

Several attempts have been made by the different persons working on this project[§] to develop such an MS/MS analytical platform, using trimmed naturally occurring decorin. Initial results could be obtained in March 2014 by Ritu Raj on a sample by Dr. Wei-Min Liu (Figure 3.8). In cartilage, the decorin proteoglycan consist of repeating chondroitin/dermatan sulfate, i.e. of repeating GlcA-GalNAc units (◇-□), connected to the protein via the tetrasaccharide linker ◇-○-○-☆-SG. As such, it was initially assumed that heparinase would cleave at the first GlcA (◇) at the reducing terminus. However, the data acquired is missing glycan fragmentations or oxonium ions, raising questions whether this is a valid identification (Figure 3.8). Despite much investment in time and efforts, none of the results could be reproduced in follow-up experiments.

[§]Elisabeth Grayson, Arghya Modak, Wei-Min Liu, Charlie Fehl, Elisabete Pires, Mylène Richard, Ritu Raj, Alexandra Gafitescu

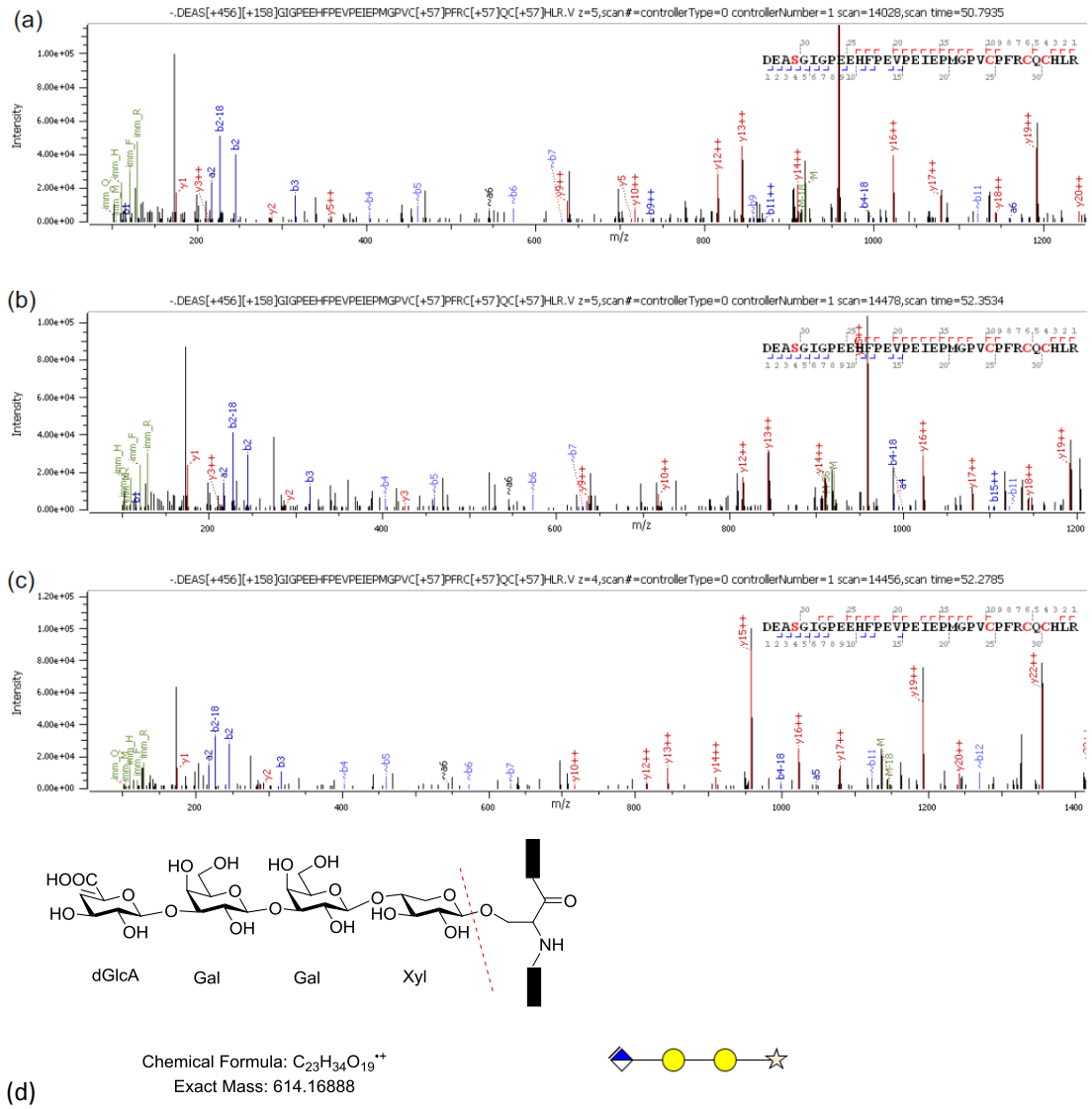


Figure 3.8. Initial results on decorin have questionable presumed glycopeptide spectra. Initial results for tryptic decorin glycopeptides obtained by Ritu Raj on a Q Exactive with HCD fragmentation, after trimming with heparinase and de-N-glycosylation with PNGaseF. Note that no oxonium ions are detected. **(a)** Glycopeptide with water loss (b4-18 peak) & intact a5 peak **(b)** Glycopeptide with water loss (b4 – 18 peak) & intact a4 peak **(c)** Glycopeptide with water loss (b4 – 18 peak) & intact a5 peak. **(d)** Structure of the glycan that was expected from the Heparinase digest. Sample from Dr. Wei-Min Liu.

In December 2018, the Zaia group published a influential study on the 'deep sequencing of complex proteoglycans'[693] in which glycan structures for the linker glycopeptides were suggested (Figure 3.9). Noteworthy, all identified linker glycopeptides are in fact hexasaccharides, suggesting that the cleavage site of heparinase is different than origin-

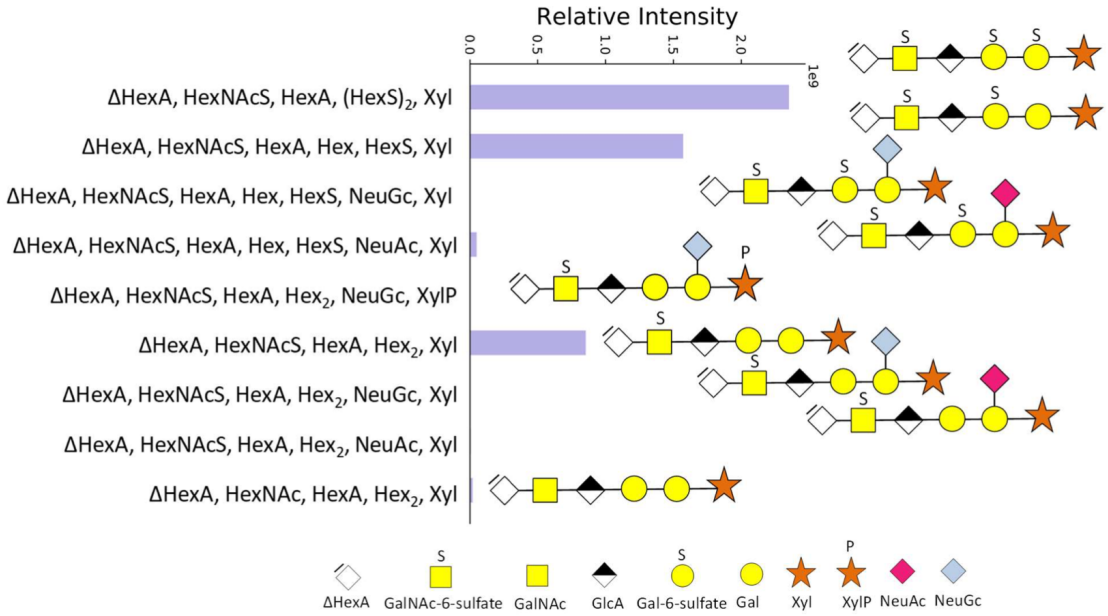


Figure 3.9. Decorin disaccharide-linker-glycopeptide compositions reported in the literature. Decorin disaccharide-linker-glycopeptide compositions. A total of 9 compositions were detected. Reproduced from[693] with permission from ©ASBMB.

ally assumed. Although the authors could not address the challenge of obtaining glycan sequencing (the MS/MS spectra contain mainly b/y ions and oxonium ions), their insights enabled us to manually curate an appropriate glycan database. In addition, this paper highlighted the importance of proline hydroxylation as a PTM on decorin.

With this knowledge, all previously acquired decorin data was re-processed. Interestingly, reprocessing of a sample from Dr. Mylène Richard run on a Q Exactive (HCD) dating back from July 2017 yielded very good identification of 3 O-glycans on decorin (see appendix 3.8.2.2): Xyl(1)Hex(2)GlcA(1)HexNAc(1)dGlcA(1); Xyl(1)Phospho(1)Hex(2) GlcA(1)HexNAc(1)dGlcA(1); Xyl(1)Hex(2)Sulfo(1)GlcA(1) HexNAc(1)dGlcA(1). The mass error on the relevant fragment ions in the Orbitap is <0.001 Da, allowing to distinguish between phosphorylation and sulfurylation. However, as visualised in Figure 3.10, no glycan sequencing information could be obtained.

These results seem to suggest that previous assumptions that heparinase would cleave at the first glucuronic acid (◈) were incorrect – it appears to cleave at the second. Moreover, the searches need to include peptide PTM such as *proline hydroxylation* as well as *sulfo-* and *phosphorylation* glycan modifications and the LC system needs to be in a suitable state for phosphorylation analysis. With this information in hand and having identified the glycans, we went back to the original intact protein LC-MS data. Using

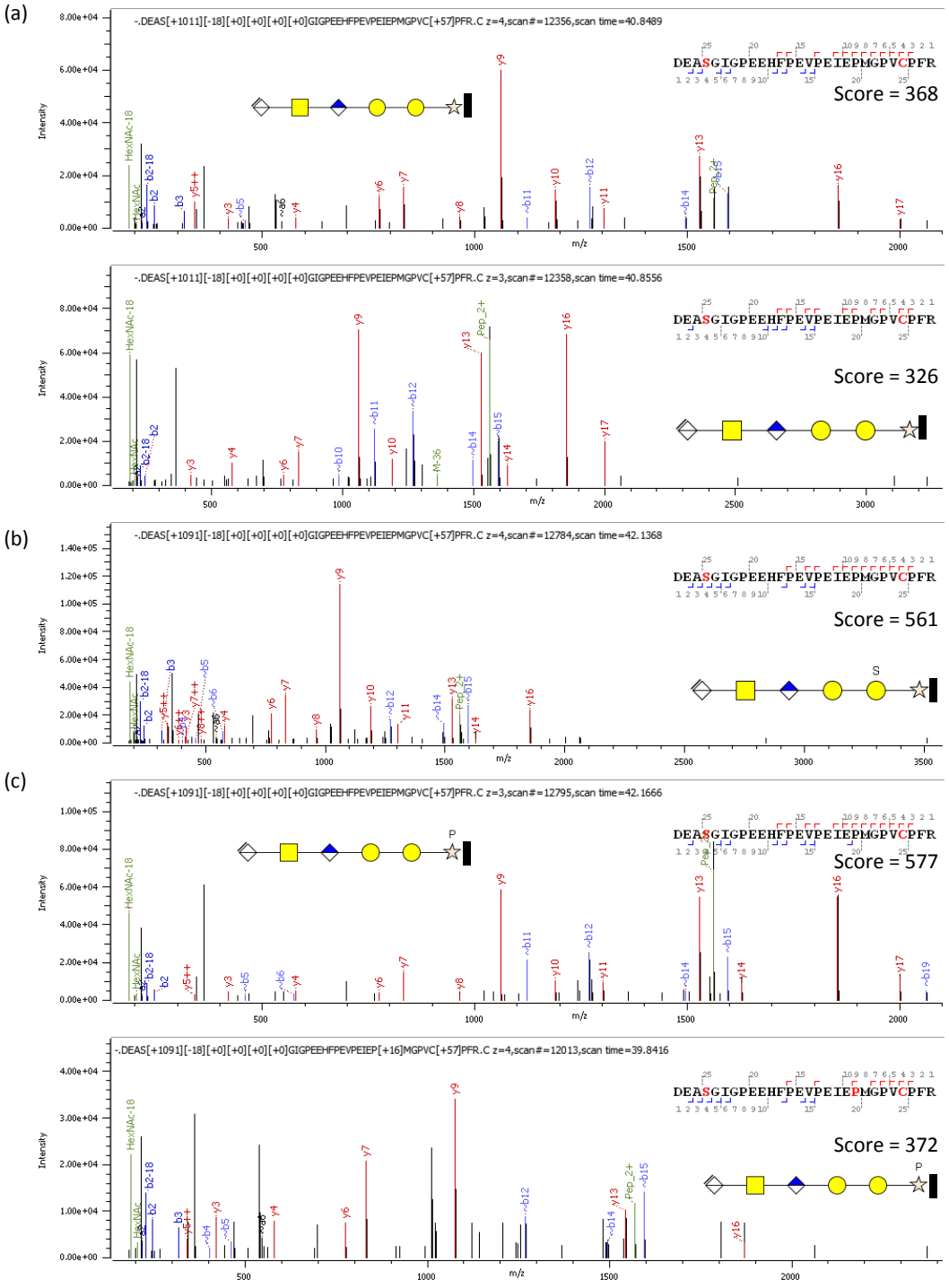


Figure 3.10. Representative spectra for the 3 identified O-glycans on decorin. Thes spectra were obtained by re-processing previously run samples. **a)** Xyl(1)Hex(2)GlcA(1)HexNAc(1)dGlcA(1). Matches the observed Intact Protein LC-MS mass of 37452. **(b)** Xyl(1)Phospho(1)Hex(2)GlcA(1)HexNAc(1)dGlcA(1). Matches the observed Intact Protein LC-MS mass of 37532. **(c)** Xyl(1) Hex(2)Sulfo(1)GlcA(1)HexNAc(1)dGlcA(1). Matches the observed Intact Protein LC-MS mass of 37532. Data obtained on a Q Exactive, standard HCD, 30% NCE.

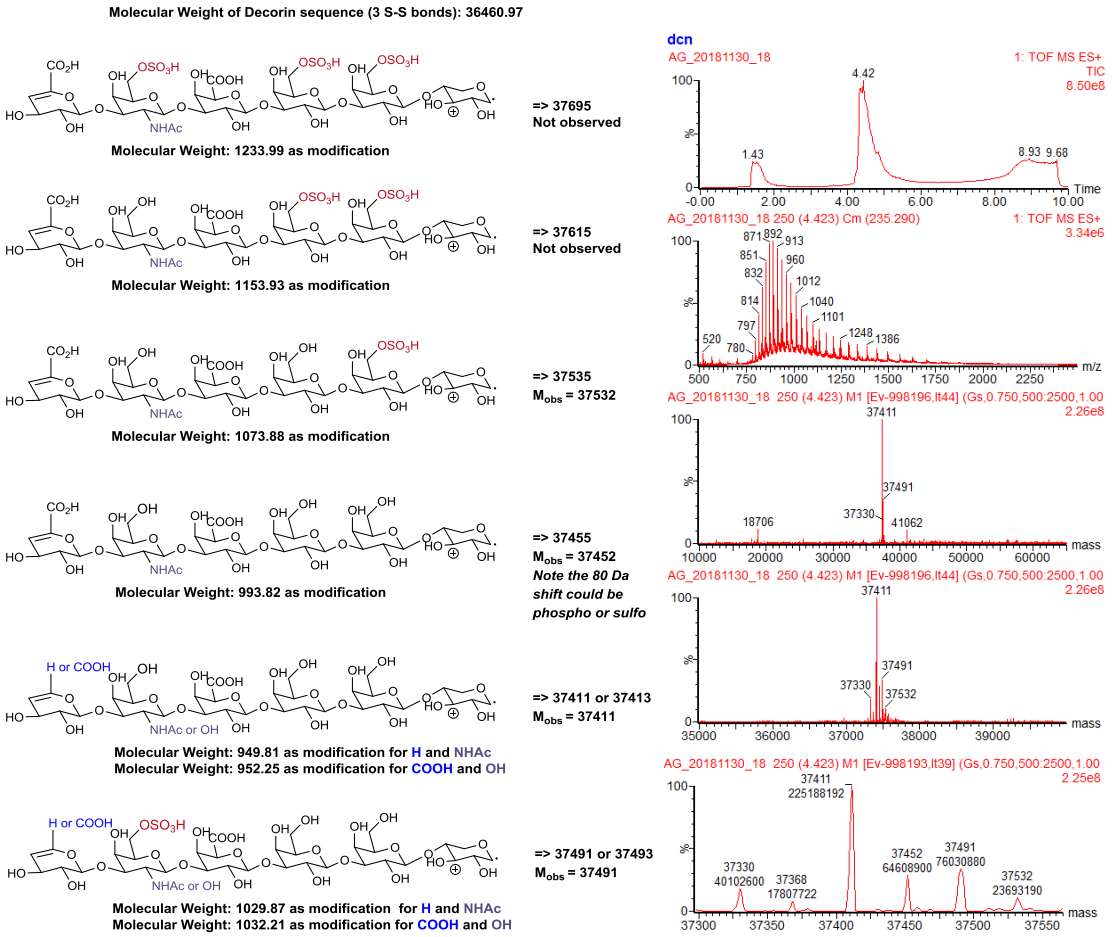


Figure 3.11. Intact protein LC-MS for isolated heparinase-trimmed decorin. Waters Xevo G2S, data from Alexandra Gafitescu. Peaks in the protein MS can be attributed to the respective glycans. A 41 Da mass shift observed suggests decarboxylation or amide loss as a modification, which can occur on either protein or glycan.

the curated sequence we use for MS/MS analysis and the identified glycans, we were able to correctly calculate the mass of the peaks in the protein LC-MS and attribute these to their glycoforms (Figure 3.11). However, the major peak is a -41 Da shift from the calculated mass of the presumed glycan, which corresponds to *decarboxylation or NHAc → OH*. As such, the data was reprocessed including the glycan modifications into the database. Equally, database entries for the hydrated double bond were included. The incorporation of these additional modifications observed in the protein LC-MS led to the identification of additional glycans in the MS/MS, using HCD fragmentation.

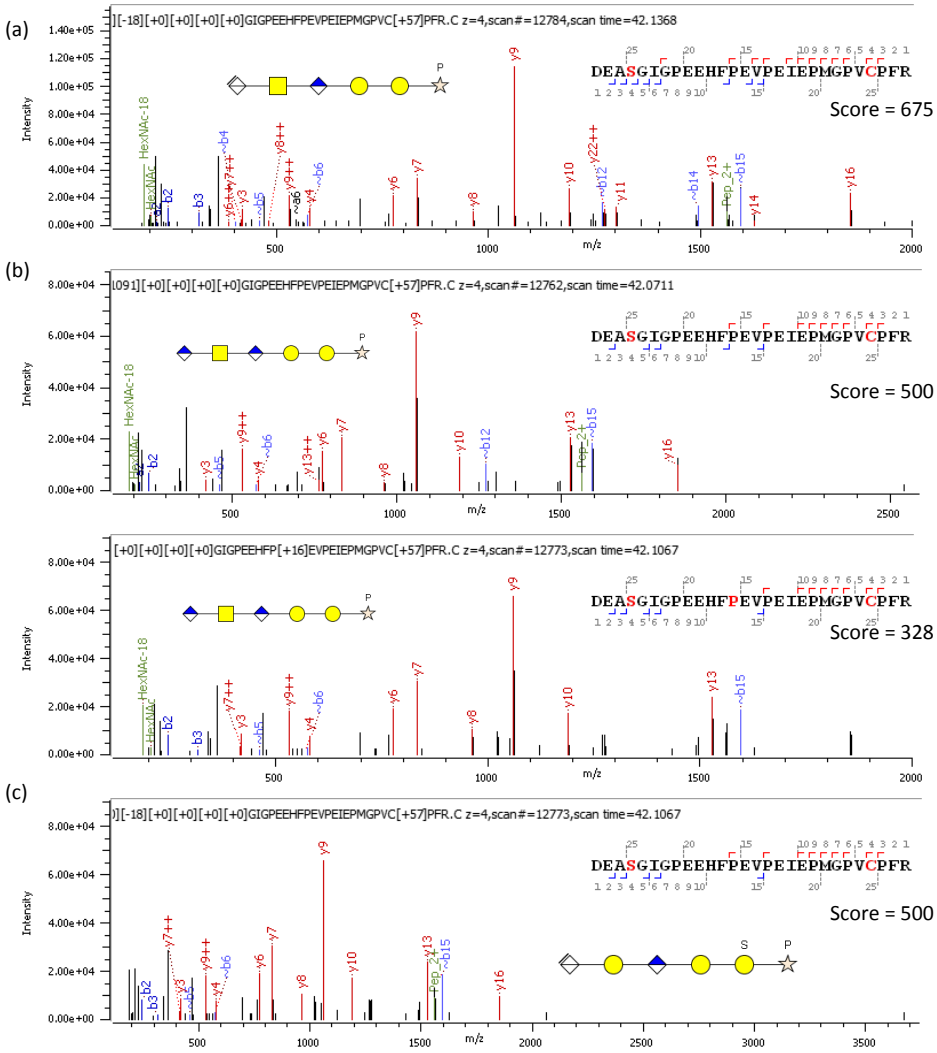


Figure 3.12. HCD spectra for additional modifications on the O-glycans of decorin.
(a) dehydrated (dGlcA) Xyl(1)Phospho(1)Hex(2)GlcA(1)HexNAc(1)dGlcA(1). **(b)** hydrated (GlcA) Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1). **(c)** Deamidation of HexNAc to Hex Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(1)dGlcA(1). (QEx, standard HCD, 30% NCE)

Next, I investigated if further modification such as double decarboxylation on both GlcA or amide loss on GlcNAc can occur, to explain the 37368, 37411 and 37491 peaks in the intact protein LC-MS. I therefore manually curated a second extensive database with all the possible modifications/combinations to investigate these possibilities. As shown in Figure 3.12, high resolution MS/MS spectra confirm amide loss (-41.0265 Da), explaining the -41 Da modification in the intact protein LC-MS. This could result either from the peptide backbone, or less likely, amide loss of HexNAc to Hex. However,

in absence of observable glycan fragmentation patterns, it is not possible to confirm any chemical information. These results emphasize the need for a glycan sequencing fragmentation method, such as resonance-activated ion-trap CID type fragmentation or lower energy HCD.

With knowledge of glycan compositions and modifications, previously acquired datasets were re-processed. As shown in appendix 3.8.2.2, it becomes apparent that the glycan originally identified by Ritu Raj is only present as a trace in the original sample, whereas the glycans that correspond to the study of Klein *et al.* are building up the bulk of the identifications.

Encouraged by these results, four different digestion protocols were tested to optimise protease digestion and thus glycopeptide properties: trypsin, chymotrypsin, GluC/trypsin and elastase. Figure 3.13 shows a comparison in peptide coverage at the site of interest for different digests of the same sample, provided by Dr. Mylène Richard. Clearly, a tryptic digest leads to the best coverage at the site of interest. This can be rationalised by the 2 additional charges coming from both the *C*-terminal arginine and the additional charge from the 2nd horseshoe (peptide length > 15 amino acids). Together with the charge on His and the *N*-terminus/1st horseshoe, this leads to the predicted charge state of 4+, observed as well in the acquired data.

The samples were then analysed with the following methods:

- Q Exactive: HCD with single NCE's of 10, 15, 20 and 30
- Q Exactive: 2-step HCD with NCE's of 10-15, 10-20, 10-30, 15-30
- Q Exactive: 3-step HCD with NCE's of 10-15-20, 10-15-30, 10-20-30, 20-30-40
- Orbitrap Elite: HCD32, HCD32-pd-ETD, HCD35-alt-ETD, EThcD35, HCD-alt-EThcD,[¶] CID, CID-pd-ETD, CID-alt-ETD, CID-pd-EThcD, CID-alt-EThcD.

Data analysis was performed with Byonic using the curated database. Unfortunately, Byonic (v 2.9) does not include a keyword for the dHexA sugar and therefore spectral annotation of glycan fragmentations does not succeed, resulting in the need for manual annotation. While this function was implemented in newer versions of Byonic, it was only released after analysis of this data. Several other post-processing and annotation

[¶]Note that due the implementation of EThcD as a 'hack' (described later in this chapter), HCD-pd-EThcD – reportedly a good method for O-glycosylation – could not be implemented. (Forbidden MS3 triggering of HCD dependent MS3 scan on HCD MS2 – as EThcD is implemented as HCD on first mass 151)

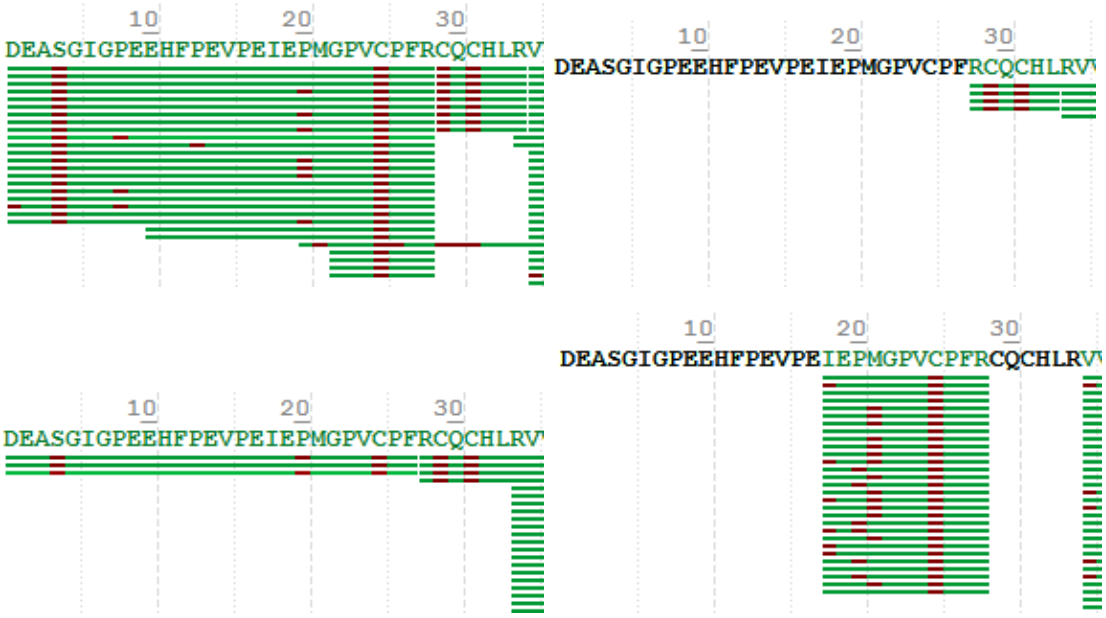


Figure 3.13. Representative HCD sequence coverage for searches with additional modifications on the *O*-glycans of decorin. (a) Tryptic digest run on a Q Exactive (standard HCD, NCE 30%) (b) Chymotryptic digest of MR307B on an Orbitrap Elite with HCD-pd-ETD. (c) Chymotryptic digest run on a Q Exactive (standard HCD, NCE 30%). (d) GluC/trypsin digest of MR307B on an Orbitrap Elite with HCD-pd-ETD. Note: elastase digest did not lead to glycan coverage (not shown).

softwares were tried using the mzML files generated by Byonic, such a GlycReSoft[344] (that includes dHexA), without succeeding in obtaining automated spectrum annotation. As such, I used GlycReSoft's glycopeptide hypothesis builder to construct myself a glycopeptide deconvoluted monoisotopic search space, i.e. a databases containing the mass of glycopeptide fragments (on both peptide and glycan including Yy and Yb fragments) to allow for speed-up manual analysis by mass matching (Figure 3.14). Concretely, this allows to quickly search the deconvoluted mass of a peak, calculated from the acquired spectrum using isotope peaks for charge determination, against a wide range of possibilities including the different possible glycan and peptide fragments.

Manual analysis of all glycan spectra was performed, including deconvolution of glycopeptides. Figure 3.15 shows the chromatogram of the digest (N-glycans were removed with PNGaseF) highlighting the 204 oxonium ion in the XIC. The rear part of the run could be attributed to decorin O-glycopeptides.

Figure 3.16, Figure 3.17 and Figure 3.19 show example pictures obtained using low SCE-HCD methods. Several features of the specta are immediately apparent. All spectra contain a block of peaks between 900 and 1300 m/z that are not annotated by Byonic

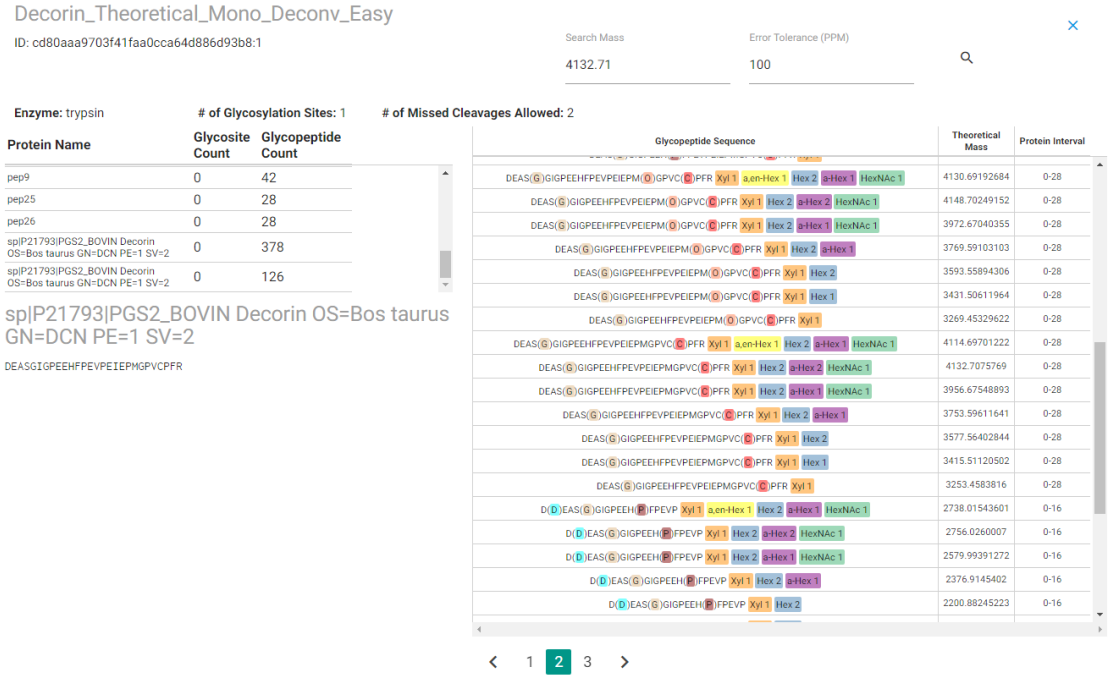


Figure 3.14. Using GlycReSoft[344] to construct a searchable mass to glycopeptide fragment database. This search space is constructed from a set of decorin peptide fragment sequences at the site of interest + glycan fragments + mass shifts of potential modifications (phosphorylation, decarboxylation, mass shift to b-ion etc.). This constructed database facilitates the task of manual analysis by allowing to rapidly interrogate peptide-glycan combinations in form of the different fragment ions.

(v2.9). These are all Y ions showing the glycopeptide fragment series. These spectra show glycan coverage from Y_0 through to Y_4, and indeed a low-abundance Y_5 is detected in the 10-15-20 spectrum. The first conclusion of these experiments is that full O-glycan sequence coverage can be achieved using low-SCE-HCD methods. This information enables unambiguous determination of the glycan sequence, and the oxonium ion intensity ratio confirms the presence of GalNAc identity.

When performing stepped HCD of 10-15 (Figure 3.17), or at a single NCE of 15 (Figure 3.24), peptide fragment ions are produced, allowing Byonic to identify the peptide. Many of these peptide fragments are not produced at a single NCE of 10, therefore the peptides are not identified by Byonic in pure NCE10 runs. These peptide fragments are either N-terminal cleavages of Pro (*Proline effect*)[696] or of Glu. Their frequency is higher at the C-terminal end, remote from the glycan. However, $\sim b$ fragment ions with loss of glycans are already observed at an NCE of 15, highlighting that the energy for O-glycopeptide fragmentation has to be tightly controlled.

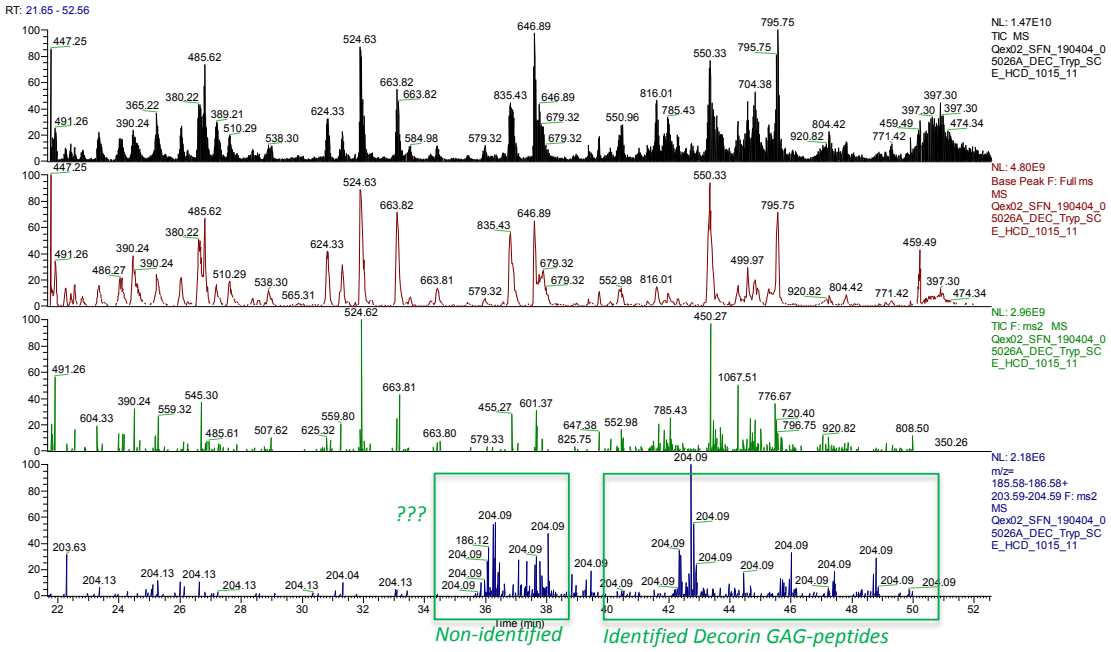


Figure 3.15. Oxonium ion trace. Chromatogram of tryptic digest run with SCE-HCD 10-15 method with XIC of HexNac and HexNAc-18 oxonium ions shows non-identified supposed glycosylations (presence of oxonium ions).

A particularly valuable asset of these methods is the generation of peptide fragment with intact glycans or (indeed rarely) the generation of Yb fragments. These fragments are not recognised by Byonic, but incredibly helpful to pinpoint the site of glycan attachment. As such, the y_{26} (loss of N-DE) and b_{12} fragments with full glycan localise the O-glycan in a stretch of 10 amino acids with only one serine. The Y_4b_{17} and Y_4b_{19} fragment ions highlight that the loss of HexNAc-dGlcA disaccharide is probably the energetically most accessible pathway, further supported by the high intensity of the B_2 oxonium ion(s), which dominates all spectra (Figure 3.18).

Phosphorylation and sulfation of proteoglycan linker peptides have been reported and there is an ongoing discussion whether phosphorylation is true or mis-interpreted.[1] Interestingly, Figure 3.17 shows formation of the $[\text{Pep}+^{0,2}X_0]^{3+}$ ion for a phosphorylated peptide, which does not appear in other, non-modified spectra. This suggest a modification on the reducing-terminal xylose (★), only phosphorylation of xylose at C-2 has been reported.[1] Therefore, it might well be that the hexasaccharide is indeed phosphorylated at xylose and the $[\text{Pep}+^{0,2}X_0]^{3+}$ ion would be a characteristic cross-ring fragmentation, made energetically accessible by transforming a hydroxyl group into a good leaving group (phosphate) thus triggering this $^{0,2}X_0$-type fragmentation.

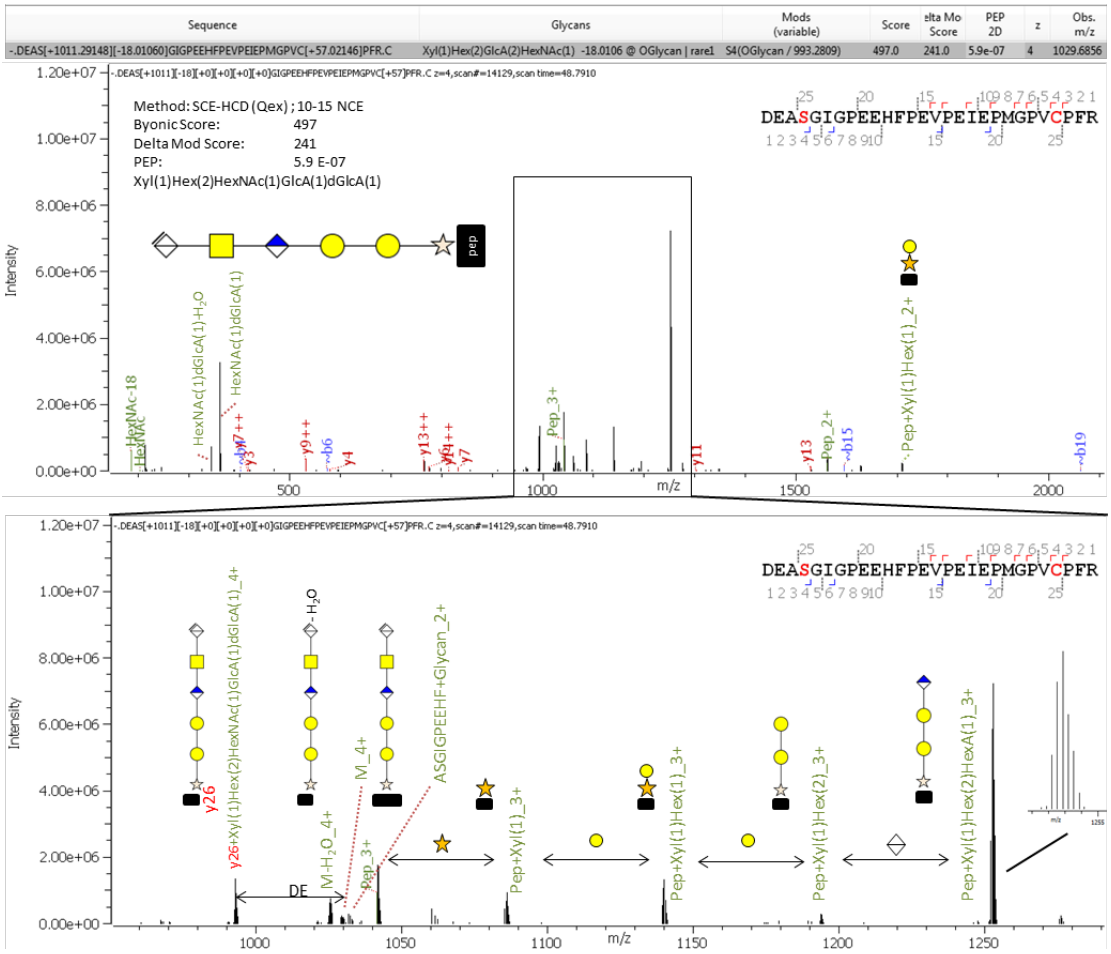


Figure 3.16. 2-step HCD with 10-15 NCE yield glycan fragmentation on decorin glycopeptides. Spectra for the Xyl(1)Hex(2)HexNAc(1)GlcA(1)dGlcA(1) glycan. Note that fragmentation of the peptide backbone by loss of *N*-terminal DE or fragmentation of the peptide backbone *N*-terminal to proline is observed as a minor fragmentation pathway, with glycan fragmentation being the major fragmentation pathway. In charge state 3+, glycan fragmentation with attached peptide at the reducing end is observed, in charge state 4+ only peaks that have not yet fragmented are detected, or the y_{26} ion (loss of DE). Interestingly, the Pep+Xyl(1)Hex(2)HexA(1) intermediate seems to be a stable intermediate in the fragmentation process.

All these described fragmentation patterns are still observable at a single NCE of 15, but many are lost at an NCE of 20, at which only few glycosylated fragments remain (Figure 3.24). This highlights the importance of precise energetic settings for *O*-glycopeptide analysis. As do others, I believe that the richer the spectrum in Y/B, intact b/y and mixed Yy and Yb ions, the better the method is suited for *O*-glycopeptide characterisation.[51] Unfortunately, studies such as Riley *et al.*[652] have mainly looked at NCE's of 20 and higher for *O*-glycan characterisation, missing the rich fragmentation

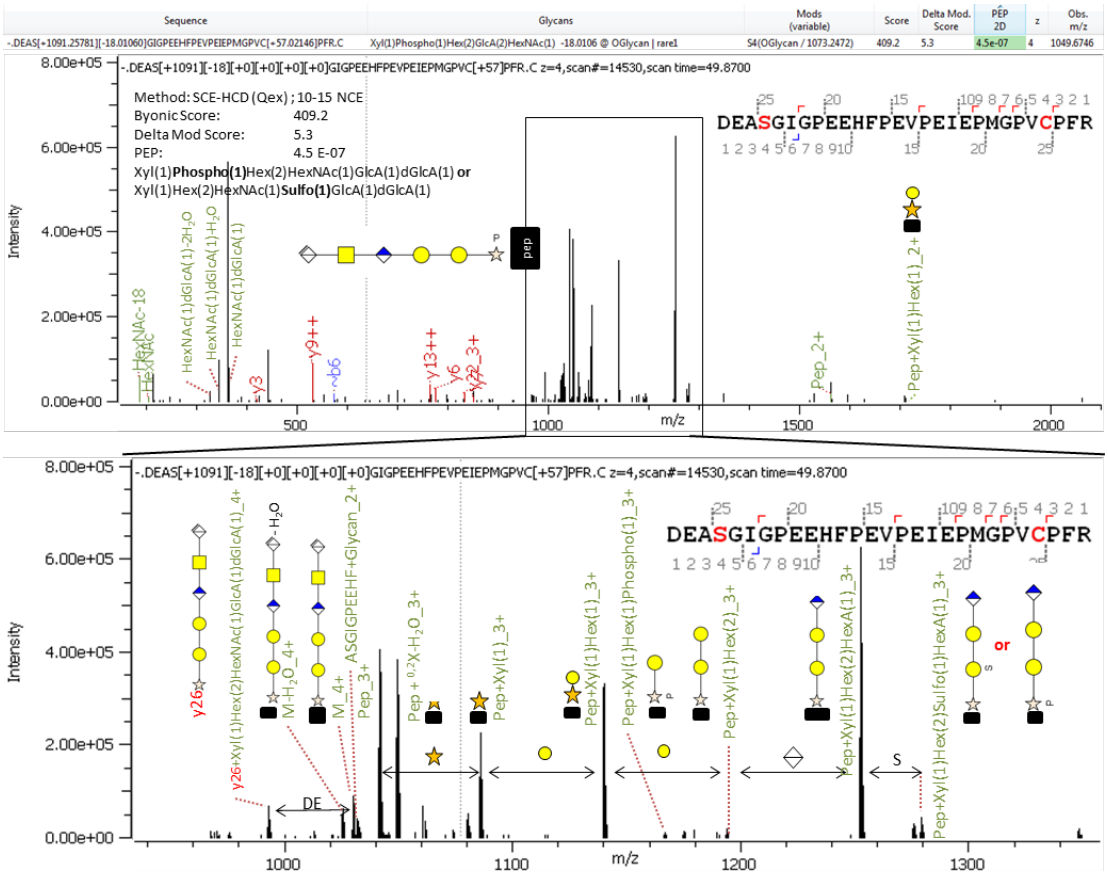


Figure 3.17. 2-step HCD with 10-15 NCE yields glycan fragmentation on decorin glycopeptides. Spectra for the Xyl(1)Phospho(1)Hex(2)HexNAc(1)GlcA(1)dGlcA(1) glycan. Note that this could be attributed to the Xyl(1)Hex(2)Sulfo(1)HexNAc(1)GlcA(1)dGlcA(1) as well, as we are not able to identify the difference from this spectrum. As the $[^{0,2}X\text{-}H_2O]^{3+}$ mass is observed, phosphorylation appears to be more likely, as such a fragment would only make sense in the case of a phosphorylated xylose. Samples were run on a phospho-friendly QEx.

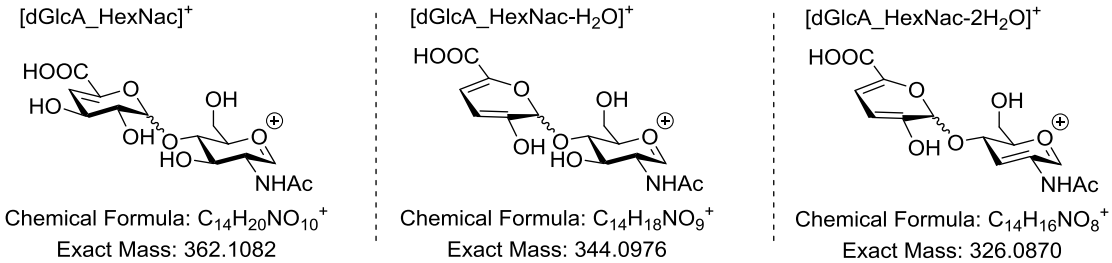


Figure 3.18. Oxonium ions characteristic for heparinase-trimmed decorin. Structures shown are hypothetical and although the masses are observed, suggested structures have not been validated.

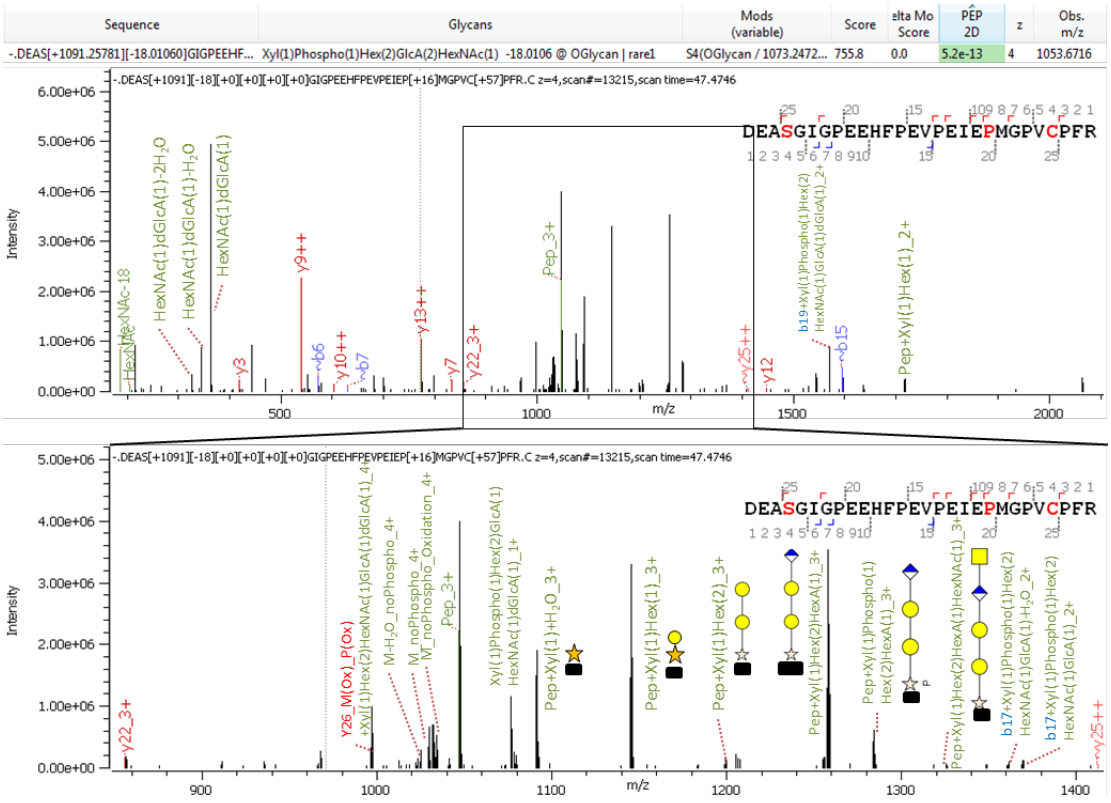


Figure 3.19. 3-step HCD with 10-15-20 NCE yields glycan fragmentation for decorin glycopeptides Of particular interest are the Y_4b_{17} and Y_4b_{19} fragment ions.

available at lower energies.

As the manual analysis is very time-demanding, Marshall Bern, Protein Metrics Inc, kindly implement support for dehydro-HexA into Byonic, facilitating the analysis. Later use of newer Byonic versions (v 3.6) showed that a substantial amount of valuable fragments such as Y-ions, X-ions, Yb ions or y/b ions with glycan attached are however still not annotated, and probably not used by the software in scoring functions (Figure 3.24). Byonic is unquestionably a leading software in O-glycopeptide analysis, this observed drawback emphasises that manual analysis is still required – this has to be considered when performing method development.

The method development outlined in this section resulted in a successful and reliable analysis of decorin O-glycopeptide, allowing the uses of this method in separate projects to monitor progress in synthetic glycosylation. However, the success of low-energy SCE-HCD methods for the decorin O-glycopeptides prompted us to analyse if these observations holds equally true for other O-glycopeptides. Also, I wanted to be able to differentiate between parameters influencing fragmentation and distinguish which

method performs better between the different stepped-HCD methods starting at NCE's of 10 or 15. Our approach placing MS^2 fragment ions as the central reference point is arguably a good way forward to assess the quality of a given method of *O*-glycopeptide MS/MS.

3.5.2. Investigating the generality of low-NCE energy-resolved fragmentation of *O*-glycopeptides

Encouraged by the success of low-energy stepped-HCD methods for decorin proteoglycan linker *O*-glycopeptide, their reproducibility on other glycoproteins with mucin-type glycosylation was assessed. This section aims at providing a qualitative glance into initial findings using this approach. Especially, we are interested which methods are particularly suited at (i) providing Y fragment ions with fragmentation patterns enabling to read out the entire glycan sequence (and perhaps connectivity), at (ii) forming b/y ions carrying the intact glycan, and (iii) at forming Yb or Yy ions showing two sites of fragmentations − if any.

3.5.2.1. Single NCE runs pinpoint the collision energy range for glycan fragmentation of *O*-glycopeptides between 10 and 20 %

Normalised collision energies provide a straightforward value to compare collision energy settings across Thermo instruments. As described in 3.2.5, studies with *N*-glycopeptides typically show an ideal collision energy range between 15 % and 25 % NCE, peaking at 20 %. The decorin data presented in 3.5.1 already suggested that the ideal range for fragmentation of *O*-linked glycopeptides could be different. Therefore, single NCE runs were performed on asialofetuin, a frequently used model glycoprotein of which sialic acid-terminal saccharides have been removed by mild acidic treatment. The rationale for the choice of asialofetuin lies in its well-studied mucin-type glycosylation. The lability of sialic acids could falsify the viewpoint at which NCE different fragmentation pathways are energetically accessible, hence the choice of this readily available asialated glycoprotein.

Several glycopeptides were identified on asialofetuin, matching the identifications described in the literature. Out of these glycopeptides, two glycopeptides were selected, namely PSVVAVPLPLHR and AAGPTPSAAGPPVASVVVGPSVVAVPLPLHR, both carrying core 1 T-antigen GalNAc(1)Gal(1) (○-□-S/T). These two examples

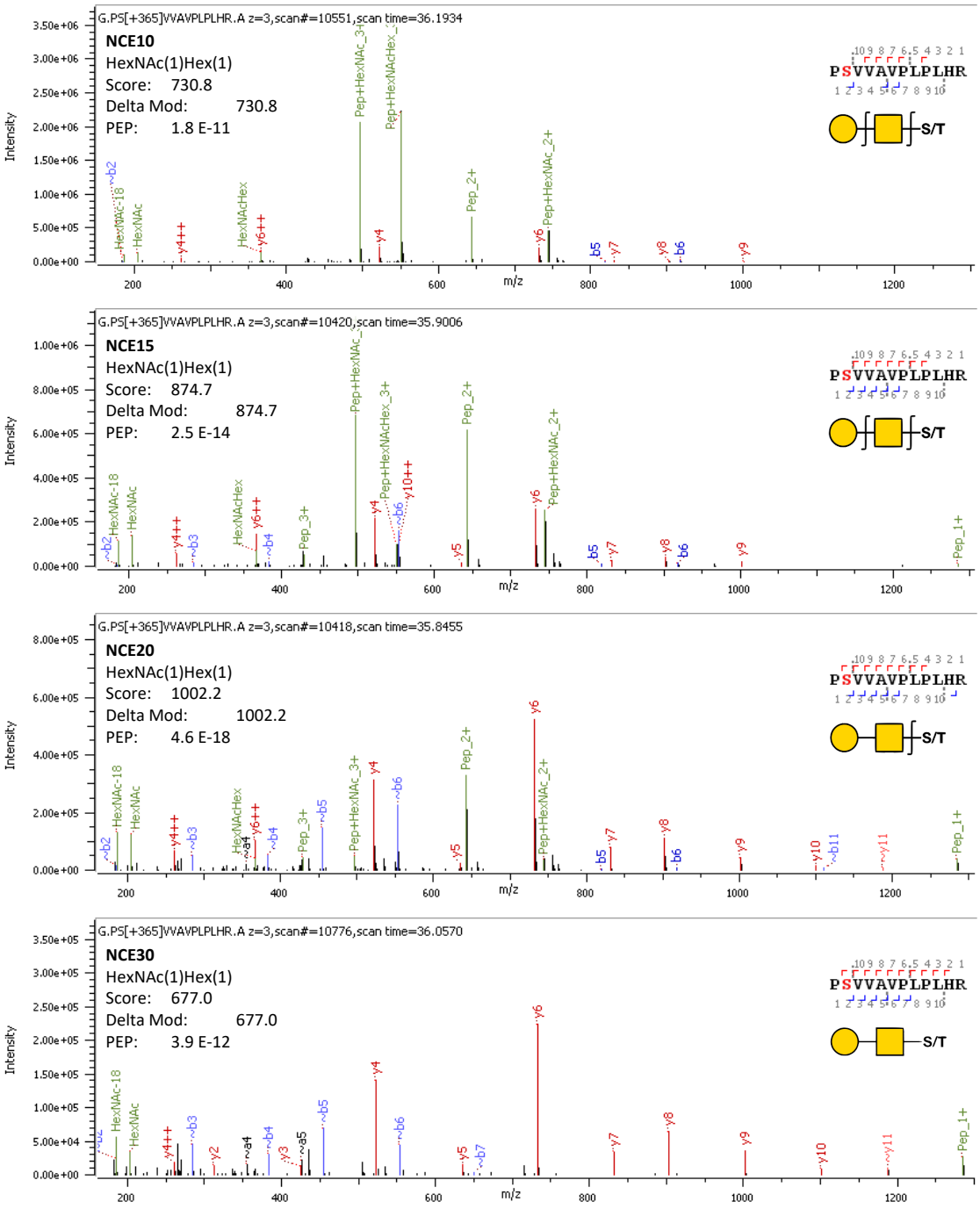


Figure 3.20. Fragmentation of the PSVVAVPLPLHR carrying HexNAcHex from bovine asialofetuin at four different NCE values. Q Exactive, $[M+3H]^{3+}$.

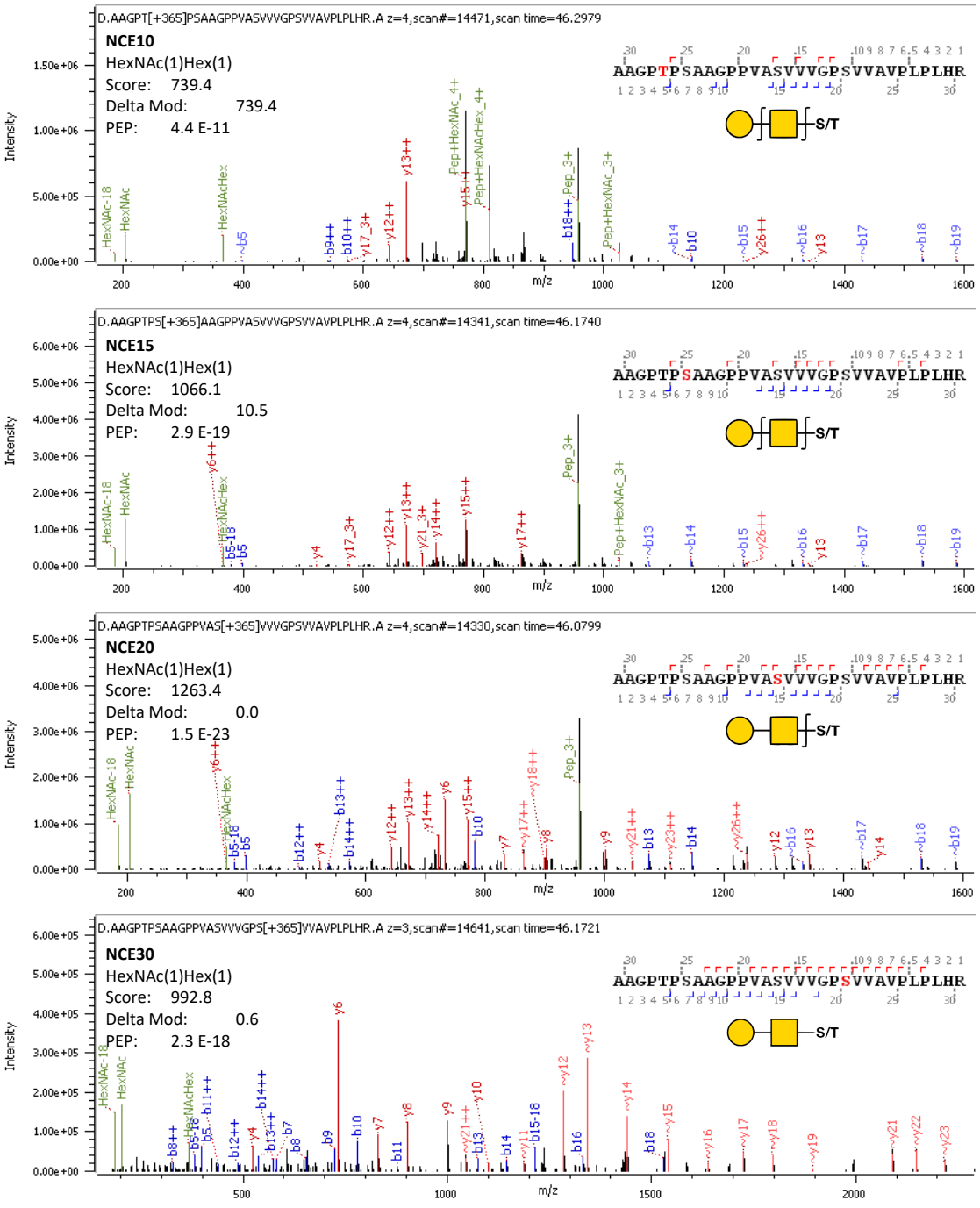


Figure 3.21. Fragmentation of the AAGPTPSAAGPPVASVVVGPS-VVAVPLPLHR carrying HexNAcHex from bovine asialofetuin at four different NCE values. Q Exactive, $[M+4H]^{4+}$.

represent extremes in glycopeptide length, the determining factor guiding glycan fragmentation aside from charge state – *O*-glycosites have similar frequencies of prevalent amino acids (mainly Pro, Ser, Thr, Val, Leu, Ala) and the effect of glycans on glycopeptide fragmentation has previously been shown to be small.[684,685]

For the shorter PSV...LHR peptide, glycan fragmentation is at NCE 10, 15 and 20, with decreasing intensity of the Y ions (Figure 3.20). Indeed, the Y_1^{3+} ion (**-S**) disappears at NCE20. No glycan fragmentation can be recovered at NCE30, only $\sim b/y$ ions with loss of glycan. For the longer AAG...LHR peptide, the Y^{4+} ions disappear already at NCE of 15. Except for the Y_0^{3+}, no Y ion is observed at NCE of 20 (Figure 3.21). The glycan is located on either Thr5 or Ser7 according to the literature, which can be conformed by NCE10 or NCE15. Intact b/y ions carrying the glycan are only observed at a NCE of 10 (i.e. b_9^{2+} and $b_1 0^{2+}$).

This difference of the exact NCE where fragmentation is observed can be explained by two reasons. Firstly, a higher charge state facilitates dissociation in CID, hence less energy is required. However, Q Exactive instruments account for this to a limited extend by the inclusion of charge factors when calculating the CE for a given m/z. The additional underlying reason is that the slope and offset of the CE-to-m/z linear correlations are different for peptides and glycopeptides (Figure 3.5). As such, in a sense "too much" energy is given to glycopeptides. This effect is exacerbated the larger the peptide is, as an m/z increase follows and the gap between the two curves widens, leading to earlier fragmentation at lower NCE for longer glycopeptides.

Two potential solutions can be thought of. One is to use a different calibration curves for glycopeptides, which is unrealistic as this would require a substantial effort from the vendor. The other solution is to use adapted SCE-methods, as is done for *N*-glycopeptides. Therefore, I set out to to understand how stepped-energy methods can be best applied.

3.5.2.2. Low-energy SCE-HCD methods are advantageous for *O*-glycopeptide glycan sequencing

For a more representative picture, four glycopeptides including the two asialofetuin glycopeptides presented in the previous section, the decorin linker peptide and a branched sialylated mucin-type hexasaccharide from fetuin were analysed using a variety of low-energy SCE-HCD methods.

The short PSV...LHR example shows glycan fragmentation at SCE 10-15 and 10-20 (Figure 3.22). All three show intact b ions enabling site-localisation of the glycan. Interestingly, SCE 10-20 shows the intact b_2 ion with glycan, enabling site-localisation by SCE-HCD. Neither SCE 10-30 nor 15-30 deliver a satisfying result for this short peptide. For this glycopeptide, the SCE 10-20 delivered the best result in terms of identification of desired fragments, followed by a single NCE 10 and by SCE 10-15.

Similar trends can be observed with the longer AAG...LHR peptide (Figure 3.23). NCE10, SCE 10-15 and SCE 10-20 observe similar Y ions. However, both NCE10 and SCE 10-15 observed the intact b_{10} ion, allowing to pinpoint the correct, literature-known (ETD) site of glycosylation (either Thr5 or Ser7). With this peptide, SCE 10-15 and SCE 10-20 give comparable results. Byonic (v 2.9) scores the latter higher as more b/y ions are formed − which is helpful considering that the software needs to pick-up the peptide in the first place, for which b/y ions are essential. However, the value of the b_{10} ion seen in SCE 10-15 for site-localisation is not accounted for.

Decorin provides an example for a linear hexasaccharide placed on a particularly long peptide (Figure 3.24). All low energy spectra have highly similar features, therefore the scoring metrics of Byonic (v 3.6) provides arguably the best indication. Byonic does not annotate Yy and Yb ions described in the previous section, these are however identified in most of the spectra but are lacking in the pure NCE20. Highest scores are obtained with SCE 10-20-30 and 10-20 followed by 10-15-20.

The last example is the core 2 hexasaccharide on the much shorter VVV...LHR peptide of fetuin (Figure 3.25). Again, the 10-20 method provides the best spectrum. Note that in this particular case, most Y ions had to be annotated manually as Byonic (v 2.9) failed to assign these, explaining the comparatively low score for high-quality *O*-glycopeptide fragmentations. Importantly, the glycan sequence and branching can be derived from the Y-ion series.

In summary, this chapter shows that low energy stepped methods such a SCE-HCD 10-20 are advantageous for the sequencing of *O*-glycopeptides. These initial findings need further investigation and validation on further model *O*-glycoproteins.

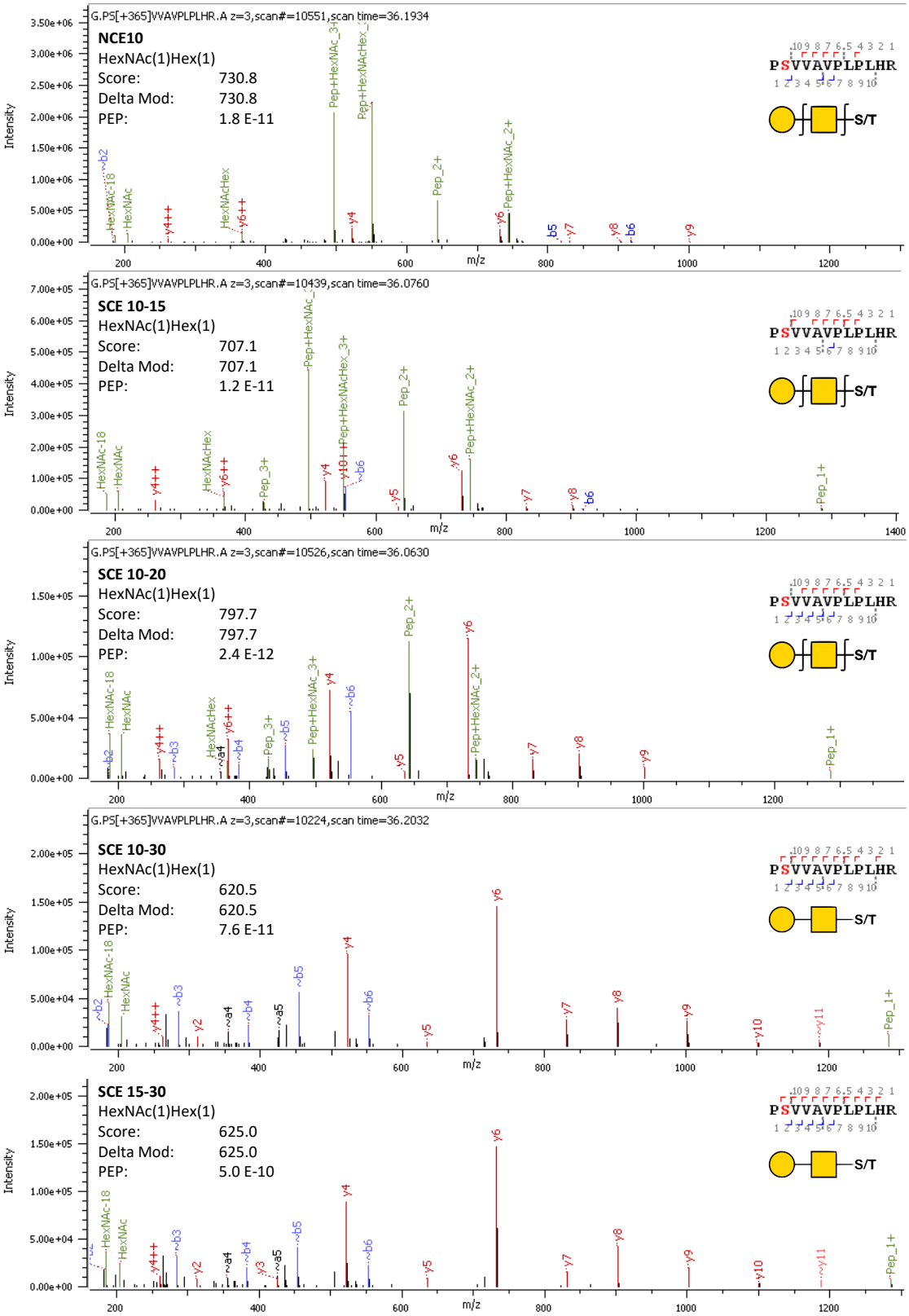


Figure 3.22. Fragmentation of the PSVVAVPLPLHR glycopeptide carrying HexNAcHex from bovine asialofetuin in four different SCE-HCD methods. The NCE10 spectrum has been included for comparision. Q Exactive, $[M+3H]^{3+}$.

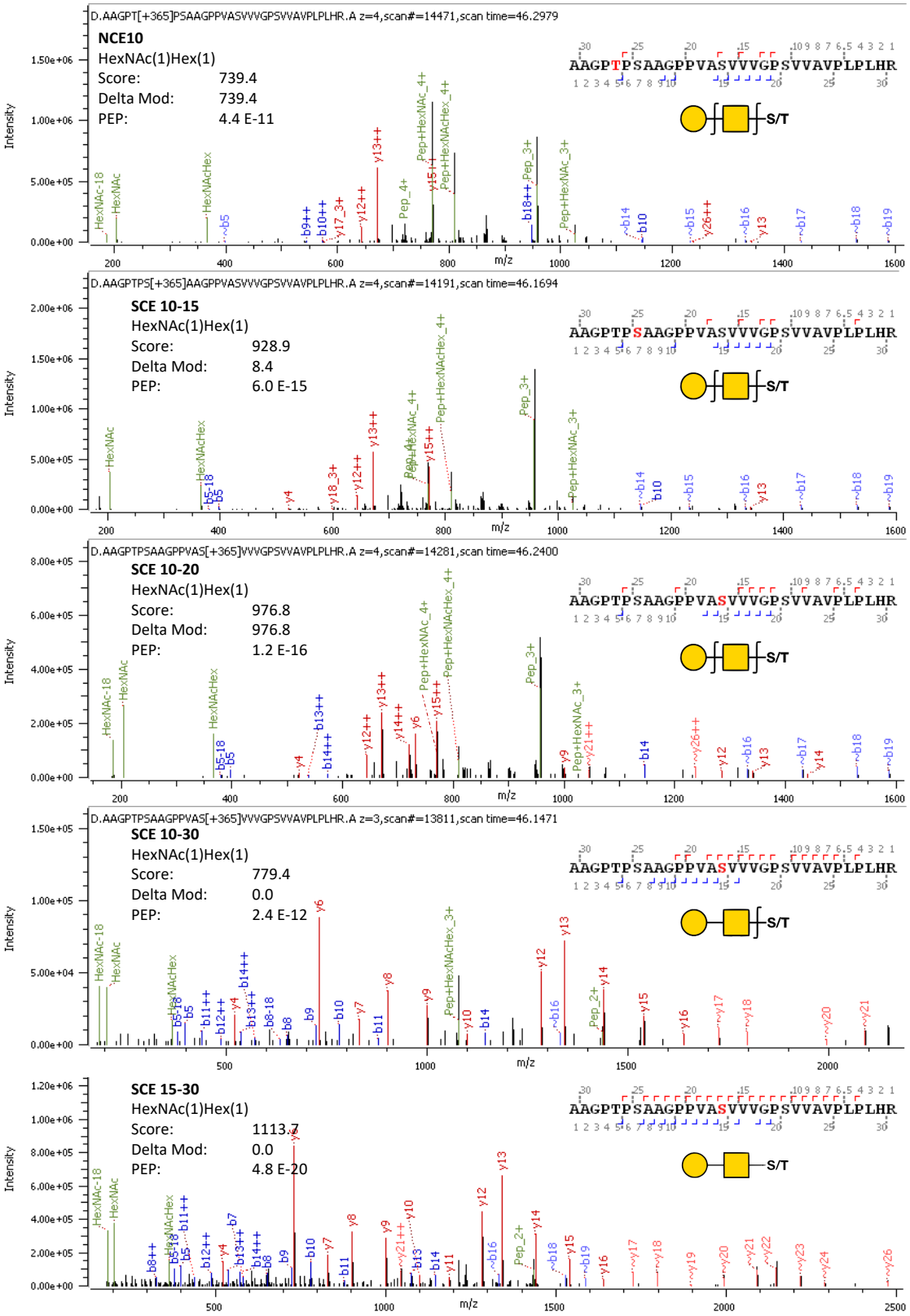


Figure 3.23. Fragmentation of the AAGPTPSAAGPPVASVVVGPSVVAVPLPLHR glycopeptide carrying HexNAcHex from bovine asialofetuin in four different SCE-HCD methods. The NCE10 spectrum has been included for comparision. Q Exactive, $[M+4H]^{4+}$.

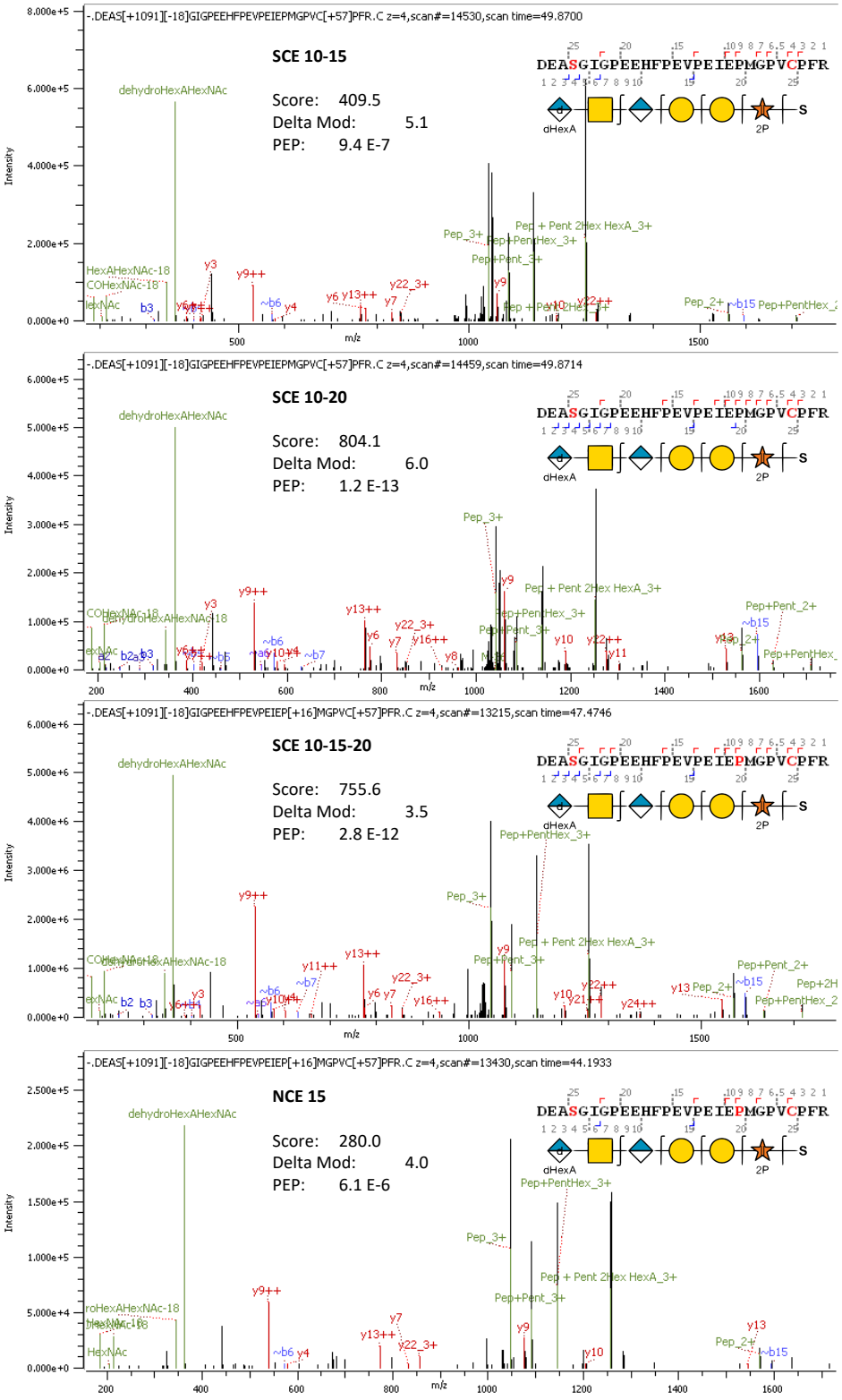


Figure 3.24. Fragmentation of the bovine decorin glycopeptide in different NCE and SCE-HCD methods. Q Exactive, $[M+4H]^{4+}$.

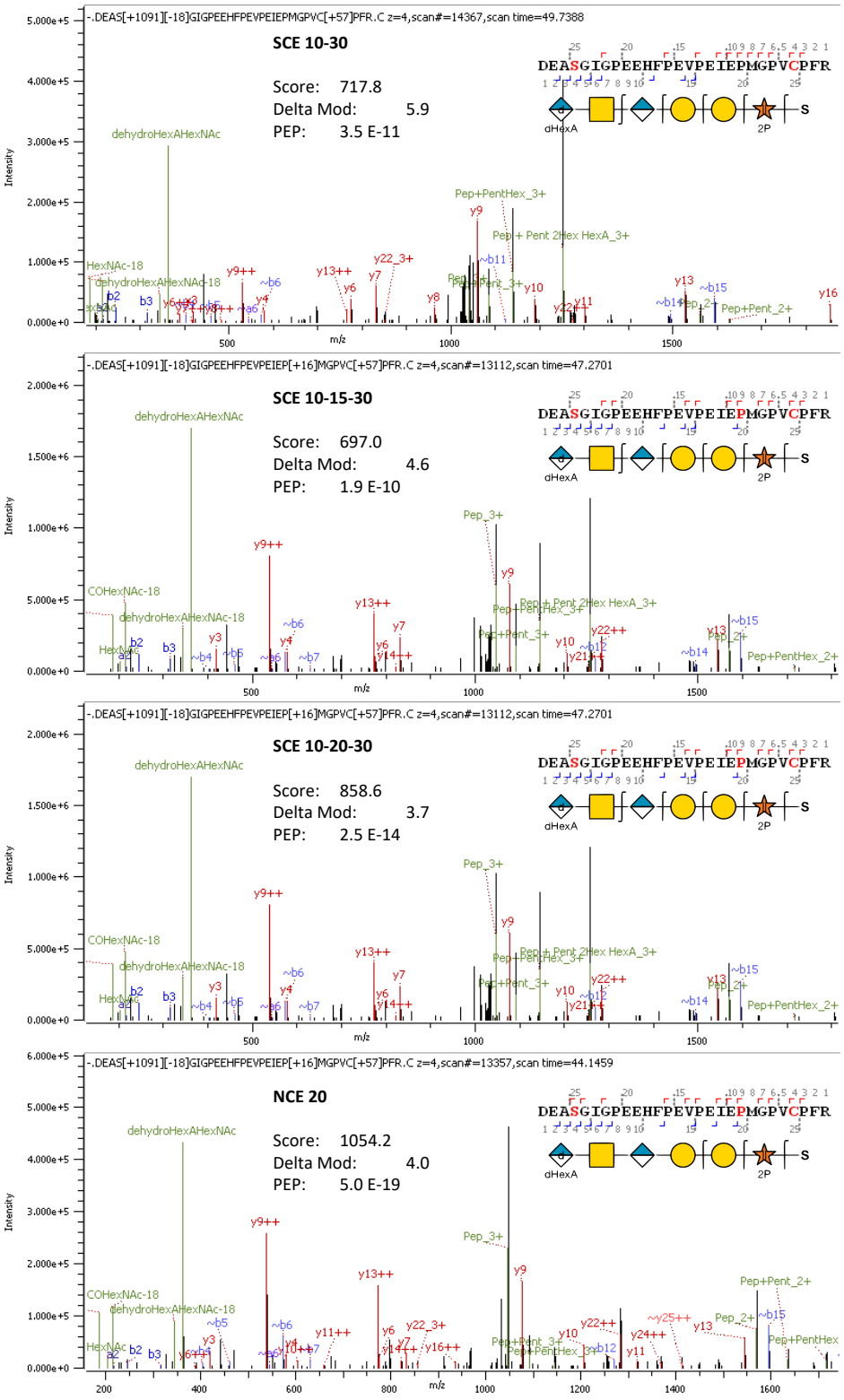


Figure 3.24. (*continued*)

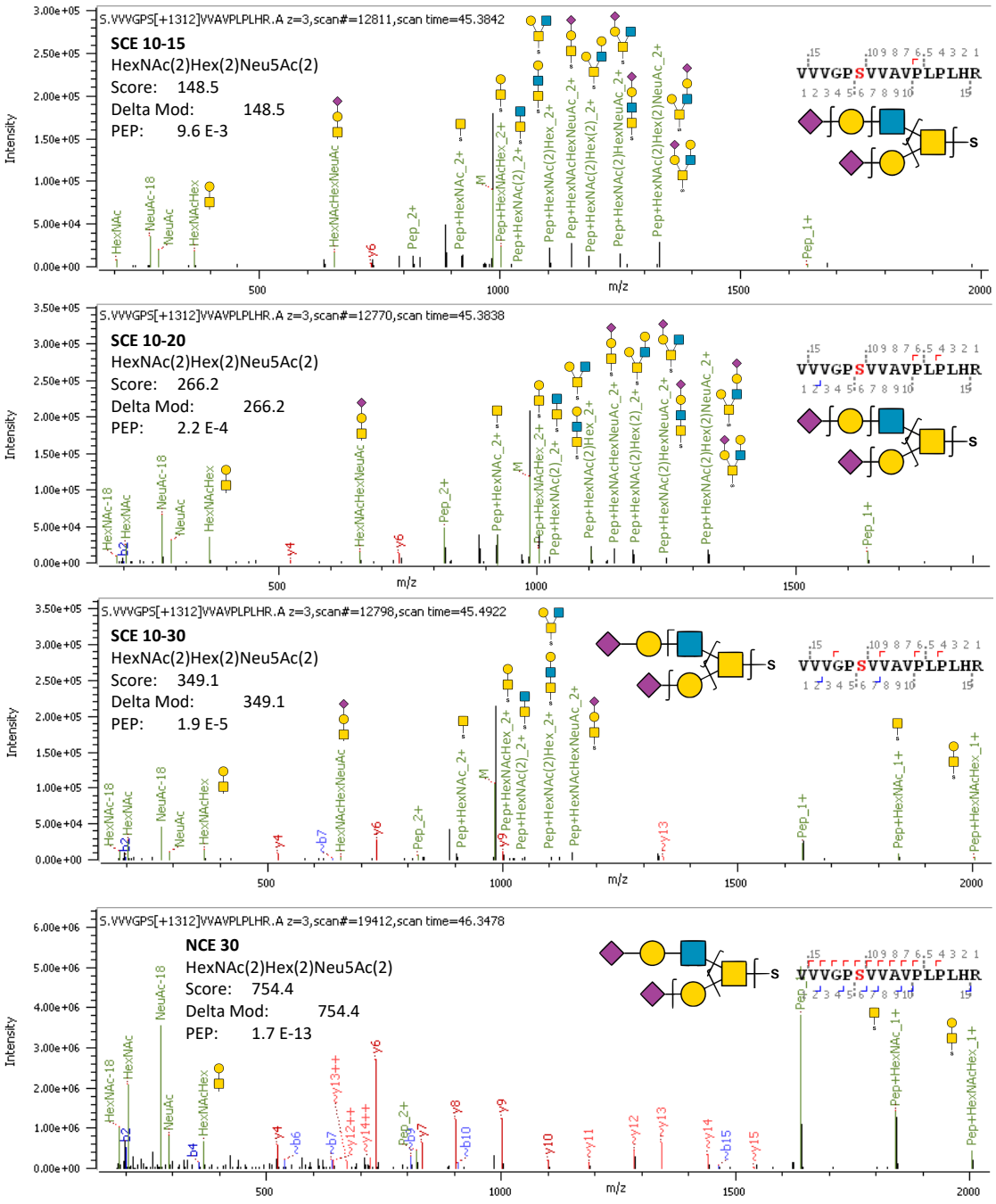


Figure 3.25. Fragmentation of the VVVGPSVVAVPLPLHR glycopeptide carrying HexNAc(2)Hex(2)Neu5Ac(2) from bovine fetuin in four different SCE-HCD methods. Q Exactive, $[M+3H]^{3+}$.

3.5.2.3. Low-energy HCD methods affect the fragmentation and scoring of other components

The previous section showed that low SCE-HCD can be beneficial for *O*-glycopeptide analysis, therefore the question I subsequently asked is whether MS/MS acquisition methods with such methods centred on *O*-glycopeptides can be used for the joint analysis of *N*-glycopeptides and peptides in a sample.

The performance of different fragmentations techniques was assessed on the sequencing of model glycopeptides used in the literature, such as IgG and RNAse B for *N*-glycosylation and fetuin for mixed *N*- and *O*-glycosylation. I did not enrich for glycopeptides as I wanted to get an impression of the effects of a given fragmentation technique on regular peptide PSMs. I investigated stepped collision energy (SCE) based HCD methods on the QExactive as well as product-dependent methods that have been suggested to be well-performing for glycans in the literature. Searches were conducted with Byonic (v 2.9) and the results listed in Table 3.3. Protein Log Score, Intensity, sequence coverage and total PSM provide useful metrics to assess the performance of the method on regular peptide identifications (see 3.4.3.3). The Byonic Score, the Byonic Delta Modifications Score, the Posterior Error Probability Score and manual analysis observations for PSMs of glycosylated peptide enable to assess the performance of the methods on different types glycopeptides. All these values are PSM-related. This choice is based on the fundamental assumption of this study that only well-fragmented and accurately annotated spectra contain the necessary information for correct assignment of glycopeptides in mass matching searches − such as Byonic does. My current opinion is that assignments based on mass difference between MS1 and backbone sequencing ID are very prone to error – this caveat of regular HCD data acquisition on QEx for glycopeptides yields many identifications but does not give 'secure' identification. Yet this methodology has a fundamental flaw which needs to be kept in mind: PSM's and the associated metrics reflects the instrument side of things and not the correctness of the assignment – as such all assignments are shown here without any manual pre- selection. It is therefore difficult to assess the presence of false positives. A comparison of metrics is therefore required, as assessment of a fragmentation technique cannot be done simply on the number of PSMs or any other metric by itself.

Table 3.3. Comparing fragmentation techniques for *N*- and *O*-glycopeptides on an Orbitrap Elite and Q Exactive. Please refer to section 3.4.3.3 for details on used metrics. Note that Byonic applies a 5% FDR at PSM level (and 1% FDR at protein level).

Method	Log-Score protein	Intensity protein; PEPrange[1]	Seq. Cover.	# PSM all pep	# PSM Glyco	Peptide Score avg (Glyco)	Delta Mod Score avg (Glyco)	Glycan fragmentation & annotation
Ig G – *N*-glycosylation								
QEx: HCD Standard 30 [2]	429.29	1.91 E+10	73.6%	773	124	177 ± 78	138 + 89	excellent
QEx: SCE-HCD 10-15	353.70	5.64 E+09	69.9 %	302	0	0	0	none
QEx: SCE-HCD 10-20	417.56	5.38 E+09	73.3 %	373	1	163.7	94.0	no b/y good Y
QEx: SCE-HCD 10-30	412.97	5.37 E+09	73.6 %	408	12	204 ± 39	170 ± 63	excellent
QEx: SCE-HCD 15-30	472.78	4.71 E+09	75.2 %	417	13	200 ± 87	159 ± 92	excellent
QEx: SCE-HCD 10-15-20	458.15	7.17 E+09	68.4 %	364	0	0	0	none
QEx: SCE-HCD 10-15-30	389.15	5.34 E+09	74.5 %	355	2	154 ± 78	154 ± 78	no b/y good Y
QEx: SCE-HCD 10-20-30	393.68	5.25 E+09	74.8 %	351	8	216 ± 45	208 ± 39	very good Y
QEx: SCE-HCD 20-30-40	379.98	3.23 E+09	73.6 %	404	34	224 ± 87	199 ± 103	excellent
ORE: HCD-pd-ETD	257.09	1.40 E+08	65.3 %	175	9	218 ± 42	140 ± 96	very good Y decent c/z ; M
ORE: EThcD	183.22	9.45 E+07	65.0 %	232	7	157 ± 58	60 ± 39	excellent ends very good c/z
ORE: CID-pd-ETD	359.67	4.40 E+08 PEP E-01	92.9 %	280	4	90 ± 56	21 ± 1.3	very good Y no ETD
ORE: CID-pd-EThcD	378.26	4.82 E+08 PEP E-01	96.3 %	289	2	97 ± 24	35 ± 14	very good Y no EThcD
RNAse B – *N*-glycosylation								
QEx: HCD Standard 30 [2]	467.02	2.18 E+11	78.0 %	1726	78	158 ± 65	79 ± 49	no b/y good Y
QEx: SCE-HCD 10-15	314.74	1.08 E+11	82.0 %	916	38	155 ± 67	8.8 ± 11	no b/y good Y
QEx: SCE-HCD 10-20	414.88	9.34 E+10	82.0 %	1049	47	169 ± 82	12 ± 18	no b/y good Y
QEx: SCE-HCD 10-30	364.47	6.51 E+10	82.0 %	1132	65	138 ± 104	23 ± 25	no b/y good Y
QEx: SCE-HCD 15-30	431.42	5.86 E+10	78.0 %	1128	66	138 ± 84	26 ± 27	no b/y good Y
QEx: SCE-HCD 10-15-20	398.05	7.87 E+10	82.0 %	1042	45	160 ± 81	11 ± 14	no b/y good Y
QEx: SCE-HCD 10-15-30	390.29	7.50 E+10	78.0 %	1060	44	150 ± 89	20 ± 24	no b/y good Y
QEx: SCE-HCD 10-20-30	425.01	6.81 E+10	82.0 %	1029	45	136 ± 98	22 ± 27	no b/y good Y
QEx: SCE-HCD 20-30-40	450.41	3.66 E+10	78.0 %	941	37	187 ± 98	54 ± 46	no b/y good Y
ORE: HCD-pd-ETD	259.33	1.01 E+09	78.0 %	366	7	65 for > 1	0 for >1	poor
ORE: EThcD	233.18	8.64 E+08	82.0 %	434	13	110 ± 62	43 ± 38	poor some c/z
ORE: CID-pd-ETD	365.46 N-glyc[4]:	3.86 E+09 PEP E-01	82.0 %	850	31	77 ± 57	27 ± 38	dirty[3], ETD for HexNAc(1)
ORE: CID-pd-EThcD	357.66 N-glyc[4]:	3.02 E+09 PEP E-01	82.0 %	828	26	60 ± 48	11 ± 10	dirty, no EThcD
Fetuin – Mixed *N*-glycosylation and *O*-glycosylation								
QEx: HCD Standard 30 [2]	1294.32 N-glyc[4]: Both[5]: O-glyc[6]:	5.50 E+11 PEP E-6^9 PEP E-6^7 E-3^18	100 %	2796	499 189 65 310	384 ± 210 445 ± 142 284 ± 126 348 ±234	17 ± 29 31 ± 20 6 ± 7 9 ± 30	excellent b/y no non-red Y very good *O*-glycan loss
QEx: SCE-HCD 10-15	420.98 N-glyc[4]: Both[5]: O-glyc[6]:	1.77 E+10 PEP E-2 PEP E-3^5 PEP E-3^5	91.1 %	433	32 all[7] 3 -N 4 both 25 -O	242 ±116 83 ± 37 176 ± 56 276 ± 110	7.8 ± 26 15 ± 17 44 ± 59 0 ± 2.5	Y red end, ~~b/y~~ good Yred ~~b/y~~ excel. Y, ~~b/y~~ bipolar[8] poor Y, b/y| great Y
QEx: SCE-HCD 10-20	600.79 N-glyc[4]: Both[5]: O-glyc[6]:	1.84 E+10 PEP E-3^6 PEP E-3^6 PEP E-3^9	99.7 %	644	56 all[7] 10 5 41	362 ± 194 173 ± 46 192 ±117 429 ±180	15 ± 32 30 ± 18 35 ± 56 9.0 ± 27	Y red end, b/y excel. Y, (b/y) excel. Y, (b/y) bipolar[8] poor Y, b/y| great Y

[1] Range of 95% of PEP values (manual). See appendix for list of PSM with PEP values.

[2] Data aquired on QEx03 while all SCE-HCD was aquired on QEx02. Same methods/parameters. NCE values are tuned across the range of Thermo instruments.

[3] Dirty = noisy or many non-assigned fragmentations or peak forest. This means that these PSM have a very high chance to be ‚untrue', depite showing a high Byonic Score.

[4] Values in this row are for *N*-glycans exclusively (no mixed N/O glycans)

[5] Values in this row are exclusively for peptides containing both types of glycosylations

[6] Values in this row are for *O*-Glycans exclusively (no mixed N/O glycans)

[7] All PSM for glycosylated peptides (all PSM)

[8] Individual O-glycosylations are well resolved and fragmented, but patches/ O-Glycan clusters not and include many ‚poor' or ‚false' identifications upon manual analysis.

(Table continued on next page.)

Table 3.3. (***continued***)

QEx: SCE-HCD 10-30	568.74	1.25 E+10	99.7 %	798	95 all[f]	246 ± 148	22 ± 38	Y red end; b/y
	N-glyc[d]:	PEP E-3^5			32	229 ± 102	42 ± 42	excel. Y; b/y
	Both[e]:	PEP E-2^3			11	129 ± 106	28 ± 54	excel. Y; b/y
	O-glyc[6]:	PEP E-3^9			52	283 ± 163	8.4 ± 23	bipolar[g] poor Y, b/y| great Y
QEx: SCE-HCD 15-30	719.84	1.31 E+10	99.7 %	796	106 all	307 ± 196	17 ± 26	huge oxonium little Y; b/y
	N-glyc[d]:	PEP E-3^6			36	215 ± 118	38 ± 23	Oxonium > Y
	Both[e]:	PEP E-3^5			8	130 ± 86	5.1 ± 4.8	Y red end; b/y
	O-glyc[6]:	PEP E-3^9			62	383 ± 204	6.6 ± 23	glyc loss or Y_i
QEx: SCE-HCD 10-15-20	563.20	2.19 E+10	92.2 %	680	51 all[f]	327 ± 197	13 ± 23	poor b/y; Y
	N-glyc[d]:	PEP E-3^5			5	159 ± 70	48 ± 32	excel Y, ~~b/y~~
	Both[e]:	PEP E-3^6			8	149 ± 100	25 ± 21	excel Y, b/y
	O-glyc[6]:	PEP E-3^8			38	386 ± 189	5.9 ± 16	dirty[a] glyc loss
QEx: SCE-HCD 10-15-30	581.92	1.65 E+10	99.7 %	659	59 all[f]	286 ± 166	21 ± 37	poor b/y; Y
	N-glyc[d]:	PEP E-3^6			15	213 ± 99	52 ± 50	excel Y, ~~b/y~~
	Both[e]:	PEP E-3^6			8	172 ± 53	27 ± 26	excel Y, b/y
	O-glyc[6]:	PEP E-3^6			36	341 ± 43	6.9 ± 20	dirty, glyc loss
QEx: SCE-HCD 10-20-30	671.86	1.23 E+10	99.7 %	617	66 all[f]	342 ± 200	22 ± 34	poor b/y; Y
	N-glyc[d]:	PEP E-3^7			22	231 ± 81	40 ± 23	excel Y, ~~b/y~~
	Both[e]:	PEP E-3			1	148	148	decent Y, ~~b/y~~
	O-glyc[6]:	PEP E-3^9			43	403 ± 217	10 ± 30	dirty, glyc loss
QEx: SCE-HCD 20-30-40	760.45	8.06 E+09	100 %	725	106 all	300 ± 187	22 ± 40	good b/y, Y
	N-glyc[d]:	PEP E-3^7			37	252 ± 112	42 ± 50	excel Y & b/y
	Both[e]:	PEP E-3^6			12	178 ± 104	11 ± 9	excel Y & b/y
	O-glyc[6]:	PEP E-3^8			57	359 ± 215	11 ± 30	dirty, glyc loss
ORE: HCD-pd-ETD	504.26	1.65 E+10	60.4 %	722	2 all[f]			
	N-glyc[d]:	0			0			
	Both[e]:	0			0	608 -HCD	69 – HCD	good b/y ; Y1
	O-glyc[6]:	PEP E2^10			1	62 - ETD	62 - ETD	good c/z, M
ORE: EThcD	File could not be searched. Unclear why							
ORE: CID-pd-ETD	590.28	3.50 E+10	100 %	1231	32	302 ± 72	18 ± 35	noisy, dirty
	N-glyc[d]:	PEP E-1^4			12	275 ± 87	28 ± 41	assigns many
	Both[e]:	PEP E-2^5			7	304 ± 51	8 ± 8	small peaks to
	O-glyc[6]:	PEP E-2^6			13	324 ± 58	14 ± 36	extensive frag

As expected, a drop in score occurs for regular peptides when using stepped methods compared to a standard HCD of 30. Standard HCD methods have more peptide spectrum matches even for glycopeptides, but this number is meaningless as many of these are incorrect. The SCE-HCD 10-15 method does not seem to be appropriate to analyse anything other than glycopeptides, as values are low in most scoring metrics. SCE-HCD 10-20 reaches the highest Byonic score for O-glycopeptides (on fetuin), followed by SCE-HCD 10-20-30 and 10-15-20. Only ETD-based methods reached good Delta Mod scores for *O*-glycopeptides, which are indicative of site localisation.

Most methods that were deemed very suited for *N*-glycopeptides performed less well on *O*-glycopeptides (SCE-HCD 20-30-40). Methods such as SCE-HCD 10-30 or 15-30 lead to very good results on *O*-glycopeptides, these did not perform as well on *N*-glycopeptides. As for previous sections, a more thorough evaluation on a larger pool of glycoproteins with a more elaborate statistical data visualisation than mean and standard deviation are required (such as violin/box plot). Nonetheless, these preliminary results seem to suggest that there might not be a 'one-size-fits-all' approach for both *N*- and *O*-glycoproteomics. This is further exarcerbated by the issue of *O*-glycan site

localisation, as will be described in the following section.

3.5.3. ETD and EThcD methods remain essential for glycan site localisation

3.5.3.1. Implementing EThcD on an Orbitrap Elite instrument for glycopeptide analysis

Upon the introduction of the EThcD fragmentation technique on Thermo's Tribrid range of instruments such as the Fusion Lumos, several groups have shown the potential of EThcD for glycopeptide fragmentation[669] in application studies on single glycoproteins[670,697] as well as complex glycosylated samples[670], showing its utility in high-throughput *N*-glycoproteomics[670] and particularly in the more challenging *O*-glycoproteomics[671], as well as for sialic acid terminal and phosphorylated glycans[669] or the *S*-GlcNAc modification[63]. Noteworthy for proteoglycans, EThcD has the ability to distinguish hydroxyproline isomers.[698]

For the purpose of this study, EThcD was implemented on an Orbitrap Elite instument. EThcD is natively integrated in Thermo's tribrid instrument series, not available to our laboratory. Therefore, the implementation described in the original publications by the Heck group was reproduced.[306] The conceptual implementation is described in 1.3.4.4. Briefly, In this setup on the Elite, EThcD is hard-coded as a HCD2 scan with first mass 151. Only with this setting is a MS^3 step in the HCD cell without precursor isolation possible i.e. defined. After precursor isolation upon MS1 in the High Pressure Cell (part of the ion trap), ETD fragmentation is conducted by electron transfer from the fluoranthene anion in the Low Pressure Cell (part of the ion trap). Fragment ions are transferred – without detection – to the HCD collision cell for further fragmentation before transfer to the Orbitrap for Fourier-transformed detection.

Performance of the method was checked and validated using BSA and *E. coli* standards. A major drop of identifications is observed for pure EThcD runs since the speed of acquisition (on an already aged Elite) is approx. 3 × slower than ETD (which is already much slower than ion-trap CID). Nevertheless, high-quality spectra are obtained, which enables qualitative studies of the use of EThcD on glycopeptides. This "low-tech" implementation should be considered when comparing our experiments to the literature.

Using our EThcD implementation, I could acquire an EThcD spectrum for IgG *N*-glycosylation (Figure 3.26). In this example, EThcD shows information about the non-reducing end of the glycan, it allows to see the large Y fragments, important for correct

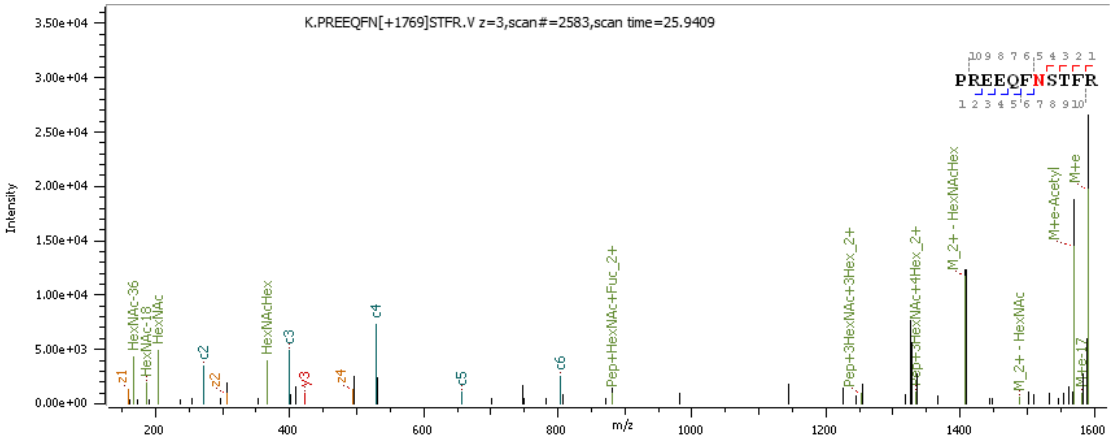


Figure 3.26. EThcD spectrum of IgG N-glycosylation shows fragmentation at the non-reducing terminus of the glycan. Tryptic digest of IgG, pure EThcD method.

sequencing. These are often missed in regular HCD fragmentations series unless SCE's are used.

Surprisingly, identifications of glycans in Byonic (v 2.9) depend on file format, as the data encoding can change the processing. In the case of IgG glycosylation, I tested the outcome or searching RAW (Thermo Fischer), mzML (PSI), mzXML (PSI, old) and mgf (Mascot) file format. Interestingly, more identifications are found in RAW files, with similar results in terms of scores as the mzML and mxXML file format. Mgf files yield less identifications, a trend observed as well on other samples (Figure 3.27). This information is important when doing file operations for manual analysis/annotation.

A note on searching mixed EThcD fragmentation data from the Orbitrap Elite This analysis applies only to mixed EThcD methods, such as for example product triggered CID-pd-EThcD or alternating ("-alt-") methods. Unfortunately, Byonic (v 2.9) does not offer to search specifically for mixed EThcD methods. Searching mixed EThcD methods with the EThcD option in Byonic (v 2.9) does not work either as the search engine is not able to find EThcD spectra and gives an error message, unless .mgf files are used. Interestingly, pure EThcD runs can be searched (see previous section). To find the best search parameters, a series of test analysis were performed on the same data file using the mixed-mode fragmentations pre-defined by Byonic (CID&ETD, CID&HCD, HCD&ETD – see Table 3.4)

Several tricks were attempted to bypass these limitations. Editing the mzML files to the correct scan header by replacing 'HCD' by 'ETD+SA' was attempted, thus removing

(a) RAW

Pos.	Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	z	Obs. m/z
168	K.[+43.00581]TKPREEQFN[+1647.6...	HexNAc(5)Hex(3)Fuc(1)	NTerm(Car...	178.5	17.1	7.24e-03	3	1110.8148
170	K.PREEQFN[+1647.61324]STFR.V	HexNAc(5)Hex(3)Fuc(1)	N7(NGlyca...	74.9	69.7	7.63e-03	3	1020.1013
170	K.PREEQFN[+1768.63952]STFR.V	HexNAc(4)Hex(5)Fuc(1)	N7(NGlyca...	228.2	105.0	7.53e-05	3	1060.4438
170	K.PREEQFN[+1606.58669]STFR.V	HexNAc(4)Hex(4)Fuc(1)	N7(NGlyca...	205.3	107.7	4.60e-04	3	1006.4253
168	K.[+43.00581]TKPREEQFN[+1606.5...	HexNAc(4)Hex(4)Fuc(1)	NTerm(Car...	114.1	17.1	8.24e-03	3	1097.1418
168	K.[+43.00581]TKPREEQFN[+1444.5...	HexNAc(4)Hex(3)Fuc(1)	NTerm(Car...	205.9	17.1	2.14e-03	3	1043.1243
170	K.PREEQFN[+1444.53387]STFR.V	HexNAc(4)Hex(3)Fuc(1)	N7(NGlyca...	89.7	89.7	7.87e-03	3	952.4073

(b) mzML

Pos.	Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	z	Obs. m/z
168	K.[+43.00581]TKPREEQFN[+1647.61324]STFR.V	HexNAc(5)Hex(3)Fuc(1)	NTerm(Car...	176.7	17.1	1.01e-02	3	1110.8137
170	K.PREEQFN[+1768.63952]STFR.V	HexNAc(4)Hex(5)Fuc(1)	N7(NGlyca...	222.1	105.0	1.97e-04	3	1060.4478
170	K.PREEQFN[+1606.58669]STFR.V	HexNAc(4)Hex(4)Fuc(1)	N7(NGlyca...	204.2	111.0	4.18e-04	3	1006.4246
168	K.[+43.00581]TKPREEQFN[+1444.53387]STFR.V	HexNAc(4)Hex(3)Fuc(1)	NTerm(Car...	217.7	17.1	4.08e-03	3	1043.1210
170	K.PREEQFN[+1444.53387]STFR.V	HexNAc(4)Hex(3)Fuc(1)	N7(NGlyca...	85.9	85.9	1.02e-02	3	952.4115

(c) mzXML

Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	z	Obs. m/z
K.[+43.00581]TKPREEQFN[+1647.6...	HexNAc(5)Hex(3)Fuc(1)	NTerm(Car...	176.7	17.1	1.01e-02	3	1110.8137
K.PREEQFN[+1768.63952]STFR.V	HexNAc(4)Hex(5)Fuc(1)	N7(NGlyca...	222.1	105.0	1.97e-04	3	1060.4478
K.PREEQFN[+1606.58669]STFR.V	HexNAc(4)Hex(4)Fuc(1)	N7(NGlyca...	204.2	111.0	4.18e-04	3	1006.4246
K.[+43.00581]TKPREEQFN[+1444.5...	HexNAc(4)Hex(3)Fuc(1)	NTerm(Car...	217.7	17.1	4.08e-03	3	1043.1210
K.PREEQFN[+1444.53387]STFR.V	HexNAc(4)Hex(3)Fuc(1)	N7(NGlyca...	85.9	85.9	1.02e-02	3	952.4115

(d) mgf

Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	z	Obs. m/z
K.PREEQFN[+1768.63952]STFR.V	HexNAc(4)Hex(5)Fuc(1)	N7(NGlyca...	228.0	105.0	2.22e-04	3	1060.4438
K.PREEQFN[+1606.58669]STFR.V	HexNAc(4)Hex(4)Fuc(1)	N7(NGlyca...	205.8	108.7	9.94e-04	3	1006.4257
K.[+43.00581]TKPREEQFN[+1444.5...	HexNAc(4)Hex(3)Fuc(1)	NTerm(Car...	205.7	32.3	3.94e-03	3	1043.1241
K.PREEQFN[+1444.53387]STFR.V	HexNAc(4)Hex(3)Fuc(1)	N7(NGlyca...	84.8	84.8	1.00e-02	3	952.4077

Figure 3.27. Effects of input file format on EThcD-based glycan identifications in Byonic. Tryptic digest of IgG, pure EThcD method.

the trace of the EThcD implementation and having the appropriate description as on Orbitrap Fusion Tribrid type instruments – without success. As such, I found that the best method to process such data (i.e. CID-alt-EThcD) is by using CID-HCD mixed search mode, but this comes at the cost of loosing the c/z annotations. Yet these can be added by manual analysis or using extensive annotation modes (such as in Peaks or MaxQuant).

3.5.3.2. Analysis of the model glycopeptides by ETD and EThcD

On all glycopeptides described in this study, ETD and EThcD methods were tried, leading to spectra for some peptides (Figure 3.28). In our engineered approach, EThcD methods did not yield convincing *O*-glycopeptide identifications in Byonic as these can-

Table 3.4. Finding appropriate search parameters for mixed CID-alt-EThcD fragmentation on ‚hacked' Orbitrap Elite. The same file, a tryptic digest of decorin run with CID-alt-EThcD was searched with different parameters.

Fragmentation type → ↓ File format	EThcD (20 ppm)	CID_ETD (0.5 Da_0.5 Da)	CID_HCD (0.5 Da_20 ppm)	ETD_HCD
.RAW	no_spectra_ to_search	# PSM = 547 no c/z annotation	# PSM = 997 no c/z annotation	# PSM = 420 ~~c/z~~
.mzML	no_spectra_ to_search	# PSM = 559 no c/z annotation	# PSM = 1012 no c/z annotation	# PSM = 420 ~~c/z~~
.mzXML	no_spectra_ to_search	# PSM = 559 no c/z annotation	# PSM = 1012 no c/z annotation	# PSM = 420 ~~c/z~~
.mgf	# PSM = 345 has c/z annotation	mixed mode not supported	mixed mode not supported	mode not supported
edited .mzML	no_spectra_ to_search	# PSM = 559 no c/z annotation	# PSM = 1012 no c/z annotation	# PSM = 420 ~~c/z~~
edited .mzXML	no_spectra_ to_search	# PSM = 604 no c/z annotation	# PSM = 559 no c/z annotation	# PSM = 49 ~~c/z~~

not be searched adequately. However, HCD-alt-EThcD methods were useful as this yields peptide identification by HCD, giving appropriate scan number for manual extraction of the EThcD spectrum (scan number + 1) and separate annotation from the RAW file using tools such as the Iterative Peptide Spectrum Annotator. However, this trick was discovered only retrospectively and many samples were not run with this method.

Nonetheless, the ETD spectra I obtained confirm the strength of the method for site-localisation of the glycan. The poor fragmentation efficiency and high signal-to-noise ratio is an issue with ETD, which could be partially due to the fact that these experiments were performed on a rather old Orbitrap Elite, partially to the low charge density of the peptides.

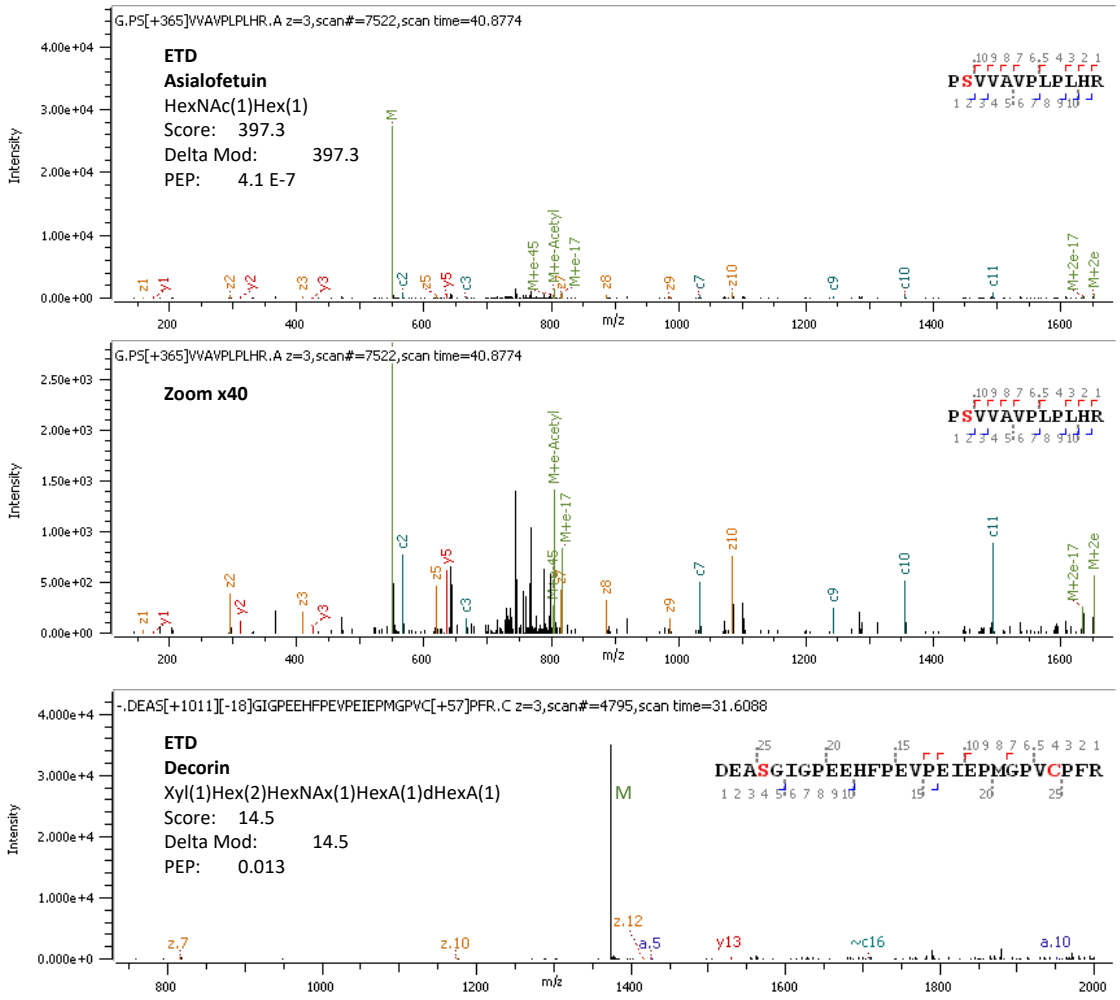


Figure 3.28. Examples of ETD fragmentation of *O*-glycopeptides presented in this chapter. Spectra acquired on an Orbitrap Elite.

3.6. Conclusion and outlook

In this chapter, I have attempted to introduce an energetic perspective into *O*-glycopeptide fragmentation and illustrated that the energetics differ from *N*-glycopeptides. Complementing scarcely distributed evidence in the literature, compiled in the introduction, our data shows that low-energy stepped-HCD methods such as 10-20% normalised collision energy (NCE) are advantageous for *O*-glycopeptide characterisation in terms of glycan sequencing (including branching and composition) and glycan attachment site localisation. Note that these values are much lower than those analysed by Riley *et al.*[652] In the case of decorin, a diagnostic ${}^{0,2}X_0$ cross-ring fragmentation observable at these low energy establishes phosphorylation at xylose instead of sulfation on the neighbouring galactose. Although all the data presented is derived from tryptic peptides, my observation on chymotryptic digests (data not shown) as well as reports on pronase digests[689] suggest this might be a generalisable principle.

In order for *O*-glycoproteomics to mature into a more widely applicable investigative technology, three key challenges in MS/MS data acquisition need solving:

- Identification of glycan attachment site / site localisation. SCE-HCD methods have enabled site-localisation of *singly* glycosylated peptides with limited Ser/-Thr attachment sites. However, more complex cases such as multiply glycosylated patches might still need support from ETD/EThcD. That being said, SCE-HCD methods can take advantage of the *proline effect*, and thus might potentially fragment among these highly glycosylated sites − this needs further investigation.
- Identification of glycan composition. Similarly as ion-trap CID methods, SCE-HCD shows extensive glycan fragmentation without loosing *b*- and *B*-ions due to the "1/3 rule". The composition is therefore straightforward to identify.
- Identification of glycan sequence and connectivity. The interpretation of glycan fragmentation pattern enables to identify branching and sequence. However, the connectivity (α1,3 vs α1,4 vs α1,6) is difficult to access without further MS^3 fragmentation or ion mobility.

Taken together, one could therefore think of combinations of SCE-HCD alternating with ETD or EThcD, perhaps depending on a decision-tree to remediate their respective disadvantage? HCD-pd-ETD methods have proven their worth for complex mixtures, therefore elaborate acquisition methods such as HCD_{30}-pd-[SCE-$HCD_{10\text{-}20}$-alt-ETD/$EThcD_{15}$-DDT] or other imaginative combinations could potentially bring the different advantages together.

In addition, two key challenges in data analysis are currently limiting progress in method development:

- Annotation of all fragment ions, including Yy and Yb ions. The use of GlycReSoft's glycopeptide hypothesis generation is useful for manual analysis. This idea could potentially be automatised and incorporated into currently existing algorithms.
- Scoring principles. Arguably, in *O*-glycoproteomics, site-localising ions, diagnostic cross-ring fragmentation or B/Y-ion formation should have a high value in the PSM's scoring. Currently, these are dominated by the peptide part. Better scoring might also help the development of better FDR methods.

Further, *O*-glycopeptide enrichment and the best possible digestion method are equally unsolved questions. Pronase digests lead to substantial search spaces, which might not be practicable for large-scale studies. The combination of mucin-type specific proteases such as StcE or OpeRATOR with trypsin, or double digests combining trypsin and another protease such as GluC or AspN more likely to cleave in the Lys/Arg-poor sequences of *O*-glycopeptides could be another avenue to explore. The balancing of cleavage specificity − the less missed cleavage, the better − with the combinatorial expansion of *O*-glycopeptide search space might however need a more extensive fine-tuning and limit multiple digestions. The enrichment of *O*-glycopeptides itself is more difficult than *N*-glycopeptides as these have a much higher variability and might not be equally well amenable to ZIC-HILIC enrichment.

Taken together, the preliminary findings presented here call for a larger study on the use of low-energy SCE-HCD methods for *O*-glycoproteomics, including their combination with other methods such as ETD and EThcD for site-localisation and matching protocols for sample preparation and data analysis.

3.7. Author contributions

Purified and trimmed decorin samples were kindly provided by Alexandra Gafitescu. Simon Nadal designed and collected data for all other research presented in this chapter.

3.8. References

1. *Essentials of Glycobiology* 3rd (eds Varki, A. *et al.*) (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2015).

51. Darula, Z. & Medzihradszky, K. F. Analysis of Mammalian O-Glycopeptides—We Have Made a Good Start, but There Is a Long Way to Go. *Molecular & Cellular Proteomics* **17,** 2–17 (2018).

63. Maynard, J. C., Burlingame, A. L. & Medzihradszky, K. F. Cysteine S-Linked N-Acetylglucosamine (S-GlcNAcylation), A New Post-Translational Modification in Mammals. *Molecular & Cellular Proteomics* **15,** 3405–3411 (2016).

114. Thaysen-Andersen, M. & Packer, N. H. Advances in LC–MS/MS-Based Glycoproteomics: Getting Closer to System-Wide Site-Specific Mapping of the N- and O-Glycoproteome. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1844,** 1437–1452 (2014).

115. Thaysen-Andersen, M., Packer, N. H. & Schulz, B. L. Maturing Glycoproteomics Technologies Provide Unique Structural Insights into the N-Glycoproteome and Its Regulation in Health and Disease. *Molecular & Cellular Proteomics* **15,** 1773–1790 (2016).

172. Stavenhagen, K., Hinneburg, H., Thaysen-Andersen, M., Hartmann, L., Silva, D. V., Fuchser, J., Kaspar, S., Rapp, E., Seeberger, P. H. & Kolarich, D. Quantitative Mapping of Glycoprotein Micro-Heterogeneity and Macro-Heterogeneity: An Evaluation of Mass Spectrometry Signal Strengths Using Synthetic Peptides and Glycopeptides. *Journal of Mass Spectrometry* **48,** 627–639 (2013).

180. Alley, W. R., Mann, B. F. & Novotny, M. V. High-Sensitivity Analytical Approaches for the Structural Characterization of Glycoproteins. *Chemical Reviews* **113,** 2668–2732 (2013).

218. Harvey, D. J. Fragmentation of Negative Ions from Carbohydrates: Part 3. Fragmentation of Hybrid and Complex N-Linked Glycans. *Journal of the American Society for Mass Spectrometry* **16,** 647–659 (2005).

219. Harvey, D. J. in *Analysis of Protein Post-Translational Modifications by Mass Spectrometry* (eds Griffiths, J. R. & Unwin, R. D.) 89–159 (John Wiley & Sons, Inc., 2016).

220. Nwosu, C. C., Strum, J. S., An, H. J. & Lebrilla, C. B. Enhanced Detection and Identification of Glycopeptides in Negative Ion Mode Mass Spectrometry. *Analytical Chemistry* **82,** 9654–9662 (2010).

237. Roepstorff, P. & Fohlman, J. Letter to the Editors. *Biomedical Mass Spectrometry* **11,** 601–601 (1984).

238. Biemann, K. Contributions of Mass Spectrometry to Peptide and Protein Structure. *Biomedical & Environmental Mass Spectrometry* **16,** 99–111 (1988).

239. Johnson, R. S., Martin, S. A., Biemann, K., Stults, J. T. & Watson, J. T. Novel Fragmentation Process of Peptides by Collision-Induced Decomposition in a Tandem Mass Spectrometer: Differentiation of Leucine and Isoleucine. *Analytical Chemistry* **59,** 2621–2625 (1987).

255. Wysocki, V. H., Tsaprailis, G., Smith, L. L. & Breci, L. A. Mobile and Localized Protons: A Framework for Understanding Peptide Dissociation. *Journal of Mass Spectrometry* **35,** 1399–1406 (2000).

256. Dongré, A. R., Jones, J. L., Somogyi, Á. & Wysocki, V. H. Influence of Peptide Composition, Gas-Phase Basicity, and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model. *Journal of the American Chemical Society* **118,** 8365–8374 (1996).

257. Jones, J. L., Dongre, A. R., Somogyi, A. & Wysocki, V. H. Sequence Dependence of Peptide Fragmentation Efficiency Curves Determined by Electrospray Ionization/Surface-Induced Dissociation Mass Spectrometry. *Journal of the American Chemical Society* **116,** 8368–8369 (1994).

267. Yang, Y.-H., Lee, K., Jang, K.-S., Kim, Y.-G., Park, S.-H., Lee, C.-S. & Kim, B.-G. Low Mass Cutoff Evasion with Qz Value Optimization in Ion Trap. *Analytical Biochemistry* **387,** 133–135 (2009).

272. Frese, C. K., Altelaar, A. F. M., Hennrich, M. L., Nolting, D., Zeller, M., Griep-Raming, J., Heck, A. J. R. & Mohammed, S. Improved Peptide Identification by Targeted Fragmentation Using CID, HCD and ETD on an LTQ-Orbitrap Velos. *Journal of Proteome Research* **10,** 2377–2388 (2011).

273. Scientific, T. F. PSB 104 - Normalized Collision Energy Technology, 2.

275. Hinneburg, H., Stavenhagen, K., Schweiger-Hufnagel, U., Pengelley, S., Jabs, W., Seeberger, P. H., Silva, D. V., Wuhrer, M. & Kolarich, D. The Art of Destruction: Optimizing Collision Energies in Quadrupole-Time of Flight (Q-TOF) Instruments for Glycopeptide-Based Glycoproteomics. *Journal of The American Society for Mass Spectrometry* **27,** 507–519 (2016).

276. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and Protein Sequence Analysis by Electron Transfer Dissociation Mass Spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 9528–9533 (2004).

277. Riley, N. M. & Coon, J. J. The Role of Electron Transfer Dissociation in Modern Proteomics. *Analytical Chemistry* **90,** 40–64 (2018).

278. Martens, J., Berden, G. & Oomens, J. Structures of Fluoranthene Reagent Anions Used in Electron Transfer Dissociation and Proton Transfer Reaction Tandem Mass Spectrometry. *Analytical Chemistry* **88,** 6126–6129 (2016).

279. Zhurov, K. O., Fornelli, L., Wodrich, M. D., Laskay, Ü. A. & Tsybin, Y. O. Principles of Electron Capture and Transfer Dissociation Mass Spectrometry Applied to Peptide and Protein Structure Analysis. *Chemical Society Reviews* **42,** 5014–5030 (2013).

289. Świerszcz, I., Skurski, P. & Simons, J. Dipole and Coulomb Forces in Electron Capture Dissociation and Electron Transfer Dissociation Mass Spectroscopy. *The Journal of Physical Chemistry A* **116,** 1828–1837 (2012).

291. Swaney, D. L., McAlister, G. C., Wirtala, M., Schwartz, J. C., Syka, J. E. P. & Coon, J. J. Supplemental Activation Method for High-Efficiency Electron-Transfer Dissociation of Doubly Protonated Peptide Precursors. *Analytical Chemistry* **79,** 477–485 (2007).

305. 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. *International Journal of Proteomics* **2012,** e560391 (2012).

306. Frese, C. K., Altelaar, A. F. M., van den Toorn, H., Nolting, D., Griep-Raming, J., Heck, A. J. R. & Mohammed, S. Toward Full Peptide Sequence Coverage by Dual Fragmentation Combining Electron-Transfer and Higher-Energy Collision Dissociation Tandem Mass Spectrometry. *Analytical Chemistry* **84,** 9668–9673 (2012).

308. Yu, J., Schorlemer, M., Gomez Toledo, A., Pett, C., Sihlbom, C., Larson, G., Westerlind, U. & Nilsson, J. Distinctive MS/MS Fragmentation Pathways of Glycopeptide-Generated Oxonium Ions Provide Evidence of the Glycan Structure. *Chemistry – A European Journal* **22,** 1114–1124 (2016).

326. Gray, C. J. *et al.* Bottom-Up Elucidation of Glycosidic Bond Stereochemistry. *Analytical Chemistry* **89,** 4540–4549 (2017).

332. Mucha, E., Stuckmann, A., Marianski, M., Struwe, W. B., Meijer, G. & Pagel, K. In-Depth Structural Analysis of Glycans in the Gas Phase. *Chemical Science* **10,** 1272–1284 (2019).

342. Bern, M., Kil, Y. J. & Becker, C. Byonic: Advanced Peptide and Protein Identification Software. *Current Protocols in Bioinformatics* **40,** 13.20.1–13.20.14 (2012).

343. Liu, M.-Q. *et al.* pGlyco 2.0 Enables Precision N-Glycoproteomics with Comprehensive Quality Control and One-Step Mass Spectrometry for Intact Glycopeptide Identification. *Nature Communications* **8,** 1–14 (2017).

344. Klein, J., Carvalho, L. & Zaia, J. Application of Network Smoothing to Glycan LC-MS Profiling. *Bioinformatics* **34,** 3511–3518 (2018).

440. Shu, Q. *et al.* Large-Scale Identification of N-Linked Intact Glycopeptides in Human Serum Using HILIC Enrichment and Spectral Library Search. *Molecular & Cellular Proteomics* **19,** 672–689 (2020).

441. Fang, P. *et al.* Multilayered N-Glycoproteome Profiling Reveals Highly Heterogeneous and Dysregulated Protein N-Glycosylation Related to Alzheimer's Disease. *Analytical Chemistry* **92,** 867–874 (2020).

442. Aguilar, H. A., Iliuk, A. B., Chen, I.-H. & Tao, W. A. Sequential Phosphoproteomics and N-Glycoproteomics of Plasma-Derived Extracellular Vesicles. *Nature Protocols* **15,** 161–180 (1 2020).

552. Kolarich, D., Jensen, P. H., Altmann, F. & Packer, N. H. Determination of Site-Specific Glycan Heterogeneity on Glycoproteins. *Nature Protocols* **7,** 1285–1298 (2012).

578. Singh, C., Zampronio, C. G., Creese, A. J. & Cooper, H. J. Higher Energy Collision Dissociation (HCD) Product Ion-Triggered Electron Transfer Dissociation (ETD) Mass Spectrometry for the Analysis of N-Linked Glycoproteins. *Journal of Proteome Research* **11,** 4517–4525 (2012).

633. Harvey, D. J. Electrospray Mass Spectrometry and Fragmentation of N-Linked Carbohydrates Derivatized at the Reducing Terminus. *Journal of the American Society for Mass Spectrometry* **11,** 900–915 (2000).

634. An, H. J. & Lebrilla, C. B. Structure Elucidation of Native N- and O-Linked Glycans by Tandem Mass Spectrometry (Tutorial). *Mass Spectrometry Reviews* **30,** 560–578 (2011).

635. Zaia Joseph. Mass Spectrometry of Oligosaccharides. *Mass Spectrometry Reviews* **23,** 161–227 (2004).

636. Sassaki, G. L. & de Souza, L. M. Mass Spectrometry Strategies for Structural Analysis of Carbohydrates and Glycoconjugates. *Tandem Mass Spectrometry - Molecular Characterization* (2013).

637. Annesley, T. M. Ion Suppression in Mass Spectrometry. *Clinical Chemistry* **49,** 1041–1044 (2003).

638. Mechref, Y. Use of CID/ETD Mass Spectrometry to Analyze Glycopeptides. *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]* **0 12,** Unit–12.1111 (2012).

639. Lee, L. Y., Moh, E. S. X., Parker, B. L., Bern, M., Packer, N. H. & Thaysen-Andersen, M. Toward Automated N-Glycopeptide Identification in Glycoproteomics. *Journal of Proteome Research* **15,** 3904–3915 (2016).

640. Stadlmann, J., Pabst, M., Kolarich, D., Kunert, R. & Altmann, F. Analysis of Immunoglobulin Glycosylation by LC-ESI-MS of Glycopeptides and Oligosaccharides. *PROTEOMICS* **8,** 2858–2871 (2008).

641. Domon, B. & Costello, C. E. A Systematic Nomenclature for Carbohydrate Fragmentations in FAB-MS/MS Spectra of Glycoconjugates. *Glycoconjugate Journal* **5,** 397–409 (1988).

642. Dodds, E. D. Gas-Phase Dissociation of Glycosylated Peptide Ions. *Mass Spectrometry Reviews* **31,** 666–682 (2012).

643. Dell, A. in *Advances in Carbohydrate Chemistry and Biochemistry* (eds Tipson, R. S. & Horton, D.) 19–72 (Academic Press, 1987).

644. Baldauf, C. & Rossi, M. Going Clean: Structure and Dynamics of Peptides in the Gas Phase and Paths to Solvation. *Journal of Physics: Condensed Matter* **27,** 493002 (2015).

645. Creese, A. J. & Cooper, H. J. Separation and Identification of Isomeric Glycopeptides by High Field Asymmetric Waveform Ion Mobility Spectrometry. *Analytical Chemistry* **84,** 2597–2601 (2012).

646. Medzihradszky, K. F., Gillece-Castroaff, B. L., Townsendaff, R. R., Burlingameaff, A. L. & Hardyaff, M. R. Structural Elucidation of O-Linked Glycopeptides by High Energy Collision-Induced Dissociation. *Journal of the American Society for Mass Spectrometry* **7,** 319–328 (1996).

647. Cocinero, E. J., Stanca-Kaposta, E. C., Gamblin, D. P., Davis, B. G. & Simons, J. P. Peptide Secondary Structures in the Gas Phase: Consensus Motif of N-Linked Glycoproteins. *Journal of the American Chemical Society* **131,** 1282–1287 (2009).

648. Sanda, M., Benicky, J. & Goldman, R. Low Collision Energy Fragmentation in Structure-Specific Glycoproteomics Analysis. *Analytical Chemistry* **92,** 8262–8267 (2020).

649. Ács, A., Ozohanics, O., Vékey, K., Drahos, L. & Turiák, L. Distinguishing Core and Antenna Fucosylated Glycopeptides Based on Low-Energy Tandem Mass Spectra. *Analytical Chemistry* **90,** 12776–12782 (2018).

650. Medzihradszky, K. F., Kaasik, K. & Chalkley, R. J. Characterizing Sialic Acid Variants at the Glycopeptide Level. *Analytical Chemistry* **87,** 3064–3071 (2015).

651. Halim, A., Westerlind, U., Pett, C., Schorlemer, M., Rüetschi, U., Brinkmalm, G., Sihlbom, C., Lengqvist, J., Larson, G. & Nilsson, J. Assignment of Saccharide Identities through Analysis of Oxonium Ion Fragmentation Profiles in LC–MS/MS of Glycopeptides. *Journal of Proteome Research* **13,** 6024–6032 (2014).

652. Riley, N. M., Malaker, S. A., Driessen, M. D. & Bertozzi, C. R. Optimal Dissociation Methods Differ for N- and O-Glycopeptides. *Journal of Proteome Research* (2020).

653. Pett, C., Nasir, W., Sihlbom, C., Olsson, B.-M., Caixeta, V., Schorlemer, M., Zahedi, R. P., Larson, G., Nilsson, J. & Westerlind, U. Effective Assignment of A2,3/A2,6-Sialic Acid Isomers by LC-MS/MS-Based Glycoproteomics. *Angewandte Chemie International Edition* **57,** 9320–9324 (2018).

654. Lee, H. K. *et al.* Selective Identification of α-Galactosyl Epitopes in N-Glycoproteins Using Characteristic Fragment Ions from Higher-Energy Collisional Dissociation. *Analytical Chemistry* **92,** 13144–13154 (2020).

655. Thaysen-Andersen, M., Wilkinson, B. L., Payne, R. J. & Packer, N. H. Site-Specific Characterisation of Densely O-Glycosylated Mucin-Type Peptides Using Electron Transfer Dissociation ESI-MS/MS. *ELECTROPHORESIS* **32,** 3536–3545 (2011).

656. Riley, N. M., Hebert, A. S., Westphall, M. S. & Coon, J. J. Capturing Site-Specific Heterogeneity with Large-Scale N-Glycoproteome Analysis. *Nature Communications* **10,** 1311 (1 2019).

657. Wu, S.-L., Hühmer, A. F. R., Hao, Z. & Karger, B. L. On-Line LC-MS Approach Combining Collision-Induced Dissociation (CID), Electron-Transfer Dissociation (ETD), and CID of an Isolated Charge-Reduced Species for the Trace-Level Characterization of Proteins with Post-Translational Modifications. *Journal of Proteome Research* **6,** 4230–4244 (2007).

658. Darula, Z. & Medzihradszky, K. F. Affinity Enrichment and Characterization of Mucin Core-1 Type Glycopeptides from Bovine Serum. *Molecular & Cellular Proteomics* **8,** 2515–2526 (2009).

659. Chan, W. Y. K. & Chan, T. W. D. Natural Structural Motifs That Suppress Peptide Ion Fragmentation after Electron Capture. *Journal of the American Society for Mass Spectrometry* **21,** 1235–1244 (2010).

660. Wong, P.-S. J., Chen, X., Deng, L., Wang, Z., Li, W., Wong, Y. L. E. & Chan, T.-W. D. Suppression of Peptide Ion Dissociation under Electron Capture: Role of Backbone Amide Hydrogen. *Rapid Communications in Mass Spectrometry* **29,** 1757–1764 (2015).

661. Alagesan, K., Hinneburg, H., Seeberger, P. H., Silva, D. V. & Kolarich, D. Glycan Size and Attachment Site Location Affect Electron Transfer Dissociation (ETD) Fragmentation and Automated Glycopeptide Identification. *Glycoconjugate Journal* (2019).

662. Darula, Z. & Medzihradszky, K. F. Glycan Side Reaction May Compromise ETD-Based Glycopeptide Identification. *Journal of The American Society for Mass Spectrometry* **25,** 977–987 (2014).

663. Hongo, Y., Nakamura, T. & Sato, A. Electron Capture Dissociation of Triantennary Complex-Type N-Glycosylated Peptides: A Case of Suppressed Peptide Backbone Cleavage. *Journal of the Mass Spectrometry Society of Japan* **55,** 77–82 (2007).

664. Mormann, M., Paulsen, H. & Peter-Katalinić, J. Electron Capture Dissociation of O-Glycosylated Peptides: Radical Site-Induced Fragmentation of Glycosidic Bonds. *European Journal of Mass Spectrometry,* 497–511 (2005).

665. Catalina, M. I., Koeleman, C. A. M., Deelder, A. M. & Wuhrer, M. Electron Transfer Dissociation of N-Glycopeptides: Loss of the Entire N-Glycosylated Asparagine Side Chain. *Rapid Communications in Mass Spectrometry* **21,** 1053–1061 (2007).

666. Savitski, M. M., Kjeldsen, F., Nielsen, M. L. & Zubarev, R. A. Hydrogen Rearrangement to and from Radical z Fragments in Electron Capture Dissociation of Peptides. *Journal of the American Society for Mass Spectrometry* **18,** 113–120 (2007).

667. Perdivara, I., Petrovich, R., Allinquant, B., Deterding, L. J., Tomer, K. B. & Przybylski, M. Elucidation of O-Glycosylation Structures of the β-Amyloid Precursor Protein by Liquid Chromato\-Graphy-Mass Spectrometry Using Electron Transfer Dissociation and Collision Induced Dissociation. *Journal of Proteome Research* **8,** 631–642 (2009).

668. Stavenhagen, K., Kayili, H. M., Holst, S., Koeleman, C. A. M., Engel, R., Wouters, D., Zeerleder, S., Salih, B. & Wuhrer, M. N- and O-Glycosylation Analysis of Human C1-Inhibitor Reveals Extensive Mucin-Type O-Glycosylation. *Molecular & Cellular Proteomics* **17,** 1225–1238 (2018).

669. Reiding, K. R., Bondt, A., Franc, V. & Heck, A. J. R. The Benefits of Hybrid Fragmentation Methods for Glycoproteomics. *TrAC Trends in Analytical Chemistry* **108,** 260–268 (2018).

670. Yu, Q., Wang, B., Chen, Z., Urabe, G., Glover, M. S., Shi, X., Guo, L.-W., Kent, K. C. & Li, L. Electron-Transfer/Higher-Energy Collision Dissociation (EThcD)-Enabled Intact Glycopeptide/Glycoproteome Characterization. *Journal of The American Society for Mass Spectrometry* **28,** 1751–1764 (2017).

671. Zhang, Y., Xie, X., Zhao, X., Tian, F., Lv, J., Ying, W. & Qian, X. Systems Analysis of Singly and Multiply O-Glycosylated Peptides in the Human Serum Glycoproteome via EThcD and HCD Mass Spectrometry. *Journal of Proteomics* **170,** 14–27 (2018).

672. Marino, F., Bern, M., Mommen, G. P. M., Leney, A. C., van Gaans-van den Brink, J. A. M., Bonvin, A. M. J. J., Becker, C., van Els, C. A. C. M. & Heck, A. J. R. Extended O-GlcNAc on HLA Class-I-Bound Peptides. *Journal of the American Chemical Society* **137,** 10922–10925 (2015).

673. Čaval, T., Zhu, J. & Heck, A. J. Simply Extending the Mass Range in Electron Transfer Higher Energy Collisional Dissociation Increases Confidence in N-Glycopeptide Identification. *Analytical Chemistry* **91,** 10401–10406 (2019).

674. Zhao, P., Viner, R., Teo, C. F., Boons, G.-J., Horn, D. & Wells, L. Combining High-Energy C-Trap Dissociation and Electron Transfer Dissociation for Protein O-GlcNAc Modification Site Assignment. *Journal of Proteome Research* **10,** 4088–4104 (2011).

675. Parker, B. L., Thaysen-Andersen, M., Fazakerley, D. J., Holliday, M., Packer, N. H. & James, D. E. Terminal Galactosylation and Sialylation Switching on Membrane Glycoproteins upon TNF-Alpha-Induced Insulin Resistance in Adipocytes. *Molecular & Cellular Proteomics* **15,** 141–153 (2016).

676. Wu, S.-W., Pu, T.-H., Viner, R. & Khoo, K.-H. Novel LC-MS2 Product Dependent Parallel Data Acquisition Function and Data Analysis Workflow for Sequencing and Identification of Intact Glycopeptides. *Analytical Chemistry* **86,** 5478–5486 (2014).

677. Jebanathirajah, J., Steen, H. & Roepstorff, P. Using Optimized Collision Energies and High Resolution, High Accuracy Fragment Ion Selection to Improve Glycopeptide Detection by Precursor Ion Scanning. *Journal of the American Society for Mass Spectrometry* **14,** 777–784 (2003).

678. Yang, H., Yang, C. & Sun, T. Characterization of Glycopeptides Using a Stepped Higher-Energy C-Trap Dissociation Approach on a Hybrid Quadrupole Orbitrap. *Rapid Communications in Mass Spectrometry* **32,** 1353–1362 (2018).

679. Trinidad, J. C., Schoepfer, R., Burlingame, A. L. & Medzihradszky, K. F. N- and O-Glycosylation in the Murine Synaptosome. *Molecular & Cellular Proteomics* **12,** 3474–3488 (2013).

680. Stadlmann, J., Taubenschmid, J., Wenzel, D., Gattinger, A., Dürnberger, G., Dusberger, F., Elling, U., Mach, L., Mechtler, K. & Penninger, J. M. Comparative Glycoproteomics of Stem Cells Identifies New Players in Ricin Toxicity. *Nature* **549,** 538–542 (2017).

681. Hare, N. J., Lee, L. Y., Loke, I., Britton, W. J., Saunders, B. M. & Thaysen-Andersen, M. Mycobacterium Tuberculosis Infection Manipulates the Glycosylation Machinery and the N-Glycoproteome of Human Macrophages and Their Microparticles. *Journal of Proteome Research* **16,** 247–263 (2017).

682. Vékey, K., Ozohanics, O., Tóth, E., Jekő, A., Révész, Á., Krenyácz, J. & Drahos, L. Fragmentation Characteristics of Glycopeptides. *International Journal of Mass Spectrometry. 80/60 : A Special Issue Honoring Keith R. Jennings and James H. Scrivens* **345-347,** 71–79 (2013).

683. Aboufazeli, F. & D. Dodds, E. Precursor Ion Survival Energies of Protonated N -Glycopeptides and Their Weak Dependencies on High Mannose N -Glycan Composition in Collision-Induced Dissociation. *Analyst* **143,** 4459–4468 (2018).

684. Kolli, V., Roth, H. A., De La Cruz, G., Fernando, G. S. & Dodds, E. D. The Role of Proton Mobility in Determining the Energy-Resolved Vibrational Activation/Dissociation Channels of N-Glycopeptide Ions. *Analytica Chimica Acta* **896,** 85–92 (2015).

685. Kolli, V. & Dodds, E. D. Energy-Resolved Collision-Induced Dissociation Pathways of Model N-Linked Glycopeptides: Implications for Capturing Glycan Connectivity and Peptide Sequence in a Single Experiment. *Analyst* **139,** 2144–2153 (2014).

686. Hoffmann, M., Pioch, M., Pralow, A., Hennig, R., Kottler, R., Reichl, U. & Rapp, E. The Fine Art of Destruction: A Guide to In-Depth Glycoproteomic Analyses—Exploiting the Diagnostic Potential of Fragment Ions. *PROTEOMICS* **18,** 1800282 (2018).

687. Scientific, T. F. *PSB 121 - Enhanced Fragmentation of Small Molecules in a Thermo Scientific LTQ Series Linear Ion Trap Mass Spectrometer Using Stepped Normalised Collision Energy*

688. Zhang, T., Jian, L., Viner, R. & Zabrouskov, V. Improving Fragmentation Efficiency of Tandem Mass Tag Labeled Peptides Using Stepped Higher Energy Collisional Dissociation. *HUPO Poster,* 1 (2009).

689. Kelly, M. I. & Dodds, E. D. Parallel Determination of Polypeptide and Oligosaccharide Connectivities by Energy-Resolved Collison-Induced Dissociation of Protonated O-Glycopeptides Derived from Nonspecific Proteolysis. *Journal of the American Society for Mass Spectrometry* **31,** 624–632 (2020).

690. Nilsson, J. Liquid Chromatography-Tandem Mass Spectrometry-Based Fragmentation Analysis of Glycopeptides. *Glycoconjugate Journal* **33,** 261–272 (2016).

691. UWPR Proteomics Resource. *Packing Capillary Columns and Pre-Columns* `https://proteo%20micsresource.washington.edu/docs/protocols05/Packing_Capillary_Columns.pdf` (2020).

692. Kovalchuk, S. I., Jensen, O. N. & Rogowska-Wrzesinska, A. FlashPack: Fast and Simple Preparation of Ultrahigh-Performance Capillary Columns for LC-MS*. *Molecular & Cellular Proteomics* **18,** 383–390 (2019).

693. Klein, J. A., Meng, L. & Zaia, J. Deep Sequencing of Complex Proteoglycans: A Novel Strategy for High Coverage and Site-Specific Identification of Glycosaminoglycan-Linked Peptides. *Molecular & Cellular Proteomics* **17,** 1578–1590 (2018).

694. Bern, M. Byonic™: O-Linked Glycopeptide Analysis. *Application Note,* 5 (2014).

695. Protein Metrics Inc. *Byonic User Manual v 2.10* 2016.

696. Breci, L. A., Tabb, D. L., Yates, J. R. & Wysocki, V. H. Cleavage N-Terminal to Proline: Analysis of a Database of Peptide Tandem Mass Spectra. *Analytical Chemistry* **75,** 1963–1971 (2003).

697. Bern, M. *et al.* Fragmentation Patterns of EThcD Spectra of Phosphopeptides and Glycopeptides. *Poster,* 1.

698. Ma, F., Sun, R., Tremmel, D. M., Sackett, S. D., Odorico, J. & Li, L. Large-Scale Differentiation and Site Specific Discrimination of Hydroxyproline Isomers by Electron Transfer/Higher-Energy Collision Dissociation (EThcD) Mass Spectrometry. *Analytical Chemistry* **90,** 5857–5864 (2018).

3.9. Appendix

3.9.1. Supplementary discussion

3.9.1.1. Supplementary discussion 1: Collision energy values in the literature and their conversion to Thermo's NCE

Linear relationship between optimal collision energy and m/z

The normalised collision energy relates the absolute collision energy in a linear relationship to the *m/z* of the isolated precursor, as an experimentally accessible proxy for the energy distribution across the degrees of freedom of the analyte.[273] Thermo Scientific implemented the NCE concept to enhance reproducibility of its MS^2 spectra across its range of instruments. The NCE relation to the actual collision energy is define for Q Exactive instruments as:[274]

$$\underset{\text{energy (eV)}}{\text{Absolute}} = \underset{\text{NCE}}{\text{settling}} \times \frac{\text{Isolation center}}{500\text{m/z}} \times \underset{\text{factor}}{\text{charge}} \left(+ \underset{\text{offset}}{\text{instrument}}\right) \tag{3.1}$$

The peptide *m/z* is defined as $[\mathrm{M} + n\mathrm{H}]^{n+}$, with the following charge factor conversion: $2^+/0.9$; $3^+/0.85$; $4^+/0.8$; $\geq 5^+/0.75$. Note that the charge factor are lower for glycopeptides than peptides.[275, Fig. 2]

This relationship was used to convert values to NCE for figures shown in the introduction.

Note that this NCE definition usually does not include the instrument offset, added in brackets. Thermo instruments are manually tweaked to deliver same spectra. It is therefore difficult to convert collision energy values across vendors.

However, collision energy equations from different vendors can be compiled from the literature, mostly from quantitative MRM publications. For the calculation of Bruker values, graphs in PDF or print were digitalised with WebPlotDigitizer (default parameters) and the data fitted by a linear trendline in Excel.

Table 3.5. Default collision energy regression equations. Data from MacLean *et al.*[699] and other indicated sources.

$CE = k * m/z + b$			
Instrument	Charge state	Slope k [V]	Intercept b [V]
Thermo Fisher Scientific			
TSQ Ultra/Vantage[700]	2	0.034	3.314
	3	0.044	3.314
TSQ Quantiva[700]	2	0.0339	2.3597
	3	0.0295	1.5123
Applied Biosystems			
4000 QTrap	2	0.043	4.756
Waters			
Xevo	2	0.034	1.314
	3	0.034	1.314
Agilent			
6400	2	0.036	-4.8
	3	0.036	-4.8
Sciex			[701,702]
Old default	1	0.0625	-3
	2	0.0575	9
	3	0.0625	-3
	4	0.0625	-5
	5	0.0625	-6
New default	1	0.05	5
	2	0.049	-1
	3	0.048	-2
	4	0.05	-2
	5	0.05	-2
Bruker (calculated)			
solariX[703]	2	0.037	-7.1
maXis HD Q-TOF[275]	all	0.066	5.668

3.9.2. Supplementary figures, tables & lists

3.9.2.1. List of Byonic keywords

Provided by Marshall Bern, Protein Metrics Inc, April 2019
Note the dHexA keyword has been added since.

```
HexNAc 203.079372605
Hex 162.0528235
HexN 161.0688075
Fuc 146.05790887
Pent 132.0422588
NeuAc 291.095416635
NeuGc 307.090331265
MurNAc_alditol 277.116138 // reduced
MurNAc 275.1004879 // non-reduced
MurNAcOAc 317.1110529
MurNAcOAc_alditol 319.126703
MurNAc6P 355.0668189
MurNAc6P_alditol 357.082469
MurNGlyc 291.095416635
MurNGlyc_alditol 293.1110529
MurN_alditol (277.116138 - 42.010565)
// de-acetylated MurNAc
MurN (275.1004879 - 42.010565)
OAcHexNAc (203.0793726058 + 42.010565)
// Acetylated HexNAc
Sodium 21.981942665
IdoA 176.0321
GlcA 176.0321
HexA 176.0321
Methyl 14.01565
Phosphate 79.966331
Sulfate 79.956815
Acetate 42.010565
Acetimide 41.0266
Pyruvate 70.0055
Phosphocholine 165.0555
Diphosphoethanolamine 202.9749
Phosphoethanolamine 123.0085
DiNAcBac 228.11143
Sialic229 229.059186
Sialic474 474.177886
Kdn 250.0688675
Gal405 405.15697
Gal160 160.03717
Pseudaminic 316.127039
dHexNAc 187.084815
Bac300Da 300.09439
Bac314Da 314.11004
Bac258Da 258.08383
HexNAzBr2OH 527.9679
// One Br-81 one Br-79
NeuAzBr2OH 615.98387
// one Br-81 one Br-79
HexNAz 244.08077
HexNAcNH2 218.09027
NeuAz 332.09681
NeuAcNH2 306.10632
HexNAz0 344.13320
HexNAz2 346.14575
HexNAz4 348.15831
HexNAz7 351.16205
HexNAz9 353.17460
NeuAz0 432.14924
NeuAz2 434.16180
NeuAz4 436.17435
NeuAz7 439.17809
NeuAz9 441.19065
NeuAI 329.11107
// New batch -- also one Br-81 and one Br-79
NeuAIBr2OH 658.03082
FucAI 156.04226
FucAIBr2OH 484.96201
```

3.9.2.2. Data for the decorin case study

1.1.1 Appendix for Decorin MS/MS

1.1.1.1 database

```
% Made by SFN & AG
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) -18.01 @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1
Xyl(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1
Xyl(1)Phospho(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1
```

Figure 1. Database of glycans used for re-processing. Note (1) that it is very well possible that other glycans could be present on Decorin, however they cannot be found as they are not in our database. Including a larger amount of glycans would have strained the computing costs even further *(→ Combinatorial Explosion of O-Glycan searches*). Note (2) that the dGlcA (deoxy-glucuronic acid) sugar resulting from Heparinase digest is defined as **dGlcA = Glc(A) -18.0106**. (3) Note that at this stage the -44 Da observed later in the intact MS, identified as decarboxylation of dGlcA hadn't been included.

Table 1. Glycan identifications after reprocessing a tryptic digest of MR307B. Q Exactive (HCD) ; SFN01073, July 4th 2017. Sorted by Glycoforms and then score.

Note that the Byonic scoring model provides 3 useful metrics. (1) *Byonic Score*. Ranging from 0 – 1000, a score > 300 is considered as ‚good', > 400 as ‚very good', >500 as ‚alsmost sure to be correct'. (2) *Byonic PEP* (Posterior Probability Error) is the probability a peptide comes from a false/decoy. (3) *Byonic Delta Mod* is an indication on the reliability of site localisation at the named site of modification. <20 is uncertain, >40 is good, >100 very good.

Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	z
-.DEAS[+1011.29148][-18.01000]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 993.2815)	368.5	368.5	1.38E-07	4
-.DEAS[+1011.29148][-18.01000]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 993.2815)	326	326	3.99E-07	3
-.DEAS[+1011.29148][-18.01000]GIGPEEHFPEVPEIEP[+15.99492]MGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 993.2815); P20(Oxidation / 15.9949)	279.6	0	1.18E-05	4
-.DEAS[+1011.29148][-18.01000]GIGPEEHFPEVPEIEP[+15.99492]MGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 993.2815); P20(Oxidation / 15.9949)	100.4	0	1.11E-03	3
-.DEAS[+1091.24829][-18.01000]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 1073.2383)	561.9	0.5	4.00E-09	4
-.DEAS[+1091.25781][-18.01000]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 1073.2478)	577.4	4.8	1.48E-09	3
-.DEAS[+1091.25781][-18.01000]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 1073.2478)	377.1	4.8	1.14E-08	3
-.DEAS[+1091.25781][-18.01000]GIGPEEHFPEVPEIEPM[+15.99492]GPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 1073.2478); M21(Oxidation / 15.9949)	62.1	4.8	5.60E-03	4
-.DEAS[+1091.25781][-18.01000]GIGPEEHFPEVPEIEP[+15.99492]MGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 1073.2478); P20(Oxidation / 15.9949)	372.4	0	1.32E-07	4
-.DEAS[+1091.25781][-18.01000]GIGPEEHFPEVPEIEP[+15.99492]MGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 1073.2478); P20(Oxidation / 15.9949)	179.3	0	1.46E-05	3
-.DEAS[+1091.25781][-18.01000]GIGPEEHFP[+15.99492]EVPEIEPM[+15.99492]GPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 1073.2478); P13(Oxidation / 15.9949); M21(Oxidation / 15.9949)	124.5	4.8	6.62E-04	4
-.DEAS[+1091.25781][-18.01000]GIGP[+15.99492]EEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 1073.2478); P8(Oxidation / 15.9949)	171.3	4.8	2.38E-04	4
-.DEAS[+1091.25781][-18.01000]GIGP[+15.99492]EEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.01 @ OGlycan | rare1	S4(OGlycan / 1073.2478); P8(Oxidation / 15.9949)	57.8	4.8	2.04E-03	3

The sequence coverage is shown in Figure 2..

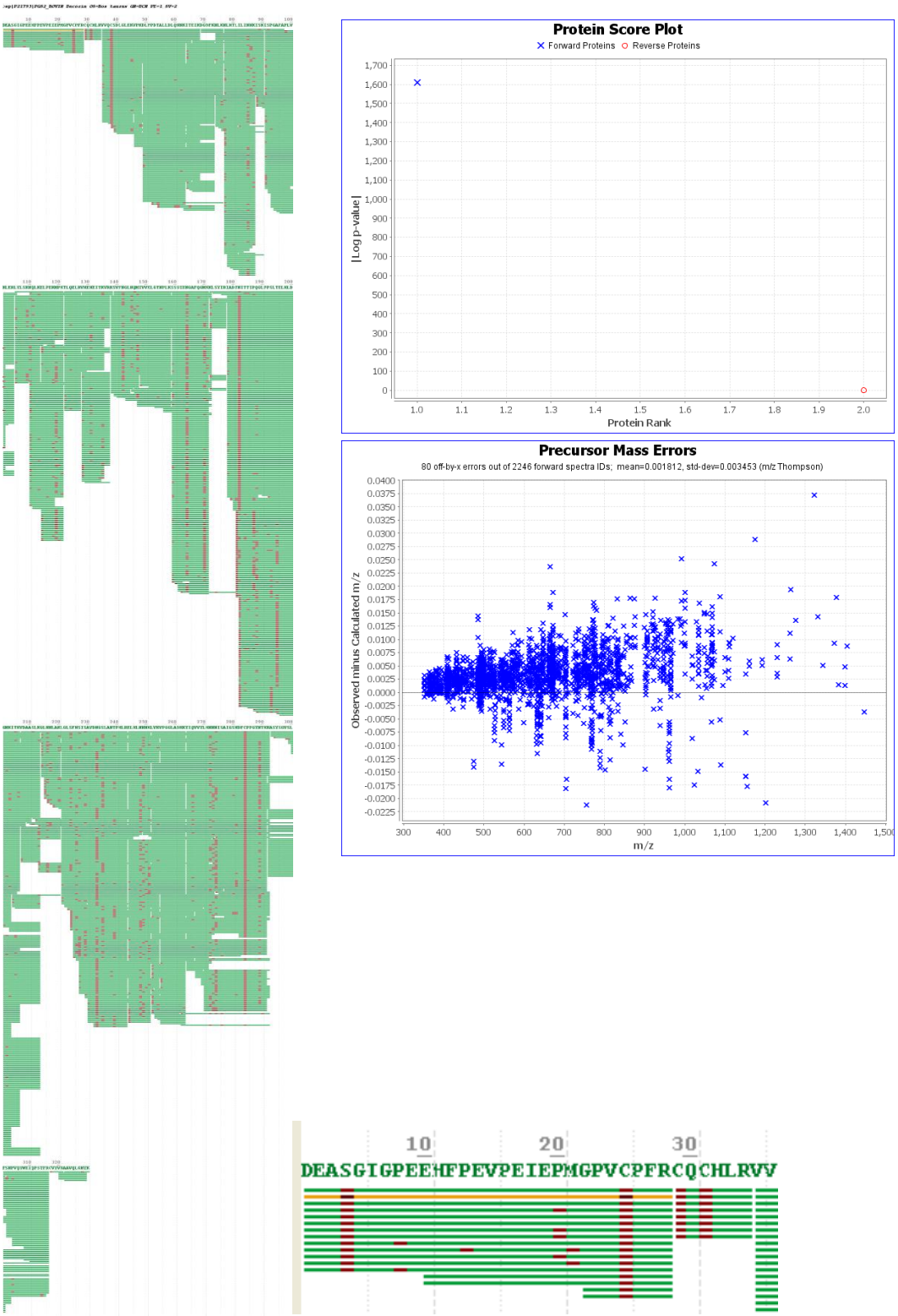


Figure 2. 100% Sequence coverage of tryptic digest MR307B with the new search parameters. Q Exactive (HCD), standard data aquisition method.

Table 2: Glycan identifications after 2nd reprocessing of a tryptic digest of MR307B including decarboxylation and hydration of the terminal glucuronic acid. Q Exactive (HCD, standard, 30% NCE) ; SFN01073, originally run on July 4th 2017. Sorted by Glycoforms and then score.

Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	z
-.DEAS[+1011.29148][-18.01060]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)GlcA(2)HexNAc(1)-18.0106 @ OGlycan | rare1	S4(OGlycan / 993.2809)	413.9	227.4	9.49E-06	3
-.DEAS[+1011.29148][-18.01060]GIGPEEHFPEVPEIEP[+15.99492]MGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)GlcA(2)HexNAc(1)-18.0106 @ OGlycan | rare1	S4(OGlycan / 993.2809); P20(Oxidation / 15.9949)	403.6	0	3.83E-07	4
-.DEAS[+1011.29148][-18.01060]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)GlcA(2)HexNAc(1)-18.0106 @ OGlycan | rare1	S4(OGlycan / 993.2809)	300	257.9	4.97E-05	4
-.DEAS[+1011.29148][-18.01060]GIGPEEHFPEVPEIEP[+15.99492]MGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)GlcA(2)HexNAc(1)-18.0106 @ OGlycan | rare1	S4(OGlycan / 993.2809); P20(Oxidation / 15.9949)	211.2	0	6.75E-05	3
-.DEAS[+1011.29148][-18.01060]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)GlcA(2)HexNAc(1)-18.0106 @ OGlycan | rare1	S4(OGlycan / 993.2809)	119.7	119.7	7.69E-05	3
-.[+43.00581]DEAS[+1011.29148][-18.01060]GIGP[+15.99492]EEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)GlcA(2)HexNAc(1)-18.0106 @ OGlycan | rare1	NTerm(Carbamyl); S4(OGlycan / 993.2809); P8(Oxidation)	107.3	32.1	5.54E-03	4
-.DEAS[+1011.29148][-18.01060]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)GlcA(2)HexNAc(1)-18.0106 @ OGlycan | rare1	S4(OGlycan / 993.2809)	90.9	90.9	3.32E-03	4
-.DEAS[+1091.24829][-62.00040]GIGP[+15.99492]EEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1)-62.0004 @ OGlycan | rare1	S4(OGlycan / 1029.2479); P8(Oxidation / 15.9949)	148.9	4.8	1.38E-03	4
-.DEAS[+1091.25781][-18.01060]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlycan / 1073.2472)	675.6	0.1	2.39E-10	4
-.DEAS[+1091.25781][-18.01060]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlycan / 1073.2472)	674.3	4.8	2.39E-10	3
-.DEAS[+1091.25781][-18.01060]GIGPEEHFPEVPEIEP[+15.99492]MGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlycan / 1073.2472); P20(Oxidation)	460.4	0	2.20E-07	4
-.DEAS[+1091.25781][-18.01060]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlycan / 1073.2472)	411	4.8	2.66E-06	3
-.DEAS[+1091.25781][-18.01060]GIGPEEHFPEVPEIEP[+15.99492]MGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlycan / 1073.2472); P20(Oxidation / 15.9949)	249.8	0	3.26E-05	3
-.DEAS[+1091.25781][-18.01060]GIGPEEHFPEVPEIEP[+15.99492]MGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlycan / 1073.2472); P20(Oxidation)	209.1	0	8.62E-05	4
-.DEAS[+1091.25781][-18.01060]GIGPEEHFPEVPEIEP[+15.99492]MGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlycan / 1073.2472); P20(Oxidation)	70.4	0	2.28E-03	3
-.DEAS[+1091.25781]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	S4(OGlycan / 1091.2578)	500.6	6.7	3.38E-06	4
-.DEAS[+1091.25781]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	S4(OGlycan / 1091.2578)	403.1	4.8	1.32E-06	3
-.DEAS[+1091.25781]GIGPEEHFPEVPEIEP[+15.99492]MGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	S4(OGlycan / 1091.2578); P20(Oxidation)	370.6	0	4.92E-07	4
-.DEAS[+1091.25781]GIGP[+15.99492]EEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	S4(OGlycan / 1091.2578); P8(Oxidation / 15.9949)	342.1	4.8	6.47E-06	4
-.DEAS[+1091.25781]GIGPEEHFP[+15.99492]EVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	S4(OGlycan / 1091.2578); P13(Oxidation)	328.3	4.8	9.50E-06	4
-.DEAS[+1091.25781]GIGPEEHFPEVPEIEPMGPVC[+57.02146]PFR.C	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	S4(OGlycan / 1091.2578)	173.7	4.8	5.95E-05	4

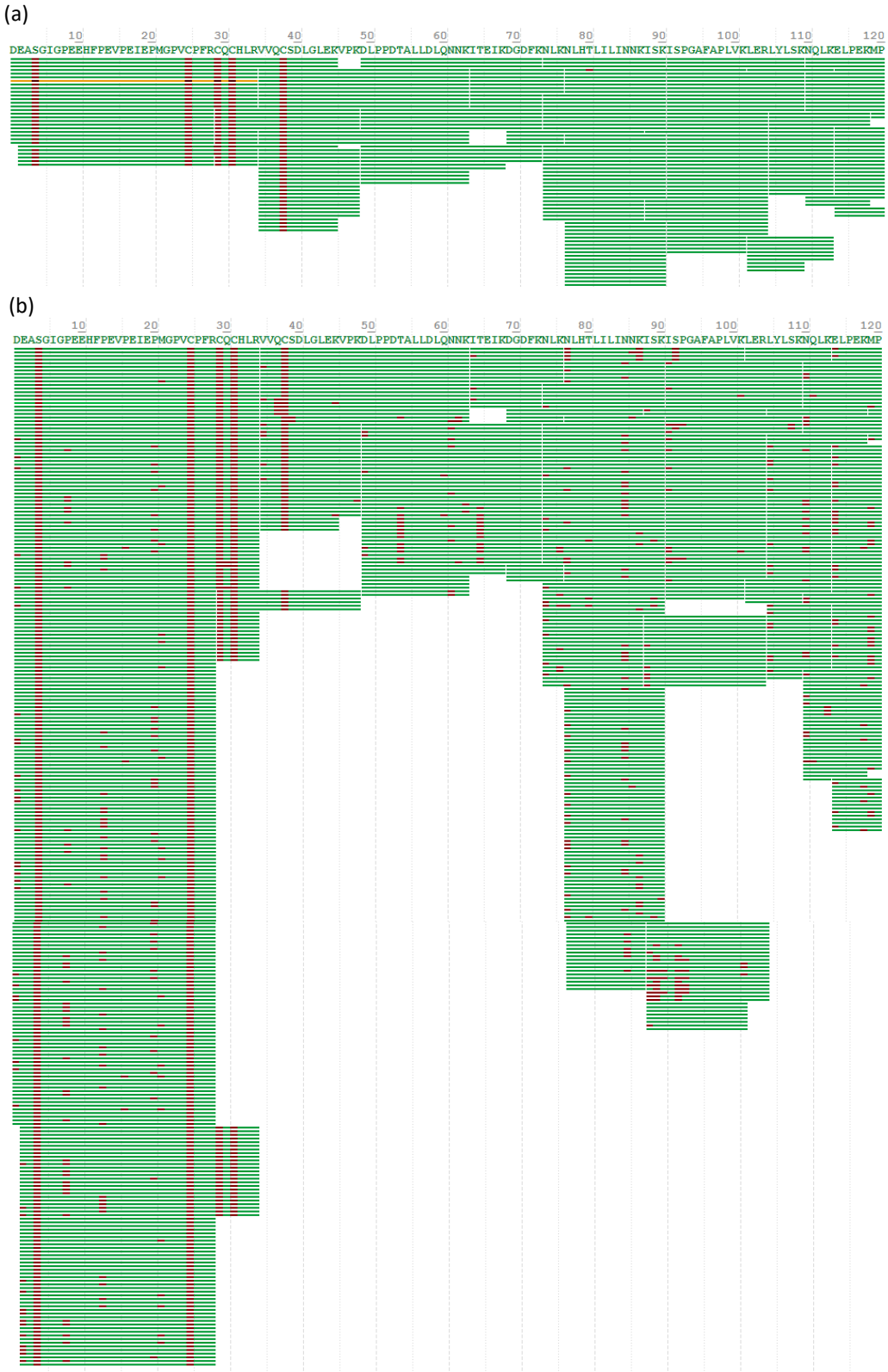


Figure 3. Reprocessing of Wei-Min Liu's original sample (MS/MS by Ritu) leads to increased peptide coverage. (a) Original search parameters. **(b)** Same sample, processed with the new search parameters. Added glycans form the majority of the identifications.

Table 3. Identifyed glycan in the original sample from Wei-Min Liu / Ritu Raj upon re-processing with the new search parameters. Sorted by Glycoforms and then score. Glycan originally found are highlighted in violet.

Glycans	Score	Delta Mod. Score	PEP 2D	z
Xyl(1)Phospho(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) @ OGlycan | rare1	230	230	3.00E-03	3
Xyl(1)Phospho(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) @ OGlycan | rare1	66.6	54.5	9.30E-03	3
Xyl(1)Phospho(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	387.3	9.7	1.00E-05	3
Xyl(1)Phospho(1)Hex(2)Sulfo(3)GlcA(1)HexNAc(1) @ OGlycan | rare1	280.4	0	9.10E-03	4
Xyl(1)Phospho(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	590	107	1.90E-07	4
Xyl(1)Phospho(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	549.1	0	3.60E-06	4
Xyl(1)Phospho(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	185.9	4.8	9.20E-03	3
Xyl(1)Phospho(1)Hex(2)Sulfo(2)GlcA(1)HexNAc(1) @ OGlycan | rare1	351.1	0	1.10E-03	4
Xyl(1)Phospho(1)Hex(2)Sulfo(2)GlcA(1)HexNAc(1) @ OGlycan | rare1	308.8	0	2.10E-03	4
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan | rare1	525.6	37.3	1.90E-07	3
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan | rare1	375.3	10.1	3.70E-03	3
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan | rare1	348	23	1.10E-04	3
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan | rare1	279	13.9	5.60E-03	3
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan | rare1	120.2	27.4	9.40E-03	3
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	290.9	0	1.80E-03	4
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	24.8	0	9.40E-03	4
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	1150	4.5	4.00E-20	3
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	791.9	3.1	6.50E-11	3
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	284.5	4.8	2.50E-03	4
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(1)HexNAc(1) @ OGlycan | rare1	85.4	4.8	8.90E-03	4
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(1) @ OGlycan | rare1	466.1	4.8	8.80E-05	4
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(1) @ OGlycan | rare1	407.6	4.8	1.20E-04	4
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(1) @ OGlycan | rare1	323.8	4.8	3.70E-03	4
Xyl(1)Phospho(1)Hex(2)Sulfo(1) @ OGlycan | rare1	721.1	721.1	2.00E-10	5
Xyl(1)Phospho(1)Hex(2)Sulfo(1) @ OGlycan | rare1	390.4	0	1.20E-04	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	942.3	2.2	4.00E-15	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	942.3	4.8	1.10E-14	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	856.4	4.8	2.00E-14	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	832.4	4.8	2.00E-13	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	796.7	4.8	4.00E-12	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	724.3	4.8	2.00E-11	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	714.1	4.8	4.00E-11	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	682.1	0	2.00E-10	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	605.3	4.8	1.10E-08	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	532.3	4.8	1.90E-07	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	507.4	2.1	1.60E-06	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	482.3	4.8	2.90E-07	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	335.6	4.8	3.90E-04	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	288.3	4.8	2.20E-04	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	270.7	0	4.20E-05	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	245.9	2.3	5.20E-03	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	42.9	0	9.40E-03	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	133.5	0	9.00E-03	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	960.1	4.8	4.00E-16	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	895.1	2.4	4.50E-14	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	876.8	4.8	4.50E-14	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	864.3	5.9	4.50E-14	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	831.9	4.8	2.00E-13	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	828.3	4.8	6.50E-12	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	827.8	4.8	4.00E-13	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	822	0	1.40E-13	5
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	818.9	3.3	9.20E-12	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	808.7	4.8	4.00E-13	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	802.8	0.4	3.20E-13	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	782.3	0	1.50E-11	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	775.9	0	1.90E-12	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	759.7	0	1.90E-12	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	751.1	0	1.90E-12	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	747.6	4.8	4.00E-11	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	731.5	0	2.10E-10	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	710.2	0	2.70E-11	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	708.8	0	1.90E-11	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	676.7	0	2.70E-10	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	659.5	0	1.90E-10	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	626.3	0	2.70E-09	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	592.3	2.2	4.00E-08	3

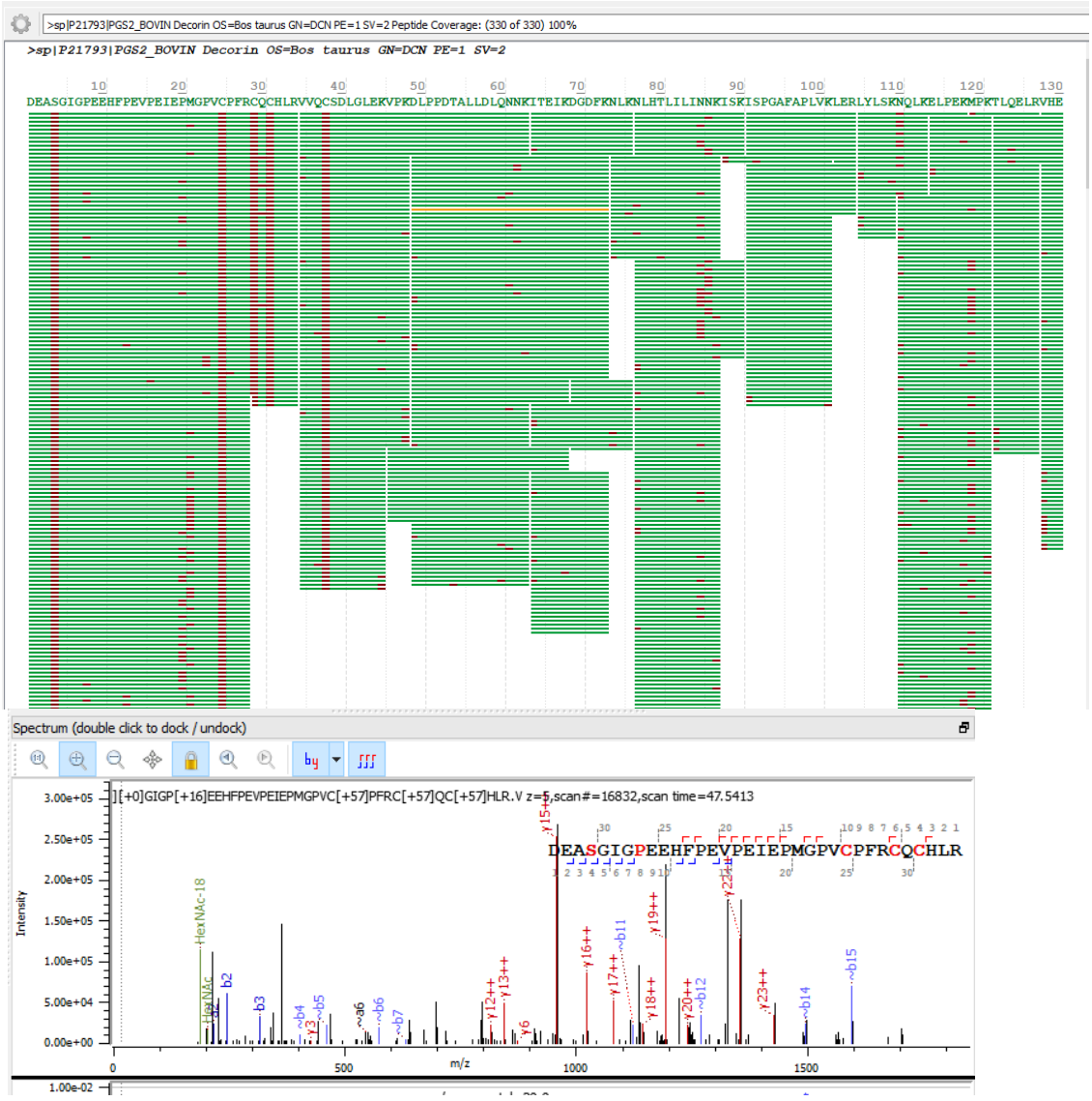
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	589.8	4.8	4.00E-08	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	583.6	0	1.90E-08	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	559.8	0	3.00E-06	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	554.1	0	4.00E-07	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	528.4	0	2.60E-07	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	492.1	0	2.30E-06	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	459.8	0.1	7.10E-06	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	434.7	0	1.80E-05	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	433.4	0	1.80E-05	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	432.5	4.8	3.00E-06	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	425	0	5.30E-07	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	408.9	4.8	2.00E-04	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	403.8	0	5.50E-04	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	356	0	4.70E-05	3
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	355.6	4.8	1.30E-04	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	339.3	4.8	2.60E-04	3
Xyl(1)Phospho(1)Hex(2)GlcA(1)HexNAc(1) @ OGlycan | rare1	392.4	4.8	8.20E-07	3
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	480.1	4.8	4.10E-06	4
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	474.4	4.8	8.80E-05	4
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	463.9	4.8	5.70E-06	4
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	450.1	4.8	1.20E-04	4
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	435.4	4.8	1.60E-04	4
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	399.6	4.8	3.80E-04	5
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	367.4	4.8	1.20E-04	5
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	351	4.8	2.90E-03	5
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	332.3	4.8	1.10E-04	3
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	275.2	4.8	6.30E-03	4
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	224.7	4.8	8.60E-03	4
Xyl(1)Phospho(1)Hex(2)GlcA(1) @ OGlycan | rare1	210.8	4.8	9.30E-03	4
Xyl(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	499.2	4.8	4.60E-07	3
Xyl(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	445.9	4.8	6.10E-07	4
Xyl(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	437.8	3.4	8.00E-07	3
Xyl(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	249.9	0	1.20E-03	4
Xyl(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	228.3	0	8.20E-03	4
Xyl(1)Hex(2)Sulfo(3)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	190.3	0	6.80E-03	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	798.3	0	2.70E-12	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	562.2	0.7	4.00E-08	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	456.9	3.4	2.40E-06	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	320.6	3.1	8.20E-04	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	673.4	0	7.90E-10	5
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	598	2.7	1.30E-07	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	582.6	4.8	1.10E-05	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	539.8	4.8	1.30E-06	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	536.3	0	7.40E-07	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	533.8	0	7.40E-07	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	532.9	4.8	4.70E-05	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	521.5	0	7.70E-07	5
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	502.1	2.6	1.90E-06	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	487.6	3.4	1.90E-06	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	482.9	4.8	1.90E-06	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	479.2	4.6	2.70E-06	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	478.9	4.8	2.40E-06	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	461.5	4.8	2.40E-06	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	428.3	4.8	1.10E-03	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	397.7	4.8	6.80E-03	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	381.6	0	5.40E-06	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	377.4	4.5	5.40E-06	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	328.4	4.8	1.30E-05	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	270.5	0	6.80E-03	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	258.5	4.8	5.70E-03	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	244.6	4.8	6.60E-05	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	82.6	0	9.30E-03	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	32.7	0	9.40E-03	4
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	14.4	4.8	9.10E-03	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	8.5	2.9	9.10E-03	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	637.1	2	4.00E-09	3
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	441.5	0	2.70E-07	4
Xyl(1)Hex(2)Sulfo(2)GlcA(1)HexNAc(1) @ OGlycan | rare1	411.4	4.8	1.20E-05	4
Xyl(1)Hex(2)Sulfo(2)GlcA(1)HexNAc(1) @ OGlycan | rare1	373.5	4.8	4.80E-06	3
Xyl(1)Hex(2)Sulfo(2)GlcA(1)HexNAc(1) @ OGlycan | rare1	328.9	4.8	4.90E-05	3
Xyl(1)Hex(2)Sulfo(2)GlcA(1)HexNAc(1) @ OGlycan | rare1	326	4.8	5.60E-05	3
Xyl(1)Hex(2)Sulfo(2)GlcA(1)HexNAc(1) @ OGlycan | rare1	110.5	4.8	9.30E-03	3

Xyl(1)Hex(2)Sulfo(2) @ OGlycan | rare1	309.6	4.8	5.90E-04	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan | rare1	691.4	0	4.00E-10	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan | rare1	688.9	4.8	2.90E-10	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan | rare1	628.6	0	3.10E-09	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan | rare1	564	1.9	4.00E-08	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan | rare1	551.3	0	4.00E-07	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	521.3	4.8	1.30E-06	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	516.9	4.8	1.70E-06	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	492.7	4.8	1.80E-04	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	484.8	4.8	1.90E-06	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	429.4	5.8	5.50E-05	5
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	413.2	4.8	2.50E-05	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	352.8	0	1.60E-04	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	345.3	4.8	1.10E-03	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	305.3	4.8	1.00E-04	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	304.9	0	2.70E-04	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	263	4.8	8.20E-03	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	252.7	4.8	6.60E-03	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	934.3	4.8	2.90E-15	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	910.2	1.4	4.00E-15	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	829.2	4.8	7.20E-12	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	808.9	4.7	3.20E-12	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	801.8	4.5	4.00E-13	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	745.9	4.8	4.00E-11	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	724.7	4.8	4.50E-11	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	720.8	2.6	4.00E-11	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	714.9	4.8	2.90E-11	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	692.8	4.8	2.90E-10	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	674.3	0	4.20E-09	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	667.6	0	4.00E-10	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	657.5	2.6	9.20E-09	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	657.4	1.5	1.50E-08	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	650	0	2.70E-09	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	597.4	0	4.00E-08	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	585.1	3.7	4.00E-08	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	553.3	0	2.60E-07	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	506	0	4.00E-07	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	486.9	0	4.80E-06	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	467.7	0	5.60E-07	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	354.5	1.6	1.80E-05	3
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	198.5	4.8	7.50E-03	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	180.3	4.8	9.10E-03	3
Xyl(1)Hex(2)Sulfo(1)GlcA(1)HexNAc(1) @ OGlycan | rare1	839	4.1	6.50E-12	3
Xyl(1)Hex(2)Sulfo(1)GlcA(1)HexNAc(1) @ OGlycan | rare1	665.2	4.8	6.50E-09	3
Xyl(1)Hex(2)Sulfo(1)GlcA(1)HexNAc(1) @ OGlycan | rare1	504.7	4.8	4.00E-07	4
Xyl(1)Hex(2)Sulfo(1)GlcA(1)HexNAc(1) @ OGlycan | rare1	472.9	4	5.20E-05	4
Xyl(1)Hex(2)Sulfo(1)GlcA(1)HexNAc(1) @ OGlycan | rare1	413.1	4.8	1.80E-04	3
Xyl(1)Hex(2)Sulfo(1) @ OGlycan | rare1	740.6	4.8	2.90E-11	3
Xyl(1)Hex(2)Sulfo(1) @ OGlycan | rare1	474.3	1	8.40E-07	3
Xyl(1)Hex(2)Sulfo(1) @ OGlycan | rare1	468.9	3.2	8.40E-07	3
Xyl(1)Hex(2)Sulfo(1) @ OGlycan | rare1	355.5	4.3	2.60E-04	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	679.4	211.2	2.00E-10	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	618.5	202.4	3.10E-09	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	345.5	56.7	2.10E-03	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	163.4	163.4	8.00E-03	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	261.7	261.7	4.30E-03	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -62.0004 @ OGlycan | rare1	25.2	18.3	9.40E-03	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	1020.2	307.9	3.40E-17	5
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	970.2	418.4	2.00E-16	5
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	968.7	393	2.00E-16	5
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	953	447.7	2.00E-16	5
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	939.9	378.2	2.00E-15	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	934.3	389.1	2.00E-15	5
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	924.8	501.1	2.80E-15	5
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	894.1	468.8	2.80E-14	5
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	888.3	482.8	2.80E-14	5
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	879.6	350	3.20E-13	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	871.9	317.4	7.20E-13	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	869.5	361.2	4.00E-14	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	858.9	366.2	4.00E-14	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	853.8	335.8	2.70E-14	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	849	366.4	4.00E-13	3

Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	846.5	349.2	4.00E-13	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	809.4	315.9	4.00E-13	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	802.8	7.4	2.80E-13	5
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	795.7	370.5	1.30E-12	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	788.7	318.9	5.00E-11	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	783	350.6	4.50E-12	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	762.4	348.4	3.20E-11	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	754.4	356.4	2.70E-12	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	744.8	515.6	2.70E-11	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	727.7	338.9	1.90E-11	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	709.8	275	6.60E-10	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	709.5	320.2	4.00E-11	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	707.1	265.2	9.20E-10	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	703.2	310.8	2.70E-11	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	700.8	286.9	4.00E-11	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	700.4	318.5	4.00E-11	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	698.9	300.2	4.00E-10	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	698.5	343.4	2.70E-10	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	688.4	260.5	2.00E-10	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	686.6	686.6	2.70E-10	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	681.1	280.4	2.00E-10	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	678.9	0	2.00E-10	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	664.2	0	2.00E-10	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	653.3	245.9	2.70E-10	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	651.1	310.1	2.70E-10	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	643	313.4	2.70E-09	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	635.7	0	2.70E-09	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	570.5	237.7	3.20E-07	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	559.1	377.7	4.00E-07	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	547.2	0	4.00E-07	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	524.8	264.5	3.10E-06	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	516.5	262.4	6.00E-07	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	473.5	199.8	1.90E-05	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	473	252.6	9.40E-07	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	458.3	292.5	1.30E-05	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	450.1	14.8	5.70E-05	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	448.8	261.6	1.90E-05	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	444.1	238.5	1.90E-05	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	433.3	183.2	2.50E-05	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	422.5	0	1.00E-04	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	416.6	0	1.20E-05	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	412.2	0	1.00E-04	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	410.1	0	3.60E-07	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	403.5	174.4	1.80E-05	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	392.9	0	4.60E-06	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	392.6	141	2.70E-04	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	389.1	389.1	4.60E-06	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	353.1	0	5.00E-05	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	272.8	269.3	3.30E-03	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	258	0	6.70E-05	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	225.5	143.4	8.50E-03	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	223.9	0	1.30E-03	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	105.5	0	9.10E-03	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	94.2	94.2	8.80E-03	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	85.4	85.4	9.10E-03	3
Xyl(1)Hex(2)GlcA(1)HexNAc(1) @ OGlycan | rare1	768.6	768.6	4.00E-12	3
Xyl(1)Hex(2)GlcA(1)HexNAc(1) @ OGlycan | rare1	594.2	594.2	4.00E-08	3
Xyl(1)Hex(2)dGlcA(1) / +614.16888 @ S | rare1	528.3	65.4	2.60E-07	4
Xyl(1)Hex(2)dGlcA(1) / +614.16888 @ S | rare1	501.6	88.7	2.80E-07	4
Xyl(1)Hex(2)dGlcA(1) / +614.16888 @ S | rare1	481	26.6	4.30E-07	3
Xyl(1)Hex(2)dGlcA(1) / +614.16888 @ S | rare1	249.4	2.6	6.60E-05	3
Xyl(1)Hex(2)dGlcA(1) / +614.16888 @ S | rare1	223.5	4.1	3.10E-03	3
Xyl(1)Hex(2)dGlcA(1) / +614.16888 @ S | rare1	69.9	0	8.60E-03	3
Xyl(1)Hex(2)dGlcA(1) / +614.16888 @ S | rare1	23.8	23.8	9.40E-03	3
Xyl(1)Hex(2)ddGlcA(1) / +570.1791 @ S | rare1	145.1	8.8	8.90E-03	3
Xyl(1)Hex(2)ddGlcA(1) / +570.1791 @ S | rare1	62.8	62.8	9.20E-03	3
Xyl(1)Hex(2) @ OGlycan | rare1	752.1	752.1	2.90E-12	3
Xyl(1)Hex(2) @ OGlycan | rare1	544.9	544.9	2.90E-07	3
Xyl(1)Hex(2) @ OGlycan | rare1	539.1	539.1	2.90E-07	3
Xyl(1)Hex(2) @ OGlycan | rare1	476.2	476.2	6.10E-07	3
Xyl(1)Hex(2) @ OGlycan | rare1	339	13.5	5.50E-04	3

1.1.1.2 Q Exactive: Trypsin – HCD standard

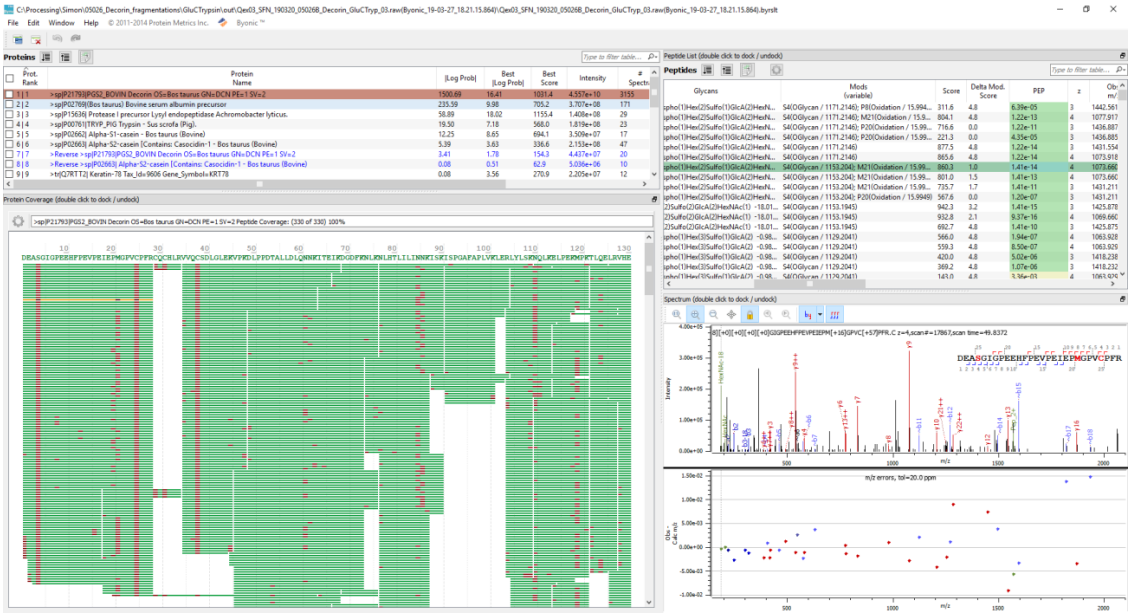
Correct glycans, 100% coverage, no glycan fragmentation



1.1.1.3 Q Exactive – Trypsin – HCD standard – 1 week @ -20

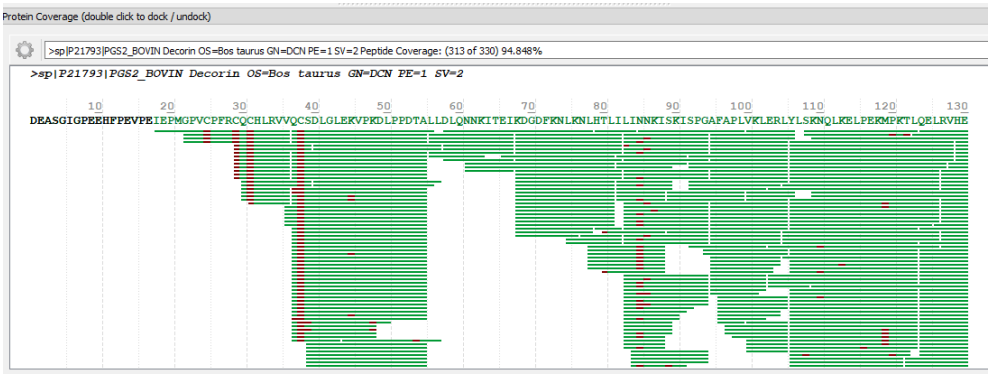
Same results. Sample seems to be stable and can be be kept at -20 without glycan hydrolysis

1.1.1.4 Q Exactive – GluC/Trypsin – HCD standard



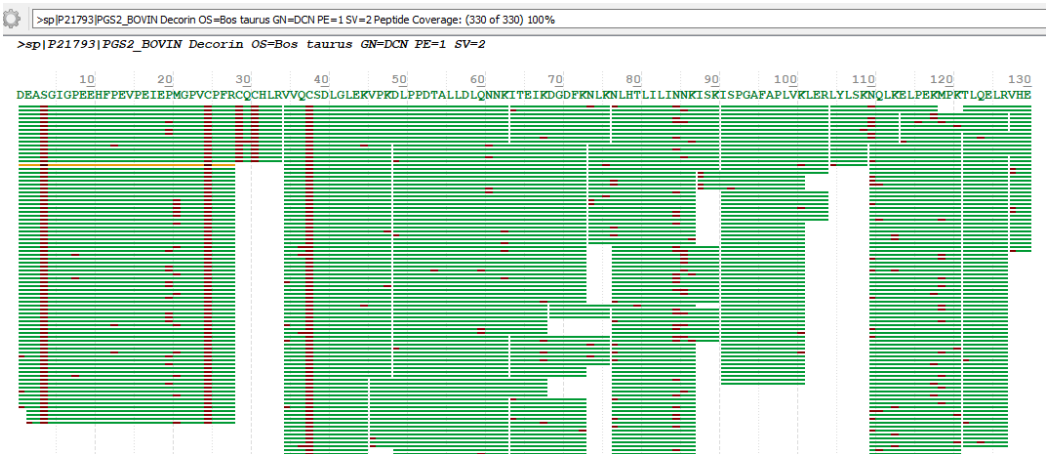
Digest works as well. Also stable 1 week.

1.1.1.5 Q Exactive – Elastase – HCD



Site of interest not covered (not found in semi-specific search with 5 missed cleavages)

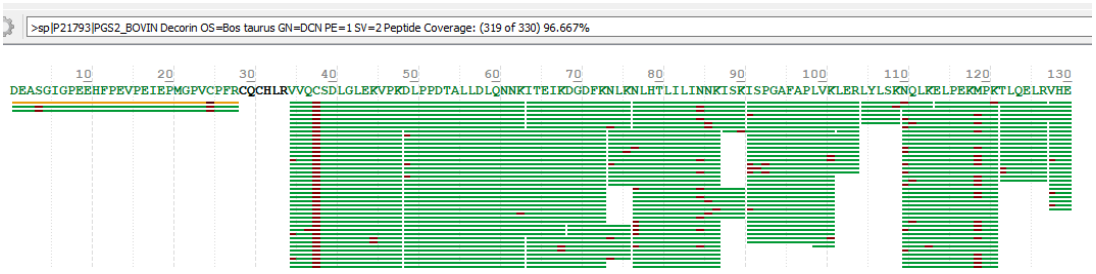
1.1.1.6 Q Exactive – Trypsin – SCE-HCD 20-30-40



Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	z
-.DEAS[+1011.29148]GIGPEEHFPEV...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) @ OGl...	S4(OGlyca...	1143.2	513.1	1.11e-19	4
-.DEAS[+1011.29148]GIGPEEHFPEV...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) @ OGl...	S4(OGlyca...	730.4	302.0	1.98e-10	4
-.DEAS[+1011.29148]GIGPEEHFPEV...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) @ OGl...	S4(OGlyca...	486.1	0.0	2.04e-06	4
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	894.6	506.9	7.08e-15	4
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	874.1	481.2	7.08e-15	4
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	572.7	214.3	1.11e-08	4
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	459.4	190.9	7.67e-07	3
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	300.5	146.5	4.17e-06	3
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	290.9	187.6	4.37e-06	3
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	55.3	55.3	6.81e-03	3
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	937.2	461.0	1.04e-14	5
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	222.6	222.6	7.26e-03	5
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	805.6	12.8	7.08e-14	4
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	511.5	0.0	7.27e-08	3
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	122.2	0.0	4.07e-03	3
-.DEAS[+1011.29148][-18.01060]Gl...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	633.1	0.0	1.04e-08	5
-.DEAS[+1050.22174][-0.98400]GlG...	Xyl(1)Hex(3)Sulfo(1)GlcA(2) -0.9840 ...	S4(OGlyca...	253.1	4.8	7.77e-03	4
-.DEAS[+1050.22174][-0.98400]GlG...	Xyl(1)Hex(3)Sulfo(1)GlcA(2) -0.9840 ...	S4(OGlyca...	35.5	0.0	7.65e-02	5
-.DEAS[+1050.23126][-0.98400]GlG...	Xyl(1)Phospho(1)Hex(3)GlcA(2) -0.98...	S4(OGlyca...	148.3	4.8	8.40e-03	4
-.DEAS[+1050.23126][-18.01060]Gl...	Xyl(1)Phospho(1)Hex(3)GlcA(2) -18.0...	S4(OGlyca...	19.7	0.0	2.11e-01	5
-.DEAS[+1091.24829][-18.01060]Gl...	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1...	S4(OGlyca...	822.4	4.1	2.34e-13	4
-.DEAS[+1091.24829][-18.01060]Gl...	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1...	S4(OGlyca...	485.8	0.0	4.76e-06	3
-.DEAS[+1091.24829][-18.01060]Gl...	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1...	S4(OGlyca...	412.7	0.0	7.57e-06	3
-.DEAS[+1091.24829][-18.01060]Gl...	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1...	S4(OGlyca...	406.5	0.0	5.63e-06	5
-.DEAS[+1091.24829][-18.01060]Gl...	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1...	S4(OGlyca...	264.1	0.0	1.20e-04	4
-.DEAS[+1091.24829][-43.98980]Gl...	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1...	S4(OGlyca...	254.0	19.2	4.35e-03	4
-.DEAS[+1091.25781]GIGPEEHFPEV...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	1080.2	4.8	7.08e-19	4
-.DEAS[+1091.25781]GIGPEEHFPEV...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	883.2	4.8	1.11e-14	4
-.DEAS[+1091.25781]GIGPEEHFPEV...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	683.0	4.8	7.08e-11	3
-.DEAS[+1091.25781]GIGPEEHFPEV...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	670.3	4.8	7.08e-11	3
-.DEAS[+1091.25781]GIGPEEHFPEV...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	677.5	0.0	1.04e-09	5
-.DEAS[+1091.25781]GIGPEEHFPEV...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	800.6	4.8	7.08e-14	4
-.DEAS[+1091.25781]GIGPEEHFPEV...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	584.2	0.0	7.08e-09	3
-.DEAS[+1091.25781]GIGPEEHFPEV...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	474.6	0.0	8.97e-08	3
-.DEAS[+1091.25781]GIGP[+15.994...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	578.9	3.8	1.98e-07	4
-.DEAS[+1091.25781]GIGP[+15.994...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	491.0	4.8	2.04e-06	4
-.DEAS[+1091.25781][-18.01060]Gl...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	1025.9	4.8	2.34e-17	4
-.DEAS[+1091.25781][-18.01060]Gl...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	922.3	4.8	7.08e-16	4

1.1.1.7 Orbitrap Elite: HCD-pd-ETD – HexNAc trigger

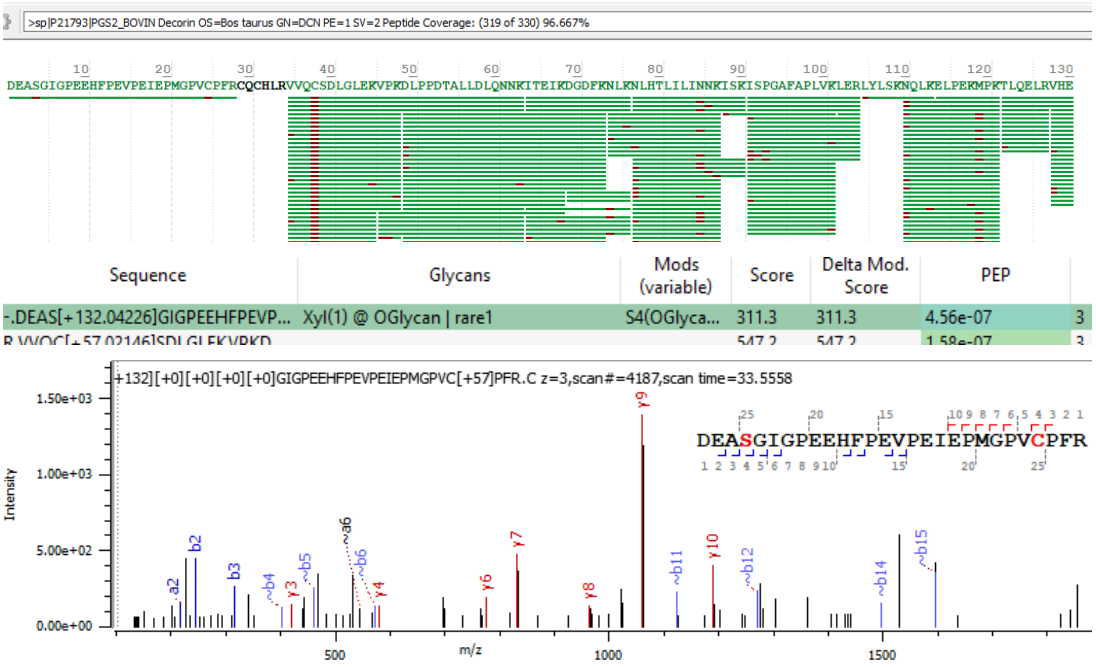
HCD methods perform poorly on the Elite compared to the QEx



Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	48.4	48.4	9.69e-03
-.DEAS[+132.04226]GIGPEEHFPEVP...	Xyl(1) @ OGlycan | rare1	S4(OGlyca...	567.0	567.0	2.21e-08

1.1.1.8 Orbitrap Elite: HCD-pd-ETD – XylHexGlcA triggers

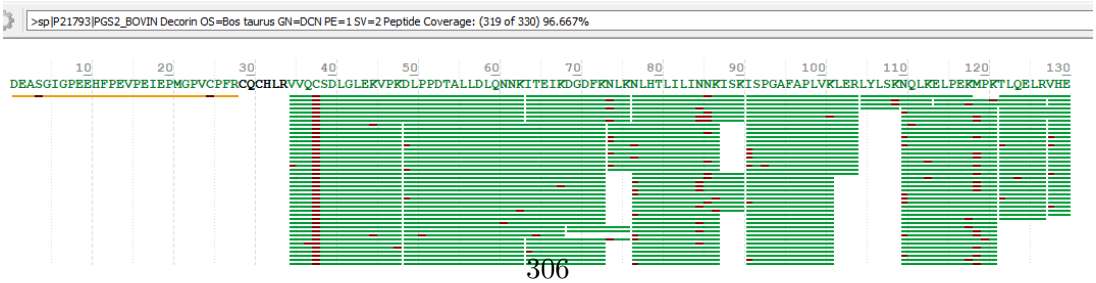
Poor



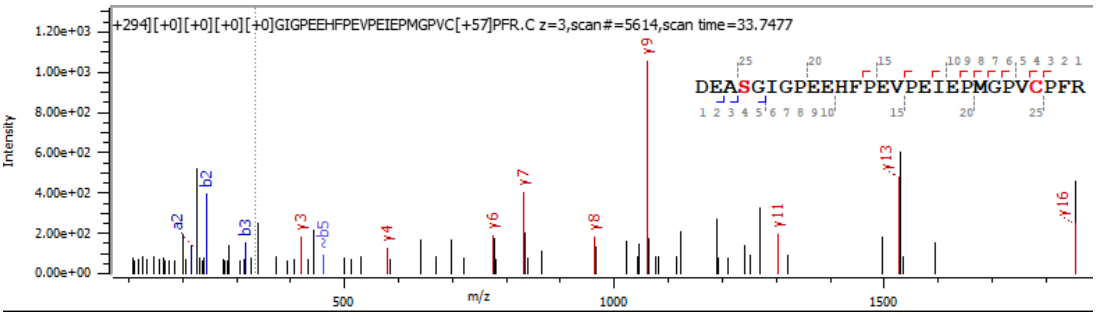
Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	
-.DEAS[+132.04226]GIGPEEHFPEVP...	Xyl(1) @ OGlycan | rare1	S4(OGlyca...	311.3	311.3	4.56e-07	3
R.VVQC[+57.02146]SDLGLEKVPKD			547.2	547.2	1.58e-07	3

1.1.1.9 Orbitrap Elite: HCD-alt-ETD

No interesting hit here

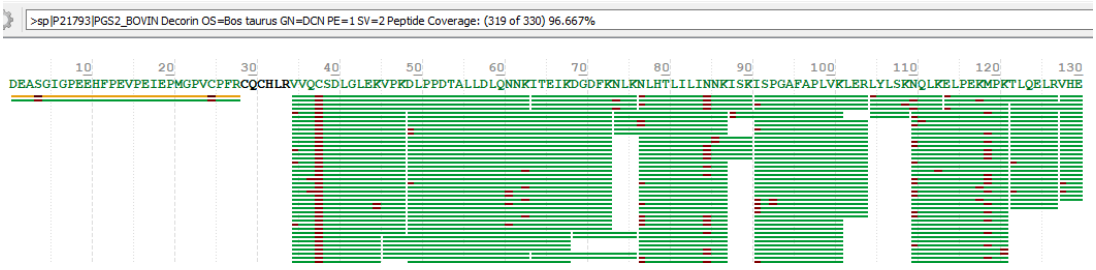


Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	381.5	381.5	3.72e-07

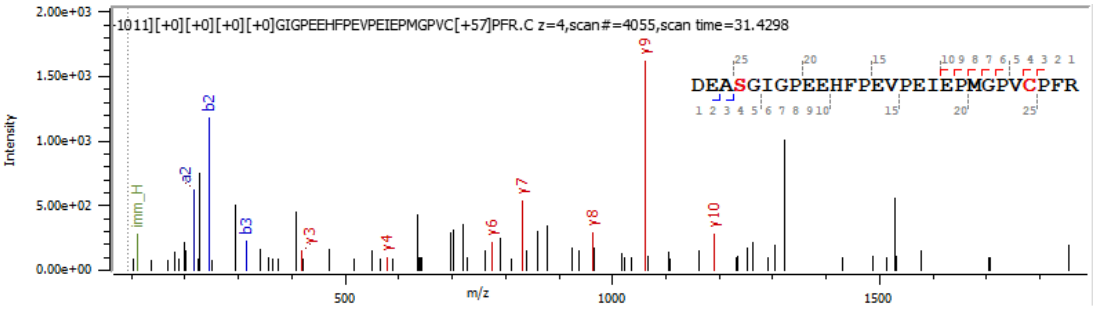


1.1.1.10 Orbitrap Elite: HCD-alt-EThcD

1.1.1.10.1 HCD & ETD search

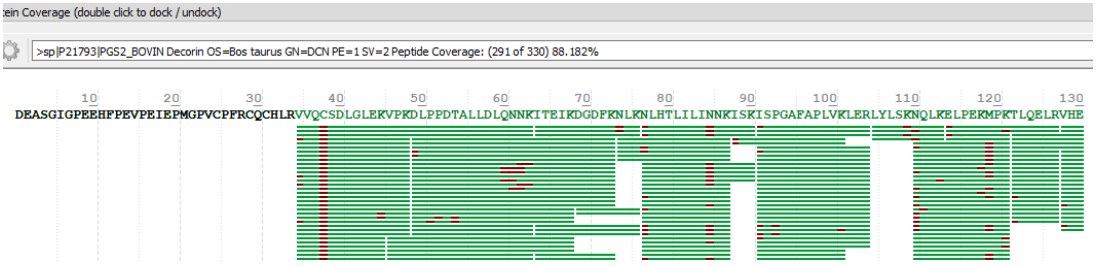


Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	z
-.DEAS[+1011.29148]GIGPEEHFPEV...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) @ OGl...	S4(OGlyca...	225.9	143.5	5.26e-05	4
-.DEAS[+1011.29148]GIGPEEHFPEV...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) @ OGl...	S4(OGlyca...	190.8	113.0	8.42e-05	4



1.1.1.10.2 EThcD search on mgf file

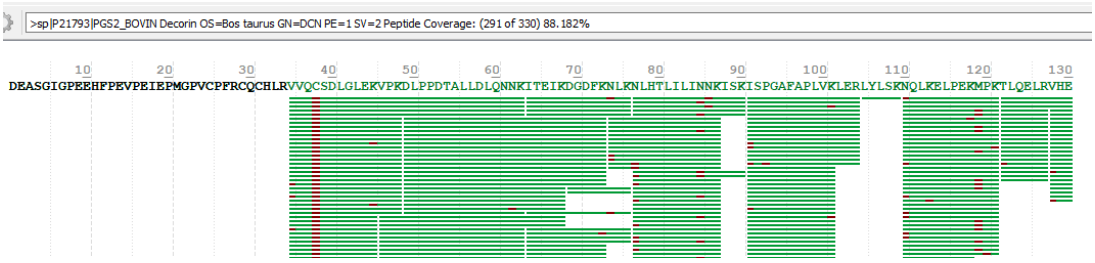
No coverage



1.1.1.11 Orbitrap Elite: EThcD

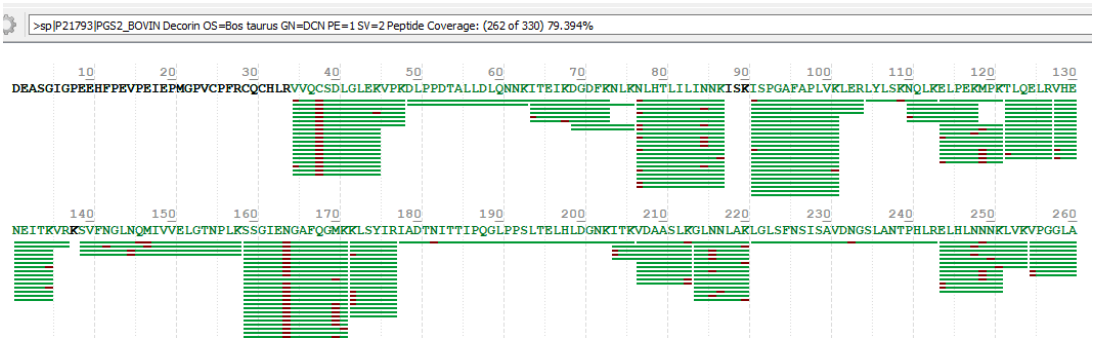
1.1.1.11.1 Search with HCD & ETD

No identification



1.1.1.11.2 Search with EThcD/mgf file

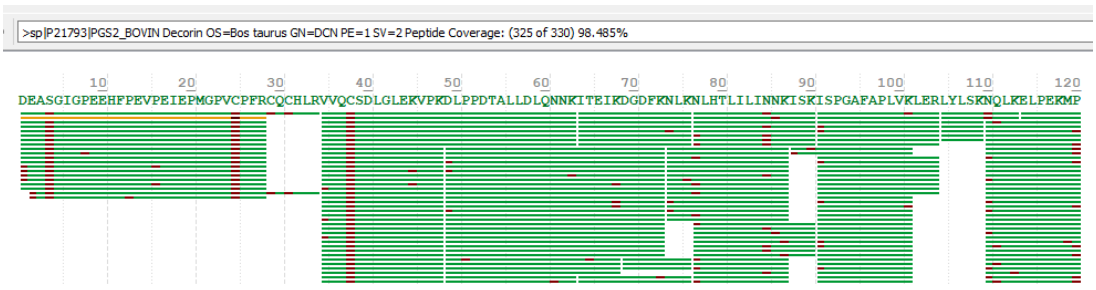
Very poor coverage compared to previous search



1.1.1.12 Orbitrap Elite: CID-pd-ETD

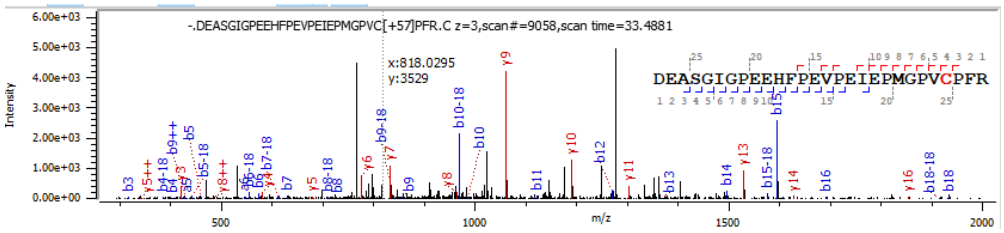
1.1.1.12.1 HexNAc triggers

Many hits but very poor quality spectra / very poor assignment (many false positives likely)

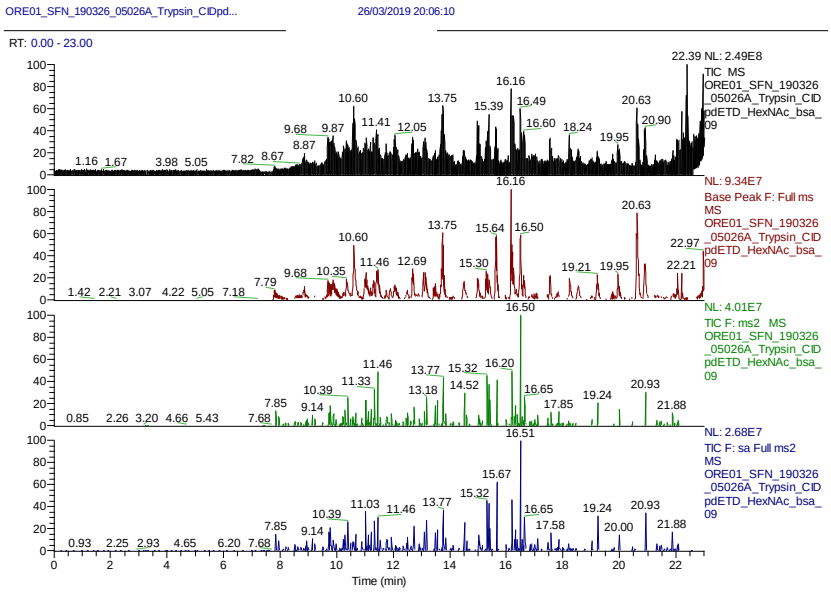


Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP
-.DEASGIGPEEHFPEVPEIEPMGPVC[...			982.6	982.6	1.45e-18
-.DEAS[+1011.29148][-18.01060]GI...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlyca...	23.7	23.7	2.41e-03
-.DEAS[+1091.25781][-18.01060]GI...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlyca...	S4(OGlyca...	177.5	0.0	2.83e-02
-.DEAS[+1130.17856][-0.98400]GIG...	Xyl(1)Hex(3)Sulfo(2)GlcA(2) -0.9840 @ OGlycan | rare1	S4(OGlyca...	194.4	4.8	5.68e-02
-.DEAS[+1130.18808][-18.01060][-0...	Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(2) -18.0106 -0.9840 @ O...	S4(OGlyca...	262.6	4.8	7.53e-05
-.DEAS[+1171.20511][-18.01060]GI...	Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | r...	S4(OGlyca...	199.1	0.6	1.17e-01
-.DEAS[+1171.21462]GIGPEEHFPEV...	Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan...	S4(OGlyca...	86.5	1.6	5.81e-02
-.DEAS[+132.04226]GIGPEEHFPEVP...	Xyl(1) @ OGlycan | rare1	S4(OGlyca...	424.7	422.5	2.25e-08
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	562.2	537.3	6.24e-09
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	350.2	350.2	2.43e-07
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	346.3	295.7	9.00e-07
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	550.5	446.3	6.81e-09
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	451.4	451.4	1.60e-08
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	189.9	189.9	8.26e-04
-.[+43.00581]DEAS[+1091.24829][-...	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | r...	NTerm(Car...	154.4	3.5	7.79e-01
-.[+43.00581]DEAS[+1130.17856][-...	Xyl(1)Hex(3)Sulfo(2)GlcA(2) -18.0106 -0.9840 @ OGlycan | rare1	NTerm(Car...	166.6	3.5	8.15e-02
-.[+43.00581]DEAS[+1130.18808]GI...	Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(2) @ OGlycan | rare1	NTerm(Car...	158.4	4.8	3.65e-02
-.[+43.00581]DEAS[+1171.20511][-...	Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -43.9898 @ OGlycan | r...	NTerm(Car...	119.0	4.8	1.32e-01
D.[+43.00581]EAS[+1130.18808][-4...	Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(2) -43.9898 @ OGlycan | ...	NTerm(Car...	80.8	4.8	3.04e-01
D.[+43.00581]EAS[+1171.21462]GIG...	Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan...	NTerm(Car...	192.5	4.8	7.65e-01
K.ASYSGVSLFSNPVQYWEIQPSTFR.C			780.3	745.7	1.25e-14

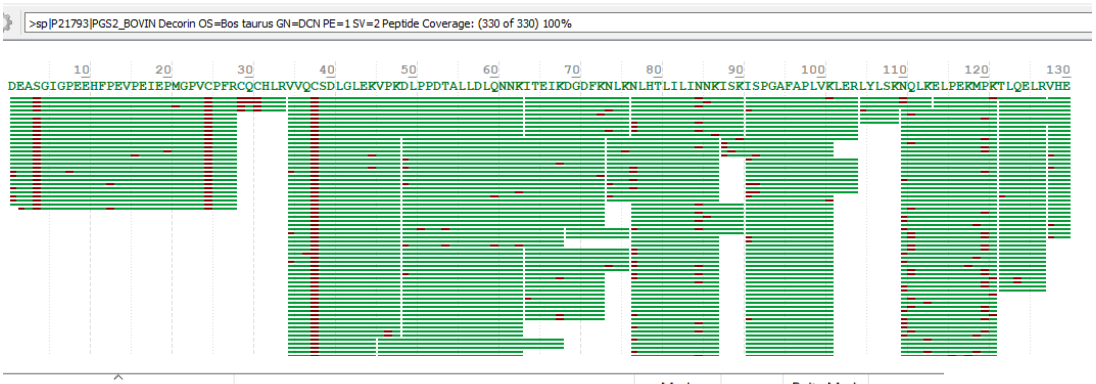
This is literally the only good CID spectra (-18 hydrolysed glycan)



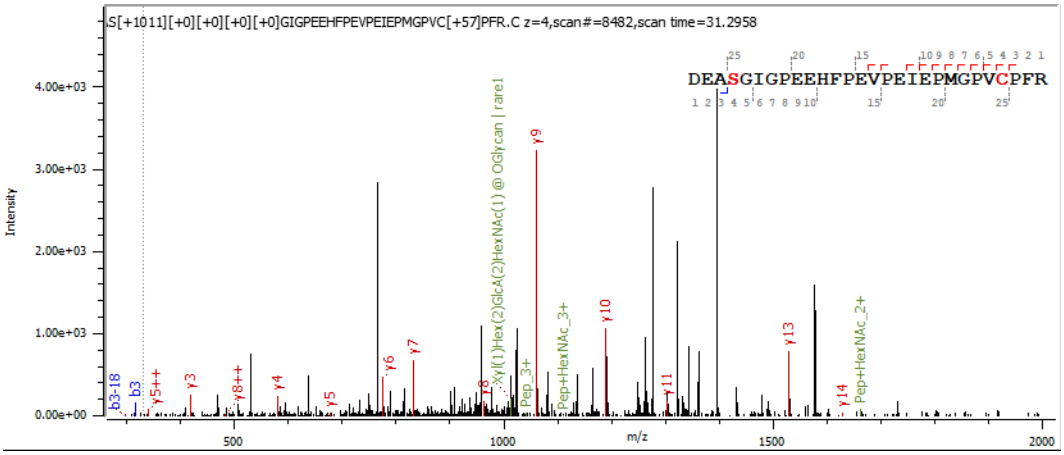
ETD was triggered but no identifications of peptide on glycosite



1.1.1.12.2 XylHexdGlcA trigger



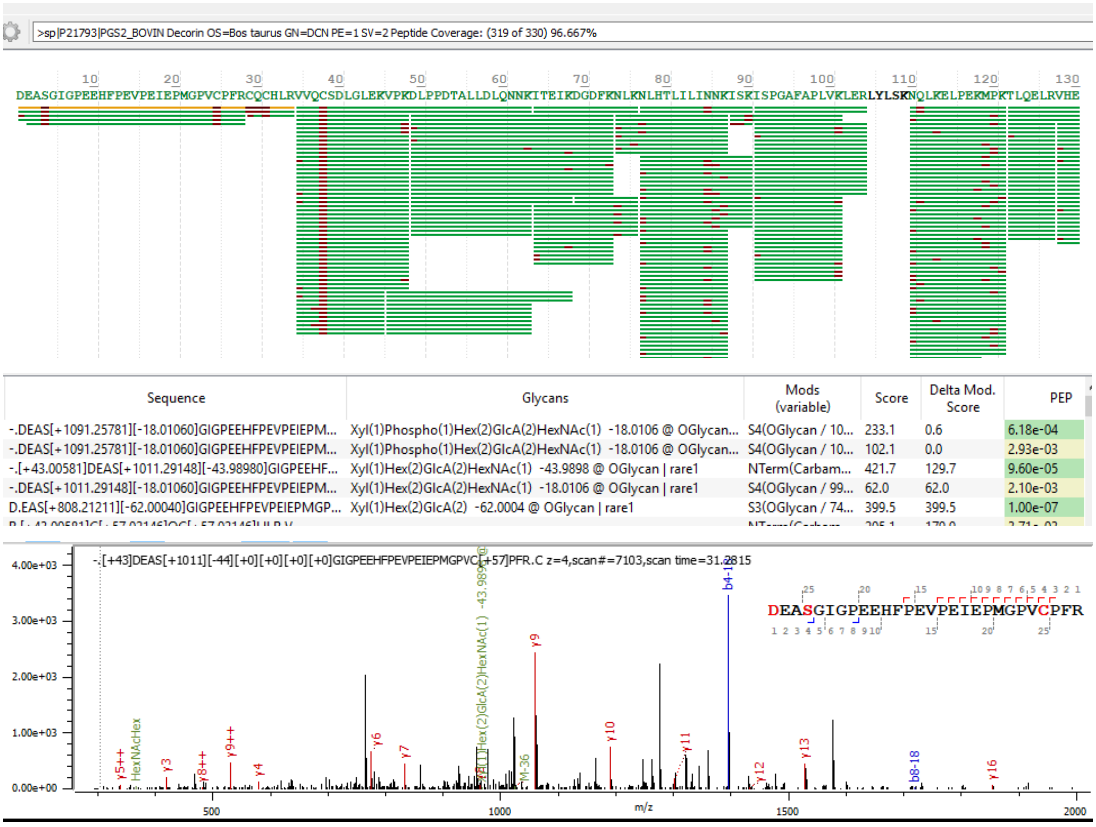
Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP
-.DEASGIGPEEHFPEVPEIEPMGPVC[...			945.9	945.9	3.68e-16
-.DEAS[+1011.29148]GIGPEEHFPEV...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	S4(OGlyca...	531.9	126.6	3.55e-08
-.DEAS[+1011.29148][-18.01060]GI...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlyca...	42.7	42.7	1.34e-01
-.DEAS[+1050.23126][-18.01060]GI...	Xyl(1)Phospho(1)Hex(3)GlcA(2) -18.0106 @ OGlycan | rare1	S4(OGlyca...	145.2	0.0	4.74e-01
-.DEAS[+1091.25781]GIGPEEHFPEV...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) @ OGlycan | rare1	S4(OGlyca...	216.3	2.7	3.85e-01
-.DEAS[+1091.25781][-18.01060]GI...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlyca...	S4(OGlyca...	265.1	0.0	5.03e-04
-.DEAS[+1091.25781][-18.01060]GI...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlyca...	S4(OGlyca...	177.4	0.0	1.12e-01
-.DEAS[+1091.25781][-87.97960]GI...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -87.9796 @ OGlyca...	S4(OGlyca...	131.1	4.8	4.74e-01
-.DEAS[+1130.18808][-18.01060]GI...	Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(2) -18.0106 @ OGlycan | ...	S4(OGlyca...	199.9	4.8	1.86e-01
-.DEAS[+1130.18808][-18.01060]GI...	Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(2) -18.0106 @ OGlycan | ...	S4(OGlyca...	135.8	4.8	4.74e-01
-.DEAS[+1130.18808][-18.01060][-0...	Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(2) -18.0106 -0.9840 @ O...	S4(OGlyca...	153.8	4.8	1.96e-01
-.DEAS[+1130.18808][-62.00040]GI...	Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(2) -62.0004 @ OGlycan | ...	S4(OGlyca...	333.2	1.5	5.02e-05
-.DEAS[+1210.14489][-0.98400]GIG...	Xyl(1)Phospho(1)Hex(3)Sulfo(2)GlcA(2) -0.9840 @ OGlycan | r...	S4(OGlyca...	84.6	4.8	6.87e-01
-.DEAS[+132.04226]GIGPEEHFPEVP...	Xyl(1) @ OGlycan | rare1	S4(OGlyca...	492.1	492.1	2.76e-08
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	763.9	757.2	3.68e-13
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	390.1	390.1	2.97e-07
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	357.3	357.3	1.89e-06
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	504.3	427.1	4.49e-08
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	415.0	409.7	1.55e-07
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	414.3	414.3	5.52e-08
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	184.7	184.7	7.68e-04
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	352.8	0.0	1.71e-06
-.[+43.00581]DEAS[+1050.23126][-...	Xyl(1)Phospho(1)Hex(3)GlcA(2) -18.0106 -0.9840 @ OGlycan |...	NTerm(Car...	84.1	4.8	4.85e-02
-.[+43.00581]DEAS[+1091.25781][-...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlyca...	NTerm(Car...	139.4	4.8	4.74e-01
-.[+43.00581]DEAS[+1091.25781][-...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlyca...	NTerm(Car...	78.9	4.8	4.88e-01
-.[+43.00581]DEAS[+1130.17856][-...	Xyl(1)Hex(3)Sulfo(2)GlcA(2) -18.0106 -0.9840 @ OGlycan | rare1	NTerm(Car...	152.4	4.8	2.19e-01
-.[+43.00581]DEAS[+1171.20511][-...	Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -43.9898 @ OGlycan | r...	NTerm(Car...	117.3	4.8	4.73e-01
D.[+43.00581]EAS[+1210.14489][-4...	Xyl(1)Phospho(1)Hex(3)Sulfo(2)GlcA(2) -43.9898 @ OGlycan | ...	NTerm(Car...	232.7	4.8	4.73e-01
K.ASYSGVSLFSNPVQYWEIQPSTFR.C			760.6	760.6	3.85e-14
K.ASYSGVSLFSNPVQYWEIQPSTFR.C			740.3	659.4	1.75e-13



As before no convincing spectra

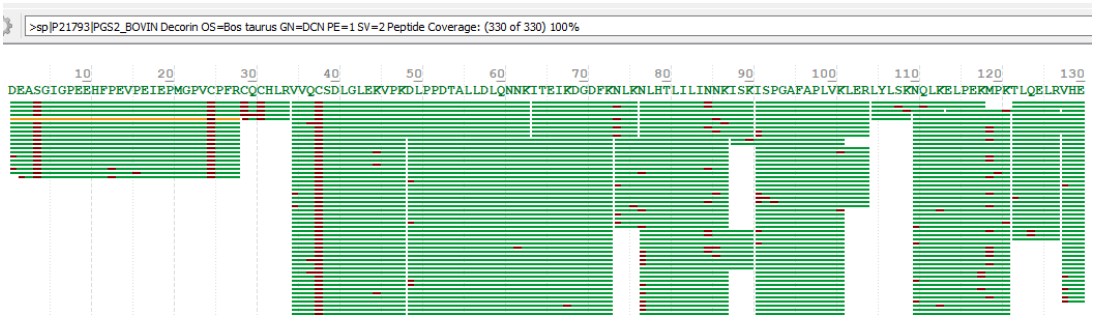
1.1.1.13 Orbitrap Elite: CID-alt-ETD

Same CID problems



1.1.1.14 Orbitrap Elite: CID-ETD-DDT

Same CID problems



Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	z
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -43.9898 @ OGlycan | r...	S4(OGlyca...	91.8	0.0	6.36e-01	4
Xyl(1)Phospho(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlyca...	S4(OGlyca...	82.5	0.0	5.19e-02	4
Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(2) -62.0004 @ OGlycan | ...	S4(OGlyca...	61.7	0.9	3.14e-01	4
Xyl(1)Phospho(1)Hex(3)GlcA(2) -18.0106 @ OGlycan | rare1	S4(OGlyca...	113.6	0.0	3.41e-01	5
Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1) -43.9898 @ OGlycan | r...	NTerm(Car...	110.4	4.8	2.18e-01	4
Xyl(1)Hex(3)Sulfo(2)GlcA(2) -18.0106 -0.9840 @ OGlycan | rare1	NTerm(Car...	79.2	3.0	4.92e-02	4
Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(2) -18.0106 @ OGlycan | ...	S4(OGlyca...	163.7	4.8	4.22e-02	4
Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan | rare1	S4(OGlyca...	46.8	1.7	2.23e-01	4
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlyca...	341.5	138.7	9.49e-05	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlyca...	330.5	88.4	2.29e-04	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlyca...	312.9	93.1	1.55e-04	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlyca...	259.2	92.5	9.36e-04	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlyca...	238.8	116.4	6.81e-04	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlyca...	118.9	77.8	4.14e-02	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlyca...	105.1	101.0	4.49e-02	3
Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.0106 @ OGlycan | rare1	S4(OGlyca...	90.4	90.4	4.49e-02	3
Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	290.0	290.0	2.34e-05	3
		893.8	893.8	1.80e-15	3
Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1) @ OGlycan...	NTerm(Car...	138.2	4.8	3.89e-01	4

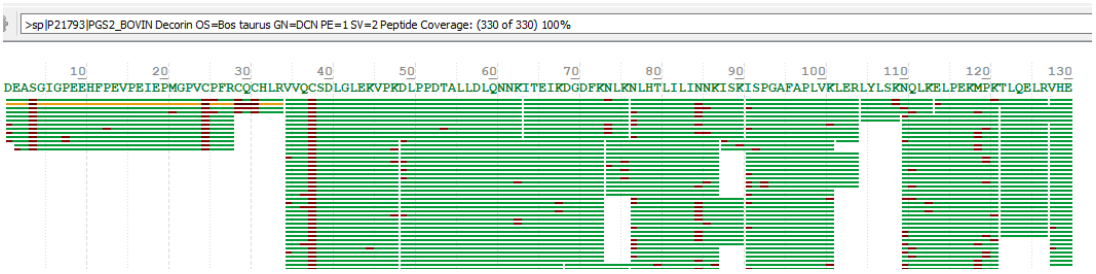
1.1.1.15 Orbitrap Elite: CID-alt-EThcD

1.1.1.15.1 CID & ETD search

One of the best CID spectra so far

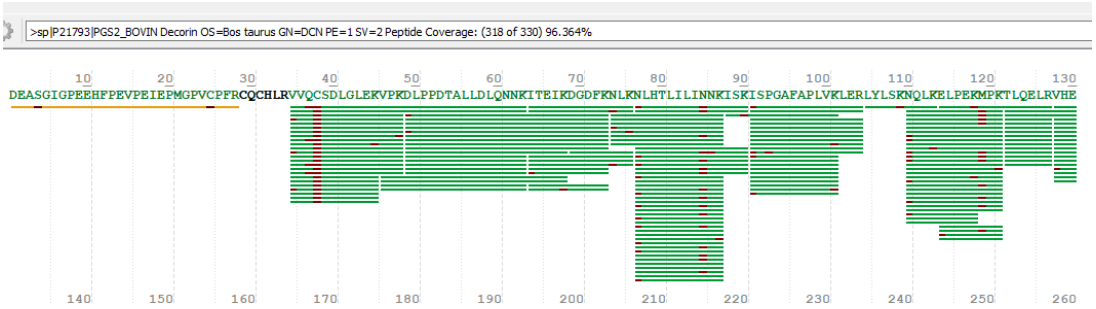
1.1.1.15.2 CID & HCD search

To check effects of hack implementation. Same results



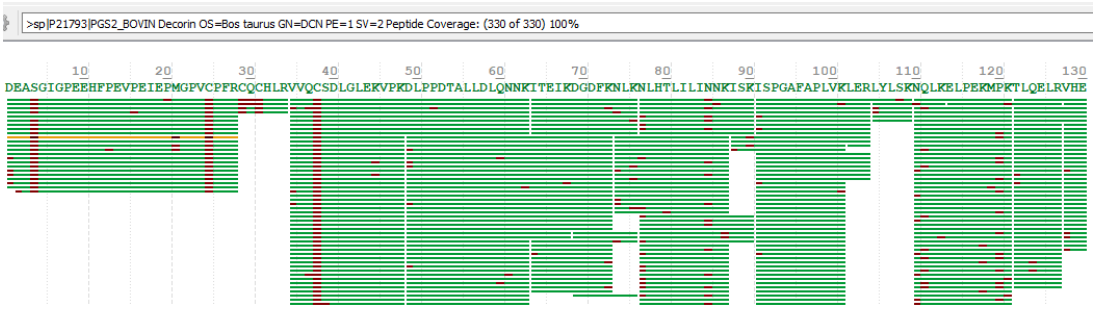
Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	
-.DEAS[+1050.23126][-18.01060][-0...	Xyl(1)Phospho(1)Hex(3)GlcA(2) -18.0...	S4(OGlyca...	90.4	1.1	7.98e-02	5
-.DEAS[+1091.25781][-18.01060]Gl...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	70.4	0.1	2.04e-01	4
-.DEAS[+1091.25781][-18.01060]Gl...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	181.2	0.0	4.31e-03	4
-.DEAS[+1130.18808][-18.01060][-0...	Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(...	S4(OGlyca...	195.4	4.5	2.68e-02	4
-.DEAS[+1171.21462][-18.01060]Gl...	Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(...	S4(OGlyca...	161.7	1.5	7.83e-03	4
-.DEAS[+1171.21462][-18.01060]Gl...	Xyl(1)Phospho(1)Hex(2)Sulfo(1)GlcA(...	S4(OGlyca...	21.5	0.1	2.40e-01	4
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	405.9	405.9	4.37e-06	3
-.DEAS[+970.26493][-18.01060]GIG...	Xyl(1)Hex(3)GlcA(2) -18.0106 @ OGly...	S4(OGlyca...	26.7	1.4	1.16e-01	4
-.[+43.00581]DEAS[+1091.24829][-...	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1...	NTerm(Car...	177.4	4.8	1.11e-01	4
-.[+43.00581]DEAS[+1091.24829][-...	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1...	NTerm(Car...	15.2	15.2	4.30e-01	4
-.[+43.00581]DEAS[+1171.20511][-...	Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1...	NTerm(Car...	49.9	4.8	6.38e-02	4
D.EAS[+1091.25781]GIGPEEHFPEVP...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S3(OGlyca...	183.2	11.6	5.17e-01	4
D.[+43.00581]EAS[+1130.18808][-4...	Xyl(1)Phospho(1)Hex(3)Sulfo(1)GlcA(...	NTerm(Car...	103.4	4.8	6.22e-01	4

1.1.1.15.3 EThcD search with mgf file



1.1.1.15.4 XylHexdGlcA – CID & ETD search

HexNAc trigger result is similar/slightly better



Sequence	Glycans	Mods (variable)	Score	Delta Mod. Score	PEP	
-.DEAS[+1011.29148][-18.01060]GI...	Xyl(1)Hex(2)GlcA(2)HexNAc(1) -18.01...	S4(OGlyca...	161.7	24.4	4.91e-03	3
-.DEAS[+1050.22174][-18.01060][-0...	Xyl(1)Hex(3)Sulfo(1)GlcA(2) -18.0106 ...	S4(OGlyca...	90.6	0.8	4.36e-01	4
-.DEAS[+1091.25781][-18.01060]GI...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	128.7	0.0	4.97e-01	4
-.DEAS[+1091.25781][-18.01060]GI...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	143.6	0.0	3.97e-01	4
-.DEAS[+1091.25781][-18.01060]GI...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	S4(OGlyca...	217.8	0.0	3.05e-03	4
-.DEAS[+1210.14489]GIGPEEHFPEV...	Xyl(1)Phospho(1)Hex(3)Sulfo(2)GlcA(...	S4(OGlyca...	135.7	8.5	7.44e-01	4
-.DEAS[+1210.14489][-0.98400]GIG...	Xyl(1)Phospho(1)Hex(3)Sulfo(2)GlcA(...	S4(OGlyca...	265.0	3.1	3.72e-03	4
-.DEAS[+132.04226]GIGPEEHFPEVP...	Xyl(1) @ OGlycan | rare1	S4(OGlyca...	631.4	631.4	4.46e-11	3
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	684.9	682.4	4.46e-12	3
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	387.7	387.7	4.77e-08	3
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	86.2	86.2	2.07e-01	3
-.DEAS[+294.09508]GIGPEEHFPEVP...	Xyl(1)Hex(1) @ OGlycan | rare1	S4(OGlyca...	427.5	4.8	5.73e-08	3
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	578.1	570.2	2.85e-09	3
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	434.8	434.8	4.50e-08	3
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	220.0	220.0	1.60e-04	3
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	213.8	213.8	3.84e-04	3
-.DEAS[+632.18001]GIGPEEHFPEVP...	Xyl(1)Hex(2)GlcA(1) @ OGlycan | rare1	S4(OGlyca...	360.2	10.6	3.54e-07	3
-.[+43.00581]DEAS[+1091.24829][-...	Xyl(1)Hex(2)Sulfo(1)GlcA(2)HexNAc(1...	NTerm(Car...	143.0	4.8	6.69e-01	4
-.[+43.00581]DEAS[+1091.25781][-...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	NTerm(Car...	144.9	4.8	6.73e-01	4
-.[+43.00581]DEAS[+1091.25781][-...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	NTerm(Car...	133.0	5.6	7.49e-01	4
-.[+43.00581]DEAS[+1171.20511][-...	Xyl(1)Hex(2)Sulfo(2)GlcA(2)HexNAc(1...	NTerm(Car...	215.7	4.8	5.77e-03	4
D.[+43.00581]EAS[+1091.25781]GIG...	Xyl(1)Phospho(1)Hex(2)GlcA(2)HexN...	NTerm(Car...	204.1	4.8	8.60e-01	4
K.ASYSGVSLFSNPVQYWEIQPSTFR.C			777.9	765.6	1.59e-14	2

3.10. Supplementary references

273. Scientific, T. F. PSB 104 - Normalized Collision Energy Technology, 2.

274. Orsburn, B. *News in Proteomics Research: Normalized Collision Energy Calculation for Q Exactive* News in Proteomics Research. http://proteomicsnews.blogspot.com/2014/06/normalized-collision-energy-calculation.html (2020).

275. Hinneburg, H., Stavenhagen, K., Schweiger-Hufnagel, U., Pengelley, S., Jabs, W., Seeberger, P. H., Silva, D. V., Wuhrer, M. & Kolarich, D. The Art of Destruction: Optimizing Collision Energies in Quadrupole-Time of Flight (Q-TOF) Instruments for Glycopeptide-Based Glycoproteomics. *Journal of The American Society for Mass Spectrometry* **27,** 507–519 (2016).

699. MacLean, B., Tomazela, D. M., Abbatiello, S. E., Zhang, S., Whiteaker, J. R., Paulovich, A. G., Carr, S. A. & MacCoss, M. J. Effect of Collision Energy Optimization on the Measurement of Peptides by Selected Reaction Monitoring (SRM) Mass Spectrometry. *Analytical chemistry* **82,** 10116–10124 (2010).

700. Thermo Fisher. iQuan: Best Practices for Peptide Quantitation on a Triple Quadrupole Mass Spectrometer. *Liquid chromatography,* 67 (2016).

701. Sciex Inc. *Optimal CE Voltages for Peptide Analysis on TripleTOF® Systems* https://sciex.com/support/knowledge-base-articles/questions-about-optimal-ce-voltages-for-peptide-analysis-on-tripletof-systems (2020).

702. Sciex Inc. *Crowdsourcing: Optimized Rolling Collision Energy Curves for ID and SWATH Acquisition* https://sciex.com/community/application-discussions/proteomics/swath/data-acquisition/crowdsourcing-optimized-rolling-collision-energy-curves-for-id-and-swath-acquisition (2020).

703. Inc, B. *FTMS-40_ebook.Pdf* https://www.bruker.com/fileadmin/user_upload/8-PDF-Docs/Separations_MassSpectrometry/Literature/literature/ApplicationNotes/FTMS-40_ebook.pdf (2020).

Part II.

Chemical proteomics for the identification of interaction partners of histone PTMs and their interaction site

The greatest teacher, failure is.

— Star Wars: The Last Jedi

4

Synthetic nucleosomes reveal interaction partners of phosphorylated and O-GlcNAcylated Histone H2B S36

Contents

4.1. Abstract . . . 317
4.2. Introduction . . . 318
4.2.1. Synthetic posttranlational modification of histones . . . 318
4.2.1.1. Background and motivation . . . 318
4.2.1.2. Searching for histone modifications using tandem mass spectrometry . . . 320
4.2.1.3. The chemical biology toolkit . . . 321
4.2.1.4. Application of chemical biology tools for the study of chromatin biology . . . 325
4.2.1.5. Probing interaction partners and downstream spreading of modifications . . . 326
4.2.2. *O*-GlcNAc is part of the histone code − the case of H2B-Ser36 . . . 328
4.2.3. Aim of the project . . . 331
4.3. Results & Discussion . . . 334
4.3.1. Synthesis of post-translationally modified histone H2B-Ser36 variants . . . 334
4.3.2. Tandem mass spectrometry of modified proteins . . . 339
4.3.3. Reconstitution of histone octamers with H2B-Ser36 variants . . . 341
4.3.4. Generation of Widom 601 nucleosome-positioning DNA . . . 343
4.3.4.1. Generation of 145bp DNA and 186 bp biotinylated DNA . . . 343
4.3.4.2. DNA degradation and subsequent purification . . . 344
4.3.5. Nucleosome assembly . . . 346

4.3.6. Interaction proteomics with synthetic nucleosomes . . . 349
4.4. Outlook . . . 355
4.5. Author contributions . . . 359
4.6. Experimental . . . 360
4.6.1. Materials . . . 360
4.6.1.1. Chemicals and enzymes . . . 360
4.6.1.2. Plasmids . . . 360
4.6.1.3. Primers . . . 361
4.6.2. Expression and purification of histone proteins . . . 362
4.6.3. Bacterial transformation . . . 364
4.6.4. Plasmid amplification . . . 365
4.6.5. Expression and purification of histone proteins . . . 365
4.6.6. Post-translational chemical mutagenesis . . . 367
4.6.6.1. Synthesis of H2B-Dha36 protein . . . 367
4.6.6.2. Synthesis of H2B-CysGlcNAc36 . . . 368
4.6.6.3. Attempts for the synthesis of thiophosphorylated histone H2B-pCys36 . . . 369
4.6.6.4. Synthesis of H2B-Pfa36 . . . 371
4.6.7. LC-MS/MS of modified histone proteins . . . 373
4.6.8. Histone octamer assembly . . . 374
4.6.9. Large scale DNA preparation for nucleosome reconstitution . . . 376
4.6.9.1. Preparation of the 145 bp "601" DNA . . . 376
4.6.9.2. Preparation of the 186 bp biotinylated "601" DNA . . . 378
4.6.9.3. Southern Blot with anti-biotin detection . . . 380
4.6.9.4. Purification of DNA fragments . . . 381
4.6.10. Nucleosome reconstitution . . . 382
4.6.11. Preparation of nuclear cell extract . . . 383
4.6.12. Interaction proteomics experiment with synthetic nucleosomes . . . 384
4.7. References . . . 387

4.1. Abstract

Post-translational modification (PTM) of histone proteins is an epigenetic mechanism regulating gene transcription. Among the many histone PTM, the *O*-GlcNAc modification is largely unexplored. In this chapter, homogeneously GlcNAcylated and phosphorylated histones and nucleosomes are generated by post-translational chemical mutagenesis and used to explore the biological function of both modification on H2B-Ser36. Mass spectrometry-based quantitative interaction proteomics reveals a direct interaction between H2B-Ser36 GlcNAcylated nucleosomes and the TCF20 transcription factor, which may provide an indication on the role of *O*-GlcNAcylation in neuronal development and diseases with neurological impairment.

4.2. Introduction

4.2.1. Synthetic posttranlational modification of histones

The scope of this section follows closely a review article published by the author during this DPhil.[704] The review has been re-written, shortened and updated for the purpose of this introduction. This section provides background information for the following two chapters.

4.2.1.1. Background and motivation

Epigenetic mechanisms allow to control and inherit changes in gene expression and cellular phenotype without changes in the underlying DNA sequence. They are contributing to cellular regulation of gene expression and can happen at any of several layers of genome organisation in the nucleus, often influencing each other (Figure 4.1.a).[705,706] One such epigenetic layer of regulation are histone post-translational modifications (PTMs).[707,708] This post-translational modification of histones is a major cellular mechanism in eukaryotes for the regulation of gene expression. The interplay of PTMs influences a complex and scarcely understood process often referred to as the 'histone code' resulting in the recruitment of transcription factors.[707]

Histone PTMs have multiple functions and regulatory tasks. As an example, H3K14ac leads to transcriptional initiation,[707] others such as H3K4me3 recruit specific transcription factors (i.e with PHD 'reader' domains) that will recognise the modification and activate a cellular signalling pathway.[709] The study of histone PTMs is hampered by several difficulties: (i) a plethora of PTMs have been discovered (Figure 4.1.b), covering a highly diverse range of chemical functionalities. (ii) Several distinct histone proteoforms (histone variants such as H2AX) are found in a single cells, increasing the variety of the histone proteome. (iii) Histone tails can have multiple modifications in close proximity, which includes a combinatorial dimension to the study of histone PTMs. Therefore, access to large quantities of homogeneously modified histones is required to study histone PTMs − which is difficult to achieve using enzymatic techniques.[710,711] Synthetic chemical biology methods have emerged in the last decade to address this ambitious challenge, allowing to generate histones with defined synthetic modifications and to assemble these into synthetic 'designer' nucleosomes.[712–721]

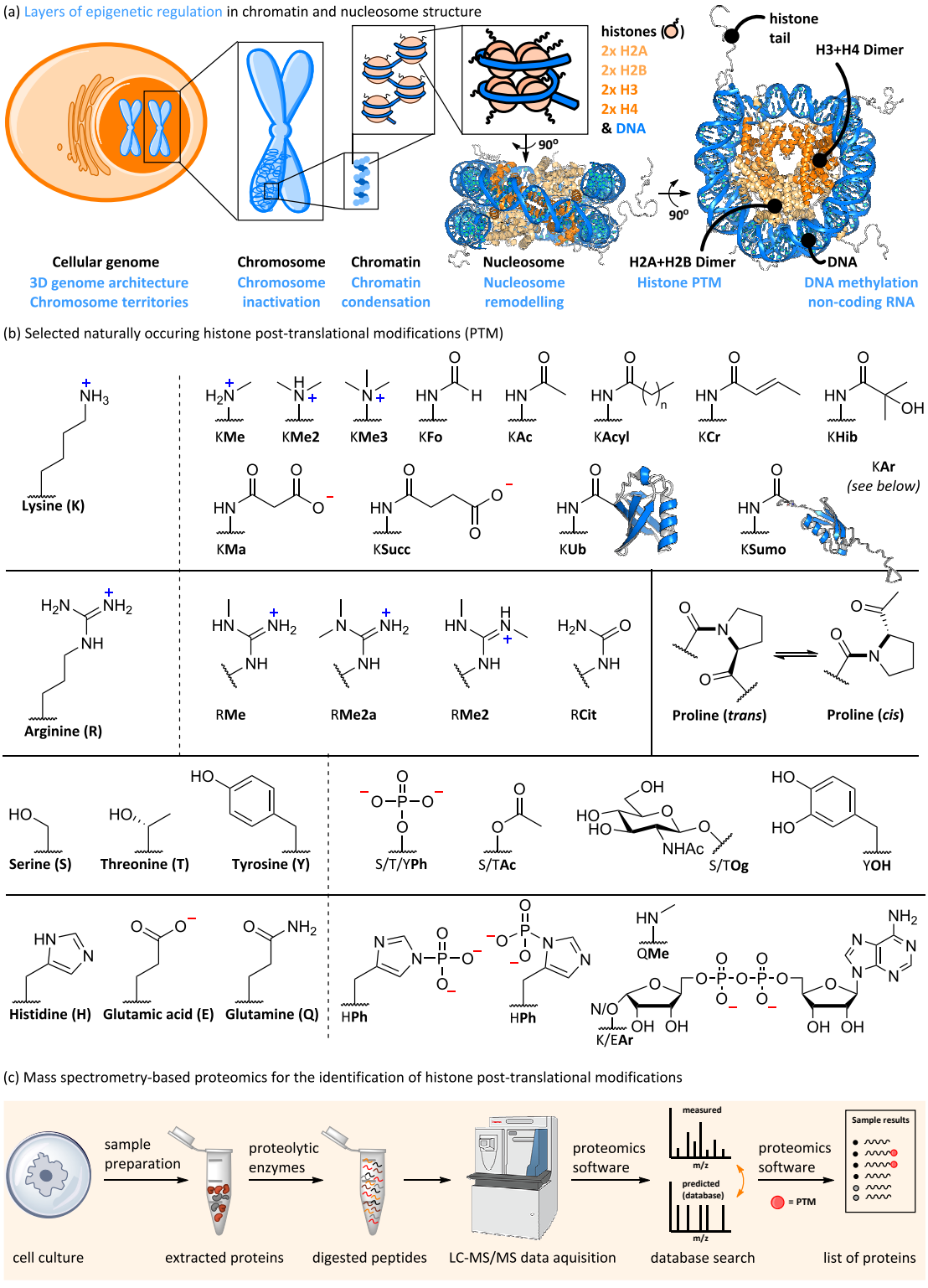


Figure 4.1. Post-translational modification of histones and their identification using mass spectrometry. (a) Cellular DNA is packaged into chromosomes, chromatin and nucleosomes. This genome organisation allows for several layers of epigenetic regulation,

Figure 4.1. (***continued***) one of which is the post-translational modification (PTM) of histones. PDB: 1KX5. **(b)** Selected histone PTMs, listed by amino acid site of modification, illustrate the broad chemical diversity. List adapted from[443]. **(c)** Overview of a bottom-up tandem mass spectrometry workflow for the identification of known PTMs. Figure adapted from[704]. Ac - acetylation, Acyl - acylation, Ar - ADP-ribosylation, Cit - citrullination, Cr - crotonylation, Fo - formylation, Hib - 2-hydroxyisobutyrylation, Ma - malonylation, Me - methylation, Og - *O*-GlcNAcylation, OH - hydroxylation, Ph - phosphorylation, Succ - succinylation, Sumo - SUMOylation, Ub - ubiquitinylation.

In the following, recent developments in chemical chromatin research are highlighted. Starting with mass spectrometry methods to identify histone PTMs, synthetic chemical biology techniques are described, and the application of the generated ('designed') constructs to probe biological function is discussed.

4.2.1.2. Searching for histone modifications using tandem mass spectrometry

Knowing which modification to study is a pre-requisite for building synthetic nucleosomes. Multiple nucleosome forms build up cellular chromatin, including distinct histone proteoforms that can carry post-translational modifications in low stoichiometry.[443] This broad and diverse population distribution poses an analytical challenge for mass spectrometry (MS)-based proteomics: both abundant marks as well as low-abundant modifications have to be reliably identified, requiring the experiments to cover a broad dynamic range.[443]

Beyond the dynamic range issue, histones are notoriously rich in basic residues leading to sub-optimal performance of trypsin (frequent cleavage). This has led to the development of specialised protocols and approaches in bottom-up proteomics to find and analyse histone post-translational modifications.[722] In particular, workflows need to be able to find novel modifications (unknown mass shift on a given peptide) and allow the determination of an empirical formula as well as chemical properties − ultimately leading to the PTM structure.[722]

In this context, the complementarity of mass spectrometry-based proteomics and chemical biology techniques has allowed both fields to benefit from each other. As such, MS-based techniques have been useful in testing the efficacy of established chemical biology techniques. As an example, it allowed probing of the deleterious effects of unnatural tags incorporated by the cellular machinery (e.g. azidohomoalanine for covalent enrichment of histones[269,723]) on cell-cycle progression.[724] Conversely, chemical biology approaches enable improvement of MS acquisition methods. Libraries of histone peptides carrying diverse PTMs have allowed improved calibration of MS detection response

factors.[725] Moreover, the development of selective enrichment techniques for modified peptides, such as phosphopeptide enrichment with TiO_2 immobilised metal oxide affinity chromatography (MOAC), further improved the identification of rare histone PTMs.[726] These examples highlight the complementary nature of both MS-based proteomics and chemical biology techniques in histone PTM research.

4.2.1.3. The chemical biology toolkit

Upon identification and validation of a modification, the elucidation of the biological role of a given histone PTM can follow several lines of investigation: biochemical assays, comparative experiments, bioinformatic correlation etc. One approach is to synthesise histones with defined PTMs and assemble these into synthetic nucleosomes or chromatin as a probe to study specific biological questions. Three main chemical biology strategies have been developed for this purpose: genetic code expansion, protein semi-synthesis and post-translational chemical mutagenesis.[704]

Genetic code expansion

Non-sense suppression mutagenesis is a powerful technique for introducing single unnatural amino acids (uAA) into cellular proteins, harnessing the cellular biosynthetic machinery. A stop codon, typically the constitutionally rare amber stop codon (TAG), is re-attributed to the uAA by incorporating both a cognate tRNA and an engineered aminoacyl tRNA synthetase (*o*-aaRS) into the host cell. The engineered synthetase recognises the cognate tRNA, carrying the amber codon, and the uAA onto it. The cellular biosynthetic machinery incorporates the uAA in response to the amber codon into proteins being synthesised (Figure 4.2.a).[727,728]

Several drawbacks limit the scope of this technology.[729] Both the orthogonal aaRS and the tRNA need to be orthogonal to the host cell biosynthetic machinery, ensuring that the tRNA is not accepted by any constitutive synthetase and that the engineered synthetase only loads the orthogonal tRNA with the uAA.[729] Engineering an aaRS-tRNA pair is thus a fine balancing act between creating orthogonality to the host cell while maintaining compatibility and recognition by the biosynthetic machinery (i.e. the ribosome).[730,731] This has intrinsically limited the chemical diversity of that can be reached by genetic code expansion.[732] A more detailed discussion, beyond the scope of this thesis, can be found in the mentioned review and references therein.[704]

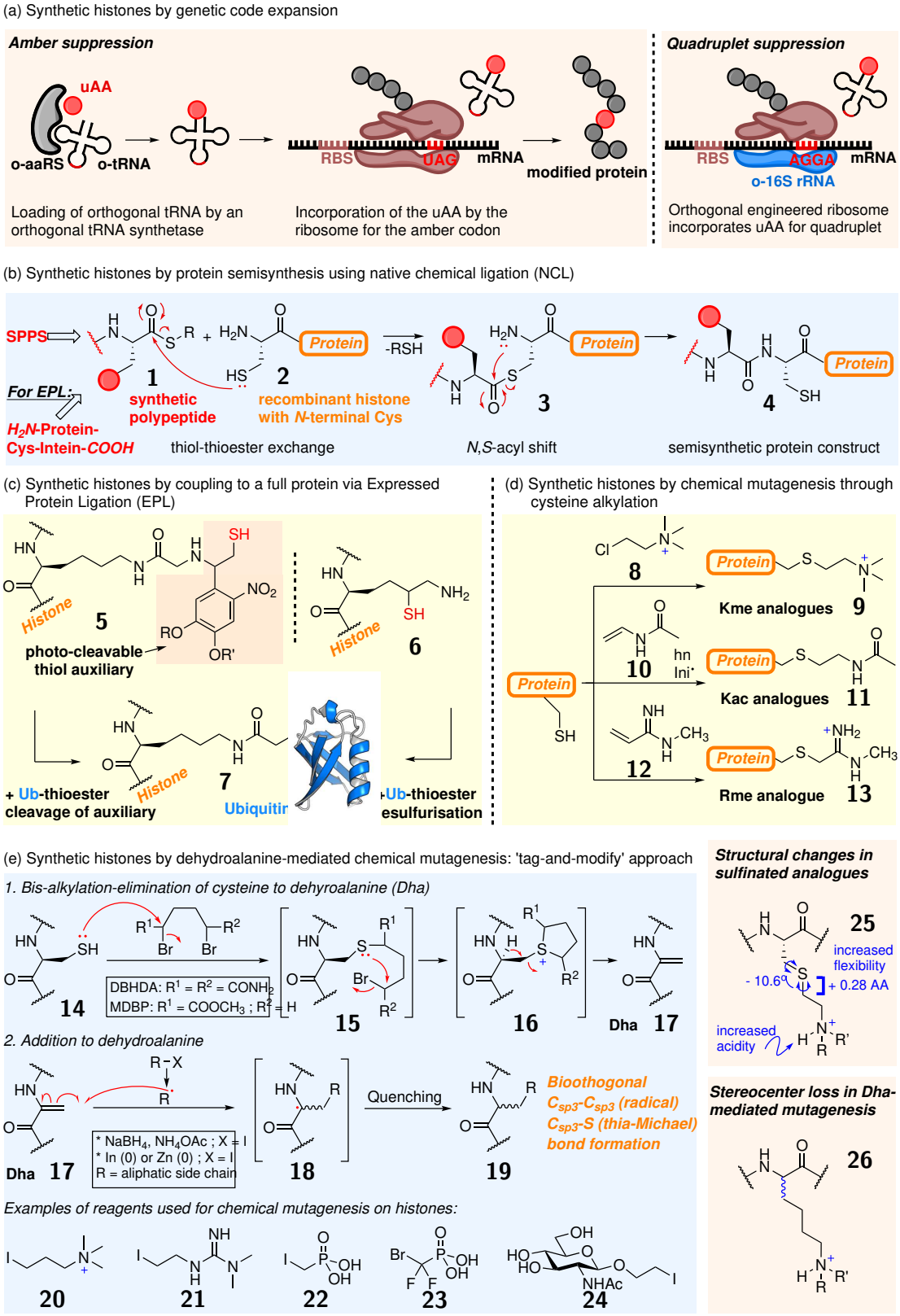


Figure 4.2. The chemical biology toolkit for synthetic histone modification.

Figure 4.2. (***continued***) **(a)** Genetic code expansion. Amber codon (TAG) suppression is the most widely used technique for targeted genetic incorporation. The use of an engineered ribosome with an orthogonal 16S ribosomal unit for quadruplet codon recognition has also been described. Adapted from[731]. **(b)** Mechanism of native chemical ligation (NCL) for protein semisynthesis. In this approach, a protein with a *C*-terminal thioester can be coupled to a protein with an *N*-terminal cysteine. The *C*-terminal thioester can be generated either via expressed protein ligation (see below) or by solid-phase peptide synthesis using appropriate resins. **(c)** Expressed protein ligation (EPL) can be used to couple proteins, shown here exemplarily with ubiquitin, to peptides/proteins containing thiol auxiliaries **5** or amino acid δ-mercaptolysine **6**. UV-irradiation or desulfurisation, respectively, can make these approaches traceless. Adapted from[717]. **(d)** Historical methods for cysteine alkylation to generate thioether-containing mimics. These mimics however lead to structural changes (see frame below, adapted from[719]). **(e)** Post-translational mutagenesis using the 'tag-and-modify' approach. An expressed protein with cysteine **14** at the site of interest can be chemically mutated in a first step to dehydroalanine (Dha, **17**) using either the DBHDA or MDBP reagents. In a second step, addition to Dha **17** is performed, by a multitude of chemistries. Here, the radical addition is exemplarily illustrated and reagents for this reaction indicated. Both methods typically generate a mixed D/L-configuration at the α-carbon. *o*-aaRS - orthogonal aminoacyl tRNA synthetase, uAA - unnatural amino acid, RBS - ribosome binding site.

Protein semisynthesis

The discovery of native chemical ligation (NCL) has unified the world of solid-phase peptide synthesis with the world of recombinant protein expression, enabling the synthesis of small proteins through elaborate synthetic schemes (Figure 4.2b-c).[718,719] The study of the small histone proteins and their chemically diverse post-translational modifications has been a key driver for the development of protein semisynthesis[712,733,734] – indeed total synthesis of histone proteins has been achieved, allowing control over each residue in the protein chain beyond the canonical amino acids.[735–738] It remains nonetheless a highly technical field, although very powerful, requiring in-depth synthetic knowledge. One of the great strengths of protein semisynthesis is the possibility to incorporate diverse modification on several sites of the peptide, such as multiple glycosylations,[739] fully-modified histone tails[735] or ubiquitin modification[740,741]. In a biological context, a large number of histone PTMs exert their full effect only in a concerted manner, for example by cooperative recruitment of transcription factors.[708] Semisynthetic strategies are thus particularly valuable to assess the multi-site PTM crosstalk that might constitute the *histone code*. A more detailed discussion, beyond the scope of this thesis, can be found in the mentioned review and references therein.[704]

Post-translational mutagenesis

Post-translational mutagenesis or chemical mutagenesis refers to the altering of protein residues by performing chemoselective chemical protein modifications reactions, which allow for the creation or conversion of amino acids (Figure 4.2d-e).[742–746] Among the proteinogenic amino acids, cysteine stands out by virtue of its unique pK_a allowing generation of the nucleophilic thiolate anion at physiological pH.[747] The reactivity of this soft nucleophile provides an entry point for reactions modifying isolated proteins in a biochemical setting − that is a pH range, temperature and aqueous buffers compatible with retention of native protein function.[747] Moreover, cysteine's low abundance in proteins ensures accessibility of synthetic strategies building on cysteines to a broad range of proteins. This is especially true for histones: only H3 contains a cyteine (Cys110), which can be mutated to alanine without loss-of-function.[748] As a result, cysteine chemistries are currently central in the chemical modification of histones.[704] In the following, innovations in the field of chemical protein modification are illustrated using synthetic histone PTMs as example, aiming at highlighting how synthetic protocols have become more elaborate and effective as mimics over time.

Early modification studies harnessed the nucleophilic character of cysteine for S_N reactions with haloalkyl reagents **8** to generate methyl-lysine ('thialysine') analogues **9**.[749] Similarly, conjugate addition with **12** was suggested for the generation of methyl-arginine analogue **13**[750] and the acetyl-lysine analogues **11** could be synthesised via strain-promoted substitution[751] or thiol-ene reaction[752] using **10**. Nonetheless, computational studies have suggested that the thioether linkages in thialysines and other thioether linked groups introduce structural distortions that may impact on their recognition (**25**), calling for further sophistication in the synthetic methodologies.[753]

One approach to allow the incorporation of a broad range of chemistries, including near native residues, is the 'tag-and-modify' approach.[754] The discovery of the efficient and selective conversion of cysteine **14** to dehydroalanine **17** via bis-alkylation **16** followed by elimination[755] enables to transform a nucleophilic residue, which has limited scope in reactivity considering the constraints of biochemical environments, to an electrophile that is accessible to a much-extended range of chemistries.[756] Conceptually, this could be described from a retrosynthetic point of view as an *'on protein' Umpolung*. Dehydroalanine is accessible for C−S,[755] C−Se,[757] C−N[758] and C_{sp3}−C_{sp3} bond[759] formation. This reactivity was expanded recently to allow for photoredox-driven incorporation of reactive side chains (chapter 5).

Post-translational chemical mutagenesis is a direct, economical and convenient method to synthetically incorporate constitutionally native PTMs into histones.[760] However, a set of limitations have prevented the widespread use of this technology. Whereas some reaction types on dehydroalanine are accessible to standard biochemistry laboratories using commercial reagents (Dha formation,[756] thiol-ene,[748] aza-Michael[758]), many do require synthetic equipment (protein-compatible photoreactor or glovebox) and access to high-resolution mass spectrometers. For many post-translational modifications (such as phosphorylation[761] or glycosylation[762]), post-translational mutagenesis is the method of choice in cases where a single modification is to be incorporated.[704] However, the incorporation of different modifications via this method is not yet practicable, neither of aromatic residues. To some extent, regioselective formation of Dha can be achieved exploiting local site accessibility or reactivity, thus enabling stepwise functionalisation.[763] The use of other precursors to dehydroalanine beyond cysteine can also be envisaged, generation of Dha from phosphoserine,[764] phenylselenocysteine[765] or alkylselenocysteine[766] occurs under different reaction conditions. The implications of the stereocenter loss at the C_α have not yet been explored.[767]

In conclusion, a plethora of synthetic possibilities exist for the incorporation of post-translational modifications of histones. Post-translational mutagenesis is a convenient choice for the incorporation of single modifications, the combination of methods allows access to many possible modified histones.[704]

4.2.1.4. Application of chemical biology tools for the study of chromatin biology

Post-translational modifications modulate the structure and function of histones in a time- and space-resolved context via two fundamental mechanisms:[707,708] changes in nucleosome structure (and therefore chromatin structure),[768] as well as recruitment of 'reader' or 'effector' proteins.[709,769] The latter can lead to downstream signalling, recruiting of further interaction partners or modification of adjacent histone side chains. In contrast to the use of native nucleosome and chromatin templates, bearing undefined mixtures of histone marks, synthetic nucleosomes enable precise investigation of single modifications or defined combinations.[704]

Probing the modulation of chromatin structure

Histone PTM can perturb the histone fold or histone-DNA interactions. Particularly charge-modulating PTMs such as acetylation, phosphorylation and ADP-ribosylation can alter electrostatic interactions between the histone proteins and the DNA wrapped around, leading to modulation of nucleosome structure and stability (Figure 4.3.a).[768] As an example, acetylation of lysine residues is generally believed to inhibit stabilising interactions between positively charged lysine residues and the negatively charged phosphates of the nucleosomal DNA, leading to a 'relaxed' chromatin state.[770] This has been explicitly illustrated in the case of H3-Lys64, using synthetic acetylated nucleosomes.[771] The effect of phosphorylation on nucleosome structure appears to be more complex: phosphorylation of H3-Thr118 leads to apparent formation of two non-canonical nucleosome structures − so-called 'nucleosome duplex' and 'altosome complex' − whose exact physiological role remains unclear.[772]

Beyond charge-modulating histone PTMs, neutral modifications can also regulate chromatin architecture. GlcNAcylation at H2A-Thr101 appears to disrupt histone-fold interactions and destabilises the nucleosome structure, as was revealed by the use of synthetic nucleosomes.[773] Acetylation or ubiquitination of H4-Lys91 perturbs the dimer-tetramer interface of the histone octamer by steric interference.[774,775] Noteworthy, small modifications such as dimethylation of H3-Arg42[776] or H3-Lys79[777] lower nucleosome stability and stimulate transcription, presumably through the loss of hydrogen bonds between histone proteins and the DNA. In summary, chemical biology methods are helpful in elucidating the structural effects of histone PTMs on chromatin dynamics.[704]

4.2.1.5. Probing interaction partners and downstream spreading of modifications

Synthetic histone proteins and designer nucleosomes have also allowed identification of binding partners in pull-down experiments (Figure 4.3.b). In their seminal contribution, Bartke *et al.*[778] used designer nucleosomes with synthetic H3K4me3, H3K9me3 and H3K27me3 for affinity enrichment of nucleosome-binding proteins from HeLa nuclear extract and identification using SILAC-based quantitative MS (see 1.3.5.4 for background on quantitative proteomics). This approach enabled the identification of preferential binding partners for individual or well-defined combinations of epigenetic marks in an unbiased manner[778] and was subsequently expanded to oligo-nucleosome templates, again using SILAC-based MS.[779] Progress in quantitative proteomics has been fuelled since then by new experimental protocols, as well as the ever-increasing of high-resolution

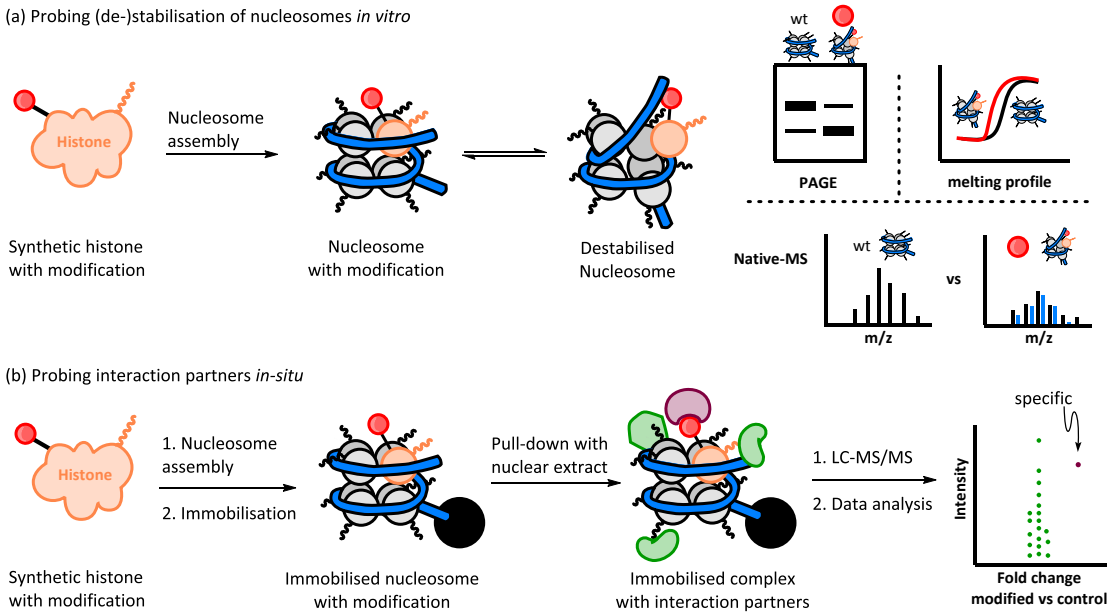


Figure 4.3. Application of chemical biology tools for understanding chromatin biology *in vitro*. **(a)** Synthetic nucleosomes can be used to probe the effect of modifications on nucleosome stability and chromatin architecture. Note that in most studies, 'designer' nucleosomes contain two modified synthetic histones per octamer (removed here for clarity). **(b)** Synthetic histones can be assembled with biotinylated DNA to a designer nucleosome that can be immobilised on a solid support. These can be used for affinity enrichment of interaction partners in pull-down experiments with a label-free quantitative (LFQ) mass spectrometry readout.

mass spectrometers in sensitivity, sequencing speed/efficiency and mass accuracy and improved software for data analysis (see 1.3.5.4). As a results, MS-based proteomics has become a mainstay for identification of chromatin-binding proteins, complementing traditional Western Blot-based methods.[443] More recently, a label-free quantitative (LFQ) approach using synthetic GlcNAcylated nucleosome has allowed the identification of subunits of the FACT complex as interacting protein partners for nucleosomes GlcNAcylated at H2B-Ser112[780] and led to a suggested updating of prior molecular mechanisms for GlcNAcylation-driven transcriptional activation from previous data based uniquely on GlcNAcylated H2B protein.[781] This case study highlights that the nucleosomal context is more representative and synthetic biology techniques should aim at building such higher order structures.[704] In this spirit, this thesis chapter presents an attempt at identifying interaction partners of H2B-Ser36 GlcNAcylation and phosphorylation using synthetic nucleosomes.

4.2.2. *O*-GlcNAc is part of the histone code – the case of H2B-Ser36

The *O*-GlcNAc modification consists in the attachment of an *N*-acetylglucosamine to serine and threonine hydroxyl moieties. Protein *O*-GlcNAcylation plays an important role in epigenetics.[782] Similarly as phosphorylation, it is both a post-translational modification on histones as well as a signalling mark to regulate chromatin-binding factors and chromatin modifying enzymes.[783] The presence of O-GlcNAc in the histone code has been first reported 10 years ago,[784] both as a flagpost cellular signal and to regulate histone or nucleosome half-life.[783] As seen in Table 4.1, several sites of histone O-GlcNAcylation have been reported to date; in some of these, a direct competition with phosphorylation is suspected.[785] Noteworthy, O-GlcNAc can be further phosphorylated at its C-6 position – although this has not yet been reported in the context of histone PTMs.

Transcriptional activation Although the study of *O*-GlcNAcylation as an epigenetic mark is at its very beginning, several principles have already emerged. Without doubt, *O*-GlcNAc is a major regulator of transcription. It is enriched at promotor and transcription start sites on chromatin[792], with a bias towards genes associated with PIP3 signaling, hexosamine biosynthesis, and lipid/carbohydrate metabolism[792] – genes that are linked to insulin-like signalling, metabolism, ageing, stress, and pathogen-response pathways.[792] Several indices indicate that the *O*-GlcNAc modification could have a preparative role in the complex biological process that is the assembly of the transcription initiation complex (i.e. a 'prepare for action' meaning). The C-terminal domain (CTD) of RNA Polymerase II is heavily O-GlcNAcylated during transcription initiation and assembly of the transcription machinery,[793] before being replaced by phosphorylation (via *O*-GlcNAc removal through OGA) for the elongation and actual mRNA translation process[794] (phosphorylation would have an 'active enzyme' meaning). *O*-GlcNAc as a histone PTM could have conceptually a similar function on select sites: H2AT101g destabilises the nucleosome and facilitates transcription[773], H2BS112g interacts with the Facilitates Chromatin Translation (FACT) complex.[780] Interestingly, H2BS112g is reduced when the energy levels are low (i.e. protein synthesis undesired) through an indirect regulation via OGT by AMPK,[789] the cell's AMP/ATP ratio sensor. The joint control mechanism of phosphorylation and glycosylation appears sensible from a metabolic point of view: phosphorylation depends on ATP, indicative of cellular energy levels, and GlcNAcylation on UDP-GlcNAc, indicative of the nutrient supply. Protein synthesis is an energy- and nutrient-expensive process and should not go ahead if cells are not in the metabolic position to do so.

Table 4.1. Sites of *O*-GlcNAc glycosylation in human histones. Reported sites in the literature are annotated for their functional roles. In cases where phosphorylation is known to occur at the same site, the functional role of phosphorylation is annotated.

#	Histone	Site	PTM	Function
1	H2A	Ser40	GlcNAc	Regulates the function of mouse trophoblastic stem cells[786]
2	H2A	Thr101	GlcNAc	Located on the *C*-terminal tail, at the interface H3·H4 tetramer with H2A·H2B dimer.[784] Destabilises the H2A·H2B dimer and facilitates transcription.[773]
3	H2AX	Ser139	Phospho.	Marker for DNA double-strand break[787] (see chapter 2)
			GlcNAc	Counteracts phosphorylation, negatively regulates DNA damage response[509]
4	H2B	Ser36	Phospho.	Phosphorylated by AMPK, nutrient-sensitive[788] Facilitates transcription of p53 target genes[788]
			GlcNAc	Direct interaction with DNA, separates tail from core. Unknown function[784]
5	H2B	Ser91	GlcNAc	unknown[781]
6	H2B	Ser112	GlcNAc	Promotes the ubiquitination of H2BK120 via BRE1A/B[781] Interacts with the FACT complex, facilitates transcription[780] Regulated by AMPK in a feedback loop with OGT.[789]
7	H2B	Ser123	GlcNAc	unknown[781]
8	H3	Ser10	Phospho.	Phosphorylated when cells enter mitosis. Regulated by Aurora B and PP1.[790]
			GlcNAc	Competes with H3S10 phosphorylation[791] Regulates H4K16ac and G2-M phase checkpoint[791]
9	H3	Thr32	Phospho.	Phosphorylated when cells enter mitosis.[790]
			GlcNAc	Inhibits the mitosis-specific phosphorylation of T32, S28 and S10.[790] Hinders the G2-M cell cycle transition[790]
10	H4	Ser47	GlcNAc	Changes during mitosis. Indirect interaction with DNA. In loop 1 of histone fold[784]

Whereas glycosylation and phophorylation appear to be transient and highly dynamic marks – unsurprisingly, given the lability of their chemistry – they can induce the writing of other histone PTM that signal more permanent changes. As an example, the H2BS112g modification has been shown to lead to H2BK110ub,[781] which is a well-known transcription-activating mark inducing a stark change in chromatin organisation.[795]

Cell cycle progression *O*-GlcNAc post-translational modification of histones influences cell cycle progression in an antiplay with phosphorylation. Both phosphorylation on H3S10p and H3T32p is necessary for cells to enter mitosis.[790] However, the

GlcNAcylation of these sites prevents G2 to M transition in the cell cycle. In comparison to the above-described interplay between phosphorylation, where glycosylation represent an intermediary assembly state between an undefined 'off' and phosphorylation as 'on' or 'go' signal, the cell cycle progression seems to have a mechanism where glycosylation represents an 'off' state, which could in theory also be of short-term nature ('wait') or a longer-term instruction ('sleep'). A more clear-cut on/off system is used for the spatial localisation of DNA damage by the interplay between H2AXS139p and H2AXS139g, where DNA damage activates the rapid signalling response by phosphorylation and the slower glycosylation subsequently counteracts, leading through this dynamic equilibrium to a spatial distribution between the two antagonistic modifications.[509]

H2BS36 – a potential nutrient-specific signalling? A central paradigm in research on *O*-GlcNAcylation is that the dependency on UDP-GlcNAc makes this PTM indicative or related to the nutrient status of the cell. Whereas hints for nutrient-regulation exist for several *O*-GlcNAc sites on histones[784], a direct connection through a molecular mechanism is missing so far. H2B-Ser36 however is known to be phosphorylated by AMP-activated kinase (AMPK),[788] thus this epigenetic mark is set when energy levels are low, and decreased in response to increasing glucose levels (*negative* correlation).[788] This modification has been further shown to occur in the promotor region as well as in transcribed regions of p53 target genes, suggesting a link to tumour-suppressing genes responsible for growth regulation.[788] Interestingly, Fujiki *et al.* have demonstrated a *positive* correlation between increased glucose levels and increased protein *O*-GlcNAcylation.[781] In light of these findings, a potential antagonism between the two modifications occupying the same site can be hypothesised, which would be dependent on the nutrient or energy status of the cell. This could in turn suggest a speculative 'three-state' response in a nutrient (e.g. glucose)-dependent manner. The study of this potential competition using synthetic nucleosomes will be the study of this thesis chapter.

4.2.3. Aim of the project

As mentioned previously, the aim of this chapter is to understand the role of glycosylation on H2B-Ser36, and investigate its relationship to phosphorylation on the same site. Histone modifications achieve their functional outcome either by altering nucleosome structure or by recruiting of interaction partners.[707] Whichever role the modification has depends on its location in the nucleosome structure and its accessibility.

H2B-Ser36 is located on the lateral surface of the histone octamer, in proximity to the nucleosomal DNA.* It constitutes a delineation point between the histone core and the protruding *N*-terminal tail of histone H2B and is located in the histone H2B repression (HBR) domain (residues 30-37 in yeast).[797] This highly basic, conserved domain has been reported to be implicated in DNA-templated epigenetic processes, such as DNA repair or gene expression.[797] Although this remains to be confirmed, the fact that the modification site lies in a loop protruding away from the octamer, the likelihood of this modification to destabilise the octamer structure is small. However, an effect on the nucleosome structure via the interaction with the nucleosomal DNA cannot be excluded, either through charge repulsion (introduced by phosphate groups) or by steric repulsion or through a change in loop conformation (both possible by the introduction O-GlcNAc). Therefore, the structural analysis suggests that only the assembly of synthetic nucleosomes would be able to reveal a structural change, if any.

*Ser36 corresponds to numbering from human UniProt sequence that includes the initial tripeptide PEP, which is the common numbering used in the literature for histone PTMs. In the core histone numbering, used more frequently for cross-species comparison, *N*-terminal methionine and initial tripeptide PEP of human H2B are not numbered. As such, H2B-S36 this would be position Ser33 in the core numbering convention, i.e. in the *Xenopus laevis* crystal structure.[797]

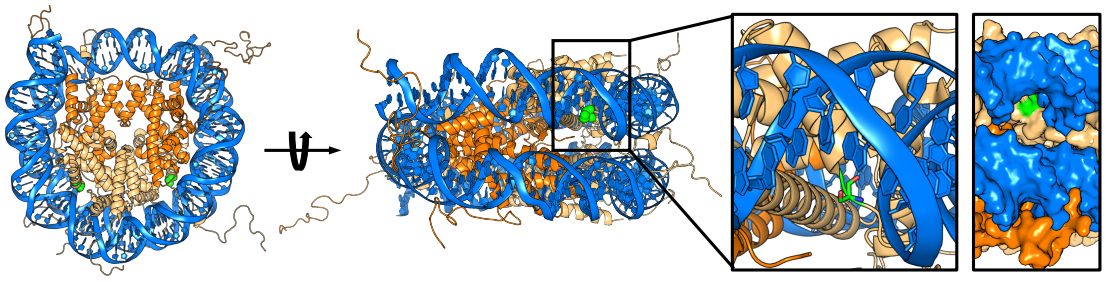


Figure 4.4. Crystal structure of the *Xenopus laevis* nucleosome highlighting the modification site. H2B-Ser36 is shown in green, the H2A-H2B dimer in beige, the H3-H4 tetramer in orange and the DNA in blue. H2B-Ser36 is located in proximity of the major groove of the DNA and is solvent-accessible. PDB 1KX5.[796]

The H2B-S36 position is however accessible to effector proteins. The generation of 'designer nucleosomes' carrying a defined synthetic PTM mimic allows for affinity enrichment of interaction partners to probe the interaction partners of a histone modification.[704,748,759,773,780] Different techniques can be used for the analysis of pull-down experiments, yet mass spectrometry stands out by its ability to reveal novel interaction partners in an unbiased manner without prior knowledge of the sample content.

Therefore, an approach was chosen to assemble synthetic nucleosomes of wildtype, or carrying a synthetic mimic of glycosylation or of phosphorylation. All three different nucleosomes are to be immobilised via a biotinylated DNA and used in pull-down experiments. Label-free quantitative mass spectrometry analysis is then used to identify possible interaction partners, as described in Figure 4.5. This project was originally started by Dr. Ritu Raj in his DPhil, who contributed one replicate of each pull-down experiment to this thesis.[798]

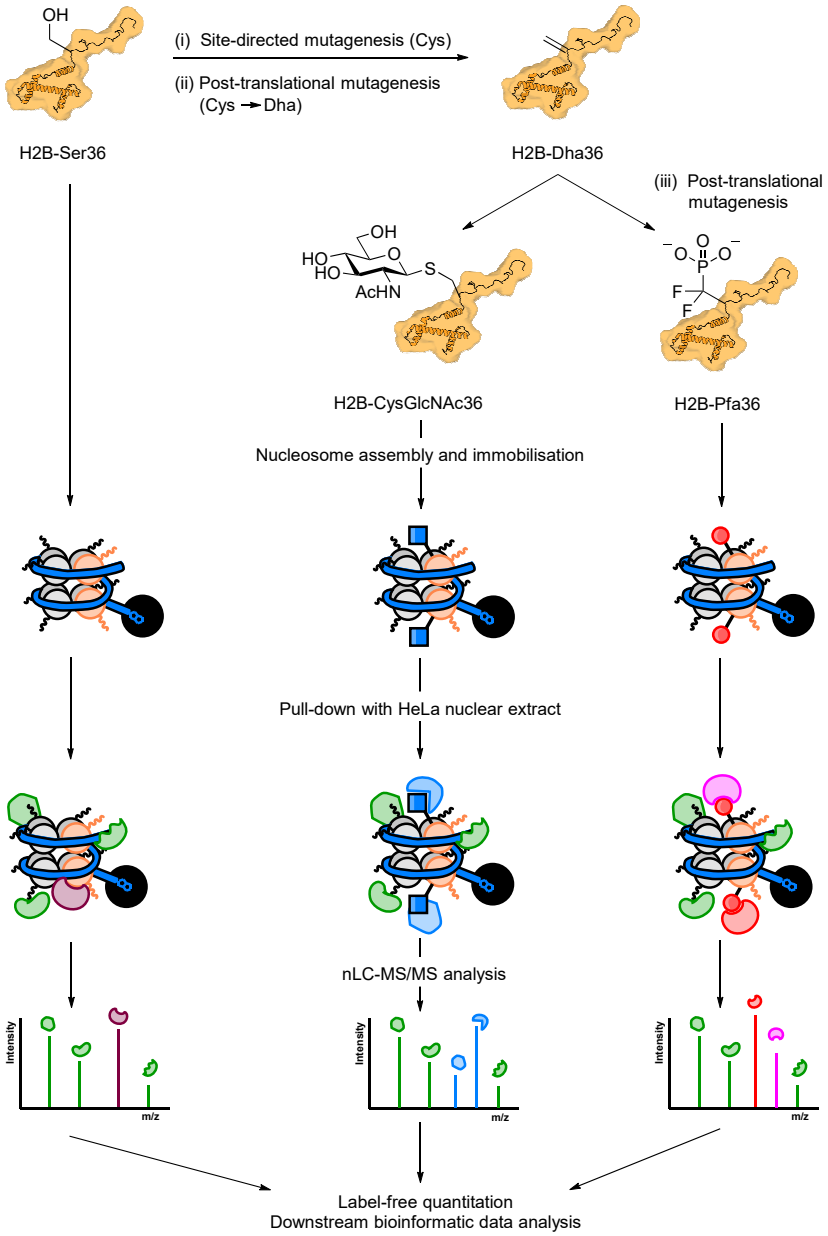


Figure 4.5. Probing the PTM-specific interaction partners of histone modifications at H2B-S36. Three different H2B variants are generated. Recombinant protein expression gives Ser36 (wildtype, left) or Cys36, which is converted by post-translational mutagenesis to H2B-Ser36 variants with glycosylation mimic (H2B-CysGlcNAc36, middle) or phosphorylation mimic (H2B-Pfa36, right) via the common dehydroalanine (H2B-Dha36) precursor. The three H2B variants were assembled with other canonical histones and biotinylated 186 bp DNA containing the Widom 601 positioning sequence into synthetic nucleosomes and immobilised on a solid support. Each nucleosome was incubated with HeLa nuclear lysate for pull-down of interaction partners. The samples were digested and analysed by LC-MS/MS separately. The MS data files were combined and processed together, identified binding partners were quantified by label-free quantitation (LFQ).

4.3. Results & Discussion

4.3.1. Synthesis of post-translationally modified histone H2B-Ser36 variants

Synthetic approaches have the advantage of generating well-defined, homogenous constructs able to precisely mimic post-translational modifications.[704] One approach termed 'tag-and-modify'[754] is to recombinantly express a protein with cysteine at the site of interest, which can selectively be converted to dehydroalanine (Dha).[755] Dehydroalanine is a functional 'tag' which can be modified by a broad range of protein-compatible chemistries.[756]

This 'tag-and-modify' strategy[754] was adopted in this work. *Xenopus laevis* H2B-Cys36 was recombinantly expressed and purified following literature procedure. Histones from *Xenopus laevis* have been the dominant recombinant proteins used for *in vitro* nucleosome studies.[797] † Conversion of cysteine to dehydroalanine can be achieved via a bis-alkylation-elimination mechanism using either the 2,5-dibromohexanediamide (DBHDA)[755] or the commercially available methyl 2,5-dibromopentanoate (MDBP) reagent.[758,799] The latter has the advantage of shorter reaction times and does not lead to overalkylation of lysine residues, yet requires careful monitoring to ensure sulfonium elimination.[756] Here, MDBP was used to convert H2B-Cys36 to H2B-Dha36 in denaturing conditions and the reaction was monitored by intact protein LC-MS. Full conversion was achieved after 1.5 h incubation at 37 °C. In addition to LC-MS, the sample was analysed by non-reducing SDS-PAGE for tracking aggregation and by MS/MS to confirm site-specificity of modification. H2B-Dha36 was used as an intermediate for the subsequent generation of H2B-Ser36-SGlcNAc, H2B-Ser36SPO$_4$ or H2B-Ser36Pfa.

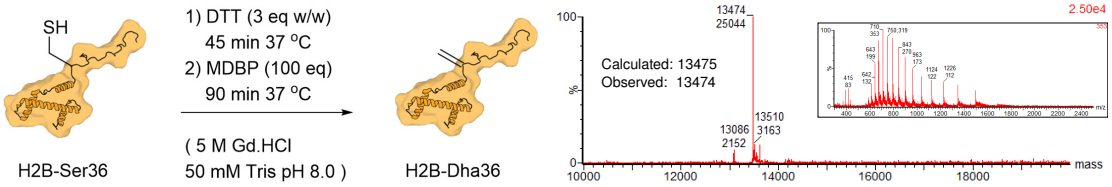


Figure 4.6. Synthesis of H2B-Dha36 from H2B-Cys36. The deconvoluted denaturing intact protein LC-MS spectrum confirms the correct mass with an insert showing the molecular ion series. The full characterisation can be found in 4.6.6.

† Histones from *Xenopus laevis* and *Homo sapiens* are highly similar. H2A differs in three positions (Human vs Xenopus): A10T, S40A, I87V. H2B differs in the N-terminal PEP in humans and 7 positions (*h.s. vs x.l.*): A18T, K27R, R28K, S35A, V36I, G57S, I66V. H3 differs only in G102A and H4 is identical.[797]

Thio-GlcNAc is a close structural mimic of *O*-GlcNAcylated serine,[800] differing only in a single atom. *S*-GlcNAcylated histone protein can by synthesised from H2B-Dha36 by a thio-Michael addition to dehydroalanine.[748] The Dha-containing protein was reacted with 400 equivalents of GlcNAc-thiol under denaturing conditions at 37 °C. Full conversion to the desired product was observed after 2 hours by LC-MS. The protein was desalted and lyophilised for later octamer assembly.

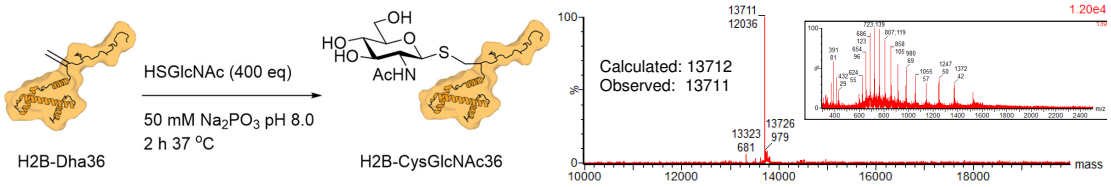


Figure 4.7. Synthesis of H2B-CysGlcNAc36 from H2B-Dha36. The deconvoluted denaturing intact protein LC-MS spectrum confirms the correct mass with an insert showing the molecular ion series. The full characterisation can be found in 4.6.6.

With GlcNAcylated protein in hand, the focus shifted to the generation of phosphorylated mimics of pSer. Phoshoproteins are at the core of several signal transduction pathways, controlled by the interplay between protein kinases and phosphatases.[362] Phosphorylation has been mimicked in different ways: in its most rudimentary form, a mimic was a negatively charged amino acid such as glutamate. However, carboxylates are poor physiological mimics,[801] having not only different pKa values (pKa ≈ 4 for COOH, vs. $pK_{a1} \sim 2$, $pK_{a2} \sim 6$ $OPO(OH)_2$)[802] but also insufficient charge (-1 vs. -2) and structural mimicry (trigonal planar vs. tetrahedral).[801] This approach has been subsequently replaced by the use of phosphorylated proteins isolated from biological samples or the use of *in vitro* enzymatic reactions. Both methods have the disadvantage of producing heterogeneous mixtures.[801, S1][759, Fig.5A] The access to pure, homogeneously phosphorylated proteins has been enabled through progress in synthetic approaches. Phosphoserine[803–806] (and phosphothreonine[730]) has been incorporated by genetic code expansion, however this requires specialised orthogonal aminoacyl-tRNA synthetase/tRNA pair and has typically low yields of the modified protein. Protein total synthesis using native chemical ligation schemes also provides a route to phosphorylated proteins, however suffer from the limitations of SPPS such as β-elimination of phosphate under basic conditions[761] and is limited when modification sites are distant from either protein termini. In cases where no natural cysteine occurs in the protein, such as with histone H2B, post-translational chemical mutagenesis via Dha is an easily accessible alternative. Two different mimics of phosphorylation can be incorporated by these means, thiophosphate[748] and phosphonodifluoromethylene alanine (Pfa)[759].

Initially, the generation of thiophosphate-modified H2B-Ser36 was attempted. Initial trials[748] by Ritu Raj incubated H2B-Dha36 at 37 °C and completion was found after 1 hour. However, these results could not be reproduced, neither by Ritu Raj nor myself, observing additional peaks by LC-MS along starting material and the desired product. Other group members having similar issues, it was speculated that this could be related to the quality of the reagent. ^{31}P-NMR investigations of commercial batches of sodium thiophosphate from either Sigma-Aldrich or Alfa Aesar identified the thiophosphate peak at 36 ppm and additional 'impurity' peak in the Sigma-Aldrich batch. This led to the recommendation of using the Alfa Aesar chemical and Ritu Raj obtained promising initial results with a Alfa Aesar batch of sodium thiophosphate using longer incubation times (overnight) at 25 °C, yielding >90% conversion.

However, remaining reproducibility issues lead to a different approach to avoid batch-to-batch variation and the resulting difficulties in optimising a precise synthetic protocol. Indeed, the colour of the batches varied: batches with grey coating were found empirically to produce poor results whereas reagents containing pure white powder performed generally speaking better. To overcome this limitation, a trituration protocol followed by drying of the reagent in a Kugelrohr was developed. The resulting white powder was analysed by ^{31}P-NMR, showing the thiophosphate peak at 36 ppm and a small impurity remaining at 19 ppm. Using this purified reagent, Ritu Raj could obtain conversion when treating H2B-Dha36 with 0.4 M sodium thiophosphate, 2 M Gd.HCl pH 8.1 at 37 °C for 2-3 hours. However, I could not reproduce this result using several different batches of purified thiophosphate, obtaining conversion ranging from 62 % to 81 % and systematically observing a +36 peak, which could not be annotated. MS/MS analysis identified peptides with sulfur dihydroxylation (+31.99) as well as further peptides with cysteine oxidation to cysteic acid (+47.98) and sulfur dioxide (+63.96). Hydrolysis of thiophoshate at longer increased temperatures is mentioned in the literature[807], which could also explain why Ritu Raj observed better results when the reaction was performed at 25 °C. Indeed, slight yellow colouring of the reaction mixture was observed upon longer incubation times at 37 °C. Best results were obtained with short incubation times, followed by desalting and re-incubation with fresh reagent. Another possible avenue could be to ensure neutral pH in the solution after addition of sodium thiophosphate. However, full conversion could not be achieved.

Having not been able to further optimise the synthetic procedure beyond 81 % conversion, it was thus tried if the synthetic phosphoprotein could be enriched by immobilised metal affinity chromatography. For this, two 1 mL Ni-NTA columns were discharged and the metal replaced to generate Ga(III)-NTA and Fe(III)-NTA 1 mL columns. Ga(III)-

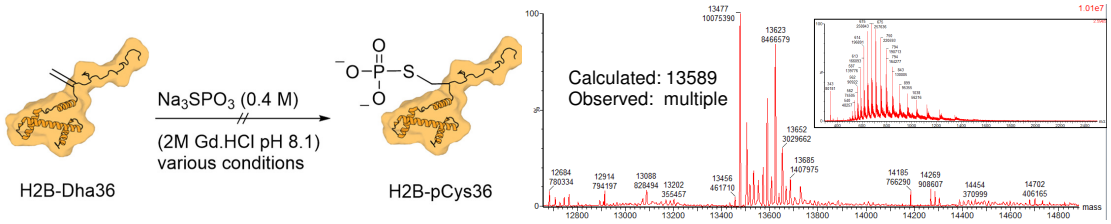


Figure 4.8. Attempted synthesis of H2B-pCys36 from H2B-Dha36. The deconvoluted denaturing intact protein LC-MS spectrum shows several proteoforms with distinct modifications. The full characterisation can be found in 4.6.6.

NTA is reportedly better at enriching for proteins,[808] multiply charged and acidophilic peptides whereas Fe(III)-NTA enriches preferably Pro-directed and basophilic phosphopeptides.[809] The enrichment of phospho-IMAC is inefficient on peptides at pI extremes (<4 and >11).[810] The sequential combination of the two columns, first Ga(III)-NTA followed by Fe(III)-NTA, reported in the literature[809], was tried on an AKTA FPLC system eluting with increasing concentration of sodium phosphate. The result showed that an enrichment is indeed possible in principle, with a small amount of phosphoprotein being retained on the column. However, the bulk of the protein, exceeding by far the conversion observed by MS, was not retained. Potentially, this could be improved by using a mixed resin loaded on gravitational column, instead of using gradient elution. Being not able to reach desired purity (ideally > 95 %) by the reaction nor by an adequate purification, it was those decided to move to an alternative mimic of phosphorylation: H3B-Pfa36.

The chemistry of dehydroalanine has remarkable versatility: Dha can both behave as an electrophile in 2e$^-$reactions, as well as a radical acceptor (SOMOphile) in 1e$^-$reactions.[756] Pioneering work has illustrated the orthogonality of 1e$^-$reactions to biological substrates[759,764] (see also chapter 5). This enables the "*on protein*" formation of carbon-carbon bonds to introduce various post-translational modifications.[759] One example of the application of this strategy for C-C bond formation is the installation of phosphonodifluoromethylene alanine (Pfa) residue on histone H3 with Ser10Pfa mutation, thus creating a functional mimic of a phosphorylated protein recognised by antibodies against H3-pSer10.[759, Fig.5] In addition to its stability to hydrolysis,[759] Pfa has been suggested to be a more suitable alternative to phosphoserine than thiophosphate, based on its better mimicry of pK_a,[811] polarity and shape.[812]

H2B-Pfa36 was synthesised from H2B-Dha36 using (bromodifluoromethyl)phosphonic acid as phosphate source and radical precursor and $NaBH_4$ as hydride donor, in a glove box at 4 °C to avoid presence of oxygen leading to protein cleavage.[759] Full conversion

was observed after 15 min by LC-MS, the reaction was found to be highly reliable.

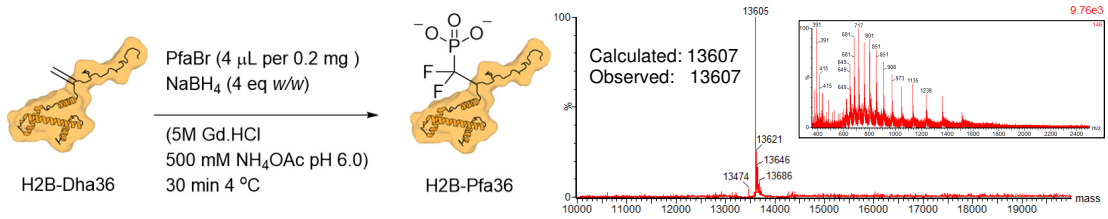


Figure 4.9. Synthesis of H2B-Pfa36 from H2B-Dha36. The deconvoluted denaturing intact protein LC-MS spectrum confirms the correct mass with an insert showing the molecular ion series. The full characterisation can be found in4.6.6.

4.3.2. Tandem mass spectrometry of modified proteins

In light of the lability of phosphorylation and glycosylation as modification in collision-induced fragmentation methods (i.e. CID and HCD), the site of modification was confirmed by LC-MS/MS using ETD-based methods on an Orbitrap Elite: HCD-pd-ETD for H2B-CysGlcNAc36 and a CID-ETD data-dependent decision tree for H2B-pCys36 and H2B-Pfa36. ETD spectra could be acquired for all modified peptides.

The ETD spectra of H2B-CysGlcNAc36 and H2B-pCys36 are particularly interesting. ECD/ETD-fragmentation of *S*-glycosylation has been previously studied among others[61] by Medzhiradsky and co-workers,[63] which noted an increased stabilty of *S*-GlcNAc compared to *O*-GlcNAc in CID − most likely because of the ability of the sulfur to better stabilise the H^+. The authors also observe significant SGlcNAc losses from the charge-reduced precursor ions.[63]

The ability of thiols, thioether and disulfides to participate in ETD fragmentation is literature-known.[280,281] Mechanisms involving either initial electron capture at sulfur or amide have been proposed.[280] For the latter, the fragmentation from the $z^\bullet$-ion or the z+1 ion forming a w-ion with N-terminal vinyl amide has been described and recently confirmed with the FELIX IR-laser[813] (see Figure 4.10.A). Elimination of the thiolated neutral loss is only observed for $z^\bullet$-ions with a cysteine residue at the cleavage site, independent of its position of cysteine in the peptide.[813] Interestingly, we observed different types of w-ions in the H2B-CysGlcNAc36 and H2B-pCys samples: the thioglycosylated sample shows formation of w+1 ions (w') whereas the thiophosphorylated sample shows formation of the regular w ion (Figure 4.10). There are three possible explanations for this observation. One possibility is that the different modifications influence the preferability of z' vs z ions formation. This would suggest that the phosphate would encourage c-ion dissociation from the intermediate complex carrying the proton (see Figure 1.12) whereas the SGlcNAc favour proton retention. The second explanation would be that the thiophosphate could dissociate from both the z and the z', but would abstract a proton while leaving the z' ion. The w-ion fragments are observed in all spectra and replace the corresponding z-ion in the ion series (see ion mass list tables in Figure 4.10). The third possibility is that z to z' ion population can increase over time when supplemental activation is used, which might influence the w ion pattern over time.

Similarly as described by Medzhiradsky and co-workers, we also observe the formation of charge-reduced precursor ions with loss of SGlcNAc (-236.0593). The charge state of 1^+ is consistent with our expectations: the precursor ion has a charge state of 3^+, the reduced precursor of 2^+ and the loss of SGlcNAc or thiophosphate could easily result

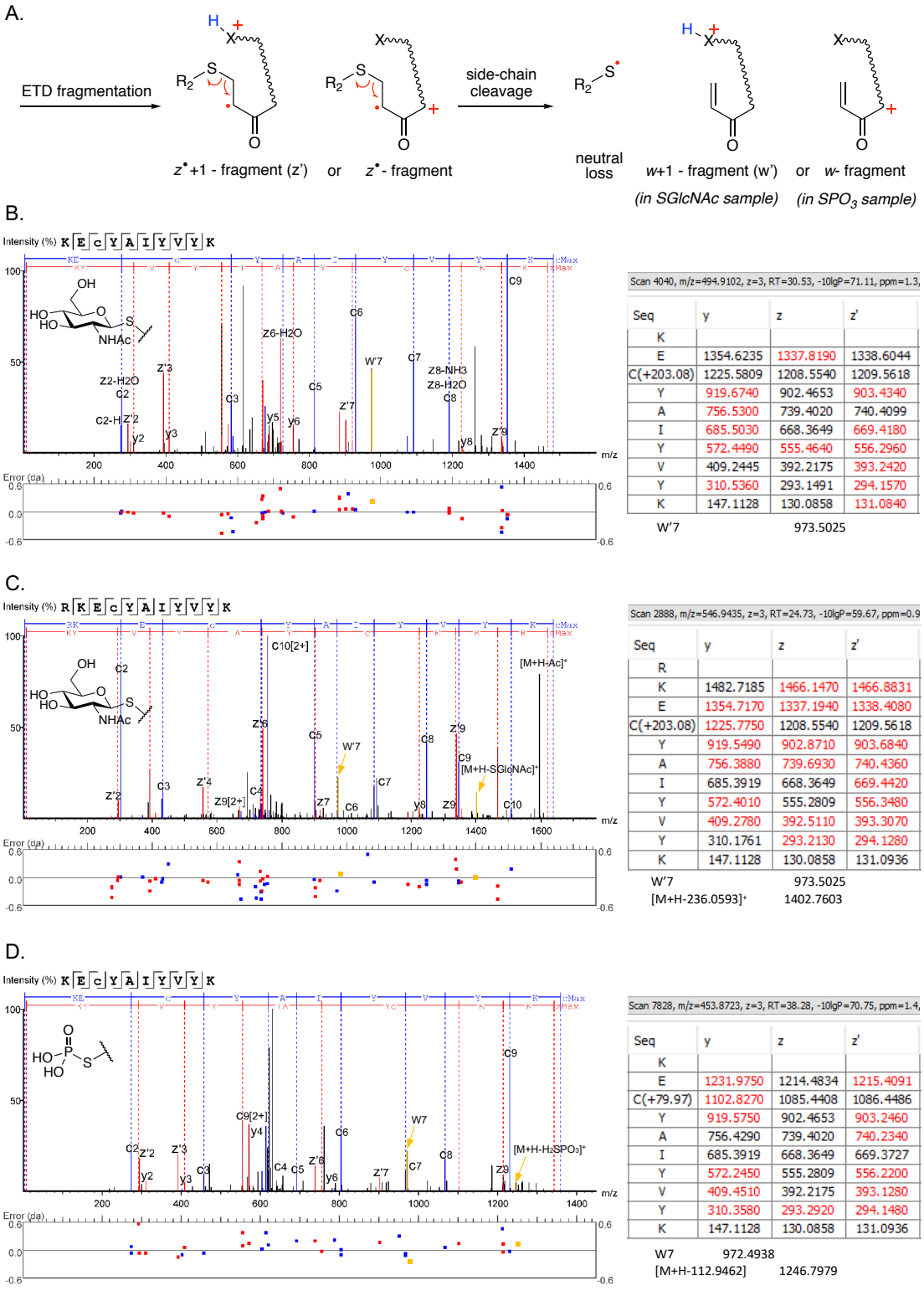


Figure 4.10. ETD MS² spectra of sulfur-linked modifications show diagnostic *w*-ions. A. Mechanism for side-chain fragmentation of sulfur-linked modifications following

Figure 4.10. (***continued***) electron capture by the peptide amide bond as described in the literature.[280,281] **B.** MS^2 spectrum of H2B-CysGlcNAc36 with 1 missed cleavage confirming site-localisation of the modification and the w_7' ion. **C.** MS^2 spectrum of H2B-CysGlcNAc36 with 2 missed cleavage. This spectrum was chosen as it highlights the charge-reduced precursor ion with loss of thioGlcNAc. Also, the ion mass list table shows z and z' ions for almost all but the Cys-GlcNAc amino acid fragment ions. Instead, the w_7‘ ion is prominent, highlighting the specificity of this diagnostic ion. **D.** MS^2 spectrum of H2B-pCys36 confirming site-localisation of the modification. In this sample, the w_7 ion is identified, although all amino acids but pCys yield z‘ ions. The charge-reduce precursor ions is less prominent in this spectrum.

in abstraction of a proton. However, this charged-reduced precursor ion fragment is not observed in all spectra.

In summary, our LC-MS/MS confirms the identified modifications and for sulphur-linked modifications, diagnostic ions are identified in the ETD MS^2 spectra.

4.3.3. Reconstitution of histone octamers with H2B-Ser36 variants

Histone octamers can be assembled from the individual histone proteins by stepwise dialysis (Figure 4.11.A).[814] The wildtype (wt) H2B-Ser36 and modified H2B-Pfa36 or H2B-SGlcNAcSer36 proteins were each mixed in equimolar ratios with the remaining canonical wt histones (H2A, H3 and H4) under denaturing conditions. Octamer assembly allows for refolding by dialysis against high-salt buffer (2M NaCl). The high salt concentration is essential to keep the octamer stable.[814] The resulting mixture of sub-complexes were separated by size-exclusion chromatography (SEC, Figure 4.11.B). The SEC trace for all octamers was highly similar, implicating no major disruption of the histone fold occurred upon modification. The retention volume for the wt octamer (66.5 ± 2.0 mL, $n = 4$) and the *S*-GlcNAcylated octamer (67.1 ± 1.2 mL, $n = 4$) match literature values, the Pfa octamer (69.2 ± 1.1 mL, $n = 4$) is slightly higher, which could be due to the charges on the phosphate leading to a slightly more compact structure in solution. Similarly, a Dha-containing octamer was assembled, removing DTT from the buffers to avoid side reaction. Pleasingly, the Dha-containing octamer showed the same SEC profile (64.7 ± 3.4 mL, $n = 2$), indicating Dha does not disturb the histone fold (traces of all octamers are shown in 4.6.8). The presence of all four histones in the octamer was confirmed by SDS-PAGE (18 % TGS gels for low MW resolution) and intact protein LC-MS.

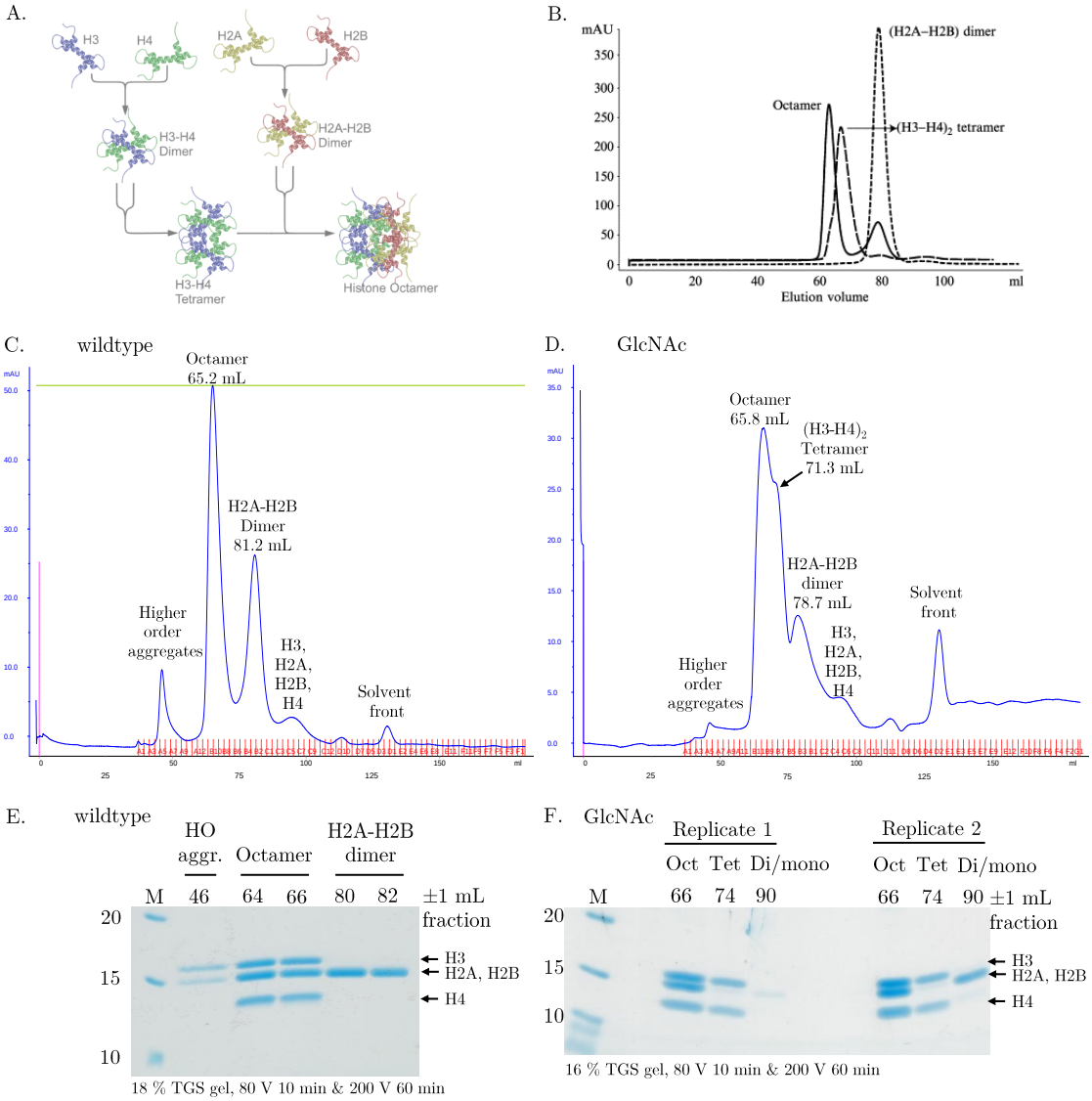


Figure 4.11. Histone octamer assembly and their purification by size-exclusion chromatography (SEC). A. Mechanism of histone octamer assembly from the individual histones. H3 and H4 assemble readily in a $(H3\text{-}H4)_2$ tetramer, to which 2 H2A-H2B dimers sequentially associate. Modified from Wikimedia Commons with permission, ©Richard Wheeler (Zephyris). **B.** Elution profile of histone subcomplexes from a Superdex S200 16-600 column. The histone octamer (solid line, MW 108.5 kDa) elutes first at ∼65 mL. The $(H3\text{-}H4)_2$ tetramer (dashed line, MW 53 kDa) elutes close to the octamer at ∼72 mL. H2A-H2B (dotted line, MW 27 kDa) elutes last at ∼80 mL. Reproduced from[814]. **C.** Annotated SEC trace of wildtype octamer (S200, 16-600). **D.** Annotated SEC trace of glycosylated octamer (S200, 16-600), showing elevated amounts of $(H3\text{-}H4)_2$ tetramer. The presence of tetramer is observed only in the glycosylated samples, with varying extend. **E.** Coomassie-stained Tris-glycine SDS (TGS)-PAGE gel resolving the histones present in the different fractions (2 mL fractions, named by midpoint). **F.** TGS-PAGE gel confirming the components of the octamer and tetramer fractions. Further characterisation of all octamers is shown in 4.6.8.

When pooling fractions, those that could contain tetramer were avoided to keep the stoichiometry as accurate as possible. Figure 4.11.C-D shows the SEC trace and the gel of the fractions of two exemplary assemblies, one with and the other without $(H3\text{-}H4)_2$ tetramer. Single histones can be resolved using adequate gel chemistries (18 % tris-glycine for example), allowing to confirm the presence of all histones in the octamer fractions, whereas the tetramer fractions lack the H2A/H2B band (Figure 4.11.E-F). The tetramer can interfere in the later nucleosome assembly. It is observed only in GlcNAcylated samples, however the extend of tetramer formation varies, from a distinct second peak (which in one assembly was the dominant species), to a shoulder such as in Figure 4.11.D to a single octamer peak where the shoulder is only slightly noticeable, leading to a longer tail. This could indicate that the GlcNAcylation would correctly assemble into the octamer, but might be slower than the other modifications.

4.3.4. Generation of Widom 601 nucleosome-positioning DNA

The aim of the later nucleosome assembly is to generate a homogenous sample of mononucleosomes. Positioning sequences of the DNA have been designed and tested to direct histone octamer assembly on the DNA to specific positions. This allows to generate homogenously defined nucleosome substrates for *in vitro* studies. In this thesis, the 'Widom 601' positioning sequence was used in two variants: A shorter 145bp DNA for the generation of homogenous nucleosomes, and a longer 186bp biotinylated DNA that enables immobilisation for pull-down assays via the incorporation of biotinylated Uracil nucleotides.

4.3.4.1. Generation of 145bp DNA and 186 bp biotinylated DNA

The 145bp could be generated by preparative scale DNA purification following the original protocols of Dyer *et al.*[814] yielding 5.7 mg. Generally speaking, the preparation did not encounter major difficulties and later nucleosome assemblies with the 145bp were reliable and successful.

The preparation of the longer 186bp biotinylated DNA however encountered more difficulties, including DNA fragmentation and degradation. This has hampered progress on nucleosome assembly of past DPhil students and myself, retrospectively explaining why most nucleosome assemblies performed in this DPhil failed (> 95 %). Therefore, the focus of the approach presented in this chapter builds on quality control of the reagents

and low-cost methods to increase purity, which were originally developed to troubleshoot the multiple failures encountered.

4.3.4.2. DNA degradation and subsequent purification

One of the main causes of DNA degradation is wrong handling, which can lead to contamination with nucleases, or DNA fragmentations if not handled with care (Figure 4.12.A). Equally, stored small DNA pieces can degrade in different ways, when kept over longer periods of time (i.e. several months, see Figure 4.12.B). According to guidance from Integrated DNA Technologies Inc, oligonucleotides are most stable when stored in TE buffer (such as Tris/EDTA 10/0.1), followed by dry storage and then storage in nuclease-free water. In the early stages of this DPhil, DNA was erroneously stored in MilliQ water at -20 °C – a practice not recommended, which can reportedly lead to depyrimidination, deamination and hydrolytic cleavage. This process is further accelerated by the presence of trace metals and demetalation by chelating agents is often suggested. Depending on the size of the DNA fragment, storage at -20 °C or -80 °C in appropriate buffers or ethanolic solutions has been proposed.

The quality of the reagents is quintessential for the success of the later nucleosome assembly. In particular, low molecular weight DNA fragments interfere in the assembly and falsify the stoichiometry of the components. Generally speaking, three different strategies exist for ensuring high quality of the DNA for nucleosome assembly – the knowledge was gathered retrospectively by speaking to the respective former DPhil students. Lukas Lercher and Ritu Raj generally tried to make the biotinylated DNA freshly before assembly, ensuring DNA has not had time to degrade. Sarah Faulkner performed titration series to compensate for degradation products by using a higher DNA concentration.

In this thesis, the approach chosen was to first develop an analytical procedure for quality control allowing to separate the 186bp DNA from the degradation products (Figure 4.12). Simple high percentage agarose gels allow separation of fragmented DNA into two strands (3.5 % agarose gel, 30 min at 100V; or 4% agarose, 50 min at 50 V) or the separation of DNA and low MW degradation products (3 % agarose, 30 min 100V).

This simple analytical gel was then used to develop a purification strategy for 186 bp DNA in which degradation has occurred, monitoring the removal of degradation products. Three approaches were evaluated:

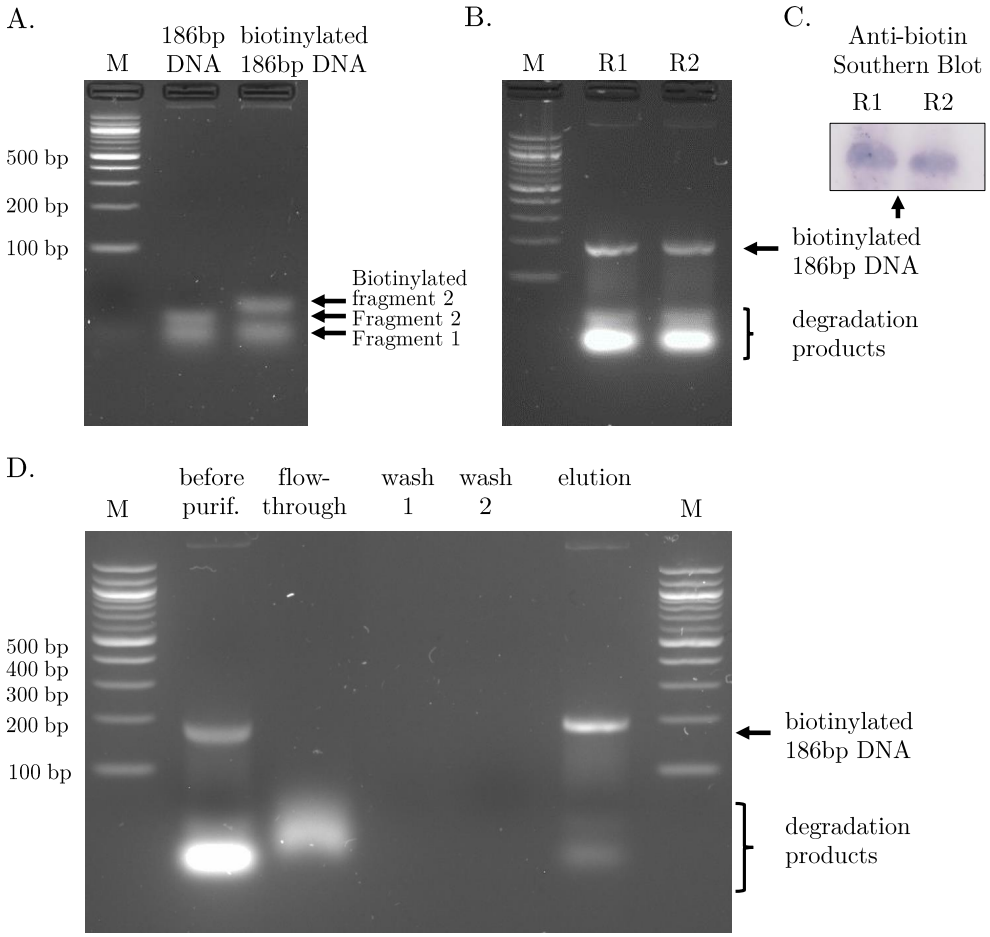


Figure 4.12. Quality control of 186bp Widom 601 DNA preparation and observed forms of DNA damage. A. Wrong handling of DNA (shearing, vortex etc.) can lead to fragmentation, analysed here by 4 % TAE agarose gel (50 V 50 min). **B.** Longer storage of DNA in MilliQ water (slightly acidic pH) at -20 °C can lead to a slow acidic hydrolysis of DNA. This can be remediated by storage in TE 10/0.1. 3 % TAE agarose gel (35 min 100 V). **C.** Successful biotinylation can be confirmed by Southern Blot followed by anti-biotin chromogenic detection. **D.** DNA with degradation product can be purified using the Zymo 100 µg PCR kit followed by EtOH precipitation. 3 % TAE agarose gel (25 min 100 V).

- Ethanol precipitation: Nucleic acid precipitation can be used to concentrate and/or purify nucleic acids. Longer nucleic acids are less soluble in ethanol than in water, addition of salt or low temperature further decreases solubility. A 2 M ammonium acetate final concentration was used, reportedly helpful for the removal of dNTPs. However, analysis of the purified pellet showed that the degradation products were still present.

- Size exclusion: DNA-grade G-25 sepharose (GE Illustra) is often used for the

removal of nucleotides, dNTPs, desalting or buffer exchange of DNA fragments larger than 10bp. This was tried on SpinTrap scale for both a degraded sample as well as the EtOH-precipitated DNA. The removal of the degradation products was unsuccessful with both samples.

- Anion exchange: PCR purification columns are anion exchangers able to retain DNA fragments of a certain length, typically > 200bp. This was initially tried on a small scale using the Macherey-Nagel NucleoSpin PCR clean-up columns and found to be successful. However, this kit is not ideal, being tailored for small amounts of DNA and longer fragments. As such, a substantial amount of the DNA was not retained in wash steps. The Zymo 100 µg PCR purification kit was then tested, as the manufacturer indicates retention >50 bp, and found to give good, reproducible results. A small amounts of degradation products remains (Figure 4.12.D).

The latter procedure enabled to purify enough DNA for nucleosome assembly. Noteworthy, it allowed to purify DNA from batches which previously failed in nucleosome assembly. It also enabled to purify DNA from batches which previously were used at 3:1 ratios for assembly. Also, it suggests that the degradation products are not single nucleotides, but rather approx. 10-20 nt long oligonucleotides, with comparable lengths to PCR primers.

In addition, a Southern Blot with anti-biotin chromogenic detection based on a HRP-conjugate was developed to confirm successful biotinylation of the DNA (Figure 4.12.C). With reagents of validated quality in hand, the nucleosome assembly could be performed.

4.3.5. Nucleosome assembly

Nucleosomes can be assembled from the octamer and a strong positioning DNA (Widom 601) or from H2A-H2B dimer and $(H3\text{-}H4)_2$ tetramer.[814] In this project, the assembly from the octamer using the salt gradient dialysis method was performed. The octamer and DNA are mixed together in high-salt buffer (2 M KCl) and are dialysed against a series of buffers with decreasing salt concentration. At 0.85 M KCl, the octamer is unstable and the $(H3\text{-}H4)_2$ tetramer binds to the DNA, followed by the H2A-H2B dimer at 0.65 M (Figure 4.13). The assembly was analysed by Native PAGE.

As mentioned previously, the success of the nucleosome assembly is highly dependent on the quality of the reagents used and of their concentrations. To recapitulate, impurities in the DNA such as low-MW degradation products falsify the ratio of reactants and

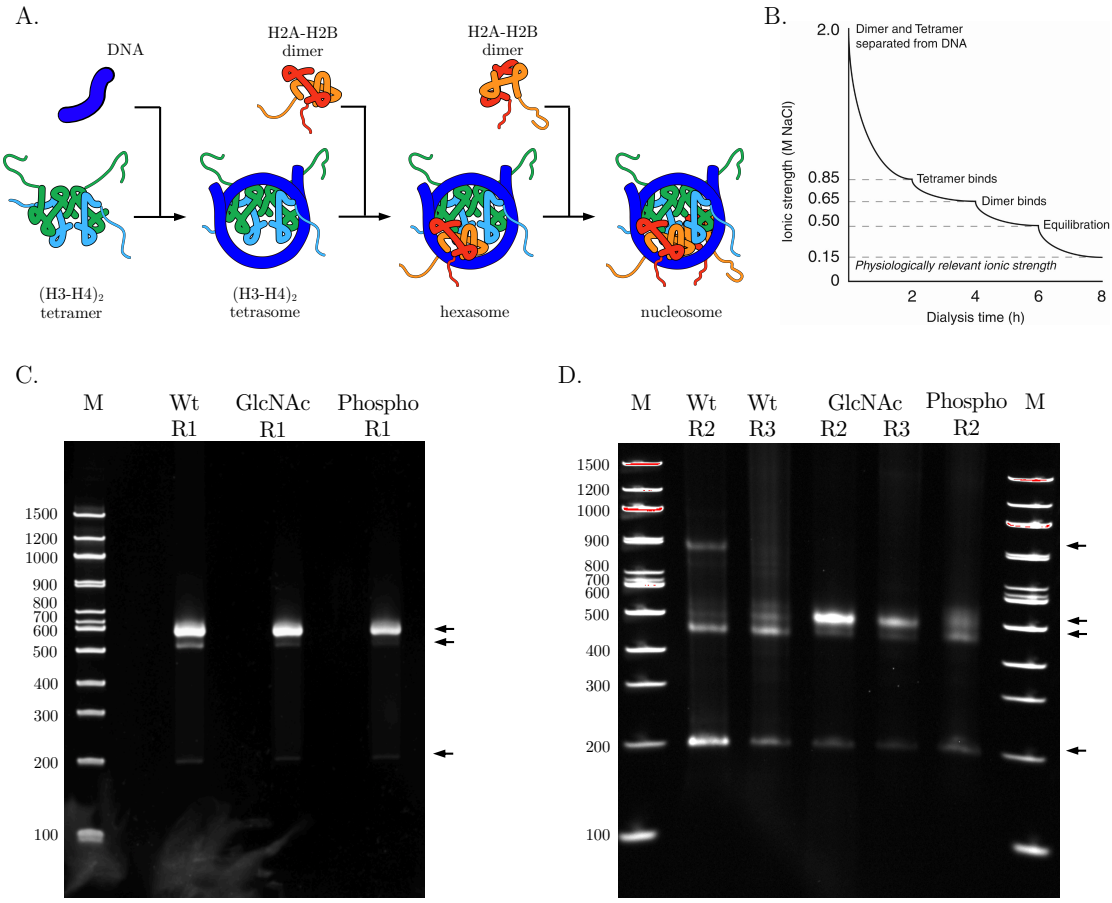


Figure 4.13. Quality control of 186bp Widom 601 DNA preparation and observed forms of DNA damage. A. Mechanism of *in vitro* nucleosome assembly by salt-gradient dialysis. The $(H3\text{-}H4)_2$ assemble first with the DNA. Subsequently, two H2A-H2B dimers associate forming a hexasome as intermediate before the fully assembled nucleosome is generated. **B.** In the salt-gradient dialysis method, the assembly coincides with dialysis steps at defined concentration of KCl. Image reproduced from[797]. **C.** Native PAGE (6 % TBE) of the three nucleosome assemblies used for pull-down experiments by Dr. Ritu Raj, reproduced from[798] with permission. These assemblies provide each one replicate. **D.** Native PAGE (6 % TBE) of the additional replicates (R2, R3) generated in this work. The different bands, corresponding to different species, are discussed in the text.

lead to failure of the assembly − observed as precipitate in the dialysis cassette due to a too high ratio in protein.[815] The presence of tetramer is thought to equally falsify the ratio and hamper the proper assembly.

Small conformational differences of nucleosomes can be resolved by native PAGE with low-salt buffer. Ideally, a unique band corresponding to a single population of nucleosome structures is obtained (Figure 4.13.C). However, native PAGE often reveals additional bands, which can come from several other nucleosome structures (Figure 4.13.D). No po-

sitioning sequence is perfect, nucleosomes with the correct stoichiometry but assembled on different parts of the DNA strands migrate differently in the native PAGE. Counter-intuitively, nucleosomes with central positioning on a specific DNA fragment can migrate more slowly than those with terminal positions.[797] This is due to centrally positioned nucleosomes having DNA projections from both sides of the histone octamer wrapped region and therefore occupy a larger effective hydrodynamic radius.[797] Another explanation for the presence of additional bands can be sub-nucleosomal structures, also named partially assembled nucleosome states (PANS), such as hexasomes lacking one H2A-H2B dimer[816] or tetrasomes lacking both H2A-H2B dimers.[797,817] The fact that homogenous assembly can be achieved (Figure 4.13.C, data from Dr. Ritu Raj) suggests that neither GlcNAcylation nor phosphorylation modification of H2B-S36 has a substantial effect on nucleosome structure − this could potentially be further analysed by variable-temperature CD spectrometry and/or native mass spectrometry.[773]

In the case of Figure 4.13.D, additional bands might be attributed to the quality of the reagents. Due to time constraints at the end of this DPhil, different sources of 186bp DNA were used for the assembly of the second and third replicates. The highest quality of DNA was used for the GlcNAcylated replicates, yielding a mostly homogenous population of nucleosomes with slight hint of an additional, unwanted band. Equally, the wt nucleosomes yielded a mostly homogenous population, with an additional band slightly higher than the main component. The reason for the difference in migration between the GlcNAcylated and Wt sample, not observed in Dr. Ritu Raj's assembly, is unclear. The phosphorylated nucleosome was assembled with a DNA of minor quality, leading to the observation of two different populations.

The assembly of the nucleosomes using the salt-gradient dialysis method is dominated by the $(H3\text{-}H4)_2$ tetramer, which assembles first.[797] By heat shifting, enough activation energy is supplied for the nucleosomes to re-explore their thermodynamic stability by sliding along DNA in a non-dissociative mechanism.[814] A potential avenue to improve the assembly would have been the exploration of a heat-shifting assay to increase the homogeneity.

Although not ideal, the presence of these additional bands should not impact substantially on the results of the pull-down experiments. The presence of false-positive interaction partners is minimised as the heterogeneity prevents their identification as statistically significant. However, low-affinity or low-abundance interaction partners might be missed (false negatives), as the actual concentration of modified H2B-containing nucleosomes is reduced by the possible presence of hexasomes or tetrasomes. This can be compensated to a small extend by adapting sample loading on the mass spectrometer.

4.3.6. Interaction proteomics with synthetic nucleosomes

Histone PTMs mediate their function through associated interaction partners, which can include 'writer', 'eraser' or 'reader' proteins. Interaction partners of wildtype, glycosylated and phosphorylated H2B-S36 were probed *in situ* in an affinity-enrichment mass spectrometry approach. The synthetic nucleosomes were immobilised on magnetic streptavidin beads for pull-down experiments of interaction partners from HeLa nuclear extract (Figure 4.5). Non-specific interaction partners were removed by washing and the remaining proteins digested with trypsin by filter-aided sample preparation (FASP)[155,156] before analysis by nano LC coupled to a hybrid quadrupole-orbitrap tandem mass spectrometer. Three independent biological replicates of this experiment were conducted, each with a distinct preparation of synthetic nucleosomes as well as nuclear extract. One replicate/nuclear extract pair was generated by Dr. Ritu Raj in his DPhil,[798] these samples were analysed by LC-MS/MS on an Orbitrap Elite. The remaining two replicates were analysed on a Q Exactive instrument. The raw data was combined for label free quantitative analysis using the MaxLFQ algorithm[381] implemented in MaxQuant[379], identifying 1784 protein groups. The associated LFQ protein intensities were used for statistical analysis with the Perseus software.[419]

In light of the different preparation and data acquisition of the samples, the experimental correlation among LFQ intensities within group (same synthetic nucleosome bait) and biological replicates were monitored. Figure 4.14.A-B illustrates a high discrepancy between replicates of each group, which can be related to the distinct preparations of the nuclear extracts. The intensity distribution histograms (Figure 4.14.C) further show a low protein content in the Wt_R2 sample. This could not be spotted by TIC and BPI monitoring during data acquisition, suggesting this might not be a mass spectrometry issue but instead a pull-down with low protein content. Generally, excellent correlation coefficients >0.9 were achieved amongst samples originating from the same nuclear extract (Figure 4.14.D). R1 and R3 correlate reasonably well with most Peason coefficients around 0.6. R2 shows lower, but acceptable correlation around 0.4. As noted, the Wt_R2 samples stands out, showing poor correlation due to low protein content.

It is clear there is a stronger correlation between replicates of same nuclear extract than of the same nucleosome constitution in the pull-down experiments. It is therefore challenging to identify proteins that might be higher enriched in a given condition. Only 86 of 1533 data points are contained in all 8 replicates, 543 in half of the replicates and 1001 in at least 2 replicates. Unsurprisingly, when no missing value treatment is performed, no protein is identified as enriched in any of the conditions. Similarly,

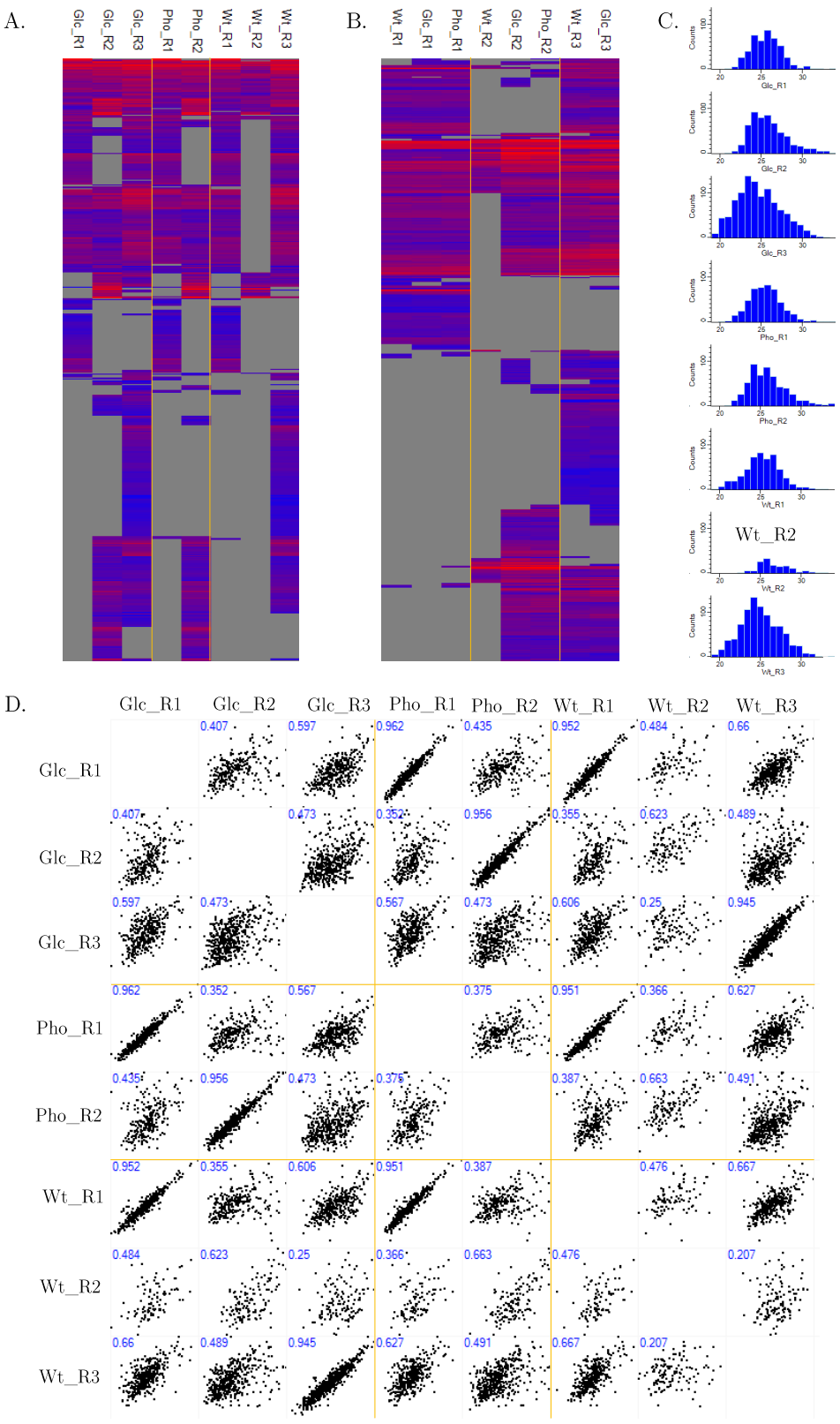


Figure 4.14. Statistical analysis of the biological replicates and individual pull-down experiments. A. Hierarchical clustering by synthetic nucleosome (bait) illustrates

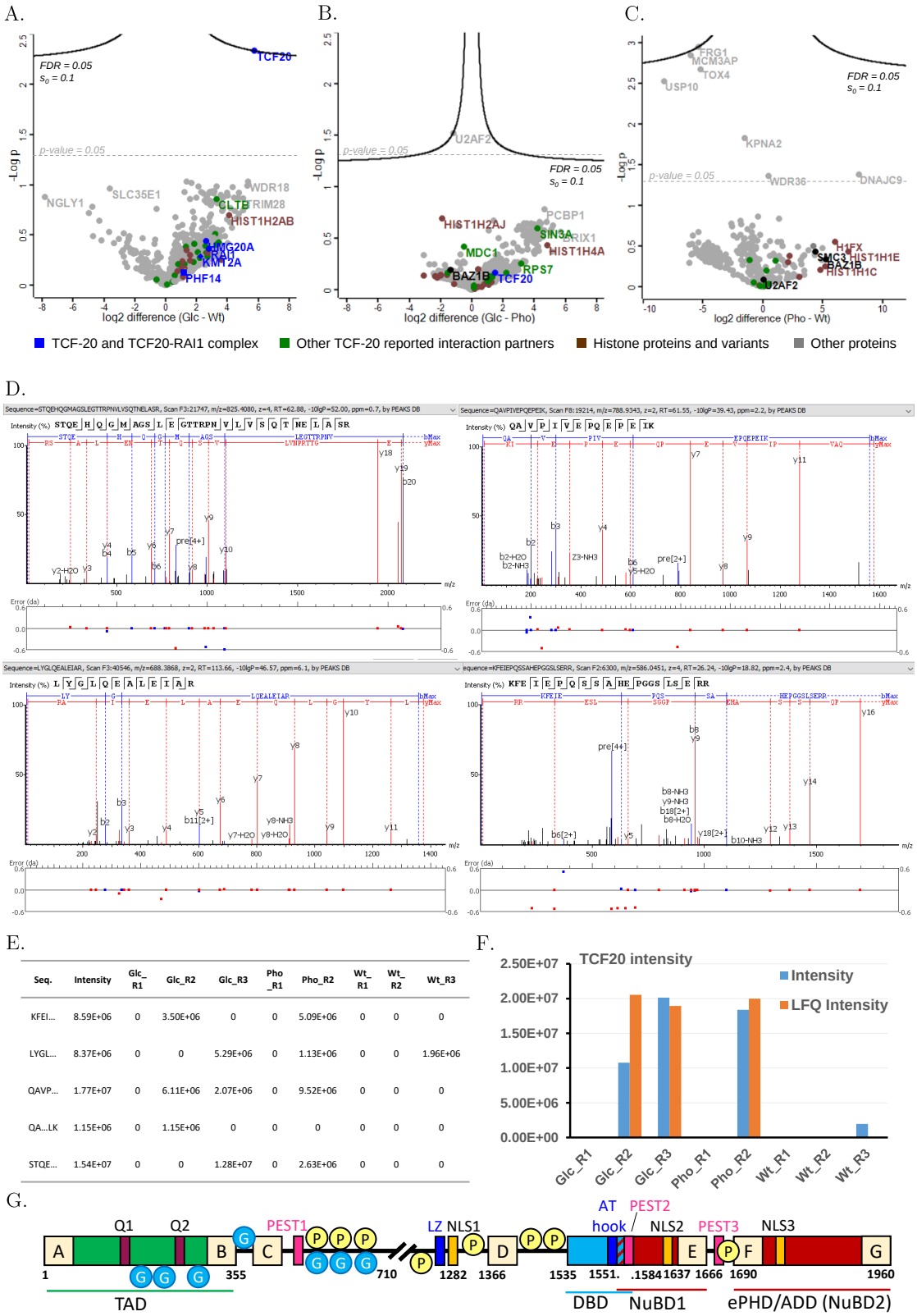
Figure 4.14. (***continued***) poor correlation between pull-downs using the same synthetic probe (= same group) but different biological replicate. Red-blue colouring of LFQ protein intensities, missing values are shown in grey. **B.** Same hierarchical clustering, sorted by biological replicates, highlighting the discrepancy between the three nuclear extract preparations. **C.** Histogram of LFQ intensities, highlighting low protein content of Wt_R2 replicate. **D.** Multi-scatter plot of LFQ protein intensities listing the Pearson correlation coefficient.

upon missing value treatment by constant, no meaningful hits was identified in the dataset containing all pulldowns, proteins appearing as enriched were all artefacts or high-abundant proteins.

AMPK is a known interaction partner of phosphorylated H2B-pSer36 and was an expected positive control for the identification.[788] When searching against the full dataset, AMPK is found only in the Pho_R1 and Glc_R1 replicates, where a single deamidated peptide belonging to the AAPK1 or AAPK2 subunits is identified. The protein is thus labelled by MaxQuant as "identified by site" and filtered out in Perseus − it would most likely not be shown as significant in further statistical analysis due to its presence in single replicates.

In order to have a chance to identify any possibly meaningful proteins in this nuclear extract-discrepant dataset, an alternative approach was chosen. The different synthetic nucleosomes were searched with MaxLFQ pair-wise against each other. This also enables to bypass the issue of having 2 replicates for the phosphorylated sample, by searching the phosphate data sing R1, Wt_R3 and Pho_R2 thus avoiding Wt_R2. Any identification would then have to be manually validated and cross-compared across samples. The results are shown in Figure 4.15.

Using the described pair-wise LFQ search, transcription factor 20 (TCF20) appears in the glycosylated sample as most enriched protein. It is however not identified as significant in a two-sample t-test (using standard parameters) as it is only identified/enriched in 2 of 3 replicates for the glycosylated nucleosomes. TCF20 is a low intensity protein, sowing only 6 % sequence coverage in this experiment and therefore low scores are reported (Protein score 15 in MaxQuant, -10logP = 113 in Peaks). As a consequence, further validation is required to ensure the validity of the result suggested by the statistical analysis. In order to further foster the evidence that the enrichment of TCF20 is not an artefact, the interaction partners of TCF20 were compiled from the IntAct, BioPlex and BioGRID databases and highlighted in green in the volcano plots, or blue for components of the RA1-TCF20 complex.[818] Figure 4.15.A shows that TCF20 interaction partners are almost exclusively enriched in the glycosylated sample. The volcano plot Figure 4.15.C (Pho/Wt) shows no enrichment in any sample.



Seq.	Intensity	Glc_ R1	Glc_R2	Glc_R3	Pho _R1	Pho_R2	Wt_ R1	Wt_ R2	Wt_R3
KFEI...	8.59E+06	0	3.50E+06	0	0	5.09E+06	0	0	0
LYGL...	8.37E+06	0	0	5.29E+06	0	1.13E+06	0	0	1.96E+06
QAVP...	1.77E+07	0	6.11E+06	2.07E+06	0	9.52E+06	0	0	0
QA...LK	1.15E+06	0	1.15E+06	0	0	0	0	0	0
STQE...	1.54E+07	0	0	1.28E+07	0	2.63E+06	0	0	0

Figure 4.15. Synthetic nucleosomes carrying the H2B-Ser36GlcNAc modification preferentially enrich the TCF20 transcription factor. A-C. Volcano plots

Figure 4.15. (***continued***) of the different LFQ search combinations, Glc vs Wt (A.), Glc vs Pho (B.) and Pho vs Wt (C.). 2 replicates were only used for the latter. The transcription factor TCF20 appears as most significantly enriched protein in the glycosylated sample. Known interaction partners of TCF20 are indicated in green, other proteins of interest in black and histone proteins in brown. **D.** MS^2 spectra of key peptides underlying the identication and quantitation of TCF20. Peptides were annotated by the PEAKS X software. **E.** Peptide MS^1 intensities as identified by MaxQuant. **D.** Protein summed intensities (blue) and normalised LFQ intensities (orange) as reported by MaxQuant. These intesities are used for statistical analysis. **G.** Schematic representation of the structure of the TCF20 transcription factor. Beige boxes indicate conserved domain with the RAI1 paralogue. TAD - transactivation domain; DBD - DNA-binding domain; NuBD - nucleosomal binding domain; Q1/Q2 - glutamine rich stretches; LZ - leucine zipper; NLS - nuclear localisation signal; ePHD - extended plant homeodomain; ADD- ATRX-DNMT3-DNMT3L domain; PEST - proline, glutamate, serine and threonine-rich sequence; G - reported site of *O*-GlcNAcylation [GlyGen]; P - reported site of phosphorylation [UniProt]. The AT-hook motif is a DNA binding sequence.

Next, the experimental evidence for TCF20 was inspected. Figure 4.15.D shows the MS/MS spectra for the identified peptides underlying the quantitation. All identifications are plausible and show good to high scores. The intensities of the peptides in each sample, used as input for the MaxLFQ algorithm, are shown in Figure 4.15.E. This illustrates the weak point of this analysis, no peptides for TCF are identified in the R1 replicates, that is the nuclear extract analysed by Dr. Ritu Raj on an Orbitrap Elite. The two other biological replicates unequivocally contain TCF20 as protein. Its peptides are enriched in the glycosylated and phosphorylated sample compared to the Wt. This translates into the Wt sample having a lower intensity, and the MaxLFQ algorithm does not provide a LFQ intensity for the Wt (Figure 4.15.F). As such the extent of enrichment of TCF20 in the phosphorylated sample remains unclear. PEAKS X equally identifies TCF20 peptides in all glycosylated and phosphorylated PD, with higher peptide AUC values for the glycosylated PD. This would suggest that the glycosylated sample strongly enriches TCF20, and the H2B-pS36 sample to a lesser extent. Recognition and binding of both phosphate and *O*-GlcNAc moiety in the same binding site has been reported previously for crystal structures of the 14-3-3 protein, an '*O*-GlcNAc reader' protein.[819]

Further supporting the identification of TCF20 are previous reports of the interaction between H2B and TCF20 in the BioPlex database, based on affinity-capture MS.[430] The schematic structure of the transcription factor is shown in Figure 4.15.G, it includes two experimentally validated nucleosome binding domains (NuBD), one of which is in close proximity to the DNA binding domain, illustrating that this TF could in principle recognise a histone PTM in close proximity to the DNA such as for H2B-S36.

Taken together, TCF20 is enriched by H2B-S36-glycosylated synthetic nucleosomes.

These results remain to be validated by orthogonal methods. Nonetheless, the MS result is plausible for a low-abundance transcription factor and possible from a biochemical perspective.

TCF20 is also known in the literature as SPBP (for Stromelysin-1 PDGF (platelet-derived growth factor)-responsive element binding protein), it is a 220 kDa nuclear chromatin binding protein located in nuclear speckles, nuclear bodies and nucleoplasm that is composed of multiple domains, including several putative histone PTM "reader" modules. It features two independent nucleosome-binding domains, a DNA binding domain and a trans-activation domain (see Figure 4.15G).[820] The C-terminal nucleosome binding domain consists of an extended PHD domain [ePHD/ADD (extended plant homeodomain/ATRX-DNMT3-DNMT3L)],[820] a known reader domain which recognises a range of histone PTMs − mostly Lys and Arg methylations / acetylations.[821] Its second nucleosome binding domain, the SPBP-(1551-1666) region, has been discovered recently and is of unknown function.[820] It was shown to be essential for nuclear localisation of TCF20.[822] This nucleosome-binding domain is located immediately adjacent to the AT hook motif of the DNA binding domain.[820,822] As mentioned beforehand, this would in theory enable this domain to recognise histone PTMs in close structural proximity to the nucleosomal DNA − as is the case of a modification on H2B-Ser36 (see Figure 4.4). Interestingly, TCF20 has been reported to be *O*-GlcNAcylated at multiple sites [GlyGen Q9UGU0].

TCF20 acts on its own or as a co-activator, it influences many other transcription factors such as JUN, SP1, PAX6 and ETS1,[823] RNF4,[824] and TopBP1.[825] It further acts as a phosphoserine-specific inhibitor of estrogen receptor (ER) alpha[826] and influences androgen receptor signalling in competition with PAX6,[827] explaining its role in colorectal cancer. In addition, TCF20 is a co-activator of the NRF2-dependent antioxidant response activated during oxidative stress (KEAP1-NRF2 signalling pathway)[828] and AhR-mediated activation of CYP1A1[829], implying an influence on oxidative metabolic processes. It's complex with RAI1, PHF14 and HMG20A has been described in HK4me-regulated gene expression in neurons.[818]

Mutations in TCF20 have all been linked to developmental delay / intellectual disability (100 %), which can exhibit different disease conditions: autism spectrum disorder (ASD) or autistic features (69%), attention disorders or hyperactivity (67%), additional neurobehavioural concerns (85%), hypotonia (63%), other neurologic or muscular concerns (85%), and minor anomalies noted on physical examination (74%).[830] In particular, the P1557L mutation located in the nucleosome binding domain 1 is linked to

ASD[831] and the Y1632T mutation is associated to loss of function.[832] TCF20 contains 7 regions with 97 % sequence similarity to its paralogue RAI1, those mutations are implicated in Smith-Magenis syndrome[832,833] and Potocki-Lupski syndrome,[834] both of which are severe neurological disorders with Mendelian disease inheritance. A phylogenetic study suggest both have been created by a gene duplication event.[822] The recognition of the H2B-S36GlcNAc modification by TCF20 could therefore potentially link the emerging importance of the *O*-GlcNAc PTM in the brain to autism and neurological impairments.

4.4. Outlook

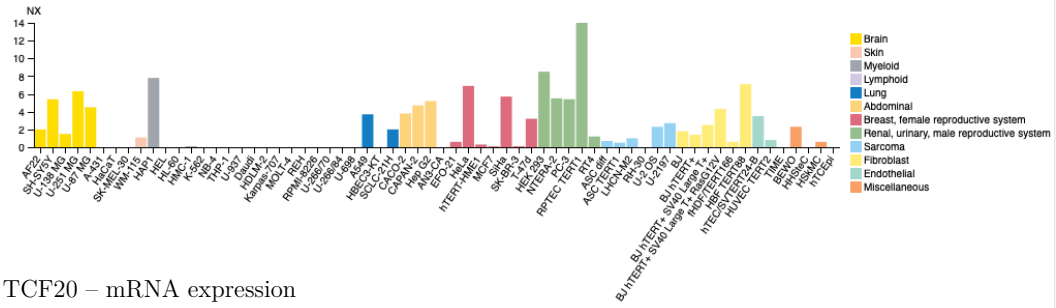
In this study, a wildtype nucleosome as well as synthetic nucleosomes carrying a glycosylation or phosphorylation mimic at H2B-Ser36 were assembled and used for label-free interaction proteomics with HeLa nuclear extract to identify PTM-specific binding partners. The transcription factor TCF20 was identified as a putative interaction partner of H2B-Ser36GlcNAc.

Synthetic nucleosome have previously proven their value in structural and biophysical studies,[773] the present study and others[773,778,780] illustrate they can also be used as precision tools to probe signalling pathways. Research in chromatin biology has heavily relied on the use of short peptides or isolated histone proteins as substrates, these are however mere fragments of true nucleosomal complex and several reports suggest these might not provide the biologically relevant context.[780,835] Synthetic nucleosomes provide a more accurate alternative but are not devoid from artefacts: despite their anticipated similarity, the recognition non-hydrolysable analogues such as Pfa of GlcNAc-thiol might differ slightly in electrostatics and flexibility, respectively.[811,812, -Pfa][800, -SGlcNAc] Nonetheless, their access is to date limited to laboratories with synthetic capabilities.

Synthetic nucleosomes thus provide a tools to precisely interrogate the role of a single histone PTM or combinations thereof in a biologically relevant context, if used appropriately. This emphasises the need for stringent quality control for the nucleosome matrix used as bait and the reagents used for the assembly, to ensure the results are biologically meaningful.[815] As such, a substantial part of this project focussed on identifying appropriate analytics and purification methods to overcome difficulties in assembly.

In addition, it might be necessary to use nuclear extracts from different cell lines as histone PTM might have differing functions depending on the cellular context. Most studies using defined, reconstituted nucleosomes – including this chapter – use HeLa

A. AMPK – mRNA expression



B. TCF20 – mRNA expression

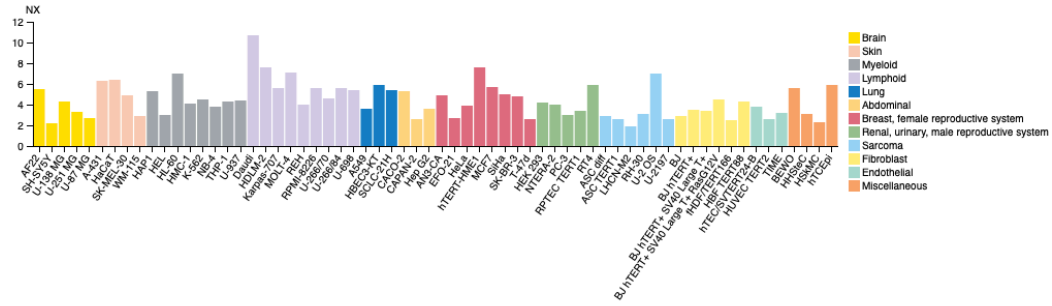


Figure 4.16. mRNA expression levels of AMPK and TCF20 in different cell lines. **A.** AMPK is a known interaction partner of H2B-Ser36.[788] The ProteomeAtlas database suggest it has tissue-enhanced overexpression, particular in heart muscle and skeletal muscle, and to a lesser extend in the kidney [Entry for PRKAA2]. As such, in addition to HeLa cells, suitable cell models could also include HEK293 or RPTEC/TERT1 where high abundance of AMPK is reported by RNA-seq [ProteomeAtlas], both of which are renal cell lines. **B.** TCF20 is widely expressed. Isoform 1 is exclusively expressed in the brain, heart, and testis and predominates in the liver and kidney. Isoform 2 is mostly expressed in the lung.

nuclear extract as the amount of generated nucleosomes is limited. The example of H2BS36p illustrates the limitations: the known interaction partner, AMPK, was not identified.[788] Its mRNA expression varies highly among cell lines – which is not indicative of the actual protein content in the cells[492,493], but highlights that this protein presumably has high variation in its tissue expression (Figure 4.16.A). TCF20 on the other hand has a comparable, ubiquitous expression (Figure 4.16.B) and was identified in this study. Other precedents include the lack of identification of BREA1A responsible for H2BS112GlcNAc-induced ubiquitination of H2BK120[781] in pull-down experiments from HeLa cells using synthetic nucleosomes.[780] Therefore a reflection on the best use of synthetic nucleosomes to probe biological processes is needed, including on the use of *Xenopus laevis* histones to probe human biology. Our results show that nuclear extract preparations can be a source of variation in itself and a QC might be needed to ensure maximal consistency amongst biological replicates. A possible avenue to overcome tissue-specific expression would be a pooled mixture of different nuclear extracts,

similarly to pooled cell extracts used in other chemical proteomics assays.[119]

Provided the findings that TCF20 recognises GlcNAcylated H2B-Ser36 with its nucleosome binding domain 1 hold true in subsequent validation experiments, this study would potentially link *O*-GlcNAc as a histone PTM to neurological disorders. The involvement of TCF20 and its paralogue RAI1 in the H3K4me writer subcomplex (containing iBRAF, PHF14, RAI1, TCF20) as an antiplayer to the LSD1-CoREST complex has linked this transcription factor family to neurodevelopmental disorders.[818,836] The importance of *O*-GlcNAc in brain diseases is already starting to emerge, and has noteworthy been characterised for Alzheimer's Disease.[837] Recently, mutations in OGT have been identified as causal in intellectual disability syndrome (ID), however, the exact mechanism leading to this disorder are unknown.[838]

Two possible, although speculative, interpretations of the results of this study can be advanced: a TCF20-centric interpretation and a H3K4me3-mediated interpretation involving the RAI1-TCF20 complex. In the TCF20-centric interpretation, glycosylation at H2B-Ser36 would recruit TCF20, which would then mediate downstream effects via the recruitment of co-activators, influencing the many downstream signalling possibilities − many pathways of which facilitate growth (Figure 4.17.A). This could hypothetically integrate in a antiplay framework with the H2B-S36 phosphorylation by AMPK reported by Bungard *et al.*[788]

The H3K4me3-mediated interpretation is suggested here because all the components of the RAI1-TCF20 complex are identified in the glycosylated pull-down: RAI1, TCF20 (= SPBP), PHF14 and iBRAF (= HMG20A). Unlike typical chromatin regulatory machineries, none of the RAI1-TCF20 complex components is a histone-modifying enzyme. However, three of the constituents, namely PHF14, RAI1 and TCF20 carry PHD or ePHD domains, which are putative methylhistone recognition domains. This complex is in antiplay with demethylation by the LSD1-Corest complex which demethylates H3K4me1/2, but not H3K4me3.[836] In their model for the role of histone methylation in neuron development, Iwase and co-worker suggest that RAI1 recognises unmethylated histone complexes with its ePHD domain − the NuBD1-nucleosome interaction appears to be independent of the H3 histone tail.[820] RAI1 does not bind to methylated H3K4me3.[836] MLL1 (KMT2A) is a H3K4me writer which, although not identified as the component for the RAI1 complex,[836] may participate in RAI1 complex through its association with iBRAF.[836] MLL1-mediated H3K4me3 stabilises the active transcriptional status of the neuron-specific genes when neurons mature. GlcNAcylation at H2B-S36 might stabilise the TCF20-RAI1 complex on the nucleosome transiently in this

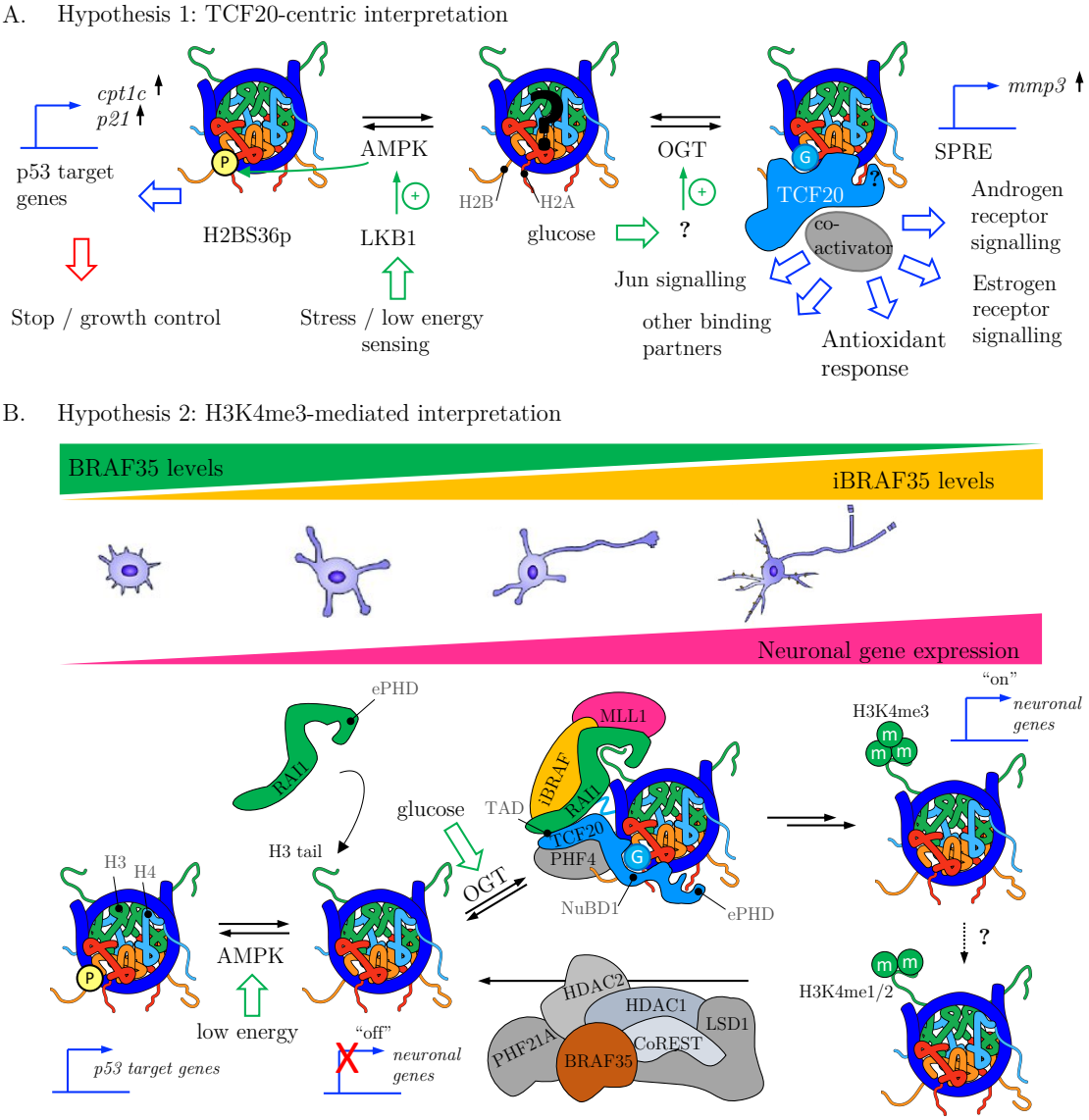


Figure 4.17. Speculative hypotheses for the role of phosphorylation and glycosylation of H2BS36. A. A TCF20-centric interpretation, in which H2BS36g recruits the TCF20 transcription factor leading to downstream signalling. The relation to H2BS36p is unclear, the expression of CPT1C leads to survival under glucose deprivation in neurons and cancer cells.[839] There could be a parallel co-existence or even competition between the two modifications. **B.** H3K4-mediated model, expanding the suggested mechanism by Iwase and co-workers[818] with the findings reported in this chapter. The LSD1-CoREST complex demethylates H3K4me leading to gene silencing. The TCF20-RAI1 complex recruits MLL1 leading to H3K4 methylation and subsequent neuron-specific gene expression. H2BS36g might facilitate or guide the binding of the TCF20-RAI1 complex, which would couple the H3K4me gene regulation to the presence of glucose as nutrient. Green - activation; red - inhibition; LKB1 - liver kinase B1; OGT - *O*-GlcNAc transferase; SPRE - Stromelysin-1 PDGF-responsive element; *mmp3* - gene for matrix metalloprotease 3; *cpt1c* - gene for Carnitine Palmitoyltransferase 1C.

process through its NuBD1 domain (Figure 4.17.B), thus coupling neuronal evolution with the presence of glucose as nutrient. The P1557L mutation in the NuBD1 domain of TCF20 which could interfere on the binding is known to cause ASD.[831]

In summary, the initial identification of H2B-Ser36GlcNAc binding to TCF20 raises more questions than it answers, and further investigation is required to assess the role of this PTM. The possible interplay with other modifications together with the high complexity of chromatin remodelling assemblies will call for a broad range of chemical biology tools and epigenetics techniques to be combined for further advances.

4.5. Author contributions

Histone plasmids were offered by Prof. Rob Klose. 186bp and 145bp plasmids were generated by Lukas Lercher. All non-modified histone proteins were expressed by Dr. Ritu Raj, who originally started the project. One pull-down experiment with synthetic nucleosomes (i.e. one biological replicate) was performed by Dr. Ritu Raj. The digest of synthetic proteins and their MS/MS data acquisition was performed by Svenja Hester. All other data presented in this chapter was generated by Simon Nadal.

Part of this work was published in:
Nadal S, Raj R, Mohammed S & Davis BG. Synthetic Post-Translational Modification of Histones. *Current Opinions in Chemical Biology* **2018**, *45*, 35.

4.6. Experimental

4.6.1. Materials

4.6.1.1. Chemicals and enzymes

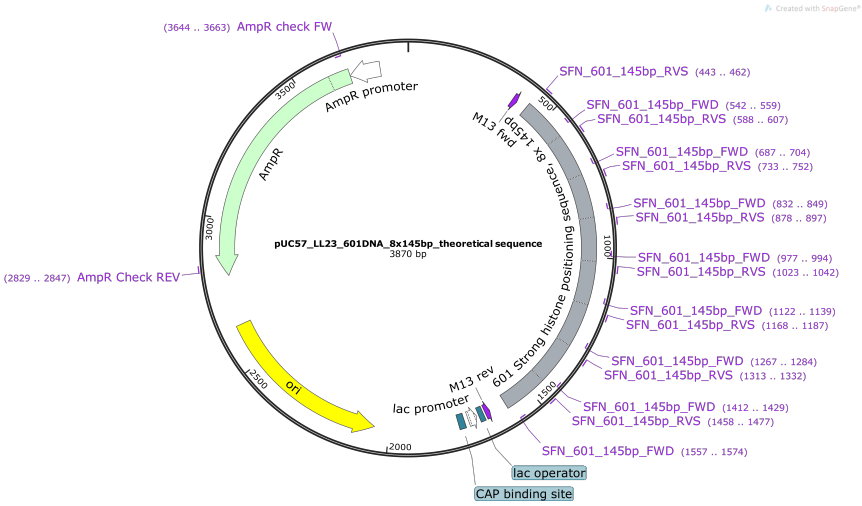
Chemicals for biological use were purchased from commercial sources (Acros Organics, Bayer, Biomol, Fluka, GE Healthcare, Merck KgaA, Sigma-Aldrich Co. LLC, Serva, Thermo Fischer Scientific, VWR, Yorkshire Biosciences Ltd) and were used without further purification unless otherwise stated. GlcNAc-thiol was kindly donated by Dr. Mylène Richard and synthesised according to the literature.[748] (Bromodifluoromethyl)phosphonic acid (Pfa) was kindly donated by Dr. Abirami Lakshminarayanan and synthesised according to the literature.[759] The following listed enzymes were used:

RNAse A	Thermo Fischer Scientific, EN0531
RNAse B	Sigma Aldrich
DNAse I	Roche Diagnostics
EcoRV-HF	New England Biolabs (NEB), R3195S
EcoRI-HF	New England Biolabs, R3101S
Klenow Fragment (3' → 5' exo-)	New England Biolabs, M0212S
Trypsin	Promega, MS grade

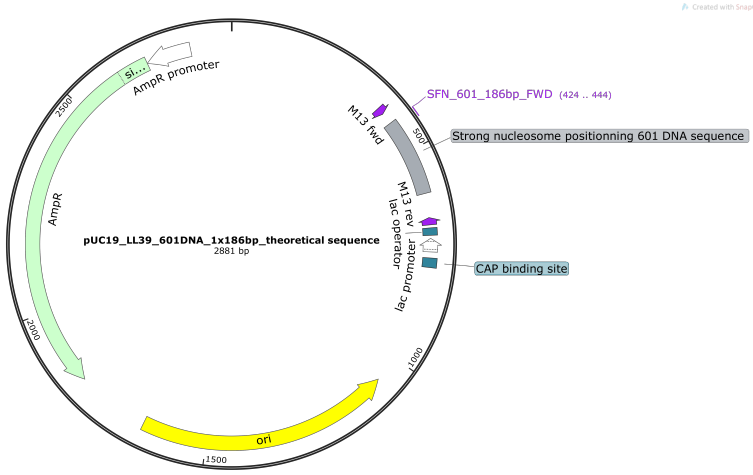
4.6.1.2. Plasmids

pET3 bacterial plasmids encoding all canonical *Xenopus laevis* histones were kindly donated by Dr Rob Klose, University of Oxford, Oxford.

pUC19:601DNA-8x145bp ; Lukas Lercher (LL23)



pUC19:601DNA-186bp ; Lukas Lercher (LL39)



4.6.1.3. Primers

SFN_601_145bp_FWD	GAG GCC GCT CAA TTG GTC
SFN_601_145bp_RVS	TAT CTG ACA CGT GCC TGG AG
SFN_601_186bp_FWD	TCG GAT CTT ACA TGC ACA GGA

4.6.2. Expression and purification of histone proteins

The expression of histone proteins was done by previous group members, Dr. Ritu Raj and Dr. Lukas Lercher. Procedures are listed here for comprehensiveness.

pET3 bacterial plasmids encoding all canonical *Xenopus laevis* histones were kindly donated by Dr Rob Klose, University of Oxford, Oxford. The following gene sequences were used for wildtype (wt) proteins:

(i) wildtype H2A

DNA sequence:
ATGTCAGGAAGAGGCAAACAAGGCGGTAAAACCCGCGCTAAGGCCAAGAC
TCGCTCATCTCGGGCTGGGCTACAGTTCCCTGTTGGCCGTGTTCACCGGCT
GTTAAGGAAAGGCAATTATGCAGAGCGGGTGGGAGCTGGAGCTCCAGTCT
ATCTGGCTGCAGTGTTGGAGTATCTGACCGCTGAGATTTTGGAATTGGCCG
GGAATGCGGCCCGTGATAACAAGAAGACTCGCATTATCCCCAGACACCTGC
AGCTCGCTGTGCGCAACGATGAGGAACTGAACAAACTGCTCGGAAGAGTC
ACTATCGCTCAGGGCGGGGTCCTGCCCAACATCCAGTCCGTGCTGCTGCCC
AAGAAAACCGAGAGTTCCAAGTCGGCCAAGAGCAAGTGATCGGATCCGGC
TGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCA
ATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTT
GCTGAAAGGAGGAACTATATCCGGATATCCACAGGACGGGTGTGGTCGCC
ATGATCGCGTAGTCGATAGTGCTCCAAGTAGCGAAGCGAGCAGGACTGGG
CGGCGGCCAAAGCGGTCGGACAGTGCTCCGAGAACGGGTGCGCATAGAAA
TTGCATCAACGCATATAG

Translated amino acid sequence:
MSGRGKQGGK TRAKAKTRSS RAGLQFPVGR VHRLLRKGNY AERVGAGAPV
YLAAVLEYLT AEILELAGNA ARDNKKTRII PRHLQLAVRN DEELNKLLGR
VTIAQGGVLP NIQSVLLPKK TESSKSAKSK

(ii) wildtype H2B

DNA sequence:
ATGGCCAAGTCCGCTCCAGCCCCGAAGAAAGGCTCCAAGAAAGCGGTGACC
AAGACTCAGAAGAAAGACGGGAAAAAGCGCAGGAAGACAAGGAAGGAGAG
TTATGCCATTTACGTGTACAAGGTGCTGAAGCAGGTGCACCCCGATACCGG

CATCTCGTCCAAGGCCATGAGCATCATGAACTCCTTTGTCAACGATGTGTT
TGAGCGCATCGCAGGGGAAGCCTCCCGCCTGGCTCATTACAACAAGCGCTC
CACCATCACCTCCCGGGAGATCCAGACCGCGGTCCGACTGCTGCTGCCTGG
GGAGTTGGCCAAACACGCCGTGTCCGAGGGCACCAAGGCTGTCACCAAGTA
CACCAGCGCCAAGTAACTGCTGCTGCCGGATCCGGCTGCTAACAAAGCCCG
AAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAG

Translated amino acid sequence:
MAKSAPAPKK GSKKAVTKTQ KKDGKKRRKT RKESYAIYVY KVLKQVHPDT
GISSKAMSIM NSFVNDVFER IAGEASRLAH YNKRSTITSR EIQTAVRLLL
PGELAKHAVS EGTKAVTKYT SAK

(iii) wildtype H3

DNA sequence:
ATGGCCCGTACCAAGCAGACCGCCCGTAAATCCACCGGAGGGAAGGCTCCC
CGCAAGCAGCTGGCCACCAAGGCAGCCAGGAAGTCCGCTCCTGCTACCGGC
GGAGTCAAGAAACCTCACCGTTACCGGCCCGGCACAGTCGCTCTCCGCGAG
ATCCGCCGCTACCAGAAATCCACCGAGCTGCTCATCCGCAAACTGCCTTTC
CAGCGCCTGGTCCGGGAGATCGCTCAGGACTTCAAGACCGACCTGCGCTTC
CAGAGCTCGGCCGTTATGGCTCTGCAGGAGGCCAGCGAGGCTTATCTGGTC
GCTCTCTTTGAGGACACCAACCTGGCCGCCATCCACGCCAAGAGGGTCACC
ATCATGCCCAAGGACATCCAGCTGGCCCGCAGAATCCGAGGCGAGAGGGCT
TAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCA
CCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGA
GGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGATATCCACAGGACGGG
TGTGGTCGCCATGATCGCGTAGTCGATAGTGGCTCCAAGTAG

Tanslated amino acid sequence:
MARTKQTARK STGGKAPRKQ LATKAARKSA PATGGVKKPH RYRPGTVALR
EIRRYQKSTEL LIRKLPFQRL VREIAQDFKT DLRFQSSAVM ALQEASEAYL
VALFEDTNLA AIHAKRVTIM PKDIQLARRI RGERA

(iv) wildtype H4

DNA sequence:
ATGTCTGGTCGTGGTAAAGGTGGTAAAGGTCTGGGTAAAGGTGGTGCTAA

ACGTCACCGTAAAGTTCTGCGTGACAACATCCAGGGTATCACCAAGCCGGC
TATCCGTCGTCTGGCTCGTCGTGGTGGTGTTAAACGTATCTCCGGTCTGAT
CTACGAAGAAACCCGCGGTGTTCTGAAAGTTTTCCTGGAAAACGTTATCCG
TGACGCTGTTACCTACACCGAACACGCTAAACGTAAAACCGTTACCGCTAT
GGACGTTGTTTACGCTCTGAAACGTCAGGGTCGTACCCTGTACGGTTTCGG
TGGTTAAAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCT
GCTGCCACCGCTGAGCAATAACTAG

Tanslated amino acid sequence:
MSGRGKGGKG LGKGGAKRHR KVLRDNIQGI TKPAIRRLAR RGGVKRISGL
IYEETRGVLK VFLENVIRDA VTYTEHAKRK TVTAMDVVYA LKRQGRTLYG
FGG

(v) H2B-Cys36

The numbering follows UniProt numbering of the human sequence (including PEP peptide). This corresponds to Cys33 in the core histone numbering.

DNA sequence:
ATGGCCAAGTCCGCTCCAGCCCCGAAGAAAGGCTCCAAGAAAGCGGTGACC
AAGACTCAGAAGAAAGACGGGAAAAAGCGCAGGAAGACAAGGAAGGAGTG
TTATGCCATTTACGTGTACAAGGTGCTGAAGCAGGTGCACCCCGATACCGG
CATCTCGTCCAAGGCCATGAGCATCATGAACTCCTTTGTCAACGATGTGTT
TGAGCGCATCGCAGGGGAAGCCTCCCGCCTGGCTCATTACAACAAGCGCTC
CACCATCACCTCCCGGGAGATCCAGACCGCGGTCCGACTGCTGCTGCCTGG
GGAGTTGGCCAAACACGCCGTGTCCGAGGGCACCAAGGCTGTCACCAAGTA
CACCAGCGCCAAGTAACTGCTGCTGCCGGATCCGGCTGCTAACAAAGCCCG
AAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAG

Tanslated amino acid sequence:
MAKSAPAPKK GSKKAVTKTQ KKDGKKRRKT RKE**C**YAIYVY KVLKQVHPDT
GISSKAMSIM NSFVNDVFER IAGEASRLAH YNKRSTITSR EIQTAVRLLL
PGELAKHAVS EGTKAVTKYT SAK

4.6.3. Bacterial transformation

The plasmid of interest was transformed in XL-10 Gold ultracompetent cells (Agilent technologies) for plasmid amplification or in BL21(DE3)pLysS (Agilent Technologies)

for protein expression. 20 µL of cells were aliquoted in a pre-chilled 5 mL BD Falcon propylene round-bottom tube. For XL-10 Gold cells, µL of β-mercaptoethanol was added to the cells. The tube was gently swirled and incubated for 10 min. 0.3 µL of plasmid DNA (50 ng/µL) was added and the content of the tube swirled gently. The cells were incubated on ice for 30 min, followed by heat-shock in a 42 °C water bath for exactly 45 seconds. The transformed mixture was immediately cooled on ice for 2 min. 250 µL of pre-warmed SOC medium was added and the cells grown at 37 °C with 180-220 rpm shaking for 60 min. The transformed cells were subsequently plated out on LB-agar plates with appropriate antibiotics, typically chloramphenicol (pLysS plasmid) and ampicillin (pET plasmid), and incubated O/N at 37 °C. After colony picking, the plates containing transformed cells were kept at 4 °C for short term storage.

4.6.4. Plasmid amplification

A single colony of transformed cells was picked to inoculate 10 mL of LB medium containing appropriate antibiotics (10 µL of 100 mg/mL ampicillin stock solution and/or 10 µL of 33 mg/mL chloramphenicol stock solution in EtOH) in a 50 mL falcon tube. The cells were grown O/N at 37 °C. The plasmid DNA was isolated using the QIAprep Spin Miniprep kit (Qiagen), performing 2 wash steps and eluting in MilliQ water. The sequence of plasmid DNA was measured by nanodrop and confirmed by Sanger sequencing.

4.6.5. Expression and purification of histone proteins

The transformation of pET3 plasmid encoding histone proteins into BL21(DE3)pLysS cells was carried out using LB-agar plate containing 100 µg/mL ampicillin and 33 µg/mL chloramphenicol. A single colony was used for inoculation of a starter culture in 50 mL LB media supplemented with the same antibiotics. The starter culture was grown at 37 °C overnight. 800 mL of 2xYT medium supplemented with antibiotics were inoculated with 10 mL of starter culture and grown at 37 °C until OD_{600} of approx. 0.4-0.6 was reached. Expression was induced with 0.5 mM IPTG and allowed to continue for 2 h (3 h for H4 histone) at 37°C. The cells were collected by centrifugation at 8000 rpm for 8 min at 4 °C. Harvested cells were resuspended in 35 mL lysis buffer (50 mM Tris pH 7.5, 100 mM NaCl, 1 mM EDTA, 5 mM β-mercapto ethanol), supplemented with a cocktail of protease inhibitors (cOmplete, EDTA-free, Roche Life Sciences) before storing them at -80 °C for further use.

For cell lysis, frozen cells were thawn in a water bath at r.t., supplemented with 5 mg of DNase I (Roche Life Sciences) and incubated at 4 °C for 20 min while gently rotating. The sonication was carried out in 5 cycles with a duration of 30 seconds (0.5 sec. on, 0.5 sec. off, 60 % amplitude) with an interval of 1 min between cycles and was followed by centrifugation (20000 rpm for 20 min at 4 °C). The supernatant was discarded and the pellet resuspended in 40 mL TW buffer (50 mM Tris pH 7.5, 100 mM NaCl, 1 mM EDTA, 5 mM β-mercapto ethanol, 1 % Triton X-100). The resulting suspension was sonicated for further 30 seconds at 60 % amplitude and centrifuged (20000 rpm for 10 min at 4 °C). The same washing step using TW buffer was carried out twice and repeated additionally using lysis buffer. The obtained pellet was resuspended in 0.2 - 0.3 mL DMSO and incubated for 10 min at r.t. Following the addition of 7-15 mL of unfolding buffer (7 M guanidinum hydrochloride, 10 mM Tris pH 7.5, 1 mM EDTA, 10 mM DTT and 1 mM benzamidine), the mixture was incubated while gently rotating for 1 h at r.t. Cell debris were removed by centrifugation (20000 rpm for 10 min at r.t.).

The supernatant was filtered and loaded onto a pre-equilibrated size exclusion column (HiLoad 26/60 Superdex 200, GE Healthcare) in SAU-1000 buffer (7 M urea, 20 mM sodium acetate pH 5.2, 1 M NaCl, 1 mM EDTA, 5 mM β-mercapto ethanol). The collected fractions were analysed by SDS-PAGE. Fractions containing purified histones were pooled together and dialysed against IEX binding buffer (7 M urea, 10 mM Tris pH 7.5, 100 mM NaCl, 1 mM EDTA, 5 mM β-mercapto ethanol).

Further purification was carried out by ion-exchange chromatography. Protein solutions were loaded onto a HiTrapSP HP column (GE Healthcare) in IEX binding buffer. After a washing step using 12 CV of IEX binding buffer to remove impurities, histones were eluted using a linear gradient (0 % to 100 %) of the elution buffer (7 M urea, 10 mM Tris pH 7.5, 1 M NaCl, 1 mM EDTA, 5 mM β-mercapto ethanol). All collected fractions were analysed using SDS-PAGE. Fractions containing purified histones were pooled together. Following a dialysis step against MQ water supplemented with 2 mM β-mercaptoethanol, purified histones were lyophilised and stored at -80°C for further use.

The purity of obtained histones was confirmed by mass spectrometry.

4.6.6. Post-translational chemical mutagenesis

4.6.6.1. Synthesis of H2B-Dha36 protein

In a typical experiment, 3.00 mg (0.22 µmol) H2B-Cys36 was dissolved in 500 µL freshly prepared reaction buffer (5 M Gd.HCl, 50 mM Tris pH 8.0). This protocol has been tried with concentrations up to 11.0 mg/mL of protein. For unknown proteins, it is recommended to test different concentrations first. 9.00 mg DTT is pre-weighed in a separate Eppendorf tube and the dissolved protein added. The reaction is incubated for 30 - 45 min, the reduction can be monitored by non-reducing SDS-PAGE or LC-MS. The reduced protein is desalted using a pre-equilibrated PD MiniTrap G-25 column, eluting in 1 mL reaction buffer. 100 eq. of MDBP (2.23 µmol, 7.0 µL) is dissolved in 50 µL DMSO and slowly added to the reaction mixture. The reaction is incubated in the dark at 37 °C. Progress is monitored by LC-MS, completion was found after 1.5 h. Upon completion, the desired H2B-Dha36 is desalted from the small molecule by passing it through a pre-equilibrated PD MidiTrap column and either stored at -20 °C, or, if required, concentrated via vivaspin 500 (5000 MWCO) to the required volume (typically 200 µL).

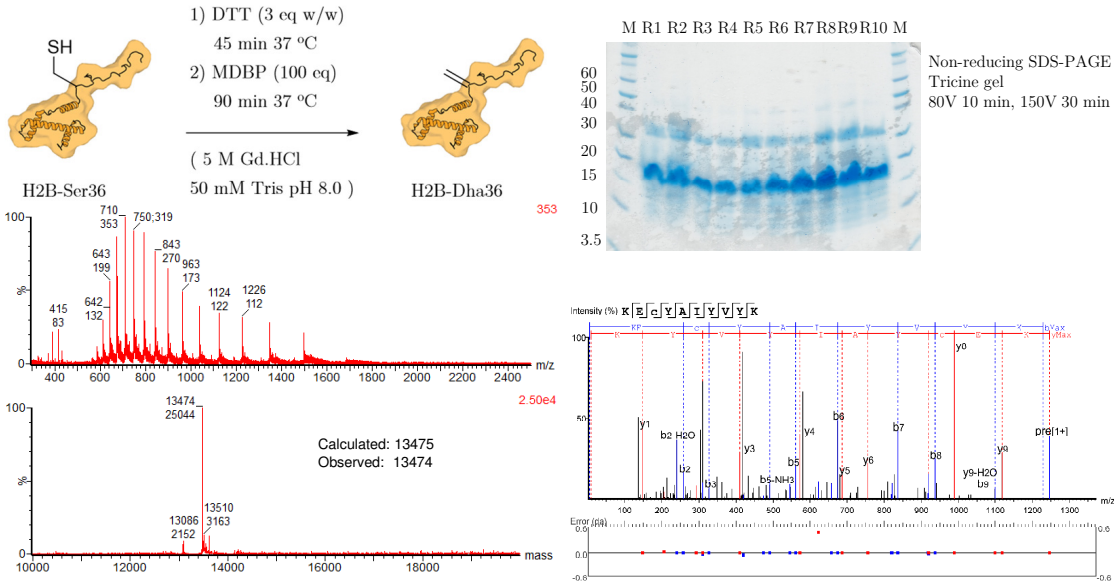


Figure 4.18. Analytical characterisation of synthesised H2B-Dha36 protein. Top left. Reaction scheme and conditions. **Top right.** Non-reducing SDS-PAGE confirms formation of the protein. A faint band corresponding to the disulphide histone H2B-Cys36

Figure 4.18. (***continued***) histone dimer is observed. **Bottom left.** Denaturing intact protein LC-MS showing the molecular ion series(top) and the deconvoluted spectrum (bottom), confirming the mass of the synthetic construct. **Bottom right.** Tandem mass spectrometry confirms the site of modification at position 36, shown here as a modification on cysteine in a HCD spectrum.

4.6.6.2. Synthesis of H2B-CysGlcNAc36

In a representative reaction, 2.00 mg (0.15 µmol) of H2B-Dha36 was dissolved in 250 µL reaction buffer (8 mg/mL, 50 mM sodium phosphate pH 8.0). GlcNAc-thiol (400 eq, 14.12 mg, 59.2 µmol) was added to the protein solution as a solid. The reaction was incubated for 2 h. Excess GlcNAc-thiol was removed by desalting using a pre-equilibrated PD MidiTrap G-25 column (GE Healthcare). The modified protein was either stored in solution in buffer or as lyophilised powder (elution in MilliQ water) at -20 °C.

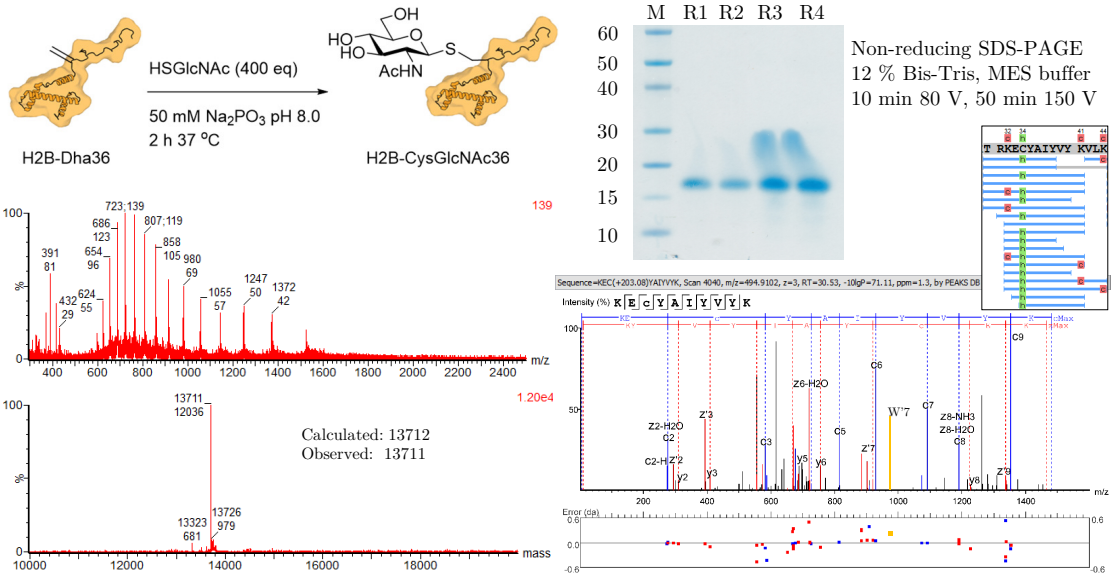
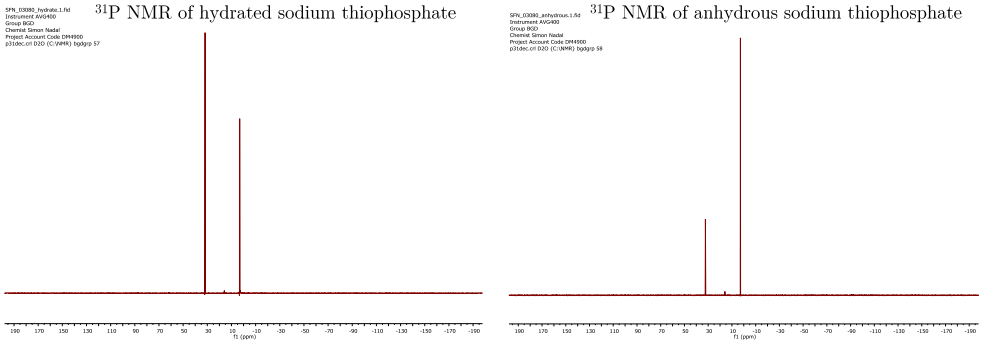


Figure 4.19. Analytical characterisation of synthesised H2B-CysGlcNAc36 protein. Top left. Reaction scheme and conditions. **Top right.** Non-reducing SDS-PAGE confirms formation of the protein. A faint band corresponding to the disulphide histone H2B-Cys36 histone dimer is observed. **Bottom left.** Denaturing intact protein LC-MS showing the molecular ion series(top) and the deconvoluted spectrum (bottom), confirming the mass of the synthetic construct. **Bottom right.** ETD tandem mass spectrometry spectrum confirms the site of modification at position 36, shown here as a modification on cysteine. The diagnostic $w7'$ ion is highlighted in orange. The insert maps the peptide coverage at the site of interest.

4.6.6.3. Attempts for the synthesis of thiophosphorylated histone H2B-pCys36

Preparation of anhydrous thiophosphate

In order to obtain reproducible results, the protein reactions were attempted with anhydrous sodium thiophosphate. The starting material thiophosphate hydrate (Santa Cruz Biotech or Alfa Aesar) was visually inspected for the presence of grey traces. Only pure white powder was processed. 5 g of anhydrous thiophosphate was triturated in dry methanol (15 mL) in an oven-dried flask under argon protective atmosphere for 30-60 min. The white solid was filtered of and subsequently dried for 1 h in a Kugelrohr (Buchi) at 100 °C, 40 rpm under vacuum (0.15 mbar); or heated for 1 h at 70 °C under high vacuum. This temperature might be too high as decomposition is known to occur at 120 - 125 °C.[840] In either case, the heating was switched of and the powder left to dry O/N under high vacuum. The remaining powder was visually inspected for its pure white colour and stored under argon. An alternative procedure described in the literature might be recrystallisation followed by freeze-drying.[807] For ^{31}P-NMR analysis, 25 mg of hydrated sodium thiophosphate was dissolved in 250 µL of 200 mM sodium phosphate buffer and 250 µL D_2O. Similarly, 5 mg of anhydrous sodium thiophosphate was dissolved in 250 µL of 200 mM sodium thiophosphate buffer and 250 µL D_2O.



Synthesis of H2B-pCys36 using anhydrous thiophosphate

Using the prepared anhydrous thiophosphate, 10 mL of reaction buffer (2 M Gd.HCl, 0.4 M Na_3SPO_3 pH 8.1) was freshly prepared. A PD SpinTrap G-25 spin column was equilibrated with this buffer and H2B-Dha36 (150 µL, 0.55 mg, 0.04 µmol) was buffer exchanged into the reaction buffer. The reaction was incubated at 37 °C. Conversion was monitored by LC-MS. The reaction was then desalted through a PD MiniTrap G-25 into 5 M Gd.HCl, 50 mM Tris pH 8.0 at stored at -20 °C.

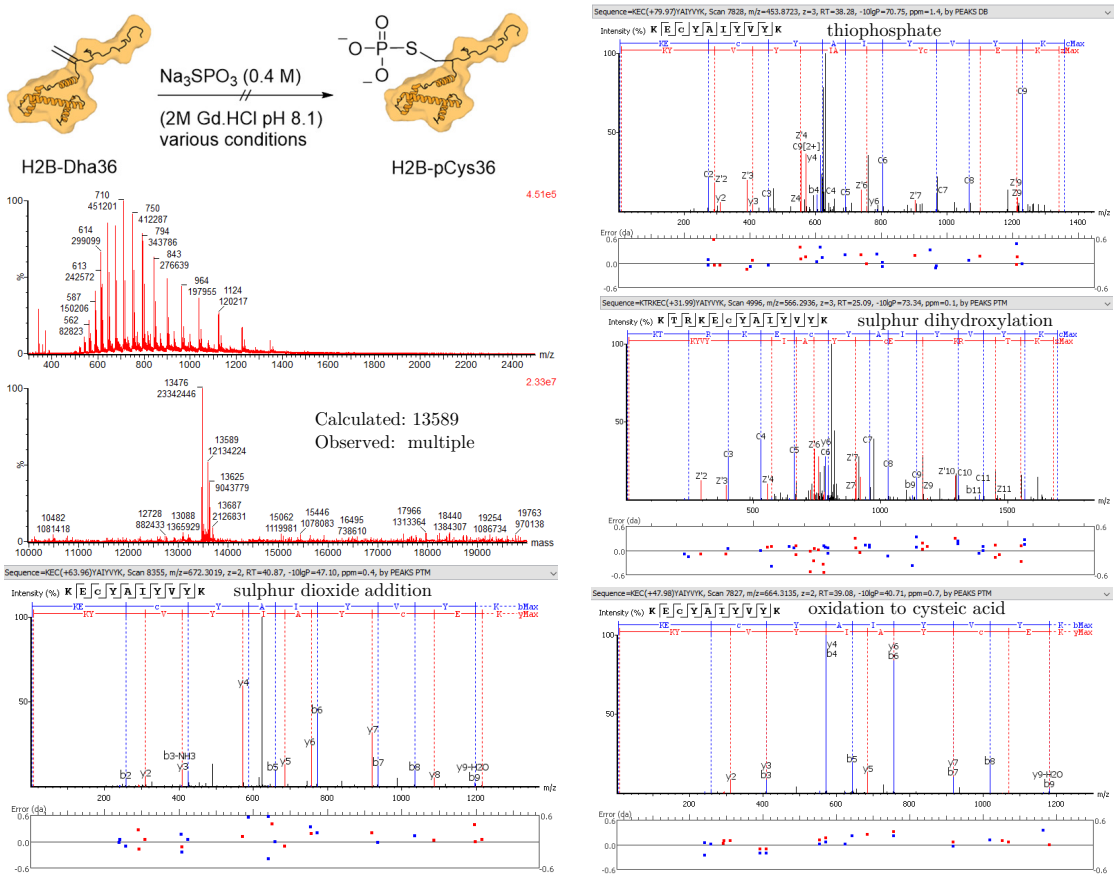


Figure 4.20. Analytical characterisation of synthesised H2B-pCys36 protein. Top left. Reaction scheme and conditions. **Mid left.** Denaturing intact protein LC-MS showing the molecular ion series(top) and the deconvoluted spectrum (bottom), showing formation of several distinct proteoforms. **Bottom and right.** Tandem mass spectrometry identifies several modifications on the H2B-pCys36 protein under the reported conditions, including sulphur dioxide addition, sulphur dihydroxylation and oxidation to cysteic acid.

Immobilised metal affinity chromatography for enrichment of thiophosphorylated protein
Having not reached > 95 % conversion, the purification of H2B-pCys36 from H2B-Dha36 was attempted using a 1 mL Ga(III)-NTA and a 1 mL Fe(III)-NTA combined column. For the generation of the columns, two 1 mL of previously used Ni-NTA columns (Hi-Trap HP, GE Healthcare) were individually discharged by washing with 10 CV MilliQ water, 15 CV stripping buffer (20 mM Na_3PO_4 pH 7.4, 500 mM NaCl, 50 mM EDTA), 15 CV binding buffer (20 mM Na_3PO_4 pH 7.4, 500 mM NaCl, 20 mM imidazole) and 15 CV MilliQ water. The discharged columns were further cleaned with 2 CV NaOH followed by 15 CV MilliQ water. The NTA resin was loaded with 5 CV charging solution (either 100mM $GaCl_3$ or 100 mM $FeCl_3$) and incubated for 30 min. The loaded column was flushed with 2 CV charging solution before rinsing with 15 CV of 1 % acetic acid

and 15 CV 20 % EtOH for storage.
For phosphoprotein enrichment using the dual Ga/Fe-NTA system, the Ga-NTA column was screwed on top of the Fe-NTA column, resulting in a 2 mL column volume which was used at a flow rate of 1 mL/min. The column was washed with 5 CV buffer A (2 M Gd.HCl, 50 mM Tris-HCl, 0.1 % acetic acid pH 5.5), followed by 5 CV buffer B (2 M Gd.HCl, 0.2 M Na_3PO_4 pH 8.0) and equilibration with 5 CV buffer A. [10 CV equilibration might be better]. The deslated reaction mixture containing H2B-Dha36 and H2B-Cys36 was injected. The column was washed with 10 CV buffer A before elution in a 5 CV gradient (0 → 100% buffer B) and a 15 CV plateau buffer B. The column was rinsed with buffer A, MilliQ and stored in 20 % EtOH. The elution profile suggests most of the protein ended in the flowthrough, but some was retained. The low amount of protein obtained did not enable conclusive analysis by SDS-PAGE or LC-MS.

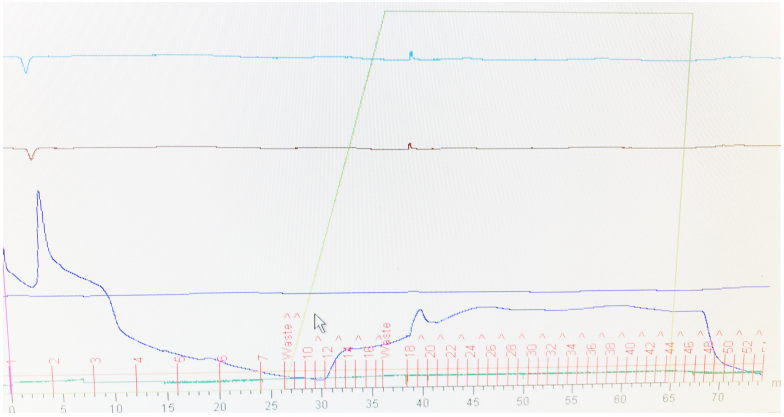


Figure 4.21. FPLC trace of the attempted enrichment of phosphorylated H2B-pCys36 using a combined Fe and Ga-IMAC.

4.6.6.4. Synthesis of H2B-Pfa36

In a typical reaction, previously synthesised H2B-Dha36 (0.2 mg, 15 nmol) was buffer-exchanged into the reaction buffer (5 M Gd.HCl, 500 mM NH_4OAc pH 6.0) and diluted to a final concentration of 1 mg/mL in a volume of 200 µL. The protein was incubated O/N at 4 °C in an anaerobic glovebox with open cap. The next day, a solution of (bromodifluoromethyl)phosphonic acid (Pfa-Br) was freshly prepared (16 mg, initially dissolved in 50 µL), its pH adjusted to 6 M and diluted to a final volume of 80 µL. A separate 1.5 mL eppendorf tube containing 0.8 mg $NaBH_4$ was prepared, minimising contact to air. The Pfa-Br solution was ported into the glovebox along with the $NaBH_4$-containing tube. 4 µL of the Pfa-Br was added to the buffered protein solution, which was then transferred to the tube containing the sodium borohydride, and incubated at

4 °C. The observation of foam indicates successful reaction. The reaction was monitored by LC-MS and completion found after 30 min. The reaction mixture was desalted using pre-equilibrated PD SpinTrap G-25 columns (GE Healthcare). The protein was either stored in solution at -20 °C or lyophilised for long-term storage.

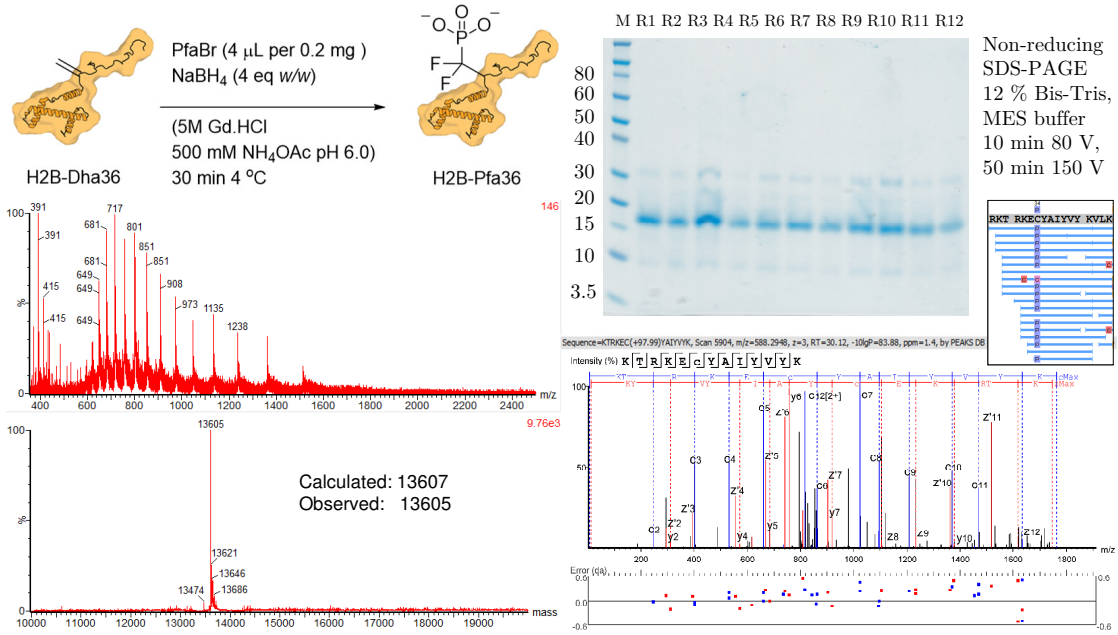


Figure 4.22. Analytical characterisation of synthesised H2B-Pfa36 protein. Top left. Reaction scheme and conditions. **Top right.** Non-reducing SDS-PAGE confirms formation of the protein. **Bottom left.** Denaturing intact protein LC-MS showing the molecular ion series(top) and the deconvoluted spectrum (bottom), confirming the mass of the synthetic construct. **Bottom right.** ETD tandem mass spectrometry spectrum confirms the site of modification at position 36, shown here as a modification on cysteine. The diagnostic $w7'$ ion is highlighted in orange. The insert maps the peptide coverage at the site of interest.

4.6.7. LC-MS/MS of modified histone proteins

The proteolytic digestion and data acquisition was performed by Ms. Svenja Hester, Advanced Proteomics Facility Oxford. The modified proteins (∼20 µg in 5 M Gd.HCl) were diluted 5× with 50 mM ammonium bicarbonate buffer. Digestion with trypsin (1:40 *w/w*) was performed O/N at 37 °C. The samples were dried in a speed-vac and resuspended in 5 % FA 5 % DMSO.

Resulting tryptic peptides 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 peptide were trapped on a guard column (75 µm i.d. × 20 mm, reprosil C18, 3 µm, 120 Å) using solvent A (0.1 % FA, 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 min, 8 % to 30 % of solvent B: 0.1 % FA, 5 % DMSO in acetonitrile, flow rate: 200 nL/min). The separated peptides were electrosprayed into the mass spectrometer operating on of the following methods.

HCD-pd-ETD (H2B-CysGlcNAc36 glycosylated sample)
The HCD-pd-ETD method has been described previously.[305,610] Full scan MS spectra (scan range 350-1500 m/z, resolution 30000, AGC target 1e6, maximum injection time 250 ms) and subsequent HCD MS/MS spectra (resolution 15000, AGC target 3e4, 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 138.0546, 168.0563 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 5e4, AGC anion target 2e5, cation maximum injection time 100 ms) were acquired in the Iontrap.

Data-dependent decision tree, H2B-pCys36 and H2B-Pfa36 phosphorylated samples
The use of a CID-ETD based data-dependent decision tree (DDDT) for phosphorylated samples has been described previously.[306] Full scan MS spectra were acquired in the Orbitrap (350-1500 m/z, resolution 120 000, AGC target 1 $\times 10^6$, maximum injection time 250 ms). CID and ETD spectra were acquired in Ion Trap (resolution 7500, AGC cation target 3 $\times 10^4$, AGC Anion target 2 $\times 10^5$, maximum injection time 100 ms). After the MS scans, the 5 most intense peaks were selected for fragmentation based on a data-dependent decision tree (DDDT). The signal threshold for fragmentation of the parent

ion was set to 500 ion counts. For CID fragmentation, the normalized collision energy and default charge state was set to 35% and 2 respectively. For ETD fragmentation, reaction time was set to 50 ms and supplemental activation was enabled. Charge state screening was enabled; parent ions with an unassigned charge state or charge state 1 were rejected. Dynamic exclusion was enabled (exclusion size list 75, exclusion duration 5 s). For the DDDT, ETD fragmentation was followed for charge states 3, 4 and 5 with m/z less than 750. For charge states greater than 5, ETD fragmentation was always performed. In all other cases, CID activation was performed.

Data was analysed using PEAKS X+, searching against the H2B-Cys36 sequence and the MaxQuant contaminant database. The data type was set to Orbi-Trap using mixed fragmentation. Precursor mass tolerance was set to 10 ppm and fragment mass tolerance to 0.6 Da. Trypsin was selected as protease with a maximum of 5 missed cleavage and non-specific cleavage at one end of the peptide. Oxidation (Methionine), Deamination (Asparagine, Glutamine), Carbamidomethylation (Lysine, peptide N-term) were set as variable modification. Depending on sample, HexNAc (Cysteine, 203.0794), phosphorylation (Cysteine, +79.9663) or Pfa (Cysteine, 97.9911) were added. A maximum of 4 variable modifications was set. A FDR of 1 % at peptide level and a *de novo* ALC of 80 was applied. All spectra and identifications were manually validated.

4.6.8. Histone octamer assembly

Histone octamers were refolded as described in the literature. Lyophilised histone proteins were dissolved in unfolding buffer (7 M Gd.HCl, 10 mM Tris pH 7.5, 1 mM DTT, 1 mM EDTA) by soft inversion and left to unfold for 1 h at r.t. The concentration was measured by nanodrop. Wt H2B-Ser36 or modified proteins were mixed with other histone proteins in equimolar ratios to a final total histone concentration of 1 mg/mL in unfolding buffer. The protein mixture was loaded into a dialysis cassette (MWCO 3500 or MWCO 6000, Pur-a-Lyzer, Sigma Aldrich) and dialysed 3 times against 600 mL refolding buffer (2 M NaCl, 10 mM Tris pH 7.5, 1 mM DTT, 1 mM EDTA) at 4 °C for at least 4 h each, one step being overnight. After dialysis, the solution was carefully transferred into a pre-cooled vivaspin 6 (MWO 5000) and concentrated (3750 rpm, 4 °C) to a volume of 1 mL. The solution is centrifuged (7000 rpm, 10 min, 4 °C) to remove white precipitate if any. Refolded octamers were purified by size-exclusion chromatography (HiLoad 16/60 Superdex S200, GE Healthcare) in refolding buffer at 4 °C, 1 mL injection volume. After chromatography, the fraction containing pure histone octamer (elution time 62-68 mL) were pooled and concentrated using vivaspin 6 (5000 MWCO,

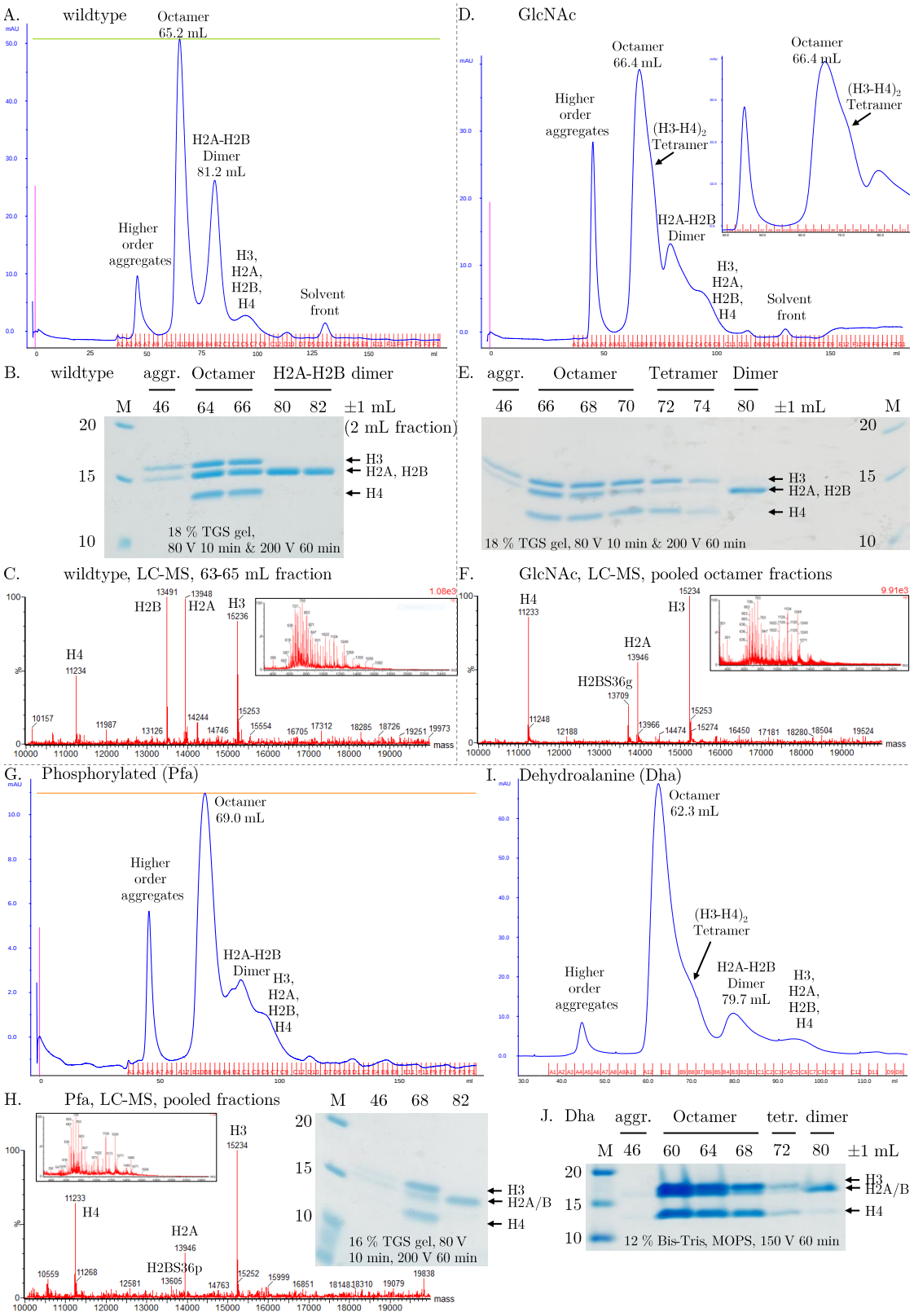


Figure 4.23. Analytical characterisation of synthetic histone octamers. The FPLC traces of the octamer are shown together with the SDS-PAGE as well as intact protein LC-MS of the fractions, confirming the presence of the individual histone proteins.

4 °C). The concentration was measured, presence of all histones confirmed by SDS-PAGE and LC-MS. Purified octamers were stored at -20 °C at -20 °C for further use. If needed, the structural integrity of the histone octamer can be later checked by analytical SEC (Tricorn Superdex S200 10/300 GL, GE Healthcare).

4.6.9. Large scale DNA preparation for nucleosome reconstitution

4.6.9.1. Preparation of the 145 bp "601" DNA

The expression and purification of the 145 bp "601" DNA was carried out as described by Dyer *et al.*[814] The pUC57 plasmid encoding 145 bp "601" DNA was transformed into XL10-Gold cells. The transformed cells were plated on LB-agar plates containing 100 µg/mL ampicillin and incubated at 37 °C overnight. 100 mL of a starter culture in 2xYT supplemented with ampicillin were inoculated using a single colony of transformed cells. The starter culture was grown at 37 °C overnight. Six cultures with 800 mL TB media containing 100 µg/mL ampicillin were inoculated with 8 mL of starter culture and grown at 37 °C for 20 h. Cells were harvested by centrifugation (8000 rpm, 8 min, 4 °C) and resuspended in 60 mL lysis buffer ALS-1 (50 mM glucose, 25 mM Tris pH 8.0, 10 mM EDTA) per litre. 120 mL alkaline lysis buffer ALS-2 (0.2 M NaOH, 1 % (w/v) SDS) were added and thoroughly resuspended. Following an incubation for 10-20 min at 4 °C, 210 mL of pre-cooled alkaline buffer ALS-3 (4 M sodium acetate, 2 N acetic acid) were gently added to the mixture. Subsequent incubation step of 20 min at 4 °C was followed centrifugation (10000 g, 20 min, 4 °C). The supernatant was filtered through miracloth (Millipore), supplemented with 0.52 volume of isopropanol and incubated for 15 min at r.t. After a centrifugation step (10000 g, 30 min, 20 °C), the supernatant was discarded and the obtained pellet air-dried for 1 h at r.t. Dried pellet was dissolved in 20 mL TE 10/50 buffer (10 mM Tris pH 8.0, 50 mM EDTA). Addition of 60 µL RNase A (20 mg/mL, NEB) was followed an incubation step at 37 °C overnight.

After centrifugation (20000 rpm, 15 min), the supernatant was transferred into a fresh tube. The plasmid DNA was precipitated from supernatant by adding 1/5 of original volume of 4 M NaCl and 2/5 of 40 % PEG 6000. The obtained mixture was shaken at 37 °C for 5 min, incubated at 4 °C for 30 min and centrifuged (3750 rpm, 20 min). Plasmid purification was validated by 1 % agarose gel. The supernatant was discarded, the pellet was dissolved in 15 mL TE 10/0.1 buffer (10 mM Tris pH 8.0, 0.1 mM EDTA) and incubated at 37 °C overnight.

Restriction digestion was carried out to excise the 145 bp DNA fragment from the plasmid DNA by EcoRV enzyme. 30 units of EcoRV-HF (NEB) per nmol of restriction site were added to plasmid DNA dissolved in 1x CutSmart buffer (NEB). The restriction digest was performed at 37 °C for 16 h. The digest was confirmed using 1 % agarose gel. Once the reaction was complete, 0.192 volume of 4 M NaCl and 0.346 volume of 40 % PEG 6000 were added to precipitate the insert. The mixture was incubated at 4 °C for 1 h. After centrifugation (27000 g, 20 min, 4 °C), the supernatant was transferred into a fresh tube. The partition of plasmid DNA and the insert was checked by agarose gel, before precipitating the insert by adding 3 volumes of ice-cold ethanol. The mixture was incubated at 4 °C for 1 h and followed by centrifugation (27000 g, 20 min, 4 °C). The obtained insert was dissolved in TE 10/0.1 and further purified by phenol extraction and ethanol precipitation: 2/5 volumes of phenol and 1/5 volume of chloroform were added to the dissolved insert solution and mixed. The mixture was spun in a benchtop centrifuge and the upper aqueous layer mixed in a 1:1 ratio with chloroform, mixed and centrifuged. This step was repeated a second time. The aqueous layer was collected in an opaque centrifuge tube, 1/10 volume of 3M sodium acetate pH 5.2 and 3 volumes of ethanol were added and the DNA precipitated overnight at -20 °C. The mixture was centrifuged at 13000 rpm for 30 min and resuspended in 0.5-1 mL TE 10/0.1 The purified DNA was analysed by agarose gel and the sequence confirmed by Sanger sequencing. The concentration and the A_{260}/A_{280} was measured by nanodrop and the DNA was stored at − 20 °C for further use.

The sequence of 145 bp "601" DNA is given below:
ATCAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCA
CCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAGGGGA
TTACTCCCTAGTCTCCAGGCACGTGTCAGATATATACATCGAT

Primer used for Sanger sequencing:
Forward: 5'-GAGGCCGCTCAATTGGTC
Reverse: 5'-TATCTGACACGTGCCTGGAG

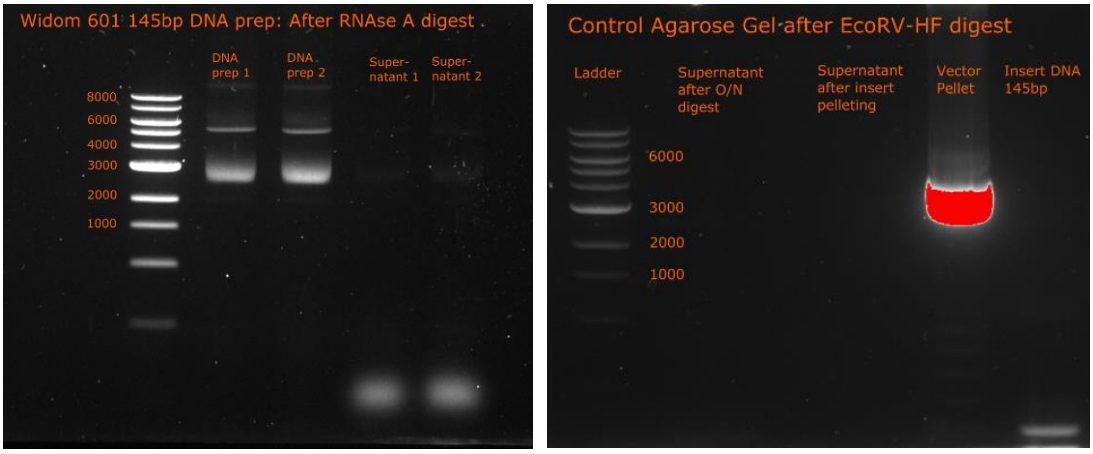


Figure 4.24. Preparation of 145bp Widom601 strong positioning DNA. 1 % agarose gel.

4.6.9.2. Preparation of the 186 bp biotinylated "601" DNA

The bacterial plasmid encoding a single repeat of 186 bp "601" DNA was kindly donated by Prof. Till Bartke, Imperial College London, UK. The procedure for large scale DNA preparation follows protocols described in literature.[778,814]

The pUC19 plasmid encoding 186 bp "601" DNA was transformed into XL10-Gold cells. The transformed cells were plated on LB-agar plates containing 100 µg/mL ampicillin and incubated at 37 °C overnight. 100 mL of a starter culture in 2xYT supplemented with ampicillin were inoculated using a single colony of transformed cells. The starter culture was grown at 37 °C overnight. Six cultures with 800 mL TB media containing 100 µg/mL ampicillin were inoculated with 8 mL of starter culture and grown at 37 °C for 20 h. Cells were harvested by centrifugation (8000 rpm, 8 min, 4 °C) and resuspended in 60 mL lysis buffer ALS-1 (50 mM glucose, 25 mM Tris pH 8.0, 10 mM EDTA) per litre. 120 mL alkaline lysis buffer ALS-2 (0.2 M NaOH, 1 % (w/v) SDS) were added and thoroughly resuspended. Following an incubation for 10-20 min at 4 °C, 210 mL of pre-cooled alkaline buffer ALS-3 (4 M sodium acetate, 2 N acetic acid) were gently added to the mixture. Subsequent incubation step of 20 min at 4 °C was followed centrifugation (10000 g, 20 min, 4 °C). The supernatant was filtered through miracloth (Millipore), supplemented with 0.52 volume of isopropanol and incubated for 15 min at r.t. After a centrifugation step (10000 g, 30 min, 20 °C), the supernatant was discarded and the obtained pellet air-dried for 1 h at r.t. Dried pellet was dissolved in 20 mL TE 10/50 buffer (10 mM Tris pH 8.0, 50 mM EDTA). Addition of 120 µL RNase A (10 mg/mL, Thermo Scientific) was followed an incubation step at 37 °C overnight.

After centrifugation (20000 rpm, 15 min), the supernatant was transferred into a fresh tube. The plasmid DNA was precipitated from supernatant by adding 1/5 of original volume of 4 M NaCl and 2/5 of 40 % PEG 6000. The obtained mixture was shaken at 37 °C for 5 min, incubated at 4 °C for 30 min and centrifuged (3750 rpm, 20 min). Plasmid purification was validated by 1 % agarose gel. The supernatant was discarded, the pellet was dissolved in 15 mL TE 10/0.1 buffer (10 mM Tris pH 8.0, 0.1 mM EDTA) and incubated at 37 °C overnight.

The excision of 186 bp insert from the plasmid DNA was carried out using EcoRV and EcoRI enzyme. For restriction digestion, 1 mg/mL plasmid and 150 µL of EcoRV and 300 u of EcoRI per mg DNA were mixed in 1x CutSmart buffer (NEB) and incubated at 37 °C overnight. The separation of the digested fragment was performed by precipitation of the plasmid DNA through addition NaCl and PEG 6000 with final concentrations of 0.5 M and 7 %, respectively. Following an incubation of 1 h at 4 °C and centrifugation (27000 g, 20 min, 4 °C), the supernatant was transferred into a fresh tube. Precipitation of the fragment was carried out by adding 2.5 volumes of ice-cold ethanol.

For biotinylation, 4 mg DNA fragment (0.407 mg/mL) were dissolved in 1x NEBuffer 2 and mixed with 100 µM dATP (NEB), 40 µM biotin-11-dUTP (York Biosciences) and 33.3 u/mL Klenow 3' to 5 exo' (NEB) and incubated at 25 °C for 2 h. The reaction was stopped by adding EDTA to a final concentration of 20 mM. The DNA was purified by phenol extraction followed by EtOH precipitation: 2/5 volumes of phenol and 1/5 volume of chloroform were added to the dissolved insert solution and mixed. The mixture was spun in a benchtop centrifuge and the upper aqueous layer mixed in a 1:1 ratio with chloroform, mixed and centrifuged. This step was repeated a second time. The aqueous layer was collected in an opaque centrifuge tube, 1/10 volume of 3M sodium acetate pH 5.2 and 3 volumes of ethanol were added and the DNA precipitated overnight at -20 °C. The mixture was centrifuged at 13000 rpm for 30 min and resuspended in 0.5-1 mL TE 10/0.1. The concentration and the A_{260}/A_{280} was measured by nanodrop and the DNA was stored at -20 °C until further use. The purified DNA was analysed by 3 % agarose gel (100 V 35 min), the sequence confirmed by Sanger sequencing and the success of the biotinylation reaction confirmed by anti-biotin Southern Blot.

The sequence of 186 bp biotinylated "601" DNA is given below: ATCGGATCTTACA TGCACAGGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCC CTTGGCGGTTAAAACGCGGGGGACAGCGCGTACGTGCGTTTAAGCGGTGC TAGAGCTGTCTACGACCAATTGAGCGGCCTCGGCACCGGGATTCTCCAGGG CGGCCGCGTAGTACTGAA**U(Biotin)U(Biotin)**

Primer used for Sanger sequencing:
Forward: 5'-TCGGATCTTACATGCACAGGA

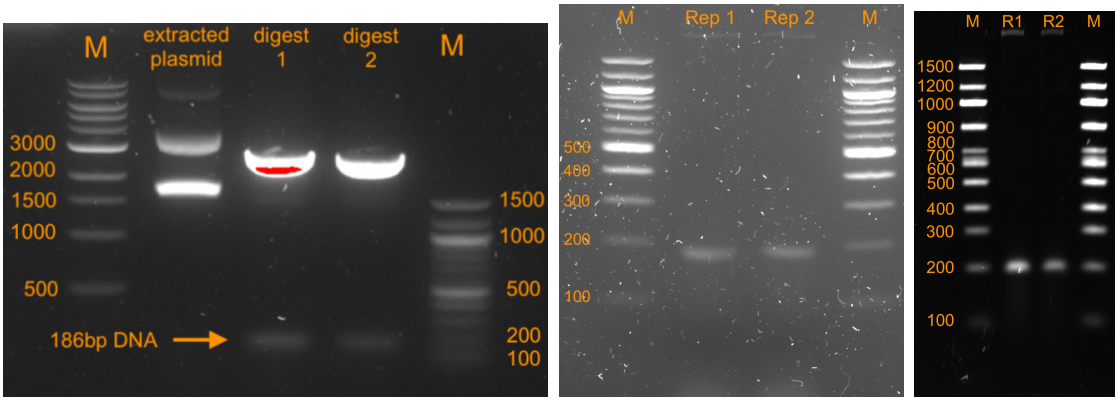


Figure 4.25. Preparation of 186bp Widom601 strong positioning DNA. Left. Double digest control, before biotinylation, 1 % TAE agarose gel. **Mid.** Final biotinylated DNA, 3 % TAE agarose gel. **Right.** Final biotinylated DNA, 6 % TBE polyacrylamide gel.

4.6.9.3. Southern Blot with anti-biotin detection

3 µL of a 1:10 dilution of the DNA (∼180 ng) and 2 µL 6× dye were diluted to a volume of 12 µL and run on a pre-equilibrated 6% TBE gel (Invitrogen) with 1× TBE buffer for 90 min at 150 V at 4 °C. The gel was stained for 30 min with SYBR Safe (Life Technologies) and destained for 15 min in MilliQ water and the image recorded on a transilluminator. The blotting procedure was performed with the iBlot dry blotting system (Invitrogen) following the manufacturer's instructions and materials. Briefly, the gel was placed between the anode stack (bottom) and the cathode stack (top) taking care to exclude air bubbles. The P8 program was used with 7 min transfer time. The transfer membrane was immersed in denaturing solution (1.5 M NaCl, 0.5 N NaOH) and incubated for 10 min on a rotary shaker for DNA denaturation. The membrane was subsequently air dryed for 5-10 min. Detection of the biotinylated membrane was performed using the Biotin Chromogenic Detection Kit (Thermo Scientific). The membrane was washed in 30 mL blocking/washing buffer for 5 min at r.t. on a rocking shaker before blocking in 30 mL of the Blocking solution for 30 min at r.t. on a rocking shaker. The membrane was then incubated with 20 mL freshly prepared Streptavidin-AP conjugate for 30 min. Subsequently, the membrane was washed 2 × with 60 mL of

Figure 4.26. Anti-biotin Southern Blot.

Blocking/Washing buffer for 15 min, followed by 20 mL of detection buffer for 10 min. The membrane was then incubated with 10 mL of freshly prepared Substrate Solution for 30 min in the dark to yield a purple-blue precipitate. The membrane was rinsed with MilliQ, air dryed and the blot recorded.

4.6.9.4. Purification of DNA fragments

In case the purity of DNA is not of satisfactory quality (low MW degradation products; longer storage etc.), further purification of the DNA might be necessary. The following techniques have been tried, the prep-scale PCR kit yielded the best result (see discussion in 4.3.4.2).

Ethanol precipitation
In an example precipitation, 50 µL of 186 bp biotinylated DNA (4.4 mg/mL) are placed in a 1.5 mL low-bind DNAse-free eppendorf tube. 1/2 volume (25 µL) of 5 M ammonium acetate was added (2 M final concentration) for the removal of dNTPs, followed by 2-3 volumes of ice-cold ethanol (150 µL, -20 °C). The solution is carefully mixed and frozen overnight at -20 °C. Alternatively, 1 h at -80 °C is also possible. The outermost edge of the tube was marked to find the pellet and the sample centrifuged (30 min, 13000 rpm, -9 °C or 4 °C). The supernatant was carefully discarded and the pellet air-dryed for 15 min or until white. The pellet was resuspended in the desired volume TE 10/0.1 and the concentration measured.

Desalting and nucleotide removal by G-25 size exclusion
The GE Illustra G-25 MicroSpin columns were used according to the manufacturer's instruction. The column was equilibrated with TE 10/0.1 and 50 µL of sample were applied.

Purification of 186 bp biotin-DNA using the Zymo 100 µg PCR purification kit
The DNA Clean & Concentrator-100 (DCC-100, Zymo Research) has a cut-off of 50 bp. Both the 500 µg and 100 µg scale were successfully tested, better results were obtained with the 100 µg scale (minimised losses). The purification was performed following the manufacturer's instructions. Briefly, 25 µL of DNA (85 µg, 3.94 mg/mL) was added to 125 µL DNA binding buffer (Zymo) and loaded onto the DCC-100 spin column. The column was centrifuged (13000 rpm, 1 min, r.t.). 600 µL of wash buffer (Zymo) was added and the columns centrifuged. The column was then centrifuged empty. For elution of the DNA, 150 L TE 10/0.1 was added to the column placed in a fresh tube and incubated for 1 min, before elution (13000 rpm, 1 min, r.t.). The DNA was then

purified by ethanol precipitation as described above, using 75 µL of 5 M ammonium acetate and 450 µL ice-cold EtOH, with precipitation for 2 h at -80 °C and pelleting (30 min, 13000 rpm, -9 °C) before air-drying and re-suspension. The resulting DNA is analysed for QC by 3 % agarose gel (100 V 35 min) or by 6 % TBE gel. Should the preparation still contain unwanted impurities, the entire procedure can be repeated a second time to increase purity.

4.6.10. Nucleosome reconstitution

The procedure for nucleosome reconstitution from purified histone octamers (3-15 mg/mL) using 145 bp DNA and biotinylated 186 bp DNA (3-6 mg/mL) follows protocols described in literature.[780,814]

For identification of the appropriate ratios of octamer to DNA for nucleosome reconstitution, small scale (10-20 µg) titration experiments were conducted using 1:0.85, 1:1 and 1:1.15 molar ratios of octamer to DNA in high salt (TCS) buffer (2 M KCl, 10 mM Tris pH 7.5, 1 mM EDTA, 1 mM DTT) at 4 °C. The final DNA concentration was 0.2 - 0.7 mg/mL and the volume kept smaller than 100 µL. Components were added on ice in the following order: TCS, DNA, 4M KCl, Octamer. The reaction mixtures were transferred to a 100 µL dialysis cups (MWCO 3500 or 6000) and dialysed at 4 °C against low salt buffers containing 0.85 M KCl, 0.65 M KCl and 0.25 M KCl in three steps with a duration of 2 h for each step. After dialysis, the samples were transferred to eppendorf tubes and centrifuged to remove any precipitation. The assembled nucleosomes were analysed by Native-PAGE using Novex 6 % TBE gels (Invitrogen). The TBE gel was pre-run in 0.5 % TBE buffer at 150 V for 90 minutes at 4 °C. 0.5-1 µg (1 µL) of nucleosome samples along with 5 µL of 6 % sucrose solution were loaded per well. The 100bp DNA ladder (NEB) and the Native-PAGE protein marker (Invitrogen) were added. The gel was run at 150 V for 90 min. The gel was stained with SYBR Gold for 30 min in the dark and destained for 15 min in water before being recorded on a transilluminator (Bio-Rad). For protein detection, the gel was stained overnight with Coomassie-G250 (20 mL stainer A, 5 mL stainer B, 20 mL MeOH, 55 mL MilliQ). The gel was destained the next morning with 8 % acetic acid followed by MilliQ water and recorded. The octamer to DNA ratio of 1:0.85 was identified as optimal for reconstitution, reconstitutions with up to 100 µg were done using the protocol described above.

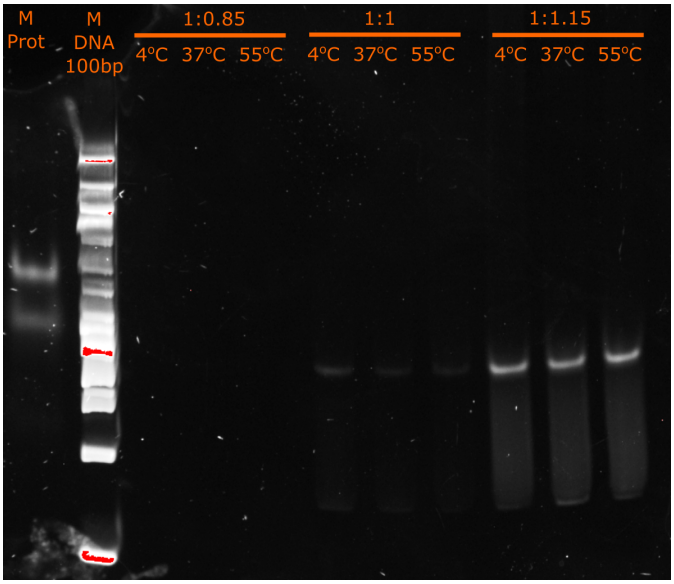


Figure 4.27. Nucleosome assembly with 145bp Widom601 strong positioning DNA. Other reconstitutions are shown in Figure 4.13.

4.6.11. Preparation of nuclear cell extract

HeLa cells tissue culture experiments were carried out under sterile conditions inside a HERAsafe KS-12 Class II biological safety cabinet (Thermo Fischer Scientific). Biological replicates were obtained using HeLa cells from three different batches. Cells collected for this experiment were 5 passages away from the certified source. HeLa cells were grown to near-confluency in a HYPERFlask™ cell culture vessel (Corning Life Sciences) in 500 mL DMEM medium supplemented with 10 % FBS. Trypsinisation and 4 mM L-glutamine ($\sim 5.710^7$ cells, 96 % alive). and subsequent washing step was performed using 10 mL of pre-cooled PBS buffer and centrifuged (300 g, 5 min, 4 °C).

The procedure for preparation of nuclear extract was followed the protocol described in Lercher *et al.*[773] and outlined below. The following inhibitor cocktail was used: 1× EDTA-free protease inhibitor cocktail (Roche), 5 mM β-glycerophosphate, 5 mM sodium fluoride, 0.5 µM PMSF, 0.2 mM Na_3VO_4 and 2 mM DTT. The sodium orthovanadate (Na_3VO_4) stock solution was prepared as described in the literature[841], i.e. a 500 mM stock solution was prepared by adjusting the pH to 10. The yellow solution is then boiled for 10 min until the colour clears, cooled to r.t. and the pH adjusted again. This is repeated until no colour change is observed.

After centrifugation cells were resuspended in 10 mL of lysis buffer (10 mM Tris pH 7.5, 10 mM NaCl, 10 mM $MgCl_2$, 250 mM sucrose, 0.1 mM EGTA, 0.5 % NP-40)

and vortexed for 10 seconds. Following an incubation of 15 min at 4 °C, cells were transferred onto a 40 mL sucrose cushion (30 % sucrose, 10 mM Tris pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$) and centrifuged (1300 g, 10 min, 4 °C) to separate nuclei from cell debris. The supernatant was discarded and the isolated nuclei were washed using 10 mL of wash buffer (10 mM Tris pH 7.5, 10 mM NaCl). The pellet containing nuclei was resuspended in 1 mL of extraction buffer (50 mM HEPES, 420 mM NaCl, 0.5 mM EDTA, 0.1 mM EGTA, 10 % glycerol and 0.1 % NP-40) and sonicated for 30 seconds. After an incubation of 30 min at 4 °C and centrifugation (13000 g, 10 min, 4 °C), the supernatant containing the nuclear extract was collected. The concentration of the nuclear extract was determined using a BCA assay and adjusted to a final concentration of 0.5 mg/mL using binding buffer (20 mM HEPES pH 7.9, 150 mM NaCl, 0.2 mM EDTA, 20 %glycerol, 0.1 % DTT supplemented with the above protease and phosphatase inhibitors). The aliquots were flash frozen in liquid nitrogen and stored at -80 °C for further use.

4.6.12. Interaction proteomics experiment with synthetic nucleosomes

The interaction proteomics experiments were carried out as described.[778,780] The following inhibitor cocktail was used: 1× EDTA-free protease inhibitor cocktail (Roche), 5 mM β-glycerophosphate, 5 mM sodium fluoride, 0.5 µM PMSF, 0.2 mM Na_3VO_4 and 2 mM DTT. The sodium orthovanadate (Na_3VO_4) stock solution was prepared as described in the literature.[841]

100 µL Dynabeads® Streptavidin MyOne™ T1 (Life Technologies) were washed with 1 mL reconstitution buffer (10 mM Tris pH 7.5, 250 mM KCl, 1 mM EDTA, 1 mM DTT, 0.1 % NP-40), followed by 4 × 100 µL reconstitution buffer. 20 µg of each nucleosome were immobilised on the Dynabeads separately, for 30 min using reconstitution buffer. The immobilisation process was monitored by measuring the absorption at 260 nm of the supernatant. Immobilised nucleosomes were washed 3-4 × using binding buffer (20 mM HEPES pH 7.9, 150 mM NaCl, 0.2 mM EDTA, 20 % glycerol, 0.1 % NP-40, 0.1 % DTT supplemented with protease and phosphatase inhibitors). Immobilised nucleosomes were incubated with 1 mL of nuclear extract (0.5 mg/mL in binding buffer) for 4 h at 4 °C under gentle mixing. In 5 × washing step with 1 mL of binding buffer, non-specifically bound protein were removed. After transfer to a fresh tube, the interaction partners were incubated with 100 µL of elution buffer (8 M urea, 2 M NaCl) for 30 min at 4 °C under gentle rotation and the supernatant collected. The protein concentration was determined by BCA assay.

Pulled interacting partners from each nucleosome were digested by trypsin using the Filter-Aided Sample Preparation (FASP) protocol.[155] Protein samples were loaded onto a pre-equilibrated 10 kDa Microcon filtration device (Millipore). Following centrifugation (14000 g, 30 min), the flow-through was discarded. The concentrates were washed twice with 200 µL 100 mM TEAB with 8 M urea (14000 g, 30 min). The concentrate was reduced in the filtration device by adding 200 µL of reduction buffer (8 M urea, 100 mM TEAB, 20 mM TCEP) and incubated at 37 °C for 30 min. Subsequently, samples were alkylated by adding 200 µL of alkylation buffer (8 M urea, 100 mM TEAB, 50 mM iodoacetamide) and incubated for 30 min at r.t. in the dark. Excess of alkylating reagent was removed by centrifugation (14000 g, 30 min). After 3 wash steps using 200 µL of 8 M urea and 50 mM ammonium bicarbonate buffer and subsequent centrifugation, the collection tube was replaced by a fresh tube. Lys-C enzyme was mixed with each concentrate in a final 1:40 (*w/w*) enzyme to protein ratio in 35 µL of 8 M urea and 50 mM TEAB and incubated for 4 h at 37 °C. Then trypsin was added (1:20 *w/w*) with 120 µL of 50 mM TEAB and incubated at 37 °C overnight. After the incubation, peptides were eluted into fresh tubes via centrifugation (14000 g, 30 min). Centrifugal units were washed with 200 µL of 0.1 % TFA collecting in the same tube, followed by a quick wash step with 200 µL of 50 % acetonitrile 0.1 % TFA solution (14000 g) to ensure the elution of all remaining peptides. The eluted peptide solution was dried in a speed-vac for 4 h.

Resulting tryptic peptides were analysed n an EASY-nLC 1000 UHPLC system (Proxeon) connected to a Q Exactive mass spectrometer (Thermo Fisher Scientific) through an EASY-Spray nano-electrospray ion source (Thermo Fisher Scientific). The peptides were trapped on an in-house packed column (75 µm i.d. x 20 mm, reprosil C18, 3 µm, 120 Å) using solvent A (0.1 % formic acid, 5 % DMSO in water) at 500 bar. The peptides were separated on an EASY-spray Acclaim PepMap® analytical column (75 µm i.d. x 500 mm, RSLC C18, 2 µm, 100 Å) using a linear gradient of 120 min from 8 % to 28 % solvent B (0.1 % formic acid, 5 % DMSO in acetonitrile) at 200 nL/min. The separated peptides were electrosprayed directly into the mass spectrometer operating in a data-dependent mode. Full scan MS spectra were acquired in the Orbitrap (scan range 350-1500 m/z, resolution 70000, AGC target 3e6, maximum injection time 100 ms). After the MS scans, the 20 most intense peaks were selected for HCD fragmentation at 30 % of normalised collision energy in the Orbitrap (resolution 17500, AGC target 5e4, maximum injection time 120 ms).

The raw data were processed using MaxQuant (version 1.6.12.0) and integrated with Andromeda search engine. Raw MS data were searched against human database (Swiss-

Prot, version 04/13) and checked against a list of common contaminants by Andromeda. Trypsin with a maximum number of two missed cleavages was selected as the protease. Acetylation (N-terminus of the protein) and oxidation (methionine) were used as variable modifications; carbamidomethylation (cysteine) was set as a fixed modification. Protein and PSM false discovery rate (FDR) were set at 0.01. Label-free quantification (LFQ) was carried out using the build-in MaxLFQ algorithm of MaxQuant[381] allowing match between runs. Further downstream analysis performed using Perseus software (version 1.6.12.0). LFQ intensities of all identified protein groups were exported; reverse, contaminants and protein groups only identified by site were removed. Protein groups identified in all three replicates of either bait or control samples were analysed. LFQ intensities were logarithmised and missing LFQ intensities were replaced by a constant, 1 unit lower than the lowest quantified value. FDR based two-sample t-test was used to generate fold-change difference as well as a p value. Protein groups with corresponding p value lower than 0.05 were considered significant. The data was visualised in an volcano plot of ratio change versus log(p value).

For the analysis of TCF20 interaction partners, the following list of gene names was compiled from the data deposited in BioGRID, IntAct, STRING and BioPlex databases. Actually identified interaction partners are highlighted in bold: TOPBP1, AR, RNF4 (we see RNF2), **HIST1H2BA**, **HIST1H4A**; ACTL8, ADARB1, **ADARB2**, AIM2, ARV1, **BAG1**, **BAG2**, BRF1, CDC37, CDK2AP1, **CLTB**, **COPA / COPB1**, ESR2, **EWSR1**, FBXO21, FKBP9, **FMR1**, GATAD2A, GATAD2B, **H2AFZ**, **H2BC1**, HARS, **HDAC1**, **HIST1H2AB**, **HIST1H2BB**, **H2H2AC**, **HIST2H2BE**, **HMG20A**, JUN, KIAA1429, **KMT2A**, MAPK8, MAPK14, **MDC1**, MTA1, **MTA2**, MTA3, MYC, NANOG, NFE2L2, NIPSNAP1, NKX2-1, NOP9, NR3C1, OBSL1, **PHF14**, POU5F1, PRDM16, **PTBP3**, PYHIN1, **RAI1**, **RBBP7**, **RPS7**, **SIN3A**, SIRT7, SLC39A10, SMAD1, SOX2, SPIRE2, TRIM25, **TRIM27**, TRIM63, USP7, **XRCC6**, YWHAH.

4.7. References

61. Stepper, J., Shastri, S., Loo, T. S., Preston, J. C., Novak, P., Man, P., Moore, C. H., Havlíček, V., Patchett, M. L. & Norris, G. E. Cysteine S-Glycosylation, a New Post-Translational Modification Found in Glycopeptide Bacteriocins. *FEBS letters* **585,** 645–650 (2011).

63. Maynard, J. C., Burlingame, A. L. & Medzihradszky, K. F. Cysteine S-Linked N-Acetylglucosamine (S-GlcNAcylation), A New Post-Translational Modification in Mammals. *Molecular & Cellular Proteomics* **15,** 3405–3411 (2016).

119. Médard, G. *et al.* Optimized Chemical Proteomics Assay for Kinase Inhibitor Profiling. *Journal of Proteome Research* **14,** 1574–1586 (2015).

155. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal Sample Preparation Method for Proteome Analysis. *Nature Methods* **6,** 359–362 (5 2009).

156. Wiśniewski, J. R. Filter Aided Sample Preparation – A Tutorial. *Analytica Chimica Acta* **1090,** 23–30 (2019).

269. Dieterich, D. C., Link, A. J., Graumann, J., Tirrell, D. A. & Schuman, E. M. Selective Identification of Newly Synthesized Proteins in Mammalian Cells Using Bioorthogonal Noncanonical Amino Acid Tagging (BONCAT). *Proceedings of the National Academy of Sciences* **103,** 9482–9487 (2006).

280. Kempkes, L.J.M. *Reaction Mechanisms of Collision and Electron Induced Peptide Dissociation Revealed by Ion Spectroscopy* (Radboud University Nijmegen, 2019).

281. Tureček, F. & Julian, R. R. Peptide Radicals and Cation Radicals in the Gas Phase. *Chemical Reviews* **113,** 6691–6733 (2013).

305. 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. *International Journal of Proteomics* **2012,** e560391 (2012).

306. Frese, C. K., Altelaar, A. F. M., van den Toorn, H., Nolting, D., Griep-Raming, J., Heck, A. J. R. & Mohammed, S. Toward Full Peptide Sequence Coverage by Dual Fragmentation Combining Electron-Transfer and Higher-Energy Collision Dissociation Tandem Mass Spectrometry. *Analytical Chemistry* **84,** 9668–9673 (2012).

362. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P. & Mann, M. Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell* **127,** 635–648 (2006).

379. Tyanova, S., Temu, T. & Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nature Protocols* **11,** 2301–2319 (12 2016).

381. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N. & Mann, M. Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics* **13,** 2513–2526 (2014).

419. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M. & Cox, J. The Perseus Computational Platform for Comprehensive Analysis of (Prote)Omics Data. *Nature Methods* **13,** 731–740 (9 2016).

430. Huttlin, E. L. *et al.* Architecture of the Human Interactome Defines Protein Communities and Disease Networks. *Nature* **545,** 505–509 (7655 2017).

443. Huang, H., Lin, S., Garcia, B. A. & Zhao, Y. Quantitative Proteomic Analysis of Histone Modifications. *Chemical Reviews* **115,** 2376–2418 (2015).

492. Liu, Y. *et al.* Multi-Omic Measurements of Heterogeneity in HeLa Cells across Laboratories. *Nature Biotechnology* **37,** 314–322 (3 2019).

493. Wang, D. *et al.* A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues. *Molecular Systems Biology* **15,** e8503 (2019).

509. Chen, Q. & Yu, X. OGT Restrains the Expansion of DNA Damage Signaling. *Nucleic Acids Research* **44,** 9266–9278 (2016).

610. Singh, N. *et al.* Dual Recognition of Phosphoserine and Phosphotyrosine in Histone Variant H2A.X by DNA Damage Response Protein MCPH1. *Proceedings of the National Academy of Sciences* **109,** 14381–14386 (2012).

704. Nadal, S., Raj, R., Mohammed, S. & Davis, B. G. Synthetic Post-Translational Modification of Histones. *Current Opinion in Chemical Biology. Molecular Imaging / Chemical Genetics and Epigenetics* **45,** 35–47 (2018).

705. Pombo, A. & Dillon, N. Three-Dimensional Genome Architecture: Players and Mechanisms. *Nature Reviews Molecular Cell Biology* **16,** 245 (2015).

706. Fraser, J., Williamson, I., Bickmore, W. A. & Dostie, J. An Overview of Genome Organization and How We Got There: From FISH to Hi-C. *Microbiology and Molecular Biology Reviews* **79,** 347–372 (2015).

707. Bannister, A. J. & Kouzarides, T. Regulation of Chromatin by Histone Modifications. *Cell Research* **21,** 381 (2011).

708. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128,** 693–705 (2007).

709. Musselman, C. A., Côté, J., Lalonde, M.-E. & Kutateladze, T. G. Perceiving the Epigenetic Landscape through Histone Readers. *Nature Structural and Molecular Biology* **19,** 1218 (2012).

710. Dhall, A. & Chatterjee, C. Chemical Approaches To Understand the Language of Histone Modifications. *ACS Chemical Biology* **6,** 987–999 (2011).

711. Allis, C. D. & Muir, T. W. Spreading Chromatin into Chemical Biology. *ChemBioChem* **12,** 264–279 (2011).

712. Shogren-Knaak, M. A. & Peterson, C. L. in *Methods in Enzymology* Supplement C vols., 62–76 (Academic Press, 2003).

713. Fierz, B. Synthetic Chromatin Approaches To Probe the Writing and Erasing of Histone Modifications. *ChemMedChem* **9,** 495–504 (2014).

714. Schwarzer, D. Chemical Tools in Chromatin Research. *Journal of Peptide Science* **16,** 530–537 (2010).

715. Nikolov, M. & Fischle, W. Systematic Analysis of Histone Modification Readout. *Molecular BioSystems* **9,** 182–194 (2013).

716. Pick, H., Kilic, S. & Fierz, B. Engineering Chromatin States: Chemical and Synthetic Biology Approaches to Investigate Histone Modification Function. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. Molecular Mechanisms of Histone Modification Function* **1839,** 644–656 (2014).

717. Fischle, W., Mootz, H. D. & Schwarzer, D. Synthetic Histone Code. *Current Opinion in Chemical Biology. Synthetic Biology • Synthetic Biomolecules* **28,** 131–140 (Supplement C 2015).

718. Holt, M. & Muir, T. Application of the Protein Semisynthesis Strategy to the Generation of Modified Chromatin. *Annual Review of Biochemistry* **84,** 265–290 (2015).

719. Müller, M. M. & Muir, T. W. Histones: At the Crossroads of Peptide and Protein Chemistry. *Chemical Reviews* **115,** 2296–2349 (2015).

720. Chuh, K. N., Batt, A. R. & Pratt, M. R. Chemical Methods for Encoding and Decoding of Posttranslational Modifications. *Cell Chemical Biology* **23,** 86–107 (2016).

721. Yang, A., Cho, K. & Park, H.-S. Chemical Biology Approaches for Studying Posttranslational Modifications. *RNA Biology* **0,** 1–14 (2017).

722. Dai, L. *et al.* Lysine 2-Hydroxyisobutyrylation Is a Widely Distributed Active Histone Mark. *Nature Chemical Biology* **10,** nchembio.1497 (2014).

723. Howden, A. J. M. *et al.* QuaNCAT: Quantitating Proteome Dynamics in Primary Cells. *Nature Methods* **10,** 343 (2013).

724. Arnaudo, A. M., Link, A. J. & Garcia, B. A. Bioorthogonal Chemistry for the Isolation and Study of Newly Synthesized Histones and Their Modifications. *ACS Chemical Biology* **11,** 782–791 (2016).

725. Lin, S. *et al.* Stable-Isotope-Labeled Histone Peptide Library for Histone Post-Translational Modification and Variant Quantification by Mass Spectrometry. *Molecular & Cellular Proteomics* **13,** 2450–2466 (2014).

726. Kulej, K., Avgousti, D. C., Weitzman, M. D. & Garcia, B. A. Characterization of Histone Post-Translational Modifications during Virus Infection Using Mass Spectrometry-Based Proteomics. *Methods (San Diego, Calif.)* **90,** 8–20 (2015).

727. Wang, L., Brock, A., Herberich, B. & Schultz, P. G. Expanding the Genetic Code of Escherichia Coli. *Science* **292,** 498–500 (2001).

728. Lang, K. & Chin, J. W. Cellular Incorporation of Unnatural Amino Acids and Bioorthogonal Labeling of Proteins. *Chemical Reviews* **114,** 4764–4806 (2014).

729. Chin, J. W. Expanding and Reprogramming the Genetic Code. *Nature* **550,** 53–60 (2017).

730. Zhang, M. S., Brunner, S. F., Huguenin-Dezot, N., Liang, A. D., Schmied, W. H., Rogerson, D. T. & Chin, J. W. Biosynthesis and Genetic Encoding of Phosphothreonine through Parallel Selection and Deep Sequencing. *Nature Methods* **14,** 729–736 (2017).

731. Des Soye, B. J., Patel, J. R., Isaacs, F. J. & Jewett, M. C. Repurposing the Translation Apparatus for Synthetic Biology. *Current Opinion in Chemical Biology. Synthetic Biology • Synthetic Biomolecules* **28,** 83–90 (2015).

732. Dumas, A., Lercher, L., Spicer, C. D. & Davis, B. G. Designing Logical Codon Reassignment – Expanding the Chemistry in Biology. *Chemical Science* **6,** 50–69 (2014).

733. He, S., Bauman, D., Davis, J. S., Loyola, A., Nishioka, K., Gronlund, J. L., Reinberg, D., Meng, F., Kelleher, N. & McCafferty, D. G. Facile Synthesis of Site-Specifically Acetylated and Methylated Histone Proteins: Reagents for Evaluation of the Histone Code Hypothesis. *Proceedings of the National Academy of Sciences* **100,** 12033–12038 (2003).

734. Shogren-Knaak, M. A., Fry, C. J. & Peterson, C. L. A Native Peptide Ligation Strategy for Deciphering Nucleosomal Histone Modifications. *Journal of Biological Chemistry* **278,** 15744–15748 (2003).

735. Maity, S. K., Jbara, M. & Brik, A. Chemical and Semisynthesis of Modified Histones. *Journal of Peptide Science* **22,** 252–259 (2016).

736. Jbara, M., Maity, S. K. & Brik, A. Palladium in the Chemical Synthesis and Modification of Proteins. *Angewandte Chemie International Edition* **56,** 10644–10655 (2017).

737. Maity, S. K., Jbara, M., Mann, G., Kamnesky, G. & Brik, A. Total Chemical Synthesis of Histones and Their Analogs, Assisted by Native Chemical Ligation and Palladium Complexes. *Nature Protocols* **12,** 2293–2322 (2017).

738. Bondalapati, S., Jbara, M. & Brik, A. Expanding the Chemical Toolbox for the Synthesis of Large and Uniquely Modified Proteins. *Nature Chemistry* **8,** 407–418 (2016).

739. Siman, P. & Brik, A. Chemical and Semisynthesis of Posttranslationally Modified Proteins. *Organic & Biomolecular Chemistry* **10,** 5684–5697 (2012).

740. McGinty, R. K., Kim, J., Chatterjee, C., Roeder, R. G. & Muir, T. W. Chemically Ubiquitylated Histone H2B Stimulates hDot1L-Mediated Intranucleosomal Methylation. *Nature* **453,** 812–816 (2008).

741. Ajish Kumar, K. S., Haj-Yahya, M., Olschewski, D., Lashuel, H. A. & Brik, A. Highly Efficient and Chemoselective Peptide Ubiquitylation. *Angewandte Chemie International Edition* **48,** 8090–8094 (2009).

742. Zioudrou, C., Wilchek, M. & Patchornik, A. Conversion of the L-Serine Residue to an L-Cysteine Residue in Peptides*. *Biochemistry* **4,** 1811–1822 (1965).

743. Polgar, L. & Bender, M. L. A New Enzyme Containing a Synthetically Formed Active Site. Thiol-Subtilisin1. *Journal of the American Chemical Society* **88,** 3153–3154 (1966).

744. Neet, K. E. & Koshland, D. E. The Conversion of Serine at the Active Site of Subtilisin to Cysteine: A "Chemical Mutation". *Proceedings of the National Academy of Sciences of the United States of America* **56,** 1606–1611 (1966).

745. Chalker, J. M. & Davis, B. G. Chemical Mutagenesis: Selective Post-Expression Interconversion of Protein Amino Acid Residues. *Current Opinion in Chemical Biology. Model Systems/Biomolecular Synthesis and Modification* **14,** 781–789 (2010).

746. Wright, T. H., Vallée, M. R. J. & Davis, B. G. From Chemical Mutagenesis to Post-Expression Mutagenesis: A 50 Year Odyssey. *Angewandte Chemie International Edition* **55,** 5896–5903 (2016).

747. Gunnoo, S. B. & Madder, A. Chemical Protein Modification through Cysteine. *ChemBioChem* **17,** 529–553 (2016).

748. Chalker, J. M., Lercher, L., Rose, N. R., Schofield, C. J. & Davis, B. G. Conversion of Cysteine into Dehydroalanine Enables Access to Synthetic Histones Bearing Diverse Post-Translational Modifications. *Angewandte Chemie International Edition* **51,** 1835–1839 (2012).

749. Simon, M. D., Chu, F., Racki, L. R., de la Cruz, C. C., Burlingame, A. L., Panning, B., Narlikar, G. J. & Shokat, K. M. The Site-Specific Installation of Methyl-Lysine Analogs into Recombinant Histones. *Cell* **128,** 1003–1012 (2007).

750. Le, D. D., Cortesi, A. T., Myers, S. A., Burlingame, A. L. & Fujimori, D. G. Site-Specific and Regiospecific Installation of Methylarginine Analogues into Recombinant Histones and Insights into Effector Protein Binding. *Journal of the American Chemical Society* **135,** 2879–2882 (2013).

751. Huang, R. *et al.* Site-Specific Introduction of an Acetyl-Lysine Mimic into Peptides and Proteins by Cysteine Alkylation. *Journal of the American Chemical Society* **132,** 9986–9987 (2010).

752. Li, F., Allahverdi, A., Yang, R., Lua, G. B. J., Zhang, X., Cao, Y., Korolev, N., Nordenskiöld, L. & Liu, C.-F. A Direct Method for Site-Specific Protein Acetylation. *Angewandte Chemie International Edition* **50,** 9611–9614 (2011).

753. Seeliger, D., Soeroes, S., Klingberg, R., Schwarzer, D., Grubmüller, H. & Fischle, W. Quantitative Assessment of Protein Interaction with Methyl-Lysine Analogues by Hybrid Computational and Experimental Approaches. *ACS Chemical Biology* **7,** 150–154 (2012).

754. Chalker, J. M., Bernardes, G. J. L. & Davis, B. G. A "Tag-and-Modify" Approach to Site-Selective Protein Modification. *Accounts of Chemical Research* **44,** 730–741 (2011).

755. Chalker, J. M., Gunnoo, S. B., Boutureira, O., Gerstberger, S. C., Fernández-González, M., Bernardes, G. J. L., Griffin, L., Hailu, H., Schofield, C. J. & Davis, B. G. Methods for Converting Cysteine to Dehydroalanine on Peptides and Proteins. *Chemical Science* **2,** 1666–1676 (2011).

756. Dadová, J., Galan, S. R. & Davis, B. G. Synthesis of Modified Proteins via Functionalization of Dehydroalanine. *Current Opinion in Chemical Biology. Synthetic Biology / Synthetic Biomolecules* **46,** 71–81 (2018).

757. Lin, Y. A., Boutureira, O., Lercher, L., Bhushan, B., Paton, R. S. & Davis, B. G. Rapid Cross-Metathesis for Reversible Protein Modifications via Chemical Access to Se-Allyl-Selenocysteine in Proteins. *Journal of the American Chemical Society* **135,** 12156–12159 (2013).

758. Dadová, J., Wu, K.-J., Isenegger, P. G., Errey, J. C., Bernardes, G. J. L., Chalker, J. M., Raich, L., Rovira, C. & Davis, B. G. Precise Probing of Residue Roles by Post-Translational β,γ-C,N Aza-Michael Mutagenesis in Enzyme Active Sites. *ACS Central Science* **3,** 1168–1173 (2017).

759. Wright, T. H. *et al.* Posttranslational Mutagenesis: A Chemical Strategy for Exploring Protein Side-Chain Diversity. *Science* **354,** aag1465 (2016).

760. Wright, T. H. & Davis, B. G. Post-Translational Mutagenesis for Installation of Natural and Unnatural Amino Acid Side Chains into Recombinant Proteins. *Nature Protocols* **12,** 2243–2250 (2017).

761. Behrendt, R., White, P. & Offer, J. Advances in Fmoc Solid-Phase Peptide Synthesis. *Journal of Peptide Science* **22,** 4–27 (2016).

762. Zhang, Y., Muthana, S. M., Farnsworth, D., Ludek, O., Adams, K., Barchi, J. J. & Gildersleeve, J. C. Enhanced Epimerization of Glycosylated Amino Acids During Solid-Phase Peptide Synthesis. *Journal of the American Chemical Society* **134,** 6316–6325 (2012).

763. Galan, S. *Site-Selective Chemical Phosphorylation and Activation of Kinases* http://purl.org/dc/dcmitype/Text (University of Oxford, 2016).

764. Yang, A., Ha, S., Ahn, J., Kim, R., Kim, S., Lee, Y., Kim, J., Söll, D., Lee, H.-Y. & Park, H.-S. A Chemical Biology Route to Site-Specific Authentic Protein Modifications. *Science* **354,** 623–626 (2016).

765. Guo, J., Wang, J., Lee, J. S. & Schultz, P. G. Site-Specific Incorporation of Methyl- and Acetyl-Lysine Analogues into Recombinant Proteins. *Angewandte Chemie International Edition* **47,** 6399–6401 (2008).

766. Wang, Z. U., Wang, Y.-S., Pai, P.-J., Russell, W. K., Russell, D. H. & Liu, W. R. A Facile Method to Synthesize Histones with Posttranslational Modification Mimics. *Biochemistry* **51,** 5232–5234 (2012).

767. Galan, S. R. G. *et al.* Post-Translational Site-Selective Protein Backbone α-Deuteration. *Nature Chemical Biology* **14,** 955 (2018).

768. Tessarz, P. & Kouzarides, T. Histone Core Modifications Regulating Nucleosome Structure and Dynamics. *Nature Reviews Molecular Cell Biology* **15,** 703 (2014).

769. Ruthenburg, A. J., Allis, C. D., Patel, D. J., Li, H. & Taverna, S. D. How Chromatin-Binding Modules Interpret Histone Modifications: Lessons from Professional Pocket Pickers. *Nature Structural and Molecular Biology* **14,** 1025 (2007).

770. Barski, A. *et al.* Combinatorial Patterns of Histone Acetylations and Methylations in the Human Genome. *Nature Genetics* **40,** 897 (2008).

771. Cerbo, V. D. *et al.* Acetylation of Histone H3 at Lysine 64 Regulates Nucleosome Dynamics and Facilitates Transcription. *eLife* **3,** e01632 (2014).

772. North, J. A., Šimon, M., Ferdinand, M. B., Shoffner, M. A., Picking, J. W., Howard, C. J., Mooney, A. M., van Noort, J., Poirier, M. G. & Ottesen, J. J. Histone H3 Phosphorylation near the Nucleosome Dyad Alters Chromatin Structure. *Nucleic Acids Research* **42,** 4922–4933 (2014).

773. Lercher, L., Raj, R., Patel, N. A., Price, J., Mohammed, S., Robinson, C. V., Schofield, C. J. & Davis, B. G. Generation of a Synthetic GlcNAcylated Nucleosome Reveals Regulation of Stability by H2A-Thr101 GlcNAcylation. *Nature Communications* **6,** ncomms8978 (2015).

774. Yan, Q., Dutt, S., Xu, R., Graves, K., Juszczynski, P., Manis, J. P. & Shipp, M. A. BBAP Monoubiquitylates Histone H4 at Lysine 91 and Selectively Modulates the DNA Damage Response. *Molecular Cell* **36,** 110–120 (2009).

775. Ye, J., Ai, X., Eugeni, E. E., Zhang, L., Carpenter, L. R., Jelinek, M. A., Freitas, M. A. & Parthun, M. R. Histone H4 Lysine 91 Acetylation: A Core Domain ModificationAssociated with Chromatin Assembly. *Molecular Cell* **18,** 123–130 (2005).

776. Casadio, F., Lu, X., Pollock, S. B., LeRoy, G., Garcia, B. A., Muir, T. W., Roeder, R. G. & Allis, C. D. H3R42me2a Is a Histone Modification with Positive Transcriptional Effects. *Proceedings of the National Academy of Sciences* **110,** 14894–14899 (2013).

777. Lu, X., Simon, M. D., Chodaparambil, J. V., Hansen, J. C., Shokat, K. M. & Luger, K. The Effect of H3K79 Dimethylation and H4K20 Trimethylation on Nucleosome and Chromatin Structure. *Nature Structural & Molecular Biology* **15,** 1122–1124 (2008).

778. Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S. C., Mann, M. & Kouzarides, T. Nucleosome-Interacting Proteins Regulated by DNA and Histone Methylation. *Cell* **143,** 470–484 (2010).

779. Nikolov, M., Stützer, A., Mosch, K., Krasauskas, A., Soeroes, S., Stark, H., Urlaub, H. & Fischle, W. Chromatin Affinity Purification and Quantitative Mass Spectrometry Defining the Interactome of Histone Modification Patterns. *Molecular & Cellular Proteomics* **10,** M110.005371 (2011).

780. Raj, R., Lercher, L., Mohammed, S. & Davis, B. G. Synthetic Nucleosomes Reveal That GlcNAcylation Modulates Direct Interaction with the FACT Complex. *Angewandte Chemie International Edition* **55,** 8918–8922 (2016).

781. Fujiki, R. *et al.* GlcNAcylation of Histone H2B Facilitates Its Monoubiquitination. *Nature* **480,** 557 (2011).

782. Dehennaut, V., Leprince, D. & Lefebvre, T. O-GlcNAcylation, an Epigenetic Mark. Focus on the Histone Code, TET Family Proteins, and Polycomb Group Proteins. *Frontiers in Endocrinology* **5** (2014).

783. Leturcq, M., Lefebvre, T. & Vercoutter-Edouart, A.-S. O-GlcNAcylation and Chromatin Remodeling in Mammals: An up-to-Date Overview. *Biochemical Society Transactions* **45,** 323–338 (2017).

784. Sakabe, K., Wang, Z. & Hart, G. W. β-N-Acetylglucosamine (O-GlcNAc) Is Part of the Histone Code. *Proceedings of the National Academy of Sciences* **107,** 19915–19920 (2010).

785. Zhao, L., Shah, J., Cai, Y., Jin, J., Zhao, L., Shah, J. A., Cai, Y. & Jin, J. 'O-GlcNAc Code' Mediated Biological Functions of Downstream Proteins. *Molecules* **23,** 1967 (2018).

786. Hirosawa, M., Hayakawa, K., Yoneda, C., Arai, D., Shiota, H., Suzuki, T., Tanaka, S., Dohmae, N. & Shiota, K. Novel O -GlcNAcylation on Ser 40 of Canonical H2A Isoforms Specific to Viviparity. *Scientific Reports* **6,** 31785 (1 2016).

787. Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. DNA Double-Stranded Breaks Induce Histone H2AX Phosphorylation on Serine 139. *Journal of Biological Chemistry* **273,** 5858–5868 (1998).

788. Bungard, D., Fuerth, B. J., Zeng, P.-Y., Faubert, B., Maas, N. L., Viollet, B., Carling, D., Thompson, C. B., Jones, R. G. & Berger, S. L. Signaling Kinase AMPK Activates Stress-Promoted Transcription via Histone H2B Phosphorylation. *Science* **329,** 1201–1205 (2010).

789. Xu, Q. *et al.* AMPK Regulates Histone H2B O-GlcNAcylation. *Nucleic Acids Research* **42,** 5594–5604 (2014).

790. Fong, J. J., Nguyen, B. L., Bridger, R., Medrano, E. E., Wells, L., Pan, S. & Sifers, R. N. β-N-Acetylglucosamine (O-GlcNAc) Is a Novel Regulator of Mitosis-Specific Phosphorylations on Histone H3. *Journal of Biological Chemistry* **287,** 12195–12203 (2012).

791. Carmena, M. & Earnshaw, W. C. The Cellular Geography of Aurora Kinases. *Nature Reviews Molecular Cell Biology* **4,** 842–854 (11 2003).

792. Love, D. C., Ghosh, S., Mondoux, M. A., Fukushige, T., Wang, P., Wilson, M. A., Iser, W. B., Wolkow, C. A., Krause, M. W. & Hanover, J. A. Dynamic O-GlcNAc Cycling at Promoters of Caenorhabditis Elegans Genes Regulating Longevity, Stress, and Immunity. *Proceedings of the National Academy of Sciences* **107,** 7413–7418 (2010).

793. Ranuncolo, S. M., Ghosh, S., Hanover, J. A., Hart, G. W. & Lewis, B. A. Evidence of the Involvement of O-GlcNAc-Modified Human RNA Polymerase II CTD in Transcription in Vitro and in Vivo. *Journal of Biological Chemistry* **287,** 23549–23561 (2012).

794. Lewis, B. A. & Hanover, J. A. O-GlcNAc and the Epigenetic Regulation of Gene Expression. *Journal of Biological Chemistry* **289,** 34440–34448 (2014).

795. Wojcik, F., Dann, G. P., Beh, L. Y., Debelouchina, G. T., Hofmann, R. & Muir, T. W. Functional Crosstalk between Histone H2B Ubiquitylation and H2A Modifications and Variants. *Nature Communications* **9,** 1394 (1 2018).

796. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution††We Dedicate This Paper to the Memory of Max Perutz Who Was Particularly Inspirational and Supportive to T.J.R. in the Early Stages of This Study. *Journal of Molecular Biology* **319,** 1097–1113 (2002).

797. Flaus, A. Principles and Practice of Nucleosome Positioning in Vitro. *Frontiers in Life Science* **5,** 5–27 (2011).

798. Raj, R. *Role of Posttranslational Modifications of Histone Proteins in Epigenetics* (University of Oxford, 2016).

799. Morrison, P. M., Foley, P. J., Warriner, S. L. & Webb, M. E. Chemical Generation and Modification of Peptides Containing Multiple Dehydroalanines. *Chemical Communications* **51,** 13470–13473 (2015).

800. Weimar, T., Kreis, U. C., Andrews, J. S. & Pinto, B. M. Conformational Analysis of Maltoside Heteroanalogues Using High-Quality NOE Data and Molecular Mechanics Calculations. Flexibility as a Function of the Interglycosidic Chalcogen Atom. *Carbohydrate Research* **315,** 222–233 (1999).

801. Chooi, K. P., Galan, S. R. G., Raj, R., McCullagh, J., Mohammed, S., Jones, L. H. & Davis, B. G. Synthetic Phosphorylation of P38α Recapitulates Protein Kinase Activity. *Journal of the American Chemical Society* **136,** 1698–1701 (2014).

802. Śmiechowski, M. Theoretical pKa Prediction of O-Phosphoserine in Aqueous Solution. *Chemical Physics Letters* **501,** 123–129 (2010).

803. Park, H.-S., Hohn, M. J., Umehara, T., Guo, L.-T., Osborne, E. M., Benner, J., Noren, C. J., Rinehart, J. & Söll, D. Expanding the Genetic Code of Escherichia Coli with Phosphoserine. *Science* **333,** 1151–1154 (2011).

804. Lee, S., Oh, S., Yang, A., Kim, J., Söll, D., Lee, D. & Park, H.-S. A Facile Strategy for Selective Incorporation of Phosphoserine into Histones. *Angewandte Chemie International Edition* **52,** 5771–5775 (2013).

805. Oza, J. P., Aerni, H. R., Pirman, N. L., Barber, K. W., ter Haar, C. M., Rogulina, S., Amrofell, M. B., Isaacs, F. J., Rinehart, J. & Jewett, M. C. Robust Production of Recombinant Phosphoproteins Using Cell-Free Protein Synthesis. *Nature Communications* **6,** 8168 (1 2015).

806. Rogerson, D. T. *et al.* Efficient Genetic Encoding of Phosphoserine and Its Nonhydrolyzable Analog. *Nature Chemical Biology* **11,** 496–503 (7 2015).

807. Pompetzki, M. & Jansen, M. Natriummonothiophosphat(V): Kristallstruktur und Natriumionenleitfähigkeit. *Zeitschrift für anorganische und allgemeine Chemie* **628,** 641–646 (2002).

808. Machida, M. *et al.* Purification of Phosphoproteins by Immobilized Metal Affinity Chromatography and Its Application to Phosphoproteome Analysis. *The FEBS Journal* **274,** 1576–1587 (2007).

809. Tsai, C.-F., Hsu, C.-C., Hung, J.-N., Wang, Y.-T., Choong, W.-K., Zeng, M.-Y., Lin, P.-Y., Hong, R.-W., Sung, T.-Y. & Chen, Y.-J. Sequential Phosphoproteomic Enrichment through Complementary Metal-Directed Immobilized Metal Ion Affinity Chromatography. *Analytical Chemistry* **86,** 685–693 (2014).

810. Biswas, S., Sarkar, A. & Misra, R. Iron Affinity Gel and Gallium Immobilized Metal Affinity Chromatographic Technique for Phosphopeptide Enrichment: A Comparative Study. *Biotechnology & Biotechnological Equipment* **31,** 639–646 (2017).

811. Zheng, W., Schwarzer, D., LeBeau, A., Weller, J. L., Klein, D. C. & Cole, P. A. Cellular Stability of Serotonin N-Acetyltransferase Conferred by Phosphonodifluoromethylene Alanine (Pfa) Substitution for Ser-205. *Journal of Biological Chemistry* **280,** 10462–10467 (2005).

812. Burton, D. J., Yang, Z.-Y. & Qiu, W. Fluorinated Ylides and Related Compounds. *Chemical Reviews* **96,** 1641–1716 (1996).

813. Kempkes, L. J. M., Martens, J., Berden, G. & Oomens, J. W-Type Ions Formed by Electron Transfer Dissociation of Cys-Containing Peptides Investigated by Infrared Ion Spectroscopy. *Journal of Mass Spectrometry* **53,** 1207–1213 (2018).

814. Dyer, P. N., Edayathumangalam, R. S., White, C. L., Bao, Y., Chakravarthy, S., Muthurajan, U. M. & Luger, K. in *Methods in Enzymology* Supplement C vols., 23–44 (Academic Press, 2003).

815. Muthurajan, U., Mattiroli, F., Bergeron, S., Zhou, K., Gu, Y., Chakravarthy, S., Dyer, P., Irving, T. & Luger, K. in *Methods in Enzymology* (ed Marmorstein, R.) 3–41 (Academic Press, 2016).

816. Arimura, Y., Tachiwana, H., Oda, T., Sato, M. & Kurumizaka, H. Structural Analysis of the Hexasome, Lacking One Histone H2A/H2B Dimer from the Conventional Nucleosome. *Biochemistry* **51,** 3302–3309 (2012).

817. Rychkov, G. N., Ilatovskiy, A. V., Nazarov, I. B., Shvetsov, A. V., Lebedev, D. V., Konev, A. Y., Isaev-Ivanov, V. V. & Onufriev, A. V. Partially Assembled Nucleosome Structures at Atomic Detail. *Biophysical Journal* **112,** 460–472 (2017).

818. Garay, P. M., Wallner, M. A. & Iwase, S. Yin–Yang Actions of Histone Methylation Regulatory Complexes in the Brain. *Epigenomics* **8,** 1689–1708 (2016).

819. Toleman, C. A., Schumacher, M. A., Yu, S.-H., Zeng, W., Cox, N. J., Smith, T. J., Soderblom, E. J., Wands, A. M., Kohler, J. J. & Boyce, M. Structural Basis of O-GlcNAc Recognition by Mammalian 14-3-3 Proteins. *Proceedings of the National Academy of Sciences* **115,** 5956–5961 (2018).

820. Darvekar, S., Johnsen, S. S., Eriksen, A. B., Johansen, T. & Sjøttem, E. Identification of Two Independent Nucleosome-Binding Domains in the Transcriptional Co-Activator SPBP. *Biochemical Journal* **442,** 65–75 (2012).

821. Sanchez, R. & Zhou, M.-M. The PHD Finger: A Versatile Epigenome Reader. *Trends in Biochemical Sciences* **36,** 364–372 (2011).

822. Darvekar, S., Rekdal, C., Johansen, T. & Sjøttem, E. A Phylogenetic Study of SPBP and RAI1: Evolutionary Conservation of Chromatin Binding Modules. *PLOS ONE* **8,** e78907 (2013).

823. Rekdal, C., Sjøttem, E. & Johansen, T. The Nuclear Factor SPBP Contains Different Functional Domains and Stimulates the Activity of Various Transcriptional Activators. *Journal of Biological Chemistry* **275,** 40288–40300 (2000).

824. Lyngsø, C., Bouteiller, G., Damgaard, C. K., Ryom, D., Sanchez-Muñoz, S., Nørby, P. L., Bonven, B. J. & Jørgensen, P. Interaction between the Transcription Factor SPBP and the Positive Cofactor RNF4 AN INTERPLAY BETWEEN PROTEIN BINDING ZINC FINGERS. *Journal of Biological Chemistry* **275,** 26144–26149 (2000).

825. Sjøttem, E., Rekdal, C., Svineng, G., Johnsen, S. S., Klenow, H., Uglehus, R. D. & Johansen, T. The ePHD Protein SPBP Interacts with TopBP1 and Together They Co-Operate to Stimulate Ets1-Mediated Transcription. *Nucleic Acids Research* **35,** 6648–6662 (2007).

826. Gburcik, V., Bot, N., Maggiolini, M. & Picard, D. SPBP Is a Phosphoserine-Specific Repressor of Estrogen Receptor α. *Molecular and Cellular Biology* **25,** 3421–3430 (2005).

827. Elvenes, J., Thomassen, E. I. S., Johnsen, S. S., Kaino, K., Sjøttem, E. & Johansen, T. Pax6 Represses Androgen Receptor-Mediated Transactivation by Inhibiting Recruitment of the Co-activator SPBP. *PLOS ONE* **6,** e24659 (2011).

828. Darvekar, S. R., Elvenes, J., Brenne, H. B., Johansen, T. & Sjøttem, E. SPBP Is a Sulforaphane Induced Transcriptional Coactivator of NRF2 Regulating Expression of the Autophagy Receptor P62/SQSTM1. *PLOS ONE* **9,** e85262 (2014).

829. Solaimani, P., Damoiseaux, R. & Hankinson, O. Genome-Wide RNAi High-Throughput Screen Identifies Proteins Necessary for the AHR-Dependent Induction of CYP1A1 by 2,3,7,8-Tetrachloro\-Dibenzo-p-Dioxin. *Toxicological Sciences* **136,** 107–119 (2013).

830. Torti, E. *et al.* Variants in TCF20 in Neurodevelopmental Disability: Description of 27 New Patients and Review of Literature. *Genetics in Medicine* **21,** 2036–2042 (9 2019).

831. Babbs, C. *et al.* De Novo and Rare Inherited Mutations Implicate the Transcriptional Coregulator TCF20/SPBP in Autism Spectrum Disorder. *Journal of Medical Genetics* **51,** 737–747 (2014).

832. Vetrini, F. *et al.* De Novo and Inherited TCF20 Pathogenic Variants Are Associated with Intellectual Disability, Dysmorphic Features, Hypotonia, and Neurological Impairments with Similarities to Smith–Magenis Syndrome. *Genome Medicine* **11,** 12 (2019).

833. Poisson, A., Nicolas, A., Bousquet, I., Raverot, V., Gronfier, C. & Demily, C. Smith-Magenis Syndrome: Molecular Basis of a Genetic-Driven Melatonin Circadian Secretion Disorder. *International Journal of Molecular Sciences* **20** (2019).

834. Mullegama, S. V., Alaimo, J. T., Fountain, M. D., Burns, B., Balog, A. H., Chen, L. & Elsea, S. H. RAI1 Overexpression Promotes Altered Circadian Gene Expression and Dyssomnia in Potocki–Lupski Syndrome. *Journal of Pediatric Genetics* **06,** 155–164 (2017).

835. Shema-Yaacoby, E. *et al.* Systematic Identification of Proteins Binding to Chromatin-Embedded Ubiquitylated H2B Reveals Recruitment of SWI/SNF to Regulate Transcription. *Cell Reports* **4,** 601–608 (2013).

836. Eberl, H. C., Spruijt, C. G., Kelstrup, C. D., Vermeulen, M. & Mann, M. A Map of General and Specialized Chromatin Readers in Mouse Tissues Generated by Label-Free Interaction Proteomics. *Molecular Cell* **49,** 368–378 (2013).

837. Zhu, Y., Shan, X., Yuzwa, S. A. & Vocadlo, D. J. The Emerging Link between O-GlcNAc and Alzheimer Disease. *Journal of Biological Chemistry* **289,** 34472–34481 (2014).

838. Pravata, V. M. *et al.* An Intellectual Disability Syndrome with Single-Nucleotide Variants in O-GlcNAc Transferase. *European Journal of Human Genetics* **28,** 706–714 (6 2020).

839. Roa-Mansergas, X., Fadó, R., Atari, M., Mir, J. F., Muley, H., Serra, D. & Casals, N. CPT1C Promotes Human Mesenchymal Stem Cells Survival under Glucose Deprivation through the Modulation of Autophagy. *Scientific Reports* **8,** 6997 (1 2018).

840. Tridot, G. & Tudo, J. Sodium and Ammonium Monothioorthophosphates. *Bulletin de la Societe Chimique de France,* 1231–1240 (1960).

841. Na3VO4 Stock Solution. *Cold Spring Harbor Protocols* **2017,** pdb.rec092940 (2017).

First Goldilocks tasted the porridge of the Great Big Bear, and that was too hot for her. Next she tasted the porridge of the Middle-sized Bear, but that was too cold for her. And then she went to the porridge of the Little Wee Bear, and tasted it, and that was neither too hot nor too cold, but just right, and she liked it so well that she ate it all up, every bit!

— The Story of the Three Bears, as in *English Fairy Tales* 1918

5

Site-specific proximity-enabled protein crosslinking allows identification of protein-protein interaction site

Contents

5.1. Introduction: Chemical crosslinking using unnatural amino acids . . . 399
5.2. Results . . . 404
5.2.1. Generation of synthetic alkylator protein eH3 containing Bromohomonorleucine unnatural amino acid . . . 404
5.2.2. Protein-protein crosslink with the unnatural amino acid bromohomonorleucine . . . 404
5.2.3. Identified crosslinks show nucleophilicty of cysteines at Zn-binding site . . . 407
5.2.3.1. Crosslinking tandem mass spectrometry for low abundance xLink 407
5.2.3.2. Parameters for xLink MS/MS data analysis . . . 408
5.2.3.3. Identification of cysteines at Zn-binding site . . . 410
5.2.3.4. Identification of inter-histone crosslinking sites show crosslinking reactivity beyond cysteine . . . 412
5.3. Conclusion & outlook . . . 414
5.4. Author contributions . . . 417
5.5. Experimental . . . 418
5.5.1. General Experimental Procedures . . . 418
5.5.2. Histone protein expression, chemical modification and crosslinking . . . 419
5.5.3. Tandem mass spectrometry . . . 429
5.6. References . . . 433
5.7. Appendix: MS/MS spectra . . . 436

5.1. Introduction: Chemical crosslinking using unnatural amino acids

Protein-protein crosslinking has been a valuable tool for structural and functional studies of protein-protein interactions (PPI), central to many biological processes.[385] The formation of a covalent bond between and within proteins allows capture of weak or transient interactions.[842] These interactions are important for the dynamic assembly of macromolecular complexes,[842] yet difficult to identify via traditionally used methods for the study of protein-protein interactions such as immunoprecipitation, pull-down-assays, yeast-two-hybrid approaches and Förster resonance energy transfer (FRET)-based assays.[843–845] Protein-protein crosslinking augments the information obtained in such assays by a structural component, as the respective sites of interactions on both proteins are linked together. Crosslinking can be performed *in vitro*, *in situ* and *in vivo* depending on a trade-off between the need for physiological conditions and the required the structural resolution. A general introduction to crosslinking tandem mass spectrometry is given in 1.3.5.5.

Several methods have been used for protein-protein crosslinking, such as (i) unspecific crosslinking with formaldehyde or (ii) crosslinking of lysine/cysteine residues using bifunctional small molecule crosslinkers such as disuccinimidyl suberate (DSS) or bis(sulfosuccinimidyl)suberate (BS3) as well as their MS-cleavable analogues such as disuccinimidyl sulfoxide (DSSO) and disuccinimidyl dibutyric urea (DSBU).[385,459] (iii) Small molecule photocrosslinkers (particularly heterobifunctional such as Sulfo-SBED) have proven their value in targeted experiments with known bait − unknown prey to find interaction partners[388], but suffer under high unspecific capture/background (detailed in 1.3.5.5).[483,846] More recently, site-specific crosslinking approaches have emerged that transform the bait protein itself into a crosslinking reagent, using (iv) unnatural amino acids (Uaa) with photoreactive moieties or (v) biocompatible chemical reactivity for proximity-enabled protein-protein crosslinking (PEPC).[847] Protein crosslinking utilising unnatural amino acids takes advantage of their site-specific incorporation to capture interactions occurring precisely at the site of interest − the leitmotiv of this chapter.

Unnatural amino acids used previously for proximity-enabled protein-protein crosslinking (PEPC) have to date exclusively been incorporated using genetic code expansion (see Figure 5.1 for a comprehensive account), the required reactive functionalities being largely incompatible with procedures used in peptide synthesis, protein semi-synthesis

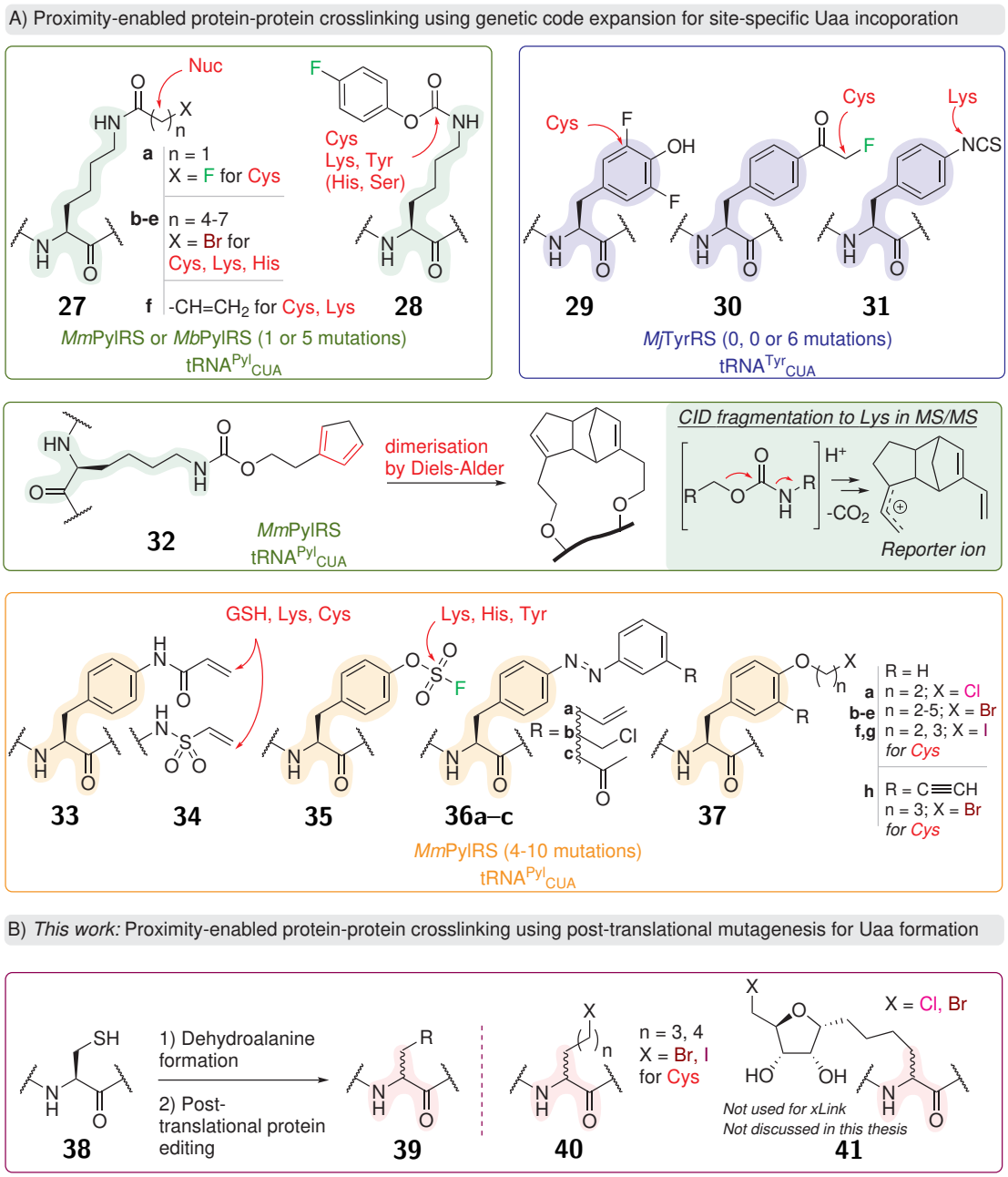


Figure 5.1. Proximity-enabled protein-protein crosslinking (PEPC) using unnatural amino acids. A. Previous examples from the literature that used genetic code expansion techniques as a mean to site-specifically incorporate unnatural amino acids. Two main systems are used, the Pyrrolysyl *Methanosarcina mazei* RNA synthetase (PylRS) or *Methanosarcina barkeri* PylRS together with the $tRNA^{Pyl}_{CUA}$, and the Tyrosyl *Methanocaldococcus jannaschi* TyrRS with the $tRNA^{Tyr}_{CUA}$, the conserved chemical core of the unnatural amino acids is highlighted in green or blue, respectively. Chemical moieties incorporated by genetic code expansion have been reviewed elsewhere[732] as have Uaa's for photo-crosslinking[847], this figure focusses on the incorporations of Uaa's used for PEPC.

Figure 5.1. (***continued***) **B.** Proximity-enabled protein-protein crosslinking by post-translational mutagenesis as introduced in this work. A recombinantly expressed protein with Cysteine **38** at its site of interest is converted in a series of 2 consecutive reactions to the unnatural amino acid **39**. Using this chemistry, Uaa's **40a**–**d** and **41** could be generated, with **40b** used in this study to demonstrate the applicability in PEPC.

and post-translational chemical mutagenesis. Examples include Uaa's with reactive alkyl halides **27a**, **27b**–**e**[848], **27c**[848,849], **27f**[850], aryl carbamate **28**[851], aryl halides **29**[852] and the α-carbonyl halide **30**[853], and aryl isothiocyanate **31**[854]. Common to all these previous amino acids is their reactivity in proximity-enabled nucleophilic substitutions. Noteworthy, a cyclopentadiene derivative of Lysine **32**[855] was used for fast and selective normal electron-demand Diels–Alder (DA) reaction using the MmPylRS sytem without further mutations. As a consequence, the incorporation of unnatural amino acids for PEPC has been mainly limited by their recognition through the orthogonal aminoacyl tRNA synthetases/tRNA pair used − thus limiting the currently available chemical space to derivates of pyrrolysine and tyrosine.

The field has long aimed to incorporate more chemical reactivity using Uaa incorporation by genetic code expansion. The PylRS system could be adapted by using a higher amount of mutations (up to 10) to incorporate Tyrosyl derivates (highlighted in orange) bypassing limitations of the MjTyrRS in its structural modification and bioorthogonality. This allowed for acrylamide **33**[850] and vinylsulfonamide **34**[850], aryl fluorosulfate **35**[856] and the photoswitchable Uaa **36a**–**c**[857] to be incorporated. Probably the most well-studied and widely used Uaa's for PEPC are the haloalkyl-bearing tyrosine derivates **37a**–**g**[858] and their bromoalkyl derivate **37h**[859] that contains an alkyne moiety for enrichment by Copper(I)-catalyzed Azide-Alkyne Cycloaddition (CuAAC) to overcome limits in detection by MS/MS. However, despite reaching very high sophistication with the genetic code expansion current systems, the limitations to Pyl and Tyr derivates remain: side chains are extended well beyond their natural length resulting in linkers with long distances (i.e. 8-13,[858] 11-14[848] or 10[859] atoms in length) between protein backbone and alkylation site. As such these cannot accurately mimic native shapes / residues, which carries strong potential for artefacts in the context of protein-protein interactions. In conclusion, there is a need for new technologies in the incorporation of unnatural amino acids amenable for the proximity-enabled protein crosslinking.

With the recently developed benign photoredox protein modification chemistry by the Davis group,[860] reactivities amenable for proximity-enabled protein crosslinking, i.e. alkylhalides such as **40** and **41** (Figure 5.1.B)) have successfully been incorporated in a range of model proteins using a radical $1e^-$ mechanism (Figure 5.2).[860] These

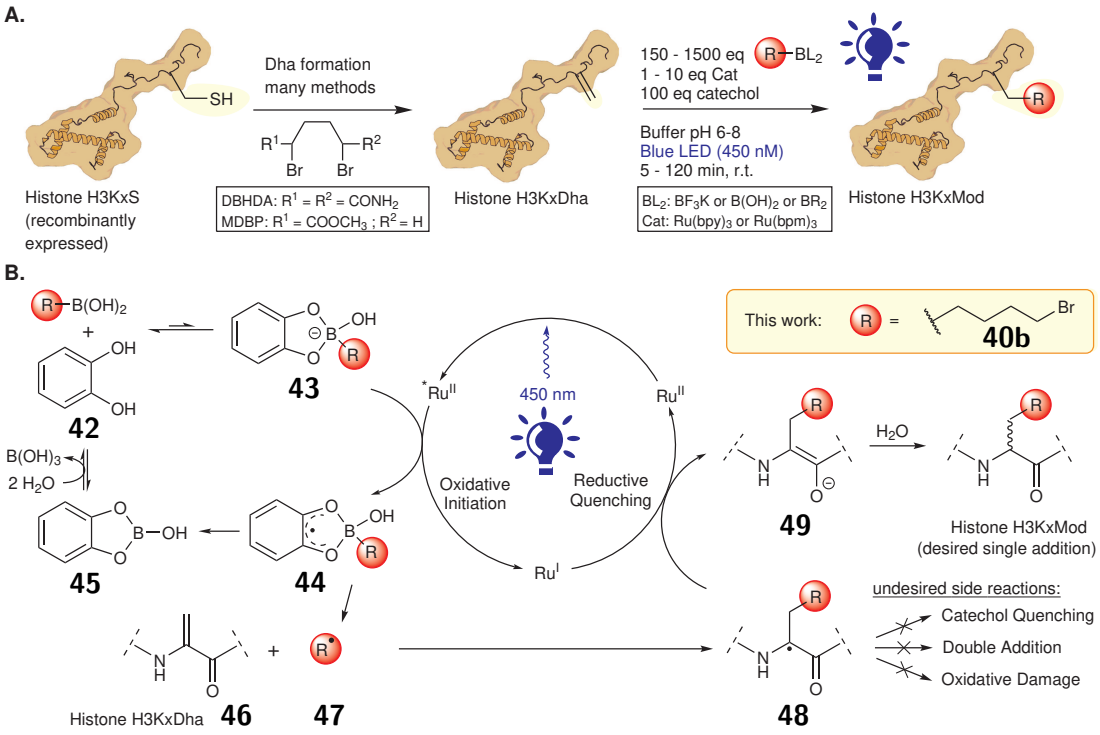


Figure 5.2. Site-specific photoredox catalysis for post-translational protein mutagenesis. A. General scheme of post-translational protein editing using a two-step reaction sequence. First, the recombinantly expressed protein, here histone H3, with Cysteine at the site of interest is converted to converted to dehydroalanine using bis-alkylation-elimination reagents. In a second step, dehydroalanine is edited using the recently developed photoredox protein chemistry by the Davis group.[860] **B.** Proposed mechanism for site-specific photoredox catalysis on proteins as suggested in Josephson *et al.*[860] The boronic acid reagent associates with catechol **42** to form the boronic acid catechol ester **43** which accepts an electron from excited $^*Ru^{II}$ to form **44**. Catechol **42** is re-generated via **45** and radical reagent **47** is generated. **47** subsequently adds to dehydroalaine **46** to give Cα radical **48** reductively quenched by Ru^{I} to enol **49**, which tautomerises to the desired unnatural amino acid.

haloalkane moieties have the advantage of being overall unreactive towards free amino acids or other biological components under physiological conditions, but when placed and kept structurally in proximity of their specific interaction partners the reaction of the Uaa with the target residue to form the desired covalent bond is facilitated. In simple terms, the electrophilic reactivity of haloalkane amino acids falls into a kinetic 'sweet spot': not too reactive to avoid unspecific reactions but fast enough to trap the specific interaction partners, a concept often referred to as 'Goldilocks' kinetics in allusion to the children's fairytale. More precisely, the enhanced reactivity (or accelerated kinetics) observed in protein complex can be explained using the concept of *kinetic effective molarity*,* that

*The *kinetic* effective molarity is the change of reaction rate as a consequence of enforced proximity, the *thermodynamic* effective molarity the change in equilibrium.[861]

is the rate acceleration of a chemical reaction as a unimolecular reaction (i.e. in a protein complex) versus as a bimolecular reaction (i.e. separated side chain reactants in solution).[862,863] It is indicated as a theoretical reagent concentration at which the observed (pseudo-first order) rate matches that of the unimolecular process.[861] Several factors explain the observable increase in effective molarity, including the increased dwell time of the reactants (i.e. k_{off} decreased through affinity / tethering),[864] appropriate linker length[863] as well as the energetical 'guiding' dictated by the supramolecular architecture and geometric complementarity.[865] The use of haloalkanes for protein-protein crosslinking has been previously shown using the Uaa's **27a** and **27b**–**e**[848,849] that covalently capture cysteines in what is thought to be a S_N2 reaction. However, these approaches have been restricted by the substrate limitations of genetic code expansion (Lys and Tyr mimics), therefore resulting in long distances between protein backbone and alkylation site. This chapter presents the Uaa's with the shortest linkers for proximity-enabled protein crosslinking to date − note that this comes at the cost of losing the stereocenter at the Cα.

This previously inaccessible Uaa chemotypes such as short bromoalkyl side chains that became available using the recently discovered photoredox protein editing chemistry (Figure 5.1.B) called for an evaluation of their potential in crosslinking experiments. Whereas such crosslinking experiments are usually evaluated with test systems such as the Z protein together with the Z_{SPA} affibody (Afb)[851,854,858,859], Fab fragments[850,851], or used in the context of protein stapling[849] or complex stabilisation for protein crystallography[848], the choice in this study was to assess its performance in a more stringent, space-limited and transient (substrate • enzyme) PPI system. Therefore, we decided to investigate the interaction of histone H3, a protein successfully used for post-translational modification with this photoredox chemistry and the histone demethylase 4A (KDM4A, also known as JMJD2A of JHDM3A), a known H3 interaction partner with a well-characterised binding site where the H3 tail enters a deep cavity on KDM4A.[866,867] Ultimately, this would allow to emphasise the applicability of proximity-enabled protein protein crosslinking in the context of studying a defined biological interaction.

This work was developed in collaboration with Dr. Charlie Fehl and Brian Josephson, who developed the chemistry and performed the crosslinking reaction, respectively. My contribution was the substrate characterisation by MS/MS, the development of a MS/MS method able to identify site-specifically crosslinked peptides and the structural analysis.

5.2. Results

5.2.1. Generation of synthetic alkylator protein eH3 containing Bromohomonorleucine unnatural amino acid

The epitope-tagged human histones eH3.1 were expressed with Lys → Cys mutations at the sites K4, K9 and K27 by Brian Josephson. Of interest, H3K4 trimethylation is a hallmark of active promoters and is normally mutually exclusive with H3K9 trimethylation, which is a mark of inactive chromatin.[868] H3K27me3 (trimethylation) is linked to transcriptional repression, both H3K27me3 and H3K4me3 determine whether a cell will remain unspecified or will differentiate eventually (see Figure 4.17.B).[869] KDM4A specifically demethylates H3K9me3 and H3K36me3, but does not act on histone H3K4 or H3K27. It neither has activity on mono- (me1) and dimethylated (me2) residues.[870] The constrained nature of the binding site of KDM4A[867] highlights the need for a precise reactive probe with the correct length of the side chain.

The cysteine mutants were converted to dehydroalanine (Dha) using the well-established bis-alkylation-elimination reagent DBHDA.[755] Dha was subsequently derivatised to the Bromohomonorleucine (Bhn) amino acid with the photoredox chemistry using bromobutyl boronic acid and $Ru(bpm)_3Cl_2$ as catalyst.[860] These protein modification reactions were performed by Charlie Fehl and Brian Josephson. The formation of Bhn was confirmed by intact protein LC-MS and bottom-up LC-MS/MS (Figure 5.3). As expected, the bromine showed the typical isotopic pattern in the MS^1 spectrum corresponding to the isotopes of ^{79}Br and ^{81}Br, which could be isolated individually and fragmented in the MS^2.

5.2.2. Protein-protein crosslink with the unnatural amino acid bromohomonorleucine

The resulting synthetically Bhn-modified histone proteins were incubated with their interaction partner KDM4A for two time ranges (15 min and 2h) and at two distinct temperatures (4 °C and 37 °C) to assess the crosslinking efficiency (Figure 5.4.A-C). A high molecular weight band corresponding to the mass of the crosslinked protein complex was observed, it is formed in comparable quantities with all three modification sites, providing evidence for the capacity of Bhn in crosslinking experiments. Noteworthy, the band appears rather weak as compared to the free histone and KDM4A bands. This could be explained by the rather low K_D for this complex and the presumably slow

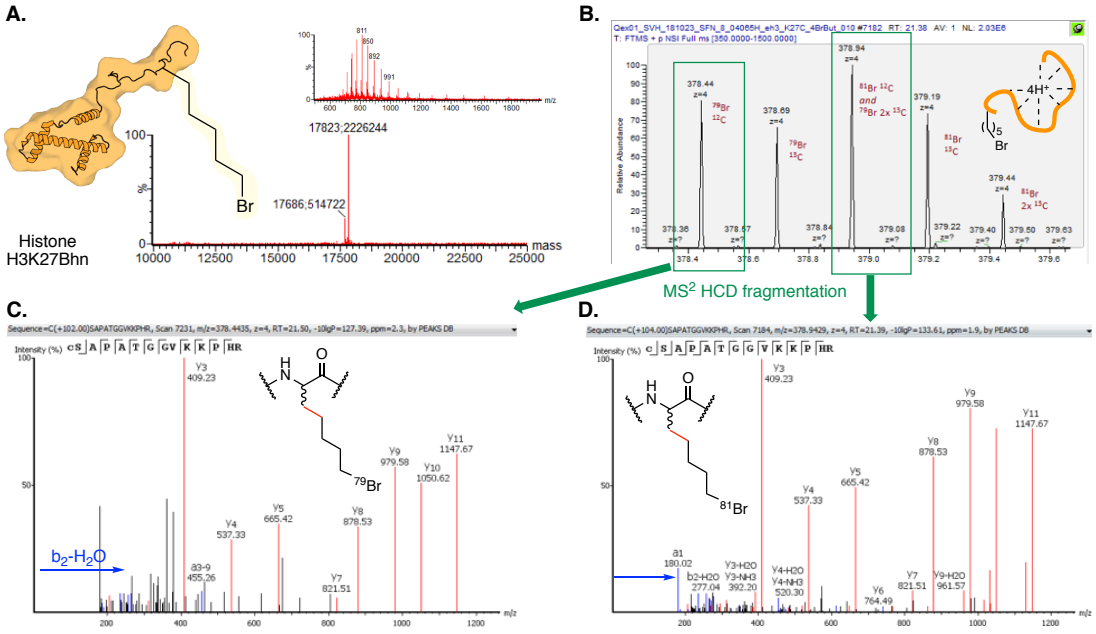


Figure 5.3. eH3 K27Bhn characterisation shows characteristic isotope pattern for brominated peptides. A. Intact protein LC-ESI-MS of H3K27Bhn confirming the expected mass (Calc: 17822, Obs: 17823). The 17686 peak corresponds to residual eH3 K27Dha. Intact Protein MS data kindly provided by Brian Josephson. **B.** MS^1 spectrum of the brominated peptide BhnSAPTGGVKKPHR (ArgC digest) shows the isotopic pattern corresponding to the isotopes of ^{79}Br, ^{81}Br, ^{12}C and ^{13}C as annotated. Both peaks framed in green were selected for MS^2 fragmentation in the top10 method (Q Exactive). **C.** MS^2 HCD spectrum of ^{79}Br and ^{12}C containing peptide confirming site of modification. **D.** MS^2 HCD spectrum of ^{81}Br and ^{12}C containing peptide confirming site of modification.

kinetics of the reaction compared to the k_{off} of the complex dissociation. Unfortunately, no intact protein ESI-MS could be obtained for the crosslinked adduct.

In contrast, no high molecular band for a crosslinked adduct is observed when incubating with H4, a nucleosomal binding partner lacking nucleophilic cysteine residues, or incubation with bovine serum albumin (BSA), a known cysteine-rich control. Notably, the H3/H4 protein-protein interaction interface in both H3-H4 dimer and tetramer does not involve Lys4, Lys9 and Lys27 whereas the H3•KDM4A PPI does.[867] The lack of reaction of BSA or H4 via non-specific Cys or non-specific binding, respectively, is consistent with previous observation in the literature that alkylhalide side chains provide reactive handles to access PPI-selective crosslinking[858].

The experiments described in the current section were performed by Brian Josephson (except for MS/MS) and are included in this chapter for comprehensiveness. Next, I set out to perform MS/MS analysis to identify the location of the crosslinking sites.

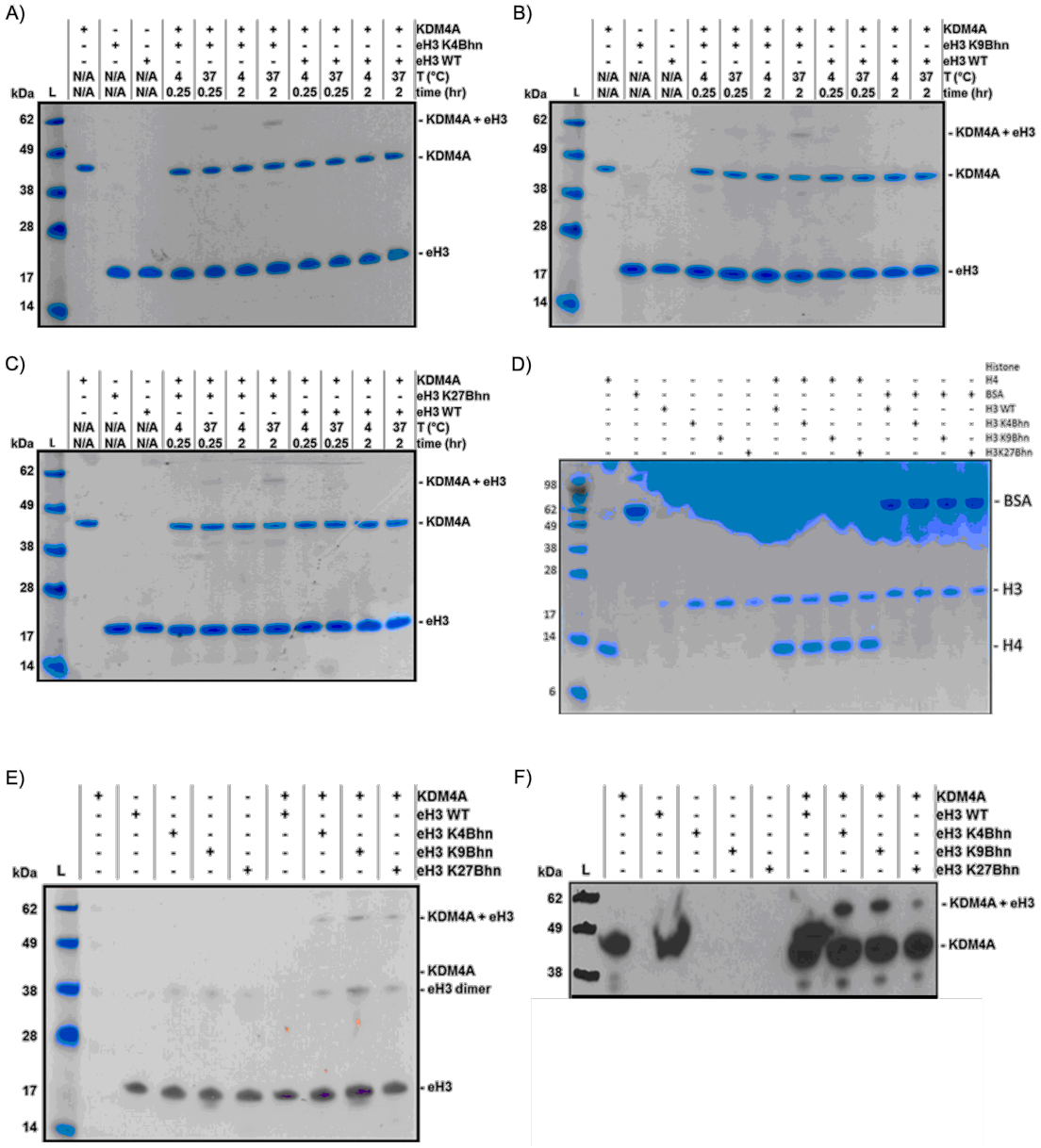


Figure 5.4. SDS-PAGE and Western Blot analysis of the crosslinking of human eH3-K4/9/27Bhn to KDM4A. Time course (15 min and 2h) and temperature dependence (4 °C and 37 °C) of Histone eH3 K4Bhn (**A.**), eH3 K9Bhn (**B.**) and eH3 K27Bhn (**C.**) crosslinking to KDM4A. SDS-PAGE shows a comparable increase in crosslinking at all sites as temperature and reaction time are increased. **D.** Control experiments attempting crosslinking between Bhn containing histones and either H4, a cysteine-free interaction partner of H3, or BSA, a cysteine-rich protein that is not an interaction partner of H3. No evidence of crosslinking events can be observed without the presence of both proximity and reactive nucleophilic residues, with 4 h incubation time at 37 °C. **E.** Western Blot with anti histone H3 antibody confirms presence of eH3 in the higher mass band, the corresponding SDS-PAGE is shown in Figure 5.5.C **F.** Western Blot with anti His_6-tag antibody confirms presence of KDM4A (with His-tag) in the higher mass band, the corresponding SDS-PAGE is shown in Figure 5.5.C. Gel & WB data kindly provided by Brian Josephson.

5.2.3. Identified crosslinks show nucleophilicty of cysteines at Zn-binding site

5.2.3.1. Crosslinking tandem mass spectrometry for low abundance xLink

Several challenges had to be overcome to obtain the desired crosslinked spectra. First, a suitable digestion protocol proved essential. Whereas we could obtain good peptide-spectrum matches (PSM) for eH3•KDM4A crosslinked peptides in a tryptic *in-solution* digest of the H3K9Bhn-KDM4A assay, this was neither the case for H3K4Bhn nor H3K27Bhn, the first yielding no identified crosslinked spectra and the latter mainly H3•H3 crosslinks of poor spectrum quality, with a single acceptable PSM for Cys-234 by means of its score and annotation. Next, a LysC in *in-solution* digestion protocol was tried, this did only yield identification of one H3•H3 PSM for K4 and no identification for K27. The ArgC *in-solution* digest was attempted for K4 but did not yield any PSM for crosslinked peptides.

As such, we hypothesised that as indicated by the poor intensity of the crosslinking band on the SDS-PAGE, the crosslinked peptides might have too low abundance to be (i) selected in top10 methods for MS^2 fragmentation or (ii) identified in the dataset. In order to reduce sample complexity, we switched to an *in-gel* digest, by cutting the respective high molecular weight band out of the SDS-PAGE, allowing purification of the crosslinked dimer and subsequent LysC digest. Unfortunately, the LysC *in gel* digests did not lead to any trustworthy crosslinked peptide PSM in any of the samples. Finally, a tryptic *in-gel* digest gave good eH3•KDM4A PSM for all modifications and proved the best sample preparation protocol (see Appendix II for full PSM list and annotated spectra).

Whereas in our case, the digestion optimisation was a trial-and-error of protocols and proteases, the approach can potentially be rationalised. Indeed, the LC-MS/MS method used for analysis of crosslinked samples differs in that it does not select 2^+ charged precursor ions for MS^2 fragmentation and runs a longer gradient in high organic elution phase. When working with Lys and Arg rich proteins such as histones, short peptides will be generated on the β-chain (histone). This implies if the α chain from KDM4A is short, the charge distribution will be of 3^+ ($2 \times$ R/K cleavage + 1 H^+ taken up by the gas-phase peptide horseshoe). As a consequence of unspecific cleavages, commonly observed with trypsin, will not be fragmented by the instrument even if abundant as they'd have a 2^+ charge state. Moreover, for pLink to identify peptide, it expects a minimal length of 6 amino acids unless otherwise stated. Whereas in our case observed

peptides correspond exactly to the cleavage length of trypsin around the modified K4 or K9 Uaa site, other sites on histones or basic proteins might need more cautious selection and fine-tuning for the MS/MS to succeed.

5.2.3.2. Parameters for xLink MS/MS data analysis

Data analysis in crosslinking tandem mass spectrometry is rendered difficult by three key obstacle well-known to the literature: the increased complexity of fragmentations, the n-square problem and the statistic evaluation of credibility/false discovery.

- Crosslinked peptides can fragment from either end of both strands connected at the branching point, leading to 4 directions of fragmentations with respective b^{1+}, b^{2+} and y^{1+}, y^{2+} ions with their neutral losses and the less common y^{3+}, a^{1+}, a^{2+}. Additionally, yb^{1+} and ya^{1+} cross-chain fragments are commonly observed (but not necessarily recognised by the software). Depending on linker chemistry, additional K-type fragments (α or β chain + linker + iminium ion, named $K^L\alpha/K^L\beta$) or L-type fragments (for cleavable linkers, $L\alpha$ and $L\beta$) are observed[386] (see Figure 1.16).

- The second problem refers to the quadratic expansion of search space as each of the n proteins in the database can potentially crosslink with n other proteins, leading to n^2 possibilities which have to be interrogated in exhaustive search strategies.[399] As such, including post-translational modifications or additional missed cleavages has a more dramatic effect on the expansion of the search space than in regular bottom-up proteomics database searches. This prevents extensive searching of the data and thus possible additional identifications can be missed.

- The third problem of statistical validation comes from the need to avoid false-positive peptide identifications. Commonly in bottom-up proteomics, this is dealt with by incorporating a false-discovery rate (FDR). Defining a FDR for crosslinked peptides is however much more challenging. This is further hampered in that the number of peptides that represent a particular interaction are non-stoichiometric, rendering quantifications based on spectral counts or AUC's particularly vulnerable. pLink 2 makes the choice of using a target-decoy approach (TDA) adapted for crosslinking datasets[406] (see Figure 1.17).

However, the situation in this project differs, crosslinking events are only allowed for the unnatural amino acids included in one very specific protein, thus avoiding the

n-square problem simply by experiment design. Additionally, our linker is presumably stable under HCD fragmentation avoiding K- and L-type fragmentations − that being said, yb^{1+} and ya^{1+} are expected, which are notoriously difficult to identify by crosslinking software and as a matter of fact not annotated/identified in the xLink MS/MS spectra that follow (pLink should labels these brown and purple). Although presumably unlikely, the possibility of $p_{(-34)}$ and $p_{(+32)}$ fragments for breakage of the C−S bond cannot be excluded.[871]

Previous xLink MS/MS studies in the literature on PEPC with Uaa's have used either Xi or Kodak, we chose to use pLink on the grounds of performance, precision and sensitivity as well as its thorough FDR treatment, with the possibility to incorporate unnatural/custom defined amino acids and linkers. Nonetheless, as this is a novel and undefined type of crosslinking technology, it is unclear which effect different MS/MS search parameters such as type of FDR, first search and fragment error tolerance have on the number and quality of the identifications. Therefore, a parameter evaluation was

Table 5.1. Parameter evaluation on the eH3K9-Bhn•KDM4A crosslink. In-solution digest, presence of free eH3 and KDM4A, pLink 2.3.5.

	Search parameters				Results			
	FDR	First search	fragment error	results filter	# Spectra	# Pep pairs	# xL sites	Notes
1	5 % specific	50 ppm	20 ppm	20 ppm	11	8	7	Score & Evalue poorly indicative, incorrect assignments
2	1 % specific	50 ppm	20 ppm	20 ppm	14	8	5	low score & Evalues poorly indicative, 1 incorrect assignment
3	1 % specific	20 ppm	20 ppm	10 ppm	14	8	5	scoring indicative of spectrum quality
4	1 % specific	10 ppm	10 ppm	10 ppm	14	8	5	scoring indicative of spectrum Quality
5	1 % global	50 ppm	20 ppm	20 ppm	14	8	5	low score & Evalues poorly indicative, 1 incorrect assignment
6	1 % global	20 ppm	20 ppm	10 ppm	16	9	6	scoring indicative of spectr. Quality, Evalue <e-3 is acceptable, <e6 good; 1 new, 2 lowest scoring 'incorrect' (ie 1 Peptide pair & 1 xL site)
7	1 % global	10 ppm	10 ppm	10 ppm	16	9	6	as above

carried out to check if data analysis with pLink is valid for the case of PEPC with our Uaa **40b** and to identify which parameters lead to the best possible identifications.

In order to maximise identifications, we evaluated the search parameters FDR, Precursor tolerance (First search), fragment mass error and results filter relating to the number of crosslinks identified and the quality of the assignment on the eH3K9Bhn-KDM4A *in solution* digest dataset (Table 5.1). The parameters 1% global FDR, 20 ppm precursor tolerance, 20 ppm fragment error, 10 ppm results filter yielded best results while taking actual instrument resolution into account. Using global FDR has further the advantage of being based on the target-decoy method with spectral/local FDR relating strongly to the scoring model.[872] These parameters were used for all subsequent xLink analysis in this chapter.

5.2.3.3. Identification of cysteines at Zn-binding site

MS/MS analysis of crosslinked gel bands revealed highly conserved crosslinking sites on KDM4A by all three Bhn-containing histones (Figure 5.5): two zinc binding cysteines, Cys234 and Cys306 were observed in the K4 and K9 modified histones, and only Cys234 for K27 (Figure 5.5.F-H). These identifications were supported by a high number of independent PSM (see Appendix II). These two cysteines are located in the Zn-binding domain of KDM4A at the PPI interface, corroborating the identifications by MS/MS (Figure 5.5.D-E). Noteworthy, no crosslink was observed to the third Cys308, being more remote (19.6 Å) and obstructed from the reactive site by the two closer Cys234 (19.0 Å) and Cys306 (16.6 Å), which react first. The crosslinking to the Zn binding site could be further confirmed by Zn ejection kinetics (Figure 5.5.I, data from Akane Kawamura group). Indeed, crosslinking to Cys234 or Cys306 leads to release of Zn^{2+}, which can be traced by fluorescence assays, showing an increase in fluorescence during the crosslinking reaction of H3K9Bhn with KDM4A (data from Akane Kawamura group).

Taken together, Bhn appears to be a representative mimic of Lys behaviour in the context of PEPC, by virtue of its near-direct (i.e. non-extended) structural analogy to Lys. The MS/MS analysis confirms that the same protein-protein interface is reached as would be by the native Lys residues, in spite of the confined nature of the H3•KDM4A PPI and thus a binding site topology challenging for crosslinking studies.

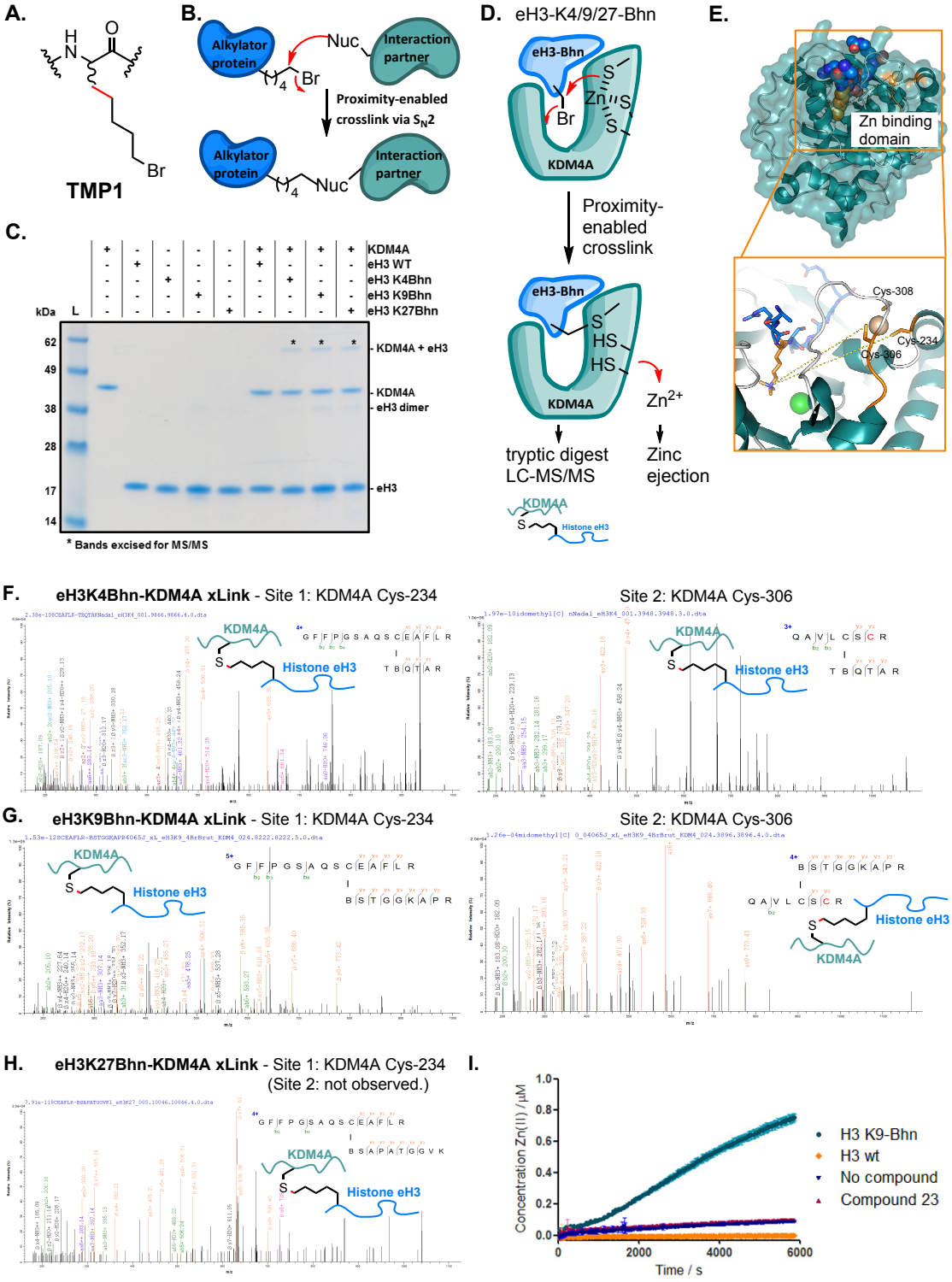


Figure 5.5. Proximity-enabled protein protein crosslinking using the unnatural amino acid Bhn 40b. A. Structure of the unnatural amino acid Bromohomonorleucine (Bhn) **40b**. **B.** Inclusion of proximity-enabled reactivities such as Bhn allows the formation of alkylating proteins for site-specific crosslinking with interaction partners containing reactive nucleophiles. **C.** SDS-PAGE for the crosslinking between KDM4A and histone

Figure 5.5. (***continued***) eH3 K4/K9/K27-Bhn including controls shows high molecular weight bands corresponding to the eH3•eH3 dimer and the eH3•KDM4A crosslink. Gel picture kindly provided by Brian Josephson. **D.** Suggested mechanism for the proximity-enabled crosslinking reaction between Bhn and the Cysteines of the Zn binding site, as identified by Zn ejection assays and LC-MS/MS. **E.** Structure of eH3 mimic peptide binding to KDM4A. Identified cysteines undergoing crosslink (Cys-306 and Cys-234) are highlighted in orange and their distance to H3K9 highlighted. **F.** Highlighted crosslinked peptides for the eH3K4 crosslink, *in-gel* digest. **G.** Highlighted peptides for the eH3K9 crosslink, *in-solution* digest. **H.** Highlighted peptides for the eH3K27 crosslink, *in-gel* digest. **I.** Zn(II) ejection upon incubation with KDM4A monitored by TSQ chelation. N=3. Data kindly provided by Akane Kawamura group, experimental details as described in[860].

5.2.3.4. Identification of inter-histone crosslinking sites show crosslinking reactivity beyond cysteine

Interestingly, the analysis of *in-solution* digest revealed additionally the presence of unspecific inter-histone crosslinking events, with different PSM identified for each K4/K9/K27Bhn modified histone. Transient association/dimerisation of histone H3 upon longer incubation times might cause a dwell time sufficient for occasional crosslink − note the band in Figure 5.5.C corresponding to the H3•H3 crosslink is much weaker than the band for the KDM4A crosslink, suggesting that the reactivity of Bhn follows the specificity or affinity of the PPI.

As H3 does not have any cysteine residues remaining, these transient crosslinking events give us insights into the reactivity of Bhn beyond cysteine. Indeed, crosslink of alkylhalides to reactivities such as Lys, His had been reported (Figure 5.1). Here we show PSM that indicate crosslink to Tyr and Ser/Thr in a Williamson-type (−C−O−C−) ether (Figure 5.6). Furthermore, no convincing loop-linked PSM could be found (terminology see Figure 1.16), that is a PSM that would exclude the possibility of HBr elimination in the acidic conditions of HCD fragmentation by having 'in loop' fragmentation, emphasising the need for specificity to allow the crosslink.

The observation of these Williamson-type in crosslinking experiments is to our knowledge wholly unprecedented.[873] The second order rate constant of this ether formation has generally been considered too low ($k_{app} < 10^{-4}$ $M.s^{-1}$)[874,875] to allow detectable cross-linking of protein-protein interactions[873] at low (nM-µM) protein concentrations. In consequence, this suggests an interaction between two eH3 proteins in presence of KDM4A, enabling the formation of a transient PPI that in turn augments the effective molarity allowing this reaction to occur. The formation of homo-multimers of KDM4A[876] via dimerisation of its JmjN domain[877] − in close structural proximity to its JmjC − might suggest an association *in vitro* that brings the eH3 tails in close prox-

imity, explaining the observed crosslinks. This result in turn highlights the importance of effective molarity as a driver for the crosslinking reaction, thus ensuring reactivity in the correct PPI context while avoiding unspecific reactions in solution.

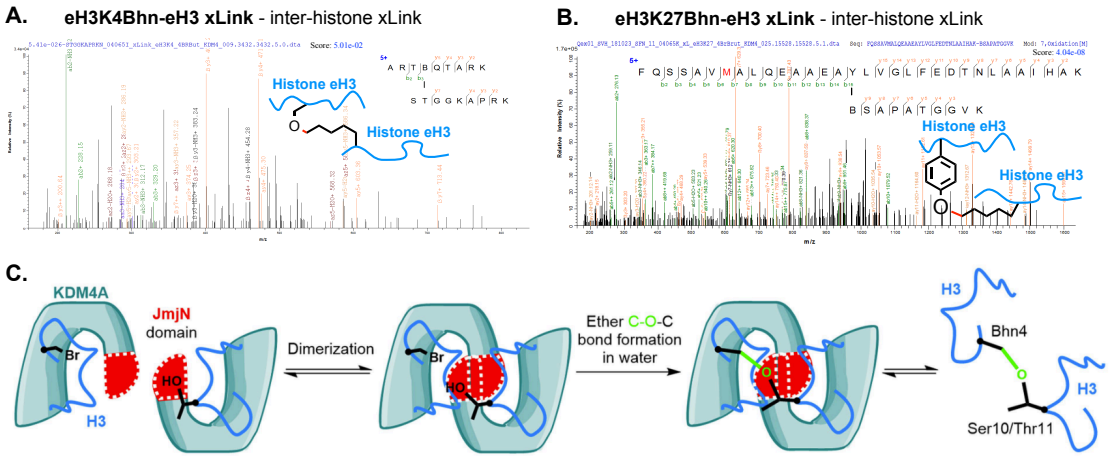


Figure 5.6. Inter-histone crosslinking peptide-spectrum matches. A. eH3•eH3 crosslink PSM of K4Bhn with Ser10 or Thr11, LysC *in solution* digest. **B.** eH3•eH3 crosslink PSM of H27Bhn with Tyr99, trypsin *in solution* digest. **C.** The formation of intermolecular eH3•eH3 Williamson-type ether bonds could be driven be effective molarity, through the *in vitro* dimerisation of KDM4A via its JmjN domain. Reproduced from[860] with permission from ©Springer Nature.

5.3. Conclusion & outlook

In this study, we presented the shortest unnatural amino acid to date for proximity-enabled protein-protein crosslinking (PEPC), bromohomonorleucine Bhn, being moreover the structurally closest Lys analogue to date for PEPC. This was allowed by the incorporation of Bhn by post-translational chemical mutagenesis, overcoming the well-known limitations to Tyr and Pyl mimics imposed on Uaa's incorporated by means of genetic code expansion. To our knowledge, this is the first incorporation of an Uaa amenable for protein-protein crosslinking by post-translational mutagenesis to date, we presume many more applications of this kind will emerge in the coming years. This will required the MS/MS methods for the analysis of chemically crosslinked peptides to be more stringently evaluated and validated, in relation to both observed fragment ions as well as statistical validation of identifications, as we have began to set out as part of this project.

Incorporating reactivities amenable for PEPC by post-translational mutagenesis in theory not only allows to incorporate a broader range of chemical functionalities than genetic code expansion methods, including structural mimics closer to their naturally occurring amino acids (with or without PTM), but this experiment design also allows *de facto* overcoming known limitations of small molecule crosslinkers.

In a recent thorough study[859], Lei Wang and his group illustrate that the widely used small molecule cross-linker BS3 fails to *in vitro* capture the interaction site between yeast PCNA (yPCNA) and its known interactor the ubiquitin-conjugating enzyme UBE2D3 by SDS-PAGE or WB, noting:

> [No crosslinking with BS3] is consistent with our experience that enzyme-substrate interactions are generally elusive to this method. (Lei Wang *and co-workers*[859])

Using their Uaa **37h**, Wang *and co-workers* could show that a crosslink occurs, illustrating the power of PEPC approaches over the small molecule-mediated crosslink. However, we note that this protein-protein interaction occurs on the protein surface at solvent-exposed sites, where the increased demands in space due to the aromatic ring in **37h** as compared to the naturally occurring Thr163, Lys164 and Glu165 of yPCNA presumably do not extensively affect the binding. Crosslinking with alkyl halide **40b** as presented in this chapter enables precise probing of buried, active sites without impairing substantially on the recognition process.

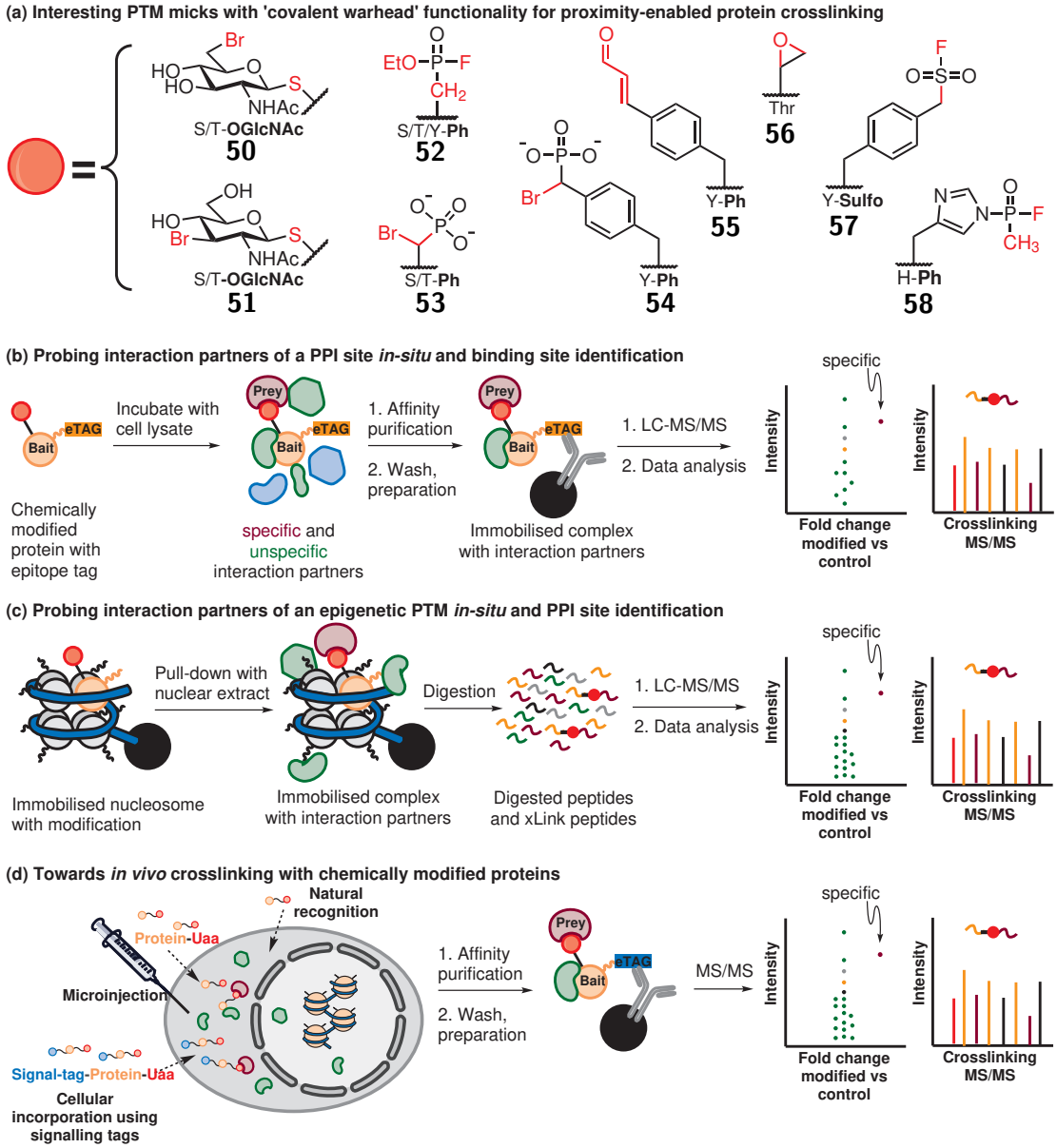


Figure 5.7. Outlook on proximity-enabled protein crosslinking with post-translationally mutated proteins. A. Suggestion of PTM mimics with reactive 'warhead' functionality for covalent proximity-enabled protein-protein crosslinking. These suggestions complement functionalities shown in Figure 5.1 and are partially inspired from covalent ABPP probes. The structural changes in these moieties should be sufficiently small/subtle to be considered 'biologically discrete' and their reactivities fall in the 'Goldilocks' kinetic window. As opposed to currently used 'non-reactive' PTM mimics used for LFQ quantification via their interaction partner's affinity, these would allow to directly identify the binding site in MS/MS studies, in a similar approach as presented in this work. **B.** Suggestion for an affinity purification mass spectrometry approach to identify interaction partners by LFQ and their binding site. This approach was attempted together with Brian Josephson and Dr. Charlie Fehl as part of this study, but yielded inconclusive results due to technical difficulties involved.

Figure 5.7. (***continued***) **C.** Pull-down experiments with synthetic nucleosomes as described in Chapter 4 could be additionally used for binding site identification. **D.** Possible strategies for the *in vivo* incorporation of post-translationally mutated proteins for the identification of binding partners. Note that in theory, signalling molecules with the desired PTMs could be incorporated by the same means, in that case a full proteome readout would be more advisable.

The restrictions of genetic code expansion[729] have so far limited the use of proximity-enabled protein crosslinking (PEPC) to residues at the protein surface, where structural changes resulting of the inclusion of bulky Pyr and Tyr mimics can easily be accommodated (see Figure 5.1). The potential of PEPC has been illustrated by many different surface PPI's (affibody/protein Z, PCNA, PAPS, CaM, different antibody Fab's, Rab1b etc.)[848,850,854,855,878]. PEPC had not yet been used to probe binding sites inside proteins with the only notable exception of the substitution of Tyr by F_2-Tyr **29** inside the active site of Cysteamine dioxygenase (ADO)[852] – a very discrete structural change indeed. Our case study emphasises the point that to probe structurally defined protein-protein recognition events, such as inside protein cavities, the structural change need to be minimal such as replacing the $-NH_2$ group by CH_2Br, keeping the chain length intact. We believe this was key to identify the Zn-binding site as site of interaction in the present study.

The emergence of increasingly benign and bio-compatible protein chemistries for post-translational protein editing in recent years opens up the possibility to incorporate previously inaccessible reactive mimics of post-translational modifications, minimizing structural and electronic alterations while introducing new functionalities. Figure 5.7.A for example speculates about possibilities that could be explored regarding GlcNAcylation and phosphorylation. Yet, the common underlying principle of 'Goldilocks kinetics' or effective molarity for all these synthetic modifications remains: the reaction kinetics need to be balanced between a sufficient reactivity to trap the interaction partners by virtue of their effective molarity but a mild to low reactivity freely in solution to limit unspecific cross-linking.

Whereas the *in-vitro* crosslinking of interactions partners as presented in this study has proof-of-concept and also investigative relevance, ultimately applications *in situ* or *in vivo* are most relevant for the discovery of interaction partners and the mapping of the PPI binding site. Genetic code expansion methods do have the advantage over synthetic protein chemistry methods to incorporate Uaa's in living cells – bioorthogonality of the tRNA/PylRS given. However, methods to incorporate expressed proteins in live cells have emerged, relying on signalling/epitope tags or technological progress such as microinjections (Figure 5.7.D). The therapeutic potential in the creation of covalent

protein drugs has been recently demonstrated,[879] exciting new avenues for proximity-enabled protein crosslinking are due to follow in the coming years.

5.4. Author contributions

This highly collaborative project compiles contributions from numerous authors. Dr. Charlie Fehl developed the photoredox protein chemistry. Brian Josephson expressed and chemically modified histone proteins, performed the crosslinking assays, analysis by SDS-PAGE and Western Blot. Simon Nadal developed & performed all LC-MS/MS methods and structural analysis. The Akane Kawamura group kindly donated KDM4A and performed Zn ejection kinetics.

This study was published in:
Charlie Fehl*, Brian Josephson*, Patrick Isenegger*, Simon Nadal, Tom H. Wright, Adeline W. J. Poh, Ben J. Bower, Andrew M. Giltrap, Lifu Chen, Christopher Batchelor-McAuley, Grace Rober, Oluwatobi Arisa, Jeroen B. I. Sap, Akane Kawamura, Andrew J. Baldwin, Shabaz Mohammed, Richard G. Compton, Veronique Gouverneur, Benjamin G. Davis. Light-driven post-translational installation of reactive protein side chains. *Nature* **2020**, *585*, 530-537.
* These authors contributed equally.

All MS/MS data underlying this work is publicly accessible in the PRIDE repository (https://www.ebi.ac.uk/pride/) under the accession number PXD019565.

5.5. Experimental

5.5.1. General Experimental Procedures

Unless otherwise noted, chemical reagents, media, and *Escherichia coli* cell stocks were obtained from commercial suppliers (Sigma-Aldrich, Fluorochem, Carbosynth, VWR, Alfa Aesar, Fisher Scientific) and used without further purification. Sonication was performed using a Fisher Scientific Model 505 Sonic Dismembrator. Proteins were purified using an Äkta FPLC System UPC-900 (GE Healthcare, UK). Gel electrophoresis was performed using Invitrogen NuPAGE 4-12 % Bis-Tris gels, Novex MiniCell tanks, and a BioRad PowerPac controller. Western blotting was performed using an iBlot gel transfer device from Invitrogen. Protein crystal structures were analysed and displayed using PyMOL v. 1.7. Synthetic gene fragments (i.e. for human histone eH3-FLAG-HA constructs) were obtained from GeneArt Gene Synthesis (Thermo Fisher Scientific).

Intact Protein Mass Spectrometry

Intact protein mass spectrometry was performed on a Waters Xevo G2-S QTof coupled to Water Acquity UPLC. Separation was achieved using a Thermo Proswift (250 mm × 4.6 mm × 5 µm) column with water + 0.1% formic acid (solvent A) and acetonitrile + 0.1% formic acid (solvent B) as the eluent system over a 10-minute linear gradient. Nitrogen was used as the desolvation gas (600 L/h) for positive electrospray ionization. Voltages used were capillary: 3000 V, cone: 160 V. Lock-spray analysis ensured continual calibration against a leucine enkephalin standard solution. Raw spectra containing multiple charged ion series were deconvoluted using MassLynx (Waters) and its maximum entropy (MaxEnt1) deconvolution algorithm (Resolution: 1.00 Da/channel, width at half height: ion series/protein dependent, minimum intensity ratios: 33% Left and Right). Spectra were deconvoluted between 10000 and 20000 Da for *Homo sapiens* Histone H3.

All **tandem mass spectrometry** methods are described in section 5.5.3.

5.5.2. Histone protein expression, chemical modification and crosslinking

Histone proteins were expressed and chemically modified by Brian Josephson or Dr. Charlie Fehl.

Expression of histone proteins with cysteine mutation at the site of interest

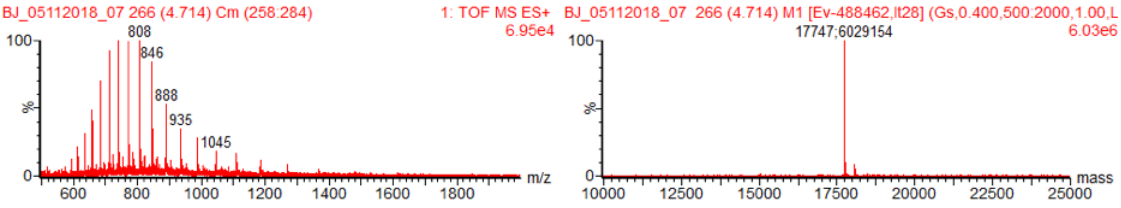
Briefly, the gene for the WT human histone H3.1 (C-terminal FLAG-HA tag, C96A and C110A) was ordered from Thermo Fisher (GeneArt service) and cloned into the pET3d expression plasmid at the NcoI and BamHI restriction enzyme sites. QuikChange mutagenesis was performed per the manufacturer's instructions (QuikChange II Site- Directed Mutagenesis Kit, Agilent) to create the desired cysteine mutants. The *Escherichia coli* strain BL21(DE3)pLysS was transformed as appropriate and selected on chloramphenicol and ampicillin. Single colonies were used to inoculate 5-20 mL starter cultures in LB broth with the same antibiotics. Flasks containing 500 mL 2xTY media were inoculated with 1% v/v of starters, grown at 37 °C until OD_{600} = 0.4-0.8. The production of histones were induced with 0.5 mM IPTG and allowed to proceed for 2 h before harvest and resuspension into 5-fold volume/weight "wash buffer" (50 mM Tris, pH 7.5, 100 mM NaCl with a protease inhibitor cocktail). Suspensions were flash-frozen and stored at -80 °C until lysis. Lysis proceeded via sonication in the presence of 1 mg DNase for 5 × 30 second bursts at 40% amplitude. The sonicate was centrifuged for 20 min at 20k rpm at 4 °C. The supernatant was discarded and the pellet resuspended in 40 mL ["wash buffer" + 1% Triton-X detergent]. Sonication was repeated once at 40% amplitude, 30 seconds, and the suspension centrifuged at 20k rpm for 10 minutes. The pellet was washed twice more as described, then once with the Triton-free "wash buffer." 1 mL of DMSO was added to the pellet and crudely mixed with a spatula to aid histone desolvation for 10 minutes. 10 mL "unfolding buffer" (7M Gdn · HCl, 20 mM Tris, pH 7.5, 10 mM DTT) was added and shaken for 1 h at rt, then the mixture was centrifuged for 10 minutes at 20k rpm at room temperature. The supernatant was loaded onto an S200 size exclusion column (GE Healthcare) pre-equilibrated with "SAU-100" buffer (7M urea, 20 mM NaOAc, pH 5.2, 100 mM NaCl, 1 mM EDTA, 10 mM DTT, 1 mM benzamidine). Protein was eluted with SAU-100, analysed by SDS-PAGE, and histone fractions were pooled and concentrated to 1-4 mL. Cation exchange chromatography was used to further purify histones (HisTrap SP 5 mL) using a linear gradient of 0-100% SAU-1000 buffer ("SAU-100" with 1000 mM NaCl final concentration). Pure fractions were pooled, dialyzed against water (with 2 mM β-mercaptoethanol) and lyophilized.

Histone sequences

In the following, histone protein sequences are described together with the result of their intact protein LC-MS analysis, showing both the molecular ion series (left) acquired on a Waters Xevo G2S Q-TOF instrument and the deconvoluted spectrum (right), as described in 5.5.1.

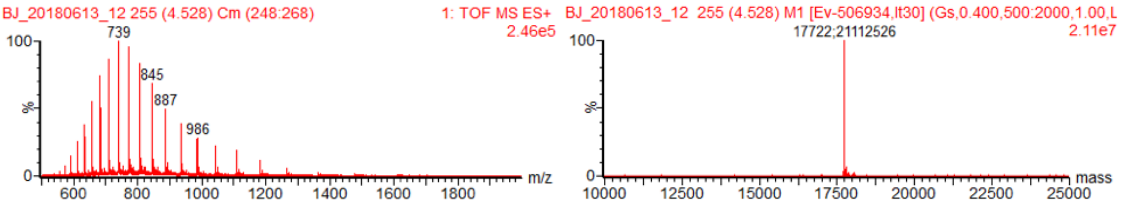
Homo sapiens Histone H3.1 WT

Mutations: C96A, C110A, C-terminal FLAG-HA epitope tags
ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIR
RYQKSTELLIRKLPFQRLVREIAQDFKTDLRFQSSAVMALQEAAEAYLVGLFED
TNLAAIHAKRVTIMPKDIQLARRIRGERAGGDYKDDDDKSAAGGYPYDVPDYA
Calculated mass: 17745 Da, Observed mass: 17747 Da



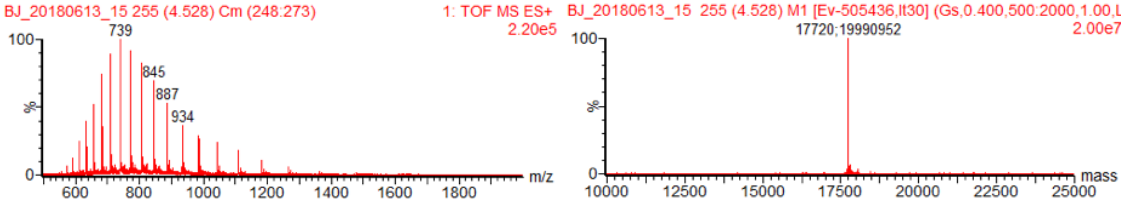
Homo sapiens Histone H3.1 K4C

Mutations: K4C, C96A, C110A, C-terminal FLAG-HA epitope tags
ARTCQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIR
RYQKSTELLIRKLPFQRLVREIAQDFKTDLRFQSSAVMALQEAAEAYLVGLFED
TNLAAIHAKRVTIMPKDIQLARRIRGERAGGDYKDDDDKSAAGGYPYDVPDYA
Calculated mass: 17720 Da, Observed mass: 17722 Da



Homo sapiens Histone H3.1 K9C

Mutations: K9C, C96A, C110A, C-terminal FLAG-HA epitope tags
ARTKQTARCSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTVALREIR
RYQKSTELLIRKLPFQRLVREIAQDFKTDLRFQSSAVMALQEAAEAYLVGLFED
TNLAAIHAKRVTIMPKDIQLARRIRGERAGGDYKDDDDKSAAGGYPYDVPDYA
Calculated mass: 17720 Da, Observed mass: 17720 Da

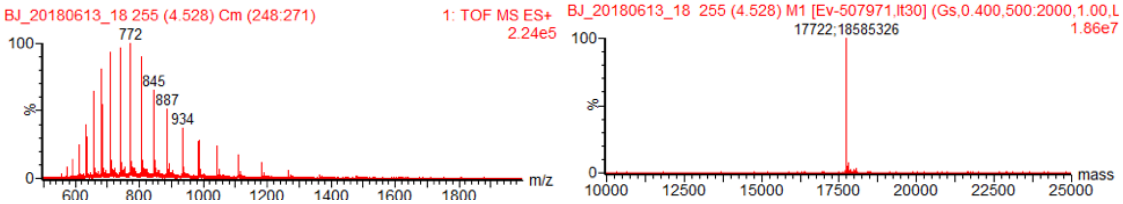


***Homo sapiens* Histone H3.1 K27C**

Mutations: K27C, C96A, C110A, C-terminal FLAG-HA epitope tags

ARTKQTARKSTGGKAPRKQLATKAARCSAPATGGVKKPHRYRPGTVALREIR
RYQKSTELLIRKLPFQRLVREIAQDFKTDLRFQSSAVMALQEAAEAYLVGLFED
TNLAAIHAKRVTIMPKDIQLARRIRGERAGGDYKDDDDKSAAGGYPYDVPDYA

Calculated mass: 17720 Da, Observed mass: 17722 Da

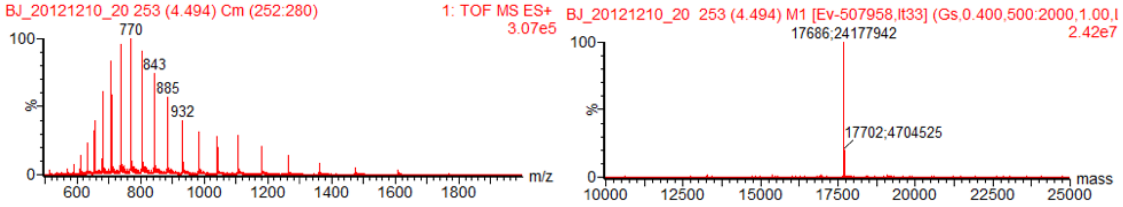


Chemical conversion of cysteine to dehydroalanine

Lyophilized Histone H3 KxC (10 mg) as added to denaturing phosphate buffer (100 mM NaPi, pH 8, 3 M Gdn · HCl, 500 µL) and mixed until fully dissolved. DTT was added (30 mg) and shaken for 30 min at rt, 500 rpm, to reduce disulfide bonds, before being removed via desalting into 1 mL of the same buffer (PD Minitrap G25, GE Healthcare). The resulting protein concentration was determined (Nanodrop) and immediately followed by the addition of DBHDA (60 eq from a freshly prepared 0.5 M DMSO stock) and subsequent shaking (500 rpm) at 25 °C for 45 min, then 37 °C for 2 h. The protein was desalted as before to remove the excess DBHDA and exchanged into the desired buffer. Protein yield and concentration was determined by Nanodrop and conversion was determined by LC-MS analysis.

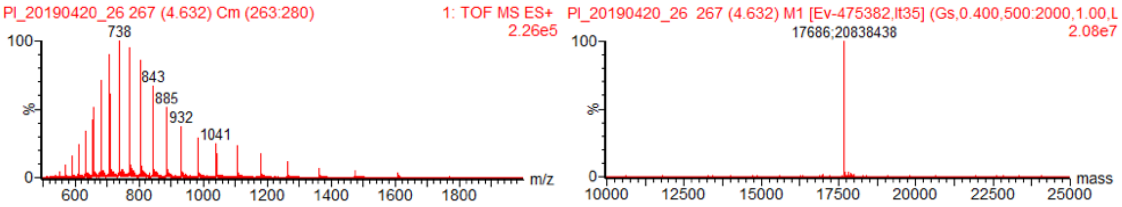
***Homo sapiens* Histone eH3.1 K4Dha - Flag epitope tag**

Calculated mass: 17686 Da ; Observed mass: 17686 Da



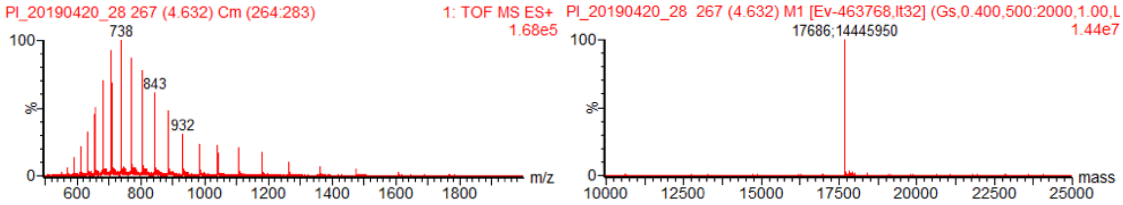
***Homo sapiens* Histone eH3.1 K9Dha - Flag epitope tag**

Calculated mass: 17686 Da ; Observed mass: 17686 Da



***Homo sapiens* Histone eH3.1 K27Dha - Flag epitope tag**

Calculated mass: 17686 Da ; Observed mass: 17686 Da



Chemical conversion of dehydroalanine to bromohomonorleucine (Bhn, 40b)

General procedure

All solutions were degassed for at least 8 h in a glovebox (<6 ppm O2). Clear glass vials (Chromacol 300 µL fixed insert vial, clear, screw top, Thermo Scientific for < 100 µL, and 2 mL CLR RAM VIAL 9MM THD, 32009-1232, Novetech for ≥ 100 µL) with gas-tight caps (Cat. No VWRI548-3298, VWR) were used. Standard reactions consisted of mixing the Dha-containing protein of choice into the desired reaction buffer in the glovebox, followed by the sequential addition of catalyst, catechol, and chemical substrates from 100× stocks prepared fresh in the glovebox. Reactions were performed on *Xenopus laevis* Histone H3 K9Dha (1 mg/mL) in denaturing buffer (500 mM NH4OAc, 5 M guanidinium chloride, pH 6.0) at a final concentration of 1 mg/mL (66 µM) in volumes of 50-200 µL. All reagents were water soluble and required no co-solvents. The reactions were then mixed thoroughly by pipette, capped, and removed from the glovebox for irradiation. 3 W blue (ca. 450 nm) LED flashlights were arranged for even irradiation of up to 20 reaction vials at a time or the photobox was used for up to 7 reactions at a time with blue LED intensities ranging from 5-50 W (Intensity readings of 1-10 on the dial, respectively). Irradiation proceeded for the desired time, afterwards which an aliquot of the reaction was diluted 25-fold for mass spectrometric analysis (2 µL in 48 µL water + 0.1% formic acid) and conversions calculated relative to intensity of deconvoluted peaks

of starting material, single, double addition, and any side reactions observed. Protein recovery was estimated at ca. 85% using PD SpinTrap G-25 (GE Healthcare) desalting columns and tracking overall protein absorbance. The modified proteins could be stored in the freezer as a crude reaction mixture for several months without any degradation or appearance of new adducts, and in several cases, incomplete reactions could be continued simply by degassing the reaction mixture again and continuing with irradiation.

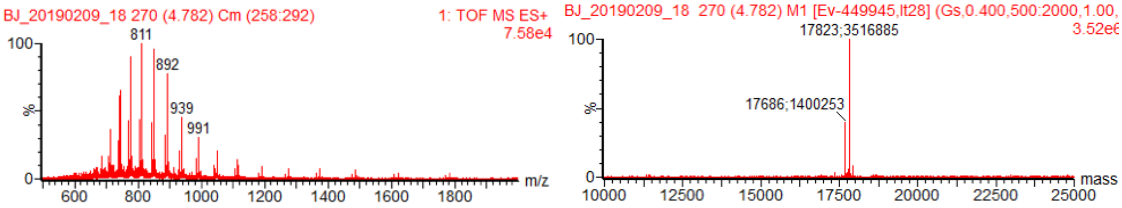
Specific procedure for *Homo sapiens* Histone H3.1 KxBhn 40b

In the glovebox, a glass HPLC vial was charged with NH4OAc buffer (500 mM, pH 6, 3M Gdn · HCl, 95 µL) containing Human Histone H3 KxDha (100 µg, final concentration of 1 mg/mL, 56 µM). After the sequential addition of Ru(bpm)3Cl2 (1 µL of a 56 mM stock prepared fresh in water, 10 eq), catechol (0.5 µL of a 560 mM stock prepared fresh in water, 50 eq) and 4-bromobutylboronic acid (5 µL of a 560 mM stock prepared fresh in buffer, 500 eq), the vial was sealed with a cap, transferred out of the glovebox and irradiated with blue LED light (50 W) for 1 h. Conversion was determined by analysis of an aliquot of the crude mixture by LC-MS.

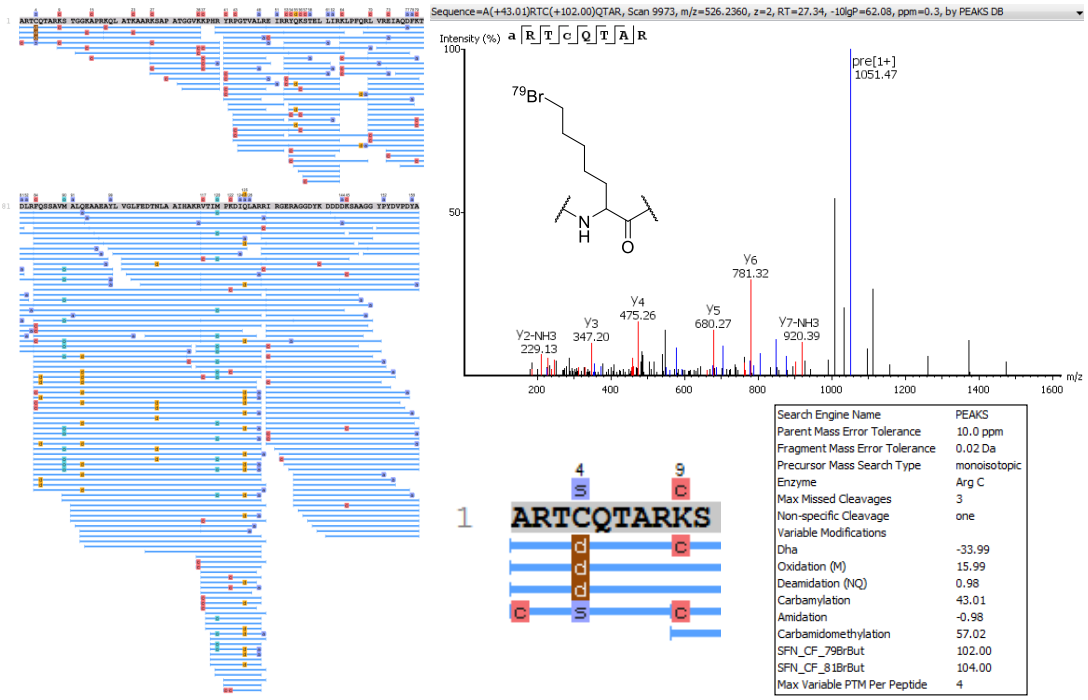
***Homo sapiens* Histone eH3.1 K4Bhn - Flag epitope tag**

Calculated mass: 17822 Da ; Observed mass: 17823 Da

Calculated Dha mass: 17686 Da ; Observed Dha mass: 17686 Da



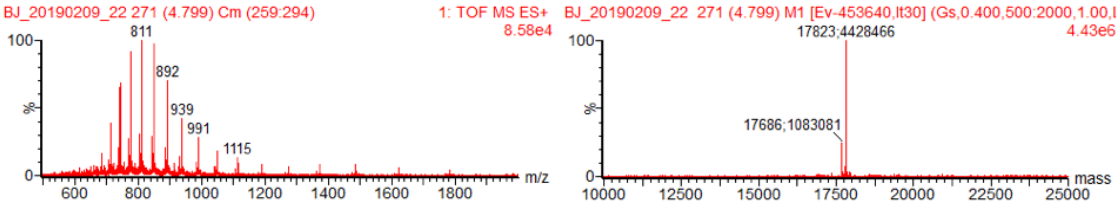
LC-MS/MS for eH3K4Bhn. ArgC digest leads to 100% sequence coverage (1% FDR). Q Exactive (HCD):



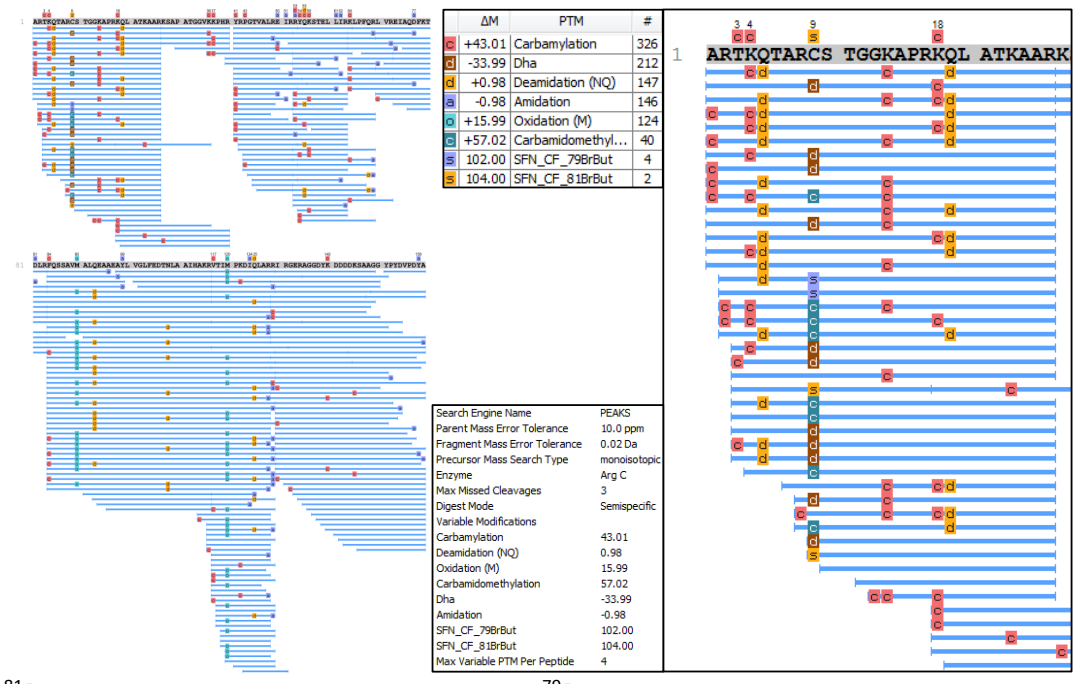
***Homo sapiens* Histone eH3.1 K9Bhn - Flag epitope tag**

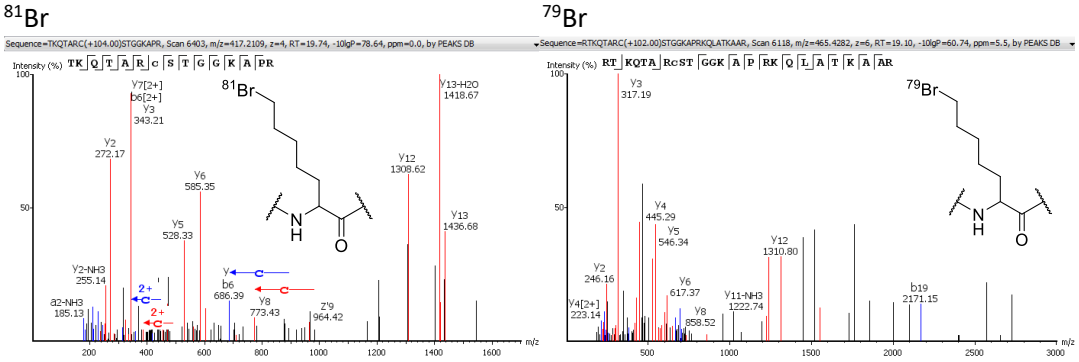
Calculated mass: 17822 Da ; Observed mass: 17823 Da

Calculated Dha mass: 17686 Da ; Observed Dha mass: 17686 Da



LC-MS/MS for eH3K9Bhn. ArgC digest leads to 100% sequence coverage (1% FDR). Q Exactive (HCD):

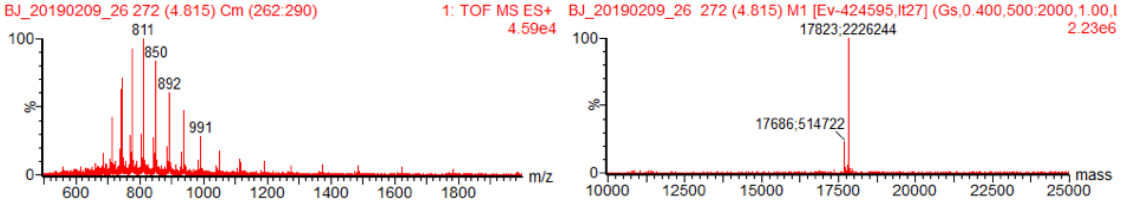




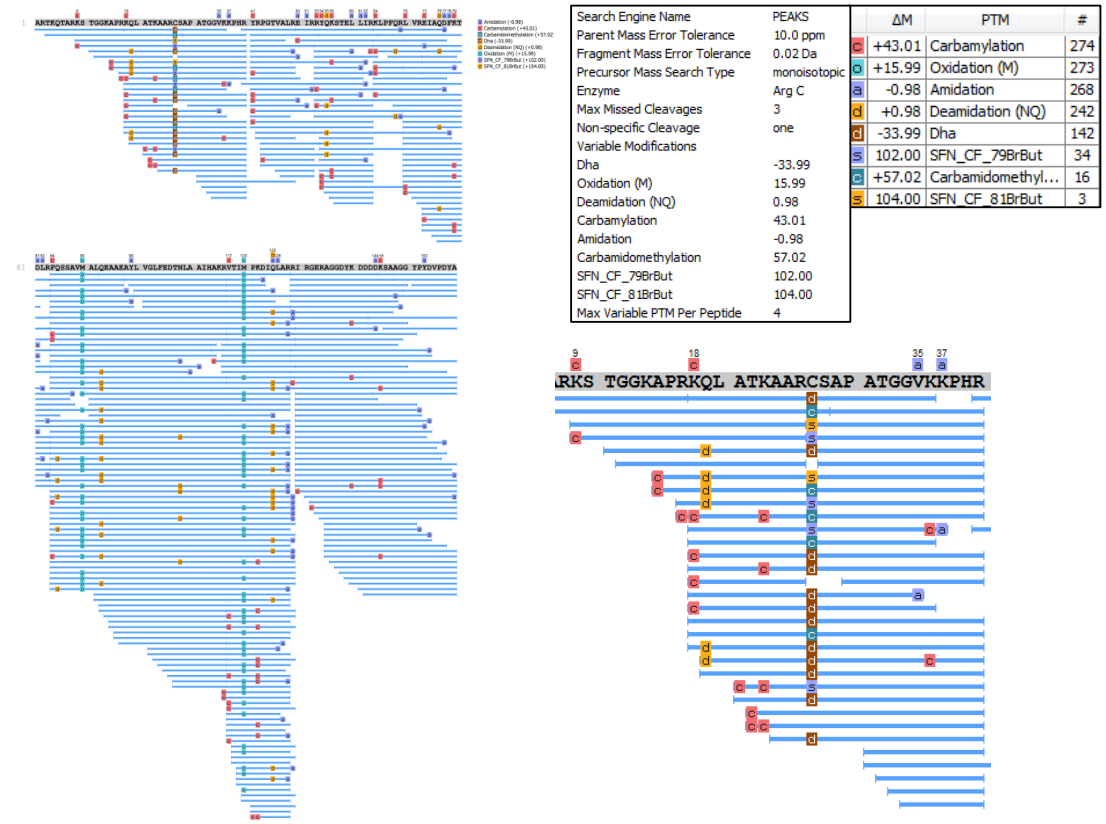
***Homo sapiens* Histone eH3.1 K27Bhn - Flag epitope tag**

Calculated mass: 17822 Da ; Observed mass: 17823 Da

Calculated Dha mass: 17686 Da ; Observed Dha mass: 17686 Da



LC-MS/MS for eH3K27Bhn. ArgC digest leads to 100% sequence coverage (1% FDR). Q Exactive (HCD):

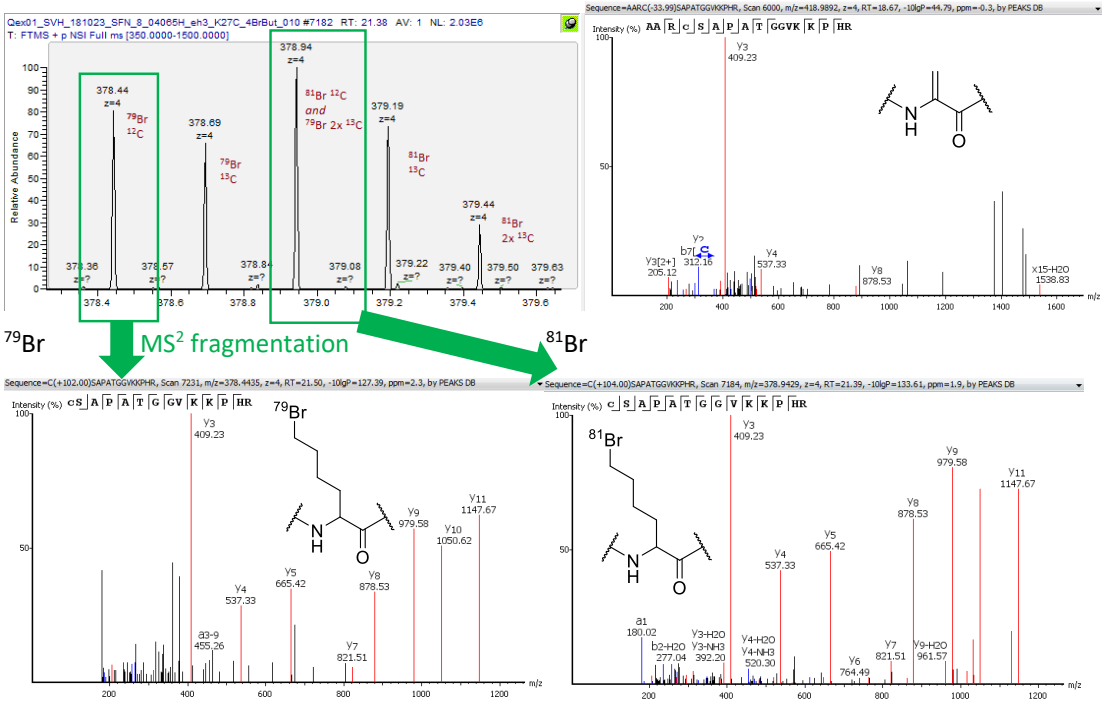


Search Engine Name	PEAKS
Parent Mass Error Tolerance	10.0 ppm
Fragment Mass Error Tolerance	0.02 Da
Precursor Mass Search Type	monoisotopic
Enzyme	Arg C
Max Missed Cleavages	3
Non-specific Cleavage	one
Variable Modifications	
Dha	-33.99
Oxidation (M)	15.99
Deamidation (NQ)	0.98
Carbamylation	43.01
Amidation	-0.98
Carbamidomethylation	57.02
SFN_CF_79BrBut	102.00
SFN_CF_81BrBut	104.00
Max Variable PTM Per Peptide	4

	ΔM	PTM	#
c	+43.01	Carbamylation	274
o	+15.99	Oxidation (M)	273
a	-0.98	Amidation	268
d	+0.98	Deamidation (NQ)	242
d	-33.99	Dha	142
s	102.00	SFN_CF_79BrBut	34
c	+57.02	Carbamidomethyl...	16
s	104.00	SFN_CF_81BrBut	3

MS[1] isotopic pattern

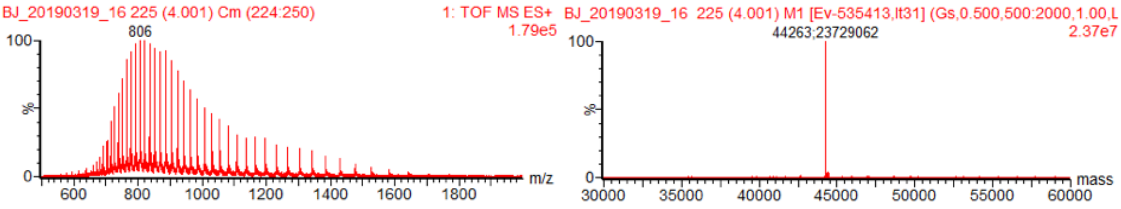
Dha



KDM4A Crosslinking to Bhn Containing Histones

The histone modifying enzyme KDM4A (2 µM) was mixed with either the modified histone eH3.1 K4/9/27Bhn or the WT control (4 µM) in HEPES buffer (50 mM, pH 7.4) and incubated at the indicated temperature and for the indicated time. The crosslinking reaction was quenched with the addition of 5× Laemmli buffer and analysed via SDS-PAGE or Western Blot. Unfortunately, no intact ESI-LC-MS could be obtained for the crosslinked proteins. LC-MS/MS binding site identification is described in subsequent sections.

KDM4A *Sequence and LC-MS Characterization*
KDM4A was kindly donated by the Akane Kawamura group, it was expressed as described in the literature.[867]
N-terminal His6 tagged truncated KDM4A 1-349: MHHHHHHSSGVDLGTENLYFQS MASESETLNPSARIMTFYPTMEEFRNFSRYIAYIESQGAHRAGLAKVVPPKEWK PRASYDDIDDLVIPAPIQQLVTGQSGLFTQYNIQKKAMTVREFRKIANSDKYCTP RYSEFEELERKYWKNLTFNPPIYGADVNGTLYEKHVDEWNIGRLRTILDLVEKE SGITIEGVNTPYLYFGMWKTSFAWHTEDMDLYSINYLHFGEPKSWYSVPPEHG KRLERLAKGFFPGSAQSCEAFLRHKMTLISPLMLKKYGIPFDKVTQEAGEFMIT FPYGYHAGFNHGFNCAESTNFATRRWIEYGKQAVLCSCRKDMVKISMDVFVR KFQPERYKLWKAGKDNTVIDHTLPTPEAAEFLKESEL
Calculated mass: 44265 Da ; Observed mass: 44263 Da



5.5.3. Tandem mass spectrometry

ArgC in-solution digest – *Homo sapiens* Histone H3.1 K9Bhn

Approx. 10 µg of modified protein sample were taken in 8M urea in 50 mM TEAB containing 20 mM methylamine to a total volume of 100 µL and denatured for 30 min at room temperature. Samples were reduced with 10 mM TCEP for 30 min at room temperature and alkylated with 50 mM chloroacetamide for 30 min at room temperature in the dark. Samples were diluted to 1M urea and digested with Arg-C (1:20 w/w) in activation buffer (50 mM TEAB, 0.2 mM EDTA, 5 mM TCEP) for 4 h – O/N at 37 °C. The reaction was stopped by addition of 10 % FA to a final concentration of 0.5 %. Samples were desalted by C18 (Oasis HLB 10 mg cartridge) and dried in a speed-vac before being resuspended in 5% FA 5% DMSO.

Tryptic in-solution digest – H3.1 K9Bhn•KDM4A crosslink

Approx. 10 µg (20 µL) of desalted & denatured modified protein sample were taken in 8 M urea in 100 mM TEAB to a total volume of 100 µL. Samples were reduced with 10 mM TCEP for 30 min at room temperature and alkylated with 50 mM chloroacetamide for 30 min at room temperature in the dark. The solution was diluted to 1 M urea with 50 mM TEAB and digested with AspN or Trypsin or Elastase 1:20 (w/w) 4h to overnight at 37 °C. %. Samples were desalted by C18 (Oasis HLB 10 mg) and dried in a speed-vac before being resuspended in 5% FA 5% DMSO.

Tryptic in-gel digest – all crosslinks

The KDM4-H3Kx crosslink was separated by SDS-PAGE and bands were excised with a scalpel and cut in 3-4 small pieces. Gel pieces were covered with 200 µL 50% acetonitrile and destained for 10 min at 37 °C. The supernatant was discarded and 200 µL of 10 mM TCEP in 100 mM TEAB were added and the gel pieces incubated for 30 min at r.t. The supernatant was discarded and the gel pieces dried 2× with 200 µL pure acetonitrile, with 10 min incubation each. 200 µL of 50 mM chloroacetamide was added and the gel pieces incubated for 30 min in the dark. Upon removal of the supernatant, the gel pieces were washed 2× with acetonitrile until completely shrunk. 10 µL of 0.5 µg/µL trypsin solution was added. After absorption of the liquid, 50 µL of 25 mM TEAB was added and the mixture incubated for 4 h at 37 °C. The supernatant was collected together with a subsequent wash of the gel pieces with 15 µL 10% FA.

Standard data acquisition – Q Exactive

Resulting peptides 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 (75 µm inner diameter x 50cm packed with ReproSil-Pur 120 C18-AQ, 1.9 µm, 120 Å, Dr. Maisch GmbH). Separation of cross-linked 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 3×10^6, maximum injection time 50 ms]. After the MS scan, the top 10 most intense peaks were selected for HCD fragmentation at 30% of normalised collision energy. HCD spectra were also acquired in the Orbitrap (resolution 17500; AGC target 5×10^4; maximum injection time 120 ms), with first fixed mass at 180 m/z.

Data acquisition for crosslinked samples – Q Exactive

Resulting peptides 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 (75 µm inner diameter x 50cm packed with ReproSil-Pur 120 C18-AQ, 1.9 µm, 120 Å, Dr. Maisch GmbH). Separation of cross-linked 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 3×10^6, maximum injection time 100 ms]. After the MS scan, the top 10 most intense peaks were selected for HCD fragmentation at 30% of normalised collision energy, excluding 1^+ and

2^+ charged species. HCD spectra were also acquired in the Orbitrap (resolution 17500; AGC target 5×10^4; maximum injection time 120 ms, scan range 200 – 2000 *m/z*), with first fixed mass at 180 *m/z*.

Standard data analysis

Searches were performed with Peaks version 8.5 (Bioinformatics Solutions Inc.)[573] for identification and *de-novo* analysis. The raw MS file was searched against the given protein sequence and a list of contaminants (generated from MaxQuant contaminants database). Samples were additionally searched with MaxQuant against the UniProt human database for confirming the purity of the sample. Precursor mass tolerance was set to 10 ppm. Fragment mass tolerances for HCD was set to 0.02 Da. The corresponding protease was selected with a maximum number of 3 missed cleavages and non-specific cleavage at one end of the peptide. Oxidation (Methionine), Deamination (Asparagine, Glutamine), Carbamidomethylation (Cysteine – except for ArgC variant 1 digest), Carbamylation (lysine, peptide N-term), Amidation (C-terminus), dehydroalanine (Cysteine, -33.9877), 79BrBut (Cysteine, 102.0011) and 89BrBut (Cysteine, 103.9990) were set as variable modifications. A maximum of 4 variable modifications was set. A FDR of 1% on peptide level and de-novo ALC of 80 was applied. All spectra and identifications were manually validated. For analysis of isotopic pattern, manual analysis with XCalibur Qual Browser was performed.

Data analysis of crosslinked samples

Crosslinked samples were searched with Peaks 8.5 as described previously to confirm presence of both proteins and to confirm sample purity. Crosslinked samples were processed with the pLink 2.3.5 software package.[406] Using the pConfig module, the amino acid 'B' of mass 205.01023 and composition H(12)C(7)N(1)O(1)Br(1) was defined. Note that the isotopic pattern doesn't matter for crosslinking analysis as HBr gets eliminated. The linker 4BrBut was defined as linking the alpha amino acid B to the beta amino acids STYCHRKWDENQ with the linker composition H(-1)Br(-1).

Name	AlphaSites	BetaSites	LinkerMass	MonoMass	Composition	MonoComposiiton
4BrBut	8	STYCHRKWDENQ	-79.926	0	H(-1)Br(-1)	H(0)

First search mass tolerance was set to 20 ppm and fragment mass tolerances at 20 ppm. A mass filter of 10 ppm was applied. E-values were computed and a global FDR of 1% set (at spectral level). Trypsin (or LysC for H3K4) was selected as a protease with a

maximum number of 3 missed cleavages. Carbamidomethylation (Cysteine), Oxidation (Methionine), Deamidation (Asparagine, Glutamine), Carbamylation (lysine, peptide N-term), Amidation (C-terminus) were set as variable modifications. RAW files were searched against a database containing the sequence of the modified eH3 construct and the expressed KDM4. Resulting spectra were manually analysed and validated using pLabel. As an empirical rule, an e-value higher than e-03 indicates a potential identification, higher than e-06 a reliable identification and higher than e-10 a very good identification.

5.6. References

385. O'Reilly, F. J. & Rappsilber, J. Cross-Linking Mass Spectrometry: Methods and Applications in Structural, Molecular and Systems Biology. *Nature Structural & Molecular Biology* **25,** 1000–1008 (2018).

386. Yang, B. *et al.* Identification of Cross-Linked Peptides from Complex Samples. *Nature Methods* **9,** 904–906 (2012).

388. *Thermo Scientific - Crosslinking Technical Handbook* `http://tools.thermofisher.com/content/sfs/brochures/1602163-Crosslinking-Reagents-Handbook.pdf` (2019).

399. Rappsilber, J. The Beginning of a Beautiful Friendship: Cross-Linking/Mass Spectrometry and Modelling of Proteins and Multi-Protein Complexes. *Journal of Structural Biology. Combining Computational Modeling with Sparse and Low-Resolution Data* **173,** 530–540 (2011).

406. Chen, Z.-L. *et al.* A High-Speed Search Engine pLink 2 with Systematic Evaluation for Proteome-Scale Identification of Cross-Linked Peptides. *Nature Communications* **10,** 3404 (2019).

459. Chavez, J. D. & Bruce, J. E. Chemical Cross-Linking with Mass Spectrometry: A Tool for Systems Structural Biology. *Current Opinion in Chemical Biology. Omics* **48,** 8–18 (2019).

483. Kleiner, P., Heydenreuter, W., Stahl, M., Korotkov, V. S. & Sieber, S. A. A Whole Proteome Inventory of Background Photocrosslinker Binding. *Angewandte Chemie International Edition* **56,** 1396–1401 (2017).

573. Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A. & Ma, B. PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Molecular & Cellular Proteomics* **11,** M111.010587 (2012).

729. Chin, J. W. Expanding and Reprogramming the Genetic Code. *Nature* **550,** 53–60 (2017).

732. Dumas, A., Lercher, L., Spicer, C. D. & Davis, B. G. Designing Logical Codon Reassignment – Expanding the Chemistry in Biology. *Chemical Science* **6,** 50–69 (2014).

755. Chalker, J. M., Gunnoo, S. B., Boutureira, O., Gerstberger, S. C., Fernández-González, M., Bernardes, G. J. L., Griffin, L., Hailu, H., Schofield, C. J. & Davis, B. G. Methods for Converting Cysteine to Dehydroalanine on Peptides and Proteins. *Chemical Science* **2,** 1666–1676 (2011).

842. Budayeva, H. G. & Cristea, I. M. A Mass Spectrometry View of Stable and Transient Protein Interactions. *Advances in experimental medicine and biology* **806,** 263–282 (2014).

843. Phizicky, E. M. & Fields, S. Protein-Protein Interactions: Methods for Detection and Analysis. *Microbiological Reviews* **59,** 94–123 (1995).

844. Berggård, T., Linse, S. & James, P. Methods for the Detection and Analysis of Protein–Protein Interactions. *PROTEOMICS* **7,** 2833–2842 (2007).

845. B-Rao, C., Subramanian, J. & Sharma, S. D. Managing Protein Flexibility in Docking and Its Applications. *Drug Discovery Today* **14,** 394–400 (2009).

846. Stahl, M. *Chemical Protein Manipulations and Their Application in Proteomics Data Mining* (Technical University of Munich, 2018).

847. Nguyen, T.-A., Cigler, M. & Lang, K. Expanding the Genetic Code to Study Protein–Protein Interactions. *Angewandte Chemie International Edition* **57,** 14350–14361 (2018).

848. Cigler, M., Müller, T. G., Horn-Ghetko, D., von Wrisberg, M.-K., Fottner, M., Goody, R. S., Itzen, A., Müller, M. P. & Lang, K. Proximity-Triggered Covalent Stabilization of Low-Affinity Protein Complexes In Vitro and In Vivo. *Angewandte Chemie International Edition* **56,** 15737–15741 (2017).

849. Chen, X.-H., Xiang, Z., Hu, Y. S., Lacey, V. K., Cang, H. & Wang, L. Genetically Encoding an Electrophilic Amino Acid for Protein Stapling and Covalent Binding to Native Receptors. *ACS Chemical Biology* **9,** 1956–1961 (2014).

850. Furman, J. L., Kang, M., Choi, S., Cao, Y., Wold, E. D., Sun, S. B., Smider, V. V., Schultz, P. G. & Kim, C. H. A Genetically Encoded Aza-Michael Acceptor for Covalent Cross-Linking of Protein–Receptor Complexes. *Journal of the American Chemical Society* **136,** 8411–8417 (2014).

851. Xuan, W., Shao, S. & Schultz, P. G. Protein Crosslinking by Genetically Encoded Noncanonical Amino Acids with Reactive Aryl Carbamate Side Chains. *Angewandte Chemie International Edition* **56,** 5096–5100 (2017).

852. Wang, Y., Griffith, W. P., Li, J., Koto, T., Wherritt, D. J., Fritz, E. & Liu, A. Cofactor Biogenesis in Cysteamine Dioxygenase: C-F Bond Cleavage with Genetically Incorporated Unnatural Tyrosine. *Angewandte Chemie International Edition* **57,** 8149–8153 (2018).

853. Xiang, Z., Ren, H., Hu, Y. S., Coin, I., Wei, J., Cang, H. & Wang, L. Adding an Unnatural Covalent Bond to Proteins through Proximity-Enhanced Bioreactivity. *Nature Methods* **10,** 885–888 (2013).

854. Xuan, W., Li, J., Luo, X. & Schultz, P. G. Genetic Incorporation of a Reactive Isothiocyanate Group into Proteins. *Angewandte Chemie* **128,** 10219–10222 (2016).

855. St. Amant, A. H. *et al.* A Diene-Containing Noncanonical Amino Acid Enables Dual Functionality in Proteins: Rapid Diels–Alder Reaction with Maleimide or Proximity-Based Dimerization. *Angewandte Chemie International Edition* **58,** 8489–8493 (2019).

856. Wang, N. *et al.* Genetically Encoding Fluorosulfate-l-Tyrosine To React with Lysine, Histidine, and Tyrosine via SuFEx in Proteins in Vivo. *Journal of the American Chemical Society* **140,** 4995–4999 (2018).

857. Hoppmann, C., Lacey, V. K., Louie, G. V., Wei, J., Noel, J. P. & Wang, L. Genetically Encoding Photoswitchable Click Amino Acids in Escherichia Coli and Mammalian Cells. *Angewandte Chemie International Edition* **53,** 3932–3936 (2014).

858. Xiang, Z., Lacey, V. K., Ren, H., Xu, J., Burban, D. J., Jennings, P. A. & Wang, L. Proximity-Enabled Protein Crosslinking through Genetically Encoding Haloalkane Unnatural Amino Acids. *Angewandte Chemie International Edition* **53,** 2190–2193 (2014).

859. Yang, B. *et al.* Spontaneous and Specific Chemical Cross-Linking in Live Cells to Capture and Identify Protein Interactions. *Nature Communications* **8,** 2240 (2017).

860. Josephson, B. *et al.* Light-Driven Post-Translational Installation of Reactive Protein Side Chains. *Nature* **585,** 530–537 (7826 2020).

861. Catalano, M. J., Price, N. E. & Gates, K. S. Effective Molarity in a Nucleic Acid-Controlled Reaction. *Bioorganic & medicinal chemistry letters* **26,** 2627–2630 (2016).

862. Page, M. I. & Jencks, W. P. Entropic Contributions to Rate Accelerations in Enzymic and Intramolecular Reactions and the Chelate Effect. *Proceedings of the National Academy of Sciences* **68,** 1678–1683 (1971).

863. Krishnamurthy, V. M., Semetey, V., Bracher, P. J., Shen, N. & Whitesides, G. M. Dependence of Effective Molarity on Linker Length for an Intramolecular Protein-Ligand System. *Journal of the American Chemical Society* **129,** 1312–1320 (2007).

864. Speltz, E. B. & Zalatan, J. G. The Relationship between Effective Molarity and Affinity Governs Rate Enhancements in Tethered Kinase–Substrate Reactions. *Biochemistry* **59,** 2182–2193 (2020).

865. Sun, H., Hunter, C. A., Navarro, C. & Turega, S. Relationship between Chemical Structure and Supramolecular Effective Molarity for Formation of Intramolecular H-Bonds. *Journal of the American Chemical Society* **135,** 13129–13141 (2013).

866. Ng, S. S. *et al.* Crystal Structures of Histone Demethylase JMJD2A Reveal Basis for Substrate Specificity. *Nature* **448,** 87–91 (2007).

867. Kawamura, A. *et al.* Highly Selective Inhibition of Histone Demethylases by *de Novo* Macrocyclic Peptides. *Nature Communications* **8,** 14773 (2017).

868. Dawson, M. A. & Kouzarides, T. Cancer Epigenetics: From Mechanism to Therapy. *Cell* **150,** 12–27 (2012).

869. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125,** 315–326 (2006).

870. Berry, W. L. & Janknecht, R. KDM4/JMJD2 Histone Demethylases: Epigenetic Regulators in Cancer Cells. *Cancer research* **73,** 2936–2942 (2013).

871. Lu, S. *et al.* Mapping Native Disulfide Bonds at a Proteome Scale. *Nature Methods* **12,** 329–331 (2015).

872. Aggarwal, S. & Yadav, A. K. in *Statistical Analysis in Proteomics* (ed Jung, K.) 119–128 (Springer New York, New York, NY, 2016).

873. Tamura, T. & Hamachi, I. Chemistry for Covalent Modification of Endogenous/Native Proteins: From Test Tubes to Complex Biological Systems. *Journal of the American Chemical Society* **141,** 2782–2799 (2019).

874. Freedman, H. H. & Dubois, R. A. An Improved Williamson Ether Synthesis Using Phase Transfer Catalysis. *Tetrahedron Letters* **16,** 3251–3254 (1975).

875. Mandal, S., Mandal, S., K. Ghosh, S., Sar, P., Ghosh, A., Saha, R. & Saha, B. A Review on the Advancement of Ether Synthesis from Organic Solvent to Water. *RSC Advances* **6,** 69605–69614 (2016).

876. Shin, S. & Janknecht, R. Diversity within the JMJD2 Histone Demethylase Family. *Biochemical and Biophysical Research Communications* **353,** 973–977 (2007).

877. Levin, M., Stark, M. & Assaraf, Y. G. The JmjN Domain as a Dimerization Interface and a Targeted Inhibitor of KDM4 Demethylase Activity. *Oncotarget* **9,** 16861–16882 (2018).

878. Wang, L. Genetically Encoding New Bioreactivity. *New Biotechnology* **38,** 16–25 (2017).

879. Li, Q. *et al.* Developing Covalent Protein Drugs via Proximity-Enabled Reactive Therapeutics. *Cell* **182,** 85–97.e16 (2020).

5.7. Appendix: MS/MS spectra

eH3K4_KDM4A Crosslink

Results of tryptic in-gel digest, 20ppm search (20ppm Frag, 10 ppm filter), 1% global FDR

Table 1 Identification of Cross-Linked Results Under 1% FDR Control.

#(Spectra)	#(Peptide Pairs)	#(Linked Sites)
25	9	8

Figure 2 All Precursor Error Distribution.

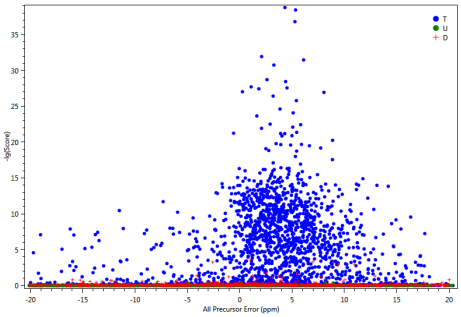
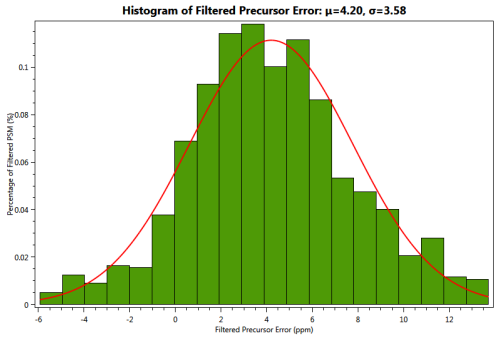


Figure 3 Filtered Precursor Error Distribution.

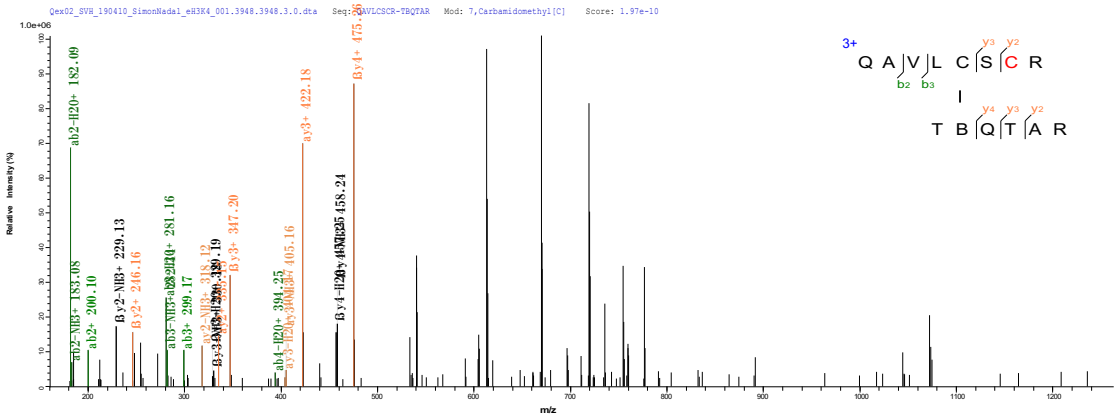


Order	Title	Charge	Precursor_Mass	Peptide	Peptide_Type	Linker	Peptide_Mass	Modifications	Evalue	Score
1	Qen02_SVH_190410_SimonNadal_eH3K4_001.3948.3948.3.0.dta	3	1636.83	QAVLCSCR(5)-TBQTAR(2)	Cross-Linked	4BrBut	1636.83	Carbamidomethyl[C](7)	1.970754e-010	4.939159e-007
2	Qen02_SVH_190410_SimonNadal_eH3K4_001.9866.9866.4.0.dta	4	2.317151e+003	GFFPGSAQSCEAFLR(10)-TBQTAR(2)	Cross-Linked	4BrBut	2.317139e+003	null	2.375866e-010	7.570014e-007
3	Qen02_SVH_190410_SimonNadal_eH3K4_001.10384.10384.4.0.dta	4	2.317149e+003	GFFPGSAQSCEAFLR(10)-TBQTAR(2)	Cross-Linked	4BrBut	2.317139e+003	null	7.891950e-008	2.574378e-006
4	Qen02_SVH_190410_SimonNadal_eH3K4_001.3723.3723.3.0.dta	3	1.636826e+003	QAVLCSCR(5)-TBQTAR(2)	Cross-Linked	4BrBut	1.636826e+003	Carbamidomethyl[C](7)	4.383559e-010	6.157091e-006
5	Qen02_SVH_190410_SimonNadal_eH3K4_001.10456.10456.4.0.dta	4	2.317148e+003	GFFPGSAQSCEAFLR(10)-TBQTAR(2)	Cross-Linked	4BrBut	2.317139e+003	null	3.534963e-004	9.439482e-006
6	Qen02_SVH_190410_SimonNadal_eH3K4_001.9876.9876.3.0.dta	3	2.317152e+003	GFFPGSAQSCEAFLR(10)-TBQTAR(2)	Cross-Linked	4BrBut	2.317139e+003	null	7.639838e-007	2.297456e-005
7	Qen02_SVH_190410_SimonNadal_eH3K4_001.3812.3812.3.0.dta	3	1.636827e+003	QAVLCSCR(5)-TBQTAR(2)	Cross-Linked	4BrBut	1.636826e+003	Carbamidomethyl[C](7)	1.906273e-007	2.806947e-005
8	Qen02_SVH_190410_SimonNadal_eH3K4_001.10407.10407.3.0.dta	3	2.317148e+003	GFFPGSAQSCEAFLR(10)-TBQTAR(2)	Cross-Linked	4BrBut	2.317139e+003	null	3.783959e-009	4.808414e-005
9	Qen02_SVH_190410_SimonNadal_eH3K4_001.3949.3949.4.0.dta	4	1.636828e+003	QAVLCSCR(5)-TBQTAR(2)	Cross-Linked	4BrBut	1.636826e+003	Carbamidomethyl[C](7)	9.646866e-004	8.932771e-004
10	Qen02_SVH_190410_SimonNadal_eH3K4_001.4710.4710.3.2.dta	3	1.679846e+003	QAVLCSCR(5)-TBQTAR(2)	Cross-Linked	4BrBut	1.679832e+003	Carbamyl[AnyN-term](0);Carbamidomethyl[C](7)	2.529051e-005	4.192111e-003
11	Qen02_SVH_190410_SimonNadal_eH3K4_001.4005.4005.4.0.dta	4	1.636828e+003	QAVLCSCR(1)-TBQTAR(2)	Cross-Linked	4BrBut	1.636826e+003	Carbamidomethyl[C](7)	5.207581e-001	4.736071e-003
12	Qen02_SVH_190410_SimonNadal_eH3K4_001.11212.11212.5.0.dta	5	4.631103e+003	VTQEAGEFMITFPYGYHAGFNHGFNCAESTNFATR(26)-TBQTAR(2)	Cross-Linked	4BrBut	4.631114e+003	Oxidation[M](9)	1.750921e-003	8.761483e-003
13	Qen02_SVH_190410_SimonNadal_eH3K4_001.4001.4001.3.0.dta	3	1.636830e+003	QAVLCSCR(5)-TBQTAR(2)	Cross-Linked	4BrBut	1.636826e+003	Carbamidomethyl[C](7)	2.586534e-002	9.075319e-003
14	Qen02_SVH_190410_SimonNadal_eH3K4_001.3735.3735.4.0.dta	4	1.636824e+003	QAVLCSCR(5)-TBQTAR(2)	Cross-Linked	4BrBut	1.636826e+003	Carbamidomethyl[C](7)	3.065555e-002	9.291043e-003
15	Qen02_SVH_190410_SimonNadal_eH3K4_001.10438.10438.4.4.dta	4	2.374171e+003	GFFPGSAQSCEAFLR(6)-TBQTAR(2)	Cross-Linked	4BrBut	2.374161e+003	Carbamidomethyl[C](10)	3.311767e-002	1.919808e-002
16	Qen02_SVH_190410_SimonNadal_eH3K4_001.11241.11241.6.1.dta	6	4.631126e+003	VTQEAGEFMITFPYGYHAGFNHGFNCAESTNFATR(26)-TBQTAR(2)	Cross-Linked	4BrBut	4.631114e+003	Oxidation[M](9)	8.261052e-005	2.839517e-002
17	Qen02_SVH_190410_SimonNadal_eH3K4_001.6752.6752.4.2.dta	4	2.281099e+003	IMTFYPTMEEFR(7)-TBQTAR(2)	Cross-Linked	4BrBut	2.281099e+003	Oxidation[M](2)	7.397563e-002	4.380056e-002
18	Qen02_SVH_190410_SimonNadal_eH3K4_001.11077.11077.5.0.dta	5	4.631134e+003	VTQEAGEFMITFPYGYHAGFNHGFNCAESTNFATR(26)-TBQTAR(2)	Cross-Linked	4BrBut	4.631114e+003	Oxidation[M](9)	6.467943e-003	6.478454e-002
19	Qen02_SVH_190410_SimonNadal_eH3K4_001.9214.9214.4.0.dta	4	2.544284e+003	GFFPGSAQSCEAFLR(10)-ARTBQTAR(4)	Cross-Linked	4BrBut	2.544277e+003	null	1.912617e-014	1.143863e-001
20	Qen02_SVH_190410_SimonNadal_eH3K4_001.10287.10287.4.3.dta	4	2.374171e+003	GFFPGSAQSCEAFLR(6)-TBQTAR(2)	Cross-Linked	4BrBut	2.374161e+003	Carbamidomethyl[C](10)	3.580825e-002	1.397161e-001
21	Qen02_SVH_190410_SimonNadal_eH3K4_001.9896.9896.4.0.dta	4	2.374176e+003	GFFPGSAQSCEAFLR(9)-TBQTAR(2)	Cross-Linked	4BrBut	2.374161e+003	Carbamidomethyl[C](10)	4.084571e-003	1.503996e-001
22	Qen02_SVH_190410_SimonNadal_eH3K4_001.4032.4032.3.1.dta	3	1.693863e+003	QAVLCSCR(1)-TBQTAR(2)	Cross-Linked	4BrBut	1.693847e+003	Carbamidomethyl[C](5);Carbamidomethyl[C](7)	6.357658e-001	1.516536e-001
23	Qen02_SVH_190410_SimonNadal_eH3K4_001.9640.9640.4.3.dta	4	2.374161e+003	GFFPGSAQSCEAFLR(6)-TBQTAR(2)	Cross-Linked	4BrBut	2.374161e+003	Carbamidomethyl[C](10)	4.433725e-001	1.551692e-001
24	Qen02_SVH_190410_SimonNadal_eH3K4_001.3708.3708.4.1.dta	4	2.026007e+003	IANSDKYCTPR(4)-TBQTAR(2)	Cross-Linked	4BrBut	2.026002e+003	Carbamidomethyl[C](8);Deamidated[Q](17)	1.000000e+000	2.817216e-001
25	Qen02_SVH_190410_SimonNadal_eH3K4_001.6924.6924.4.0.dta	4	2.281100e+003	IMTFYPTMEEFR(7)-TBQTAR(2)	Cross-Linked	4BrBut	2.281099e+003	Oxidation[M](2)	3.780599e-002	4.265043e-001

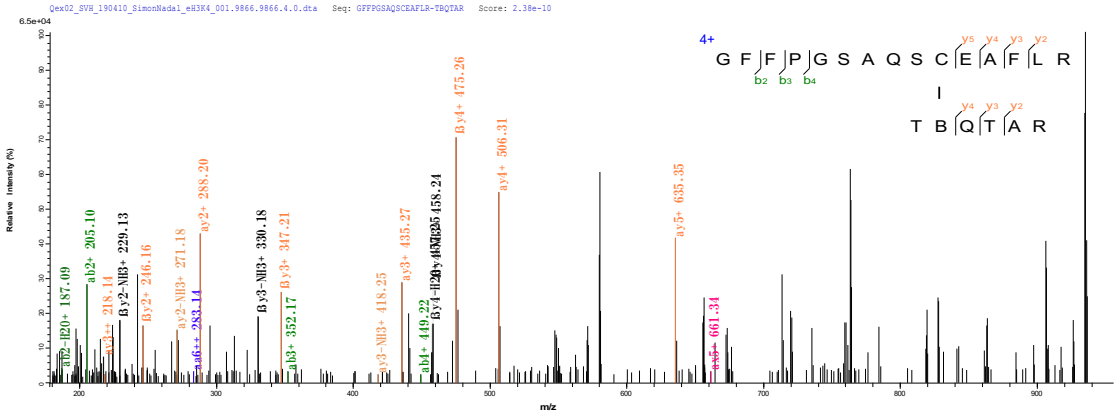
Sorted by table entry.

As an empirical rule, an e-value higher than e-03 indicates a potential identification, higher than e-06 a reliable identification and higher than e-10 a very good identification.

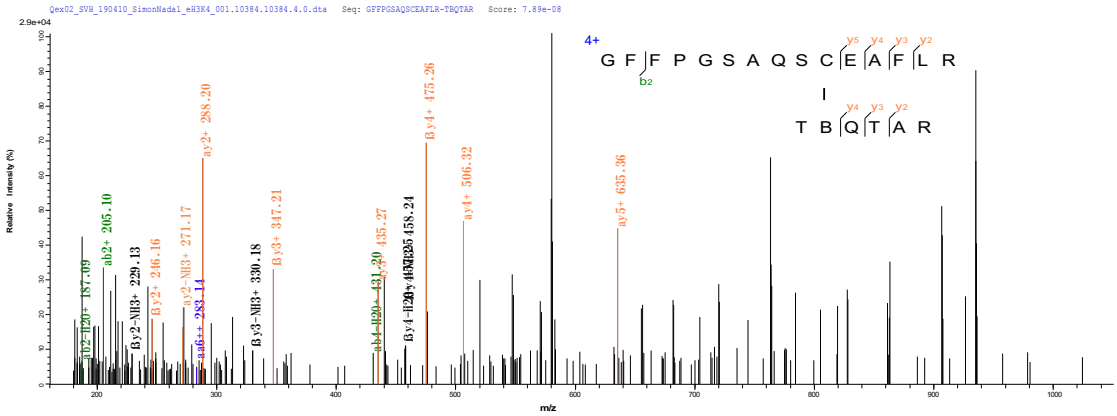
Entry 1 – (Spectrum# 3948) – Evalue: 1.970754e-010 ; Score: 4.939159e-007. KDM4A crosslink



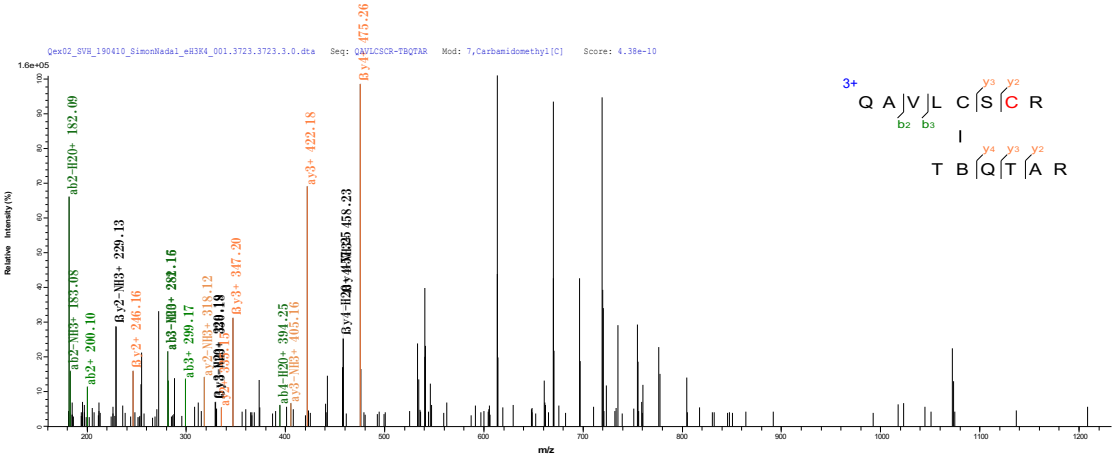
Entry2 – (Spectrum# 9866) – Evalue: 2.375866e-010 ; Score: 7.570014e-007. KDM4A crosslink



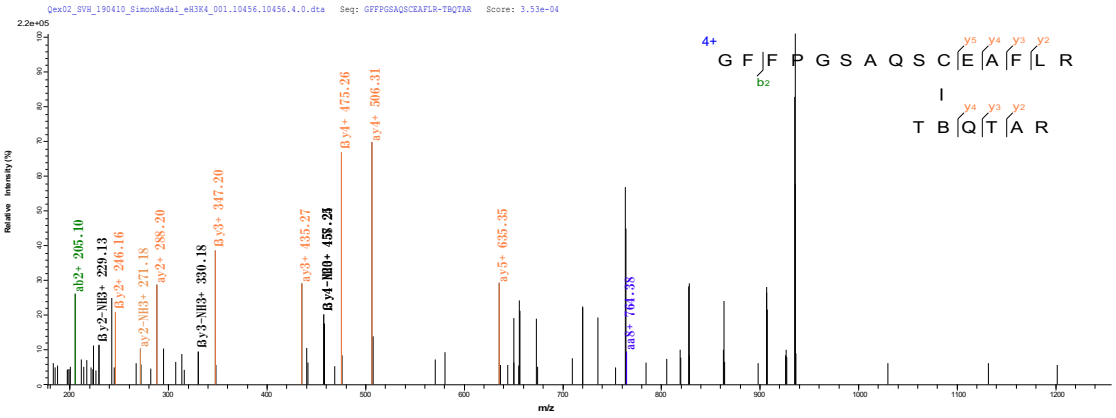
Entry 3 – (#Spectrum 10384) – Evalue: 7.891950e-008 ; Score: 2.574378e-006. KDM4A crosslink



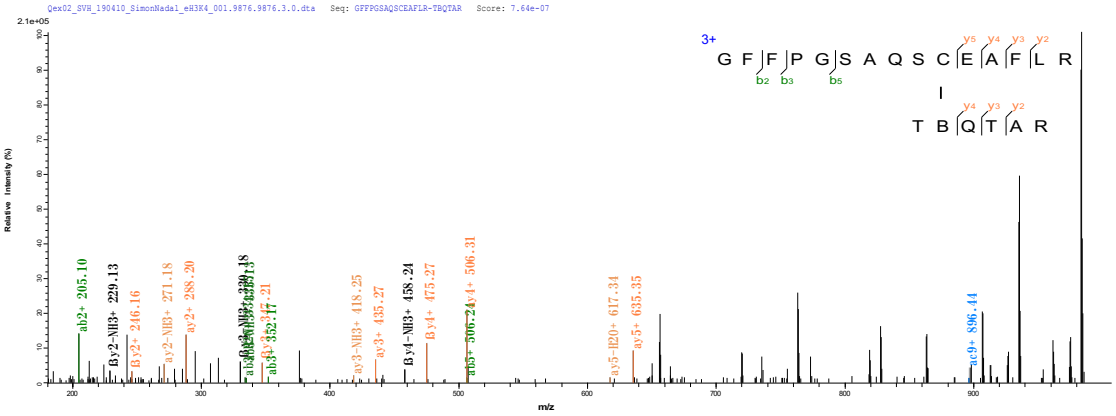
Entry 4 – (Spectrum# 3723) – Evalue: 4.383559e-010 ; Score: 6.157091e-006. KDM4A Crosslink



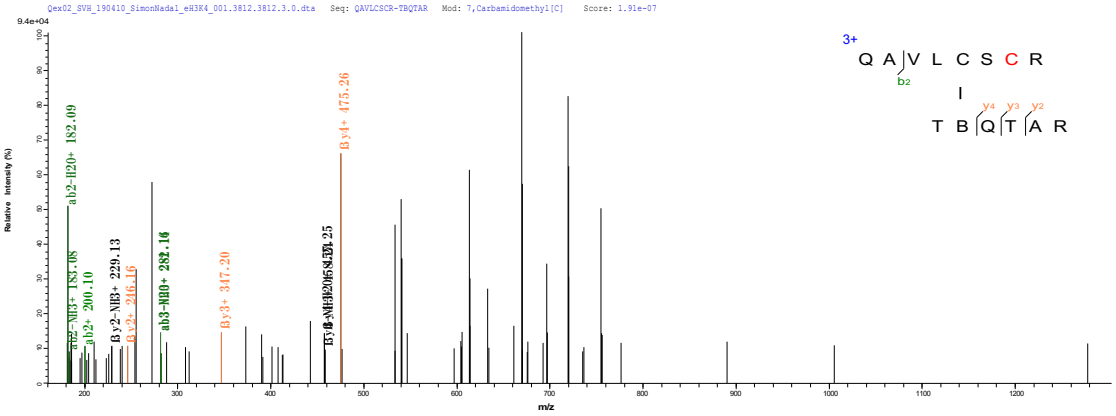
Entry 5 – (Spectrum# 10456) – Evalue: 3.534963e-004 ; Score 9.439482e-006. KDM4A Crosslink



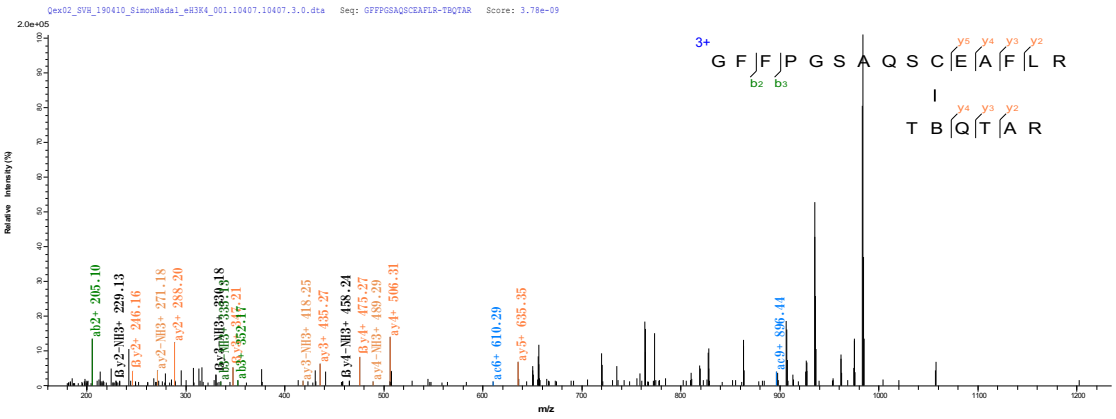
Entry 6 – (Spectrum# 9876) – Evalue: 7.639838e-007 ; Score: 2.297456e-005. KDM4A crosslink



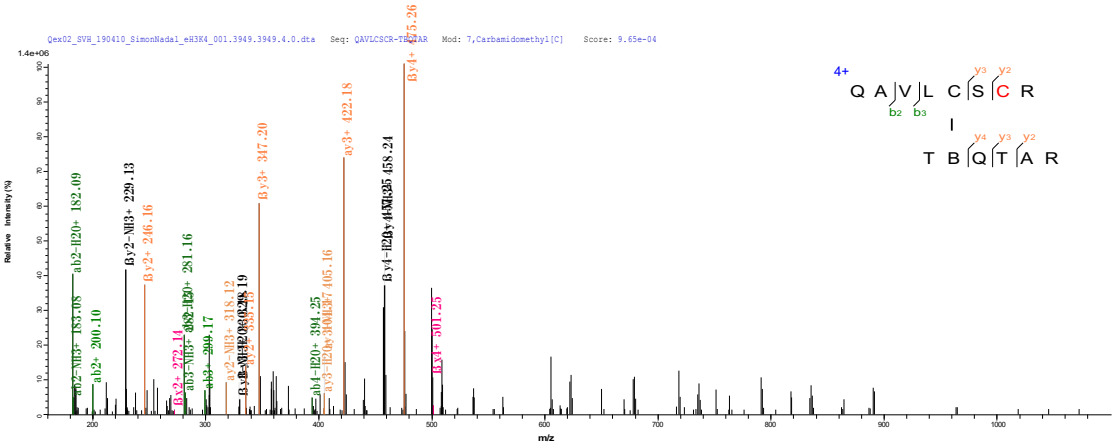
Entry 7 – (Spectrum# 3812) – Evalue: 1.906273e-007 ; Score: 2.806947e-005. KDM4A Crosslink



Entry 8 – (Spectrum# 10407) – Evalue: 3.783959e-009 ; Score: 4.808414e-005. KDM4A Crosslink



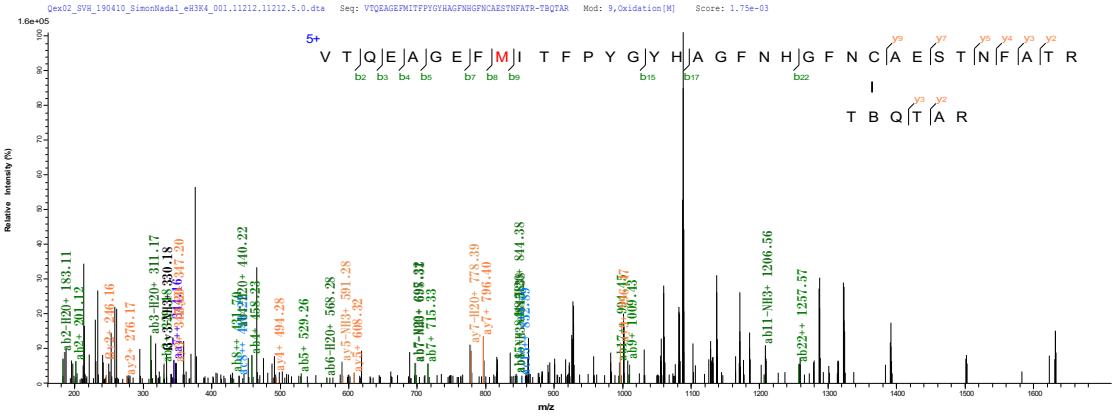
Entry 9 – (Spectrum# 3949) – Evalue: 9.646866e-004 ; Score: 8.932771e-004. KDM4A crosslink



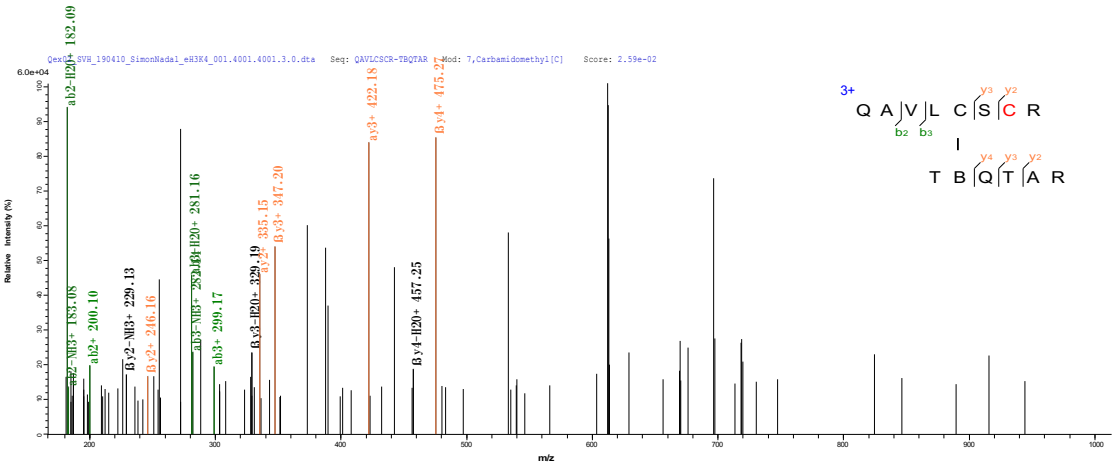
Entry 10 – no fragments of α-arm identified.

Entry 11 – low E-value.

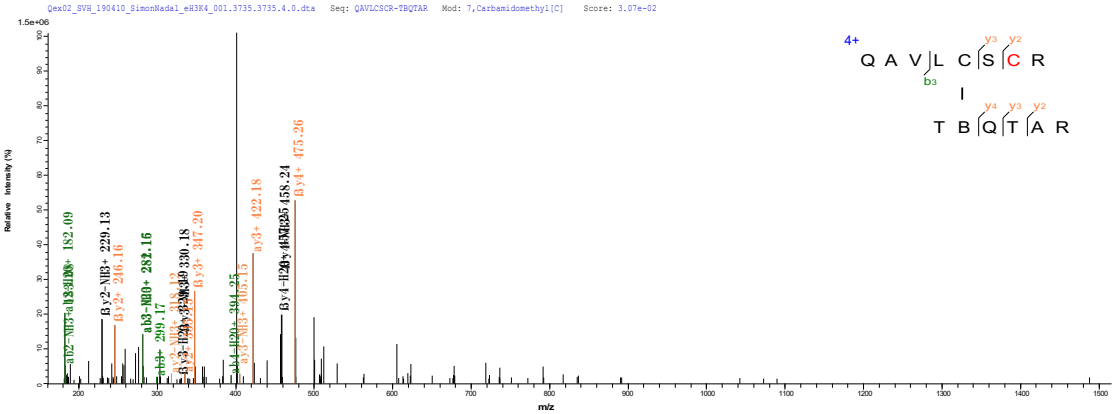
Entry 12 – (Spectrum# 11212) – Evalue: 1.750921e-003 ; Score: 8.761483e-003. KDM4A crosslink



Entry 13 – (Spectrum# 4001) – Evalue: 2.586534e-002 ; Score: 9.075319e-003. KDM4A crosslink

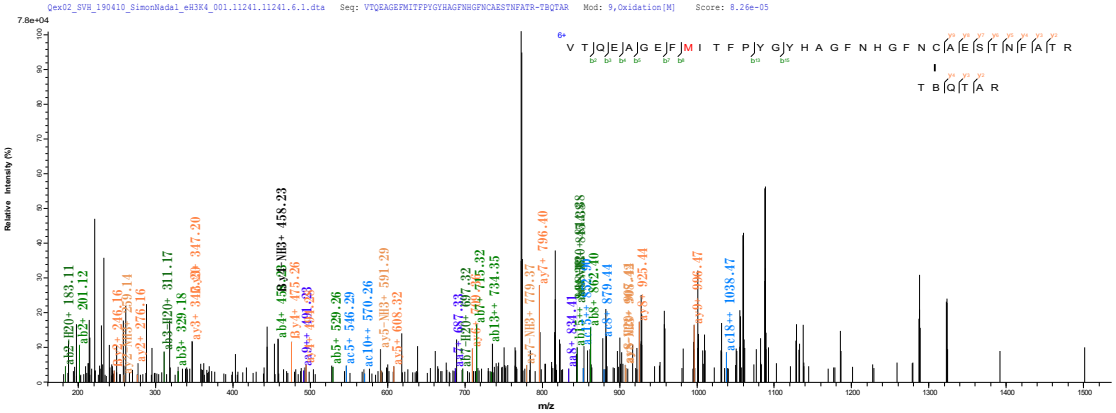


Entry 14 – (Spectrum# 3735) – Evalue: 3.065555e-002 ; Score: 9.291043e-003. KDM4A crosslink



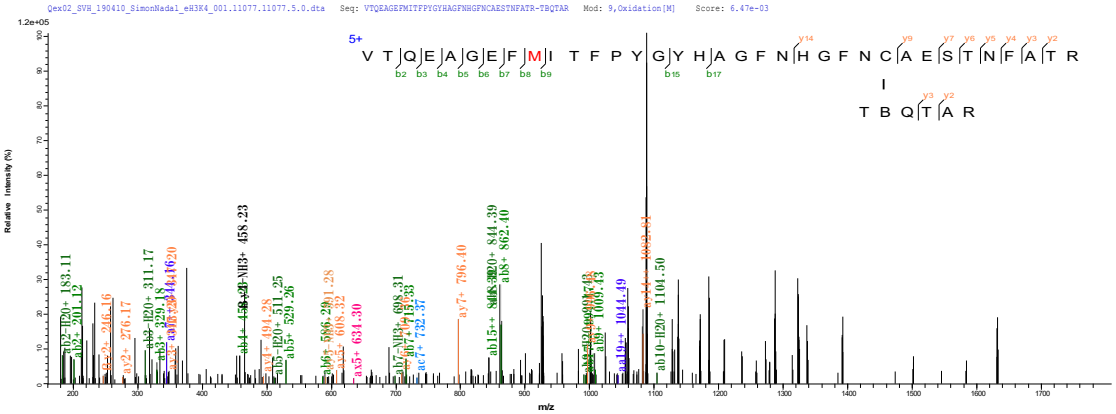
Entry 15 – low E-value.

Entry 16 – (Spectrum# 11241) – Evalue: 8.261052e-005 ; Score: 2.839517e-002. KDM4A Crosslink



Entry 17 – low E-value

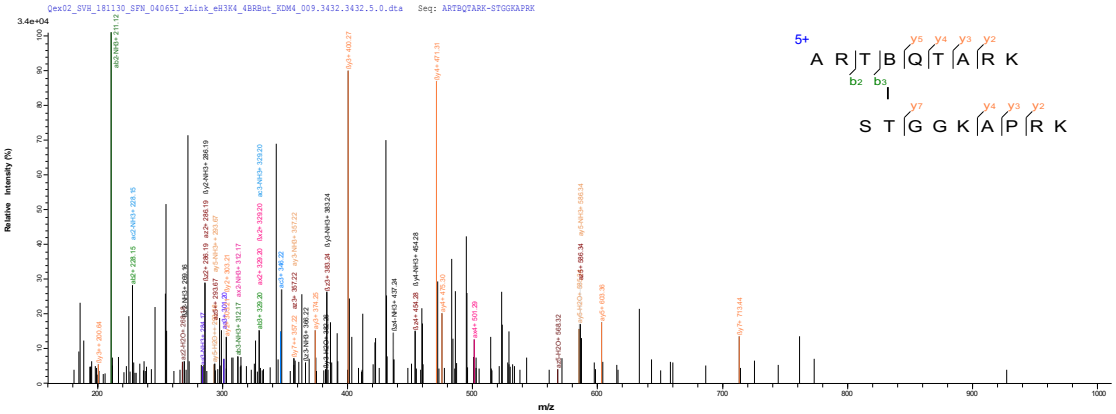
Entry 18 – (Spectrum# 11077) – Evalue: 6.467943e-003 ; Score: 6.478454e-002. KDM4A Crosslink



Further entries above e-value 1.0 e-03 are not listed.

Results of LysC in-solution digest, 20ppm search (20ppm Frag, 10 ppm filter), 1% global FDR

Inter-histone crosslink:

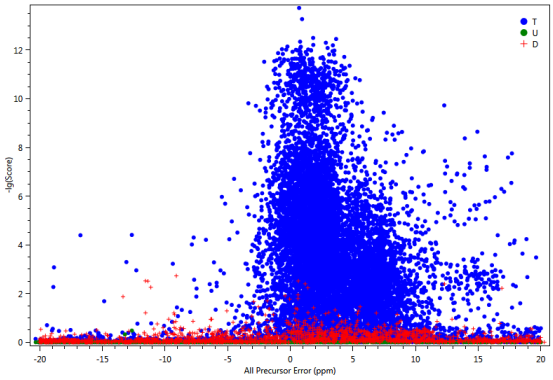


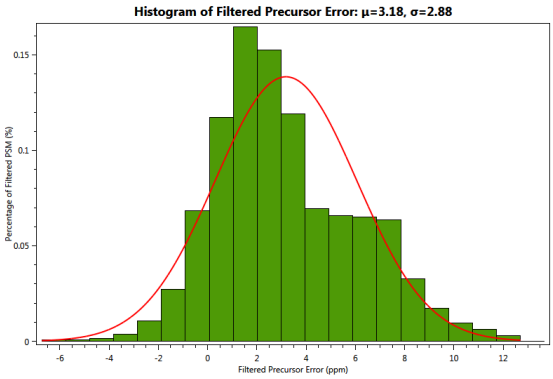
eH3K9_KDM4A crosslink

Results of tryptic in-solution digest, 20ppm search (20ppm Frag, 10 ppm filter), 1% global FDR

Table 1 Identification of Cross-Linked Results Under 1% FDR Control.

#(Spectra)	#(Peptide Pairs)	#(Linked Sites)
16	9	6



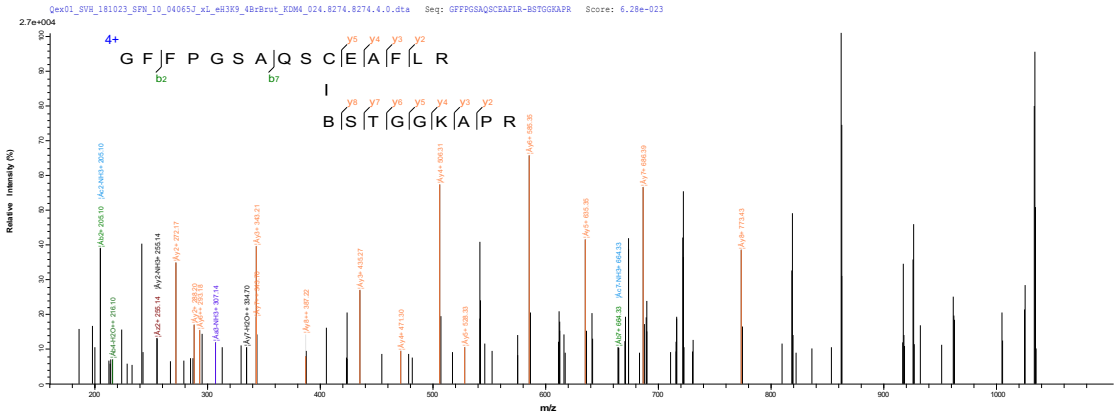


Order	Title	Charge	Precursor_Mass	Peptide	Peptide_Type	Linker	Peptide_Mass	Modifications	Evalue	Score
1	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.8274.8274.4.0.dta	4	2514.26	GFFPGSAQSCEAFLR(10)-BSTGGKAPR(1)	Cross-Linked	4BrBut	2514.26	null	6.277745e-023	3.721426e-006
2	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.9086.9086.4.0.dta	4	2.190073e+003	GFFPGSAQSCEAFLR(10)-BSTGGK(1)	Cross-Linked	4BrBut	2.190065e+003	null	7.957494e-010	3.889212e-006
3	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.8232.8232.4.0.dta	4	2.514260e+003	GFFPGSAQSCEAFLR(10)-BSTGGKAPR(1)	Cross-Linked	4BrBut	2.514256e+003	null	4.085251e-019	4.965458e-006
4	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.8271.8271.5.0.dta	5	2.514258e+003	GFFPGSAQSCEAFLR(10)-BSTGGKAPR(1)	Cross-Linked	4BrBut	2.514256e+003	null	6.058832e-014	7.539856e-006
5	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.3896.3896.4.0.dta	4	1.833941e+003	BSTGGKAPR(1)-QAVLCSCR(5)	Cross-Linked	4BrBut	1.833942e+003	Carbamidomethyl[C](19)	9.967531e-007	8.993495e-006
6	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.3972.3972.4.0.dta	4	1.833944e+003	BSTGGKAPR(1)-QAVLCSCR(6)	Cross-Linked	4BrBut	1.833942e+003	Carbamidomethyl[C](17)	1.491346e-005	3.483864e-004
7	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.8222.8222.5.0.dta	5	2.514258e+003	GFFPGSAQSCEAFLR(10)-BSTGGKAPR(1)	Cross-Linked	4BrBut	2.514256e+003	null	1.528679e-012	3.729139e-004
8	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.8331.8331.4.0.dta	4	2.514261e+003	GFFPGSAQSCEAFLR(10)-BSTGGKAPR(1)	Cross-Linked	4BrBut	2.514256e+003	null	4.064013e-010	1.320267e-003
9	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.3981.3981.5.0.dta	5	1.833939e+003	BSTGGKAPR(1)-QAVLCSCR(6)	Cross-Linked	4BrBut	1.833942e+003	Carbamidomethyl[C](17)	2.423092e-003	6.781247e-003
10	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.3633.3633.4.0.dta	4	1.637849e+003	QAVLCSCRK(7)-BSTGGK(1)	Cross-Linked	4BrBut	1.637846e+003	Carbamidomethyl[C](5)	1.054979e-002	7.243578e-003
11	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.3669.3669.5.0.dta	5	1.962040e+003	QAVLCSCRK(5)-BSTGGKAPR(1)	Cross-Linked	4BrBut	1.962037e+003	Carbamidomethyl[C](7)	1.061905e-002	1.066675e-002
12	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.9075.9075.3.0.dta	3	2.190065e+003	GFFPGSAQSCEAFLR(9)-BSTGGK(1)	Cross-Linked	4BrBut	2.190065e+003	null	7.630947e-004	1.345166e-002
13	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.8439.8439.5.0.dta	5	2.514259e+003	GFFPGSAQSCEAFLR(10)-BSTGGKAPR(1)	Cross-Linked	4BrBut	2.514256e+003	null	8.951725e-004	1.503848e-002
14	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.3638.3638.3.0.dta	3	1.637850e+003	QAVLCSCRK(7)-BSTGGK(1)	Cross-Linked	4BrBut	1.637846e+003	Carbamidomethyl[C](5)	6.845501e-002	9.194286e-002
15	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.3705.3705.5.0.dta	5	1.962039e+003	QAVLCSCRK(6)-BSTGGKAPR(1)	Cross-Linked	4BrBut	1.962037e+003	Carbamidomethyl[C](5)	1.344572e-001	2.606775e-001
16	Qex01_SVH_181023_SFN_10_040651_xL_eH3K9_4BrBrut_KDM4_024.3991.3991.4.0.dta	4	1.509746e+003	QAVLCSCR(1)-BSTGGK(1)	Cross-Linked	4BrBut	1.509751e+003	Carbamidomethyl[C](7)	7.170196e-001	3.652874e-001

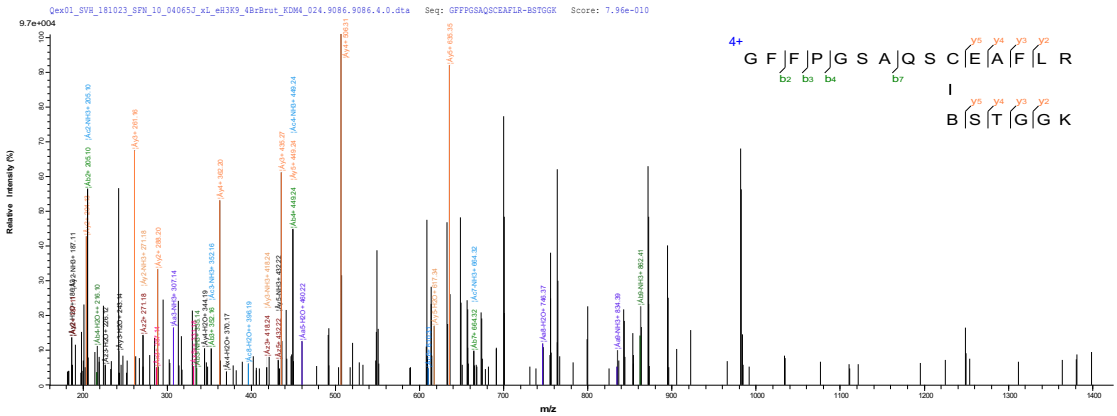
Sorted by table entry.

As an empirical rule, an e-value higher than e-03 indicates a potential identification, higher than e-06 a reliable identification and higher than e-10 a very good identification.

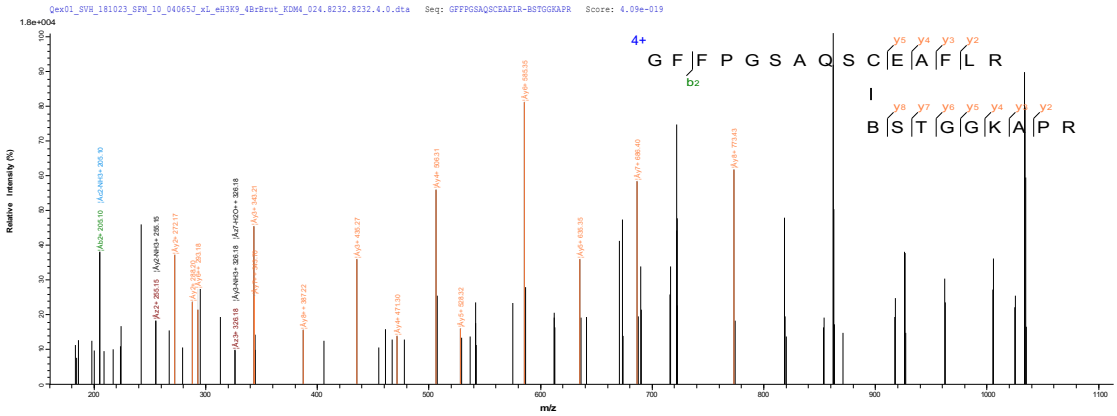
Entry 1 (8274) – Evalue: 6.277745e-023 ; Score: 3.721426e-006. KDM4A crosslink



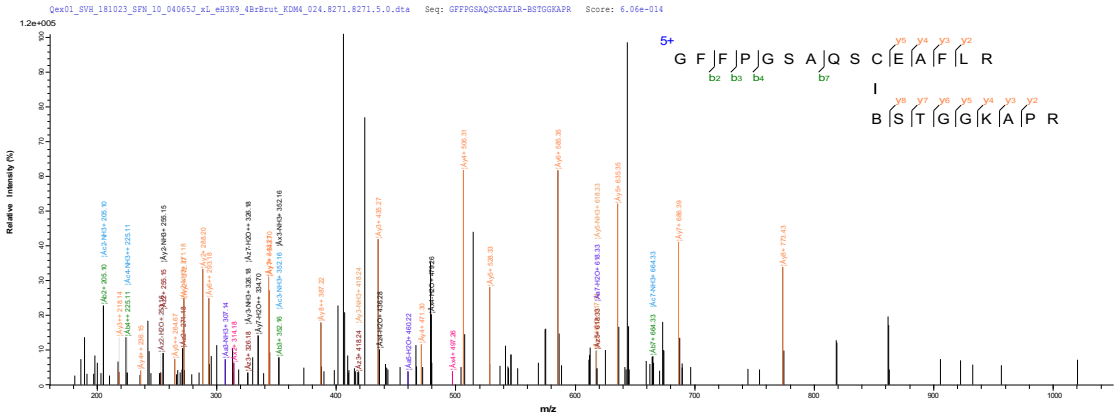
Entry 2 (9086) – Evalue: 7.957494e-010 ; Score: 3.889212e-006. KDM4A crosslink



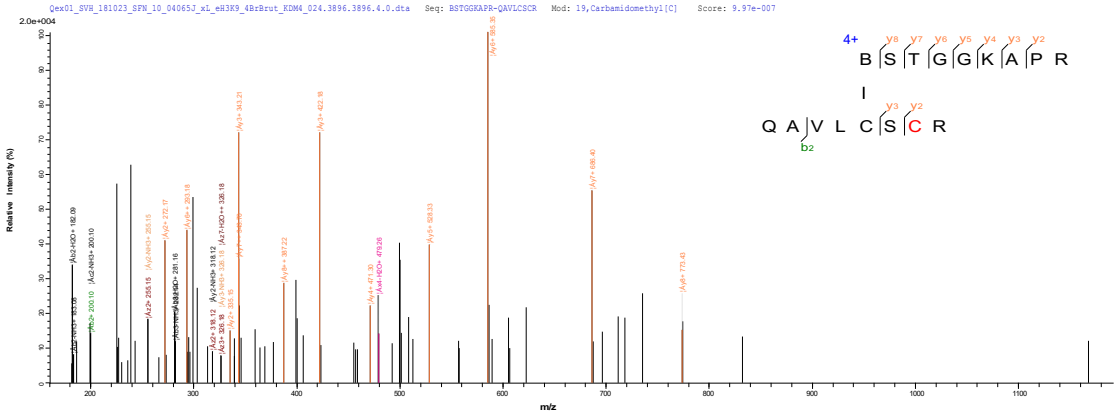
Entry 3 - (8232) – Evalue: 4.085251e-019 ; Score: 4.965458e-006. KDM4A crosslink



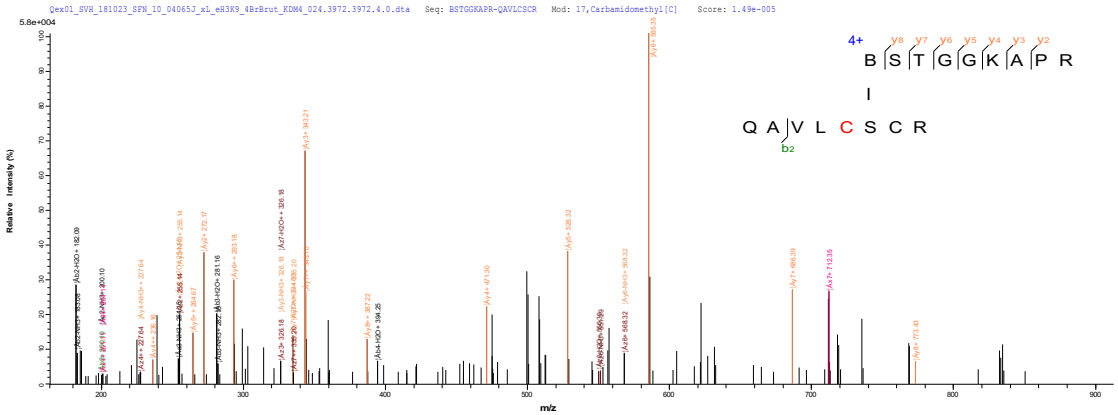
Entry 4 – (8271) – Evalue: 6.058832e-014 ; Score: 7.539856e-006. KDM4A crosslink



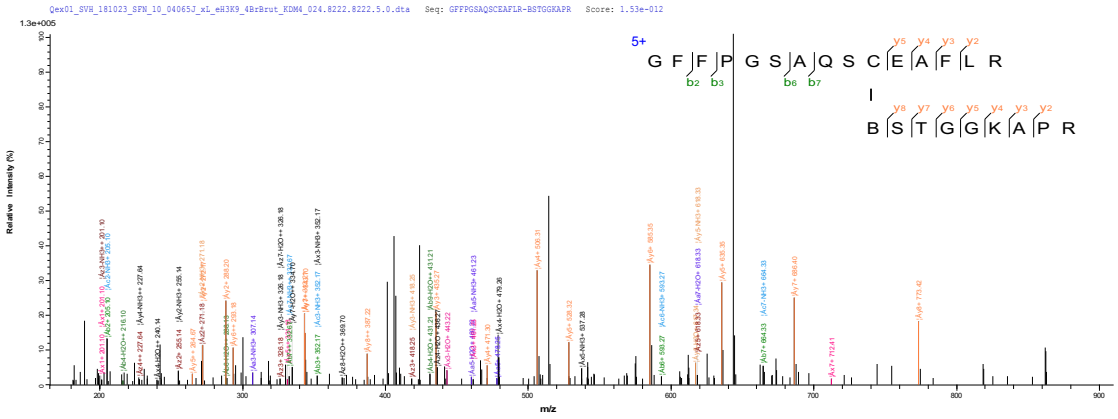
Entry 5 (3896) – Evalue: 9.967531e-007 ; Score: 8.993495e-006. KDM4A crosslink



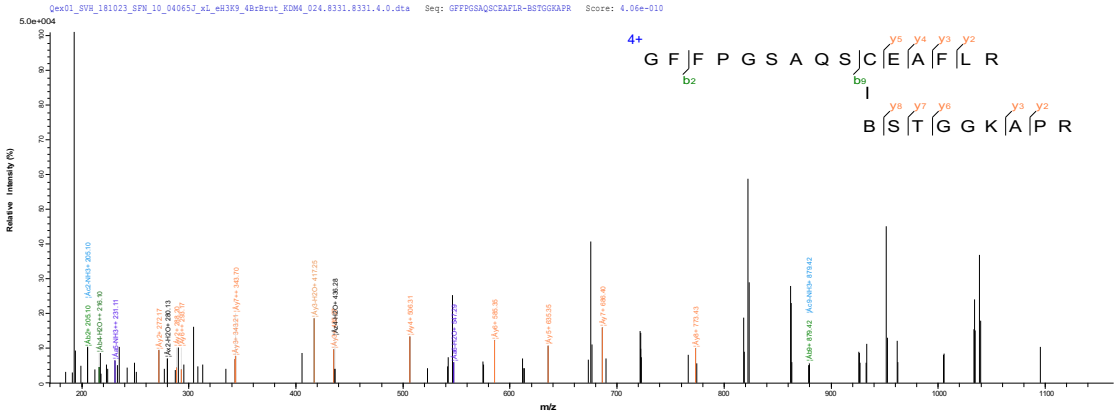
Entry 6 – (3972) – Evalue: 1.491346e-005 ; Score: 3.483864e-004. KDM4A crosslink



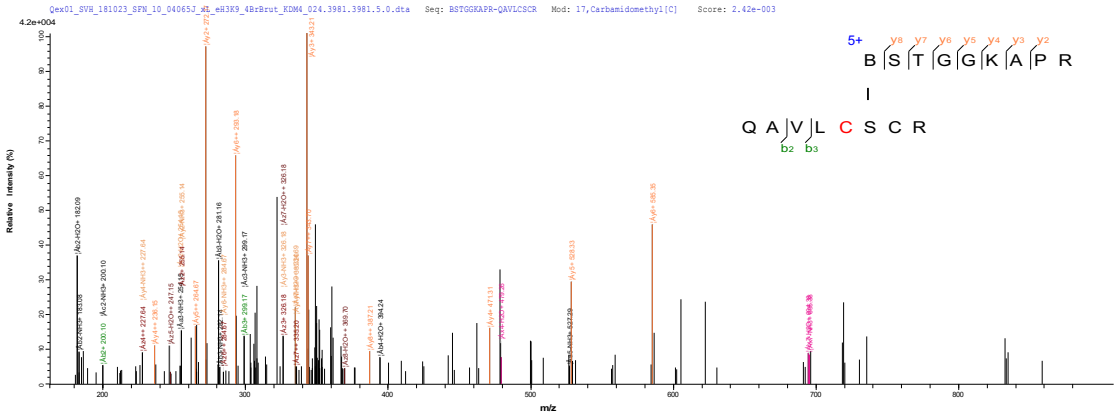
Entry 7 (8222) – Evalue: 1.528679e-012 ; Score: 3.729139e-004. KDM4A crosslink



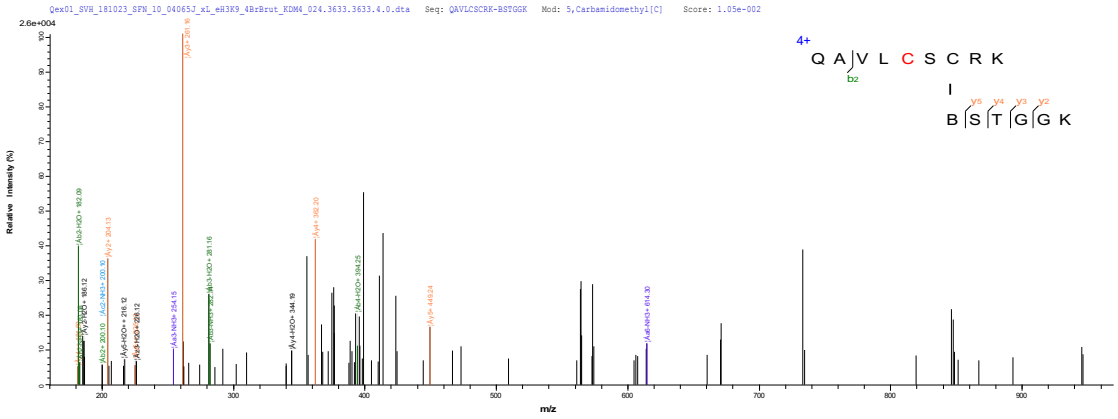
Entry 8 – (8331) – Evalue: 4.064013e-010 ; Score: 1.320267e-003. KDM4A crosslink



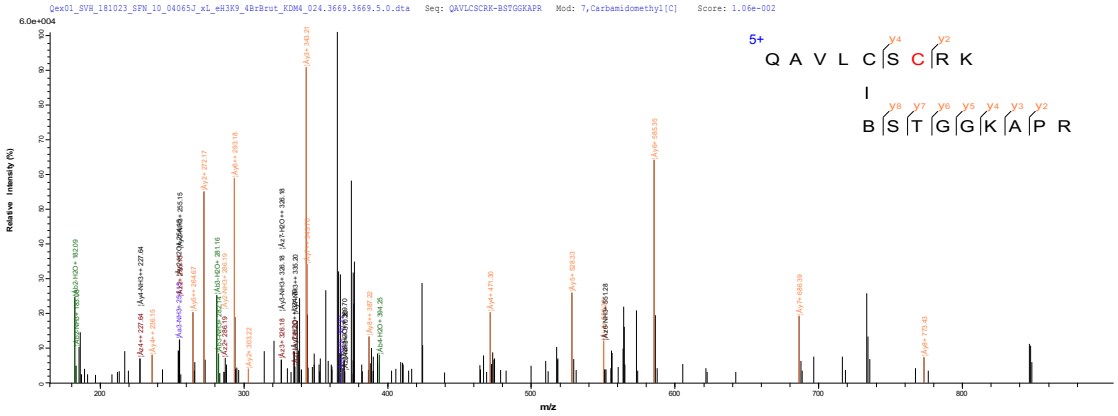
Entry 9 – (3981) – Evalue: 2.423092e-003 ; Score: 6.781247e-003. KDM4A crosslink



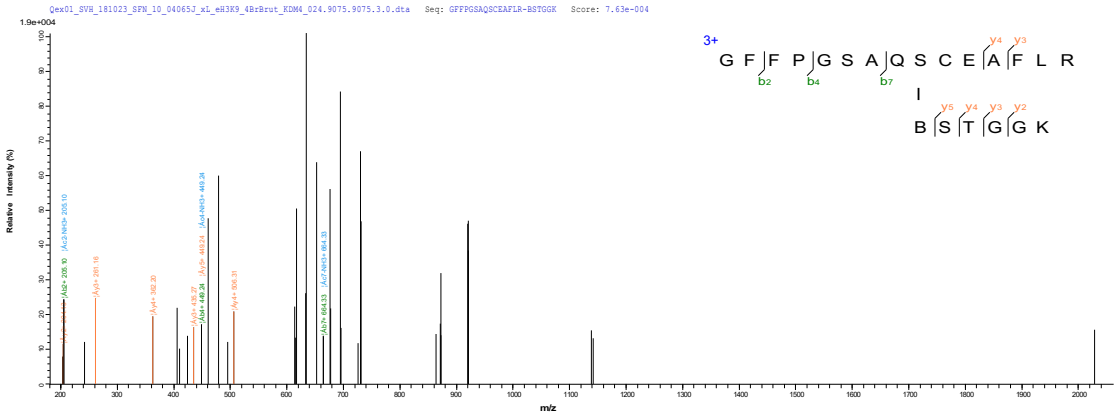
Entry 10 – (3633) – Evalue: 1.054979e-002 ; Score: 7.243578e-003. KDM4A crosslink



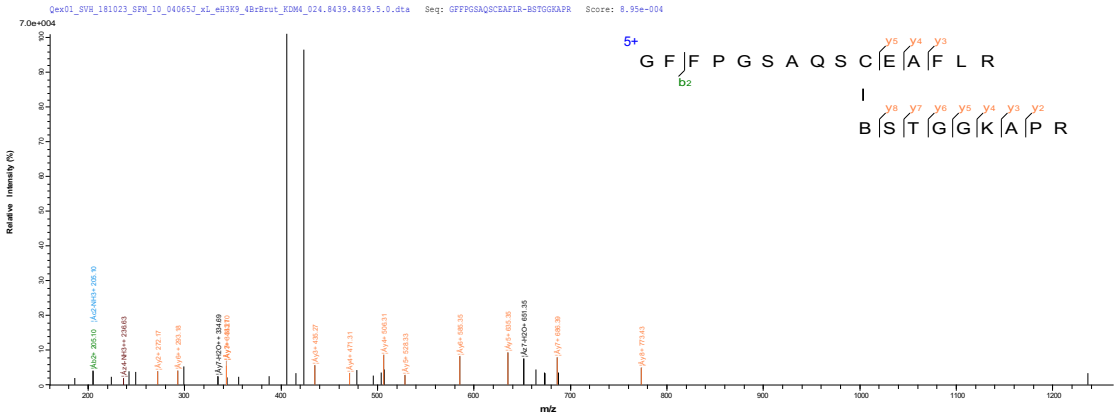
Entry 11 - (3669) – Evalue: 1.061905e-002 ; Score: 1.066675e-002. KDM4A crosslink



Entry 12 (9075) – Evalue: 7.630947e-004 ; Score: 1.345166e-002. KDM4A crosslink



Entry 13 (8439) – Evalue: 8.951725e-004 ; Score: 1.503848e-002. KDM4A crosslink



Further entries above e-value 1.0 e-03 are not listed.

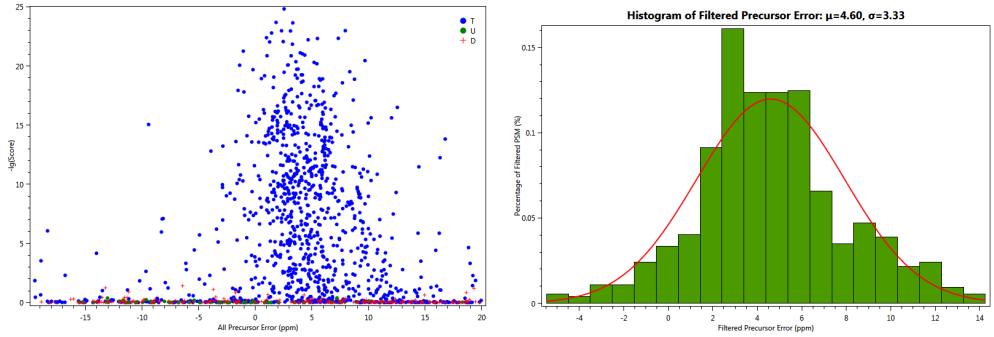
eH3K27_KDM4A crosslink

Results of trypstic in-gel digest, 20ppm search (20ppm Frag, 10 ppm filter), 1% global FDR

Table 1 Identification of Cross-Linked Results Under 1% FDR Control.

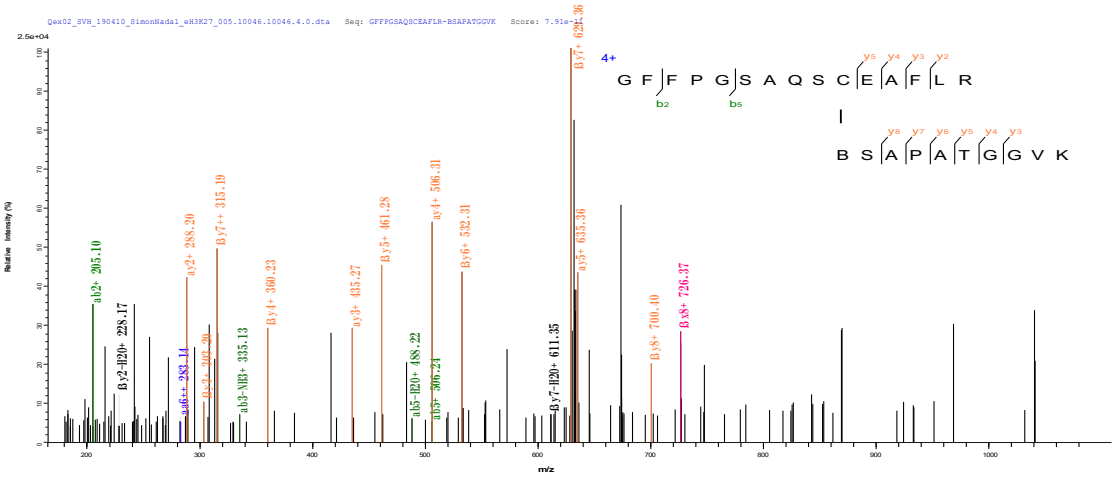
#(Spectra)	#(Peptide Pairs)	#(Linked Sites)
12	9	7

Figure 2 All Precursor Error Distribution.

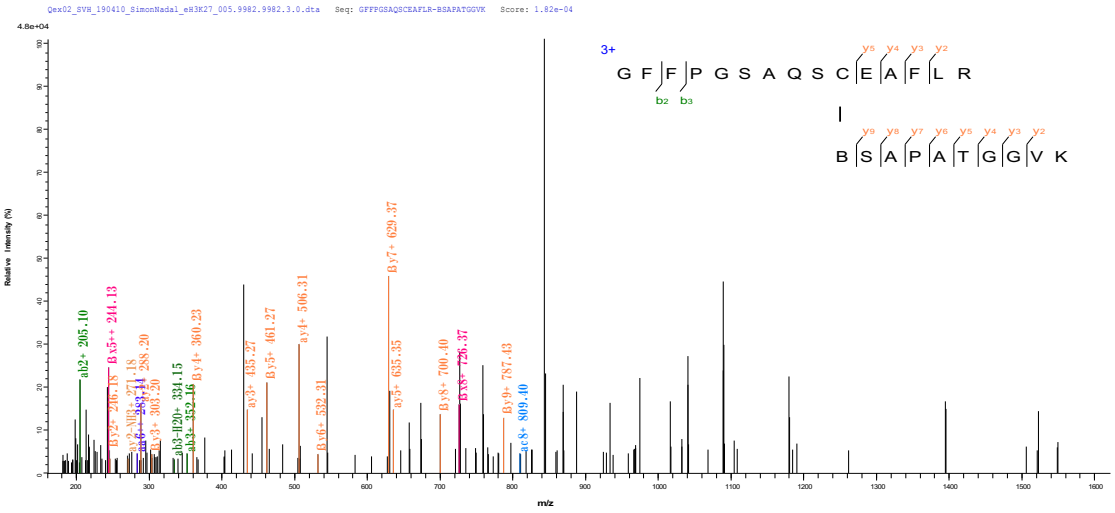


Order	Title	Charge	Precursor_Mass	Peptide	Peptide_Type	Linker	Peptide_Mass	Modifications	Evalue	Score
1	Qex02_SVH_190410_SimonNadal_eH3K27_005.10046.10046.4.0.dta	4	2528.27	GFFPGSAQSCEAFLR(10)-BSAPATGGVK(1)	Cross-Linked	4BrBut	2528.26	null	7.911960e-011	5.733951e-006
2	Qex02_SVH_190410_SimonNadal_eH3K27_005.9982.9982.3.0.dta	3	2.528266e+003	GFFPGSAQSCEAFLR(10)-BSAPATGGVK(1)	Cross-Linked	4BrBut	2.528260e+003	null	1.816204e-004	4.124760e-004
3	Qex02_SVH_190410_SimonNadal_eH3K27_005.3836.3836.3.0.dta	3	2.179123e+003	IANSDKYCTPR(8)-BSAPATGGVK(1)	Cross-Linked	4BrBut	2.179117e+003	null	1.000000e+000	3.155650e-003
4	Qex02_SVH_190410_SimonNadal_eH3K27_005.9969.9969.4.0.dta	4	2.528269e+003	GFFPGSAQSCEAFLR(9)-BSAPATGGVK(1)	Cross-Linked	4BrBut	2.528260e+003	null	2.795416e-008	3.332581e-003
5	Qex02_SVH_190410_SimonNadal_eH3K27_005.9024.9024.4.0.dta	4	2.826443e+003	GFFPGSAQSCEAFLR(9)-AARBSAPATGGVK(4)	Cross-Linked	4BrBut	2.826435e+003	null	9.694642e-010	5.414799e-003
6	Qex02_SVH_190410_SimonNadal_eH3K27_005.9027.9027.5.0.dta	5	2.826443e+003	GFFPGSAQSCEAFLR(10)-AARBSAPATGGVK(4)	Cross-Linked	4BrBut	2.826435e+003	null	4.360045e-006	6.801347e-003
7	Qex02_SVH_190410_SimonNadal_eH3K27_005.11300.11300.4.1.dta	4	2.571286e+003	GFFPGSAQSCEAFLR(10)-BSAPATGGVK(1)	Cross-Linked	4BrBut	2.571266e+003	Carbamyl[AmyN-term](0)	9.651211e-002	3.843255e-002
8	Qex02_SVH_190410_SimonNadal_eH3K27_005.7049.7049.4.0.dta	4	2.492237e+003	IMTFYPTMEEFR(7)-BSAPATGGVK(1)	Cross-Linked	4BrBut	2.492220e+003	Oxidation[M](2)	3.818307e-001	9.714301e-002
9	Qex02_SVH_190410_SimonNadal_eH3K27_005.7719.7719.4.0.dta	4	2.996481e+003	IMTFYPTMEEFRNFSR(3)-BSAPATGGVK(1)	Cross-Linked	4BrBut	2.996464e+003	Oxidation[M](8)	4.306393e-001	1.805050e-001
10	Qex02_SVH_190410_SimonNadal_eH3K27_005.9037.9037.3.1.dta	3	2.826440e+003	GFFPGSAQSCEAFLR(10)-AARBSAPATGGVK(4)	Cross-Linked	4BrBut	2.826435e+003	null	3.599128e-004	2.558521e-001
11	Qex02_SVH_190410_SimonNadal_eH3K27_005.3885.3885.4.3.dta	4	2.146140e+003	AARBSAPATGGVK(4)-QAVLCSCR(1)	Cross-Linked	4BrBut	2.146122e+003	Carbamidomethyl[C](23)	2.965600e-002	3.120087e-001
12	Qex02_SVH_190410_SimonNadal_eH3K27_005.11587.11587.3.0.dta	3	2.322180e+003	IANSDKYCTPR(5)-BSAPATGGVK(1)	Cross-Linked	4BrBut	2.322151e+003	Carbamyl[K](6);Carbamidomethyl[C](8);Carbamyl[K](24)	4.228145e-002	7.912126e-001

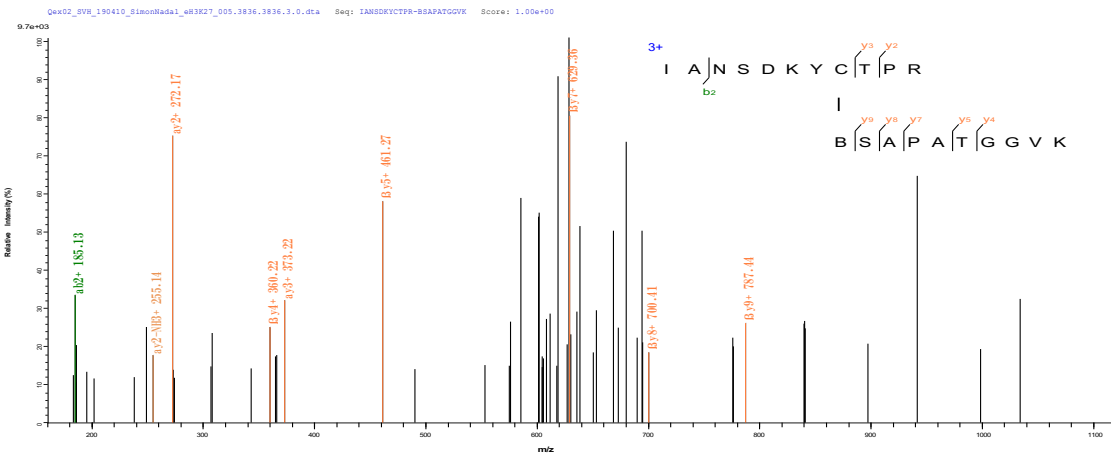
Entry 1 (Spectrum# 9045) – Evalue: 7.911960e-011; Score: 5.733951e-006. KDM4A crosslink



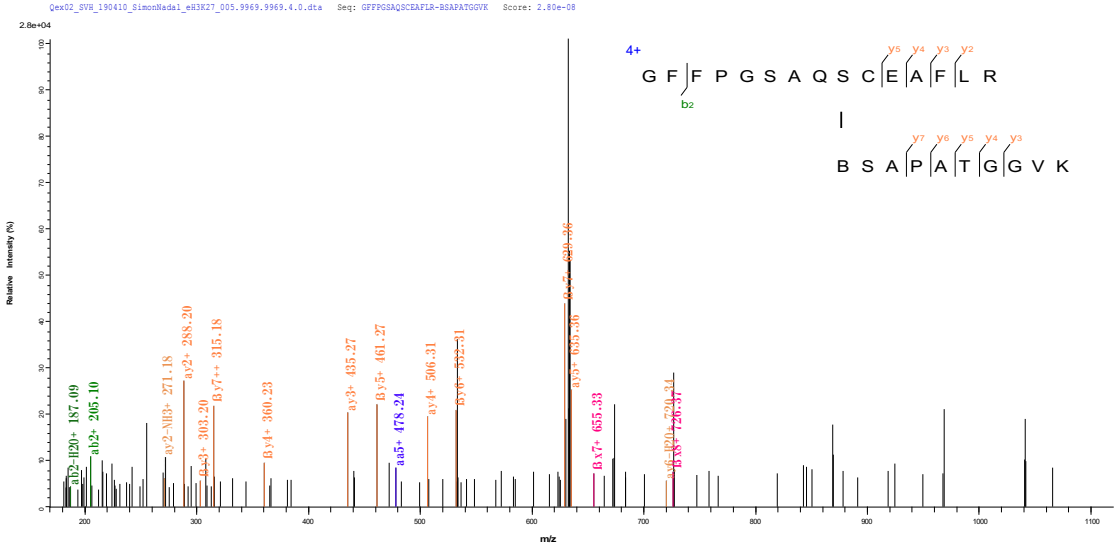
Entry 2 (Spectrum# 9982) – Evalue: 1.816204e-004; Score: 4.124760e-004. KDM4A crosslink.



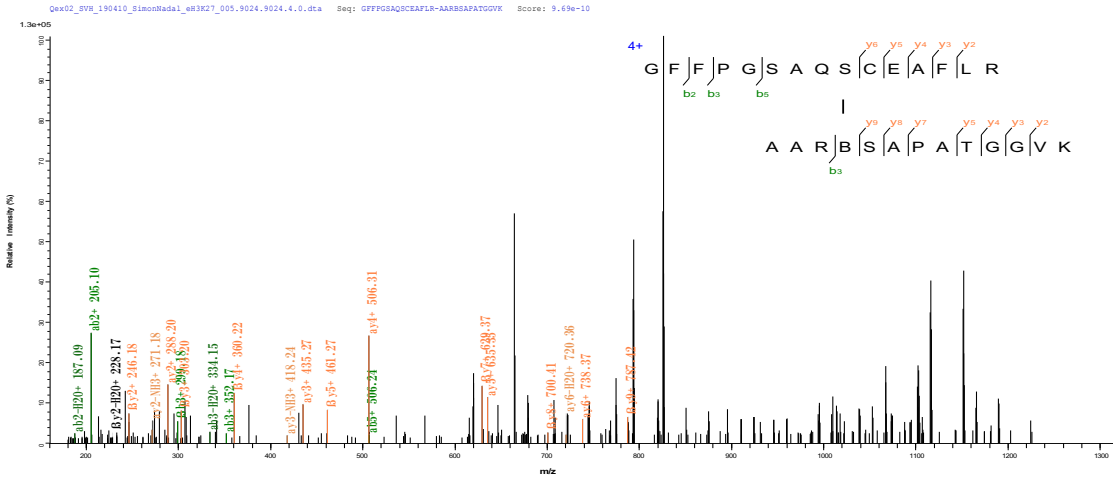
Entry 3 (Spectrum# 3836) – Evalue: none ; Score: 3.155650e-003. KDM4A crosslink.



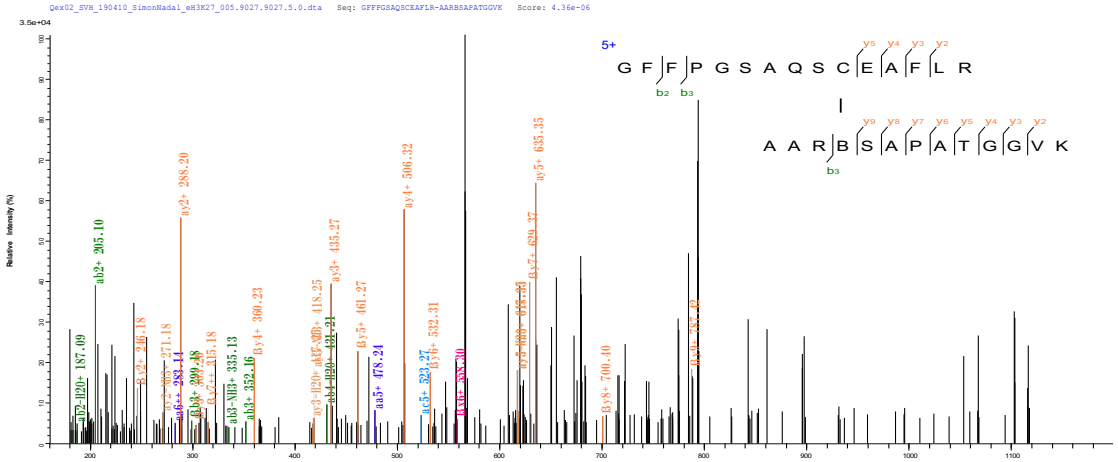
Entry 4 (Spectrum# 9969) – Evalue: 2.795416e-008 ; Score: 3.332581e-003. KDM4A crosslink.



Entry 5 (Spectrum# 9024) – Evalue: 9.694642e-010 ; Score: 5.414799e-003. KDM4A crosslink.



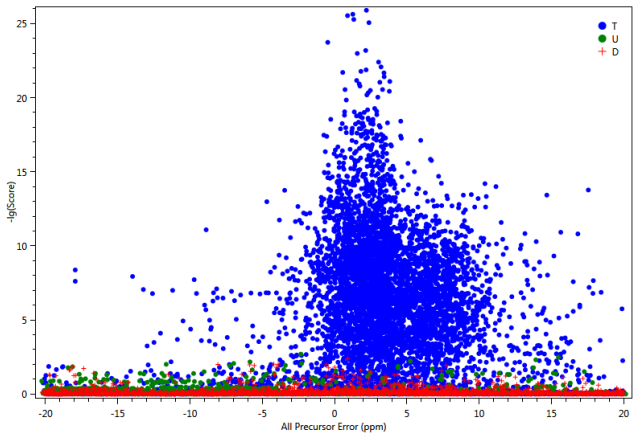
Entry 6 (Spectrum# 9027) – Evalue: 4.360045e-006 ; Score: 6.801347e-003. KDM4A crosslink.

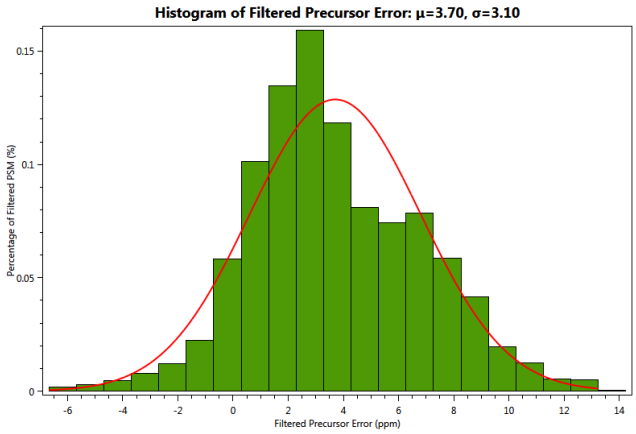


Results of trypstic in-solution digest, 20ppm search (20ppm Frag, 10 ppm filter), 1% global FDR

Table 1 Identification of Cross-Linked Results Under 1% FDR Control.

#(Spectra)	#(Peptide Pairs)	#(Linked Sites)
12	12	10



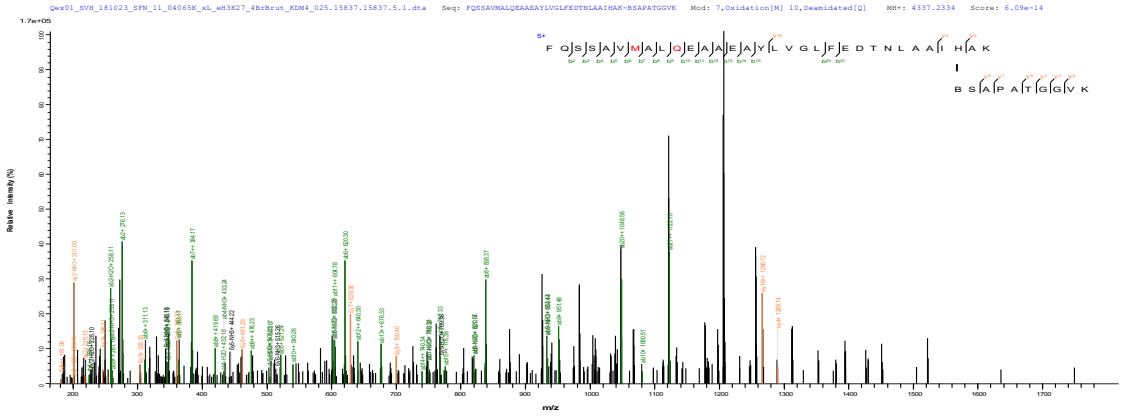


Order	Title	Charge	Precursor_Mass	Peptide	Peptide_Type	Linker	Peptide_Mass	Modifications	Evalue	Score
1	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.15837.15837.5.1.dta	5	4337.23	FQSSAVMALQEAAEAYLVGLFEDTNLAAIHAK(30)-BSAPATGGVK(1)	Cross-Linked	4BrBut	4337.2	Oxidation[M](7);Deamidated[Q](10)	6.094873e-014	8.996254e-003
2	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.15528.15528.5.1.dta	5	4.336204e+003	FQSSAVMALQEAAEAYLVGLFEDTNLAAIHAK(16)-BSAPATGGVK(1)	Cross-Linked	4BrBut	4.336217e+003	Oxidation[M](7)	4.042429e-008	1.410139e-002
3	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.15641.15641.5.0.dta	5	4.336234e+003	FQSSAVMALQEAAEAYLVGLFEDTNLAAIHAK(22)-BSAPATGGVK(1)	Cross-Linked	4BrBut	4.336217e+003	Oxidation[M](7)	8.474083e-008	5.953388e-002
4	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.16730.16730.7.0.dta	7	5.266750e+003	FQSSAVMALQEAAEAYLVGLFEDTNLAAIHAK(3)-AARBSAPATGGVKKPHR(4)	Cross-Linked	4BrBut	5.266706e+003	Deamidated[Q](2);Carbamyl[K](32);Carbamyl[AnyN-term](35);Carbamyl[K](48)	2.496320e-001	9.143983e-002
5	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.16644.16644.6.0.dta	6	5.265751e+003	FQSSAVMALQEAAEAYLVGLFEDTNLAAIHAK(2)-AARBSAPATGGVKKPHR(4)	Cross-Linked	4BrBut	5.265722e+003	Carbamyl[AnyN-term](0);Carbamyl[K](32);Carbamyl[K](48)	1.835616e-001	9.187609e-002
6	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.9045.9045.4.0.dta	4	2.528270e+003	GFFPGSAQSCEAFLR(10)-BSAPATGGVK(1)	Cross-Linked	4BrBut	2.528260e+003	null	1.525090e-019	9.440228e-002
7	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.16589.16589.6.2.dta	6	5.266747e+003	FQSSAVMALQEAAEAYLVGLFEDTNLAAIHAK(4)-AARBSAPATGGVKKPHR(4)	Cross-Linked	4BrBut	5.266706e+003	Deamidated[Q](2);Carbamyl[K](32);Carbamyl[AnyN-term](35);Carbamyl[K](48)	1.000000e+000	9.716230e-002
8	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.16721.16721.6.0.dta	6	5.265757e+003	FQSSAVMALQEAAEAYLVGLFEDTNLAAIHAK(10)-AARBSAPATGGVKKPHR(4)	Cross-Linked	4BrBut	5.265722e+003	Carbamyl[K](32);Carbamyl[AnyN-term](35);Carbamyl[K](48)	1.000000e+000	9.729572e-002
9	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.15297.15297.5.1.dta	5	4.320203e+003	FQSSAVMALQEAAEAYLVGLFEDTNLAAIHAK(10)-BSAPATGGVK(1)	Cross-Linked	4BrBut	4.320222e+003	null	1.774971e-006	1.362491e-001
10	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.9463.9463.7.0.dta	7	6.949574e+003	EIAQDFKTDLRFQSSAVMALQEAAEAYLVGLFEDTNLAAIHAKR(27)-QLATKAARBSAPATGGVKK(9)	Cross-Linked	4BrBut	6.949590e+003	Oxidation[M](18);Deamidated[N](36);Carbamyl[K](43);Carbamyl[K](52);Carbamyl[K](65);Carbamyl[K](66)	1.000000e+000	1.451585e-001
11	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.15719.15719.6.5.dta	6	4.988624e+003	FQSSAVMALQEAAEAYLVGLFEDTNLAAIHAKR(32)-AARBSAPATGGVKK(4)	Cross-Linked	4BrBut	4.988605e+003	Carbamyl[AnyN-term](36);Carbamyl[K](49)	2.183911e-001	1.537704e-001
12	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.8744.8744.7.0.dta	7	6.822510e+003	EIAQDFKTDLRFQSSAVMALQEAAEAYLVGLFEDTNLAAIHAKR(11)-QLATKAARBSAPATGGVKK(9)	Cross-Linked	4BrBut	6.822541e+003	Oxidation[M](18);Deamidated[Q](21);Deamidated[N](36);Deamidated[Q](48);Carbamyl[K](66)	1.000000e+000	1.615094e-001

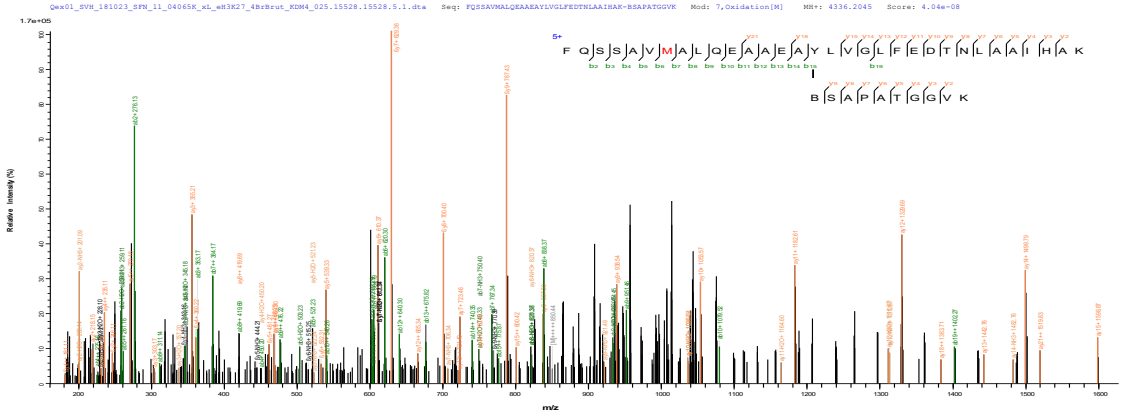
Sorted by table entry.

As an empirical rule, an e-value higher than e-03 indicates a potential identification, higher than e-06 a reliable identification and higher than e-10 a very good identification.

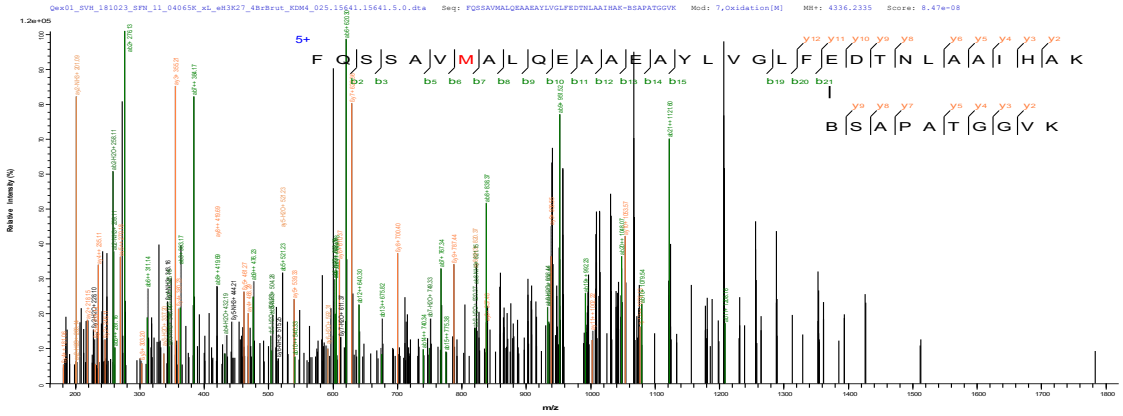
Entry 1 (15837) – Evalue: 6.094873e-014 ; Score: 8.996254e-003. Inter-Histone crosslink



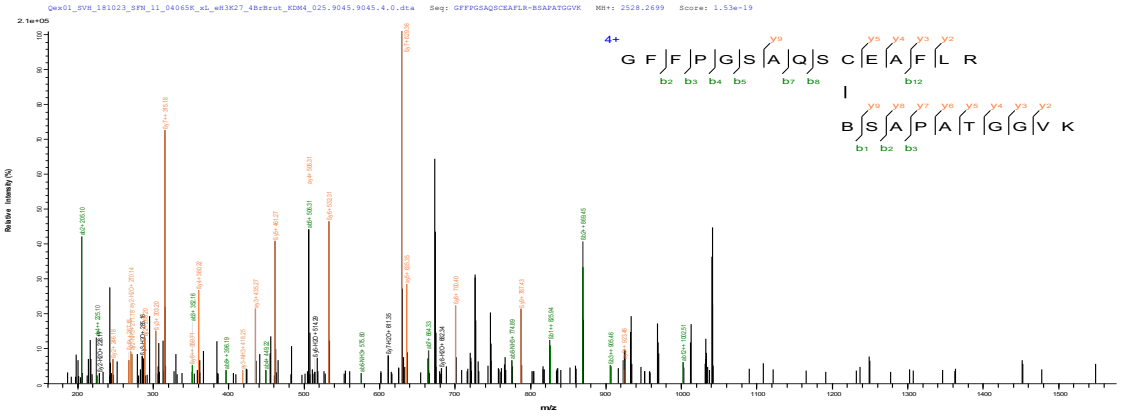
Entry 2 (15528) – Evalue: 4.042429e-008 ; Score: 1.410139e-002. Inter-Histone crosslink



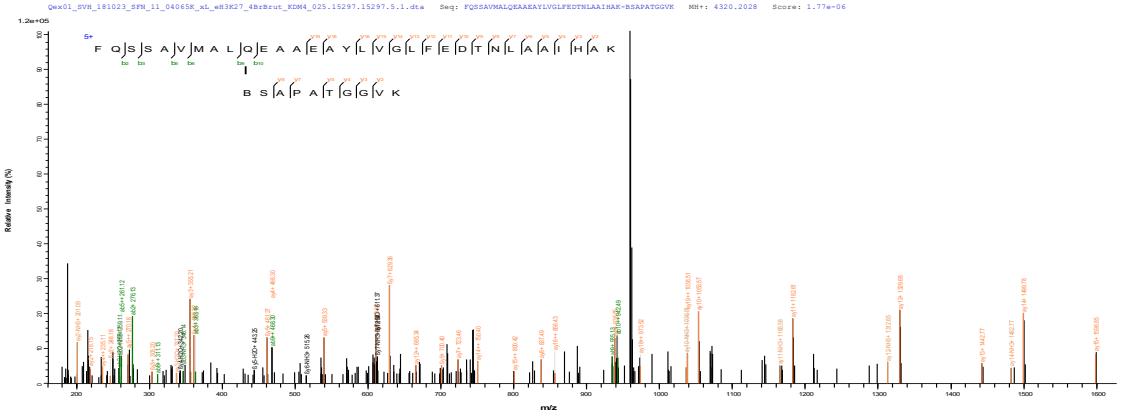
Entry 3 (15641) – Evalue: 8.474083e-008 ; Score: 5.953388e-002. Inter-Histone crosslink



Entry 6 (9045) – Evalue: 1.525090e-019 ; Score: 9.440228e-002. KDM4A crosslink



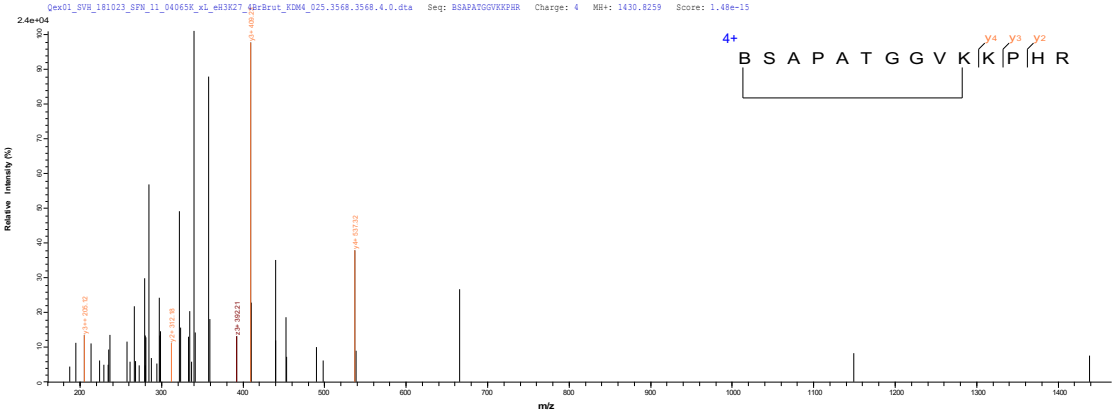
Entry 9 (15297) – Evalue: 1.774971e-006 ; Score: 1.362491e-001. Inter-Histone crosslink



Further entries above e-value 1.0 e-03 are not listed.

Loop-linked spectra (self-crosslinked by proximity inside 1 histone)

Order	Title	Charge	Precursor_Mass	Peptide	Peptide_Type	Linker	Peptide_Mass	Modifications	Evalue	Score
1	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.3602.3602.4.0.dta	4	1430.83	BSAPATGGVKKPHR(1)(2)	Loop-Linked	4BrBut	1430.82	null	6.272141e-022	3.768712e-004
2	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.3568.3568.4.0.dta	4	1.430826e+003	BSAPATGGVKKPHR(1)(2)	Loop-Linked	4BrBut	1.430823e+003	null	1.480489e-015	4.163075e-004
3	Qex01_SVH_181023_SFN_11_04065K_xL_eH3K27_4BrBrut_KDM4_025.6210.6210.4.2.dta	4	1.753006e+003	QLATKAARBSAPATGGVK(4)(9)	Loop-Linked	4BrBut	1.752997e+003	Deamidated[Q](1)	7.447221e-001	9.875189e-002



6

General appendix

6.1. Abbreviations

Note: Standard abbreviations for SI units and amino acid are used. SNFG abbreviations for monosaccharides are listed separately in section 6.4.

ABPP	activity-based protein profiling
AC	alternating current
Ac	acetyl
ACN	acetonitrile
ADCC	antibody-dependent cellular cytotoxicity
AE-MS	affinity enrichment mass spectrometry; synonym of AP-MS
Afb	affibody
AGC	automatic gain control
AI-ETD	activated ion ETD
ALC	average local confidence
AML	acute myeloid leukaemia
AMPK	5' AMP-activated protein kinase [protein]
API	active pharmaceutical ingredient
AP-MS	affinity purification mass spectrometry
Ar	aromatic substituent

ATM	ataxia-telangiectasia mutated [protein]
ATR	ATM- and RAD3-related [protein]
AUC	area under the curve
Bhn	bromohomonorleucine
BLAST	Basic Local Alignment Search Tool
Bn	benzyl
Boc	*tert*-butoxycarbonyl
bp	base pairs
bpm	2,2'-bipyrimidine
BRCT	BRCA1 C-Terminus domain
BS3	bissulfosuccinimidyl suberate
BSA	bovine serum albumin
Bu	butyl
CAD	collision-activated dissociation
CATDS	concentration- and target-dependent selectivity
CAZy	Carbohydrate-Active Enzymes [database]
CBM	carbohydrate-binding molecule
CCCP	compound-centric chemoproteomics
CCS	collision cross section
CDG	congenital discorder of glycosylation
CDR	complementarity determining region
CE	capillary electrophoresis *or* collision energy
CETSA	cellular thermal shift assay
CFG	Consortium for Functional Glycomics
CID	collision-induced dissociation
CMP	cytidine monophosphate
ConA	concavalin A
COSY	correlation spectroscopy
cryo-EM	cryo-electron microscopy
CTD	*C*-terminal domain
CuAAC	copper(I)-catalysed azide-alkyne cycloaddition
Cyc	cyclic ring system
Da	Dalton
DARTS	drug affinity responsive target stability
DBHDA	2,5 dibromohexanediamide
DCM	dichloromethane

DDA	data-dependent acquisition
DDDT	data-dependent decision tree
DFT	density functional theory
Dha	dehydroalanine
DIA	dataindependent acquisition
DIPEA	*N*,*N*-diisopropylethylamine
DMAP	4-*N*,*N*-dimethylaminopyridine
DMSO	dimethyl sulfoxide
DMF	*N*,*N*-dimethylformamide
DNA	desoxyribonucleic acid
DNA-PK	DNA-dependent protein kinase
Dol-P	dolichol phosphate
DSB	DNA double-strand break
DSBU	disuccinimidyl dibutyric urea
DSS	disuccinimidyl suberate
DSSO	disuccinimidyl sulfoxide
DTT	dithiothreitol
e^-	electron
EC	Enzyme Commission
ECD	electron capture dissotiation
ECM	extracellular matrix
EDTA	ethylenediamine tetraacetate
EGF	epidermal growth factor
EHT	Extended Hückel Theory
EI	electron impact
ELISA	enzyme-linked immunoabsorbent assay
EM	electron microscopy
EPO	erythropoietin
eq	equivalents
ER	endoplasmatic reticulum
ERAD	endoplasmatic-reticulum-associated degradation
ERT	enzyme replacement therapy
ESI	electron spray ionisation
Et	ethyl
ETD	electron transfer dissociation
EThcD	electron-transfer/higher-energy collision dissociation

EtOAc	ethyl acetate
EtOH	ethanol
FA	formic acid
FAB	fast atom bombardment
F_{ab}	fragment, antigen-binding
FACT	facilitates chromatin transcription [protein complex]
FASP	filter-aided sample preparation
F_C	fragment, crystallisable
FDR	false-discovery rate
FLR	false localisation rate
FR	framework region
FRET	Förster resonance energy transfer
FSH	follicle-stimulating hormone
FT	Fourier transform
F_V	fragment, variable
FWHM	full width at half maximum
GAG	glucosaminoglycan
GALK	galactokinase
GALT	galactose-1-phosphate uridyltranserase
GC	gas chromatography
GO	gene ontology
GPI	glycosylphosphatidylinositol
GSEA	gene set enrichment
HAT	histone acetyltransferase
HBR	histone H2B repression [protein domain]
HBTU	*N*,*N*,*N*′,*N*′-tetramethyl-*O*-(1*H*-benzotriazol-1-yl)uronium hexafluorophosphate
HCD	higher-energy collision dissociation, *formerly:* higher-energy C-trap collision dissociation
HDAC	histone deacetylase
HDACi	histone deacetylase inhibitor
HDX	hydrogen deuterium exchange
HECID	high energy CID [not the same as HCD]
Hex	hexyl
HFR	hydroxyl-radical footprinting
HILIC	hydrophilic interaction liquid chromatography

HIV	human immunodeficiency virus
HMBC	heteronuclear multiple-bond correlation spectroscopy
HOMO	highest occupied molecular orbital
HpH	high-pH reverse phase chromatography
HPLC	high performance liquid chromatography
HSQC	heteronuclear single quantum coherence spectroscopy
Hyp	hydroxyproline
ICR	ion cyclotron resonance
IdoA	iduronic acid
Ig	immunoglobulin
IMAC	immobilised metal affinity chromatography
IMGT	international ImMunoGeneTics information system for immunoglobulins or antibodies
IR	infrared
IRMPD	infrared multiphoton dissociation
iST	in-StageTips
IUPAC	International Union of Pure and Applied Chemistry
J	coupling constant
JNK1	c-Jun N-terminal kinase 1 [protein]
KDM	lysine demethylase
KEGG	Kyoto Encyclopedia of Genes and Genomics
LC	liquid chromatography
LC-MS/MS	liquid chromatography - tandem mass spectrometry
LDN	LacdiNAc
LDS	lithium dodecyl sulfate
LFQ	label-free quantitation
LiP	limited proteolysis
LIT	linear ion trap
LOPIT	localisation of organelle protein by isotope tagging
LUMO	lowest unoccupied molecular orbital
mAb	monoclonal antibody
MALDI	matrix-assisted laser desorption/ionisation
MAPK	mitogen-activated protein kinase
max	maximum
MD	molecular dynamics
MDBP	methyl 2,5-dibromopentanoate

Me	methyl
MEA	modular enrichment analysis
MeOH	methanol
min	minutes
MM	molecular mechanics
MM-GBSA	molecular mechanics/Generalized Born surface area
MM-PBSA	molecular mechanics/Poisson-Boltzmann surface area
mOGT	mitochondrial OGT
MOPS	3-(*N*-morpholino)propanesulfonic acid
MRI	magnetic resonance imaging
MRM	multiple reaction monitoring
mRNA	messenger RNA
MS	mass spectrometry
MS/MS	tandem mass spectrometry
MWCO	molecular weight cut-off
MWI	microwave irradiation
NAD	nicotinamide adenine dinucleotide
NADP	nicotine adenine dinucleotide phosphate
NCBI	National Center for Biotechnology Information
NCE	normalised collision energy
NCL	native chemical ligation
ncOGT	nucleocytoplasmic OGT
NGS	Next Generation Sequencing
NLS	nuclear localisation sequence
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
NSAF	normalised spectral abundance factor
OGA	*O*-GlcNAcase
OGT	*O*-GlcNAc transferase
p	quintet
PANS	partially assembled nucleosome states
PCP	protein correlation profiling
PD	pulldown
PDB	Protein Data Bank
PDGF	platelet-derived growth factor [protein]
PE	petroleum ether

PEP	posterior error probability
PEPC	proximity-enabled protein-protein crosslinking
PGC	porous graphitised carbon
Ph	phenyl
PHD	plant homeodomain
PNGase F	peptide-*N*-glycosidase F
PP	pulse proteolysis
PPI	protein-protein interaction
ppm	parts per million
PSGL-1	P-selectin glycoprotein ligand 1
PSM	peptide-spectrum match
PTM	post-translational modification
PylRS	Pyrrolysyl RNA synthetase
QEx	Q exactive
QM	quantum mechanics
Q-TOF	quadrupole-time of flight
quant.	quantitative
RCSB	Research Collaboratory for Structural Bioinformatics
REMD	replica-exchange molecular dynamics
RF	radio frequency
RMSD	root mean square deviation
RNA	ribonucleic acid
RNAi	RNA interference
RNA-seq	RNA sequencing
ROS	reactive oxygen species
rpm	rotations per minute
RT	retention time
r.t.	room temperature
SA	supplementary activation
SAAV	single amino acid variants
sat.	saturated
SAX	strong anion exchange chromatography
SAXS	small-angle X-ray scattering
SCE	stepped collision energy
SCX	strong cation exchange
SDS PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis

SEA	singular enrichment analysis
SEC	size-exclusion chromatography
sec	secondary
Sia	sialic acid
SILAC	stable isotope labelling with amino acids in cell culture
SNFG	Symbol Nomenclature for Glycans
SNP	single nucleotide polymorphism
SOMO	singly occupied molecular orbital
SP3	solid-phase enhanced sample preparation
SPBP	Stromelysin-1 PDGF-responsive element binding protein [= TCF20]
SPR	surface plasmon resonance
SPROX	stability of proteins from rates of oxidation
SRM	selected reaction monitoring
Sulfo-SBED	sulfo-*N*-hydroxysuccinimidyl-2-(6-[biotinamido]-2-(p-azido benzamido)-hexanoamido) ethyl-1,3'-dithioproprionate
SWATH	sequential windowed acquisition of all theoretical fragment ion mass spectra
TCEP	tris(2-carboxyethyl)phosphine
TCF20	transcription factor 20
TDA	target-decoy approach
TEA	triethylamine
TEAB	triethylammonium bicarbonate
TFA	trifluoroacetic acid
tert	tertiary
TGN	*trans*-Golgi network
THF	tetrahydrofuran
TLC	thin layer chromatography
TMT	tandem mass tag
TNF	tumour necrosis factor
TOF	time of flight
TPP	thermal proteome profiling
TPR	tetratricopeptide repeat
TRADD	tumour necrosis factor receptor 1-associated DEATH domain [protein]
TSP	thrombospondin

TSQ	6-methoxy-8-*p*-toluenesulfonamido-quinoline
TSR	thrombospondin repeats
TyrRS	Tyrosyl RNA synthetase
Uaa	unnatural amino acid
UPLC	ultra-performance liquid chromatography
UV	ultraviolet
UVPD	ultraviolet photodissociation
VBA	Virtual Basic
Wt	wild type
WB	Western Blot
WGA	wheat germ agglutinin
XIC	extracted ion chromatogram
xLink	crosslink
XL-MS	crosslinking mass spectrometry
ZIC-HILIC	zwitterionic HILIC

6.2. List of Figures

1.1. Common classes of animal glycans and glycoconjugates. 4
1.2. Glycan-protein linkages reported in nature 7
1.3. Mammalian glycosylation cores with examples. 10
1.4. Common extensions, epitopes and antigens on human glycans. 17
1.5. Simplified overview of a proteomics workflow. 24
1.6. Recently discovered *O*-glycosylation-dependent proteases. 27
1.7. PNGase F resistant *N*-glycopeptides. 28
1.8. Selected *N*-glycopeptide enrichment principles. 32
1.9. Overview of mass analysers relevant to this thesis. 38
1.10. Peptide fragment ion nomenclature. 42
1.11. CID fragmentation mechanism. 44
1.12. ETD fragmentation mechanism. 46
1.13. Ion routing scheme in a Q Exactive and Orbitrap Elite. 49
1.14. The target-decoy approach. 55
1.15. Quantification methods used in proteomics. 58
1.16. Fragment ions in crosslinking mass spectrometry. 61
1.17. False discovery rate in crosslinking mass spectrometry. 64

2.1. Structure of monoclonal IgG antibodies and their genetic encoding. 113
2.2. Hybridoma sequencing using Absolute Antibody's RNAseq workflow. . . . 114
2.3. Objective 120
2.4. Reducing denaturing intact protein LC-MS of the anti-γH2AX mAb. . . 123
2.5. Flowchart of Peaks 8.5 software package. 125
2.6. ALPS algorithm. 126
2.7. Flowchart of PeaksAB 2.0 *de novo* sequencing algorithm. 128
2.8. De novo sequencing of Herceptin light chain − PSM-DD alignment 136

2.9. De novo sequencing of Herceptin light chain − PSM-DD alignment 137

2.10. *De novo* sequencing of Herceptin light chain − PSM-DD alignment 138

2.11. *De novo* sequencing of Herceptin light chain − PSM-DDS alignment . . . 138

2.12. *De novo* sequencing of Herceptin light chain by PeaksAB. 139

2.13. Sequence coverage of the Herceptin heavy chain in the combined dataset. 140

2.14. Contig alignment to the Herceptin heavy chain 141

2.15. *De novo* sequencing of Herceptin by PeaksAB 142

2.16. Alignment of the contig to the highest score antibody database 143

2.17. Validation and inspection of the anti-γH2Ax light chain template 144

2.18. Solving the CDR-3 region of anti-γH2Ax light chain. 146

2.19. *De novo* sequencing summary of the anti-γH2AX light chain. 147

2.20. *De novo* sequencing of the anti-γH2Ax mAb light chain by PeaksAB. . . . 149

2.21. Alignment of the three templates sequences from the antibody database . 151

2.22. variable 155

2.23. Sequence coverage of the consensus template. 157

2.24. Investigating the CMITDF sequence coverage gap in the heavy chain. . . 159

2.25. Investigation of the C-terminus of the anti-γH2Ax heavy chain. 162

2.26. Summary of *de novo* sequencing of the anti-γH2AX heavy chain. 164

2.27. *De novo* sequencing of the anti-γH2Ax mAb heavy chain by PeaksAB . . 165

2.28. *N*-terminal pyroglutamate formation on the anti-γH2Ax heavy chain. . . . 167

2.29. Glycosylation analysis of the murine anti-γH2Ax monoclonal antibody. . . 172

2.30. Analysis of the second glycopeptide of the murine anti-γH2Ax monoclonal antibody. 174

2.31. Homology modelling followed by antibody docking suggests a phosphoserine-specific antibody-antigen recognition. 176

2.32. Analysis of the antibody-antigen interactions as predicted by ligand docking. 178

2.33. Structural genomic analysis of the murine anti-γH2AX mAb light chain. . 181

2.34. Structural genomic analysis of the murine anti-γH2AX mAb heavy chain. 183

2.35. Building the full antibody model of the murine anti-γH2AX 186

2.36. Results of the humanisation of the anti-γH2AX antibody. 187

2.37. Homology modelling of the murine F_V. 197

3.1. Glycopeptide fragment ion nomenclature following the Roepstorff-Fohlman system for peptides and the Domon-Costello system for glycans. 218

3.2. Glycopeptide fragmentation in collision-induced dissociation down to the different oxonium ions. 220

3.3. Glycopeptide-specific fragmentation pathways observed in ETD. 226

3.4. The main dissociation methods in glycoproteomics. 229

3.5. Peptides and glycopeptides differ in their fragmentation energetics. 232

3.6. Energy-resolved fragmentation of *N*-glycopeptides. 234

3.7. Energy-resolved fragmentation of *O*-glycopeptides. 236

3.8. Initial results on decorin have questionable presumed glycopeptide spectra.250

3.9. Decorin disaccharide-linker-glycopeptide compositions reported in the literature. 251

3.10. Representative spectra for the 3 identified O-glycans on decorin. 252

3.11. Intact protein LC-MS for isolated heparinase-trimmed decorin. 253

3.12. HCD spectra for additional modifications on the O-glycans of decorin. . 254

3.13. Representative HCD spectra for additional modifications on the O-glycans of decorin 256

3.14. Using GlycReSoft to construct a searchable mass to glycopeptide fragment database. 257

3.15. Oxonium ion trace. 258

3.16. 2-step HCD with 10-15 NCE yield glycan fragmentation on decorin glycopeptides. 259

3.17. 2-step HCD with 10-15 NCE yields glycan fragmentation on decorin glycopeptides. 260

3.18. Oxonium ions characteristic for heparinase-trimmed decorin. 260

3.19. 3-step HCD with 10-15-20 NCE yields glycan fragmentation for decorin glycopeptides 261

3.20. Fragmentation of the PSVVAVPLPLHR carrying HexNAcHex from bovine asialofetuin at four different NCE values. 263

3.21. Fragmentation of the AAGPTPSAAGPPVASVVVGPSVVAVPLPLHR carrying HexNAcHex from bovine asialofetuin at four different NCE values. . 264

3.22. Fragmentation of the PSVVAVPLPLHR glycopeptide carrying HexNAcHex from bovine asialofetuin in four different SCE-HCD methods. 267

3.23. Fragmentation of the AAGPTPSAAGPPVASVVVGPSVVAVPLPLHR glycopeptide carrying HexNAcHex from bovine asialofetuin in four different SCE-HCD methods. 268

3.24. Fragmentation of the bovine decorin glycopeptide in different NCE and SCE-HCD methods. 269

3.25. Fragmentation of the VVVGPSVVAVPLPLHR glycopeptide carrying HexNAc(2) Hex(2)Neu5Ac(2) from bovine fetuin in four different SCE-HCD methods. 271

3.26. EThcD spectrum of IgG N-glycosylation shows fragmentation at the non-reducing terminus of the glycan. 276

3.27. Effects of input file format on EThcD-based glycan identifications in Byonic. 277

3.28. Examples of ETD fragmentation of *O*-glycopeptides presented in this chapter. 279

4.1. Post-translational modification of histones and their identification using mass spectrometry. 319

4.2. The chemical biology toolkit for synthetic histone modification. 322

4.3. Application of chemical biology tools for understanding chromatin biology *in vitro*. 327

4.4. Crystal structure of the *Xenopus laevis* nucleosome highlighting the modification site. 331

4.5. Probing the PTM-specific interaction partners of histone modifications at H2B-S36. 333

4.6. Synthesis of H2B-Dha36 from H2B-Cys36. 334

4.7. Synthesis of H2B-CysGlcNAc36 from H2B-Dha36. 335

4.8. Attempted synthesis of H2B-pCys36 from H2B-Dha36. 337

4.9. Synthesis of H2B-Pfa36 from H2B-Dha36. 338

4.10. ETD MS^2 spectra of sulfur-linked modifications show diagnostic w-ions. . 340

4.11. Histone octamer assembly and their purification by size-exclusion chromatography (SEC). 342

4.12. Quality control of 186bp Widom 601 DNA preparation and observed forms of DNA damage. 345

4.13. Quality control of 186bp Widom 601 DNA preparation and observed forms of DNA damage. 347

4.14. Statistical analysis of the biological replicates and individual pull-down experiments. 350

4.15. Synthetic nucleosomes carrying the H2B-Ser36GlcNAc modification preferentially enrich the TCF20 transcription factor. 352

4.16. mRNA expression levels of AMPK and TCF20 in different cell lines. . . . 356

4.17. Speculative hypotheses for the role of phosphorylation and glycosylation of H2BS36. 358

4.18. Analytical characterisation of synthesised H2B-Dha36 protein. 367

4.19. Analytical characterisation of synthesised H2B-CysGlcNAc36 protein. . . 368

4.20. Analytical characterisation of synthesised H2B-pCys36 protein. 370

4.21. FPLC trace of the attempted enrichment of phosphorylated H2B-pCys36 using a combined Fe and Ga-IMAC. 371

4.22. Analytical characterisation of synthesised H2B-Pfa36 protein. 372

4.23. Analytical characterisation of synthetic histone octamers. 375

4.24. Preparation of 145bp Widom601 strong positioning DNA. 378

4.25. Preparation of 186bp Widom601 strong positioning DNA 380

4.26. Anti-biotin Southern Blot. 380

4.27. Nucleosome assembly with 145bp Widom 601 strong positioning DNA. . . 383

5.1. Proximity-enabled protein-protein crosslinking using unnatural amino acids. 400

5.2. Site-specific photoredox catalysis for post-translational protein mutagenesis. 402

5.3. eH3 K27Bhn characterisation shows characteristic isotope pattern for brominated peptides. 405

5.4. xLink gels 406

5.5. Proximity-enabled protein protein crosslinking using the unnatural amino acid Bhn 411

5.6. Inter-histone crosslinking PSM 413

5.7. Outlook on proximity-enabled protein crosslinking with post-translationally mutated proteins. 415

6.3. List of Tables

1.1. Characteristics of selected proteases relevant to this thesis 26

3.1. Commonly observed oxonium ions in collision-induced fragmentations. . . 221

3.2. Commonly observed *B*-ions in collision-induced fragmentations. 223

3.3. Comparing fragmentation techniques for *N*- and *O*-glycopeptides on an Orbitrap Elite and Q Exactive. 273

3.4. Finding appropriate search parameters for mixed CID-alt-EThcD fragmentation on ‚hacked' Orbitrap Elite. 278

3.5. Default collision energy regression equations 291

4.1. Sites of *O*-GlcNAc glycosylation in human histones. 329

5.1. Parameter evaluation on the eH3K9-Bhn•KDM4A crosslink 409

6.1. SNFG monosaccharide symbol nomenclature 470

6.4. SNFG monosaccharide symbols & codes

SNFG monosaccharide symbol nomenclature

Table 6.1. SNFG monosaccharide symbol nomenclature.

SHAPE	White	Blue	Green	Yellow	Orange	Pink	Purple	Light Blue	Brown	Red
Filled Circle	Hexose	Glc	Man	Gal	Gul	Alt	All	Tal	Ido	
Filled Square	HexNAc	GlcNAc	ManNAc	GalNAc	GulNAc	AltNAc	AllNAc	TalNAc	IdoNAc	
Crossed Square	Hexosamine	GlcN	ManN	GalN	GulN	AltN	AllN	TalN	IdoN	
Divided Diamond	Hexuronate	GlcA	ManA	GalA	GulA	AltA	AllA	TalA	IdoA	
Filled Triangle	Deoxyhexose	Qui	Rha		6dGul	6dAlt		6dTal		Fuc
Divided Triangle	DeoxyhexNAc	QuiNAc	RhaNAc			6dAltNAc		6dTalNAc		FucNAc
Flat Rectangle	Di-deoxyhexose	Oli	Tyv		Abe	Par	Dig	Col		
Filled Star	Pentose		Ara	Lyx	Xyl	Rib				
Filled Diamond	Deoxynonulosonate		Kdn				Neu5Ac	Neu5Gc	Neu	Sia
Flat Diamond	Di-deoxynonulosonate		Pse	Leg		Aci		4eLeg		
Flat Hexagon	Unknown	Bac	LDManHep	Kdo	Dha	DDManHep	MurNAc	MurNGc	Mur	
Pentagon	Assigned	Api	Fruc	Tag	Sor	Psi				

The symbols presented here follow the Third Edition of the *Essentials of Glycobiology*. https://www.ncbi.nlm.nih.gov/glycans/snfg.html

KEGG / SNFG full list of monosaccharide codes & abbreviations

Hexose (six-carbon sugar)		
Hex	Hexose	
Glc	D-Glucose	D-Glucopyranose
Galf	D-Galactose	D-Galactofuranose
Gal	D-Galactose	D-Galactopyranose
LGal	L-Galactose	L-Galactopyranose
Man	D-Mannose	D-Mannopyranose
All	D-Allose	D-Allopyranose
LAlt	L-Altrose	L-Altropyranose
Gul	D-Gulose	D-Gulopyranose
LIdo	L-Idose	L-Idopyranose
Tal	D-Talose	D-Talopyranose
Pentose (five-carbon sugar)		
Ribf	D-Ribose	D-Ribofuranose
Rib	D-Ribose	D-Ribopyranose
Araf	D-Arabinose	D-Arabinofuranose
Ara	D-Arabinose	D-Arabinopyranose
LAraf	L-Arabinose	L-Arabinofuranose
LAra	L-Arabinose	L-Arabinopyranose
Xyl	D-Xylose	D-Xylopyranose
Lyx	D-Lyxose	D-Lyxopyranose
Tetrose (four-carbon sugar)		
Ery	D-Erythrose	D-Erythrofuranose
Tho	D-Threose	D-Threofuranose
Heptose (seven-carbon sugar)		
Hep	Heptose	
Lgro-manHep	L-glycero-D-manno-Heptose	L-glycero-D-manno-Heptopyranose (LDmanHep)
gro-manHep	D-glycero-D-manno-Heptose	D-glycero-D-manno-Heptopyranose (DDmanHep)
Deoxysugar (hexose or pentose without a hydroxyl group at the 6-position or the 2-position)		
L6dAlt	6-Deoxy-L-altrose	6-Deoxy-L-altropyranose
6dGul	6-Deoxy-D-gulose	6-Deoxy-D-gulopyranose
6dTal	6-Deoxy-D-talose	6-Deoxy-D-talopyranose
Fuc	D-Fucose	6-Deoxy-D-galactopyranose
LFuc	L-Fucose	6-Deoxy-L-galactopyranose
Rha	D-Rhamnose	6-Deoxy-D-mannopyranose
LRha	L-Rhamnose	6-Deoxy-L-mannopyranose
Qui	D-Quinovose	6-Deoxy-D-glucopyranose
2dGlc	2-Deoxyglucose	2-Deoxy-D-arabino-hexose
dRib	2-Deoxyribose	2-Deoxy-D-erythro-pentose
Di-deoxysugar (hexose without hydroxyl groups at the 6-position and another position)		
Oli	Olivose	2,6-Dideoxy-D-arabino-hexopyranose (2,6daraHex)
Tyv	Tyvelose	3,6-Dideoxy-D-arabino-hexopyranose (3,6daraHex)
Asc	Ascarylose	3,6-Dideoxy-L-arabino-hexopyranose (L3,6daraHex)
Abe	Abequose	3,6-Dideoxy-D-xylo-hexopyranose (3,6dxylHex)

Par	Paratose	3,6-Dideoxy-D-ribo-hexopyranose (3,6dribHex)
Dig	Digitoxose	2,6-Dideoxy-D-ribo-hexopyranose (2,6dribHex)
Col	Colitose	3,6-Dideoxy-L-xylo-hexopyranose (L3,6dxylHex)
Amino sugar (hexose with an amino group at the 2-position)		
GlcN	D-Glucosamine	2-Amino-2-deoxy-D-glucopyranose
GalN	D-Galactosamine	2-Amino-2-deoxy-D-galactopyranose
ManN	D-Mannosamine	2-Amino-2-deoxy-D-mannopyranose
AllN	D-Allosamine	2-Amino-2-deoxy-D-allopyranose
LAltN	L-Altrosamine	2-Amino-2-deoxy-L-altropyranose
GulN	D-Gulosamine	2-Amino-2-deoxy-D-gulopyranose
LIdoN	L-Idosamine	2-Amino-2-deoxy-L-idopyranose
TalN	D-Talosamine	2-Amino-2-deoxy-D-talopyranose
Amino sugar (hexose with an N-acetylated amino group at the 2-position)		
GlcNAc	N-Acetyl-D-glucosamine	2-Acetamido-2-deoxy-D-glucopyranose
GalNAc	N-Acetyl-D-galactosamine	2-Acetamido-2-deoxy-D-galactopyranose
ManNAc	N-Acetyl-D-mannosamine	2-Acetamido-2-deoxy-D-mannopyranose
AllNAc	N-Acetyl-D-allosamine	2-Acetamido-2-deoxy-D-allopyranose
LAltNAc	N-Acetyl-L-altrosamine	2-Acetamido-2-deoxy-L-altropyranose
GulNAc	N-Acetyl-D-gulosamine	2-Acetamido-2-deoxy-D-gulopyranose
LIdoNAc	N-Acetyl-L-idosamine	2-Acetamido-2-deoxy-L-idopyranose
TalNAc	N-Acetyl-D-talosamine	2-Acetamido-2-deoxy-D-talopyranose
FucNAc	N-Acetyl-D-fucosamine	2-Acetamido-2,6-dideoxy-D-galactopyranose
LFucNAc	N-Acetyl-L-fucosamine	2-Acetamido-2,6-dideoxy-L-galactopyranose
LRhaNAc	N-Acetyl-L-rhamnosamine	2-Acetamido-2,6-dideoxy-L-mannopyranose
QuiNAc	N-Acetyl-D-quinovosamine	2-Acetamido-2,6-dideoxy-D-glucopyranose
L6dAltNAc	N-Acetyl-6-deoxy-L-altrosamine	2-Acetamido-2,6-dideoxy-L-altropyranose
6dTalNAc	N-Acetyl-6-deoxy-D-talosamine	2-Acetamido-2,6-dideoxy-D-talopyranose
Uronic acid (hexose with a negatively charged carboxylate at the 6-position)		
GlcA	D-Glucuronic acid	D-Glucopyranuronic acid
GalA	D-Galacturonic acid	D-Galactopyranuronic acid
ManA	D-Mannuronic acid	D-Mannopyranuronic acid
AllA	D-Alluronic acid	D-Allopyranuronic acid
LAltA	L-Altruronic acid	L-Altropyranuronic acid
GulA	D-Guluronic acid	D-Gulopyranuronic acid
LGulA	L-Guluronic acid	L-Gulopyranuronic acid
LIdoA	L-Iduronic acid	L-Idopyranuronic acid
TalA	D-Taluronic acid	D-Talopyranuronic acid
Sialic acid (nine-carbon acidic sugar)		
Sia	Sialic acid	
Neu	Neuraminic acid	5-Amino-3,5-dideoxy-D-glycero-D-galacto-non-2-ulosonic acid
Neu5Ac	N-Acetylneuraminic acid	5-Acetamido-3,5-dideoxy-D-glycero-D-galacto-non-2-ulosonic acid
Neu5Gc	N-Glycolylneuraminic acid	5-Glycolylamido-3,5-dideoxy-D-glycero-D-galacto-non-2-ulosonic acid
Sugar alcohol		
Ery-ol	Erythritol	
Ara-ol	Arabinitol	
Xyl-ol	Xylitol	
Rib-ol	Ribitol	

Glc-ol	Glucitol	
Gal-ol	Galactitol	
Man-ol	Mannitol	
Ketose		
Psi	D-Psicose	D-ribo-Hex-2-ulopyranose
Fruf	D-Fructose	D-arabino-Hex-2-ulofuranose (D-fructofuranose)
Fru	D-Fructose	D-arabino-Hex-2-ulopyranose
LSor	L-Sorbose	L-xylo-Hex-2-ulopyranose
Tag	D-Tagatose	D-lyxo-Hex-2-ulopyranose
Xul	D-Xylulose	D-threo-Pent-2-ulopyranose
Sed	D-Sedoheptulose	D-altro-Hept-2-ulopyranose
Others		
Api	Apiose	3-C-(Hydroxymethyl)-D-erythofuranose
Bac	Bacillosamine	2,4,6-Trideoxy-2,4-diamino-D-glucopyranose
The	Thevetose	6-Deoxy-3-O-methyl-D-glucose (Qui3Me)
Aco	Acofriose	3-O-Methyl-D-rhamnose (Rha3Me)
LAco	L-Acofriose	3-O-Methyl-L-rhamnose (LRha3Me)
Cym	Cymarose	2,6-Dideoxy-3-methyl-D-ribo-hexose (Dig3Me)
Mur	Muramic acid	2-Amino-3-O-[(R)-1-carboxyethyl]-2-deoxy-D-glucopyranose
MurNAc	N-Acetylmuramic acid	2-Acetamido-3-O-[(R)-carboxyethyl]-2-deoxy-D-glucopyranose
MurNGc	N-Glycolylmuramic acid	2-Glycolylamido-3-O-[(R)-1-carboxyethyl]-2-deoxy-D-glucopyranose
Dha	3-Deoxy-lyxo-heptulosaric acid	3-Deoxy-D-lyxo-hept-2-ulopyranosaric acid
Kdo	Ketodeoxyoctonic acid	3-Deoxy-D-manno-oct-2-ulopyranosonic acid
Kdn	Ketodeoxynononic acid	3-Deoxy-D-glycero-D-galacto-non-2-ulopyranosonic acid
Pse	Pseudaminic acid	5,7-Diamino-3,5,7,9-tetradeoxy-L-glycero-L-manno-non-2-ulopyranosonic acid
Aci	Acinetaminic acid	5,7-Diamino-3,5,7,9-tetradeoxy-L-glycero-L-altro-non-2-ulopyranosonic acid
Leg	Legionaminic acid	5,7-Diamino-3,5,7,9-tetradeoxy-D-glycero-D-galacto-non-2-ulopyranosonic acid
4eLeg	4-Epilegionaminic acid	5,7-Diamino-3,5,7,9-tetradeoxy-D-glycero-D-talo-non-2-ulopyranosonic acid

Compound abbreviations: Dol, Dolichol; Cer, Ceramide; Ino, Inositol; P, Phosphoric acid; S, Sulfuric acid; EtN, Ethanolamine; Cho, Choline; Ac, Acetyl group; Py, 1-Carboxyethane-1,1-diyl group; Me, Methyl group; Gro, Glycerol

Configurational prefixes: D, D-form; L, L-form; gro, glycero; ery, erythro; thr, threo; rib, ribo; ara, arabino; xyl, xylo; lyx, lyxo; alt, altro; glc, gluco; man, manno; gul, gulo; gal, galacto; tal, talo

Other prefixes and suffixes: d, deoxy; en, unsaturated; f, furanose; p, pyranose; ol, alditol; onic, aldonic acid

Monosaccharide codes in SNFG: In SNFG absolute configuration is assumed by default L for Ara, Fuc, Ido, IdoA, Rha, Alt, AltA, Sor, Api and D for other monosaccharides. In contrast, default configuration is always D in KEGG. Thus, for example, Ara in SNFG is LAra in KEGG.

Source: https://www.genome.jp/kegg/catalog/codes2.html
Last updated: February 26, 2020