Fast detection and quantification of proteoforms in therapeutic proteins using intact protein mass spectrometry

Manasi Gaikwad

A thesis presented for the acquisition of the academic degree
Doctor rerum naturalium
Dr. rer. nat.

at

University of Hamburg
Faculty of Mathematics, Informatics and Natural Sciences
Department of Chemistry

2021

Research for this thesis was carried out from April 2018 until June 2021 in the Mass Spectrometric Proteomics group of Prof. Dr. Hartmut Schlüter, located at the University Medical Center Hamburg-Eppendorf.

Evaluators for the dissertation:
Prof. Dr. Hartmut Schlüter
Dr. Maria Riedner

Date of oral disputation:
13th August 2021

My research work was carried out as a part of A4B – Analytics for Biologics project. A4B consortium is a Europe-wide innovative training network (ITN), funded by the Horizon 2020, Marie Sklodowska-Curie Action ITN 2017, of the European Commission (H2020-MSCA-ITN-2017). The A4B project consortium was composed of 15 Ph.D. students working on dedicated topics of therapeutic protein production, purification, and analysis.

List of Publications

Laura Heikaus, Siti Hidayah, **Manasi Gaikwad**, Marta Kotasinska, Verena Richter, Marcel Kwiatkowski, Hartmut Schlüter 'Sample displacement batch chromatography of proteins', in *Methods in Molecular Biology* (2021) doi: 10.1007/978-1-0716-0775-6_19.

Kyowon Jeong, Jihyung Kim, **Manasi Gaikwad**, Siti Nurul Hidayah, Laura Heikaus, HartmutSchlüter, Oliver Kohlbacher 'FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics', Cell Systems (2020) doi: 10.1016/j.cels.2020.01.003.

Siti Nurul Hidayah, **Manasi Gaikwad**, Laura Heikaus and Hartmut Schlüter 'Preparing Proteoforms of Therapeutic Proteins for Top-Down Mass Spectrometry', in *Proteoforms - Concept and Applications in Medical Sciences* (2020) doi: 10.5772/intechopen.89644.

Ramin Fazel, Yudong Guan, Behrouz Vaziri, Christoph Krisp, Laura Heikaus, Amirhossein Saadati, Siti Nurul Hidayah, **Manasi Gaikwad**, Hartmut Schlüter 'Structural and In Vitro Functional Comparability Analysis of Altebrel^{TM}, a Proposed Etanercept Biosimilar: Focus on Primary Sequence and Glycosylation', *Pharmaceuticals*, 12(1), p. 14 (2019) doi: 10.3390/ph12010014.

Yudong Guan, Min Zhang, **Manasi Gaikwad**, Hannah Voss, Ramin Fazel, Samira Ansari, Huali Shen, Jigang Wang, Hartmut Schlüter 'An integrated strategy reveals complex glycosylation of erythropoietin using top-down and bottom-up mass spectrometry'(Preprint) doi: https://doi.org/10.1101/2021.02.09.430553

List of posters

Manasi Gaikwad, Siti Nurul Hidayah, Jihyung Kim, Kyowon Jeong, Oliver Kohlbacher, Hartmut Schlüter 'Fast quantification of proteoforms at intact level from a commercial ovalbumin fraction' Bioprocessing Summit Europe 2020 (Online)

Manasi Gaikwad, Benjamin Dreyer, Hartmut Schlüter 'Simple Workflow for Proteoform Separation and Analysis Using 2D Gel Electrophoresis High Resolution Mass Spectrometry', 22nd International Mass Spectrometry Conference, 2018, Florence, Italy

List of conferences and presentations

Analytics for Biologics (A4B)-Network meeting 3 including presentation, 2021 (Online).

Bioprocessing Summit Europe 2021 (Online).

American Society for Mass Spectrometry (ASMS) reboot 2020 (Online).

Analytics for Biologics (A4B)-Network meeting 2 including presentation, 2020, Leiden, Netherlands.

Analytics for Biologics (A4B)-Network meeting 1 including presentation, 2019, Stockholm, Sweden.

Contents

List of abbreviations

%	Percent
°C	Degree Celsius
Å	Angstrom
ACN	Acetonitrile
AGC	Automatic gain control
AmAc	Ammonium acetate
AUC	Area under curve
BCA	Bicinchoninic acid
CE	Capillary electrophoresis
DDA	Data-dependent acquisition
DIA	Data independent acquisition
DTT	Dithiothreitol
E	Eluate fraction from SDBC experiment
EIF	Extracted ion flowgram
ELISA	Enzyme-linked immunosorbent assay
ESI	Electrospray ionization
F	Fucose
FA	Formic acid
Fab	Fragment antigen-binding region
Fc	Fragment crystallizable region
FIA	Flow injection analysis
FT	Flow-through fraction from SDBC experiment
G	Galactose
G0F	Galactose with no fucose
G1F	Galactose with one fucose
H	Hexose
HpH	High pH fractionation
HPLC	High Performance Liquid Chromatography
IdeS	FabRICATOR enzyme
IAA	Iodoacetamide
ISCID	In-source collision-induced dissociation
LBA	Ligand binding assay
LC	Liquid chromatography
LOD	Limit of detection
LOQ	Limit of quantification
m/z	Mass to charge ratio
mAb	Monoclonal antibody- Adalimumab
mg	Milligram
min	Minutes
mL	Millilitres

mM	Millimolar
Mo	Monomer of Ovalbumin
MS	Mass spectrometry
MS^1	Full scan spectra
MS^2	Fragment spectra
MS/MS	Tandem mass spectrometry
N	N-Acetylhexosamine
NaCl	Sodium chloride
NANA	Acetylneuraminic acid.
ng	Nano gram
Ori	Original sample
P	Phosphorylation
pH	pH value
PTM	Post translational modifications
R	Resolution
R^2	coefficient of determination
RP	Reverse phase
RPLC	Reverse phase liquid chromatography
RT	Retention time
s	Seconds
SDBC	Sample displacement batch chromatography
SDC	Sodium deoxycholate
SDS PAGE	Sodium dodecyl sulphate–polyacrylamide gel electrophoresis
SN	Signal to noise
TOF	Time of flight
TP(s)	Therapeutic protein(s)
Tr	C terminal truncated Ovalbumin form
UPLC	Ultra-high performance liquid chromatography
v/v	Volume / volume
µg	Microgram
µL	Microliter
µm	Micrometres
z	Charge

1 Abstract / Zusammenfassung

Post translation modifications, amino acid substitutions, and truncations of therapeutic proteins (TPs) are very common and responsible for a large heterogeneity, which can alter the functional efficacy of the TPs. Detailed qualitative and quantitative analysis of these multiple protein species - now referred to as “proteoforms” - is critical for the regulatory approval of TPs. Considering the growing demand for TPs and the significance of proteoforms, this thesis aimed at establishing a fast method for the quantification of proteoforms in TPs. Systematic but fast distinctions and quantification of proteoforms are however challenging tasks for most analytical techniques, because of the highly similar physiochemical properties shared by the proteoforms. The state-of-the-art high-resolution "intact protein mass spectrometry (MS)" was employed and optimized in the present work for achieving proteoform quantification.

First, the classic 'reverse-phase liquid chromatography coupled to MS' (RPLC-MS) method was studied, especially concerning the recovery of proteoforms. Some degree of on-column loss of proteoforms from the model protein Ovalbumin was observed while using the monolithic RP column for analysis. For circumventing the problem of loss of proteoforms on stationary phases of liquid chromatography columns, flow injection analysis coupled to MS (FIA-MS) was investigated as an alternative method for fast proteoform quantification. Factors contributing towards sensitive proteoform detection in the FIA-MS method were systematically optimized. With an analysis time of only 4 mins per sample and improved specificity of proteoform detection achieved, the eminence of FIA-MS method over the RPLC-MS method was proven. Further on, the FIA-MS based detection of lower abundant proteoforms in the sample could be improved by applying "in-solution supercharging" with sulfolane as an additive supercharging agent.

Until the beginning of this thesis, quantification of full-length proteoforms from isotopically unresolved mass spectral signals was underexplored. Therefore, both-isotopically resolved and isotopically unresolved signals of proteoforms were examined for developing a proteoform quantification strategy. Results of proteoform quantifications by ‘extracted ion flowgram (EIF) strategy' & ‘deconvoluted spectrum-based strategy’ were compared and assessed based on accuracy and precision. Further on, the correlation of quantification results, between multiple deconvolution tools for isotopically unresolved proteoforms was shown for the first time. The optimized FIA-MS method and quantification strategy was further applied for analysis of proteoforms from the clinically

relevant TP-Adalimumab. The specific detection of Adalimumab proteoforms was more challenging because of the overlapping m/z distribution and high dynamic range of these proteoforms. To accomplish detection of maximum proteoforms from Adalimumab, an offline 'sample displacement batch chromatography’ (SDBC) approach was used for fractionation of proteoforms from this TP. By applying the strategy of deconvoluted spectrum-based proteoform quantification, the success of SDBC method for fractionation of proteoforms was validated.

In conclusion, a novel approach for fast proteoform detection using the FIA-MS method with total analysis time of fewer than 4 mins per sample as well as an accurate proteoform quantification strategy was developed and proven with an Adalimumab sample. Fast FIA-MS-based detection & deconvoluted spectrum-based quantification, can be broadly applied to various TPs for quick proteoform overview.

Zusammenfassung

Substitutionen in der Aminosäuresequenz, Trunkierungen und posttranslationale Modifikationen sind für eine Vielzahl von Protein-Spezies und damit für eine große Heterogenität von therapeutischen Proteinen (TPs) verantwortlich. Die Heterogenität der TPs kann die funktionelle Wirksamkeit des Arzneimittelprodukts verändern. Eine detaillierte quantitative Analyse der Protein-Spezies der TPs - im Folgenden als "Proteoformen" bezeichnet, ist für die behördliche Zulassung von TPs äußerst wichtig. In Anbetracht der wachsenden Nachfrage nach TPs und die Bedeutung von Proteoformen der TPs, hatte diese Arbeit die Etablierung einer schnellen Methode zur Quantifizierung von Proteoformen von TPs zum Ziel. Die Quantifizierung von Proteoformen ist jedoch für die meisten analytischen Techniken eine große Herausforderung, da Proteoformen sehr ähnliche physiko-chemische Eigenschaften aufweisen. In der vorliegenden Arbeit wurde die derzeitige hochauflösende "Intakt-Protein-Massenspektrometrie" für die Quantifizierung von Proteoformen eingesetzt und optimiert. Für die Entwicklung einer schnellen Quantifizierungsmethode wurde zunächst die klassische "Umkehrphasen-Flüssigkeits-chromatographie gekoppelt an die Massenspektrometrie" (RPLC-MS) untersucht, insbesondere im Hinblick auf die Wiederfindung der Proteoformen. Hier wurde ein Verlust einiger Proteoformen des Modellproteins Ovalbumin auf der monolithischen RP-Säule festgestellt. Daher wurde die Fließinjektionsanalyse gekoppelt an MS (FIA-MS) als alternative Methode für eine schnelle Proteoform-Detektion gewählt und im Detail untersucht. Systematisch wurden die Faktoren optimiert, die zur sensitiven Detektion der Proteoformen in der FIA-MS-Methode beitragen. Die Überlegenheit der FIA-MS-Methode gegenüber der RPLC-MS-Methode wurde mit einer Analysezeit von nur 4 min pro Probe nachgewiesen. Weiterhin konnte die Detektion von gering abundanten Proteoformen in der Probe durch die Anwendung von "in-solution supercharging" mit Sulfolan verbessert werden. Die nächste Fragestellung der Arbeit bestand darin, ein Ansatz für die Quantifizierung von Proteoformen in voller Länge zu entwickeln. Derzeit ist vor allem die Quantifizierung von Volllängen-Proteoformen aus isotopisch unaufgelösten Massenspektraldaten ein enormer Engpass und noch wenig erforscht. Daher wurde die Quantifizierung von Proteoformen sowohl für isotopisch aufgelöste als auch für isotopisch unaufgelöste Massenspektraldatensätze untersucht. Proteoform-Quantifizierungen durch die "extrahierte Ionenflussgramm (EIF)-Strategie" und die "dekonvolutierte spektrenbasierte Strategie" wurden verglichen und auf Basis der Genauigkeit und Präzision bewertet. Darüber hinaus konnte erstmals die Korrelation der Quantifizierungsergebnisse zwischen verschiedenen Dekonvolutions-Tools für isotopisch unaufgelöste Spektren von Proteoformen gezeigt werden. Das optimierte FIA-MS-Protokoll sowie die mit Modellproteinen entwickelte Quantifizierungsstrategie wurden auch für die Analyse der Proteoformen des klinisch relevanten TP-Adalimumab angewendet. Die Detektion der weniger häufig vorkommenden Adalimumab-Proteoformen stellte eine größere Herausforderung dar, da die m/z-Verteilung dieser Proteoformen im Massenspektrum überlappt. Für die Fraktionierung der weniger häufig vorkommenden Proteoformen in der TP-Probe wurde die Anwendung eines Offline-Ansatzes der "Sample Displacement Batch Chromatography (SDBC)" getestet. Der Erfolg der Anwendung der SDBC-Methode konnte mittels der Quantifizierung der fraktionierten Adalimumab-Proteoformen gezeigt werden. Zusammenfassend lässt sich sagen, dass ein neuartiger Ansatz zur schnellen Quantifizierung von Proteoformen mittels FIA-MS-Methode mit einer Gesamtanalysezeit von weniger als 4 min pro Probe entwickelt wurde und deren Brauchbarkeit mit einer Adalimumab-Probe nachgewiesen wurde. Die in dieser Arbeit entwickelte schnelle Quantifizierungsmethode kann auf verschiedene TPs angewendet werden.

2 Introduction

2.1 Concept of proteoforms

The central dogma of molecular biology explains the flow of genetic information from DNA to RNA to proteins which are composed of amino acid residues. A simplified model of ‘one gene to one protein’ was hypothesized back in 1941 by US geneticist George Beadle (Beadle and Tatum, 1941). However, with advances in bioanalytical techniques, a discrepancy between the number of genes and the resulting number of proteins was soon realized. The genome was recognized to be nearly static. On the other hand, the proteome-defined as a set of proteins expressed limited set of genes (Wilkins *et al.*, 1996), gained importance as key factors governing the cell-specific functions. Moreover, the proteome was recognized to be highly dynamic and complex. Cascading events of transcription and translation were recognized to be generating numerous variants in resultant protein products (Lalley and Shows, 1974). Natural loss of function or gain of function mutation in a gene resulted in altered proteins with frameshifts, amino acid changes (Cohen, 1988) (Liu, Watson and Zhang, 2015) (Albalat and Cañestro, 2016). Events like mRNA pre-processing and alternate splicing lead to more protein isoforms/species (Chang *et al.*, 1999) (Climente-González *et al.*, 2017). More protein species encoded by the same gene but harbouring varied post translational modifications (PTMs) were recognized (Hirn *et al.*, 1983) (Mann and Jensen, 2003). The scientific community now has accepted and acknowledged that one gene encodes for multiple protein products. Altered protein products arising from the same gene were referred to in multiple ways as “protein isoforms” (Cohen, 1988) (Misek *et al.*, 2005) (Blakeley *et al.*, 2010), protein species (Schlüter *et al.*, 2009), or even protein variants. There was an inconsistency in the nomenclature of protein products resulting from various mutations and modifications (Agarwal *et al.*, 1975) (Parekh *et al.*, 1989) (Blakeley *et al.*, 2010).

To harmonize the nomenclature system, the term “proteoforms” was coined in 2013. Proteoforms are defined as “all of the different molecular forms in which the protein product of a single gene can be found, including changes due to genetic variations, alternatively spliced RNA transcripts and post-translational modifications” (Smith and Kelleher, 2013). The concept of proteoforms is identical to its predecessor term- “protein species” (Schlüter *et al.*, 2009). A detailed account of the influence of various modifications in the diversification of human proteoform profile is detailed by Nielsen, Savitski and Zubarev, 2006 and Jungblut *et al.*, 2008.

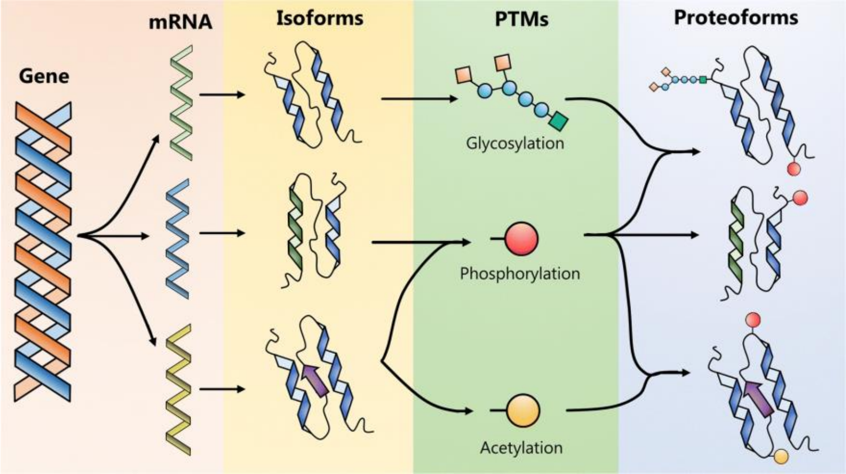


Figure 1: Scheme shows proteoforms encoded by a typical gene. Adapted from (Gil et al., 2019)

Nielsen and colleagues, in their research, forecasted that approximately 50 proteoforms exist behind every gene (Nielsen, Savitski and Zubarev, 2006). With the advances in analytical technologies, the complexity of human proteoforms is estimated to be far more. It was predicted that approximately 250,000 proteoforms exist per cell type, estimating proteoforms in human proteome about one billion (Kelleher, 2012).

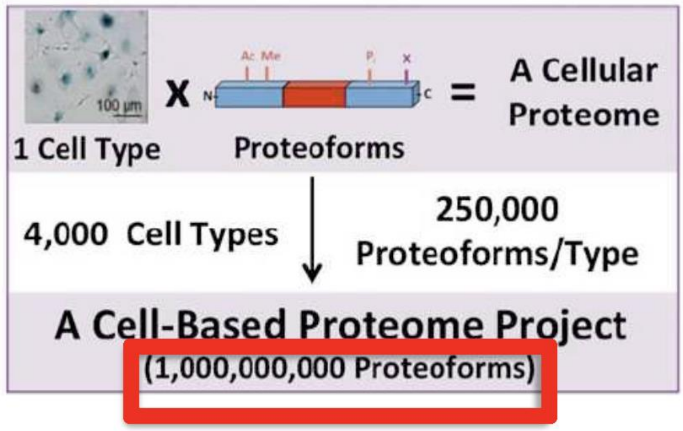


Figure 2: An estimated total number of human proteoforms in non-diseased cells that can be detected with the technological advances in mass spectrometry (assuming there are 250,000 proteoforms per cell type). Figure adapted from Kelleher, 2012

With increasing research in the field, the functional significance of proteoforms in molecular processes is being realized more evidently. Roles of proteoform level changes and disease progression have been reported in the field of oncology, neurodegeneration, cardiovascular disorders (Climente-González *et al.*, 2017) (Tucholski *et al.*, 2020). For example- A dysregulation of phosphorylated proteoforms in cardiac troponin complex can cause altered muscle contractility, which is further associated with heart failure (Peng *et al.*, 2014). Many more relevant cases for proteoform as disease diagnostic markers have been reviewed and summed up by Steffen *et al.*, 2016. For an understanding of key factors in biomolecular process and thereby enable the design of targeted therapies, not only a proteome level but also a deeper proteoform level understanding is crucial (Tiambeng *et al.*, 2019).

2.2 Therapeutic proteins

Artificial transfection of a gene encoding for 'protein of interest' in an expression host system was made possible with the use of recombinant DNA technology. Combined with advances in bioprocessing, this technology facilitated large-scale production and purification of clinically relevant recombinant proteins. Recombinant human insulin became the first protein used as 'therapeutic protein' (TP) for the treatment of diabetes (Nossal, 1980). Since its introduction in the 1980s, the production and use of therapeutic proteins (TPs) have increased exponentially. Today, TPs are available in form of growth factors, hormones, enzymes, coagulation factors, plasma proteins, and many other formats, but majorly dominated by monoclonal antibodies (Ecker, Jones and Levine, 2015).
However even to date, developing a TP for clinical use is a very complex process starting from host engineering for acquiring high protein yields up to controlled and regulated bioprocessing operations for obtaining safe drug products (Jamrichová *et al.*, 2017).

Briefly, the manufacturing process of TPs starts with lab-scale culturing engineered host cell lines producing TP of interest. Glycoengineering of host cell lines to obtain the desired glycosylation pattern on TP is a field of growing interest in bioprocess development. *Escherichia coli*, *P. pastoris*, *S. cerevisiae*, Chinese Hamster ovary (CHO) are among the most popularly used engineered host cell systems (Jayapal *et al.*, 2007) (Rosano and Ceccarelli, 2014). The culturing step is further scaled up to large production bioreactors, followed by series of bioprocessing steps for harvesting, purifying, and formulating of desired TP.

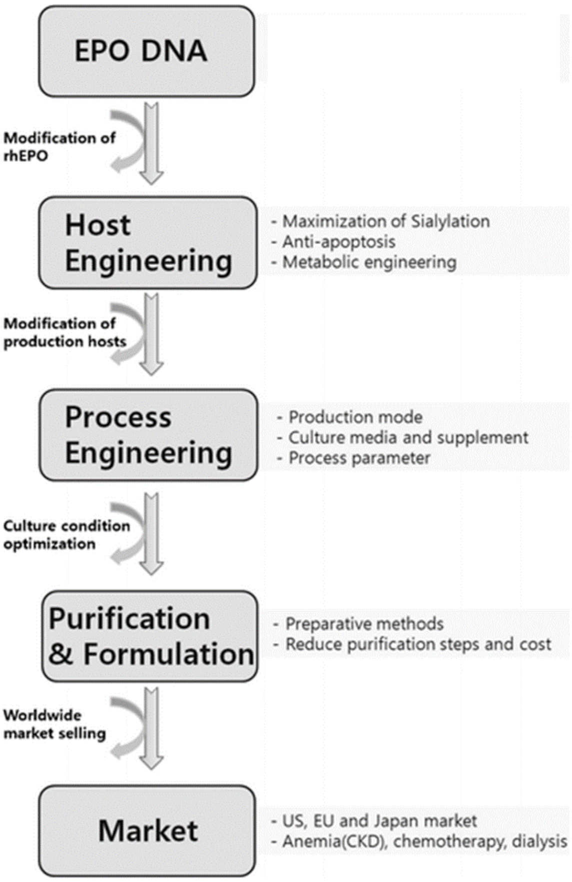

EPO DNA
Modification of rhEPO
Host Engineering
- Maximization of Sialylation
- Anti-apoptosis
- Metabolic engineering
Modification of production hosts
Process Engineering
- Production mode
- Culture media and supplement
- Process parameter
Culture condition optimization
Purification & Formulation
- Preparative methods
- Reduce purification steps and cost
Worldwide market selling
Market
- US, EU and Japan market
- Anemia(CKD), chemotherapy, dialysis

Figure 3: Key steps in the production of recombinant erythropoietin (in blocks), along with critical factors that need to be regulated at respective steps. Figure reproduced from (Lee et al., 2012).

2.2.1 Proteoforms in therapeutic proteins

Unlike, chemically synthesized small molecule drugs, TPs yielded by engineered host cells are extremely heterogeneous. An inherent heterogeneity in produced TPs can be due to biomolecular processes supported by the respective host cell machinery. This heterogeneity is reflected in proteoforms with varied disulphide bond formations or with varied post translational modifications (PTMs) patterns (Zhang, Moo-Young and Chou, 2010) (Yu *et al.*, 2020). An example of varied N glycosylation patterns obtained in recombinant human monoclonal antibodies is demonstrated in the figure below.

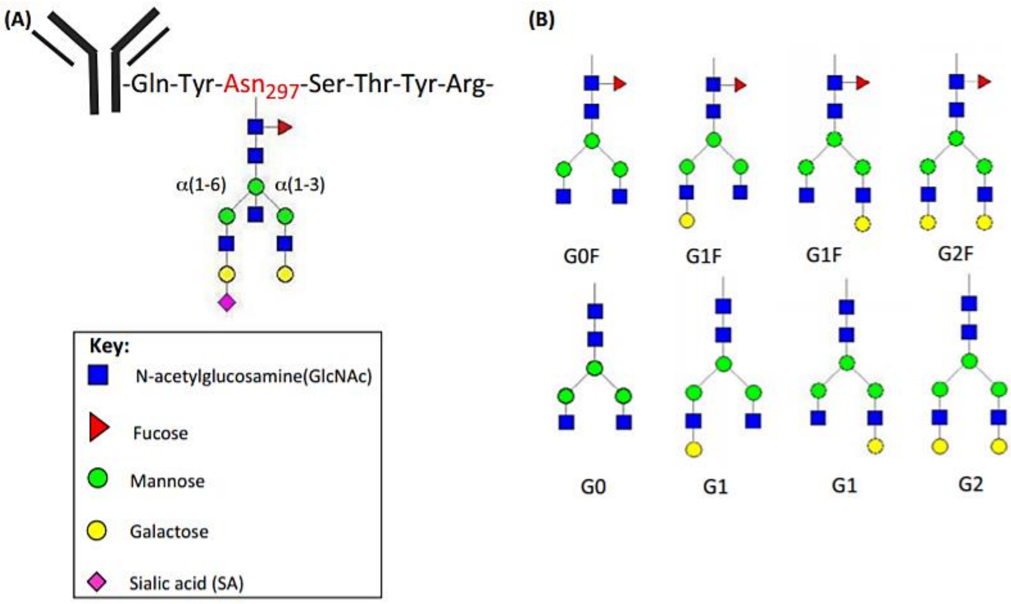


Figure 4: Scheme of therapeutic monoclonal antibody A) showing glycosylation at asparagine 297 with the respective glycan structure. B) Commonly observed N glycans structures in recombinantly produced monoclonal antibody (IgG1). Adapted from (Sha et al., 2016).

The heterogeneity profile of recombinant TPs is also influenced by conditions in bioprocess operations like viral inactivation with low pH hold, buffers in protein capturing, and purification (Chung et al., 2018). Some prominent bioprocess conditions responsible for influencing the heterogeneity of TP are detailed in following table.

Table 1: Proteoforms present in recombinant therapeutic proteins due to the different conditions in the production process (upward arrow ↑ signifies increase, downward arrow ↓ signifies decrease). Content reproduced from Chung et al., 2018.

Modification in recombinant TP	**Bioprocess conditions**	**Reference**
Acidic variants/ proteoforms		
Glycation	Glucose in culture media→ ↑	Yuk et al. (2011)
Deamidation	Hold temperature in downstream processing steps→ ↑	Diepold et al. (2012)
	Hold pH in downstream processing steps→ ↑	Pace et al. (2013)
	pH in downstream processing steps → ↑; Temperature → ↑	Yang et al. (2016)
	pH in downstream processing steps → ↑	Xie et al. (2016)
	Sucrose in culture media → ↓	Stratton et al. (2001)
Oxidation	Hold temperature in downstream processing steps → ↑	Lam et al. (1997)
	Manganese in culture media→ ↓	Hazeltine et al. (2016)
	Tryptophan in culture media→ ↓	
	Cysteine in culture media→ ↑	
	Copper in culture media → ↓	
	Iron in culture media → ↑	Vijayasankaran et al. (2013)
	Epigallocatechin gallate in culture media→ ↓	Hossler et al. (2015)
	Rutin in culture media → ↓	
Cysteine variants	Copper in culture media → ↓	Trexler-Schmidt et al. (2010)
	pH in downstream processing steps → ↑	Xie et al. (2016)
Basic variants/ proteoforms		
C-terminal lysine truncations	Zinc in culture media → ↓	Luo et al. (2012)
	Copper in culture media → ↑	
	Temperature in production bioreactor → ↓	Zhang et al. (2015)
C-terminal amidation	Copper in culture media → ↑	Kaschak et al. (2011)

Scientific literature published especially in bioprocessing area has addressed heterogeneity associated with TP with multiple terms like protein isoforms (Sugihara *et al.*, 2018), protein charge variants (Ebersold and Zydney, 2004), protein species (Rosenlöcher *et al.*, 2016), micro-heterogeneities in proteins (Beck and Liu, 2019). All the various terms stated above can be collectively addressed under the umbrella term- proteoforms of TPs, and is used so in the rest of the thesis.

2.3 Importance of proteoform quantification in therapeutic proteins

Modifications associated with TP like phosphorylations, oxidations, acetylations, and especially glycosylations among others, can play a critical role in governing the efficacy of respective TP (Kayser *et al.*, 2011) (Pawlowski *et al.*, 2018). Controlling the proteoform profile of TPs or regulating critical quality attributes of proteoforms in TPs, throughout the scaled-up process is a huge bottleneck in biopharmaceutical operations. To guarantee a safe and effective TP product, having a homogenous proteoform composition of a desired active TP formulation would be ideal case scenario. For example, all protein copies with a core fucosylated biantennary N-glycan at position 297 on each of the heavy chain arm, no undesired deamidation in complementarity-determining regions (CDRs) would be the most desired proteoform composition of recombinant Adalimumab (Raju, 2008). To achieve this ideal state, multiple purification operations capable of removing of all other undesired proteoforms would be needed. An example of common downstream purification steps used in production of recombinant monoclonal antibody is shown in the figure below.

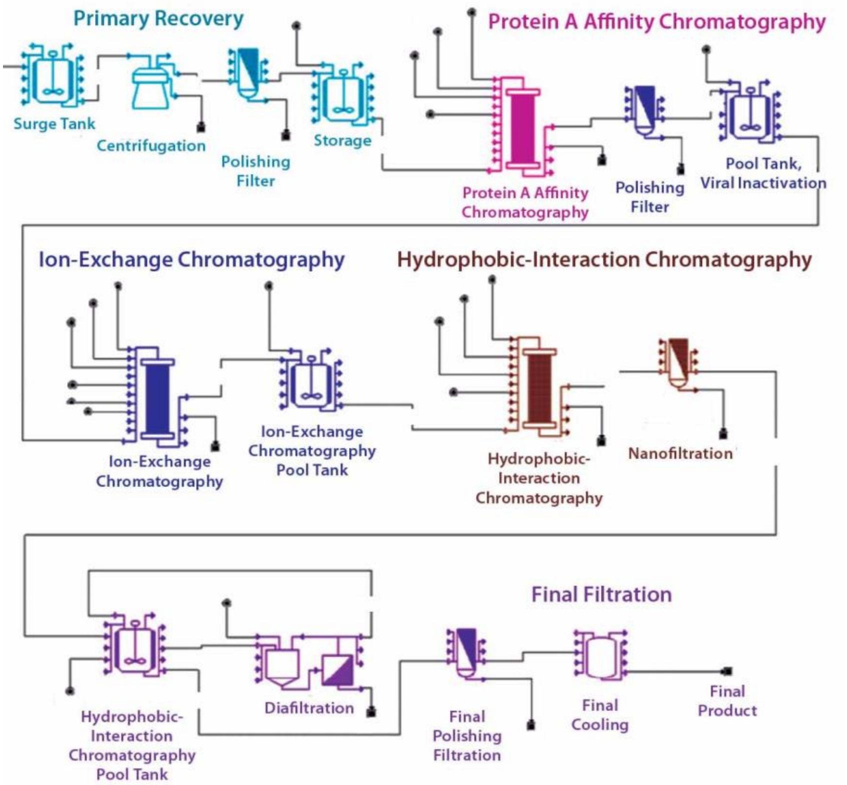

Primary Recovery
Surge Tank
Centrifugation
Polishing Filter
Storage
Protein A Affinity Chromatography
Protein A Affinity Chromatography
Polishing Filter
Pool Tank, Viral Inactivation
Ion-Exchange Chromatography
Ion-Exchange Chromatography
Ion-Exchange Chromatography Pool Tank
Hydrophobic-Interaction Chromatography
Hydrophobic-Interaction Chromatography
Nanofiltration
Hydrophobic-Interaction Chromatography Pool Tank
Diafiltration
Final Filtration
Final Polishing Filtration
Final Cooling
Final Product

Figure 5: Flowchart of commonly used processes in the purification of recombinant monoclonal antibody (Franzreb, Muller and Vajda, 2014)

In the biopharmaceutical industry, the basis for the design of effective purification processes is the differences in physiochemical, biochemical properties of TP and other contaminating side products or also its proteoforms. For recombinant monoclonal antibodies, the first step in purification is generally a specific protein A affinity chromatography (Duhamel *et al.*, 1979). It is followed by ion-exchange chromatography for removal of other contaminating proteins, DNA impurities, highly acidic proteoforms in TP, and then by hydrophobic-interaction chromatography for removal of proteoforms based on hydrophobicity differences(Fraud *et al.*, 2009).

Though the above-mentioned purification steps keep the major contaminants in check, these purification steps cannot eliminate all undesired proteoforms. This is because capturing small differences in physiochemical properties of proteoforms from TPs for selective purification is extremely challenging (Walsh and Jefferis, 2006). The resultant efficacy of TPs is strongly affected if excess amounts of undesired proteoforms are present in patient-administered doses. For example, excess of mAb proteoforms bearing high mannose glycans leads to reduced therapeutic efficacy of the TP due to high clearance rate of mAb from serum (Goetze *et al.*, 2011). Excess of trastuzumab proteoforms harbouring isomerized aspartate 102 on their heavy chain is proved to lower the potency of the administered TP (Harris *et al.*, 2001). Further on, severe immunogenic side effects are also reported in patients administered with recombinant TP containing higher allelic variants (Leal *et al.*, 2013).

In all, this suggests that amounts of proteoforms need to be strictly monitored throughout bioprocessing and especially in formulated TP products. To determine the success of the bioprocessing and purification steps, not only qualitative but an accurate quantitative proteoform profile of TP is a must. To achieve this purpose, ultimately, the necessity of a sensitive and fast bioanalytical technique for proteoforms quantification is highlighted.

2.4 Conventional bioanalytical techniques for primary protein analysis

Traditionally ligand binding assays (LBA) have been the gold standard in the biopharmaceutical industry for qualitative and quantitative analysis of TPs (Mayer and Hottenstein, 2016). Generic ligand binding assays like enzyme-linked immunosorbent assay (ELISA) or western blots use specific antigens to detect epitopes on the protein of interest (Stubenrauch, Wessels and Lenz, 2009). LBA especially have many applications in immunogenicity testing during drug development as well as in pharmacokinetics and pharmacodynamics studies (Marini *et al.*, 2013). The quantitative results given by LBA, however, may be biased, depending on the affinities of binding partners and there are fewer possibilities of validating these quantification results (Neubert *et al.*, 2018).

Gel-based electrophoresis techniques like sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS PAGE) is commonly used for qualitative protein analysis, also in between purification steps of TPs. SDS PAGE is generally followed by densitometry analysis for quantification of recombinant proteins. However, the quantification is mostly limited only to proteoforms or proteins showing a large difference in molecular weight (Miles and Saul, 2005). Two-dimensional gel electrophoresis offers an upper hand in the sensitivity of analysis because differences in both charges, as well as the size of proteins, are utilized in the analytical process. However, ultimately the quantification with gel electrophoresis or western blot relies on non-regulated workflows of densitometry analysis (Gassmann *et al.*, 2009). UV or fluorescence detection-based capillary electrophoresis (CE) is yet another popular technique in protein analysis, especially TPs (Swinney and Bornhop, 2000). The quantitative peak area response with UV detection is however comparatively low and affected by the interference from buffer components used. The fluorescence-based TP detection in CE is widely used but largely dependent on dependent on chromophore derivatization (Guzman *et al.*, 1992).

On other hand, mass spectrometry (MS) based detection forms of the gold standard in terms of TP analysis in both qualitative and quantitative aspects (Lewis *et al.*, 1994) (Gong *et al.*, 2014). MS-based analysis is also proven to outclass conventionally used analytical techniques in terms of speed & accuracy (Kopp *et al.*, 2020). In mass spectrometry, different proteins or even proteoforms can be discriminated based on mass differences.

2.5 Mass spectrometry for protein analysis

MS analysis of proteins is usually performed in positive mode, wherein ionizing molecules acquire positive charge and detected ions are reflected in form of mass to charge (m/z) ratios in the mass spectrum. Soft ionization techniques like matrix-assisted laser desorption/ionization and electrospray ionization (ESI) are most popularly used for protein analytics (Hillenkamp and Karas, 1990) (Fenn *et al.*, 1989). The ionization process is followed by the separation of these charged ions in the mass analysers, under high vacuum pressure. Further on, the charged protein ions (called precursor ions) can be acquired in full scan mode referred to as MS^1 spectrum. These precursor ions can be further dissociated into fragment ions, in so-called MS^2 scans. Fragment ions generated in MS/MS spectrum allow the analyst to perform identification of the ionized TP molecules (Hoffmann, 1996).

Mass spectrometry-based TP analysis can be performed as a bottom-up approach, middle down approach and top-down approach.

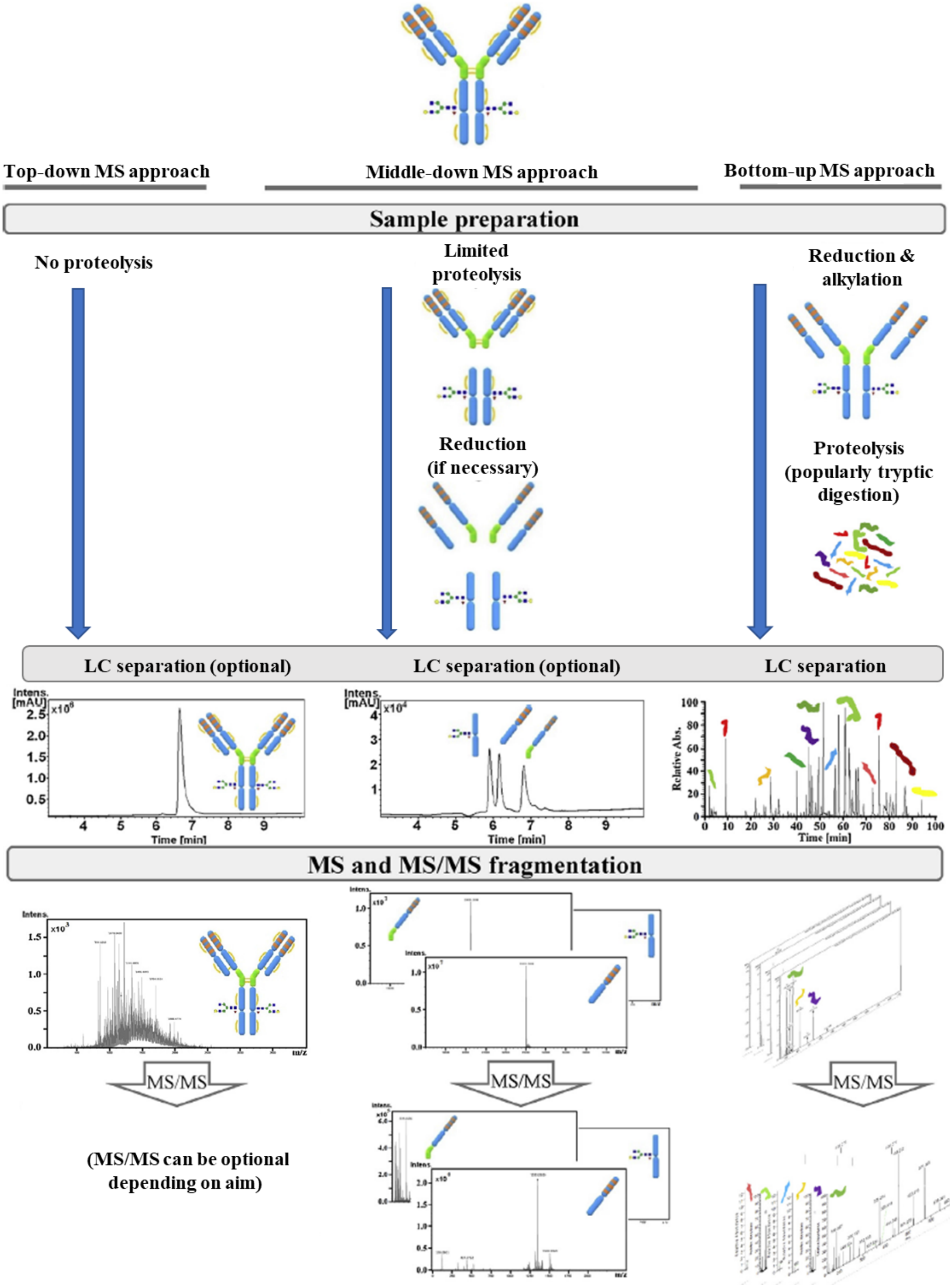


Figure 6: Scheme showing different MS-based approaches used in protein analysis. Figure reproduced from Háda et al., 2018.

2.5.1 ‘Bottom-up’ MS approach

In the bottom-up approach, proteins to be analysed are digested to smaller peptides prior to analysis. The digestion of proteins can be achieved with several proteases, some of them being chymotrypsin, LysC, ArgC, AspN, and GluC, but most popularly performed with trypsin. A set of unique surrogate peptides (minimum 2) are used to represent the information for the whole protein.

Bottom-up MS approach is the most used as well as the validated approach in MS for qualitative as well as quantitative TP analysis. It is especially a state-of-the-art technique to confirm the amino acid sequence of the generated recombinant TP (Alexandridou *et al.*, 2009) (Song *et al.*, 2017).

2.5.2 ‘Middle-down’ MS approach

In middle-down MS approach, proteins are digested to longer polypeptides (more than 20-25 amino acid residues) prior to MS analysis. This approach is similar to the bottom-up MS approach involving digestion, but the size of the polypeptides is usually more than 4kDa (Boyne *et al.*, 2009). Often alternative proteases like LysC (leaves C-terminal side of lysine), GluC (C-terminal side of glutamic acid and aspartic acid residues), IdeS (consecutive glycine residues at the hinge region of immunoglobulin G) are used in middle-down MS approach (Pandeswari and Sabareesh, 2019).

Middle down MS is popularly used for qualitative and quantitative analysis of large TPs having complex PTM profile, free sulfhydryl groups, methionine oxidation (Faid *et al.*, 2018) (Pipes *et al.*, 2010) (Young *et al.*, 2010).

2.5.3 ‘Top-down’ or ‘intact protein MS’ approach

‘Top-down MS’ or ‘intact protein MS’ approach involves analysis of full-length proteins/proteoforms without digestion into peptides. Theoretically, this MS approach provides wholesome information of protein along with its associated PTMs. Additionally, sample preparation in ‘intact protein MS’ approach is less tedious and has fewer steps than bottom-up or middle down MS workflow. However, unlike peptides which ionize at lower m/z range, intact proteoforms mainly require MS instruments with higher mass range. Moreover, the transmission of heavier proteoform ions requires increased pressures in the ion guides of MS system (Chernushevich and Thomson, 2004).

‘Top-down MS’ or ‘intact protein MS’ approach is typically used in biopharmaceutical industries only for intact mass conformation of produced TPs. MS/MS fragmentation of intact proteins has been demonstrated, but offers less protein coverage compared to the peptide fragmentation in bottom-up MS approach (Haverland *et al.*, 2017). This is currently a huge limitation for its wider application of 'intact protein MS' approach.

2.6 Quantification of proteoforms in therapeutic proteins

2.6.1 Limitations of conventional bioanalytical techniques for proteoform quantification

The significance of quantifying proteoforms in TPs was detailed in section 2.3. As stated earlier, the commonly used bioanalytical techniques for TP like ligand binding assays, gel electrophoresis offer only limited success in terms of proteoform specific quantification. Given the homologous amino acid sequences shared by proteoforms in therapeutic proteins, having specific proteoform-specific quantitative LBA is extremely difficult. With exception of special case scenarios, the quantitation reported in LBA is generally a total protein amount rather than of specific proteoforms. Two-dimensional western blotting is reported for proteoform detection and quantification but often suffers from drawbacks of tedious workflows and concerns about quantitative accuracy (Herzog *et al.*, 2020). Densitometry analysis based proteoform quantification used in ELISA or western blotting can only perform accurately when known reference standards are available for comparison and interpolation of quantitative data(Gassmann *et al.*, 2009). Quantification results in CE with UV or florescence-based detection also have limitations. Non-MS hyphenated proteoform quantification in CE can be affected by disturbances in the electroosmotic flow and can be rationalized only with the use of specific internal standards (Guzman *et al.*, 1992). The required proteoform-specific full-length internal standards are not available for most of the TPs.

For accurate distinction and quantification of proteoforms, the bioanalytical technique must be capable of detecting the minor differences in physiochemical or biochemical properties of proteoforms. Specifically, for quantification, ideally, the quantitative response obtained from the technique must give a direct correlation to proteoform concentrations and must work over a wide dynamic range.

Among other bioanalytical techniques, MS analysis offers the most regulated and accurate quantitative data evaluations. The linear dependence of the mass spectral signal on the injected concentration of analytes forms the basis of ESI-MS-based quantification (Whitehouse *et al.*, 1985).

2.6.2 Conventional quantitative mass spectrometry techniques applied for proteoforms

The current state of art mass spectrometry-based quantification is established and validated for the bottom-up MS approach. In bottom-up MS approach, quantification of proteoforms is given by the quantification of a typical surrogate peptide or unique peptide (Hagman *et al.*, 2008). Surrogate peptide-based proteoform quantification has been also extensively used in pharmacokinetic studies of TPs (Jenkins *et al.*, 2015), wherein the surrogate peptide chosen possesses the PTM of interest, representing typical proteoforms. Conventionally, surrogate peptide-based proteoform quantification uses tandem mass spectrometry techniques or the MS/MS level fragment ion data.

MS/MS data for bottom-up MS-based proteoform quantification can be acquired from precursor peptides, in either data-dependent acquisition (DDA) mode or data-independent acquisition (DIA) mode. In DDA mode, as the name suggests, the selection of precursor ions for fragmentation is dependent on the signal intensity of peptide ions. On the other hand, in DIA mode, the precursor ions falling within certain isolation windows are fragmented, and this process is independent of the precursor intensities in MS^1 scan (Egertson *et al.*, 2015). Such quantification techniques are majorly used in differential proteomics studies, for example, to report significant changes in proteome for biomarker discovery (Song *et al.*, 2017).

When precursor ions of interest are known, the list of ions to be fragmented can also be submitted to MS as an ‘inclusion list (Kalli *et al.*, 2013). The analyst defined peptide ion-based quantification, which can be performed as SRM i.e. selected reaction monitoring or MRM i.e. multiple reaction monitoring, or PRM i.e. parallel reaction monitoring (Ronsein *et al.*, 2015).

Further on, incorporation of internal standards, typically a stable isotope labelled peptides also allows absolute quantification of protein with bottom-up MS approach (Tu *et al.*, 2014). Absolute quantification of the trastuzumab proteoforms has been widely reported quantification of signature/surrogate peptides containing modification of interest like-deamidation, isoaspartate, intermediate succinimide respectively (Bults *et al.*, 2016). A sample workflow for such proteoform quantification is seen in figure below.

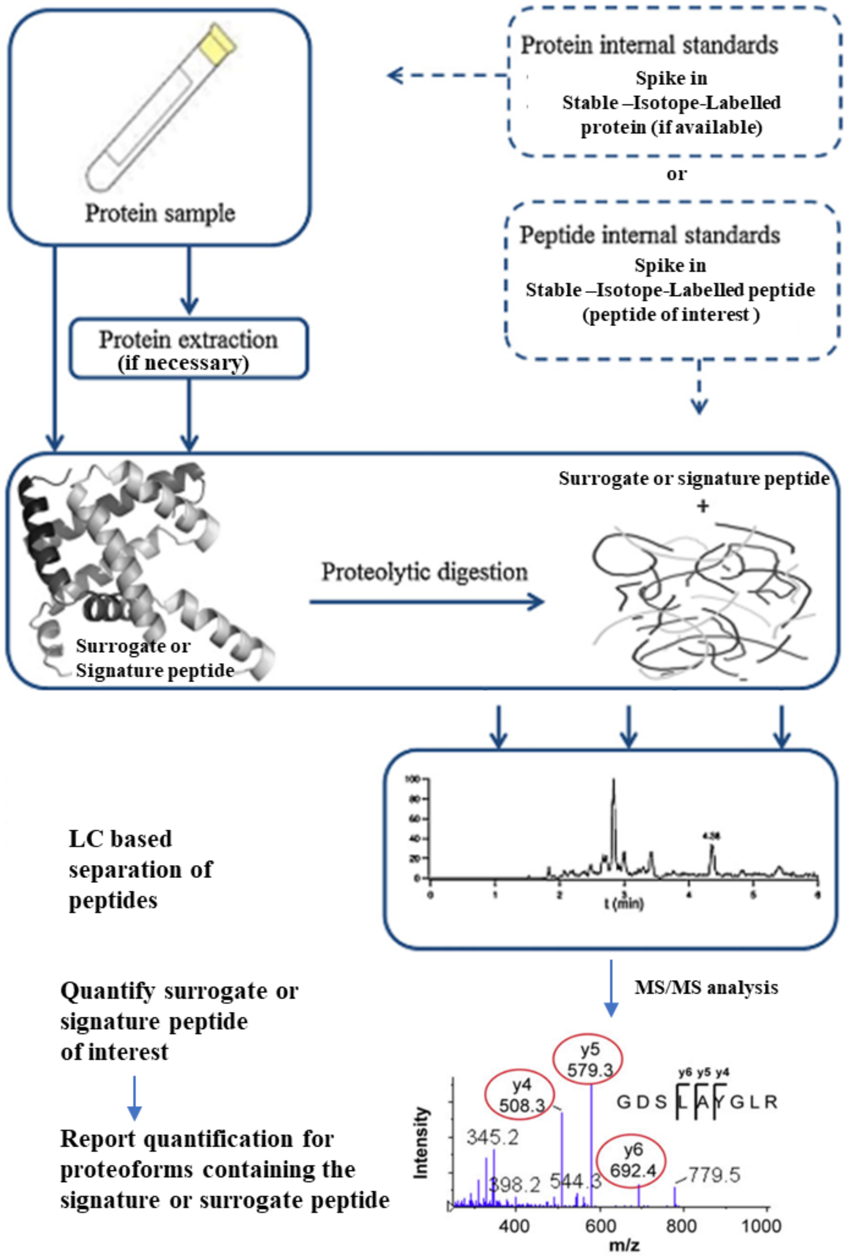


Figure 7: Workflow for surrogate peptide-based absolute quantification of therapeutic protein using liquid chromatography coupled to tandem mass spectrometry. Figure reproduced from Bronsema, Bischoff and Van de Merbel, 2012.

MS^1 ion current based reliable surrogate peptide-based proteoform quantification is now gaining popularity, due to advances in the resolution capacity for bottom-up MS approach (Tu *et al.*, 2014). However, such a surrogate peptide-based quantification involves considerable effort and time in chromatographic method development. It is necessary to guarantee that surrogate/signature peptide is quantified without any interference from other unwanted peptides from the same or other contaminating proteins (Kamiie *et al.*, 2008).

2.6.3 Significance of ‘intact protein MS’ approach for proteoform quantification

In most cases proteoforms are decorated with multiple PTMs and a set of proteoforms can often share a common PTM. The presence of typical PTM associated with TP can have a complementary downstream effect on the presence of other PTM(s) on different sites of same proteoform. For example, Bush et al. demonstrated that relative amounts of deamidation were related to methionine loss as well as glycan structure in human interferon-β1 (Bush *et al.*, 2016). The glycan structure associated with recombinant human interferon-β1, further influences the serum half-life and thereby efficacy of the TP. In such cases, it becomes necessary to quantify proteoform retaining information of multiple PTMs associated with it.

Bottom-up MS based quantification approaches being dependent on surrogate peptide detection, fail to preserve the wholesome information associated with a proteoform. It is not possible to trace the origin of a surrogate peptide to the full-length proteoform. (Song *et al.*, 2020). On other hand, intact protein MS provides a more accurate report of the inherent nature of the protein in the sample (Zhang *et al.*, 2018). Thus, intact protein MS becomes a requisite for accurate proteoform quantification.

2.7 Challenges in ‘intact protein MS’ approach for proteoform quantification

Challenges in ‘intact protein MS’ for sensitively detecting proteoforms start already from the ionization of proteoforms ions. Incomplete desolvation leads to the formation of wider ions signals from intact proteoforms than the theoretically expected isotopic distributions (McKay *et al.*, 2006) (Lu *et al.*, 2015). Effective ionization and transmission are important prerequisites for proteoforms quantification, but difficult for many conventionally used MS systems (Heck and Van Den Heuvel, 2004) (Wang *et al.*, 2017). Another level of challenge for proteoform detection is the lower abundance or low copy number of these proteoforms (Aebersold *et al.*, 2018). The limited sampling rate of proteoforms is also an important factor in ‘intact protein MS’ methods.

‘Intact protein MS’ approaches applied for proteoform identification and quantification primarily rely upon mass difference as a measure for distinguishing or resolving proteoforms. However, the resolution of intact proteoform ions in mass spectra is another bottleneck in ‘intact protein MS’(Lössl, Snijder and Heck, 2014). The resolution acquired by proteoforms signals strongly relies on the molecular weight of proteoform in consideration and kind of mass analyser in MS systems. The resolution capacity of different mass analysers for intact proteoforms is depicted in the figure below.

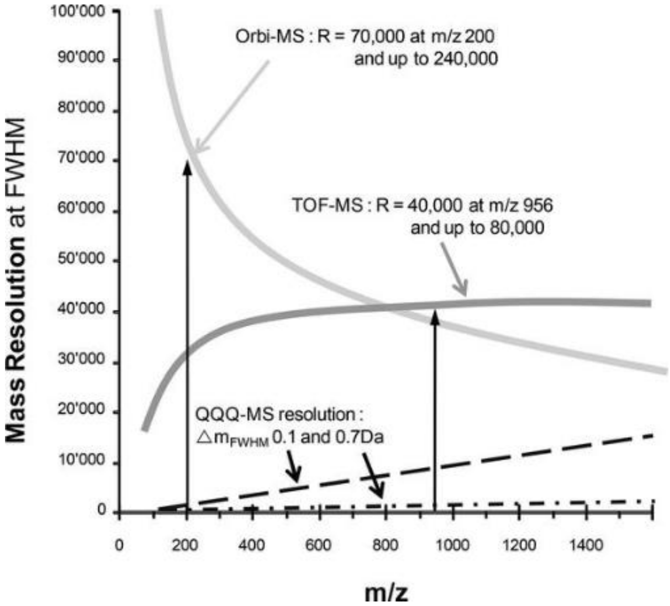


Figure 8: Plot representing mass resolution as a function of m/z range for three most commonly used mass analysers in MS instruments. Adapted from (Rochat, 2019).

The MS systems have more resolution capacity for proteoforms ionizing in the lower m/z range. The resolution capacity especially suffers above 3000m/z for large proteoforms, for most conventional mass analysers (Rochat, 2019).

An additional level of challenges is also present for analysis of ‘intact protein MS’ data especially in the case of ESI-MS, as signals of proteoforms are distributed across multiple charge states. The multiple charges acquired by a typical proteoform in ESI-MS is commonly referred to as the charge envelope of the proteoform. Each charge state in charge envelope can provide a measurement of proteoform mass and respective intensity (Lu *et al.*, 2015). In presence of an organic solvent, a mAb can acquire more than +60 charges. The data analysis of proteoform suffers due to the spread of charge state distribution and differs based on the number charge states considered in the analysis. An example of ESI-MS analysed mAb, showing the spread of proteoform signals across multiple charge states is shown in the figure below.

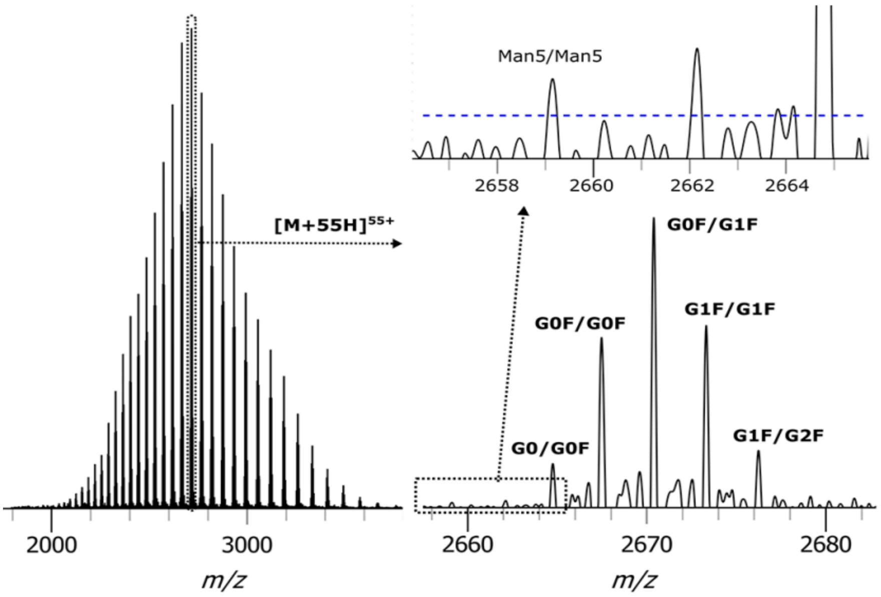


Figure 9: Mass spectrum of an intact monoclonal antibody with analysed ESI-MS highlighting charge state distribution obtained in the analysis. The expanded view of charge state 55 shows signals from different proteoforms. Adapted from (Srzentić et al., 2020)

In terms of ‘intact protein MS’ data analysis, the process of converting the multi-charge m/z data obtained in ESI-MS analysis of proteoforms to a zero-charge mass value is called deconvolution. A simple scheme of deconvolution process for intact proteoform MS data is shown in the figure below.

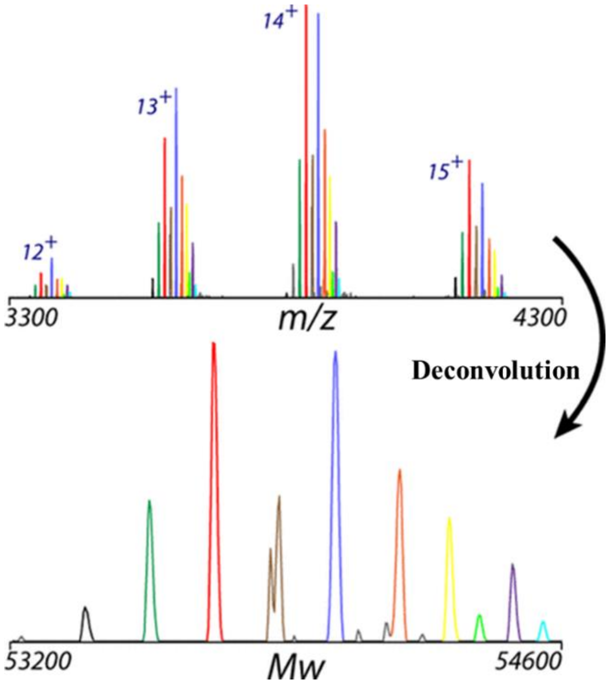


Figure 10: Deconvolution process of m/z signals from intact protein full scan mass spectrum into mass. Adapted from (Bern et al., 2018)

Deconvolution is a complex process involving series of steps like deisotoping, decharging, pattern matching (optional for limited algorithms), feature finding (Jeong *et al.*, 2020). The number of deconvolution tools/softwares providing reliable intact MS data analysis is currently limited. Moreover, the quantification process in these softwares lacks transparency. Thereby there is less possibility to validate the accuracy of quantified intact MS data. It is necessary to resolve or address these challenges for intact protein MS-based proteoform quantification.

Table 2: Table briefing the challenges towards achieving quantification of proteoforms using intact protein mass spectrometry.

Challenges towards achieving intact proteoform quantification using mass spectrometry		
Level	**Challenge**	**Cause**
Mass spectrometry analysis		
Sensitivity	Signal to noise ratio of proteoforms is very low Dynamic range among proteoforms in sample	Incomplete desolvation in ESI-MS, dilution of signal intensity across multiple charge states
Specificity	Proteoform signals interfering in same m/z window; Varying peak width across charge state	Limited by the capacity of mass analyser to isotopically resolve (especially large) proteoforms
Data analysis		
Deconvolution softwares	Commercial softwares are black-box model & expensive; Limited open-source softwares options	Unable to handle isotopically unresolved MS data
Quantification	No universally accepted quantification methodology; Less options for validating & scoring results	Limited research published on underlying algorithms; The complexity of the MS data

3 Aim

With the increased demand for therapeutic proteins (TPs), there is also an increased need to provide speed and accuracy in analytical testing methods for TPs. Setting up a fast quantitative analytical method can aid in quick decision-making for ongoing bioprocessing operations of TP in production. Thus, the major aim of this work was to establish, a sensitive but fast ‘intact protein MS or top-down MS’ method for the quantification of full-length proteoforms in therapeutic proteins. Initially, the fast MS analysis method was targeted to be developed with model protein Ovalbumin.

Quantification of all proteoforms in the TP sample, without any losses in the testing process, is the most ideal bioanalytical scenario. Thus, the focus of fast MS method establishment was also to quantify maximum proteoforms in TP sample having distinct masses, irrespective of the detailed structural identification. Later, for data obtained in the fast MS method, assessing the accuracy of proteoform quantification across different data processing strategies was yet another objective of the thesis. The quantitative accuracy was also aimed to be tested for different proteoforms giving isotopically resolved and isotopically unresolved MS data.

The ‘fast MS’ method developed with model protein was intended to be transferred and applied for proteoform quantification in therapeutic protein-Adalimumab. Finally, theorized application of sample displacement batch chromatography (SDBC) for proteoform fractionation was also planned to be evaluated by using the fast quantitative MS method established during this work.

4 Materials

Table 3: List of materials and consumables used for the experiments.

Chemicals	**Manufacturer**
Acetonitrile	Merck KGaA, (Darmstadt, Germany)
Ammonium acetate	Merck KGaA, (Darmstadt, Germany)
Dithiothreitol	Sigma Aldrich (St. Louis, Missouri, USA)
Eshmuno CPX resin	Merck Millipore (Burlington, Massachusetts, USA)
FabRICATOR (IdeS)	Genovis (Lund, Sweden)
Formic acid	Fluka, Fisher Scientific GmbH (Schwerte, Germany)
Iodoacetamide	Sigma Aldrich (St. Louis, Missouri, USA)
Sequencing grade modified trypsin	Promega (Madison, Wisconsin, USA)
Sodium chloride	Sigma-Aldrich, Merck KGaA (Darmstadt, Germany)
Sodium deoxycholate Sigma	Aldrich (St. Louis, Missouri, USA)
Sulfolane	Sigma-Aldrich, Merck KGaA (Darmstadt, Germany)
Triethylammonium bicarbonate	Thermo Fisher (Waltham, Massachusetts, USA)
MS grade water	Merck KGaA (Darmstadt, Germany)
Chromatography columns	**Manufacturer**
Acquity UPLC® Peptide BEH C18 Column, 75 µm x 200 mm	Waters (Milford, Massachussets, USA)
Acquity UPLC® Symmetry C18 Trap Column, 180 µm x 20 mm	Waters (Milford, Massachussets, USA)
ProSwift™ RP 4H analytical, 1 x 250 mm	Thermo Fisher (Waltham, Massachusetts, USA)
ProSwift™ RP 4H analytical, 1 x 50 mm	Thermo Fisher (Waltham, Massachusetts, USA)
Disposables	**Manufacturer**
Amicon Ultra 0.5 ml 10K centrifugal filters	Merck Millipore (Billerica, Massachussets, USA)
Amicon Ultra 4 ml 10K centrifugal filters	Merck Millipore (Billerica, Massachussets, USA)
Pipette tips	Sarstedt (Nümbrecht, Germany)
Reaction tubes	Sarstedt (Nümbrecht, Germany)
Total recovery sample vials	Waters (Milford, Massachussets, USA)
Instruments	**Manufacturer**
Acquity UPLC	Waters (Milford, Massachussets, USA)
Agilent 1200 series	Agilent (Santa Clara, California, USA)
Äkta prime plus fractionator	GE Healthcare (Chicago, Illinois, USA)
Analytical scale ALS 120	Kern & Sohn GmbH (Balingen, Germany)
Centrifuge 5424	Eppendorf (Hamburg, Germany)

Dionex Ultimate 3000	Thermo Fisher (Waltham, Massachusetts, USA)
Elute LC	Bruker Daltonics Inc. (Billerica, Massachusetts, USA)
maXis II™	Bruker Daltonics Inc. (Billerica, Massachusetts, USA)
Microplate reader	Tecan Life Sciences (Männedorf, Switzerland)
Q Exactive™ Hybrid Quadrupole Orbitrap™ Mass Spectrometer	Thermo Fisher (Waltham, Massachusetts, USA)
Speedvac	Thermo Fisher (Waltham, Massachusetts, USA)
Thermomixer compact	Eppendorf (Hamburg, Germany)
Proteins	**Providers**
Adalimumab	University of Natural Resources & Life Sciences (Vienna, Austria)
Filgrastim	CinnaGen Co. (Tehran, Iran)
Erythropoietin	CinnaGen Co. (Tehran, Iran)
Ovalbumin (A5503)	Sigma-Aldrich, Merck KGaA (Darmstadt, Germany)
Softwares	**Developer**
FLASHDeconv	OpenMS (Kohlbacher lab)
MetaUniDec	UniDec suite (Michael T. Marty Lab group)
ProteomeDiscoverer 2.0	Thermo Fisher (Waltham, Massachusetts, USA)
ReSpect™,	Thermo Scientific (Waltham, Massachusetts, USA)
Skyline 20.1	MacCoss Lab Group
UniDec	UniDec suite (Michael T. Marty Lab group)

5 Methods

5.1 RPLC-MS for intact proteoform analysis

RPLC-MS for intact protein analysis was demonstrated for model protein Ovalbumin on a polymeric monolithic analytical RP ProSwift™ RP 4H, 1 x 250 mm (Thermo Scientific™). A stock solution of 10000ng/µL Ovalbumin was prepared in MS grade water which was further diluted to have 1000ng/µL working stock. 3000ng sample (3 µL of working stock) was injected onto the monolithic RP column. Solvents used to create binary gradient were 0.1% FA water as mobile phase A and 0.1%FA acetonitrile (ACN) as mobile phase B. The column was operated at 30°C according to the recommendations of the column manufacturer. The blanks following Ovalbumin injection comprised injection of 1uL of mobile phase A solution i.e., 0.1% FA water. A shorter gradient was used for blank injections following protein elution. The gradients used for the elution of protein and the following blank, are presented in Table 4a & Table 4b respectively.

Table 4: Binary gradient used in RPLC-MS analysis a) used for elution of protein b) used for blank runs following protein injection.

a) Gradient used for elution of protein		
Time (min)	%B	Flow rate(mL/min)
0.00	2	0.2
1.00	2	0.2
1.10	15	0.2
8.50	38	0.2
9.50	38	0.2
16.00	70	0.2
17.00	70	0.2
17.10	2	0.2
23.00	2	0.2

b) Gradient used for following blanks		
Time (min)	%B	Flow rate(mL/min)
0.00	2	0.2
1.00	2	0.2
1.10	30	0.2
8.50	70	0.2
10.00	70	0.2
10.50	2	0.2
15.00	2	0.2

The elute from RP column is directed into Q Exactive™ Hybrid Quadrupole Orbitrap mass spectrometer (Thermo Scientific™) where full scan mass spectrum was acquired in positive polarity with ESI. The scan range for Ovalbumin was 900 to 2800 m/z. The rest data acquisition parameters on Q Exactive™ MS were electrospray voltage 3.5 kV, S lens RF level 75 units, 35000 Orbitrap resolution, 3 microscans and AGC target 1e6. The RPLC-MS data was visualized in Xcalibur™ Software (Thermo Fisher Scientific). The intact protein MS data was further deconvoluted using a computational suite UniDec (ver.

4.2 from Marty et. al.). Data processing parameters for deconvolution in UniDec were set as charge range 7 to 32, mass range 38000 Da to 48000 Da, sample every 0.1Da. Background subtraction and smoothing of charge state distribution were enabled. Binning function was disabled. For peak selection and annotations of deconvoluted spectrum the peak detection range was set to 10 Da and the peak detection threshold was set to 0.06 units.

5.1.1 Inclusion list for MS/MS fragmentation

In a follow-up RPLC-MS run for Ovalbumin, signals eluting at 7 mins were further fragmented to find the identity of eluting species. To capture fragment ions, the scan range was readjusted from 500 to 4500m/z. The MS/MS spectra were acquired for specific signals using an inclusion list for fragment ions. The inclusion list provided for fragmentation was as follows 953.47 m/z (z=5), 971.23 m/z (z=4), 1056.77 m/z (z=4), 1089.03 m/z (z=4), 1094.52 m/z (z=4). The generated fragment spectra were further analysed using ProSightLite from Northwestern University (Thomas *et al.*, 2014).

5.2 Quantification of Ovalbumin recovered from monolithic RP column using bottom-up MS approach

5.2.1 Experimental setup for studying recovery of proteoforms from reverse phase column

Relative quantitation of Ovalbumin recovered form monolithic RP column was performed using the bottom-up MS approach. An Agilent 1200 series HPLC setup equipped with UV detector was used for this experiment. Two sample conditions were evaluated- 1) Specified amount of Ovalbumin eluting from RP column using binary gradient (labelled as RP column eluted Ovalbumin) and 2) Specified amount of Ovalbumin spiked in blank gradient eluting from RP column (labelled as Spiked Ovalbumin control). 15ug of Ovalbumin was injected onto the ProSwift™ RP 4H, 1 x 50 mm RP column operating at 30°C. The whole gradient for eluting Ovalbumin (presented in Table 5) was collected in a falcon tube containing 2mL of MS grade water. Additionally, blank gradients eluting from the RP column were also collected in separate falcon tube and 15ug Ovalbumin was spiked into this solution. Both these processes were performed in triplicates.

Table 5: Gradient used for elution of proteoforms from monolithic RP column.

Time(min)	%B	Flow rate(mL/min)
0.10	20	0.1
1.00	20	0.1
11.00	70	0.1
14.00	70	0.1
14.10	20	0.1
20.00	20	0.1

5.2.2 Sample preparation for tryptic digestion of Ovalbumin

Amicon ultra centrifugal filters (4mL) were used to buffer exchange samples into water and reduce the sample volume for further processing. The retentate from the respective Amicon (2x3) filters was collected in six different 1.5mL eppendorf tubes. For tryptic digestion, samples were made upto 100µL total volume with Sodium deoxycholate (SDC) buffer. Reduction of disulphide bonds in the protein was induced by addition of 1µL of 1M dithiothreitol (DTT) to have final concentration 10 mM DTT in respective samples. Samples were incubated for 30 min at 56°C in an Eppendorf 5355 Thermomixer R. 4µL of 0.5M iodoacetamide (IAA) was further added to samples to have final concentration 20mM IAA in samples. Samples with IAA were incubated in dark for 30 min at room temperature for blocking the reduced disulphide bonds. For enzymatic cleavage of Ovalbumin, 50ng trypsin was added to respective sample and incubated over night at 37°C. After 24 hrs, formic acid was added to sample solution (2% final concentration in sample solution) to stop the reaction of trypsin and precipitate SDC. Samples were later centrifuged for 5min at 14,000g and obtained supernatant (consisting of Ovalbumin tryptic peptides) was transferred in a new tube. Prior to LC-MS/MS analysis, collected supernatants were dried in a vacuum centrifuge and reconstituted with 30µL of 0.1% FA water.

5.2.3 High pH reverse phase fractionation of Ovalbumin tryptic peptides

As one of the steps for generating (DDA data) spectral library, offline high pH fractionation (HpH) of tryptic peptides was performed. For high pH fractionation, total 50µg of tryptic peptides (5µL tryptic peptides pooled from each of the 6 samples) were injected onto a monolithic ProSwift RP 4-H 25 cm column. The elution of tryptic peptides was achieved using mobile A comprising of 10mM ammonium bicarbonate in water (pH

8) and mobile phase B comprising of 10mM ammonium bicarbonate in 90% acetonitrile (pH 8). The elution gradient for the HpH fractionation of tryptic peptides was as presented in Table 6.

Table 6: Gradient used for elution of tryptic peptides in HpH fractionation.

Time (min)	%B	Flow rate (mL/min)
0.1	3.3	0.2
5	3.3	0.2
25	38.5	0.2
26	95	0.2
36	95	0.2
37	3.3	0.2
45	3.3	0.2

29 fractions were collected per min of elution gradient, each fraction constituting of 200uL volume. The collected fractions of tryptic peptides were systematically concatenated (as presented in Table 7) to ultimately have 13 samples for DDA analysis (labelled F-3 to F10). Pooled fractions (F-3 to F10) were dried in a SpeedVac and redissolved in 0.1 % FA prior to LC-MS/MS analysis.

Table 7: Table representing scheme used to concatenate tryptic peptides obtained from HpH fractionation.

Fraction number used for pooling	Final fraction labelled as
1 + 2 + 3 fractions pooled	F-3
4 + 5 + 6 fractions pooled	F-2
7 + 8 + 9 fractions pooled	F-1
10 + 20 fractions pooled	F1
11 + 21 fractions pooled	F2
12 + 22 fractions pooled	F3
13 + 23 fractions pooled	F4
14 + 24 fractions pooled	F5
15 + 25 fractions pooled	F6
16 + 26 fractions pooled	F7
17 + 27 fractions pooled	F8
18 + 28 fractions pooled	F9
19 + 29 fractions pooled	F10

5.2.4 LC-MS/MS analysis of tryptic peptides

For the LC-MS/MS analysis, nanoAcquity UPLC system equipped with trapping column (Acquity UPLC® Symmetry C18; 100 Å pore size, 5µm particle diameters, 180µm x 20 mm) and analytical column (Acquity UPLC® Peptide BEH C18; 130 Å pore size, 1.7µm particle diameters, 75µm x 200mm). The peptide trapping was performed at higher flow rate 15µL/min with 99% mobile phase A (0.1% FA in water). Mobile phase B used herein was 0.1%FA ACN. The elution of tryptic peptides was achieved using a gradient of 2%B to 30%B in 60 mins, followed by 95%B for 2 mins, and equilibration of column in next 10 mins at 2%B.

The tryptic peptides were analysed on Q Exactive™ Hybrid Quadrupole Orbitrap mass spectrometer in positive mode. Data is acquired in two modes- DDA and DIA respectively Spectral library was generated using DDA data for the (fractionated &) concatenated tryptic peptides.

For DDA mode in LC-MS/MS analysis, full MS spectrum acquired in range of 400 to 1200m/z at 70000 Orbitrap resolution setting. Precursor ions were dissociated into fragment ions using 25% normalized collision-induced dissociation (CID). MS/MS spectra were acquired in topN 15, isolation window 2m/z, automatic gain control (AGC) target set at 1e5 and Orbitrap resolution set at 17500.

For DIA mode in LC-MS/MS analysis, full MS spectrum acquired in range of 390 to 1210m/z at 70000 Orbitrap resolution setting and 1e6 AGC target. Precursor ions were dissociated into fragment ions using 28% normalized collision-induced dissociation (CID). MS/MS spectra were acquired using 25 m/z isolation window, loop count 16, AGC target set at 1e5 and Orbitrap resolution set at 17500.

5.2.5 Processing LC-MS/MS data from tryptic peptides for quantification

The MS/MS data from DDA files were analysed in Proteome Discoverer 2.0 software (Thermo Scientific™). The required spectral library was generated using chicken FASTA file (obtained from UniProt) and Proteome Discoverer result files. The detailed MS/MS data analysis & quantification (total area fragment based) from the DIA files was performed in Skyline 20.1 (MacCoss Lab Group).

5.3 Sample preparation for various experimental setups

5.3.1 Preparation of protein dilution series for testing FIA-MS as quantitative method

Quantitation using FIA-MS method was evaluated for three different proteins namely Ovalbumin, Filgrastim and Erythropoietin. The stock solution of 10000ng/µL and working solution of 1000ng/ µL was prepared in MS grade water, for each of the three proteins. The dilution series had protein concentrations 25,50,100,250,500,1000 ng/µL respectively. Limit of detection (LOD) was the minimum concentration detected in FIA-MS method (5 continuous charge states in mass spectrum of intact protein). Limit of quantification (LOQ) was calculated based on the ten times the value of signal to noise obtained from the calibration curve.

5.3.2 Sample preparation for investigating effect of non-volatile salt adducts on proteoform ionization

50µL of Eshmuno CPX resin was taken in an (1.5mL) Eppendorf tube and applied with 1000µg Ovalbumin prepared in 1mL of 25mM acetate buffer pH 5. The protein was eluted from the resin after 30 mins incubation using 1M NaCl prepared in 25mM acetate buffer pH 5. The eluate of Ovalbumin in 1M NaCl sample buffer was frozen overnight. Next day the sample was thawed at room temperature and buffer exchanged using 10kDa Amicon centrifugal filter to MS grade water (5 times). The resulting Ovalbumin solution in water was analysed with the optimized FIA-MS parameters settings.

5.3.3 Sample preparation for investigating in-solution supercharging

Desalting effect of supercharging protein was evaluated for a 18kDa therapeutic protein Filgrastim. Filgrastim sample (18mg/mL) acquired from CinnaGen Co. (Tehran, Iran) was in a sodium salt-based sample buffer. A working solution of 1000ng/µL was prepared by diluting the original sample with MS grade water. 20µL of filgrastim sample (1000ng/µL) was spiked with 5%v/v sulfolane supercharger.20µL filgrastim sample without supercharger was used as control sample. Each of these samples were transferred to separate total recovery clear glass sample vials (Waters™) for further analysis with FIA-MS method.

For evaluating effect of supercharger sulfolane on protein with higher molecular, Adalimumab was evaluated with FIA-MS. 500ng/µL Adalimumab was prepared in 60mM ammonium acetate with few drops of 1M NaOH (pH of final sample solution was 8.3). 5% v/v sulfolane was spiked in 20uL of (500ng/µL) Adalimumab solution prepared as above. Adalimumab in sample solution, without sulfolane supercharger was used as a positive control. Samples were transferred to separate total recovery clear glass sample vials (Waters™) for further FIA-MS analysis.

1000ng/µL Ovalbumin sample in pure MS grade water was prepared as working stock. Two sets of dilution series of Ovalbumin with concentrations of 31.25, 62.50, 125, 250, 500, 1000ng/µL respectively were prepared in MS grade water. Second set of Ovalbumin dilution series was spiked with 5% v/v sulfolane supercharger. Samples were transferred to separate total recovery clear glass sample vials (Waters™) for further FIA-MS analysis in triplicates.

5.3.4 Preparation of samples with two proteoforms in defined ratios

Samples consisting of two proteins namely Filgrastim and Myoglobin were prepared in defined ratios. Four eppendorf tubes with 20µL of 2000ng/µL Filgrastim each were prepared in MS grade water. Four Myoglobin samples with concentration 1000ng/ µL, 500ng/µL, 250ng/µL, 125ng/µL respectively were prepared in MS grade water. 20µL of these four respective Myoglobin samples (with concentration 1000ng/ µL, 500ng/µL, 250ng/µL, 125ng/µL) were added to four eppendorf tubes consisting of 20µL of (2000ng/µL) Filgrastim. Ultimately, resulting four samples consisted of Filgrastim:Myoglobin in ratio of 2:1, 4:1, 8:1 and 16:1 respectively.

5.3.5 Preparation for dilution series from proteoform samples

Two set of protein were used to demonstrate quantification with proteoforms-Filgrastim and Ovalbumin. Dilution series of Filgrastim was prepared for working stock solution of 1000ng filgrastim dissolved in MS grade water. The concentrations used for the dilution series were 31.25, 62.50, 125, 250,500,1000ng/µL respectively (prepared purely in MS grade water).

Similarly, to demonstrate quantification with isotopically unresolved spectra, dilution series consisting of 31.25, 62.50, 125, 250,500,1000ng/µL Ovalbumin respectively was used (prepared purely in MS grade water). The optimised FIA-MS method was used to analyse dilution series from these intact proteins. Q Exactive™ Hybrid Quadrupole Orbitrap mass spectrometer (Thermo Scientific™) was used as detector for FIA method.

5.3.6 Steps in sample displacement batch chromatography (SDBC) based fractionation

SDBC experiment used for offline prefractionation of Ovalbumin was performed in 10 segments using 10 Eppendorf tubes (batch mode). 50µL of Eshmuno CPX resin (with binding capacity 2µg/µL resin) was suspended into 10 Eppendorf tubes in equal amounts. 1000µg Ovalbumin sample prepared in 1mL of 25mM acetate buffer (pH 5) was loaded into the first tube. The sample loaded resin was continuously shaken in rotational shaker for 30 min. After 30 mins incubation, the supernatant separated from sedimented resin was transferred to the second Eppendorf tube, incubated, and shaken. This process was repeated for all of 10 tubes. The sample bound resin was washed thrice with 25mM acetate buffer (pH 5) prior to elution step. The supernatant from washing step was discarded. For elution of proteoforms from the resin, 200µL of 1M NaCl prepared in 25mM acetate buffer (pH 5) was used. The sample were incubated with this elution buffer and shaken for 30 minutes. After sedimentation of the resin, the supernatant now called the eluate fraction, was collected from each segment. The eluate of Ovalbumin was then desalted in 10kDa cut-off Amicon centrifugal filters and the total amount of protein was measured using a BCA test. FIA-MS method was used to analyse and quantify proteoforms from each of these eluted SDBC fractions.

5.4 FIA-MS method for analysis of full-length proteoforms

5.4.1 Establishing FIA-MS setup for analysis of Ovalbumin proteoforms

The FIA-MS setup used autosampler of ACQUITY UPLC System (Waters™) as an injector and Q Exactive™ Hybrid Quadrupole Orbitrap mass spectrometer as a detector. FIA is a no column, direct injection-based analysis. The column compartment of the UPLC system (maintained at 30°C) had no column, but only a PEEK tubing (PEEK, 1/16" x 0.13 mm ID, red) connecting directly to the ion source of MS system. Sample was injected by the autosampler action to a stream of spray solvent under laminar flow conditions. The total run time of FIA-MS method was 4 mins wherein initial 1.5 mins operated at flow rate of 0.075mL/min, followed by 0.1mL/min flush for 1 min and returned to 0.075mL/min at 2.5 mins of the run.

The optimization of spray solvent utilized three different solutions namely 40% ACN with 0.1% FA, pure MS grade water, and 150mM ammonium acetate (AmAc). 1µL of 1000ng/µL Ovalbumin sample dissolved in (pure MS grade) water was injected for each analysis.

For optimization of sample application solution, 1000ng/µL Ovalbumin was prepared in respective sample application buffer containing 0, 25, 50 and 100mM AmAc. 1000ng Ovalbumin was injected in this set of FIA-MS method optimization.

Full scan mass spectrum was acquired in positive polarity with ESI on Q Exactive™ Hybrid Quadrupole Orbitrap mass spectrometer. The scan range for Ovalbumin was 1200 to 3600m/z. Except for specific MS parameter optimization, the constant settings on

QExactive™ MS for FIA-MS of Ovalbumin were electrospray voltage 3.5 kV, S lens RF level 75 units, 30eV ISCID, 17500 Orbitrap resolution, 4 microscans and AGC target 5e6.

5.4.2 FIA-MS setup for analysis of Filgrastim proteoforms

The FIA-MS for Filgrastim was performed with ACQUITY UPLC System (Waters™) coupled to Q Exactive™ Hybrid Quadrupole Orbitrap mass spectrometer as a detector. The spray solvent used was pure MS grade water at isocratic flow rate 0.075mL/min for a run time of total 4 mins per sample. Intact Filgrastim data was acquired in positive polarity for ESI and in full scan MS mode. The scan range for Filgrastim was 800 to 3000 m/z. The constant settings on Q Exactive™ MS for FIA-MS of Filgrastim were electrospray voltage 3.5 kV, S lens RF level 75, 0eV ISCID, 140000 Orbitrap resolution, 4 microscans and AGC target 5e6.

5.4.3 FIA-MS setup for analysis of Adalimumab proteoforms

The FIA-MS analysis was performed with Elute LC (Bruker Daltonics Inc.) coupled to maXis II™ (Bruker Daltonics Inc.) as MS detector. The spray solvent used was pure MS grade water at isocratic flow rate 0.075mL/min for a run time of total 4 mins per sample.
Intact adalimumab data was acquired in positive polarity for ESI and in full scan MS mode. The ion source parameters were 4500V capillary voltage, dry temperature 200°C, dry gas flow rate of 8.0mL/min and end plate offset of 500V. The scan range was set to 800 to 6000m/z, ISCID to 100eV, ion energy to 4eV, collision energy 8eV, collision RF to 3800Vpp, pre pulse storage to 20µs and transfer time to 200µs. For better acquisition of proteoforms, the MS parameters were fined tuned to ISCID energy of 150eV, ion energy 6eV, collision energy 12eV, pre pulse storage of 40µs (transfer time was maintained at 200µs and the collision RF to 3800Vpp). The scan range was shifted to 2500 to 6100 m/z. The intact protein mass spectra were deconvoluted using computational suite UniDec (ver. 4.2 from Marty et. al.) or from Bruker Compass Data Analysis 5.1

5.5 Data processing parameters for deconvolution and quantification of proteoforms

5.5.1 Processing parameters in UniDec for deconvolution of Ovalbumin data

The MS data from Ovalbumin proteoforms was deconvoluted using a computational suite called UniDec (ver. 4.2 from Marty et. al.) that operates on a Bayesian algorithm.

Background subtraction and smoothing of charge state distribution was enabled. Binning function was disabled.

Data processing parameters for Ovalbumin deconvolution in UniDec were set as charge range 7 to 32, mass range 38000 Da to 48000 Da, sample every 0.1Da. For peak selection, annotations and quantification of masses detected in deconvoluted spectrum, the peak detection range was set to 10 Da and peak detection threshold was set to 0.06 units. Intensity reported by UniDec deconvolution was used for reporting deconvoluted spectrum-based quantification. The AUC value obtained from manual integration was used to represents the EIF based quantification of proteoforms.

5.5.2 Processing parameters in UniDec for deconvolution of Filgrastim data

The MS data from Filgrastim proteoforms was deconvoluted using a computational suite called UniDec (ver. 4.2 from Baldwin *et al.*, 2015). Data processing parameters included charge range set to 7 to 25, mass range from 18000 to 19500Da, sample every 0.1Da. For peak selection, annotations and quantification of masses detected in deconvoluted spectrum, the peak detection range was set to 5 Da and peak detection threshold was set to 0.03 units.

For deconvoluted spectrum-based quantification, results from UniDec deconvolution (using above parameters) were used. For the extracted ion flowgram (EIF) based quantification, specific charge state in charge envelope from each of the protein was used for attaining area under the curve (AUC) value. The AUC value obtained from manual integration was used to represents the EIF based quantification of proteoforms.

5.5.3 Processing parameters in UniDec for deconvolution for Adalimumab data

The MS data from Adalimumab proteoforms was deconvoluted in UniDec (ver. 4.2 from Baldwin *et al.*, 2015). Processing parameters were set as charge range 20 to 60, mass range 146000 Da to 153000 Da, sample every 0.1 Da. For peak annotations, quantification of masses detected in deconvoluted spectrum, the peak detection range was set to 15 Da and peak detection threshold was set to 0.02 units. Intensity reported by UniDec deconvolution was used for reporting deconvoluted spectrum-based quantification.

6 Results

6.1 Proteoform samples under investigation in the current thesis

6.1.1 Model protein- Ovalbumin

Majority of the method development part in the current thesis was performed with a non-therapeutic protein-Ovalbumin. Ovalbumin is an acetylated, di-phosphorylated glycoprotein derived from chicken. The phosphorylation is reported at serine position 69 and 345, one N linked glycan at 293 and acetylation at 2nd position glycine (Retrieved from UniProt, https://www.uniprot.org/uniprot/P01012).

Ovalbumin makes an excellent model protein because it has a complex proteoform profile ideal for study but also is cheaply available in bulk amounts. A major part of Ovalbumin heterogeneity is due to its glycan component. The complex proteoform profile of Ovalbumin was utilized in previous studies for heterogeneity assessment with native MS (Yang *et al.*, 2013). The commercially available Ovalbumin is a lyophilised powder derived from probably thousands of chicken egg white portions. The proteoform complexity can thereby also be contributed from the isoforms derived from multiple chicken genomes.

6.1.2 Therapeutic protein Filgrastim

Filgrastim is a recombinant form of human granulocyte colony-stimulating factor or GCSF. Filgrastim is a18k Da therapeutic protein, used for the treatment of low neutrophil counts or neutropenia. Filgrastim is popularly used in cancer patients receiving chemotherapy which reduces the white blood cells (Dale, 1998).

It was chosen as one of the test therapeutic proteins in this thesis as it is a small protein with a comparatively less complicated proteoform profile reported in the literature. The non-pegylated filgrastim (used in this thesis work) is a non-complex protein with only methionine oxidation reported as modifications. There are 4 sites reported for these methionine oxidations are Met^{1}, Met^{122}, Met^{127}, and Met^{138} (Holzmann *et al.*, 2013).

6.1.3 Therapeutic protein Adalimumab

Adalimumab (mAb) is a ≈148kDa human recombinant form of immunoglobulin G1 anti-TNF monoclonal antibody. This TP constitutes 1330 amino acids with one glycosylation site at asparagine 297, which is a major source of heterogeneity in the commercially produced recombinant protein. Other common modifications associated with Adalimumab are lysine truncations, asparagine deamidation, glycation, succinimide formation (Füssl *et al.*, 2019). Adalimumab sold under the brand name “Humira” is used for the treatment of chronic inflammatory diseases induced by releases of proinflammatory cytokine tumor necrosis factor. Adalimumab tops the charts as the most selling therapeutic mAb with sales of around 20 billion US dollars for the year 2020 (Urquhart, 2021).

It was chosen as one of the major test proteins in the current thesis as it belongs to the class of IgG1 proteins which is a popularly used class of monoclonal antibody among TPs. Additionally, Adalimumab is a large and complex protein to be handled for full-length proteoform quantification in intact protein MS.

6.2 RPLC-MS for analysis of full-length Ovalbumin proteoforms

Initial experiments were performed to test the suitability of RPLC-MS as a fast method for quantitative analysis of proteoforms. RPLC-MS analysis was performed with a monolithic RP column for the model protein Ovalbumin (described in method section 5.1). The RPLC-MS run of 20 mins, yielded only one main chromatographic peak (Figure 11). No hydrophobicity-based separation of proteoforms along the retention time (RT) was seen. Further on, the mass spectrum obtained at the main peak was very convoluted and represented overlapping signals of multiple charge envelopes from Ovalbumin proteoforms.

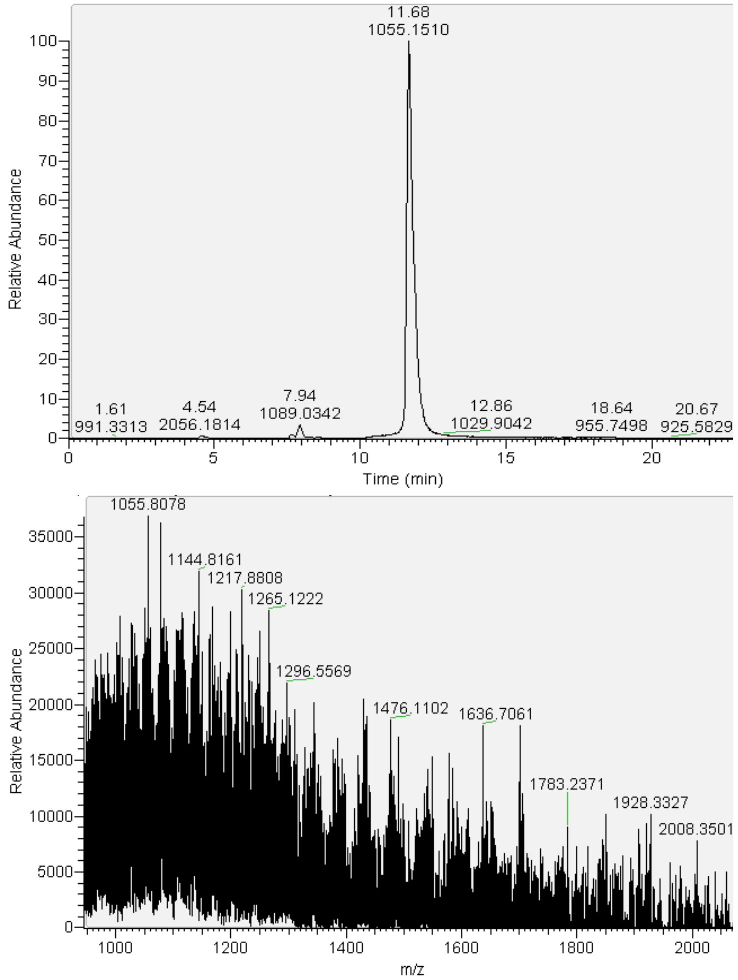


Figure 11: Results of RPLC-MS analysis of Ovalbumin. The upper section shows the RP chromatogram with the main peak at 11.7 min. The section below shows the mass spectrum corresponding to the main peak representing highly convoluted signals of co-eluting Ovalbumin proteoforms. m/z value annotated overhead of respective signal.

The mass spectrum behind the chromatographic peak seen at 7.94 mins in RPLC-MS, represented a peptide signal with a charge (z) value of four (Figure 12a & Figure 12b). To locate the identity of this peptide, a MS/MS analysis was performed using collision-induced dissociation (CID)-based fragmentation. For attaining a good fragmentation of the desired signals (like m/z =1089.04, z=4), an inclusion list of m/z signals to be fragmented was used in MS/MS experiment. The fragment ions data, along with the expected protein FASTA sequences was submitted to a tool called ProSightLite (http://prosightlite.northwestern.edu) for further data analysis.

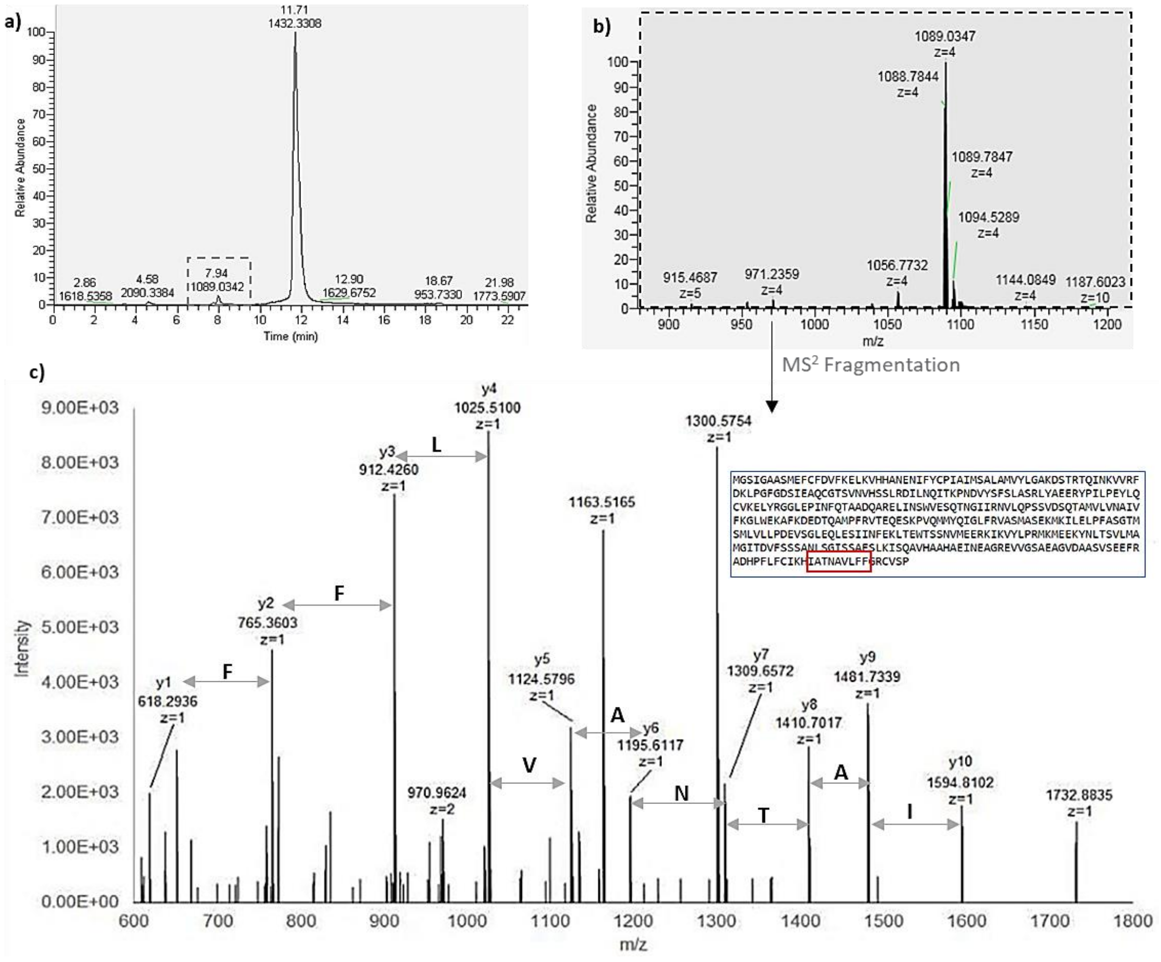


Figure 12: Results of MS^2 analysis of peptide at 971.23 m/z in the mass spectrum of Ovalbumin a) Chromatogram from RPLC-MS analysis of Ovalbumin highlighting peak eluting at 7.94 min. b) Dotted outline shows the mass spectrum behind peak at 7.94 min. c) MS^2 fragments were obtained for precursor at 971.23 m/z and annotated y ion series. The blue inset shows the FASTA sequence of Ovalbumin, with peptide identified in MS2 analysis highlighted in red.

The results of MS/MS data analysis revealed that the peptides detected at 7.94 mins, represented 40-amino acids, a C terminal fragment of Ovalbumin (Figure 12c). Detailed information on the position and sequence of this C terminal fragment is given in Table 8.

Table 8: The peptide fragment seen in RPLC-MS was identified to be a C terminal fragment using ProSight Lite (Northwestern University)

Theoretical peptide mass	4352.93 Da
Experimental peptide mass	4352.13 Da
Position	348-386
Peptide sequence	EAGVDAASVSEEFRADHPFLFCIKHIATNAVLFFGRCVSP

Further on, downstream blank runs to 3000ng Ovalbumin injection in RPLC-MS analysis exhibited an unusual peak at approximately 4.92 mins (Figure 13). The chromatographic peak seen in the blank run was eluting in mid gradient, at approximately 50% ACN.

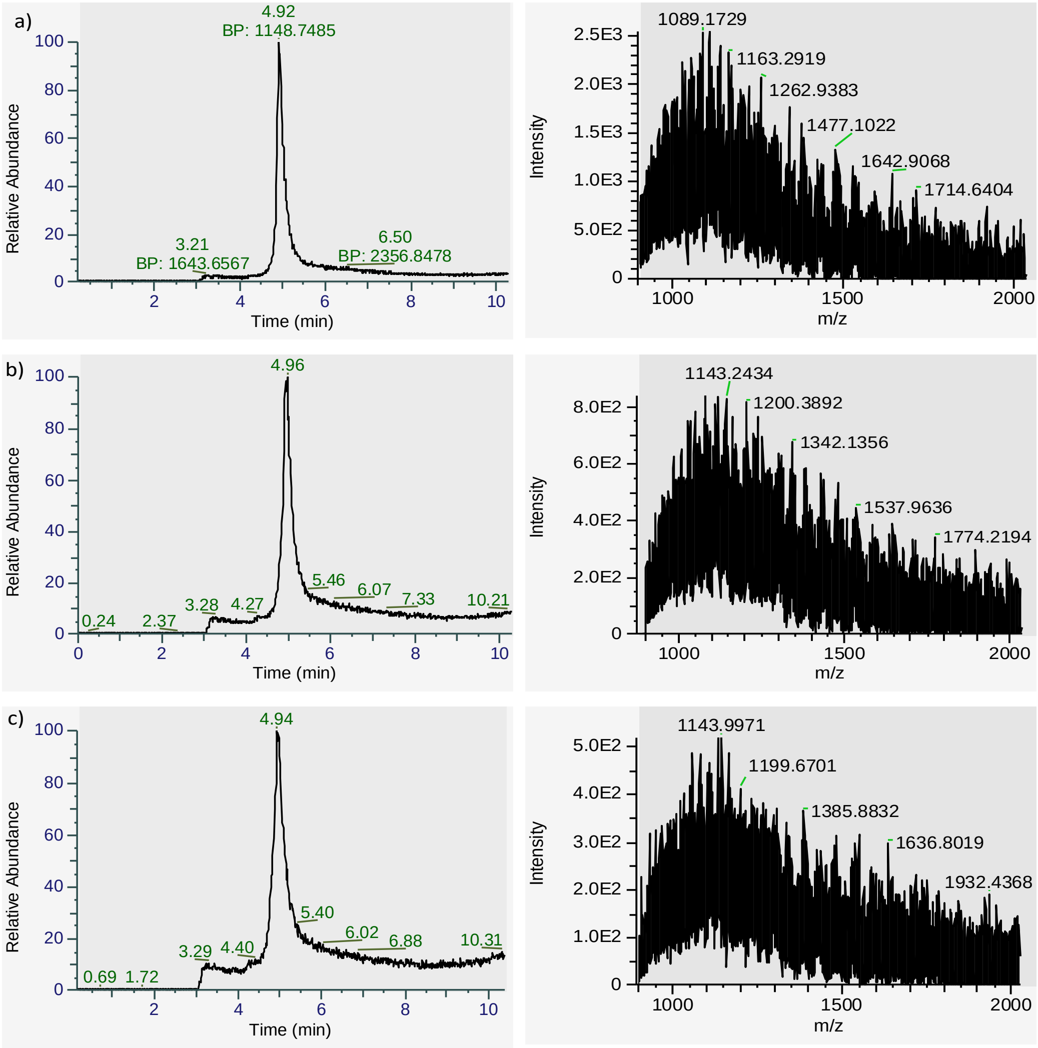


Figure 13: Protein retention seen in the blank runs following Ovalbumin injection on the monolithic RP column. The respective MS spectra (right side) corresponding to the dominant chromatographic peak are representative of Ovalbumin signals. a) First blank injection after Ovalbumin injection b) Second blank injection after Ovalbumin injection c) Third blank injection following Ovalbumin injection.

The mass spectra associated with the peak at 4.92 mins in each of the blanks (Figure 13), revealed signals of Ovalbumin, that were not recovered from previous injections. The previous run to blank used 3000ng Ovalbumin, and this injected amount was well under the loading capacity (90µg) of the monolithic RP column used for analysis. Nevertheless, a huge amount of Ovalbumin was retained back onto the column that eventually decreased with more blank injections.

6.2.1 Evaluation of the recovery of Ovalbumin from the RP column

After recognizing the incomplete recovery of Ovalbumin from the monolithic RP column, a further experiment was performed to quantify the loss of Ovalbumin onto RP column. The experimental workflow used for quantification of Ovalbumin recovering from the RP column is shown in Figure 14. The operating conditions of RPLC used in this experiment are presented in section 5.2.1.

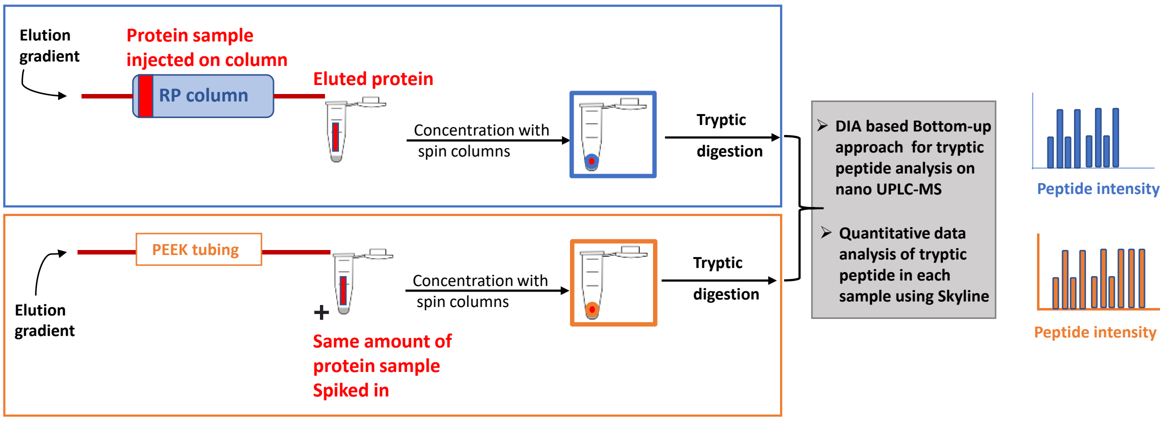

Elution gradient
Protein sample injected on column
RP column
Eluted protein
Concentration with spin columns
Tryptic digestion
PEEK tubing
Elution gradient
+
Same amount of protein sample Spiked in
Concentration with spin columns
Tryptic digestion
DIA based Bottom-up approach for tryptic peptide analysis on nano UPLC-MS
Quantitative data analysis of tryptic peptide in each sample using Skyline
Peptide intensity
Peptide intensity

Figure 14: Scheme of the experimental setup used to estimate the recovery of protein from RP column. The blue box depicts the workflow where a monolithic RP column is used for proteoform elution. The orange box represents the same workflow carried out without a RP column.

The Ovalbumin sample that passed the monolithic RP column was labelled as ‘RP column eluted Ovalbumin’ and the control sample which did not pass-through RP column was labelled as ‘Spiked control Ovalbumin’. As seen in Figure 14, samples underwent the same treatment steps and were digested by trypsin for bottom-up LC-MS/MS analysis. To capture maximum ions, data-independent acquisition (DIA) was performed for tryptic peptides from each condition. The results of summed peptide intensity for respective samples (achieved using data analysis tool-Skyline) are presented in Figure 15a. Details of the peptides detected in respective samples and their respective intensities are also presented in Figure 15b.

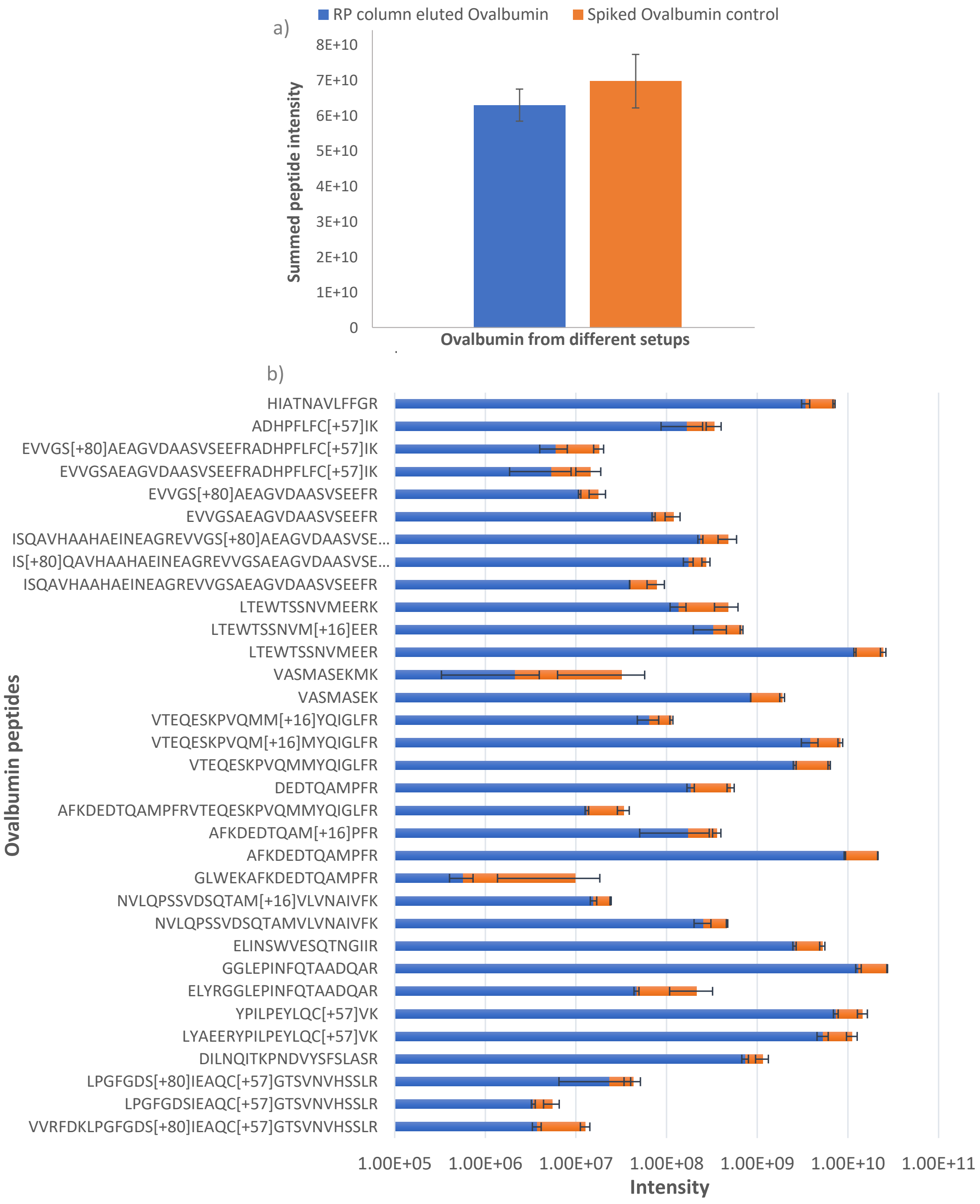


Figure 15: Quantification results of Ovalbumin proteoforms recovered from RP column & analysed using bottom-up MS approach. a) Summed intensity of all the tryptic peptides identified in Ovalbumin that eluted from monolithic RP column (blue) and Spiked Ovalbumin that did not pass the monolithic RP column (orange). b) Intensities of Ovalbumin peptides (with up to 2 missed cleavages) identified respectively in each of the tested samples.

It was evident that the summed peptide intensity from the sample which passed a monolithic RP column was almost 7-fold less than the sample that did not pass the RP column. This implies that some proteoforms do not survive the RPLC analysis before reaching the detector.

6.3 Optimization of FIA-MS as a fast proteoform detection method

With the results obtained from RPLC-MS analysis (section 6.2.1) that indicated incomplete recovery of injected Ovalbumin from the RP column, a suitable alternative was needed. With the main aim of having a fast MS method, flow injection analysis coupled to the mass spectrometer (FIA-MS) was chosen and evaluated for quantification of proteoform from TPs. Flow injection analysis (FIA) setup consists of an injection device, a stream of spray solvent to deliver a sample, and a detector (mass spectrometer in this case). Typically, the autosampler of the HPLC/UPLC system serves as an injection device (Figure 16). As opposed to column chromatography, the sample is guided directly into the mass spectrometer through narrow PEEK tubing. Thus, the resulting elution profile in FIA does not have a perfect bell-shaped Gaussian distribution and is referred to as flowgram (Delabrière *et al.*, 2017).

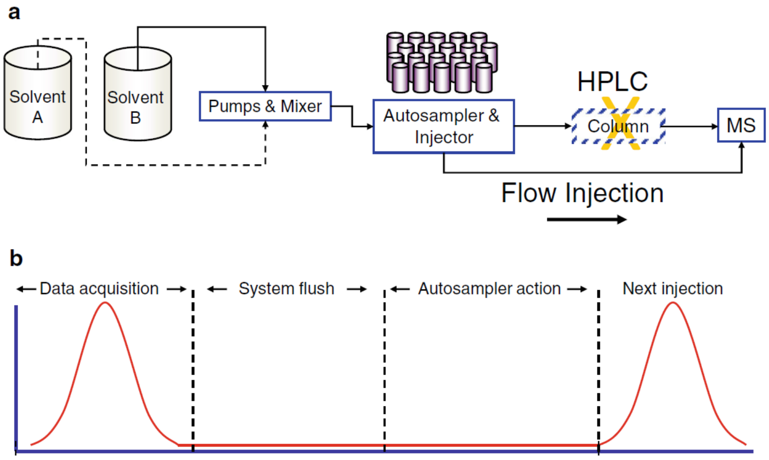


Figure 16: Schematic illustration for flow injection analysis coupled to mass spectrometer. a) the setup b) the cycle from between two data acquisitions. Adapted from (Nanita and Kaldon, 2016)

The efficiency of proteoform ionization can be governed by multiple factors in the flow injection-based method. Before using TPs, elementary factors in FIA-MS, influencing the ionization of proteoforms were evaluated for model protein Ovalbumin. The points considered in the establishment of FIA-MS approach were as follows:

1. Investigating optimal spray solvent in FIA-MS approach.
2. Investigating optimal sample application solution in FIA-MS approach.
3. Investigating optimal MS resolution setting in the FIA approach for proteoform detection.
4. Investigating optimal in-source CID (ISCID) setting in FIA-MS approach.
5. Testing comparability of RPLC-MS and the optimised FIA-MS approach for proteoform detection.

6.3.1 Determination of optimal spray solvent in FIA-MS of Ovalbumin

In FIA-MS mode, the choice of spray solvent is one of the critical factors dictating the sensitivity of proteoform detection. Initially, the effect of organic and aqueous spray solvents respectively on the detectability of proteoforms was studied. For organic spray solvent, Ovalbumin was sprayed in an isocratic flow of 40% ACN solution with 0.1% FA (FIA-MS setup detailed in method section 5.4.1). The 40% ACN was chosen here considering the approximate percentage at which Ovalbumin eluted from the RP column. The spectral signals of Ovalbumin proteoforms seen in this FIA-MS setup had overlapping charge envelopes (Figure 17). This Ovalbumin mass spectrum in FIA with 40% ACN as spray solvent looked alike to the mass spectrum obtained in RPLC-MS analysis (Figure 11). This phenomenon indicated that the convoluted spectra of Ovalbumin proteoforms are a result of the presence of an organic solvent.

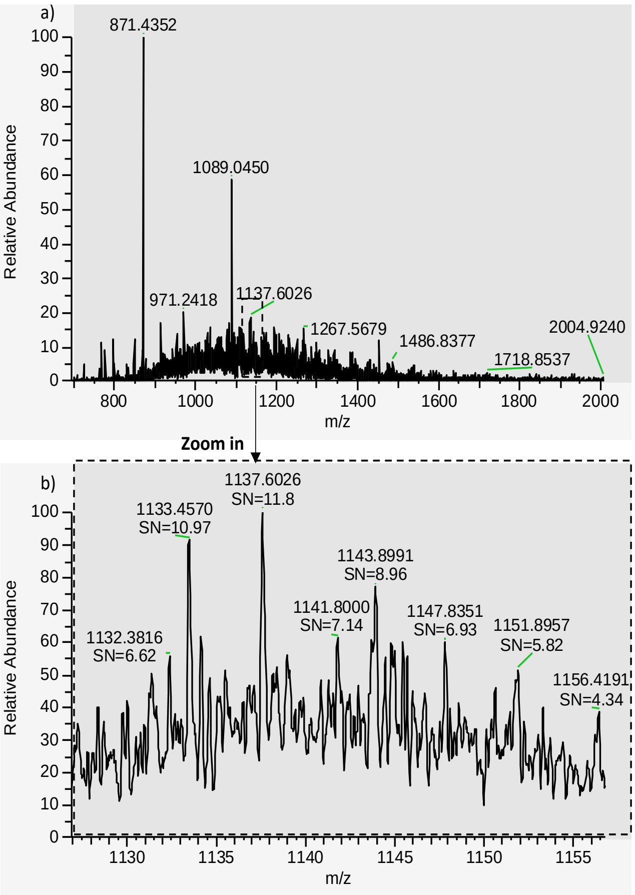


Figure 17: Results of Ovalbumin FIA-MS with 40% ACN used as spray solvent. a) Mass spectrum of Ovalbumin showing widespread charge distribution between 800-1800m/z encompassing more than 25 charge states. Signal to noise ratios obtained for proteoforms annotated as SN. Annotated signals at 871.43 m/z, 1089.04 m/z respectively represent truncated C terminal peptide fragments of Ovalbumin. b) Zoomed view of the apex charge state in charge envelope of Ovalbumin (black dotted outline in 5a) shows the obtained SN for proteoforms ranging between 5 to 11

The signal to noise ratio (SN) obtained at the mass spectral level is one of the most important factors for the detection & quantification of proteoforms. The SN ratio obtained for lower abundant proteoforms was ranging approximately between 4 to 12 (zoomed-in section Figure 17b). Such a low SN ratio makes the proteoforms not suitable for quantification.

To simplify the proteoform spectra, the choice of spray solvent to be used in the FIA-MS setup was changed to native volatile salt-based solvents. Ammonium acetate salt dissolved in water can mimic the physiological conditions proteins are found in (Konermann, 2017). Thus, with intention of switching to milder conditions for FIA-MS, 150 mM ammonium acetate was tested as the next spray solvent. The resulting mass spectrum was highly simplified and possessed fewer charge states in the charge envelope (Figure 18a). The main advantage is seen with 150 mM ammonium acetate as a spray solvent was the increased spacing between the charge states of detected proteoforms. However, the SN ratio did not improve significantly (Figure 18b).

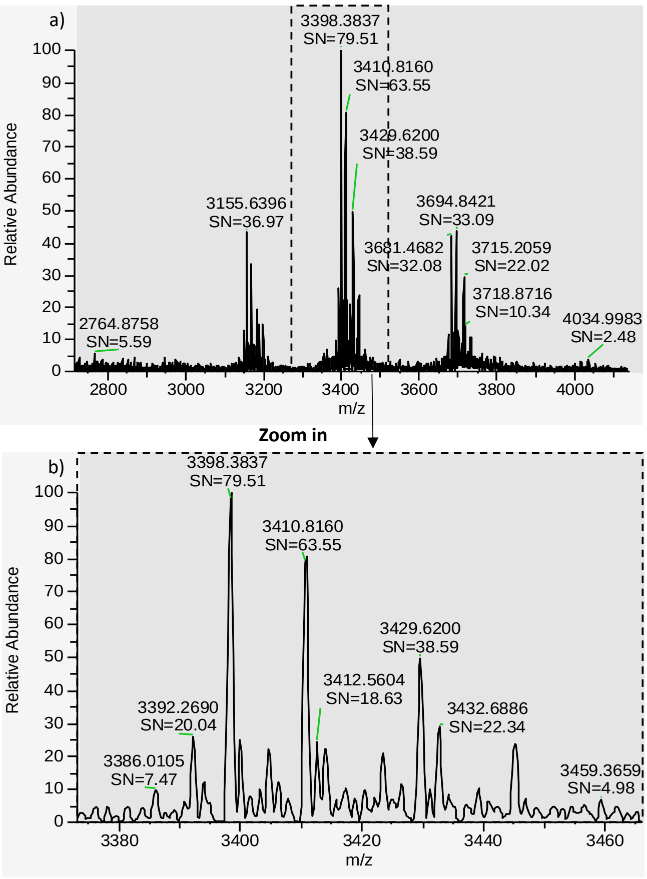


Figure 18: Results of Ovalbumin FIA-MS with 150mM ammonium acetate used as spray solvent. A) Mass spectrum of Ovalbumin showing narrow charge distribution with three charge states. m/z value and signal to noise ratio (SN) annotated overhead of respective proteoform signal. b) Zoomed view of the apex charge state in charge envelope (black dotted outline) shows the SN for proteoforms ranging between 4 to 79.

To adhere to the benefits of the aqueous solutions, pure (MS grade) water was further tested as a spray solvent in the FIA-MS setup. The resulting mass spectrum (seen in Figure 19a) displayed cleaner, non-convoluted signals of Ovalbumin proteoforms. Compared to the mass spectrum obtained with 40% ACN (Figure 17a), fewer charge states are visible in Ovalbumin mass spectrum with water as spray solvent. With signal divided over few numbers of charge states, a subsequent increase in the signal to noise ratio for individual Ovalbumin proteoforms was seen (Figure 19b). With water as a spray solvent for FIA, there was minimum background noise, and lower abundant Ovalbumin proteoforms were visible.

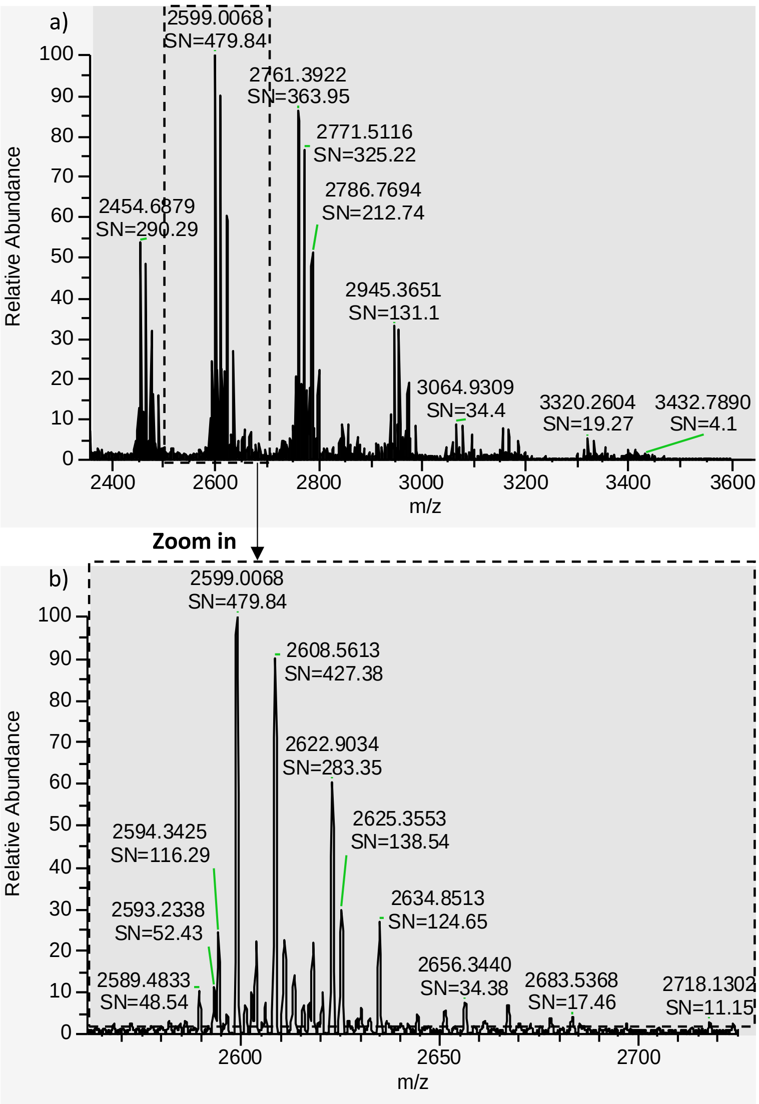


Figure 19: Results of Ovalbumin FIA-MS with pure water as spray solvent. a) Mass spectrum of Ovalbumin showing charge distribution between 2400 to 3600 m/z encompassing six charge states. m/z value and signal to noise ratio (SN) annotated overhead of respective proteoform signal b) Zoomed view of the apex charge state in charge envelope (black dotted outline) shows the SN for proteoforms ranging between 11 to 479.

Additionally, for the same amount of Ovalbumin injection, the SN ratio obtained in water is 5-7-fold higher than in ammonium acetate & 40-45 times higher than in 40% ACN. Further on, compared to ammonium acetate as a spray solvent, water gave a 3-fold higher total proteoform intensity (Figure 20). Between water and 150mM ammonium acetate, water provides better acidification to the proteoforms. This acidification provided by the solvent is crucial in the positive mode ionization of ESI-MS. With the results above, pure MS grade water was decided to be used as spray solvent of choice in all further FIA-MS experiments.

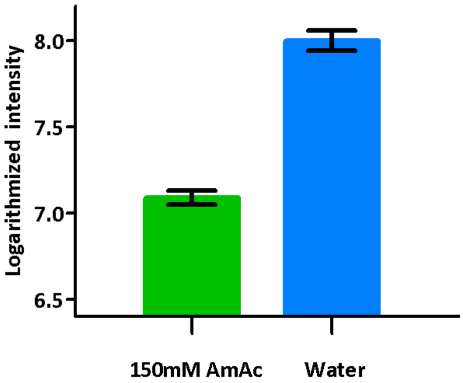


Figure 20: Intensity-based comparison for the Ovalbumin signals obtained in FIA-MS with 150mM ammonium acetate (green) vs water (blue) as spray solvent. The intensity is calculated as the area under the curve for respective flowgrams obtained in FIA-MS. Intensity values are logarithmized to the base 10.

6.3.2 Determination of optimal sample application solution for FIA-MS of Ovalbumin

Considering the SN ratios obtained at the spectral level for Ovalbumin proteoforms, the choice of sample application solution was confined to aqueous solvents. Different concentrations of ammonium acetate were tested in sample application solution (solution in which sample is dissolved, for FIA-MS method). The amount of Ovalbumin used for this experiment kept constant at 500ng per injection. A comparative result of total proteoform signal intensities, with different sample application solutions, is seen in Figure 21. The X-axis represents the concentration of ammonium acetate salt present in the sample application solution. The Y-axis represents Ovalbumin intensity obtained from the area under the curve (flowgram in the current case). It was evident that signal intensity for Ovalbumin serially decreased with the increased presence of ammonium acetate in the sample application solution.

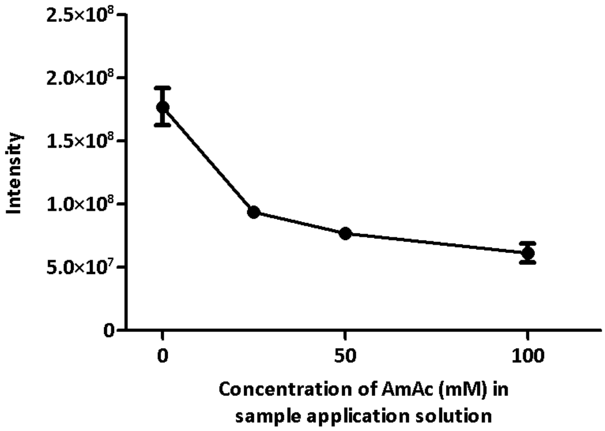


Figure 21: Intensity-based comparison for FIA-MS of Ovalbumin, which was applied in sample application solution containing different concentrations of ammonium acetate. (Water as spray solvent and 500ng of Ovalbumin injection was constant for each sample herein).

The intensity obtained with 25mM ammonium acetate was 2-fold less than the intensity obtained with water (no ammonium acetate used) as a sample application solution. Thus, water was used as a sample application solvent in all further experiments of FIA-MS.

6.3.3 Determination of optimal MS resolution setting for the FIA of Ovalbumin

As there is no online front-end separation of proteoforms in the proposed FIA-MS method, the resolution obtained at the MS level is critical for the distinction of proteoforms. Different resolution settings 17,000; 35,000; and 70,000 @m/z 200 were tested to choose the best fit. Resultant mass spectra at various MS resolution settings are presented in Figure 22. The complexity of the mass spectra is seen to be increased with a higher resolution setting on MS. The signals of the low abundant proteoforms (orange circles overhead) are seen to move closer to noise in a higher resolution of 70,000 (Figure 22c). Further on, as seen in the zoomed section of Figure 22, at 3077.53 m/z, the SN ratio of the representative lower abundant proteoform diminishes from 27 at 17,000 resolution setting, to seven at 70,000 resolution setting. It is also evident from zoomed sections of Figure 22 that the proteoform signals were not isotopically resolved even with higher resolution. Thus, in the interest of SN ratios obtained, it was decided to have 17,000 as the resolution setting for analysing Ovalbumin.

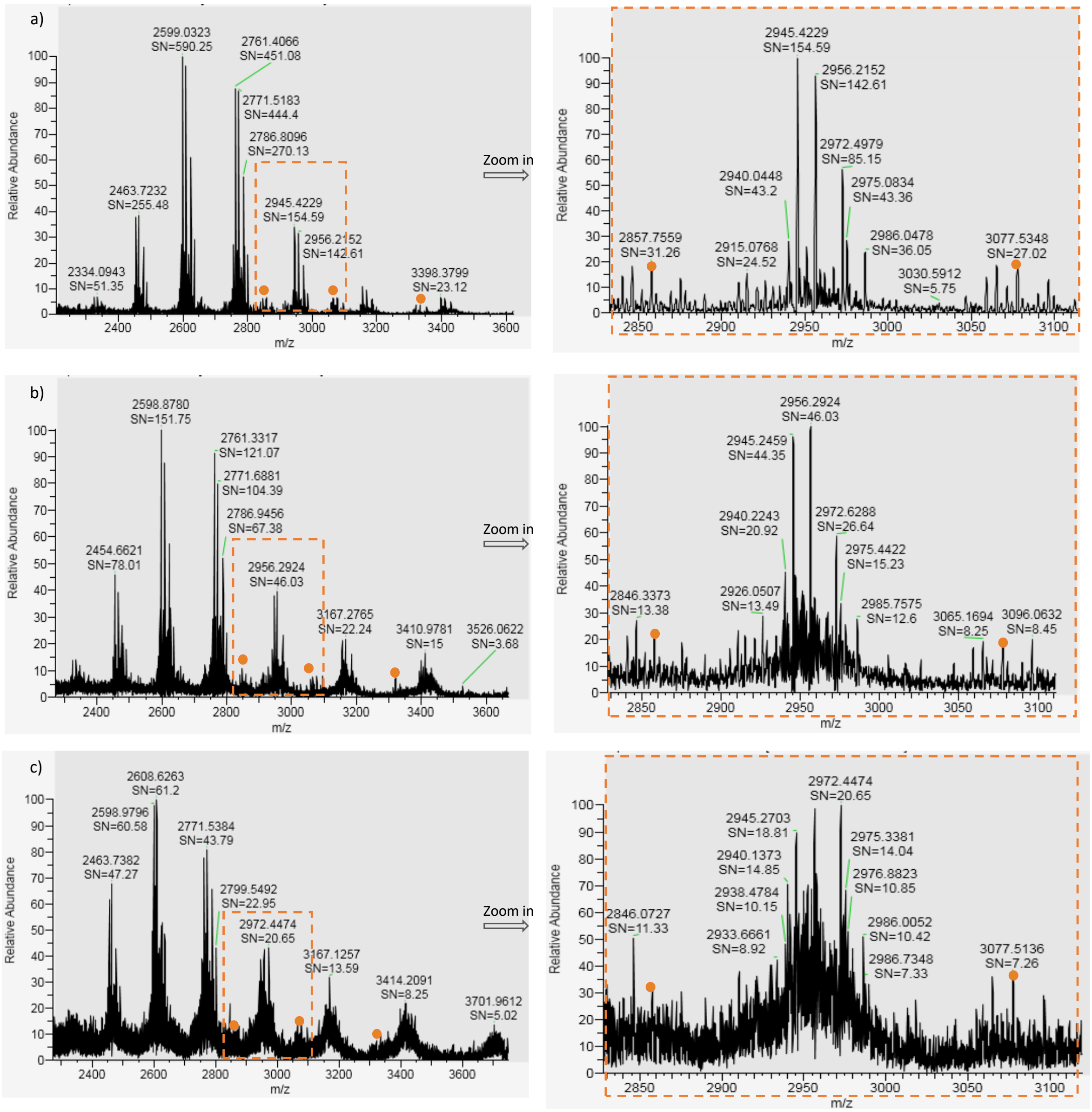


Figure 22: Mass spectra & respective zoomed section for Ovalbumin analysed by FIA-MS method, at different MS resolution settings. m/z value and signal to noise ratio (SN) annotated overhead of respective proteoform signal a) For MS resolution set at 17k b) 35k c) 70k on a hybrid Quadrupole-Orbitrap (Q Exactive) mass spectrometer. The SN for low abundant proteoforms represented in orange circles is seen to diminish with an increase in the resolution setting on the MS instrument.

6.3.4 Determination of optimal in-source CID (ISCID) setting for the FIA-MS of Ovalbumin

For full-length proteoform MS analysis and quantification, a mild clean-up is recommended within the MS using in-source fragmentation (Kilpatrick and Kilpatrick, 2017). However, higher values set for ISCID can fragment the proteoforms, generating additional proteoforms induced by experimental conditions. Thus, a fine balance needs to be maintained while setting this parameter, to not fragment the proteoform in the MS ion source. In the case of Ovalbumin, we tested a range of ISCID from 2eV to 80eV and the resultant spectra are as shown in Figure 23 (1000ng Ovalbumin injection for each FIA-MS run).

The fragments marked in the green outline in Figure 23 were present irrespective of the ISCID applied, indicating that these fragments are originally present in the Ovalbumin sample. These fragments (1089.03 m/z, z=4) matched the C terminal fragment ions of Ovalbumin, seen in RPLC-MS analysis (Figure 12b). From the orange outline in Figure 23, it was noticeable that at the ISCID setting of 40eV, low abundant proteoforms are more distinguishable from the background. The ionization of proteoforms improved with increased ISCID and this was also seen at 80eV ISCID setting. However, at 80eV certain additional fragments (1905.95 m/z, z=2) appeared in the protein spectrum, marked in Figure 23 with a blue outline. This additional fragmentation seen at 80eV ISCID was undesired. Thus, it was decided to use 40eV as an ISCID setting for further Ovalbumin experiments.

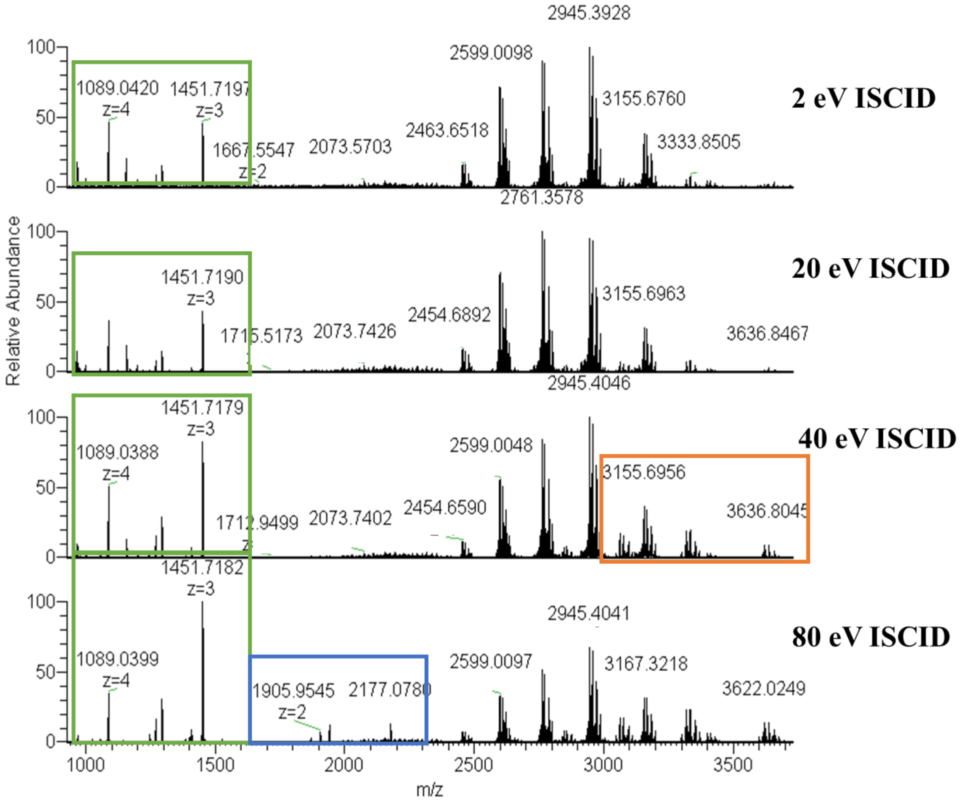


Figure 23: Effect of increasing in-source collision-induced dissociation (ISCID) on the mass spectra of Ovalbumin analysed via FIA-MS. m/z value and charge(z) annotated overhead of respective signal The fragments seen in the green inset were present irrespective of the ISCID setting. With ISCID 40 eV and above, the ionization & thereby SN of low abundant Ovalbumin proteoforms was improved (orange inset). The blue inset at 80 eV of ISCID represents the fragments generated due to higher values of in-source CID.

6.3.5 Comparability of RPLC-MS and the optimised FIA-MS approach for identification of Ovalbumin proteoforms

Comparison of results obtained in RPLC-MS and FIA-MS method was based on proteoform detection efficacy & total time for analysis. Figure 24 presents the results of Ovalbumin proteoforms in RPLC-MS analysis. In RPLC-MS analysis, the run time per sample (20 mins in this case) is long due to the time required for gradient elution and column equilibration. With only one major peak in the RP chromatogram, no separation of Ovalbumin proteoforms is seen in RPLC-MS analysis.

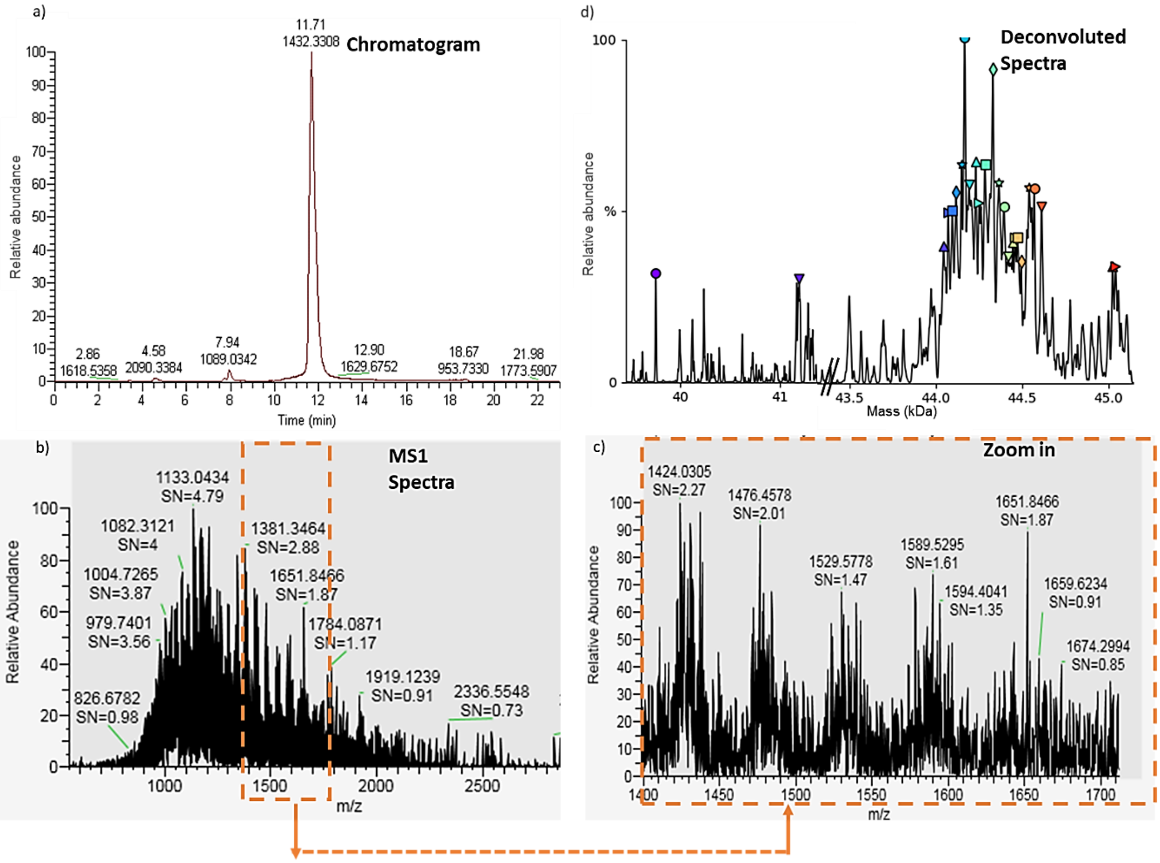


Figure 24: Detailed results of Ovalbumin analysed via RPLC-MS. a) Chromatogram from RPLC-MS analysis of Ovalbumin shows no significant separation of proteoforms but only one main peak. b) The mass spectrum for the peak at 11 min representing multiple co-eluting proteoforms. m/z value and signal to noise ratio (SN) annotated overhead of respective proteoform signal c) Orange inset shows zoomed section of the mass spectrum at 1400 to 1700 m/z depicting low SN of proteoforms obtained in RPLC-MS. d) UniDec deconvoluted zero charge spectrum depicts lot of distorted noise signals along with the colour annotated Ovalbumin proteoforms.

The mass spectrum shows a typical pattern for protein charge envelope, but for more than one proteoforms overlapping & interfering at given 900 to 2500m/z. The overlapping signals were responsible for the crowded mass spectrum. A 300 m/z unit zoom-in section of the Ovalbumin mass spectrum presented in Figure 24c, shows five charge states accommodated in 1400 to 1700m/z range. The m/z signals show annotation of SN ratios obtained for proteoforms at the respective charge state. The SN ratio is seen to be below 10 for all annotated proteoforms. As the SN ratio of proteoforms is diluted, the efficacy of distinguishing a proteoform signal is also reduced. These low SN ratios of proteoforms further affected the data analysis i.e., for the deconvolution process. Figure 24d, represents the deconvolution results of Ovalbumin proteoforms in RPLC-MS analysis. Successfully detected Ovalbumin proteoforms are annotated in different coloured symbols. The non-annotated signals in the deconvoluted spectrum (Figure 24d), represent noise or deconvolution artefacts. Especially, there were only a few annotations for lower abundant Ovalbumin proteoforms in the deconvoluted spectrum (39- 40.5kDa range).

On the other hand, the results from FIA-MS presented in Figure 25, had many positive arguments in comparison to RPLC-MS analysis.

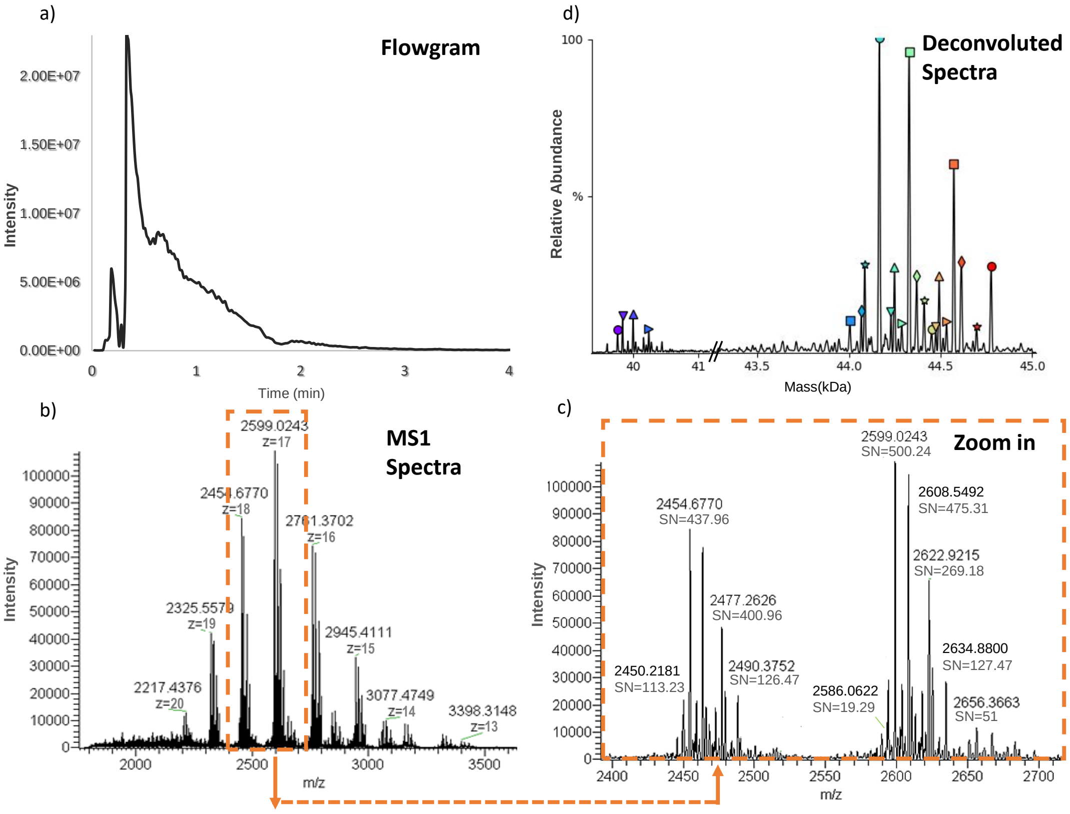


Figure 25: Detailed results of Ovalbumin analysed with FIA-MS method. a) Flowgram from FIA-MS analysis of Ovalbumin. b) The raw mass spectrum representing simplified charge distribution of Ovalbumin proteoforms. m/z value and charge(z) annotated overhead of respective signal c) Orange inset shows zoomed section of mass spectrum at 2400 to 2700 m/z depicting higher signal to noise ratio (SN) of proteoforms obtained in Ovalbumin FIA-MS analysis. d) UniDec deconvoluted zero charge spectrum depicts cleaner deconvolution with the colour annotated Ovalbumin proteoform masses.

Figure 25a, depicts the flowgram obtained in FIA-MS analysis. The flowgram, was not a perfect bell-shaped as expected, because FIA is a no column analysis. The mass spectrum obtained in FIA (seen in Figure 25b), depicts that the proteoform charge envelope is shifted to a higher m/z range (1800 to 3500 m/z). The overlap of proteoform charge envelopes is present but less severe as compared to Ovalbumin mass spectrum obtained in RPLC-MS analysis. Parallelly, compared to RPLC-MS, the distance between consecutive charge states of Ovalbumin proteoforms is also increased in FIA-MS. A 300 m/z unit zoomed section of the Ovalbumin mass spectrum obtained with FIA is seen in Figure 25c. Zoomed view shows two charge states that are accommodated in the 2400 to 2700 m/z range. From the annotated signals in the mass spectrum, it is evident that the SN value for representative proteoforms improved significantly by almost 400-500 times in FIA-MS, then in RPLC-MS analysis of Ovalbumin. Further on, the deconvoluted spectrum (Figure 25d), clearly represents distinct Ovalbumin proteoforms detected in FIA. In comparison to Figure 24d, less unannotated signals i.e., less noise is seen in the deconvoluted spectrum of Figure 25d.

It must be also noted that for the same amount of sample injection as in RPLC-MS, the FIA-MS method could detect a greater number of lower abundant Ovalbumin proteoforms at 39-40.5kDa. The improved SN ratios for proteoforms in the FIA-MS level are responsible for the effective deconvolution process. Additionally, if the speed of data analysis or run time of software for data analysis is considered, the processing of FIA-MS proteoform data speeded up by 71 % than the RPLC-MS proteoforms data. For the required objective of a fast and reliable quantification method, guaranteeing good signal intensities for proteoform is prime. Thus, in the comparative analysis, the FIA-MS method proved to be better than RPLC-MS for the current study. The identity of Ovalbumin proteoforms detected in the FIA-MS method is presented in Table 9.

Table 9: Identity of Ovalbumin proteoforms (and the modifications associated with respective proteoform). Tr=C terminal truncated form, Mo=Monomer, P=Phosphorylation, H=Hexose, N=N-Acetylhexosamine, F=Fucose, NANA=Acetylneuraminic acid. All forms identified here are N terminal acetylated.

Ovalbumin proteoform identity				
Experimental Masses (Da)	**Form**	**Glycan**	**Theoretical mass (Da)**	**Error (ppm)**
39753.70	Tr+P (Hydrolysed)	H3N3	39752.06	41.11
39831.10	Tr+P+P (Hydrolysed)	H3N3	39832.05	23.93
39994.00	Tr+P+P (Hydrolysed)	H3N3+NANA	39994.11	2.69
40237.10	Tr+P+P (Hydrolysed)	H4N4	40238.21	27.64
43875.80				
44004.70	Mo+P+P	H4N2F1	44004.00	15.84
44067.80	Mo+P+P+Sodium	H4N2	44068.01	4.75
44086.00	Mo+P	H3N3	44086.06	1.47
44122.30				
44166.40	Mo+P+P	H3N3	44166.05	7.78
44201.50				
44230.10				
44248.80	Mo+P+P	H4N3	44248.11	15.66
44287.30	Mo+P	H6N2	44289.14	41.62
44328.50	Mo+P+P	H3N3+NANA	44328.11	8.85
44369.80	Mo+P+P	H6N2	44369.13	15.02
44410.30	Mo+P+P	H4N3F1	44410.16	3.15
44450.60	Mo+P+P	H4N4	44451.19	13.18
44492.30	Mo+P	H4N4F1	44492.22	1.73
44532.40	Mo+P+P	H3N4+NANA	44531.19	27.20
44572.50	Mo+P+P	H4N4F1	44572.21	6.43
44613.80	Mo+P+P	H5N4	44613.24	12.57
44645.89				
44696.10	Mo+P	H4N5F1	44695.30	17.87
44775.30	Mo+P+P	H4N5F1	44775.29	0.17

6.4 Assessment of FIA-MS as quantification method using dilution series of various proteoforms

Initially, a simple assessment was performed to check the validity of FIA-MS method for quantification of proteoforms. A dilution series of proteoform samples was analysed with the FIA-MS method for this quantitative assessment. Regression lines were plotted using area under the curve (AUC) of the respective flowgrams resulting from the analysis. The result of the linearity assessment with Ovalbumin dilution series is presented in Figure 26, where the X-axis represents logarithmized values of Ovalbumin concentrations used in the dilution series and the Y-axis represents the respective intensity. Ovalbumin quantitation yielded a linear regression line with a coefficient of determination (R^2) value of 0.9962. The limit of detection (LOD) and limit of quantification (LOQ) obtained for Ovalbumin in the FIA-MS approach was 25ng/µL and 83.3ng/uL, respectively. The standard deviation between analysed triplicate samples- represented as error bars on the plot, was negligibly small.

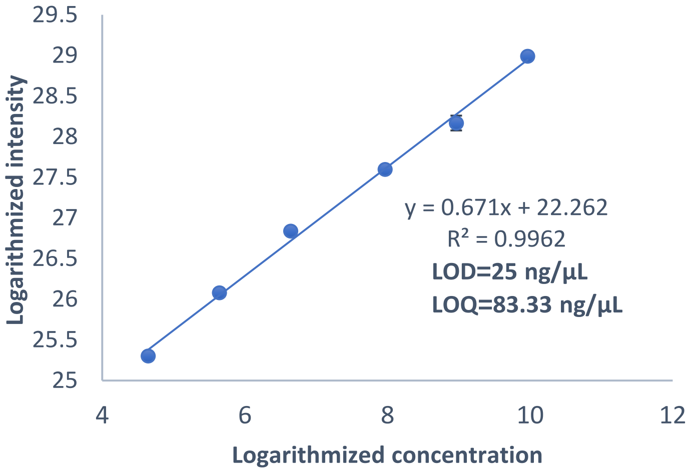


Figure 26: Regression curve in log (2) scale for a dilution series of Ovalbumin comprising of all proteoforms detected in FIA-MS analysis. (The error bars presenting standard deviation between triplicates are present but relatively very small to be visible).

After linear regression response obtained for Ovalbumin sample, quantitative assessment of FIA-MS method was further extended to two other proteins namely Filgrastim and Erythropoietin. The resulting regression lines for Filgrastim and Erythropoietin are shown in Figure 27 and Figure 28 respectively. Alike Ovalbumin, a linear response was observed for quantitation of both therapeutic proteins that were analysed by the FIA-MS method. The R^2 value obtained for the linear regression line of Filgrastim was 0.9944; while the R^2 value obtained for Erythropoietin was 0.9787, respectively. For Filgrastim, the LOD and

LOQ values obtained were 9ng/µL and 30ng/uL, respectively. For Erythropoietin, the LOD obtained was 25ng/µL while LOQ was 83ng/µL.

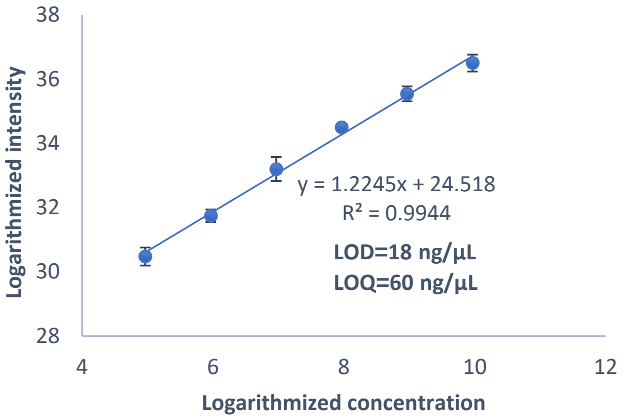


Figure 27: Regression curve in log scale for a dilution series of Filgrastim analysed by FIA-MS. A good fit of linearity is attained with R^2 value of 0.99. Error bars represent the standard deviation between triplicates.

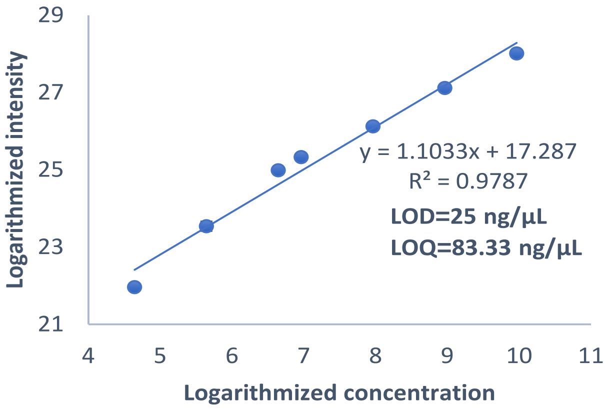


Figure 28: Regression curve in log scale for a dilution series of Erythropoietin analysed in FIA-MS mode. A good fit of linearity is attained with R^2 value of 0.97. Error bars represent the standard deviation between triplicates.

6.4.1 Assessing repeatability of proteoform signals obtained in FIA-MS method

The repeatability of signals obtained in FIA-MS method was examined by evaluating intensity from repeated Ovalbumin injections. The intensity values were obtained using values for the AUC of the respective flowgrams. Figure 29 represents results in terms of the logarithmized intensity values from five repetitive (500ng) Ovalbumin injections in the FIA-MS method. The relative standard deviation from the repeatability tests was calculated to be 6.5% and is seen to be within the acceptance criteria (<15%) for an analytical method establishment.

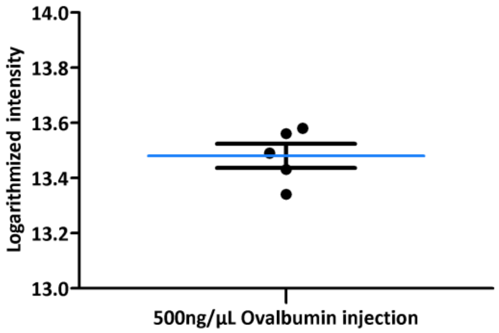


Figure 29: Results of repeatability testing for FIA-MS method with five consecutive injections of 500ng Ovalbumin, from the same vial. Values obtained for the area under the curve are under 6.5% standard deviation.

Further on, the reproducibility of signals obtained in FIA-MS was also examined by overlaying the original mass spectra obtained in the dilution series experiment (Figure 30). Irrespective of the proteoform concentration injected, a fit of signals was observed in overlaid mass spectra.

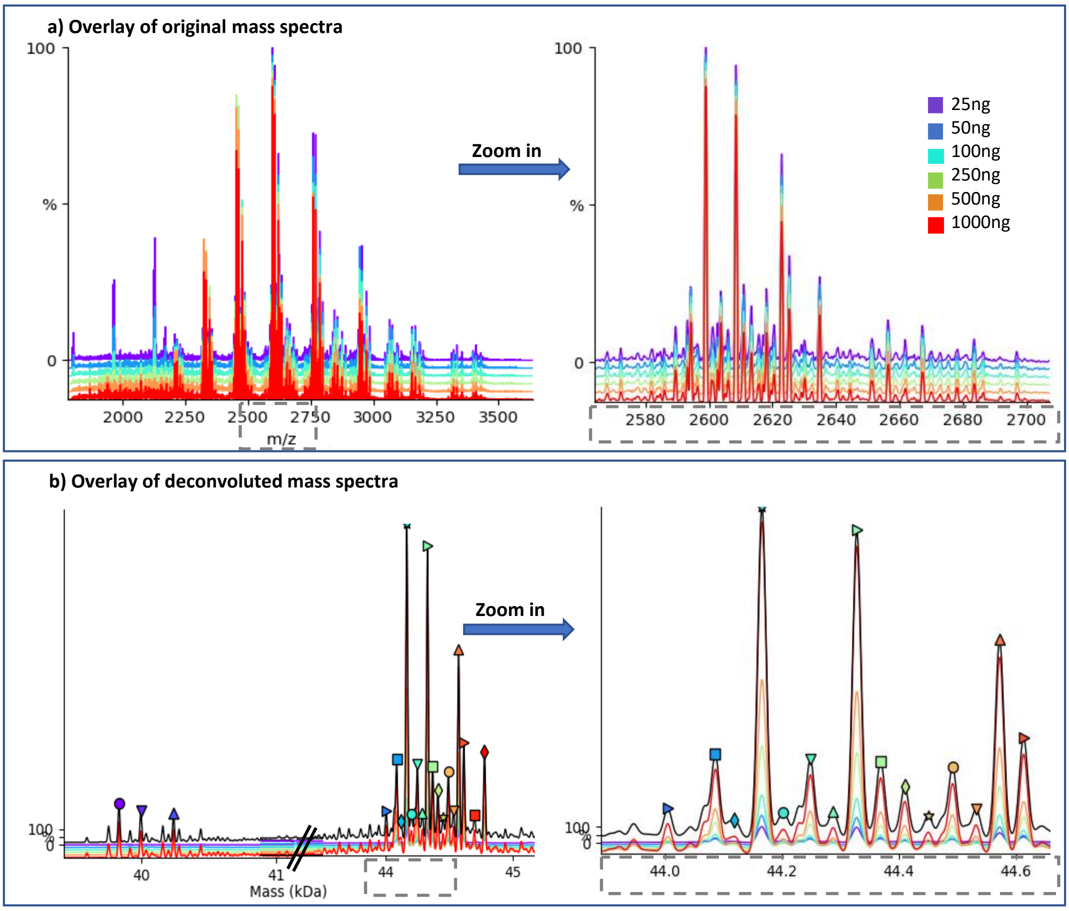


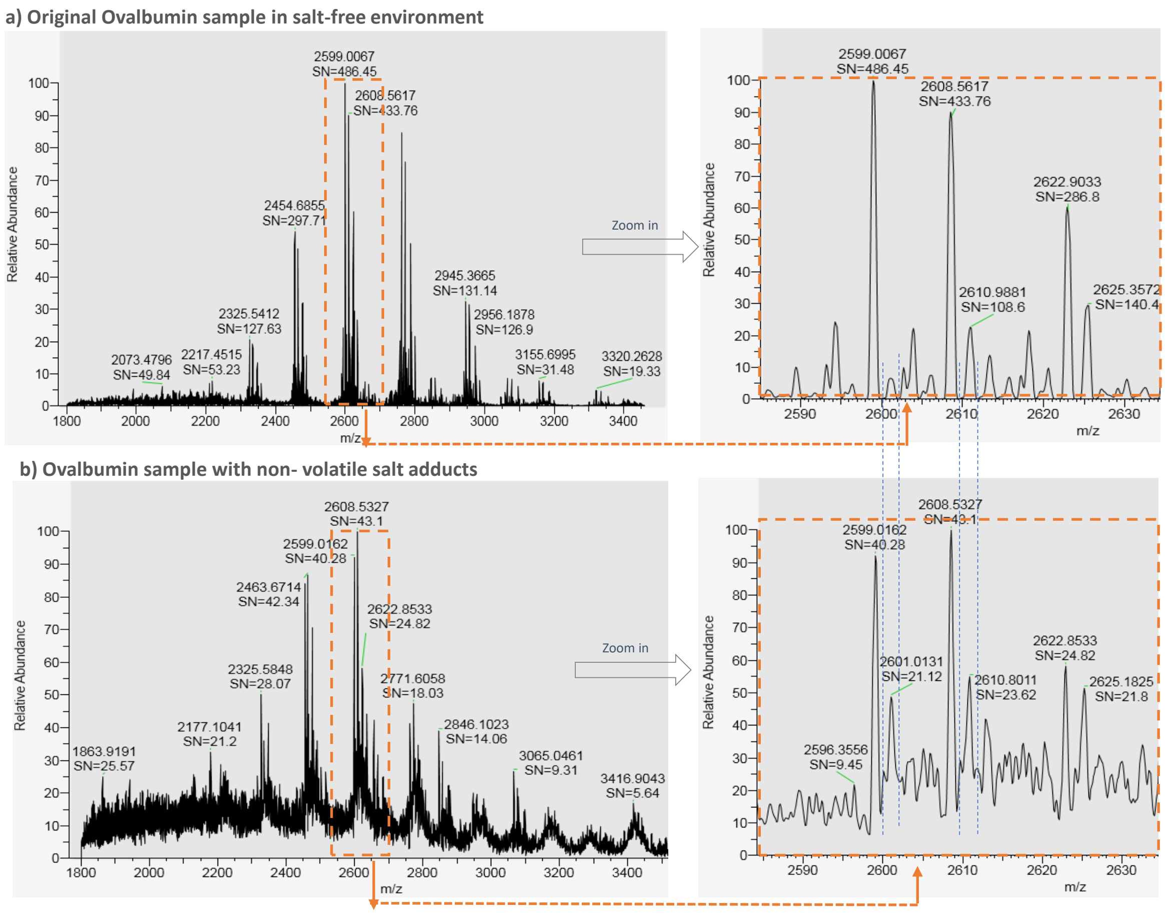
Figure 30: Reproducibility of Ovalbumin signals analysed with FIA-MS at the level of original & deconvoluted spectrum, respectively. a) Overlay of original mass spectra from different amounts of Ovalbumin injection denotes reproducibility of signals obtained in FIA-MS setup. b) Overlay of UniDec deconvoluted zero charge spectra representing masses of Ovalbumin proteoforms. Fitting overlay of deconvoluted signals across the different concentrations of Ovalbumin injected, signifies consistency of results analysed by FIA-MS and as reported by UniDec.

The direct injected-based FIA-MS method, thus, is seen to provide reproducible signals for proteoform analysis. The consistency of proteoform masses obtained with FIA-MS was better visualized after deconvolution of the original mass spectra. Thus, the overlay and fit of masses in the deconvoluted spectrum were also tested (Figure 30b). The coloured annotated symbols denote Ovalbumin proteoforms detected across 25ng to 1000ng of injection. The overlay of deconvoluted spectra displayed consistency in the proteoform masses obtained with the FIA-MS method.

6.5 Assessing specificity of proteoform detection in FIA-MS approach

6.5.1 Investigating the effect of non-volatile salt adducts on proteoform mass spectra

Prefractionation of Ovalbumin (as a part of enriching proteoforms) involved the use of non-volatile salt namely sodium chloride (NaCl). Thus, it was necessary to study the impact of non-volatile salts on the resultant proteoform spectra obtained in the FIA-MS method. Two kinds of samples were considered for FIA-MS analysis- Ovalbumin sample exposed to NaCl (Ovalbumin in 1M NaCl solution for 30 mins and later buffered exchanged to water) and Ovalbumin sample not exposed to NaCl. The mass spectra obtained in FIA-MS analysis for each of the conditions are presented in Figure 31. The resultant mass spectrum of Ovalbumin exposed to NaCl (Figure 31b) and did not resemble the resultant mass spectrum of Ovalbumin not exposed to NaCl treatment (Figure 31a). The mass spectrum of Ovalbumin exposed to NaCl represented crowded signals with higher noise and baseline raise. The zoomed section of this mass spectrum revealed multiple additional signals (represented by blue dotted lines in Figure 31). When the differences in m/z units were examined closely, these additional signals appeared to be from Na ion adduction to multiple proteoforms.



a) Original Ovalbumin sample in salt-free environment

b) Ovalbumin sample with non- volatile salt adducts

Figure 31: Influence of non-volatile salt adducts on Ovalbumin proteoforms seen in FIA-MS analysis. m/z value and signal to noise ratio (SN) annotated overhead of respective proteoform signal a) Ovalbumin mass spectrum in absence of salt adducts. b) Crowded spectral signals of Ovalbumin proteoforms bearing non-volatile salt adducts. On the right side, zoomed view of the respective mass spectrum is seen in an orange outlined box. Additional signals appearing in the mass spectrum due to non-volatile salt adduction to Ovalbumin proteoforms (bottom) are represented with blue dotted lines.

Another characteristic feature seen in the mass spectrum of NaCl treated Ovalbumin was decreased SN ratios (zoomed section of Figure 31b). The SN ratio at apex 2599.01 m/z was seen to be reduced more than 10 times (from 486 to 40) for Ovalbumin sodium salt adducts. The resultant Ovalbumin mass spectral complexity due to adducted sodium molecules is further a hindrance to ionization of lower abundant proteoforms as well as a limitation for accurate peak detection in the deconvolution processes.

6.6 Effects of supercharging on proteoforms bearing non-volatile salt adducts

To circumvent the likelihood of losing proteoform signals due to the copresence of non-volatile salt adducts, the addition of a supercharging agent was considered. The choice of sulfolane as an additive supercharger was made due to its reported effectiveness for charging protein compared to other supercharging agents (Going, Xia and Williams, 2015). Effect of sulfolane based supercharging in FIA-MS was evaluated for two different proteins:

1. Filgrastim
2. Adalimumab

6.6.1 Investigating desalting effects of supercharger sulfolane on Filgrastim proteoforms

A protein with no complex PTMs like phosphorylation or glycosylation would present ideal case scenario to trace the loss of non-volatile salt ions among the ionized proteoforms seen in. Thus, to demonstrate the in-solution desalting phenomenon of supercharger sulfolane, a small 18 kDa therapeutic protein- Filgrastim was used initially as a model protein. The FIA-MS analysis of Filgrastim, in the presence and absence of supercharger sulfolane, is represented in Figure 32.

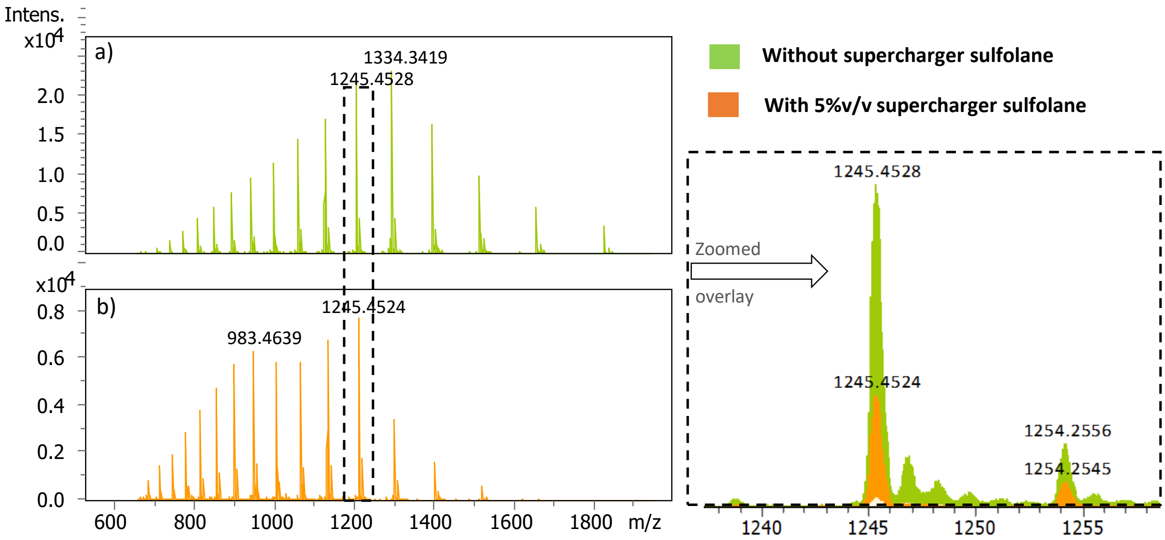
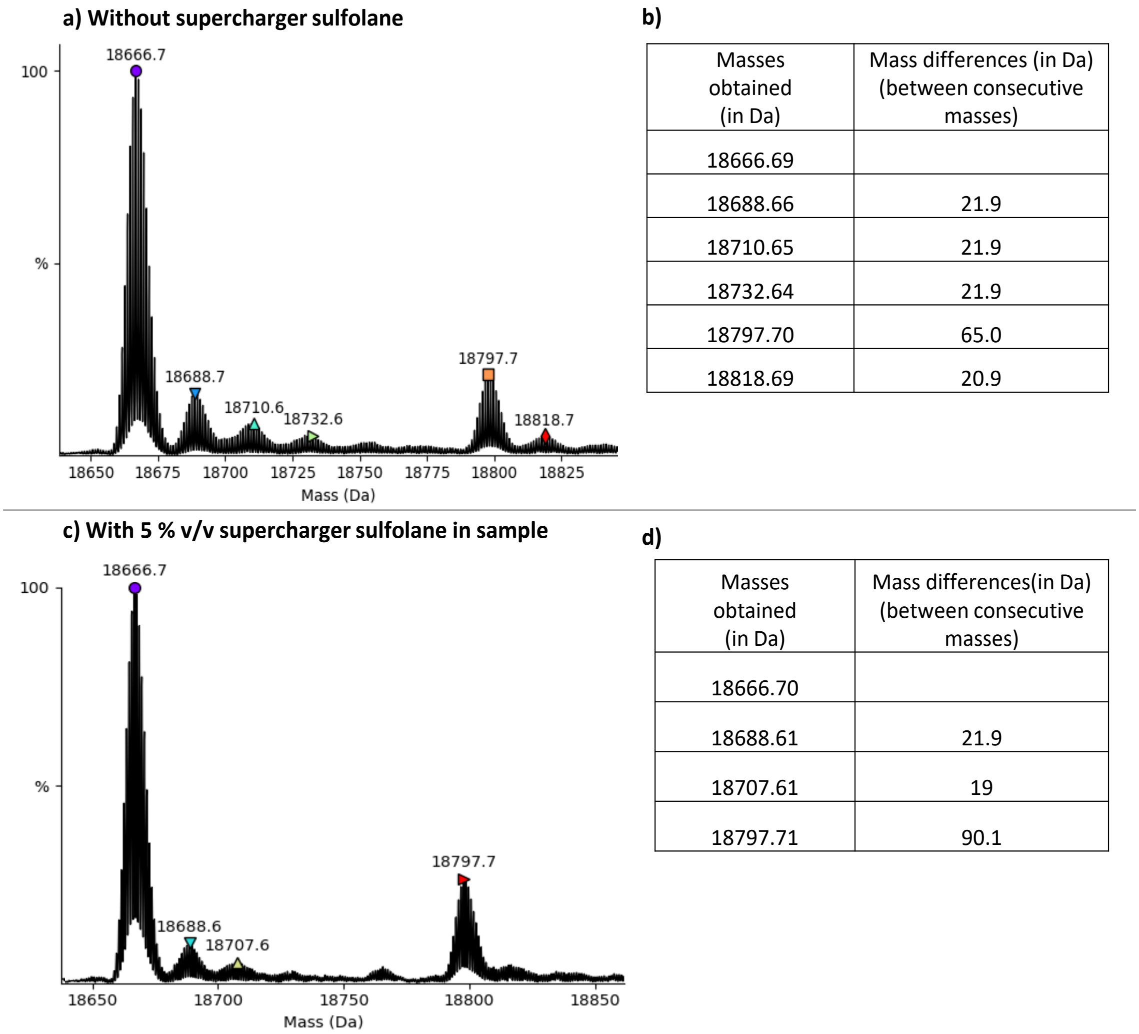


Figure 32: Mass spectra from FIA-MS of Filgrastim a) without the presence of supercharger sulfolane and b) in presence of 5%v/v supercharger sulfolane. The zoomed in and overlaid image on the right, beneficial loss of salt adducted signals in presence of supercharger sulfolane (orange) is seen.

The change noticed at the original mass spectrum of Filgrastim in presence of 5% v/v sulfolane (Figure 32) was shifting of charge envelope towards lower m/z. This was a direct consequence of the increased number of charges acquired by Filgrastim in presence of supercharger sulfolane. A zoomed overlay of a typical charge state for FIA-MS spectra, in the respective conditions (with and without supercharging) denoted the loss of certain m/z signals in the supercharged Filgrastim (orange coloured spectrum in Figure 32).

The loss of signals represented the loss of adducted Na forms. Clear confirmation of this phenomenon was visualized at the level of the deconvoluted mass spectrum (Figure 33).



a) Without supercharger sulfolane

b)

Masses obtained (in Da)	Mass differences (in Da) (between consecutive masses)
18666.69	
18688.66	21.9
18710.65	21.9
18732.64	21.9
18797.70	65.0
18818.69	20.9

c) With 5 % v/v supercharger sulfolane in sample

d)

Masses obtained (in Da)	Mass differences(in Da) (between consecutive masses)
18666.70	
18688.61	21.9
18707.61	19
18797.71	90.1

Figure 33: UniDec deconvoluted spectrum with proteoform masses annotated in different coloured symbols for a) Filgrastim without the presence of supercharger sulfolane. c) Filgrastim in presence of 5%v/v supercharger sulfolane. Table of proteoform masses and mass difference between consecutive proteoforms b) in non-supercharged Filgrastim, 3 Na adducted species (mass difference 21.99 Da) detected d) in supercharged Filgrastim, only one Na adducted species detected.

It was evident from Figure 33a & Figure 33c, that supercharging the proteoform sample with sulfolane, significantly reduced the relative intensity of Na adducted signals, if not eliminated. Table b & c in Figure 33 presents Filgrastim masses detected in respective conditions along with the mass difference between consecutive masses. The obtained mass difference of 21.99 Da represents adducted Na ion. On supercharging Filgrastim sample, three Na adducted forms were reduced to only one detectable Na adducted form. The other significant mass difference was seen between two main proteoforms- 18797.7 Da & 18666.7 corresponds to 131Da, representing methionine loss. It is also evident from Figure 33c that the relative abundance of 18797 Da proteoform was increased with a corresponding decrease in Na adducted forms.

6.6.2 Investigating desalting effects of supercharger sulfolane on adalimumab proteoforms

After demonstrating the effect of supercharger sulfolane on a comparatively small protein Filgrastim, supercharging was further tested on a large protein namely a monoclonal antibody-Adalimumab. To evaluate the desalting effect of supercharger sulfolane, Adalimumab (mAb) sample solution was prepared with trace amounts of Na (addition of 5-7 drops of 1M NaOH). The mAb spectra obtained with FIA-MS in the absence and presence of the supercharger sulfolane are presented in Figure 34.

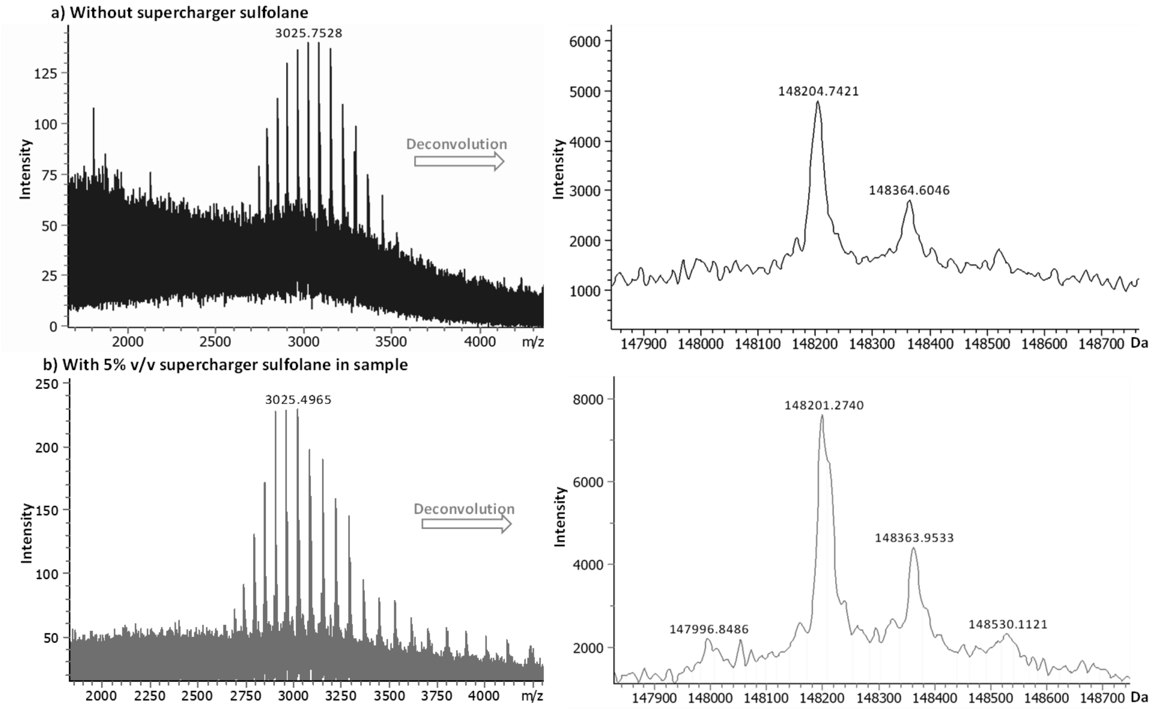


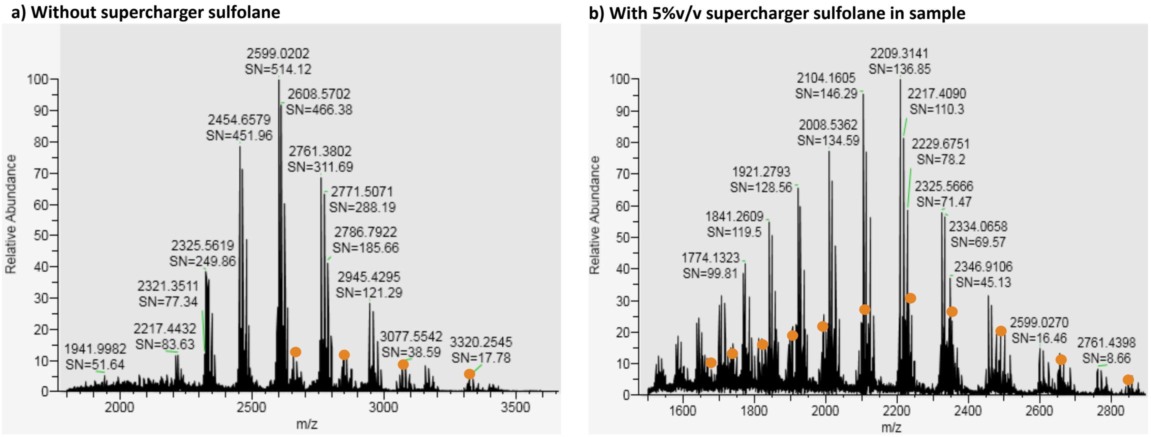
***Figure 34:** Mass spectra from FIA (on left) & deconvoluted spectrum (on right) for Adalimumab in presence of trace amounts of Na ions in sample solution a) In absence of supercharger sulfolane, the mass spectrum shows a rise in baseline & noise. Deconvoluted spectrum detected two proteoforms masses with a high baseline. b) In presence of 5%v/v supercharger sulfolane improved ionization is seen in the original mass spectrum. Four proteoform masses were detected with higher intensity & SN in the deconvoluted spectrum.*

For non-supercharged mAb, a raise in baseline and background noise is seen in the mass spectrum due to the presence of non-volatile salts (Figure 34a). Deconvolution of the subsequent mass spectrum (Figure 34a) resulted in the detection of only two proteoforms: 148203 Da and 148364 Da. Alike the raw mass spectrum, the deconvoluted spectrum resulting from non-supercharged mAb showed an increased noise and high baseline. In contrast, the mass spectrum from supercharged mAb was seen to have reduced background noise. Due to the supercharging effect of sulfolane on the mAb, more charge states were visible in the mass spectrum (Figure 34b). Further on, the deconvolution results obtained from the supercharged mAb had more distinct proteoforms signals compared to the non-supercharged mAb. The relative abundance of mAb proteoforms- 148203 Da & 148364 Da was increased in the case of supercharged mAb. Although with a low SN ratio, two additional proteoforms with mass 147996.8 Da & 148530.1 Da respectively, could be detected in the deconvoluted spectrum of supercharged mAb. With high baseline noise, only two mAb proteoforms were annotated in non-supercharged mAb.

6.7 Effects of Ovalbumin supercharging on detection of its low abundant proteoforms

As the next step, the effect of supercharging phenomenon on the detectability of lower abundant proteoforms was studied. The study was performed on a complex proteoform pool of Ovalbumin using the FIA- MS method. Figure 35 represents the changes seen in the Ovalbumin spectrum with and without the addition of supercharger sulfolane to the sample.

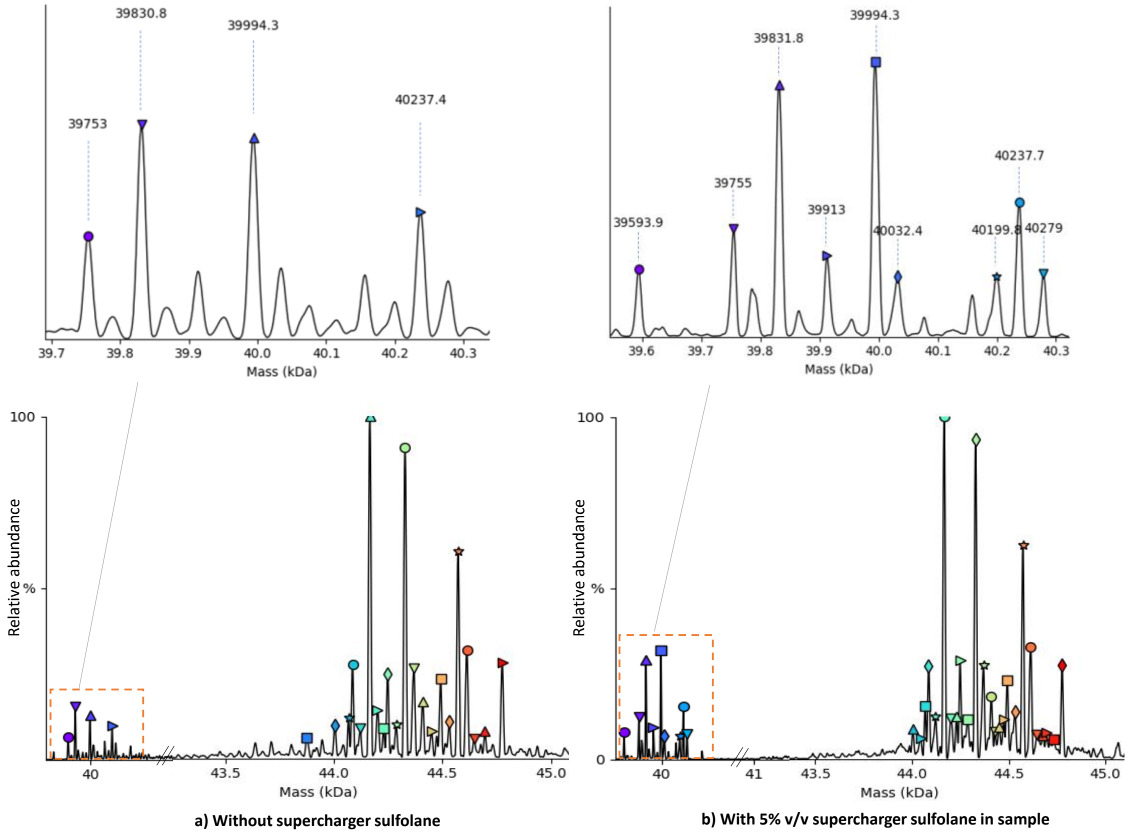
In presence of supercharger sulfolane, the charge state distribution of Ovalbumin was shifted towards lower m/z. The number of charge states observed increased from seven in non-supercharged Ovalbumin, to fourteen in supercharged Ovalbumin. It was also apparent that the ionization of lower abundant proteoform was improved in presence of supercharger sulfolane with sample (orange dots in Figure 35b).


a) Without supercharger sulfolane

b) With 5%v/v supercharger sulfolane in sample

Figure 35: Mass spectrum of Ovalbumin obtained in FIA-MS a) without supercharger sulfolane added to Ovalbumin. b) with 5%v/v supercharger sulfolane in Ovalbumin sample solution. m/z value and signal to noise ratio (SN) annotated overhead for respective signal. Signals from lower abundant Ovalbumin proteoforms annotated with orange dots are noticeably increased in presence of supercharger sulfolane.

Improved ionization and detection of lower abundant proteoforms was more evident at the level of deconvoluted spectra, which is presented in Figure 36.


a) Without supercharger sulfolane

b) With 5% v/v supercharger sulfolane in sample

Figure 36: UniDec deconvoluted spectra of Ovalbumin a) without supercharger sulfolane denotes lower relative intensities for 39kDa Ovalbumin proteoforms. b) In presence of supercharger sulfolane, the detectability & relative abundance of Ovalbumin proteoforms in the 39k Da range is significantly increased. The number of Ovalbumin proteoforms (coloured annotations) also increased from 25 to 34 in presence of supercharger sulfolane.

Due to the usage of supercharger sulfolane as an additive to proteoform sample, the number of Ovalbumin proteoforms detected increased from 25 to 34. The orange inset in Figure 36 shows that not only the number, but also the relative abundancies of proteoforms were increased due to supercharging of the Ovalbumin sample. It is also noticeable that ion signals of specifically lower abundant proteoforms were intensified.

6.7.1 Effects of Ovalbumin supercharging on quantification of its proteoforms

As a next step, the impact of supercharging on quantification of Ovalbumin proteoforms was examined. For testing limits of detection and linearity response, dilution series of supercharged Ovalbumin with known concentrations was evaluated with the FIA-MS method. The linear regression curve obtained using the AUC values of flowgrams is presented in Figure 37 (logarithmized values of both X and Y-axis). The high value for the coefficient of determination (R^2 =0.9931) was indicative of a good linearity response from dilution series of supercharged Ovalbumin. Additionally, an improvement of LOD & LOQ was obtained in comparison to non-supercharged Ovalbumin (Figure 26). The LOD for supercharged Ovalbumin was 9ng/µL & LOQ was 30ng/µL.

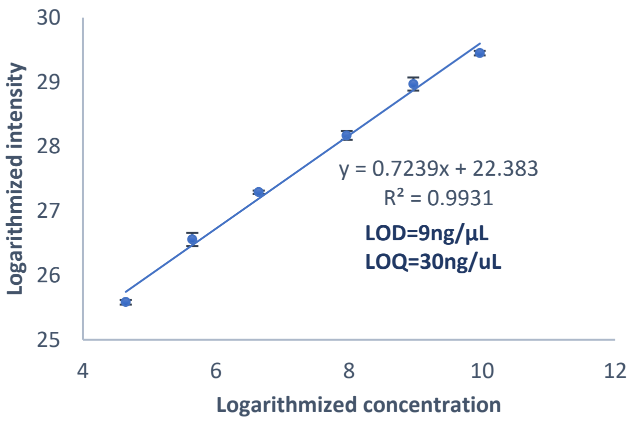


Figure 37: Regression curve with FIA-MS method for supercharged Ovalbumin showing excellent fit for linearity with a R^2 value of 0.99 and negligible standard deviation among triplicates. The X and Y-axis representing concentration and intensity respectively are logarithmized to a scale of 2.

To assess the benefits of supercharging for quantitation of proteoforms in detail, rather than relative abundancies, absolute intensity values obtained for proteoforms were compared. Among 35 proteoforms detected, two representative proteoforms with masses- 39994 Da and 44166 Da, were used to represent the effect of supercharging on lower abundant and higher abundant proteoforms, respectively (Figure 38). In total, when all proteoforms of Ovalbumin were considered for extracted ion flowgram (EIF) based quantification, the values obtained for supercharged Ovalbumin were higher than non-supercharged Ovalbumin irrespective of the amount injected (Figure 38a). At the proteoform level, for supercharged Ovalbumin (Figure 38 b & c), there was an increase in the signal intensity of the low abundant proteoform 39994 Da. In the case of supercharged Ovalbumin, for the lowest injection amount i.e., 25ng, a 1.3-fold signal increase was detected for the lower abundant proteoform 39994 Da. At the same time, no increase in signal intensity was observed for a higher abundant proteoform 44166 Da. These results demonstrated the benefits of a supercharging for the detection of lower abundant proteoforms from a complex proteoform pool of a medium sized protein.

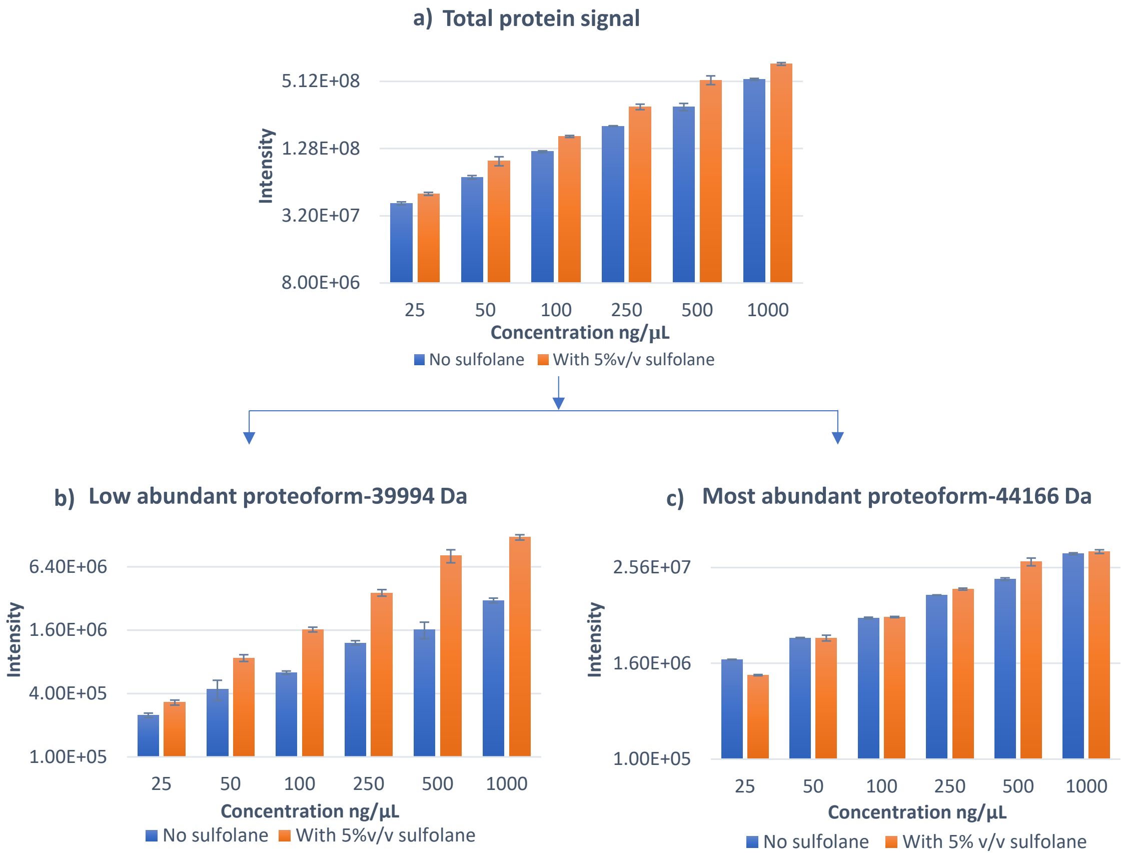


Figure 38: Intensity-based comparison for FIA-MS of Ovalbumin without supercharger sulfolane (blue) & in presence of 5%v/v sulfolane, across different concentrations in dilution series. a) Summed intensity for Ovalbumin proteoforms is seen to be higher in presence of sulfolane supercharger. b) The intensity of lower abundant proteoform with mass 39994 Da is increased in presence of supercharger sulfolane. c) The intensity of abundant proteoform (in sample) with mass 44166 Da, is not affected by the presence of supercharger relative to the lower abundant proteoform.

6.8 Data processing strategies for full-length proteoform quantification in MS

In intact protein MS, there is no universal method to obtain reliable quantification values for individual proteoforms. Hence, there is a need to evaluate data processing strategies for proteoform quantification. Figure 39 shows steps used in the current thesis for evaluating data analysis strategies for attaining proteoform quantification.

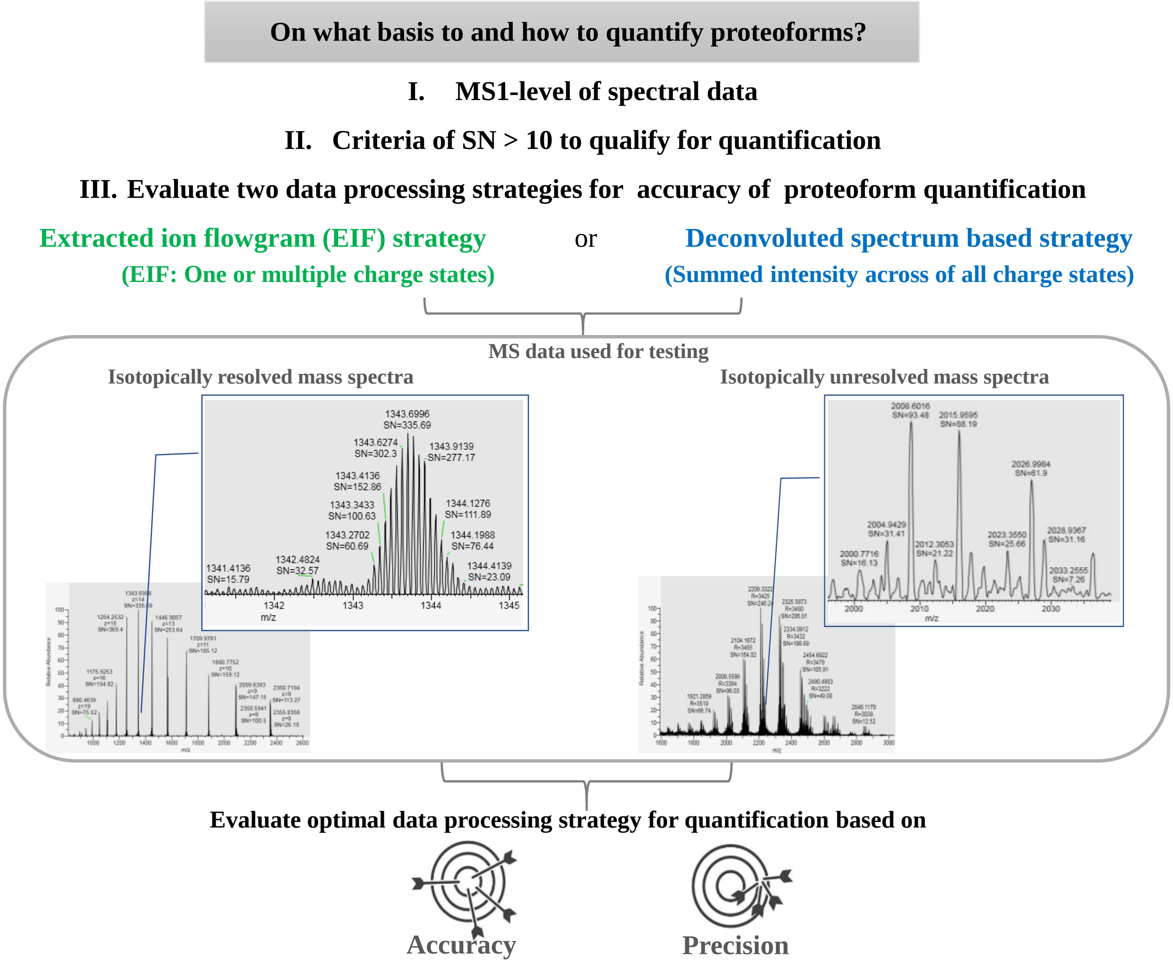

On what basis to and how to quantify proteoforms?
I. MS1-level of spectral data
II. Criteria of SN > 10 to qualify for quantification
III. Evaluate two data processing strategies for accuracy of proteoform quantification
Extracted ion flowgram (EIF) strategy
(EIF: One or multiple charge states)
or
Deconvoluted spectrum based strategy
(Summed intensity across of all charge states)
MS data used for testing
Isotopically resolved mass spectra
Isotopically unresolved mass spectra
1343.6996 SN=335.69
1343.6274 SN=302.3
1343.9139 SN=277.17
1343.4136 SN=152.86
1343.3433 SN=100.63
1344.1276 SN=111.89
1343.2702 SN=60.69
1344.1988 SN=76.44
1342.4824 SN=32.57
1341.4136 SN=15.79
1344.4139 SN=23.09
1342
1343
1344
1345
m/z
2008.6016 SN=93.48
2015.9595 SN=88.19
2026.9984 SN=61.9
2004.9429 SN=31.41
2012.3053 SN=21.22
2023.3550 SN=25.66
2028.9367 SN=31.16
2000.7716 SN=16.13
2033.2555 SN=7.26
2000
2010
2020
2030
m/z
Evaluate optimal data processing strategy for quantification based on
Accuracy
Precision

Figure 39: Scheme followed for evaluating the suitable data processing strategy for quantification of proteoforms.

The first basis stated was quantification at MS1 level data or using full scan mass spectrum. The second criterion was that only proteoforms showing signal to noise ratio (SN) above 10 at the original spectrum would qualify for quantification. The third basis was to evaluate the most suitable data processing strategy for reporting proteoform quantification. Deconvoluted spectrum-based quantification may suffer in accuracy for isotopically unresolved spectrum. Thus, extracted ion chromatogram (EIF) &

deconvolution-based quantification was evaluated for both-isotopically resolved and non-resolved spectra. Filgrastim was opted to study isotopically resolved spectra and Ovalbumin for the assessment of the quantification of isotopically unresolved intact protein spectra. Quantification data from respective data processing strategy was compared & evaluated (to choose the optimal data processing strategy) based on accuracy and precision.

6.8.1 Quantification of two most abundant proteoforms in sample- comparing data processing strategies

Before looking at the low abundant proteoform level, quantitative evaluation was initially performed for the two most abundant proteoforms in samples consisting of defined ratios of two proteins, namely Filgrastim and Myoglobin (method section 5.3.4). Four sample solutions each consisting of Filgrastim, and Myoglobin mixed in 2:1, 4:1, 8:1, 16:1 proportion were evaluated with FIA-MS in triplicates. The respective data processing strategies namely extracted ion flowgram and deconvoluted spectrum-based quantification were applied on proteoform data obtained in the FIA-MS method. EIF based quantification value was derived from manual integration of the apex charge state for obtaining the AUC value. Deconvoluted spectrum-based quantification used intensity value as reported by UniDec software. The mean intensity value (over triplicates) from respective quantification strategies was obtained. Ratios between mean values of each of five consecutive samples were drawn out. The experimentally obtained ratios of Filgrastim:Myoglobin were compared to known solution-phase ratios (Table 10).

Table 10: Results given by different data processing strategies for quantifying Filgrastim & Myoglobin samples (mixed in defined ratios). First column represents the actual ratio of Filgrastim: Myoglobin present in sample solution (expected ratio). Quantification results based on single charge EIF data processing (second column) are deviated from the expected ratios. Quantification results based on deconvoluted mass spectrum (third column) represent more accurate ratios closer to the expected ratios in column 1

Expected ratios Filgrastim: Myoglobin	Calculated ratios Single charge EIF-based quantification	Calculated ratios Deconvoluted spectrum-based quantification
2	4	2
4	10	4
8	21	10
16	35	20

The quantified values for single charge state EIF method (column 2, Table 10) gave ratios that were off from the in-solution ratios, for all evaluated samples. Values obtained from deconvoluted spectrum-based quantification (column 3, Table 10) closely represented the expected values (column 1, Table 10).

6.9 Quantification of proteoforms from isotopically resolved Filgrastim mass spectrum

6.9.1 Isotopically resolved mass spectrum of intact Filgrastim proteoforms

Figure 40 shows the mass spectrum obtained for FIA analysis of intact Filgrastim. The X-axis represents m/z signals while the Y-axis shows the relative abundance of respective m/z values. Between 800-2500 m/z, the charge envelope of Filgrastim shows 14 charge states. At an Orbitrap resolution of 140k at 200 m/z, (on Q Exactive Hybrid Quadrupole-Orbitrap MS), baseline resolved m/z signals were obtained for 18kDa Filgrastim. However, the experimental resolution at proteoform level is not 140k (set resolution) but ranges from 45000-58000 units.

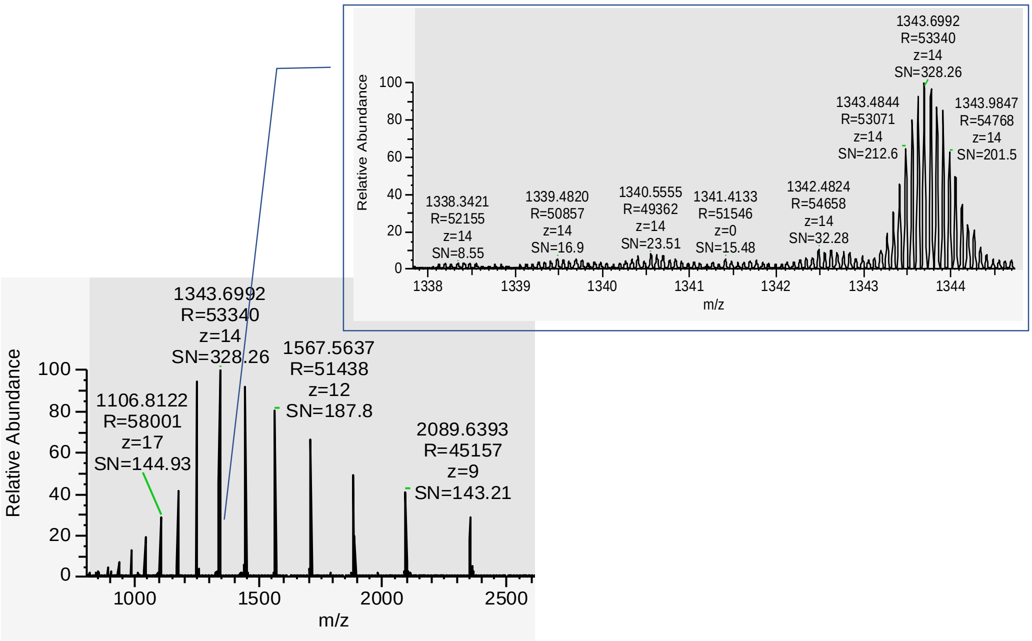


Figure 40: Isotopically resolved mass spectrum obtained for Filgrastim in FIA-MS approach. m/z value, resolution(R), charge (z), signal to noise ratio (SN) annotated overhead for respective proteoform signal. Seen below in the figure is the charge envelope of Filgrastim showing multiple charge states(z) ranging from 9 to 20. The blue inset shows zoomed section, depicting isotopically resolved peaks at charge state 14.

A zoomed view of the apex charge state of Filgrastim shows isotopically resolved peaks of Filgrastim proteoforms. Proteoforms were identifiable from the raw mass spectrum, but only at very low relative abundance. Moreover, from the blue inset of Figure 40, a continuous peak pattern for isotopic distribution of low abundant Filgrastim proteoforms is seen.

6.9.2 Selection of charge states for EIF based quantification of Filgrastim proteoforms

Figure 41 represents the details of Filgrastim mass spectrum, for the selection of charge states for reporting EIF based quantification. The three charge states (highlighted in Figure 41a) were used to obtain multiple charge-based EIF values for respective proteoform quantification. The apex charge state highlighted in light green (Figure 41a) was used to calculate single charge-based EIF values for respective proteoform quantification. Figure 41b depicts a zoomed section of the apex charge state where the isotopically resolved peaks of Filgrastim proteoforms are shown. Proteoforms for Filgrastim considered for quantification (SN ratios > 10) are annotated with numbers 1-4 in the raw mass spectrum (Figure 41b).

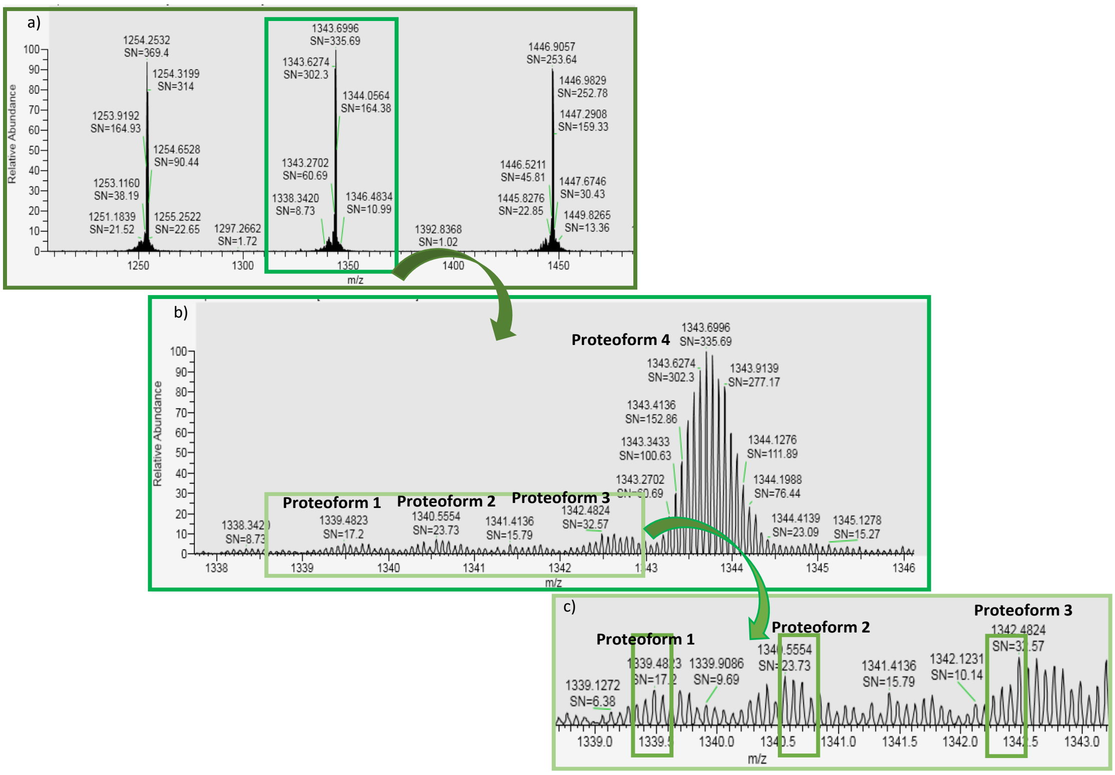


Figure 41: Detailing on Filgrastim mass spectrum obtained in FIA-MS analysis for the EIF based quantification. m/z value and signal to noise ratio (SN) annotated overhead for respective proteoform signal a) Three most abundant charge states in the mass spectrum of Filgrastim that are used for three charge state based EIF quantification of proteoforms. b) Apex charge state used for single charge state-based EIF quantification of proteoforms. Zoomed view of apex charge state shows main Filgrastim proteoform numbered 4 and its lower abundant proteoforms (numbered 1-3) to left in the florescent green inset. c) Isotopically resolved peaks of lower abundant Filgrastim proteoforms. The light green inset depicts the m/z window used for EIF based quantification per proteoform.

Further zoomed portion of Figure 41b is highlighted in section c of Figure 41. The overlapping isotopic distributions of low abundant proteoforms are evident. The start and end of isotopic distributions are not distinguishable for lower abundant proteoforms. However, the choice of the EIF window is an important factor for the correct quantification of overlapping proteoform signals. Thus, as represented in Figure 41c, an EIF window of 4 isotopic peaks (0.05 m/z peak width for each isotopic peak) at a particular charge state was chosen.

6.9.3 Deconvoluted spectrum-based quantification for Filgrastim proteoforms

Signals used for obtaining deconvoluted spectrum-based quantification are presented in this section. Figure 42a, depicts the charge envelope of Filgrastim, while Figure 42b depicts the results of UniDec based deconvolution. Unlike EIF based quantification, information of all states, represented in Figure 42a is used to generate deconvoluted spectra and thus to obtain quantitative values for each proteoform in Figure 42b. Thereby, deconvoluted spectrum-based quantification is the intensity value reported by the algorithm, for respective proteoforms. Quantification values are represented on the Y-axis of Figure 42b as % relative abundances.

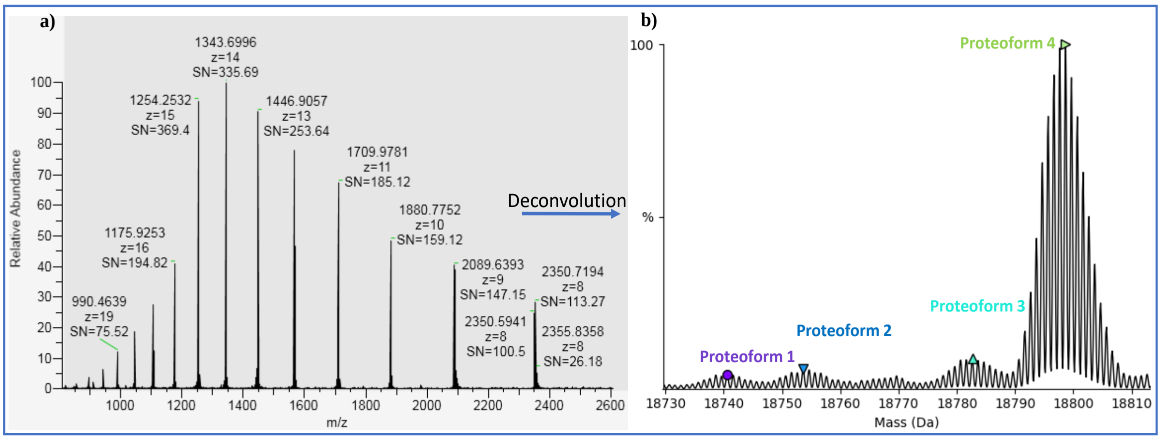
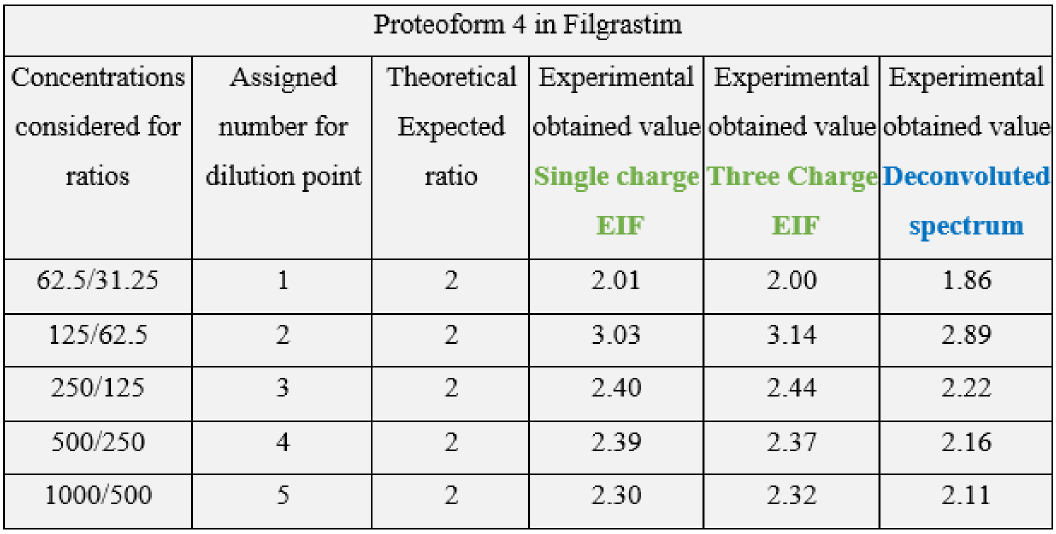


Figure 42: Deconvolution results of Filgrastim proteoforms for deconvoluted spectrum-based quantification a) Full scan mass spectrum for Filgrastim showing the charge state distribution & SN at the respective charge state b) UniDec processed deconvoluted spectrum encompassing Summed intensity of Filgrastim proteoforms across all charge states.

6.9.4 Comparing data processing strategies for quantification of proteoforms from isotopically resolved mass spectra

The data processing strategies for proteoform quantification were compared in regard to the accuracy and precision of reported values. For Filgrastim, the accuracy and precision for quantification were investigated over a dilution series with known concentration. Detailed information on the concentrations considered and numbers assigned for each dilution point is presented for one of the Filgrastim proteoforms (numbered as proteoform 4) in Table 11. When ratios between consecutive dilutions were calculated, the theoretically expected ratio is two (column 3, Table 11). Ratios obtained with the area under curve values for single charge integration and three charge states respectively, were compared for EIF quantification (columns 4, 5, of Table 11). Ratios obtained with intensity values from UniDec deconvolution are presented in column 6 of Table 11.

Table 11: Ratios calculated using different data processing strategies for dilution series of Filgrastim. Quantification results for dilution point 2 are seen to have deviated across all three data processing strategies.



Proteoform 4 in Filgrastim					
Concentrations considered for ratios	Assigned number for dilution point	Theoretical Expected ratio	Experimental obtained value **Single charge EIF**	Experimental obtained value **Three Charge EIF**	Experimental obtained value **Deconvoluted spectrum**
62.5/31.25	1	2	2.01	2.00	1.86
125/62.5	2	2	3.03	3.14	2.89
250/125	3	2	2.40	2.44	2.22
500/250	4	2	2.39	2.37	2.16
1000/500	5	2	2.30	2.32	2.11

Only two proteoforms-annotated 3 & 4 in Figure 41, passed the criteria of SN ratio above 10, required as LOQ. Hence the quantitative accuracy & precision calculations were considered only for proteoform 3 & 4. In Figure 43 and Figure 44, the X-axis represents different data processing strategies, while the Y-axis shows the ratios obtained at different dilution points. Accuracy -defined here as the value closest to the expected ratio (i.e., value closest to 2) is represented as a blue dotted line. According to the FDA regulation, for analytical testing, values with ±20% variance of expected values, are under acceptance criteria (FDA, 2018). The quantified ratios at dilution point 2 (ratio of 125ng to 62.5 ng/µL protein) presented an outlier. This outlier (orange dots in Figure 43 and Figure 44) was detected across all data processing strategies, indicating a systematic error at dilution point 2 in dilution series preparation. As visualized within blue circles in Figure 43 & Figure 44, for deconvoluted spectrum-based quantification, maximum dilution points were under the criteria for bias less than 20% of the expected value (2 ±0.4, for a ratio-based calculation).

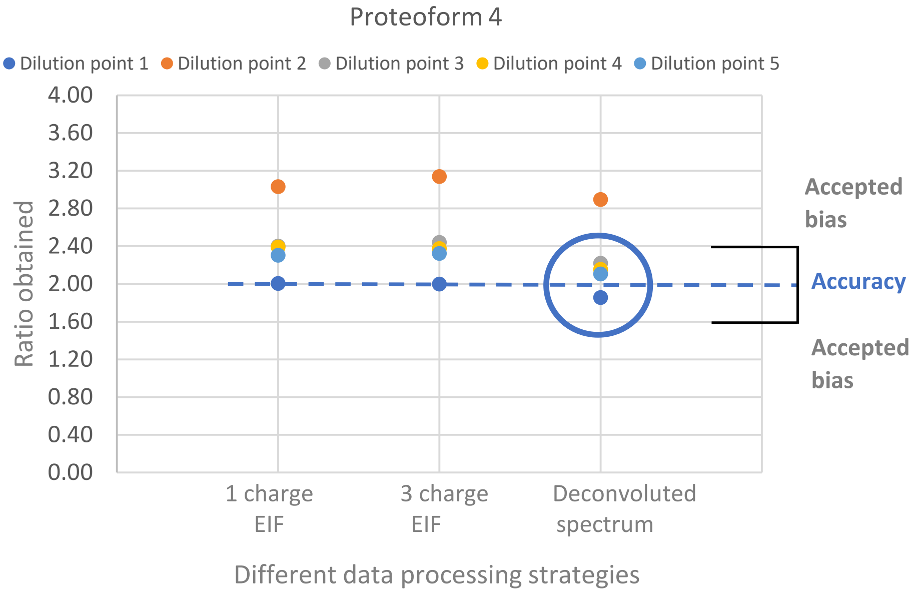


Figure 43: Accuracy & precision plot showing quantification results for Filgrastim proteoform 4, as given by 3 different data processing strategies, respectively. The coloured dots are ratios of quantification value obtained in consecutive points of dilution series (from 31ng/µL to 1000ng/µL). The accurate ratio for each dilution point is 2 and is represented by a blue dotted line. Limits of the accepted deviation in quantified ratios are denoted in square bracket.

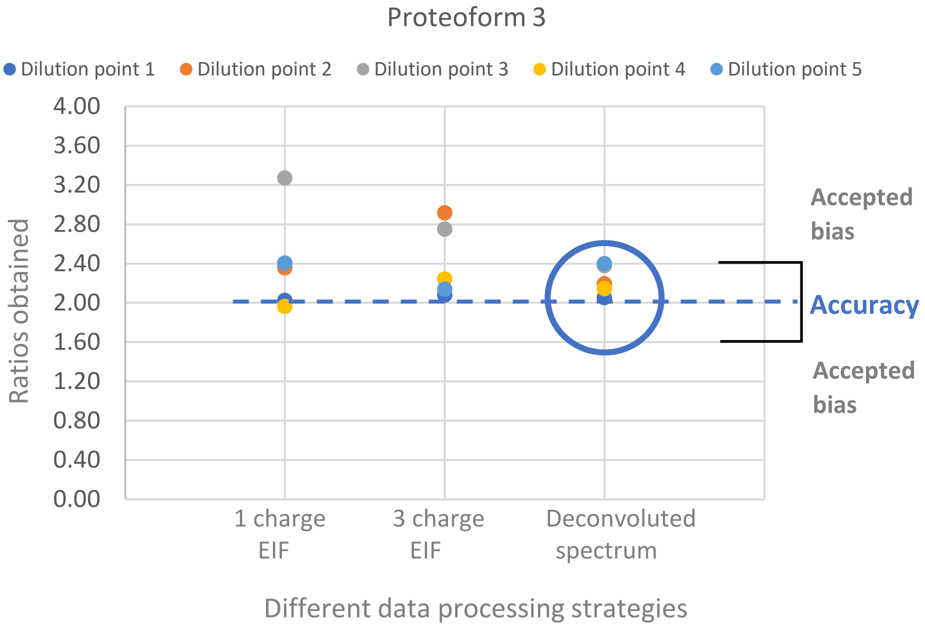


Figure 44: Accuracy & precision plot showing quantification results for Filgrastim proteoform 3, as given by 3 different data processing strategies, respectively. The coloured dots are ratios of quantification value obtained in consecutive points of dilution series (from 31ng/µL to 1000ng/µL). The accurate ratio for each dilution point is 2 and is represented by a blue dotted line. Limits of the accepted deviation in quantified ratios are denoted in square bracket.

The next comparison was made for quantification results across all Filgrastim proteoforms.

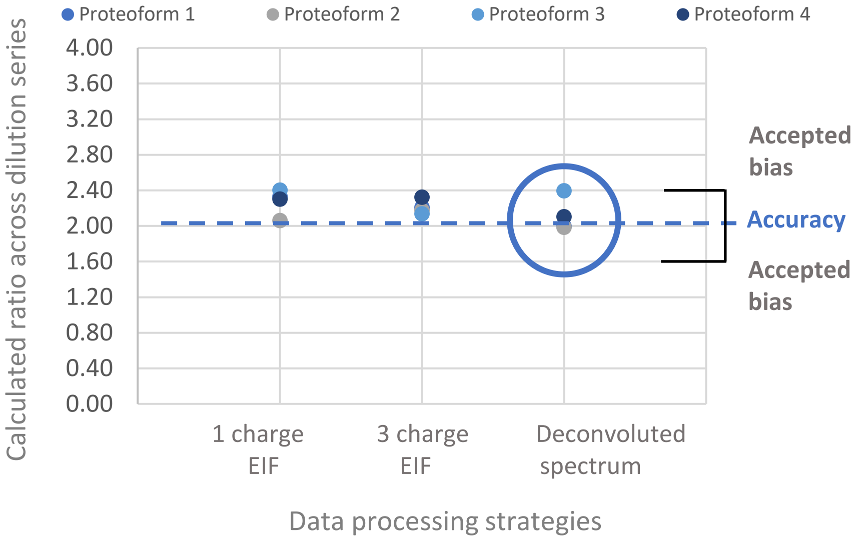


***Figure 45:* Accuracy & precision plot showing quantification results for all Filgrastim proteoforms, as given by 3 different data processing strategies, respectively. The coloured dots are representing Filgrastim proteoforms numbered 1 -4. The expected value is represented by a blue dotted line (ratio obtained for 1000ng/µL to 500ng/µL). Limits of the accepted deviation in quantified ratios are denoted in square bracket.**

As seen in Figure 45, each of the four proteoforms passed the precision criteria (< 20% deviation from the expected value) in all data processing strategies (for the ratio of quantification values at 1000 & 500 ng/µL dilution point). In between single charge state-based EIF & three charge states-based EIF approach, it is shown, that the higher the number of charge states considered, the more experimental values represented expected ratios. The best fit for accuracy and precision was observed for deconvoluted spectrum-based quantification.

6.10 Quantification of proteoforms from isotopically unresolved Ovalbumin mass spectrum

Four proteoforms of supercharged Ovalbumin harbouring different PTMs and thereby possessing different SN ratios were chosen for the comparison of EIF and deconvolution-based quantification. The four chosen Ovalbumin proteoforms were 44086 Da (monophosphorylated, acetylated Ovalbumin with H3N3 glycan), 44166 Da (diphosphorylated, acetylated Ovalbumin with H3N3 glycan), 44328 Da (diphosphorylated, acetylated Ovalbumin with H3N3+NANA glycan), and 44369 Da (diphosphorylated, acetylated Ovalbumin with H6N2 glycan) respectively. The identity of the 4 proteoforms is presented in Table 9.

6.10.1 Isotopically resolved mass spectrum of intact Ovalbumin proteoforms

Figure 46 represents the details of Ovalbumin mass spectrum obtained from the FIA-MS method. At an Orbitrap resolution setting of 17k at 200m/z on a hybrid Quadrupole-Orbitrap (QExactive) mass spectrometer, an isotopic resolution could not be obtained for Ovalbumin proteoforms. The experimental resolution achieved at proteoform level was approximately between 2000-4000 units (Blue inset of Figure 46). With this resolution & high heterogeneity of samples, the best strategy for quantification of proteoforms may differ from quantification strategy in isotopically resolved proteoforms.

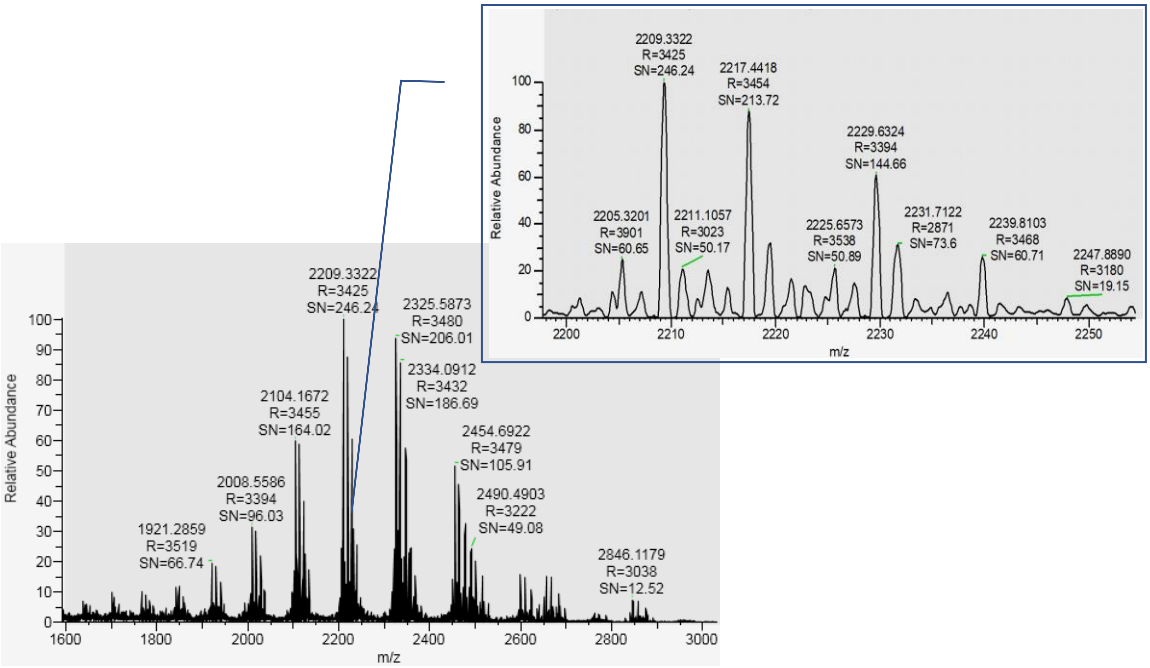


Figure 46: Isotopically unresolved mass spectrum obtained for Ovalbumin in FIA-MS. Seen below in the figure is the charge envelope of Ovalbumin showing multiple charge states. Resolution (R) and signal to noise ratio (SN) annotated overhead for respective proteoform signal. The blue inset shows a zoomed view denoting the isotopically unresolved peaks at the apex charge state.

6.10.2 Selection of spectral signals for EIF based quantification of Ovalbumin proteoforms

Figure 47 represents the details of Ovalbumin raw spectrum and denotes the peaks used for EIF based quantification. Figure 36a shows the MS1 spectrum with apex charge state highlighted in green outline. A zoomed section of the green outlined box is shown in

Figure 47b. The coloured symbols annotated on the peaks in Figure 47b represent multiple Ovalbumin proteoforms. Four annotations with masses overhead represent the peak used to generate AUC value, which is used for EIF-based quantification.

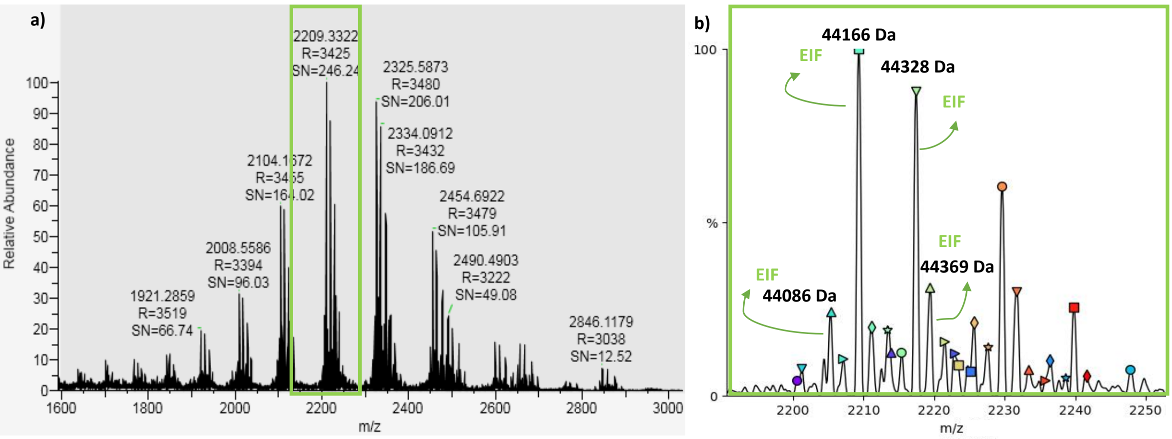


Figure 47: Scheme of EIF data processing strategy for quantification of Ovalbumin proteoform obtained in FIA-MS analysis. a) The original mass spectrum of Ovalbumin with a green box highlighting the apex charge state used for EIF quantification. b) Zoomed view of apex charge state showing coloured annotation for different Ovalbumin proteoform signals. The green arrows indicate signals used for EIF generation for the respective proteoform.

Deconvoluted spectrum-based quantification of these proteoforms used the values directly reported by UniDec software.

6.10.3 Comparing data processing strategies for quantification of proteoforms from isotopically unresolved mass spectra

Respective data processing strategies were compared for a dilution series of Ovalbumin. Two regression lines per chosen proteoform were drawn out from the values obtained for EIF and deconvolution-based quantification, respectively. The linear regression lines obtained with respective quantification strategies (with logarithmized intensity on Y-axis and logarithmized concentration values on X-axis) are presented Figure 48. The linear regression line obtained with EIF values is presented in green, while the linear regression line obtained with deconvoluted spectrum values is presented in blue. The linearity of the regression line is a measure for the accuracy of the very quantification method.

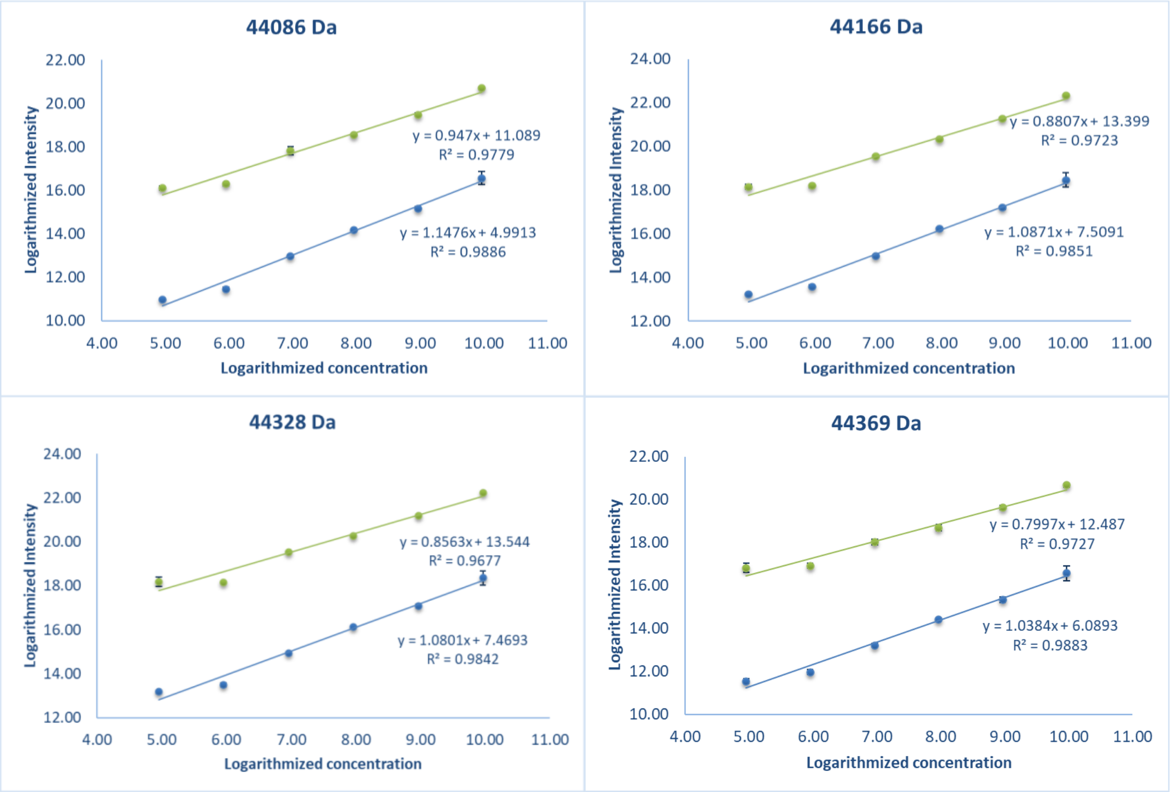


Figure 48: Regression curves for four different Ovalbumin proteoforms calculated with a single charge based EIF quantification (green) & UniDec deconvoluted spectrum-based quantification (blue). Fit of linearity represented by the coefficient of determination (R^2)

For all four analyzed proteoforms, a higher intensity was obtained from EIF based quantification. Nevertheless, both EIF & deconvoluted spectrum-based calculations gave linear trends. However, with regards to the coefficient of determination (R^2), a higher value (R^2 close to 1) was obtained for deconvoluted spectrum-based quantification (Figure 48). The second criteria for the judgment were precision values i.e the concordance between triplicates. For plotting precision, standard deviations were presented in form of a radar plot in Figure 49. The vertices of the hexagon in the radar plot (Figure 49) represent the amount of Ovalbumin injected (ng). The innermost to outermost hexagon represents 5%, 10%, 15% & 20% variances for triplicates. Both, the EIF and deconvoluted spectrum-based calculations were within the 20% variance limit. However, the single charge EIF method revealed a higher variance for different proteoforms. The variance between triplicates in EIF based method was also seen to be dependent on the concentration of proteoform sample injected (vertices of the hexagon in the radar plot).

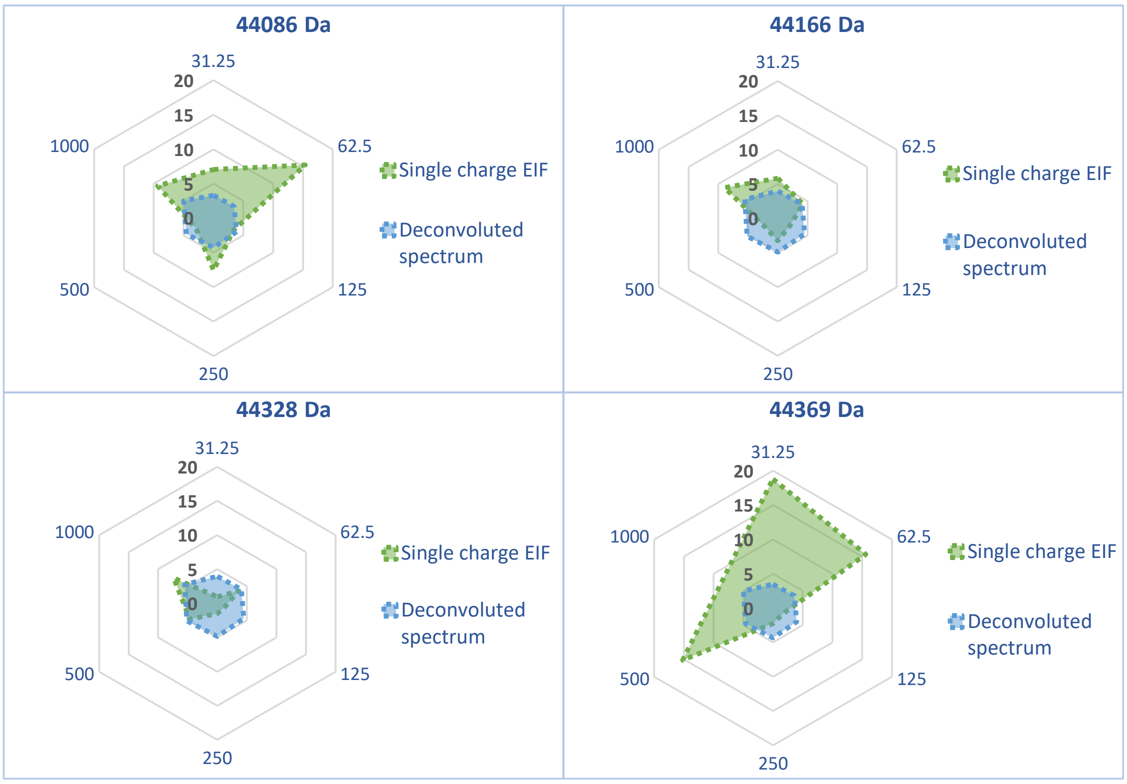


Figure 49: Precision calculated across triplicates for four different Ovalbumin proteoforms is represented as a radar plot. The vertices of the hexagon represent the amount of Ovalbumin injected (ng). Innermost to outer hexagon represent 5% to 20% deviation in quantified values. Deviation for single charge EIF based quantification values in green color & deviation obtained with UniDec deconvoluted spectrum-based quantification values in blue.

The deconvoluted spectrum-based quantification, on the other hand, showed a constant variance of approximately 5%, irrespective of proteoforms and concentration of Ovalbumin injected. Thus, based on accuracy and precision as judgment criteria, a deconvoluted spectrum presents the best strategy for reporting the quantification of proteoforms for FIA-MS-based quantification.

6.11 Deconvoluted spectrum-based quantification across different deconvolution softwares

6.11.1 Investigation of deconvoluted spectrum based quantification for samples with known concentration

From the results of previous sections (6.9.4 & 6.10.3), deconvolution-based quantification revealed higher precision and accuracy regarding proteoform quantification. For all experiments until, the UniDec tool was used for deconvolutions and deconvoluted spectrum-based quantification. To further optimize the experimental setting, UniDec was compared to different deconvolution tools concerning relative quantification of proteoforms. For comparison, the choice of four deconvolution tools was made based on the literature survey & the reported efficacy to handle isotopically unresolved spectrum. ReSpect™ (licensed by Thermo Scientific) is a deconvolution tool for isotopically unresolved spectra under the BioPharma Finder™ software package. The other two deconvolution tools namely-UniDec and MetaUnidec, belonged to the same open-source software package and operated on the principle of Bayesian deconvolution. The major difference is that MetaUniDec tool can be used for batch processing of multiple files, while UniDec tool requires individual files as an input. Also, the underlying peak extraction phenomenon for MetaUniDec and UniDec is slightly different. FLASHDeconv was considered as yet another open-source deconvolution tool used for comparison. FLASHDeconv algorithm was developed as a part of this project consortium with a major collaboration in Eberhard Karls University of Tübingen, (Jeong *et al.*, 2020).

Intensity values of four Ovalbumin proteoforms quantified by 4 different tools are shown in Figure 50 (calculated over FIA-MS spectra for 1000ng injection of supercharged Ovalbumin).

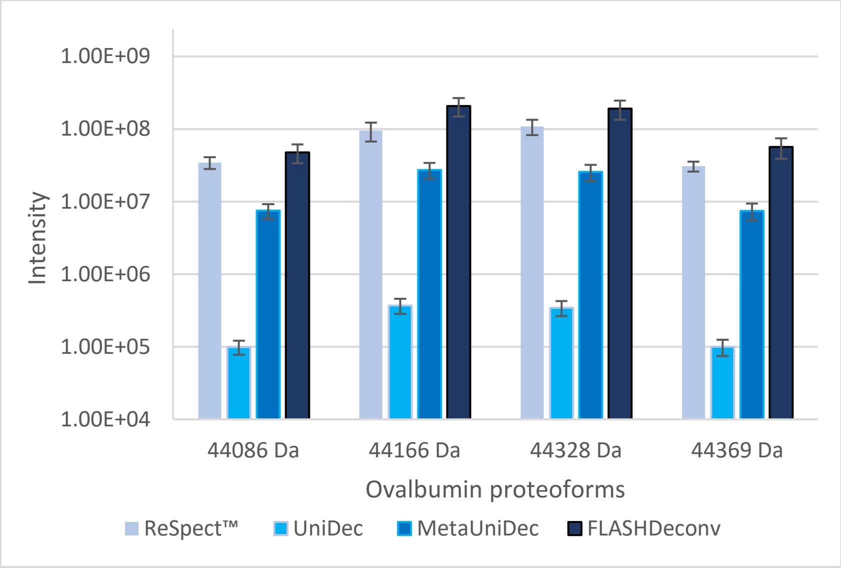


Figure 50: Comparison of intensity reported by different deconvolution tools (deconvoluted spectrum-based proteoform quantification) for same Ovalbumin analysed via FIA-MS. The absolute value (reported on Y-axis) is seen to be different for each deconvolution tool tested.

It is not possible to conclude which algorithm shows the highest accuracy concerning proteoform quantification. Herein the focus was not on the absolute quantification but only the relative quantitative results of proteoforms. In the presented Figure 50, each algorithm reported proteoforms 44086 Da & 44369 Da with similar abundance. Likewise, the relative quantities of proteoform 44166 Da & 44328 Da are also reported to be similar by all four algorithms. Thus, it could be concluded that irrespective of the deconvolution algorithm, relative quantification for a sample yields similar results.

6.11.2 Investigation of deconvoluted spectrum based quantification for samples with unknown proteoform amounts

After analysing samples with known amounts & concentrations, further analysis was performed on Ovalbumin samples with unknown proteoform quantitates. SDBC pre-fractionated Ovalbumin samples were used as samples of unknown proteoform amounts. Relative quantification across different deconvolution algorithms was evaluated. The comparative results of relative quantification for SDBC fractions with four Ovalbumin proteoforms and four deconvolution tools are presented in Figure 51.

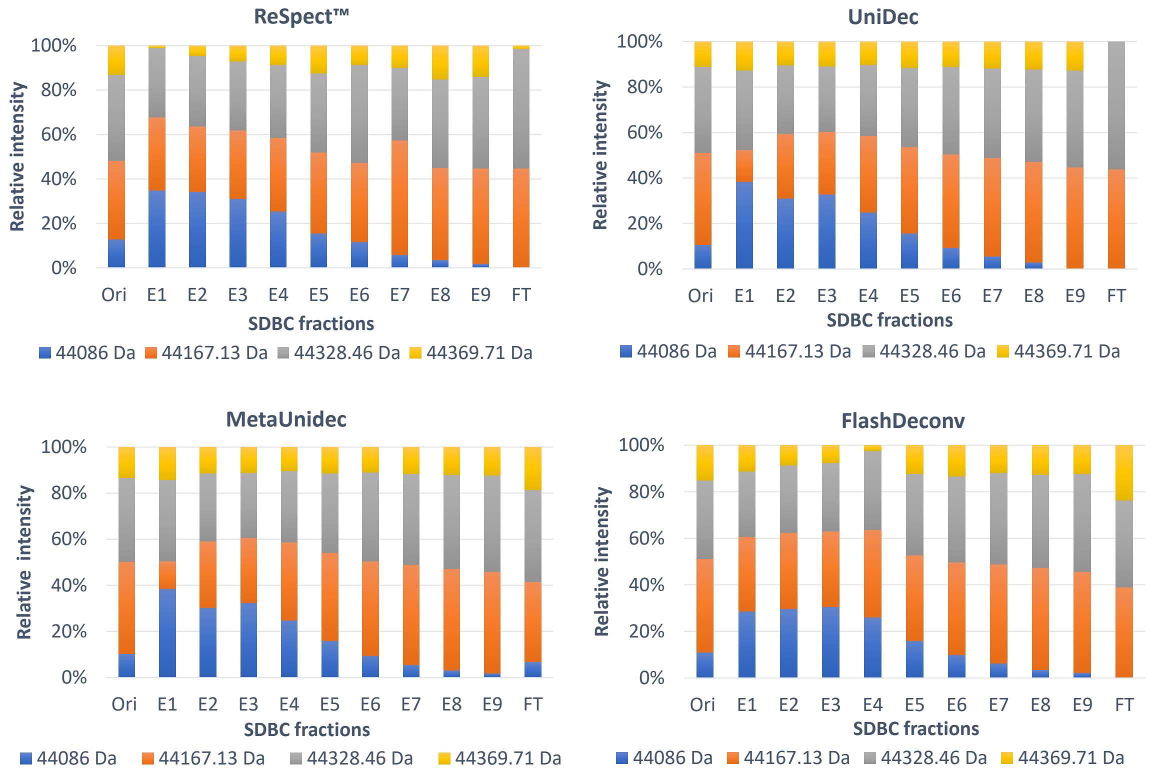


Figure 51: Relative quantification of SDBC pre-fractionated Ovalbumin samples with different deconvolution tools namely ReSpect™, UniDec, MetaUniDec, and FLASHDeconv. Relative intensities of four different proteoforms are presented on Y-axis with different four colours respectively.

The X-axis of the figure represents nine eluates of SDBC fractionation, abbreviated as E1 to E9. The abbreviation FT stands for flowthrough and Ori stands for the original sample.

The Y-axis represents the relative abundance calculated from the absolute values obtained from each of the used deconvolution tools. From figure 41 it is evident, that the relative intensity, reported by different deconvolution tools are very similar. However, at a closer look, a discrepancy among reported results can be identified for lower abundant proteoforms 44086 Da in E9 & flow-through (FT) fraction. Also noticeable is that the relative quantification given for the 44369 Da proteoform varies across four tools especially for FT fraction.

The results from UniDec (individual file processing), resembled results given by the ReSpect™ algorithm (standard algorithm for comparison). Additionally, the peak detection function in UniDec allows analysts to trace back the reported masses to the m/z signals in the raw mass spectrum. UniDec allows the analyst to decide the eligibility of a proteoform for quantification by looking at the SN ratio of that proteoform in the raw mass spectrum. Due to the possibility of validating results in open-source software, UniDec was preferred for deconvoluted spectrum-based and quantification of proteoforms in all further experiments.

6.12 Application of fast FIA-MS method for quantification of Adalimumab proteoforms

6.12.1 Optimization of FIA-MS for detection of lower abundant proteoforms from Adalimumab

With the FIA-MS method & UniDec based quantification established, the method was further applied for quantitative analysis of proteoforms from the therapeutic protein-Adalimumab. The effect of supercharging on the raw Adalimumab sample (without non - volatile salt contaminants) sample was initially tested. The mass spectrum obtained with the FIA of Adalimumab in presence of a supercharger is represented in Figure 52.

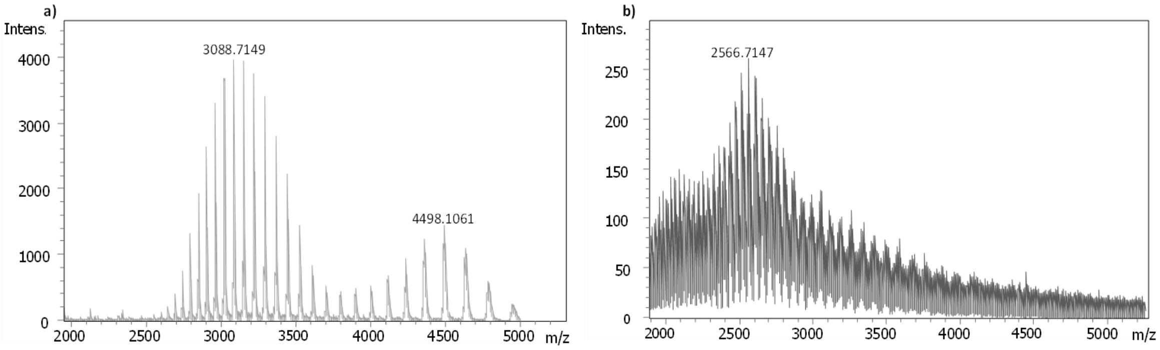


Figure 52: Mass spectra from FIA-MS of Adalimumab on MaXis II™ (Bruker Daltonics Inc.) a) without supercharger showing charge distribution at 2500 to 5000m/z. b) with supercharger spiked in sample showing charge distribution shifted to lower m/z in 2000 to 4000 range.

Supercharging of mAb sample resulted in shifting of charge state distribution to a lower m/z range of 2000 to 4000 (Figure 52b). Additional charges are acquired by the mAb sample due to supercharging phenomenon. On the other hand, non-supercharged mAb (Figure 52a) resulted in a cleaner spectrum with more distinct charge states and higher SN ratios for the proteoforms. Thus, mAb analysis was further performed without supercharging.

To improve the detection of mAb proteoforms in this FIA-MS setup, the MS parameters on MaXis II™ (Bruker Daltonics Inc.) were further optimized (for example ISCID was increased to 120eV). Additional m/z signals of the lower abundant proteoforms could be seen at the level of the raw mass spectrum (Figure 53). There was no indication of in-source fragmentation of mAb with increased energy settings applied herein.

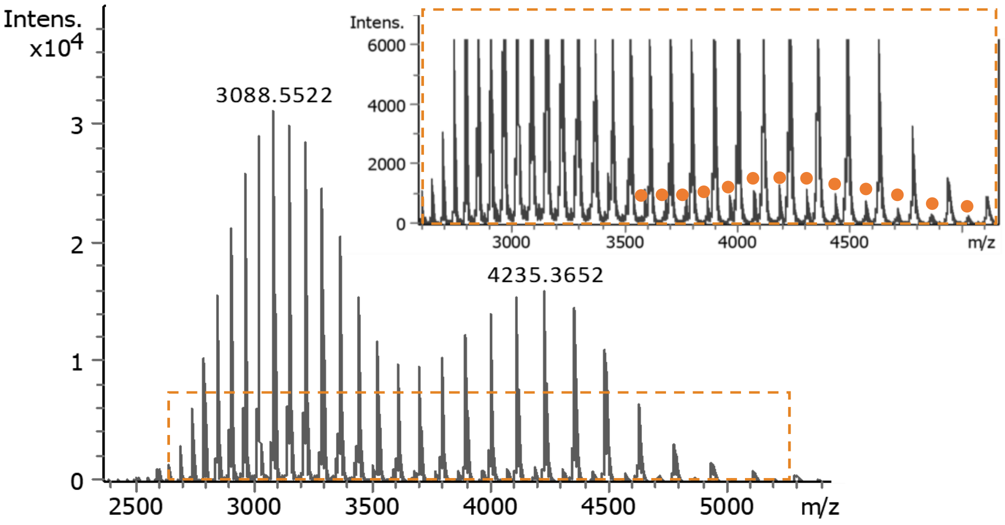


Figure 53: Improved Adalimumab mass spectrum after fine-tuning of MS parameters in FIA-MS approach. The orange inset shows the zoomed view of 2500 to 5000m/z. Signals of lower abundant proteoforms are presented in orange dots.

The deconvolution result of Adalimumab obtained with optimized FIA-MS approach is represented in Figure 53

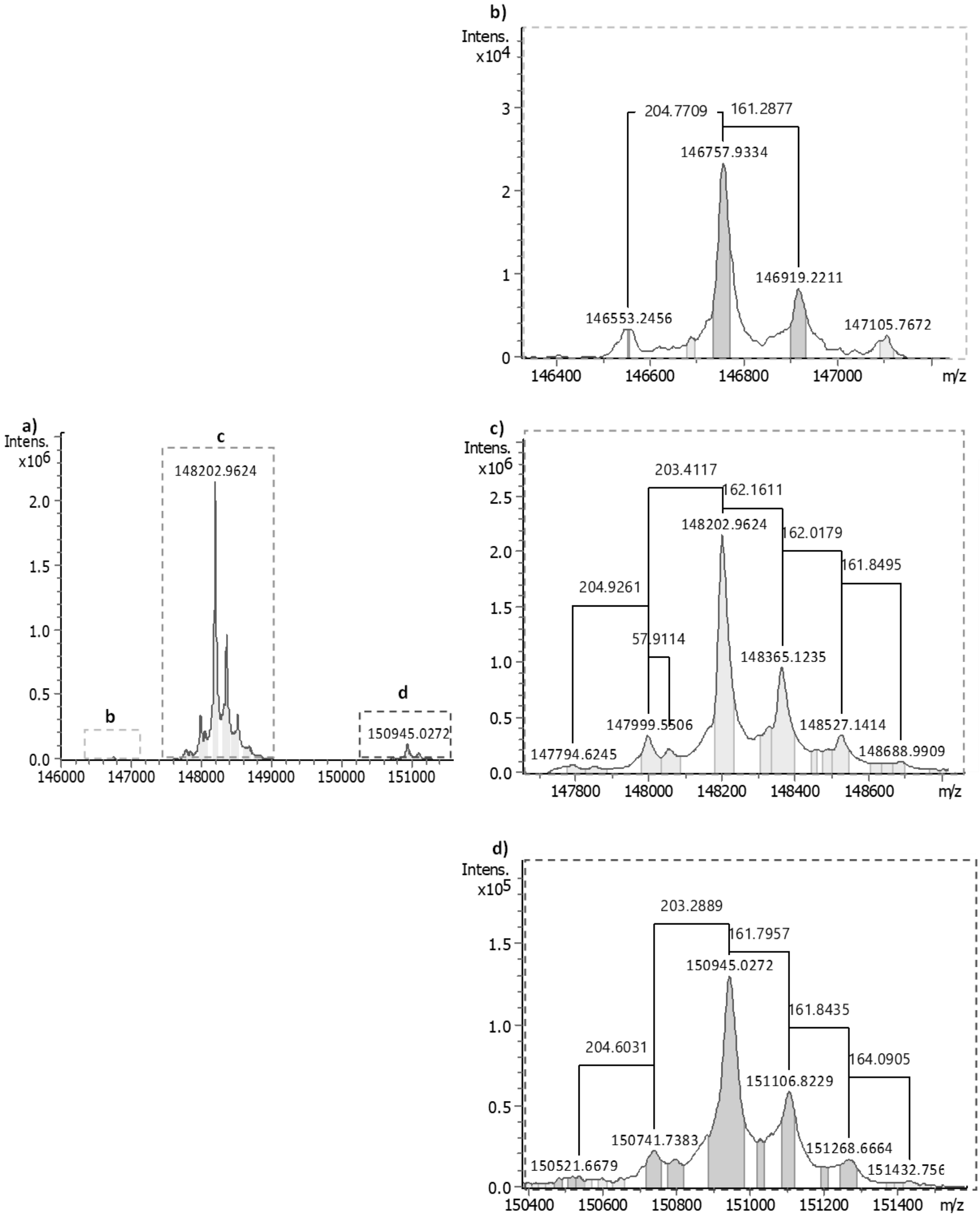


Figure 54: Detailed overview of proteoforms detected in Adalimumab sample via FIA-MS method. a) Deconvoluted spectrum of Adalimumab. b) Zoomed section of the deconvoluted spectrum at 145757 Da revealing lower abundant proteoforms c) Zoomed section of the deconvoluted spectrum at 148203 Da and associated proteoforms d) section of the deconvoluted spectrum at 150945 Da and its lower abundant proteoforms. The number overhead between two masses indicates the mass difference in Da.

The fine-tuning of MS parameters (method section 5.4.3) improved the detection of mAb proteoforms. Fourteen proteoforms were detected from the raw Adalimumab sample in the 2-min FIA-MS method. 148203 Da, 150945 Da, and 146757 Da were the main proteoform masses detected. The details of the masses identified are presented in the following section.

6.12.2 Identity of the higher molecular weight proteoform detected in Adalimumab sample

Referring to the mass of deglycosylated mAb & the mass of N glycans in mAb, 148203 Da-the most abundant proteoform detected in FIA-MS was identified as a G0F-G0F form. The 162 Da mass increment to 148203 Da detected for masses- 148365 Da & 148527 Da was associated with additional hexose. These proteoforms 148365 Da & 148527 Da were thereby identified as G0F-G1F & G1F-G1F respectively. The mass difference between the most abundant G0F-G0F form (148203 Da) & lower abundant proteoforms (146757 Da) corresponded to a loss of H3N4F1 glycan (-18 Da for loss of water). Thus, 146757 Da represented mAb with a single G0F residue. The mass difference between the 150945 Da proteoform & the most intense G0F-G0F proteoform (148203 Da), however, did not correspond to any identified N-glycan.

6.12.2.1 Middle-down analysis of Adalimumab sample for identification of higher molecular weight proteoform

A middle down approach was tested to further identify whether the detected 150945 Da mass was an Adalimumab proteoform. IdeS digestion was performed as indicated by the manufacturer to break mAb into a F(ab')$_2$ & Fc region (Figure 55).

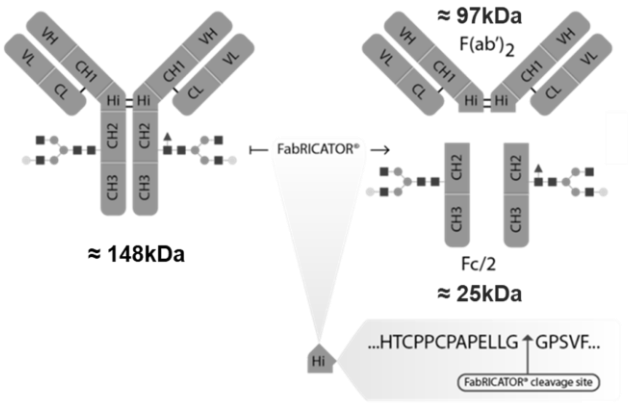


Figure 55: Scheme of mAb fragments generated after digestion with FabRICATOR enzyme. Figure adapted from Genovis https://www.genovis.com/products/igg-proteases/fabricator/

The masses identified in middle down mAb analysis using the FIA-MS method are represented in Figure 56.

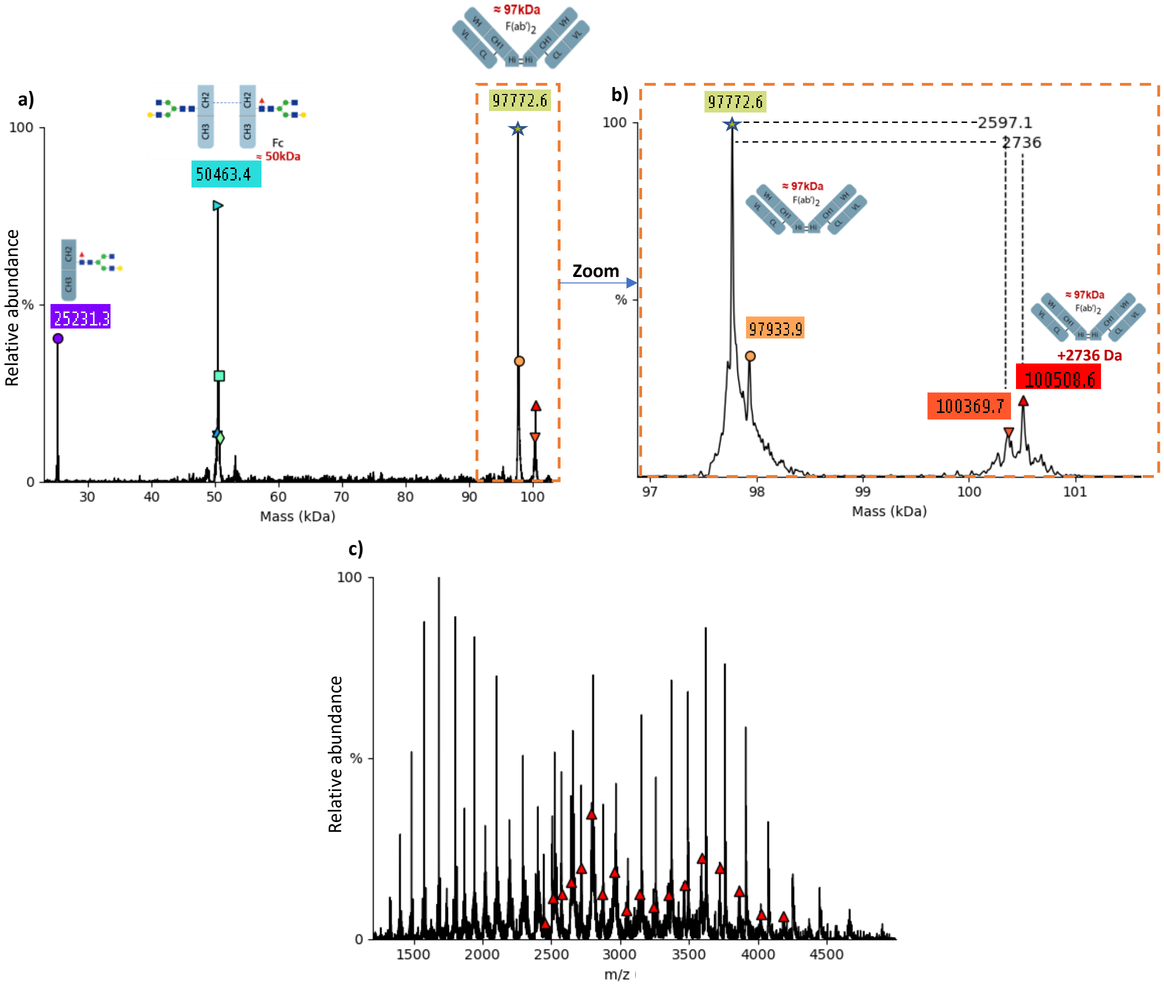


Figure 56: Detailed overview of IdeS digested Adalimumab analysed via FIA-MS. a) Deconvoluted spectrum of IdeS digested Adalimumab shows three distinct masses namely 25231.3 Da, 50463.4 Da, 97772.6 Da, each annotated in a different colour. B) Zoom in on the higher molecular weight section marked in an orange inset in a) shows two other significant masses- 100508.6 Da & 100369.7 Da respectively. The mass difference between the masses is denoted in the overhead dotted lines. c)The m/z signals contributing to mass 100508.6 Da annotated in the raw MS^1 spectrum with red triangles.

Two of the masses identified after IdeS digestion 97772.6 Da & 25231.3 Da closely represented expected masses of F(ab')$_2$ & reduced Fc/2 region. The 50463.4 Da signal matched the mass of (non-covalently attached) Fc fragments of the mAb. However, the additionally detected mass of 100508.6 Da was not reported in the literature for the middle-down analysis of Adalimumab. As seen from Figure 56c, the detected 100508.6 Da mass was not a deconvolution artifact. The signals for this mass (100508.6 Da) could be traced back to the raw mass spectrum of IdeS digested mAb (annotated as red triangles). Taking into consideration the disulphide bonds in mAb, this 100508.6 Da mass detected in middle down analysis, complemented the results of intact mAb FIA analysis wherein 150945 Da proteoform was detected (100508.6+50463.4-18-8= 150946 Da). The detected mass of 150945 Da is thus confirmed to be an Adalimumab proteoform (suggested to be mAb with intact signal peptide at the F(ab')$_2$ region).

6.12.3 Quantification of proteoforms in SDBC fractionated Adalimumab

The optimized FIA-MS settings for mAb were further applied for analysis of SDBC fractionated Adalimumab. SDBC fractionation was performed (offline to MS) using two sets of buffer systems at pH 6 and pH 9, respectively. The SDBC fractionated Adalimumab were ten eluate fractions (labelled E1 to E10) comprising unknown amounts of respective mAb proteoforms. Flow-through fraction is labelled as Ft while the original untreated sample is labelled as Ori. Relative quantification was performed using intensities of deconvoluted spectra obtained with the UniDec tool. The quantification results of SDBC fractionated Adalimumab proteoforms at pH 6 and 9 respectively, are presented in Figure 57 and Figure 58.

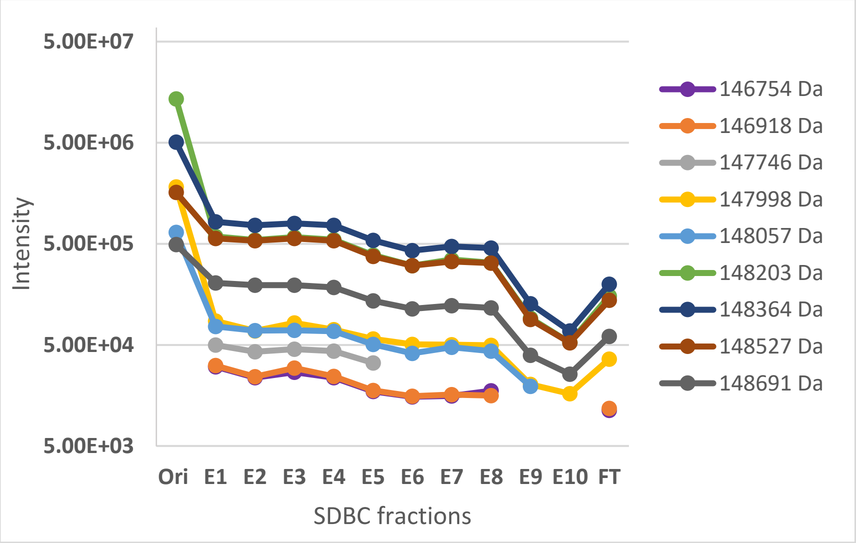


Figure 57: Relative quantification of Adalimumab proteoforms identified across 10 SDBC fractions, wherein SDBC was performed at pH 6. Eluates from SDBC fractions are labelled E1-E10, Ori- Original sample, and FT -flow through fraction, and are represented on X-axis. The intensity for each proteoform was obtained from UniDec deconvolution software (Y-axis).

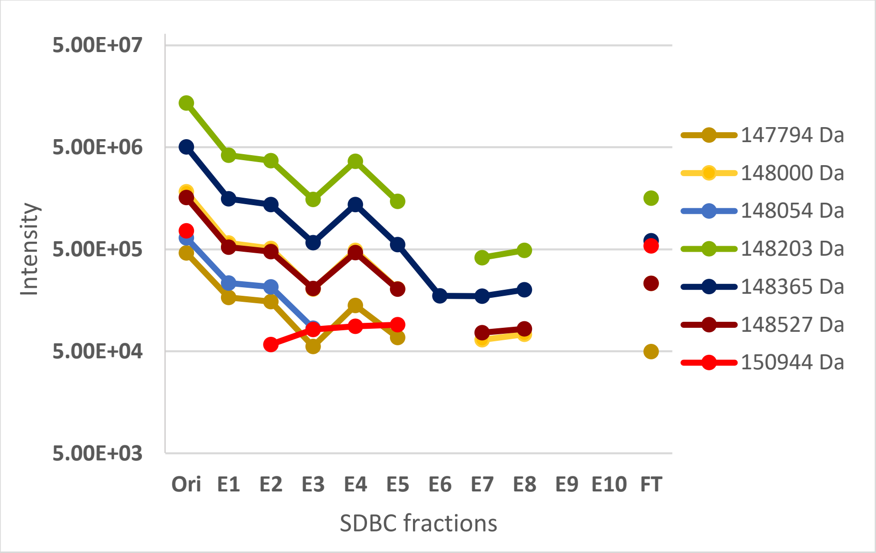


Figure 58: Relative quantification of Adalimumab proteoforms identified across 10 SDBC fractions, wherein SDBC was performed at pH 9. Eluates from SDBC fractions are labelled E1-E10, Ori- Original sample, and FT -flow through fraction, and are represented on X-axis. The intensity for each proteoform was obtained from UniDec deconvolution software (Y-axis).

Nine proteoforms were detected in all the SDBC eluate fractions, fractionated at pH 6. As expected, with the increasing fractionation steps, the relative abundance of proteoforms decreases. It is evident, that the proteoform with mass 147746 Da was not detected after fraction E5. The proteoform 147746 Da identified in SDBC fractionated samples was not identified with a quantifiable SN ratio in the raw/original sample. Additionally, proteoforms 146754 Da & 146918 Da were quantified only up to fraction E8 and then recovered from FT fraction (Figure 57).

The results from SDBC samples fractionated at pH 9 reveal seven proteoforms, that were detected across all the fractions. The concentrations of all proteoforms are decreasing with increasing fraction numbers, except for fraction E4. Only one proteoform (148365 Da), was detected in the E6 fraction that matched the proteoforms present in other fractions. From Figure 58, it is also evident that the high molecular weight mAb proteoform- 150945 Da, is detected only up to eluate 5 (E5). As opposed to other fractions, 150945 Da proteoform also showed an increasing quantitative trend in SDBC fraction (red line in Figure 58). The increased relative abundance of 150945 Da proteoform is also observed in the flow-through fraction. This quantitative result suggests that 150945 Da proteoform was enriched by SDBC at (basic pH) pH 9 fractionation.

7 Discussion

7.1 RPLC-MS as fast method for proteoform quantification

Initial efforts for establishing fast proteoform quantification were carried with commercially available Ovalbumin (grade 5) as model protein and using RPLC-MS as a method for analysis. At the chromatographic level of RPLC-MS analysis, no separation of Ovalbumin proteoforms was observed even with longer gradient elution. This is because most Ovalbumin proteoforms possibly have similar hydrophobicity values. Similar results pointing to difficulty in separating proteoforms based on hydrophobicity was presented by Bartonek, Braun and Zagrovic, 2020.

Additionally, problems regarding incomplete recovery of proteoforms from the RP column (Figure 13), as well as difficulties in unambiguous proteoform detection (as seen in Figure 24) were noticed in the RPLC-MS analysis setup. Incomplete recovery of proteoforms from the RP column will translate to biased quantitative analysis of only recovered proteoforms and does not suit our purpose of total proteoforms quantification.

A prominent reason for incomplete recovery observed herein could be on-column (RP column) precipitation of certain full-length proteoforms, due to the use of organic solvents. This limitation in the use of RPLC for analysis of intact proteoforms, especially with an increased molecular weight of proteoforms, has been also reported early on by Welinder, Sørensen and Hansen, 1987. Another prominent reason for incomplete proteoform recovery from the RP column could be secondary interactions of certain (charge bearing) proteoforms with the stationary phase matrix. Cases of secondary interactions are comparatively severe in silica-based RP columns and can go as far as forming a quasi-irreversible interaction (Mathé *et al.*, 2013). The degree of proteoform retention onto the RP column depends on the types and the number of interactions involved. The RP column used in this study was a polymer-based phenyl monolithic RP column. Among the other types of RP columns, polymer-based monolithic RP columns have been proven to be superior in terms of total protein recovery, also for proteins as big as 150kDa monoclonal antibody (Fekete *et al.*, 2012). Butyl ligand monolithic RP columns were further reported to have more likely adsorption effects than phenyl ligand-based monolithic columns (Aasim *et al.*, 2018). Thus, a phenyl ligand-based monolithic RP column used in our experiment was already an advantage point with attaining minimum on-column proteoform adsorption effects.

Though monolithic RP columns specifically present a lot of advantages for efficient protein recovery, certain proteoforms can still adsorb via strong hydrophobic interactions (Rizvi, 2010) as can also be seen in our experimental setup.

Proteoforms recovering from the monolithic RP column were quantified in this thesis, using the bottom-up MS approach. Based on my literature review, the DIA-based quantitation performed in this thesis for studying intact proteoform fraction surviving the RPLC analysis is the first of its kind. Figure 15 indicated lower amounts of 'total proteoform fraction' recovering from the monolithic RP column in comparison to the proteoform sample that never faced the RP column. However, these results represent peptides quantified with the tandem MS approach. Peptides with reduced intensities in the "RP column eluted Ovalbumin" sample, were not associated to typical modifications or regions of intact Ovalbumin. Consequently, decreased intensity in peptides could not be extrapolated to the loss of an entire proteoform containing that peptide. Also, as proteoforms can share multiple peptides, it is impossible to trace the peptide to its original proteoforms with such a quantification approach in the 'bottom-up MS' approach. The sequence coverage obtained for the protein in the bottom-up analysis also is an important factor in grading these results. Nevertheless, a relative quantitative comparison of respective protein fractions performed in section 6.2.1, is still indicative of the fact that not all proteins or specially proteoforms will survive the RPLC analysis before detection by the respective detector.

Similar to results in my work, the on-column adsorption of intact protein/proteoforms onto the monolithic RP column can also be supported by the findings of Aasim and colleagues. By studying the chemistries of surface energies of protein on a monolithic RP column, the authors concluded that some proteins might present reversible to strongly irreversible interaction due to the dehydrated state of the proteoform (Aasim *et al.*, 2018). The advances in column chemistries with end-capping agents already offer a good improvement in the recovery of proteins from RP columns. However, it does not eliminate the possibility of some proteoforms retaining back on the RP column (Vailaya and Horváth, 1998) (Kopp *et al.*, 2020).

7.2 FIA-MS approach and its sensitivity in fast proteoform detection

The approach of flow injection analysis coupled to MS (FIA-MS) used in this thesis, was put forth as an alternative to RPLC-MS for fast quantitative analysis of proteoform. FIA-MS has been reported previously for qualitative (Allen *et al.*, 2003) as well as quantitative analysis (Nanita, 2013) but only for metabolites or for small protein with negligible proteoform complexity (Roberts, Green and Morris, 1997). FIA-MS method established herein is mainly aimed to be applied for fast quantitation of full-length proteoforms from purified TP (like monoclonal antibodies) or TP at different stages of production & downstream purification.

The flow injection analysis method established herein is set for 4 min/sample, where the initial 2 mins were dedicated for proteoform data acquisition, and the rest 2 mins were dedicated to (higher flow rate) flushing step to avoid carryover from the sample. A 2-min proteoform detection time was considered in the current setup (75uL/min flow rate) to allow enough ion sampling rate for the heterogeneous proteoform sample. The method can be easily reduced to a minute per sample or less, provided that enough data points are acquired for proteoform analysis. Considering 4 mins analysis time per sample, 360 runs can be achieved within a day. This presents a higher analytical speed than most other fast gradient LC methods (average 5 min for RPLC-MS, average 20 mins for SEC-MS, average 30 mins for IEX-MS) currently used in the analysis (Regl *et al.*, 2019) (Bondarenko *et al.*, 2009) (Haberger *et al.*, 2016) (Leblanc *et al.*, 2017). Higher sampling rates like 15 sec/sample are now eventually possible, but only with more complex and expensive robotic handling system like the Agilent RapidFire system (Sawyer *et al.*, 2020).

Along with the speed of analysis, the sensitivity of proteoform detection was also addressed in the current thesis, which is an important factor towards achieving reliable quantification (significance detailed in section 2.7). The sensitivity of proteoform detection in ESI-MS is in turn related to efficient ionization of sample. Multiple factors like basicity, volatility of solvents affect the efficacy of ionization process in ESI-MS (Kiontke *et al.*, 2016). In terms of volatility of solvents, organic solvents like methanol, acetonitrile, or isopropanol are generally preferred in ESI-MS-based proteomics, because of their higher GB (Generalized Born). The higher GB implies that these solvents would evaporate faster in the desolvation process of ESI-MS (Iavarone, Jurchen and Williams, 2000). However, these above-mentioned organic solvents also tend to denature full-length proteoforms considered in our analysis. Additionally, results from a study from Griebenow and Klibanov established that, when the aqueous component of solvent diminishes in ESI droplet, the tendency of an intact protein/proteoform to denature is even greater than in a binary aqueous-organic solvent (Griebenow and Klibanov, 1996). Binary aqueous-organic solvent i.e., 40%ACN experimented in this thesis work for FIA-MS approach, provided similar (denaturing) conditions and strongly impacted the SN ratio of ionized full-length proteoforms (as seen in Figure 24).

For the choice of solvents in the FIA-MS approach, solvents aiding narrow charge state distribution were examined. The narrower proteoform charge state distribution/envelope in ESI-MS, the more would be the charge state resolution seen among ionized proteoforms. This narrow charge envelope of proteoform will ultimately help in the specificity of detection and thereby quantification. SN ratios of proteoforms were increased with the use of water or ammonium acetate as spray solvents for FIA-MS. These results in my work are also supplementary to the findings of Kafader *et al.*, 2020 and Donnelly *et al.*, 2019.

Another important factor addressed in this thesis (Figure 20) was the influence of solution basicity provided by water on ionizing proteoforms, in comparison to ammonium acetate solution (widely used solvent in native MS for intact protein/proteoform analysis). Proteoform ionization using water as spray or sample application solvent in the FIA-MS approach produced higher protein signal intensities than 150mM ammonium acetate. Results presented by Uetrecht *et al.*, 2019 also demonstrated that higher protein ion intensity was obtained at least concentration of ammonium acetate solution used for MS analysis for intact proteins/proteoforms. This is due to the higher acidity provided by water over ammonium acetate solution (Figure 20). Though water has neutral pH, the ESI water droplets tend to be more acidic than water in solution (Iavarone, Jurchen and Williams, 2000).

The optimized FIA-MS method, with water as spray solvent was suggested as a superior alternative to the popularly used RPLC-MS method for fast analysis of single protein (comprising multiple proteoforms) sample.

7.3 Supercharging for improved proteoform detection

The results in Figure 31, point out that the presence of even millimolar concentrations of non-volatile salts can strongly suppress proteoform ionization, specifically compromising the detection of low abundant proteoforms. It was necessary to resolve this problem to achieve quantification of all or maximum proteoforms present in the sample.

‘In-solution supercharging’ was successfully evaluated in the current thesis for the elimination of non-volatile salt adducted to proteoforms (Figure 32) & improving the detection of lower abundant proteoforms in a sample (Figure 36). The use of supercharger sulfolane has been previously demonstrated for some proteins (Cassou and Williams, 2014), but in this thesis application of supercharger was explored specifically for effective detection and quantification of lower abundant full-length proteoforms.

For purpose of supercharging, 5 % v/v sulfolane was spiked in the sample solution prior to FIA-MS injections, according to reference with literature (Miladinović *et al.*, 2012) (Cassou and Williams, 2014) (Peters, Metwally and Konermann, 2019). Among the three proteins evaluated in the thesis, the efficacy of proteoform desalting due to supercharging agent sulfolane was demonstrated clearly for Filgrastim proteoforms (Figure 32). As Filgrastim was comparatively smaller with no complex PTMs like phosphorylation or glycosylation, the mass spectrum obtained in FIA presented a smaller number of proteoforms. Thus, it was easier to follow the loss of non-volatile salt adducts from Filgrastim proteoforms in presence of supercharger sulfolane. Unlike Filgrastim, the elimination of adducted Na ions was not seen at the mass spectrum level for bigger proteins like mAb. This is because the MS instrument and associated analyser used herein, does not have resolution powers to distinguish the small mass difference like Na ion on a 148k Da mAb. This limitation of MS analysers to resolve large proteoforms specially with Na adducted forms was also reported by Lössl, Snijder and Heck, 2014.

Another important factor indicated in this thesis was that supercharging induced in-droplet denaturation of full-length protein/proteoforms is a protein-specific phenomenon. The ambiguity in the mechanism of protein supercharging and resultant protein denaturation was also presented by Konermann *et al.*, 2019. In the current work, the results from supercharging denoted that the use of 5 %v/v supercharger did not lead to total denaturation of Ovalbumin proteoforms (Figure 35), but the denaturation for mAb proteoforms due to supercharging was seen to be higher. The additional mAb proteoforms ionizing because of supercharging effect resulted in more overlapping charge envelopes (at lower m/z range). This phenomenon was disadvantageous for the following deconvolution and quantification process of supercharged mAb proteoforms (Figure 52).

7.4 Establishing quantification for full-length proteoforms in FIA-MS approach

Accurate quantification of full-length proteoforms was described to be a challenging process due to the complexity of MS data (Labowsky, Whitehouse and Fenn, 1993). The molecular weight of proteoforms, peak width of ionized proteoform seen in the mass spectrum, resolution achieved, are among others, some important factors affecting the accuracy of full-length proteoform quantification (Ruan *et al.*, 2011) (Pohl et al., 2020) (Kellie *et al.*, 2020). Some of these challenges towards achieving proteoform quantification are briefed also in section 2.7 of the introduction. As in this thesis, most of the MS-based proteoform quantifications use full scan mass spectrum (Schaffer *et al.*, 2019) (Donnelly *et al.*, 2019). Only a few studies are detailing the use of fragment ion data for quantification for intact proteoforms. For example, Holt and colleagues reported the quantification of specific H4 histone proteoforms by using fragment ions from ETD fragmentation, wherein specific precursor ions were submitted as inclusion list and isolated with a 1m/z window. MS^2 level proteoform quantification was used in this case, as the MS^1 level quantification did not resolve the isobaric acetylated and trimethylated histone proteoforms (Holt, Wang and Young, 2019). Further on, constraints in using conventional techniques like multiple reaction monitoring (MRM) for proteoform quantification are also reported and are mostly associated with limited transmission capacities of currently available triple quadrupole mass spectrometers for heavy intact protein ions (Wang *et al.*, 2017).

For achieving accurate masses and quantification values from proteoform data, it is necessary to consider the quality of the full-scan spectral signals like the signal to noise (SN). SN ratio of full-length proteoforms in ESI-MS analysis is very low compared to the counterpart surrogate peptide signals acquired in the bottom-up MS approach and is a bottleneck for proteoform quantification. Improving the SN ratio of ionizing proteoforms was also a section that was focused on in this thesis, ultimately for achieving accurate quantification results. To obtain higher SN ratios for proteoforms, optimizations were performed in the FIA-MS approach, which is already addressed in previous sections 6.3.5, 6.5, 6.6 of the thesis. Later, SN>10 at the original mass spectrum was the basis for quantitative data processing for proteoforms in the approach followed (Figure 39).

Further on, data processing strategies for proteoform quantification can be based either on one or more charge states or the deconvoluted spectrum. These possibilities of data processing for obtaining proteoforms quantification were evaluated in detail in this (result section 6.8, 6.9, 6.10) thesis. Initially, the efficacy of charge state-dependent quantification, based on extracted ion flowgram (EIF) strategy, was evaluated (result section 6.9.2). Quantification based on the most intense charge state has been previously documented by Roman and Murphy, 2017 (using extracted ion chromatogram), but only for the main proteoform and not for associated lower abundant proteoforms in the sample. For quantification of proteoforms using an isotopically resolved dataset, the efficacy of quantification was not varied by processing one or multiple isotope peaks for a typical proteoform. A similar observation was put forth by Kellie *et al.*, 2017. Nevertheless, it is necessary to be consistent in processing the same number of isotopes across different datasets for accurate quantification (specifically when the generation of EIF is a manual process, like in current thesis work). Under the EIF strategy, the accuracy of quantification values was noticed to be better for three charge states-based quantification, rather than single charge state quantification, especially for proteoforms in lower concentrations range (Figure 43 & Figure 44). A similar observation was made by Qiu *et al.*, 2018, wherein the inclusion of more charge states for quantification was reported to give more accurate data.

The other data processing strategy considered for achieving quantification was deconvoluted spectrum-based proteoform quantification, which is an algorithm-specific approach. The intensities reported in deconvoluted spectrum-based quantification, are associated with the height of deconvoluted peaks. This is in turn, calculated based on the summed intensity of all charge states associated with respective proteoforms. Deconvoluted spectrum-based quantification was found to perform better than EIF based quantification for isotopically resolved as well as unresolved proteoform signals. A similar observation was made by Pohl *et al.*, 2020. Deconvoluted spectrum was also used by Bern et al, for reporting quantification of glycoproteoforms varying in sialic acid content (Bern *et al.*, 2018).

As the next step, details of deconvoluted spectrum-based proteoform quantification were evaluated. The complexity and challenges associated with deconvolution of intact proteoform MS data are briefed in section 2.7 of the thesis. Different data analysis software/deconvolution tools have different underlying deconvolution algorithms. Thus, the quantification value given by each software will depend on the pick picking/feature extraction and associated deconvoluting process used in the respective software. Most of the proteoform quantification is currently dependent on expensive licensed software using iterative charge-based algorithms like ReSpect™ (Thermo Scientific™), MaxEnt (MaxEnt Solutions Ltd, Cambridge, UK) (Bern *et al.*, 2018).

One of the objectives in the current thesis was to also evaluate whether the relative quantification values reported by different open-source, (free) deconvolution tools correlated to the value reported by licensed deconvolution tool- ReSpect™. The relative quantity reported by ReSpect™ -Thermo Scientific was considered as standard for comparison as it a licensed deconvolution tool reported in many studies for the quantification of therapeutic proteins (Füssl *et al.*, 2019) (Wohlschlager *et al.*, 2018).

According to my fullest knowledge, this is one of the first reports where deconvolution tool-based comparison is presented for quantification of full-length proteoforms. The quantification values given by four deconvolution tools considered in the thesis were found to be complementary. Other known deconvolution tools like- TopFD, MS-Deconv, ProMex, and others (part of open-source data analysis suite - Mash Explorer) were not included in the current thesis for algorithm comparison because these tools provide accurate proteoform quantification only for isotopically resolved mass spectrum. Thus, deconvolution tools within Mash Explorer Suite (Wu *et al.*, 2020) are not best suitable for quantification of isotopically unresolved mAb proteoforms- which is our TP of interest.

Among the four deconvolution tools compared in this thesis, (at the current stage) FLASHDeconv algorithm, is one of the fastest algorithms currently available for intact protein or top-down MS data (developed together with collaborators in Kohlbacher lab, as part of the current A4B project consortium). However, FLASHDeconv lacks the scoring function required for the validation of proteoform quantifications. MetaUniDec-a batch file processing deconvolution tool (also from the UniDec software package) was not most suitable for quantification of fractionated proteoform samples. The peak extraction threshold setup for batch file processing in MetaUniDec can often over or underestimate the quantification results for individual fractionated samples (overestimated results seen in the current study). Ultimately it must be noted that deconvolution and thereby quantification results for full-length proteoforms are also hugely dependent on the data processing parameters set in the deconvolution algorithm. The importance of appropriately fine-tuning processing parameters for analysis of complex intact proteoform mass spectrum is also presented by Cleary *et al.*, 2018.

UniDec was preferred as the deconvolution tool of choice for reporting proteoform quantifications due to the speed of data processing, abilities to tackle/reduce deconvolution artifacts & the possibility of validating results with the scoring function (Marty, 2019) (Marty, 2020). The superiority of simulated/fitting algorithms (like UniDec) over other algorithms for deconvolution of proteoform data is also highlighted by Peris-Díaz *et al.*, 2020.

8 Conclusion and outlook

The aim of establishing a fast quantitative detection method for proteoforms from therapeutic proteins (TPs) was successfully fulfilled with the FIA-MS method. The total analysis time of 4 mins/sample in our proposed FIA-MS method, outweighed the analysis time offered by the conventionally used RPLC-MS or SEC-MS method (size exclusion chromatography)- for analysis of TPs. The fast FIA-MS method reported herein offers an economical setup for sensitive proteoform detection, as it simply uses the existing robotic sampling mechanism of the UPLC system for delivering the sample to MS. Additionally, improved detection of lower abundant proteoforms was successfully demonstrated with the use of sulfolane for in-solution supercharging. The detailed evaluation of proteoform quantification strategies in ESI-MS data was a major part of the thesis and relative proteoform quantification using deconvoluted spectrum was proven to be the most efficient quantification strategy for both- isotopically resolved and unresolved MS data sets.

However, proteoforms possessing only a slight mass difference between one another are still challenging targets for accurate quantification due to the complex peak detection involved in the process. The scheme of deconvoluted spectrum-based proteoform quantification reported in this thesis is also limited to quantification of only non-isobaric proteoforms. The MS/MS fragmentation necessary to solve this ambiguity of full-length isobaric proteoforms is currently underdeveloped and is also limited by fragmentation techniques available on the MS system. A combination of successful proteoform sample fractionation, sensitive mass spectrometry analysis, and data processing are necessities for accurate proteoform quantification in the ‘intact protein MS or top-down MS’ approach. Accessible technological advances like modifications within the MS systems for enhanced transmission and fragmentation of heavier ions would further help to improve proteoform detection and thereby quantification.

9 References

Aasim, M. *et al.* (2018) ‘Protein adsorption onto monoliths: A surface energetics study’, *Engineering in Life Sciences*. doi: 10.1002/elsc.201700097.

Aebersold, R. *et al.* (2018) ‘How many human proteoforms are there?’, *Nature Chemical Biology*. doi: 10.1038/nchembio.2576.

Agarwal, K. C. *et al.* (1975) ‘Purine Nucleoside Phosphorylase. Microheterogeneity and Comparison of Kinetic Behavior of the Enzyme from Several Tissues and Species’, *Biochemistry*. doi: 10.1021/bi00672a013.

Albalat, R. and Cañestro, C. (2016) ‘Evolution by gene loss’, *Nature Reviews Genetics*. doi: 10.1038/nrg.2016.39.

Alexandridou, A. *et al.* (2009) ‘UniMaP: Finding unique mass and peptide signatures in the human proteome’, *Bioinformatics*. doi: 10.1093/bioinformatics/btp516.

Allen, J. *et al.* (2003) ‘High-throughput classification of yeast mutants for functional genomics using metabolic footprinting’, *Nature Biotechnology*. doi: 10.1038/nbt823.

Baldwin, A. J. *et al.* (2015) ‘Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles’, *Analytical Chemistry*, 87(8), pp. 4370–4376. doi: 10.1021/acs.analchem.5b00140.

Bartonek, L., Braun, D. and Zagrovic, B. (2020) ‘Frameshifting preserves key physicochemical properties of proteins’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1911203117.

Beadle, G. W. and Tatum, E. L. (1941) ‘Genetic Control of Biochemical Reactions in Neurospora’, *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.27.11.499.

Beck, A. and Liu, H. (2019) ‘Macro- and Micro-Heterogeneity of Natural and Recombinant IgG Antibodies’, *Antibodies*. doi: 10.3390/antib8010018.

Bern, M. *et al.* (2018) ‘Parsimonious Charge Deconvolution for Native Mass Spectrometry’, *Journal of Proteome Research*. doi: 10.1021/acs.jproteome.7b00839.

Blakeley, P. *et al.* (2010) ‘Investigating protein isoforms via proteomics: A feasibility study’, *Proteomics*. doi: 10.1002/pmic.200900445.

Bondarenko, P. V. *et al.* (2009) ‘Mass Measurement and Top-Down HPLC/MS Analysis of Intact Monoclonal Antibodies on a Hybrid Linear Quadrupole Ion Trap-Orbitrap Mass Spectrometer’, *Journal of the American Society for Mass Spectrometry*. doi: 10.1016/j.jasms.2009.03.020.

Boyne, M. T. *et al.* (2009) ‘Tandem mass spectrometry with ultrahigh mass accuracy clarifies peptide identification by database retrieval’, *Journal of Proteome Research*. doi:

10.1021/pr800635m.

Bronsema, K. J., Bischoff, R. and Van de Merbel, N. C. (2012) ‘Internal standards in the quantitative determination of protein biopharmaceuticals using liquid chromatography coupled to mass spectrometry’, *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*. doi: 10.1016/j.jchromb.2012.02.021.

Bults, P. *et al.* (2016) ‘LC-MS/MS-Based Monitoring of in Vivo Protein Biotransformation: Quantitative Determination of Trastuzumab and Its Deamidation Products in Human Plasma’, *Analytical Chemistry*. doi: 10.1021/acs.analchem.5b04276.

Bush, D. R. *et al.* (2016) ‘High Resolution CZE-MS Quantitative Characterization of Intact Biopharmaceutical Proteins: Proteoforms of Interferon-β1’, *Analytical Chemistry*. doi: 10.1021/acs.analchem.5b03218.

Cassou, C. A. and Williams, E. R. (2014) ‘Desalting protein ions in native mass spectrometry using supercharging reagents’, *Analyst*. doi: 10.1039/c4an01085j.

Chang, A. C. Y. *et al.* (1999) ‘Alternative splicing regulates the production of ARD-1 endoribonuclease and NIPP-1, an inhibitor of protein phosphatase-1, as isoforms encoded by the same gene’, *Gene*. doi: 10.1016/S0378-1119(99)00435-7.

Chernushevich, I. V. and Thomson, B. A. (2004) ‘Collisional Cooling of Large Ions in Electrospray Mass Spectrometry’, *Analytical Chemistry*. doi: 10.1021/ac035406j.

Chung, S. *et al.* (2018) ‘Industrial bioprocessing perspectives on managing therapeutic protein charge variant profiles’, *Biotechnology and Bioengineering*. doi: 10.1002/bit.26587.

Cleary, S. P. *et al.* (2018) ‘Extracting Charge and Mass Information from Highly Congested Mass Spectra Using Fourier-Domain Harmonics’, *Journal of the American Society for Mass Spectrometry*. doi: 10.1007/s13361-018-2018-7.

Climente-González, H. *et al.* (2017) ‘The Functional Impact of Alternative Splicing in Cancer’, *Cell Reports*. doi: 10.1016/j.celrep.2017.08.012.

Cohen, P. T. W. (1988) ‘Two isoforms of protein phosphatase 1 may be produced from the same gene’, *FEBS Letters*. doi: 10.1016/0014-5793(88)80378-8.

Dale, D. C. (1998) ‘The discovery, development and clinical applications of granulocyte colony-stimulating factor.’, *Transactions of the American Clinical and Climatological Association*.

Delabrière, A. *et al.* (2017) ‘proFIA: A data preprocessing workflow for flow injection analysis coupled to high-resolution mass spectrometry’, *Bioinformatics*. doi: 10.1093/bioinformatics/btx458.

Donnelly, D. P. *et al.* (2019a) ‘Best practices and benchmarks for intact protein analysis for top-down mass spectrometry’, *Nature Methods*. Springer US, 16(7), pp. 587–594. doi: 10.1038/s41592-019-0457-0.

Donnelly, D. P. *et al.* (2019b) ‘Best practices and benchmarks for intact protein analysis for top-down mass spectrometry’, *Nature Methods*. Springer US, 16(7), pp. 587–594. doi: 10.1038/s41592-019-0457-0.

Duhamel, R. C. *et al.* (1979) ‘pH gradient elution of human IgG1, IgG2 and IgG4 from protein A-Sepharose’, *Journal of Immunological Methods*. doi: 10.1016/0022-1759(79)90133-9.

Ebersold, M. F. and Zydney, A. L. (2004) ‘Separation of protein charge variants by ultrafiltration’, *Biotechnology Progress*. doi: 10.1021/bp034264b.

Ecker, D. M., Jones, S. D. and Levine, H. L. (2015) ‘The therapeutic monoclonal antibody market’, *mAbs*. doi: 10.4161/19420862.2015.989042.

Egertson, J. D. *et al.* (2015) ‘Multiplexed peptide analysis using data-independent acquisition and Skyline’, *Nature Protocols*. doi: 10.1038/nprot.2015.055.

Faid, V. *et al.* (2018) ‘Middle-up analysis of monoclonal antibodies after combined IgdE and IdeS hinge proteolysis: Investigation of free sulfhydryls’, *Journal of Pharmaceutical and Biomedical Analysis*. doi: 10.1016/j.jpba.2017.11.046.

Fekete, S. *et al.* (2012) ‘Impact of mobile phase temperature on recovery and stability of monoclonal antibodies using recent reversed-phase stationary phases’, *Journal of Separation Science*. doi: 10.1002/jssc.201200297.

Fenn, J. B. *et al.* (1989) ‘Electrospray ionization for mass spectrometry of large biomolecules’, *Science*. doi: 10.1126/science.2675315.

Franzreb, M., Muller, E. and Vajda, J. (2014) ‘Cost estimation for protein a chromatography: An in silico approach to mab purification strategy’, *BioProcess International*.

Fraud, N. *et al.* (2009) ‘Hydrophobic-Interaction Membrane Chromatography for Large-Scale Purification of Biopharmaceuticals’, *BioProcess International*.

Füssl, F. *et al.* (2019) ‘Comprehensive characterisation of the heterogeneity of adalimumab via charge variant analysis hyphenated on-line to native high resolution Orbitrap mass spectrometry’, *mAbs*. doi: 10.1080/19420862.2018.1531664.

Gassmann, M. *et al.* (2009) ‘Quantifying Western blots: Pitfalls of densitometry’, *Electrophoresis*. doi: 10.1002/elps.200800720.

Gil, J. *et al.* (2019) ‘Clinical protein science in translational medicine targeting malignant melanoma’, *Cell Biology and Toxicology*. doi: 10.1007/s10565-019-09468-6.

Goetze, A. M. *et al.* (2011) ‘High-mannose glycans on the Fc region of therapeutic IgG antibodies increase serum clearance in humans’, *Glycobiology*. doi: 10.1093/glycob/cwr027.

Gong, C. *et al.* (2014) ‘Development and validation of an LC-MS/MS assay for the quantitation of a PEGylated anti-CD28 domain antibody in human serum: Overcoming interference from antidrug antibodies and soluble target', *Bioanalysis*. doi: 10.4155/bio.14.181.

Griebenow, K. and Klibanov, A. M. (1996) 'On protein denaturation in aqueous-organic mixtures but not in pure organic solvents', *Journal of the American Chemical Society*. doi: 10.1021/ja961869d.

Guzman, N. A. *et al.* (1992) 'Effect of buffer constituents on the determination of therapeutic proteins by capillary electrophoresis', *Journal of Chromatography A*. doi: 10.1016/0021-9673(92)87124-Q.

Haberger, M. *et al.* (2016) 'Rapid characterization of biotherapeutic proteins by size-exclusion chromatography coupled to native mass spectrometry', *mAbs*. Taylor & Francis, 8(2), pp. 331–339. doi: 10.1080/19420862.2015.1122150.

Háda, V. *et al.* (2018) 'Recent advancements, challenges, and practical considerations in the mass spectrometry-based analytics of protein biotherapeutics: A viewpoint from the biosimilar industry', *Journal of Pharmaceutical and Biomedical Analysis*. doi: 10.1016/j.jpba.2018.08.024.

Hagman, C. *et al.* (2008) 'Absolute quantification of monoclonal antibodies in biofluids by liquid chromatography-tandem mass spectrometry', *Analytical Chemistry*. doi: 10.1021/ac702115b.

Harris, R. J. *et al.* (2001) 'Identification of multiple sources of charge heterogeneity in a recombinant antibody', *Journal of Chromatography B: Biomedical Sciences and Applications*. doi: 10.1016/S0378-4347(00)00548-X.

Haverland, N. A. *et al.* (2017) 'Defining Gas-Phase Fragmentation Propensities of Intact Proteins During Native Top-Down Mass Spectrometry', *Journal of the American Society for Mass Spectrometry*. doi: 10.1007/s13361-017-1635-x.

Heck, A. J. R. and Van Den Heuvel, R. H. H. (2004) 'Investigation of intact protein complexes by mass spectrometry', *Mass Spectrometry Reviews*. doi: 10.1002/mas.10081.

Herzog, R. *et al.* (2020) 'Improved Alignment and Quantification of Protein Signals in Two-Dimensional Western Blotting', *Journal of Proteome Research*. doi: 10.1021/acs.jproteome.0c00061.

Hillenkamp, F. and Karas, M. (1990) 'Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization', *Methods in Enzymology*. doi: 10.1016/0076-6879(90)93420-P.

Hirn, M. *et al.* (1983) 'Molecular heterogeneity and structural evolution during cerebellar ontogeny detected by monoclonal antibody of the mouse cell surface antigen BSP-2', *Brain Research*. doi: 10.1016/0006-8993(83)91337-9.

Hoffmann, E. de (1996) 'Tandem Mass Spectrometry: a Primer', *Journal of Mass Spectrometry*.

Holt, M. V., Wang, T. and Young, N. L. (2019) ‘High-Throughput Quantitative Top-Down Proteomics: Histone H4’, *Journal of the American Society for Mass Spectrometry*. doi: 10.1007/s13361-019-02350-z.

Holzmann, J. *et al.* (2013) ‘Top-down MS for rapid methionine oxidation site assignment in filgrastim’, *Analytical and Bioanalytical Chemistry*. doi: 10.1007/s00216-013-7138-0.

Iavarone, A. T., Jurchen, J. C. and Williams, E. R. (2000) ‘Effects of solvent on the maximum charge state and charge state distribution of protein ions produced by electrospray ionization’, *Journal of the American Society for Mass Spectrometry*. doi: 10.1016/S1044-0305(00)00169-0.

Jamrichová, D. *et al.* (2017) ‘How to approach heterogeneous protein expression for biotechnological use: An overview’, *Nova Biotechnologica et Chimica*. doi: 10.1515/nbec-2017-0001.

Jayapal, K. P. *et al.* (2007) ‘Recombinant protein therapeutics from CHO Cells - 20 years and counting’, *Chemical Engineering Progress*.

Jenkins, R. *et al.* (2015) ‘Recommendations for Validation of LC-MS/MS Bioanalytical Methods for Protein Biotherapeutics’, *AAPS Journal*. doi: 10.1208/s12248-014-9685-5.

Jeong, K. *et al.* (2020) ‘FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics’, *Cell Systems*. doi: 10.1016/j.cels.2020.01.003.

Jungblut, P. R. *et al.* (2008) ‘The speciation of the proteome’, *Chemistry Central Journal*. doi: 10.1186/1752-153X-2-16.

Kafader, J. O. *et al.* (2020) ‘Multiplexed mass spectrometry of individual ions improves measurement of proteoforms and their complexes’, *Nature Methods*. doi: 10.1038/s41592-020-0764-5.

Kalli, A. *et al.* (2013) ‘Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: Focus on LTQ-orbitrap mass analyzers’, *Journal of Proteome Research*. doi: 10.1021/pr3011588.

Kamiie, J. *et al.* (2008) ‘Quantitative atlas of membrane transporter proteins: Development and application of a highly sensitive simultaneous LC/MS/MS method combined with novel in-silico peptide selection criteria’, *Pharmaceutical Research*. doi: 10.1007/s11095-008-9532-4.

Kayser, V. *et al.* (2011) ‘Glycosylation influences on the aggregation propensity of therapeutic monoclonal antibodies’, *Biotechnology Journal*. doi: 10.1002/biot.201000091.

Kelleher, N. L. (2012) ‘A cell-based approach to the human proteome project’, *Journal of the American Society for Mass Spectrometry*. doi: 10.1007/s13361-012-0469-9.

Kellie, J. F. *et al.* (2020) ‘Intact Protein Mass Spectrometry for Therapeutic Protein Quantitation, Pharmacokinetics, and Biotransformation in Preclinical and Clinical Studies: An Industry Perspective’, *Journal of the American Society for Mass Spectrometry*. doi:

10.1021/jasms.0c00270.

Kiontke, A. *et al.* (2016) ‘Electrospray ionization efficiency is dependent on different molecular descriptors with respect to solvent pH and instrumental configuration’, *PLoS ONE*. doi: 10.1371/journal.pone.0167502.

Kopp, J. *et al.* (2020) ‘Development of a generic reversed-phase liquid chromatography method for protein quantification using analytical quality-by-design principles’, *Journal of Pharmaceutical and Biomedical Analysis*. doi: 10.1016/j.jpba.2020.113412.

Labowsky, M., Whitehouse, C. and Fenn, J. B. (1993) ‘Three-dimensional deconvolution of multiply charged spectra’, *Rapid Communications in Mass Spectrometry*. doi: 10.1002/rcm.1290070117.

Lalley, P. A. and Shows, T. B. (1974) ‘Lysosomal and microsomal glucuronidase: Genetic variant alters electrophoretic mobility of both hydrolases’, *Science*. doi: 10.1126/science.185.4149.442.

Leal, M. T. A. *et al.* (2013) ‘Immunogenicity of recombinant proteins consisting of Plasmodium vivax circumsporozoite protein allelic variant-derived epitopes fused with Salmonella enterica serovar typhimurium flagellin’, *Clinical and Vaccine Immunology*. doi: 10.1128/CVI.00312-13.

Leblanc, Y. *et al.* (2017) ‘Charge variants characterization of a monoclonal antibody by ion exchange chromatography coupled on-line to native mass spectrometry: Case study after a long-term storage at +5 °C’, *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*. Elsevier B.V., 1048, pp. 130–139. doi: 10.1016/j.jchromb.2017.02.017.

Lee, J. S. *et al.* (2012) ‘Current state and perspectives on erythropoietin production’, *Applied Microbiology and Biotechnology*. doi: 10.1007/s00253-012-4291-x.

Lewis, D. A. *et al.* (1994) ‘Characterization of Humanized Anti-TAC, an Antibody Directed Against the Interleukin 2 Receptor, Using Electrospray Ionization Mass Spectrometry by Direct Infusion, LC/MS, and MS/MS’, *Analytical Chemistry*. doi: 10.1021/ac00077a003.

Liu, M., Watson, L. T. and Zhang, L. (2015) ‘HMMvar-func: A new method for predicting the functional outcome of genetic variants’, *BMC Bioinformatics*. doi: 10.1186/s12859-015-0781-z.

Lössl, P., Snijder, J. and Heck, A. J. R. (2014) ‘Boundaries of mass resolution in native mass spectrometry’, *Journal of the American Society for Mass Spectrometry*. doi: 10.1007/s13361-014-0874-3.

Lu, J. *et al.* (2015) ‘Improved Peak Detection and Deconvolution of Native Electrospray Mass Spectra from Large Protein Complexes’, *Journal of the American Society for Mass Spectrometry*, 26(12), pp. 2141–2151. doi: 10.1007/s13361-015-1235-6.

Mann, M. and Jensen, O. N. (2003) ‘Proteomic analysis of post-translational modifications', *Nature Biotechnology*. doi: 10.1038/nbt0303-255.

Marini, J. C. *et al.* (2013) 'Development and validation of ligand- binding assays to support the bioanalysis of therapeutic', in *Bioanalysis of Biotherapeutics*. doi: 10.4155/EBO.13.345.

Marty, M. T. (2019) 'Eliminating Artifacts in Electrospray Deconvolution with a SoftMax Function', *Journal of the American Society for Mass Spectrometry*. doi: 10.1007/s13361-019-02286-4.

Marty, M. T. (2020) 'A Universal Score for Deconvolution of Intact Protein and Native Electrospray Mass Spectra', *Analytical Chemistry*. doi: 10.1021/acs.analchem.9b05272.

Mathé, C. *et al.* (2013) 'Structural determinants for protein adsorption/non-adsorption to silica surface', *PLoS ONE*. doi: 10.1371/journal.pone.0081346.

Mayer, A. P. and Hottenstein, C. S. (2016) 'Ligand-Binding Assay Development: What Do You Want to Measure Versus What You Are Measuring?', *AAPS Journal*. doi: 10.1208/s12248-015-9855-0.

McKay, A. R. *et al.* (2006) 'Mass measurements of increased accuracy resolve heterogeneous populations of intact ribosomes', *Journal of the American Chemical Society*. doi: 10.1021/ja061468q.

Miladinović, S. M. *et al.* (2012) 'In-spray supercharging of peptides and proteins in electrospray ionization mass spectrometry', *Analytical Chemistry*. doi: 10.1021/ac300845n.

Miles, A. P. and Saul, A. (2005) 'Quantifying recombinant proteins and their degradation products using SDS-PAGE and scanning laser densitometry.', *Methods in molecular biology (Clifton, N.J.)*. doi: 10.1385/1-59259-922-2:349.

Misek, D. E. *et al.* (2005) 'A wide range of protein isoforms in serum and plasma uncovered by a quantitative intact protein analysis system', *Proteomics*. doi: 10.1002/pmic.200500103.

Nanita, S. C. (2013) 'Quantitative mass spectrometry independence from matrix effects and detector saturation achieved by flow injection analysis with real-time infinite dilution', *Analytical Chemistry*. doi: 10.1021/ac402567w.

Nanita, S. C. and Kaldon, L. G. (2016) 'Emerging flow injection mass spectrometry methods for high-throughput quantitative analysis', *Analytical and Bioanalytical Chemistry*. doi: 10.1007/s00216-015-9193-1.

Neubert, H. *et al.* (2018) '2018 White Paper on Recent Issues in Bioanalysis: Focus on immunogenicity assays by hybrid LBA/LCMS and regulatory feedback Part 2 - PK, PD & ADA assays by hybrid LBA/LCMS & regulatory agencies' inputs on bioanalysis, biomarkers and immunogenicity', in *Bioanalysis*. doi: 10.4155/bio-2018-0285.

Nielsen, M. L., Savitski, M. M. and Zubarev, R. A. (2006) 'Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics', *Molecular and Cellular Proteomics*. doi: 10.1074/mcp.M600248-MCP200.

Nossal, G. J. V. (1980) 'Human insulin through recombinant DNA technology', *Medical Journal of Australia*. doi: 10.5694/j.1326-5377.1980.tb76995.x.

Pandeswari, P. B. and Sabareesh, V. (2019) 'Middle-down approach: a choice to sequence and characterize proteins/proteomes by mass spectrometry', *RSC Advances*. doi: 10.1039/C8RA07200K.

Parekh, R. B. *et al.* (1989) 'Cell-Type-Specific and Site-Specific N-Glycosylation of Type I and Type II Human Tissue Plasminogen Activator', *Biochemistry*. doi: 10.1021/bi00445a021.

Pawlowski, J. W. *et al.* (2018) 'Influence of glycan modification on IgG1 biochemical and biophysical properties', *Journal of Pharmaceutical and Biomedical Analysis*. doi: 10.1016/j.jpba.2017.12.061.

Peng, Y. *et al.* (2014) 'Top-down mass spectrometry of cardiac myofilament proteins in health and disease', *Proteomics - Clinical Applications*. doi: 10.1002/prca.201400043.

Peris-Díaz, M. D. *et al.* (2020) 'Mass Spectrometry-Based Structural Analysis of Cysteine-Rich Metal-Binding Sites in Proteins with MetaOdysseus R Software', *Journal of Proteome Research*. doi: 10.1021/acs.jproteome.0c00651.

Peters, I., Metwally, H. and Konermann, L. (2019) 'Mechanism of Electrospray Supercharging for Unfolded Proteins: Solvent-Mediated Stabilization of Protonated Sites during Chain Ejection', *Analytical Chemistry*. doi: 10.1021/acs.analchem.9b01470.

Pipes, G. D. *et al.* (2010) 'Middle-down fragmentation for the identification and quantitation of site-specific methionine oxidation in an IgG1 molecule', *Journal of Pharmaceutical Sciences*. doi: 10.1002/jps.22158.

Pohl, K. *et al.* (2020) *A New Level of Compliant-Ready Intact Biotherapeutic Protein Quantification using Reconstructed Masses*.

Qiu, X. *et al.* (2018) 'Quantitation of intact monoclonal antibody in biological samples: Comparison of different data processing strategies', *Bioanalysis*. Future Medicine Ltd., 10(13), pp. 1055–1067. doi: 10.4155/bio-2018-0016.

Raju, T. S. (2008) 'Terminal sugars of Fc glycans influence antibody effector functions of IgGs', *Current Opinion in Immunology*. doi: 10.1016/j.coi.2008.06.007.

Regl, C. *et al.* (2019) 'Dilute-and-shoot analysis of therapeutic monoclonal antibody variants in fermentation broth: a method capability study', *mAbs*. doi: 10.1080/19420862.2018.1563034.

Rizvi, S. S. H. (2010) *Separation, Extraction and Concentration Processes in the Food, Beverage and Nutraceutical Industries, Separation, Extraction and Concentration*

Processes in the Food, Beverage and Nutraceutical Industries. doi: 10.1533/9780857090751.

Roberts, N. B., Green, B. N. and Morris, M. (1997) ‘Potential of electrospray mass spectrometry for quantifying glycohemoglobin’, *Clinical Chemistry*. doi: 10.1093/clinchem/43.5.771.

Rochat, B. (2019) ‘Quantitative and Qualitative LC-High-Resolution MS: The Technological and Biological Reasons for a Shift of Paradigm’, in *Recent Advances in Analytical Chemistry*. doi: 10.5772/intechopen.81285.

Roman, G. T. and Murphy, J. P. (2017) ‘Improving sensitivity and linear dynamic range of intact protein analysis using a robust and easy to use microfluidic device’, *Analyst*. doi: 10.1039/c6an02518h.

Ronsein, G. E. *et al.* (2015) ‘Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics’, *Journal of Proteomics*. doi: 10.1016/j.jprot.2014.10.017.

Rosano, G. L. and Ceccarelli, E. A. (2014) ‘Recombinant protein expression in Escherichia coli: Advances and challenges’, *Frontiers in Microbiology*. doi: 10.3389/fmicb.2014.00172.

Rosenlöcher, J. *et al.* (2016) ‘Recombinant glycoproteins: The impact of cell lines and culture conditions on the generation of protein species’, *Journal of Proteomics*. doi: 10.1016/j.jprot.2015.08.011.

Ruan, Q. *et al.* (2011) ‘Strategy and its implications of protein bioanalysis utilizing high-resolution mass spectrometric detection of intact protein’, *Analytical Chemistry*. doi: 10.1021/ac201540t.

Sawyer, W. S. *et al.* (2020) ‘High-throughput antibody screening from complex matrices using intact protein electrospray mass spectrometry’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1917383117.

Schaffer, L. V. *et al.* (2019) ‘Identification and Quantification of Proteoforms by Mass Spectrometry’, *Proteomics*. doi: 10.1002/pmic.201800361.

Schlüter, H. *et al.* (2009) ‘Finding one’s way in proteomics: A protein species nomenclature’, *Chemistry Central Journal*. doi: 10.1186/1752-153X-3-11.

Sha, S. *et al.* (2016) ‘N-Glycosylation Design and Control of Therapeutic Monoclonal Antibodies’, *Trends in Biotechnology*. doi: 10.1016/j.tibtech.2016.02.013.

Smith, L. M. and Kelleher, N. L. (2013) ‘Proteoform: A single term describing protein complexity’, *Nature Methods*. doi: 10.1038/nmeth.2369.

Song, E. *et al.* (2017) ‘Targeted proteomic assays for quantitation of proteins identified by proteogenomic analysis of ovarian cancer’, *Scientific Data*. doi: 10.1038/sdata.2017.91.

Song, K. *et al.* (2020) ‘Glycosylation Heterogeneity of Hyperglycosylated Recombinant

Human Interferon-β (rhIFN-β)’, *ACS Omega*. doi: 10.1021/acsomega.9b04385.

Srzentić, K. *et al.* (2020) ‘Interlaboratory Study for Characterizing Monoclonal Antibodies by Top-Down and Middle-Down Mass Spectrometry’, *Journal of the American Society for Mass Spectrometry*. doi: 10.1021/jasms.0c00036.

Steffen, P. *et al.* (2016) ‘Protein species as diagnostic markers’, *Journal of Proteomics*. doi: 10.1016/j.jprot.2015.12.015.

Stubenrauch, K., Wessels, U. and Lenz, H. (2009) ‘Evaluation of an immunoassay for human-specific quantitation of therapeutic antibodies in serum samples from non-human primates’, *Journal of Pharmaceutical and Biomedical Analysis*. doi: 10.1016/j.jpba.2009.01.030.

Sugihara, T. *et al.* (2018) ‘Isolation of recombinant human antithrombin isoforms by Cellufine Sulfate affinity chromatography’, *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*. doi: 10.1016/j.jchromb.2018.07.001.

Swinney, K. and Bornhop, D. J. (2000) ‘Detection in capillary electrophoresis’, in *Electrophoresis*. doi: 10.1002/(SICI)1522-2683(20000401)21:7<1239::AID-ELPS1239>3.0.CO;2-6.

Thomas, P. M. *et al.* (2014) ‘ProSight Lite: Graphical software to analyze top-down mass spectrometry data’, *Proteomics*, 15(7), pp. 1235–1238. doi: 10.1002/pmic.201400313.

Tiambeng, T. N. *et al.* (2019) ‘Analysis of cardiac troponin proteoforms by top-down mass spectrometry’, in *Methods in Enzymology*. doi: 10.1016/bs.mie.2019.07.029.

Tu, C. *et al.* (2014) ‘Systematic assessment of survey scan and MS2-based abundance strategies for label-free quantitative proteomics using high-resolution MS data’, *Journal of Proteome Research*. doi: 10.1021/pr401206m.

Tucholski, T. *et al.* (2020) ‘Distinct hypertrophic cardiomyopathy genotypes result in convergent sarcomeric proteoform profiles revealed by top-down proteomics’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.2006764117.

Uetrecht, C. *et al.* (2019) ‘Native mass spectrometry provides sufficient ion flux for XFEL single-particle imaging’, *Journal of Synchrotron Radiation*. doi: 10.1107/S1600577519002686.

Urquhart, L. (2021) ‘Top companies and drugs by sales in 2020’, *Nature Reviews Drug Discovery*. doi: 10.1038/d41573-021-00050-6.

Vailaya, A. and Horváth, C. (1998) ‘Retention in reversed-phase chromatography: Partition or adsorption?’, *Journal of Chromatography A*. doi: 10.1016/S0021-9673(98)00727-4.

Walsh, G. and Jefferis, R. (2006) ‘Post-translational modifications in the context of therapeutic proteins’, *Nature Biotechnology*. doi: 10.1038/nbt1252.

Wang, E. H. *et al.* (2017) 'Investigation of Ion Transmission Effects on Intact Protein Quantification in a Triple Quadrupole Mass Spectrometer', *Journal of the American Society for Mass Spectrometry*. doi: 10.1007/s13361-017-1696-x.

Welinder, B. S., Sørensen, H. H. and Hansen, B. (1987) 'Reversed-phase high-performance liquid chromatography of human growth hormone', *Journal of Chromatography A*. doi: 10.1016/S0021-9673(01)96517-3.

Whitehouse, C. M. *et al.* (1985) 'Electrospray Interface for Liquid Chromatographs and Mass Spectrometers', *Analytical Chemistry*. doi: 10.1021/ac00280a023.

Wilkins, M. R. *et al.* (1996) 'From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis', *Bio/Technology*. doi: 10.1038/nbt0196-61.

Wohlschlager, T. *et al.* (2018) 'Native mass spectrometry combined with enzymatic dissection unravels glycoform heterogeneity of biopharmaceuticals', *Nature Communications*, 9(1), pp. 1–9. doi: 10.1038/s41467-018-04061-7.

Wu, Z. *et al.* (2020) 'MASH Explorer: A Universal Software Environment for Top-Down Proteomics', *Journal of Proteome Research*. doi: 10.1021/acs.jproteome.0c00469.

Yang, Y. *et al.* (2013) 'Analyzing protein micro-heterogeneity in chicken ovalbumin by high-resolution native mass spectrometry exposes qualitatively and semi-quantitatively 59 proteoforms', *Analytical Chemistry*. doi: 10.1021/ac403057y.

Young, N. L. *et al.* (2010) 'Collective mass spectrometry approaches reveal broad and combinatorial modification of high mobility group protein a1a', *Journal of the American Society for Mass Spectrometry*. doi: 10.1016/j.jasms.2010.01.020.

Yu, L. *et al.* (2020) 'Analysis of Molecular Heterogeneity in Therapeutic IFNα2b from Different Manufacturers by LC/Q-TOF', *Molecules*. doi: 10.3390/molecules25173965.

Zhang, L. *et al.* (2018) 'Top-down LC-MS quantitation of intact denatured and native monoclonal antibodies in biological samples', *Bioanalysis*, 10(13), pp. 1039–1054. doi: 10.4155/bio-2017-0282.

Zhang, L., Moo-Young, M. and Chou, C. P. (2010) 'Effect of aberrant disulfide bond formation on protein conformation and molecular property of recombinant therapeutics', in *Pure and Applied Chemistry*. doi: 10.1351/PAC-CON-09-01-06.

10 Appendix

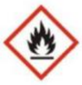
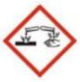
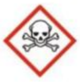
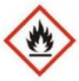
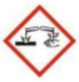
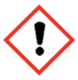
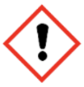
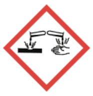
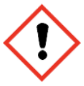
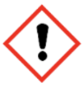
10.1 Statement of contribution by others

Certain set of experiments presented in this thesis were done in in collaboration with Analytics for Biologics (A4B) consortium members or colleagues.

- Establishing the “Sample displacement batch chromatography” (SDBC) method for Ovalbumin and Adalimumab was predefined task specifically attributed to Siti Nurul Hidayah, a fellow consortium member of the A4B project network and colleague from Prof. Dr. Schluter’s lab, University Medical Center Hamburg-Eppendorf, Germany.
- Developing the bioinformatics framework of FLASHDeconv algorithm for deconvolution and relative quantification of intact proteins was predefined task specifically attributed to by Jihyung Kim, a fellow consortium member of the A4B project network from Prof. Dr. Oliver Kohlbacher, Eberhard Karls University of Tübingen, Germany.
- Lab scale production of the Adalimumab sample used in the study was predefined task specifically attributed to Daniel Komuzcki, a fellow consortium member of the A4B project network from Prof. Dr. Alois Jungbauer, University of Natural Resources and Life Sciences Vienna, Austria.
- The glycan list for Ovalbumin used for identity of proteoforms was generated by Dr. Yudong Guan, a fellow lab colleague from Prof. Dr. Schluter’s lab, University Medical Center Hamburg-Eppendorf, Germany.

10.2 Risk and safety statements

Pictograms of potentially hazardous chemicals used throughout this study, based on the Globally Harmonized System of Classification and Labelling of Chemicals (GHS), GHS hazard and precautionary statements.

Chemicals	GHS symbol	GHS hazard statement	GHS precautionary statements
Formic acid (FA)	GHS02  GHS05  GHS06 	H226 H302 H314 H331 EUH071	P210 P280 P301 + P330 + P331 P304+P340 P305+P351+P338 P308 + P310
Acetic acid	GHS02  GHS05 	H226 H314	P210, P280 P301+P330+P331 P305+P351+P338 P308+P310
Sodium acetate		H319	P264, P280
2-(N-morpholino) ethanesulfonic acid (MES)		H315 H319	P264, P280, P321 P302+P352 P305+P351+P338 P332+P313 P337+P313 P362+P364
Sulfolane	 	H302 H315 H319 H335 H402	P261, P264, P270, P271, P273, P280 P330, P405, P501, P332+P313, P337+P313, P362+P364, P403+P233
Ammonium acetate		H315 H319 H335 H402	P261, P264, P271, P273, P280, P312, P302+P352 P305+P351+P338 P332+P313 P337+P313

Acetonitrile (ACN)	GHS02 GHS07	H226 H314	P210 P280 P301+P330+P331 P305+P351+P338 P308+P310
Ammonium bicarbonate (ABC)	GHS07	H302	P301+P312 P330
Dithiothreitol (DTT)	GHS07	H302 H312 H332	P261 P280 P301+P310 P304+P340 P361 P501
Iodoacetamide (IAA)	GHS06 GHS08	H301 H317 H334	P261 P280 P301 + P310 P342 + P311
Trypsin	GHS07 GHS08	H315 H319 H334 H335	P302 + P352 P304+P340 P305+P351 P342+P311

10.3 CMR list

No cancerogenic, mutagenic or reprotoxic substances (CMR substances) from the GHS category 1A or 1B were used in the work.

11 Acknowledgement

First and foremost, I would express my deep gratitude to Prof. Dr. Hartmut Schlüter, for giving me an opportunity to carry out my doctoral research project at University Medical Centre Hamburg-Eppendorf. Hartmut, your guidance, eye to details, criticism, has improved my critical thinking and helped me expand the horizons in field of proteomics. It was insightful journey to explore this topic as a part of the A4B consortium and I thank you very much for it. Thanks to the funding organisation- European Commission for supporting this work through Horizon 2020, Marie Sklodowska-Curie Action ITN 2017.

The A4B consortium members – consisting of 14 other the fellow PhD students along with their mentors made a great team. Sincere thanks to consortium members for the help & appreciation for taking forwards the projects successfully amid the pandemic. Special thanks to our prior project manager and colleague Dr. Heikaus, current project manager and colleague Dennis, soft skill trainer- Dr. Schütte, and all the organisers of the A4B training program who helped a great deal throughout the doctoral studies. Speaking of great team, I'm grateful for my A4B colleagues and friends -Siti Nurul Hidayah & Jihyung Kim, without whom the project would not have been the same. Thanks a ton for long scientific meetings and many other things you both have helped me through. Also, cannot forget great help from our main collaborators- Dr.Kyowon Jeong, Prof. Dr. Oliver Kohlbacher, and Dr. Marty for giving me the much needed bioinformatics related insights for completion of this project.

Special thanks to my colleagues turned friends from mass spectrometric proteomics group,- Chris, Beni, Hannah, Lorena, Min, Manu, Manka, Sönke, Yudong for accommodating me in the group and being there for all the help required at work. Special mention to my constants- Chris and Hannah,- pandemic felt much bearable with company of you two. Definitely cannot forget my master student -Alice and Bachelor students-Aqila and Ali who contributed to experiments and were a great company.

A huge shoutout to Chan, Luis, Hanna and friends in Clinical Chemistry department. You have been a solid support throughout and I'm so thankful for having you all around.

I cannot thank enough my housemate – Kai, my guardian here in Hamburg. In last 3 years, you have really been the Robin to my Batman, and at times Leonard to my Sheldon! I will be forever grateful for your kind-hearted nature, your father-like encouragement, the smallest to biggest help amid pandemic, surgeries and zillion other things. Vielen Vielen Dank Kai!

And finally loads of gratitude to my rock-solid loving family! You all have been biggest support from 5000 miles apart. I have seldom expressed this, but Mumma & Pappa, my dearest little brother, I cannot thank you all enough, for putting up with my moods, pulling me up through my lows and for the trust you have in me. Lots of love to you guys. Also, huge thanks to my extended fam- uncle, grandparents who are always there through thick and thin. I’m highly indebted to have such loving family and owe this journey and success to you all. Thank you!

Declaration

I hereby declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research. I have not used other than the acknowledged resources and aids. The submitted written version corresponds to the version on the electronic storage medium.

Hiermit versichere ich an Eides statt, die vorliegende Dissertation selbst verfasst und keine anderen als die angegebenen Hilfsmittel benutzt zu haben. Die eingereichte schriftliche Fassung entspricht der auf dem elektronischen Speichermedium. Ich versichere, dass diese Dissertation nicht in einem früheren Promotionsverfahren eingereicht wurde

Hamburg, 11th June 2021

Place and date Signature