EXPOSING STRUCTURAL FEATURES OF PROTEIN ASSEMBLIES USING MASS SPECTROMETRY SEM TAMARA 2019

EXPOSING STRUCTURAL FEATURES OF PROTEIN ASSEMBLIES USING MASS SPECTROMETRY

SEM TAMARA

EXPOSING STRUCTURAL FEATURES OF PROTEIN ASSEMBLIES USING MASS SPECTROMETRY

SEM TAMARA

EXPOSING STRUCTURAL FEATURES OF PROTEIN ASSEMBLIES USING MASS SPECTROMETRY

Structurele kenmerken van eiwit complexen bestuderen met behulp van massaspectrometrie

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 27 november 2019 des middags te 12.45 uur

door

Sem Tamara

geboren op 3 november 1992
te Kholmsk, Rusland

ISBN: 978-90-393-7209-8

Cover: Ksenia Vist, www.ksvist.com

Lay-out: Sem Tamara

Print by: Gildeprint - Enschede, www.gildeprint.com

Thermo Fisher Scientific financially supported the printing of this thesis.

Promotors: Prof.dr. A.J.R. Heck
Prof.dr. A.A. Makarov

Co-promotor: Dr. R.A. Scheltema

TABLE OF CONTENTS

CHAPTER

1

TANDEM MASS SPECTROMETRY FOR IDENTIFICATION OF PROTEOFORMS AND STRUCTURAL ANALYSIS OF MOLECULAR MACHINES

INTRODUCTION

FROM GENES TOWARDS MULTIPROTEOFORM ASSEMBLY

The primary products of gene expression, proteins, mediate all the essential processes observed in living cells. The path from a gene toward a functioning protein is not always as straightforward as translating a polynucleotide sequence into a polypeptide chain but is a much more complex multi-staged process. Molecular perturbations can occur at many points in this process, producing a completely different protein than would be expected from the polynucleotide sequence alone. This process can be influenced by, *e.g.* alternative splicing or errors in transcription. Even subtle changes can affect aspects of the final protein structure and, therefore, function. Hence, gene-protein connectivity is only to a limited extent defined by to the gene sequence, as even formation of distinct sequence variants, *i.e.* isoforms, during alternative splicing can result in altered function and interactions[1]. The complexity expansion does not stop at the level of sequence variants or – in more conventional terms – isoforms[2]. Each newly synthesized polypeptide chain must mature through one or several stages of molecular transformations, including post-translational processing events and modifications. The result of such maturation is a proteoform, which represents a product of a specific gene attributed by distinct amino acid sequence and a distinct set of post-translational modifications (PTMs)[3,4]. Endogenously, cells are often operated by protein machines that represent assemblies of non-covalently bound proteoforms, nowadays termed multiproteoform complexes (MPCs)[3,5], which enable functionality far exceeding that of the sum of the individual components. Altogether, the expansion of variation on the path from gene toward protein provides an extraordinarily heterogeneous and dynamic landscape of proteoforms and their complexes.

Proteoforms

The complexity of proteoforms is immense as it is a result of all combinatorial variations at the level of DNA, RNA, and even proteins themselves[6] (Figure 1A). For example, human DNA contains ~20,000 genes, which are substantially affected by coding single nucleotide polymorphisms (cSNPs) and mutations. At the level of RNA, the major variability is attributed to alternative splicing events with ~40,000 splice isoforms encountered in the Human SwissProt Database (version 2019). Additionally, transcriptional and translational mistakes are not uncommon, although most of them are corrected by various repair mechanisms or do not result in an alteration of the final protein sequence[7]. Finally, multiple post-translational protein processing mechanisms significantly boost the number of potential proteoforms introducing > 50,000 post-translationally modified sites. Altogether these perturbations lead to an expansion of the proteoform space into several million, especially when considering distinct exacerbating factors like somatic recombination detected in certain cell types[8]. While the identification of a protein can be achieved by simply determining a portion of its amino acid sequence, the characterization of proteoforms requires determination of the primary sequence and processing steps, including all applied post-translational modifications.

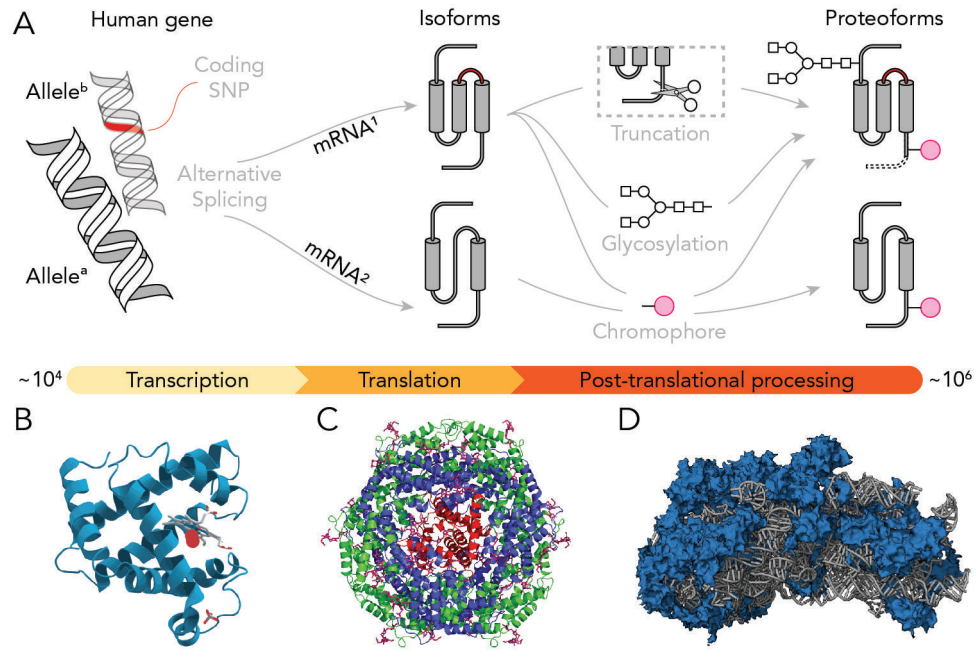


Figure 1 | (A) The chain of events that may occur during the transition from a gene towards a mature proteoform. Non-covalent complexes of (B) myoglobin with non-covalently attached heme group, (C) light-harvesting B-phycoerythrin comprised of α, β, and γ subunits, and (D) 40S human ribosomal particle intertwined with RNA.

Multiproteoform assemblies

The expansion of functionality introduced by MPCs comprised of the sets of proteoforms is further extended through a manifold of intertwined structural features: the sequence together with various post-translational modifications (PTMs) form the primary structure, conformation of the protein backbone defines its secondary and tertiary structure, and, finally, non-covalently associated proteoforms encompass the quaternary structure. While each of the primary, secondary, tertiary, and quaternary structural features contribute to the final protein function, proteins are additionally affected by substoichiometric subunits and transient interactors. Further functional diversification for the multiproteoform assemblies is achieved through binding of co-factors such as metal ions[9], ligands[10], RNA[11,12] or DNA[13], and lipids[14] (Figure 1B-D).

To understand the phenotypic behavior observed in living cells, each protein of interest must, therefore, be studied in detail at all levels of its molecular complexity. Amongst the numerous available methods, mass spectrometry (MS) has recently expanded from being a method of choice for high-throughput protein identification and quantification towards a multi-functional pillar for the interrogation of all aspects of both the primary structure as well as the higher-order protein structure. Combined, this positioned MS as a powerful supplement to other biochemical and biophysical tools.

METHODS TO STUDY PROTEIN STRUCTURE

Many techniques are available to study various aspects of multiproteoform assemblies. While some of them provide a high-resolution structural snapshot of a protein complex, others produce large sets of disintegrated low-resolution information that can be used to infer protein sequence and its structural features. The first category is primarily formed by conventional high-resolution structural biology techniques like NMR[15,16], x-ray crystallography[17], and cryo-EM[18–20]. The second category includes low-resolution methods like circular dichroism[21,22] and numerous mass spectrometry (MS) approaches that interrogate protein structure at the level of peptides, intact proteoforms, or entire MPCs. All these methods have distinct features and principles of operation. For instance, they differ in the required amount and/or state of the analyte, *e.g.* solid for x-ray crystallography and cryo-EM, liquid/solid for MS and NMR, etc. Among all techniques, structural mass spectrometry as a set of different MS-based methods has emerged as a versatile and powerful technique for interrogation of protein structure and interactions both in solution as well as in the gas phase.

Investigating the primary structure of proteoforms

The primary structure of a proteoform is an essential piece of information as, together with the harbored PTMs, it significantly affects both protein conformation and protein function. To characterize a protein sequence, the identity and order of amino acids in this sequence should be determined. Typically, the sequence is described *de novo*, or it can be inferred from available genome databases[23]. Among protein sequencing methods, Edman degradation for a long time was the technique of choice, however, since the introduction of tandem mass spectrometry it has become nearly fully substituted although sometimes it is still used in combination[24]. The most basic MS-based method used to characterize the protein sequence is the so-called shotgun proteomics. Here, the protein is first digested, either chemically or enzymatically, producing a mixture of peptides that are then typically analyzed using liquid chromatography-tandem mass spectrometry (LC-MS/MS). This approach is fast allowing for high throughput that enables for *de novo* investigation of complex protein mixtures (*e.g.* cell lysates), although it suffers from sensitivity issues and requires complex computational solutions to decipher the data, often only delivering complete sequence for purified proteins.

Alternative techniques for the determination of the primary protein structure include the emerging sub-nanopore protein sequencing that has the potential to circumvent sensitivity issues by sequencing single protein molecules[25,26]. However, this is currently still in the early stages of development and has not yet been tested on complex protein mixtures. As for high-resolution structural techniques, although cryo-EM and x-ray diffraction can be utilized for the determination of the primary protein structure, these methods typically find it harder to uncover small PTMs and substoichiometric proteoforms. The advancing free-electron X-ray laser (XFEL) technology is another example of a promising single-molecule structural tool[27]. Notwithstanding the newly emerging tools and approaches, tandem MS remains so far the

golden standard for high-throughput proteoform characterization. A more detailed overview of its principles is provided in the next sections.

Investigating the higher-order structure of MPCs with low-resolution techniques

Proteoform composition, subunit stoichiometry, conformation, topology, and interaction sites, are all examples of higher-order structural features of protein assemblies. While high-resolution structural approaches like cryo-EM, x-ray, and NMR are considered the primary methods to study protein structure and provide near-atomic resolution, each of these techniques also has its limitations (Table 1). One of the drawbacks of high-resolution approaches is that they require significant structural averaging, prompting some transient features to be *averaged out*. As such, they neglect less prominent assembly variants, which often originate from heterogeneity at the proteoform level. In contrast, some emerging technologies, despite having lower resolution, provide a means to detect all the micro- and macro-heterogeneities in the composition and structure of MPCs. These new methods appear as prospective and complementary tools for the structural analysis of macromolecular assemblies.

In particular MS-based approaches proved to be complementary to various high-resolution methods as they enable to independently target distinctive structural features of MPCs, *e.g.* composition, stoichiometries, and interactions. Although mass spectrometry is not able to directly interrogate protein structure and does not always achieve complete coverage of its analytes, due to inherent limitations of mass measurements, it has the potential to characterize very fine structural details in a high-throughput or highly targeted manner. Moreover, the versatility of MS-based approaches allows researchers to interrogate protein systems that are challenging for most of the high-resolution methods, *e.g.* intrinsically disordered proteins and membrane proteins for x-ray crystallography, small proteins for cryo-EM, and highly heterogeneous MPCs for NMR (see table 1 for an overview of some of the strengths and weaknesses of the different approaches). The prominence of structural mass spectrometry will further increase in the era of emerging integrative MS approaches, which combine benefits of different MS-based techniques or integrate MS with orthogonal tools for the structural investigation of proteoforms and MPCs. Among the integrative MS approaches, some employ the high-throughput nature of MS-based proteomics to supplement the analysis of protein structures in solution, *e.g.* by performing surface labeling or chemical cross-linking in solution and analysis in the gas phase. At the same time, others allow researchers to investigate protein gas-phase structures, *e.g.* via the detection of differential ion mobility in a neutral gas[28,29], or the investigation of protein fragmentation patterns[30,31].

Ion mobility-mass spectrometry (IM-MS) is an established and powerful technique that allows for separation of proteins based on a combination of charge and shape, uncovering various structural features, *e.g.* composition and topology of protein assemblies. For most IM implementations, ions travel through a 'bath' of neutral gas molecules that create resistance, resulting in longer travel times for larger ions. The way ions are manipulated is slightly different for available technologies, for example ions can be dragged by a potential applied across the drift tube (DT) or

Table 1 | Overview of the common techniques for analysis of protein higher-order structure.

	NMR	Cryo-EM	X-ray Crystallography	Circular Dichroism	Tandem MS	Ion Mobility MS	HDX/Covalent Labelling MS	Cross-linking MS
Analyte phase	Liquid (or solid)	Solid	Solid	Liquid (gas; solid)	Gas	Gas	Liquid-Gas	Liquid-Gas
Pros	- Preserves analyte - Atomic resolution - Solution-phase analysis ensures native-like structures - Can be used to study conformational dynamics	- Nearly atomic resolution structures of large macromolecular assemblies	- Preserves analyte Direct structural information at atomic resolution	- Preserves analyte - Fast analysis of specific structural features	- Sensitive - Determination of stoichiometry - Identification of proteoforms and substoichiometric interactors - Reveals compositional and structural heterogeneity	- Sensitive - Analysis of mixtures - Reveals conformational heterogeneity - Determination of stoichiometry	- Sensitive - Labeling occurs in solution - Works with disordered proteins - Possible to analyze mixtures and intact proteins	- Cross-linking occurs in-solution - Highly throughput (proteomes) - Works with disordered proteins
Cons	- Purified and concentrated samples - Low upper-mass limit (~250 kDa)	- Purified and concentrated samples - Hampered by lower-mass limit (~50 kDa) - Destructive	- Purified and concentrated samples - Might be destructive - Many analytes appear difficult to crystallize	- Difficult to obtain quantitative data - Very low resolution	- Destructive - Low resolution - Transfer to the gas-phase may affect the conformation - Challenging data integration and structure inference	- Destructive - Transfer to the gas-phase may affect the conformation - Relies on molecular dynamics to infer structure from CCS	- Destructive - Low resolution - Back exchange (HDX only)	- Destructive - Low resolution - Low sensitivity - Only provides structural restraints
Works best for	Mixtures of small molecules, small proteins	Large purified macromolecular assemblies	Small to medium structured proteins	Elucidation of protein secondary structure	In-depth analysis of compositional and structural heterogeneity within protein assemblies	Analysis of structural conformers in the gas phase	Analysis of conformational dynamics in solution or gas phase	Large protein assemblies without available structures

manipulated using the traveling wave (TW) technology. While DT design allows researchers to obtain collisional cross-sections (CCS) of molecules directly, TW often requires a calibration curve for accurate determination of CCS but provides higher resolution. Over the last decade, IM-MS has proven valuable for obtaining snapshots of conformational landscapes[32,33], which is rather challenging to achieve with other techniques.

Another popular low-resolution technique for the analysis of higher-order protein structure is hydrogen-deuterium exchange (HDX). HDX exploits the exchange of labile hydrogens in the protein backbone with hydrogens in solution[34]. When the solution is comprised of D_2O, proteins undergo H-to-D (H/D) exchange. The technique exploits the behavior, in which exposed or disordered regions have a higher degree of H/D exchange compared to tightly folded regions or binding interfaces of non-covalently interacting proteins. Initially, HDX experiments were monitored with nuclear magnetic resonance (NMR). However, the field has since also coupled HDX with MS to benefit from higher throughput and unlimited mass range. HDX MS is useful for the analysis of protein conformation and binding interfaces, typically through digestion of proteins into peptides. Recently, the method has been extended for the structural analysis of distinctive MPC variants in their native state[35–37].

Chemical cross-linking (XL MS), a technique based on covalently linking specific functional groups in a protein or between proteins provides yet another means to obtain further structural information. Current advances in MS instrumentation extended XL MS from a niche method to one of the most promising approaches in structural mass spectrometry. In the past decade, the technique has been applied to map protein-protein interactions at the proteome level[38] as well as to the analysis of the structures of purified protein assemblies. This has provided highly complementary information to other structural techniques; especially for the analysis of challenging protein systems, *e.g.* membrane proteins and intrinsically disordered proteins.

A so far less explored avenue of MS-based structural analysis is represented by integrating multiple applications of tandem mass spectrometry. This approach, boosted by an array of complementary activation methods, uncovers various structural features of proteins by in-depth analysis of their dissociation products. Recently tandem MS was applied not only for the characterization of intact proteoforms but also for the investigation of entire protein assemblies and their distinctive variants[2,39]. The following sections will focus on the fundamentals of tandem mass spectrometry as applied to the structural analysis of proteoforms and MPCs.

ENABLING STRUCTURAL MASS SPECTROMETRY

The technique of mass spectrometry (MS) emerged more than a century ago[40,41]. It is mainly based on the principle of detecting the mass-to-charge (*m/z*) ratio of atoms or molecules in the gas phase through the manipulation of electric and magnetic fields. J.J. Thomson was the first who utilized these principles in his work on cathode rays and, later, the unit for *m/z* was coined in his memory (Thomson "Th")[42].

Initially, the technique was primarily used for isotope analysis of various small molecules in the fields of physics and chemistry, as the available ionization techniques were too harsh for biomolecules, like peptides or oligonucleotides. In the 1980s, however, the co-development of milder ionization techniques and more advanced mass analyzers opened up new avenues, especially for biomolecular MS, making it a popular technique for cell biologists and biochemists, with the latter being interested in its use for the structural analysis of biomacromolecules.

The introduction of fast atom bombardment (FAB)[43] in the early 1980s originally circumvented the need for the equally tedious and harsh chemical derivatization step required for older ionization methods like electron ionization (EI) and chemical ionization (CI). Although FAB already allowed for the ionization of peptides and small proteins[44], explosive growth in the utilization of mass spectrometry for structural analysis of proteoforms and MPCs occurred after the invention of the "true" soft ionization methods: electrospray ionization (ESI)[45] and matrix-assisted laser desorption/ionization[46]. Both techniques were developed in the 1980s and allowed researchers to transfer intact biomolecules into the gas phase without fragmentation, albeit by distinctively different mechanisms. While MALDI requires biomolecules to be immobilized in a crystalline matrix with concomitant ionization through short-duration laser pulses, ESI achieves ionization directly from the solution the analytes are dissolved in by producing small charged liquid droplets in a strong electrical field. Distinctively from ESI, that produces highly charged ions, MALDI primarily leads to the formation of singly charged ionized species. This ionization behavior sometimes facilitates mass detection for complex ion mixtures, although it largely limits dissociation of peptides and proteins, which often depends on the number of charges carried by the ion[47,48]. It also prevents direct coupling to pre-fractionation techniques, whereby aqueous solutions are used as the analyte carrier, *e.g.* liquid chromatography. Reassuringly, the ability of MALDI to produce protein ions directly from untreated tissues or cell samples defined its prominence in MS-based clinical diagnostic microbiology and imaging[49–51]. At the same time ESI, which was further improved by the development of nano-ESI (nESI)[52], requiring less sample input, became the ionization method of choice for LC-coupled mass spectrometry and tandem mass spectrometry analyses. Moreover, because of a very delicate ion formation mechanism, nano-ESI preserves most of the non-covalent interactions, enabling native mass spectrometry (native MS), which aims to analyze intact proteoforms as well as large MPCs in their native state[53].

Another essential component of any mass spectrometer is the mass analyzer, which must meet specific requirements for allowing the structural investigation of intact proteins and their complexes. In principle, mass analyzers achieve separation of ions based on their mass-to-charge ratio (*m/z*) by manipulation with electrical and/or magnetic fields. Over the years, clearly defined mechanisms were introduced for ion separation based on their *m/z*. In magnetic or electric sector-type mass analyzers this is achieved by manipulating, respectively, the molecular momentum or the kinetic energy of an ion[54]. Quadrupole (Q) mass analyzers and radio frequency ion traps (RF IT or Paul IT) separate ions based on their path stability in oscillating electrical fields[55]. The velocity of ions is used to derive the *m/z* values in time-of-flight (TOF) instruments[56]. Finally, variations of ion trap technologies, namely the Penning trap exemplified by the Fourier transform ion cyclotron resonance (FT-ICR)[57], and the Kingdon trap exemplified by the Orbitrap[58] mass analyzers, discern between ions' *m/z* based on their angular frequencies. The scope of MS applications is defined by the various characteristics of the mass analyzers, *e.g.* mass range, resolution, and sensitivity to name a few. (see Table 2).

Table 2 | Characteristics of the most commonly used mass analyzers.

	RF IT	Q	TOF[59]	FT-ICR[60,61]	Orbitrap[62,63]
Mass accuracy	Low (100 ppm)	Low (100 ppm)	Good (5-2 ppm)	Excellent (100 ppb)	Good (2 ppm)
Resolving power @200*m/z*	Low (<10,000)	Low (<10,000)	Moderate (<100,000)	The highest (>2,000,000)	High (> 500,000)
Ion manipulation	RF and DC electric fields	RF and DC electric fields	Electric fields (static and pulsed)	Magnetic field and electric field	Electrostatic field
Separation principle	Path stability	Path stability	Flight time $t= k \sqrt{(m/z)}$	Angular frequency $\omega=B/(m/z)$	Angular frequency $\omega=\sqrt{(k/(m/z))}$
Mass range	Moderate	Moderate	The highest	High	High
Ion Storage	+	-	-	+	+
Sensitivity	Femtomole	Attomole	Femtomole	Femtomole	Femtomole
Speed	Fast	Moderate	Fast	Slow	Moderate
Cost	*	**	***	****	***
Major drawbacks	Low mass accuracy and limited ion detection	Low mass accuracy and limited ion detection	Low mass accuracy, large size, low resolution, and limited ion storage	Difficult to maintain, low dynamic range, large size, and very high cost	Requires ultra-high vacuum; space charge effects
Best for	Multi-stage mass selection in MS^n-type experiments and ion chemistry	Mass selection for tandem MS analysis	Mass detection of intact proteins and protein assemblies	Mass detection of intact proteins, protein assemblies, and their fragments	Mass detection of intact proteins, protein assemblies, and their fragments

Sector mass analyzer represents the oldest design, which nonetheless is capable of providing fast, sensitive, and accurate mass detection, reaching a high dynamic range. The sector mass analyzers are perfect for precision analysis, but due to their large size, high cost, and limited throughput, they are no longer a popular choice for the analysis of proteins. In this domain, FT-ICR, which inherits benefits of the sector mass analyzers, has significant advantages in terms of ion storage and mass resolution, facilitating the analysis of intact biomacromolecules. While the highest measurement speed is primarily achieved on more simplistic quadrupole or RF ion trap platforms, low resolving power and low mass accuracy prevented broad applicability of these technologies. After three decades of biomolecular MS, Orbitrap, FT-ICR, and TOF mass analyzing technologies have become extremely valuable for the analysis of intact proteins and protein assemblies primarily due to their theoretically unlimited mass range, high resolving power, and mass accuracy, although

TOF mass analyzers have a substantially lower mass accuracy when compared to the other two platforms. At the same time, mass analyzers with lower resolving power and mass accuracy, but featuring higher scan rates (RF IT or TOF) and/or ability to filter analytes based on physicochemical properties (Q and RF IT), were incorporated in so-called hybrid instruments to further optimize the transmission pipelines and capabilities of the mass spectrometer.

Box 1 | Mass resolution, mass determination, and mass range

High mass resolution is essential to resolve isotopologues of intact proteins and their fragments, enabling the detection of ion charge, which is especially valuable for tandem mass spectrometry (MS/MS or MS^2) where complex mixtures of dissociation products are typically present at multiply charged states[64]. In such spectra, the determination of the ion charge state is essential to obtain the monoisotopic mass. Alternatively, the average mass of all detected isotopologues can be readily determined even at a lower mass resolution for proteoforms and MPCs as their ions are present in more than one charge state, as is the case in ESI[65]. In the case of increased molecular heterogeneity of the analyzed protein, *e.g.* glycosylated proteins, high resolving power is often required to resolve all the proteoforms or assembly variants, for which the mass differences are minimal on the *m/z* scale. As for the mass range, native proteins and protein complexes carry significantly fewer charges when compared to ions formed under denaturing ESI conditions, as a result of the decreasing number of putative protonation sites due to the more globular conformation of the proteins in the native solution[66]. Moreover, for assemblies containing DNA/RNA, negative charges of the phosphorylated polynucleotide backbone result in ions with a substantially higher *m/z* than for complexes of comparable mass assembled solely of proteins[67].

Taken together, an ionization source capable of transferring protein ions unperturbed (primarily ESI) into the gas-phase combined with a high-resolution mass analyzer (*e.g.* TOF, FT-ICR or Orbitrap) provides a minimalistic setup for researchers to detect the intact proteoforms or MPCs.

Mass measurement alone has proven capable of providing valuable information about sample content and revealing some structural aspects of analytes due to the intrinsic properties of ESI (see also Box 1). Distinct mechanisms for ion formation of denatured and native proteins in the process of ESI were proposed long ago and recently were supported by Molecular Dynamics simulations[68]. For proteins retaining a more globular, native-like conformations ionization occurs primarily via the charge residue model (CRM). As for proteins analyzed under denaturing conditions, hence with more extended conformations, ions are formed preferentially through a charge ejection model (CEM) (Figure 2).

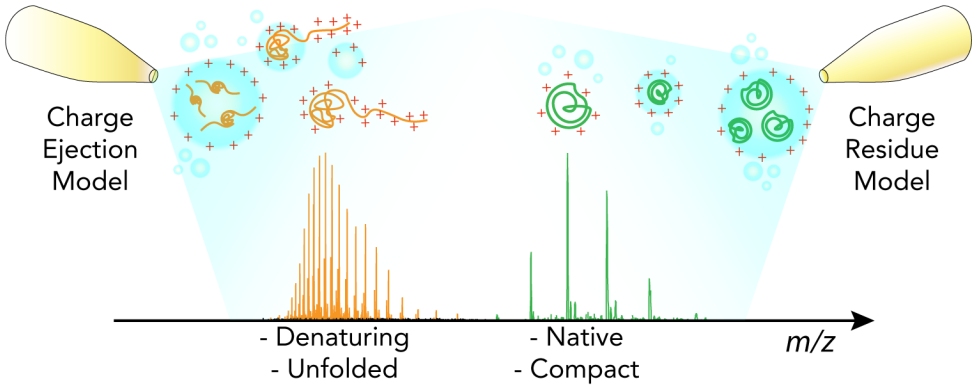


Figure 2 | Electrospray ionization in denaturing and native mass spectrometry occur via the charge ejection model and charge residue model, respectively.

The number of charge states, as well as the number of charges per charge state, is, therefore, indicative of the protein conformation. More globular proteins have fewer available protonation sites (*e.g.* Arg, Lys and His residues) resulting in the formation of ions with fewer charges in contrast to highly-charged ions of denatured proteins that have most putative protonation sites exposed. This behavior also dictates the expected *m/z* range for detection of denatured proteins to be lower than required for native proteins. By inducing protein unfolding in solution, for instance by changing the pH from neutral towards acidic or increasing the temperature of the protein incubation, researchers can readily gain insights into dynamics of protein unfolding or determine whether a protein is part of a non-covalent assembly. In all cases, a single molecule is represented by multiple charge states upon ESI, providing a means for precise mass determination by averaging the many mass detections of the same molecule. This feature of ESI shapes the central principle behind intact mass experiments, *i.e.* detection of intact proteins under denaturing conditions, and native MS experiments. The detection of masses can be used for the preliminary assessment of PTM patterns, in case of intact mass measurements, as well as for the determination of subunit stoichiometries and composition of protein assemblies in native MS by matching the observed masses to the theoretical masses.

Notwithstanding the importance of accurate and precise mass determination, even with the ultra-high mass resolutions achieved by FT-ICR instruments, unambiguous determination of the elemental composition of peptide ions based on exact masses, thus far, could only be achieved for peptides with a mass below ~1 kDa[69]. Therefore, because the mass of intact protein ions is not directly characteristic of its primary structure, it was crucial to develop alternative approaches. The versatility of mass spectrometry for the structural analysis of proteins was extended mainly by tandem mass spectrometry (MS/MS or MS^2), whereby certain precursor ions are mass (*m/z*) selected, typically using a quadrupole, and fragmented with simultaneous detection of the dissociation product masses. In instruments equipped with an RF IT or FT-ICR, this process could be repeated providing multi-stage MS/MS capabilities, termed MS^n, where the specific product(s) of a dissociation step can be further mass selected and fragmented. Since accurate mass measurements do not allow for direct inference of protein sequence, to say nothing of secondary-to-quaternary structural levels, activation of proteins in the gas-phase triggering the formation of informative fragment ions became the method of choice for the characterization of protein primary structures and, recently, was extended to the investigation of higher-order structural features of intact MPCs.

VERSATILITY OF TANDEM MASS SPECTROMETRY

Tandem mass spectrometry can interrogate protein structure and function in three conceptually distinct ways: at the level of peptides, intact proteoforms, and their complexes (Figure 3). The first tier is based on the fragmentation of peptides and is predominantly utilized for protein identification. The investigation of peptides is centered around the widely-used and heavily optimized methodology of "shotgun" or bottom-up mass spectrometry[23]. The technique is primarily used for high-throughput analysis of complex protein mixtures, reaching the level of entire

proteomes; however, it is also useful in the analysis of purified protein complexes providing a detailed overview of the subunits and co-purified interactors alongside with an (incomplete) snapshot of the PTMs they harbor. By virtue of available protein sequence databases, derived from sequenced genomes or transcriptomes, *e.g.* Uniprot[70], and sophisticated software solutions for the analysis of large protein datasets, the bottom-up approach can identify nearly all known proteins present in a sample as complex as a whole lysate. The main principle behind this tier of tandem mass spectrometry is digestion of proteins with proteases, specific or unspecific, resulting in a mixture of peptides that are then analyzed with liquid chromatography (LC) coupled online to nESI-MS/MS. The generated peptide patterns are straightforward to predict from available protein sequences and known site-specificities of the proteases. Furthermore, only a few peptides are required to confidently identify and – in most cases – relatively quantify a protein, hence the approach works for various samples, providing a high dynamic range and has virtually no limitation in protein size or other physicochemical properties. While the average size of peptides generated by trypsin, the most commonly used protease in proteomics, is below three kDa, certain proteases can produce significantly larger peptides. When these proteases are employed, the approach is distinguished from bottom-up proteomics and has been termed "middle-down proteomics"[71]. The main drawback of proteolytic approaches is the loss of peptide-to-proteoform connectivity, which hampers direct characterization of proteoform structures and determination of combinatorial PTM patterns[72], although successful attempts to restore this connection by hybrid MS approaches have been reported[73,74]. Overall, bottom-up MS is not only an essential tool for protein identification and quantification, PTM discovery and localization, but it is also used as the primary companion of many valuable structural techniques like surface covalent and non-covalent labeling as well as cross-linking mass spectrometry[75].

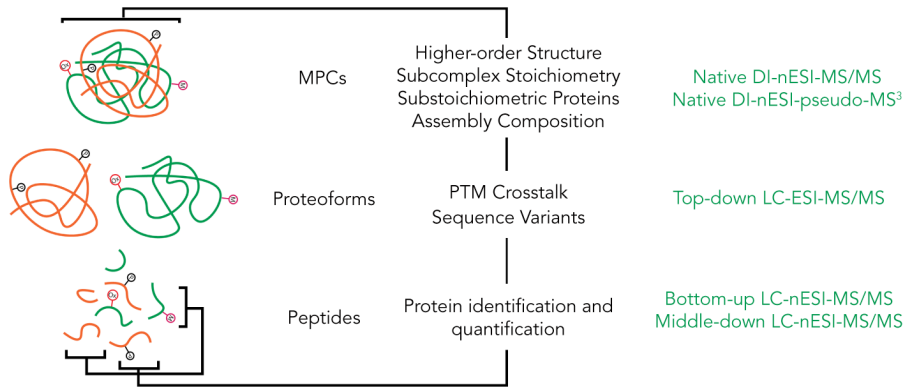


Figure 3 | Three tiers of tandem mass spectrometry facilitate analysis of protein samples at the level of peptides, proteoforms, and their assemblies.

Top-down mass spectrometry is the second tier of common tandem mass spectrometry approaches. This technique allows for the direct identification of intact proteoforms with a molecular weight (Mw) of up to ~100 kDa[76] and nearly complete characterization of proteoforms with Mw below 30 kDa with resolutions up to the residue-level. Like bottom-up MS, conventional top-down MS analyzes mixtures of

proteins and, therefore, requires a prefractionation step prior to MS analysis (Figure 4A). The most common examples include LC and less widely-used capillary electrophoresis (CE[77]/CZE[78]). These techniques allow researchers to analyze samples of medium complexity and often require an additional offline prefractionation step to reduce the complexity from the level of the proteome down to a mixture of a few dozen proteoforms. For this purpose, various implementations of gel-based separation[79] or LC fractionation, *e.g.* size-exclusion chromatography (SEC) or ion-exchange chromatography (IEX), have been used. Fractions of proteoforms are typically reduced and denatured prior to LC-ESI-MS/MS analysis.

Notwithstanding the benefits of top-down MS for structural investigations of proteoforms, the technique has significant drawbacks compared to bottom-up MS, in terms of lower sensitivity and considerably more challenging ionization and MS/MS analysis. Moreover, although often capable of providing an in-depth characterization of primary structures, top-down MS mainly operates with denatured proteins, hampering the analysis of their higher-order structural features. The advances in instrumentation and online processing algorithms, especially in the domains of FT-ICR[80] and Orbitrap-based mass spectrometers[81,82] (discussed in **Chapter 5**), have allowed developments to be made in the past decade that substantially increase the throughput and MS/MS capabilities for the analysis of intact proteoforms by top-down MS.

The final tier of tandem mass spectrometry approaches is native top-down mass spectrometry, which allows the structural analysis of intact proteoforms and MPCs in their native states, *i.e.* preserving non-covalent interactions, like metal ions, ligands, and solution-like conformations. In contrast to conventional top-down MS, native top-down MS operates with purified protein samples introduced into the mass spectrometer by means of direct infusion (DI) using a static ESI source. Mainly this is because online separation techniques operating with native buffers are not yet well-established. Nevertheless, several attempts have been made, mostly succeeded by intact mass analysis[83–85]. Furthermore, direct infusion-nESI-MS/MS is more suitable for the interrogation of the charge-reduced folded ions of proteins and their assemblies, as they fragment less efficiently than their denatured counterparts, producing lower abundant fragments that require longer acquisition times. Among all the tandem mass spectrometry approaches, native top-down MS is the only one that allows for the direct interrogation of higher-order structural information. Already at the beginning of this century, fragmentation of native proteins with in-source electron-capture dissociation (ECD) MS was pioneered by Breuker *et al.* for the investigation of protein folding dynamics[86] and protein higher-order structure[87,88]. However, obtaining covalent fragmentation of large multiproteoform assemblies is challenging due to the drastically different nature of dissociation products, spanning from low *m/z* covalent backbone fragments to low-to-medium *m/z* ejected intact subunits to high *m/z* non-dissociated residual precursors. Non-covalent partitioning alone, discussed in detail in the following sections, makes native top-down MS a powerful tool to interrogate composition, topologies, and structure of multiproteoform assemblies[89,90]. On the other hand, the ability to gain covalent fragments directly from intact multimeric macromolecules is highly useful for the in-

vestigations of endogenous protein complexes in a discovery mode[91] as well as for in-depth structural analysis[92]. With recent "thrusts" in instrumental developments, multi-stage tandem MS approaches are gaining popularity for native top-down MS experiments[93–95] (Figure 4B).

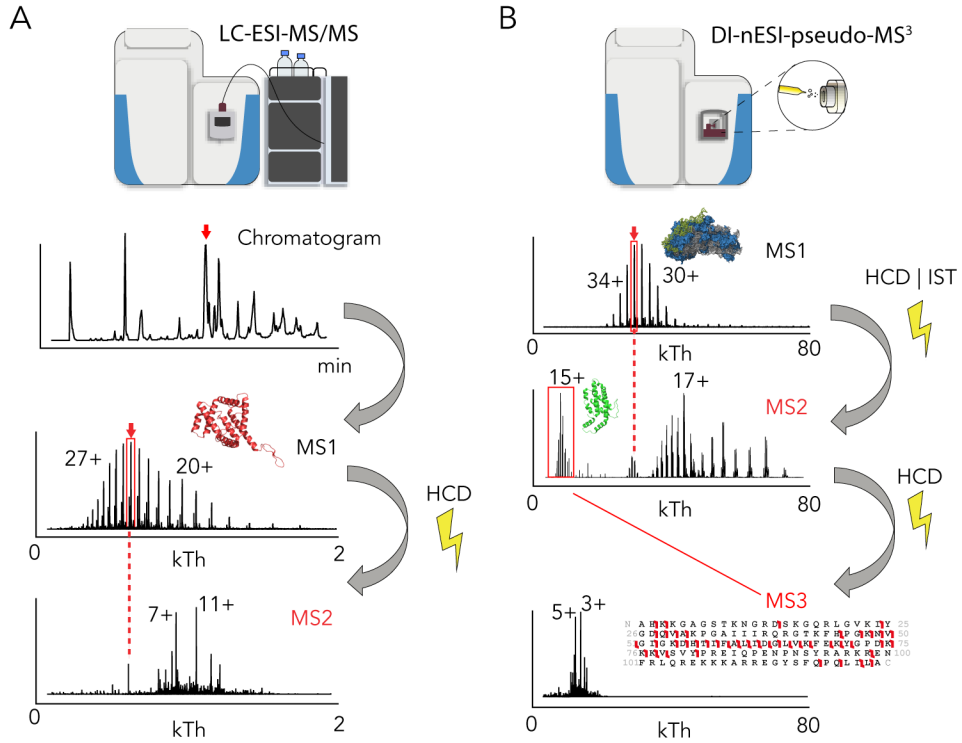


Figure 4 | Examples of experimental workflow for (A) top-down LC-ESI-MS/MS analysis and (B) native top-down DI-nESI-pseudo-MS3.

Altogether, the three tiers of tandem mass spectrometry provide researchers with a toolset for an all-level MS-based structural interrogation of multiproteoform assemblies. Bottom-up MS is a sensitive tool to probe protein content and potentially occurring PTMs, *e.g.* glycosylation, chromophorylation, and many others. Top-down MS is a powerful method for capturing the landscape of protein heterogeneity by characterizing and quantifying primary structures of sequence variants and their proteoforms. The third tier, native top-down MS, allows for direct identification of protein assemblies and interrogation of distinct variants of MPCs as well as their structural features. Combination of the highly complementary data obtained with all these tiers provides novel and invaluable insights into the function of the protein systems as will be highlighted in this thesis.

Over the past three decades, several methods were invented to facilitate tandem MS and ion activation in the gas-phase. Combined with state-of-the-art instrumentation for ionization and mass detection, these methods produce highly informative MS/MS spectra containing thousands of mass features, enabling to uncover multiple layers of protein structure.

GAS-PHASE ACTIVATION OF PROTEINS

Covalent fragmentation of proteoforms

To untangle protein sequences, ions must be activated in the gas-phase yielding the products of unimolecular decomposition with, primarily, cleavage of the peptide backbone. These products provide diagnostic mass differences that correspond to characteristic masses of the 20 naturally occurring amino acid residues or corresponding derivatives. Although most residue masses are easily distinguished from others on a high-resolution instrument, there are a few that represent a challenge for mass spectrometry because they are isobaric, *i.e.* having identical mass (Leucine and Isoleucine, m = 113.084064 Da), or having masses that differ very little (Glutamine and Lysine, m = 128.058578 and m = 128.094963). Fragmentation of the backbone can occur through several different pathways resulting in distinctive fragment types. A systematic nomenclature for peptide fragmentation pathways has been introduced by Roepstorff *et al.*[96] (Figure 5A).

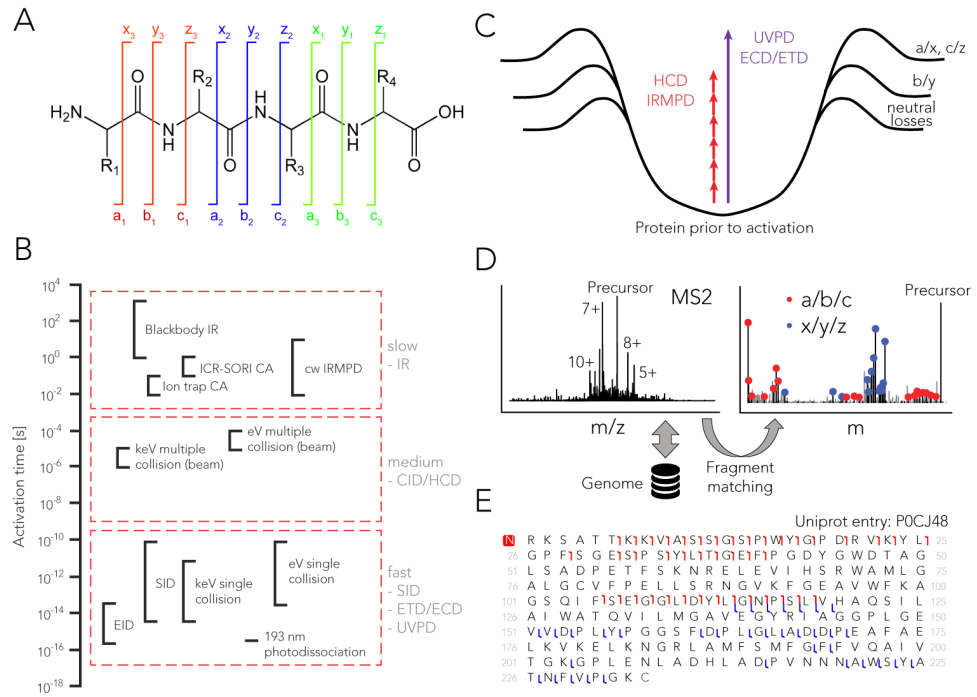


Figure 5 | (A) Nomenclature of peptide and protein fragment ions. (B) The timescale of the common gas-phase activation methods in proteomics. (C) Schematic of energy potentials and covalent dissociation pathways for proteins activated in the gas phase. (D) Fragment ions of an unknown protein are typically mass matched to the theoretically generated ions of all proteoforms derived from the genome of a specific organism(s). (E) Fragmentation map of an identified proteoform with N-terminal acetylation supported by N-terminal fragments (in red).

In this nomenclature, a fragment ion is described by a letter and a number. Lower-case alphabet letters are used to specify the type of bond cleaved upon activation and the number represents the position within the protein or peptide backbone

relative to the N- or C-terminus. The first letters of the Latin alphabet (*a*, *b*, *c*) denote N-terminal fragments and the last letters of the Latin alphabet (*x*, *y*, *z*) represent C-terminal fragments. Distinctive fragmentation pathways stem from different activation mechanisms that can be divided into two major categories of ergodic and non-ergodic processes.[97,98] (I) The first category encompasses slower low-energetic fragmentation techniques, *e.g.* the most widely-used collision-induced dissociation (CID) and infrared multiphoton dissociation (IRMPD) (Figure 5B, C). For these methods, the internal energy of the subjected ions is slowly increased through collisions with the neutral gas molecules or absorption of low energy IR photons. Then the deposited energy is distributed along the backbone via intra-vibrational redistribution (IVR) and, finally, released through cleavage of the most thermally labile amide bonds (C-N). This process results in the preferred formation of *b*- and *y*-type fragment ions with marginal amounts of *a*-type fragments. Furthermore, low-energetic fragmentation processes can produce fragments with neutral losses of H_2O or NH_3. (II) The second category includes more energetic methods, whereby polypeptides undergo, primarily, site-specific cleavage without IVR processes. The most prominent examples are electron- and UV photon-based activation (Figure 5B, C). In electron-capture dissociation (ECD)[99] or electron-transfer dissociation (ETD)[100] preferentially the N-C_α bonds in polypeptide chains are cleaved producing *c*- and *z*-type fragment ions.

Utilizing the fluence of energetic UV photons, *e.g.* at 193 nm wavelength, causes polypeptides to dissociate with cleavage of C_α-C bonds, preferentially resulting in *a*- and *x*-type fragment ions, often alongside less abundant other fragment ion types: *b*-, *c*-, *y*-, and *z*-type fragments[101]. Additionally, products of unimolecular dissociation can undergo further fragmentation by cleavage of the amino acid side chain or another peptide bond, yielding *d*-/*w*-type fragments and internal fragments[102], respectively. However, these fragments are rarely used in database searches as they complicate the analysis of MS/MS spectra, *e.g.* by introducing a combinatorial explosion of options in the case of internal fragments, except a few niche cases[103]. While covalent backbone fragmentation of proteoforms allows for the characterization of the primary structure, identification of a protein is typically achieved through matching masses of the detected fragments against an *in silico* generated database of fragments for theoretical proteoforms derived from a sequenced genome or transcriptome of a given organism (Figure 5D, E).

Because activation methods induce fragmentation through distinctive mechanisms, they often yield information covering different regions of the protein sequence and/or PTMs. There are several factors driving diversification of fragmentation patterns, including amino acid content, the charge state of an ion and/or the presence of disulfide bridges, as well as other PTMs. These effects become even more pronounced when proteins are fragmented in their native state, retaining interactions responsible for protein folding. While these interactions are largely disrupted following IVR – as encountered in the low-energetic activation approaches – they often survive when fragmentation is induced in a more energetic and site-specific manner with electron- or UV photon-induced fragmentation. This effect is most prominent for ETD and ECD of native proteins, wherein an electron captured by

the low-charged protein ions often results in cleavage of the N-C_α bond and neutralization of one charge state. However, Coulombic repulsion of the formed fragments is not sufficient to set them apart, therefore resulting in the detection of the charge-reduced precursor ion and absence of any fragments. Among solutions for circumventing this electron-transfer-no-dissociation (ETnoD) behavior, hybrid fragmentation methods appeared where ETD is followed by some type of low-energetic supplemental activation. One of the most prominent examples include EThcD[104], where ETD is supplemented with higher-energy collisional dissociation (HCD) available on Orbitrap-based mass spectrometers or activated ion ETD (AI-ETD)[105], where the protein/peptide ions are, first, activated by IRMPD followed by the ETD reaction of the ions with readily disrupted non-covalent interactions.

Fragmentation of native proteins can also be hampered by the presence of disulfide bridges, complicating the detection of fragments in-between the bridged cysteines. It has been reported that in electron-based methods the electron is captured with high affinity by a disulfide bond, which is formed by the cysteines. In the case of disulfide bridged peptides, ETD preferably cleaves the S-S bridge and releases the two bridged peptides as the primary dissociation pathway[106]. In the next step, supplemental activation is typically applied to sequence the detached peptides further. ECD has very similar behavior to ETD, although it has been so far limited mainly to the most complex FT-ICR systems because the technique requires manipulating a beam of thermal electrons. However, with recent developments in instrumentation, ECD was implemented on a benchtop Orbitrap Q Exactive mass spectrometer[108]. In another recent development, it was shown that UVPD is also able to induce preferential cleavage of the disulfide bridges with the S-S bond acting as a UV-absorbing chromophore[109].

The fact that some fragmentation methods allow for retention of non-covalent interactions while producing site-specific backbone cleavages has significant benefits for the analysis of non-covalent protein-protein and protein-ligand complexes. By interrogating fragmentation patterns obtained with ETD/ECD/UVPD of such complexes, one can determine ligand binding as well as provide a means for positioning the interaction site as was demonstrated on several heme-binding and metal-binding proteins[91,110–113]. Moreover, the techniques mentioned above allow for the detection of non-covalent interactions[30,114] as well as charge localization[115] via in-depth analysis of the generated covalent fragments. For the overview of discussed fragmentation techniques see Table 3.

Non-covalent partitioning of protein assemblies

Advances in native mass spectrometry brought many exciting avenues for analytical chemists to explore. With ever-expanding *m/z* ranges and solutions for efficient cooling and transmission of large ions, the size of non-covalent assemblies eligible for mass spectrometric analysis was extended up to viral particles in the mega-Dalton (MDa) range. However, as is the case for MS analysis of isolated proteoforms, obtaining masses of native complexes alone often provides too little information

Table 3 | Overview of commonly used activation methods in the gas-phase analysis of proteins.

	CID/HCD	IRMPD	SID	ECD	ETD	EThcD/AI-ETD	UVPD
Timescale	Slow	Slow	Fast	Fast	Fast	Medium	Fast
Principle	Collision with neutral gas molecules	Absorption of multiple low-energetic IR photons (*e.g.* λ=10.3μm)	Impact with surface	Capture of near-thermal electron (E<0.2eV)	Transfer of electron from reactive radical anion to protein	Electron transfer and collision with neutral gas molecule	Absorption of high-energetic UV photon (*e.g.* λ=193nm)
Energy deposition	Low	Low	Very High	High	High	High + Low	High
Predominant ion types	b, y	b, y	b, y	c, z	c, z	b, c, y, z	a, b, c, x, y, z
Protein sequence coverage	Low	Low	Low	Medium-High	Medium-High	High	High
Disrupts non-covalent interactions	+	+	+/-	-	-	+/-	+/-
Primary dissociation pathway for protein complexes	Asymmetric partitioning	Asymmetric partitioning	Symmetric partitioning	Surface exposed fragments	Surface exposed fragments	Covalent fragmentation	Mixed dissociation pathway
Structural information	Stoichiometry, composition	Stoichiometry, composition	Topology, stoichiometry, composition, ligand binding	Structurally exposed regions, positioning of labile PTMs	Structurally-exposed regions, positioning of labile PTMs	Primary structure	Primary structure, ligand binding
Best for	Peptide sequencing, supplemental activation[104]	Peptide sequencing, supplemental activation[107]	Investigation of subunit topologies and binding interfaces	Denatured proteins with labile PTMs	Denatured proteins with labile PTMs	Sequencing of denatured and native proteins	Denatured and native proteins and protein complexes

for structural characterization of MPCs, especially for heterogeneous assemblies. Hence, gas-phase activation of non-covalent assemblies has gained interest in the past two decades. As a result, different fragmentation techniques were explored in pursuit of finding the optimal tools for the structural interrogation of such assemblies.

While gas-phase activation of a single proteoform primarily leads to the production of covalent fragments, for an MPC multiple dissociation scenarios exist in parallel. Alongside with covalent cleavage of backbone bonds, disruption of non-covalent interactions in the binding interfaces leads to partitioning of the non-covalent assemblies into the comprising subunits. For CID, non-covalent dissociation of multimeric assemblies involves a very intricate landscape of structural rearrangements, ultimately resulting in a pathway termed asymmetric charge or complex partitioning. In brief, a single subunit of a given multimeric assembly undergoes substantial unfolding with concomitant charge rearrangement that is driven by Coulombic repulsion and exposure of the new protonation sites on the unfolding subunit chain. This process is followed by the ejection of the most unfolded, *i.e.* highly-charged, subunit, resulting in a characteristic dissociation pattern with low *m/z* highly-charged ions of the ejected subunit and the high *m/z* charge-reduced residual precursor missing the single ejected subunit. Such non-covalent dissociation behavior proved to be very useful in the determination of the composition of an assembly, subunit stoichiometries, as well as, some structural aspects of multiproteoform assemblies[89,116–118]. Although heavily utilized for more than two decades, only recently insights into the understanding of the mechanics underlying asymmetric charge partitioning were described[119].

Distinctive non-covalent complex partitioning was observed with higher-energetic fragmentation methods such as surface-induced dissociation (SID)[120] and, more recently, UVPD[121]. The technique of SID was developed mainly in the group of Vicki Wysocki and applied to the structural analysis of non-covalent assemblies[122]. SID operates by forcefully impacting native protein complexes against a solid surface, resulting in rapid deposition of the energy and complex partitioning prior to IVR and concomitant unfolding of the subunits. Contrastingly to CID, SID can provide a snapshot of assembly components in their nearly-native state. In the example of tetrameric streptavidin, instead of leading to the ejection of a single unfolded subunit, as is observed in CID, SID results in the following partitioning schema. In one dissociation pathway, a single subunit is ejected carrying a nearly symmetrical number of charges relative to the charge of the complete assembly. Concomitantly, the residual trimer retains a larger portion of charges than the ejected monomer. In an alternative dissociation pathway, more interestingly, the complex partitions into intact dimers of subunits (Figure 6). From these observations it can be inferred that the assembly represents a dimer of dimers exhibiting a non-uniform strength of binding interfaces, rather than a uniformly assembled tetramer. Later, UVPD was shown to produce similar dissociation patterns as observed for both SID and CID (see **Chapter 2**).

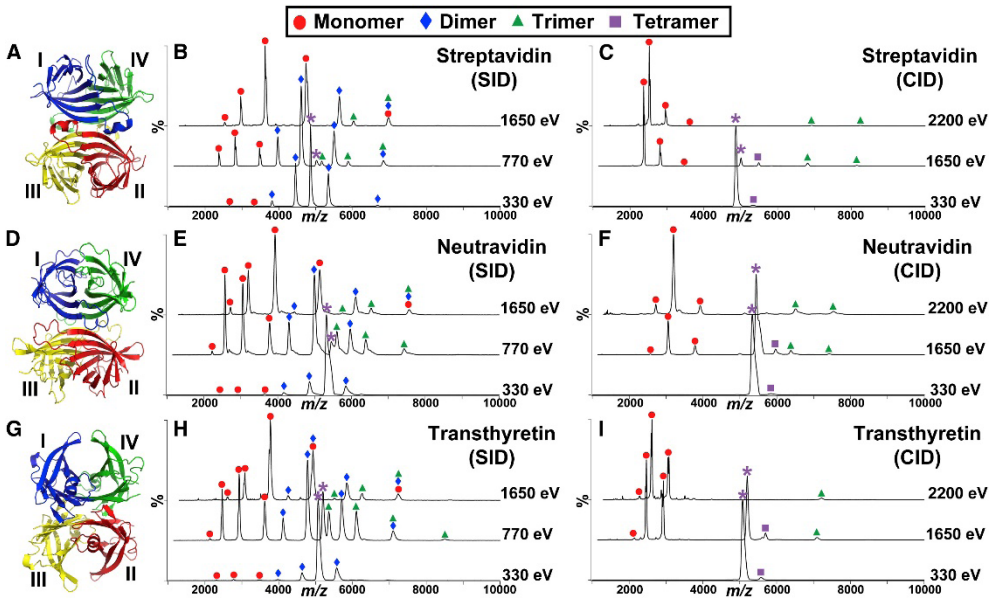


Figure 6 | Tandem MS-induced non-covalent partitioning of the tetrameric protein complexes streptavidin (A-C), Neutravidin (D-F), and Transthyretin (G-I) with CID (C, F, I) and SID (B, E, H). Figure adapted from Quintyn, R.S., Yan, J., and Wysocki, V.H. (2015). Surface-Induced Dissociation of Homotetramers with D2 Symmetry Yields their Assembly Pathways and Characterizes the Effect of Ligand Binding. Chem. Biol. 22, 583–592, with permission from Elsevier.

NOVEL ORBITRAP-BASED SOLUTIONS FOR THE TOP-DOWN ANALYSIS OF PROTEINS

Even though TOF and FT-ICR mass analyzers both have seen much progress in their applications for structural analysis of proteins, a great deal of experimental work in this thesis was performed on the Orbitrap-based platforms. Hence, in this section, the most prominent features of the Orbitrap mass spectrometers that facilitate tandem MS of intact proteoforms and MPCs are described. The technological developments of the Orbitrap platform include new prospective fragmentation techniques for more efficient structural investigation of proteoforms, as well as features facilitating the analysis of MPCs with ever-expanding mass. Only a few options of these mass spectrometric designs for top-down and native top-down MS have reached a commercial status, with most of them confined to research laboratories.

Orbitrap-based instruments used for top-down and native top-down MS in this thesis include the tribrid Orbitrap Fusion Lumos (Lumos) and several modified versions of the benchtop Orbitrap Q Exactive (QE) mass spectrometers. The former instrument is equipped with a useful toolset for conventional top-down MS methods of proteins with Mw < 50 kDa, providing several fragmentation options (CID, HCD, ETD, EThcD, and UVPD) and multi-stage tandem MS (MS^n) capabilities (Figure 7A) allowed by the two integrated linear ion traps. The QE series of instruments (Figure 7B-E) lack multiple fragmentation methods as well as MS^n capabilities when compared to the Tribrid Lumos instrument. Notwithstanding, they proved indispensable for native MS with many useful solutions built on the Orbitrap QE platform. Additionally, novel data processing algorithms and optimizations for intact protein analysis were recently introduced[79] (also see method section of **Chapter 5** and **Chapter 6**), allowing for throughput top-down MS analysis on QE-type instruments with fragmentation limited so far to HCD only.

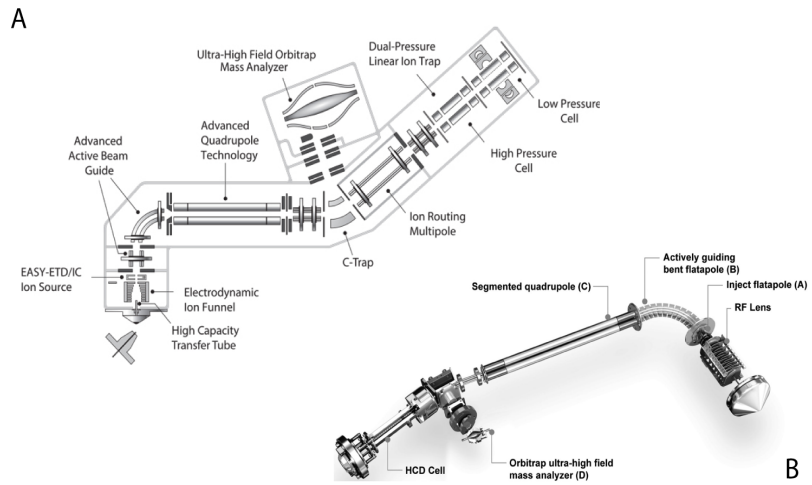


Figure 7 | Schemes of several different Orbitrap-based instrumental setups. (A) Schematics of the Tribrid Orbitrap Fusion Lumos instrument featuring a quadrupole and a linear ion trap among other compartments. (B) Schematics of an Orbitrap Q Exactive HF equipped with a quadrupole and an HCD cell. (C) Schematics of an Orbitrap Q Exactive mass spectrometer equipped with Excimer laser (C) for 193 nm UVPD and (D) with an ECD cell from e-MSion. (E) Orbitrap Q Exactive with ultra-high mass range (UHMR) capabilities enables for multi-stage sequencing of MPCs using front-end in-source trapping (IST) and back-end HCD.

Modified Orbitrap-based instruments for the analysis of entire macromolecular assemblies emerged recently[123] and featured an extended mass range quadrupole for transmitting and selecting high-*m/z* ions, lower RF frequencies of ion optics, as well as optimized ion cooling essential for the analysis of large non-covalent assemblies that often suffer from incomplete desolvation.

The missing ingredient for direct structural investigation of MPCs was the ability to obtain backbone fragmentation directly from intact assemblies. It was later achieved in two manners: either by using high energetic fragmentation methods like UVPD or ECD as applied to intact MPCs, or by performing multi-stage dissociation. The former solution was achieved by coupling already available instruments to UV lasers (discussed in **Chapter 2**) or supplementing a component on the ion transmission path with an ECD cell (Figure 7C, D). For the latter, the instruments were modified to allow for collisional activation in the front-end of the instrument, known as in-source trapping (IST) in Orbitrap mass spectrometers, with subsequent mass selection and fragmentation of released subunits in the HCD cell[94,124] (Figure 7E) (see **Chapters 5** and **6**).

THE ERA OF HYBRID APPROACHES

The multiple techniques available in the arsenal of structural mass spectrometry are largely complementary, providing an exceptionally powerful toolset for the analysis of MPCs. Native MS provides information on entire protein assemblies, including insights into composition, topologies, and stoichiometries; however, it struggles to characterize every subunit in such multimeric assemblies or to pinpoint all the decorating PTMs. In a complementary manner, top-down proteomics identifies the primary proteoforms present in the sample, but typically loses information about their higher-order assembly. For even more sensitive and detailed analysis, bottom-up MS sheds light on the positions and structure of more complex PTMs and detects contaminants and low abundant substoichiometric proteins. Only when optimally combined, these methods do allow to obtain a larger detailed picture of the protein assembly structure and function.

The combination of bottom-up MS with intact mass or native MS proved beneficial for characterizing the intrinsically heterogeneous glycoproteins and their glycan content[74,125]. Similarly, the benefits of combining native MS with top-down proteomics were explored in the characterization of antibody-drug conjugates (ADCs)[126]. By integrating native MS with gas-phase separation of proteins based on their mobility through neutral gas molecules[33] conformers and certain assembly states could be investigated independently. With advances in algorithms for the analysis of IM-MS data, new avenues for the analysis of MPCs are being explored using the technique of collision-induced unfolding (CIU) (see **Chapter 3**). The latter is especially useful as a tool for interrogating higher-order structures, stabilities as well as binding affinities within protein assemblies[127]. Additionally, an extra layer of information is obtained by combining MS with methods that allow probing protein structures in solution, like various surface labeling techniques[34,128] or chemical cross-linking mass spectrometry[129]. It is not uncommon to integrate techniques even further as was demonstrated

by combining in-solution and/or even gas-phase surface labeling with IM-MS[130,131].

Moreover, the complementarity of MS reaches far outside of the field of analytical chemistry, positioning it as a powerful supplement to other biophysical methods[132]. Bottom-up mass spectrometry is already extensively utilized as a reliable tool that can complement high-throughput and low-throughput methods for the analysis of MPCs with cryo-EM, NMR, and many other techniques. Native MS is also becoming more extensively utilized as a fast screening step prior to more elaborate structural analysis. Maturation of top-down MS, native top-down MS as well as hybrid tandem MS holds the promise of more informative methods for quality control and exploration of complex mixtures with an aim to define their structures.

At first glance, the complexity of multi-modal approaches to untangle protein structure might seem excessively laborious; however, with current advances in instrumentation, experimental design and software solutions for integrative data analysis and structural modeling the development of ever more elaborate hybrid methods is becoming increasingly promising.

SOFTWARE SOLUTIONS FOR PROTEOFORM IDENTIFICATION

Whether a proteoform is detected by an intact mass alone or detected and fragmented with tandem mass spectrometry, its mass, as well as fragment masses (if available), must be converted into the amino acid sequence decorated with PTMs to provide an identification. Although there is no uniform algorithmic recipe for proteoform identification, multiple solutions are available to perform this task by utilizing distinctive computational approaches.

Proteoforms can be identified by directly matching their theoretical masses to experimental masses provided that the protein sequences and any PTMs they harbor are known. Such an identification requires that the difference in theoretical and experimental mass is marginal, typically in the range of a few ppm. This solution can be expanded to define proteoform groups based on the mass differences observed between the experimental proteoform masses and the theoretical mass of the known protein sequence(s). This approach was used for identifying proteoform groups in ribosomal proteins, as described in **Chapter 5**. Intact mass matching and proteoform mass differences are utilized by tools like Proteoform Suite[133], which cluster mass features into proteoform families (*i.e.* groups) based on whether or not they differ by a specific mass corresponding to a PTM or mutation. Although this provides valuable information, such an approach lacks analytical depth as proteoforms with identical modifications occupying different amino acid positions, known as positional isomers, have identical masses and therefore cannot be discerned. Similarly, changing the order of amino acids in a sequence or introducing isobaric substitutions does not result in a difference in the masses of the proteoforms. Therefore, identifications based on intact mass matching must be verified with either tandem MS of intact proteoforms, *i.e.* top-down MS, or by digestion and bottom-up MS. This much more extensive approach of intact mass matching with verification

from bottom-up MS was employed in **Chapter 6** to identify B-phycoerythrin proteoforms carrying multiple chromophorylation sites.

In a top-down MS experiment, whereby proteoforms are detected and fragmented, proteoform identification can be performed in multiple ways when a database of theoretical protein sequences is available. The most naïve approach would be to generate all possible theoretical proteoforms out of the list of protein sequences with known PTMs/polymorphisms and, for each of them, perform *in silico* fragmentation producing all theoretical fragments for the fragmentation technique of choice (*e.g. b*/*y*-fragment ions for HCD; *c*/*z* and *b*/*y*-fragment ions for EThcD, *etc.*). Next, the experimentally observed fragment masses are searched against the *in silico* generated fragment masses during 'spectral alignment' resulting in several candidate proteoform spectrum matches (PrSMs). The candidate PrSMs can be graded with various metrics, *e.g.* E-value, which is calculated by multiplying the probability of getting at least the number of matching fragments due to chance by the number of candidate proteoforms. This approach is at the core of multiple software solutions, ranging from commercially available ProSigthPC[134] to free tools such as MSPathFinder[135] and MASH Suite Pro[136]. Provided that the correct PTMs and sequences are known for the protein of interest and that a sufficient amount of identified experimental fragments are obtained, this approach results in successful proteoform identification. However, in cases where modifications or exact amino acid sequences are unknown, the database search ends up with no result or false positive identification. One way to circumvent this problem requires generating all possible sequence variants and decorating them with all possible PTMs. However, in this case search space is expanded to such a degree that extremely long processing times are required, especially for large databases, leading additionally to an increased number of false positive identifications.

Addressing these limitations requires several computational optimizations. First, protein filtering can circumvent the computationally expensive steps, *e.g.* generating all possible proteoforms[137]. The filtering could be performed based on a proteoform mass or, more efficiently, on sequence information derived *de novo* from the fragmentation spectrum. The latter comes in the form of consecutive fragment masses that encode an amino acid subsequence or a "sequence tag". Only protein sequences that match to the theoretical mass or have the observed sequence tag are selected for further processing, significantly simplifying the proteoform search. Secondly, proteoform identification could be achieved with enhanced spectral alignment where the sequence is aligned to the spectrum in a manner that each pair of experimental peaks is converted into their mass difference. These differences ideally match to the mass of an amino acid residue with any of the PTMs available in a modification database like UniMod (http://www.unimod.org). This method allows for the prediction of unknown modifications and consequent detection of sequence discrepancies as implemented in, *e.g.* MS-Align-E[138] or TopPIC[141]. In more advanced implementations a proteoform can be directly reconstructed from the spectrum as a sequence of annotated mass shifts observed between experimental peaks, commonly referred to as a mass graph. The most recent software solutions implement this concept, *e.g.* TopMG[139]. Finally, proteoform identifications must be scored to

discern between true and false hits. Several advanced scoring strategies have been introduced for estimating statistical significance of the identified proteoforms spanning from C-score[143], which builds upon the E-value by taking into account precursor mass matching, cleavage site propensities and characterization of the observed modifications, to Markov chain Monte Carlo methods[144], which provide even more nuanced estimations of statistical significance at the level of proteoforms.

The primary advantage of these tailored proteoform identification algorithms helps in the case of top-down datasets that are searched against large databases. Throughout this thesis we primarily analyze small subsets of proteoforms comprising purified multiproteoform assemblies often with a focus on a single known protein. Hence, in our identification strategy we typically either use the available conventional software solutions, like ProSightPC[134] and ProSigth Lite[140] (**Chapter 5**), or in-house generated scripts (in R and C#) for specific tasks. Non-exhaustively these tasks include: fragment matching (**Chapter 2 and 4**), unbiased screening of unknown mass shifts between theoretical fragment masses and experimental masses to detect sequence processing and unknown PTMs, and dynamic PTM positioning to probe all putative modification sites in an unbiased manner. These software tools were successfully used to identify new fragment types as observed in UVPD (**Chapter 2**), and to identify and position PTMs for ribosomal proteins (**Chapter 5**) and B-phycoerythrin proteoforms (**Chapter 6**). Additionally, in **Chapter 6** we use sequence tags detected in the intact mass measurements of B-phycoerythrin proteoforms to filter large proteoform databases.

As the field of top-down proteomics is rather in its infancy, the available software suites are still undergoing maturation. Moreover, bioinformatics solutions are required to be frequently optimized in parallel with the new analytical tools for top-down proteomics (*e.g.* new mass analyzers, new fragmentation methods) that are being introduced by industry and academia, as also described in this thesis. The *ad hoc* solutions to the emerging analytical challenges are being gradually implemented by the generic software tools, ultimately contributing to the optimal toolbox for MS-based proteoform identification.

THESIS OVERVIEW

The **Chapter 2** to **6** in this thesis illustrate the power of tandem mass spectrometry for investigating various structural aspects of non-covalent assemblies at all levels of molecular complexity, from the primary structure of proteoforms to binding interfaces and other higher-order structural features of MPCs. The applicability of these methods is demonstrated on recombinant and endogenous analytes, including homo- and heteromeric assemblies, protein-peptide complexes, various ribosomal particles, and photosynthetic light-harvesting assemblies.

In the first part of this thesis (**Chapter 2** to **4**), tandem mass spectrometry of intact protein assemblies is explored as a multi-functional tool for extracting useful information for all levels of molecular complexity. In **Chapter 2,** the power of a newly emerging fragmentation technique, UVPD, is explored for the structural analysis of various non-covalent protein complexes. The chapter provides an in-depth overview

of the observed dissociation pathways of the native proteins and protein complexes upon UVPD and HCD. The effects of metal ion binding as well as disulfide bridges on the observed fragmentation patterns are discussed. An integrative structural MS approach combining native MS and IM-MS is applied in **Chapter 3** to investigate the stabilities of the homologous co-chaperonin GroES and gp31 complexes, that both are required to cap GroEL and enable it to fold substrate proteins. The results obtained for these homomeric assemblies in the gas phase by tandem mass spectrometry and ion mobility are compared to the stabilities observed in solution, thus providing insights into structural rearrangements upon the transition to the gas phase. An unconventional approach to gas-phase covalent labeling for structural analysis of non-covalent complexes is explored in **Chapter 4**. By utilizing the phenomenon of gas-phase phosphate migration and multi-stage EThcD-MS^3, we were able to pinpoint the binding site in a phosphate-mediated non-covalent complex.

The second part of this thesis (**Chapter 5** and **6**) focuses on integrative tandem mass spectrometry and its ability to dissect structural and compositional heterogeneities in MPCs. The presented data depicts the power of hybrid tandem MS in obtaining a larger snapshot of complex molecular machines, providing insights not only into the structure but also function of these protein systems. In **Chapter 5**, by utilizing a three-tiered MS approach incorporating bottom-up, top-down, and native MS, we explored the heterogeneities in assembly variants and characterize numerous proteoforms in four ribosomal samples purified from different organisms, providing a detailed overview of diversity observed for ribosomal particles. In addition, we explore novel instrumental capabilities of a benchtop Orbitrap instrument, including improved charge detection and data-dependent analysis in top-down MS experiments. Finally, in **Chapter 6**, by using an entire arsenal of structural tandem MS we explore the intricate nature of the light-harvesting subcomplex B-phycoerythrin from red algae *P. cruentum* and discuss the implications of determined structural and compositional assembly variants on the processes of the efficient energy transfer in the light-harvesting machinery of the phycobilisome.

REFERENCES

1. Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G.M., *et al.* (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell *164*, 805–817.
2. Toby, T.K., Fornelli, L., and Kelleher, N.L. (2016). Progress in Top-Down Proteomics and the Analysis of Proteoforms. Annu. Rev. Anal. Chem. *9*, 499–519.
3. Smith, L.M., and Kelleher, N.L. (2013). Proteoform: A single term describing protein complexity. Nat. Methods *10*, 186–187.
4. Smith, L.M., and Kelleher, N.L. (2018). Proteoforms as the next proteomics currency. Science (80-.). *359*, 1106–1108.
5. Skinner, O.S., Havugimana, P.C., Haverland, N.A., Fornelli, L., *et al.* (2016). An informatic framework for decoding protein complexes by top-down mass spectrometry. Nat. Methods *13*, 237–240.
6. Aebersold, R., Agar, J.N., Amster, I.J., Baker, M.S., *et al.* (2018). How many human proteoforms are there? Nat. Chem. Biol. *14*, 206–214.
7. Picardi, E., D'Erchia, A.M., Lo Giudice, C., and Pesole, G. (2017). REDIportal: a comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res. *45*, D750–D757.
8. Glanville, J., Zhai, W., Berka, J., Telman, D., *et al.* (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc. Natl. Acad. Sci. *106*, 20216–20221.
9. Shi, W., and Chance, M.R. (2011). Metalloproteomics: Forward and Reverse Approaches in Metalloprotein Structural and Functional Characterization. Curr. Opin. Chem. Biol. *15*, 144–148.
10. Rix, U., and Superti-Furga, G. (2009). Target profiling of small molecules by chemical proteomics. Nat. Chem. Biol. *5*, 616–624.
11. Yu, Y., Ji, H., Doudna, J.A., and Leary, J.A. (2009). Mass spectrometric analysis of the human 40S ribosomal subunit: Native and HCV IRES-bound complexes. Protein Sci. *14*, 1438–1446.
12. van de Waterbeemd, M., Tamara, S., Fort, K.L., Damoc, E., *et al.* (2018). Dissecting ribosomal particles throughout the kingdoms of life using advanced hybrid mass spectrometry methods. Nat. Commun. *9*, 2493.
13. Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., *et al.* (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat. Struct. Mol. Biol. *18*, 529–536.
14. Marty, M.T., Hoi, K.K., Gault, J., and Robinson, C. V. (2016). Probing the Lipid Annular Belt by Gas-Phase Dissociation of Membrane Proteins in Nanodiscs. Angew. Chemie - Int. Ed. *55*, 550–554.
15. Tuttle, M.D., Comellas, G., Nieuwkoop, A.J., Covell, D.J., *et al.* (2016). Solid-state NMR structure of a pathogenic fibril of full-length human α-synuclein. Nat. Struct. Mol. Biol. *23*, 409–415.
16. Liu, J.J., Horst, R., Katritch, V., Stevens, R.C., *et al.* (2012). Biased signaling pathways in β2-adrenergic receptor characterized by 19F-NMR. Science (80-.). *335*, 1106–1110.
17. Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., *et al.* (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature *181*, 662–666.
18. Kuhlbrandt, W. (2014). The Resolution Revolution. Science (80-.). *343*, 1443–1444.
19. Sirohi, D., Chen, Z., Sun, L., Klose, T., *et al.* (2016). The 3.8 Å resolution cryo-EM structure of Zika virus. Science *352*, 467–70.
20. Cheng, Y., Glaeser, R.M., and Nogales, E. (2017). How Cryo-EM Became so Hot. Cell *171*, 1229–1231.
21. Micsonai, A., Wien, F., Kernya, L., Lee, Y.-H., *et al.* (2015). Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. Proc. Natl. Acad. Sci. U. S. A. *112*, E3095-103.
22. Ranjbar, B., and Gill, P. (2009). Circular dichroism techniques: Biomolecular and nanostructural analyses- A review. Chem. Biol. Drug Des. *74*, 101–120.

23. Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. *422*.

24. Miyashita, M., Presley, J.M., Buchholz, B.A., Lam, K.S., *et al.* (2001). Attomole level protein sequencing by Edman degradation coupled with accelerator mass spectrometry. Proc. Natl. Acad. Sci. *98*, 4403–4408.

25. Restrepo-Pérez, L., Joo, C., and Dekker, C. (2018). Paving the way to single-molecule protein sequencing. Nat. Nanotechnol. *13*, 786–796.

26. Kolmogorov, M., Kennedy, E., Dong, Z., Timp, G., *et al.* (2017). Single-molecule protein identification by sub-nanopore sensors. PLoS Comput. Biol. *13*, e1005356.

27. Spence, J.C.H. (2017). XFELs for structure and dynamics in biology. IUCrJ *4*, 322–339.

28. Warnke, S., Hoffmann, W., Seo, J., Genst, E. De, *et al.* (2016). From Compact to String — The Role of Secondary of Gas-Phase Proteins. J. Am. Soc. Mass Spectrom. *28*, 638-646

29. Ruotolo, B.T., Benesch, J.L.P., Sandercock, A.M., Hyung, S., *et al.* (2008). Ion mobility – mass spectrometry analysis of large protein complexes. *3*, 1139–1152.

30. Zhang, Z., Browne, S.J., and Vachet, R.W. (2014). Exploring salt bridge structures of gas-phase protein ions using multiple stages of electron transfer and collision induced dissociation. J. Am. Soc. Mass Spectrom. *25*, 604–613.

31. Lermyte, F., and Sobott, F. (2015). Electron transfer dissociation provides higher-order structural information of native and partially unfolded. Proteomics *15*, 2813–2822.

32. Jeanne Dit Fouque, K., Garabedian, A., Leng, F., Tse-Dinh, Y.C., *et al.* (2019). Microheterogeneity of Topoisomerase IA/IB and Their DNA-Bound States. ACS Omega *4*, 3619–3626.

33. Lanucara, F., Holman, S.W., Gray, C.J., and Eyers, C.E. (2014). The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. Nat. Chem. *6*, 281–294.

34. Konermann, L., Pan, J., and Liu, Y.H. (2011). Hydrogen exchange mass spectrometry for studying protein structure and dynamics. Chem. Soc. Rev. *40*, 1224–1234.

35. Mistarz, U.H., Brown, J.M., Haselmann, K.F., and Rand, K.D. (2016). Probing the Binding Interfaces of Protein Complexes Using Gas-Phase H/D Exchange Mass Spectrometry. Structure *24*, 310–318.

36. Mistarz, U.H., Chandler, S.A., Brown, J.M., Benesch, J.L.P., *et al.* (2019). Probing the Dissociation of Protein Complexes by Means of Gas-Phase H/D Exchange Mass Spectrometry. J. Am. Soc. Mass Spectrom. *30*, 45–57.

37. Kostyukevich, Y., Ovchinnikov, G., Kononikhin, A., Popov, I., *et al.* (2018). Thermal dissociation and H/D exchange of streptavidin tetramers at atmospheric pressure. Int. J. Mass Spectrom. *427*, 100–106.

38. Liu, F., Rijkers, D.T.S., Post, H., and Heck, A.J.R. (2015). Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat. Methods *12*, 1179–1184.

39. Shaw, J.B., Li, W., Holden, D.D., Zhang, Y., *et al.* (2013). Complete Protein Characterization Using Top-Down Mass Spectrometry and Ultraviolet Photodissociation. J. Am. Chem. Soc. *135*, 12646–12651.

40. McLafferty, F.W. (2011). A Century of Progress in Molecular Mass Spectrometry. Annu. Rev. Anal. Chem. *4*, 1–22.

41. Yates, J.R. (2011). A century of mass spectrometry: From atoms to proteomes. Nat. Methods *8*, 633–637.

42. Thomson, J.J. (1913). Rays of positive electricity and their application to chemical analyses (London ; New York: Longmans, Green).

43. M. Barber R. D. Sedgwick. A. N. Tyler, R.S.B. (1981). Fast atom bombardmment of solids : a new ion source for mass specxtrometry. J. Chem. Soc. Chem. Commun *101*, 325–327.

44. Biemann, K. (1988). Contributions of mass spectrometry to peptide and protein structure. Biol. Mass Spectrom. *16*, 99–111.

45. Fenn, J.B., Man, M., Meng, C.K., Wong, S.F., *et al.* (1990). Ectrospray Ionization for Mass Spectrometry of Large Biomolecules. Science (80-.). *21*, 64–71.

46. Hillenkamp, F., Karas, M., Beavis, R.C., and Chait, B.T. (1991). Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry of Biopolymers. Anal. Chem. *63*, 1193A-1203A.

47. Liu, J., Gunawardena, H.P., Huang, T.-Y., and McLuckey, S.A. (2008). Charge-dependent dissociation of insulin cations via ion/ion electron transfer. Int. J. Mass Spectrom. *276*, 160–170.

48. Reid, G.E., Wu, J., Chrisman, P.A., Wells, J.M., *et al.* (2001). Charge-State-Dependent Sequence Analysis of Protonated Ubiquitin Ions via Ion Trap Tandem Mass Spectrometry. Anal. Chem. *73*, 3274–3281.

49. Ucal, Y., Durer, Z.A., Atak, H., Kadioglu, E., *et al.* (2017). Clinical applications of MALDI imaging technologies in cancer and neurodegenerative diseases. Biochim. Biophys. Acta *1865*, 795–816.

50. Opota, O., Prod'hom, G., and Greub, G. (2016). Applications of MALDI-TOF Mass Spectrometry in Clinical Diagnostic Microbiology. FEMS Microbiol. Rev. *36*, 55–92.

51. Stauber, J., MacAleese, L., Franck, J., Claude, E., *et al.* (2010). On-tissue protein identification and imaging by MALDI-Ion mobility mass spectrometry. J. Am. Soc. Mass Spectrom. *21*, 338–347.

52. Karas, R.M., Bahr, U., and Dülcks, T. (2000). Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine. Fresenius J Anal Chem *366*, 669–676.

53. Leney, A.C., and Heck, A.J.R. (2017). Native Mass Spectrometry: What is in the Name? J. Am. Soc. Mass Spectrom. *28*, 5–13.

54. Boyd, R.K., and Beynon, J.H. (1977). Scanning of sector mass spectrometers to observe the fragmentations of metastable ions. Org. Mass Spectrom. *12*, 163–165.

55. Dawson, P.H. (1976). Quadrupole mass spectrometry and its applications.

56. Wiley, W.C., and McLaren, I.H. (1955). Time-of-Flight Mass Spectrometer with Improved Resolution. Rev. Sci. Instrum. *26*, 1150–1157.

57. Marshall, A.G., Hendrickson, C.L., and Jackson, G.S. (1998). Fourier transform ion cyclotron resonance mass spectrometry: A primer. Mass Spectrom. Rev. *17*, 1–35.

58. Makarov, A. (2000). Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. Anal. Chem. *72*, 1156–1162.

59. Giles, R., Sudakov, M., and Wollnik, H. (2017). Time-of-flight mass spectrometer and a method of analysing ions in a time-of-flight mass spectrometer. 45 pp.

60. Smith, D.F., Podgorski, D.C., Rodgers, R.P., Blakney, G.T., *et al.* (2018). 21 Tesla FT-ICR Mass Spectrometer for Ultrahigh-Resolution Analysis of Complex Organic Mixtures. Anal. Chem. *90*, 2041–2047.

61. Shaw, J.B., Lin, T.Y., Leach, F.E., Tolmachev, A. V, *et al.* (2016). 21 Tesla Fourier transform ion cyclotron resonance mass spectrometer greatly expands mass spectrometry toolbox. J Am Soc Mass Spectrom *27*, 1929–1936.

62. Denisov, E., Damoc, E., Lange, O., and Makarov, A. (2012). Orbitrap mass spectrometry with resolving powers above 1,000,000. Int. J. Mass Spectrom. *325–327*, 80–85.

63. Eliuk, S., and Makarov, A. (2015). Evolution of Orbitrap Mass Spectrometry Instrumentation. Annu. Rev. Anal. Chem. *8*, 61–80.

64. Xu, G., Stupak, J., Yang, L., Hu, L., *et al.* (2018). Deconvolution in mass spectrometry based proteomics. Rapid Commun. Mass Spectrom. *32*.

65. Mann, M., Meng, C.K., and Fenn, J.B. (1989). Interpreting Mass Spectra of Multiply Charged Ions. Anal. Chem. *61*, 1702–1708.

66. Tong, W., and Wang, G. (2018). How can native mass spectrometry contribute to characterization of biomacromolecular higher-order structure and interactions? Methods *144*, 3–13.

67. Van De Waterbeemd, M., Fort, K.L., Boll, D., Reinhardt-Szyba, M., *et al.* (2017). High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. Nat. Methods *14*, 283–286.

68. Konermann, L., Ahadi, E., Rodriguez, A.D., and Vahidi, S. (2013). Unraveling the Mechanism of Electrospray Ionization. Anal. Chem. *85*, 2–9.

69. He, F., Emmett, R.M., Håkansson, K., Hendrickson, C.L., *et al.* (2003). Theoretical and Experimental Prospects for Protein

Identification Based Solely on Accurate Mass Measurement. J. Proteome Res. *3*, 61–67.

70. UniProt Consortium, T.U. (2008). The universal protein resource (UniProt). Nucleic Acids Res. *36*, D190-5.

71. Cristobal, A., Marino, F., Post, H., van den Toorn, H.W.P., *et al.* (2017). Toward an Optimized Workflow for Middle-Down Proteomics. Anal. Chem. *89*, 3318–3325.

72. Compton, P.D., Kelleher, N.L., and Gunawardena, J. (2018). Estimating the Distribution of Protein Post-Translational Modification States by Mass Spectrometry. J. Proteome Res. *17*, 2727–2734.

73. Yang, Y., Liu, F., Franc, V., Halim, L.A., *et al.* (2016). Hybrid mass spectrometry approaches in glycoprotein analysis and their usage in scoring biosimilarity. Nat. Commun. *7*, 1–10.

74. Franc, V., Yang, Y., and Heck, A.J.R. (2017). Proteoform profile mapping of the human serum Complement component C9 reveals unexpected new features of N- , O- and C-glycosylation. Anal. Chem. *89*, 3483–3491.

75. Gingras, A.-C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. Nat. Rev. Mol. Cell Biol. *8*, 645–54.

76. Tran, J.C., Zamdborg, L., Ahlf, D.R., Lee, J.E., *et al.* (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature *480*, 254–8.

77. Li, Y., Compton, P.D., Tran, J.C., Ntai, I., *et al.* (2014). Optimizing capillary electrophoresis for top-down proteomics of 30-80 kDa proteins. Proteomics *14*, 1158–1164.

78. McCool, E.N., Lubeckyj, R.A., Shen, X., Chen, D., *et al.* (2018). Deep Top-Down Proteomics Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry: Identification of 5700 Proteoforms from the Escherichia coli Proteome. Anal. Chem. *90*, 5529–5533.

79. Fornelli, L., Durbin, K.R., Fellers, R.T., Early, B.P., *et al.* (2017). Advancing Top-down Analysis of the Human Proteome Using a Benchtop Quadrupole-Orbitrap Mass Spectrometer. J. Proteome Res. *16*, 609–618.

80. Anderson, L.C., DeHart, C.J., Kaiser, N.K., Fellers, R.T., *et al.* (2017). Identification and Characterization of Human Proteoforms by Top-Down LC-21 Tesla FT-ICR Mass Spectrometry. J. Proteome Res. *16*, 1087–1096.

81. Fornelli, L., Durbin, K.R., Fellers, R.T., Early, B.P., *et al.* (2017). Advancing Top-down Analysis of the Human Proteome Using a Benchtop Quadrupole-Orbitrap Mass Spectrometer. J. Proteome Res. *16*, 609–618.

82. Cleland, T.P., Dehart, C.J., Fellers, R.T., Vannispen, A.J., *et al.* (2017). High-Throughput Analysis of Intact Human Proteins Using UVPD and HCD on an Orbitrap Mass Spectrometer. J. Proteome Res. *16*, 2072-2079

83. Shen, X., Kou, Q., Guo, R., Yang, Z., *et al.* (2018). Native Proteomics in Discovery Mode Using Size-Exclusion Chromatography–Capillary Zone Electrophoresis–Tandem Mass Spectrometry. Anal. Chem. *90*, 10095–10099.

84. Haberger, M., Leiss, M., Heidenreich, A.-K., Pester, O., *et al.* (2016). Rapid characterization of biotherapeutic proteins by size-exclusion chromatography coupled to native mass spectrometry. MAbs *8*, 331–339.

85. Yan, Y., Liu, A., Wang, S., Daly, T.J., *et al.* (2018). Ultrasensitive Characterization of Charge Heterogeneity of Therapeutic Monoclonal Antibodies Using Strong Cation Exchange Chromatography Coupled to Native Mass Spectrometry. Anal. Chem. *90*, 13013–13020.

86. Breuker, K., and McLafferty, F.W. (2005). The thermal unfolding of native cytochrome c in the transition from solution to gas phase probed by native electron capture dissociation. Angew. Chemie - Int. Ed. *44*, 4911–4914.

87. Breuker, K., and McLafferty, F.W. (2003). Native Electron Capture Dissociation for the Structural Characterization of Noncovalent Interactions in Native Cytochrome c. Angew. Chemie - Int. Ed. *42*, 4900–4904.

88. Oh, H., Breuker, K., Sze, S.K., Ge, Y., *et al.* (2002). Secondary and tertiary structures of gaseous protein ions characterized by electron capture dissociation mass spectrometry and photofragment spectroscopy. Proc. Natl. Acad. Sci. U. S. A. *99*, 15863–15868.

89. Benesch, J.L.P., Aquilina, J.A., Ruotolo, B.T., Sobott, F., *et al.* (2006). Tandem Mass Spectrometry Reveals the Quaternary Organization of Macromolecular Assemblies. Chem. Biol. *13*, 597–605.

90. Quintyn, R.S., Yan, J., and Wysocki, V.H. (2015). Surface-Induced Dissociation of Homotetramers with D2 Symmetry Yields their Assembly Pathways and Characterizes the Effect of Ligand Binding. Chem. Biol. *22*, 583–592.

91. Skinner, O.S., Haverland, N.A., Fornelli, L., Melani, R.D., *et al.* (2018). Top-down characterization of endogenous protein complexes with native proteomics. Nat. Chem. Biol. *14*, 36–41.

92. Li, H., Nguyen, H.H., Loo, R.R.O., Campuzano, I.D.G., *et al.* (2018). An integrated native mass spectrometry and top-down proteomics method that connects sequence to structure and function of macromolecular complexes. Nat. Chem. *10*, 139–148.

93. Ben-Nissan, G., Belov, M.E., Morgenstern, D., Levin, Y., *et al.* (2017). Triple-Stage Mass Spectrometry Unravels the Heterogeneity of an Endogenous Protein Complex. Anal. Chem. *89*, 4708–4715.

94. Belov, M.E., Damoc, E., Denisov, E., Compton, P.D., *et al.* (2013). From protein complexes to subunit backbone fragments: A multi-stage approach to native mass spectrometry. Anal. Chem. *85*, 11163–11173.

95. Mehaffey, M.R., Sanders, J.D., Holden, D.D., Nilsson, C.L., *et al.* (2018). Multistage Ultraviolet Photodissociation Mass Spectrometry To Characterize Single Amino Acid Variants of Human Mitochondrial BCAT2. Anal. Chem. *90*, 9904–9911.

96. Roepstorff, P., and Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed. Mass Spectrom. *11*, 601.

97. Brodbelt, J.S. (2016). Ion Activation Methods for Peptides and Proteins. Anal. Chem. *88*, 30–51.

98. Qi, Y., and Volmer, D.A. (2017). Electron-based fragmentation methods in mass spectrometry: An overview. Mass Spectrom. Rev. *36*, 4–15.

99. Zubarev, R.A., Kelleher, N.L., and McLafferty, F.W. (1998). Electron capture dissociation of multiply charged protein cations. A nonergodic process. J. Am. Chem. Soc. *120*, 3265–3266.

100. Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J., *et al.* (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc. Natl. Acad. Sci. *101*, 9528–9533.

101. Brodbelt, J.S. (2014). Photodissociation mass spectrometry: new tools for characterization of biological molecules. Chem. Soc. Rev. *43*, 2757–83.

102. Durbin, K.R., Skinner, O.S., Fellers, R.T., and Kelleher, N.L. (2015). Analyzing Internal Fragmentation of Electrosprayed Ubiquitin Ions During Beam-Type Collisional Dissociation. J Am Soc Mass Spectrom *26*, 782–787.

103. Lebedev, A.T., Damoc, E., Makarov, A.A., and Samgina, T.Y. (2014). Discrimination of Leucine and Isoleucine in Peptides Sequencing with Orbitrap Fusion Mass Spectrometer. Anal. Chem. *86*, 7017–7022.

104. Frese, C.K., Altelaar, A.F.M., Toorn, H. Van Den, Nolting, D., *et al.* (2012). Toward Full Peptide Sequence Coverage by Dual Fragmentation Combining Electron-Transfer and Higher-Energy Collision Dissociation Tandem Mass Spectrometry. Anal. Chem. *84*, 9668–9673.

105. Riley, N.M., and Coon, J.J. (2017). The Role of Electron Transfer Dissociation in Modern Proteomics. Anal. Chem. *90*, 40–64.

106. Liu, F., van Breukelen, B., and Heck, A.J.R. (2014). Facilitating protein disulfide mapping by a combination of pepsin digestion, electron transfer higher energy dissociation (EThcD), and a dedicated search algorithm SlinkS. Mol. Cell. Proteomics *13*, 2776–86.

107. Riley, N.M., Westphall, M.S., and Coon, J.J. (2015). Activated Ion Electron Transfer Dissociation for Improved Fragmentation of Intact Proteins. Anal. Chem. *87*, 7109–7116.

108. Fort, K.L., Cramer, C.N., Voinov, V.G., Vasil'Ev, Y. V., *et al.* (2018). Exploring ECD on a Benchtop Q Exactive Orbitrap Mass

Spectrometer. J. Proteome Res. *17*, 926–933.

109. Soorkia, S., Dehon, C., Kumar, S.S., Pedrazzani, M., *et al.* (2014). UV Photofragmentation Dynamics of Protonated Cystine: Disulfide Bond Rupture. J. Phys. Chem. Lett. *5*, 1110–1116.

110. Cammarata, M.B., and Brodbelt, J.S. (2015). Structural characterization of holo- and apo-myoglobin in the gas phase by ultraviolet photodissociation mass spectrometry. Chem. Sci. *6*, 1324–1333.

111. O'Brien, J.P., Li, W., Zhang, Y., and Brodbelt, J.S. (2014). Characterization of Native Protein Complexes Using Ultraviolet Photodissociation Mass Spectrometry. J. Am. Chem. Soc. *136*, 12920–12928.

112. Yin, S., and Loo, J.A. (2010). Elucidating the site of protein-ATP binding by top-down mass spectrometry. J. Am. Soc. Mass Spectrom. *21*, 899–907.

113. Sahasrabuddhe, A., Hsia, Y., Busch, F., Sheffler, W., *et al.* (2018). Confirmation of intersubunit connectivity and topology of designed protein complexes by native MS. Proc. Natl. Acad. Sci. U. S. A. *115*, 1268–1273.

114. Schneeberger, E.-M., and Breuker, K. (2017). Native Top-Down Mass Spectrometry of TAR RNA in Complexes with a Wild-Type tat Peptide for Binding Site Mapping. Angew. Chemie *129*, 1274–1278.

115. Morrison, L.J., and Brodbelt, J.S. (2015). Charge site assignment in native proteins by ultraviolet photodissociation (UVPD) mass spectrometry. Analyst *166*, 166–176.

116. Niu, S., and Ruotolo, B.T. (2015). Collisional unfolding of multiprotein complexes reveals cooperative stabilization upon ligand binding. Protein Sci. *24*, 1272–1281.

117. Aquilina, J.A. (2009). The major toxin from the Australian Common Brown Snake is a hexamer with unusual gas-phase dissociation properties. Proteins Struct. Funct. Bioinforma. *75*, 478–485.

118. Heck, A.J.R., and Van Den Heuvel, R.H.H. (2004). Investigation of intact protein complexes by mass spectrometry. Mass Spectrom. Rev. *23*, 368–389.

119. Popa, V., Trecroce, D.A., McAllister, R.G., and Konermann, L. (2016). Collision-Induced Dissociation of Electrosprayed Protein Complexes: An All-Atom Molecular Dynamics Model with Mobile Protons. J. Phys. Chem. B *8*, 5114–5124.

120. Jones, C.M., Beardsley, R.L., Galhena, A.S., Dagan, S., *et al.* (2006). Symmetrical gas-phase dissociation of noncovalent protein complexes via surface collisions. J. Am. Chem. Soc. *128*, 15044–15045.

121. Morrison, L.J., and Brodbelt, J.S. (2016). 193 nm ultraviolet photodissociation mass spectrometry of tetrameric protein complexes provides insight into quaternary and secondary protein topology. J. Am. Chem. Soc. *138*, 10849–10859.

122. Quintyn, R.S., Zhou, M., Yan, J., and Wysocki, V.H. (2015). Surface-Induced Dissociation Mass Spectra as a Tool for Distinguishing Different Structural Forms of Gas-Phase Multimeric Protein Complexes. Anal. Chem. *87*, 11879–11886.

123. Rosati, S., Rose, R.J., Thompson, N.J., Duijn, E. Van, *et al.* (2012). Exploring an Orbitrap Analyzer for the Characterization of Intact Antibodies by Native Mass Spectrometry. Angew. Chemie - Int. Ed. *51*, 12992–12996.

124. Fort, K.L., van de Waterbeemd, M., Boll, D., Reinhardt-Szyba, M., *et al.* (2017). Expanding the structural analysis capabilities on an Orbitrap-based mass spectrometer for large macromolecular complexes. Analyst *143*, 100–105.

125. Yang, Y., Liu, F., Franc, V., Halim, L.A., *et al.* (2016). Hybrid mass spectrometry approaches in glycoprotein analysis and their usage in scoring biosimilarity. Nat. Commun. *7*, 1–10.

126. Botzanowski, T., Erb, S., Hernandez-Alba, O., Ehkirch, A., *et al.* (2017). Insights from native mass spectrometry approaches for top- and middle- level characterization of site-specific antibody-drug conjugates. MAbs *9*, 801–811.

127. Ben-Nissan, G., and Sharon, M. (2018). The application of ion-mobility mass spectrometry for structure/function investigation of protein complexes. Curr. Opin. Chem. Biol. *42*, 25–33.

128. Zhang, H., Cui, W., and Gross, M.L. (2014). Mass spectrometry for the biophysical characterization of therapeutic monoclonal antibodies. FEBS Lett. *588*, 308–317.

129. Sinz, A. (2018). Cross-Linking/Mass Spectrometry for Studying Protein Structures and Protein–Protein Interactions: Where Are We Now and Where Should We Go from Here? Angew. Chemie - Int. Ed. *57*, 6390–6396.

130. Beveridge, R., Migas, L.G., Payne, K.A.P., Scrutton, N.S., *et al.* (2016). Mass spectrometry locates local and allosteric conformational changes that occur on cofactor binding. Nat. Commun. *7*, 12163.

131. Pacholarz, K.J., Burnley, R.J., Jowitt, T.A., Ordsmith, V., *et al.* (2017). Hybrid Mass Spectrometry Approaches to Determine How L-Histidine Feedback Regulates the Enzyme MtATP-Phosphoribosyltransferase. Structure *25*, 730–738.

132. Snijder, J., Schuller, J.M., Wiegard, A., Lössl, P., *et al.* (2017). Structures of the cyanobacterial circadian oscillator frozen in a fully assembled state. Science (80-.). *355*, 1181–1184.

133. Cesnik, A.J., Shortreed, M.R., Schaffer, L. V, Knoener, R.A., et al. (2017). Proteoform Suite: Software for Constructing, Quantifying, and Visualizing Proteoform Families. J. Proteome Res. 17, 568-578.

134. Zamdborg, L., LeDuc, R.D., Glowacz, K.J., Kim, Y.-B., et al. (2007). ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. Nucleic Acids Res. 35, W701-6.

135. Park, J., Piehowski, P.D., Wilkins, C., Zhou, M., et al. (2017). Informed-Proteomics: open-source software package for top-down proteomics. Nat. Methods 14, 909-914.

136. Cai, W., Guner, H., Gregorich, Z.R., Chen, A.J., et al. (2016). MASH Suite Pro: A Comprehensive Software Tool for Top-Down Proteomics. Mol. Cell. Proteomics 15, 703–14.

137. Kou, Q., Wu, S., and Liu, X. (2018). Systematic Evaluation of Protein Sequence Filtering Algorithms for Proteoform Identification Using Top-Down Mass Spectrometry. Proteomics 18, 1–12.

138. Liu, X., Hengel, S., Wu, S., Tolic, N., et al. (2013). Identification of ultramodified proteins using top-down tandem mass spectra. J. Proteome Res. 12, 5830–8.

139. Kou, Q., Wu, S., Tolic, N., Paša-Tolic, L., et al. (2016). A mass graph-based approach for the identification of modified proteoforms using top-down tandem mass spectra. Bioinformatics 33, 1309-1316.

140. Fellers, R.T., Greer, J.B., Early, B.P., Yu, X., et al. (2014). ProSight Lite : Graphical Software to Analyze Top - Down Mass Spectrometry Data. Proteomics 15, 1235–8.

141. Kou, Q., Xun, L., and Liu, X. (2016). TopPIC: A software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics 32, 3495–3497.

142. LeDuc, R.D., Fellers, R.T., Early, B.P., Greer, J.B., et al. (2019). Accurate Estimation of Context-Dependent False Discovery Rates in Top-Down Proteomics. Mol. Cell. Proteomics 18, 796–805.

143. Leduc, R.D., Fellers, R.T., Early, B.P., Greer, J.B., et al. (2014). The C-Score: A bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. J. Proteome Res. 13, 3231–3240.

144. Kou, Q., Wang, Z., Lubeckyj, R.A., Wu, S., et al. (2019). A Markov Chain Monte Carlo Method for Estimating the Statistical Significance of Proteoform Identifications by Top-Down Mass Spectrometry. J. Proteome Res. 18, 878–889.

PART I

GAS-PHASE ACTIVATION REVEALS STRUCTURAL FEATURES OF PROTEIN ASSEBMLIES

CHAPTER 2

SYMMETRY OF CHARGE PARTITIONING IN COLLISIONAL AND UV PHOTON-INDUCED DISSOCIATION OF PROTEIN ASSEMBLIES

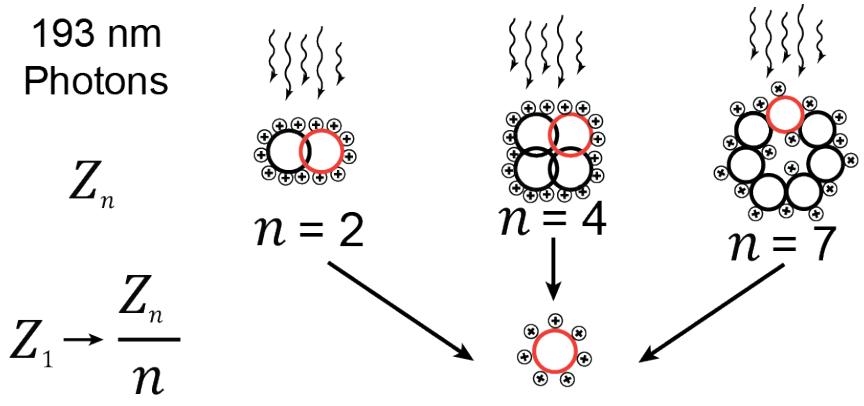
Sem Tamara[†,‡], Andrey Dyachenko[†,‡], Kyle L. Fort[†], Alexander Makarov[†,#], Richard A. Scheltema[†] and Albert J.R. Heck[†]

† Utrecht University, Utrecht, The Netherlands
# Thermo Fisher Scientific, Bremen, Germany
‡ Contributed equally

J. Am. Chem. Soc. 2016, 138 (34) 10860-10868
DOI: 10.1021/jacs.6b05147

ABSTRACT

Tandem mass spectrometry can provide structural information on intact protein assemblies, generating mass fingerprints indicative of the stoichiometry and quaternary arrangement of the subunits. However, in such experiments, collision- induced dissociation yields restricted information due to simultaneous subunit unfolding, charge rearrangement, and subsequent ejection of a highly charged unfolded single subunit. Alternative fragmentation strategies can potentially overcome this and supply a deeper level of structural detail. Here, we implemented ultraviolet photodissociation (UVPD) on an Orbitrap mass spectrometer optimized for native MS and benchmark its performance to HCD fragmentation using various protein oligomers. We investigated dimeric β-lactoglobulin, dimeric superoxide dismutase, dimeric and tetrameric concanavalin A, and heptameric GroES and Gp31; ranging in molecular weight from 32 to 102 kDa. We find that, for the investigated systems, UVPD produces more symmetric charge partitioning than HCD. While HCD spectra show sporadic fragmentation over the full protein backbone sequence of the subunits with a bias toward fragmenting labile bonds, UVPD spectra provided higher sequence coverage. Taken together, we conclude that UVPD is a strong addition to the toolbox of fragmentation methods for top-down proteomics experiments, especially for native protein assemblies.



INTRODUCTION

Most biological processes in cells involve in time and space regulated non-covalent interactions between proteins. These interactions functionalize molecular machines, providing more complex behavior than the sum of the individual parts would allow.[1,2] An estimated 80% of all expressed proteins engage in such interactions, which for the human proteome with its 20,000 genes may result into 650,000 protein-protein interactions within the cell.[3,4] This rich and highly dynamic level of complexity in cellular processes can be investigated with high throughput methods like affinity-purification combined with mass spectrometry (AP-MS), producing large catalogues of interacting proteins[5–7] in short timeframes.[8] Although very useful to provide a snapshot of active protein interactions, these methods identify the involved proteins with varying degrees of confidence and provide very global insight into the detected interactions. A higher level of detail requires use of different structure based methods such as X-ray crystallography, electron microscopy and NMR. However, these methods exhibit inherent restrictions as they typically require high amounts of analyte and have low throughput. Thus, alternative methods would be beneficial.

Native mass spectrometry is a complementary technique that allows for the investigation of proteins and their interaction partners in their native state.[9–11] With this technique, it is possible to extract information with regard to the quaternary structure and subunit stoichiometry of the assembly,[9] as well as, sequence information from individual subunits.[12] While stoichiometry can often be determined by the intact mass spectrum through the use of high-resolution mass spectrometers, determination of quaternary structure requires additional techniques like MS fragmentation methods. Subjecting the protein assembly ion to increasing collisional energies leads to the sequential ejection of individual subunits, providing insight into the quaternary structure of the assembly.[13] From the available data so far, it has been shown that collision based fragmentation methods almost exclusively eject a protein monomer regardless of the protein assembly size, structure or subunit organization.[14–19] Moreover, the ejected monomer is visible in the mass spectrum at high charge states, indicative of protein monomer unfolding and charge relocalization prior to its ejection,[20] which leads to a loss of information about the structure of the precursor protein assembly.[21,22] It has been postulated that the monomer unfolding occurs due to an increase in internal energy followed by transfer of several protons to relieve coulombic repulsion, which in turn promotes further unfolding and finally monomer ejection.[15,23,24] The remaining (n − 1)-meric assembly is present in the mass spectrum with the remainder of the charges, leading to the assumption that it remains in a more folded state.[15] Overall, this results in charge partitioning that is asymmetric with respect to the mass of each product, as the charges are distributed roughly proportional to the surface area.[23,25] This mechanism of fragmentation limits its usefulness in determining structural information as the unfolded monomer does not retain significant amounts of topological information.[23] Additionally, the highly charged nature of the monomer can limit our ability to resolve structural heterogeneity which may be present within the primary sequence of the protein and, in some cases, drive the ions outside the mass and/or transmission range of the mass spec-

trometer due to the limited *m/z* window of the ion optics.[26] Previous investigations have indicated that the asymmetric dissociation process depends on several factors including the charge state of the molecule, but also on the timescales involved in the fragmentation process.[17] This suggests that alternative and faster means of activating protein assemblies may be beneficial, especially when they open up dissociation channels other than the ejection of a single, unfolded monomer.

An interesting alternative fragmentation method explored is surface induced dissociation (SID). Noteworthy, SID was shown to produce preferably symmetric charge partitioning for protein assemblies, which was attributed to its more prompt, high-energy fragmentation mechanism.[25,27,28] Another newcomer in the field of protein assembly dissociation techniques is ultraviolet photodissociation (UVPD), which utilizes the natural chromophores present in the backbone of peptides and proteins to absorb highly energetic photons (λ = 193 nm) emitted from a laser.[29–35] The technique may provide prompt fragmentation at an activation time scale close to the energy deposition in SID,[36] and could thus potentially prove to be a beneficial method for probing of assembly composition and topology. In this study, we report the modification of an Orbitrap-based mass spectrometer with Extended Mass Range (EMR) capabilities to support UVPD. We introduce further optimizations to make our previously reported UVPD strategy work for these large assemblies.[35,37] The new UVPD capabilities of the instrument were applied to the investigation of multimeric protein assemblies in their native state and compared to the dissociation behavior with HCD. Studied systems include dimeric β-lactoglobulin (β-Lac), dimeric Cu, Zn-superoxide dismutase (Cu,Zn-SOD), dimeric and tetrameric concanavalin A (ConA), and heptameric GroES and Gp31 assemblies. These protein assemblies range in mass from 32 to 102 kDa. Each system was subjected to a range of collision energies, which were selected in such a way that the lowest value is the onset of precursor depletion while the highest value completely depletes the precursor. For these systems, HCD fragmentation data, in agreement with literature, shows largely asymmetric charge partitioning and ejection of monomers. However, we also find that the asymmetric charge partitioning of HCD is diminished when structural constraints like disulfide bonds are present.[15,38] In contrast, UVPD experiments, where photons are absorbed by the precursor ions during the fixed short 5 ns pulse of the laser, lead more to the ejection of a compact, low-charge monomer via a high energy deposition pathway, similar to SID. Investigation of these phenomena as a function of system size, degree of stabilizing interactions, and complexity revealed that there is likely a natural limit to UVPD in its utility for symmetric partitioning based on the size of the subunits and the stability of the binding interfaces between the subunits. Additionally, we find that UVPD for all investigated systems outperforms HCD in terms of backbone sequence coverage. Overall, these data demonstrate that UVPD provides a simple, versatile method for the structural analysis of protein assemblies by native mass spectrometry, adding to the toolbox for top-down proteomics.

RESULTS AND DISCUSSION

Implementation of UVPD on the Orbitrap based EMR

For a detailed description of the instrument modifications please refer to the Methods section. In brief, we modified a standard Orbitrap-based Exactive Plus mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) to support the analysis of large protein assemblies in native conditions[39]. The implementation of the UV laser is largely similar to the one described previously.[35] In addition, to compensate for the pressure drop caused by the removal of the electrometer, a CaF2 Teflon-sealed viewport was constructed, which sealed the HCD cell and served as an optical aperture for collimation of the laser beam (Figure 1; inset).

The performance of UVPD on our system was benchmarked on the often used model system ubiquitin, for which our setup produces fragments covering 100% of the sequence and results in significantly better n-terminal backbone coverage when compared to HCD (Supporting Figure S1.1). With these modifications, protein fragments generated for native state proteins with precursor masses of up to at least 100 kDa can be transmitted and analyzed (Supporting Figure S1.2). Following this successful benchmark testing, we subsequently set out to investigate the performance of HCD and UVPD on native non-covalently bound protein assemblies, starting from dimers to higher oligomers, as described in the following paragraphs.

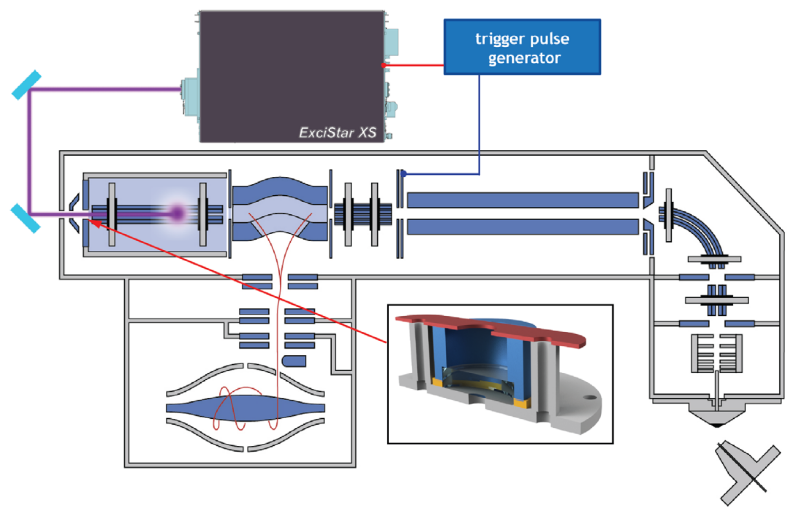


Figure 1 | Implementation of UVPD on the modified Orbitrap-based Exactive Plus mass spectrometer. The laser is guided into the HCD cell through a viewport fitted to the back-end of the HCD cell. Triggering of the laser is done through read-out of the split lens.

Dimeric protein assemblies

As an initial system to investigate the difference between HCD and UVPD for dissociation of protein assemblies, the lectin Concanavalin A (ConA) was selected. This lectin naturally occurs as both a homodimer (51 kDa) and a homotetramer (102

kDa) and has previously been investigated by tandem MS using CID[40], SID[41], and electron transfer dissociation (ETD)[42]. The dimer was investigated first as it offers the least non-covalent intermolecular interactions between the subunits making it the simpler system. During tandem MS analysis, the 15+ charge state (z = 15+) of the homodimer was isolated and subjected to either HCD or UVPD fragmentation, at a variety of collision energies and laser energies with values of 10-140 V and 0.5 – 2.5 mJ pulse^{-1}, respectively.

When subjected to low collision energy HCD (10 V), the tandem mass spectrum (Figure 2a) is dominated by the remaining precursor and shows monomeric dissociation products with z = 8+ and 7+. Interestingly, these products are consistent with symmetric charge partitioning of the homodimer, indicating that the initially accessed fragmentation pathway is the disruption of the noncovalent interactions between the two subunits. When the collision energy is increased to 100 V, the tandem mass spectrum shows the emergence of two distinct charge state envelopes (Figure 2a, bottom; red and orange dashed lines) and several covalent fragments present at m/z values of < 2000 Th. The first envelope, denoted by the orange dashed line, comprises of the monomeric dissociation products at z = 8+ and 7+, which are consistent with the symmetric dissociation products observed at low collision energies. This envelope also shows the presence of the monomeric dissociation products at z = 9+ and 6+, suggesting that charge relocalization and thus asymmetric dissociation is starting to occur. This is further supported by the presence of a second charge state envelope (denoted by the red dashed line), which is centered on the monomeric dissociation product at z = 11+ and is consistent with the full onset of asymmetric dissociation and monomer unfolding. This transition from symmetric to asymmetric dissociation is clearly visible by studying the dissociation products as a function of collision energy (Figure 2b). At the lowest collision energy, the symmetric dissociation pathway dominates, as shown by the high intensity of the monomeric dissociation product at z = 8+. At a collision energy of 80 V, the asymmetric dissociation pathway is increasingly accessed as shown by the appearance of the monomeric dissociation product at z = 11+, reaching maximum intensity at 100 V collision energy, and a sharp decrease in intensity of z = 8+. At collision energies higher than 100 V, both the 11+ and 8+ charge states are reduced in relative intensity and there is an increase in sequence coverage, which indicates the onset of covalent bond cleavage. Taken together we hypothesize that the dissociation pathway for dimeric ConA with HCD displays the following energy dependent order. At low energies, the protein assembly undergoes symmetric dissociation. As the collision energy is increased, the protein assembly dissociates in a more asymmetric fashion. Further increase of the collision energy enhances the asymmetric dissociation behavior and finally leads to covalent bond cleavage.

In contrast to HCD, the UVPD fragmentation mass spectra show only symmetric dissociation prior to covalent bond cleavage. At the lowest laser energy investigated, the formation of the monomer dissociation products at z = 8+ and 7+ are dominant (Figure 2c, top). As the laser energy is increased to 2 mJ pulse^{-1}, these monomeric product ions remain the predominant pathway of dissociation, as shown by the stable charge state envelope over the range of energies (green dashed line), and there is an absence of charge states consistent with asymmetric dissociation (Figure 2c, bottom). Moreover, covalent bond cleavages are readily occurring at almost all laser energies, indicating that symmetric dissociation leads directly to backbone fragmentation with high sequence coverage, while the asymmetric dissociation pathway is not accessed at any appreciable amount. At all investigated laser energies, the symmetric dissociation products are the dominant pathway as suggested by an absence of the monomeric dissociation product at z = 11+. Although, as the energy increases, the relative intensity of the monomeric dissociation product at z = 8+ decreases and there is a small increase in sequence coverage after 1 mJ pulse^{-1} (Figure 2d). These data demonstrate that UVPD does not lead to appreciable monomeric unfolding upon fragmentation for this dimeric assembly.

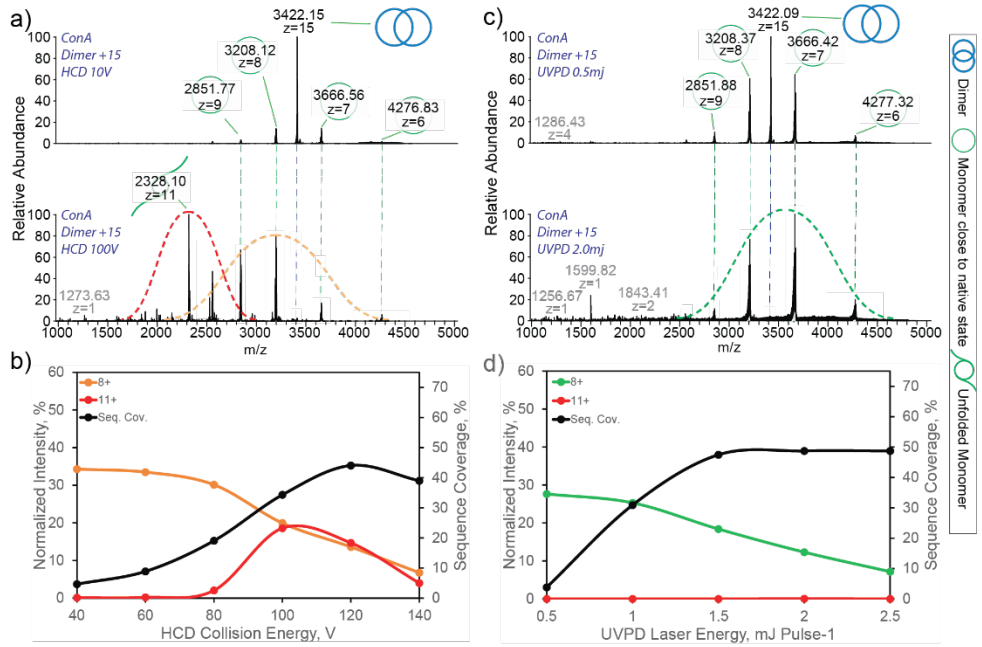


Figure 2 | Annotated spectra representing (a) low and high energy HCD and (c) low and high energy UVPD of dimer ConA (z = 15+). Symbols corresponding to the form of ConA are described on the right side of the figure. Normalized intensities of the z = 8+ and 11+, representations of the symmetric and asymmetric dissociation pathway, respectively, are plotted as a function of (b) collisional energies for HCD and (d) laser energies for UVPD. Sequence coverage (seq. cov.) is provided on the right-hand vertical axis. The dashed lines indicate the different peak envelopes.

For the ConA dimer, UVPD somewhat outperforms HCD in backbone coverage, covering 50% and 45% of the sequence, respectively. The differences in activation, however, do result in differences in the observed fragmentation patterns (Figure S2.1). The observed sporadic cleavages over the full backbone for HCD are likely caused by the preference of collision- induced dissociation techniques toward cleaving the labile bonds first, resulting in more readily detectable fragments associated with those bonds. UVPD on the other hand tends to provide most of the cleavages from the surface-exposed regions, which are known to be both termini in the case of dimeric ConA (PDB reference: dimeric concanavalin A, 1GKB). Such behavior

was previously reported for ETD and electron capture dissociation (ECD),[12] however these methods are hampered, in the analysis of proteins under native conditions, by their strong charge dependence. When combining single HCD and UVPD deconvoluted spectra the backbone sequence coverage reaches 66%, an increase of 32% and 46% as compared to UVPD and HCD alone, respectively (Figure S2.1). Comparison of the UVPD fragmentation patterns between the monomeric and dimeric form provides insights into the binding interface. Based on a previously reported spectroscopy study, the regions stabilized by hydrogen bonds tend to favor proton transfer, which can potentially lead to a lower degree of fragmentation.[43] Here, we indeed find that UVPD of the monomer provided longer N-terminal fragments than for the dimer, highlighting that this region is potentially involved in subunit interaction. The UVPD data of the monomer additionally showed reduced C-terminal coverage, suggesting that this part is surface-exposed upon binding.

The differences in fragmentation pathways between HCD and UVPD for dimeric protein assemblies have been attributed to the unfolding of the monomeric subunit prior to ejection with HCD versus the ejection of the folded monomer with UVPD.[44] As such, it is of interest to explore the differences in fragmentation pathways for an assembly that possesses structural constraints restricting its ability to unfold.[45] The protein assembly β-Lactoglobulin (β-Lac) is present as a homodimer (36 kDa); however, unlike ConA, β-Lac contains 2 intra-subunit disulfide bridges that restricts the conformational flexibility and the ability to unfold.[46] The 13+ charge state of the homodimer was isolated and subjected to HCD and UVPD fragmentation at a variety of collision and laser energies with values of 25–150 V and 0.5–2.5 mJ pulse−1, respectively. At the lowest collision energy (25 V) the spectrum is dominated by monomeric dissociation products at z =7+ and z = 6+, which are consistent with symmetric charge partitioning (Figure 3a, top). At 100 V collision energy, the production of these ion species remains the dominant pathway and the formation of backbone fragmentation products starts to occur (Figure 3a, bottom). Consistent with the onset of backbone fragmentation, the monomeric dissociation product at z = 7+ decreases in relative intensity at energies higher than 100 V (Figure S3.1, top panel). UVPD fragmentation of the β- Lac homodimer at the lowest laser energy of 0.5 mJ pulse−1 shows minor dissociation products, nevertheless, the mono- meric dissociation products at z = 7+ and 6+, indicative of symmetrical charge partitioning, are already visible (Figure 3b, top panel). As the energy is increased to 2 mJ pulse−1, the amount of precursor depletion increases but the symmetrical dissociation products remain dominant (Figure 3b, bottom panel) and backbone fragments below 2000 m/z start to appear. For variable laser energy studies, the symmetric dissociation pathway remains dominant at all investigated laser energies, as shown by the minor decrease in normalized intensity of the monomeric dissociation product at z =7+ (Figure S3.1, bottom panel).

Comparison of covalent bond cleavages generated by HCD and UVPD fragmentation show that for both methods there are very few fragments within the region demarked by the disulfide bridges on the monomer and the majority of assigned fragments correspond to the unrestricted N-terminal region of the protein (Figure 3c). Within this N-terminal region, UVPD produces greater sequence coverage than that produced by HCD. Additionally, UVPD shows some fragment ions within the

disulfide protected region, indicating that the disulfide bond is cleaved during laser irradiation, consistent with previously reported results by O'Brien et al.[44] and subsequent covalent bond cleavages occur. This process is largely absent for HCD fragmentation. Collectively, the fragmentation data for the β-Lac homodimer generated with HCD and UVPD appear largely similar in terms of symmetric versus asymmetric dissociation, indicating that the structural rigidity supplied by the disulfide bonds limits the extent of monomeric unfolding that occurs with HCD. Similar behavior was also observed for the superoxide dismutase (Cu–Zn– SOD) dimer, which also contains a disulfide bridge offering it structural rigidity (Figure S3.2).

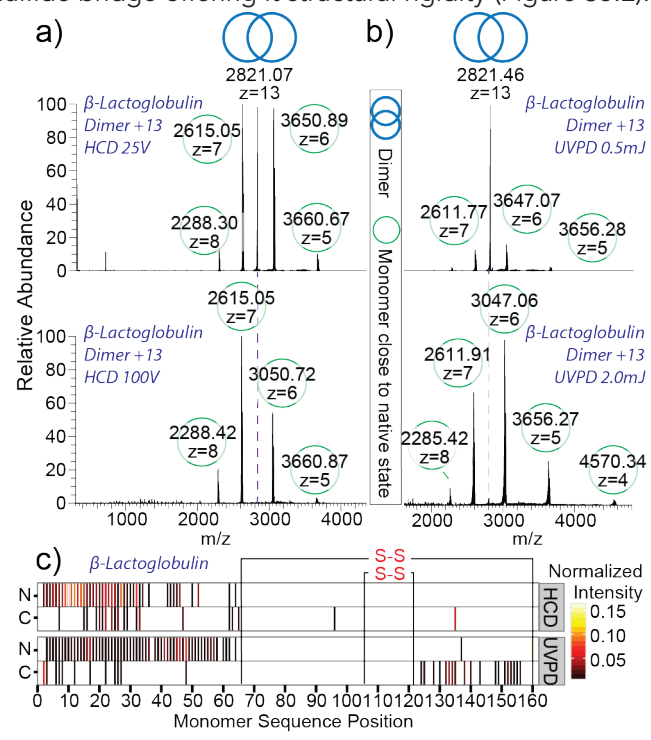


Figure 3 | (a) HCD and (b) UVPD of dimeric β-Lactoglobulin are represented by spectra recorded at low (top) and high (bottom) collisional and laser energies. Fragmentation heat maps (c) display cleavage positions produced via HCD and UVPD as well as normalized intensities of the respective fragments.

Tetrameric protein assemblies

The tetrameric assembly of ConA (102 kDa) provides enhanced stability between the subunits owing to increased numbers of noncovalent intermolecular interactions, potentially resulting in differences between dissociation with UVPD and HCD as compared to the homodimer. When tetrameric ConA at z = 21+ is analyzed with UVPD at 3 mJ pulse−1, the resulting mass spectrum shows two charge envelopes (Figure 4a). The first envelope shows monomeric dissociation products ranging from z =6+to 13+, while the second envelope corresponds to the complementary trimeric dissociation products ranging from z = 8+ to 12+. These data indicate that the preferred pathway of dissociation of the tetramer is the ejection of a monomeric

subunit as opposed to formation of two dimers. Moreover, the range of charge states for the ejected monomer suggests that a combination of symmetric (z = 6+ and 7+) dissociation, as well as, asymmetric dissociation (z = 13+, 12+, and 11+) occurs and that structural rearrangement may potentially also occur with UVPD. In comparison, the HCD generated mass spectrum contains more highly charged monomeric dissociation products (Figure 4b). It is of interest to note that the HCD mass spectrum shows no monomeric dissociation products at charge states that correspond to the symmetric dissociation pathway. Additionally, HCD appears to produce a bimodal charge state envelope for the monomer (red and orange dashed lines). The elevated charge states in the second distribution suggest that an additional structural transition occurs leading to a likely less compact structural gas-phase conformation of the monomer.[47] In contrast, UVPD results in an envelope with a smaller average charge, suggesting a more compact gas-phase conformation. The fragmentation map for tetrameric ConA shows few N- terminal fragments (Figure S4.1), potentially explained by additional stabilization by the Ca2+ and Mn2+ bound at the N- terminus.[48]

Heptameric protein assemblies

As a next step in our investigation we analyzed the GroES heptamer, which is a molecular co-chaperonin found for instance in *E. coli*. In complex with the chaperonin GroEL it acts as a macro- molecular machine whose main function is to assist the correct folding of the proteins in the cell.[49,50] Under physiological conditions the 10.4 kDa monomers of GroES assemble into stable ring-shaped heptamers with a molecular weight of 73 kDa.[51] The z = 18+ charge state of heptameric GroES was isolated and subjected to fragmentation at a range of collision and laser energies; for HCD: 20−200 V, and for UVPD: 0.5−4 mJ pulse−1. At low energies, both photon-induced activation and collisional activation resulted in ejection of the monomer with charge states ranging from z = 4+ to 8+ displaying a remarkable bimodal charge distribution (Figure S5.1). The bimodal distribution, observed for the heptameric GroES dissociation spectra is similar to that observed for the tetrameric ConA HCD spectra, which suggests a gas-phase conformational change of the ejected monomer. At higher laser energies the photon-induced activation prompted the ejection of a low- charged monomer (z = 3+) that was never observed in the HCD spectra (Figure 5a). This represents a unique UVPD dissociation pathway that is consistent with symmetric charge partitioning upon dissociation. The subunits that were ejected following this pathway were likely to partly retain their tertiary structure, indicative of the fast deposition of a large amount of energy into the ion.[25,52]

Gp31 is a bacteriophage T4 structural homologue of GroES, which following infection of E. coli competes with GroES for binding to GroEL to favor the folding of the bacteriophage proteins.[53] The three-dimensional structure of Gp31 closely resembles that of the GroES with slightly larger subunits and thus a higher molecular weight of the intact heptamer of 84 kDa.[54] We found that the Gp31 heptamer exhibits lower stability compared to GroES both in solution and in the gas phase. Both UVPD and HCD activation of the isolated Gp31 heptamer at z = 21+ leads to ejection of a monomer displaying asymmetric charge partitioning (Figure 5b). However, UVPD resulted in a lower average charge, as the highly charged mono-

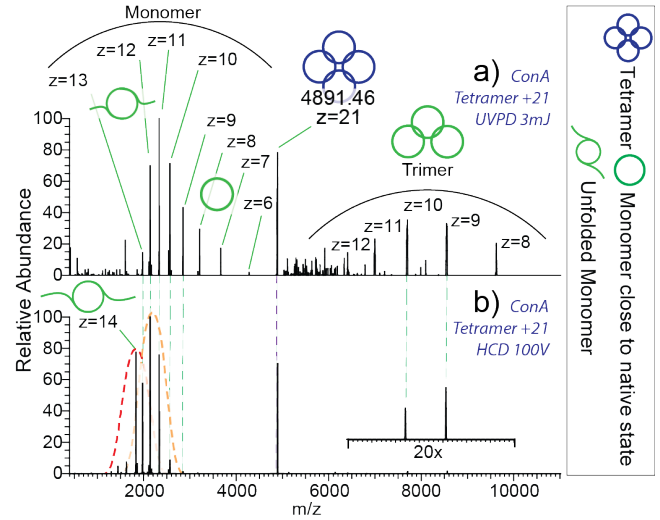


Figure 4 | (a) UVPD spectrum represent partitioning of tetrameric ConA of z = 21+ into monomers of broad charge range and complementary trimers. (b) HCD spectrum is dominated by highly charged monomers. The dotted lines indicate the peak envelopes.

mer dissociation products observed in the HCD spectra (z = +9, + 8) are absent in the UVPD spectra. This is indicative of a relatively more compact state of the subunits dissociated via UVPD. Interestingly, due to the weak intersubunit interactions ejection of a compact monomer from the Gp31 complex becomes energetically more favorable relative to the subunit unfolding. For this co-chaperone system both HCD and UVPD display symmetrical charge partitioning pathway, producing z = 3+ monomers from the z = 21+ heptameric precursor. This highlights the role of intersubunit interactions in the interplay between the monomer ejection and unfolding upon collision- and photon-induced activation, but also reveals that dissociation pathways of very alike systems (GroEL and Gp31 heptamers) may be distinct, and indicative of the biochemical properties of their native precursors. At the higher energies, UVPD of Gp31 resulted in improved sequence coverage compared to HCD fragmentation. As described above, the average charge states for UVPD are lower, indicating that this method is capable of retaining a higher degree of structural stability for weakly interacting subunits, and generate covalent fragments from a more compact state of the molecule (Figure S6.1).

Comparison of experimental charge partitioning with theoretical predictions

To describe the charge partitioning upon assembly dissociation, a number of models have been proposed that suggest that the charge state of the ejected subunit can serve as a predictor of the degree of its unfolding.[24] Additionally, the ejected subunit and the remaining (n − 1)-mer divide the number of charges roughly proportional to their exposed surface areas,[17] which has been shown to often be the case with SID fragmentation.[25] Thus, if dissociation occurs on a time scale shorter than the time scale of gas-phase conformational rearrangement, the fraction of the

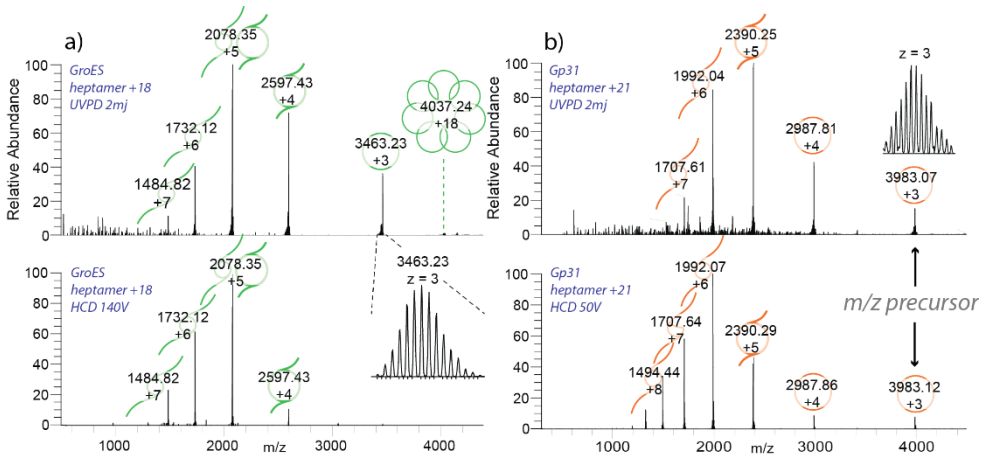


Figure 5 | (a) Dissociation spectra of heptameric GroES (z = +18) activated by UVPD at 2 mJ pulse^{-1} (top) and with HCD at 140 V (bottom). (b) Dissociation spectra of gp31 heptamer (z = +21) activated by UVPD at 2 mJ pulse^{-1} (top) and with HCD at 50 V.

precursor ion charge retained by the ejected monomer can be roughly estimated as $S_{mon}/(S_{mon} + S_{(n-1)mer})$, where S_{mon} and $S_{(n-1)mer}$ are exposed surface areas of the ejected monomer and the remaining (n − 1)-mer, respectively. Our implementation of UVPD allows energy deposition on a time scale close to that of SID,[36] enabling it to more readily achieve symmetric charge partitioning. We indeed find that UVPD produces ejected monomers with charge states more consistent with symmetric charge partitioning as compared to those generated by HCD, which is especially true for the dimers (Figure 6). For assemblies with more than two subunits we however cannot expect fully symmetric charge distributions, as the smaller subunit has a larger surface area relative to its mass than the remaining (n − 1)-mer. Overall, we find that in the case of UVPD the ConA dimer displays fully symmetrical charge partitioning. For the tetramers and the heptamers the average charge of the ejected monomer is higher than expected for the symmetric partitioning, indicating that the energy deposited by the UV photons also leads to subunit unfolding, although to a smaller extent than HCD.

CONCLUSIONS

Here, we report new modifications to an Exactive Plus mass spectrometer with EMR capabilities, enabling it to perform both HCD and UVPD fragmentation on native protein assemblies with molecular weights up to at least 100 kDa. We compare the UVPD performance to the built-in HCD fragmentation capabilities on a set of oligomeric protein assemblies, ranging from dimers to heptamers, and in mass from 32 to 102 kDa. As expected, HCD leads to mostly asymmetric dissociation products, consistent with structural unfolding during the dissociation process. However, UVPD showed more symmetrical dissociation behavior, resembling, in some cases, SID-like behavior.

While UVPD did lead to a higher degree of symmetric dissociation for all systems

investigated, we also show that UVPD depends on both the size of the protein complex as well as the stability of the intersubunit interface. This is reflected in that higher laser energies were required to produce more symmetric dissociation products. However, this intersubunit stability dependence of UVPD may be structurally informative for certain oligomeric protein complexes. For the same oligomeric state of GroES and Gp31, we show that the protein complex with the lower intersubunit stability, Gp31, exhibits more symmetric dissociation products than GroES with UVPD. Furthermore, the ability of UVPD to offer both symmetric dissociation products, while at the same time producing significant backbone coverage makes the technique an attractive one-stop method for simultaneous probing protein assembly structure and stability and subunit sequence. This will provide further confidence in protein identification and ligand/PTM site localization. Collectively, our results demonstrate that UVPD is poised to become a strong addition to the top-down proteomics toolbox as it produces higher subunit backbone coverage, a high percentage of symmetric dissociation products as compared to HCD, and that the partitioning between symmetric and asymmetric pathways may be reflective of the biochemical and biophysical nature of that particular protein complex.

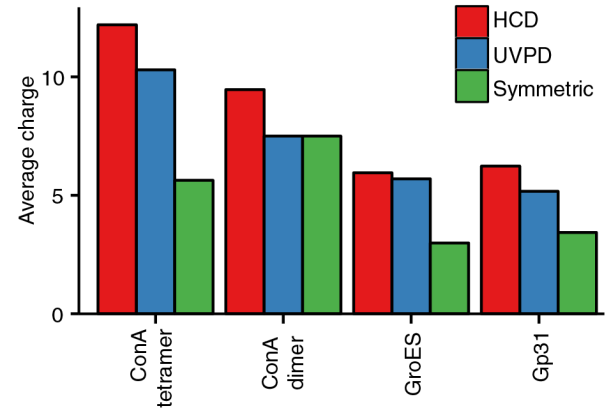


Figure 6 | Average charges of the ejected monomers produced by HCD and UVPD, compared to the expected charge in symmetric dissociation. The symmetric values were calculated from the X-ray structure based on the exposed surface areas following deletion of individual subunits.

METHODS

Instrument Modifications

As previously described, an Exactive Plus mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) was optimized for transmission and detection of ions with m/z up to 50 kTh. [37,39,55] A dedicated gauge was installed controlling the pressure of the collisional gas for more efficient cooling and desolvation of heavy ions. Furthermore, the operating frequencies of the front-end RF guides and the HCD cell were lowered to improve ion transmission and reduce loss of ions during activation at

high energies; a preamplifier with lower high-pass filter cutoff was used to improve transmission of lower frequency image current signals originating from ions with higher m/z values. For isolation we added a standard quadrupole mass filter from a Q Exactive instrument (Thermo Fisher Scientific, Bremen, Germany) with a modified electronic board featuring a decreased resonance frequency of 284 kHz enabling an upper mass-selection limit above 20 kTh.

Introduction of the laser into the mass spectrometer was done as described before.[35] A parallel coherent beam of 193 nm UV photons was generated with an ExciStar XS 500 series excimer laser (Coherent, Santa Clara, CA) filled with an ArF gas mixture. The laser produces 5 ns pulses at a maximum repetition rate of 500 Hz, with functional energies ranging from 0.5 to 5 mJ pulse^{-1} (~10−60 photons per nm^2). The laser beam was guided into the high vacuum region of the mass spectrometer via a periscope assembly, equipped with 45° UV mirrors (Edmund Optics, Barrington, NJ) mounted on micropositioners. The back end flange of the HCD cell was modified by replacing the equipped electrometer with a fused silica vacuum viewport (Kurt J. Lesker Company, Hustings, England). The laser beam was aligned to the longitudinal axis of the HCD cell to maximize the overlap with the trapped ion cloud and avoid irradiation of the ion optics components. Energy transmission through the viewport was measured as 3.5% at the energy range of 2−10 mJ pulse^{-1}. Removal of the electrometer breaks the AGC functionality of the mass spectrometer, which we resolved by optimizing fixed injection time for each protein system. It also removes the vacuum seal between the HCD cell and the high-vacuum chamber. To mitigate the resulting loss of pressure, we designed a custom viewport with a CaF_2 window and Teflon ring (Thorlabs, Newton, NJ) to completely seal the opening. The measured energy transmission through the viewport was 93% at the energy range of 2−10 mJ pulse^{-1}. To synchronize the laser pulses with the presence of the trapped ion cloud inside the HCD cell a purpose-built trigger pulse generator (TPG) was designed. By reading the pulse sequence from the split lens of the mass spectrometer the TPG determines the moment when injection of the ions in the HCD cell is completed. It then generates one (or multiple) 50 µs TTL pulse(s) that trigger(s) the laser emission. The number of pulses as well as the delay between the end of the injection and the trigger pulse can be adjusted.

Studied Proteins

All proteins were purchased from Sigma (Zwijndrecht, The Netherlands) unless otherwise stated; acetonitrile (ACN) was purchased from Biosolve (Valkenswaard, The Netherlands). Gp31 and GroES were recombinantly expressed in E. coli and purified as previously described.[56,57] Native MS analysis was performed on three dimeric protein assemblies: bovine β-lactoglobulin (β-Lac), concanavalin A (ConA), and Cu, Zn-superoxide dismutase (Cu, Zn-SOD); a tetrameric protein assembly: ConA; and two heptameric protein assemblies: Gp31 and GroES. Lyophilized protein assemblies were dissolved to stock concentration of 1 mg/mL in aqueous ammonium acetate (10−300 mM) with pH ranging from 6.5 to 7.5 depending on the most stable conditions reported for each protein assembly. Proteins were desalted in centrifugal filters (Amicon Ultra, Merck, Germany) with 10 kDa molecular weight cutoff. Prior to mass spectrometric analysis stock solutions of protein assemblies were diluted

in aqueous ammonium acetate solution to final monomer concentrations of 5 µM (β-Lac, Cu,Zn-SOD, GroES, and ConA) and 7 µM (Gp31).

Data Acquisition

Electrospray ionization for native MS on the modified Exactive Plus mass spectrometer was performed using in-house pulled borosilicate capillaries coated with gold using a static nanoESI source. Capillary voltage, source fragmentation voltage, front-end transfer parameters, and injection times were optimized for each analyte individually. Nitrogen pressure inside HCD cell was optimized indirectly by monitoring the ultrahigh vacuum (UHV) read-out in the Orbitrap chamber. For all tandem MS experiments the resolution of the Orbitrap mass analyzer was set to 140 000 at 400 m/z. The most abundant charge state of each protein assembly was isolated with a 1−10 Th window for subsequent analysis via UVPD or HCD. All data were collected as a single scan of 500 microscans.

Data Analysis

Native protein fragmentation spectra were deconvoluted with Xtract[58] incorporated into Protein Deconvolution 4.1 (Thermo Fisher Scientific, Bremen, Germany), with the following settings: a signal-to-noise ratio (S/N) threshold of 2, a fit factor of 80%, and a remainder threshold of 25%. The resulting deconvoluted spectra were further processed with the in-house developed intact protein data analysis environment top-down lab (Brunner et al., publication in preparation). Shortly, as a first step, we determined which ions both fragmentation techniques produce for non-covalently bound assemblies by looking at frequently found mass differences to reference points for each amino acid position in the fragmentation spectrum. We calculated these reference points for N-terminal as b − H; and for C-terminal as y +H. After binning the found mass differences to the reference points in 40 ppm bins, the number of occurrences in each bin was calculated. With this unbiased method we found the predominant fragment ions for HCD: y and b (validating the approach); and for UVPD: x, y, a, a+, b, and c (Figure S7.1). After configuring the environment with these fragment ion types, dynamic mass calibration is applied to each spectrum individually based on all annotatable fragment ions for the used fragmentation technique at ±20 ppm. The median of the mass deviation of all annotated fragment ions is consequently used as correction factor. After calibration the global mass cutoff for all spectra is dynamically calculated by estimating the boundaries of the normally distributed mass deviations; the resulting narrow mass filter prevents false positives in assignment. Further statistical analysis of the resulting peptide fragment annotations was performed in R, extended by ggplot2 for data visualization.[59,60] The protein exposed surface areas were calculated using POPS algorithm.[61]

Notes

The authors declare the following competing financial interest(s): A.M. declares competing financial interests as he is an employee of Thermo Fisher Scientific, the manufacturer of the Exactive Plus instrument used in this research.

SUPPLEMENTARY MATERIAL

S1.1 UVPD and HCD spectra recorded on our setup of denatured Ubiquitin

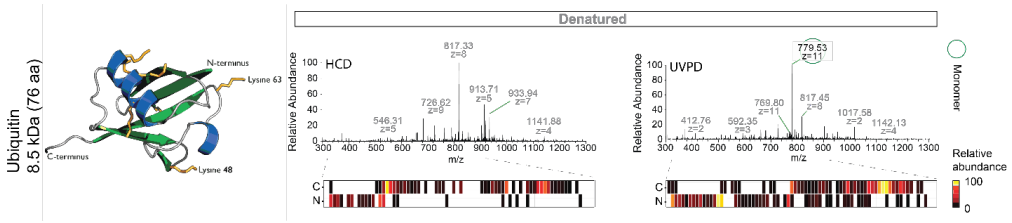


Figure S1.1 | Dissociation of ubiquitin z = 11+ ions under denaturing conditions with HCD (left spectrum) and UVPD (right spectrum). Backbone cleavage maps demonstrate that UVPD (right) provides faster sequencing with N-terminal fragments than HCD (left) using exactly the same 5 µscan (IT = 10 ms) acquisition times. The backbone sequence coverage was in both cases 100%, but UVPD clearly provide a wider coverage of the N-terminal fragment ions.

S1.2 Fragmentation of Ubiquitin with UVPD before and after sealing the electrometer slot

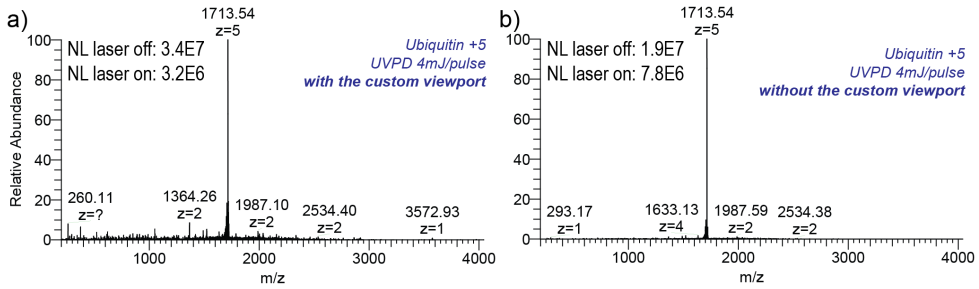


Figure S1.2 | Spectrum of UVPD of ubiquitin z = 5+ demonstrating better transmission of precursor and fragments after sealing the electrometer slot with a custom viewport (a) than before the modification (b) with other parameters unchanged. The fragmentation mass spectrum prior to the viewport installation (right) shows a dearth of signals corresponding to covalent fragments. However, when the modified viewport including the Teflon ring is used (left), the covalent fragments (shown as peaks occurring at < 1600 Th and > 1800 Th) are present, consistent with increased transmission efficiency of the smaller fragments. The "normalized largest" (NL) intensity values, shown for each spectrum, indicate better precursor signal depletion, indicative of better ion cooling and radial confinement thanks to improved gas regime.

S2.1 Unique stretches of the sequence are covered in HCD spectra of dimeric ConA and the UVPD spectra of the ConA monomer and dimer

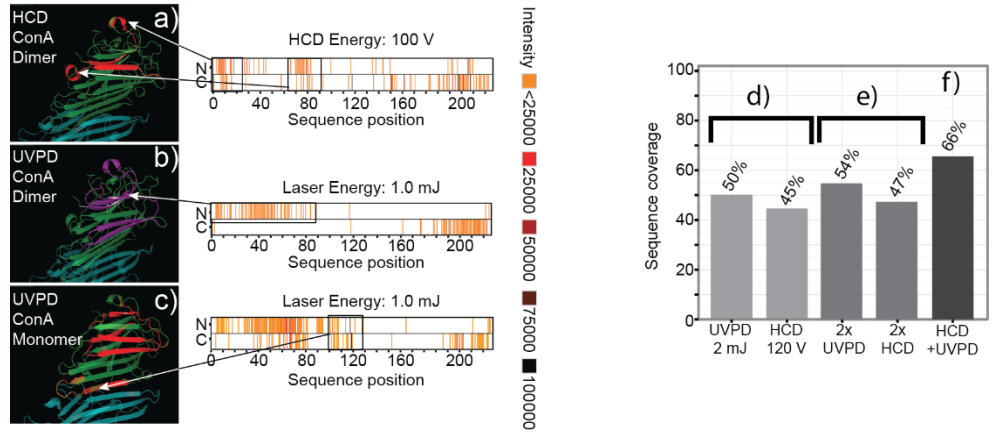


Figure S2.1 | Backbone cleavage maps are plotted for the C- and N-terminal fragments obtained by HCD of z = 15+ dimeric ConA (a), UVPD of z = 15+ dimeric ConA (b), and UVPD of z = 10+ monomeric ConA (c). Specific abundant fragmentation patterns are observed at regions highlighted with colors on the deposited PDB structure 1gkb that is used for reference. In (d) the ConA dimer z = 15+ backbone sequence coverage is shown obtained in the best distinct UVPD (50%) and HCD (45%) experiments. Combining two replicate experiments in (e) for UVPD (54%) and HCD (47%) enhances the coverage. However, as shown in (f) the combination of a single HCD and a single UVPD experiment provides the highest backbone sequence coverage highlighting the complementary nature of both fragmentation methods.

S3.1 Relative intensity profiles for HCD and UVPD of the dimeric β-Lactoglobulin (z = 7+) precursor

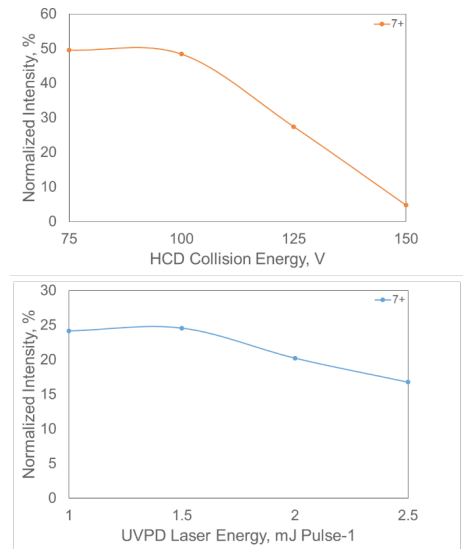


Figure S3.1 | Relative intensity profiles of the β-Lactoglobulin precursor ions at z = 7+ for HCD (top panel) and UVPD (bottom panel). During HCD, the z=7+ symmetric monomeric subunit shows decreased intensity after collision energies of 100 V, indicating the onset of covalent fragmentation. In contrast, the production of this symmetric dissociation product remains the dominant pathway with UVPD as shown by a more consistent production over all laser energies.

S3.2 Symmetrical dissociation of the Cu, Zn-SOD dimer (z=11) using HCD and UVPD

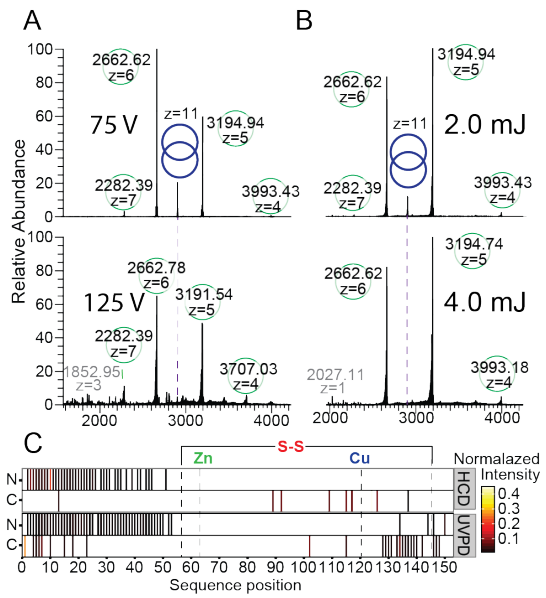


Figure S3.2 | UVPD (a) and HCD (b) spectra of Superoxide Dismutase dimer (z = 11+) ions at energies partially affecting (top panel) and completely depleting (lower panel) the precursor ions. Backbone cleavage map plotted as cleavage sites at monomer sequence positions for Cu,Zn-SOD (c). Color represents fragment intensity; disulfide bridge and metal ligands positions are mapped at the positions as described previously in the literature.

Similarly to the case of β-lactoglobulin, photon-induced and collisional activation of Cu, Zn SOD both result in completely symmetrical charge partitioning. As described previously, the presence of structural constraints such as disulfide bond and metal ion affect the dissociation pathways of Cu, Zn SOD, making the ejection of the compact subunit energetically more favorable process compared to subunit unfolding.

S4.1 UVPD and HCD spectra of tetrameric ConA

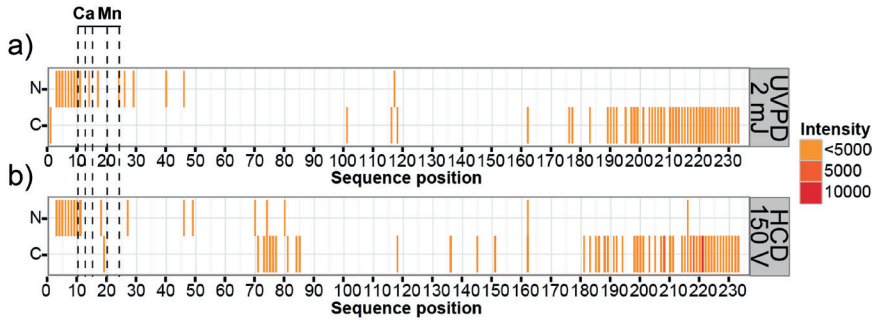


Figure S4.1 | Backbone cleavage maps are plotted for the C- and N-terminal fragments for UVPD (2 mJ pulse^{-1}) (a) and HCD (150 V) (b) of tetrameric ConA (z = 21+) ions. For both UVPD and HCD, most assigned covalent fragments occur at the C-terminal region of the ion and there is an absence of ion fragments corresponding to regions where the Ca^{2+} and Mn^{2+} ions coordinate, the N-terminus. This data suggest that the metal ions provide increased stabilization to the N-terminus, consistent with minimal fragmentation here.

S5.1 UVPD and HCD spectra of heptameric GroES

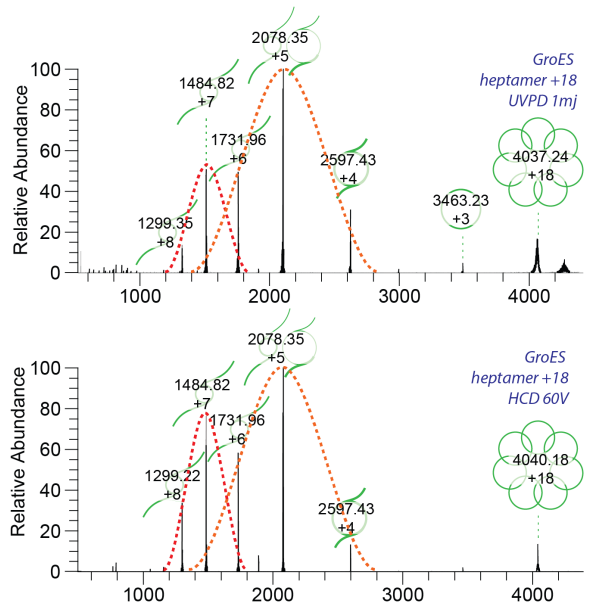


Figure S5.1 | Dissociation spectra of the GroES heptamer (+18) activated by UVPD at 1 mj/pulse (top) and with HCD at 60 V collision energy (bottom).

S6.1 Backbone fragmentation and monomer sequence coverage in HCD and UVPD of Gp31

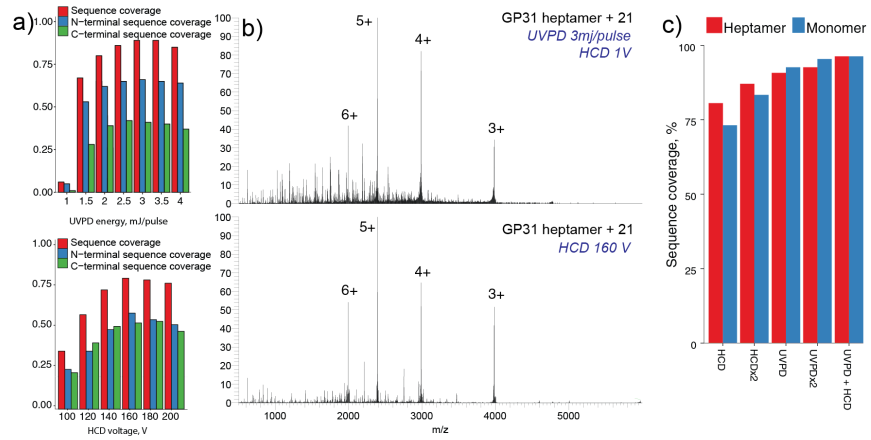


Figure S6.1 | Sequence coverage obtained for the monomer when performing UVPD and HCD on the Gp31 (21+) heptamer. (a) Sequence coverage obtained for Gp31 activated by UVPD ranging from 1 to 4 mJ/pulse (top) and by HCD ranging from 100 to 200 V. Blue and green bars represent sequence coverage calculated from the N- and C-terminal ions, respectively, whereas the red bar indicates the cumulative sequence coverage. (b) Dissociation spectra of the Gp31 heptamer (+21) activated by UVPD at 3 mJ/pulse (top) and with HCD at 160 V (bottom). (c) Gain in sequence coverage obtained from fragmenting the heptaperic (red) and monomeric (blue) Gp31. Post-acquisition combination of the HCD and UVPD data results again in the best overall sequence coverage.

S7.1 Determination of fragment ion types

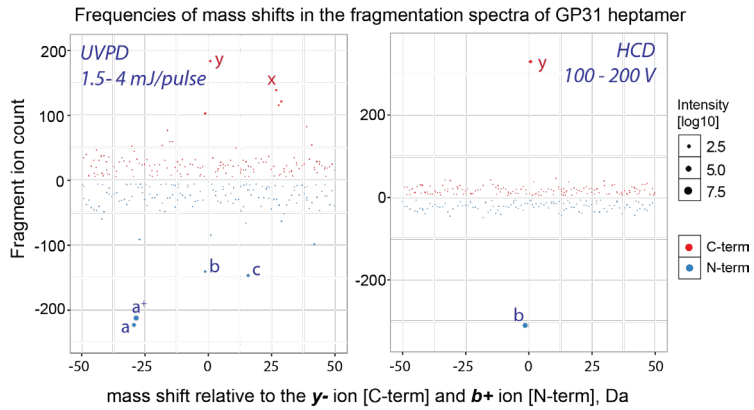


Figure S7.1 | Use of unbiased mass differences for each peak, of the decharged spectral data, in its neighborhood +/- 100 Da, in 40 ppm bins. For the N-terminal ions the difference to the b + H is calculated. For the C-terminal ions the difference to y - H is calculated. The counts for each bin are expressed on the y-axis and the mass difference on the x-axis; the masses with high counts are specific fragment types and the cloud in the middle corresponds to internal fragments and other non-specific fragments. From these calculations we are able to annotate for UVPD: x, y, a, a+, b, and c fragment ions, and for HCD: b and y fragment ions. It reveals clearly that the b and y ions dominate the HCD spectra, while the UVPD data are especially rich in x, y, a, a+, b, and c fragment ions. This fits very well with expected fragments for HCD, for which standardized software exclusively utilizes this pair.

ACKNOWLEDGEMENTS

We thank the group of Jennifer Brodbelt (University of Texas, Austin, TX) for advice in implementing the UV laser in our setup. We also thank our colleagues in the BioMS group, especially Arjan Barendregt for fruitful discussions and technical support. A.J.R.H. and A.M. acknowledge support through the European Union Horizon 2020 program FET-OPEN project MSmed, Project Number 686547. This work forms part of the Roadmap Initiative Proteins@Work (Project Number 184.032.201) financed by The Netherlands Organisation for Scientific Research (NWO).

REFERENCES

1. Alberts, B. (1998). The cell as a collection of protein machines: Preparing the next generation of molecular biologists. Cell *92*, 291–294.
2. Robinson, C. V, Sali, A., and Baumeister, W. (2007). The molecular sociology of the cell. Nature *450*, 973–82.
3. Stumpf, M.P.H., Thorne, T., de Silva, E., Stewart, R., An, H.J., *et al.* (2008). Estimating the size of the human interactome. Proc. Natl. Acad. Sci. U. S. A. *105*, 6959–64.
4. Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., *et al.* (2009). An empirical framework for binary interactome mapping. Nat. Methods *6*, 83–90.
5. Gingras, A.-C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. Nat. Rev. Mol. Cell Biol. *8*, 645–54.
6. Mellacheruvu, D., Wright, Z., Couzens, A.L., Lambert, J.-P., St-Denis, N.A., *et al.* (2013). The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nat. Methods *10*, 730–6.
7. Hein, M.Y., Hubner, N.C., Poser, I., Cox, J., Nagaraj, N., *et al.* (2015). A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. Cell *163*, 712–723.
8. Hosp, F., Scheltema, R.A., Eberl, H.C., Kulak, N.A., Keilhauer, E.C., *et al.* (2015). A Double-Barrel Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) System to Quantify 96 Interactomes per Day. Mol. Cell. Proteomics *14*, 2030–2041.
9. Heck, A.J.R. (2008). Native mass spectrometry: a bridge between interactomics and structural biology. Nat. Methods *5*, 927–933.
10. Snijder, J., and Heck, A.J.R. (2014). Analytical approaches for size and mass analysis of large protein assemblies. Annu. Rev. Anal. Chem. *7*, 43–64.
11. Mehmood, S., Allison, T.M., and Robinson, C. V. (2015). Mass Spectrometry of Protein Complexes: From Origins to Applications. Annu. Rev. Phys. Chem. *66*, 453–474.
12. Zhang, H., Cui, W., Wen, J., Blankenship, R.E., and Gross, M.L. (2010). Native electrospray and electron-capture dissociation in FTICR mass spectrometry provide top-down sequencing of a protein component in an intact protein assembly. J. Am. Soc. Mass Spectrom. *21*, 1966–8.
13. Van den Heuvel, R.H.H., van Duijn, E., Mazon, H., Synowsky, S., Lorenzen, K., *et al.* (2006). Improving the Performance of a Quadrupole Time-of-Flight Instrument for macromolecualr Mass Spectrometry. Anal. Chem. *78*, 7473–7483.
14. Felitsyn, N., Kitova, E.N., and Klassen, J.S. (2001). Thermal decomposition of a gaseous multiprotein complex studied by blackbody infrared radiative dissociation. Investigating the origin of the asymmetric dissociation behavior. Anal. Chem. *73*, 4647–4661.
15. Jurchen, J.C., Garcia, D.E., and Williams, E.R. (2004). Further studies on the origins of asymmetric charge partitioning in protein homodimers. J. Am. Soc. Mass Spectrom. *15*, 1408–1415.
16. Benesch, J.L.P., Aquilina, J.A., Ruotolo, B.T., Sobott, F., and Robinson, C. V. (2006). Tandem Mass Spectrometry Reveals the Quaternary Organization of Macromolecular Assemblies. Chem. Biol. *13*, 597–605.
17. Wanasundara, S.N., and Thachuk, M. (2007). Theoretical Investigations of the Dissociation of Charged Protein Complexes in the Gas Phase. J. Am. Soc. Mass Spectrom. *18*, 2242–2253.
18. Light-Wahl, K.J., Springer, D.L., Winger, B.E., Edmonds, C.G., Camp, D.G., *et al.* (1993). Observation of a small oligonucleotide duplex by electrospray ionization mass spectrometry. J. Am. Chem. Soc. *115*, 803–804.
19. Versluis, C., van der Staaij, A., Stokvis, E., Heck, A.J., and de Craene, B. (2001). Metastable ion formation and disparate charge separation in the gas-phase dissection of protein assemblies studied by orthogonal time-of-flight mass spectrometry. J. Am. Soc. Mass Spectrom. *12*, 329–36.
20. Sinelnikov, I., Kitova, E.N., and Klassen,

J.S. (2007). Influence of Coulombic repulsion on the dissociation pathways and energetics of multiprotein complexes in the gas phase. J. Am. Soc. Mass Spectrom. *18*, 617–31.

21. Sharon, M. (2010). How far can we go with structural mass spectrometry of protein complexes? J. Am. Soc. Mass Spectrom. *21*, 487–500.

22. Benesch, J.L.P., and Robinson, C. V (2006). Mass spectrometry of macromolecular assemblies: preservation and dissociation. Curr. Opin. Struct. Biol. *16*, 245–51.

23. Jurchen, J.C., and Williams, E.R. (2003). Origin of Asymmetric Charge Partitioning in the Dissociation of Gas-Phase Protein Homodimers. J. Am. Chem. Soc, 2817–2826.

24. Sciuto, S. V., Liu, J., and Konermann, L. (2011). An electrostatic charge partitioning model for the dissociation of protein complexes in the gas phase. J. Am. Soc. Mass Spectrom. *22*, 1679–1689.

25. Beardsley, R.L., Jones, C.M., Galhena, A.S., and Wysocki, V.H. (2009). Noncovalent protein tetramers and pentamers with "n" charges yield monomers with n/4 and n/5 charges. Anal. Chem. *81*, 1347–56.

26. Sobott, F., and Robinson, C. V. (2002). Protein complexes gain momentum. Curr. Opin. Struct. Biol. *12*, 729–734.

27. Jones, C.M., Beardsley, R.L., Galhena, A.S., Dagan, S., Cheng, G., *et al.* (2006). Symmetrical gas-phase dissociation of noncovalent protein complexes via surface collisions. J. Am. Chem. Soc. *128*, 15044–15045.

28. Wysocki, V.H., Joyce, K.E., Jones, C.M., and Beardsley, R.L. (2009). Surface-Induced Dissociation of Small Molecules, Peptides, and Non-covalent Protein Complexes. J Am Soc Mass Spectrom *19*, 190–208.

29. Brodbelt, J.S. (2014). Photodissociation mass spectrometry: new tools for characterization of biological molecules. Chem. Soc. Rev. *43*, 2757–83.

30. Halim, M.A., Girod, M., MacAleese, L., Lemoine, J., Antoine, R., *et al*. (2016). 213 nm Ultraviolet Photodissociation on Peptide Anions: Radical-Directed Fragmentation Patterns. J. Am. Soc. Mass Spectrom. *27*, 474–86.

31. Joly, L., Antoine, R., Broyer, M., Dugourd, P., and Lemoine, J. (2007). Specific UV photodissociation of tyrosyl-containing peptides in multistage mass spectrometry. J. Mass Spectrom. *42*, 818–24.

32. Ko, B.J., and Brodbelt, J.S. (2011). Ultraviolet photodissociation of carboxylate-derivatized peptides in a quadrupole ion trap. J. Am. Soc. Mass Spectrom. *22*, 49–56.

33. Madsen, J.A., Kaoud, T.S., Dalby, K.N., and Brodbelt, J.S. (2011). 193-nm photodissociation of singly and multiply charged peptide anions for acidic proteome characterization. Proteomics *11*, 1329–34.

34. Reilly, J.P. (2009). Ultraviolet photofragmentation of biomolecular ions. Mass Spectrom. Rev. *28*, 425–47.

35. Fort, K.L., Dyachenko, A., Potel, C.M., Corradini, E., Marino, F., *et al*. (2016). Implementation of UV-photodissociation on a benchtop Q Exactive mass spectrometer and its application to phospho-proteomics. Anal. Chem. *88*, 2303–2310.

36. Mcluckey, S.A., and Goeringer, D.E. (1997). Slow Heating Methods in Tandem Mass Spectrometry. J. Mass Spectrom. *32*, 461–474.

37. Dyachenko, A., Wang, G., Belov, M.E., Makarov, A., de Jong, R.N., *et al*. (2015). Tandem native mass-spectrometry on antibody-drug conjugates and sub-million Da antibody-antigen protein assemblies on an Orbitrap EMR equipped with a high-mass quadrupole mass selector. Anal. Chem. *87*, 6095–102.

38. Hall, Z., Hernández, H., Marsh, J.A., Teichmann, S.A., and Robinson, C. V. (2013). The role of salt bridges, charge density, and subunit flexibility in determining disassembly routes of protein complexes. Structure *21*, 1325–1337.

39. Rose, R.J., Damoc, E., Denisov, E., Makarov, A., and Heck, A.J.R. (2012). High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. Nat. Methods *9*, 1084–6.

40. Van Dongen, W.D., and Heck, A.J.R. (2000). Binding of selected carbohydrates to apo-concanavalin A studied by electrospray ionization mass spectrometry : Biological mass spectrometry. Analyst *125*, 583–589.

41. Zhou, M., Dagan, S., and Wysocki, V.H. (2013). Impact of charge state on gas-phase behaviors of noncovalent protein complexes in collision induced dissociation and surface induced dissociation. Analyst *138*, 1353–62.

42. Lermyte, F., Williams, J.P., Brown, J.M., Martin, E.M., and Sobott, F. (2015). Extensive Charge Reduction and Dissociation of Intact Protein Complexes Following Electron Transfer on a Quadrupole-Ion Mobility-Time-of-Flight MS. J. Am. Soc. Mass Spectrom. *26*, 1068–1076.

43. Crespo-Otero, R., Mardykov, A., Sanchez-Garcia, E., Sander, W., and Barbatti, M. (2014). Photo-stability of peptide-bond aggregates: N-methylformamide dimers. Phys. Chem. Chem. Phys. *16*, 18877.

44. O'Brien, J.P., Li, W., Zhang, Y., and Brodbelt, J.S. (2014). Characterization of Native Protein Complexes Using Ultraviolet Photodissociation Mass Spectrometry. J. Am. Chem. Soc. *136*, 12920–12928.

45. Hall, Z., Hernández, H., Marsh, J.A., Teichmann, S.A., and Robinson, C.V. (2013). The Role of Salt Bridges, Charge Density, and Subunit Flexibility in Determining Disassembly Routes of Protein Complexes. Structure *21*, 1325–1337.

46. Qin, B.Y., Bewley, M.C., Creamer, L.K., Baker, H.M., Baker, E.N., *et al.* (1998). Structural basis of the Tanford transition of bovine beta-lactoglobulin. Biochemistry *37*, 14014–23.

47. Kükrer, B., Barbu, I.M., Copps, J., Hogan, P., Taylor, S.S., *et al.* (2012). Conformational isomers of calcineurin follow distinct dissociation pathways. J. Am. Soc. Mass Spectrom. *23*, 1534–43.

48. Kaushik, S., Mohanty, D., and Surolia, A. (2009). The role of metal ions in substrate recognition and stability of concanavalin A: a molecular dynamics study. Biophys. J. *96*, 21–34.

49. Hayer-Hartl, M., Bracher, A., and Hartl, F.U. (2016). The GroEL-GroES Chaperonin Machine: A Nano-Cage for Protein Folding. Trends Biochem. Sci. *41*, 62–76.

50. Gruber, R., and Horovitz, A. (2016). Allosteric Mechanisms in Chaperonin Machines. Chem. Rev. *116*, 6588–6606.

51. Xu, Z., Horwich, A.L., and Sigler, P.B. (1997). The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex. Nature *388*, 741–750.

52. Jones, C.M., Beardsley, R.L., Galhena, A.S., Dagan, S., Cheng, G., *et al.* (2006). Symmetrical gas-phase dissociation of noncovalent protein complexes via surface collisions. J. Am. Chem. Soc. *128*, 15044–5.

53. Clare, D.K., Bakkes, P.J., van Heerikhuizen, H., van der Vies, S.M., and Saibil, H.R. (2006). An expanded protein folding cage in the GroEL-gp31 complex. J. Mol. Biol. *358*, 905–11.

54. Hunt, J.F., van der Vies, S.M., Henry, L., and Deisenhofer, J. (1997). Structural Adaptations in the Specialized Bacteriophage T4 Co-Chaperonin Gp31 Expand the Size of the Anfinsen Cage. Cell *90*, 361–371.

55. Belov, M.E., Damoc, E., Denisov, E., Compton, P.D., Horning, S., *et al.* (2013). From protein complexes to subunit backbone fragments: A multi-stage approach to native mass spectrometry. Anal. Chem. *85*, 11163–11173.

56. van der Vies, S.M. (2000). Purification of the Gp31 Co-chaperonin of bacteriophage T4. Methods Mol. Biol. *140*, 51–61.

57. Quaite-Randall, E., and Joachimiak, A. (2000). Purification of GroES from an overproducing E. coli strain. Methods Mol. Biol. *140*, 41–9.

58. Zabrouskov, V., Senko, M.W., Du, Y., Leduc, R.D., and Kelleher, N.L. (2005). New and automated MSn approaches for top-down identification of modified proteins. J. Am. Soc. Mass Spectrom. *16*, 2027–2038.

59. Wickham, H. (2009). ggplot2 - Elegant Graphics for Data Analysis | Hadley Wickham | Springer (New York).

60. Ihaka, R., and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. J. Comput. Graph. Stat. *5*, 1991.

61. Cavallo, L. (2003). POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. Nucleic Acids Res. *31*, 3364–3366.

PART I

GAS-PHASE ACTIVATION REVEALS STRUCTURAL FEATURES OF PROTEIN ASSEBMLIES

3

CHAPTER

DISTINCT STABILITIES OF THE STRUCTURALLY HOMOLOGOUS HEPTAMERIC CO-CHAPERONINS GROES AND GP31

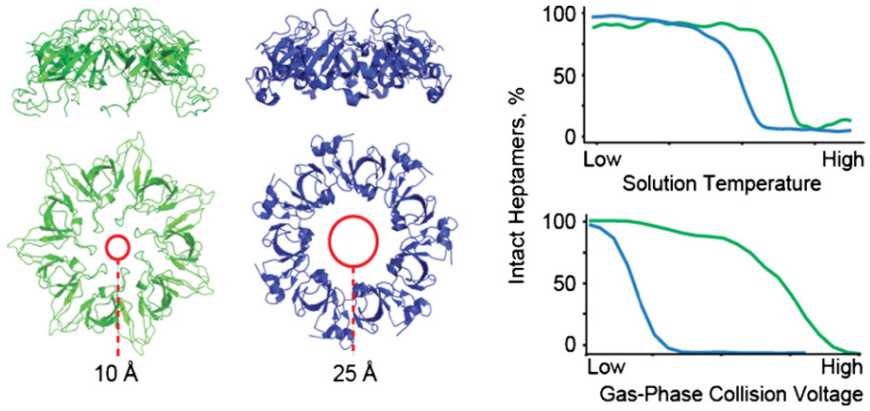
Andrey Dyachenko†, Sem Tamara†, Albert J. R. Heck†

† Utrecht University, Utrecht, The Netherlands

J. Am. Soc. Mass Spectrom. 2018, 30 (1) 7–15
DOI: 10.1007/s13361-018-1910-5

ABSTRACT

The GroES heptamer is the molecular co-chaperonin that partners with the tetradecamer chaperonin GroEL, which assists in the folding of various nonnative polypeptide chains in *E. coli*. Gp31 is a structural and functional analogue of GroES encoded by the bacteriophage T4, becoming highly expressed in T4-infected *E. coli*, taking over the role of GroES, favoring the folding of bacteriophage proteins. Despite being slightly larger, gp31 is quite homologous to GroES in terms of its tertiary and quaternary structure, as well as in its function and mode of interaction with the chaperonin GroEL. Here, we performed a side-by-side comparison of GroES and gp31 heptamer complexes by (ion mobility) tandem mass spectrometry. Surprisingly, we observed quite distinct fragmentation mechanisms for the GroES and gp31 heptamers, whereby GroES displays a unique and unusual bimodal charge distribution in its released monomers. Not only the gas-phase dissociation, but also the gas-phase unfolding of GroES and gp31 were found to be very distinct. We rationalize these observations with the similar discrepancies we observed in the thermal unfolding characteristics and surface contacts within GroES and gp31 in solution. From our data, we propose a model that explains the observed simultaneous dissociation pathways of GroES and the differences between GroES and gp31 gas-phase dissociation and unfolding. We conclude that, although GroES and gp31 exhibit high homology in tertiary and quaternary structure, they are quite distinct in their solution and gas-phase (un)folding characteristics and stability.



INTRODUCTION

The molecular heptameric co-chaperonin GroES forms a cap on top of the *Escherichia coli* molecular tetradecameric GroEL chaperone that assists the folding and refolding of nonnative polypeptide chains (Figure 1). The chaperonin GroEL, GroES and nonnative protein undergo a *binding -> folding -> release* cycle regulated by ATP hydrolysis [1–4]. In this process, GroES acts as a lid that covers the inner cavity of the GroEL chaperonin following the binding of the substrate and ATP inside the Anfinsen cage of GroEL. Formation of GroEL-GroES complex is only possible when GroEL is in the open conformation (ATP bound). When substrate proteins are absent, GroEL primarily binds one GroES molecule. In the presence of substrate, however, GroEL could accommodate one or two GroES molecules, forming either asymmetric bullet-shaped or symmetric US-football-shaped complexes, respectively [5]. These complexes are in a folding-active state, providing nano-environment wherein the substrate is free of nonnative interactions that can lead to aggregation [6,7]. Structurally, GroES is a homo-heptamer with a ring symmetry, constituted of 10.4 kDa monomeric subunits [8] (Figure 1a, b, c; green). Each subunit comprises a rigid beta-barrel structure and an unstructured loop that forms contacts with the apical domain of a corresponding GroEL subunit upon GroEL-GroES complex formation [5,9].

Gp31 is a structural homologue of GroES (Figure 1a, b, c; blue) encoded by the *E. coli* bacteriophage T4, which is essential for the folding of the T4 major capsid protein gp23 [10,11]. Gp31 mimics the action of GroES by competitively binding to GroEL and acting as a co-chaperonin. In vitro, the GroEL-gp31 complex is capable of folding substrates such as citrate synthase and Rubisco with an efficiency similar to that of GroEL-GroES [11–13]. In vivo, the GroEL-gp31 complex can substitute the *E. coli* folding machinery and fold all the *E. coli* proteins that normally rely on the GroEL-GroES system [14]. Therefore, gp31 can functionally substitute GroES, and additionally can assist in folding of the T4 major capsid protein precursor gp23, which at ~56 kDa is slightly larger than the substrates that can be accommodated within the GroEL-GroES Anfinsen cage [15].

Although GroES and gp31 have amino acid sequence identity of only 14%, their tertiary and quaternary structures are similar [16] (Figure 1a, b, c). Like GroES, gp31 in solution forms ring-shaped heptamers composed of identical subunits. The fold of the gp31 monomer also consists of a β-barrel and a mobile loop (Figure 1c). There are, however, a few notable differences. The higher molecular weight (MW) of the gp31 monomer (12 kDa versus 10 kDa for the GroES monomer), together with longer flexible loops (22 residues, GroES has 16 residues) result in GroEL-gp31 complexes having a slightly larger internal chamber volume as compared to GroEL-GroES. The direct measurements performed *via* cryo-electron microscopy (cryo-EM) revealed a size increase of 8%, from 194,000 Å^3 to 210,000 Å^3 [17]. Additionally, gp31 does not contain the roof loops that form the top of the dome in the GroEL-GroES complex (Figure 1a), leaving a wider opening and potentially allowing part of the polypeptide substrate chain to stick out. These subtle differences are generally believed to make the GroEL-gp31 complex capable of folding the larger

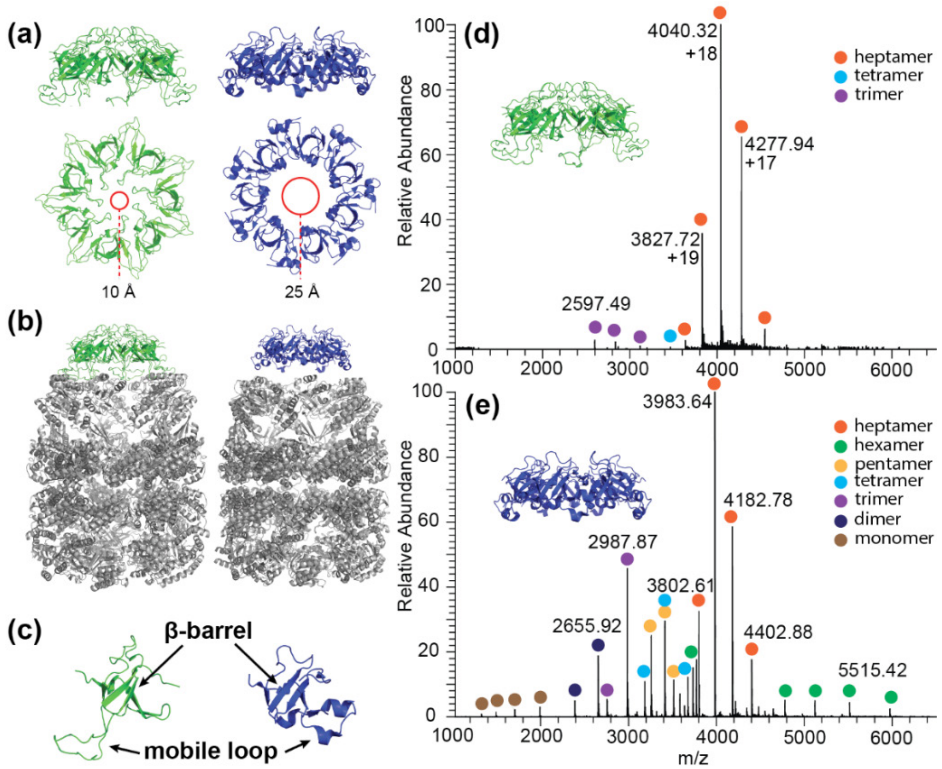


Figure 1 | Structural differences and similarities between GroES and gp31. (a) Side and top views of the GroES (green) and gp31 (blue) crystal structures (PDB accession codes 1AON and 2CGT, respectively). Red circles highlight the difference in the diameter of the central hole in the gp31 and GroES ring when complexed to GroEL; (b) comparison of GroES and gp31 in their complexes with GroEL. 3D representation of the complexes from the side. Figure 1a and 1b are adopted from [17]; (c) similarities of GroES and gp31 monomers; (d) native electrospray ionization mass spectra of GroES and (e) gp31 revealing the preferential heptameric stoichiometry in aqueous ammonium acetate.

gp23 substrate protein.

Along with development of biomolecular mass spectrometry, it has been heavily debated whether biomolecules retained their native structural properties upon transition into the gas phase [18–20]. Primary argument against unaltered gas-phase transition points at the requirement of water molecules to stabilize and maintain the native structure [21]. Additionally, in the early days of the gas-phase analysis researches often observed false positives, i.e. formation of non-specific complexes upon transfer into the mass spectrometer [22]. However, with improvements in instrumentation (e.g. nano-electrospray ionization), as well as advances in the fields of native mass spectrometry (native MS) and ion mobility-mass spectrometry (IM-MS) data supporting unharmed gas-phase transition started to accumulate [23,24]. Furthermore, non-ergodic fragmentation techniques that preferentially preserve non-covalent interactions (e.g. electron-transfer dissociation) revealed similarities between

native-like structures of proteins observed in solution and in the gas-phase [25–28]. Taken together, the mass of evidence indicates that it is possible to relate gas- and solution-phase properties of proteins and protein complexes directly.

GroES and gp31 have been previously examined with gas-phase techniques, both alone [29–31] and in complex with GroEL [32–34]. It has been established that both GroES and gp31 can be maintained in the gas phase as stable heptameric complexes. Here we present a side-by-side comparison of GroES and gp31 gas-phase behavior, revealing substantial differences in their collision induced dissociation (CID) and collision induced unfolding (CIU) behavior. We attribute these differences to the differing structural features of the two complexes in solution and in the gas phase.

RESULTS

Distinct gas-phase collision induced dissociation of the GroES and gp31 heptamers

To initiate investigations on the gas-phase dissociation behavior of GroES and gp31 we first analyzed both complexes by native MS. GroES and gp31 were dissolved in aqueous ammonium acetate buffer adjusted to a pH ~6.8. Samples with a final concentration of the monomer of 10 μM were sprayed from the nano-electrospray (nano-ESI) ion source into Q-Exactive EMR Orbitrap mass spectrometer, modified as described earlier [35,36]. The resulting native mass spectra of GroES and gp31 are shown in the Figure 1d, e. In case of GroES the majority is present in the form of heptamers, with only marginal quantities of the trimer and the tetramer (Figure 1d). Gp31 is present in a substantially higher variety of oligomeric forms (Figure 1e), which may hint at a potentially weaker stability of the gp31 heptamers. We conclude from these data that both GroES and gp31 are most stable as heptamers both in solution and in the gas phase.

To investigate gas-phase stability of the two heptamers, we used a high-mass quadrupole mass filter to isolate individual charge states of each complex ion, and subjected them to collisional activation. It has been previously reported that, within a given charge-state envelope, the lower-charged ions resemble the proteins native state better than higher-charged ions [37]. Hence, we chose the lowest detectable, albeit still relatively abundant, charge states of GroES (+17) and gp31 (+19) for the comparison of their response to collisional activation (Figure 2a, c). The breakdown curve for each heptameric complex is plotted against the collision voltage, which was applied at the entrance of higher-energy collisional dissociation (HCD) cell (Figure 2b). In order to reduce bias introduced by comparison of different charge states we also produced breakdown curves for the same charge state (+19) of GroES and gp31 (Figure S1). Direct comparison of the two heptamers highlights differences in the gas-phase stability of GroES and gp31 upon collisional activation. First, GroES requires substantially higher voltages to dissociate than gp31.

Considering that MW of gp31 is 20% higher than MW of GroES, these results imply a significantly weaker inter-subunit interactions and/or lower stability of gp31 as compared to GroES. Second, the breakdown curve of the GroES follows a clear

double sigmoidal shape, as opposed to a single sigmoidal curve obtained for gp31 (Figure 2b). The double sigmoidal breakdown curve suggests (at least) two co-existing dissociation pathways, whose relative contributions depend on the voltage and hence the kinetic energy of the ion. Alike data on the other charge states of GroES and gp31 (data not shown) corroborated these findings, wherein at low collisional energies gp31 consistently demonstrated unimodal distribution of released monomers, while GroES displayed relatively less abundant bimodal monomer distribution.

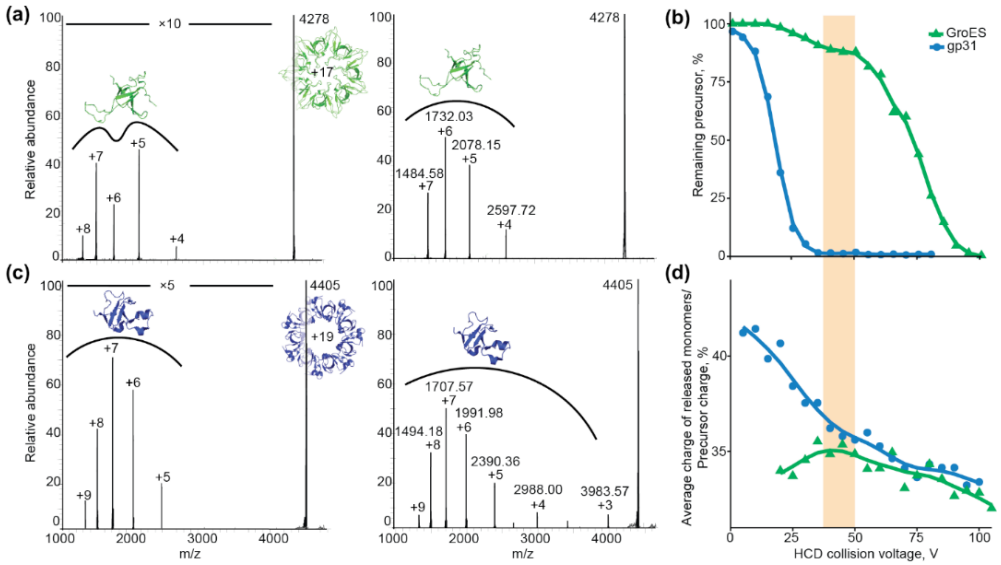


Figure 2 | Tandem mass spectra and breakdown curves of mass-selected GroES and gp31 heptamers. Tandem MS of (a) GroES^{17+}, HCD voltage 40 V (left) and HCD voltage 75 V (right), and (c) gp31^{19+}, HCD voltage 20 V (left) and HCD voltage 25 V (right); ×5 and ×10 are magnification factors for indicated region; (b) breakdown curves of GroES^{17+} and gp31^{19+} against range of applied collision voltages; (d) average charge of monomers released from GroES^{17+} and gp31^{19+} as percentage of precursor charge plotted against the applied HCD voltages.

Tandem MS spectra can offer a greater level of detail about the dissociation process by allowing for a qualitative comparison of the dissociation spectra at varying energy. The tandem MS spectra taken before and after the first plateau of the GroES breakdown curve display strikingly different dissociation patterns. At low voltages, the distribution of dissociation product intensities adopt an atypical bimodal form, with extrema around charges +7 and +5 of the released monomers (Figure 2a; left). At the higher energies, the monomer intensity distribution is more unimodal (Figure 2a; right). Distinct monomer distributions at low collision energies might indicate the co-occurrence of two distinct dissociation mechanisms, with only one becoming prominent at higher collision energies. In contrast, intensities of the monomers released from gp31 follow a unimodal distribution throughout all applied collision voltages, pointing at a more facile and uniform dissociation pathway (Figure 2c).

The charge states of protein ions produced *via* electrospray are known to correlate with the degree of unfolding [38,39], as more extended conformations offer more surface area to accommodate protons. Similarly, more unfolded subunits released from protein complexes upon collision-induced activation in the gas phase harbor more protons than subunits ejected with more compact structures [40]. Hence, the bimodal charge distributions hint at the presence of distinct unfolding states of the GroES monomers ejected from the complex upon collisional activation. This bimodal charge distribution of released monomers is not observed for gp31. To illustrate these differences, we plotted the average charge state of the monomers ejected from GroES and gp31 heptamers as fraction of precursor charge at different HCD collision voltages (Figure 2d).

At low HCD voltages, the normalized average charge of ejected monomers for GroES is lower than that for gp31, indicating that at low collision energies the ejected subunits experience a relatively low degree of unfolding. With elevation of collision energies, distribution of released monomers for GroES start to shift toward higher charge states, indicating that monomer unfolding prior to ejection starts to prevail. At a HCD voltage of ~40 V there is a transition point, after which both GroES and gp31 display similar dissociation behavior. The transition point in the GroES curve coincides with the plateau of the breakdown curve (Figure 2b, d; highlighted with light orange), indicating that from this point on the second dissociation pathway is taking over. Consistent decrease of normalized average charge of released monomers for gp31 (Figure 2d) agrees with the predicted uniform dissociation pathway.

Distinct collision induced unfolding of the GroES and gp31 heptamers revealed by IM-MS

To further investigate the interplay between dissociation and unfolding of heptameric GroES and gp31, we next performed ion mobility (IM-MS) experiments. First, we examined the conformational changes of the heptameric ring complexes upon collisional activation through collision induced unfolding (CIU) [41,42]. As before, we chose the lowest detectable charge states of both GroES (+17) and gp31 (+19) also to enable a direct comparison between their behavior under CIU and CID conditions (Figure 3). Prior to IM-MS measurements, the ions were subjected to activation by collisions with an inert gas (Ar) at varying energies. The degree of unfolding was monitored by following the changes in ion drift times. Unfolding of a protein in the gas phase relates generally to an increase of its geometrical cross-section [43], resulting in a shift in arrival time distribution (ATD) in IM-MS.

The 2D heat maps of GroES and gp31 display several notable differences (Figure 3a, b). Unfolding of GroES is mainly represented with a sharp shift of its ATD (Figure 3a, c; red arrow), followed by less dramatic subsequent second and third unfolding events. Prior to the first major shift of the ATD, the GroES heptamer displays a broadening of the driftogram (Figure 3c; black arrow). This change at low collision energy is accompanied by released monomers with relatively low average charge (Figure 2d), which hints at their partly retained folded state. At this point precursor dissociation occurs to a small extent probably due to strong inter-subunit interactions. The sharp shift of ATD for the GroES heptamer is likely associated with disruption of the inter-subunit interface and opening of the heptameric ring (Figure

3a, c; red arrow). At higher collision voltages, the GroES heptamers are present in two co-existing relatively abundant extended conformations. These smaller conformational changes can be attributed to unfolding of the terminal monomers of the resulting extended structure, prior to their elimination from the heptamer.

In contrast, the gp31 heptamer does not display any sharp shifts in ATD upon activation. Elevation of collision voltages leads to gradual increase of the drift time and significant broadening of the ATD (Figure 3b, d). This behavior can be best explained by gradual unfolding of one or several subunits that are still forming a heptameric ring. We detected all the discussed IM-MS features also for the other relatively abundant charge states of the GroES and gp31 heptamers (Figure S2).

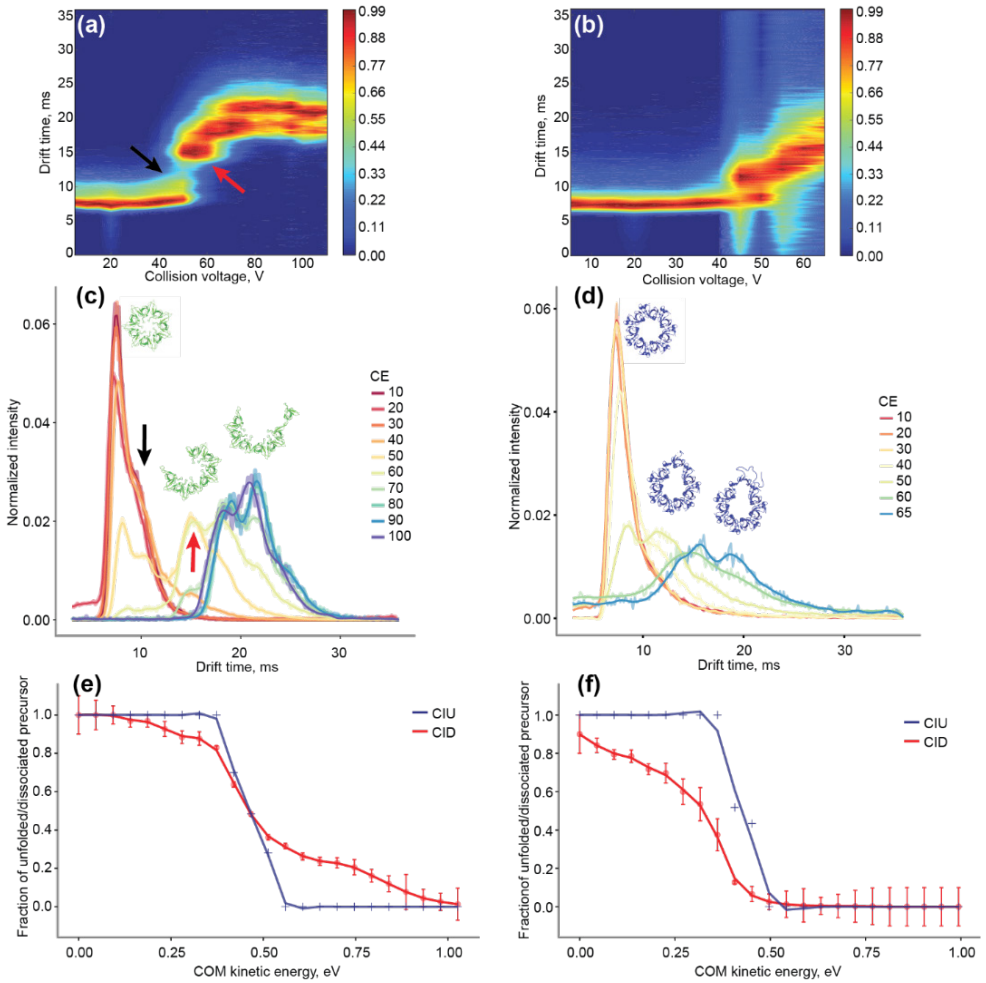


Figure 3 | Collision induced dissociation (CID) and collision induced unfolding (CIU) characteristics of GroES and gp31 heptamers. (a) 2D heatmap representing the unfolding of $GroES^{17+}$; (b) 2D heatmap representing the unfolding of $gp31^{19+}$; (c) arrival time distributions (ATD) of $GroES^{17+}$ ions at various collision energies (CE); (d) ATDs of $gp31^{19+}$ ions at different collision energies; (e) comparison between the CIU and CID of $GroES^{17+}$ ions; (f) comparison between the CIU and CID of $gp31^{19+}$ ions. Structures in Figures 3c and 3d depict predicted unfolding events.

The distinct behavior of the two heptamers upon both CID and CIU becomes clearer upon plotting their unfolding curves (percentage of folded heptamer) along with the breakdown curves (percentage of intact precursor) plotted against the center-of-mass energy of the ion (E_{kin}^{COM}) (Figure 3e, f). In case of GroES there are again two separate regions: the low energy region ($E_{kin}^{COM} < 0.5$ eV) and the higher energy region ($E_{kin}^{COM} > 0.5$ eV) (Figure 3e). At the lower energies the majority of the heptamers have mostly still retained their original compact conformation.

All deposited energy at this point goes into disruption of the inter-subunit interfaces, leading to ejection of relatively folded monomers. However, this occurs only to a fraction of the heptameric precursor. At higher energies, the heptamer adopts a more extended conformation, which likely happens prior to dissociation and is associated with unfolding of the GroES subunit after disruption of the ring structure. The dissociation products in this energy regime are expected to be more unfolded. Interplay of these two dissociating mechanisms for GroES is reflected in the bimodal monomer charge distribution observed at 40 V collision energy (Figure 2a).

Despite our efforts to preserve the intact state of gp31 heptamer, we still observed a certain degree of dissociation even at the lowest energy (Figure 3f). However, the narrow ATD of the precursor heptamer (Figure 3d) assured that the remaining part of the heptameric precursor is still largely in the original intact native-like form. This observation could be explained by relatively weak inter-subunit interfaces, which are likely to be disrupted simultaneously. Interestingly, the high average charge of the released monomers (> 40% of precursor charge) at low collision energy (Figure 2d) indicates that this energy is not only enough to disrupt the binding of the subunit within the complex, but additionally is sufficient to unfold the subunit prior to ejection. With the increase of energy, the precursor dissociates further, while the remaining intact precursor retains a compact state. Broadening of the ATD with elevation of collisional energy is accompanied by significant decrease of the overall precursor intensity. We hypothesize that at this point in parallel with disruption of binding interfaces and unfolding of the ejected single subunits, multiple subunits might undergo unfolding competing for charges, as recently proposed by *in silico* simulations for tetrameric complexes [40]. More stochastic ejection of variously unfolded subunits, rather than release of the most unfolded monomer, would explain presence of low charged monomers down to 3+ (~15% of precursor charge state) along with higher-charged monomers at high collision energy (Figure S1b).

Stability of GroES and gp31 in solution

The striking behavior we observed in the gas phase whereby GroES heptamers dissociate at much more elevated activation energies than the structural homologue gp31 was somewhat unanticipated. Therefore, we set out to test the stability of these two heptamers in solution. We performed a thermal unfolding assay by recording circular dichroism (CD) spectra at variable temperatures. Upon heating, both GroES and gp31 assemblies experience clear unfolding transitions (Figure 4). Additionally, we used the first derivative of the ellipticity *versus* the temperature to confirm the transition point where a maximum rate of the ellipticity change is observed (Figure S3). For both assemblies, the transition represented a sharp in-

crease of the unstructured content. In these assays in solution, GroES displayed a significantly higher stability, undergoing a sharp unfolding transition at 71 °C (in accordance with published data [44]), whereas for gp31 the unfolding transition point was observed at 60 °C (Figure 4; dashed lines). This strongly suggests that also in solution more energy is required to unfold GroES than gp31. Seemingly this order of stability in solution is retained in the gas phase. Next, we sought to further explain this behavior inspecting the inter-subunit interfaces within the GroES and gp31 heptamers.

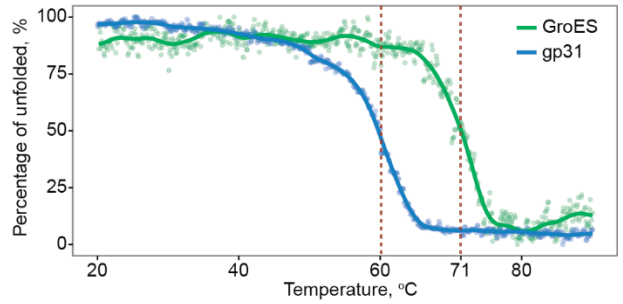


Figure 4 | Thermal unfolding curves of the GroES and gp31 heptamers in solution, as monitored by changes in the circular dichroism (CD) spectra. The extracted melting transitions are 71 and 60 °C, respectively for GroES and gp31.

Inter-subunit binding is more stabilized in GroES than in gp31

It has previously been argued that the chemical nature of the binding interface (solvent accessible surface area, number of salt bridges, amount and strength of hydrogen bonding events, etc.) can be correlated to the dissociation behavior of protein complexes upon activation in the gas phase [45–47]. Zooming in on the binding interfaces of GroES and gp31, using the available crystal structures, revealed several differences (Table 1). The contact area between the subunits was calculated using the "Contact Surface" script, that calculates the solvent accessible surface area (SASA) buried in the interface between the interacting molecules. The contact areas within the GroES heptamer are 7% larger than in gp31. Considering that GroES is about 15% smaller than gp31, this provides a strong indication of tighter inter-subunit interactions within GroES. Next, we determined the number of inter-subunit hydrogen bonds, using a distance cutoff and bond angle restriction of 3.2 Å and 55°, respectively. Again, we observed that the GroES heptamer is stabilized by more hydrogen bonds than gp31 (Table 1).

Table 1 | Summary of the characteristics of the contact areas between the subunits in heptameric GroES and gp31 extracted from the available crystal structures (PDB 1AON (GroES), 1G31 (gp31)).

	Contact area, Å^2	H-bonds	AH-bond distance, Å	Salt bridges
GroES	813 ± 11	7 ± 1	2.7 ± 0.2	R37-E76
gp31	754 ± 8	6 ± 1	2.8 ± 0.2	R77-E44

As in both assemblies we identified only one inter-subunit salt bridge, in both cases between an arginine and a glutamic acid, we argue that this does not contribute to

the difference. Overall, our analysis of the interaction surface between monomers within the GroES and gp31 heptamers indicates that it is more stabilized in GroES, when compared to gp31, likely explaining the higher melting temperature observed for GroES, and the higher resistance to gas-phase dissociation, when compared to gp31.

GroES and gp31 unfold and dissociate in the gas phase *via* distinct mechanisms

Based on the experimental data, we propose the following model to describe the distinctive gas-phase behavior of GroES and gp31. When the internal energy of the ionized heptameric complexes increases because of collisions with the buffer gas, its inter-subunit and intra-subunit noncovalent interactions begin to weaken. For most protein assemblies reported to date this process would lead to the specific unfolding of one of the subunits, followed by its elimination from the assembly [48], with unfolding typically happening at lower energies than dissociation [49]. Distinctively, GroES heptamer dissociation seems to proceed through a combination of two mechanisms. The first mechanism involves the partial unfolding of an individual subunit that leads to disruption from neighboring inter-subunit interfaces and dissociation of the monomer. The partial unfolding can be seen by the shoulder being formed at the right edge of the ATD of GroES (Figure 3c; black arrow). This mechanism is prevalent at lower activation energies. As the activation energy increases, the second mechanism starts to kick in and take over. More energetic collisions begin to destabilize inter-subunit interfaces which leads to disruption of the ring. Further redistribution of energy likely causes destabilization and unfolding of two terminal subunits of the resulting extended structure, which is reflected in two minor ATD shifts that happen directly after the first major shift (Figure 3c). Finally, part of the resulting extended heptamers dissociate into a monomer and a hexamer, completing the second dissociation mechanism. The GroES monomers, produced *via* different dissociation mechanisms, become unfolded to a different degree, which explains the bimodal charge distribution of the released monomer in the low collisional energy regime (Figure 2a). Both the breakdown curves (Figure 2b; green) and combined CIU/CID plots (Figure 3e) suggest that there are two energy zones that can be roughly separated by the point where the precursor ATD undergoes a sharp shift (at E_{kin}^{COM} ~0.48 eV, CID voltage ~50V). Considering the above explanation, it is likely that the first mechanism is prevalent at the lower energies and the second – in the high-energy regime.

Gp31, in contrast, does not change its dissociation behavior over the entire activation energy range. Due to the intrinsic weaker inter-subunit interface and less stable tertiary structure, even a marginal increase of energy leads to destabilization of the fold of individual subunits and facile elimination of a monomer. With increase of collisional energy, more than one subunit undergoes unfolding with subsequent release from the complex. That likely explains why the monomeric dissociation products are unfolded to a various degree, which is manifested in a wide distribution of the monomer charge states at elevated collision voltages (Figure 2c; right).

MATERIALS AND METHODS

GroES and gp31

Both co-chaperonins were over-expressed in *E. coli* strain MC1009 [50] and purified as described previously [51]. The aliquots were flash-frozen in liquid nitrogen and stored at -80 °C.

Thermal unfolding

For the thermal unfolding assay, the proteins were dissolved in a 50 mM phosphate buffer at pH 6.8 to a final concentration of 0.15 mg/ml. The measurements were taken on a J-810 Spectropolarimeter (Jasco Inc, Easton, MD, USA) using 250µL cuvette. Both GroES and gp31 were scanned from 20 to 90 °C at the rate of 0.5 °C/min. Ellipticity was measured at 203 nm and 207 nm for GroES and gp31, respectively. Doubly-averaged CD spectra were taken with 10 degree steps to control the state of the sample. For the thermal unfolding plots, the ellipticity values were normalized on the maximal value for each sample and plotted against the temperature. The transition temperatures were defined as the temperature value at which half of the analyte was unfolded.

Mass spectrometry

MS and tandem MS data were collected on the Thermo Scientific Orbitrap Exactive Plus mass spectrometer modified and optimized for transmission and detection of ions with m/z up to 50 kTh as described previously [35]. For ion isolation we used a standard quadrupole mass filter from a Q Exactive instrument with a modified electronic board featuring a decreased resonance frequency of 284 kHz enabling an upper mass-selection limit above 20 kTh [36,52]. The samples were buffer-exchanged into 10 mM ammonium acetate at pH 6.8 and diluted to a final concentration of 10 µM immediately before the experiment. All acquisitions were collected at the 64 ms transient times equivalent to 17500 resolution at m/z 200. For each final spectrum a minimum of 10 scans was combined, containing 10 µscans at 100 ms injection times. In all MS and IM-MS experiments, the samples were sprayed from gold-coated borosilicate glass needles produced in-house[53].

Ion mobility-mass spectrometry

IM-MS measurements were performed in positive ion mode using an electrospray ionization quadrupole ion mobility time-of-flight (ESI-Q-IM-TOF) instrument (Synapt HDMS, Waters, UK) equipped with a Z-spray nano-electrospray ionization source. To retain complexes intact in the gas phase and improve collisional cooling and transmission of ions, pressure in the source region was elevated to 8 mBar. Argon was used as a collisional gas in the trap region at a pressure of 3×10^{-2} mBar. The ion mobility cell was filled with nitrogen (7.5×10^{-1} mBar). Wave height of 12 V and 15 V and wave velocity of 300 m/s and 500 m/s were used for the analyzed proteins in heptameric and monomeric forms, respectively.

Data processing

The Orbitrap MS data were automatically processed using in-house built software. The IM-MS data were manually extracted using DriftScope 3.0 (Waters, UK) and converted to text format using MSConvert utility from the ProteoWizard 3 suite [54]. Calculations of SASA were performed by using GETAREA scripts from http://curie.utmb.edu/getarea.html. "Contact Surface" script was taken from https://pymolwiki.org/index.php/Contact_Surface. Number of H-bonds were calculated with PyMOL script *list_hbonds.py* from http://pldserver1.biochem.queensu.ca/~rlc/work/pymol/. All further data analysis was performed using R [55] and ggplot2 [56]. The drift time and collision cross-section heatmap plots were produced using modified versions of the CIUSuite scripts[42].

CONCLUSION

The GroES and gp31 protein heptamers are functional and structural homologues. Both protein complexes act as molecular co-chaperonins partnering with the tetradecamer GroEL to assist in the folding of nonnative polypeptide chains in *E. coli*. Here, we first observed that in the gas phase the GroES heptamers are strikingly more stable than the gp31 heptamers, resulting in very distinctive breakdown curves as obtained upon collisional activation as well as distinctive collision induced unfolding patterns, which we determined by IM-MS. Subsequently, we probed the stability of GroES and gp31 heptameric complexes in solution using thermal unfolding assays. GroES showed a sharp melting/unfolding transition at 71 °C, whereas for gp31 this transition occurs at 60 °C. Conclusively, gp31 is less stable than GroES both in the gas phase and in solution. Analyzing the available high-resolution structures of the GroES and gp31 heptamers we could deduce that the subunit interfaces within GroES are larger and harbor more hydrogen bonds comparatively to gp31. These findings underscore the higher stability of GroES when compared to gp31 in solution. Overall, our data reveal that although the GroES and gp31 heptamers are functionally and structurally homologous, they exhibit striking differences in stability and unfolding. Finally, we conclude that solution phase structural properties of protein complexes stay partially unharmed upon transfer into the gas phase, which makes mass spectrometry-based approaches useful for complementing solution-phase analysis of protein complex stabilities.

ACKNOWLEDGMENTS

We thank the members of the Heck laboratory for support, especially Arjan Barendregt. We also thank Jonas Dorr for help with CD experiments and their interpretation and Aneika Leney for critical reading of the manuscript. This research was performed within the framework of The Netherlands Organization for Scientific Research (NWO) and supported by the large-scale proteomics facility Proteins@Work (project 184.032.201) embedded in The Netherlands Proteomics Centre. This project received additional funding from the European Union's Horizon 2020 research and innovation program under grant agreement 686547 (MSMed) for AJRH.

SUPPLEMENTARY MATERIAL

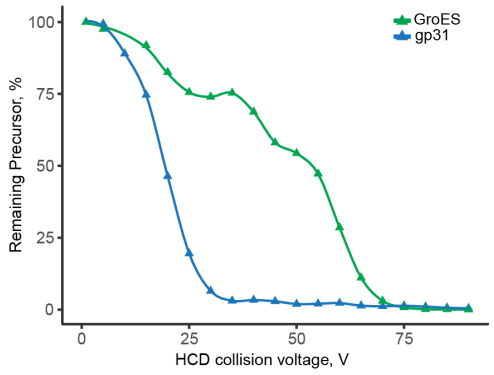


Figure S1 | Breakdown curves of 19+ charge state of heptameric GroES and gp31

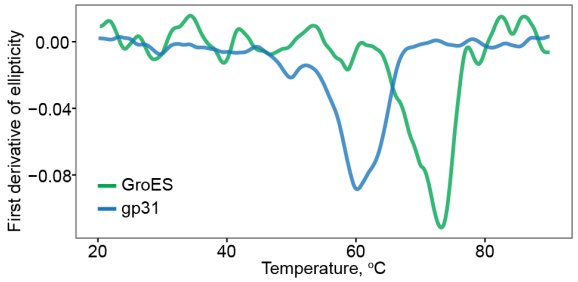


Figure S2 | IM-MS 2D heatmaps for several charge states of GroES and gp31.

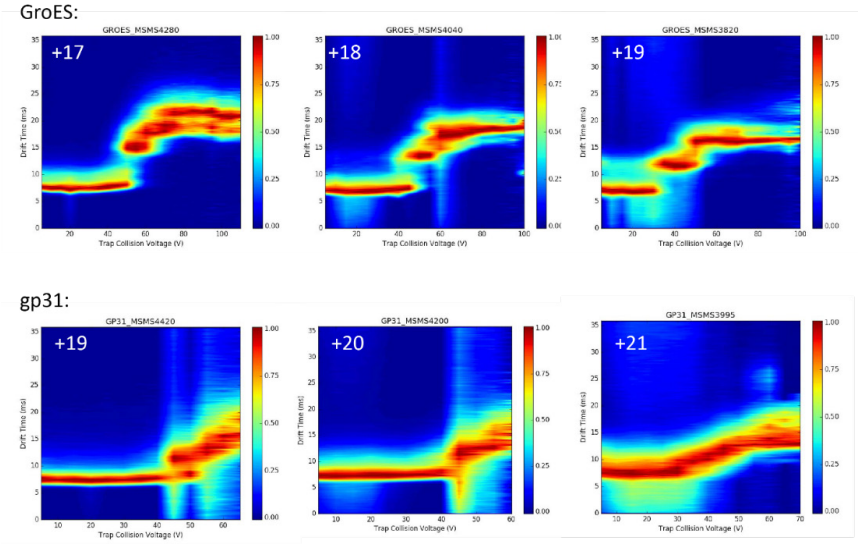


Figure S3 | First derivative of the molar ellipticity of GroES and gp31 plotted versus the temperature.

REFERENCES

1. Gruber, R., and Horovitz, A. (2016). Allosteric Mechanisms in Chaperonin Machines. Chem. Rev. *116*, 6588–6606.
2. Hayer-Hartl, M., Bracher, A., and Hartl, F.U. (2016). The GroEL-GroES Chaperonin Machine: A Nano-Cage for Protein Folding. Trends Biochem. Sci. *41*, 62–76.
3. Hartl, F.U., and Hayer-Hartl, M. (2002). Molecular chaperones in the cytosol: from nascent chain to folded protein. Science *295*, 1852–8.
4. Saibil, H.R., Fenton, W.A., Clare, D.K., and Horwich, A.L. (2013). Structure and allostery of the chaperonin GroEL. J. Mol. Biol. *425*, 1476–1487.
5. Fei, X., Ye, X., LaRonde, N.A., and Lorimer, G.H. (2014). Formation and structures of GroEL:GroES2 chaperonin footballs, the protein-folding functional form. Proc. Natl. Acad. Sci. U. S. A. *111*, 12775–80.
6. Horwich, A.L., and Fenton, W.A. (2009). Chaperonin-mediated protein folding: using a central cavity to kinetically assist polypeptide chain folding. Q. Rev. Biophys. *42*, 83–116.
7. Taguchi, H. (2005). Chaperonin GroEL meets the substrate protein as a "load" of the rings. J. Biochem. *137*, 543–9.
8. Hunt, J.F., Weaver, a J., Landry, S.J., Gierasch, L., and Deisenhofer, J. (1996). The crystal structure of the GroES co-chaperonin at 2.8 A resolution. Nature *379*, 37–45.
9. Xu, Z., Horwich, A.L., and Sigler, P.B. (1997). The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex. Nature *388*, 741–750.
10. Keppel, F., Lipinska, B., Ang, D., and Georgopoulos, C. (1990). Mutational analysis of the phage T4 morphogenetic 31 gene, whose product interacts with the Escherichia coli GroEL protein. Gene *86*, 19–25.
11. van der Vies, S.M., Gatenby, A.A., and Georgopoulos, C. (1994). Bacteriophage T4 encodes a co-chaperonin that can substitute for Escherichia coli GroES in protein folding. Nature *368*, 654–6.
12. Richardson, A., Schwager, F., Landry, S.J., and Georgopoulos, C. (2001). The importance of a mobile loop in regulating chaperonin/co-chaperonin interaction: Humans versus Escherichia coli. J. Biol. Chem. *276*, 4981–4987.
13. Richardson, A., Van Der Vies, S.M., Keppel, F., Taher, A., Landry, S.J., *et al.* (1999). Compensatory changes in GroEL/Gp31 affinity as a mechanism for allele- specific genetic interaction. J. Biol. Chem. *274*, 52–58.
14. Keppel, F., Rychner, M., and Georgopoulos, C. (2002). Bacteriophage-encoded co-chaperonins can substitute for Escherichia coli's essential GroES protein. EMBO Rep. *3*, 893–898.
15. Clare, D.K., Bakkes, P.J., van Heerikhuizen, H., Vies, S.M. van der, Saibil, H.R., *et al.* (2009). Chaperonin complex with a newly folded protein encapsulated in the folding chamber. Nature *457*, 107–113.
16. Hunt, J.F., van der Vies, S.M., Henry, L., and Deisenhofer, J. (1997). Structural Adaptations in the Specialized Bacteriophage T4 Co-Chaperonin Gp31 Expand the Size of the Anfinsen Cage. Cell *90*, 361–371.
17. Clare, D.K., Bakkes, P.J., van Heerikhuizen, H., van der Vies, S.M., and Saibil, H.R. (2006). An expanded protein folding cage in the GroEL-gp31 complex. J. Mol. Biol. *358*, 905–11.
18. Ruotolo, B.T., and Robinson, C. V. (2006). Aspects of native proteins are retained in vacuum. Curr. Opin. Chem. Biol. *10*, 402–408.
19. Jurneczko, E., and Barran, P.E. (2011). How useful is ion mobility mass spectrometry for structural biology? The relationship between protein crystal structures and their collision cross sections in the gas phase. Analyst *136*, 20–28.
20. Pierson, N.A., Chen, L., Valentine, S.J., Russell, D.H., and Clemmer, D.E. (2011). Number of solution states of bradykinin from ion mobility and mass spectrometry measurements. J. Am. Chem. Soc. *133*, 13810–13813.
21. Wolynes, P.G. (1995). Commentary Biomolecular folding in vacuo !!!(?). Proc. Natl. Acad. Sci. U. S. A. *92*, 2426–2427.

22. Smith, R.D., Loo, J. a, Edmonds, C.G., Barinaga, C.J., and Udseth, H.R. (1990). New developments in biochemical mass spectrometry: electrospray ionization. Anal. Chem. *62*, 882–899.

23. Bush, M.F., Hall, Z., Giles, K., Hoyes, J., Robinson, C. V, *et al.* (2010). Collision Cross Sections of Proteins and Their Complexes : A Calibration Framework and Database for Gas-Phase Structural Biology. Anal. Chem. *82*, 9557–9565.

24. Hall, Z., Hernández, H., Marsh, J.A., Teichmann, S.A., and Robinson, C. V. (2013). The role of salt bridges, charge density, and subunit flexibility in determining disassembly routes of protein complexes. Structure *21*, 1325–1337.

25. Tamara, S., Scheltema, R.A., Heck, A.J.R., and Leney, A.C. (2017). Phosphate Transfer in Activated Protein Complexes Reveals Interaction Sites. Angew. Chemie - Int. Ed. *56*, 13641–13644.

26. Zhang, Z., Browne, S.J., and Vachet, R.W. (2014). Exploring salt bridge structures of gas-phase protein ions using multiple stages of electron transfer and collision induced dissociation. J. Am. Soc. Mass Spectrom. *25*, 604–613.

27. Morrison, L.J., and Brodbelt, J.S. (2015). Charge site assignment in native proteins by ultraviolet photodissociation (UVPD) mass spectrometry. Analyst *166*, 166–176.

28. Zhang, J., Malmirchegini, G.R., Clubb, T., and Loo, J.A. (2015). Native top-down mass spectrometry for the structural characterization of human hemoglobin. Eur. J. Mass Spectrom. *21*, 221–231.

29. Geels, R.B.J., van der Vies, S.M., Heck, A.J.R., and Heeren, R.M. a (2006). Electron capture dissociation as structural probe for noncovalent gas-phase protein assemblies. Anal. Chem. *78*, 7191–6.

30. Tamara, S., Dyachenko, A., Fort, K.L., Makarov, A.A., Scheltema, R.A., *et al.* (2016). Symmetry of Charge Partitioning in Collisional and UV Photon-induced Dissociation of Protein Assemblies. J. Am. Chem. Soc. *138*, 10860–10868.

31. Donald, L.J., Stokell, D.J., Holliday, N.J., Ens, W., Standing, K.G., *et al.* (2005). Multiple equilibria of the Escherichia coli chaperonin GroES revealed by mass spectrometry. Protein Sci. *14*, 1375–1379.

32. Van Duijn, E., Barendregt, A., Synowsky, S., Versluis, C., and Heck, A.J.R. (2009). Chaperonin complexes monitored by ion mobility mass spectrometry. J. Am. Chem. Soc. *131*, 1452–1459.

33. van Duijn, E., Bakkes, P.J., Heeren, R.M.A., van den Heuvel, R.H.H., van Heerikhuizen, H., *et al.* (2005). Monitoring macromolecular complexes involved in the chaperonin-assisted protein folding cycle by mass spectrometry. Nat. Methods *2*, 371–6.

34. van Duijn, E., Heck, A.J.R., and van der Vies, S.M. (2007). Inter-ring communication allows the GroEL chaperonin complex to distinguish between different substrates. Protein Sci. *16*, 956–65.

35. Rose, R.J., Damoc, E., Denisov, E., Makarov, A., and Heck, A.J.R. (2012). High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. Nat. Methods *9*, 1084–6.

36. Dyachenko, A., Wang, G., Belov, M.E., Makarov, A., de Jong, R.N., *et al.* (2015). Tandem native mass-spectrometry on antibody-drug conjugates and sub-million Da antibody-antigen protein assemblies on an Orbitrap EMR equipped with a high-mass quadrupole mass selector. Anal. Chem. *87*, 6095–102.

37. Meyer, T., de la Cruz, X., and Orozco, M. (2009). An atomistic view to the gas phase proteome. Structure *17*, 88–95.

38. Konermann, L., Silva, E.A., and Sogbein, O.F. (2001). Electrochemically induced pH changes resulting in protein unfolding in the ion source of an electrospray mass spectrometer. Anal. Chem. *73*, 4836–4844.

39. Kaltashov, I.A., and Eyles, S.J. (2002). Studies of biomolecular conformations and conformational dynamics by mass spectrometry. Mass Spectrom. Rev. *21*, 37–71.

40. Popa, V., Trecroce, D.A., McAllister, R.G., and Konermann, L. (2016). Collision-Induced Dissociation of Electrosprayed Protein Complexes: An All-Atom Molecular Dynamics Model with Mobile Protons. J. Phys. Chem. B *8*, 5114–5124.

41. Niu, S., Rabuck, J.N., and Ruotolo, B.T. (2013). Ion mobility-mass spectrometry of intact protein-ligand complexes for pharmaceutical drug discovery and development. Curr. Opin. Chem. Biol. *17*, 809–817.

42. Eschweiler, J.D., Rabuck-Gibbons, J.N., Tian, Y., and Ruotolo, B.T. (2015). CIU-Suite: A Quantitative Analysis Package for Collision Induced Unfolding Measurements of Gas-Phase Protein Ions. Anal. Chem. *87*, 11516–11522.

43. Ruotolo, B.T., Benesch, J.L.P., Sandercock, A.M., Hyung, S.-J., and Robinson, C. V (2008). Ion mobility-mass spectrometry analysis of large protein complexes. Nat. Protoc. *3*, 1139–52.

44. Boudker, O., Todd, M.J., and Freire, E. (1997). The structural stability of the co-chaperonin GroES. J. Mol. Biol. *272*, 770–9.

45. Song, Y., Nelp, M.T., Bandarian, V., and Wysocki, V.H. (2015). Refining the Structural Model of a Heterohexameric Protein Complex: Surface Induced Dissociation and Ion Mobility Provide Key Connectivity and Topology Information. ACS Cent. Sci. *1*, 477–487.

46. Hall, Z., Hernández, H., Marsh, J.A., Teichmann, S.A., and Robinson, C.V. (2013). The Role of Salt Bridges, Charge Density, and Subunit Flexibility in Determining Disassembly Routes of Protein Complexes. Structure *21*, 1325–1337.

47. Dodds, E.D., Blackwell, A.E., Jones, C.M., Holso, K.L., O'Brien, D.J., *et al.* (2011). Determinants of gas-phase disassembly behavior in homodimeric protein complexes with related yet divergent structures. Anal. Chem. *83*, 3881–3889.

48. Sinelnikov, I., Kitova, E.N., and Klassen, J.S. (2007). Influence of Coulombic repulsion on the dissociation pathways and energetics of multiprotein complexes in the gas phase. J. Am. Soc. Mass Spectrom. *18*, 617–31.

49. Niu, S., and Ruotolo, B.T. (2015). Collisional unfolding of multiprotein complexes reveals cooperative stabilization upon ligand binding. Protein Sci. *24*, 1272–1281.

50. Richardson, A., and Georgopoulos, C. (1999). Genetic analysis of the bacteriophage T4-encoded cochaperonin Gp31. Genetics *152*, 1449–1457.

51. van der Vies, S.M. (2000). Purification of the Gp31 Co-chaperonin of bacteriophage T4. Methods Mol. Biol. *140*, 51–61.

52. Belov, M.E., Damoc, E., Denisov, E., Compton, P.D., Horning, S., *et al.* (2013). From protein complexes to subunit backbone fragments: A multi-stage approach to native mass spectrometry. Anal. Chem. *85*, 11163–11173.

53. Rosati, S., Yang, Y., Barendregt, A., and Heck, A.J.R. (2014). Detailed mass analysis of structural heterogeneity in monoclonal antibodies using native mass spectrometry. Nat. Protoc. *9*, 967–976.

54. Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., *et al.* (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nat. Biotechnol. *30*, 918–920.

55. R Core Team (2012). R: A Language and Environment for Statistical Computing.

56. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer New York).

PART I

GAS-PHASE ACTIVATION REVEALS STRUCTURAL FEATURES OF PROTEIN ASSEBMLIES

CHAPTER

4

PHOSPHATE TRANSFER IN ACTIVATED PROTEIN COMPLEXES REVEALS INTERACTION SITES

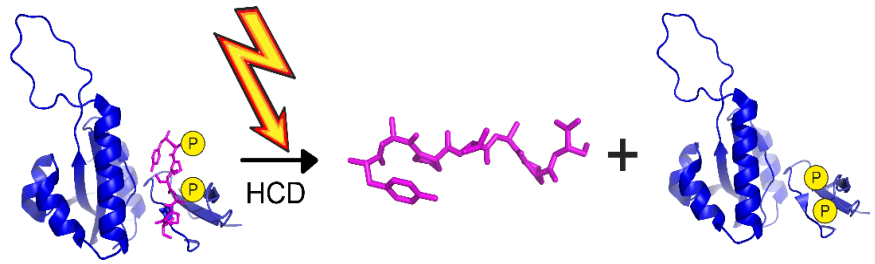
Sem Tamara†, Richard A. Scheltema†, Albert J. R. Heck†, Aneika C. Leney†

† Utrecht University, Utrecht, The Netherlands

Angew. Chemie Int. Ed. 2017, 129 (44) 13829-13832
DOI: 10.1002/anie.201706749

ABSTRACT

For many proteins, phosphorylation regulates their interaction with other biomolecules. Herein, we describe an unexpected phenomenon whereby phosphate groups are transferred non-enzymatically from one interaction partner to the other within a binding interface upon activation in the gas phase. Providing that a high affinity exists between the donor and acceptor sites, this phosphate transfer is very efficient and the phosphate groups only ligate to sites in proximity to the binding region. Consequently, such phosphate-transfer reactions may define with high precision the binding site between a phosphoprotein and its binding partner, as well as reveal that the binding site in this system is retained in the phase transfer from solution to the gas phase.



INTRODUCTION

The transfer of a phosphate group to a protein is a key regulator in protein function[1,2]. Phosphorylation can be a strict pre-requisite for protein interactions resulting in the switching on or off of signaling cascades. This is exemplified by proteins containing SH2 domains. In these, the SH2 domain is crucial for interactions with phosphorylated tyrosine residues regulating signaling in receptor tyrosine kinase pathways[3]. Another example is E3 ligase substrates that harbor phosphodegrons, whereby when the degron sequence is phosphorylated the substrate interacts with the ligase, becomes ubiquitinated, and is targeted to the proteasome for degradation[4]. Although the frequency of phosphorylation-mediated interactions in cells is high, fundamental knowledge is often lacking as to how phosphorylation governs this interaction.

One, less characterized interaction, is that between the peptidyl-prolyl cis-trans isomerase (PPIase), Pin1 and its phosphoprotein substrates. Pin1 comprises two domains; a N-terminal WW domain and a C-terminal PPIase domain[5]. The presence of a WW domain in Pin1, a protein module which facilitates binding to phosphorylated motifs, makes it unique within the PPIase family[6]. Thus, Pin1 catalyzes the cis-trans isomerization of specifically phosphorylated Ser/Thr-Pro bonds. This isomerization in turn can regulate protein dephosphorylation since many phosphatases only act on substrates comprising a specific prolyl peptide bond conformation[7]. This sequence of events catalyzed by Pin1 plays an important role in the regulation of transcription and pre-mRNA processing[8]. Here, Pin1 binds to the phosphorylated C-terminus of RNA polymerase (RNAP CTD), modulating its phosphorylation status and thus its ability to transcribe genes during the cell cycle[9]. Pin1 also acts to regulate transcription in response to TGF-beta signaling through its interaction with phosphorylated Smad3[10,11]. Interestingly, Pin1 is not only involved in transcriptional regulation, but also linked to Tau hyper-phosphorylation[12], a phenomenon that correlates with Alzheimer's disease progression and pro-survival signaling in eosinophils, a defined feature of Asthma[13]. Therefore, it is critical to investigate the mechanism by which Pin1, in a phosphorylation-dependent manner, binds its interaction partners.

To gain further understanding of phosphorylation-dependent interactions, methods are needed to monitor when and under which conditions these interactions occur. Ideally, these methods would monitor both complex formation, the number of phosphorylation sites required for binding, and localize precisely the phosphorylation sites of interest[14]. Mass spectrometry (MS) is an ideal tool for the detection and localization of protein post-translational modifications. In standard, bottom-up proteomics approaches, modified proteins are enzymatically digested into peptides, the peptides separated by LC and their sequence and modified sites confirmed by MS/MS[15,16]. These data report quantitatively on the specific sites of modification, however, lack structural information on the original native state of the proteins in solution (i.e. whether these modified proteins are involved in non-covalent complex formation or function as single entities). Native mass spectrometry (native MS), a technique whereby biomolecules are analyzed in their non-denatured state[17], provides this additional information and thereby provides insights into phosphoryla-

tion-dependent non-covalent protein complex formation[14,18]. In addition, through combination with tandem MS, native MS can provide information on the stoichiometry of proteins within large macromolecular complexes and their interaction networks[19]. These experiments rely on the products of dissociation in the gas-phase accurately reflecting the assembly partners in solution. Indeed, this has proven highly successful in the analysis of many large protein complexes such as V-type ATPases[20], ribosomes[21] and the 19S proteosome[22].

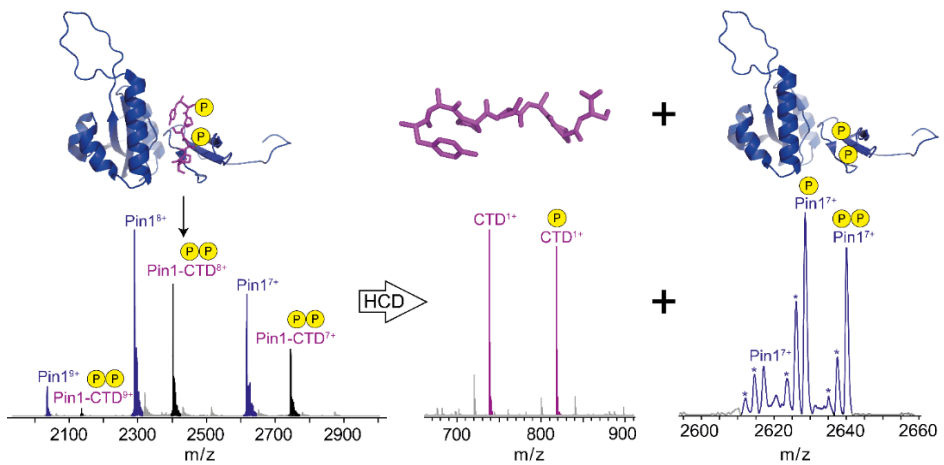


Figure 1 | Native (tandem) MS spectra of the Pin1-RNAP CTD (CTD) complex (left) and its dissociation products (right) when subjected to higher-energy collisional induced dissociation (HCD). Phosphate moieties transfer from phosphorylated RNAP-CTD to Pin1 resulting in the formation of doubly- and singly-phosphorylated Pin1 and a non-phosphorylated RNAP CTD peptide.

RESULTS AND DISCUSSION

Here, in an investigation focusing on high-affinity Pin1-phosphopeptide complexes, we stumbled upon an unanticipated phenomenon, whereby the products of gas-phase dissociation no longer reflected the original constituents in solution. Instead, phosphate moieties moved from the original phosphopeptides to proximate acceptor sites on the interacting Pin1 protein, providing information on the location of the binding site. We first observed this phenomenon in experiments whereby Pin1 was incubated with a doubly phosphorylated peptide mimicking its known protein binding partner, termed RNAP CTD (Table S1). As expected, based on earlier data[23,24], we observed by native MS a 1:1 Pin1:RNAP CTD complex (Figure 1, Figure S1A, D). The affinity estimated by native MS (Kd = 29 ±12 μM) is consistent with the reported affinity of RNAP CTD to the WW domain of Pin1[24]. Next, a single charge state $[M+8H]^{8+}$ of the Pin1-RNAP CTD complex was subjected to higher-energy collisional induced dissociation (HCD). We expected that in these HCD experiments the non-covalent interaction would break resulting in the complex dissociating into its original constituents, i.e. Pin1 and the doubly phosphorylated RNAP CTD peptide. However, the most dominant peaks observed correspond to Pin1 with one or two phosphates covalently bound suggesting that phosphate groups have been transferred from the RNAP CTD to Pin1 within the complex (Figure 1, Figure S2A). In line with this observation, we also observed the complementary fragment ions corresponding to the singly phosphorylated and unphosphorylated RNAP CTD (Figure 1, Figure S2A). The formation of phosphorylated Pin1 correlates well with the dephosphorylation of RNAP CTD, when monitored as a function of normalized HCD energy (Figure S3A). Thus, phosphate transfer fully precludes dissociation upon Pin1-RNAP CTD activation (Figure 1) with only baseline levels of free Pin1 and doubly phosphorylated peptide observed in the tandem MS spectra.

Intrigued by our observations, we sought whether this phenomenon was peptide independent. Thus, another phosphorylated peptide was chosen corresponding to residues 202-215 of Smad3 (Table S1) and incubated with Pin1. Comparable to data for the Pin1-RNAP CTD complex, a 1:1 Pin1-Smad3 complex was observed in native MS with an estimated somewhat smaller dissociation constant of 57 ± 20 μM suggesting Smad3 is also bound to the WW domain of Pin1 (Figure S1B, D). Likewise, upon HCD, phosphate transfer between phosphorylated Smad3 and Pin1 was observed, albeit to a lesser extent (Figure S2B, S3B).

We wanted to eliminate the possibility that the phosphate transfer had occurred already in solution prior to gas-phase analysis. Thus, the Pin1-RNAP CTD complex was formed in solution and subsequently dissociated prior to MS analysis (Figure S2D). In these experiments, no phosphorylation was detected on Pin1. Thus, we conclude that phosphate transfer exclusively occurs upon gas-phase activation of the Pin1-phosphopeptide complex and not in solution.

We hypothesized that the extent of phosphate transfer upon HCD (Figure S2) might correlate with the binding affinities between the Pin1-phosphopeptide complexes, occurring thus predominantly within high affinity, specific Pin1-phosphopeptide complexes. To test this, RNAP CTD was incubated with cytochrome c; a protein with an alike MW as Pin1, but with no known specific interaction with RNAP CTD. The native MS data revealed weak ion signals corresponding to a low abundant cytochrome c:RNAP CTD complex, likely formed by non-specific interactions that can occur during the ESI process[25,26] (Figure S1C, D). Consistent with our hypothesis, upon HCD of this low-affinity non-covalent complex no phosphate transfer was observed between cytochrome c and RNAP CTD. Instead, we observed the predicted formation of cytochrome c and double phosphorylated RNAP CTD as product ions (Figure S2C, S3C). Thus, we conclude that phosphate transfer is likely specific to protein-phosphopeptide complexes whereby the phosphopeptide is bound tightly within the binding site. Consistent with our data, phosphate transfer has previously been observed in the gas-phase within phosphopeptides and during the dissociation of phosphopeptide dimers[27]. However, to our knowledge, this is the first instance of phosphate transfer within a non-covalent protein complex, our findings having significant implications on the analysis of structural biology-based MS experiments on phosphoprotein-protein complexes.

In the reported crystal structure of the Pin1-RNAP CTD complex, the phosphoserine at Ser5 on RNAP CTD interacts strongly to Pin1 through hydrogen bonding to Ser16, Arg17 and Tyr23[24]. If such a structure would be largely retained in the gas-

phase, we would expect the phosphate group on Ser5 to most likely migrate to one of these aforementioned Pin1 acceptor residues. To validate this hypothesis, further top-down fragmentation (i.e. MS3) was performed on the doubly phosphorylated Pin1 fragment ions formed following HCD activation (Figure S4A). Fragments were observed throughout the entire Pin1 sequence (up to 75 % of backbone cleavages) enabling us to accurately pin-point the phosphorylation sites within the WW domain. Short singly-phosphorylated fragments exclusive for the WW domain together with long doubly-phosphorylated fragments spanning across both domains locate the phosphorylation sites in Pin1 to residues in between 16-23 (Figure S4B). Upon comparison of all possible transfer sites within the Pin1 sequence, we found that all these possible phosphosites on Pin1 are within 10 Å of pSer5 in the crystal structure. Interestingly, Ser16, Arg17 and Tyr23 in Pin1 are the closest residues to pSer5 phosphate and consistently display in our data a high number of characteristic phosphorylated fragments (Figure 2), thus we anticipate that these are the most likely transfer sites on Pin1. For the transfer of the second phosphate (pSer2), it is likely that the phosphate migrates to multiple sites within the binding region since this side chain is more flexible and less stabilized in the crystal structure.

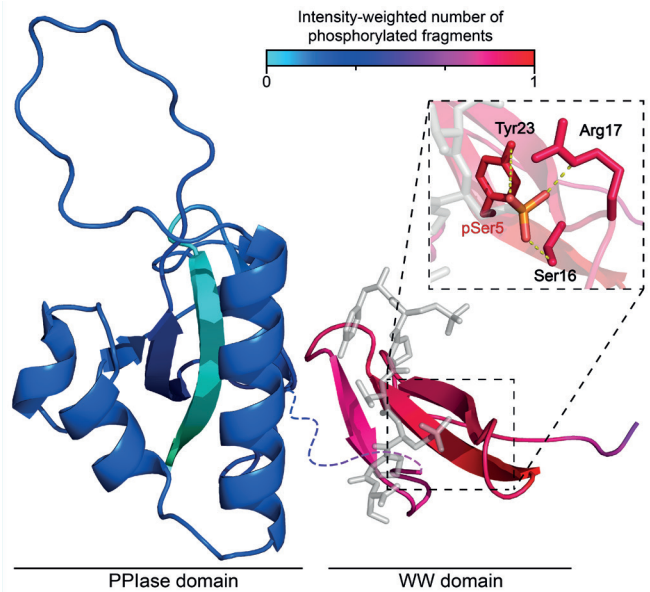


Figure 2 | Crystal structure of Pin1-RNAP CTD (PDB entry: 1f8a) color coded corresponding to the intensity-weighted number of observed phosphorylated fragments. The RNAP CTD mimicking peptide is shown in grey. The interaction of pSer5 (CTD) with Tyr23, Arg17, and Ser16 (Pin1) is displayed in the inset. The dotted line corresponds to residues absent in the crystal structure.

In summary, we show that phosphate groups within a non-covalent complex can transfer from one to the other binding partner upon gas-phase activation. Such a rearrangement indicates a major structural change within the complex. This phosphate transfer seems to occur exclusively in tightly interacting complexes whereby the phosphate group is present at the binding interface and crucial for the high affinity interaction. Thus, since this phosphate transfer only occurs to phosphate

receptor residues in close proximity, location of these phosphate receptor sites in combination with the location of the original phosphosite, can together provide valuable information on the location of protein-protein interaction interfaces. We anticipate that this finding is not unique to Pin1 and could have broader implications in the context of other high affinity phosphorylation-dependent biomolecular interactions.

EXPERIMENTAL SECTION

For complex formation, Pin1/cytochrome C (5 µM) was incubated with a 5-fold excess of either a phosphopeptide mimicking the C-terminal domain of RNA polymerase or the SMAD3 protein in 50 mM ammonium acetate pH 6.8 on ice for 30 min. Binding affinities were calculated at different ligand concentrations and non-specific binding corrected for using the reference protein method[25]. Mass spectra were acquired by direct infusion using a nanoESI source coupled to either and Orbitrap EMR or Orbitrap Fusion Lumos mass spectrometer. To monitor the phosphate transfer reactions, the most abundant charge state (8+) corresponding to the Pin1-phosphopeptide complex was selected and subjected systematically to HCD using a normalized collision energy of 5-30. For phosphate transfer-site localization, the singly and doubly phosphorylated Pin1 fragment ions formed following HCD were mass selected and further subjected to EThcD fragmentation. All Pin1 fragments were assigned using an in-house developed data analysis software[18]. More detailed experimental details are available in the Supporting Information.

ACKNOWLEDGEMENTS

We thank Dr. G. E. Folkers for assistance with Pin1 expression and purification. This work was supported by the Roadmap Initiative Proteins@Work (project number 184.032.201), funded by the Netherlands Organization for Scientific Research (NWO) and through the European Union Horizon 2020 programme FET-OPEN project MSmed, Project 686547.

SUPPLEMENTARY MATERIAL

Pin1 expression and purification

$Pin1_{his}$ was a gift from Dustin Maly (Addgene plasmid # 40773). The $Pin1_{his}$ plasmid was transformed into BL21 (DE3) Rosetta™ 2 *E.coli* cells (Novagen) to enable protein overexpression. $Pin1_{his}$ was overexpressed using autoinduction[28] and purified using nickel-affinity chromatography. The hexa-histidine tag was subsequently removed by incubation overnight with recombinant hexa-histidine tagged TEV protease. TEV protease and any remaining $Pin1_{his}$ were then separated from Pin1 using nickel-affinity chromatography and Pin1 further purified by gel filtration chromatography. Pin1 was stored in 50 mM HEPES, 300 mM NaCl, 1 mM DTT, 10 % glycerol, pH 7.4 at -80 °C and buffer exchanged into 50 mM ammonium acetate pH 6.8 prior to mass spectrometry analysis.

Protein-ligand complex formation and analysis

Peptides were synthesized in the group of H. Ovaa at Leiden University Medical Centre. The sequences used and their corresponding molecular weights are detailed in Table S1.

Table S1 | Overview of phosphopeptides used. The RNA polymerase peptide (termed RNAP CTD) sequence represents one of the 52 heptad repeats present in the C-terminal domain of the human RNA polymerase. The Smad3 peptide covers residues 202-215 of human Smad3.

Peptide Sequence	Protein	Molecular Weight (Da)
YpSPTpSPS	RNA polymerase	897.6
AGpSPNLSPNPMSPA	Smad3	1419.4

To form the Pin1-phosphopeptide complexes, Pin1 (5 µM) was incubated with a 5-fold excess of phosphopeptide for 30 min in 50 mM ammonium acetate pH 6.8 at 4 °C. Complex formation was monitored by native MS performed on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific) coupled with a nanoESI source (Figure S1A, B). In-house pulled gold-coated borosilicate capillaries were used and all measurements were performed in positive mode. A capillary voltage of 1-1.3 kV was used, the source fragmentation set to zero, and the instrument operated in standard pressure mode using nitrogen as a collision gas. Ions were detected in the Orbitrap using a resolution of 120, 000 at 400 m/z. The data was processed using Xcalibur v2.2 (Thermo Fisher Scientific).

Non-specific binding between proteins and ligands can occur during the ESI process[29]. To verify the protein-phosphopeptide complexes observed by native MS are specific and reflective of the complexes present in solution, the binding of cytochrome C (purchased from Sigma-Aldrich) to RNAP CTD, an interaction not known to occur in solution, was monitored by native MS. Briefly, cytochrome C (5 µM) was incubated with a 5-fold excess of RNAP CTD for 30 min in 50 mM ammonium acetate pH 6.8 at 4 °C and native MS carried out as previously described. Less than 5 % of the total protein signal was attributed to the formation of cytochrome c-RNAP CTD complex, i.e. non-specific binding (Figure S1C, D). Thus, the detected Pin1-RNAP CTD and Pin1-Smad3 complexes are tight binders and do not result as an artifact of the ESI process.

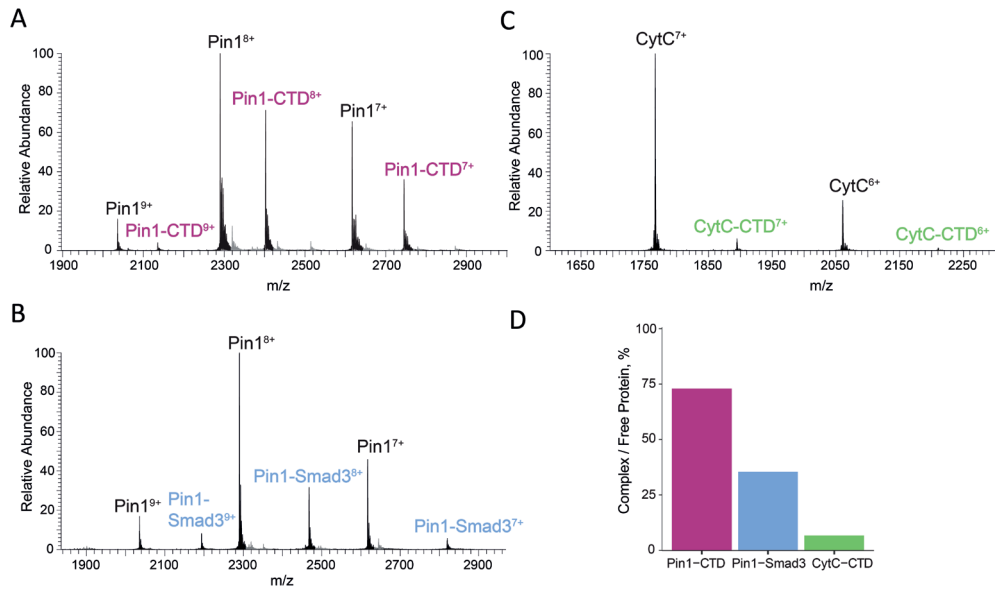


Figure S1 | Pin1 binds phosphopeptide mimics of RNAP CTD and Smad3. Native MS spectra of non-covalent protein-phosphopeptide complexes formed between (A) Pin1 and a doubly phosphorylated RNAP CTD (CTD), and (B) Pin1 and a singly phosphorylated Smad3 peptide. (C) Native MS showing a low-abundant, non-specific complex formed between Cytochrome C (CytC) and the RNAP CTD peptide. (D) A summary of complex abundance, which is substantially lower for the non-specific CytC-RNAP CTD complex in comparison with the specific Pin1-RNAP CTD and Pin1-Smad3 complexes formed.

The reference protein ESI-MS assay was used to quantify the binding of RNAP CTD and Smad3 to Pin1[29]. Binding measurements were carried out using a fixed Pin1 concentration of 5 µM and 4 different ligand concentrations (5 µM, 10 µM, 20 µM and 50 µM). Cytochrome c was used as a reference protein to correct for non-specific interactions. All data was acquired on a Orbitrap EMR mass spectrometer (Thermo Fisher Scientific) equipped with a nanoESI source. The capillary voltage was set to 1.4 kV, source fragmentation 0, trap gas 4. The injection flatapole, inter flatapole lens, bent flatapole DC and transfer multiple were set to 8, 7, 6 and 4, respectively. CsI was used for calibration and all spectra were acquired at a resolution of 15,000. All data were processed using MassLynx software v.4.1 (Waters). The dissociation constant (K_d) was calculated as described earlier using Equation 1[30–32].

$$K_d = \frac{[L]_0 - \frac{R}{1+R}[P]_0}{R} \quad (1)$$

where $[P]_0$ and $[L]_0$ are the initial Pin1 and phosphopeptide concentrations, respectively, and R is the abundance ratio of the Pin1-phosphopeptide complex to free Pin1 as measured by native MS. The error represents the standard deviation between the binding constants calculated at different ligand concentrations.

Gas phase transfer of phosphate from phosphopeptides to Pin1

For phosphate transfer studies, the most abundant charge state of the Pin1-phosphopeptide complex (8+) was isolated and subjected to HCD fragmentation using a Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific) (Figure S2A, B). The normalized collision energy (NCE) was systematically increased from 5 to 30 using 5 NCE increments. All mass spectra were deconvoluted using Protein Deconvolution 4.1 (Thermo Fisher Scientific). The relative abundance of the dissociation products (Pin1, phosphorylated Pin1, peptide and phosphopeptide ions) were then plotted as a function of NCE (Figure S3A, B). All measurements were performed in triplicate.

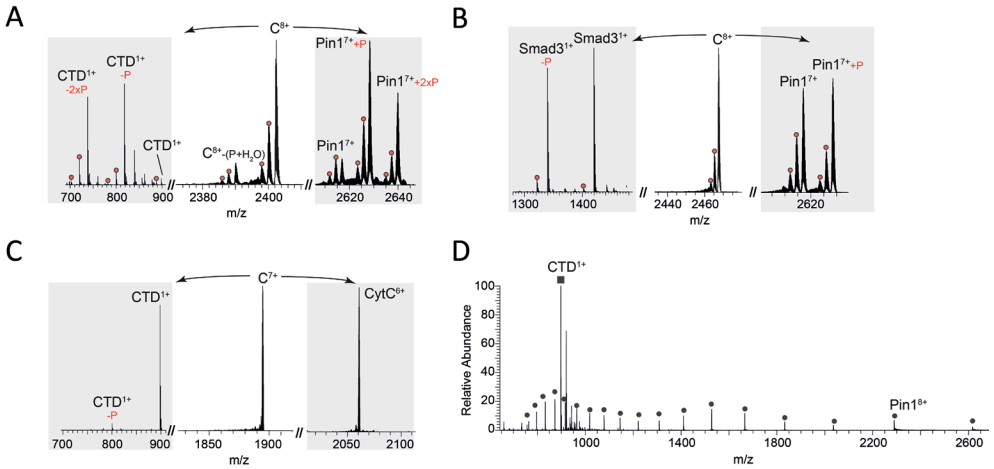


Figure S2 | Phosphate transfer occurs with high affinity Pin1-phosphopeptide complexes in the gas-phase and not in solution. (A) HCD (25 NCE) on the isolated Pin1-RNAP CTD^{8+} complex (C^{8+}) resulted in its dissociation into doubly-, singly-, and non-phosphorylated Pin1 and RNAP CTD peptide (CTD). (B) HCD (25 NCE) on the isolated Pin1-Smad3^{8+} complex (C^{8+}) led to the formation of singly- and non-phosphorylated Pin1 and Smad3 peptide. Peaks corresponding to HCD-related neutral water losses are annotated by a red circle. For these high affinity, specific complexes the products of gas-phase complex dissociation are not reflective of their initial complex constituents. (C) HCD (20 NCE) on the non-specific cytochrome C-RNAP CTD^{7+} complex (C^{7+}) showed dissociation occurs without phosphate transfer. (D) Mass spectra of Pin1-RNAP CTD complex (formed from incubating 5 μM Pin1 with 15 μM RNAP CTD) after in solution dissociation displayed no ions corresponding to phosphorylated Pin1 demonstrating that phosphate transfer does not occur in solution. Thus, phosphate transfer is a gas-phase specific process observed only at elevated collisional energies.

To verify whether phosphate transfer only occurs within high affinity Pin1-phosphopeptide complexes present in solution, the non-specific cytochrome C-RNAP CTD complex (7+ charge state) was isolated and subjected to HCD fragmentation (Figure S2C). Upon increasing HCD energy (10-25 NCE), no phosphate transfer was observed between RNAP CTD and cytochrome C (Figure S2C, S3C).

To confirm that phosphate transfer occurs only upon activation in the gas phase, the Pin1-RNAP CTD complex was formed in solution (as described earlier) and subsequently dissociated in solution by addition of formic acid to a final concentration of 1 %. MS analysis of the in-solution dissociated complex was performed on an

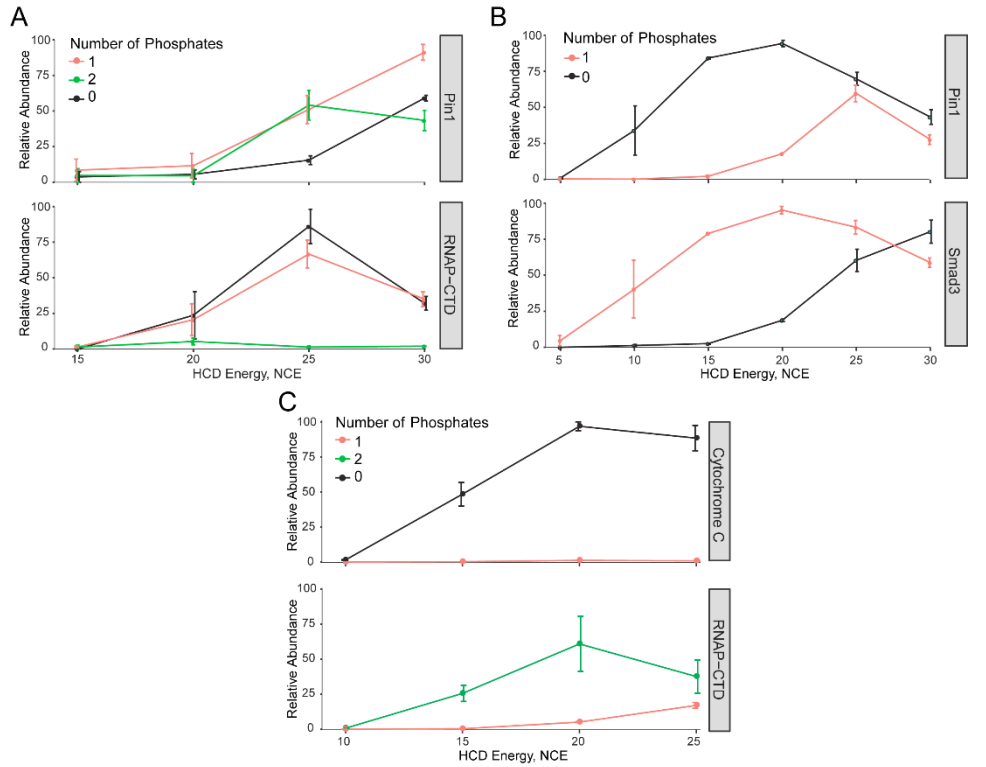


Figure S3 | Energy-resolved dissociation profiles of the Pin1-phosphopeptide complexes. (A) Relative abundance of singly-, doubly-, and non-phosphorylated Pin1 (top) and RNAP-CTD (bottom) as a function of normalized collisional energy (NCE) highlight the complementary release of phosphorylated Pin1 and the non-phosphorylated peptide from the Pin1-RNAP CTD complex. (B) Dissociation of the Pin1-Smad3 complex showed phosphate transfer occurs at high HCD energy in parallel with the typical protein-peptide dissociation pathway. (C) HCD activation of non-specific cytochrome C-RNAP CTD complex followed the standard dissociation pathway at all analyzed normalized collisional energies with no phosphorylated cytochrome C observed in the spectra. Error bars represent standard error of mean value calculated from three technical replicates.

Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific) using the parameters described previously. No phosphorylated Pin1 was detected showing the phosphate transfer observed does not occur in solution.

Identification of phosphate location on Pin1

All experiments were performed on a Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). To localize the phosphorylation sites on Pin1, ions corresponding to the 8+ charge state of the Pin1-RNAP CTD complex were initially subjected to HCD fragmentation using a NCE of 27.5. The unmodified, singly phosphorylated and doubly phosphorylated Pin1^{8+} ions formed were then further isolated and subjected to top-down fragmentation using electron transfer dissociation supplemented with HCD (EThcD)[33]. The ETD reaction time was set to 75 ms and the fragment ions from two NCE's (22.5 and 25) combined to localize

the phosphorylation site on Pin1. An ion injection time of 1000 ms was used and over 100 µscans were averaged in the time domain. Fragment ions were detected in the Orbitrap using a resolution of 120,000 at 400 m/z. The acquired data was deconvoluted using Protein Deconvolution 4.1 (Thermo Fisher Scientific) applying the Xtract[34] algorithm with the parameters: signal/noise threshold = 2, fit factor = 80%, remainder threshold = 25%. The fragment ions were assigned using an in-house developed program. Within this program the fragment ions were mapped on the Pin1 amino acid sequence, assigning them to the best matching theoretical fragments considering all b, c, y and z-type ions. The masses of experimental and theoretical fragment ions were compared and ions with a mass difference outside the set +/- 2.5 ppm range discarded. The phosphorylation sites were then localized by adding one or two phosphate moieties to Ser/Thr/Tyr residues in the N- or C-terminus of the Pin1 sequence. The sequence stretches covered by fragment ions were then reported showing at which position a new phosphorylation site occurs (Figure S4). The data was visualized in R with ggplot2 package[35]. The phosphosite locations on Pin1 were visualized in PyMOL v1.8 (Schrödinger, LLC) (PDB:1F8A).

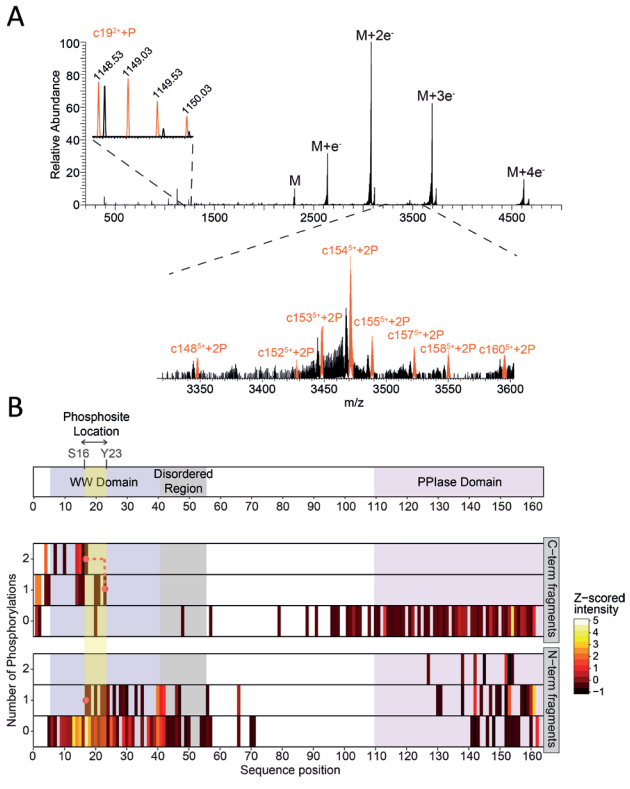


Figure S4 | Positioning of the transferred phosphate moieties on Pin1. (A) MS3 EThcD spectrum of the mass-selected $[M+H]^{8+}$ ions of doubly-phosphorylated Pin1 using a ETD reaction time of 75 ms with 17.5 % supplemental activation. M and P correspond to HPO_3 and $Pin1^{8+}$ with two phosphates bound, respectively. The mass spectrum is dominated by charge-reduced non-dissociated products. An example of some of the c-type fragment ions observed are shown. (B) Fragmentation heat map revealed no short fragments in the PPIase domain (C-terminal fragments) containing phosphate moieties. Short N-terminal fragments containing a single phosphorylation site and long C-terminal fragments containing 2 phosphorylation sites provide direct constraints for the localization of the phosphosites within the WW domain i.e. between Ser16 and Y23 (highlighted in yellow).

REFERENCES

1. Cohen, P. (2000). The regulation of protein function by multisite phosphorylation – a 25 year update. Trends Biochem. Sci. *25*, 596–601.
2. Ptacek, J., and Snyder, M. (2006). Charging it up: global analysis of protein phosphorylation. Trends Genet. *22*, 545–554.
3. Schlessinger, J. (1994). SH2/SH3 signaling proteins. Curr. Opin. Genet. Dev. *4*, 25–30.
4. Holt, L.J. (2012). Regulatory modules: Coupling protein stability to phopsho-regulation during cell division. FEBS Lett. *586*, 2773–2777.
5. Lu, K.P., and Zhou, X.Z. (2007). The prolyl isomerase PIN1: a pivotal new twist in phosphorylation signalling and disease. Nat. Rev. Mol. Cell Biol. *8*, 904–916.
6. Liou, Y.-C., Zhou, X.Z., and Lu, K.P. (2011). Prolyl isomerase Pin1 as a molecular switch to determine the fate of phosphoproteins. Trends Biochem. Sci. *36*, 501–514.
7. Mayfield, J.E., Fan, S., Wei, S., Zhang, M., *et al.* (2015). Chemical Tools To Decipher Regulation of Phosphatases by Proline Isomerization on Eukaryotic RNA Polymerase II. ACS Chem. Biol. *10*, 2405–2414.
8. Xu, Y.-X., Hirose, Y., Zhou, X.Z., Lu, K.P., *et al.* (2003). Pin1 modulates the structure and function of human RNA polymerase II. Genes Dev. *17*, 2765–2776.
9. Xu, Y.-X., and Manley, J.L. (2007). Pin1 modulates RNA polymerase II activity during the transcription cycle. Genes Dev. *21*, 2950–2962.
10. Nakano, A., Koinuma, D., Miyazawa, K., Uchida, T., *et al.* (2009). Pin1 down-regulates transforming growth factor-beta (TGF-beta) signaling by inducing degradation of Smad proteins. J. Biol. Chem. *284*, 6109–6115.
11. Matsuura, I., Chiang, K.-N., Lai, C.-Y., He, D., *et al.* (2010). Pin1 Promotes Transforming Growth Factor-β-induced Migration and Invasion. J. Biol. Chem. *285*, 1754–1764.
12. Butterfield, D.A., Abdul, H.M., Opii, W., Newman, S.F., *et al.* (2006). Pin1 in Alzheimer's disease. J. Neurochem. *98*, 1697–1706.
13. Shen, Z.-J., Esnault, S., Schinzel, A., Borner, C., *et al.* (2009). The peptidyl-prolyl isomerase Pin1 facilitates cytokine-induced survival of eosinophils by suppressing Bax activation. Nat. Immunol. *10*, 257–265.
14. van de Waterbeemd, M., Lössl, P., Gautier, V., Marino, F., *et al.* (2014). Simultaneous Assessment of Kinetic, Site-Specific, and Structural Aspects of Enzymatic Protein Phosphorylation. Angew. Chemie Int. Ed. *53*, 9660–9664.
15. Mann, M., and Jensen, O.N. (2003). Proteomic analysis of post-translational modifications. Nat. Biotechnol. *21*, 255–261.
16. Witze, E.S., Old, W.M., Resing, K.A., and Ahn, N.G. (2007). Mapping protein post-translational modifications with mass spectrometry. Nat. Methods *4*, 798–806.
17. Leney, A.C., and Heck, A.J.R. (2017). Native Mass Spectrometry: What is in the Name? J. Am. Soc. Mass Spectrom. *28*, 5–13.
18. Lössl, P., Brunner, A.M., Liu, F., Leney, A.C., *et al.* (2016). Deciphering the Interplay among Multisite Phosphorylation, Interaction Dynamics, and Conformational Transitions in a Tripartite Protein System. ACS Cent. Sci. *2*, 445–455.
19. Hernández, H., and Robinson, C. V. (2007). Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. Nat. Protoc. *2*, 715–726.
20. Zhou, M., Morgner, N., Barrera, N.P., Politis, A., *et al.* (2011). Mass Spectrometry of Intact V-Type ATPases Reveals Bound Lipids and the Effects of Nucleotide Binding. Science (80-.). *334*, 380–385.
21. van de Waterbeemd, M., Fort, K.L., Boll, D., Reinhardt-Szyba, M., *et al.* (2017). High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. Nat. Methods *14*, 283–286.
22. Politis, A., Stengel, F., Hall, Z., Hernández, H., *et al.* (2014). A mass spectrometry-based hybrid method for structural

modeling of protein complexes. Nat. Methods *11*, 403–406.

23. O'Brien, J.P., Li, W., Zhang, Y., and Brodbelt, J.S. (2014). Characterization of native protein complexes using ultraviolet photodissociation mass spectrometry. J. Am. Chem. Soc. *136*, 12920–8.

24. Verdecia, M.A., Bowman, M.E., Lu, K.P., Hunter, T., *et al*. (2000). Structural basis for phosphoserine-proline recognition by group IV WW domains. Nat. Struct. Mol. Biol. *7*, 639–643.

25. Sun, J., Kitova, E.N., Wang, W., and Klassen, J.S. (2006). Method for Distinguishing Specific from Nonspecific Protein−Ligand Complexes in Nanoelectrospray Ionization Mass Spectrometry. Anal. Chem. *78*, 3010–3018.

26. Kitova, E.N., El-Hawiet, A., Schnier, P.D., and Klassen, J.S. (2012). Reliable Determinations of Protein–Ligand Interactions by Direct ESI-MS Measurements. Are We There Yet? J. Am. Soc. Mass Spectrom. *23*, 431–441.

27. Gonzalez-Sanchez, M.-B., Lanucara, F., Hardman, G.E., and Eyers, C.E. (2014). Gas-phase intermolecular phosphate transfer within a phosphohistidine phosphopeptide dimer. Int. J. Mass Spectrom. *367*, 28–34.

28. Studier, F.W. (2005). Protein production by auto-induction in high-density shaking cultures. Protein Expr. Purif. *41*, 207–234.

29. Sun, J., Kitova, E.N., Wang, W., and Klassen, J.S. (2006). Method for Distinguishing Specific from Nonspecific Protein−Ligand Complexes in Nanoelectrospray Ionization Mass Spectrometry. Anal. Chem. *78*, 3010–3018.

30. Kitova, E.N., El-Hawiet, A., Schnier, P.D., and Klassen, J.S. (2012). Reliable determinations of protein-ligand interactions by direct ESI-MS measurements. Are we there yet? J. Am. Soc. Mass Spectrom. *23*, 431–441.

31. Jørgensen, T.J.D., Roepstorff, P., and Heck, A.J.R. (1998). Direct Determination of Solution Binding Constants for Noncovalent Complexes between Bacterial Cell Wall Peptide Analogues and Vancomycin Group Antibiotics by Electrospray Ionization Mass Spectrometry. Anal. Chem. *70*, 4427–4432.

32. Daniel, J.M., Friess, S.D., Rajagopalan, S., Wendt, S., *et al.* (2002). Quantitative determination of noncovalent binding interactions using soft ionization mass spectrometry. Int. J. Mass Spectrom. *216*, 1-27.

33. Frese, C.K., Altelaar, A.F.M., Toorn, H. Van Den, Nolting, D., *et al.* (2012). Toward Full Peptide Sequence Coverage by Dual Fragmentation Combining Electron-Transfer and Higher-Energy Collision Dissociation Tandem Mass Spectrometry. Anal. Chem. *84*, 9668–9673.

34. Zabrouskov, V., Senko, M.W., Du, Y., Leduc, R.D., *et al.* (2005). New and automated MSn approaches for top-down identification of modified proteins. J. Am. Soc. Mass Spectrom. *16*, 2027–2038.

35. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer New York).

PART II

ANALYSIS OF COMPOSITIONAL AND STRUCTURAL DIVERSITY IN PROTEIN ASSEBMLIES

5

CHAPTER

DISSECTING RIBOSOMAL PARTICLES THROUGHOUT THE KINGDOM OF LIFE USING ADVANCED HYBRID MASS SPECTROMETRY METHODS

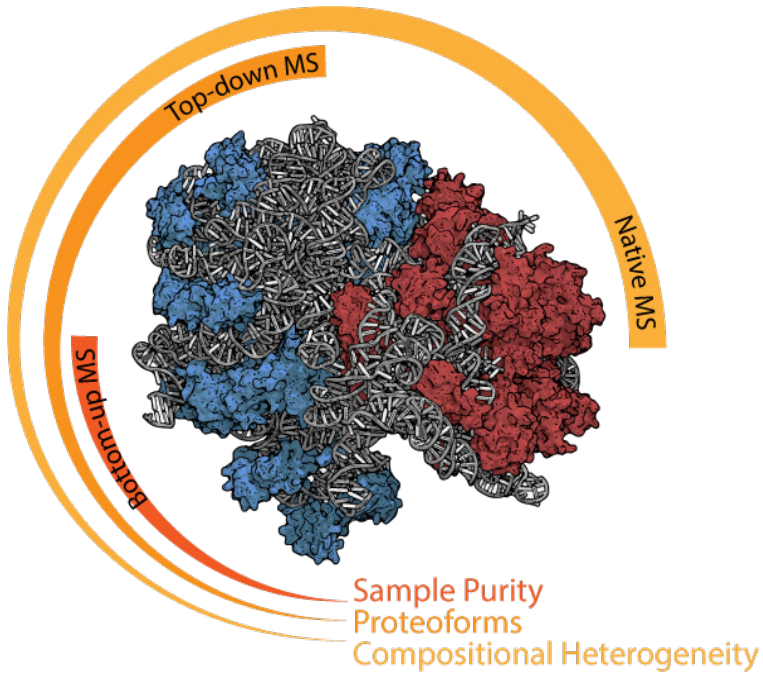
Michiel van de Waterbeemd[†,‡], Sem Tamara[†,‡], Kyle L Fort[†,‡,#], Eugen Damoc[#], Vojtech Franc[†], Philipp Bieri[$], Martin Itten[$], Alexander Makarov[†,#], Nenad Ban[$], and Albert JR Heck[†]

† Utrecht University, Utrecht, The Netherlands
$ ETH Zurich, Zurich, Switzerland
# Thermo Fisher Scientific, Bremen, Germany
‡ Contributed equally

Nat Commun. 2018, 9 (1) 2493
DOI: 10.1038/s41467-018-04853-x

ABSTRACT

Biomolecular mass spectrometry has matured strongly over the past decades and has now reached a stage where it can provide deep insights into the structure and composition of large cellular assemblies. Here, we describe a three-tiered hybrid mass spectrometry approach combining the latest developments in bottom-up, top-down and native mass spectrometry, that enables the dissection of macromolecular complexes in order to complement structural studies. To demonstrate the capabilities of the approach we have investigated ribosomes, large ribonucleoprotein particles responsible for protein synthesis and consisting of more than 50 protein subunits and multiple strands of RNA. Our results identify the sites of sequence processing, protein post-translational modifications and the assembly and stoichiometry of individual ribosomal proteins in four distinct ribosomal particles of bacterial, plant and human origin. Amongst others, we report extensive cysteine methylation in the zinc finger domain of the human S27 protein, the heptameric stoichiometry of the stalk complex in chloroplastic ribosomes, the heterogeneous composition of human 40S ribosomal subunit and its association to the CrPV and HCV internal ribosome entry site RNA elements.



INTRODUCTION

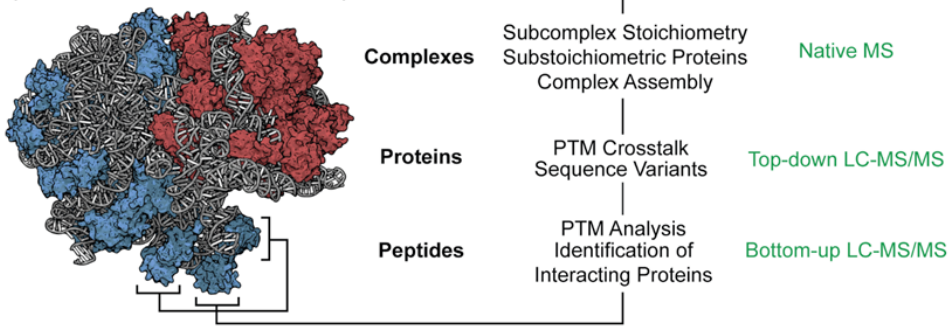
Biomolecular mass spectrometry (MS) has matured substantially over the past decades finding applications in among others biochemistry, molecular and structural biology and systems biology[1]. With its ability to analyze biological systems at multiple levels – whether its metabolites, RNA and DNA, proteins, protein complexes or entire proteomes – the MS toolbox has proven invaluable in a life science research environment[2–4]. While methods like (phospho)-proteomics and metabolomics have firmly settled in the field of cell biology, the use of MS methods for characterizing protein complexes in a structural biology setting is less matured. With recent developments in data analysis for cross-linking MS, this technique is rapidly gaining popularity among structural and systems biologists for its ability to map protein-protein interactions on a global scale[5–8]. However, other MS approaches can also complement structural biology techniques and provide highly useful insight into the assembly and composition of macromolecular assemblies[9]. Here, we describe a three-tiered MS approach for the detailed characterization of protein complexes and highlight its use by characterizing various ribosomal particles from different organisms and organelles.

Ribosomes are large ribonucleoprotein complexes responsible for the translation of messenger RNA (mRNA) into proteins. Their composition and architecture vary along the phylogenetic tree, from eukaryotes to bacteria as well as among the organellar ribosomes, although their functional elements catalyzing the key reactions like the decoding of the mRNA and the formation of the peptide bond are highly conserved[10,11]. Recent developments in structural biology techniques, notably X-ray crystallography and cryo-electron microscopy (cryo-EM), have provided insight into the structure and function of many ribosomal complexes and in combination with biophysical and biochemical data led to a detailed understanding of the translation mechanism[12]. Yet, even with structures of ribosomes from many kingdoms of life and different organelles resolved [13–16], small but potentially important features of ribosomal particles have been mostly overlooked. These features, including specific post-translational modifications (PTMs), sequence variations, binding of protein co-factors or sub-stoichiometric presence of ribosomal proteins, can be elusive to standard structural biology techniques and therefore require the use of complementary approaches, such as mass spectrometry (MS).

Our three-tiered MS approach makes use of a set of MS techniques, which provide information on the composition, assembly and activity of ribosomal particles (Figure 1a). First, bottom-up liquid chromatography-tandem mass spectrometry (LC-MS/MS), a MS technique commonly used in proteomics research, provides the ability to identify and quantify the ribosomal proteins and their PTMs[17]. Moreover, it can determine the presence of ribosome-interacting factors, which have remained bound to the ribosomal particles during their purification[18]. Common bottom-up LC-MS/MS workflows start with unfolding of the proteins followed by their digestion into peptides. These peptide mixtures are separated using high-performance liquid chromatography (HPLC) and sequenced by a mass spectrometer. For the second tier, top-down LC-MS/MS, proteins are denatured but kept intact and separated

by a HPLC system. The intact masses of the different proteins and their co-occurring proteoforms are measured by the mass spectrometer, providing an overview of all the different versions of the gene products such as proteins carrying multiple PTMs. Proteins are identified in top-down LC-MS/MS through top-down sequencing, which can additionally localize PTMs[19–21]. In this way, top-down LC-MS/MS can potentially provide information on proteoforms of ribosomal proteins that would have been lost upon digestion into peptides like the crosstalk between different PTMs[21,22]. The third tier in the approach, native MS, omits even the unfolding step and introduces the intact ribonucleoprotein complexes into the mass spectrometer, after which their masses are measured[23–26]. Because non-covalent interactions are generally preserved in native MS, accurate mass measurements of the complexes can provide detailed insight into their composition, including the stoichiometry of the protein or nucleic acid subunits[27–29]. Moreover, using the latest innovations in mass-analyzers, native MS can resolve and characterize co-occurring assemblies, such as ribosomes with and without an interacting protein or with substoichiometric presence of a ribosomal protein, as was recently demonstrated for prokaryotic ribosomal particles[30].

a



b

	Ec70S	So70S	Hs40S	Hs60S
Organism	*E. coli*	*S. oleracea*	*H. sapiens*	*H. sapiens*
Subcellular location	Cytosol	Chloroplast	Cytosol	Cytosol
Number of proteins	55	57	33	47
rRNA fragments	23S, 5S, 16S	16S, 23S, 4.5S, 5S	18S	28S, 5S, 5.8S
Approximate Mass	2302 kDa	2435 kDa	1212 kDa	2688 kDa

Figure 1 | Three-tiered mass spectrometry-based approach for the dissection and characterization of ribosomal particles. a) The three-tiered hybrid mass spectrometry approach described here can provide insight into multiple molecular levels, ranging from the amino acid sequence of individual ribosomal proteins to the stoichiometric composition of the intact ribonucleoprotein particles. Which tier in the approach provides information on which level is indicated in green. The structural model of the E. coli 70S ribosome (PDB 4YBB) displays the rRNA in grey, the proteins of the 30S subunit in red and of the 50S subunit in blue[77]. b) Information on the ribosomal particles investigated here.

Here, we demonstrate how this three-tiered MS approach can provide in-depth characterization of four distinct ribosomal particles: *E. coli* cytosolic 70S ribosomes (Ec70S), chloroplastic 70S ribosomes from spinach (So70S), and human cytosolic 40S (Hs40S) and 60S (Hs60S) ribosomal subunits (Figure 1b).

RESULTS

Before describing the in-depth analysis of the ribosomal particles by our three-tiered MS approach, we first describe some novel workflows, hardware and software we used for top-down LC-MS/MS making use of a recently introduced mass analyzer, the Orbitrap™HF-X[31].

Improved top-down LC-MS/MS through on-the-fly deconvolution

Standard top-down LC-MS/MS experiments encounter a number of challenges that limit the analytical depth of the analyses. When full MS spectra are acquired to determine which proteoforms are selected for subsequent top-down sequencing events, there is an information redundancy that comes from the repeated selection and fragmentation of the different charge states of the same proteoform. Additionally, current workflows regularly employ full MS scans at high resolving power (R = 120,000 at *m/z* = 200) to determine the proteoform charge state from the isotopic distributions, followed by top-down sequencing also performed at high resolving power (so-called *High-High* workflows). However, on Orbitrap™ mass spectrometers, the short-lived transients of large proteins (>30 kDa) are generally suppressed when competing with the longer-lived transients of smaller proteins, introducing a negative bias in the identification of larger proteins. Additionally, to isotopically resolve the high molecular weight proteins (>60 kDa), longer transient times are required (>500 ms) that are generally less compatible with the LC timescales of top-down LC-MS/MS analysis. By acquiring the full MS spectra at medium resolution (R = 7500 at *m/z* = 200), but the MS/MS data at high resolving power (termed *Medium-High* workflows), the bias is removed and both large and small proteins can be identified by top-down sequencing. At this resolution, the isotopic distributions of the proteoforms remain unresolved. This information is used to calculate their charge states, an essential part of the experiment since this information is required for the determination of fragmentation parameters (e.g. collision voltage), database searching (calculation of the measured mass) and the sequencing of co-eluting proteins (exclusion of different charge states of the same proteoform). We solved these redundancy and bias issues by actively (on-the-fly) deconvoluting the medium-resolution full MS spectra, assigning charge and mass to every peak in the spectrum (Materials and Methods section).

We benchmarked the combination of these *Medium-High* and *High-High* workflows by identifying all 55 ribosomal proteins from the Ec70S ribosome and an additional 12 ribosome-associated proteins (Supplementary Figure 1). The advantage of the *Medium-High* workflow over the *High-High* workflow for high molecular weight proteins is immediately evident when medium- and high-resolution full MS scans of the ribosomal protein S1 (61 kDa) and the peptide chain release factor 2 (prfB, 41 kDa), a ribosomeinteracting protein, are compared (Supplementary Figure 2). While the medium-resolution scans contain clearly resolved charge states, the high-resolution scans suffer from poor signal-to-noise ratio hampering identification. Moreover, the *Medium-High* method allows for shorter duty cycles (Supplementary Figure 3a) and faster processing time of the deconvolution algorithm (Supplementary Figure 3b).

On the other hand, the high-resolution full MS of *High-High* workflows provides a higher number of reliably deconvoluted proteoforms for lower molecular weight proteins (Supplementary Figure 3c) making the two approaches somewhat complementary (Supplementary Figure 4). Therefore, we used a combination of optimized *Medium-High* and *High-High* workflows to perform in-depth top-down LC-MS/MS analysis of the So70S, Hs40S and Hs60S ribosomal particles.

Overview of bottom-up and top-down LC-MS/MS on ribosomes

In Figure 2 we provide an overview of the results of the bottom-up (2a-c) and top-down (2d-f) LC-MS/MS experiments on So70S, Hs40S and Hs60S ribosome samples. On the left, the bottom-up LC-MS/MS data is presented by Intensity Based Absolute Quantification (iBAQ) plots, which rank the detected proteins in the samples by their estimated abundance [32]. Although these plots are not directly suitable for the determination of protein stoichiometry in the complexes, they provide an accurate prediction of the protein abundances in the investigated samples. On the right, the base peak intensity chromatograms of the top-down LC-MS/MS runs are shown, where peaks represent proteins or mixtures of chemically similar proteins eluting from the column and introduced intact into the mass spectrometer, in which *Medium-High* or *High-High* workflows are used to measure their intact mass and perform top-down sequencing. In this way, we could identify all expected 57 ribosomal proteins of the So70S ribosome, 47 ribosomal proteins of the Hs60S and 33 ribosomal proteins of the Hs40S ribosomal subunit. Moreover, we detected several non-ribosomal proteins that were co-purified with the ribosomal particles, e.g. translation factor pY (pY) and ribosome recycling factor (RRF) in the So70S sample. Such a high identification rate lays the foundation for an in-depth investigation of the co-occurring proteoforms in these ribosomal assemblies.

Information on the purity of the sample can be extracted from both the bottom-up and top-down LC-MS/MS measurements. According to the bottom-up LC-MS/MS data, the 50 most abundant proteins in the So70S sample are either ribosomal proteins of the 30S and 50S subunit, which comprise the 70S ribosome, or translation factors (pY and RRF). The main impurities seem to be non-ribosomal, generally high abundant chloroplastic proteins (Figure 2a and Supplementary Table 1). Additionally, a small set of low abundant proteins of the cytosolic 80S ribosome are detected. The majority of the proteins in the Hs40S sample belong to the 40S ribosomal subunit, followed by a number of mitoribosomal proteins of the 39S large subunit (Figure 2b). The Hs60S sample however seems to be most pure since there is a much larger gap between the abundance of the 60S ribosomal proteins and other co-purified proteins (Figure 2c). Additionally, the main impurities here include ribosomal proteins of the 40S subunit and not mitoribosomal proteins. The top-down LC-MS/MS runs show highly comparable results when inspecting the identified proteins. In the human 40S sample, around half of the ribosomal proteins identified were mitochondrial. For the 60S sample, only 12% was mitochondrial while 32% was identified as 40S ribosomal protein.

In this way, other, non-ribosomal protein complexes can also be identified. For instance, top-down LC-MS/MS showed that the purified human 40S ribosome sample

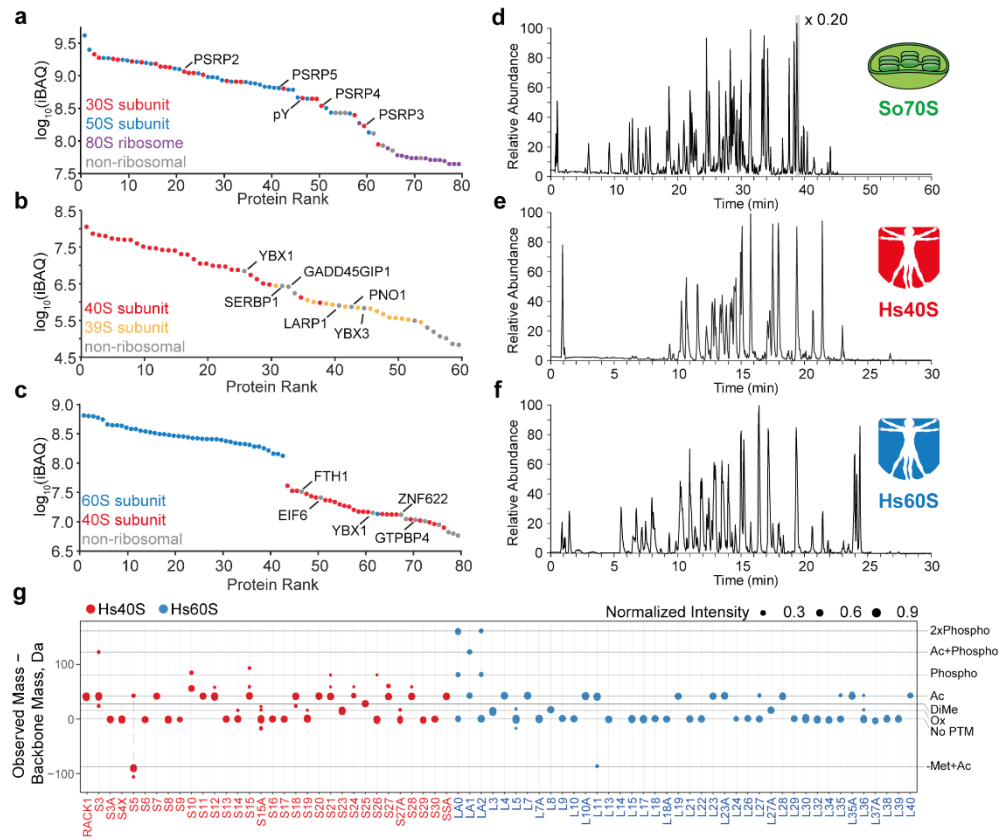


Figure 2 | Dissecting ribosomal particles from different organisms and organelles by bottom-up and top-down LC-MS/MS analyses. a-c) Relative quantification of both ribosomal and non-ribosomal proteins present in the preparations of So70S, Hs40S and Hs60S ribosomal particles, respectively. Protein abundance was estimated by using the intensity based absolute quantification (iBAQ) values of each identified protein. d-f) Top-down LC-MS/MS base peak chromatograms of all proteins in the So70S, Hs40S and Hs60S ribosomal particles, respectively. Top-down LC-MS/MS analysis allows for identification of both ribosomal and non-ribosomal proteins and their distinct proteoforms. g) Relative abundance distributions of proteoforms detected in ribosomal proteins of the 40S and 60S subunit identified by database searching. Each dot represents a proteoform of the gene products listed on the x-axis with increasing size representing increasing relative abundance. The position of the dots along the y-axis shows the deviation of measured mass of the proteoform from the mass calculated from the amino acid sequence. Several commonly occurring post-translational modifications (e.g. acetylation, phosphorylation) are annotated with their corresponding mass shift. The data-point for L14 is missing since the observed mass shifts introduced by a varying number of inserted alanine repeats is outside the scale of the plot (namely: +213 and +355 Da). To stay consistent we adopted the ribosomal protein names from Uniprot entries.

contained multiple protein subunits from the spliceosome (snRNPs E, F, G, SM D1 and SM D2). Characterization of these proteins is not only limited to identification; the snRNP SM D1 was found to be dimethylated 9 times in its glycine- and arginine-rich C-terminus and the fragmentation maps of the snRNP E protein suggest that its initiator methionine is removed and the protein is both acetylated and dimethylated (Supplementary Figure 5)[33]. The combination of top-down and

bottom-up LC-MS/MS allows for better coverage of the proteome of these purified complexes. Bottom-up LC-MS/MS provides great depth to characterize lower abundant proteins and has bias towards identification of larger proteins with no real upper limit in protein length and molecular weight while top-down LC-MS/MS provides a more complete view of the different co-occurring proteoforms of proteins within approximately 5-60 kDa mass range (Supplementary Figure 6).

An overview of proteoforms detected and characterized with top-down LC-MS/MS of ribosomal proteins in the Hs40S and Hs60S subunits is displayed in Figure 2g. For only half of the proteins the intact measured mass agreed with the mass determined directly from the gene sequence. We could identify several instances of PTMs and sequence variations, which are poorly described in most protein databases. These include the removal of initiator methionines (L11, L19, L23, L35a, L30, L36a, S5 and S25) as well as N-terminal acetylation (L23, L35a). Absence of this information from the protein database can in turn lead to incorrect interpretation by standard database searching software, as is exemplified by the S25 protein (incorrect assignment of an N-terminally acetylated proline prevented the identification of a so far unreported dimethylated state of this S25 protein (Supplementary Figure 6). Furthermore, available protein databases such as UniProt and neXtProt mostly lack detailed information on disulfide bridges in human ribosomal proteins. In our LC-MS/MS approach we could chromatographically separate oxidized and reduced forms of human ribosomal proteins L5, S21, S27 and S27a as well as detect their subtle mass differences (Supplementary Figure 7).

Determining sequences of plastid ribosomal proteins

An advantage of bottom-up LC-MS/MS for characterizing proteins is that their initial identification can proceed by detecting just a single or a few unique peptides. Even when the exact full sequences of the proteins are not known or available, for instance because the proteome of the source organism is poorly described, a proteome from a closely related organism can be sufficient. The So70S proteins are partially encoded on the chloroplast genome while some others are encoded on the nuclear genome and have to be imported into the chloroplast. These imported proteins use a transit peptide at the N-terminus to pass the chloroplast membranes, after which this peptide is cleaved off[34]. Identification of these proteins in bottom-up LC-MS/MS is not really hampered by these events since it can be based on the part of the sequence that is not part of the transit peptide. Top-down identification however requires more accurate information because sequence variations in the termini prevent matching of fragment ions with their theoretical expected masses. On the other hand, top-down LC-MS/MS can exactly characterize the fully processed form of the ribosomal proteins, which may be beneficial in fitting the electron density maps of cryo-EM reconstructions.

The spinach proteome is rather poorly described. At present Uniprot contains only 286 reviewed entries from *S. oleracea*, compared to 20,214 from human or 15,423 from the plant *A. thaliana*. Although a significant part of the So70S proteins is described in the database, particularly proteins encoded in the nuclei have incomplete or missing protein sequences[35–37]. For these proteins the Beta Vulgaris Resource

(BvSeq), which contains a set of spinach genome sequencing data, is a helpful source[38]. However, this database does not contain the sequences of a subset of the So70S proteins encoded on the chloroplast genome. Additionally, gene sequences from the two different sources are not always identical and BvSeq does not distinguish between the transit peptide and the protein product sequence.

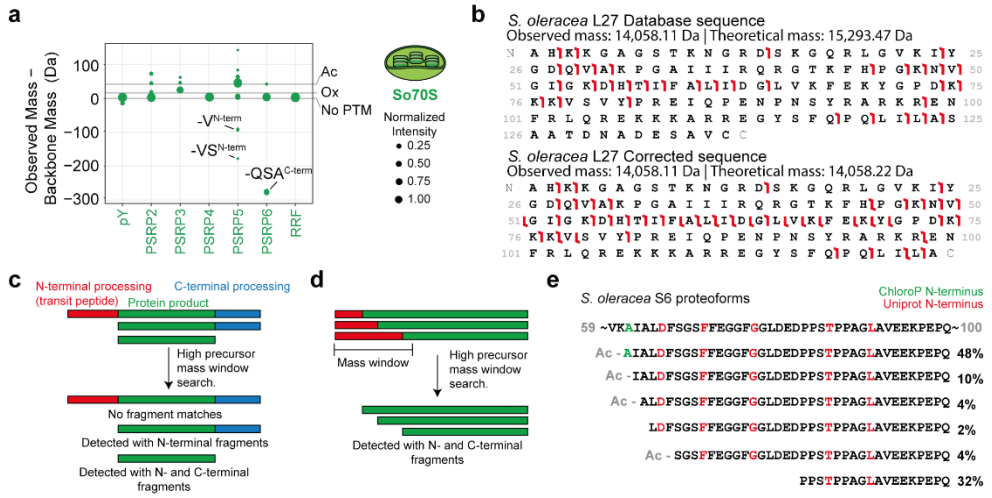


Figure 3 | Top-down analysis using high precursor mass window searches identifies processing of the plastid ribosomal proteins. a) For the plastid-specific ribosomal proteins (PSRPs) of the chloroplastic 70S ribosome, the translation factor pY and the ribosome recycling factor (RRF), several co-occurring proteoforms could be identified and quantified, resulting from post-translational modification and/or processing of the termini. In the corresponding cryo-EM structure of the chloroplastic 70S ribosome, RRF was not modelled due to absence of electron density [15]. b-c) The chloroplastic ribosomal protein L27 is processed at both the N- and C-terminus and is therefore not detected through top-down sequencing when searching for the intact gene product. Corrections to either the N- or C-terminal sequence provides fragment ions from the C- and N-terminus, respectively, as long as the precursor mass window is larger than the mass of the sequence correction. d-e) In a similar way, variable processing at the N-terminus of the chloroplastic ribosomal protein S6 could be detected by using C-terminal fragments to first identify which intact masses belong to S6 and subsequently correct the sequence until the theoretical and measured masses matched.

We analyzed our top-down LC-MS/MS data with Proteome Discoverer Absolute Mass Search, which features database searching with an extended precursor mass window[19]. A sufficiently large window allows the identification of plastid ribosomal proteins even when N-terminal processing is not taken into account, by making use of fragments from the C-terminus. Furthermore, if processing of the N-terminal transit peptide is taken into account it can also identify processing on the C-terminal side correctly. In this way, we attempted to determine the N- and C-terminal processing of the ribosomal proteins in our So70S preparation, as well as their potential PTMs. We managed to identify all 57 ribosomal proteins of the chloroplastic 70S ribosome including five plastid-specific ribosomal proteins (PSRPs) as well as the translation factor pY (formerly PSRP1) and the ribosome recycling factor (RRF) using top-down sequencing and assembled their protein product sequences (Figure 3a, Supplementary Table S2). Notably, in the reported cryo-EM structure of So70S, no

electron density was annotated for RRF while it is readily identified in our top-down LC-MS/MS runs[15].

Although in the majority of the cases the correct sequences of the So70S proteins could be extracted from the combination of Uniprot and BvSeq databases, some proteins required more extensive manual interpretation. Protein L27 is encoded in the nucleus and identification of the protein only occurred when the first 58 amino acids (determined previously to be the transit peptide) were removed (Figure 3b, c). However, removal of these amino acids still left a ~1.2 kDa mass difference between the measured mass and the sequence mass and no C-terminal fragments could be detected. By removing amino acids 182-194 this mass difference could be accounted for and 10 fragment ions from the C-terminus were newly identified. Interestingly, the online transit peptide prediction tool ChloroP also predicts the presence of a cleavage site between amino acids 181 and 182, suggesting similar processing enzymes may have performed the cleavage[39]. Protein S6 is encoded in the nucleus and is reported to exist in multiple forms resulting from distinct transit peptide cleavages (Figure 3d,e). In line with this, the protein was identified with up to 9 kDa mass differences, only C-terminal fragment ions and in 5 different forms with ranging abundance. Uniprot reports 5 different N-termini at positions 66, 71, 77, 86 and 91 while ChloroP predicts transit peptide cleavage sites between amino acids 61 and 62. By combining the intact masses of the S6 proteoforms with their corresponding fragmentation maps we could determine the N-terminal processing in this ribosomal protein. The most abundant form of S6 has Ala-61 as N-terminal amino acid, as predicted by ChloroP, and has additional N-terminal acetylation. Three additional S6 variants could be detected, which all have a single amino acid less at the N-terminus, suggesting they are likely a result of exo-proteolytic processing rather than differential transit peptide cleavages. A fifth form starts with Ser-68 and is N-terminally acetylated while the last, second most abundant form, which features Pro-83 at the N-terminus, lacks acetylation. The data on chloroplastic ribosomal protein S6 nicely illustrate the richness of the ribosomal proteome through co-occurring proteoforms and the benefit of top-down characterization.

Novel cysteine methylations in human ribosomal proteins

The advantage of top-down LC-MS/MS over bottom-up LC-MS/MS is not just limited to the detection of unknown sequence processing events. Ribosomal proteins can also exist in multiple proteoforms as the result of PTMs. These modifications can be detected in bottom-up MS but by digesting the protein information on any potential crosstalk between modified sites is lost. In top-down LC-MS/MS, all proteoforms with distinct masses can be detected based on their intact mass. Information on the site and occupancy of the modifications can be gathered from the fragmentation spectra. A good example of this is the human ribosomal protein S27 that we found to be multiply methylated (Figure 4). Initially, S27 was identified with a mass around 42 Da higher than the mass predicted from the sequence, hinting at acetylation. However, fragmentation maps indicated removal of the initiator methionine leaving an N-terminal proline, making N-terminal acetylation unlikely[40]. Close inspection of the proteoforms identified as S27 revealed an

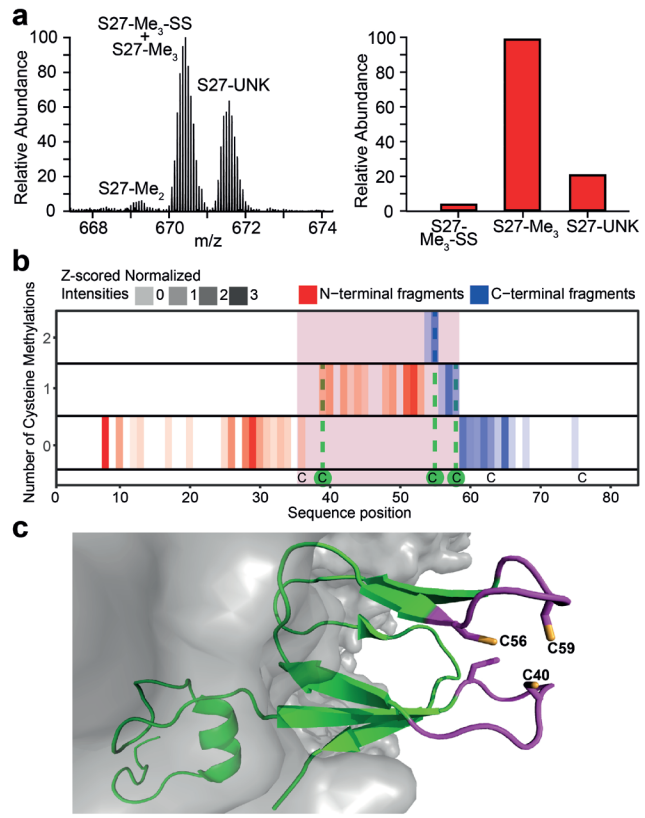


Figure 4 | Human S27 harbors several methylated cysteines in its Zn-finger domain. a) Intact mass spectra of S27 reveals three proteoforms with distinct masses (left). Based on the mass differences from the sequence mass of S27 these are identified as di- and tri-methylated forms of the protein (S27-Me_2 and S27-Me_3) and a third proteoform carrying an unknown modification (S27-UNK). An additional lower abundant disulfide-linked form was detected in fragmentation maps of the trimethylated form (S27-Me_3-SS). b) Top-down analysis of the S27-Me_3 fragmentation spectra pinpoints the methylation to the cysteine residues 40, 56 and 59. Fragments are mapped along the protein sequence (vertical bars) and methylated amino acids are identified by a mass shift of 14 Da. c) Structure of the human 40S subunit with protein S27 in green and its cysteine containing zinc finger domain in magenta (PDB 5A2Q). The zinc-finger cysteines are shown as sticks and the methylated cysteine residues are labeled.

additional variant with a mass difference around 28 Da, albeit lower abundant. Since this suggested methylation rather than acetylation we performed an in-depth analysis of the top-down fragmentation spectra. By comparing the masses of fragments from both the N- and the C-terminus with their theoretical masses based on the S27 sequence, we found fragments of the protein carrying between one and three modifications with a mass of 14 Da. We assigned these modifications as methylations to three cysteines at positions 39, 55 and 58 (Figure 4b). Interestingly, these cysteines are all part of a C4-type zinc finger domain that protrudes from the side of the 40S subunit (Figure 4c). Methylation likely prevents the binding of a zinc ion, which agrees well with the electron density map [41]. Although methylation of cysteine residues in S27 has been reported in yeast ribosomes, methylation was never reported

for human ribosomes, nor to this extent [42]. This is likely caused by the fact that in standard bottom-up LC-MS/MS workflows cysteine methylation is not commonly included in the database search. The role of this modification is unknown although methylation of cysteines in TGF-beta-activated kinases prevented zinc binding and recognition of ubiquitin chains[43]. We could not detect any cysteine methylation in the other C4-type zinc finger proteins present in human cytosolic ribosomes (L37, L37A, S27A and S29) making it unlikely the observed modifications are artifacts of sample handling.

Stoichiometry and composition of ribosomal stalk complexes

Up to now, we focused on the benefits of bottom-up and top-down LC-MS/MS in the characterization of the composition of ribosomal particles. However, with these methods it is hard to gather any information on the stoichiometry of ribosomal proteins in the particles, an area where native MS can play an important role. The ribosomal L7/L12 stalk is a sub-complex within the large ribosomal subunit involved in the binding of several translation factors[44,45]. It is composed of a single copy of the L10 protein bound by multiple dimers of the L7/L12 protein, where L7 is the N-terminally acetylated form of L12. The stoichiometry of the proteins in the ribosomal stalk depends on the number of binding sites for L7/L12 dimers on the flexible tail of L10 and can be predicted based on sequence alignment (Figure 5a)[46]. Additionally, the stoichiometry can be determined unambiguously using a specific native MS experiment, as has been shown previously for ribosomal stalks of *T. maritima* and *T. thermophiles* (heptameric, L10 [L7/L12]$_6$) and *B. subtilis* and *E. coli* (pentameric, L10 [L7/L12]$_4$) [47,48]. Due to the endosymbiotic origin of chloroplasts, the chloroplastic 70S ribosomes have a prokaryotic evolutionary ancestor and we set out to determine the oligomeric state of the chloroplastic stalks. Therefore, we made use of the in-source trapping activation of the recently described QE-UHMR mass spectrometer to release stalk complexes from the chloroplastic 70S ribosomes and measured their mass as being 103.4 kDa (Figure 5b), revealing a heptameric stoichiometry[30,49]. Isolation of ions of the intact stalk complex in the quadrupole and subsequent HCD fragmentation ejected a single L12 subunit that further confirmed this assignment (Figure 5b). We also used top-down LC-MS/MS to identify potential PTMs or processing of the stalk proteins in the Ec70S, So70S and Hs60S ribosomal particles. This revealed that unlike stalks in *E. coli*, chloroplastic ribosomes contain nearly no L7-form of the L12 protein and additionally no protein methylation could be detected (Figure 5c). In human 60S subunits, stalks consist of the equivalent proteins P0 and dimers of P1 and P2. Our data reveals that these proteins are present as different phospho-isoforms harboring either 0, 1 or 2 phosphorylations in about equal abundance in our preparations (Figure 5d).

Monitoring intact ribosomes and their association with RNA

Above we readily showed how native MS can assist in the determination of the oligomeric state of ribosomal stalk protein complexes. However, using the recently introduced QE-UHMR mass spectrometer, we can also perform accurate high-resolution mass measurements of intact ribosomal particles, despite their high RNA

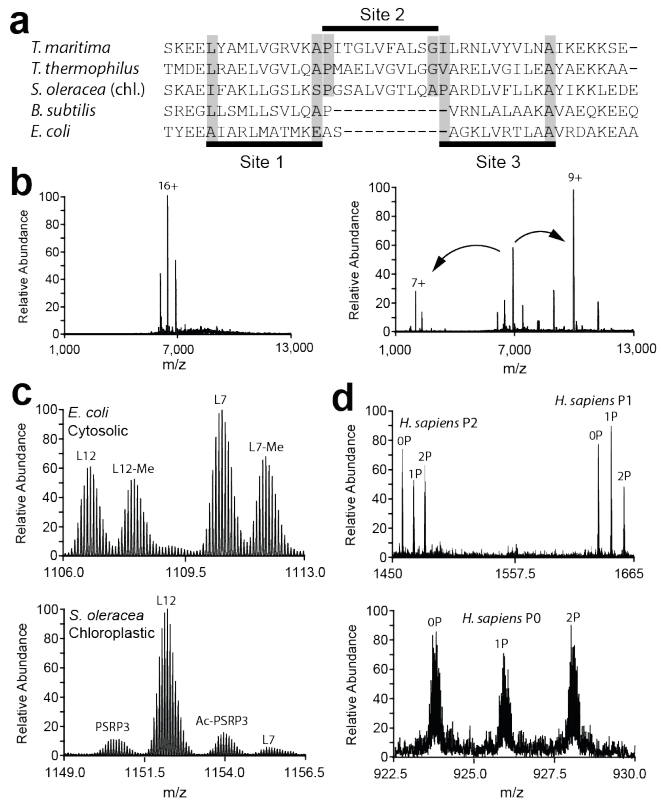


Figure 5 | Characterization of ribosomal stalk complexes; composition, post-translational modification and stoichiometry a) Sequence alignment of the flexible tail region of the ribosomal protein L10 from two thermophiles (*T. maritima* and *T. thermophilus*) and two prokaryotes (*B. subtilis* and *E. coli*) with chloroplastic L10 from *S. oleracea*. The presence of three L7/L12 dimer binding sites predicts chloroplast ribosomal stalks to have a heptameric stoichiometry (L10 [L7/L12]$_6$). b) Determination of the chloroplastic ribosome stalk stoichiometry using a pseudo-MS3 experiment. Chloroplastic 70S ribosomes are introduced into the gas phase and activated in the source region to release intact stalk complexes (top) with a mass that corresponds well with the predicted heptameric stalks. Isolation and further fragmentation of these complexes releases a single copy of the L12 protein confirming the assignment of the oligomeric state. c) A magnificiation of a representative charge state of top-down LC-MS/MS analysisof bacterial (*E. coli)* and chloroplastic (*S. oleracea*) ribosomal stalk proteins shows that unlike its bacterial counterpart, chloroplastic ribosomes contain almost no L7 protein and methylation of L7 or L12 is absent. d) Human stalk complexes in the 60S ribosome consist of the phosphoproteins P0, P1 and P2. A magnificiation of a representative charge state of top-down LC-MS/MS analysis reveals that all three proteins are present in their unphosphorylated (0P), singly phosphorylated (1P) and double phosphorylated forms (2P).

content (~50%) and molecular weights in the MDa range[30,49]. For the human 40S ribosomal subunit we detected several well-resolved charge state distributions around 23,000 *m/z* ratio (Figure 6a). The theoretical mass of the 40S subunit based on the protein and rRNA sequences is 1,209,602 Da. The mass of the most abundant species in this spectrum, 1,215,347 ± 125 Da, deviates significantly (5.7 kDa) from this theoretical mass, more than is expected based on the high resolution of the mass spectrum and the error in our measurements. Inspection of the top-down

LC-MS/MS data of the 40S subunit showed that the ribosomal protein L41 of the large 60S subunit is present at relatively high abundance (Supplementary Figure 9). Furthermore, in EM reconstructions of the 40S subunit, L41 was observed to be associated to the 40S as well[41]. Addition of the mass of the L41 protein (3456.35 Da) to the theoretical mass of the 40S particle already lowered the mass deviation to 1.7 kDa. The remaining deviation can be partly explained by the presence of metal ion or other small molecules to the ribosomes and incomplete desolvation of the ions inside the mass spectrometer. By making use of the high resolving power of the QE-UHMR mass spectrometer, we could confidently identify other, lower abundant 40S ribosomal particles (Figure 6a, red and purple labels) with molecular weights of 1,196,399 +/- 90 Da and 1,201,827 +/- 117 Da, respectively. We assign these to 40S ribosomal particles lacking either a copy of the ribosomal protein S10 (18.9 kDa) or S25 (13.6 kDa), respectively. Interestingly, both of these proteins were also identified as substoichiometric components of human ribosomal particles in a recent selective reaction monitoring study by Shi et al[50]. S25 was found to be mainly absent in polysomes and translation of mRNA transcripts involved in vitamin B12 pathways was influenced by the presence of this ribosomal protein. Regardless, compared to other small ribosomal subunits like Ec30S and So30S, the heterogeneity of the human 40S subunit remains rather low (Supplementary Figure 10) [30].

Motivated by our ability to collect high-resolution native mass spectra of intact human 40S ribosomal subunits and to detect the presence or absence of individual proteins we wanted to see if we could also detect binding of functional RNA molecules to the ribosome. Therefore, we reconstituted complexes of human 40S subunits with internal ribosome entry site (IRES) RNA elements from cricket paralysis virus (CrPV) and hepatitis C virus (HCV). IRES elements are specific RNA sequences that by binding to ribosomal particles in a specific way allow for translation start while skipping certain steps of the canonical initiation process. The elements are frequently present within viral RNAs to stimulate expression of viral proteins over host cell proteins[41,51–53]. We collected native mass spectra of mixtures of free and IRES bound 40S ribosomal particles and monitored the increase in mass upon binding of the RNA, which we determined as 66 kDa for CrPV IRES and 106 kDa for HCV IRES (Figure 6b). Since the sequences of the IRES RNA fragments were not exactly known, we also collected high resolution native mass spectra of the free IRES RNA structures to determine their mass. These masses (CrPV: 64,054.27 Da, HCV: 103,726.19 Da) were in good agreement with the increase in mass of the 40S subunits upon IRES binding, indicating that the ribosomal particles do not undergo a significant change in composition upon binding of the viral RNA elements. Furthermore, these data unambiguously demonstrate the 1:1 stoichiometry of IRES binding to the 40S human ribosome.

DISCUSSION

Structural biology has experienced remarkable developments due to technical advances in cryo-EM over the past years, culminating in being awarded the Nobel Prize in chemistry in 2017[54–56]. This sparked an increase in the number of publications featuring high-resolution structures of many important macromolecular machines[57], and especially expanded our structural knowledge about ribosomal complexes [14,15,58]. At this point, researchers have been able to describe many aspects of the translation process with significant structural detail.

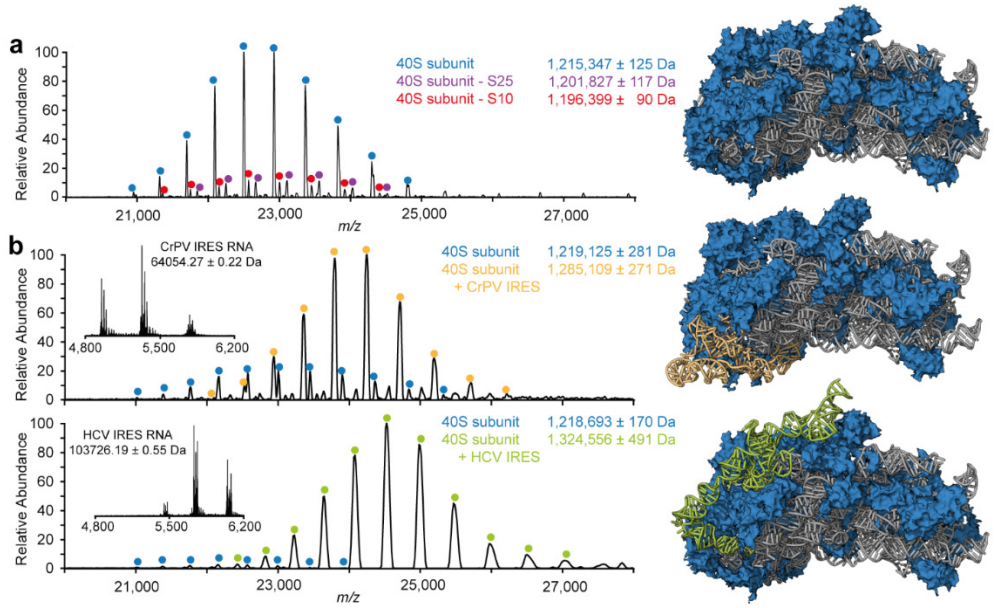


Figure 6 | High resolution native mass spectra of free and IRES bound human 40S subunits. a) Native mass spectrum of human 40S subunit acquired with the recently introduced QE-UHMR mass spectrometer. The well-resolved charge states of three distinct forms of the ribosomal subunit could be detected. The most abundant fully assembled 1.2 MDa 40S particles are labeled in blue, while lower abundant particles lacking either the protein S25 or S10 are labeled in magenta and red respectively. b) Monitoring of the formation of a complex containing human 40S ribosomes and Internal Ribosome Entry Site (IRES) RNA fragments of Cricket Paralysis Virus (CrPV) and Hepatits C virus (HCV). Mass spectra of the RNA fragments alone (insets) provide the accurate mass of the IRES elements. This mass corresponds well with the observed increase in mass of the 40S ribosome upon binding of the RNA indicating that the particles do not undergo a significant change in composition. Structures of the free 40S ribosomes (PDB entry 5A2Q) and particles bound by CrPV (PDB entry 4V91) and HCV IRES (PDB entry 5A2Q) are shown, with the ribosomal proteins in blue, the rRNA in gray and the IRES elements in yellow and green respectively[41,52].

Traditionally, structural biology techniques such as X-ray crystallography and NMR spectroscopy mostly required the production of artificially tagged recombinant proteins in bacterial systems. Cryo-EM, partly through its sensitivity, advanced structural biology into the field of endogenously expressed protein complexes, as is clearly exemplified by the ability of cryo-EM single particle analysis to study macromolecular assemblies purified from native sources. Assemblies from more complex (e.g. mammalian) systems and purified from native sources display a strong increase in structural heterogeneity, both in protein and/or RNA composition and stoichiometry, but especially in the presence of chemical modifications on either the protein or nucleic acid subunits. Although some structural studies paid special attention to the importance of PTMs, the various proteoforms of the protein subunits of large protein complexes are often not detected or ignored, even though their importance

is widely acknowledged[59,60]. Instances of mapping and identifying small chemical modifications in the electron density maps have been reported in literature but they required significant resolution, something which is still rarely reached using current technologies [58]. Similarly, the determination of the stoichiometry of protein and/or RNA components requires large numbers of homogenous particles images and can be hampered by compositional heterogeneity and conformational flexibility of the particle[61,62]. The techniques in the mass spectrometry toolbox can provide crucial information on all these aspects, leading to a more complete picture and understanding of macromolecular assemblies that are the subject of structural studies.

Here, we describe the in-depth characterization of four different ribosomal particles by MS using a three-tiered approach. By making use of the latest technological advances in the field of MS, most notably in top-down LC-MS/MS and native MS, our approach probes several aspects of ribosomal particles generally considered elusive to traditional structural biology methods. Above we highlighted several examples where we used the approach to gain novel insight into protein PTMs, co-occurring proteoforms and processing of ribosomal proteins, protein and RNA stoichiometry and assembly of ribosomal particles. As such, the data described serve as a proof of principle that novel, interesting aspects of large ribonucleoprotein assemblies, as shown here for diverse ribosomal particles, can be discovered using hybrid MS approaches. It lays a foundation for future studies aiming to completely characterize macromolecular complexes. For instance, the cysteine methylations of human S27 could be followed through the different stages of translation or in different organisms. Furthermore, possibly in combination with cryo-EM, more effort should be invested into fully characterizing all the proteoforms of ribosomal proteins, the exact location of these modifications, and their impact on ribosome assembly and function, as has been recently attempted for ribosomal rRNA modifications [58,63–65].

Technological and methodological improvements are continuously made in the field of structural biology and mass spectrometry is moving in parallel. Instruments such as the Q Exactive HF-X and the Q Exactive UHMR used here improve the depth and detail of the information that can be extracted from ribosomal particle analysis. In the future, the approach described here can be supplemented with other parts of the mass spectrometry toolbox. Mainly Top-down LC-MS/MS can benefit from the use of alternative fragmentation techniques such as Electron Transfer Dissociation or Ultraviolet Photodissociation[21,66,67]. These techniques have the potential to prevent proteins with labile modifications such as glycosylation or phophorylation to escape identification and PTM localization and further improve the sequence coverage of other proteins. The presence of PTMs can greatly increase the complexity of the protein mixtures introduced in the mass spectrometer and advanced LC separations (for instance ion-exchange columns[68]) can help tackle these challenges. Additionally, advancements in data analysis software may improve the interpretation of both the raw mass spectra and large amount of information coming from all 3 tiers in the approach. This will make hybrid mass spectrometry approaches the ideal partner for the field of structural biology to completely unravel all details of ribosomal particles and other molecular machines[1,69]. It is evident that structure function relationships of life's cellular machineries can only be fully elucidated by the use of multiple or hybrid technologies, whereby mass spectrometry may become an indispensable pillar.

MATERIALS AND METHODS

Names of ribosomal proteins

Throughout the manuscript we used a short version of the Uniprot entries for ribosomal proteins. Although the nomenclature for ribosomal proteins suggested by Ban et al [70] has been adopted in most recent structural studies, the frequent use of database searching in Uniprot in this manuscript prevented us from using this naming system.

Purification of Ec70S, So70S, Hs40S and Hs60S ribosomal particles

E. coli 70S ribosomes were purchased from New England Biolabs.

Chloroplastic 70S ribosomes were purified as previously described in detail by Bieri et al[15]. In brief, chloroplasts were extracted from fresh leaves of spinach (*S. oleracea*) and lysed by gentle stirring in lysis buffer (10 mM TrisHCl pH 7.6, 25 mM KCl, 25 mM $MgCl_2$, 2 mM DTT, 0.1 mM PMSF, 2 mM spermidine, 0.05 mM spermine, 2% (w/v) Triton X-100). The lysate was cleared by centrifugation (25,350 *g*, 30 min, 4°C) using a Beckman Type 45Ti rotor (Beckman-Coulter). The supernatant was loaded onto 50% (w/v) sucrose cushion and centrifuged (101,390 *g*, 15 h, 4°C) using a Beckman Type 45Ti rotor (Beckman-Coulter). The ribosome pellets were dissolved in monosome buffer (25 mM TrisHCl pH 7.6, 25 mM KCl, 25 mM $MgOAc_2$, 2 mM DTT, 2 mM spermidine, 0.05 mM spermine), layered onto 10–40% (w/v) sucrose gradient and centrifuged (51,610 *g*, 15 h, 4°C) using a Beckman Type SW-32Ti rotor (Beckman-Coulter). The fractions containing the chloroplastic 70S ribosomes were pooled and concentrated using Amicon Ultra-4 centrifugal filter units with 100 kDa molecular weight cutoff (Merck Millipore). Aliquots of the So70S sample were flash-frozen in liquid nitrogen and shipped for MS analysis by a dry shipper (Taylor-Wharton).

Human 40S and 60S ribosomal subunits were purified similarly to previously described protocols to isolate human 80S ribosomes[41] and human 40S ribosomal subunits[71]. In brief, approximately $8.5x10^9$ frozen HEK293-6E cells were thawed and lysed by gentle stirring in lysis buffer (50 mM HEPESKOH pH 7.6, 300 mM NaCl, 6 mM $MgAc_2$, 0.5% (w/v) NP-40, 5 µM E-64, 20 µM Leupeptin, 20 µM Bestatin, 5 µM Pepstatin A, 1 mM PMSF and 2 mM DTT). The lysate was cleared by centrifugation (45,000 g, 20min, 4°C) using a SS-34 rotor (Sorvall). The supernatant was loaded onto 60% (w/v) sucrose cushion and centrifuged (257,000 g, 20 h, 4°C) using a Beckman Type 70Ti rotor (Beckman-Coulter). The ribosome pellets were dissolved in resuspension buffer (50 mM HEPESKOH pH 7.6, 150 mM KCl, 6 mM $MgAc_2$, 2 mM DTT), layered onto 12-48% (w/v) sucrose gradient prepared with dissociation buffer (50 mM HEPESKOH pH 7.6, 500 mM KCl, 6 mM $MgAc_2$, 2 mM DTT) and centrifuged (78,000 g, 18.5h, 4°C) using a Beckman Type SW-32Ti rotor (Beckman-Coulter). The bands containing the human 60S and 40S ribosomal subunits were extracted using a syringe and the fractions from several gradients were pooled and concentrated using Amicon Ultra-15 centrifugal filter units with 100 kDa molecular

weight cutoff (Merck Millipore). Aliquots of the Hs40S and Hs60S ribosome samples were flash-frozen in liquid nitrogen and shipped for MS analysis by a dry shipper (Taylor-Wharton).

Purification of IRES RNA sequences

The CrPV IRES RNA and the HCV IRES RNA were obtained similarly as previously described by Quade et al[41]. In brief, the IRES RNAs were produced by *in vitro* transcription of a linearized plasmid containing the CrPV IRES and HCV IRES sequence, respectively, followed by denaturing polyacrylamide gel electrophoresis. The IRES RNAs were extracted from the gel and the urea containing buffer was exchanged to water using Amicon Ultra-15 centrifugal filter units with 10 kDa molecular weight cutoff (Merck Millipore). Aliquots of the IRES RNA samples were flash-frozen in liquid nitrogen and shipped for MS analysis by a dry shipper (Taylor-Wharton).

Sample Preparation for bottom-up LC-MS/MS analysis

Hs40S, Hs60S and So70S ribosome preparations were reduced with 5 mM DTT at 56 °C for 30 minutes and alkylated with 15 mM iodoacetamide at room temperature for 30 minutes in the dark. The excess of iodoacetamide was quenched by adding 5 mM DTT. Digestion of intact proteins was performed at 37 °C with Lys-C for 4 hours followed by overnight digestion with trypsin at an enzyme-to-protein-ratio of 1:100 (w/w). All proteolytic digests were desalted, dried and dissolved in 40 µL of 0.1% FA prior to LC-MS/MS analysis.

Bottom-up LC-MS/MS analysis

Separation of digested protein samples was performed on an Agilent 1290 Infinity HPLC system (Agilent Technologies, Waldbronn. Germany). Samples were loaded on a 100 µm x 20 mm trap column (in-house packed with ReproSil-Pur C18-AQ, 3 µm) (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) coupled to a 50 µm x 500 mm analytical column (in-house packed with Poroshell 120 EC-C18, 2.7 µm) (Agilent Technologies, Amstelveen, The Netherlands). A 2 µL sample of peptides was used, corresponding to ~0.1 µg of material. The LC-MS/MS run time was set to 60 min with flow rate of 300 nL/min. Mobile phases A (water/0.1% formic acid) and B (80% ACN/0.1% formic acid) were used for 45 minutes gradient elution: 13 to 40% B for 35 minutes, and 40 to 100% B over 3 minutes. Samples were analyzed on a Thermo Scientific Q Exactive™ HF quadrupole-Orbitrap instrument. Nano-electrospray ionization was achieved using a coated fused silica emitter (New Objective, Cambridge, MA) biased to 2 kV. The mass spectrometer was operated in positive ion mode and the spectra were acquired in the data-dependent acquisition mode. Full MS scans were acquired with 60,000 resolution (at 200 *m/z*) and at a scan mass range of 375 to 1,600 *m/z*. Automatic Gain Control (AGC) target was set to 3e6 with maximum injection time of 20 ms. Data dependent-MS/MS (dd-MS/MS) scan was acquired at 30,000 resolution (at 200 *m/z*) and with mass range of 200 to 2,000 *m/z*. AGC target was set to 1e5 with maximum of injection time defined at 50 ms. 1 µscan was acquired both for full MS and dd-MS/MS scans. Data dependent method was set to isolation and fragmentation of the 12 most intense peaks defined in full

MS scan. Parameters for isolation/fragmentation of selected ion peaks were set as follows: isolation width = 1.4 *Th*, HCD normalized collision energy (NCE) = 27%.

Preparation of ribosomal proteins for top-down LC-MS/MS

Approximately 150 µg of ribosomal proteins and ribosome associated proteins were separated from the ribosomal RNA by glacial acetic acid precipitation according to Hardy et al[72]. Briefly, to vigorously shaken ribosomal particles magnesium acetate was added to around 100 mM. Immediately afterwards 2 volumes of glacial acetic acid were added. This solution was left shaking at 4 °C for 60 minutes followed by centrifugation for 10 minutes at 20,000x g. The supernatant was moved to a new vial and the pellet was washed with 66% glacial acetic acid containing 100 mM magnesium acetate. After repeating the centrifuge step the two supernatants were combined and the buffer was exchanged to buffer A (0.1% v/v formic acid in water).

Top-down LC-MS/MS analysis

Chromatographic separation of intact protein samples was conducted on a Thermo Scientific Vanquish Horizon UHPLC system equipped with MAbPac RP 2.1 mm x 50 mm column. Around 0.5-5 µg of material was loaded on the column heated to 80 °C. LC-MS/MS runtime was set to 40 or 70 minutes with flow rate of 250 µL/min. Gradient elution was performed using mobile phases A (water/0.1% formic acid) and B (ACN/0.1% formic acid): 10 to 50% B for 30 or 60 minutes, and 50 to 80% B over additional 4 minutes.

All top-down MS experiments were performed on a Thermo Scientific Q Exactive HF-X instrument[31]. The instrument provides an array of new features facilitating top-down analysis of complex samples. Among them, Advanced Peak Detection (APD) algorithms, that allow for on-the-fly deconvolution of monoisotopic or average masses with improved charge detection. Along with extended charge state range (up to z = 100+) for calculation of optimum HCD collision energy, APD provides more efficient peak picking and fragmentation of highly-charged ions of intact proteins. Improved ion optics in the front-end of the instrument result in brighter ion source as compared with other instruments of the Q Exactive series, which provides means for less sample loads. Additionally, transmission of higher molecular weight ions (*i.e.* intact proteins) is improved through modified electronics and gas regime in the back-end of the instrument. LC-MS/MS data were collected with instruments set to the Intact Protein Mode.

For analysis of intact proteins two primary methods were employed, *Medium-High* and *High-High*. The former method implies switching of resolution parameter from 7,500 (at 200 *m/z*) for full MS scan to 120,000 (at 200 *m/z*) for dd-MS/MS. The *High-High* approach defines the resolution parameter at 120,000 (at 200 *m/z*) for both full MS and dd-MS/MS. Full MS scans in both methods were acquired for the range of 400 to 2,000 *m/z* with AGC target set to 3e6. Maximum of injection time was defined at 250 ms with 5 µscans recorded in *High-High* approach and at 50 ms with 10 µscans – in *Medium-High*. Data dependent strategy was focused on the three most intense proteoforms defined in full MS scan by Advanced Peak Detection algorithms. Shortly, masses corresponding to charge states of the same proteo-

form were deconvoluted with only the single most intense charge state selected for isolation/fragmentation in dd-MS/MS and all other charge states excluded from candidate list for user-defined exclusion time, which was optimized for each sample individually. Selected ions were isolated with 2 *Th* window. Collision energy applied in dd-MS/MS was normalized for the *m/z* and charge state of selected ion with final setting of 30-35%. All the dd-MS/MS scans were recorded at the mass range of 200 to 2,000 *m/z* with AGC target set to 3e6 and maximum of injection time defined at 250 ms. A total of 5 µscans was recorded per scan.

Preparation of IRES bound human 40S ribosomal particles

Folding of IRES RNA fragments and formation of complexes with Hs40S ribosomal particles was performed according to Quade et al[41]. Briefly, IRES RNA fragments of hepatitis C virus and cricket paralysis virus were diluted to 5 µM in folding buffer (20 mM Tris pH 7.5, 100 mM potassium acetate, 2.5 mM magnesium acetate and 250 µM spermidine) and folded by two cycles of heating to 95 °C and cooling on ice for 1 minute each. Human 40S ribosomal particles were buffer exchange into binding buffer (20 mM HEPES pH 7.6, 100 mM KCl, 5 mM $MgCl_2$ and 2 mM DTT) through 2 cycles of concentration and dilution using a 100 kDa molecular weight cutoff centrifuge filter. Hs40S particles were mixed with a 2-fold molar access of IRES RNA in binding buffer and incubated for 5 minutes at 37 °C and stored on ice for further use.

Preparation of ribosomal particles and IRES RNA fragments for native MS

Hs40S, Hs40S-IRES and So30S particles were prepared for native mass spectrometry by buffer exchanging into native MS buffer using cycles of concentration and dilution with a 10 kDa molecular weight cutoff centrifuge filter. For Hs40S particles, 1 M ammonium acetate pH 7.6 with 0.5 mM magnesium acetate was used. For So30S and Hs40S-IRES particles, 150 mM ammonium acetate pH 7.6 with 0.5 mM magnesium acetate was used. Samples were introduced into the mass spectrometer at a concentration of 100 nM after dilution with native MS buffer which contained 25 mM triethylammonium acetate pH 7.6 Hs40S and Hs40S-IRES particles.

Folded IRES RNA fragments were buffer exchanged to 150 mM ammonium acetate pH 7.6 using a 6 kDa Bio-Rad micro Bio-Spin column and introduced into the mass spectrometer at a concentration of 0.5 µM.

Native MS analysis using a QE-UHMR mass spectrometer

Samples were introduced into the Q Exactive mass spectrometer with Ultra High Mass Range detection capability (QE-UHMR) mass spectrometer with gold-coated borosilicate capillaries prepared in house[30,49]. The following mass spectrometer settings were typically used. Capillary voltage: 1350 V in positive ion mode. Collision gas: Xenon. Automatic gain control (AGC) mode: Fixed. Noise level parameter: 2. Ion transfer optics (injection flatapole, inter-flatapole lens, bent flatapole, transfer multipole and C-trap entrance lens) and voltage gradients throughout the instrument were tuned for every analyte specifically. Instrument calibration was performed using cesium iodide clusters up to 11,000 *m/z*. For Hs40S and Hs40S-IRES

complexes, HCD voltage was between 250 and 300 V. For IRES-RNA fragments, HCD voltage was between 80 and 120 V. For the detection of stalk complexes of So70S ribosomes, the in-source-trapping activation voltage was optimized for maximal release and transmission of the stalks without fragmenting them further.

Spectra were viewed in Xcalibur QualBrowser software (Thermo Fisher Scientific). Masses were determined manually by minimization of the error over the charge-state envelope from different charge-state assignments. The theoretical mass of the Hs40S particle was calculated by extracting the protein sequences from UniProtKB database (http://www.uniprot.org/uniprot/) and the rRNA sequences from NCBI.

Data analysis for bottom-up LC-MS/MS

Raw LC-MS/MS data was interpreted with the Byonic software suite (Protein Metrics Inc.)[73]. The following parameters were used for data searches: precursor ion mass tolerance, 10 ppm; product ion mass tolerance, 20 ppm; fixed modification: Cys carbamidomethyl; variable modification: Met oxidation. Enzymatic specificity was set to trypsin. Searches were made against UniProtKB/Swiss-Prot human and spinach proteome sequence databases. Intensity-based absolute quantification (iBAQ) values were obtained with MaxQuant software (version 1.5.6.0)[74].

Database generation for top-down LC-MS/MS analysis of Ec70S, So70S, Hs40S and Hs60S

Database searching for top-down LC-MS/MS analysis of Ec70S, Hs40S and Hs60S was performed using the Escherichia coli (strain K12) and Human XML format proteomes from UniProtKB.

Database searching for So70S ribosomal proteins was performed using a custom database assembled by combining sequences from Spinacia Oleracea in UniProtKB and the BvSeq resource (http://bvseq.molgen.mpg.de). The sequences were combined in FASTA format without processing of the transit peptides.

Databases imported from XML format files in ProsightPC (Ec70S, Hs40S and Hs60S) were treated as follows. Initiator methionine removal and N-terminal acetylation was allowed as well as other PTMs, up to 13 features or 70 kDa of features in mass per sequence. For Spinacia oleracea databases, no PTMs or other modifications were included in the search space.

Data analysis for top-down LC-MS/MS

Isotopically-resolved and unresolved spectra obtained in top-down LC-MS/MS experiments of intact ribosomal proteins were deconvoluted using Xtract[75] or ReSpect algorithms (Thermo Fisher Scientific, Bremen, Germany), respectively. Automatic searches were made in Thermo Proteome Discoverer software (version 2.2.0.388) with use of ProSightPD nodes for *Medium-High* and *High-High* experimental workflows. Parameters for *Medium-High* method were set as follows. ReSpect parameters: precursor *m/z* tolerance – 0.2 *Th*; relative abundance threshold – 10%; precursor mass range – 5-100 kDa; precursor mass tolerance – 30 ppm; charge state range – 5-100. Xtract parameters: signal-to-noise (S/N) threshold – 3; *m/z* range –

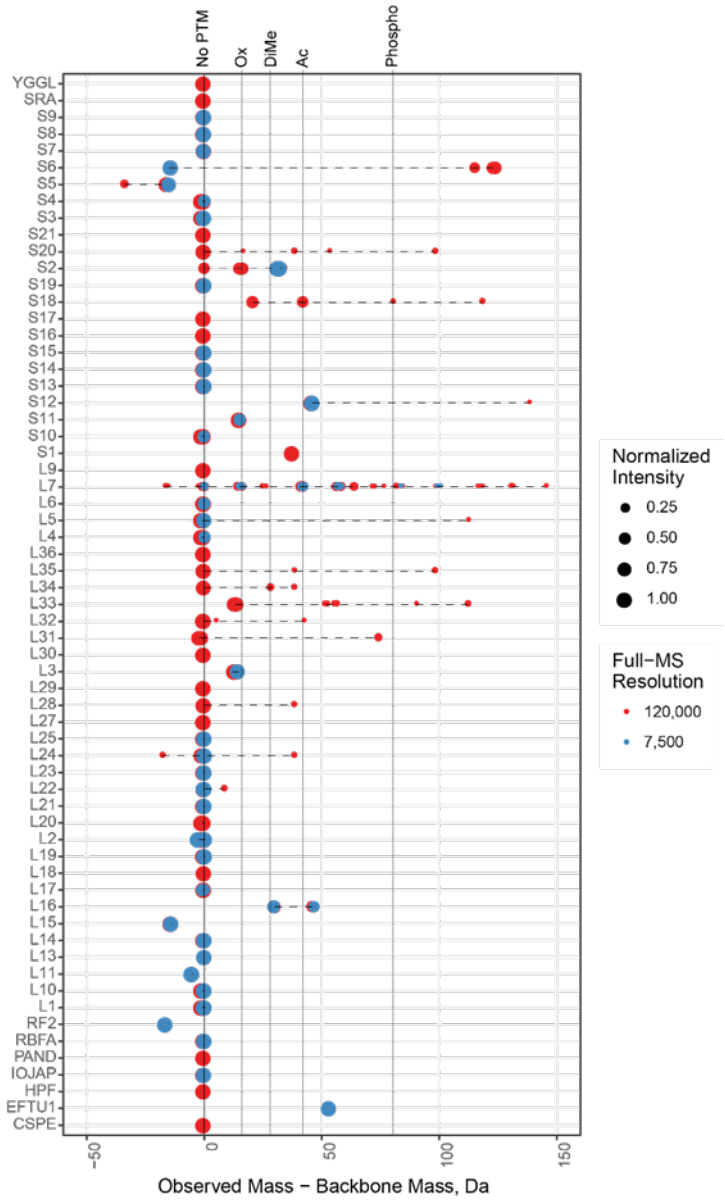
200-2000 *Th*. Absolute mass search parameters: precursor mass tolerance – 500 Da; fragment mass tolerance – 10 ppm. For *High-High* searches ReSpect parameters were not defined, instead Xtract with identical parameters was used to deconvolute spectra in both full MS and dd-MS/MS scans.

For validation of novel PTMs, HCD-MS/MS scans of the same proteoform were manually combined and fragments were assigned using in-house built fragment matching software. Intensities of assigned fragments were z-scored, where mean intensity was subtracted from each fragment's intensity and resulting value was divided by standard deviation of the population. The same approach was employed to characterize ribosomal proteins not detected with automated database searches. Data visualization was conducted in R extended with ggplot2 package[76]. For proteoform overview plots, monoisotopic or average masses of proteins were extracted from the precursor mass lists in automated database searches and matched with deconvoluted mass lists from MS only LC-MS experiments (with a mass tolerance window of 1 Da). Mass differences, observed mass – backbone mass (derived from UniProt sequences), were used to represent proteoforms of ribosomal proteins. Proteoform intensity was normalized on sum of proteoform intensities for each protein.

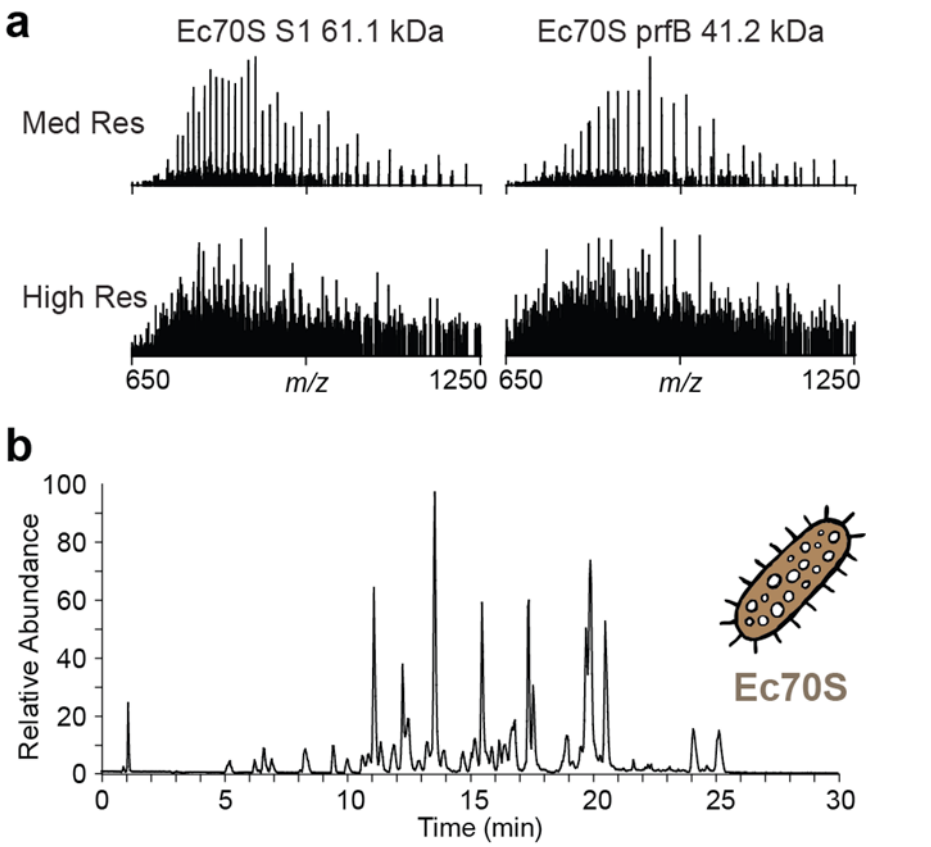
Estimation of false discovery rates in Top-down LC-MS/MS database searches

In order to estimate false discovery rates (FDR) in Ec70S, Hs40S and Hs60S ribosomal particles, parallel searches were performed against both normal and shuffled proteome databases of *E.Coli* (strain K12) and *Homo Sapiens*, respectively. For So70S a normal and shuffled customized database including 77 proteoforms was used. In the cases of Ec70S and Hs40S ribosomal particles there were no protein spectral matches (PrSMs) observed when searched against reshuffled databases. For Hs60S and So70S there were 4 and 2 false positive PrSMs detected against reshuffled databases, respectively. Further analysis showed the false discovery rate to be below 0.25% (Supplementary Figure 8).

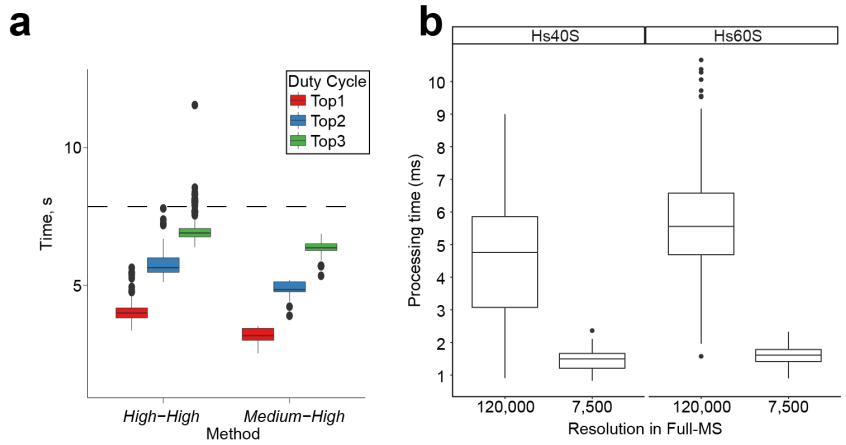
SUPPLEMENTARY MATERIAL

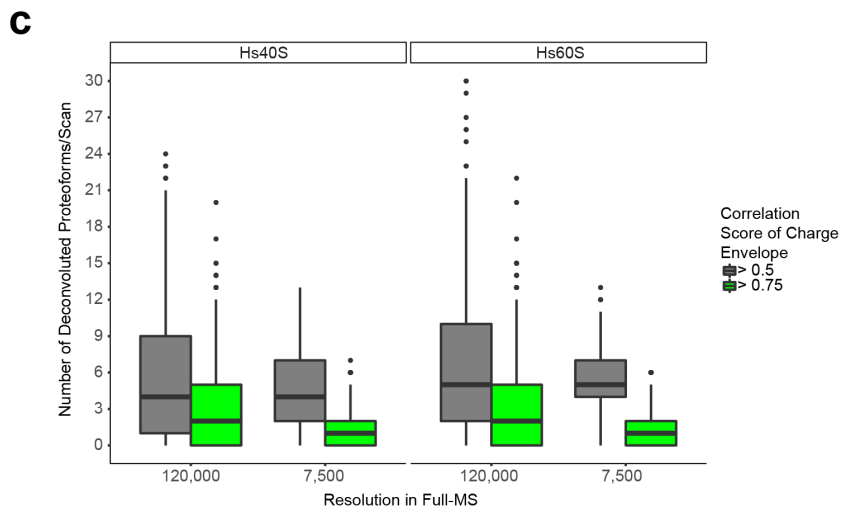


Supplementary Figure 1 | Overview of post-translational modifications observed in the top-down analysis of *E. coli* ribosomal particles. The abundances of different proteoforms of ribosomal and non-ribosomal proteins detected by top-down LC-MS/MS in the Ec70S sample are indicated by sized filled circles.

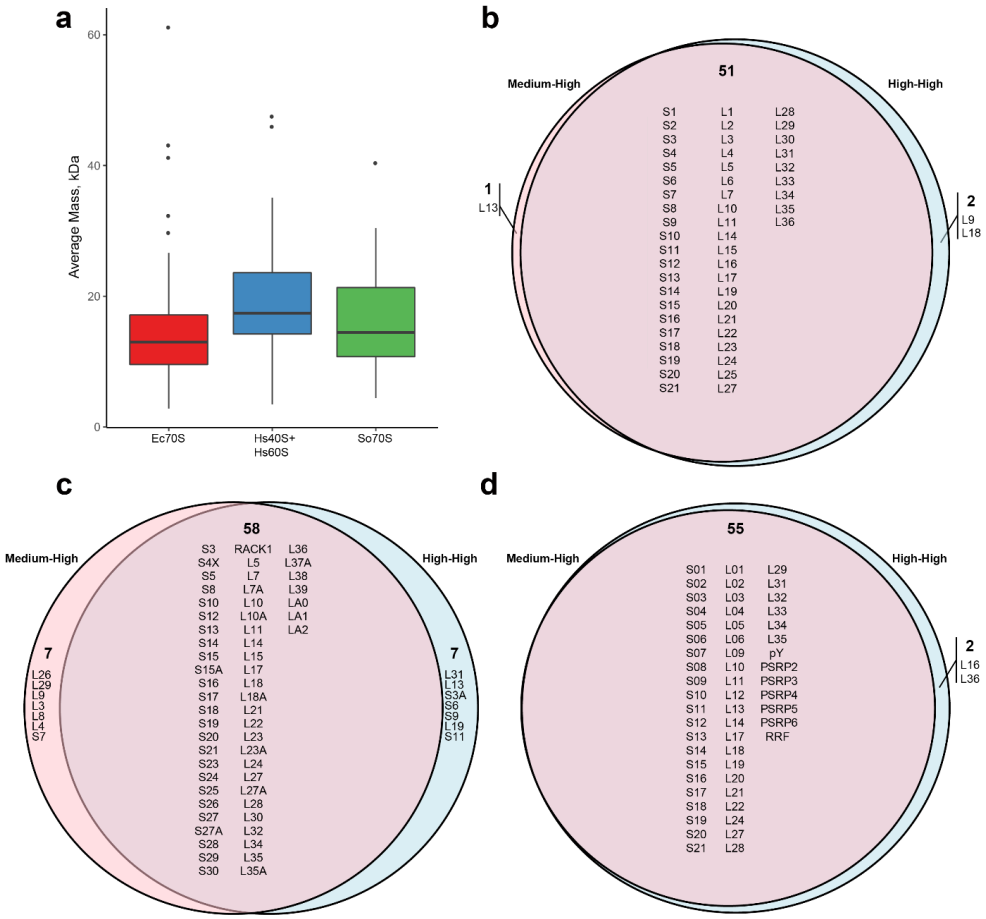


Supplementary Figure 2 | Performance comparison of *Medium-High* and *High-High* workflows in top-down analysis of Ec70S ribosomes. a) Top-down LC-MS/MS workflows with medium resolution MS1 and high resolution MS2 (*Medium-High* analysis) provide better coverage of the ribosomal proteome. Short transient scans, although they have lower resolution, provide better signal to noise ratio than long transient, high resolution scans. b) This allows larger proteins to be selected for fragmentation more frequently, covering the entire *E. coli* 70S ribosome proteome, as can be seen in a typical base peak chromatogram.

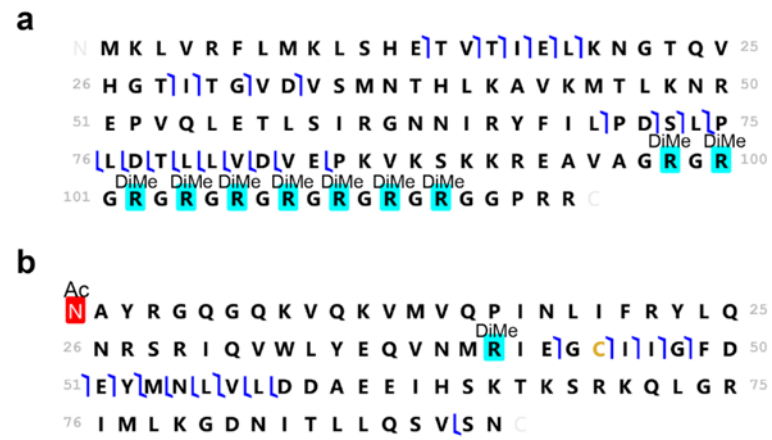




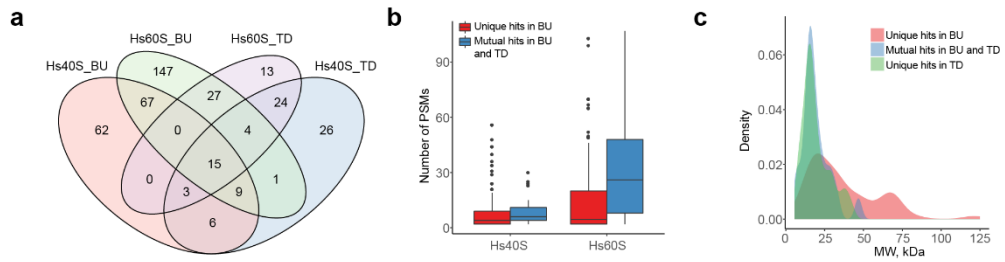
Supplementary Figure 3 | Performance comparison of *Medium-High* and *High-High* workflows in top-down analysis of Hs40S and Hs60S ribosomes. a) Duty cycle times for TopN (N = 1, 2, or 3) *High-High* and *Medium-High* methods compared to the shortest elution window observed for human ribosomal proteins. The data indicate that the *Medium-High* method provides means for faster sequencing of proteoforms in samples of higher complexity. b) Processing time of on-the-fly deconvolution algorithms for isotopically-resolved full MS spectra (recorded at 120,000 resolution setting) and unresolved spectra (recorded at 7,500 resolution setting) for human ribosomal particles. The processing time of the algorithm for high-resolution spectra increases with increase of sample complexity, while the deconvolution algorithm for unresolved spectra proves to be faster and shows less dependence on sample complexity. c) The detected number of deconvoluted proteoforms per scan demonstrates that high-resolution full MS is more sensitive to picking more reliable proteoforms. The correlation score reflects how well the observed charge envelope resembles the predicted charge envelope.



Supplementary Figure 4 | Comparison of detected proteins by *Medium-High* and *High-High* workflows in top-down analysis. a) Distribution of the theoretical average masses of the proteins in the analyzed ribosomal particles. b-d) Venn diagrams showing the ribosomal proteins identified by database searching with Prosight PD by using either *High-High* (blue) or *Medium-High* workflows (pink). The results are shown for the distinct ribosomal particles: b) Ec70S, c) Hs40S and Hs60S, and d) So70S.

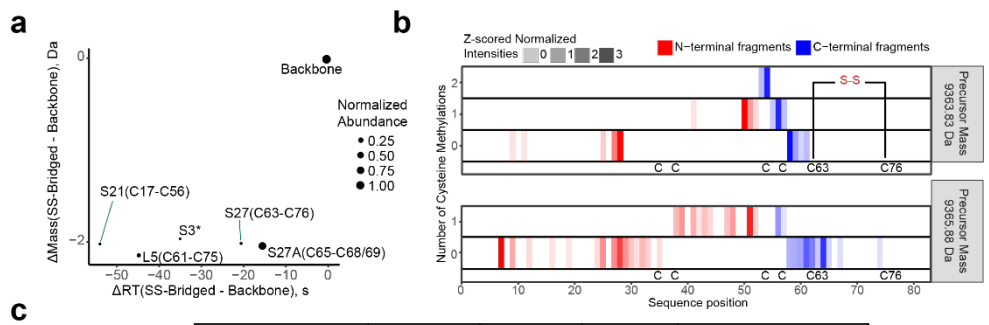


Supplementary Figure 5 | Fragmentation maps of two human spliceosome proteins co-purified in the human ribosomal purification. a) Small ribonucleoprotein D1 with manually assigned 9 arginine dimethylations, localized in the Gly-Arg rich C-terminus. b) Small ribonucleoprotein E with removed methionine and unreported dimethylation. For both proteins, unambiguous determination of the exact site locations of the post-translational modifications requires more extensive analysis with a higher number of fragments.



Supplementary Figure 6 | Comparison of protein identifications in top-down and bottom-up analyses of human ribosomal samples. a) Venn diagram of proteins identifications in bottom-up (BU) and top-down (TD) experiments of Hs40S and Hs60S samples. b) Proteins identified in top-down comprise only fraction of bottom-up hits with high number of peptide spectrum matches (PSMs), while proteins identified with fewer PSMs were detected uniquely in bottom-up. c) Density plot for mass distribution of proteins mutually or uniquely identified in bottom-up and top-down LC-MS/MS. For all comparisons single bottom-up LC-MS/MS run results were compared with merged results from 3-6 top-down LC-MS/MS runs. Only proteins with at least two unique peptides identified and high confidence of identification were taken into account.

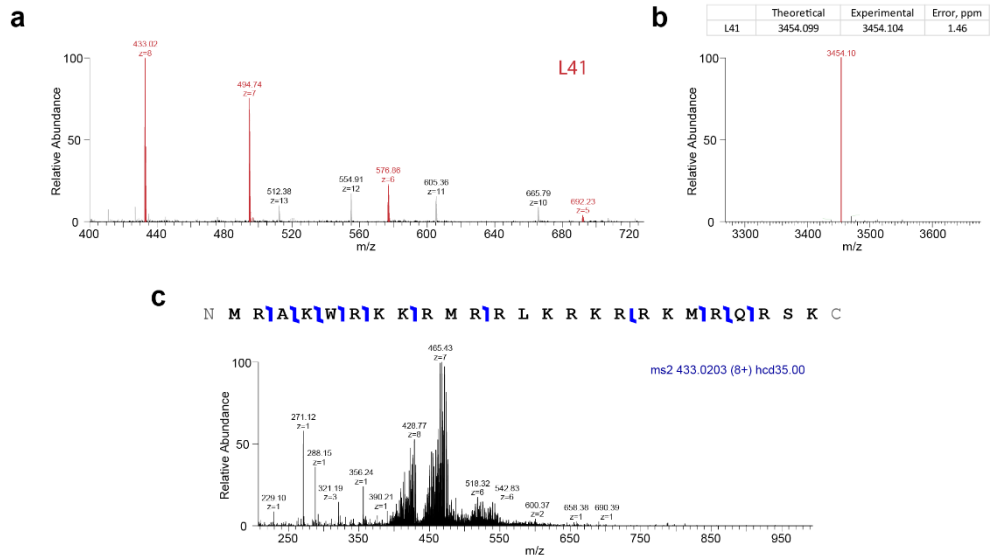
Supplementary Figure 7 | Manual annotation reveals Human S25 to be dimethylated at Lys-3. a) Incorrectly assigned proteoform of the human S25 as N-terminally acetylated in Prosight PD *versus* b) correct positioning of an unreported dimethylation at Lysine 3. Upon manual annotation the P-score for identification of S25 protein improved from 2e-11 to 1.1e-52 for the most intense protein spectrum match.



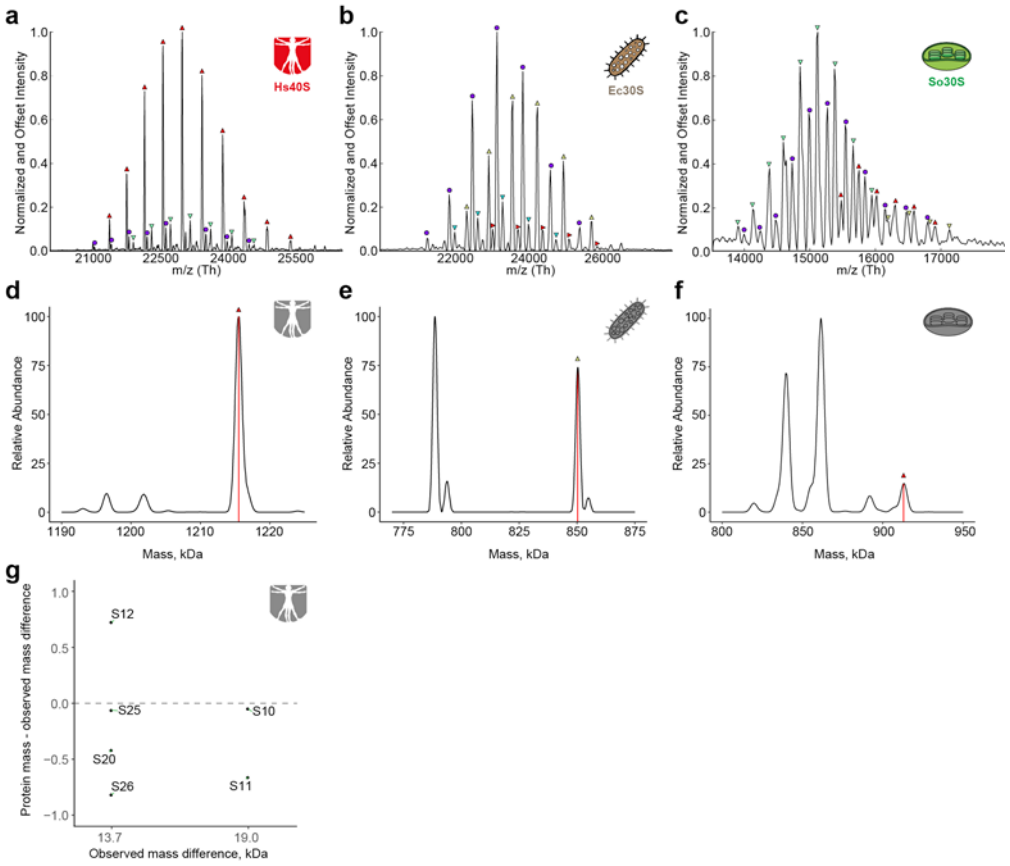
c

	Observed	Theoretical	Error, ppm	Composition
3CysMet	9365.8762	9365.8422	3.64	C411H671N121O115S7
3CysMet+S-S Bridge	9363.8493	9363.8266	2.44	C411H669N121O115S7

Supplementary Figure 8 | Top-down analysis reveals co-occurrence of reduced and non-reduced disulfide bridges in human ribosomal proteins. a) Disulfide linked proteoforms can be distinguished from their reduced forms by shifts in retention time and intact mass in Top-down LC-MS/MS as is shown for 5 human ribosomal proteins (S27, S3, S21, S27A and L5). The dot size displays abundance relative to abundance of the primary proteoform. b) Fragmentation of LC-separated disulfide-bridged and primary proteoforms of S27 human ribosomal protein. c) Accuracy of deconvoluted precursor masses of S27 with and without disulfides further supports the unambiguous identification and characterization.



Supplementary Figure 9 | Presence of L41 ribosomal protein in Hs40S ribosomal purification confirmed by top-down LC-MS/MS. a) Full MS spectrum of human ribosomal protein L41. b) The extracted deconvoluted mass of L41. c) Fragmentation map and HCD MS2 spectrum of isolated z = 8+ ions of L41.

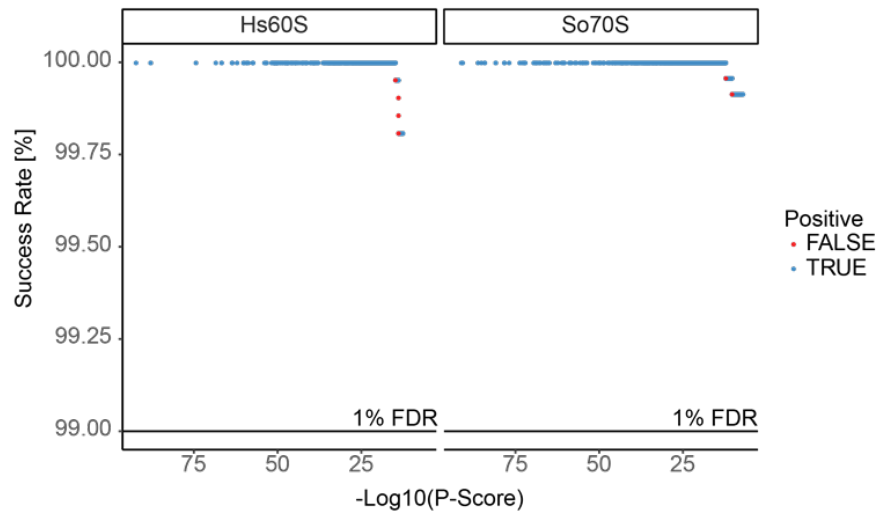


Supplementary figure 10 | Heterogeneity in small ribosome particles revealed by native MS. a-c) Native mass spectra of Hs40S, Ec30S and So30S particles with distinct charge state distributions labelled. The Ec30S data were originally reported by van de Waterbeemd et al[30]. d-c) Deconvoluted native mass spectra of Hs40S, Ec30S and So30S particles. The mass that is considered a fully assembled particle is labelled with a red line. From this data it is evident that the level of compositional heterogeneity varies significantly over the different particles. The most heterogeneous particle, So30S, requires further investigation by native MS to fully understand the different sources of the heterogeneity. g) The 13.7 and 19 kilo-Dalton mass differences in the Hs40S native MS spectrum can be confidently linked to the proteins S25 and S10.

a

	Full Proteome		Shuffled Full Proteome	
	Protein Groups	PSMs	Protein Groups	PSMs
Ec70S	56	989/1054	0	0/0
Hs40S	135	1749/2384	0	0/0
Hs60S	98	1409/1660	2	1/4
So70S	65	2497/2615	2	2/2

b



Supplementary Figure 11 | Estimated false discovery rates in top-down database searches. a) Identification statistics for database searches of top-down LC-MS/MS of Ec70S, Hs40S-Hs60S, and So70S ribosomal particles against full proteomes of *E. coli* (strain K12), *H. sapiens*, and *S. oleracea*, respectively. PSMs are represented as follows: Unambiguous PSMs/All PSMs. b) Cumulative error based on false positive protein spectrum match calculated against decreasing $-\log_{10}$(P-Score) for Hs60S and So70S ribosomal particles, for which false positive hits against shuffled databases were observed. This estimates the false discovery rate to be around 0.25%, well below the generally accepted 1%.

ACKNOWLEDGEMENTS

M.v.d.W., S.T., K.L.F., V.F. and A.J.R.H. are funded by the large-scale proteomics facility Proteins@Work (Project 184.032.201) embedded in the Netherlands Proteomics Centre and supported by the Netherlands Organization for Scientific Research (NWO). A.J.R.H. acknowledges support by the Spinoza Prize of NWO (Project SPI.2017.028). M.v.d.W. and A.J.R.H. are also supported by a grant (12PR3303-2) from Fundamenteel Onderzoek der Materie (FOM). A.M. and A.J.R.H. acknowledge additional support through the European Union Horizon 2020 program FET-OPEN project MSmed, Project 686547. P.B. and M.I. were supported by the Swiss National Science Foundation (SNSF) and the National Center of Competence in Research (NCCR) RNA and disease programme of the Swiss National Science Foundation (SNSF). We thank Nick Quade for his support in preparation of the IRES RNAs.

AUTHOR CONTRIBUTIONS

M.v.d.W., S.T., K.L.F., E.D. and V.F. performed the experiments. M.v.d.W., S.T., and A.J.R.H. wrote the manuscript. P.B., M.I., and N.B. provided Ribosomal complexes and IRES RNA and contributed to discussions on Ribosome biology. M.v.d.W., A.M. and A.J.R.H. designed the study.

COMPETING INTERESTS

K.L.F., E.D. and A.M. are employees of Thermo Fisher Scientific, the manufacturer and supplier of Orbitrap-based mass spectrometers.

REFERENCES

1. Lössl, P., van de Waterbeemd, M., and Heck, A.J.R. (2016). The diverse and expanding role of mass spectrometry in structural and molecular biology. EMBO J. *16*, 155–166.
2. Altelaar, A.F.M., Munoz, J., and Heck, A.J.R. (2012). Next-generation proteomics: towards an integrative view of proteome dynamics. Nat. Rev. Genet. *14*, 35–48.
3. Grimsrud, P.A., Swaney, D.L., Wenger, C.D., Beauchene, N.A., and Coon, J.J. (2010). Phosphoproteomics for the masses. ACS Chem. Biol. *5*, 105–19.
4. Tran, J.C., Zamdborg, L., Ahlf, D.R., Lee, J.E., Catherman, A.D., Durbin, K.R., Tipton, J.D., Vellaichamy, A., Kellie, J.F., Li, M., *et al.* (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature *480*, 254–8.
5. Liu, F., Rijkers, D.T.S., Post, H., and Heck, A.J.R. (2015). Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat. Methods *12*, 1179–1184.
6. Liu, F., Lössl, P., Scheltema, R., Viner, R., and Heck, A.J.R. (2017). Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. Nat. Commun. *8*, 15473.
7. Fernandez-Martinez, J., Kim, S.J., Shi, Y., Upla, P., Pellarin, R., Gagnon, M., Chemmama, I.E., Wang, J., Nudelman, I., Zhang, W., *et al.* (2016). Structure and function of the nuclear pore complex cytoplasmic mRNA export platform. Cell *167*, 1215-1228.
8. Leitner, A., Faini, M., Stengel, F., and Aebersold, R. (2016). Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. Trends Biochem. Sci. *41*, 20–32.
9. Heck, A.J.R. (2008). Native mass spectrometry: a bridge between interactomics and structural biology. Nat. Methods *5*, 927–933.
10. Petrov, A.S., Bernier, C.R., Hsiao, C., Norris, A.M., Kovacs, N.A., Waterbury, C.C., Stepanov, V.G., Harvey, S.C., Fox, G.E., Wartell, R.M., *et al.* (2014). Evolution of the ribosome at atomic resolution. Proc. Natl. Acad. Sci. U. S. A. *111*, 10251–6.
11. Roberts, E., Sethi, A., Montoya, J., Woese, C.R., and Luthey-Schulten, Z. (2008). Molecular signatures of ribosomal evolution. Proc. Natl. Acad. Sci. U.S.A *105*, 13953–13958.
12. Ramakrishnan, V. (2014). The ribosome emerges from a black box. Cell *159*, 979–984.
13. Khatter, H., Myasnikov, A.G., Natchiar, S.K., and Klaholz, B.P. (2015). Structure of the human 80S ribosome. Nature *520*, 640–645.
14. Fischer, N., Neumann, P., Konevega, A.L., Bock, L. V., Ficner, R., Rodnina, M. V., and Stark, H. (2015). Structure of the E. coli ribosome–EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. Nature *520*, 567–570.
15. Bieri, P., Leibundgut, M., Saurer, M., Boehringer, D., and Ban, N. (2017). The complete structure of the chloroplast 70S ribosome in complex with translation factor pY. EMBO J. *36*, 475–486.
16. Greber, B.J., Bieri, P., Leibundgut, M., Leitner, A., Aebersold, R., Boehringer, D., and Ban, N. (2015). The complete structure of the 55S mammalian mitochondrial ribosome. Science (80-.). *348*, 303–308.
17. Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.-C., Yates, J.R., and III (2013). Protein analysis by shotgun/bottom-up proteomics. Chem. Rev. *113*, 2343–94.
18. Dunham, W.H., Mullin, M., and Gingras, A.-C. (2012). Affinity-purification coupled to mass spectrometry: Basic principles and strategies. Proteomics *12*, 1576–1590.
19. Zamdborg, L., LeDuc, R.D., Glowacz, K.J., Kim, Y.-B., Viswanathan, V., Spaulding, I.T., Early, B.P., Bluhm, E.J., Babai, S., and Kelleher, N.L. (2007). ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. Nucleic Acids Res. *35*, W701-6.
20. Liu, X., Sirotkin, Y., Shen, Y., Anderson, G., Tsai, Y.S., Ting, Y.S., Goodlett, D.R., Smith, R.D., Bafna, V., and Pevzner, P.A. (2012).

Protein identification using top-down. Mol. Cell. Proteomics *11*, M111.008524.

21. Brunner, A.M., Lossl, P., Liu, F., Huguet, R., Mullen, C., Yamashita, M., Zabrouskov, V., Makarov, A., Altelaar, A.F.M., and Heck, A.J.R. (2015). Benchmarking multiple fragmentation methods on an orbitrap fusion for top-down phospho-proteoform characterization. Anal. Chem. *87*, 4152–4158.

22. Catherman, A.D., Skinner, O.S., and Kelleher, N.L. (2014). Top down proteomics: facts and perspectives. Biochem. Biophys. Res. Commun. *445*, 683–93.

23. Leney, A.C., and Heck, A.J.R. (2017). Native Mass Spectrometry: What is in the Name? J. Am. Soc. Mass Spectrom. *28*, 5–13.

24. Marcoux, J., and Robinson, C.V. (2013). Twenty Years of Gas Phase Structural Biology. Structure *21*, 1541–1550.

25. McKay, A.R., Ruotolo, B.T., Ilag, L.L., and Robinson, C. V. (2006). Mass measurements of increased accuracy resolve heterogeneous populations of intact ribosomes. J. Am. Chem. Soc. *128*, 11433–11442.

26. Rostom, a a, Fucini, P., Benjamin, D.R., Juenemann, R., Nierhaus, K.H., Hartl, F.U., Dobson, C.M., and Robinson, C. V (2000). Detection and selective dissociation of intact ribosomes in a mass spectrometer. Proc. Natl. Acad. Sci. U. S. A. *97*, 5185–5190.

27. Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R., *et al.* (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat. Struct. Mol. Biol. *18*, 529–536.

28. Sakata, E., Stengel, F., Fukunaga, K., Zhou, M., Saeki, Y., Förster, F., Baumeister, W., Tanaka, K., and Robinson, C.V. (2011). The Catalytic Activity of Ubp6 Enhances Maturation of the Proteasomal Regulatory Particle. Mol. Cell *42*, 637–649.

29. Skinner, O.S., Havugimana, P.C., Haverland, N.A., Fornelli, L., Early, B.P., Greer, J.B., Fellers, R.T., Durbin, K.R., Vale, L.H.F. Do, Melani, R.D., *et al.* (2016). An informatic framework for decoding protein complexes by top-down mass spectrometry. Nat. Methods *13*, 237-240.

30. van de Waterbeemd, M., Fort, K.L., Boll, D., Reinhardt-Szyba, M., Routh, A., Makarov, A., and Heck, A.J.R. (2017). High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. Nat. Methods *14*, 283–286.

31. Kelstrup, C.D., Bekker-Jensen, D.B., Arrey, T.N., Hogrebe, A., Harder, A., and Olsen, J. V. (2017). Performance evaluation of the Q Exactive HF-X for shotgun proteomics. J. Proteome Res., *17*, 727-738.

32. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

33. Gonsalvez, G.B., Tian, L., Ospina, J.K., Boisvert, F.-M., Lamond, A.I., and Matera, A.G. (2007). Two distinct arginine methyltransferases are required for biogenesis of Sm-class ribonucleoproteins. J. Cell Biol. *178*, 733–740.

34. Bruce, B.D. (2000). Chloroplast transit peptides: structure, function and evolution. Trends Cell Biol. *10*, 440–7.

35. Yamaguchi, K., von Knoblauch, K., and Subramanian, A.R. (2000). The Plastid Ribosomal Proteins. J. Biol. Chem. *275*, 28455–28465.

36. Yamaguchi, K., and Subramanian, A.R. (2000). The Plastid Ribosomal Proteins. J. Biol. Chem. *275*, 28466–28482.

37. Yamaguchi, K., and Subramanian, A.R. (2003). Proteomic identification of all plastid-specific ribosomal proteins in higher plant chloroplast 30S ribosomal subunit. Eur. J. Biochem. *270*, 190–205.

38. Dohm, J.C., Minoche, A.E., Holtgräwe, D., Capella-Gutiérrez, S., Zakrzewski, F., Tafer, H., Rupp, O., Sörensen, T.R., Stracke, R., Reinhardt, R., *et al.* (2013). The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature *505*, 546–549.

39. Emanuelsson, O., Nielsen, H., and Heijne, G. Von (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. *8*, 978–984.

40. Goetze, S., Qeli, E., Mosimann, C., Staes, A., Gerrits, B., Roschitzki, B., Mohanty, S., Niederer, E.M., Laczko, E., Timmerman, E., *et al.* (2009). Identification and functional characterization of N-terminally acetylated proteins in Drosophila melanogaster. PLoS Biol. *7*, e1000236.

41. Quade, N., Boehringer, D., Leibundgut, M., van den Heuvel, J., and Ban, N. (2015). Cryo-EM structure of Hepatitis C virus IRES bound to the human ribosome at 3.9-Å resolution. Nat. Commun. *6*, 7646.

42. Young, B.D., Weiss, D.I., Zurita-lopez, C.I., Webb, K.J., Clarke, S.G., and Mcbride, A.E. (2012). Identi fi cation of Methylated Proteins in the Yeast Small Ribosomal Subunit: A Role for SPOUT Methyltransferases in Protein Arginine Methylation. Biochemistry *51*, 5091−5104.

43. Zhang, L., Ding, X., Cui, J., Xu, H., Chen, J., Gong, Y.-N., Hu, L., Zhou, Y., Ge, J., Lu, Q., *et al.* (2011). Cysteine methylation disrupts ubiquitin-chain sensing in NF-κB activation. Nature *481*, 204–208.

44. Diaconu, M., Kothe, U., Schlünzen, F., Fischer, N., Harms, J.M., Tonevitsky, A.G., Stark, H., Rodnina, M. V., and Wahl, M.C. (2005). Structural Basis for the Function of the Ribosomal L7/12 Stalk in Factor Binding and GTPase Activation. Cell *121*, 991–1004.

45. Wahl, M.C., and Möller, W. (2002). Structure and function of the acidic ribosomal stalk proteins. Curr. Protein Pept. Sci. *3*, 93–106.

46. Davydov, I.I., Wohlgemuth, I., Artamonova, I.I., Urlaub, H., Tonevitsky, A.G., and Rodnina, M. V. (2013). Evolution of the protein stoichiometry in the L12 stalk of bacterial and organellar ribosomes. Nat. Commun. *4*, 1387.

47. Gordiyenko, Y., Videler, H., Zhou, M., Mckay, A.R., Fucini, P., Biegel, E., Mu, V., and Robinson, C. V (2010). Mass Spectrometry Defines the Stoichiometry of Ribosomal Stalk Complexes across the Phylogenetic Tree. Mol. Cell. Proteomics *9*, 1774–1783.

48. Ilag, L.L., Videler, H., McKay, A.R., Sobott, F., Fucini, P., Nierhaus, K.H., and Robinson, C. V. (2005). Heptameric (L12)6/L10 rather than canonical pentameric complexes are found by tandem MS of intact ribosomes from thermophilic bacteria. Proc. Natl. Acad. Sci. *102*, 8192–8197.

49. Fort, K.L., van de Waterbeemd, M., Boll, D., Reinhardt-Szyba, M., Belov, M.E., Sasaki, E., Zschoche, R., Hilvert, D., Makarov, A.A., and Heck, A.J.R. (2017). Expanding the structural analysis capabilities on an Orbitrap-based mass spectrometer for large macromolecular complexes. Analyst *143*, 100-105.

50. Shi, Z., Fujii, K., Kovary, K.M., Genuth, N.R., Röst, H.L., Teruel, M.N., and Barna, M. (2017). Heterogeneous Ribosomes Preferentially Translate Distinct Subpools of mRNAs Genome-wide. Mol. Cell *67*, 71-83.e7.

51. Yu, Y., Ji, H., Doudna, J.A., and Leary, J.A. (2009). Mass spectrometric analysis of the human 40S ribosomal subunit: Native and HCV IRES-bound complexes. Protein Sci. *14*, 1438–1446.

52. Fernández, I.S., Bai, X.-C., Murshudov, G., Scheres, S.H.W., and Ramakrishnan, V. (2014). Initiation of Translation by Cricket Paralysis Virus IRES Requires Its Translocation in the Ribosome. Cell *157*, 823–831.

53. Hellen, C.U., and Sarnow, P. (2001). Internal ribosome entry sites in eukaryotic mRNA molecules. Genes Dev. *15*, 1593–612.

54. Cheng, Y., Glaeser, R.M., and Nogales, E. (2017). How Cryo-EM Became so Hot. Cell *171*, 1229–1231.

55. Cressey, D., and Callaway, E. (2017). Cryo-electron microscopy wins chemistry Nobel. Nature *550*, 167–167.

56. Kuhlbrandt, W. (2014). The Resolution Revolution. Science (80-.). *343*, 1443–1444.

57. Fernandez-Leiro, R., and Scheres, S.H.W. (2016). Unravelling biological macromolecules with cryo-electron microscopy. Nature *537*, 339–346.

58. Natchiar, S.K., Myasnikov, A.G., Kratzat, H., Hazemann, I., and Klaholz, B.P. (2017). Visualization of chemical modifications in the human 80S ribosome structure. Nature *551*, 472–477.

59. Walls, A.C., Tortorici, M.A., Frenz, B., Snijder, J., Li, W., Rey, F.A., DiMaio, F., Bosch, B.-J., and Veesler, D. (2016). Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. Nat. Struct. Mol. Biol. *23*, 899–905.

60. Zhang, S., Chang, L., Alfieri, C., Zhang, Z.,

Yang, J., Maslen, S., Skehel, M., and Barford, D. (2016). Molecular mechanism of APC/C activation by mitotic phosphorylation. Nature *533*, 260–264.

61. Martinez-Rucobo, F.W., Kohler, R., van de Waterbeemd, M., Heck, A.J.R., Hemann, M., Herzog, F., Stark, H., and Cramer, P. (2015). Molecular Basis of Transcription-Coupled Pre-mRNA Capping. Mol. Cell *58*, 1079–1089.

62. Snijder, J., Schuller, J.M., Wiegard, A., Lössl, P., Schmelling, N., Axmann, I.M., Plitzko, J.M., Förster, F., and Heck, A.J.R. (2017). Structures of the cyanobacterial circadian oscillator frozen in a fully assembled state. Science (80-.). *355*, 1181–1184.

63. Sharma, S., and Lafontaine, D.L.J. (2015). 'View From A Bridge': A New Perspective on Eukaryotic rRNA Base Modification. Trends Biochem. Sci. *40*, 560–575.

64. Roundtree, I.A., Evans, M.E., Pan, T., and He, C. (2017). Dynamic RNA Modifications in Gene Expression Regulation. Cell *169*, 1187–1200.

65. Cantara, W.A., Crain, P.F., Rozenski, J., McCloskey, J.A., Harris, K.A., Zhang, X., Vendeix, F.A.P., Fabris, D., and Agris, P.F. (2011). The RNA modification database, RNAMDB: 2011 update. Nucleic Acids Res. *39*, D195–D201.

66. Shaw, J.B., Li, W., Holden, D.D., Zhang, Y., Griep-raming, J., Fellers, R.T., Early, B.P., Thomas, P.M., Kelleher, N.L., and Brodbelt, J.S. (2013). Complete Protein Characterization Using Top-Down Mass Spectrometry and Ultraviolet Photodissociation. J. Am. Chem. Soc. *135*, 12646–12651.

67. Riley, N.M., and Coon, J.J. (2018). The Role of Electron Transfer Dissociation in Modern Proteomics. Anal. Chem. *90*, 40-64.

68. Valeja, S.G., Xiu, L., Gregorich, Z.R., Guner, H., Jin, S., and Ge, Y. (2015). Three Dimensional Liquid Chromatography Coupling Ion Exchange Chromatography/ Hydrophobic Interaction Chromatography/Reverse Phase Chromatography for Effective Protein Separation in Top-Down Proteomics. Anal. Chem. *87*, 5363–5371.

69. Steven, A.C., and Baumeister, W. (2008). The future is hybrid. J. Struct. Biol. *163*, 186–195.

70. Ban, N., Beckmann, R., Cate, J.H.D., Dinman, J.D., Dragon, F., Ellis, S.R., Lafontaine, D.L.J., Lindahl, L., Liljas, A., Lipton, J.M., *et al.* (2014). A new system for naming ribosomal proteins. Curr. Opin. Struct. Biol. *24*, 165–9.

71. Weisser, M., Schäfer, T., Leibundgut, M., Böhringer, D., Aylett, C.H.S., and Ban, N. (2017). Structural and Functional Insights into Human Re-initiation Complexes. Mol. Cell *67*, 447-456.e7.

72. Hardy, S.J., Kurland, C.G., Voynow, P., and Mora, G. (1969). The ribosomal proteins of Escherichia coli. I. Purification of the 30S ribosomal proteins. Biochemistry *8*, 2897–905.

73. Bern, M., Kil, Y.J., and Becker, C. (2012). Byonic: advanced peptide and protein identification software. Curr. Protoc. Bioinforma. *Chapter 13*, Unit13.20.

74. Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

75. Zabrouskov, V., Senko, M.W., Du, Y., Leduc, R.D., and Kelleher, N.L. (2005). New and automated MSn approaches for top-down identification of modified proteins. J. Am. Soc. Mass Spectrom. *16*, 2027–2038.

76. Wickham, H. (2009). ggplot2 - Elegant Graphics for Data Analysis | Hadley Wickham | Springer (New York).

PART II

ANALYSIS OF COMPOSITIONAL AND STRUCTURAL DIVERSITY IN PROTEIN ASSEBMLIES

6

CHAPTER

A COLORFUL PALETTE OF B-PHYCOERYTHRIN PROTEOFORMS EXPOSED BY A MULTIMODAL MASS SPECTROMETRY APPROACH

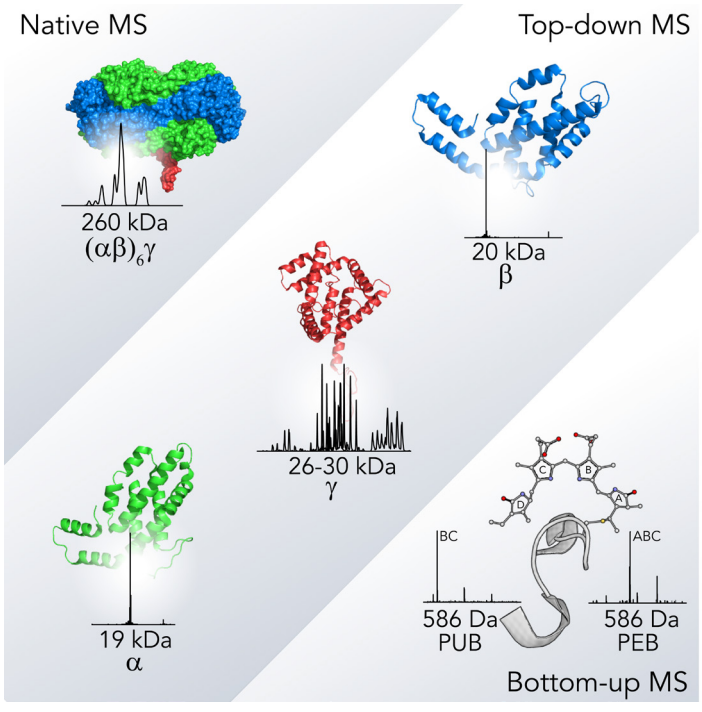
Sem Tamara[†], Max Hoek[†], Richard A. Scheltema[†], Aneika C. Leney[$] and Albert J. R. Heck[†]

† Utrecht University, Utrecht, The Netherlands
$ University of Birmingham, Birmingham, UK

Chem 2019, 5 (5) 1302-1317
DOI: 10.1016/j.chempr.2019.03.006

SUMMARY

Cyanobacteria and red algae represent some of the oldest lifeforms on the planet. During billions of years of evolution they have fine-tuned the structural details of their light-harvesting antenna, called phycobilisomes, that represent one of the most efficient systems for light-harvesting and energy transfer. Yet, the exact details of phycobilisome assembly and energy transfer are still under investigation. Here, we employed a multi-modal mass spectrometric approach to unravel the molecular heterogeneity within B-phycoerythrin, the major phycobiliprotein in the red algae *P. cruentum*. B-phycoerythrin consists of 12 subunits $(\alpha\beta)_6$ arranged in ring with the central cavity housing a linker (γ) subunit which is crucial for stabilizing B-phycoerythrin within the phycobilisome. Using top-down MS we unravel the heterogeneity in the γ proteoforms, characterizing the distinct γ chains and multiple isobaric chromophores they harbor. Our data highlight the key role γ plays in phycobilisome organization that enables optimal light transmission.



HIGHLIGHTS

- B-phycoerythrin assembly is structurally and chemically exceptionally heterogeneous
- Complex heterogeneity unraveled by combining different tiers of mass spectrometry
- The γ subunit is present in 4 distinct isoforms carrying 3 to 5 chromophores each
- MS/MS allows unambiguous distinction of attached isobaric PEB and PUB chromophores

THE BIGGER PICTURE

Some of the most efficient light-harvesting machineries present on earth are found in red algae and cyanobacteria. These systems, termed phycobilisomes, are comprised of numerous proteins decorated with a plethora of chromophores. The precise arrangement of all proteins and chromophores in the phycobilisome assembly form the basis of the extremely efficient energy transfer. Here we combine different mass spectrometric methods enabling the structural investigation of all components of the B-phycoerythrin sub-complex in a highly-detailed manner. This includes identifying all proteoforms present in the assembly, as well as distinguishing the various (isobaric) chromophores they harbor. Together this information leads to fundamental insights into the arrangement and chemical heterogeneity of the phycobilisome. Better understanding of the architecture of this complex is essential for the future design of even more efficient light-harvesting machineries.

INTRODUCTION

Phycobilisomes are large light-harvesting antennas that facilitate the conversion of light into chemical energy in different species of cyanobacteria and red algae[1,2]. These MDa protein assemblies are formed by a morphologically distinct core complex and rod-like assemblies that are attached to the core[3]. Both structural units consist of stacked disc-shaped phycobiliproteins (PBP), which are themselves multichain protein complexes with distinct photochemical properties. The core of the phycobilisome is primarily comprised of allophycocyanin (APC; λ_{max} 651 nm)[4] while the rods incorporate phycocyanin (PC; λ_{max} 620 nm)[5] and phycoerythrin (PE; λ_{max} 565 nm)[6] that are situated proximal (PC) or distal (PE) to the core. Specific topologies of PBPs within phycobilisomes facilitate spontaneous excitation energy flow as energy transitions decrease from the rods to the core[7]. The distinct photochemical properties of these PBP types are largely defined by tetrapyrrole prosthetic groups (called bilins) that are covalently attached to the cysteine residues of the polypeptide chains[8].

The phycoerythrin family of PBPs are unique to red algae and cyanobacteria and have the most pronounced fluorescent and colorant properties of all PBPs with fluorescence quantum yield (Q) in the range of 0.82-0.98[9]. As such, phycoerythrins (PEs) have numerous biotechnological applications as dyes and fluorescent tags [10,11]. One of the most studied PEs is B-phycoerythrin (B-PE; Q = 0.98)[12–15], which is the most abundant PBP (~42% of all colorant proteins) in the red algae *Porphyridium cruentum*[16]. B-PE is known to be a hetero-13-mer that contains six α, six β, and one linker protein subunit termed γ[6]. The primary architecture of the B-PE assembly consists of two overlaid disc-shaped $(\alpha\beta)_3$ hexamers which form a central cavity that is filled by a single γ subunit[17]. The B-PE complex has two types of bilin prosthetic groups that are covalently bound to cysteine residues: phycoerythrobilin (PEB; λ_{max} 550 nm) and phycourobilin (PUB; λ_{max} 498 nm)[18]. The maximum absorbance of B-PE is at 565 nm, which originates from B-PE assemblies harboring a high content of PEB molecules. It has been well-documented that each 17.8 kDa α chain carries 2 PEB prosthetic groups, while the 18.5 kDa β chain harbors 3 PEB molecules. Typically, bilins are connected *via* a single thioester bond to the cysteine residue, however one of the bilin prosthetic groups of the β chain is connected through two thioester linkages[19]. Compared to the available knowledge about the α and β chains, the nature of the γ chain and the chromophores it carries has so far remained much more elusive.

The γ chain is important as it stabilizes the tertiary structure of phycoerythrins by holding the discs of $(\alpha\beta)_3$ hexamers together[15,17]. Moreover, the bilin prosthetic groups the γ chain harbors enhance the light-absorbance properties without increasing the spacing of phycobiliproteins[2]. Additionally, it has been proposed that the γ chain provides energetic decoupling protecting the photosynthetic reaction center from damage induced by excessive photoexcitation[20]. Initially, in biochemical studies of PEs, the γ subunit was identified as a single band on sodium dodecyl sulfate (SDS) gels and was assumed to be a single protein[6]. Later, reversed phase liquid chromatography (RP-LC) revealed that the γ subunit is represented by at least three distinct polypeptide chains in B-PE assemblies[16,21], however, the exact amino acid sequences and positions of attached bilins were not determined. Overall, the γ subunits are expected to have molecular weights in the range of 27-35 kDa based on the sequences of predicted genes with likely up to four bilins attached to them, supposedly two PEBs and two PUBs[6]. In the related R-PE complex the γ subunit was long proposed to harbor 4 chromophores as well, however five distinct chromphorylated peptides were detected, which was rationalized by presence of several distinct γ subunits[8,22]. Recently, a structural model was reported for the entire phycobilisome from the red alga *Griffithsia pacifica* based on cryo-EM data revealing more details on the structure and conformation of the γ linker subunit[2]. Predicted structures outlined the presence of a chromophore-binding domain on γ that could carry up to 5 bilin molecules. However, due to the variations of the γ subunit sequences within single and, in particular, different algae strains, it is likely that not all γ subunits are identically modified. Moreover, since extensive "class" averaging was performed to obtain the cryo-EM images, heterogeneity in these γ protein sequences present within the PE core and their bilin modifications are difficult to resolve. Thus, alternative methods are indispensable to distinguish and analyze B-PE variants and γ proteoforms separately, allowing the extent of post-translational processing events to be individually characterized and quantified.

Mass spectrometry (MS) is a rapidly emerging tool to monitor protein isoforms present within protein complexes[23–25]. Due to their difference in mass, proteoforms and modifications contained within can be rapidly distinguished and quantified. Indeed, preliminary MS work has already been utilized to reveal details on the bilin architecture and amino acid sequence of the γ subunit in R-phycoerythrin[26,27]. The combination of multiple MS approaches provides complimentary information that is not obtainable from a single MS method, which proved to be advantages for analysis of highly heterogeneous proteins and protein complexes[25,28–30]. Here, we use a combination of bottom-up, top-down, and native MS to explore the structural heterogeneity present within the protein assembly B-PE and its constituent subunits in the red algae *P. cruentum*. We unequivocally determine the co-occurrence of multiple variants of the B-PE assembly and link each of these variants to distinct proteoforms of the α, β, and γ subunits. In our work four distinct polypeptide chains of the γ subunit are identified (one more than previously reported for B-PE from *P. cruentum*[13]), quantified, and fully characterized. These chains harbor different number of bilins ranging from 3 to 5 including both PEB and PUB molecules (more heterogeneous than previously reported for B-PE from *P. cruentum*[21]). Thus, the complete B-PE assembly can carry up to 35 bilin chromophores. In our work, by linking fragment signatures to the structures of isobaric prosthetic groups we unambiguously characterize and position each PEB or PUB bilin on the γ chains. Such information is relevant as, ultimately, the photochemical properties of the B-PE assembly are a result of the interplay between all co-assembled proteins and the chromophore groups they harbor. Moreover, because topologies of phycobiliproteins in a phycobilisome are influenced by a linker protein and number/type of carried bilin molecules[2] different γ subunits define ordering of B-PE within phycobilisome rods.

RESULTS

Heterogeneity of B-phycoerythrin probed by native MS

As a first step to unravel the structural heterogeneity within B-PE, we measured high-resolution native mass spectra of the fluorescent assembly. These spectra indicated that the B-PE complex is heterogeneous, with several co-occurring charge states present, all originating from assemblies with molecular weights between 260 and 270 kDa (Figure S1A), in line with literature and as expected for $\alpha_6\beta_6\gamma$ assemblies[15]. To further investigate the factors influencing the heterogeneity observed in the native mass spectrum of B-PE, we then performed native tandem MS (native top-down MS/MS) experiments on this assembly. The z = 37+ ions of different B-PE variants were isolated and subsequently subjected to collisional activation. These tandem mass spectra revealed the ejection of α and β subunits and residual complexes of B-PE wherein α, β or combinations of the two subunits had been eliminated (Figure 1). Upon collisional dissociation of different precursor ions originating from different assembly variants, the released α and β subunits always had identical masses while the residual high molecular weight fragment complexes, formed by the loss of an α and/or β subunit, exhibited clear mass differences that, thus, can be attributed to the presence of different forms of the γ subunit.

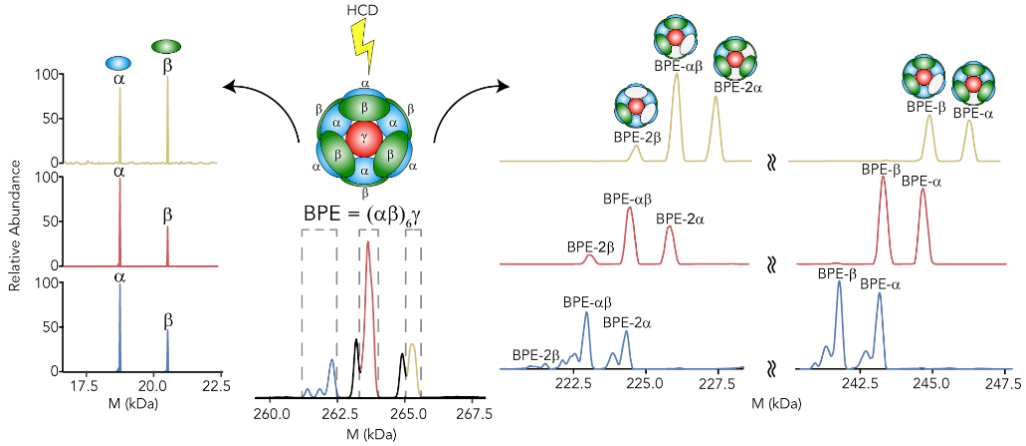


Figure 1 | Native Top-down MS/MS of B-phycoerythrin Assembly Variants. Deconvoluted mass spectra observed following HCD fragmentation of B-phycoerythrin (B-PE) precursor ions sprayed under native conditions. Precursor ions (z = + 37) corresponding to B-PE species enclosed with dashed boxes were subjected individually to HCD, which resulted in the ejection of α and β monomeric subunits (left) and the concomitant formation of residual fragmented complexes (right) missing a single α or β subunit or combinations of subunits (2α, 2β, or αβ).

The Native top-down dissociation of B-PE was found to be limited to the ejection of maximally two subunits. We did not detect the γ subunit to be ejected. This may have multiple reasons; namely its lower stoichiometry, broader structural heterogeneity and topological lower accessibility, as the γ subunit is buried inside the cavity of the $(\alpha\beta)_3$ hetero-hexamers. If γ is present in a large number of isoforms/proteoforms its ion signals would spread over numerous peaks, hampering detection.

We additionally performed pseudo-MS3 experiments (Supplemental Experimental Procedures) to test whether it is possible to detect the γ subunit detaching from the residual B-PE complexes formed upon collisional activation. For this, following in-source activation residual complexes corresponding to B-PE assembly lacking 2α, αβ, or 2β subunits were mass-selected and fragmented with HCD (Figure S1B). Alongside the α/β monomeric products of dissociation in this experiment we observed dimers and trimers of α/β subunits, however still we did not detect intact γ subunits (Figure S1C-D).

Characteristics of proteins comprising the B-phycoerythrin assemblies

To identify the heterogeneity behind the γ subunit as revealed by native MS (here and earlier by Leney et al.[15]), we next denatured B-PE, digested it into peptides and the resulting peptides were analyzed by bottom-up LC-MS/MS. The data revealed that the most abundant proteins in the sample were as expected the α and β chains of B-PE. Identifying any γ subunits present, however, is more challenging since their mature sequences as well as positions, type, and number of bilins have not been explicitly reported. Thus, a database was set up incorporating all of the possible sequences obtained from the *P. cruentum* genome[31]. By comparing the peptides identified with this dedicated protein database, the presence of 4 different γ subunits could be revealed in the list of the most abundant identified proteins (Figure 2A and Table S1). Next to these we also detected peptides originating from other linker protein(s) that did not carry any bilin molecules, albeit typically all at a much lower abundance. Such linker proteins are more prominent for PBPs in the proximal parts of the rods in relation to the core of the phycobilisome[32]. Along with B-PE-related proteins bottom-up LC-MS/MS revealed subunits of R-PE detected albeit with significantly lower abundances, likely because of their similar biochemical properties that resulted in their co-purification.

To verify the presence of multiple γ subunits within B-PE, we next denatured B-PE and separated the intact proteins using reversed-phase HPLC. Consistent with the bottom-up results, the data showed two abundant signals corresponding to the α (~9 min retention time) and β subunits (12.5 min) as well as several lower abundant peaks with shorter retention times (2.5-5 min) (Figure 2B). Peak splitting observed for β subunit was attributed to shifted retention time of oxidized proteoforms. Additionally, we observed a peak at 5.5 min retention time that was assigned to the colorless linker protein (Figure S2; fraction A06). The shorter retention times of the γ subunits can be rationalized by these proteins harboring more hydrophilic residues than the α and β subunits. To verify whether the eluting proteins are chromophorylated subunits of B-PE we measured the absorbance spectra of the fractions (Figure 2C). Clearly, three distinct absorption profiles were observed that resembled previously reported absorbance spectra of the B-PE subunits[13]. Additionally, we measured the absorbance for fractions corresponding to different γ chains (Figure S2; fractions A02-A04), whereby we observed that all these fractions resulted in alike absorbance profiles. However, such absorption data cannot directly reveal the number and positions of the bilin prosthetic groups these subunits harbor. Therefore, we next set out to further characterize all proteoforms of the γ subunits.

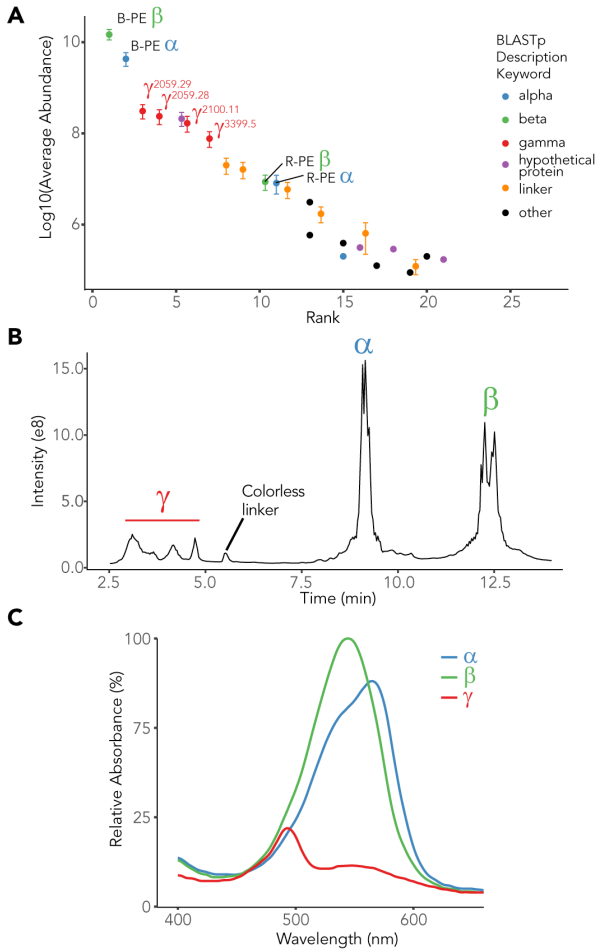


Figure 2 | Identification of Distinctive B-phycoerythrin Subunits. Overview of all proteins, with distinct physico- and photochemical properties, identified in the B-phycoerythrin sample. (A) Proteins identified in the bottom-up LC-MS/MS analysis and ranked based on the combined abundance of the respective peptides in LC-MS. The error bars represent standard error of the mean abundance. (B) Reversed-phase (RP) LC separation of the intact subunit proteins in the B-phycoerythrin sample represented by base peak intensities against the retention time. (C) Absorbance spectra of fractions collected after RP-LC, corresponding to the α (blue), β (green), and four γ (red) subunits.

Characterizing γ Subunit with "Sequence Tags" and Mass Matching

Although traditional shotgun LC-MS/MS methods provide means for fast and sensitive identification of proteins, the information on mature proteoforms and their chromophorylation stoichiometries is lost because of protein digestion. Top-down MS circumvents this problem by analyzing the proteins intact and, therefore, potentially allows for the identification of all proteoforms present.

In top-down MS experiments of non-modified proteins the backbone fragments typically provide direct sequence information. However, the γ proteoforms studied here harbor various bilin modifications. These chromophores heavily influence and

complicate the observed dissociation patterns (Figure S3A). Moreover, co-isolation of co-eluting proteoforms limit further the straightforward retrieval of sequence information. Therefore, we here characterized the γ proteins using an alternative approach, wherein the proteoforms were identified by mass matching to the theoretical proteoform masses that were further verified by using bottom-up LC-MS/MS.

Using fast low resolution (7,500 at 200 m/z) recording of mass spectra, several γ proteoforms were successfully resolved by top-down LC-MS (Figure S3B; Data S1), consistent with the broad elution peaks observed in the total ion chromatogram corresponding to the γ subunits (Figure 2B). The intact masses alone provide significant insight into the heterogeneity present within the γ subunit with multiple molecular weights being identified for each proteoform (Figure S3C). Elucidating their sequences from intact masses only, however, is challenging. This is due to two reasons. Firstly, the protein sequences of B-PE γ subunits from *P. cruentum* are missing from the conventionally used protein databases (*e.g.* UniProt). Secondly, information on the bilin content and localization in the γ protein sequences is incomplete. For these reasons, we designed an unbiased screening approach by building a custom database that incorporated all of the sequences from the *P. cruentum* genome with varying number of chromophorylations. However, in our initial attempts no matches were found between the experimental and theoretical masses. Thus, we hypothesized that considering the number of γ proteoforms identified, post-translational sequence processing events could have occurred.

While MS scans of the α and β subunits were dominated by a single most abundant proteoform (Figure S4), for the γ subunit we observed the co-elution of various proteoforms differing in masses likely originating from the addition or deletion in the sequence of a few amino acid residues (Figure 3A, B). Excitingly, by searching these "sequence tags" against the custom database of the γ subunit proteoforms we identified one of the γ subunits, which we named $\gamma^{2059.29}$ wherein the superscript refers to the contig number annotated in the DNA sequencing[31]. The observed position of the "sequence tag" indicated a N-terminal processing of this chain, which previously was proposed for γ subunits of R-PE, based on the fact that they require a transit peptide for transfer into the chloroplasts[33]. Interestingly, upon further analysis, top-down LC-MS revealed truncated sequence variants for all the γ subunits detected showing that all of them in the final phycobilisome complex require cleavage of the transit peptide prior to complex assembly. Indeed, consistent with these results, no peptides in the N-terminal transit peptides of the γ subunits were detected by bottom-up (data not shown).

To investigate γ protein processing and chromophorylation sites, the sequences of the γ subunits were aligned making use of the MUSCLE algorithm[34]. This alignment revealed that all the γ subunits contain a conserved region, which has been recently pointed out as the chromophore binding domain of the γ subunits[2]. In three out of four γ chains this domain contained 5 conserved cysteine residues, which can be regarded as potential sites for bilin attachment ($\gamma^{3399.5}$ is missing 2 out of total 5 conserved cysteines) (Figure 3C). We extended our custom database of theoretical proteoforms with truncated forms of γ based on the experimentally detected "sequence tags", which indicated at mature N-termini.

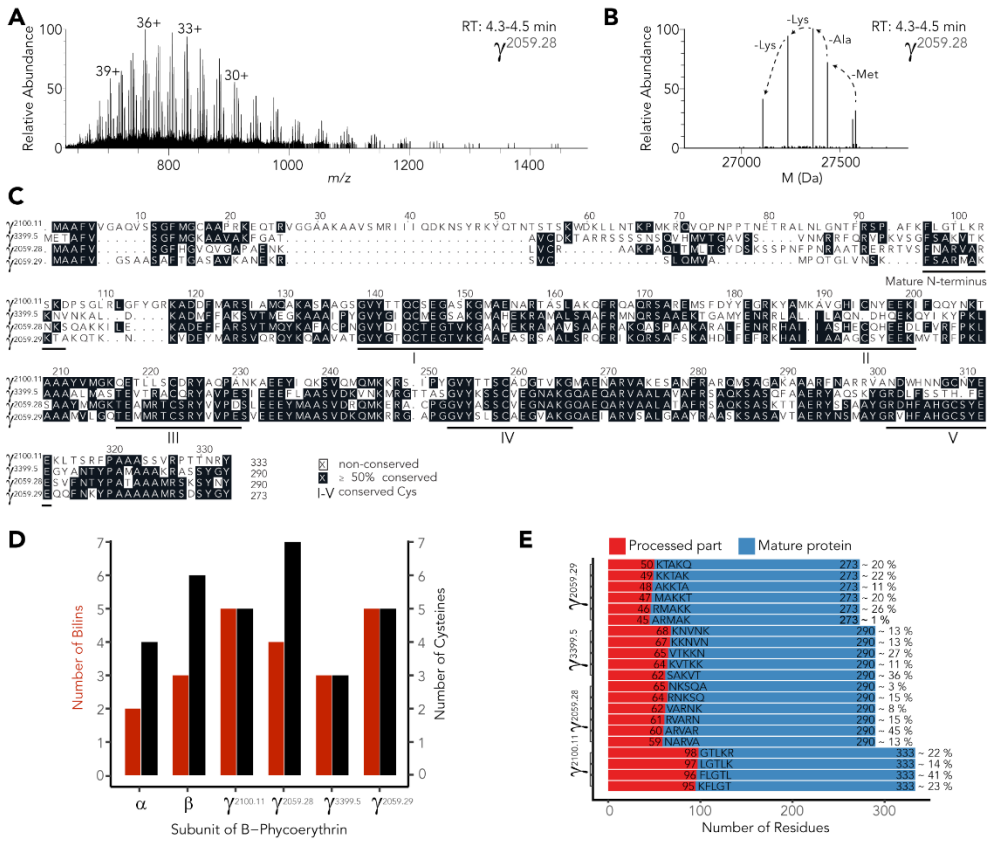


Figure 3 | **Determination of the Pallet of γ Subunit Proteoforms Identified by Top-down LC-MS.** (A) A full LC-MS scan displaying a mixture of charge envelopes for different co-eluting proteoforms of $\gamma^{2059.29}$. (B) Deconvolved mass spectrum of (A) reveals several proteoforms with mass differences that originate from the sequential deletion of specific amino acid residues due to protein processing. (C) Predicted sequences of the γ subunits aligned by using the MUSCLE algorithm. Conserved regions corresponding to potential chromophore-binding sites are annotated with roman numbers. (D) The number of chromophorylations observed (red bars) on the most abundant proteoform and total number of cysteines on the detected B-PE subunits (black bars). (E) Relative abundances of each of the processed sequence variants for each γ subunit.

Using the extended custom database, we identified multiple sequence variants of the B-PE subunits with varying number of occupied chromophorylation sites (Figure 3D). This is in contrast to the recent cryo-EM study, wherein only complete occupancy of cysteine residues was reported, highlighting the advantages mass spectrometry can provide in revealing the structural heterogeneity within the B-PE assembly. Using top-down LC-MS, we were able to assess the relative abundance of each of the γ proteoforms, the data revealing both the extent of protein processing and the variable bilin occupancies for each γ subunit (Figure 3D, E). Two out of four γ subunits were primarily represented by proteoforms having all their cysteine residues fully occupied by bilins. The polypeptide chain of $\gamma^{2059.28}$, which has the highest number of cysteine residues, was not detected in a form fully saturated with chromophores. Finally, $\gamma^{3399.5}$, which lacks two conserved cysteines, was observed harboring only 1, 2, or 3 bilin molecules. In agreement with previously reported data, the α and β subunits predominantly carry 2 and 3 bilins per subunit, respectively, resulting in half of the total number of cysteines being occupied (Figure 3D). The heterogeneity observed for the γ subunit in our work for B-PE from *P. cruentum* is in agreement with the cryo-EM data recently reported for the phycobilisome from *G. pacifica*, which has R-PE as a primary type of phycoerythrin. For *G. pacifica* 5 distinct isoforms of the γ subunit were detected, of which four harbored 5 chromophores and one had 4 bilin molecules attached[2]. The complete list of matched proteoforms of α, β, and the various γ subunits of B-PE from *P. cruentum* can be found in the Data S1. Additionally, to verify the most prominent γ proteoform within the B-PE assembly we targeted the RP-LC fraction corresponding to the $\gamma^{2059.29}$ isoform and performed EThcD MS/MS. All 5 predicted bilins with isobaric masses of 586.279 Da were detected by corresponding fragments and could all be positioned at the conserved chromophorylation sites (Figure S5 and Supplemental Experimental Procedures).

Reconstruction of the Native Mass Spectrum of B-PE from the Qualitative and Quantitative Data on All α, β, and γ proteoforms

Mass matching of features extracted from the top-down LC-MS runs resulted in the identification of a wide variety of proteoforms of all B-PE subunits. To validate this, we used a recently developed computational approach[30] to recreate a native mass spectrum based on the intensities of mass features detected in the top-down LC-MS data. Thus, the mass of the γ proteoforms based on their average abundances from the intact LC-MS data were plotted alongside the mass of the B-PE complex as determined by native MS whereby the mass of $(\alpha\beta)_6$ had been subtracted.

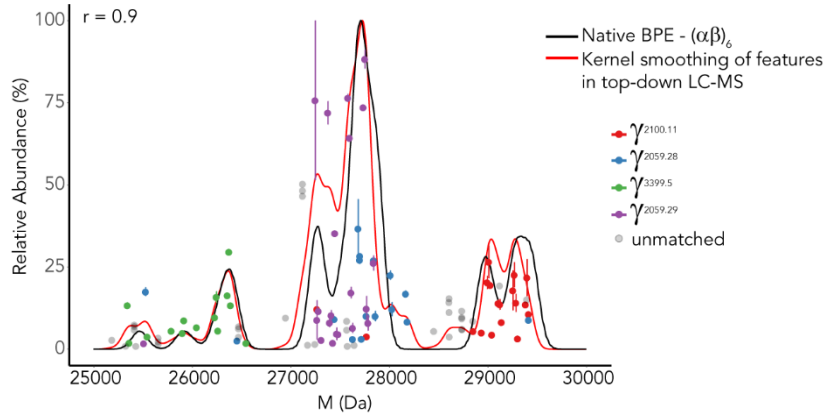


Figure 4 | **Reconstruction of the Features Observed in the Native Mass Spectrum of the Intact B-phycoerythrin Assembly Based on the Proteoforms Detected in the Top-down LC-MS Analysis.** Summing over all mass features detected in the top-down LC-MS runs provides a mass profile that correlates well with the profile observed in the native mass spectra of B-phycoerythrin (here displayed by subtraction of the $(\alpha\beta)_6$ mass). Mass features are color-coded in accordance with the matching γ proteoform masses from the custom database. Error bars represent standard error of the mean calculated for three technical replicates.

Direct comparison of these two profiles showed a high correlation of 0.9 (Figure 4)

indicating that the γ subunits that participate in formation of different B-PE variants have been explicitly and correctly identified by our top-down mass spectrometry approach. Additionally, it confirms that the γ subunits are the dominant factor contributing to the mass heterogeneity within the full B-PE assembly. Based on this analysis we conclude that the most abundant B-PE assembly is formed by $(\alpha\beta)_6\gamma^{2059.29}$ with 35 attached chromophores in total (Figure 4 and Figure S6).

Further molecular diversity introduced by the isobaric PEB and PUB chromophores

The absorbance maxima observed for the B-PE complex and its subunits (Figure 2C, Figure S7) indicate the presence of two types of chromophores, namely, phycoerythrobilin (PEB, absorbance maximum at 550 nm) and phycourobilin (PUB, absorbance maximum at 498 nm). In the PubChem database, the molar mass of PEB (CID 5289229) is 586.689 g/mol while the molar mass of PUB (CID 5289229) is 590.721 g/mol. Because the chromophore moiety binds loosely to the cysteine residue – being readily detached upon HCD in bottom-up or top-down LC-MS/MS experiments – the mass of the bilin molecule could be determined. Our data showed that masses of the majority of chromophorylated peptides from B-PE subunits indicate at mass shift of 586 Da corresponding to PEB (Figure S8). For some positions we also observed several PEB derivatives that displayed the addition or deletion of 1-2 hydrogens, however not a single peptide was observed with mass shift of 590 Da, which is the theoretical mass of PUB. Based on this evidence we conclude that contrary to the theoretically expected masses both PUB and PEB bilin moieties when attached to B-PE subunits are isobaric and have a monoisotopic mass of 586.279 Da. Having identical masses, the distinct difference in absorbance of PEB and PUB can be reasoned as the chromophores have different π conjugation systems. Thus, taking into account that the double carbon-carbon bond is nearly twice as strong as the single bond (*D* = 602 kJ/mol and 346 kJ/mol, respectively)[35], we hypothesized that upon MS/MS the extended π conjugation system of PEB should prevent formation of fragments containing two inner pyrroles (annotated as BC in Figure 5A, B), thus producing distinctive fragmentation signatures different than for PUB. Indeed, MS/MS spectra of the chromophorylated peptides of the B-PE subunits were dominated by either three-pyrrole (*m/z* 466.23^{+1} and 464.22^{+1}) or two-pyrrole (*m/z* 343.17^{+1}) fragment ions (Figure 5B) in the low mass region indicative for either PEB or PUB chromophore, respectively. Moreover, peptides chromophorylated with PEB displayed fragment ions consistent with three pyrroles closest to the attachment site (*m/z* 466.23^{+1} annotated as ABC in Figure 5A, B). This fragmentation pattern agrees with the proposed conjugation system in the PEB molecule. Notably, the slightly higher abundance of the tri-pyrrole fragment BCD *versus* ABC is supported by the bond energetics with a dissociation energy of 618 kJ/mol for ABC and 602 kJ/mol for BCD calculated by summing the dissociation energies of the respective bonds (C-S and C-C single bonds for ABC; C=C double bond for BCD)[35].

Fragmentation signature peaks allowed us to unambiguously characterize all the high-stoichiometric chromophorylations in each of the B-PE subunits by quantifying the abundance of characteristic bilin fragments. First, chromophorylated peptides

from α/β subunits displayed similar bilin fragmentation and were clustered together based on four characteristic fragments supporting presence of PEB on all sites (Figure 5C). Hierarchical clustering directly provided two groups of chromophorylation sites in agreement with expected presence of two bilin types suggested by the absorbance profiles of B-PE subunits (Figure 2B). Distinctively, doubly-linked PEB on β subunit (Cys50-Cys61) produced less ABC fragment upon dissociation compared to the singly-linked PEB molecules (Figure S9), ultimately allowing us to distinguish Cys-PEB, Cys-PUB and Cys-PEB-Cys linked bilins.

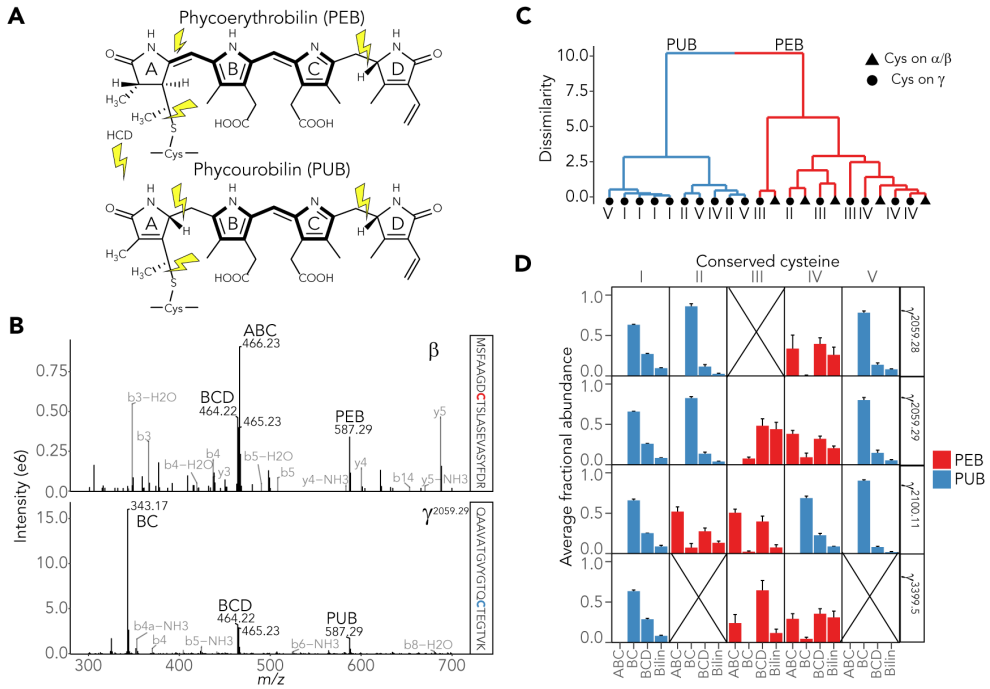


Figure 5 | Differentiating between isobaric phycoerythrobilin (PEB) and phycourobilin (PUB) moieties attached to the B-PE subunits. (A) Chemical structures of the isobaric PEB and PUB moieties attached to cysteine residues with the most prominent fragmentation channels observed upon HCD indicated. (B) Examples of tandem mass spectra of peptides chromophorylated with PEB (β subunit) or PUB ($\gamma^{2059.29}$ subunit). (C) Dendrogram representing the hierarchical clustering of the chromophorylation sites based on the abundances of four characteristic fragments (ABC, BC, BCD, and Bilin). Each dot or triangle represents the chromophorylation site on the α/β or γ subunit, respectively. (D) Abundances of bilin fragments calculated for each of the conserved cysteines of the γ chains, color-coded based on the hierarchical clustering in (C). Error bars represent standard error of the means.

In the γ subunits several conserved cysteine residues – annotated by roman numbers I-V in Figure 3C – displayed a distinct preference for either PEB or PUB moieties (Figure 5D). Recently, it was proposed that for the red algae *G. pacifica* the bilins of the γ subunits connect with those of the β subunit to allow efficient energy transfer within rods of the phycobilisome[2]. Using the recently published structural model of the phycobilisome from *G. pacifica* we color-coded in Figure 6 bilins of the γ and β subunits of phycoerythrins in accordance with the bilin types revealed for B-PE subunits from *P. cruentum* in our study. Consistent with *G. pacifica*, our

identified PUB sites on the γ subunit within B-PE in *P. cruentum* also connect with the PEB molecules on the β subunits allowing them to transfer energy efficiently when the initial chromophore is excited. The data acquired within this work, therefore, offer an explanation as to how we determined that PUB molecules of B-PE carried by the most abundant γ subunits are participating in the shortest energy transfer pathways indicated by the black arrow lines in the Figure 6A. In a phycobilisome, such a layout of chromophores in the phycoerythrin, away from the core parts of the rods, can extend the spectral range for more efficient light-harvesting and allow for more energetic excitation-relaxation transitions of PUB molecules to happen prior to less energetic primarily PEB-mediated energy transfer within phycoerythrins in the proximal parts of the rods of a phycobilisome. To a lesser investigated extent PUB groups of the γ subunits might act as light quenchers by absorbing less energetic emission from excited PEB molecules. The overall structure of stacked phycoerythrin complexes reveals that the γ subunits from one phycoerythrin complex link to another *via* its N-terminus, being in close proximity to bilin groups on conserved cysteines III and IV of the γ subunit from subsequent phycoerythrin (Figure 6B). Diverse bilin combinations at these positions (Figure 5D) may regulate the ordering of B-PE complexes in the rods of the phycobilisome.

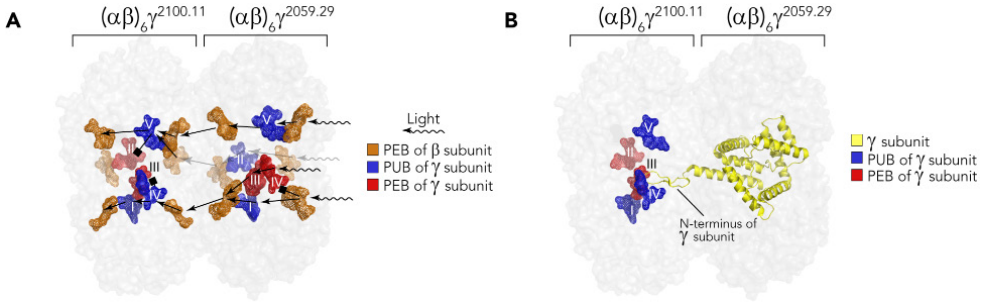


Figure 6 | γ Subunit Facilitates Energy Transfer and Linking of Phycoerythrins in Phycobilisome. (A) Energy transfer pathways within phycoerythrin complexes as stacked in the structural model of the intact phycobilisome (PDB entry: 5Y6P, phycobilisome from *G. pacifica*). The bilin chromophores on the γ subunits are colored to resemble bilin types determined in the current study for $\gamma^{2059.29}$ and $\gamma^{2100.11}$. Rings of $(\alpha\beta)_6$ are represented as the grey transparent surface. The black lines schematically represent the shortest energy transfer pathways through the phycobilisome complex. (B) N-terminus of the γ subunit from one phycoerythrin lies within the complex in close proximity to bilin groups on conserved cysteines III and IV (Figure 3C) of the γ subunit from a neighboring phycoerythrin complex.

DISCUSSION

The distinct photochemical properties of each phycobiliprotein drive an increasing demand for their industrial utilization. However, progress in discovering the molecular details of how these protein complexes function and – thus – opening new biotechnological capabilities has been hindered due to their high complexity and structural heterogeneity. Moreover, it is only when we can identify how the phycobilisome components function individually, that we can attempt to unravel the mechanistic details behind how the intact phycobilisome operates. One of the ways to extend understanding of the light harvesting machineries and improve their exploitation is through explicit characterization of phycobiliprotein variants and constituent proteoforms. Here, by using different tiers of mass spectrometric analysis we were able to determine the heterogeneity of B-PE in unprecedented detail. Native MS allowed us to detect multiple variants of the intact B-PE assembly and gain insights into its stoichiometry. Top-down LC-MS on the intact subunits revealed the heterogeneity within the B-phycoerythrin subunits and provided means for characterization and quantification of the prominent proteoforms. Lastly, bottom-up LC-MS/MS facilitated identification and localization of prosthetic groups on each of the B-PE subunits. Taken together, B-phycoerythrin was detected as mixture composed of six α, six β and one of four distinct γ subunits: $\gamma^{2059.29}$, $\gamma^{2059.28}$, $\gamma^{3399.5}$, and $\gamma^{2100.11}$ whereby 35 bilin molecules decorate on average each of the B-PE protein complexes. Finally, we demonstrated that by using bottom-up LC-MS/MS it was possible to unambiguously distinguish between isobaric tetra-pyrrole chromophores attached to each of the modified cysteine residues.

Together, our work reveals high levels of structural heterogeneity present within B-PE. Interestingly, this heterogeneity is confined to the γ subunit; the subunit that links B-PE's together and is essential for its stability. Our results indicate that 4 distinct γ subunits are present which is in agreement with the cryo-EM structure whereby γ subunits are required to link the individual PE complexes in the rods of the phycobilisome[2]. Furthermore, the different N-terminal regions of the distinct γ subunits that participate in linker-linker contacts might influence the absorbance and emission properties of the involved prosthetic groups, as the protein microenvironment influences the fluorophore-mediated light transmission[36]. Additionally, we show the most abundant γ subunits ($\gamma^{2059.29}$ and $\gamma^{2100.11}$) have all 5 cysteines saturated with bilins. However, it is important to note that this is not the case for all γ subunits as $\gamma^{2059.28}$ and $\gamma^{3399.5}$ carry 4 and 3 bilins, respectively. The γ subunit is the only phycobiliprotein within phycoerythrin to contain phycourobilin (PUB). The nature of PUB is crucial for the absorption of 495 nm light through the phycobilisome. Here, we locate different combinations of PUB and PEB groups on γ subunits wherein two chromophore positions that participate in linking of phycoerythrins demonstrate unique chromophorylation patterns. Thus, we speculate that varying chromophorylation patterns and distinct primary structures of linker subunits drive joining and ordering of phycoerythrins for efficient light transmission throughout the rods of phycobilisomes. Overall, we expect that the detailed molecular knowledge gathered here will provide strong foundation for further investigations into how these large macromolecular machines function and add important detail about how energy may be most efficiently transferred through these light-harvesting complexes. Moreover, unravelling the complexity of the phycobilisome will prove essential for the further applications of such systems in science and industry.

EXPERIMENTAL PROCEDURES

Sample Preparation

B-phycoerythrin (B-PE) was purchased from Thermo Fischer Scientific. For bottom-up LC-MS/MS analysis B-phycoerythrin was reduced with 20 mM TCEP at room

temperature for 30 minutes and alkylated with 20 mM chloroacetamide for 30 minutes in the dark. Digestion of proteins was done overnight at 37 °C with trypsin (Promega Benelux, Leiden, The Netherlands) at a protein-to-enzyme ratio of 50:1 (weight/weight). Samples were kept at pH > 7 prior to LC-MS/MS in order to prevent interconversion between phycourobilin and phycoerythrobilin.

For top-down LC-MS/MS protein samples were buffer exchanged into 0.1% formic acid by using 3 kDa molecular weight cutoff centrifuge filters (Amicon Ultra, Merck KGaA, Darmstadt, Germany) and then diluted to 1 µg/µl final concentration.

B-PE sample was prepared for native MS experiments by several cycles of buffer exchange into aqueous ammonium acetate. Centrifugal filters (Amicon, Ultra Merck KGaA, Darmstadt, Germany), which were used in the buffer exchange procedure, had a molecular weight cutoff at 10 kDa. The final concentration of the ammonium acetate was 300 mM and the pH was adjusted to 7.5.

Bottom-up LC-MS/MS Analysis

Separation of the peptides from the digested B-PE was performed on an Agilent 1290 Infinity HPLC system (Agilent Technologies, Waldbronn, Germany). Samples were loaded on a 100 µm x 20 mm trap column (in-house packed with ReproSil-Pur C18-AQ, 3 µm) (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) coupled to a 50 µm x 500 mm analytical column (in-house packed with Poroshell 120 EC-C18, 2.7 µm) (Agilent Technologies, Amstelveen, The Netherlands). A 2-5 µL injection of peptides was used, corresponding to ~0.05 µg of material. The LC-MS/MS run time was set to 40 min with flow rate of 300 nL/min. Mobile phases A (water/0.1% formic acid) and B (80% ACN/0.1% formic acid) were used for gradient time of 35 minutes: 13 to 44% B for 20 minutes, and 44 to 100% B over 3 minutes. Samples were analyzed on a Thermo Scientific Orbitrap Fusion™ Lumos™ Tribrid™ Mass Spectrometer. Nano-electrospray ionization was achieved using a coated fused silica emitter (New Objective, Cambridge, MA, USA) biased to 2 kV. The mass spectrometer was operated in positive ion mode and the spectra were acquired in the data-dependent acquisition mode. Full MS scans were acquired with resolution setting set to 60,000 (200 m/z) and at a scan mass range of 375 to 2,000 m/z. Automatic Gain Control (AGC) target was set to 4e5 with maximum injection time of 50 ms. Data dependent-MS/MS (dd-MS/MS) scans were acquired at 30,000 resolution (at 200 m/z) and with mass range of 200 to 2,000 m/z. AGC target was set to 5e4 with maximum of injection time defined at 54 ms. 1 µscan was acquired both for full MS and dd-MS/MS scans. Data dependent method was set to isolation and fragmentation for the cycle time set to 5 seconds. Parameters for isolation/fragmentation of selected ion peaks were set as follows: isolation width – 1.6 Th; HCD normalized collision energy (NCE) – 28%; mass analyzer – Orbitrap.

Top-down LC-MS/MS Analysis

Chromatographic separation of intact protein samples was conducted on a Thermo Scientific Vanquish Flex UHPLC system equipped with MAbPac RP 2.1 mm x 50 mm column. 2 µg of material was loaded on the column heated to 80 °C. LC-MS runtime was set to 22 minutes with flow rate of 250 µL/min. Gradient elution was performed

using mobile phases A (water/0.1% formic acid) and B (ACN/0.1% formic acid): 25 to 46% B for 14 minutes.

All top-down MS experiments were performed on a Thermo Scientific Q Exactive HF-X instrument (Thermo Fisher Scientific, Bremen, Germany)[37]. LC-MS data were collected with instrument set to the Intact Protein Mode. For analysis of intact proteins, a resolution of 7,500 at 200 *m/z* was used. Full MS scans were acquired for the range of 150 to 2,000 m/z with AGC target set to 3e6. Maximum of injection time was defined at 16 ms with 1 µscan recorded.

Absorbance Measurements

B-PE subunits separated with RP-LC were collected at the time of elution following loading of 25 µg of material on the Thermo Scientific MAbPac RP LC column (Thermo Fisher Scientific). Absorbance spectra were measured for the range 400-750 nm on the Thermo Scientific Multiscan GO spectrophotometer (Thermo Fisher Scientific, Ratastie, Finland). Fractions of 250 µl corresponding to each subunit were loaded into 96-well plate. Spectra were recorded in the precision mode and corresponding absorbance values were exported with Thermo Scientific SkanIt Software (Thermo Fisher Scientific). Background was measured as of the respective buffer and subsequently subtracted from absorbance values of the samples.

Native Top-down MS/MS on QE-UHMR Mass Spectrometer

B-PE at concentration of ~2 µM was introduced into Q Exactive mass spectrometer with Ultra High Mass Range (QE-UHMR, Thermo Fisher Scientific, Bremen, Germany) via in-house pulled gold-coated borosilicate capillaries. Sample was sprayed at capillary voltage set to 1.3 kV in positive ion mode. The following mass spectrometer parameters were used: collision gas – Nitrogen; AGC mode – fixed; noise level – 2. Ion transmission settings were as follows: S-lens voltage – 25 V, inject flatapole offset – 10 V, bent flatapole DC – 4 V, gate lens voltage – 3. Resolution setting was 8,750 (at 200 *m/z*) and ion injection time was set to 100 ms. Instrument calibration was performed using cesium iodide clusters up to 11,000 *m/z*. Scan mass range was between 300 and 20,000 Th for all experiments. For measurements of intact complex, source trapping voltage was set to 25 V and HCD voltage was defined at 10 V. For native MS/MS experiments peaks of interest were isolated with 8-10 Th width, ion injection time was increased to 500 ms, and HCD voltage was elevated to 150 V. Each spectrum was obtained by averaging ~100 microscans in the time domain. Pseudo-MS3 analysis of B-PE complexes that have already ejected one or more subunits, are described in the Supplemental Experimental Procedures.

Identification of B-PE Proteoforms

First, the gamma subunits were identified through matching of the sequence tags observed in the full MS. A gamma proteoform database of all possible sequence truncations was created, based on the in literature described distinct sequences and here detected sequence tags, with the addition of a variable number of chromophores, up to the number of available cysteines. Then, the masses of these created proteoforms were matched with 2 Th tolerance to the mass features in triplicate

LC-MS runs. All the proteoforms matched in at least two out of three runs were manually verified with information available from the bottom-up LC-MS/MS data regarding the maximum number of detected chromophores per γ subunit and respective sequence coverage provided by the detected peptides. The most abundant proteoform of the γ subunit was further investigated by direct injection on a Thermo Scientific Orbitrap Fusion™ Lumos™ Tribrid™ Mass Spectrometer and sequenced by using EThcD MS/MS (see Supplemental Experimental Procedures).

Data Analysis

Raw bottom-up LC-MS/MS data was analyzed with Proteome Discoverer 2.2 (Thermo Fisher Scientific) equipped with Byonic nodes (Protein Metrics, Cupertino, USA). Following parameters were used for database search. Protease: Trypsin (full). Variable modifications: Met oxidation; Cys carbamidomethyl; Cys chromophorylations of 586.279 and 590.31 Da. Protein sequence database was generated based on recently published genome of *P. purpureum* (*P. cruentum*)[31]. Sequence alignment was performed in R with the use of "msa" package[38] and MUSCLE algorithm[34]. Top-down LC-MS raw files were deconvoluted by Sliding Window ReSpect algorithm available in Protein Deconvolution 4.0 software package (Thermo Fisher Scientific). Zero charged mass distribution profiles were obtained from raw native mass spectra with UniDec[39]. Structural visualization of phycoerythrin complexes was done in PyMOL (Schrödinger). Chemical structures of bilin molecules were drawn in ChemDraw (PerkinElmer). All additional data analysis was performed in R; hierarchical clustering was performed using algorithm that implements Ward's criterion[40]; data was visualized with ggplot2 package[41].

Data and software availability

The data have been deposited to the ProteomeXchange Consortium via the PRIDE[42] partner repository with the dataset identifier PXD011275. The native MS data relevant to the study is available upon request.

SUPPLEMENTARY MATERIAL

Table S1 | B-phycoerythrin proteins identified by bottom-up peptide-centric LC-MS/MS. The relative abundance is based on the sum of the abundances of the detected unique peptides.

	Description	Average Relative Abundance	Top BLAST hit
1	B-phycoerythrin beta chain OS=Porphyridium purpureum GN=cpeB PE=1 SV=1	1.51E+10	B-phycoerythrin beta chain OS=Porphyridium purpureum GN=cpeB PE=1 SV=1
2	B-phycoerythrin alpha chain OS=Porphyridium purpureum GN=cpeA PE=1 SV=1	4.43E+09	B-phycoerythrin alpha chain OS=Porphyridium purpureum GN=cpeA PE=1 SV=1
3	evm.model.contig_2059.29	3.17E+08	gamma 31 kDa subunit of phycoerythrin precursor
4	evm.model.contig_2100.11	2.44E+08	gamma 31 kDa subunit of phycoerythrin precursor
5	evm.model.contig_3415.11	2.16E+08	hypothetical protein
6	evm.model.contig_2059.28	1.71E+08	gamma 31 kDa subunit of phycoerythrin precursor
7	evm.model.contig_3399.5	7.89E+07	gamma 31 kDa subunit of phycoerythrin precursor
8	evm.model.contig_2146.16	2.06E+07	Phycobilisome 27.9 kDa linker polypeptide, phycoerythrin-associated, rod
9	evm.model.contig_3466.4	1.66E+07	phycoerythrin-associated linker protein
10	R-phycocyanin-1 beta chain OS=Porphyridium purpureum GN=rpcB PE=1 SV=1	8.87E+06	R-phycocyanin-1 beta chain OS=Porphyridium purpureum GN=rpcB PE=1 SV=1
11	R-phycocyanin-1 subunit alpha OS=Porphyridium purpureum GN=rpcA PE=1 SV=1	8.33E+06	R-phycocyanin-1 subunit alpha OS=Porphyridium purpureum GN=rpcA PE=1 SV=1
12	evm.model.contig_725.2	6.06E+06	phycoerythrin-associated linker protein
13	evm.model.contig_2287.2	3.18E+06	conserved unknown protein
14	evm.model.contig_3399.1	1.77E+06	phycobilisome linker polypeptide
15	evm.model.contig_3693.7	6.59E+05	phycoerythrin-associated linker protein
16	evm.model.contig_2400.2	6.01E+05	evm.model.contig_2400.2
17	evm.model.contig_2313.4	4.01E+05	serine/threonine protein kinase
18	evm.model.contig_4496.2	3.21E+05	hypothetical protein
19	evm.model.contig_2141.3	2.97E+05	hypothetical protein
20	Allophycocyanin alpha subunit OS=Porphyridium purpureum GN=apcA PE=3 SV=1	2.06E+05	Allophycocyanin alpha subunit OS=Porphyridium purpureum GN=apcA PE=3 SV=1
21	evm.model.contig_4398.4	2.05E+05	NA
22	evm.model.contig_481.3	1.76E+05	hypothetical protein
23	evm.model.contig_3427.10	1.28E+05	polyphosphate kinase
24	evm.model.contig_604.1	1.25E+05	Phycobilisome linker polypeptide:CpcD phycobilisome linker-like
25	Ribulose bisphosphate carboxylase small subunit OS=Porphyridium purpureum GN=rbcS PE=4 SV=1	9.10E+04	Ribulose bisphosphate carboxylase small subunit OS=Porphyridium purpureum GN=rbcS PE=4 SV=1
26	evm.model.contig_2020.12	2.57E+04	evm.model.contig_2020.12

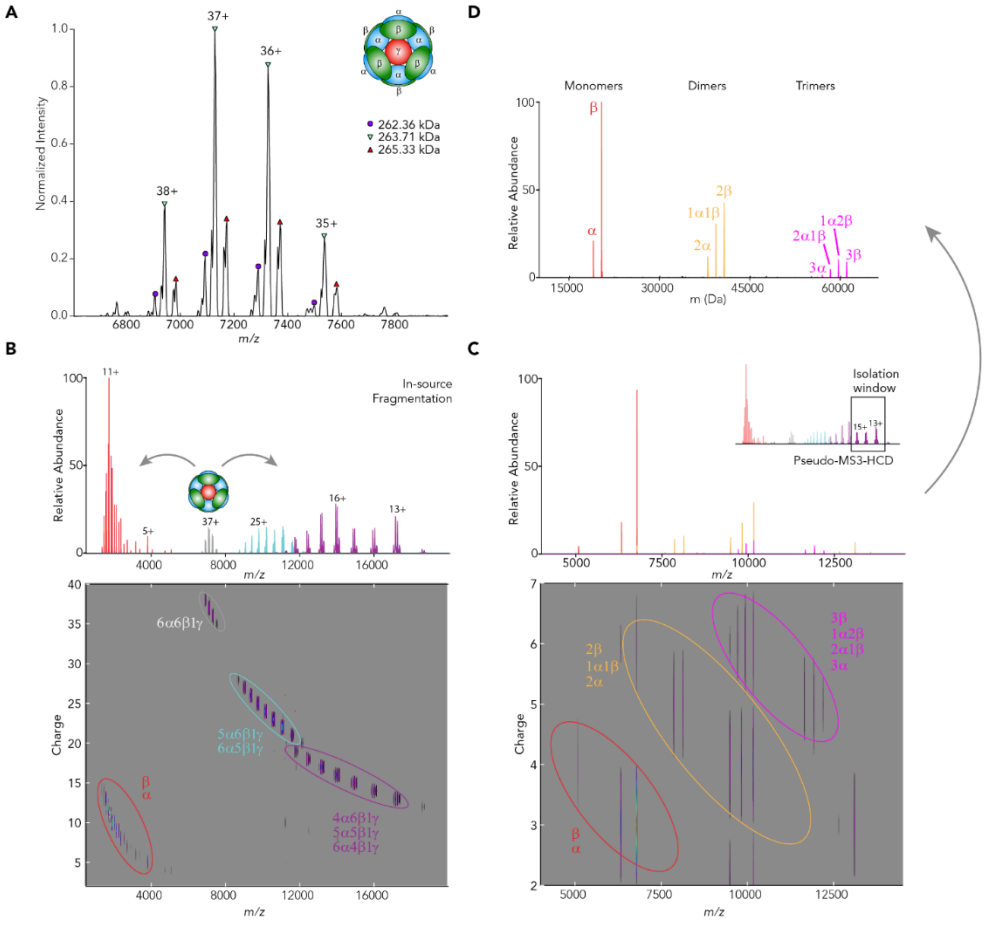
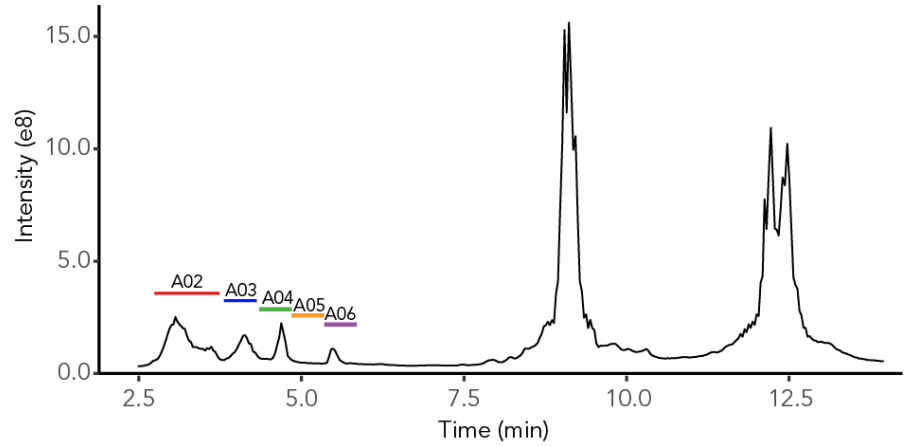


Figure S1 | (A) The high-resolution native mass spectrum of B-phycoerythrin (B-PE) reveals that the 13-mer protein assembly is present in different co-occurring forms exhibiting molecular weights of around 260 kDa. The deconvoluted masses of the distinct assembly states are given in the inset and the peaks used for these mass calculations are depicted. Three of the most abundant assemblies have approximate MWs of 262.36, 263.71 and 265.33 kDa, respectively, but the spectra also reveal several additional assemblies. (B) In-source fragmentation of B-PE assembly envelope on the QE UHMR (ST: 125 V) leads to the ejection of up to 2 monomeric subunits (top) of α and/or β. Significant charge state losses upon dissociation indicate that charge redistribution and monomer unfolding occurred prior to ejection (bottom). (C) Pseudo-MS3-HCD spectrum of residual complexes with charge states 13–15+ isolated in the range of 14500-18000 *Th* (inset) after in-source fragmentation leads to the formation of monomers, dimers, and trimers of α and/or β subunits. However, complementary residual complexes with γ were not observed. See details in the Supplemental Experimental Procedures. (D) The zero-charge deconvoluted spectrum of (C).



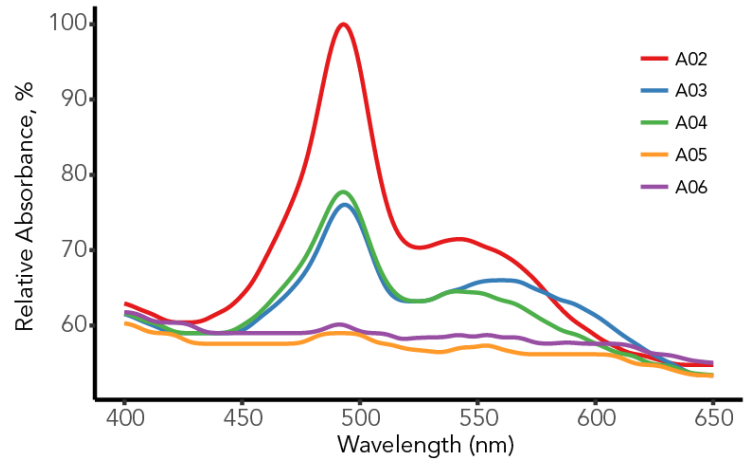


Figure S2 | Absorption spectra for individual fractions of distinctively eluting γ subunits and linker protein(s). These fractions containing distinct γ subunits demonstrate alike absorbance profiles, also similar to the absorption spectrum of the pooled fractions shown in Figure 2 of the main text of the manuscript.

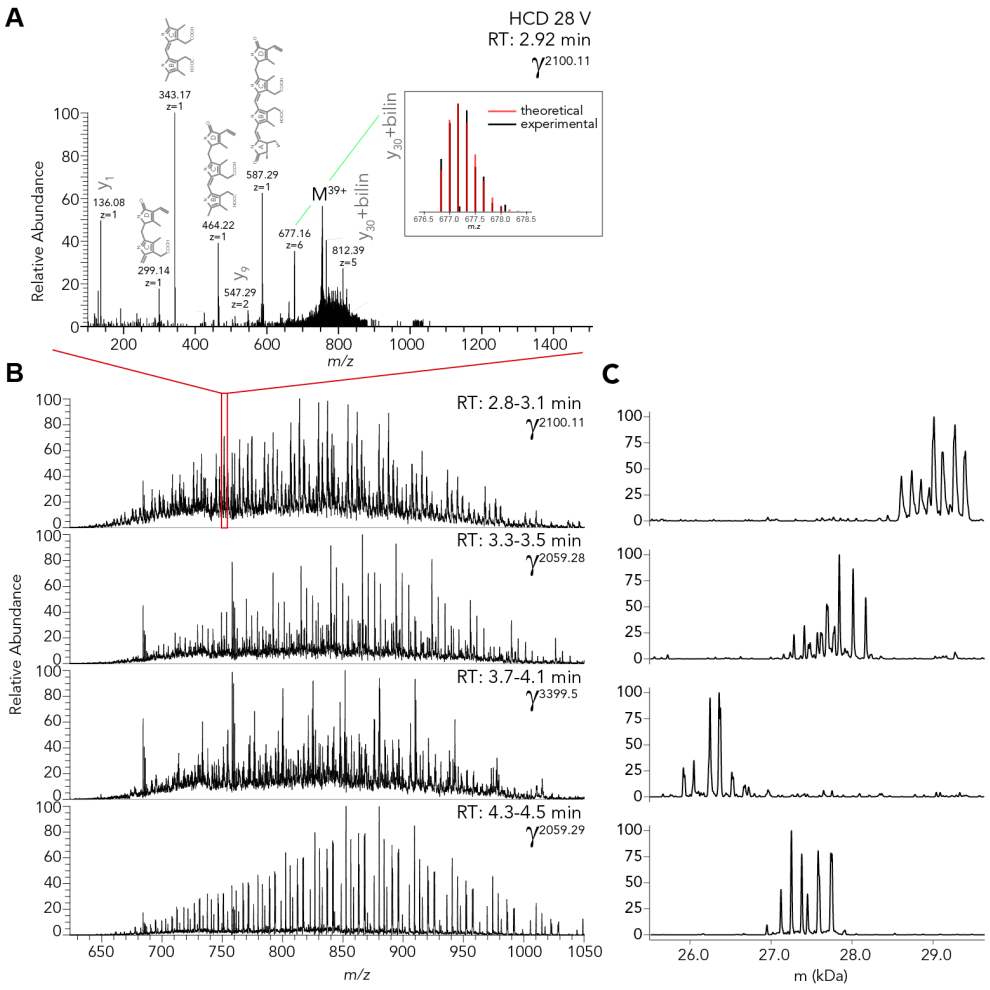


Figure S3 | (A) HCD MS/MS spectrum of the $\gamma^{2100.11}$ (39+) proteoform. (B) Mass spectra of various proteoforms of the γ subunits separated by reversed-phase LC and (C) corresponding deconvoluted zero-charge mass spectra.

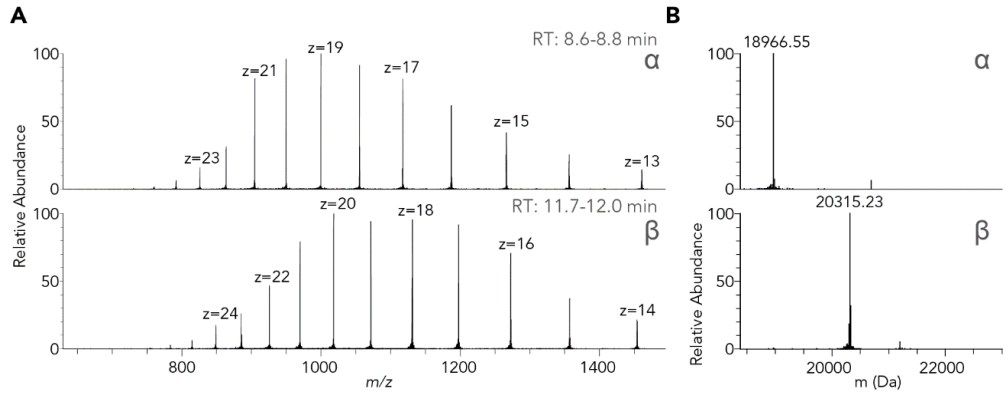


Figure S4. (A) Electrospray ionization mass spectra and (B) deconvoluted monoisotopic masses of the α and β subunits of B-PE.

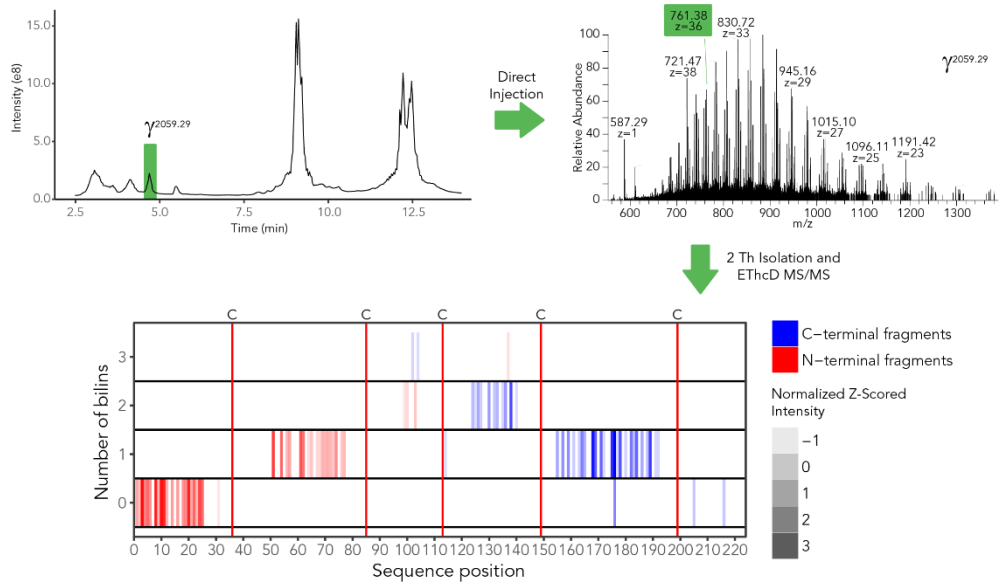


Figure S5 | Fragmentation map of the most abundant $\gamma^{2059.29}$ proteoform. RP-LC fraction corresponding to $\gamma^{2059.29}$ was directly injected into the Fusion Lumos Orbitrap mass spectrometer and the 36+ charged ion peak was subjected to EThcD (ETD reaction time: 2 ms, Supplemental HCD: 15 %). Fragments carrying 1, 2, and 3 bilin molecules were observed for both termini as graphically depicted at the bottom. See details in the Supplemental Experimental Procedures.

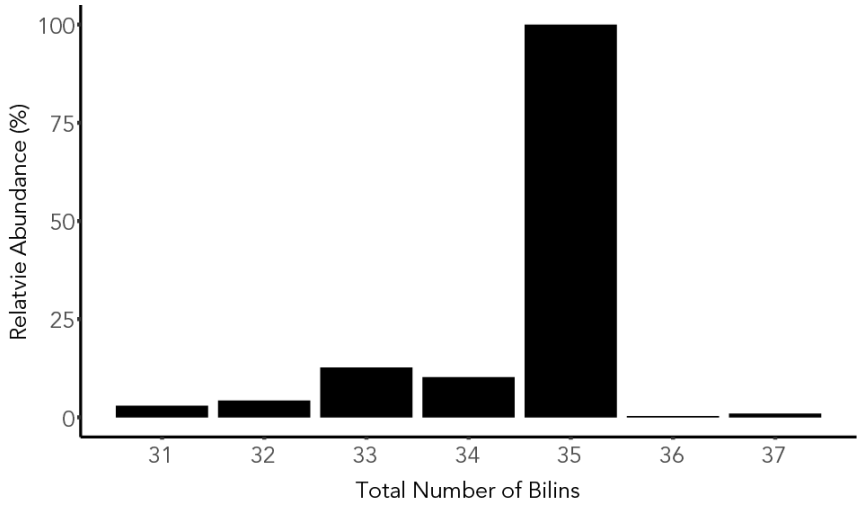


Figure S6 | Distribution of the total number of bilin chromophores attached to the proteins in B-phycoerythrin. Most of the $\alpha_6\beta_6\gamma$ assemblies harbor 35 bilins. Through variation in the γ chain some B-PE assemblies carry less bilin chromophores.

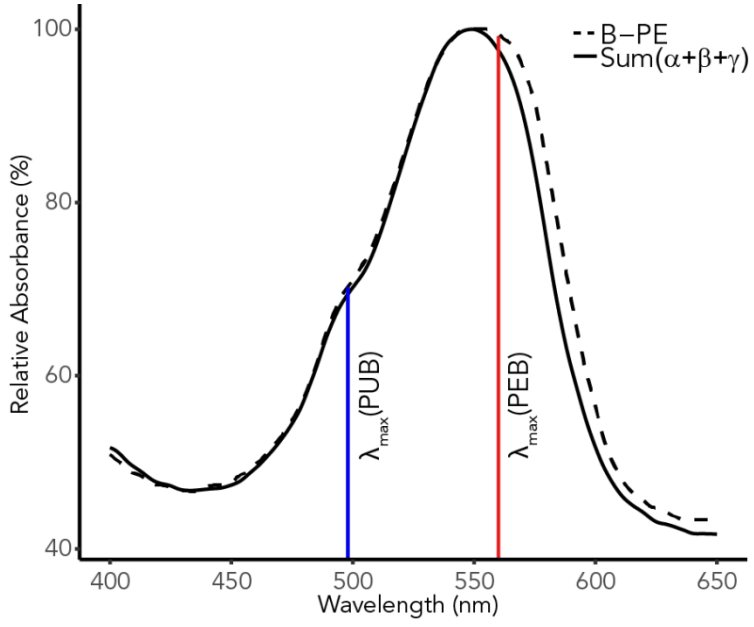


Figure S7 | Overlay of the absorbance spectrum of B-phycoerythrin (dashed line) and sum of the absorbance spectra of the B-PE subunits (solid line). Maximum absorbance wavelengths are indicated in red and blue for PEB and PUB, respectively.

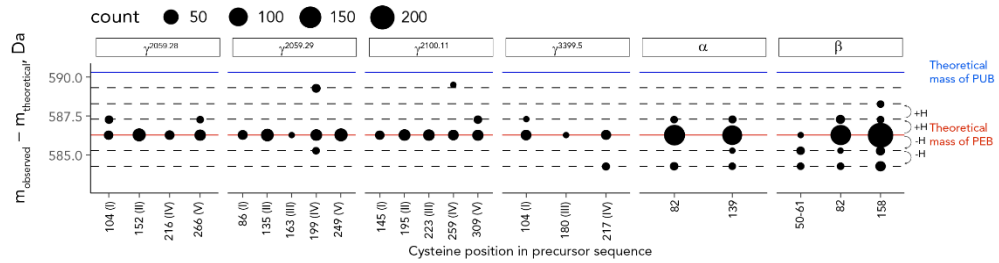


Figure S8 | Differences between experimental masses of the chromophorylated peptides detected in bottom-up LC-MS/MS experiments and the masses of theoretical, unmodified peptides of B-PE subunits. Mass differences were binned with 0.1 Da window and counted. Theoretical masses of phycoerythrobilin and phycourobilin modifications are indicated with horizontal solid lines while hydrogen(s) gain and loss is indicated with dash lines.

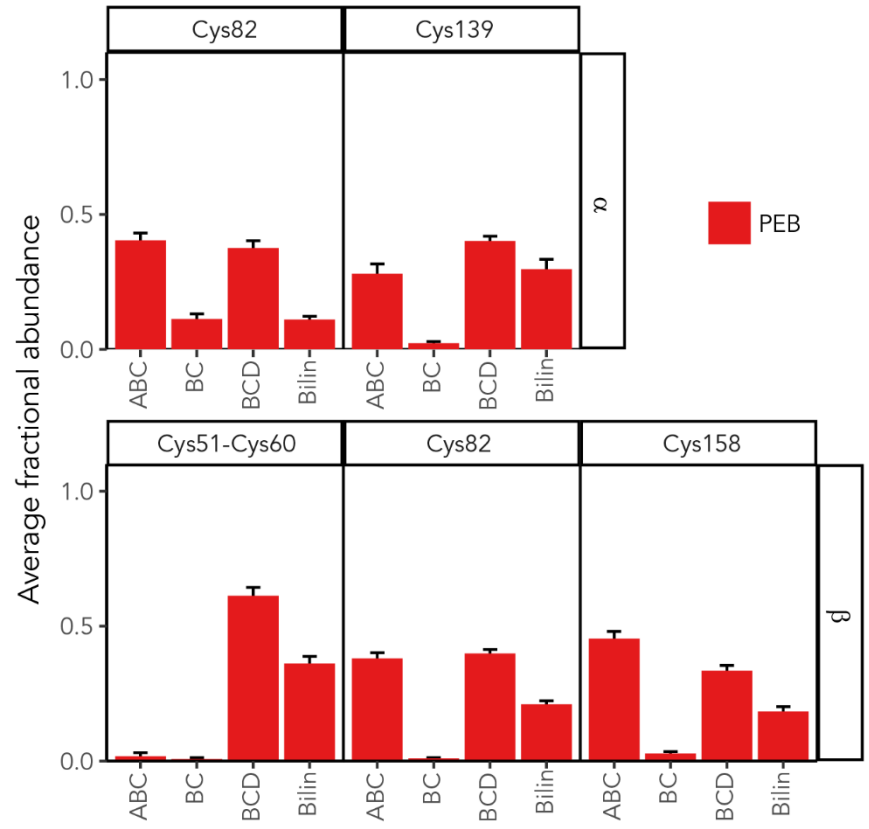


Figure S9 | Abundances of bilin fragments observed in tandem mass spectra of peptides from α and β subunits of B-PE, color-coded by the cluster from Figure 5c in the manuscript. Error bars represent standard error of the mean value.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Sample Handling for Native MS

Prior to experiments B-phycoerythrin sample was stored at 4 °C in 60% saturated ammonium sulfate, 50 mM potassium phosphate, and pH 7.0. For native MS experiments sample was transferred into Ammonium Acetate solution (300 mM at pH 7.5) without addition of extra salts/buffer.

Column Specifications for Top-down LC-MS

The MAbPac column media represents polymeric resin with large pour size (supramacroporous particles of 4 µm). These particles are inherently hydrophobic and thus there is no alkyl ligand typically used in reversed-phase separations. In our experience, the media allows for similar separation quality achieved with C4 columns.

Top-down LC-MS/MS on QE HF-X Mass Spectrometer

HF-X instrument provides an array of new features facilitating top-down analysis. Among them, Advanced Precursor Determination (APD) algorithms, that allow for on-the-fly deconvolution of monoisotopic or average masses with improved charge detection. Additionally, transmission of higher molecular weight ions (i.e. intact proteins) is improved through modified electronics and gas regime in the back-end of the instrument. LC-MS data were collected with instrument set to the Intact Protein Mode.

For top-down LC-MS/MS analysis resolution was defined as 7,500 at 200 *m/z* in full MS and 120,000 at 200 *m/z* in MS/MS. Data dependent (DD) strategy was focused on the 3 most intense proteoforms detected and on-the-fly deconvoluted in full MS scan by APD algorithms. Only the single most intense charge state (CS) was selected for isolation/fragmentation in dd-MS/MS per deconvoluted peak array with other CSs excluded from candidate list for defined exclusion time, which was set to 6 seconds. Selected ions were isolated with 2 *Th* width. Collision energy applied in dd-MS/MS was normalized for the *m/z* and charge state of selected ion with final setting of 28 %. All the dd-MS/MS scans were recorded without specifying first *m/z* value with AGC target at 3e6. Final MS/MS scan was a time-domain average of 5 µscans.

Pseudo-MS3 on QE UHMR Mass Spectrometer

B-PE at concentration of ~2 µM was introduced into Q Exactive mass spectrometer with Ultra High Mass Range (QE-UHMR, Thermo Fisher Scientific, Bremen, Germany) via in-house pulled gold-coated borosilicate capillaries. Sample was sprayed at capillary voltage set to 1.3 kV in positive ion mode. The following mass spectrometer parameters were used: collision gas – Nitrogen; AGC mode – fixed; noise level – 2. Ion transmission settings were as follows: S-lens voltage – 25 V, inject flatapole offset – 10 V, bent flatapole DC – 4 V, gate lens voltage – 3. Resolution setting was 8,750 (at 200 *m/z*) and ion injection time was set to 100 ms. Instrument calibration was performed using cesium iodide clusters up to 11,000 *m/z*. Scan mass range was between 300 and 20,000 *Th* for all experiments. For the detection of high-mass B-PE dissociation products, the in-source trapping voltage as well as Nitrogen gas pressure were optimized. For B-PE optimal in-source trapping voltage was set to 125-150 V and pressure setting was set to 3. Prior to pseudo-MS3 spectra were recorded without HCD activation to verify efficient transmission of dissociation products. To achieve pseudo-MS3 fragmentation high-m/z dissociation products were isolated with 2000 *Th* width, ion injection time was increased to 500-1000 ms, and HCD voltage was elevated to 250-300 V. Final spectra were obtained by averaging ~1000 microscans in the time domain.

Direct injection EThcD MS/MS analysis of $\gamma^{2059.29}$

25 µg of B-PE was loaded onto the MAbPac column and separated by using a RP-LC gradient. The fraction corresponding to $\gamma^{2059.29}$ was collected in the retention time range 4.2-4.8 min. MWCO (10 kDa) centrifugal filters (Amicon, Ultra Merck KGaA, Darmstadt, Germany) were used to concentrate sample with 2 cycles of dilution and concentration by aqueous solution with 10% formic acid. Resulting sample at concentration of approximately 2 µM was directly infused into the Fusion Lumos Tribrid Mass Spectrometer via in-house pulled gold-coated borosilicate capillaries. Full MS spectrum was recorded in the range from 300 to 2000 *m/z* with the instrument set to Intact Protein mode and resolution setting of 120,000 at 200 *m/z*. Ion peak with charge 36+ was isolated with 2 *Th* width and activated with EThcD. ETD reaction time was defined at 2 ms and HCD supplemental activation set to 15 %. AGC target for EThcD MS/MS was increased to 5e6 with maximum injection time of 1000 ms. Final spectrum was obtained by averaging of at least 100 scans in the time-domain. For visualization of a fragmentation map, only fragments matched within 5 ppm mass deviation were used and intensities of assigned deconvoluted fragments were z-scored, i.e. mean intensity was subtracted from intensities of all peaks and final value was divided by standard deviation. Fragmentation map was visualized in R with ggplot2 package.

ACKNOWLEDGMENTS

The Netherlands Organization for Scientific Research (NWO) supported this research through funding of the large-scale proteomics facility *Proteins@Work* (project 184.032.201) embedded in the Netherlands Proteomics Centre, and through the Spinoza Award SPI.2017.028 for AJRH. Additional support came through the European Union Horizon 2020 program FET-OPEN project MSmed (Project 686547), and the European Union Horizon 2020 program INFRAIA project Epic-XS (Project 823839). We thank Aline Tschanz for help in acquiring preliminary data for the project.

AUTHOR CONTRIBUTIONS

Conceptualization, S.T., A.C.L., and A.J.R.H.; Methodology, S.T., A.C.L., and A.J.R.H.; Investigation, S.T. and M.H.; Software, S.T. and R.A.S.; Formal Analysis, S.T.; Visualization, S.T.; Writing – Original Draft, S.T., A.C.L., and A.J.R.H.; Writing – Review & Editing, S.T., M.H., A.C.L., A.J.R.H.; Funding Acquisition, A.J.R.H.; Resources, A.J.R.H.; Supervision, A.C.L., R.A.S., and A.J.R.H.

REFERENCES

1. Grossman, A.R., Schaefer, M.R., Chiang, G.G., and Collier, J.L. (1993). The phycobilisome, a light-harvesting complex responsive to environmental conditions. Microbiol. Rev. *57*, 725–49.
2. Zhang, J., Ma, J., Liu, D., Qin, S., Sun, S., Zhao, J., and Sui, S.F. (2017). Structure of phycobilisome from the red alga Griffithsia pacifica. Nature *551*, 57–63.
3. Adir, N. (2005). Elucidation of the molecular structures of components of the phycobilisome: Reconstructing a giant. Photosynth. Res. *85*, 15–32.
4. Murakami, A., Mimuro, M., Ohki, K., and Fujita, Y. (1981). Absorption spectrum of allophycocyanin isolated from Anabaena cylindrica: Variation of the absorption spectrum induced by changes of the physico-chemical environment. J. Biochem. *89*, 79–86.
5. Glazer, A.N., Fang, S., and Brown, D.M. (1973). Spectroscopic properties of C-phycocyanin and of its alpha and beta subunits. J. Biol. Chem. *248*, 5679–5685.
6. Glazer, A.N., and Hixson, C.S. (1977). Subunit Structure and Chromophore Composition of Rhodophytan Phycoerythrins. J. Biol. Chem. *252*, 32–42.
7. Glazer, A.N. (1989). Light guides. Directional energy transfer in a photosynthetic antenna. J. Biol. Chem. *264*, 1–4.
8. Nagy, J.O., Bishop, J.E., Klotz, A. V., Glazer, A.N., and Rapoport, H. (1985). Bilin attachment sites in the α, ß, and y subunits of R-phycoerythrin. Structural studies on singly and doubly linked phycourobilins. J. Biol. Chem. *260*, 4864–4868.
9. Oi, V.T., Glazer, A.N., and Stryer, L. (1982). Fluorescent phycobiliprotein conjugates for analyses of cells and molecules. J. Cell Biol. *93*, 981–986.
10. Torres-acosta, M.A., Monterrey, D., Monterrey, C., Eugenio, A., Sada, G., Ruiz-ruiz, F., Monterrey, D., Monterrey, C., Eugenio, A., Sada, G., *et al.* (2016). Economic Analysis of Pilot-Scale Production of B-Phycoerythrin. Biotechnol Prog. *32*, 1472–1479.
11. Fleurence, J. (2003). R-Phycoerythrin from red macroalgae: Strategies for extraction and potential application in Biotechnology. Appl. Biotechnol. Food Sci. Policy *1*, 63–68.
12. Tang, Z., Zhao, J., Ju, B., Li, W., Wen, S., Pu, Y., and Qin, S. (2016). One-step chromatographic procedure for purification of B-phycoerythrin from Porphyridium cruentum. Protein Expr. Purif. *123*, 70–74.
13. Bermejo, R., Talavera, E.M., and Alvarez-Pez, J.M. (2001). Chromatographic purification and characterization of B-phycoerythrin from Porphyridium cruentum - Semipreparative high-performance liquid chromatographic separation and characterization of its subunits. J. Chromatogr. A *917*, 135–145.
14. Munier, M., Jubeau, S., Wijaya, A., Morançais, M., Dumay, J., Marchal, L., Jaouen, P., and Fleurence, J. (2014). Physicochemical factors affecting the stability of two pigments: R-phycoerythrin of Grateloupia turuturu and B-phycoerythrin of Porphyridium cruentum. Food Chem. *150*, 400–407.
15. Leney, A.C., Tschanz, A., and Heck, A.J.R. (2018). Connecting color with assembly in the fluorescent B-phycoerythrin protein complex. FEBS J. *285*, 178–187.
16. Redlinger, T., and Gantt, E. (1981). Phycobilisome structure of Porphyridium cruentum: polypeptide composition. Plant Physiol. *68*, 1375–1379.
17. Ficner, R., and Huber, R. (1993). Refined crystal structure of phycoerythrin from Porphyridium cruentum at 0.23-nm resolution and localization of the y subunit. FEBS J. *218*, 103–106.
18. Sepúlveda-ugarte, J., Brunet, J.E., Matamala, A.R., Martínez-oyanedel, J., and Bunster, M. (2011). Spectroscopic parameters of phycoerythrobilin and phycourobilin on phycoerythrin from Gracilaria chilensis. J. Photochem. Photobiol. A Chem. *219*, 211–216.
19. Lundell, D.J., Glazergt, A.N., I, R.J.D., and Douglas, M. (1984). Bilin Attachment Sites in the α and ß subunits of B-phycoerythrin. J. Biol. Chem. *259*, 5472–5480.
20. Liu, L.N., Elmalk, A.T., Aartsma, T.J., Thomas, J.C., Lamers, G.E.M., Zhou, B.C., and

Zhang, Y.Z. (2008). Light-induced energetic decoupling as a mechanism for phycobilisome-related energy dissipation in red algae: A single molecule study. PLoS One *3*, e3134.

21. Swanson, R. V., and Glazer, A.N. (1990). Separation of phycobiliprotein subunits by reverse-phase high-pressure liquid chromatography. Anal. Biochem. *188*, 295–299.

22. Klotz, A. V., and Glazer, A.N. (1985). Characterization of the bilin attachment sites in R-phycoerythrin. J. Biol. Chem. *260*, 4856–4863.

23. Skinner, O.S., Haverland, N.A., Fornelli, L., Melani, R.D., Vale, L.H.F. Do, Seckler, H.S., Doubleday, P.F., Schachner, L.F., Srzentic, K., Kelleher, N.L., *et al.* (2018). Top-down characterization of endogenous protein complexes with native proteomics. Nat. Chem. Biol. *14*, 36–41.

24. Wu, D., Struwe, W.B., Harvey, D.J., Ferguson, M.A.J., and Robinson, C. V. (2018). N-glycan microheterogeneity regulates interactions of plasma proteins. Proc. Natl. Acad. Sci. *115*, 8763–8768.

25. Franc, V., Zhu, J., and Heck, A.J.R. (2018). Comprehensive Proteoform Characterization of Plasma Complement Component C8αβγ by Hybrid Mass Spectrometry Approaches. J. Am. Soc. Mass Spectrom. *29*, 1099–1110.

26. Vásquez-Suárez, A., Lobos-González, F., Cronshaw, A., Sepúlveda-Ugarte, J., Figueroa, M., Dagnino-Leone, J., Bunster, M., and Martínez-Oyanedel, J. (2018). The γ33 subunit of R-phycoerythrin from Gracilaria chilensis has a typical double linked phycourobilin similar to γ subunit. PLoS One *13*, e0195656.

27. Nair, D., Krishna, J.G., Panikkar, M.V.N., Nair, B.G., Pai, J.G., and Nair, S.S. (2018). Identification, purification, biochemical and mass spectrometric characterization of novel phycobiliproteins from a marine red alga, Centroceras clavulatum. Int. J. Biol. Macromol. *114*, 679–691.

28. Wohlschlager, T., Scheffler, K., Forstenlehner, I.C., Skala, W., Senn, S., Damoc, E., Holzmann, J., and Huber, C.G. (2018). Native mass spectrometry combined with enzymatic dissection unravels glycoform heterogeneity of biopharmaceuticals. Nat. Commun. *9*, 1–9.

29. Franc, V., Yang, Y., and Heck, A.J.R. (2017). Proteoform profile mapping of the human serum Complement component C9 reveals unexpected new features of N- , O- and C-glycosylation. Anal. Chem. *89*, 3483–3491.

30. Yang, Y., Liu, F., Franc, V., Halim, L.A., Schellekens, H., and Heck, A.J.R. (2016). Hybrid mass spectrometry approaches in glycoprotein analysis and their usage in scoring biosimilarity. Nat. Commun. *7*, 1–10.

31. Bhattacharya, D., Price, D.C., Chan, C.X., Qiu, H., Rose, N., Ball, S., Weber, A.P.M., Cecilia Arias, M., Henrissat, B., Coutinho, P.M., *et al.* (2013). Genome of the red alga Porphyridium purpureum. Nat. Commun. *4*, 1941.

32. Liu, L.N., Chen, X.L., Zhang, Y.Z., and Zhou, B.C. (2005). Characterization, structure and function of linker polypeptides in phycobilisomes of cyanobacteria and red algae: An overview. Biochim. Biophys. Acta - Bioenerg. *1708*, 133–142.

33. Apt, K.E., Hoffman, N.E., and Grossman, A.R. (1993). The γ subunit of R-phycoerythrin and its possible mode of transport into the plastid of Red algae. J Biol Chem. *268*, 16208–15.

34. Edgar, R.C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

35. Blanksby, S.J., and Ellison, G.B. (2003). Bond dissociation energies of organic molecules. Acc. Chem. Res. *36*, 255–263.

36. Mancini, J.A., Sheehan, M., Kodali, G., Chow, B.Y., Bryant, D.A., Dutton, P.L., and Moser, C.C. (2018). De novo synthetic biliprotein design, assembly and excitation energy transfer. J. R. Soc. Interface *15*, 20180021.

37. Kelstrup, C.D., Bekker-Jensen, D.B., Arrey, T.N., Hogrebe, A., Harder, A., and Olsen, J. V. (2017). Performance evaluation of the Q Exactive HF-X for shotgun proteomics. J. Proteome Res. *17*, 727–738.

38. Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., and Hochreiter, S. (2015). Msa: An R package for multiple sequence alignment. Bioinformatics *31*, 3997–3999.

39. Marty, M.T., Baldwin, A.J., Marklund, E.G., Hochberg, G.K.A., Benesch, J.L.P., and Robinson, C. V. (2015). Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. Anal. Chem. *87*, 4370–4376.

40. Murtagh, F., and Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J. Classif. *31*, 274–295.

41. Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer New York).

42. Vizcaino, J.A., Csordas, A., Griss, J., Lavidas, I., Mayer, G., Perez-riverol, Y., Reisinger, F., Ternent, T., Xu, Q., Wang, R., *et al.* (2016). 2016 update of the PRIDE database and its related tools. *44*, 447–456.

7

CHAPTER

SUMMARY, SAMMENVATTING, PERSPECTIVE AND OUTLOOK

SUMMARY

Most cellular processes depend on a multitude of molecular machines, predominantly comprised of proteins. The diverse set of tasks performed by these machines, known as multi-proteoform complexes (MPCs), is driven by interactions between heterogeneous sets of proteoforms, products of the multi-stage process of protein synthesis and maturation. To understand the biological processes driven by the MPCs, researchers attempt to characterize both the involved proteoforms individually and the entire protein assemblies in-depth with various techniques capable of uncovering structural details. These techniques reveal, for example, the primary sequence of proteoforms and higher-order structural features of the individual subunits as well as the full assemblies. As reviewed in **Chapter 1**, different structural methods have distinctive stronger and weaker points. While some provide a high-resolution and near-atomic structural snapshot, others allow researchers to infer specific structural features in either a high-throughput or a targeted fashion at a lower resolution. Mass spectrometry has emerged in the last decades as a versatile and highly complementary technique with respect to established high-resolution structural methods. With mass spectrometry, it is possible to transfer proteins from their natural in-solution environment to the gas-phase, predominantly by electrospray ionization (ESI), and separate them based on their mass-to-charge (*m/z*) ratio. In a secondary, or tandem, step ions selected based on their *m/z* ratios can be activated, which results in the disruption of covalent and/or non-covalent bonds, producing characteristic dissociation products.

In this thesis, the versatility and power of mass spectrometry applied to intact MPCs are demonstrated by several experimental approaches, sub-divided in two major sections: **(I)** the structural analysis of MPCs through gas-phase activation, and **(II)** the in-depth characterization of sophisticated multi-proteoform assemblies with hybrid MS-based approaches. For **(I)** we have expanded the variety of dissociation methods with a technique that offers distinctive and highly complementary fragmentation patterns to conventional collision-based methods, by equipping a benchtop Orbitrap mass spectrometer with a 193 nm UV laser (ultraviolet photodissociation; UVPD). In **Chapter 2,** the differences of UVPD to a widely established collisional dissociation technique are investigated on six different MPCs, which range from dimers up to heptamers. As opposed to collisional dissociation, which predominantly results in unfolding and partitioning of the single highly-charged subunit, UVPD leads to the ejection of compact as well as denatured subunits. This behavior is system-dependent with a higher degree of subunit unfolding upon increasing binding interfaces, providing insight about higher-order subunit arrangement within the full assembly. Additionally, UVPD results in a higher degree of covalent backbone fragmentation, leading up to a more confident identification of the involved proteoforms.

Activation of MPCs can also uncover the stability of protein assemblies in the gas-phase. As discussed in **Chapter 3,** collisional activation of protein assemblies leads to an array of molecular rearrangements that result in a gradual unfolding of the monomer(s) prior to ejection from the assembly. By monitoring both the size of an

activated assembly as well as the produced dissociation products by ion mobility MS as a function of varying activation energy, we were able to assess and compare stabilities of functionally and structurally homologous, yet distinctive, homomeric co-chaperonins GroES and gp31. These two look-alike assemblies demonstrated distinctive unfolding patterns and fragmentation yields, indicating that GroES exhibits a more stable structural organization. Both in-solution experiments, as well as available crystal structures, further corroborated the conclusions derived from the gas-phase data, ultimately indicating that the assemblies at least partly retain their native structural features upon transition into the gas-phase.

Probing of binding interfaces between molecules is a difficult task with immense biological implications. Such knowledge is critical for the development of binding disruptors or enhancers, depending on the clinical need. Often creative experimental designs are required for elucidating this structural feature. In **Chapter 4**, we use MS to uncover the interaction interface of a non-covalent, phosphorylation mediated protein-phosphopeptide complex. To achieve this, we used a two-step dissociation procedure. In the first step, we dissociated the complex, which was accompanied by the transfer of phosphorylation groups to the protein. Selecting the liberated protein with the transferred phosphorylation groups and dissociating the covalent backbone allowed for localization of the transferred phosphorylation groups and, hence, exposed the interaction interface.

Although activation of intact protein complexes is a powerful tool for the investigation of higher-order structure, at increasing molecular sizes, the amount of information tends to decrease. Such a situation is further aggravated by the presence of non-protein components as is the case for RNA-intertwined mega-Dalton (MDa) ribosomal particles. For increasing the information content, multiple MS-based approaches can be combined. In **Chapter 5**, we use an integrative MS-based method to disentangle ribosomal particles purified from bacteria, plant leaves, and human cells. MS analysis on the digested proteins allowed us to investigate the protein content in each sample, analysis of the intact, albeit denatured proteins elucidated the major proteoforms of ribosomal subunits, and finally, analysis by native MS provided mass information on the complete structure and integrity of the distinctive ribosomal particles. Combined, these three tiers of MS analyses yielded a broad view on the compositional and structural diversity in the analyzed ribosomal samples, identifying novel PTM sites as well as revealing distinctive assembly variants.

In **Chapter 6,** we characterized the protein assembly B-phycoerythrin (B-PE), which is part of the light-harvesting machinery in red algae and cyanobacteria. B-PE is a highly heterogeneous heteromeric protein assembly decorated with multiple chromophores covalently attached to cysteines. From the native MS data, we discovered that B-PE is present as a heterogeneous mixture of MPCs, which are not readily explained from the information available in the literature. Gas-phase activation of the B-PE assemblies was used and allowed us to attribute this heterogeneity predominantly to the low stoichiometry the single γ subunit that exhibited high proteoform variety. To unravel all the unique proteoforms of the B-PE subunits, we obtained MS data on the denatured assemblies. As a result, we identified four distinct gene products, each represented by multiple sequence variants decorated with a

varying number of chromophores. In the final step, MS data on the digested proteins provided the location and identity of the isobaric chromophores, which were distinguished by their distinct absorption and characteristic fragmentation patterns.

SAMENVATTING

De meeste cellulaire processen zijn afhankelijk van moleculaire machines die voornamelijk bestaan uit eiwitten. De diverse taken die door deze machines worden uitgevoerd, bekend als multi-proteovormcomplexen (MPC's), worden gedreven door interacties tussen heterogene collecties van proteovormen die het product zijn van eiwit-synthese en -rijping. Onderzoekers die biologische processen proberen te begrijpen, bestuderen zowel de betrokken proteovormen afzonderlijk als de gehele eiwitassemblages met verschillende technieken die structurele details kunnen blootleggen. Deze technieken onthullen bijvoorbeeld de primaire sequentie van proteovormen, de hogere orde structurele kenmerken van de afzonderlijke subeenheden evenals de volledige assemblages, en vele andere details. Zoals besproken in **hoofdstuk 1** hebben verschillende structurele methoden sterkere en zwakkere punten. Terwijl sommige een hoge resolutie en bijna-atomaire structurele momentopname bieden, laten anderen onderzoekers specifieke structurele kenmerken afleiden met een hoge doorvoer van monsters of op een gerichte manier met een lagere resolutie. Massaspectrometrie is in de afgelopen decennia naar voren gekomen als een veelzijdige en complementaire methode voor bestaande structurele methoden die een hoge resolutie bieden. Met massaspectrometrie is het mogelijk om eiwitten van hun natuurlijke omgeving in oplossing naar de gasfase over te brengen, voornamelijk via elektrospray ionisatie (ESI), en ze te scheiden op basis van hun massa-lading (m/z) verhouding. In een secundaire of tandem stap kunnen ionen worden gefragmenteerd die zijn geselecteerd op basis van hun m/z, wat resulteert in de verstoring van covalente en/of niet-covalente bindingen, waardoor karakteristieke dissociatieproducten worden geproduceerd.

In dit proefschrift worden de veelzijdigheid en kracht van massaspectrometrie voor de studie van intacte MPC's aangetoond door verschillende experimentele benaderingen, onderverdeeld in twee overkoepelende secties: **(I)** de structurele analyse van MPC's door gasfase-activering, en **(II)** de karakterisering van geavanceerde multi-proteovorm-assemblages in detail met hybride MS-gebaseerde benaderingen. Voor **(I)** hebben we de beschikbare dissociatiemethoden uitgebreid met een techniek die complementaire fragmentatiepatronen biedt aan conventionele methoden, door een benchtop Orbitrap massaspectrometer uit te rusten met een 193 nm UV-laser (ultraviolette fotodissociatie; UVPD). In **hoofdstuk 2** worden de verschillen tussen UVPD en een conventionele dissociatie techniek onderzocht op zes verschillende MPC's, die variëren van dimeren tot heptameren. In tegenstelling tot de conventionele techniek, die voornamelijk resulteert in het ontvouwen en partitioneren van de hoogst geladen subeenheid, leidt UVPD tot het uitwerpen van zowel compacte als gedenatureerde subeenheden. Dit gedrag is systeemafhankelijk, met een hogere mate van zich ontvouwende subeenheden bij sterker bindende interfaces, waardoor inzicht wordt verkregen in de rangschikking van hogere orde

subeenheden. Bovendien resulteert UVPD in een hogere mate van covalente ruggengraatfragmentatie, wat leidt tot een verbeterde identificatie van de betrokken proteovormen.

Activering van MPC's kan ook de stabiliteit van eiwitassemblages in de gasfase blootleggen. Zoals besproken in **hoofdstuk 3**, leidt activering van eiwitassemblages tot een reeks moleculaire herschikkingen die resulteren in een geleidelijke ontplooiing van de monomeren voorafgaand aan hun uitstoting. Door de grootte van de geproduceerde dissociatieproducten te monitoren door ionenmobiliteit MS als een functie van variërende activeringsenergie was het mogelijk de stabiliteit te vergelijken van de functioneel en structureel homologe co-chaperoninen GroES en gp31. Deze twee sterk op elkaar lijkende assemblages vertoonden onderscheidende ontplooiingspatronen en fragmentatie producten, wat er op duidde dat GroES een stabielere structurele organisatie heeft. Onze experimenten in oplossing alsmede de beschikbare kristalstructuren bevestigden de conclusies afgeleid uit de gasfase gegevens, wat aangeeft dat de assemblages ten minste gedeeltelijk hun structurele kenmerken behouden bij de overgang naar de gasfase.

Het ophelderen van de bindingsinterfaces tussen moleculen is een moeilijke taak met immense biologische implicaties. Dergelijke kennis is van cruciaal belang voor de ontwikkeling van binding verstoorders of versterkers, afhankelijk van de klinische behoefte. Vaak zijn creatieve experimentele ontwerpen vereist om dit structurele kenmerk op te helderen. In **hoofdstuk 4** gebruiken we MS om de interface van een niet-covalent, fosforylatie-gemedieerd eiwit-fosfopeptide complex aan het licht te brengen. Om dit te bereiken hebben we een tweestaps dissociatieprocedure gebruikt. In de eerste stap hebben we het complex gedissocieerd, wat gepaard ging met de overdracht van fosforylatiegroepen naar het eiwit. Het selecteren van het vrijgemaakte eiwit met de overgebrachte fosforylatiegroepen en het dissociëren van de covalente ruggengraat maakte lokalisatie van de overgebrachte fosforylatiegroepen mogelijk en legde de interactie-interface bloot.

Hoewel activering van intacte eiwitcomplexen een krachtig middel is voor het onderzoeken van hogere orde structuren, neemt de hoeveelheid informatie bij toenemende molecuulgrootte af. Een dergelijke situatie wordt verder verergerd door de aanwezigheid van niet-eiwitcomponenten, zoals het geval is voor de RNA-verweven mega-Dalton (MDa) ribosomen. Om de informatie-inhoud te vergroten kunnen meerdere op MS gebaseerde benaderingen worden gecombineerd. In **hoofdstuk 5** gebruiken we een integrale MS-gebaseerde methode om ribosomen te karakteriseren die gezuiverd zijn uit bacteriën, plantenbladeren en menselijke cellen. MS-analyse van de verteerde eiwitten stelde ons in staat om het eiwitgehalte in elk monster te onderzoeken; analyse van de intacte, zij het gedenatureerde eiwitten, lichtte de belangrijkste proteovormen van ribosomale subeenheden uit; en ten slotte leverde analyse door native MS massa informatie over de volledige structuur en integriteit van de verschillende ribosomen. Gecombineerd gaven deze drie niveaus van MS-analyses een uitgebreid beeld van de samenstelling en structurele diversiteit in de geanalyseerde ribosomale monsters, waarbij nieuwe PTM-locaties werden geïdentificeerd en assemblagevarianten werden onthuld.

In **hoofdstuk 6** hebben we de eiwitassemblage B-phycoerythrin (B-PE) beschreven, die deel uitmaakt van de lichtverzameling moleculaire machines in rode algen en cyanobacteriën. B-PE is een zeer heterogene heteromeer met meerdere chromoforen die covalent gehecht zijn aan cysteïnen. Uit de MS-gegevens hebben we ontdekt dat B-PE aanwezig is als een heterogeen mengsel van MPC's, die niet gemakkelijk worden verklaard uit de beschikbare informatie in de literatuur. Gasfase-activering van B-PE werd gebruikt om deze heterogeniteit voornamelijk toe te schrijven aan de in lage stoichiometrie aanwezige γ subeenheid die een hoge proteovorm-variëteit vertoonde. Om alle unieke proteovormen van de B-PE-subeenheden te ontrafelen, hebben we MS-gegevens verzameld van gedenatureerde assemblages. Hiermee identificeerden we vier verschillende genproducten, elk resulterend in meerdere sequentievarianten met een variërend aantal chromoforen. In een laatste stap verschaften MS-gegevens opgenomen van gedigesteerde eiwitten de locatie en identiteit van de isobare chromoforen, die konden werden onderscheiden door hun verschillende absorptie- en karakteristieke fragmentatiepatronen.

PERSPECTIVE AND OUTLOOK

As vividly portrayed in this thesis, tandem mass spectrometry is an extremely potent tool for the characterization of proteins, largely owing to the advances in instrumentation that enable to ionize and dissociate large biomolecules such as intact proteins. The ever-increasing sensitivity, speed, and resolving power of mass analyzers allow researchers to detect even the slightest mass deviations, stemming from diversity introduced by distinctive amino acid sequences and PTM patterns, not only at the proteoform level but also at the level of protein assemblies. The emergence of novel fragmentation techniques, experimental designs, and software solutions is continuously pushing the boundaries of conventional tandem MS. Promising results have already been achieved in the characterization of proteoforms with high molecular weights and/or complexity of PTM patterns; *e.g.* characterization of the histone code[1], complex biotherapeutics[2,3], and large proteoforms in fractionated human proteome[4]. Likewise, tandem MS has been successfully applied for the structural investigation of multiproteoform assemblies, recombinant[5] as well as endogenous[6].

Further improvements should extend the applicability of structural tandem MS to even more challenging systems, like membrane proteins[7] or macromolecular protein complexes with molecular weight (Mw) in the MDa range[8]. The role of mass spectrometry in structural biology will likely expand in the coming decades. To illustrate emerging possibilities of tandem mass spectrometry, an outlook on future applications is provided below.

Role of proteoform profiling in integrative structural approaches

While MS-based chemical cross-linking and surface labeling already have a proven record of accomplishment as complementary and informative methods to high-resolution structural analysis, top-down proteomics is still at the developmental stage and so far has not reached significant utilization in the field of structural biology. The ability to detect sequence variants and position all the PTMs for each distinctive

proteoform in a sample has, however, immense potential, as even low abundant substoichiometric proteoforms can have significant biological implications with truncations and PTMs altering protein structure and function.

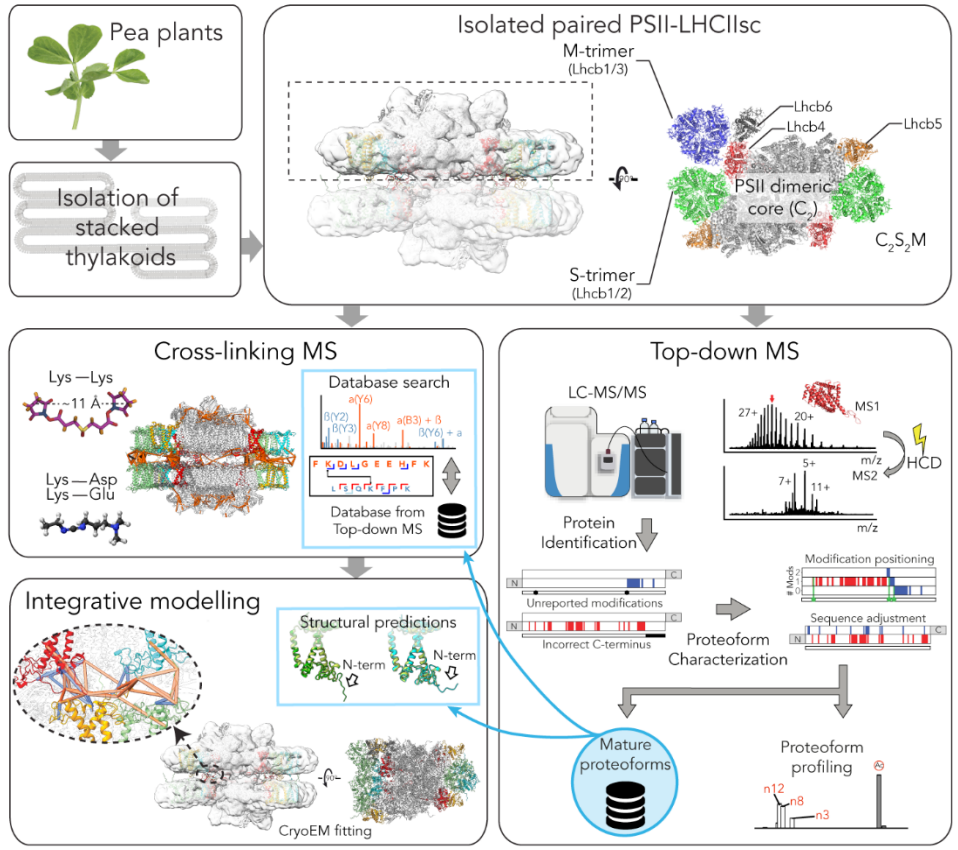


Figure 1 | Integrative structural analysis of isolated PSII-LHCII supercomplexes from *P. sativum* (pea). Paired PSII-LHCII supercomplexes were purified following isolation of stacked thylakoids from pea plants (Top boxes). Samples were analyzed in parallel with chemical cross-linking MS and top-down MS. Top-down MS allowed the identification of distinctive sequence variants and their primary proteoforms for all the subunits forming the paired PSII-LHCII supercomplexes (Bottom right box). The determined sequence variants and their proteoforms were used in the interpretation of the cross-linking MS data (Middle left box), as well as for structural modeling and fitting of subunit structures into the cryo-EM maps (Bottom left box). The figure is adapted from Albanese, P., Tamara, S., Saracco, G., Scheltema, R. A., Pagliano, C. (2019). How Paired PSII-LHCII Supercomplexes Mediate the Stacking of Plant Thylakoid Membranes Unveiled by Integrative Structural Mass Spectrometry. *Submitted*.

The prospects of proteoform profiling for the structural analysis of challenging protein systems is vividly illustrated in our investigation of supercomplexes comprising Photosystem II (PSII) and light-harvesting complex II (LHCII) (PSII-LHCII supercomplex; manuscript under review). These supercomplexes are embedded into the thylakoid membranes of higher plants, where they pair to form sandwich-like structures throughout grana stacks (Figure 1; top right panel). We utilized top-down MS as a part of the integrative structural approach to uncover the sequence variants

and their proteoforms involved in the formation of PSII-LHCII supercomplexes (Figure 1; bottom right panel). The resulting mature sequences of PSII-LHCII subunits were used to analyze the data collected in parallel cross-linking MS experiments (Figure 1; middle left panel). Next, the restraints detected for PSII-LHCII subunits were used to model the final proteoforms and fit them into existing cryo-EM maps (Figure 1; bottom left panel). The final structural data indicated that N-terminal truncations combined with N-terminal acetylation, the major PTMs observed in purified PSII-LHCII proteins, play an important role in the pairing of PSII-LHCII supercomplexes.

This combined approach outlines a new hybrid direction in structural proteomics, integrating top-down and cross-linking mass spectrometry for the investigation of challenging protein assemblies. Whereas it was largely ignored in the past, proteoform profiling lays a solid foundation for interpreting the structural information obtained with alternative structural techniques, like cross-linking MS. These benefits will meet its full potential in the examination of large endogenous protein assemblies, for which complex structural datasets require external characterization of the highly heterogeneous building blocks.

Extending the mass range and sample complexity in native tandem mass spectrometry

Historically, the amount of structural information extracted with native MS experiments on protein assemblies was mainly limited to the subunit stoichiometry and, if the analyte was of known origin, the composition. Practically, the scope of structural native MS analysis was for a long time hampered by the limitations of conventional collisional fragmentation techniques, which predominantly result in the ejection of a single unfolded subunit. Additionally, the majority of high-resolution instruments struggle to simultaneously detect disparate products of dissociation, as low-*m/z* covalent backbone fragments and high-*m/z* non-dissociated residual assemblies exhibit widely different properties and require distinctive measuring conditions. The situation is further impaired for multimeric macromolecular assemblies, whereby gas-phase activation results in the formation of ultra-high-*m/z* products outside the mass boundaries of most available instruments. Advanced instrumentation and novel fragmentation techniques are actively being developed with the aim to extend the scope of informative dissociation products and enable their simultaneous detection.

The new instruments featuring extended mass range and optimized parameters for the analysis of high-*m/z* ions have recently allowed not only for the successful detection of large macromolecular protein complexes like ribosomes and viral particles[9] but also for the analysis of all the products of their gas-phase dissociation. We exemplify this by using an Orbitrap mass spectrometer with ultra-high mass range (UHMR) capabilities for the native tandem MS analysis of the human 40S ribosomal subunit (HS40S) with a Mw of 1.215 MDa. By mass selecting and activating the most abundant charge state of this assembly (z = 74+) with an isolation window of 100 Th, highly informative fragmentation products of this very large macromolecular assembly can be achieved. Ramping up activation energies allowed us to monitor the

ejection of several distinctive subunits, identified using masses of low-*m/z* products or mass losses in high-*m/z* products. Consequently, the data provided structural insights about the identity and localization of these subunits in regard to the bulk of the assembly (Figure 2).

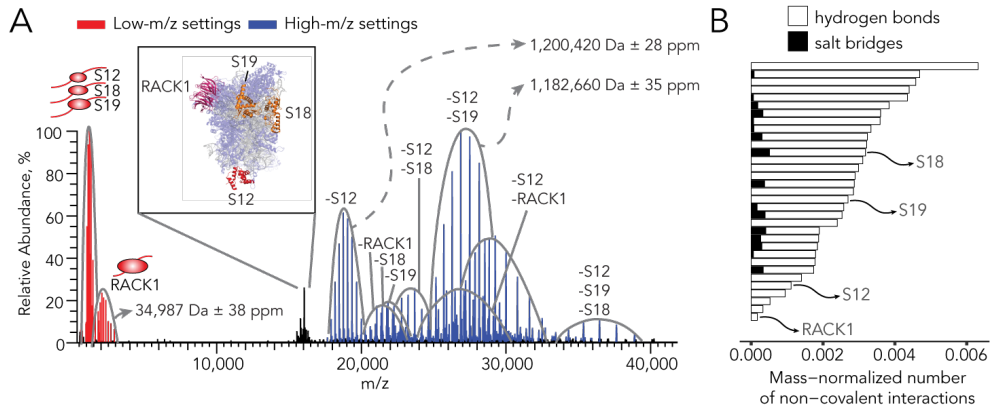


Figure 2 | (A) Collisional dissociation of the most abundant charge state (z = 74+) of the human 40S ribosomal particle recorded with distinct parameters for detection of either high-*m/z* (blue) or low-*m/z* (red) dissociation products. The inset displays the crystal structure of the 40S ribosome (PDB ID: 5A2Q) whereby the ejected subunits are highlighted in red (S12), pink (RACK1), and orange (S12 and S18). The high-*m/z* fragments are annotated by the names of the ejected ribosomal subunits. (B) Distribution of non-covalent interactions observed between ribosomal subunits and RNA molecule in the reported crystal structure (PDB ID: 5A2Q) normalized to the subunit mass. The ejected subunits are indicated.

Overall, the high-resolution fragmentation spectra obtained for HS40S enabled the detection of accurate masses with minimal errors for most of the observed dissociation products, which ultimately enabled the identification of the ejected subunits. Further analysis indicated that the two primary ejected ribosomal proteins, S12 and RACK1, were among the most exposed subunits (Figure2A; inset) with the least number of non-covalent bonds in the reported crystal structure (Figure 2B; PDB ID: 5A2Q). The subsequent ejection of the other two subunits, S19 and S18, can be explained by structural rearrangements taking place after the ejection of the most exposed subunits. Interpreting spectra of such complexity is challenging because the gas-phase dissociation pathways of multimeric macromolecular assemblies produce a highly heterogeneous mixture of ions. Nevertheless, this example demonstrates how technological advances can provide a platform for utilizing conventional collisional fragmentation to uncover structural features of large macromolecular assemblies in the mega-Dalton mass range. To utilize the technique to its fullest potential, mass spectrometers must be further optimized to enable the simultaneous detection of all the products of collisional dissociation.

One of the prospective ways to simultaneously capture multiple fragmentation products is by using novel fragmentation approaches that provide fast energy deposition and site-specific backbone cleavages, alongside conventional dissociation pathways. Recently, we have probed the potency of UVPD as a fragmentation method for native tandem MS of protein systems that extend the boundaries in the mass

range as well as the complexity of current analytes. In this work, we analyzed several macromolecular assemblies, including a virus-like particle with a Mw of ~1 MDa, a multimeric ribonucleoprotein complex CRISPR-Cas Csy, and a light-harvesting subcomplex B-PE decorated by a multitude of chromophores. The data indicated that native top-down UVPD-MS is capable of producing highly informative fragmentation spectra exposing the stoichiometry, the composition, and providing sufficient sequence information to identify distinct proteoforms within the analyzed protein assemblies with a Mw of up to 1 MDa (Figure 3A). Unlike collisional dissociation, UVPD of large protein complexes has an advantage of faster energy deposition through absorption of highly-energetic UV photons by the protein backbone yielding covalent peptidic fragments in parallel with the processes of intra-vibrational energy redistribution. As a result, simultaneous detection of peptidic fragments (Figure 3B), ejected subunits (Figure 3C), and residual complexes (Figure 3E) is achieved.

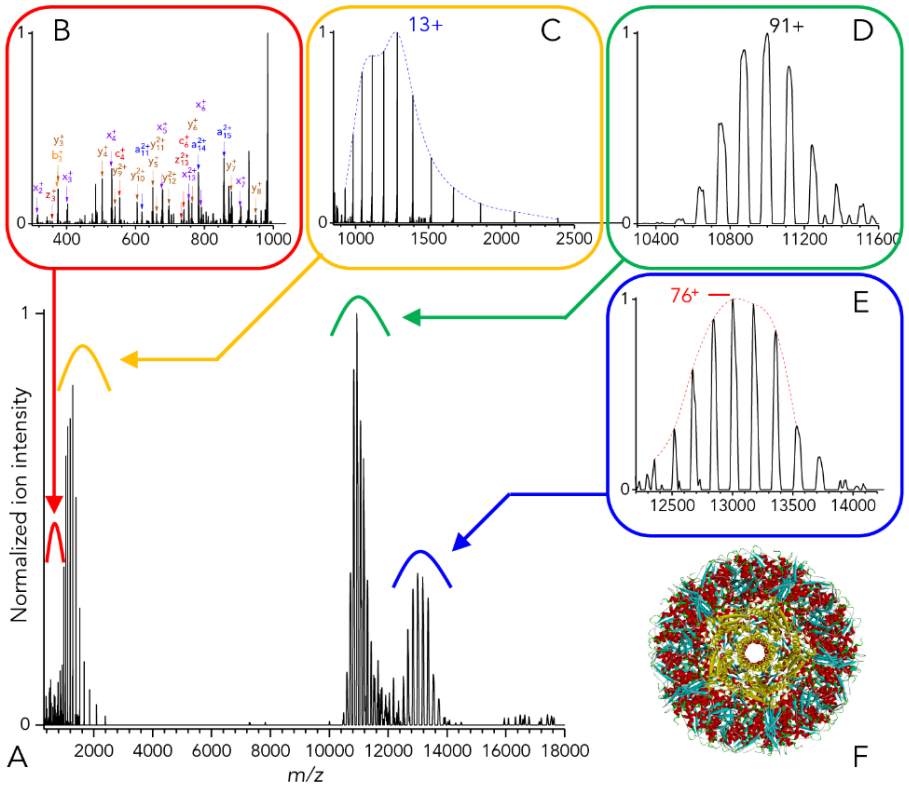


Figure 3 | (A) Native top-down mass spectrometry of a virus-like particle (Mw ~1 MDa) using UVPD. (B) Low mass backbone fragments and (C) intact subunits resulting from the dissociation of the (D) precursor ions, together with the formation of (E) residual complexes of the viral assembly. (F) Structure model of the analyzed viral particle (PDB ID: 5MPP). The figure is reproduced from Greisch, J.-F., Tamara, S., Scheltema, R.A., Maxwell, H.W.R., Fagerlund, R.D., Fineran, P.C., Tetter, S., Hilvert, D., and Heck, A.J.R. (2019). Expanding the mass range for UVPD-based native top-down mass spectrometry. Chem. Sci. 10 (30) 7163-7171.

Further improvements in instrumentation, software, optimization of dissociation approaches, and deepening our understanding of the gas-phase behavior of large macromolecular assemblies have the potential to cement native tandem MS as a powerful tool for screening unknown protein assemblies, interrogating structures of known and unknown protein complexes, and analyzing distinctive assembly variants through the mass selection and dissociation. All these prospects will expand the scope of native MS when applied as an individual tool or as a supplemental approach to high-resolution structural techniques.

Altogether, the versatility provided by distinct modes of tandem mass spectrometry is steadily growing, and the scope of prospective applications is continuously expanding[10]. Ultimately, tandem MS is on its way to becoming an indispensable ingredient in any structural study of heterogeneous macromolecular assemblies. There currently is no other method that can outperform MS in speed, cost, and specificity towards minor structural alterations, which are reflected in the exact mass of the molecule and its characteristic fragmentation patterns.

REFERENCES

1. Greer, S.M., and Brodbelt, J.S. (2018). Top-Down Characterization of Heavily Modified Histones Using 193 nm Ultraviolet Photodissociation Mass Spectrometry. J. Proteome Res. *17*, 1138–1145.
2. Fornelli, L., Srzentic, K., Huguet, R., Mullen, C., Sharma, S., *et al.* (2018). Accurate Sequence Analysis of a Monoclonal Antibody by Top-Down and Middle-Down Orbitrap Mass Spectrometry Applying Multiple Ion Activation Techniques. Anal. Chem. *90*, 8421–8429.
3. Wang, Z., Liu, X., Muther, J., James, J.A., Smith, K., *et al.* (2019). Top-down Mass Spectrometry Analysis of Human Serum Autoantibody Antigen-Binding Fragments. Sci. Rep. *9*, 1–9.
4. Fornelli, L., Durbin, K.R., Fellers, R.T., Early, B.P., Greer, J.B., *et al.* (2017). Advancing Top-down Analysis of the Human Proteome Using a Benchtop Quadrupole-Orbitrap Mass Spectrometer. J. Proteome Res. *16*, 609–618.
5. Li, H., Nguyen, H.H., Loo, R.R.O., Campuzano, I.D.G., and Loo, J.A. (2018). An integrated native mass spectrometry and top-down proteomics method that connects sequence to structure and function of macromolecular complexes. Nat. Chem. *10*, 139–148.
6. Skinner, O.S., Haverland, N.A., Fornelli, L., Melani, R.D., Vale, L.H.F. Do, *et al.* (2018). Top-down characterization of endogenous protein complexes with native proteomics. Nat. Chem. Biol. *14*, 36–41.
7. Brown, K.A., Chen, B., Guardado-Alvarez, T.M., Lin, Z., Hwang, L., *et al.* (2019). A photocleavable surfactant for top-down proteomics. Nat. Methods *16*, 417–420.
8. Greisch, J.-F., Tamara, S., Scheltema, R.A., Maxwell, H.W.R., Fagerlund, R.D., *et al.* (2019). Expanding the mass range for UVPD-based native top-down mass spectrometry. Chem. Sci. *10*, 7163-7171.
9. Van De Waterbeemd, M., Fort, K.L., Boll, D., Reinhardt-Szyba, M., Routh, A., *et al.* (2017). High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. Nat. Methods *14*, 283–286.
10. Allison, T.M., and Bechara, C. (2019). Structural mass spectrometry comes of age: new insight into protein structure, function and interactions. Biochem. Soc. Trans. *47*, 317-327.

A

CURRICULUM VITAE,
LIST OF PUBLICATIONS,
ACKNOWLEDGEMENTS

CURRICULUM VITAE

I was born on the 3rd of November 1992 in the city called Kholmsk on the Russian far-east island Sakhalin. In 2008 I moved out of my parents' house to study in the Specialized Educational Scientific Center for Physics, Mathematics, Chemistry and Biology of Novosibirsk State University (SESC NSU) in Akademgorodok, Novosibirsk. There, I spent the last two years of high school with a bias in chemistry and biology. After finishing high school in 2010, I started education at Novosibirsk State University where I pursued a specialist degree program in Biology at the Department of Molecular Biology. During the training, I conducted a research project in the Laboratory of Developmental Genetics at the Institute of Cytology and Genetics. The primary research for the specialist thesis was performed in the Group of Proteomics and Metabolomics at the International Tomography Center under the supervision of Lyudmila V. Yanshole. The goal of the project was to apply mass spectrometry and metabolomics to the analysis of factors that influence cataract formation in the human lens. Following graduation in 2015, I moved to Utrecht, the Netherlands, and started a Ph.D. in the Biomolecular Mass Spectrometry and Proteomics group at Utrecht University under the supervision of Albert J. R. Heck, Richard A. Scheltema and Alexander A. Makarov. During this time, I worked on the characterization of protein assemblies with native top-down proteomics and integrative mass spectrometry approaches. The results of the research performed during my PhD are presented in this thesis.

LIST OF PUBLICATIONS

1. **Tamara, S.***, Dyachenko, A.*, Fort, K.L., Makarov, A.A., Scheltema, R.A., and Heck, A.J.R. (2016). Symmetry of Charge Partitioning in Collisional and UV Photon-Induced Dissociation of Protein Assemblies. *J. Am. Chem. Soc.* 138 (34) 10860–10868.

2. **Tamara, S.**, Scheltema, R.A., Heck, A.J.R., and Leney, A.C. (2017). Phosphate Transfer in Activated Protein Complexes Reveals Interaction Sites. *Angew. Chemie Int. Ed.* 129 (44) 13829-13832.

3. Dyachenko, A., **Tamara, S.**, and Heck, A.J.R. (2019). Distinct Stabilities of the Structurally Homologous Heptameric Co-Chaperonins GroES and gp31. *J. Am. Soc. Mass Spectrom.* 30 (1) 7–15.

4. van de Waterbeemd, M.*, **Tamara, S.***, Fort, K.L., Damoc, E., Franc, V., Bieri, P., Itten, M., Makarov, A., Ban, N., and Heck, A.J.R. (2018). Dissecting Ribosomal Particles throughout the Kingdoms of Life Using Advanced Hybrid Mass Spectrometry Methods. *Nat. Commun.* 9 (1) 2493.

5. **Tamara, S.**, Hoek, M., Scheltema, R.A., Leney, A.C., and Heck, A.J.R. (2019). A Colorful Pallet of B-Phycoerythrin Proteoforms Exposed by a Multimodal Mass Spectrometry Approach. *Chem* 5 (5) 1302–1317.

6. Greisch, J.-F., **Tamara, S.**, Scheltema, R. A., Maxwell, H., Fagerlund, R., Fineran, P., Tetter, S., Hilvert, D., Heck, A. J. R. (2019). Expanding the Mass Range for UVPD-based Native Top-down Mass Spectrometry. *Chem. Sci.* 10 (30) 7163-7171.

7. Albanese, P.*, **Tamara, S.***, Saracco, G., Scheltema, R. A., Pagliano, C. (2019) How Paired PSII-LHCII Supercomplexes Mediate the Stacking of Plant Thylakoid Membranes Unveiled by Integrative Structural Mass Spectrometry. *Submitted.*

* contributed equally

ACKNOWLEDGEMENTS

Dear reader,

Writing a doctoral thesis could not possibly be accomplished without external help. My case is not an exception. These were memorable four years, and here I will try my best to acknowledge the people who supported me along the way, although inevitably I might miss someone…

Albert and Richard, I would like to thank you first, for inviting me to attempt a PhD in such a big and versatile laboratory. I am glad you dared to introduce me, a naive Siberian student, to the world of international science. It was a great pleasure to do a PhD and work on the very topical projects with state-of-the-art equipment under your supervision. I want to thank you for being demanding supervisors yet allowing me to be creative and flexible in choosing my projects. Besides that, I have been fortunate to work on method development projects under the supervision of Alexander Makarov. I value all the meetings and numerous trips to Bremen, where we were able to test new MS instruments. These projects resulted in fruitful publications, which made a large part of this thesis.

Next, I want to thank the people whom I worked closely with in the first year as they had a significant impact on my work. I learned a lot from you, and I would definitely have had a more challenging start without your help. Andrey, thank you for teaching me to operate mass spectrometers and to fragment proteins by all means. Moreover, you showed me how essential data analysis skills are. Learning R and scripting was, arguably, one of the most valuable experiences that helped me throughout the entire program. Additionally, you introduced me to the bouldering and aspired for doing sport persistently, which was crucial for keeping me stable even in the most challenging moments. I am truly grateful for all the help, so thank you Andrey, a lot! Kyle, I also want to thank you for providing valuable input and helping throughout all these four years. I definitely benefited from your professional advice.

In addition to the people who closely worked with me on the projects, it would be an oversight not to mention the technicians who provided with all of the necessary knowledge regarding the wide array of techniques employed in the lab. Anja, Arjan, Dominique, Mirjam, and Harm, thank you for teaching the basics and assisting in native, denaturing, and other shades of proteomics. Special acknowledgment to Anja and Arjan for the Native MS training. Your uplifting attitude was often very supportive, and I am glad you were – in case of Arjan, continues to be – friendly and helpful in the lab. Congratulations on becoming a happy family with two beautiful daughters!

Another source of support for me was in the fellow PhD students with whom we started the program. I want to thank you for all the occasional coffees, beers and Spar trips where we shared our ups and downs. Furthermore, it was inspiring to proceed through years seeing you struggle and succeed alongside; the journey would feel lonely without you. Tomislav, thank you for being the most idiosyncratic fellow PhD student. Your passion and affection toward the GlycoUniverse is addictive. I also want to thank Jing, Elmo, Wei and Sander for fruitful lunch meetings and discussions.

It was aspiring to work alongside successful senior PhD students and post-docs throughout this time. Among others, thank you Michiel and Aneika for bringing me along to work on the great projects and sharing your expertise in biochemistry and structural biology. Additionally, I want to thank all the participants of native meetings and journal clubs for sharing your knowledge. All those discussions definitely helped to learn more about the field and to make sense of my own results. Thanks to all PhDs who started in the recent years: Inge, Julia, Max, Maurits, Tobi, and many others, your fresh ideas and enthusiasm were contagious. Good luck with successfully finishing your doctoral theses!

During these four years, it was a great pleasure to have the opportunity of collaborating with many visiting scientists. Either unraveling disulfide bridges with Cristian, studying toxic venom proteins with Jure or shedding light onto the photosystem supercomplexes with Pascal – all were valuable and fun experiences. It really makes a big difference to have people with such varying backgrounds visiting the lab, although it is also sad to see you leave so soon. I am glad you returned, Pascal, and congratulations on becoming a father!

This lab for the first time allowed me to network and explore the world through visiting numerous conferences, symposia, summer schools and other events in Europe and beyond. So many wonderful and, at first, scary experiences, presentations, talks and posters in Rolduc, Dubrovnik, San Diego, Stockholm, and many other places around the globe. It was a pleasure to explore them together with colleagues and friends.

I want to thank all the people who helped me stay motivated and supported me during academic and sport endeavors, whether it was running, climbing, biking or anything else. I am super lucky to work alongside such an active bunch of scientists. Special thanks to Barbara and Johannes for being the best running buddies and good friends! Also, thank you Charlotte, Gadi, Franziska, Julia, Kelly, Matina and many others for being around at certain sportive and social moments. Thanks to all my Russian friends for supporting me on the distance and for all the occasional visits. And thank you, Corine, for being an indispensable part of this lab!

Finally, my PhD definitely wouldn't have been the same without Vojtech. Thank you for being a great officemate and a true friend. All the best to you, Linsey and Isa in your happy family project! Dear Valeriya, thank you for being with me throughout the last two years of my PhD. You, definitely, made this time more delightful and special.

And, most importantly, I want to thank my family: Мама, Папа, Соня, Лена, Даша, Бабушка and others, for supporting me while being over 8000 km away. Everything I am is only possible because of you! I miss you and hope we see each other more often in the future. Please stay safe and healthy.

Sorry, if I haven't mentioned you personally, and thank you for reading my thesis!

Best,

Семён (Sem)