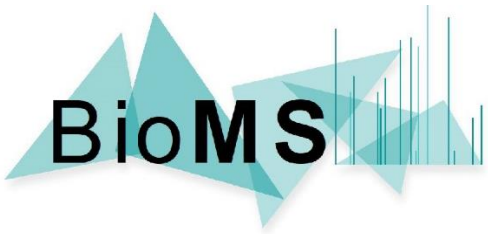
Utrecht University



Accepting the complexity: advance in denaturing top-down proteomics in complex biological systems

Author: Nadzeya Staliarova (7838799)

MSc programme 'Drug Innovation', Faculty of science

Supervisor: Dr. Kelly Stecker

Reviewer: Dr. Karli Robert Reiding

Utrecht, August 2022

Table of Contents

Abstract

Proteins act as central players in molecular events and are subject to variations at the DNA, RNA and PTM levels. These variations result in highly complex and dynamic proteoforms that form the human proteome. Since the traditional bottom-up approach suffers from 'peptide-to-protein inference' problem, proteoforms are studied with the top-down technique which analyses intact proteins. Comprehensive information obtained from proteoform level top-down proteomics addresses clinically relevant research questions. Here we reviewed current technical limitations of top-down proteomics and recent solution to address them with emphasis on top-down application in a complex biological context.

1. Introduction

Proteins are the essential molecules of any living organism and the main driving forces in biology. A huge variety of protein forms, structures, sizes and functions have been revealed. Proteins link the genome to the vast diversity of phenotypes in various states of health and disease. The variety of proteins far exceeds the number of genes. A single gene can give rise not just to one linear polypeptide chain heteropolymer but to many different protein molecules called proteoforms (protein species, protein variants)[1]. This complexity results from genetic variations such as mutations and coding single-nucleotide polymorphisms (cSNPs), alternative splicing RNA transcripts, mistranslation events and post-translational modifications (PTMs)[2] (Fig. 1). PTMs are changes in the chemical structure in the side chains of amino acid residues associated with biochemical processes in the cell. A protein can be post-translationally modified at several distinct residues. For example, P53 has over 100 PTM sites. However, it is still unknown how many of these modifications can occur simultaniously and play a role in PTM cross-talk. More than 400 different types of PTM are known to date[3]. Some types of PTMs have intrinsic complexity themselves. For instance, ubiquitination is the reversible attachment of ubiquitin, a small 76 amino acid globular protein with a mass of 8.5-kDa. Ubiquitin can be attached as monoubiquitin and in chains of various lengths and architectures, with a connection through any of its seven lysine residues creating mixed and branched chains[4,5]. Ubiquitination can form chains of up to 25 ubiquitin molecules, adding up to ~200 kDa and substantially increasing the molecular weight of proteoform[3].

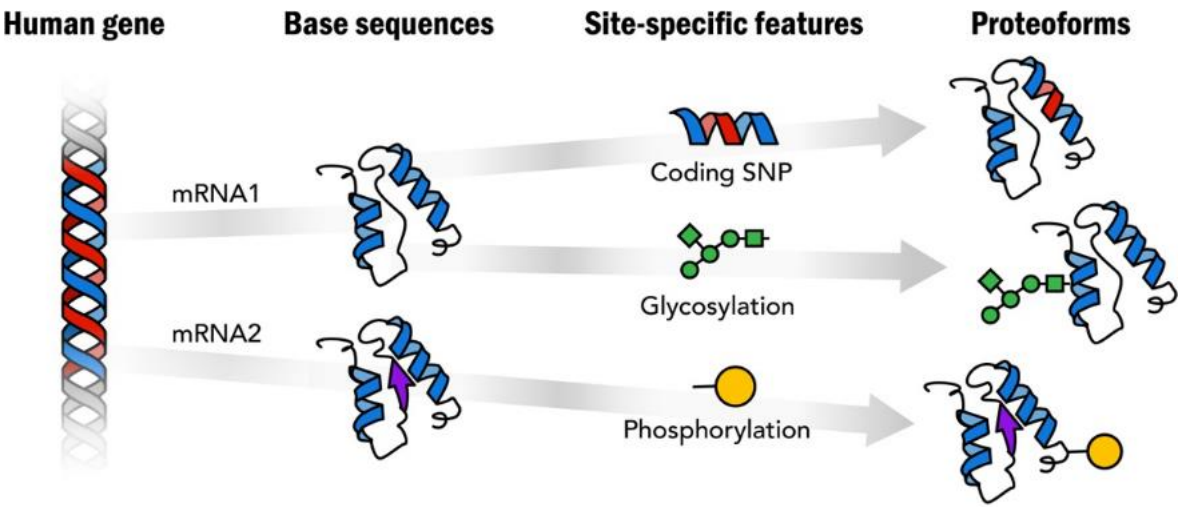


Fig. 1 | *A visual representation of sources in protein variability with the generation of three proteoforms from a single human gene. On the left, one gene and two of its isoforms with a difference of several amino acids in the protein backbone, which can be obtained due to alternative splicing of RNA and from the usage of different locations for translational start or promoters. In addition to changes in the primary amino acids sequence, site-specific features are also encountered, for example, coding single-nucleotide polymorphisms (SNP) and co- or post-translational modifications like N-glycosylation or phosphorylation, respectively. From Fig. 1 in Ref.*[6]

The data based on theoretical prediction suggests that around 6 million proteoforms arise from the 20,300 human genes[7]. Mapping all protein variations and understanding their functions is undoubtedly a challenging goal for fundamental and translational research. However, several research groups have already joined the effort to develop an ambitious "The Human Proteoform Project"[8], which aims to cover the whole human proteome and has two main directions nowadays. The first is focused on the study of proteoforms for medical and clinical applications in areas such as cardiovascular health, infectious disease, cancer, immunology and neurodegenerative disorders (Fig. 2). In many cases, along with identification, it is necessary to obtain absolute or relative quantitative information about the protein content since the concentrations of specific proteoforms may vary depending on the state of the said biological system. The second main objective is developing new as well as improving already existing technologies and their widespread applications that will accelerate the complete proteome coverage, thereby creating a "proteoform atlas"[6].

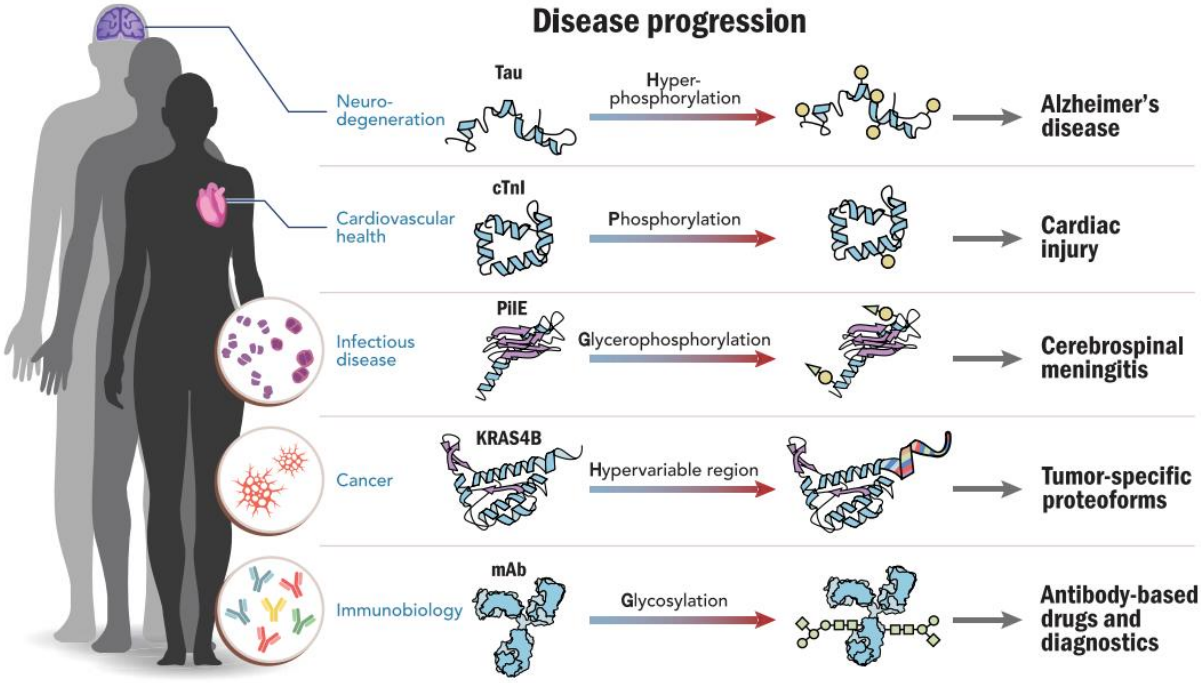


Fig. 2 | Targeted study of proteoforms. New clinical opportunities in five areas of translational research with examples of proteoforms leading to the development of human diseases. mAb, monoclonal antibody. From Fig. 2 in Ref.[6]

The resulting comprehensive knowledge of proteoforms will open new horizons in understanding fundamental biological processes and practical advancements in medicine and pharmacology; creating new therapeutic opportunities such as new drug targets and disease biomarkers.

However, identifying and characterising proteoforms is an analytical challenge for traditional molecular biology methods such as enzyme-linked immunosorbent assay (ELISA) and Western blotting. Nowadays, significant success has been achieved with the mass spectrometry (MS) analysis of proteins. The most general approach for identifying proteins is the so-called bottom-up proteomics, which is based on the LC-MS analysis of short peptides obtained by specific hydrolysis of protein chains[9]. The most commonly used enzyme trypsin is specific for lysine and arginine residues, although other alternative methods can be applied. Since the frequency of lysine and arginine occurrence in mammalian proteins is 5 and 6%, respectively, there are on average 11 trypsin cleavages per 100 peptide bonds. Therefore, it creates peptides with favourable physicochemical properties for the subsequent mass spectrometry analysis since most tryptic peptides have a length of 6-15 amino acids and molecular weight in the range of 1000–2000 Da. This is a suitable range for MS sequencing and mass

spectra recording under electrospray ionisation (ESI) and Matrix-Assisted Laser Desorption-Ionization (MALDI) conditions[10,11].

Another advantage of tryptic peptides is the formation of mainly doubly charged ions during ESI, with the charges located at opposite ends of the peptide. The most favourable protonation sites of peptides are basic amino acids (Arg, Lys, His), as well as the nitrogen atom of the N-terminal amino groups. This ensures locations for two protons, the first is the N-terminus and the second is the C-terminus of tryptic peptides as it is always basic amino acids (Arg or Lys). Therefore, series of fragment ions observed in tandem MS spectra of tryptic peptides, which ensures high coverage of the peptide sequence.

Despite the widespread and high efficiency of bottom-up proteomics, it has a major limitation. Enzymatic cleavage leads to a set of peptides, and identification of the original protein is based on only a few of peptides measured in MS resulting in incomplete coverage of amino acid sequence. Taking into account that proteins are characterised by a wide range of isoforms, with myriad PTMs and possible mutations, this identification method cannot be wholly correct and precise. As some peptides have amino acid sequences shared between isoforms and even between different proteins, there is a problem of logical inference ('peptide-to-protein inference' problem)[12,13]. Thus bottom-up proteomics rarely allows precise identification of proteoform, as minor differences in specific protein regions remain unidentified. Often, isoforms are not included in proteomic databases as individual compounds. Therefore, bottom-up proteomics identifies a specific protein group rather than a biomolecule with a fully determined structure. Usually, numerous isoforms are present on two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) after the protein separation procedure. However, the bottom-up method results in all isoforms identified as one protein.

Along with the typical bottom-up approach, another method is being developed based on the LC-MS analysis of intact whole proteins - top-down proteomics (TDP). In this method, the molecular weight of the whole protein is determined, and then different fragmentation techniques are used to generate fragment-ions. The main advantage of the top-down approach is the ability to elucidate a protein's entire amino acid sequence,

including different PTM and their localisation. Thus, this approach is specialised in studying proteomes at the proteoform level[14].

However, numerous technical limitations hamper its widespread use today[15]. In this review, we highlight the current limitations of TDP and possible solutions to overcome them in the context of proteoform study in complex biological samples. Since proteoforms are formed under the influence of numerous specific factors and conditions, preserving the link between the altered proteoform population and its biological context is crucial. Results obtained from targeted analysis of proteoforms can address most research questions in the biomedical field and have applications in clinical practice.

2. Proteoform analysis with mass spectrometry

Unlike the bottom-up method based on the analysis of peptides after protein digestion, the top-down approach skips the digestion stage and directly analyses the intact protein molecule[14] (Fig. 3). Proteins are thermolabile, non-volatile compounds. Mass spectrometry analysis of these compounds requires soft ionisation methods, which convert protein molecules into the gas phase without breaking covalent bonds. There are currently two main ionisation methods – Matrix-Assisted Laser Desorption-Ionization (MALDI) and ElectroSpray Ionisation (ESI), for the creation of which the Nobel Prize in Chemistry was awarded in 2002[16]. In ESI sources, ions are formed by spraying a solution of substances in the presence of strong electric fields at atmospheric pressure. A modification of electrospray, called "nanoelectrospray", operates with the flow rate of the analysed solution of several nl per minute, which increases the efficiency of the formation of ions of the analyte. This is especially important for protein analysis since it significantly reduces the amount of analyte required to acquire a high-quality mass spectrum. Such a source can be easily docked with a liquid chromatography (LC) system, allowing additional mixture separation during the experiment[9,17,18].

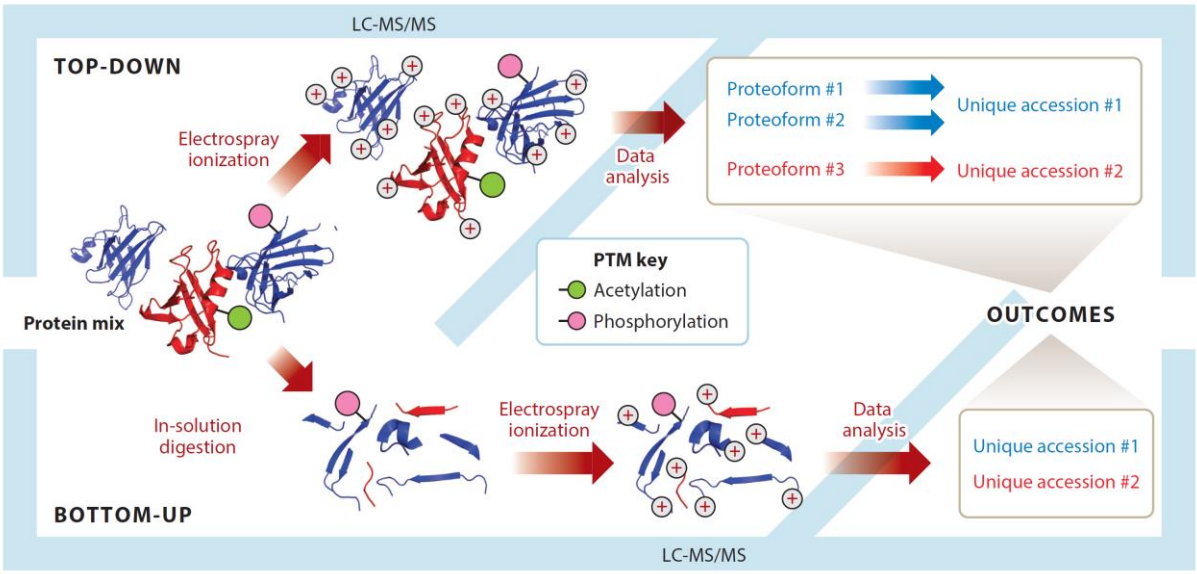


Fig. 3 | Difference between top-down and bottom-up proteomics workflows. In the bottom-up technique, proteins are digested into peptides for subsequent LC-MS/MS analyses. While in the top-down approach digestion phase is skipped, and intact proteins are directly analysed in LC-MS/MS, elucidating the complete protein sequence, including PTM. Abbreviations: LC-MS/MS, liquid chromatography coupled online with tandem mass spectrometry; PTM, post-translational modification. From Fig. 3 in Ref.[14]

As droplets move from the ESI source towards the entrance of the mass spectrometer, they are reduced in size due to evaporation of the solvent and then "explode" with the formation of smaller droplets. This process repeats, resulting in final droplets containing only one charged molecule, which enters the gas phase after evaporation of the residual solvent[19]. Thus, denatured proteins acquire multiple positive charges on the basic amino acid residues due to ESI. These unsolvated positively charged precursor ions are collected in an ion trap with a diverse set of ion optics and then reach the mass analyser. Furthermore, these proteoform cations undergo dissociation, usually collisional or electron-based, forming various fragment ions. Then fragments' mass to charge ratio (m/z) is measured with high accuracy in the mass analyser, and so we acquire tandem MS spectrum (MS/MS, MS2). If the fragmentation step yields the formation of a complete set of different complementary fragments of the proteoform, then 100% sequence coverage of the protein molecule, including PTM, is determined. Therefore, top-down analysis is indispensable when it is necessary to unambiguously establish the complete protein structures to confirm the identification of known proteins and for de novo sequencing[14].

However, despite the crucial importance of studying proteoforms, this method has many limitations that have not allowed it to become a high throughput technique. One major limitation is an escalation of charge state due to ESI with increasing protein size. Therefore, the spectra become too difficult to analyse since many peaks of multiply charged cations are located in a relatively small area range (m/z from 500 to 4000). Other limitations include poor solubility of protein molecules, the low content of most proteoforms in complex samples and low sensitivity due to broad dynamic range. Moreover, the complexity of data analysis due to the variations in proteoform structure and intricacy of fragment ions slows down the development of robust bioinformatics tools[13,15,20]. In the following sections, we focus on the constrains that arise from sample complexity and discuss possible solutions to overcome them.

3. Prospects and challenges

3.1 Sample preparation for top-down proteomics

Since only a limited amount of material is often available for studying the proteome, a reproducible and effective protein extraction method is needed to achieve protein isolation from complex biological samples before any type of proteomics study. One-third of human proteome are hydrophobic membrane proteins that have poor solubility, but at the same time, they play a vital role in biology and are of great interest for research, for example, membrane receptors and transport proteins like ion channels[21]. In addition, large proteins (>70 kDa) also suffer from poor solubility[22]. Low solubility hampers TDP since proteins have to be solubilised before ESI. It brings an additional challenge to TDP compared to bottom-up because, in the latest, not the whole poorly soluble proteins but their peptides have to be solubilised and sprayed for ionisation. Taking into account that resulting peptides have better solubility and more favourable physicochemical properties for ESI compared to the whole hydrophobic protein, TDP need to tackle this challenge. Therefore, several approaches aim to achieve efficient protein solubility. An essential aspect during protein extraction, solubility and all sample preparation steps in proteomics is to preserve all PTM and avoid using reagents that can artificially introduce modifications, as it will be impossible to distinguish whether the modification was endogenous or acquired during sample preparation.

Several methods are commonly used in proteomics for cell lysis and protein extraction. Disruption of membranes by physical processes such as sonication is carried out in buffers containing non-volatile salts incompatible with MS analysis, as they result in ion suppression forming adducts[23]. Among the widely used chemical reagents for cell lysis are ionic detergents (also called surfactants) such as Sodium Dodecyl Sulfate (SDS). Although SDS has a strong denaturing ability and is effective for cell lysis, it is incompatible with MS analysis. SDS causes severe ion suppression even in 0.01% w/v concentration and must be removed before MS analysis[23–25]. Several methods have been developed to deplete incompatible reagents from TDP samples, among them precipitation[25,26], membrane ultrafiltration[17,25] and single-spot solid-phase sample

preparation using paramagnetic beads (SP3)[27]. Membrane ultrafiltration on spin cartridges with a specific molecular weight cutoff allows retaining proteins of specific size while exchanging incompatible buffers before MS analysis[17,25]. Another approach is based on protein precipitation and their subsequent recovery from pellet[25,26]. Finally, in the SP3 method, proteins are subjected to non-selective binding, followed by a washing step and subsequent protein elution[27]. Yang et. al. compared these three techniques for removing SDS and reported that membrane ultrafiltration allows MS compatibility and yields higher unbiased protein recovery with better reproducibility compared to chloroform-methanol precipitation and SP3[25]. However, all these three approaches introduce an additional step in sample preparation that inevitably leads to protein loss, variation and poor reproducibility, and increases the experiment's time. Mild non-ionic surfactants such as n-Dodecyl-β-D-maltopyranoside (DDM) or N-octyl β-D-glucopyranoside (OG) are directly compatible with MS analysis when used in low concentration. However, they are less effective in protein extraction and solubility than ionic detergents[28].

Recent work has demonstrated the use of an efficient and MS-compatible surfactant, 4-Hexylphenylazosulfonate (Azo)[29]. This anionic surfactant decomposes into parts under the action of UV radiation, and the resulting by-products do not interfere with MS analysis. Thus, no additional steps are required to remove it, which ensures reproducibility and prevents protein loss. Moreover, Brown et. al. demonstrated that Azo has comparable performance for protein extraction and solubility to commonly used SDS[29]. These Azo's favourable characteristics have been proven in experiments with protein extraction from cardiac tissues and solubility of integral membrane proteins such as phospholamban, receptor-expressing enhancing protein, succinate dehydrogenase cytochrome b560 with transmembrane domains, and every subunit of the ATP synthase complex[29]. Finally, Azo does not require complex synthesis[29], thus making its wide application in proteomics practice feasible.

Thus, while extraction of proteins and their delivery for ESI presents certain technical limitations, there are ways to overcome them with their pros and cons. Accordingly, the choice of method should be made based on the experiment's objectives and the availability of each technique.

3.2 Prefractionation and separation of complex proteoform mixture

The human proteome is very complex and has a high dynamic range from just a single copy of the least abundant proteoform to several million orders of magnitude for the most abundant, which significantly complicates proteoform profiling by MS due to several technical limitations[3]. Firstly, the coelution of different protein species complicates spectrum interpretation and reduces the signal intensity, identifying only highly abundant and low molecular weight (MW) proteoforms[14]. Secondly, as the protein size increases, the protein charge increases during ESI under denaturing conditions, which exponentially reduces the signal-to-noise ratio due to signal broadening between numerous charge states and isotope forms[30]. These limitations make it difficult to analyse complex mixtures of proteins in top-down MS, leading to the identification of only highly abundant and low MW proteins. To overcome these limitations, sample complexity has to be reduced before top-down MS analysis. Despite the significantly lower number of proteins compared to the number of peptides obtained from them after digestion in the bottom-up proteomics, reverse phase liquid chromatography (RPLC) does not allow to achieve the same effective separation of proteins as for peptides. These difficulties in protein separation are due to the wide variety of their physicochemical properties, such as size, charge, hydrophobicity, etc[23]. This section summarises recent developments in front-end separation and fractionation methods to reduce sample complexity before top-down MS analysis.

In bottom-up proteomics, analysed peptide mixtures are obtained after digestion with for example trypsin and always have arginine or lysine cleavage sites. Therefore resulting tryptic peptides have a relatively narrow mass range, mostly between 0.6 to 3 kDa[12]. While top-down proteomics analyses intact proteins that have a considerable variation in mass in a complex biological samples, ranging from a few kDa to several MDa (titin proteoforms ~4 MDa[31]). As detection of high molecular weight proteins is a challenging goal due to the exponential decrease of S/N ratio with increasing protein mass, thus even slight interference of low MW proteins (5-20 kDa) has a detrimental effect on top-down MS analysis[30]. To separate high MW from low MW proteins, size-based protein separation techniques such as Gel-Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE)[32], Passively Eluting Proteins from Polyacrylamide gels as

Intact species for MS (PEPPI-MS)[33] and Size-Exclusion Chromatography (SEC) are used[34].

Tran and Doucette proposed an offline GELFrEE separation in which proteins are run through a polyacrylamide gel medium with varying concentrations of acrylamide in different columns, achieving high resolution in mass separation[32]. The method is applied to small amounts of protein ranging from low microgram to miligram[32]. Nowadays commercial fractionation stations and cartridges are available for this method, for example GELFREE 8100 fractionation system[12,35]. Takemori recently introduced the PEPPI-MS method for protein separation[33]. PEPPI-MS is carried out on SDS-PAGE equipment that widely used and available in laboratories and thus can be broadly applied in practice[33]. The disadvantages of these methods include the use of detergents, such as SDS, that are incompatible with MS analysis, and therefore require additional steps to remove them. SEC uses a pore column to separate proteoforms of different size. The main limitation of SEC that it has a relatively low resolution. Cai developed a method using sequentially several columns with different pore sizes of the polymer material, thus achieving a good separation of low MW proteins from high MW proteins in complex samples.

In contrast to separation based on protein size, there are several methods that separate proteins based on charge, including Ion Exchange Chromatography (IEC), Capillary Zone Electrophoresis (CZE) and capillary Isoelectric Focusing (cIEF). IEC uses electrostatic interaction to separate proteoforms, and buffer salt solutions for their elution. However, added salts can form adducts and negatively affect the MS spectra. Although IEC poses good orthogonal properties with RPLC[36]. The development of the sheath flow and sheathless interface has made it technically possible to directly connect cIEF and CZE electrophoretic separation methods with ESI-MS[37,38]. CZE separates proteoforms based on their size-to-charge ratios. CZE has better resolution and sensitivity compared to traditional RPLC for separating proteoforms. In a recent study, the CZE method showed a theoretical plate count up to 10E6 for myoglobin[39]. However, CZE only allows a small sample volume, which makes it challenging to analyse low-abundant proteoforms in complex samples. Another drawback of this technique is the fast separation, which results in a narrow separation window[40,41]. This has a

negative effect on the analysis of complex mixtures since there is not enough time for the mass spectrometer to acquire a large number of spectra, resulting in low proteoform coverage. Another technique cIEF allows proteins to be separated based on their isoelectric point in a pH gradient achieved using ampholytes in an electric field. However, the presence of ampholytes causes an ion suppression issue when directly coupled to MS, limiting the application of this method. Xu recently developed an automated cIEF-MS method for proteoforms identification from complex biological samples[42]. However, this method was unabled to characterise proteoforms with highly basic pI>10 and highly acidic pI<3[42].

In addition to already described techniques based on the separation of proteins by their size or charge, the following category of separation methods based on protein hydrophobicity. This group includes Reversed-Phase Liquid Chromatography (RPLC) and other polarity-based methods like Hydrophobic Interaction Chromatography (HIC) and Hydrophilic Interaction Chromatography (HILIC). RPLC is the predominant final dimension of separation online coupled to MS. RPLC is a type of partition chromatography that uses particles linked with nonpolar alkyl chains (C3, C4, C8 and C18) as a stationary phase and a polar solvent as a mobile phase. RPLC columns with C18 material are traditionally used to separate peptide mixture in bottom-up proteomics. However, in the top-down approach, C18 can lead to irreversible binding of hydrophobic (e.g., membrane) proteins, leading to their loss and decrease in their signal intensity. Therefore, resin with short alkyl chains, C3 or C4, is typically used to separate intact proteins in the top-down proteomics[43]. In general, RPLC alone makes it possible to identify only a small number of proteoforms, about 100, and mainly low MW (up to 30 kDa) proteins [44].

All the separation methods described above have their own distinct advantages and limitations, but none of them alone can provide the efficient separation of the complex protein mixture. Therefore, to achieve an effective separation, several orthogonal methods are applied together, complementing each other. Various combinations of separation modes have led to the development of a large number of 2-dimensional (2D) and 3-dimensional (3D) separation approaches. In these multidimensional methods, fractions are collected after offline protein separation in the first dimension, allowing

buffers exchange to remove any MS-incompatible reagents (for example, SDS) and sample concentration. Collected fractions are then separated further in the second dimension and a large number of the resulting fractions are further analysed by LC-MS/MS[13,36]. However, the multidimensional separation technique is low-throughput, labour intensive, time-consuming, introduces variabilities and requires a large amount of initial material, considering possible losses at each step. Nevertheless, despite all the drawbacks of the multidimensional separation approach, it can significantly improve the detection and characterisation of proteoforms in complex mixtures. For example, Cai applied a two-dimensional separation strategy to analyse heart tissue[44]. First, a complex mixture containing proteins of a wide MW range was subjected to serial size exclusion chromatography (sSEC) using successive columns with different porous. Then the resulting factions were further separated in a second dimension RPLC coupled online to MS (Fig. 4).

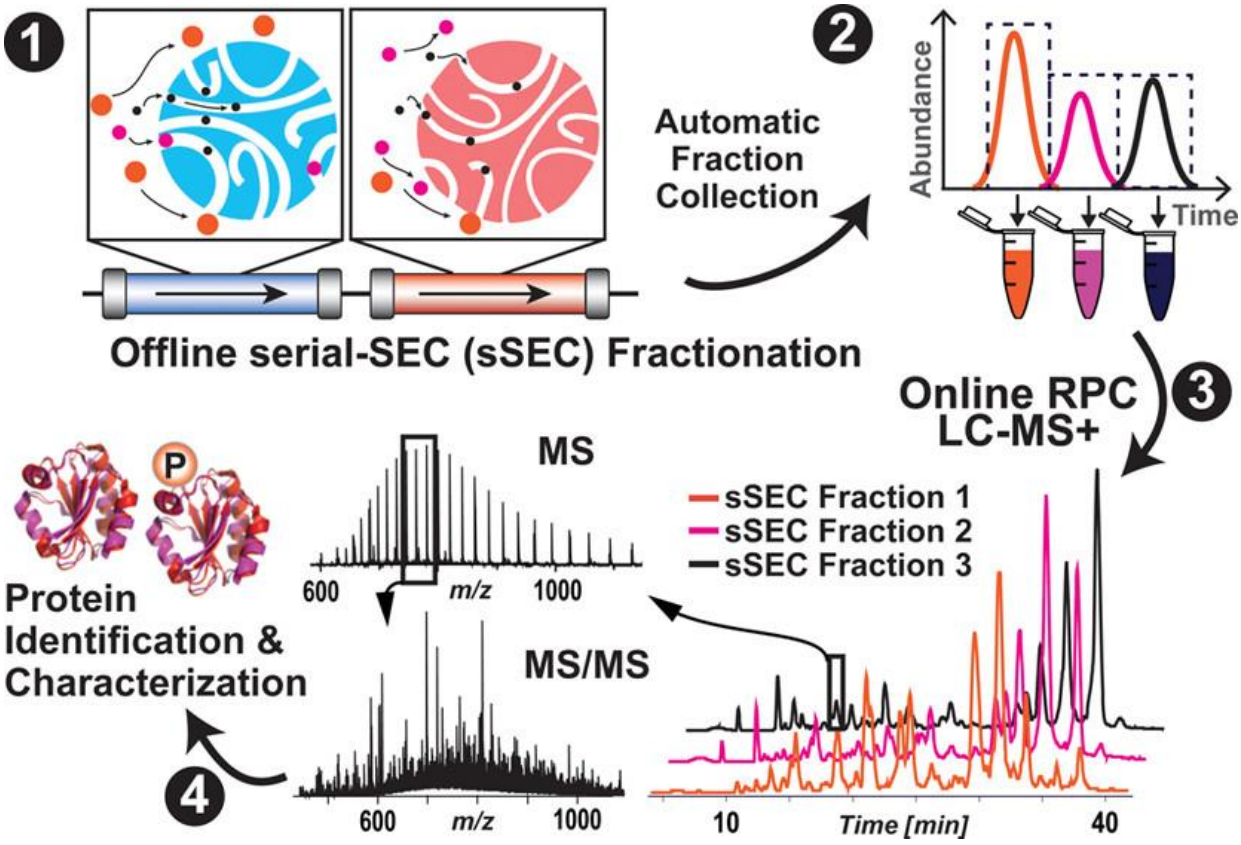


Fig. 4 | 2-dimensional sSEC-RPC workflow for denatured top-down proteomics. Complex samples were offline size-based fractionated in the first dimension with serial size exclusion chromatography on two sSEC columns with different pore size, and then the specific fractions were further separated in the second dimension based on protein hydrophobicity by reverse phase chromatography. Finally, proteins were identified and characterised by high-resolution tandem MS. 2D fractionation increased proteome coverage detecting low-abundant proteoforms with PTM. Figure adapted from abstract in Ref. [44]

This approach allowed the separation of proteins with an extremely wide range of MW from 10 to 223 kDa, with a significantly improved capture of high MW proteins (> 60 kDa). Moreover, this 2D separation method enabled the detection of low-abundant proteoforms with PTM. In quantitative respect, the 2D approach allowed to detect 4044 more proteoforms than using only 1D separation with RPLC[44]. Overall, despite the recent advances and the evolving innovations in separation methods, the high complexity of proteoforms in biological samples still presents an obstacle to developing an effective and universally applicable separation workflow for deep proteoform characterisation in top-down MS. Therefore for TDP, it is necessary to reduce the sample complexity even further by first enriching the target protein and thus adding another layer of separation on top of the standard one.

3.3 Protein enrichment for detection of low-abundant proteoforms

The human proteome is highly dynamic, considering overall protein turnover rate, all possible variations in amino acid sequence, and reversible PTM changes in response to a myriad of stimuli. For example, even a single mistranslation event results in a proteoform that is possibly found only as low as a single copy per cell or even a group of cells. Such low abundant proteoforms are still below the detection limit of modern mass spectrometers. Although mass spectrometers successfully identify and characterise the highly abundant proteoforms with top-down MS, many scientists are interested in specific proteins that are usually scarce in the complex biological matrix but can be potential drug targets or disease biomarkers. For example, many biomarkers present in serum are at much lower levels than dominant protein - human serum albumin[13]. At the same time, the concentration of proteoforms associated with specific disease or dysfunction and have clinical value is much higher in the location of the pathology[45]. However, some locations have limitations in terms of the accessibility and quantity of tissue material available for biopsy. Sample complexity reduction remains a crucial step as interference from other proteins have a detrimental effect on top-down MS analysis. Thus, it is essential to enrich proteins of interest from a complex biological sample before top-down MS to characterise proteoforms for translational research[23].

The challenge of the high dynamic range of the proteome led to development of enrichment strategies for proteomics research. Antibody-based affinity purification

has been widely used in biomedical research to isolate and purify proteins[46]. However, antibody-based techniques suffer from low specificity and batch-to-batch variations of antibodies, ultimately leading to low reproducibility of results. Moreover, antibodies have low stability, and their production is costly[47,48]. Therefore, considerable effort is put into developing alternative reagents for affinity purification of proteins.

To gain insights into cellular biology enriching proteoforms with specific PTM is essential to study its localisation and function. For example, the isolation of glycoproteins is traditionally achieved by lectin affinity purification[49]. Some success has been achieved in the targeted study of intact phosphorylated proteins with enrichment based on the physicochemical properties of the phosphate groups attached to protein[23]. Roberts et. al. developed robust nanoproteomics platform to enrich endogenous phosphoproteins from highly complex proteome[50]. They employed superparamagnetic iron oxide (Fe_3O_4) nanoparticles that were surface silanised with an affinity ligand dinuclear Zn(II)-dipicolylamine (Zn-DPA) complex[51]. This specific and efficient enrichment method in combination with LC-MS/MS allowed the capture and characterisation of low abundant phosphoproteins from complex protein mixture of human cardiac tissue[50].

Recently Tiambeng et. al. employed a nanoproteomics approach using these superparamagnetic nanoparticles coupled with a peptide ligand for affinity enrichment of a cardiac tropin I (cTnI). cTnI is a biomarker of cardiac injury with many its proteoforms identified carrying diverse PTM such as phosphorylation, acetylation, O-GlcNAcylation, citrullination and oxidation[45,52,53]. Using short (12 amino acids long) linear peptide as the most suitable epitope for all proteoforms with different PTM they achieved high affinity enrichment of various proteoforms of cTnI[54]. Characterisation of captured proteoforms revealed that this protein was highly modified, and analysis of changes in PTM could provide new insight into the pathophysiology of cardiovascular diseases. Moreover, this method allowed for cTnI enrichment from the complex protein mixture in serum in concentrations of the trace level, as low as <1 ng/mL[54]. Therefore, such nanoparticles have great potential, as their size allows them to be easily distributed in a complex mixture of proteins, and they have a large area-to-volume ratio, which contributes to effective interaction with the target proteins[13,54]. However, all

these advancements are possible only if nanoparticles are coupled with a reagent that has a high affinity to a target protein.

3.4 Intelligent data acquisition for advanced identification and characterisation of intact proteins

An advantage of top-down proteomics is the ability to characterise the entire intact protein elucidating its sequence with possible truncations and all endogenous PTM[55]. However, for proteoform profiling in complex samples, efficient fragmentation techniques in combination with high resolution and mass accuracy mass spectrometers are needed. Proteins are complex biomolecules consisting of hundreds and thousands of atoms. Therefore, the number of possible ways of fragmentation of their molecular ions is enormous and is determined by the distinct structure of each proteoform. However, certain dissociation patterns are found to be common to all proteins. They include breaks along the protein backbone with the formation of a series of characteristic ions[56,57]. Fig. 5 illustrates the types of ions produced by various activation and fragmentation methods for proteins.

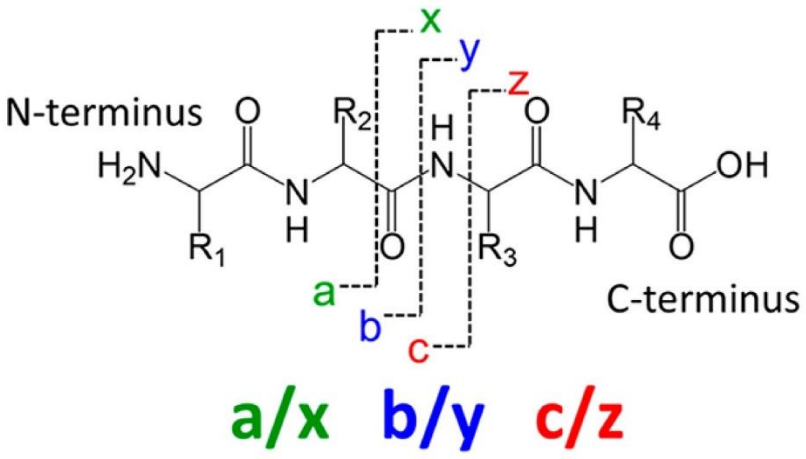


Fig. 5 | Types of ions generated during fragmentation of proteins. From Fig. 1 in Ref. [57]

A large number of activation and fragmentation methods are used for protein analysis by MS. Among the most commonly used are collisional activation techniques, which include collision induced dissociation (CID) and higher-energy collisional dissociation (HCD). The predominant number of tandem mass spectra of proteins has been obtained with these fragmentation methods. Although, these activation techniques have a significant limitation, as they break the most energetically favourable parts (e.g. labile PTM phosphorylation). This often leads to a spectrum with only a few ions, which does

not allow for establishing the entire amino acid subsequence of proteoform. On the other hand, electron-based methods such as electron capture dissociation (ECD) and electron transfer dissociation (ETD) break protein backbone randomly and thus increase sequence coverage. Further development of hybrid methods, including activated ion-ETD (AI-ETD), ETD followed by additional HCD (EThcD), and ETD followed by additional CID (ETciD), allowed reaching better sequence coverage and fragmentation efficiency. Another activation method is ultraviolet photodissociation (UVPD) which is based on the absorption of ultraviolet photons. Although it is less often used, it is very effective[23,57,58]. The complete characterisation of proteoforms can be achieved by using multiple methods of activation and fragmentation simultaneously or in combination[57]. As a result, numerous different fragmentation ions can be formed that reveal information about the amino acid sequence of proteoforms and PTM localisation[59].

Several challenges in fragmentation and data acquisition hamper complete proteoform characterisation. Difficulties in fragmentation arise for high MW proteins and localisation of labile PTMs (e.g. phosphorylation). Dissociation of proteins leads to the formation of hundreds or thousands of fragment ions with a low signal-to-noise ratio[30]. Due to the low intensity of the resulting fragment ions, more starting material is required to characterise intact proteins in the top-down approach. Moreover, high resolution is essential to separate isotopic peaks and obtain monoisotopic masses via mass deconvolution, which is more accurate than the average masses obtained from low-resolution data.

ESI of proteins in denaturing conditions results in multiple charge states, that further impede the detection and characterisation of low abundant proteoforms in a complex mixture. High MW proteins have more charges which lead to many fragmentation channels that result in a highly complex spectrum and reduce the number of precursor ions that can accumulate in the ion trap. The ion trap has a defined ion capacity and can get only a limited number of charges during the scan[30]. This results in a small number of detectable ions and a low signal-to-noise ratio. Therefore, a longer acquisition time is applied to improve the signal-to-noise ratio. However, this extended acquisition time during the traditional Data Dependent Acquisition strategy results in only a few the

most abundant precursor ions being fragmented (usually only the top 2 or 3 most abundant in top-down, which is much less than the top 10 in the bottom-up approach). This often leads to the repeated fragmentation of the same highly abundant precursor ions, while most of the proteoforms in a complex sample remain undissociated, and most of the MS1 data is not fully utilised. Moreover, applying the exclusion list based on the mass-to-charge ratio (m/z) does not prevent the selection of the same proteoform (m). Overall, these obstacles significantly reduce the number of proteoforms identifications in complex samples. Thus, the development of efficient acquisition methods and fragmentation techniques is needed to extend the detection and characterisation of proteoforms in complex samples.

Durbin introduced Autopilot, an acquisition software for top-down proteomics that detects masses of intact proteins in real-time by deconvoluting MS1 spectra. It allows during MS run to determine and select protein species for further fragmentation. Proteoforms that are fully characterised are excluded from further analysis, allowing the characterisation of other species and increasing the number of identifications in a complex protein mixture. Also, this intelligent data acquisition system allows determining the optimal fragmentation method for generating the MS2 spectrum and thus expands the sequence coverage of detected proteoforms. Proteins that are not fully characterised are fragmented again but with a new fragmentation mode with adjusted parameters. This intelligent data acquisition software made it possible to increase the number of identified proteoforms and expanded the sequence coverage, leading to better efficiency of top-down analysis of complex samples. However, the large amount of complex data processed during the relatively short duty cycle of MS instrument represents a significant limitation of this Intelligent Data Acquisition (IDA) approach[60].

The further development of fast and accurate algorithms for deconvolution of mass spectra in real time for guided data acquisition can significantly improve the quality and efficiency of top-down MS analysis. In 2022 Jeong developed FLASHIda, an IDA algorithm that combines ultrafast real-time spectrum deconvolution and a machine learning-based algorithm for precursor selection. Compared to the typical DDA

approach, FLASHIda almost doubled (from 800 to 1500) detection of proteoforms from *E. coli* lysate. It is also compatible with Thermo instruments through iAPI interface[61].

The described IDA algorithms can be further upgraded for the targeted detection of proteoforms of interest in complex samples. However, such tailored data acquisition will require complete information about the proteoform, such as amino acid sequence, possible PTM and existing homologous isoforms.

4. Conclusion and future outlook

The number of canonical human proteins is 20 398, according to the latest information from the Uniprot database (August 2022)[62]. However, the number of proteoforms is not yet known, but it is much larger, considering all alternative splicing, SNP, PTM, etc[3]. Undoubtedly, the proteoforms comprise the human proteome, not a protein's encoded amino acid sequence. Moreover, cellular signalling events are primarily driven by protein's PTM, which arise at proteoforms level. Therefore, although studying the vast number of complex proteoforms is difficult, it carries great potential. Many methods have been used in top-down proteomics to analyse proteoforms of interest and for the general exploration of proteoforms, using which it was possible to detect thousands of unique proteoforms in complex human samples. Future work should maximise the benefits and minimise the impact of limiting factors in these existing methods and approaches. In addition, the development of standard unified criteria for the application of available methods and mutually recognisable criteria for assessing the quality of the obtained results would allow us to combine and summarise the efforts of many research groups working on the complexity of the human proteome. Furthermore, the widespread use of top-down proteomics for proteoform research is limited by the complexity of this approach and requires well-trained experts in MS analysis and expensive equipment, which can also be greatly improved. Although a complete analysis of proteoforms in a sample of interest is still an unattainable goal, this should not stop researchers from working in this direction and looking for possible breakthrough solutions.

Proteoforms are the ultimate critical players in determining molecular events. Therefore, uncovering the biological function of proteoforms is an essential criterion in understanding the pathogenesis of diseases and is applicable in translational research to find optimal biomarkers and drug targets. Such a targeted study of the functions of proteoforms would benefit greatly and be accelerated if proteoforms, with their unique sequence of amino acids and PTM, can be created or introduced into the studied systems at the desired cellular locations and specific concentrations. Moreover, a structured database containing comprehensive information on proteoforms, including phenotypic characteristics, would help uncover the complex human proteome and

establish the role of proteoforms. For example, it would be possible to establish the function of distinct proteoforms in different cell types and disease states, which would play a vital role in developing effective drugs.

Continuous improvements in each of the stages in the top-down workflow make it possible to expand the top-down application to solve clinical problems and proved that top-down proteomics has excellent potential in discovering new disease biomarkers and highly specific drug targets[15]. Examples of successful top-down approaches include the analysis of cardiac tissues and tumour biopsies, where the specific proteoforms signatures of cTnI and KRAS were identified to heart failure and tumour development, respectively[52]. The RAS family of genes play a role in the signalling pathways regulating cell proliferation and growth (PI3K, MAPK). Thus, mutations in the RAS genes are responsible for cancer development in many cases. Ntai applied an immunoaffinity enrichment approach with top-down MS to compare WT and mutated samples and revealed the effect of mutations on the altered proteoform levels. This was followed by discovering the function of distinct PTM, C-terminal carboxymethylation, of KRAS4b in proteoforms' membrane attachment and colorectal cancer progression[63]. Therefore, elucidating the functions of proteoforms opens up new possibilities for the development of tailored therapies and potent drugs with minimal side effects.

Quantifying proteoforms as disease-specific fingerprints is an essential aspect of estimating disease progression. Although to date, accurate and reproducible quantification at the top-down level is challenging. Therefore, researchers can use the advantages of robust and high throughput bottom-up proteomics by building up targeted assay (selected reaction monitoring, parallel reaction monitoring) based on analysis of proteoform's unique and specific peptides[64,65]. In brief, first, comprehensive information about proteoforms of interest will be obtained using top-down proteomics. Then, targeted bottom-up proteomics assay will be employed for sensitive detection, and targeted quantification. Thus, accurate and reproducible results can be obtained by taking the strengths from both proteomics approaches.

Overall, top-down analysis of proteoforms is possible in complex samples, but it requires finding the optimal combination of sample enrichment and separation methods that are most appropriate for the proteoform of interest based on its

physicochemical properties. Nevertheless, numerous examples described in recent works demonstrate feasibility in mapping proteoforms in specific conditions and disease states, show a gradual expansion of the technical boundaries of the top-down approach and serve as a source of inspiration for future work. Once all difficulties and limitations of the top-down approach are addressed and the exact structure of the proteoform is found, it will be possible to detect proteoforms in a wide variety of clinical samples using a targeted bottom-up approach. This can speed up the extensive studies involving many patients since it will allow high throughput experimental design and significantly reduce costs, which is essential in clinical practice. Thus, the complementary application of the potentials and strengths of the top-down and bottom-up approaches is the foundation for successful analysing proteoforms in complex samples.

In the long perspective of top-down proteomics, future work is needed to address the obstacles and technical challenges described in this review, aimed at improving sensitivity and establishing platforms for automatic sample preparation, implementing intelligent data acquisition approaches, and developing tailored and intuitive software for data analysis. Nevertheless, I am confident that this progress in top-down proteomics can be achieved and will lead to its broad application in translational research to analyse complex clinical samples.

5. References

1. Smith, L. M. & Kelleher, N. L. Proteoform: A single term describing protein complexity. *Nat. Methods* **10**, 186–187 (2013).
2. Smith, Lloyd M; Kelleher, N. L. Proteoforms as the next proteomics currency. *Science (80-.).* **359**, 1106–1108 (2018).
3. Aebersold, R. *et al.* How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206–214 (2018).
4. Ge, Z. *et al.* Integrated Genomic Analysis of the Ubiquitin Pathway across Cancer Types Resource Integrated Genomic Analysis of the Ubiquitin Pathway across Cancer Types. 213–226 (2018).
5. Vere, G., Kealy, R., Kessler, B. M. & Pinto-Fernandez, A. Ubiquitomics: An overview and future. *Biomolecules* **10**, 1–22 (2020).
6. Smith, L. *et al.* The Human Proteoform Project: A Plan to Define the Human Proteome. *Preprint* 1–18 (2020).
7. Ponomarenko, E. A. *et al.* The Size of the Human Proteome: The Width and Depth. *Int. J. Anal. Chem.* **2016**, (2016).
8. Smith, L. M. *et al.* The human proteoform project: Defining the human proteome. *Sci. Adv.* **7**, 1–9 (2021).
9. Gao, Y. & Yates, J. R. Protein analysis by shotgun/bottom-up proteomics. *Mass Spectrom. Chem. Proteomics* 1–38 (2019) doi:10.1002/9781118970195.ch1.
10. Chait, B. T. Mass Spectrometry: Bottom-Up or Top-Down? *Science (80-.).* **314**, 66–67 (2006).
11. Schaffer, L. V., Millikin, R. J., Shortreed, M. R., Scalf, M. & Smith, L. M. Improving Proteoform Identifications in Complex Systems through Integration of Bottom-Up and Top-Down Data. *J. Proteome Res.* **19**, 3510–3517 (2020).
12. Toby, T. K. *et al. A comprehensive pipeline for translational top-down proteomics from a single blood draw*. *Nature Protocols* vol. 14 (2019).
13. Melby, J. A. *et al.* Novel Strategies to Address the Challenges in Top-Down Proteomics. *J. Am. Soc. Mass Spectrom.* **32**, 1278–1294 (2021).
14. Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. ofAnalytical Chem.* 499–521 (2016) doi:10.1146/annurev-anchem-071015-041550.
15. Brown, K. A., Melby, J. A., Roberts, D. S. & Ge, Y. Top-down proteomics: challenges, innovations, and applications in basic and clinical research. *Expert Rev. Proteomics* **17**, 719–733 (2020).
16. Walther, T. C. & Mann, M. Mass spectrometry-based proteomics in cell biology. *J. Cell Biol.* **190**, 491–500 (2010).
17. Donnelly, D. P. *et al.* Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nat. Methods* **16**, 587–594 (2019).
18. Cai, W., Tucholski, T. M., Gregorich, Z. R. & Ge, Y. Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. *Expert Rev. Proteomics* **13**, 717–730 (2016).
19. Fenn, J. B., Mann, M., Meng, C. K. A. I., Wong, S. F. & Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science (80-.).* **246**, (1989).
20. Catherman, A. D., Skinner, O. S. & Kelleher, N. L. Top Down proteomics: Facts and perspectives. *Biochem. Biophys. Res. Commun.* **445**, 683–693 (2014).
21. Dobson, L., Reményi, I. & Tusnády, G. E. The human transmembrane proteome. *Biol. Direct* **10**, 1–18 (2015).
22. Gregorich, Z. R. & Ge, Y. Top-down proteomics in health and disease: Challenges and opportunities. *Proteomics* **14**, 1195–1210 (2014).

23. Chen, B., Brown, K. A., Lin, Z. & Ge, Y. Top-Down Proteomics: Ready for Prime Time? *Anal. Chem.* **90**, 110–127 (2018).

24. Kachuk, C. & Doucette, A. A. The benefits (and misfortunes) of SDS in top-down proteomics. *J. Proteomics* **175**, 75–86 (2018).

25. Yang, Z., Shen, X., Chen, D. & Sun, L. Toward a Universal Sample Preparation Method for Denaturing Top-Down Proteomics of Complex Proteomes. *J. Proteome Res.* **19**, 3315–3325 (2020).

26. Doucette, A. A., Vieira, D. B., Orton, D. J. & Wall, M. J. Resolubilization of precipitated intact membrane proteins with cold formic acid for analysis by mass spectrometry. *J. Proteome Res.* **13**, 6001–6012 (2014).

27. Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019).

28. Speers, A. E. & Wu, C. C. Proteomics of integral membrane proteins - Theory and application. *Chem. Rev.* **107**, 3687–3714 (2007).

29. Brown, K. A. *et al.* A photocleavable surfactant for top-down proteomics. *Nat. Methods* **16**, 417–420 (2019).

30. Compton, P. D., Zamdborg, L., Thomas, P. M. & Kelleher, N. L. On the scalability and requirements of whole protein mass spectrometry. *Anal. Chem.* **83**, 6868–6874 (2011).

31. Zhu, C. & Guo, W. MethodsX Detection and quanti fi cation of the giant protein titin by SDS-agarose gel electrophoresis. *MethodsX* **4**, 320–327 (2017).

32. Tran, J. C. & Doucette, A. A. Gel-Eluted Liquid Fraction Entrapment Electrophoresis : An Electrophoretic Method for Broad Molecular Weight Range Proteome Separation. **80**, 1568–1573 (2008).

33. Takemori, A. *et al.* PEPPI-MS: Polyacrylamide-Gel-Based Prefractionation for Analysis of Intact Proteoforms and Protein Complexes by Mass Spectrometry. (2020) doi:10.1021/acs.jproteome.0c00303.

34. Nickerson, J. L. *et al.* Recent advances in top - down proteome sample processing ahead of MS analysis. (2021) doi:10.1002/mas.21706.

35. Witkowski, C. & Harkins, J. Using the GELFREE 8100 Fractionation System for molecular weight-based fractionation with liquid phase recovery. *J. Vis. Exp.* 1–2 (2010) doi:10.3791/1842.

36. Valeja, S. G. *et al.* 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 (2015).

37. Maxwell, E. J., Zhong, X., Zhang, H., Van Zeijl, N. & Chen, D. D. Y. Decoupling CE and ESI for a more robust interface with MS. *Electrophoresis* **31**, 1130–1137 (2010).

38. Moini, M. Simplifying CE-MS operation. 2. Interfacing low-flow separation techniques to mass spectrometry using a porous tip. *Anal. Chem.* **79**, 4241–4246 (2007).

39. Lubeckyj, R. A., Basharat, A. R., Shen, X., Liu, X. & Sun, L. Large-Scale Qualitative and Quantitative Top-Down Proteomics Using Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry with Nanograms of Proteome Samples. *J. Am. Soc. Mass Spectrom.* **30**, 1435–1445 (2019).

40. Dai, J., Lamp, J., Xia, Q. & Zhang, Y. Capillary Isoelectric Focusing-Mass Spectrometry Method for the Separation and Online Characterization of Intact Monoclonal Antibody Charge Variants. *Anal. Chem.* **90**, 2246–2254 (2018).

41. Xiaojing Shen, Zhichang Yang, Elijah N. McCool, Rachele A. Lubeckyj, Daoyang Chen, L. S. Capillary zone electrophoresis-mass spectrometry for top-down proteomics. *Trends Anal. Chem.* **72**, (2019).

42. Xu, T. *et al.* Automated Capillary Isoelectric Focusing-Tandem Mass Spectrometry for Qualitative and Quantitative Top-Down Proteomics. *Anal. Chem.* **92**, 15890–15898 (2020).

43. Shen, Y. *et al.* High-resolution ultrahigh-pressure long column reversed-phase liquid chromatography for top-down proteomics. *J. Chromatogr. A* **1498**, 99–110 (2017).

44. Cai, W. *et al.* Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion Chromatography Strategy. *Anal. Chem.* **89**, 5467–5475 (2017).

45. Zhang, J. *et al.* Top-down quantitative proteomics identified phosphorylation of cardiac troponin i as a candidate biomarker for chronic heart failure. *J. Proteome Res.* **10**, 4054–4065 (2011).

46. Lollo, B., Steele, F. & Gold, L. Beyond antibodies: New affinity reagents to unlock the proteome. *Proteomics* **14**, 638–644 (2014).

47. Bradbury A, Lyon, O. M. F. Standardize antibodies Used in Research. *Nature* **518**, 27–29 (2015).

48. Oliver, K. Blame it on the antibodies. *Phys. World* **22**, 48 (2009).

49. Wu, D., Li, J., Struwe, W. B. & Robinson, C. V. Probing: N -glycoprotein microheterogeneity by lectin affinity purification-mass spectrometry analysis. *Chem. Sci.* **10**, 5146–5155 (2019).

50. Roberts, D. S. *et al.* Reproducible large-scale synthesis of surface silanized nanoparticles as an enabling nanoproteomics platform: Enrichment of the human heart phosphoproteome. *Nano Res.* **12**, 1473–1481 (2019).

51. Hwang, L. *et al.* Specific enrichment of phosphoproteins using functionalized multivalent nanoparticles. *J. Am. Chem. Soc.* **137**, 2432–2435 (2015).

52. Soetkamp, D. *et al.* The continuing evolution of cardiac troponin I biomarker analysis: From protein to proteoform. *Expert Rev. Proteomics* **14**, 973–986 (2017).

53. Dong, X. *et al.* Augmented phosphorylation of cardiac troponin I in hypertensive heart failure. *J. Biol. Chem.* **287**, 848–857 (2012).

54. Tiambeng, T. N. *et al.* Nanoproteomics enables proteoform-resolved analysis of low-abundance proteins in human serum. *Nat. Commun.* **11**, 1–12 (2020).

55. Schaffer, L. V. *et al.* Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **19**, 1–15 (2019).

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

57. Macias, L. A., Santos, I. C. & Brodbelt, J. S. Ion activation methods for peptides and proteins. *Anal. Chem.* **92**, 227–251 (2020).

58. Shaw, J. B. *et al.* Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation. *J. Am. Chem. Soc.* **135**, 12646–12651 (2013).

59. Zenaidee, M. A. *et al.* Internal Fragments Generated from Different Top-Down Mass Spectrometry Fragmentation Methods Extend Protein Sequence Coverage. *J. Am. Soc. Mass Spectrom.* **32**, 1752–1758 (2021).

60. Durbin, K. R., Fellers, R. T., Ntai, I., Kelleher, N. L. & Compton, P. D. Autopilot: An online data acquisition control system for the enhanced high-throughput characterization of intact proteins. *Anal. Chem.* **86**, 1485–1492 (2014).

61. Jeong, K., Kim, J. & Jensen, O. N. FLASHIda enables intelligent data acquisition for top – down proteomics to boost proteoform identi fi cation counts. (2022) doi:10.1038/s41467-022-31922-z.

62. Bateman, A. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

63. Ntai, I. *et al.* Precise characterization of KRAS4b proteoforms in human colorectal cells and tumors reveals mutation / modification cross-talk. *PNAS* **115**, 4140–4145 (2018).

64. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: Workflows, potential, pitfalls and future directions. *Nat. Methods* **9**, 555–566 (2012).

65. Wolf-Yadlin, A., Hautaniemi, S., Lauffenburger, D. A. & White, F. M. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 5860–5865 (2007).