EXPLORING THE FUNCTION OF PLASTOGLOBULES USING TOP-DOWN AND BOTTOM-UP PROTEOMICS BY CAPILLARY ZONE ELECTROPHORESIS – MASS SPECTROMETRY

By

Qianjie Wang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Chemistry – Doctor of Philosophy
Biochemistry and Molecular Biology – Dual Major

2024

ABSTRACT

Plastoglobules (PGs) are lipoprotein particles with dynamic morphology and composition responding to abiotic stress and senescence, and their functions are influenced by post-translational modifications (PTMs) of PG proteome. Capillary zone electrophoresis – tandem mass spectrometry (CZE-MS/MS) based bottom-up proteomics (BUP) and top-down proteomics (TDP) approaches study proteins with their unique PTMs status (i.e., proteoforms) in complex samples.

Chapter 2 presents the first large-scale TDP analysis on *Arabidopsis thaliana* leaf and chloroplast samples, where 3198 and 1836 proteoforms were identified respectively. Notable 1024 and 363 proteoforms exhibited mass shifts from the theoretical mass. Among them, proteoforms with phosphorylation and acetylation were validated by their electrophoretic mobility shifts. Additionally, this analysis also provided direct evidence of N- and C- terminal sequencing that precisely delineated the true transit peptide cleavage sites, offering valuable insights to plant biologists. Despite these advancements, no PG-localized proteins were identified, likely due to challenges related to protein size and hydrophobicity. Thus, the subsequent chapters are dedicated to developing methods to address these issues.

In Chapter 3, we developed the CZE-MS method for TDP analysis of integral membrane proteoforms (IMPs) enriched from mouse brains. We employed a sample buffer containing 30% (v/v) formic acid and 60% (v/v) methanol to solubilize IMPs and a separation buffer composed of 30% (v/v) acetic acid and 30% (v/) methanol to maintain solubility of IMPs during CZE separation. Single-shot CZE-MS/MS identified 51 IMPs. Coupling size-exclusion chromatography (SEC)-CZE-MS enabled the identification of 276 IMPs with 1-4 transmembrane domains. This proof-of-concept work demonstrates the high potential of CZE-MS/MS for the large-scale TDP of IMPs.

In Chapter 4, we tackle the issues related to separation resolution and reproducibility in CZE for TDP, which stem from non-specific protein adsorption in linear polyacrylamide (LPA) coated capillaries. We developed a simple method for applying a cationic, poly(acrylamide-co-(3-acrylamidopropyl) trimethylammonium chloride [PAMAPTAC]) to the capillaries. This PAMAPTAC coating significantly improves the resolution of proteoform separation and achieves consistent measurements across both standard and complex samples, such as yeast cell lysates. The coating enables the detection of large proteoforms (≥30 kDa) without prefractionation and

allows for accurate prediction of proteoform mobility, establishing PAMAPTAC for high-resolution and reproducible TDP analysis.

In Chapter 5, we pioneered the native proteomics measurement of large proteoforms or protein complexes up to 400 kDa from a complex proteome via online coupling of native capillary zone electrophoresis (nCZE) to an ultra-high mass range Orbitrap mass spectrometer (UHMR). The nCZE-MS technique enabled the measurement of a 115-kDa standard protein complex while consuming only about 100 pg of protein material. nCZE-MS analysis of an *E. coli* cell lysate detected 76 proteoforms or protein complexes in a mass range of 30-400 kDa in a single run while consuming only 50-ng protein material. The mass distribution of detected proteoforms or protein complexes agreed well with that from mass photometry measurement. This work represents a technical breakthrough of native proteomics for measuring complex proteomes.

In Chapter 6, we summarized the current challenges in TDP and discussed the advancements made in this dissertation. Additionally, we explore two future directions for advancing the field. The first direction involves cross-laboratory collaborations to enhance reproducibility and broaden the application of CZE-MS-based TDP techniques. The second direction proposes combining BUP and TDP to leverage the advantages of each method, integrating their strengths to yield more comprehensive information about proteoforms and their functions. For example, combining BUP and TDP in a PG shaving experiment aims to deepen the understanding of how proteins are localized and recruited on PGs, providing insights into their functional dynamics under various biological conditions.

ACKNOWLEDGEMENTS

I would like to express my profound appreciation to my advisor, Professor Liangliang Sun for his continuous support and guidance throughout my research journey. His patience in explaining fundamental concepts and his diligence in troubleshooting instruments have been invaluable. I vividly remember his help in refining my presentations and his detailed explanations of collision-associated fragmentation as we walked to the conference room. His optimism and encouraging spirit have been crucial, especially in fostering a trusting and solution-focused environment for collaborative projects. His advice to stay focused and to read more has been a cornerstone of my academic development. I am deeply thankful for his mentorship, his recommendations, and for being an exemplary role model.

My heartfelt thanks also go to my co-advisor Professor Peter Lundquist, whose support and encouragement have significantly broadened my horizons. Working on the Agspectrum project under his guidance was an enriching experience, delving into photosynthesis and maize research. The summertime spent in the cornfield was not only memorable but also tasty. I treasure the afternoons we spent discussing our first publication sentence by sentence – an experience that helped me grow as a researcher. Graduating as the first Ph.D. student from Lundquist lab is an honor I deeply cherish.

I extend my appreciations to my committee members, Professors Jian Hu and Dana Spence for their insightful feedback and discussion both inside and outside of the classroom. Prof. Spence's expertise in statistics and Prof. Hu's knowledge in protein structure and function have been indispensable in shaping my research projects.

I am also grateful to Professors Gary Blanchard and David Arnosti for their support as I pursued the dual degree and help during the past six years, and to Alissa Cohen for her training and suggestion in teaching.

Thanks to my collaborators, Zihao, William and Professor Vicky Wysocki for their dedication to the native CZE-SID project. I also appreciate Prof. Xiaowen Liu for his prompt response and guidance on TopPIC, and Dr. James Xia from CMP Scientific for his collaborative efforts. Thanks are also due to Prof. Chen Chen for providing the mouse samples for our experiment, and to Prof. Julian Whitelegge at UCLA for the enriching discussions.

I am fortunate to have been part of the Sun group, and I am particularly grateful to Xiaojing who introduced me to this fantastic team and taught me so much. Thank you to Eli and

Rachele for their sharing and support. Thank you to Daoyang and Zhichang for their engaging discussions and suggestions on job seeking. A special mention goes to Tian, who has always been a reliable source of help and support in both research and life. I appreciate Qianyi and Dr. Fang for their significant lab contributions and companionship over the years. Working with all the fellows has been a remarkable learning experience. Thanks also to the current members, Jorge, Amir Hossain, Olivia, Mehrdad, Maryam, Bahar, Guangyao and Lance. It has been a joy to work with and learn from each member of the Sun Lab. Reuniting at ASMS and celebrating your achievements has always been a highlight of the year.

In the Lundquist lab, I have gained valuable insights from everyone on the biochemistry side. Thanks to Drs. Hiser and Devadasu for generously sharing the samples, Drs. Shivaiah and Espinoza-Corral for their insightful suggestions, and Dr. Ying for working together in the field. I have also greatly benefited from the knowledge and camaraderie of current members, especially Febri, whose expertise I've valued and company I've enjoyed. Additionally, I have learned a lot from Shannon and Mohit in different fields. I wish all the best to the Lundquist group and cherish our lab gatherings and PRI meetings.

Thanks to my mentor Dr. Blevins for helping me build industrial connections and supporting my research during summer internship at Genentech.

I am immensely grateful for the support from my friends during the past six years, through the ups and downs, and for the many joyful moments we have shared. While it is impossible to list everyone here, I must give special thanks to Jia, Shan, Xuanye, and Xue who have been with me since our college days, and to Jiaqi, Linqing, Yuhan, Xiaotong, Huan, Nian, and Sara for their companionship in East Lansing.

Finally, my deepest thanks to my family – my parents and grandparents – whose unconditional love and support have been the bedrock of my life.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AA	Acetic acid
ABC	Ammonium bicarbonate
ABC1	The absence of bc1 complex
AC	Alternate current
ACN	Acetonitrile
AGC	Automatic gate control
ADC	Antibody-drug conjugate
AOC	Allene oxide cyclase
AOS	Allene oxide synthase
BGE	Background electrolyte
BUP	Bottom-up proteomics
CA	Carbonic anhydrase II
CE	Capillary electrophoresis
CEM	Chain ejection mode
CRM	Charge residue mode
CID	Collision-induced dissociation
CRC	Colorectal cancer
cTP	Chloroplast transit peptide
CZE	Capillary zone electrophoresis
DC	Direct current
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DTT	Dithiothreitol
E. coli	Escherichia coli
ECD	Electron capture dissociation
EOF	Electroosmotic flow
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FA	Formic acid
FAIMS	High-field asymmetric waveform ion mobility spectrometry

FASS	Field-amplified sample stacking
FASP	Filter-aided sample preparation
FDR	False discovery rate
FT	Fourier transform
FTICR	Fourier transform ion cyclotron resonance
GELFrEE	Gel-eluted liquid fraction entrapment electrophoresis
GO	Gene ontology
HCD	Higher energy collisional dissociation
HILIC	Hydrophilic interaction chromatography
HETP	Height equivalent theoretical plate
IAA	Iodoacetamide
ID	Identification
IEM	Ion evaporation mode
IEX	Ion exchange chromatography
IMP	Integral membrane proteins
IRMPD	Infrared multiphoton dissociation
LC	Liquid chromatography
LFQ	Label free quantification
LOD	Limit of detections
LPA	Linear polyacrylamide
LTQ	Linear ion trap
luTP	Lumenal targeting peptide
mAb	Monoclonal antibody
MAPK	Mitogen-activated protein kinase
MCP	Methanol-chloroform precipitation
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
mTP	Mitochondrial transit peptide
Myo	Myoglobin
MWCO	Molecular weight cutoff
PEPPI	passive eluting proteins from polyacrylamide gels as intact species

PES	Phytol ester synthase
pI	Isoelectric point
PSII	Photosystem II complex
PTM	Post-translational modification
PG	Plastoglobule
PQ-9	Plastoquinone-9
RbcL	Rubisco large subunit
RNA	Ribonucleic acid
RPLC	Reversed-phase liquid chromatography
RSD	Relative standard deviation
SA	Streptavidin
SID	Surface induced dissociation
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
SNP	Single nucleotide polymorphism
SCP	Single-cell proteomics
TAG	Triacylglycerols
TCEP	Tris(2-carboxyethyl) phosphine
TDP	Top-down proteomics
tITP	Transient isotachophoresis
TMD	Transmembrane domain
TMT	Tandem Mass Tag
TOF	Time-of-flight
UV	Ultraviolet
UVPD	Ultraviolet photodissociation
UHMR	Ultra high mass range

CHAPTER 1. Introduction

1.1 Mass spectrometry-based proteomics

Proteins are the direct functional participants in cells that link genotype to phenotype by executing and regulating biological functions as encoded. Proteomics studies all proteins in a biological system (e.g., cells, tissue, biofluid). Mass spectrometry (MS) based proteomics is a powerful analytical technique for comprehensively characterizing and quantifying proteins in complex mixtures, which allows scientists to study the modifications, interactions, and ultimately understand the function of the proteins [1]. After proteins are extracted and purified from the biological sample, a series of sample preparations and orthogonal separations are conducted prior to a mass spectrometer to reduce the sample complexity and concentrate low-abundant analytes. Then the mass spectrometer measures the mass-to-charge (m/z) ratio and abundance of peptides or proteins in the gas phase for peptide or proteins identification and quantification [2].

1.1.1 Mass spectrometry

1.1.1.1 Electrospray ionization

Electrospray ionization (ESI) represents a pivotal advancement in MS for transforming nonvolatile analytes from liquid to gas phase [3]. This soft ionization technique for large biomolecular measurement, pioneered by John Fenn and coworkers, employs a high voltage to induce the formation of charged droplets from the Taylor cone at the sharp tip, as illustrated in **Figure 1.1**.

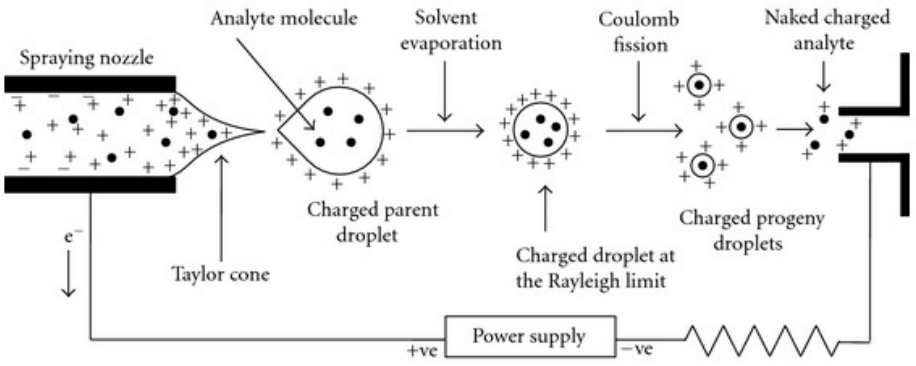


Figure 1.1. Schematic representation of the electrospray ionization process. The figure is reprinted with permission from reference [4].

As the solvent evaporates and jet fission occurs, the droplets shrink in size and their charge density increases. Eventually, the droplets reach their Rayleigh limit, where the surface

tension can no longer sustain the Coulomb force of repulsion, then they explode into smaller highly charged droplets, namely Coulomb fission [4].

Different mechanisms have been proposed for the final stage of ESI [5]: The Ion Evaporation Model (IEM) is an established mechanism for small ions undergoing ESI. When the electric field at the surface is strong enough, it induces the direct emission of solvated ions. Experimental evidence has confirmed IEM's applicability to Na^+ and small proteins such as ubiquitin [6]. For folded proteins and native macromolecules, Charge Residue Model (CRM) is suggested [7]. In this model, the charged droplets undergo Coulomb fission multiple times, eventually drying into gas phase ions by solvent and small ions evaporation [8]. In this process, excess charges are emitted, and the native-like conformations are preserved. In contrast, the Chain Ejection Model (CEM) is proposed for unfolded protein ions with chain-line structures under denaturing conditions [7]. Here, hydrophobic amino acids migrate to the surface to minimize solvent interaction, leading to the sequential ejection of the protein chain from the droplet into the gas phase. Given all proposed mechanisms, ESI is particularly noted for producing multiply charged ions in the gas phase with preserved labile modifications and noncovalent interactions, which is critical for the analysis of proteins and other large biomolecules [9]. The multiply charged ions have an effectively reduced mass-to-charge (m/z) ratio compared to singly charged forms, enhancing the detectability by different mass analyzers.

1.1.1.2 Mass analyzer

Charged gas phase ions are introduced into the mass spectrometer through vacuum and voltage differences, and then further directed to a mass analyzer where they are separated by their m/z ratio. The resolution between two peaks is defined as the ratio of their average m/z ratio ($\bar{\mathrm{m}}$) to the difference in their m/z ratio (Δm)

$$R = \frac{\bar{\mathrm{m}}}{\Delta m} \qquad \text{Equation 1.1}$$

Different types of mass analyzers are utilized depending on the specific requirements of sensitivity, resolution, mass accuracy and scan speed. Low-resolution analyzers, include quadrupole and ion trap. High-resolution analyzers like Fourier transform ion cyclotron resonance (FT-ICR), orbitrap, and time-of-flight (TOF) provide high mass accuracy and precise mass measurement [2,10]

The quadrupole mass analyzer consists of two pairs of parallel metal rods opposing each other and connected to a combination of direct current (DC) and alternate current (AC). Only

certain ions of a particular m/z ratio reach the electron multiplier detector depending on applied DC and AC voltages. By continuously varying the DC and AC voltages, analytes in a certain range of m/z ratio can be measured. Quadrupole is known for its robustness but suffers from limited mass range and mass resolution (10^3) [10]. Quadrupole can also act as a mass filter to select ions and a collision cell to fragment ions in hybrid mass spectrometers. The ion trap mass analyzer is similar to the quadrupole, using an oscillating radio frequency (RF) and DC electric field to trap ions between electrodes and selectively eject ions with different m/z ratios [11]. Despite the low resolving power (10^3), ion traps can accumulate ions over time to improve the sensitivity and perform multiple rounds of mass spectrometry on the trapped ions (MS^n).

TOF analyzers measure the time it takes for ions to travel in the flight tube. Lighter or more highly charged ions reach the multichannel plate detector faster, allowing TOF analyzers to handle a wide range of m/z ratios (20 – 500,000) and providing high scan speed (>1000 Hz) and a fair resolution (10^4) [12]. FTICR stands out with the highest mass resolution (10^6) and mass accuracy (< 1 ppm) [13]. Ions are trapped in a strong magnetic field with electric trapping plates. Ions circulate in a plane perpendicular to the magnetic field by the Lorenz force, and their angular frequency in radians is related to the m/z ratio and the strength of the magnetic field. A broadband RF field is applied using the orthogonal excitation plates. When the cyclotron frequencies of the ions match with the RF frequency, the ions resonate, achieving higher velocities and large cyclotron radii. The resulting image current signal in the time domain is detected and then Fourier transformed to frequency and subsequently into the mass spectra. The high resolving power can resolve the isotopic peaks of large proteins (50 kDa), and the high precision measurement could identify the proteins based on their accurate intact mass. However, the high cost of an FTICR mass spectrometer limits its accessibility for broad applications [10,11].

Similar to FTICR, the orbitrap is another type of analyzer that captures the image current and uses Fourier transform to convert it into frequency [14,15]. Instead of the magnetic field, orbitrap consists of a central spindle electrode and a coaxial outer barrel-like electrode. A static electrostatic field is created between the electrodes to trap ions. Once injected, the ions oscillate harmonically along the axis of the central electrode while spiraling around it. Different ions are separated according to their frequencies. The orbitrap has high resolution (10^5) and high mass accuracy (< 5 ppm). Unlike FTICR where the decrease in resolving power is inversely

proportional to the m/z ratio, the decrease in resolving power of orbitrap is inversely proportional to the square root of the m/z ratio. This property gives the orbitrap advantages when analyzing high molecular weight compounds. Orbitrap-based instruments have been widely used for different proteomic profiling studies [2].

1.1.1.3 Tandem mass spectrometry

Tandem mass spectrometry, also known as MS/MS or MS^2, utilizes two or more mass analyzers consecutively. The first analyzer selects ions with a specific m/z. The selected ions, so-called precursor ions, are then fragmented into product ions which are analyzed by a second mass analyzer. The intact ion mass from MS1 and corresponding fragment ions from MS2 enable the identification of peptides and proteins [11]. Generally, there are two strategies for precursor ion selection in MS-based proteomics: data-dependent acquisition (DDA) and data-independent acquisition (DIA). DDA selects precursor ions based on their abundances in MS1 scans, usually the most abundant N (N being the number of the selected ions). DIA fragments all precursor ions within a m/z window, regardless of their intensity, while the m/z window is systematically scanned across a mass range [16]. DIA is typically searched using spectral libraries, whereas DDA is commonly searched with a sequence database. Both modes are widely used in peptide analysis. DIA has a reported higher reproducibility compared to DDA because it captures the complete fragmentation of all ions and uses the spectral library-based identifications [16]. However, the complex mass spectra generated by DIA require significant computational power and depend on the quality of spectral libraries. For protein analysis, DDA is more frequently used for its direct database search and high-quality MS/MS data.

A variety of gas-phase fragmentation techniques are used in MS/MS to fragment protein backbones [17]. Collision-based fragmentation techniques (i.e., collision-induced dissociation CID and higher energy collision dissociation HCD) convert the kinetic energy into internal energy by colliding the ions with neutral gas molecules. The internal energy accumulates, typically causing cleavage of the peptide bond C-N and generating b-type and y-type ions as shown in **Figure 1.2**. Compared to conventional CID in the ion trap mass spectrometer, HCD in the orbitrap-based instrument accelerates ions in a separate collision cell and usually has a higher energy and shorter activation time [18]. Generally, HCD preferred to generate singly charged y-type ions with lower charge states, smaller fragment ions, enhanced cleavage at the N-terminal of hydrophobic residues, and more internal fragments [18–21]. For electron-based fragmentation

techniques (i.e., electron transfer dissociation ETD and electron capture dissociation ECD), electrons are captured or transferred to multiply charged precursor ions and redistribute, leading to the fragmentation of the N-C_α bond and generating c-type and z-type ions as shown in **Figure 1.2** [17,19]. Photon-based fragmentation includes infrared multiphoton dissociation (IRMPD) and ultraviolet photodissociation (UVPD) [22]. IRMPD uses relatively low-energy photons to excite precursor ions, predominantly generating b/y ions, whereas UVPD generates a/x ions in addition to b/y and c/z ions utilizing high-energy photons (e.g., 193 nm and 213 nm). Surface-induced dissociation (SID) is an activation method where analytes collide against a surface. The rapid energy deposition process can dissociate the subunits from the protein complex without unfolding [23].

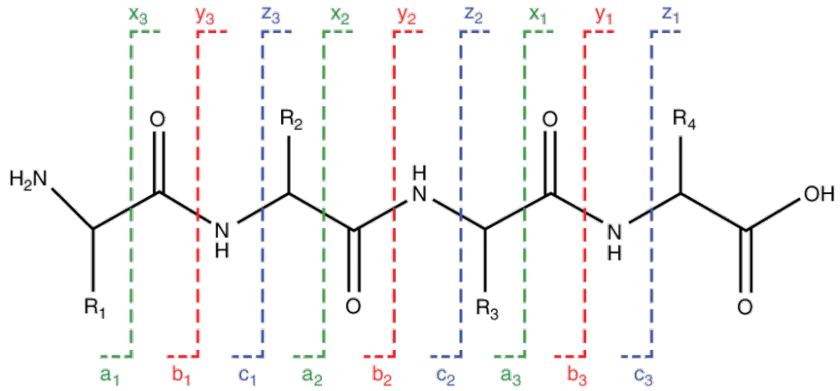


Figure 1.2. The nomenclature for peptide fragmentation. The figure is reprinted with permission from reference [17].

1.1.1.4 Orbitrap-based Hybrid Instruments

Orbitrap-based hybrid instruments have significantly evolved since their commercial introduction in 2005 by Thermo Fisher Scientific [24]. In the last decade, orbitrap-based instruments have become dominant in cutting-edge proteomics for their high resolution, mass accuracy, and sensitivity [22]. The high-filed orbitrap mass analyzer increases the frequencies of ion oscillation, thereby increasing resolving power over a fixed acquisition time [25]. The schematic of a quadrupole high field orbitrap (Q Exactive HF) mass spectrometer is shown in **Figure 1.3**. Ions enter a transmission capillary, and pass through a stacked-ring ion guide (S-Lens), a pre-filler injection flatapole, and an actively guiding bent flatapole, which prevent contamination and improve the ion transmission [26]. After isolation in the quadrupole, ions pass through an octapole into the C-trap, and then transferred to the HCD cell. Fragmentation occurs by adjusting the offset of the RF rods and the axial field to provide collision energy. Following

fragmentation, ions are transferred back into the C-trap and ejected into the orbitrap mass analyzer. Simultaneously, new ions are continuously introduced into the C-trap or HCD cell. Completing a full top10 method (MS1 + 10 MS2, resolution 70,000 and 12,500 at m/z 200, respectively) takes approximately 1 second for the Q Exactive HF, making it compatible with online separation methods, such as liquid chromatograph (LC) or capillary electrophoresis (CE) [26,27].

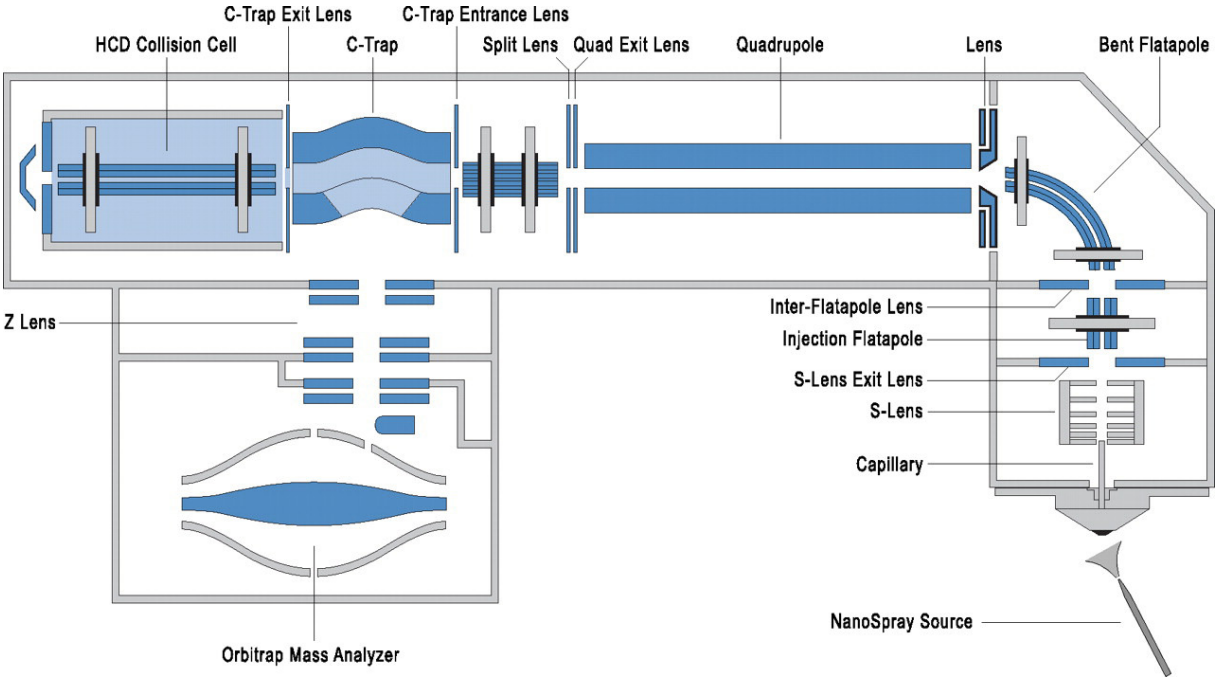


Figure 1.3. Construction details of the Q Exactive orbitrap mass spectrometer. The figure is reprinted with permission from reference [26]. Note that the drawing is not to scale.

Different models of orbitrap-based hybrid instruments have emerged in the proteomics field [22]. For example, the Q-Exactive ultra high mass range (UHMR) has a mass range up to m/z 80,000 for intact protein and native MS. Orbitrap Exploris, introduced in 2020, features a smaller re-designed Q-Orbitrap platform with easier maintenance. The Exploris 480 enables varied scan speeds (up to 40 Hz) and high resolution (480, 000 at m/z 200) for high-throughput proteomic analysis [28]. Flexible fragmentation modalities (ETD, HCD, CID, UVPD) are equipped in the Tribrid mass spectrometers (Fusion, Ascend, etc.).

1.1.2 Bottom-up Proteomics

Proteoforms are all the different molecular forms of protein products generated from a single gene [29]. For the approximately 20,000 protein-coding genes in the human genome, single-nucleotide polymorphisms (SNPs), alternative splicing, and post-translational

modifications (PTMs) contribute to an extremely complicated proteome, resulting in over 1 million proteoforms as illustrated in **Figure 1.4 A** [30]. As **Figure 1.4 B** shows, these variations significantly increase the complexity of the proteome. Bottom-up proteomics (BUP) is a conventional MS-based proteomics strategy. A typical workflow for BUP includes protein extraction, denaturation, reduction and alkylation, and enzymatic digestion followed by offline/online separation coupled with ESI-MS/MS [31,32]. The database search is processed by matching the tandem mass spectra of peptides with the theoretical mass spectra, followed by protein inference and grouping based on their peptides.

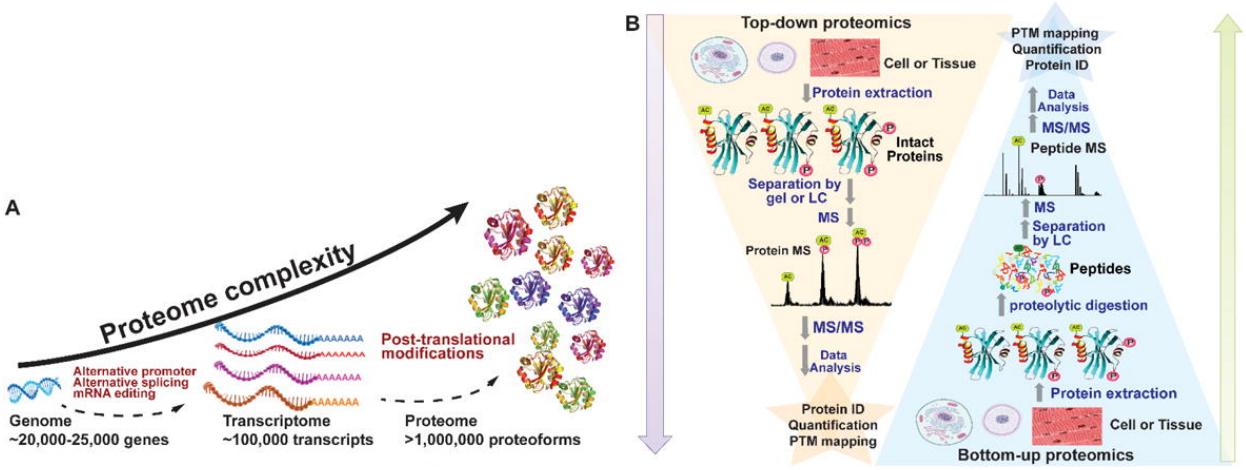


Figure 1.4. Proteomic complexity and strategies. (A). The exponential increase of proteome complexity from the genome. (B). Schematic depicting the workflow for top-down proteomics as compared with that of bottom-up proteomics. The figure is reprinted from the reference [31].

BUP has been widely used for global proteome analysis because of favorable peptide ionization, efficient peptide separation, sufficient fragmentation, and advanced bioinformatic tools [33]. Trypsin, a serine protease, specifically cleaves proteins at the carboxyl side of lysine and arginine, resulting in peptides with an average size of 600 ~ 1000 Da. Peptides thus contain at least two charges (N-terminus of the peptide and C-terminus of lysine or arginine), which fall within the ideal m/z range for MS analysis. The few charge states of tryptic peptides help keep the m/z signal undiluted, achieving high sensitivity and ~1 zmol detection limits [34]. Additionally, sufficient fragmentation and predictable fragmentation patterns of tryptic peptides aid in the localization of PTMs. Reversed-phase LC (RPLC) and CE are commonly used for high-resolution separation, while multi-dimensional approaches enhance peak capacity and reduce peptide coelution.

Quantitative analysis in BUP is crucial for understanding protein changes and addressing biological questions. Isotope labeling (e.g., tandem mass tags TMT) and label-free quantification (LFQ) are two widely used quantification methods [35]. TMT involves tagging peptides with an isobaric label consisting of a reporter ion, a mass balancer, and a reactive group. While the reporter differs in mass, the mass balancer ensures that the intact isobaric labels have the same mass. Upon peptide isolation and fragmentation, the isobaric mass tags break down to release reporter ions of different masses. The intensities of these reporter ions are then measured, allowing accurate quantification of multiple samples (up to 16) in a single experiment. LFQ, on the other hand, quantifies proteins based on the signal intensity of peptides in chromatograms. The area of extracted chromatographic peak represents protein abundance. This approach avoids additional sample preparation but demands high reproducibility and advanced bioinformatics tools.

The Human Proteome Project, which aims to picture the human proteome, has credibly detected 19,778 predicted proteins encoded in the human genome, mostly relying on BUP [36]. Combining with six proteases, heavy fractionation (24 - 80 fractions), and three fragmentation methods, a latest study identified a million unique peptides from 17,174 protein groups, achieving a median sequence coverage of 80% [37]. The comprehensive study allows for a global detection of human proteome. Despite the robust identification and quantification of peptides, BUP often suffers from incomplete protein characterization. The presence of an entire protein is inferred using a small number of peptides and sequence coverage can be limited, especially for membrane proteins. Protein inference introduces ambiguity because some peptides are shared among a protein cluster [33]. There is limited information about proteoform identification, especially regarding the combination of PTMs.

1.1.3 Top-down Proteomics

The chemical diversity of proteins, proteoform, is foundational for biological complexes and cellular function. The Human Proteoform Project aims to create a human map proteoform atlases and discover the proteoform biomarker related to disease [30,38]. The introduction of ESI expanded MS analysis toward larger biomolecules, enabling the direct sequence of intact proteoforms. As **Figure 1.4 B** shows, top-down proteomics (TDP) deciphers the combination of PTMs and amino acid variations in individual proteoforms [29,31]. Unlike BUP, the workflow of TDP retains proteoforms in their intact forms for separation and analysis, allowing for the

identification based on accurate masses of precursor and fragment ions through database searching. A subdiscipline of TDP known as native TDP (native proteomics or nTDP) is performed under near-physiological conditions without protein denaturation, preserving native protein structures and interactions [39–41].

Most quantitative TDP studies employ a label-free approach for relative proteoform quantification by comparing proteoform intensity across different biological conditions [42,43]. The label-free approach is simple and applicable to various biological samples. Using LFQ, differentially expressed phosphorylated proteoforms of important cancer-related genes in well-known colorectal cancer (CRC) cell lines SW620 and SW480 are disclosed [44]. Functional characterization of a KRAS4b proteoform with truncation at C185 residue revealed its loss of the ability to interact with the plasma membrane, resulting in decreased activation of downstream mitogen-activated protein kinase (MAPK) signaling [42]. TMT labeling has been developed for high-throughput quantitative TDP of complex samples. With over 90% labeling efficiency, hundreds of proteoforms per LC-MS/MS run were quantified [45]. However, the labeling efficiency variation depends on the structure and size of the proteoforms, affecting proteoform identification and quantification. Fragmentation energy of TMT labels and proteoform backbones differs, necessitating specific bioinformatic tools to process the complicated mass spectra.

Compared to the fingerprint mass spectra generated in BUP, TDP produces highly complex mass spectra due to larger precursor ions. Thus, a deconvolution algorithm is used to represent fragment ions as singly charged or neutral masses, integrating the signal intensity of charge and isotopic variants. Currently, available database search software includes [46]: Prosight – Uses Poisson matching to identify modifications in the UniProtKB XML database; TopPIC – Uses spectral alignment based on mass ladders and nonspecific mass shifts; pTop – uses *de novo* sequencing to short list proteins and variable PTMs; MSPathFinderT – Uses dynamic programming for spectral alignment and sequence graph filtering. These pipelines have features favoring different biological research scenarios and various outputs of proteoform identification.

1.1.4 Challenges in Top-down Proteomics

During the past decades, TDP has made substantial progress toward proteoform landscape measurement. However, several significant challenges persist in this field:

First, current large-scale TDP studies mainly focus on proteoforms smaller than 30 kDa [44]. Complete characterization of large proteoforms is difficult due to coelution in separation, low sensitivity in MS detection, and insufficient gas-phase fragmentation. For complex cell lysates, proteoform coelution in separation causes ionization suppression of large and low abundance proteoforms. The signal-to-noise ratio (S:N) of proteoforms from MS detection decreases exponentially as the proteoform mass increases due to a much broader charge state and isotopic distribution [47]. Proteins constitute multiple amino acids that could carry charge during ionization, resulting in various charge states. The ionization usually produces a Gaussian-like envelope of peaks reflecting the distribution of the different charge states, as shown in **Figure 1.5 B**. The continuous charge distribution is critical for accurately deconvoluting the mass of large proteins and protein complexes from their m/z measurement [48]. Isotopic patterns are commonly used to determine peptide and protein mass [49]. Similarly, isotopic peaks of large proteins are more complicated and require extremely high resolution to distinguish the isotopic peaks. Compared to the few charge states in peptides shown in **Figure 1.5 C**, the S:N ratio decrease greatly for large proteins, as shown in **Figure 1.5 A** [47]. Additionally, limited backbone cleavage coverage of large proteoforms using conventional collision-based fragmentation methods constantly causes problems in precise sequence ID and PTM localization [50]. In TDP, the analysis of large proteins often generates complex internal fragment that contain neither N- nor C-terminus of proteoforms. These internal fragments are particularly challenging to identify because they are not recognized by most software, especially when a combination of fragmentation is employed, leading to a reduced number of identifiable proteins and limited sequence coverage [21,51]. Although a specific bioinformatics tool is developed and produces a 75% sequence coverage of the whole NIST mAb, it requires high mass accuracy and special attention in controlling false positive hits [52,53].

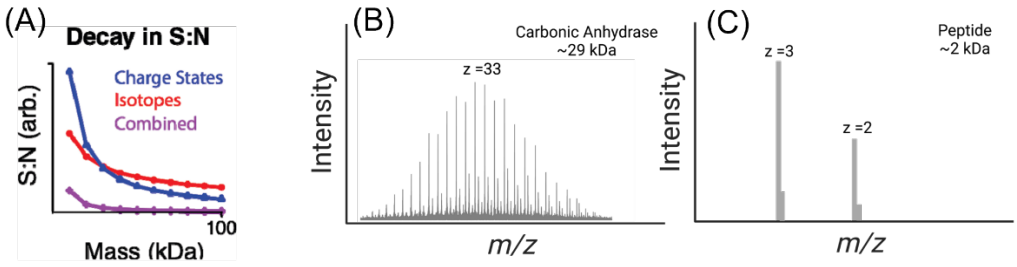


Figure 1.5. The signal-to-noise (S:N) decay for large proteins. (A). The reduction in signal-to-noise ratio for the isotopic distribution and charge states over the protein size.

Figure 1.5. (cont'd)
The figure is reprinted with permission from reference [47]. (B). The mass spectrum of carbonic anhydrase with the charge state distribution. (C). The mass spectrum of an example peptide with the charge state distribution.

Second, TDP of membrane proteins is challenging due to their low solubility in MS-compatible buffers. Despite membrane protein-coding genes being projected to constitute around 30% of the human proteome, membrane proteins, which play critical roles in cellular function and are therapeutic targets, tend to be underrepresented in most global proteomic studies. A typical top-down MS analysis of membrane proteoforms requires detergent micelles, organic solvent, ionic liquid, or membrane mimetics (i.e., nanodiscs, lipid vesicles, bicelles) to keep the solubility prior to MS [54,55].

Third, delineating heavily modified proteoforms (e.g., histones, spike proteins) requires significant effort to improve separations, fragmentations, and bioinformatics tools for accurate PTM determination and localization. Heavily modified proteoforms share similar physical and chemical properties, necessitating high-resolution separation techniques. Various PTM combinations also alter their mass and charge state distribution, complicating the mass spectra interpretation.

Fourth, achieving the identification (ID) of proteoforms of thousands of protein-coding genes in a single study is challenging for MS-intensive TDP due to the extremely high sample complexity. It is estimated that more than one million proteoforms are present in the human body [30,38]. Multiple sample preparation methods are necessary for fractionating the proteome. Additionally, less than one-third of the tandem mass spectra are identified in current TDP studies, stating that further improvement in bioinformatics tools for proteoform ID and quantification to advance TDP in the large-scale studies [46].

1.2 Sample preparation and separation before MS

1.2.1 Sample preparation and separation

Protein extraction is typically performed using different buffer-containing detergents (e.g., Triton X-100, sodium dodecyl sulfate SDS) or chaotropic agents (e.g., urea, guanidine hydrochloride) to disrupt the cell membrane and solubilize proteins. The extraction buffer should regulate both pH and ionic strength, to maximize protein solubility. The addition of protease inhibitors and phosphatase inhibitors prevents protein degradation [56]. Mechanical disruption methods such as homogenization, ultrasonication, and bead beating are commonly applied to

facilitate cell lysis. Detergents and chaotic reagents need to be removed before MS analysis as ion suppression [31]. Buffer exchange is usually performed by membrane ultrafiltration or protein precipitation to remove incompatible reagents and concentrate proteins.

Membrane ultrafiltration is a universal preparation method in BUP as the filter-aided sample preparation (FASP) to remove SDS before enzymatic digestion [57]. Briefly, protein is solubilized in 4% SDS, then retained and concentrated into microliter volumes in a molecular weight cutoff (MWCO) filter. Molecules smaller than the cutoff (3 kDa /10 kDa /30 kDa, etc.) are depleted after several washes. Reduction using dithiothreitol (DTT) or tris(2-carboxyethyl) phosphine (TCEP) and alkylation using iodoacetamide (IAA) is performed on the filter to prevent disulfide bond formation. After trypsin digestion, the peptides are collected from the filter unit and desalted for downstream analysis. In TDP, reduction and alkylation can enhance the identification of cysteine-containing proteoforms by breaking disulfide bonds. However, the elevated temperature often used in reduction can lead to increased identification of truncated proteoforms. Also, the disulfide bond information, which is crucial for understanding protein structure is lost during this process [58]. Therefore, reduction and alkylation are not always necessary in TDP. Membrane ultrafiltration is also a widely used buffer exchange method in TDP with high protein recovery (~86%), no protein bias, good reproducibility (relative standard deviation, RSD < 12%) and great compatibility with follow-up MS analysis [55,59]. Briefly, the SDS-extracted protein sample is washed three times with 8M urea which disrupts the protein-SDS interaction, and then further washed with the desired sample buffer (e.g., ammonium acetate). For native protein complexes, denaturation and reduction are avoided. A Bio Spin column or gel filtration column is used for buffer exchange, ensuring the proteins remain in their native state with proper salt ion strength [60]. Common precipitation protocols use organic solvents, such as methanol-chloroform precipitation (MCP) to agglomerate proteins and remove salts and detergents. Protein precipitation enables protein concentrating in a quick and easy manner but faces the challenge of re-solubilization. Acetone precipitation has the advantage of leaving many proteins folded but can modify proteins with +98 Da adducts [55].

Gel electrophoresis not only separates proteins by their molecular weight but also prepares proteins for downstream MS analysis. Smaller proteins migrate faster through the gel matrix. In-gel digestion is a standard sample preparation method in BUP allowing the extraction of peptides from specific protein bands. As for TDP, GELFrEE (gel-eluted liquid fraction

entrapment electrophoresis) elutes proteins out of the gel into a collection chamber, facilitating the subsequent MS analysis. The Kelleher group utilized a GELFrEE (gel-eluted liquid fraction entrapment electrophoresis)-CZE-MS/MS platform to improve the detection of large proteoforms [61]. Thirty proteins in the mass range of 30-80 kDa from the *Pseudomonas aeruginosa* whole-cell lysate were identified. The Takemori group developed a method called passive eluting proteins from polyacrylamide gels as intact species for MS ("PEPPI-MS") [62]. The PEPPI is a gel-based electrophoresis method for high-resolution separation of proteoforms based on their masses and recovery of proteoforms from polyacrylamide gels. A 68% median recovery rate for < 100 kDa proteins and a 57% median recovery rate for proteins larger than 100 kDa were reported previously. Coupling PEPPI to RPLC-MS and MS/MS identified nearly 60 proteoforms higher than 30 kDa from an *E. coli* cell lysate. Although PEPPI efficiently separates large proteoforms, the PEPPI workflow includes protein cleanup steps (protein precipitation) before MS analysis. More careful evaluations are needed for comprehensive proteoform ID and accurate quantification using the PEPPI-based approach [63].

Different sample preparation methods influence proteoform identification in TDP and need to be carefully selected for samples in different sizes, hydrophobicity, PTMs, and relative abundance [56,58]

1.2.2 Liquid chromatography (LC)

Both BUP and TDP are heavily dependent on the separation technologies (LC and CE) to provide large-scale proteome coverage, higher throughput, and a dynamic concentration range [33]. Current BUP can rapidly profile proteomic mixtures, up to 200 unique protein identifications per minute. Spectral overlap is more pronounced in TDP given the multiple charge states and isotopic distribution from higher mass compounds [56]. Highly efficient separation reduces ionization suppression and affords enough time for MS analysis of complex peptides or proteins. LC is a prevalent separation technique due to its high capacity and efficiency.

In separation science, theoretical plates represent a hypothetical zone in which two phases reach equilibrium. The separation efficiency is often measured by the number of theoretical plates (N), which is dependent on the length of the column L, and the height equivalent theoretical plate (HETP or H) [64].

$$N = \frac{L}{H} \qquad \text{Equation 1.3}$$

A small HETP value indicates a higher number of theoretical plates and higher separation efficiency. HETP is influenced by various factors, as described in van Deemter equation.

$$H = A + \frac{B}{\mu} + C\mu \quad \text{Equation 1.4}$$

Here, *A* is the eddy diffusion parameter relating to the multiple flow paths due to the different packing particles. Using smaller and more uniform particles minimizes the variation in path lengths and results in smaller HETP. *B* is the longitudinal diffusion parameter arising from the dispersion of molecules along the column axis due to concentration difference. *B* is related to the analytes' diffusion coefficient. Longitudinal diffusion accounts for band broadening in the mobile phase and is inversely proportional to the linear velocity, μ. Increasing the velocity of the mobile phase will shorten the time of analyte longitudinal diffusion. *C* is the resistance to mass transfer parameter representing the time for analytes to transfer between the mobile and stationary phases. Using low-viscosity mobile phases, packed columns with very fine particles, and thin films of the stationary phase will minimize HETP. High velocity of the mobile phase leads to incomplete mass transfer between phases and peak tailing. Thus, optimizing the flow rate to minimize the HETP is important. As **Figure 1.6** shows, the optimal flowrate given minimum HETP is $\mu_{opt} = \sqrt{\frac{B}{C}}$.

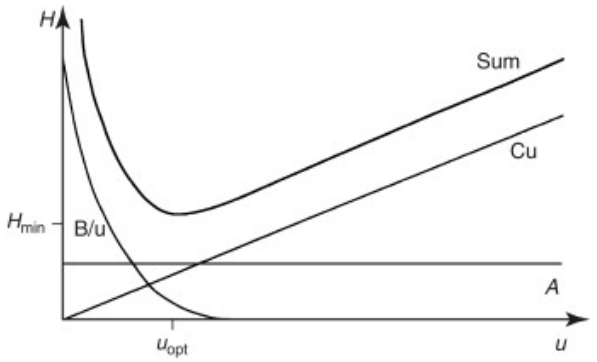


Figure 1.6. The van Deemter curve describes the dependence of the height equivalent theoretical plate H, and the linear velocity. The figure is reprinted with permission from reference [64].

In summary, the separation efficiency of LC can be improved through the application of smaller and more uniform packing particles, flow rate, and column size. Smaller particle size leads to a higher pressure. A narrow inner diameter of the column is compatible with ultra-high pressure and ESI flow rate, reducing dilution, improving heat dissipation, and increasing sensitivity. With sub-2μm particles packed into a 25-cm-long capillary column with 75 μm inner

diameter (i.d.), operating at 10^3 bar, the number of theoretical plates could reach 10^5, exhibiting high separation power [65]. Based on the principle of separation, LC includes several techniques: RPLC, which separates analytes based on hydrophobicity; Size Exclusion Chromatography (SEC), which separates analytes according to size; Ion Exchange Chromatography (IEX) which separates ions based on their net surface charge; hydrophilic interaction chromatography (HILIC) which separates analytes based on their polarity.

RPLC is the most commonly used separation approach in high-throughput BUP and TDP due to its large sample loading capacity, buffer compatibility with MS, and high separation efficiency. Proteins bind with the non-polar stationary phase and are eluted and separated based on their hydrophobicity by gradually increasing the concentration of organic solvents (e.g., acetonitrile) in the mobile phase. Peak capacity is calculated based on the average peak width in the separation time. Longer columns provide better separation efficiency and high peak capacity. Using a 2-meter-long capillary with 50 μm i.d and 3-μm porous particle packing column, a high peak capacity of 1500 is achieved in BUP analysis [66]. To improve sensitivity, the inner diameter of the RPLC column is reduced to the micrometer scale and the flow rate is reduced to nano-liter per minute, so-called nanoRPLC. Around 1000 protein groups from 75 pg peptides are identified using a 2-μm-i.d. column with a picoliter-scale flow rate, representing the sensitivity of BUP by LC-MS [67]. In TDP analysis, short alkyl bonded phases, such as C4 columns, are chosen instead of the C18 used in BUP. Shorter alkyl chains outperformed C18 in terms of recovery and separation efficiency of proteins. This is attributed to increased partitioning into shorter chains as opposed to adsorption/desorption with long chains [68]. Coupling with high-resolution FTICR with nanoRPLC (75-μm-i.d. capillary column packed with PLRP-S or C8 resin), over three thousand unique proteoforms are identified in 40 LC-MS/MS experiments, with 372 proteoforms over 30 kDa and isotopically resolved [69]. Using a MAbPac capillary column (1,500 Å pore size, 4-μm particle size, 150-μm-i.d.) at 2 μL/min, large proteoforms (myosin heavy chain, around 223 kDa) are detected from 100 ng of protein extracted from single muscle fiber [70]. Monolithic columns require a lower backpressure compared to packed columns. An ethane-bridged hybrid monolith column (120 cm × 75 μm i.d.) with well-defined mesopores was applied for TDP analysis of *E. coli* lysates [71]. A high peak capacity of 646 was achieved within a 240 min gradient. However, due to hydrophobic interactions, there was a sample carryover between runs. It requires longer time to balance the column in nanoRPLC

given the high back pressure, which affects the throughput. Membrane proteins with high hydrophobicity could irreversibly adsorb to the column, causing sample loss and affecting column performance. Given the high complexity of proteoforms, more liquid-phase separation with better separation efficiency is important for large-scale TDP analysis.

SEC is an alternative separation based on analyte size, specifically hydrodynamic volume. The pore size ranges from 300 Å to 1000 Å, accommodating the size of proteins. Smaller proteoforms diffuse into the pores more readily, while larger proteoforms are less likely to enter the pores and elute first. Coupling SEC with RPLC/CE-MS/MS enables large-scale proteoform characterization of up to tens of thousands of proteoforms [44,71–73]. The resolution of one dimension SEC is usually limited, and the mobile phase may have a high concentration of non-volatile salt. Thus, SEC serves more as a fractionation method for proteoforms by size to reduce sample complexity. A serial SEC (sSEC) links multiple columns of different porosity to more effectively separate smaller proteoforms from large proteoforms, enabling high-resolution fractionation of proteoforms (10- 223 kDa) [48]. Two-dimensional sSEC-RPLC allowed for the identification of over 4000 unique proteoforms and a 15-fold increase in the detection of proteins above 60 kDa, compared to one-dimensional RPLC. By using neutral hydrophilic mobile and stationary phases, SEC can separate proteins and protein complexes under native conditions and can be utilized for protein buffer exchange [41,74].

1.2.3 Capillary Zone Electrophoresis (CZE)

Electrophoretic methods are promising for large proteoform separations. Capillary zone electrophoresis (CZE) separates analytes under an electric field according to their charge-to-size ratios. As CZE performs separation in an open tubular fused silica capillary containing no stationary phase, there is no eddy diffusion or resistance to mass transfer term in van Deemter equation. The number of theoretical plates (N) in CZE is represented as follows:

$$N = \frac{\mu V}{2D} = \frac{(\mu_{ep} + \mu_{eof})E}{2D} \qquad \text{Equation 1.5}$$

Where μ is the mobility of analytes (velocity under unit electric field), V is the voltage applied across the capillary, and D is the analytes' diffusion coefficient. CZE achieves higher separation efficiency and less sample loss for large proteoforms than many conventional LC techniques. First, large molecules such as proteins have low diffusion coefficients in solution, leading to ultra-high separation efficiency (~10^6 theoretical plates/m) [75]. Second, CZE-MS has shown 10-100 times higher sensitivity than RPLC for measuring peptides and proteoforms, especially in

mass-limited amounts [76,77]. CZE requires only the nanoliter-scale of injection volumes (at nanogram scale amounts) [75]. Third, CZE has high separation efficiency for proteoforms with PTMs such as phosphorylation and acetylation. Three phosphoproteoforms of cardiac troponin I (cTnI) were not resolved by RPLC but were baseline-separated by CZE [78]. Lastly, the mobility of peptides and proteoforms can be accurately predicted, further boosting proteoform and PTM identification [79,80].

The mobility of the analyte is the sum of electrophoretic mobility (μ_{ep}) and electroosmotic mobility (μ_{eof}), as shown in **Equation 1.5**. The electrophoretic mobility (μ_{ep}) is determined by the charge-to-size ratio of the analyte ions, as **Equation 1.6** shows

$$\mu_{ep} = \frac{q}{6\pi\eta r} \quad \text{Equation 1.6}$$

where q is the net charge, r is the analyte radius, and η is the buffer viscosity. As **Figure 1.7** shows, smaller and highly charged ions migrate forward faster while neutrally charged ions do not separate [81]. On the other hand, electroosmotic mobility is the same for all ions. The silanol group on the inner wall of the bare fused silica capillary carries negative charge at pH above 2, forming -SiO^-. Positively charged cations are attracted to form a double layer: one fixed layer binding tightly to the surface and one diffuse layer loosely attached to the fixed layer. Under an electric field, the positively charged diffuse layer migrates towards the cathode carrying the bulk solution, producing the electroosmotic flow.

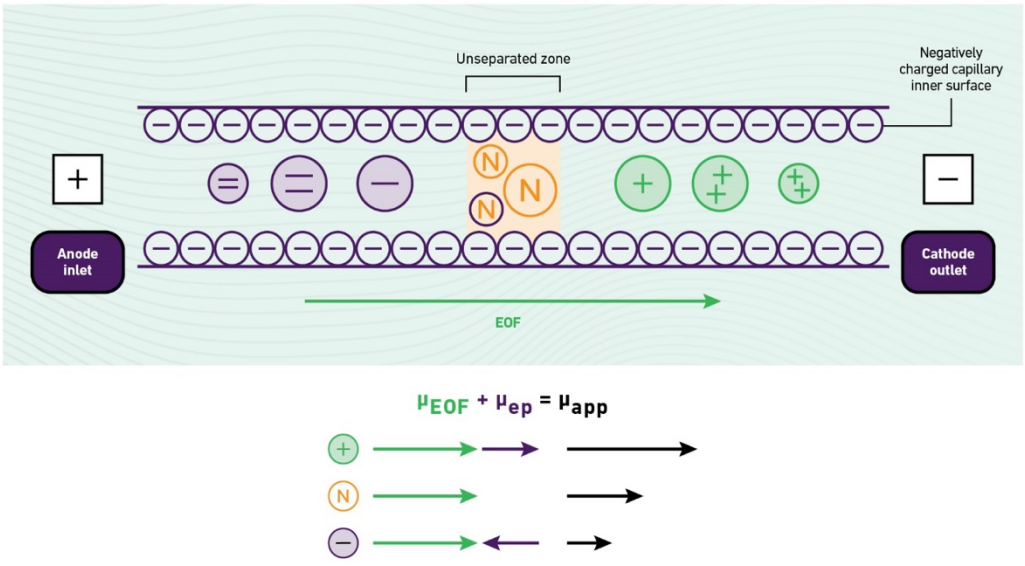


Figure 1.7. Scheme of the ion migration based on electrophoretic and electroosmotic mobility in capillary zone electrophoresis. The figure is reprinted from Technology Networks [81].

The electroosmotic mobility is defined as

$$\mu_{eof} = \frac{\varepsilon\zeta}{4\pi\eta} \quad \text{Equation 1.7}$$

where ε is the buffer dielectric constant, ζ is the zeta potential (potential of the diffuse layer at the slipping plane relative to the bulk), and η is the viscosity of the buffer. Zeta potential is proportional to the charge on the capillary wall and the thickness of the double layer. Thus, pH and ionic strength alternate EOF by influencing the charge density on the wall and the thickness of the double layer.

EOF in bare fused silica drives all analytes forward, leading to a narrow separation window of around 30 minutes. Although this is suitable for high throughput analysis, limited MS and MS/MS spectra are acquired within a peak. Neutral and hydrophilic coatings (e.g., linear polyacrylamide, LPA) are used to eliminate EOF and minimize protein adsorption [82]. A 90-min separation window for proteoforms is achieved in a 1-meter-long LPA-coated capillary [83]. Cationic coatings have also been used to generate a countercurrent EOF and minimize the protein adsorption on the wall for proteomics [61,84,85].

Large-scale CZE-MS-based proteomics is limited by the low sample loading capacity, which is around 1% of the total capillary volume to maintain the high separation efficiency. In that way, for a 1-meter-long capillary with 50-µm i.d., a 20 nL sample injection is suggested. Low loading capacity requires a high concentration of the sample for detection and limits the detection of low-abundant peptides and proteins. Online concentration methods, including field-amplified sample stacking (FASS), transient isotachophoresis (tITP), and dynamic pH junction are developed to increase the loading capacity [56,86].

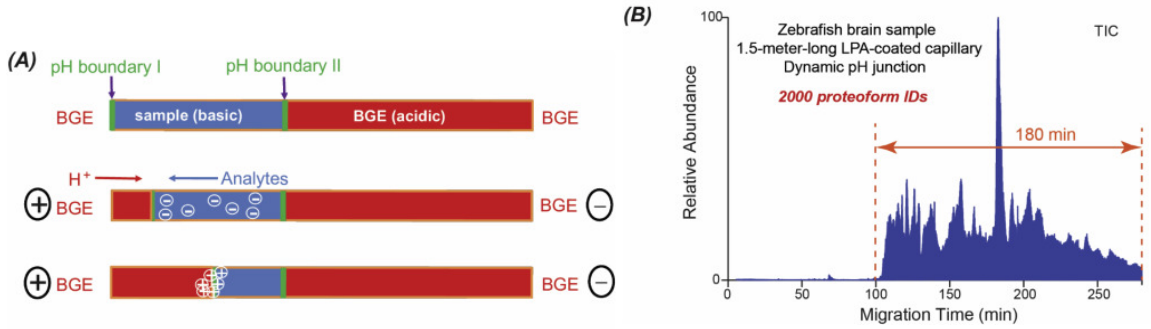


Figure 1.8. Dynamic pH junction. (A). The simplified diagram of the dynamic pH junction method. (B). The electropherogram of a 180-min separation window in a 1.5-meter-long LPA coated capillary using dynamic pH junction method. The figure is reprinted with permission from reference [73].

Dynamic pH junction utilizes the pH differences between the sample and the background electrolyte (BGE). As illustrated in **Figure 1.8 A**, the basic sample buffer (e.g., ammonium bicarbonate at pH 8) forms two boundaries with the acidic BGE (e.g., 5% acetic acid at pH 2.4). Negatively charged proteoforms in the basic buffer move toward the anode. Simultaneously, hydrogen ions from the BGE migrate towards the cathode. As pH boundary I is slowly titrated by hydrogen ions, the sample zone is concentrated, causing proteoforms to become positively charged and move toward the cathode. A microliter scale sample loading capacity (25-75% of the capillary volume) is achieved for large-scale BUP and TDP [83,87,88]. **Figure 1.8 B** shows the electropherogram of 66.7% loading capacity (2 µL) of the total capillary volume, achieving a 180-min separation window [75].

1.3 Capillary Zone Electrophoresis – Mass Spectrometry

Although LC-ESI-MS was introduced at the same time as CE-ESI-MS in the 1980s, LC-ESI-MS has been more widely adopted. Given the high sensitivity and low flow rate of CZE (i.e., 50-300 nL/min), the interface coupling CZE with MS is crucial to exploit the advantages of both analytical techniques.

1.3.1 CE-ESI-MS Interface

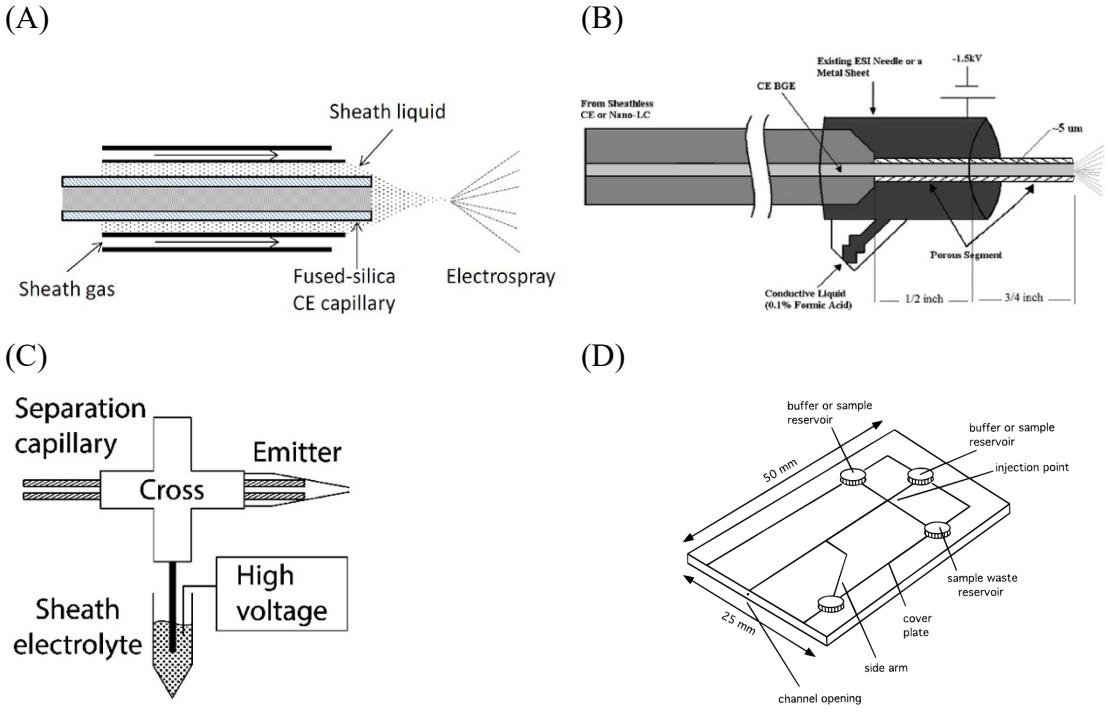


Figure 1.9. The schematic designs of four commercialized CE-ESI-MS interfaces [56]. (A). Coaxial sheath flow interface, commercialized by Agilent. The figure is reprinted from reference [89].

Figure 1.9. (cont'd)
(B). Sheathless interface using a porous capillary tip as ESI emitter commercialized by Sciex. The figure is reprinted with permission from reference [90]. (C). Electro-kinetically pumped sheath flow interface commercialized by CMP Scientific. The figure is reprinted with permission from reference [77]. (D). Glass microfluidic device with integrated CE and CE-MS using electroosmotic pumping commercialized by 908 Devices. The figure is reprinted from reference [91].

The interface must maintain the continuous circuit of the CZE current and provide a stable low flow rate to ensure efficient and reproducible sample introduction to the MS [89]. During the last decades, CE-MS has gained broader recognition and wider application in proteomics due to the development and commercialization of the CE-ESI-MS interface. **Figure 1.9 A-D** shows four commercialized CE-ESI-MS interfaces with relatively robust performance summarized in reference [56]. The coaxial sheath flow interface (**Figure 1.8 A**) developed by Smith et al. implements the capillary inside two coaxially larger tubes, with sheath liquid and nebulizing gas (i.e., nitrogen) facilitating ESI [92]. The voltage is applied to the first metal tube and delivered via the conductive sheath liquid at the tip. The sheath liquid operates in a split flow mode at 1 to 20 μL/min while the flow rate of CZE is typically between 20 and 100 nL/min. This difference in flow rate allows the use of MS-incompatible reagents in CZE, such as SDS, but also causes significant sample dilution, resulting in a lower sensitivity. Later improvement reduces the flow rate of sheath liquid control to improve the sensitivity. On the other hand, the sheathless interface (**Figure 1.8 B**) avoids signal dilution by inserting the capillary outlet with a porous tip into the metal ESI needle filled with conductive liquid. The electrical connection is built by ion transportation through the porous tip. Higher sensitivity is achieved but the separation buffer must support ESI. Internal pressure from the inlet end of the capillary (~100 mbar) is usually applied for a stable spray.

A comparable high-sensitivity sheath flow interface is developed by Dovichi group in 2010, called electrokinetically pumped sheath flow interface [93]. The interface design is based on the electroosmotic flow at around nL/min without a pump or nebulizer gas. As **Figure 1.8 C** shows, the separation capillary is placed inside a tapered glass emitter, and the liquid flows over the end of the separation capillary, closing the circuit and mixing with the capillary effluent. Third-generation design employs a larger diameter emitter orifice with a very short distance between the capillary tip and emitter orifice (20 μm) [77]. A 10% relative standard deviation in peak area, and low carry-over (much less than 1%) were observed. A low zmole detection limit

for peptides was achieved using this interface, demonstrating ten times higher sensitivity compared to RPLC in protein intensity [94]. A later study compared the sensitivity of a simplified nanoflow sheath liquid interface with a sheathless interface, showing similar performance in terms of sensitivity [95]. Using the sheathflow interface, internal pressure is not required, and diverse background electrolytes are compatible. Particularly, the sheath liquid design not only allows ampholytes inside of the capillary, but also serve as the chemical mobilizers in capillary isoelectric focusing method [96,97]. Electroosmotic pumped ionization is also applied at the corner of a rectangular glass microchip CE, eluting analytes to the spray tip without additional dead volume **(Figure 1.8 D)** [91,98]. Recently, a few different ionization setups have been developed for various purposes. For example, nanoCEasy includes separation capillary and sheath liquid capillary in a nanoflow sheath liquid CE-MS interface to prevent contaminants from entering the MS and keep a robust high sensitivity [99,100].

1.3.2 Advancements of CZE-MS in multilevel proteomics

Technological advancements in capillary coating, online concentration methods, and CE-MS interface have significantly increased the attention towards CE-MS. This field has experienced tremendous growth over the last decade, leading to broader applications to various biological studies [56,86].

Microliter-scale loading capacity and two to three hours of separation window are achieved for high-resolution and high sensitivity in CE-MS for both BUP and TDP levels [75,87]. Single-shot CZE-MS/MS identified nearly 600 proteoforms from *E. coli*. Triplicate single-shot CZE-MS/MS experiments quantitatively discovered 700 differentially expressed proteoforms between two regions in the zebrafish brain. Coupling with multidimensional separation increases both the scale and sensitivity of proteomic measurement. The nanoRPLC-CZE-MS/MS-based BUP identified 7500 proteins and 60,000 peptides using 5-μg proteome digest [101]. The high sensitivity of mass-limited samples indicates the potential of CZE-MS/MS in single-cell proteomics (SCP). Coupling reversed-phase based solid-phase microextraction (RP-SPME)-CZE-MS analysis for 0.1 ng sample injection, when only 0.25 ng HeLa digest in the sample vial, identified 257 ± 24 proteins and 523 ± 69 peptides (N = 2) [102]. Proteoform analysis of picogram-level injection is introduced from the pressure difference of ESI in TDP [103]. Over 200 proteoforms were detected from 50 pg of *E. coli* lysate and 100-400 proteoforms were detected from fewer than 50 cells. On-capillary cell lysis enabled the

identification of 40-88 proteoforms from 7 ± 2 mammalian cancer cells and 23-50 proteoforms from single Hela cells [104]. At the same time, large-scale proteomics using SEC-RPLC-CZE-MS/MS shows high peak capacity (~ 4000) and a large number of proteoform identifications (~ 6000) from *E. coli* [105]. Over 20, 000 proteoforms are identified from two colorectal cancer cell lines, unveiling differentially expressed proteoforms between two cell lines [44]. The long-term reproducibility of the CZE-MS/MS system in TDP studies, along with an efficient capillary cleanup procedure using 0.5% ammonium hydroxide shows high qualitative and quantitative reproducibility after initial decrease [106]. These technological advancements will benefit the understanding of important biomarkers associated with specific functions and diseases, as well as single-cell heterogeneity.

Open-tubular CZE eliminates the interaction between the analytes and the stationary phase, enabling the separation of proteins and protein complexes under native conditions. Native CZE-MS has high separation efficiency and high detection sensitivity for protein complexes compared with native SEC-MS. Native CZE-MS has been applied to analyzing low-complexity protein samples, i.e., monoclonal antibodies [107], large protein complexes like GroEL (near 1MDa) [108–110], ribosomes [111], and nucleosomes [112] as well as the complex proteome. Native SEC fractionation and online nCZE-MS analysis of an *E. coli* cell lysate identified 672 proteoforms and 23 protein complexes smaller than 30 kDa, representing the first native proteomics study of a complex proteome using online liquid-phase separation-MS [41]. In summary, CZE-MS/MS plays an important role in the characterization of multilevel proteomics.

1.4 Exploring the Plastoglobule Proteome

1.4.1 Plastoglobule (PG): Lipoprotein Particle with Dynamic Morphology

Plastoglobuli (PGs) are ubiquitous lipoprotein particles surrounded by a membrane lipid monolayer in different plastids (e.g., chloroplast) of plant tissues [113]. In chloroplast, PGs are attached to the thylakoid membrane, where the light-dependent reactions occur. Photons are absorbed and high-energy electrons are transferred via an electron transport chain at the thylakoid membrane. As **Figure 1.10 A** shows, PGs are around 25-75 nm in diameter and are located mostly at the high curvature of the thylakoid membrane [114,115]. PGs blister from the outer leaflet of the thylakoid membrane. Mostly under oxidative stress conditions, PGs form interconnected linkage groups surrounded by a single continuous half-lipid bilayer, as seen in **Figure 1.10 B**. This half-lipid bilayer constitutes a super hydrophobic core of PG, containing an

abundance of neutral lipids (such as triacylglycerols [TAG]), tocopherols and quinones (in particularly plastoquinone 9 [PQ-9], the electron transporter), and various carotenoids and derivatives [113]. Thus, the enzymatic and structural proteins are located on the surface of PGs, as illustrated in **Figure 1.10 C**.

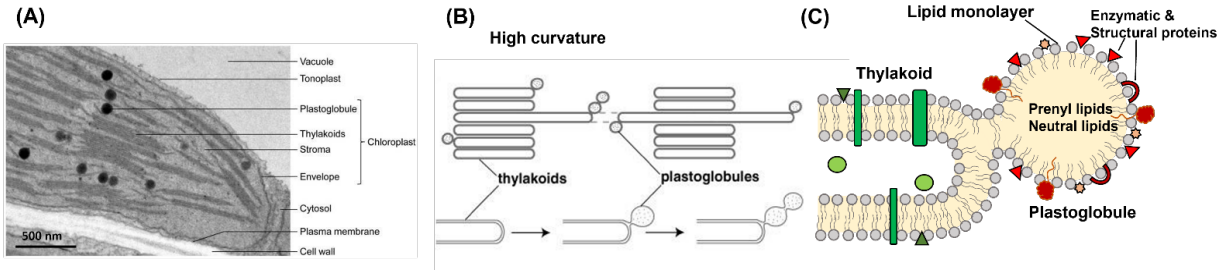


Figure 1.10. The electron micrograph of plastoglobules and the illustration model of plastoglobules. (A). Electron micrograph of an Arabidopsis leaf chloroplast showing plastoglobules in proximity with thylakoids. The figure is reprinted with permission from reference [114]. (B). Plastoglobules form on thylakoid membranes at the areas of high curvature and blister from the outer leaflet of the thylakoid membrane. PGs form interconnected linkage groups surrounded by a single continuous half-lipid bilayer. The figure is reprinted with permission from reference [115]. (C). The half-lipid bilayer that surrounds the plastoglobule is studded with structural and enzymatic proteins. The figure is adapted with permission from reference [115].

PGs have dynamic morphology and composition, rapidly increasing or decreasing in size under abiotic stress and during developmental transitions. During leaf senescence, the thylakoid membrane is dismantled, and the diameter of PGs increases more than 10-fold to 600 nm [116], as shown in **Figure 1.10 E-F**. PG size rapidly increases under light stress or nitrogen-limiting conditions and decreases back upon stress relief, as shown in **Figure 1.10 A-D** [113,117]. The size of PG and the extent of OsO_4 staining suggest that the metabolite exchange occurs between the thylakoid membrane and PGs. Thus, PGs may play an important functional role in senescence and stress response. Understanding the function and regulation of PGs will provide knowledge about plant stress resilience and metabolic engineering of biofuels.

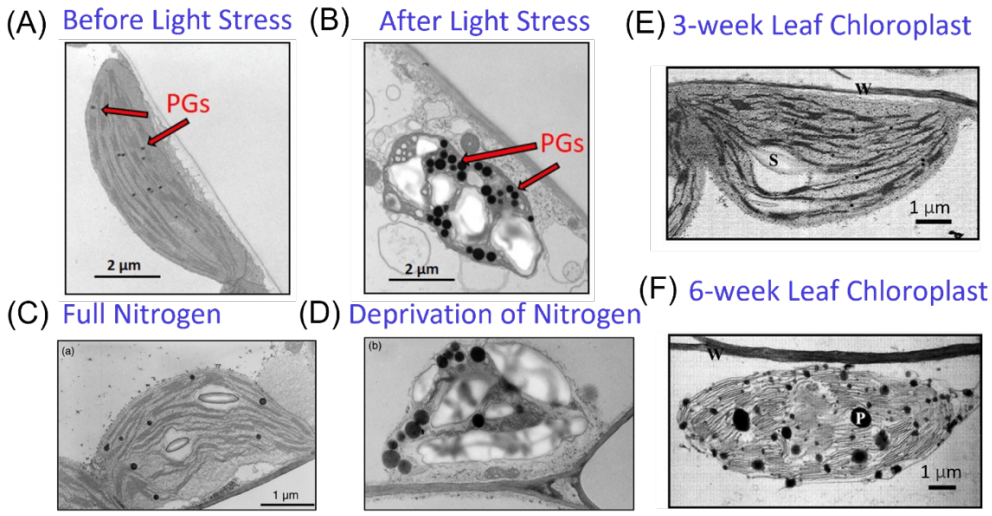


Figure 1.11. Dynamic morphology of PGs. (A) and (B) are the Arabidopsis leaf chloroplast before (120 µE) and after three days of light stress (520 µE), respectively. The figure is reprinted with permission from reference [113]. (C) and (D) is the electron microscopy of chloroplast from leaf mesophyll cells of Arabidopsis WT plants at full nutrition and deprived of nitrogen, respectively. The figure is reprinted with permission from reference [117]. (E) and (F) are electron micrographs of chloroplasts in young and senescing rosette leaves. S: starch grain; W: cell wall; P: plastoglobule. The figure is reprinted with permission from reference [116].

1.4.2 Plastoglobule proteome

PGs were initially considered as a lipid reservoir until bottom-up proteomics profiled the PG proteome purified from *Arabidopsis thaliana* chloroplasts using nanoLC-MS/MS in 2006[118]. Approximately 30 proteins are exclusively localized in PGs and quantified using LFQ[119]. The fibrillin family (FBN, also known as plastid-lipid associated protein [PAP]) is the most abundant protein family in chloroplast PGs and is proposed to play a role in stabilizing the globules. The absence of bc1 complex (ABC1) atypical kinase family is the second most abundant protein family. Seven fibrillin proteins and six ABC1K proteins constitute more than 70% of PG protein mass. Other proteins include enzymes exclusively localized on PG, such as tocopherol cyclase (VTE1) which catalyzes the cyclization of quinone substrates such as PQ-9 to tocochromanol products, phytol ester synthase 1 (PES1) and PES2 which are involved in the formation of phytyl esters following cleavage of the phytol tail from chlorophyll and cleavage of free fatty acids from galactolipids, NAD(P)H quinone dehydrogenase C1 (NDC1) which reduces oxidized PQ-9 to PQ-9-H_2 within PGs, and carotenoid cleavage dioxygenase (CCD4). And some enzymes involved in the plant hormone jasmonate are recruited to PGs in the double mutant of ABC1K1 and ABC1K3 (k1k3) but not wild type PGs, such as Lipoxygenase 2, 3, 4 (LOX2,

LOX3, LOX4) and Allene oxide cyclase 1 and 2 (AOC1, AOC2), and Allene oxide synthase (AOS) [120]. Some enzymes related to senescence and jasmonic acid biosynthesis are recruited under light stress, such as Plastoglobular M48 protease, Esterase/lipase/thioesterase (ELT4) [121]. These dynamic recruitments suggest that PGs are participating in thylakoid disassembly and jasmonate production. A functional network of the *Arabidopsis* PG proteome and genome-wide coexpression reveals four distinct modules that PG proteins are participating in: chlorophyll degradation/senescence, isoprenoid biosynthesis, plastid proteolysis, and redox regulators and phosphoregulators of electron flow [119].

Although the direct function and substrate of the ABC1K proteins in PGs are unclear, the hypothesis is that the ABC1K proteins regulate prenyl quinone metabolism via phosphorylation of enzymes in the pathway. Studies of loss-of-function mutants showed that ABC1K1 and ABC1K3 affect prenyl lipid content in PGs and their response to high light stress, as shown in **Figure 1.12** [120].

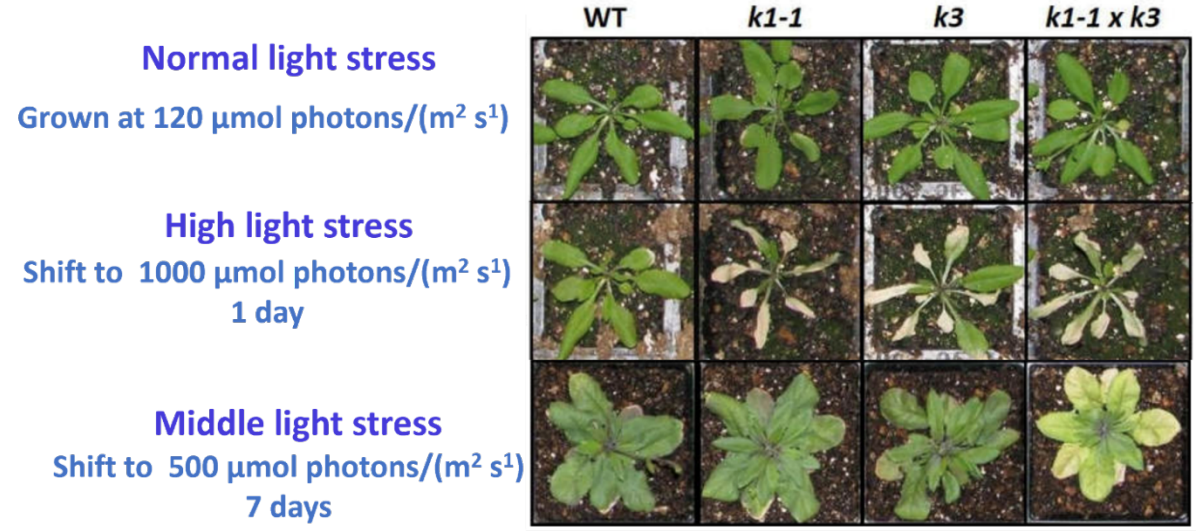


Figure 1.12. The loss of ABC1K1 and ABC1K3 mutant and double mutant have different responses to light stress. Mature 3-week-old plants grown under normal light stress were transferred to middle and high light stress at the described time points. WT: wild type. The figure is reprinted with permission from [120].

The original ABC1K proteins in *Saccharomyces* (Abc1p) and *E. coli* (UbiB) regulate the ubiquinone metabolism via phosphorylation of related enzymes [119,122,123]. A meta-analysis of published phosphoproteomics studies in *A. thaliana* shows that 16 of the 30 PG core proteins have reported phosphorylation sites, totaling more than 70 phosphorylation sites in PG-localized

and recruited proteins [124]. Among these, VTE1 was reported to be phosphorylated in vitro during the incubation with recombinant ABC1K1 or ABC1K3 [125,126]. A reduced activity of VTE1 was observed in the k1k3 mutant while the protein level remains unchanged [120]. So far, the in vivo phosphorylation lacks direct support, and the phosphorylation site is undetermined. To understand the function and regulation of ABC1K proteins, it is important to map the proteins with their PTMs, especially phosphorylation in PGs.

1.4.3 Research question and hypothesis

Current profiling of PGs has been performed only by BUP, offering high sensitivity and better localization of PTMs. However, a reported triple phosphorylated peptide of VTE1 was mapped inside the predicted chloroplast transit peptide (cTP) region, which is predicted to be cleaved after the protein is transferred into chloroplast [124,127]. In contrast, TDP characterizes the proteins in their intact sequence with PTMs, theoretically allowing for the identification of both the correct cTP cleavage site and the multiply phosphorylated forms of proteins. Despite this potential, there are some challenges in applying TDP directly on PG proteome:

First, PG core proteins localize at the periphery of PGs and may possess hydrophobic domains to anchor on the lipid core. Due to the monolayer structure of PGs, there is no transmembrane domain in PG-localized proteins despite of the hydrophobicity. Sixteen FBN protein products (including two truncation isoforms) are partitioned based on their hydrophobicity (GRAVY Index) and isoelectric point (pI). The seven PG-localized FBNs have low pIs and higher hydrophobicity compared to thylakoid-localized FBNs, as shown in **Figure 1.13 A**. PG-localized FBNs harbor an amphipathic helix at the lip of their β-barrel that is necessary for proper PG association [128]. Thus, the hydrophobic nature of PG localized proteins requires a reasonable solubility for effective separation.

Second, most of the PG core proteins exceed 35 kDa, as illustrated in **Figure 1.13 B**, suffering from the low signal-to-noise (S/N) as the proteoform mass increases due to a much broader charge state and isotopic distribution [47]. Additionally, the identification of large proteoforms requires effective separation with coeluted small proteins. Method optimization of TDP for PGs is therefore crucial to address these issues.

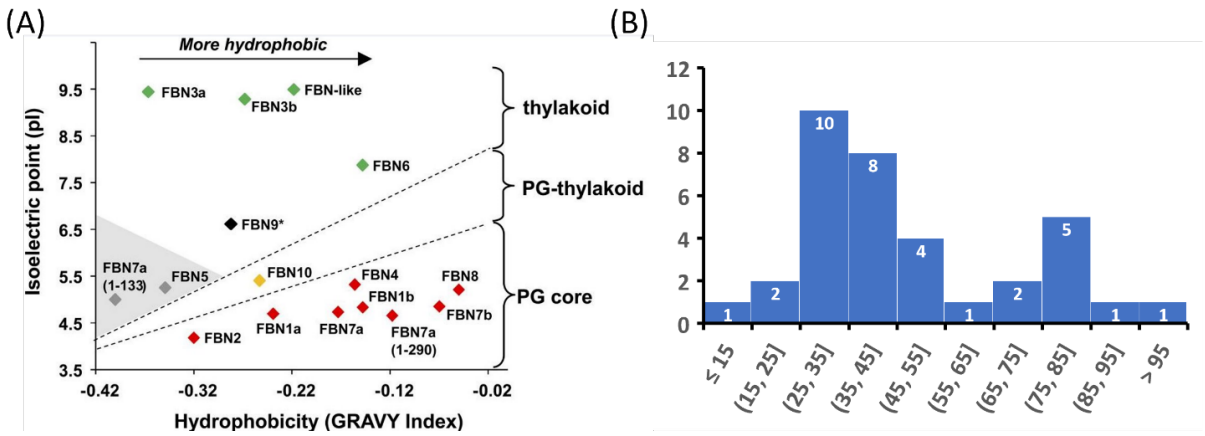


Figure 1.13. The challenges of analyzing PG protein using TDP. (A). The isoelectric point and hydrophobicity of fibrillin proteins. The figure is reprinted from reference [119]. (B). The molecular weight distribution of 35 PG core proteins, data collected from [129].

1.5 Summary

This chapter introduced the technology background of mass spectrometry-based proteomics for multilevel proteomic studies and the biological research background of PG and their proteome.

Bottom-up proteomic, top-down proteomic, and native proteomics each study peptides, proteoforms, and protein complexes in complex samples, respectively, with their own advantages and limitations. For example, BUP profiles multiple phosphorylation sites in plastoglobules to find the substrate targets of ABC1 kinase. However, the digested peptide information from BUP may lose the context of multiple phosphorylation on different peptides and cannot not distinguish the chloroplast transit peptide sequence from the mature protein. By combining multilevel proteomic approaches, studying the proteome at different perspectives, we can achieve a more comprehensive understanding of biological functions.

To advance TDP studies of PG proteome, several challenges must be addressed, such as the analysis of large and hydrophobic proteoforms. Effective separation is the key to overcoming these challenges. CZE-MS/MS shows high potential high potential as an analytical technique for proteomics due to its high sensitivity and lack of a stationary phase. Improvements in CZE-MS/MS, such as enhanced capillary coatings and dynamic pH concentration, have significantly increased its applicability. The subsequent chapters in this dissertation will discuss the method development of CZE-MS/MS for analyzing large and hydrophobic proteins, as well as the application of top-down proteomics in chloroplasts and PGs. These advancements aim to

improve the characterization of proteoforms and enhance our understanding of PGs in plant stress resilience and metabolic processes.

REFERENCES

[1] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003; 422:198–207.

[2] Yates JR, Ruse CI, Nakorchevsky A. Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. Annual Review of Biomedical Engineering. 2009; 11:49–79.

[3] Wilm M. Principles of Electrospray Ionization. Mol Cell Proteomics. 2011;10:M111.009407.

[4] Banerjee S, Mazumdar S. Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte. Int J Anal Chem. 2012; 2012:282574.

[5] Konermann L, Ahadi E, Rodriguez AD, et al. Unraveling the Mechanism of Electrospray Ionization. Anal Chem. 2013; 85:2–9.

[6] Aliyari E, Konermann L. Formation of Gaseous Proteins via the Ion Evaporation Model (IEM) in Electrospray Mass Spectrometry. Anal Chem. 2020; 92:10807–10814.

[7] Pimlott DJD, Konermann L. Using covalent modifications to distinguish protein electrospray mechanisms: Charged residue model (CRM) vs. chain ejection model (CEM). International Journal of Mass Spectrometry. 2021; 469:116678.

[8] Hogan CJ, Carroll JA, Rohrs HW, et al. A Combined Charged Residue-Field Emission Model of Macromolecular Electrospray Ionization. Anal Chem. 2009; 81:369–377.

[9] Fenn JB, Mann M, Meng CK, et al. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. Science. 1989; 246:64–71.

[10] Haag AM. Mass Analyzers and Mass Spectrometers. In: Mirzaei H, Carrasco M, editors. Modern Proteomics – Sample Preparation, Analysis and Practical Applications [Internet]. Cham: Springer International Publishing; 2016 [cited 2024 May 14]. p. 157–169.

[11] De Hoffmann E, Stroobant V, De Hoffmann E. Mass spectrometry: principles and applications. 3. ed., reprinted. Chichester Weinheim: Wiley; 2011.

[12] Chernushevich IV, Loboda AV, Thomson BA. An introduction to quadrupole–time-of-flight mass spectrometry. Journal of Mass Spectrometry. 2001; 36:849–865.

[13] Philipsen MH, Haxen ER, Manaprasertsak A, et al. Mapping the Chemistry of Hair Strands by Mass Spectrometry Imaging—A Review. Molecules. 2021; 26:7522.

[14] Hu Q, Noll RJ, Li H, et al. The Orbitrap: a new mass spectrometer. J Mass Spectrom. 2005; 40:430–443.

[15] Perry RH, Cooks RG, Noll RJ. Orbitrap mass spectrometry: Instrumentation, ion motion and applications. Mass Spectrometry Reviews. 2008; 27:661–699.

[16] Fernández-Costa C, Martínez-Bartolomé S, McClatchy DB, et al. Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results. J Proteome Res. 2020; 19:3153–3161.

[17] Hao Z, Hong Q, Zhang F, et al. Current Methods for the Characterization of Posttranslational Modifications in Therapeutic Proteins Using Orbitrap Mass Spectrometry. Protein Analysis using Mass Spectrometry. John Wiley & Sons, Ltd; 2017. p. 21–34.

[18] Shao C, Zhang Y, Sun W. Statistical characterization of HCD fragmentation patterns of tryptic peptides on an LTQ Orbitrap Velos mass spectrometer. Journal of Proteomics. 2014; 109:26–37.

[19] Brodbelt JS. Ion Activation Methods for Peptides and Proteins. Anal Chem. 2016; 88:30–51.

[20] Michalski A, Neuhauser N, Cox J, et al. A Systematic Investigation into the Nature of Tryptic HCD Spectra. J Proteome Res. 2012; 11:5479–5491.

[21] Wei B, Zenaidee MA, Lantz C, et al. Towards Understanding the Formation of Internal Fragments Generated by Collisionally Activated Dissociation for Top-Down Mass Spectrometry. Anal Chim Acta. 2022; 1194:339400.

[22] Peters-Clarke TM, Coon JJ, Riley NM. Instrumentation at the Leading Edge of Proteomics. Anal Chem. 2024; 96, 20, 7976-8010.

[23] Harvey SR, Ben-Nissan G, Sharon M, et al. Surface-Induced Dissociation for Protein ComplexProtein complexes Characterization. In: Sun L, Liu X, editors. Proteoform Identification: Methods and Protocols [Internet]. New York, NY: Springer US; 2022 [cited 2024 May 17]. p. 211–237. Available from: https://doi.org/10.1007/978-1-0716-2325-1_15.

[24] Eliuk S, Makarov A. Evolution of Orbitrap Mass Spectrometry Instrumentation. Annual Review of Analytical Chemistry. 2015; 8:61–80.

[25] Makarov A, Denisov E, Lange O. Performance Evaluation of a High-field Orbitrap Mass Analyzer. Journal of the American Society for Mass Spectrometry. 2009; 20:1391–1396.

[26] Michalski A, Damoc E, Hauschild J-P, et al. Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer*. Molecular & Cellular Proteomics. 2011;10:M111.011015.

[27] Scheltema RA, Hauschild J-P, Lange O, et al. The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-filter, High-performance Quadrupole and an Ultra-high-field Orbitrap Analyzer. Mol Cell Proteomics. 2014; 13:3698–3708.

[28] Denisov E, Damoc E, Makarov A. Exploring frontiers of orbitrap performance for long transients. International Journal of Mass Spectrometry. 2021; 466:116607.

[29] Smith LM, Kelleher NL. Proteoform: a single term describing protein complexity. Nat Methods. 2013; 10:186–187.

[30] Aebersold R, Agar JN, Amster IJ, et al. How many human proteoforms are there? Nat Chem Biol. 2018; 14:206–214.

[31] Cai W, Tucholski TM, Gregorich ZR, et al. Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. Expert Rev Proteomics. 2016; 13:717–730.

[32] Zhang Y, Fonslow BR, Shan B, et al. Protein Analysis by Shotgun/Bottom-up Proteomics. Chem Rev. 2013; 113:2343–2394.

[33] Dupree EJ, Jayathirtha M, Yorkey H, et al. A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. Proteomes. 2020; 8:14.

[34] Sun L, Zhu G, Yan X, et al. Capillary zone electrophoresis for bottom-up analysis of complex proteomes. Proteomics. 2016; 16:188–196.

[35] O'Connell JD, Paulo JA, O'Brien JJ, et al. Proteome-Wide Evaluation of Two Common Protein Quantification Methods. J Proteome Res. 2018; 17:1934–1942.

[36] Omenn GS, Lane L, Overall CM, et al. Progress Identifying and Analyzing the Human Proteome: 2021 Metrics from the HUPO Human Proteome Project. J Proteome Res. 2021; 20:5227–5240.

[37] Sinitcyn P, Richards AL, Weatheritt RJ, et al. Global detection of human variants and isoforms by deep proteome sequencing. Nat Biotechnol. 2023;1–11.

[38] Smith LM, Agar JN, Chamot-Rooke J, et al. The Human Proteoform Project: Defining the human proteome. Science Advances. 2021; 7:eabk0734.

[39] Habeck T, Brown KA, Des Soye B, et al. Top-down mass spectrometry of native proteoforms and their complexes: a community study. Nat Methods. 2024;1–9.

[40] Skinner OS, Haverland NA, Fornelli L, et al. Top-down characterization of endogenous protein complexes with native proteomics. Nat Chem Biol. 2018; 14:36–41.

[41] Shen X, Kou Q, Guo R, et al. Native Proteomics in Discovery Mode Using Size-Exclusion Chromatography–Capillary Zone Electrophoresis–Tandem Mass Spectrometry. Anal Chem. 2018; 90:10095–10099.

[42] Adams LM, DeHart CJ, Drown BS, et al. Mapping the KRAS proteoform landscape in colorectal cancer identifies truncated KRAS4B that decreases MAPK signaling. J Biol Chem. 2023; 299:102768.

[43] Melani RD, Gerbasi VR, Anderson LC, et al. The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. Science. 2022; 375:411–418.

[44] McCool EN, Xu T, Chen W, et al. Deep top-down proteomics revealed significant proteoform-level differences between metastatic and nonmetastatic colorectal cancer cells. Sci Adv. 8:eabq6348.

[45] Guo Y, Yu D, Cupp-Sutton KA, et al. A benchmarking protocol for intact protein-level Tandem Mass Tag (TMT) labeling for quantitative top-down proteomics. MethodsX. 2022; 9:101873.

[46] Tabb DL, Jeong K, Druart K, et al. Comparing Top-Down Proteoform Identification: Deconvolution, PrSM Overlap, and PTM Detection. J Proteome Res. 2023; 22:2199–2217.

[47] Compton PD, Zamdborg L, Thomas PM, et al. On the Scalability and Requirements of Whole Protein Mass Spectrometry. Anal Chem. 2011; 83:6868–6874.

[48] Marty MT, Baldwin AJ, Marklund EG, et al. Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. Anal Chem. 2015; 87:4370–4376.

[49] Slawski M, Hussong R, Tholey A, et al. Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. BMC Bioinformatics. 2012; 13:291.

[50] Cai W, Tucholski T, Chen B, et al. Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion Chromatography Strategy. Anal Chem. 2017; 89:5467–5475.

[51] Rolfs Z, Smith LM. Internal Fragment Ions Disambiguate and Increase Identifications in Top-Down Proteomics. J Proteome Res. 2021; 20:5412–5418.

[52] Lantz C, Zenaidee MA, Wei B, et al. ClipsMS: An Algorithm for Analyzing Internal Fragments Resulting from Top-Down Mass Spectrometry.

[53] Wei B, Lantz C, Liu W, et al. Added Value of Internal Fragments for Top-Down Mass Spectrometry of Intact Monoclonal Antibodies and Antibody–Drug Conjugates. Anal Chem. 2023; 95:9347–9356.

[54] Robinson CV. Mass spectrometry: From plasma proteins to mitochondrial membranes. Proc Natl Acad Sci USA. 2019; 116:2814–2820.

[55] Donnelly DP, Rawlins CM, DeHart CJ, et al. Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. Nat Methods. 2019; 16:587–594.

[56] Chen D, McCool EN, Yang Z, et al. Recent advances (2019-2021) of capillary electrophoresis-mass spectrometry for multilevel proteomics. Mass Spectrom Rev. 2023; 42:617–642.

[57] Wiśniewski JR, Zougman A, Nagaraj N, et al. Universal sample preparation method for proteome analysis. Nat Methods. 2009; 6:359–362.

[58] Tholey A, Kaulich P, Jeong K, et al. Influence of Different Sample Preparation Approaches on Proteoform Identification by Top-Down Proteomics. 2024; doi: 10.21203/rs.3.rs-3990966/v1\

[59] Yang Z, Shen X, Chen D, et al. Toward a Universal Sample Preparation Method for Denaturing Top-Down Proteomics of Complex Proteomes. J Proteome Res. 2020;19:3315–3325.

[60] Laganowsky A, Reading E, Hopper JTS, et al. Mass spectrometry of intact membrane protein complexes. Nat Protoc. 2013;8:639–651.

[61] Li Y, Compton PD, Tran JC, et al. Optimizing capillary electrophoresis for top-down proteomics of 30–80 kDa proteins. PROTEOMICS. 2014;14:1158–1164.

[62] Takemori A, Butcher DS, Harman VM, et al. PEPPI-MS: Polyacrylamide-Gel-Based Prefractionation for Analysis of Intact Proteoforms and Protein Complexes by Mass Spectrometry. J Proteome Res. 2020; 19:3779–3791.

[63] Takemori A, Kaulich PT, Cassidy L, et al. Size-Based Proteome Fractionation through Polyacrylamide Gel Electrophoresis Combined with LC-FAIMS-MS for In-Depth Top-Down Proteomics. Anal Chem. 2022; 94:12815–12821.

[64] Meyer VR. CHROMATOGRAPHY | Principles. In: Worsfold P, Townshend A, Poole C, editors. Encyclopedia of Analytical Science (Second Edition) [Internet]. Oxford: Elsevier; 2005 [cited 2024 May 25]. p. 98–105. Available from: https://www.sciencedirect.com/science/article/pii/B0123693977000893.

[65] Jorgenson JW. Capillary liquid chromatography at ultrahigh pressures. Annu Rev Anal Chem (Palo Alto Calif). 2010; 3:129–150.

[66] Xie F, Smith RD, Shen Y. Advanced proteomic liquid chromatography. Journal of Chromatography A. 2012; 1261:78–90.

[67] Xiang P, Zhu Y, Yang Y, et al. Picoflow Liquid Chromatography–Mass Spectrometry for Ultrasensitive Bottom-Up Proteomics Using 2-μm-i.d. Open Tubular Columns. Anal Chem. 2020; 92:4711–4715.

[68] Sorensen MJ, Anderson BG, Kennedy RT. Liquid chromatography above 20,000 PSI. TrAC Trends in Analytical Chemistry. 2020;124:115810.

[69] Anderson LC, DeHart CJ, Kaiser NK, et al. Identification and Characterization of Human Proteoforms by Top-Down LC-21 Tesla FT-ICR Mass Spectrometry. J Proteome Res. 2017; 16:1087–1096.

[70] Melby JA, Brown KA, Gregorich ZR, et al. High sensitivity top–down proteomics captures single muscle cell heterogeneity in large proteoforms. Proceedings of the National Academy of Sciences. 2023; 120:e2222081120.

[71] Wang C, Liang Y, Zhao B, et al. Ethane-Bridged Hybrid Monolithic Column with Large Mesopores for Boosting Top-Down Proteomic Analysis. Anal Chem. 2022; 94:6172–6179.

[72] Melby JA, Roberts DS, Larson EJ, et al. Novel Strategies to Address the Challenges in Top-Down Proteomics. J Am Soc Mass Spectrom. 2021; 32:1278–1294.

[73] Shen X, Yang Z, McCool EN, et al. Capillary zone electrophoresis-mass spectrometry for top-down proteomics. Trends Analyt Chem. 2019; 120:115644.

[74] Campuzano IDG, Li H, Bagal D, et al. Native MS Analysis of Bacteriorhodopsin and an Empty Nanodisc by Orthogonal Acceleration Time-of-Flight, Orbitrap and Ion Cyclotron Resonance. Anal Chem. 2016; 88:12427–12436.

[75] Lubeckyj RA, Basharat AR, Shen X, et al. 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. 2019; 30:1435–1445.

[76] Wang Y, Fonslow BR, Wong CCL, et al. Improving the Comprehensiveness and Sensitivity of Sheathless Capillary Electrophoresis–Tandem Mass Spectrometry for Proteomic Analysis. Anal Chem. 2012; 84:8505–8513.

[77] Sun L, Zhu G, Zhang Z, et al. Third-Generation Electrokinetically Pumped Sheath-Flow Nanospray Interface with Improved Stability and Sensitivity for Automated Capillary Zone Electrophoresis–Mass Spectrometry Analysis of Complex Proteome Digests. J Proteome Res. 2015; 14:2312–2321.

[78] Drown BS, Jooß K, Melani RD, et al. Mapping the Proteoform Landscape of Five Human Tissues. J Proteome Res. 2022; 21:1299–1310.

[79] Chen D, Lubeckyj RA, Yang Z, et al. Predicting Electrophoretic Mobility of Proteoforms for Large-Scale Top-Down Proteomics. Anal Chem. 2020; 92:3503–3507.

[80] Krokhin OV, Anderson G, Spicer V, et al. Predicting Electrophoretic Mobility of Tryptic Peptides for High-Throughput CZE-MS Analysis. Anal Chem. 2017; 89:2000–2008.

[81] An Introduction to Capillary Electrophoresis: Theory, Practice and Applications [Internet]. Analysis & Separations from Technology Networks. [cited 2024 May 27]. Available from: http://www.technologynetworks.com/analysis/articles/an-introduction-to-capillary-electrophoresis-theory-practice-and-applications-378737.

[82] Zhu G, Sun L, Dovichi NJ. Thermally-initiated free radical polymerization for reproducible production of stable linear polyacrylamide coated capillaries, and their application to proteomic analysis using capillary zone electrophoresis-mass spectrometry. Talanta. 2016; 146:839–843.

[83] Lubeckyj RA, McCool EN, Shen X, et al. Single-Shot Top-Down Proteomics with Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry for Identification of Nearly 600 Escherichia coli Proteoforms. Anal Chem. 2017; 89:12059–12067.

[84] Han X, Wang Y, Aslanian A, et al. Sheathless Capillary Electrophoresis-Tandem Mass Spectrometry for Top-Down Characterization of Pyrococcus furiosus Proteins on a Proteome Scale. Anal Chem. 2014; 86:11006–11012.

[85] Giorgetti J, Beck A, Leize-Wagner E, et al. Combination of intact, middle-up and bottom-up levels to characterize 7 therapeutic monoclonal antibodies by capillary electrophoresis - Mass spectrometry. J Pharm Biomed Anal. 2020; 182:113107.

[86] Gomes FP, Yates III JR. Recent trends of capillary electrophoresis-mass spectrometry in proteomics research. Mass Spectrometry Reviews. 2019; 38:445–460.

[87] Chen D, Shen X, Sun L. Capillary zone electrophoresis-mass spectrometry with microliter-scale loading capacity, 140 min separation window and high peak capacity for bottom-up proteomics. Analyst. 2017; 142:2118–2127.

[88] Zhu G, Sun L, Yan X, et al. Bottom-Up Proteomics of Escherichia coli Using Dynamic pH Junction Preconcentration and Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry. Anal Chem. 2014; 86:6331–6336.

[89] Wu H, Tang K. Highly Sensitive and Robust Capillary Electrophoresis-Electrospray Ionization-Mass Spectrometry: Interfaces, Preconcentration Techniques and Applications. Reviews in Analytical Chemistry. 2020; 39:45–55.

[90] Moini M. Simplifying CE−MS Operation. 2. Interfacing Low-Flow Separation Techniques to Mass Spectrometry Using a Porous Tip. Anal Chem. 2007; 79:4241–4246.

[91] Ramsey RS, Ramsey JM. Generating Electrospray from Microchip Devices Using Electroosmotic Pumping. Anal Chem. 1997; 69:1174–1178.

[92] Smith RD, Barinaga CJ, Udseth HR. Improved electrospray ionization interface for capillary zone electrophoresis-mass spectrometry. Anal Chem. 1988; 60:1948–1952.

[93] Wojcik R, Dada OO, Sadilek M, et al. Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. Rapid Commun Mass Spectrom. 2010; 24:2554–2560.

[94] MCCOOL EN, SUN L. Comparing nanoflow reversed-phase liquid chromatography-tandem mass spectrometry and capillary zone electrophoresis- tandem mass spectrometry for top-down proteomics. Se Pu. 2019; 37:878–886.

[95] Höcker O, Montealegre C, Neusüß C. Characterization of a nanoflow sheath liquid interface and comparison to a sheath liquid and a sheathless porous-tip interface for CE-ESI-MS in positive and negative ionization. Anal Bioanal Chem. 2018; 410:5265–5275.

[96] Xu T, Shen X, Yang Z, et al. Automated Capillary Isoelectric Focusing-Tandem Mass Spectrometry for Qualitative and Quantitative Top-Down Proteomics. Anal Chem. 2020; 92:15890–15898.

[97] Xu T, Han L, Thompson AMG, et al. An improved capillary isoelectric focusing-mass spectrometry method for high-resolution characterization of monoclonal antibody charge variants. Anal Methods. 2022; 14:383–393.

[98] Mellors JS, Gorbounov V, Ramsey RS, et al. Fully Integrated Glass Microfluidic Device for Performing High-Efficiency Capillary Electrophoresis and Electrospray Ionization Mass Spectrometry. Anal Chem. 2008; 80:6881–6887.

[99] Höcker O, Knierman M, Meixner J, et al. Two capillary approach for a multifunctional nanoflow sheath liquid interface for capillary electrophoresis-mass spectrometry. ELECTROPHORESIS. 2021; 42:369–373.

[100] Schlecht J, Stolz A, Hofmann A, et al. nanoCEasy: An Easy, Flexible, and Robust Nanoflow Sheath Liquid Capillary Electrophoresis-Mass Spectrometry Interface Based on 3D Printed Parts. Anal Chem. 2021; 93:14593–14598.

[101] Yang Z, Shen X, Chen D, et al. Microscale Reversed-Phase Liquid Chromatography/Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Deep and Highly Sensitive Bottom–Up Proteomics: Identification of 7500 Proteins with Five Micrograms of an MCF7 Proteome Digest. Anal Chem. 2018; 90:10479–10486.

[102] Colón Rosado JA, Sun L. Solid-Phase Microextraction-Aided Capillary Zone Electrophoresis-Mass Spectrometry: Toward Bottom-Up Proteomics of Single Human Cells. J Am Soc Mass Spectrom [Internet]. 2024 [cited 2024 May 27]; Available from: https://doi.org/10.1021/jasms.3c00429.

[103] Zhao Z, Guo Y, Chowdhury T, et al. Top-Down Proteomics Analysis of Picogram-Level Complex Samples Using Spray-Capillary-Based Capillary Electrophoresis–Mass Spectrometry. Anal Chem [Internet]. 2024 [cited 2024 May 27]; Available from: https://doi.org/10.1021/acs.analchem.4c01119.

[104] Johnson KR, Gao Y, Greguš M, et al. On-capillary Cell Lysis Enables Top-down Proteomic Analysis of Single Mammalian Cells by CE-MS/MS. Anal Chem. 2022;94:14358–14367.

[105] McCool EN, Lubeckyj R, Shen X, et al. Large-scale Top-down Proteomics Using Capillary Zone Electrophoresis Tandem Mass Spectrometry. J Vis Exp. 2018;

[106] Sadeghi SA, Chen W, Wang Q, et al. Pilot Evaluation of the Long-Term Reproducibility of Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Top-Down Proteomics of a Complex Proteome Sample. J Proteome Res. 2024;23:1399–1407.

[107] Shen X, Liang Z, Xu T, et al. Investigating native capillary zone electrophoresis-mass spectrometry on a high-end quadrupole-time-of-flight mass spectrometer for the characterization of monoclonal antibodies. Int J Mass Spectrom. 2021;462:116541.

[108] Marie A-L, Georgescauld F, Johnson KR, et al. Native Capillary Electrophoresis–Mass Spectrometry of Near 1 MDa Non-Covalent GroEL/GroES/Substrate Protein Complexes. Advanced Science. n/a:2306824.

[109] Jooß K, McGee JP, Melani RD, et al. Standard procedures for native CZE-MS of proteins and protein complexes up to 800 kDa. ELECTROPHORESIS. 2021; 42:1050–1059.

[110] Belov AM, Viner R, Santos MR, et al. Analysis of Proteins, Protein Complexes, and Organellar Proteomes Using Sheathless Capillary Zone Electrophoresis - Native Mass Spectrometry. J Am Soc Mass Spectrom. 2017; 28:2614–2634.

[111] Mehaffey MR, Xia Q, Brodbelt JS. Uniting Native Capillary Electrophoresis and Multistage Ultraviolet Photodissociation Mass Spectrometry for Online Separation and Characterization of Escherichia coli Ribosomal Proteins and Protein Complexes. Anal Chem. 2020; 92:15202–15211.

[112] Jooß K, Schachner LF, Watson R, et al. Separation and Characterization of Endogenous Nucleosomes by Native Capillary Zone Electrophoresis–Top-Down Mass Spectrometry. Anal Chem. 2021; 93:5151–5160.

[113] van Wijk KJ, Kessler F. Plastoglobuli: Plastid Microcompartments with Integrated Functions in Metabolism, Plastid Developmental Transitions, and Environmental Adaptation. Annual Review of Plant Biology. 2017; 68:253–289.

[114] Nacir H, Bréhélin C. When Proteomics Reveals Unsuspected Roles: The Plastoglobule Example. Front Plant Sci [Internet]. 2013 [cited 2019 Sep 23];4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3635846/.

[115] Austin JR, Frost E, Vidi P-A, et al. Plastoglobules Are Lipoprotein Subcompartments of the Chloroplast That Are Permanently Coupled to Thylakoid Membranes and Contain Biosynthetic Enzymes. The Plant Cell. 2006; 18:1693–1703.

[116] Kaup MT, Froese CD, Thompson JE. A Role for Diacylglycerol Acyltransferase during Leaf Senescence. Plant Physiology. 2002; 129:1616–1626.

[117] Gaude N, Brehelin C, Tischendorf G, et al. Nitrogen deficiency in Arabidopsis affects galactolipid composition and gene expression and results in accumulation of fatty acid phytyl esters. Plant Journal. 2007; 49:729–739.

[118] Ytterberg AJ, Peltier JB, van Wijk KJ. Protein profiling of plastoglobules in chloroplasts and chromoplasts. A surprising site for differential accumulation of metabolic enzymes. Plant Physiology. 2006; 140:984–997.

[119] Lundquist PK, Poliakov A, Bhuiyan NH, et al. The functional network of the Arabidopsis plastoglobule proteome based on quantitative proteomics and genome-wide coexpression analysis. Plant Physiol. 2012; 158:1172–1192.

[120] Lundquist PK, Poliakov A, Giacomelli L, et al. Loss of Plastoglobule Kinases ABC1K1 and ABC1K3 Causes Conditional Degreening, Modified Prenyl-Lipids, and Recruitment of the Jasmonic Acid Pathway. The Plant Cell. 2013; 25:1818–1839.

[121] Espinoza-Corral R, Schwenkert S, Lundquist PK. Molecular changes of Arabidopsis thaliana plastoglobules facilitate thylakoid membrane remodeling under high light stress. The Plant Journal. 2021; 106:1571–1587.

[122] Do TQ, Hsu AY, Jonassen T, et al. A Defect in Coenzyme Q Biosynthesis Is Responsible for the Respiratory Deficiency in Saccharomyces cerevisiae abc1Mutants *. Journal of Biological Chemistry. 2001; 276:18161–18168.

[123] Poon WW, Davis DE, Ha HT, et al. Identification of Escherichia coli ubiB, a Gene Required for the First Monooxygenase Step in Ubiquinone Biosynthesis. Journal of Bacteriology. 2000; 182:5139–5146.

[124] Lohscheider JN, Friso G, van Wijk KJ. Phosphorylation of plastoglobular proteins in Arabidopsis thaliana. J Exp Bot. 2016; 67:3975–3984.

[125] Martinis J, Glauser G, Valimareanu S, et al. A Chloroplast ABC1-like Kinase Regulates Vitamin E Metabolism in Arabidopsis. Plant Physiology. 2013; 162:652–662.

[126] Martinis J, Glauser G, Valimareanu S, et al. ABC1K1/PGR6 kinase: a regulatory link between photosynthetic activity and chloroplast metabolism. Plant J. 2014; 77:269–283.

[127] Emanuelsson O, Nielsen H, Brunak S, et al. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 2000; 300:1005–1016.

[128] Shivaiah, KK., Boren, DM., Tequia-Herrera, A., Vermaas, J., Lundquist, PK. An amphipathic helix drives interaction of Fibrillins with plastoglobuli lipid droplets. Biorxiv, 2023.09.28.559984.

[129] Sun Q, Zybailov B, Majeran W, et al. PPDB, the Plant Proteomics Database at Cornell. Nucleic Acids Res. 2009; 37:D969-974.

CHAPTER 2.* Large-scale top-down proteomics of the Arabidopsis thaliana leaf and chloroplast proteomes

2.1 Introduction

Proteins often undergo modifications following their translation, such as proteolytic processing or addition of covalent linkages that are critical for proper function. Because of such modifications, a bewildering array of potential proteoforms, each with a distinct chemical structure, can arise from a single genetic locus [1]. The combination of such modifications, ultimately resulting in a mature proteoform, can profoundly influence protein function, stability, interaction, structure, localization, or activity by altering physico–chemical properties of the protein. For example, a subset of the *Arabidopsis thaliana* light-harvesting complex II subunit pool is dynamically phosphorylated in response to light quality shifts, promoting their migration within the thylakoid membrane [2]. Similarly, the N-terminal residue of a sequence influences protein function and dictates protein stability through the so-called N-end rule [3, 4].

Proteolytic processing of proteins often serves to remove N-terminal sequence tags used to target a protein to its proper sub-cellular localization. These N-terminal extensions can range from 15 to 162 amino acids in length, and are proteolytically removed after import [5, 6] and, at least in some cases, are likely further processed at the N-terminus by unidentified peptidases [7]. The chloroplast alone is estimated to harbor approximately 3000 nuclear-encoded proteins, targeted using an obligatory N-terminal chloroplast transit peptide (cTP). A subset of such chloroplast-targeted proteins is subsequently further targeted to the lumen of the internal thylakoid membrane, requiring a second cleavable protein sequence called the lumenal targeting peptide (luTP), immediately downstream of the cTP. Likewise, signal peptides (SPs) are necessary for proper targeting to the secretory pathway including the endoplasmic reticulum and Golgi, while mitochondrial transit peptides (mTPs) are necessary for targeting to the mitochondria. The different sub-cellular localization signals have broadly distinct characteristics that can facilitate their prediction from primary protein sequences. Multiple algorithms have been developed to predict sub-cellular localization sequences and propose cleavage sites of the localization signal [5, 8-11]. However, sequence conservation among targeting signals is almost

* *This chapter is partially adapted with permission from Wang, Q., Lundquist, PK., Sun, L. Large-scale top-down proteomics of the Arabidopsis thaliana leaf and chloroplast proteomes. Proteomics. 2022; 23(3-4):2100277.*

wholly non-existent, and exceptions to the general sequence patterns abound. This has made the prediction of targeting signals difficult [5].

Numerous bottom-up proteomics (BUP) studies have been undertaken for the large-scale determination of N- and C-termini of mature protein sequences. A comprehensive proteomics analysis of the chloroplast was performed in part to establish the N-termini of chloroplast proteins and thereby propose cTP cleavage sites [12]. However, information on N-termini was limited to the subset of N-terminally acetylated tryptic peptides. Subsequent efforts employed a covalent tagging approach that greatly expanded the coverage of identified N-termini of chloroplast proteoforms [7]. This work identified a clear enrichment for N-terminal residues of Ala, Val, Thr, and Ser. However, the bottom-up nature of the study limited the ability to characterize N-termini in the context of full-length sequence or possible covalent modifications.

In contrast, top-down proteomics (TDP) directly characterizes the primary, intact sequence of different proteoforms. In 2002, Whitelegge et al. applied intact mass measurements to the chloroplast grana proteome, in which one of the first single-pass membrane proteoforms was defined [13]. Since then, the subunits of the cytochrome b6f complex [14, 15], the photosystem II complex (PSII) [16], and the 26S proteasome [17] have been investigated using TDP. Novel insights, such as the presence of palmitoylation, phosphorylation and distinct lipid modifications have been gleaned [18], expanding our understanding of the composition and assembly of large protein complexes of the plant cell. TDP also provides an effective strategy to determine the mature (i.e., post-transit peptide cleavage) proteoform identities of a proteome while avoiding extra sample handling steps and artificial covalent modifications [19]. Smith et al. have established a five-level classification system that assesses the ambiguity a given proteoform identification concerning the PTM localization, PTM identification, amino acid sequence, and gene, ranging from no ambiguity (Level 1) to ambiguity among all four categories (Level 5) [20]. Among the first applications of TDP to chloroplast samples was the use of three-dimensional Fourier transform MS [21]. Of the 22 molecular weight values found (from 9 to 26 kDa), seven proteins were fully characterized, in comparison to 97 identified by BUP. The application of TDP could delineate similar proteins differing only by 12 residues, differentiate proteins with and without N-methylation, and correct the cleavage site of transit peptides. While the TDP applications from these early studies represent pioneering efforts that provided significant biological insights, characterization of sequence tags was performed manually, and

protein separation was performed offline with direct infusion [21], or by reverse-phase liquid chromatography (RPLC) separation [13].

Compared with RPLC, capillary-zone electrophoresis (CZE) is known for its high separation efficiency for large biomolecules and high sensitivity for intact protein characterization [22, 23]. The advanced CZE-MS/MS interface [24, 25], capillary coating [26], and online stacking methods enabled identification of nearly 600 intact proteoforms from an Escherichia coli cell lysate in a single shot CZE- MS/MS [27]. Furthermore, 5700 proteoforms were identified from *E. coli* lysate by combining size exclusion chromatography (SEC) with RPLC pre-fractionation [28]. Orthogonal to SEC and RPLC, CZE separates proteoforms according to their different electrophoretic mobility (μ_{ef}), which is directly related to the size and charge of the proteoform. Thus, charge-modified PTMs should alter mobility of proteoforms in a predictable manner, unlike migration in RPLC [29].

With the substantial advancements in informatics tools for proteoform identification, such as ProSight [30, 31] and TopPIC suite [32], and the advancement of multi-dimensional separations, numerous TDP experiments have been successfully applied to human and animal samples [33]. In contrast, large-scale TDP studies in plants have been broadly lacking since the initial studies of the early 2000s. Given that, we performed a large-scale TDP analysis of A. thaliana leaf and chloroplast samples using two-dimensional orthogonal separations of SEC followed by CZE-MS/MS. A. thaliana was selected for our study as it represents the foremost model plant species with a high-quality annotated genome. We identified 3198 and 1836 proteoforms from the total leaf and the chloroplast sample, respectively, and a total of 4782 unique proteoforms across the two samples. We identified numerous PTMs, established protein N-termini, and corrected predicted sub-cellular localization signals. New chloroplast protein targets of Trp oxidation, indicative of singlet oxygen retrograde signaling, were found. This work fills a significant gap in plant proteoform characterization, demonstrates the advancement of TDP methods, and provides the foundation for future developments in the characterization of intact protein species of plant proteomes.

2.2 Experimental section

2.2.1 Materials and Reagents

Acrylamide was purchased from Acros Organics (NJ, USA). Ammonium bicarbonate (NH_4HCO_3), urea, dithiothreitol (DTT), iodoacetamide (IAA) and 3-(Trimethoxysilyl) propyl

methacrylate, Tris-HCl, Hepes, $MgCl_2$, phosphatase inhibitors (NaF, Na-Orthovanadate, Na-Pyrophosphate·$10H_2O$, β-Glycerophosphate·$2Na·5H_2O$), and protease inhibitors (Chymostatin, Antipain·2HCl, Bestatin, E-64, Leupeptin (hemisulfate), P-ramidon·2Na, AEBSF, Aprotinin) were purchased from Sigma-Aldrich (St. Louis, MO). LC/MS grade water, acetonitrile (ACN), methanol, and formic acid were purchased from Fisher Scientific (Pittsburgh, PA). Aqueous mixtures were filtered with Nalgene Rapid-Flow Filter units (Thermo Scientific) with 0.2 μm CN membrane and 50 mm diameter. Fused silica capillaries (50 μm i.d./360 μm o.d.) were obtained from Polymicro Technologies (Phoenix, AZ). Complete, mini protease inhibitor cocktail (provided in EASYpacks) was bought from Roche (Indianapolis, IN).

2.2.2 Sample collection

Total leaf samples: Two trays of eight-week-old *A. thaliana* (ecotype Columbia-0) were grown at 16/8 light/dark photoperiod, 20 °C. Leaves were cut from 64 plants, pooled and flash frozen in a mortar with liquid nitrogen, and thoroughly ground. The powder was then mixed with a lysis buffer containing 50 mM Tris-HCl (pH 8.0), 2% SDS and protease inhibitor cocktail by pipetting, and then vortexed for 25 s. After centrifugation (13,000 rpm for 2.5 min), the supernatant containing the extracted proteins was collected and stored in -80 °C. Whole chloroplast samples: 4 trays (128 plants) of 39-day-old *A.* thaliana Col-0 were grown at 10/14 light/dark photoperiod. Leaves were cut and washed in pre-cold water, and ground in isolation buffer (330 mM sorbitol, 20 mM Hepes, 13 mM Tris-HCl, 3 mM $MgCl_2$, 0.1% fat-free BSA, 5 mM ascorbic acid, and 5 mM reduced cysteine, and phosphatase inhibitors) using a Waring blender with medium intensity for 10 seconds. Lysate was filtered through 1 layer of gauze, and then centrifuged for 5 minutes at 1500g. The supernatant was discarded, and the pellet was resuspended in the wash buffer (330 mM sorbitol and 50 mM Hepes with phosphatase inhibitors). Pellet was washed and re-collected with 5 min centrifugation at 1500 g, after which the pellet was resuspended in 6 mL osmotic shocking buffer (0.6 M sucrose, 1 mM EDTA, 10 mM Tricine, protease inhibitors and phosphatase inhibitors), sitting on ice for 30 minutes. Samples were collected in 15 mL falcon tubes, lyophilized, and stored in -80 °C. To prepare samples for MS/MS, they were thawed on ice, sonicated with a Branson Sonifier 250 (VWR Scientific, Batavia, IL) on ice for 10 mins, and then resuspended in 2% SDS with protease inhibitor cocktail. The protease inhibitors with final concentrations are listed below: 50 μg/ml Antipain·2HCl, 40 μg/ml Bestatin, 10 μg/ml Chymostatin, 10 μg/ml E-64, 5 μg/ml Leupeptin

(hemisulfate), 10 µg/ml P-ramidon· 2Na, 50 µg/ml AEBSF, 2 µg/ml Aprotinin. The concentration of phosphatase inhibitors is: 50 mM NaF, 25 mM β-Glycerophosphate·2Na·5H_2O, 1 mM Na-Orthovanadate, 10 mM Na-Pyrophosphate·10H_2O.

2.2.3 Sample preparation

A 4:1 (v/v) ratio of acetone was added to solubilized protein samples (both chloroplast and total leaf) with overnight precipitation. A 10,000 x g centrifugation removed the supernatant, and the protein pellet was resuspended in 8 M urea and 100 mM ammonium bicarbonate (pH 8.0), denatured at 37 °C for 30 minutes, reduced with dithiothreitol (DTT) at 37 °C for 30 minutes and alkylated with iodoacetamide (IAA) at room temperature without light for 20 minutes. Then, samples were desalted by a 30 kDa molecular weight cut off centrifugal filter (Millipore Sigma, Inc.) washed with 100 mM NH_4HCO_3 (pH 8.0). Finally, sample was diluted into 50 mM NH_4HCO_3 (pH 8.0).

2.2.4 Size exclusion chromatography separation

Samples were fractionated by size exclusion chromatography (SEC) in preparation for CZE-MS/MS analysis. For total leaf sample, the SEC column was 4.6 x 300 mm, 3 µm particles, 300 Å pores from Agilent, the mobile phase was 0.1% (v/v) FA, and the flow rate was 0.25 mL/min. The column temperature was kept at 40 °C. We collected 6 fractions from 10-22 min (2 min for each fraction) from 120 µL of 1mg/mL total leaf sample input. For chloroplast sample, the Bio SEC-5 column (4.6 x 300 mm, 3 µm particles, 300 Å pores) from Agilent was used. The mobile phase was 0.4% (v/v) FA, and the flow rate was 0.25 mL/min. The column temperature was kept at 40 °C. We collected 9 fractions from 10-36 min (4 min for each fraction except for 2 min for the 3rd, 4th, 5th, and 6th fractions) from 60 µL of 2 mg/mL chloroplast protein sample input.

2.2.5 CZE-MS/MS

An automated ECE-001 CE autosampler and a commercialized electro-kinetically pumped sheath flow CE−MS interface from CMP Scientific (Brooklyn, NY) [24, 25] was coupled to a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific). The protocol of a 100-min CZE-MS/MS and parameters of QE-HF were according to McCool, et al [28]. Briefly, a fused silica capillary (50 µm i.d., 360 µm o.d., 1 m) was coated with linear polyacrylamide (LPA) and etched with hydrofluoric acid at the end near the CE-MS interface to reduce the outer diameter of the capillary. Acetic acid, 10% (v/v), was used as the background electrolyte (BGE). The sheath buffer was 0.2% (v/v) formic acid containing 10% (v/v) methanol. Sample injection

was carried out by applying pressure (5 psi) at the sample injection end, and the injection periods were calculated based on Poiseuille's law for different sample loading volumes [34]. For the total leaf sample, the 500 nl injection volume was adopted with ~500 ng of protein in each sample assuming equal distribution of protein among each SEC fraction. For chloroplast samples which have a relatively smaller proteome, the injection volume was 250 nL (~160 ng of protein assuming equal distribution among SEC fractions) except 200 nL for fraction 5 and 6 due to the higher protein abundance. A dynamic pH junction was used to concentrate sample in acidic BGE to accommodate the large injection volumes [35-37, 23]. A 30 kV high voltage was applied at the injection end of the capillary and around 2.0–2.2 kV was applied for electrospray. MS parameters are listed as follows. Microscan is 3 for both full MS and MS/MS. For full MS, the resolution was 240,000 at m/z 200, and AGC target value was 1E6 with 50 ms maximum injection time. The scan range was 600–2000 m/z and the top 5 ions of the highest intensity in full MS were isolated with a 4 m/z isolation window and fragmented with a 20% normalized collision energy. The resolution for MS/MS is 120,000 at m/z 200, and the AGC target was 1E5 with 200 ms maximum injection time. Intact protein mode and exclude isotopes settings were on. Proteins with 1–5 charge state were excluded and the dynamic exclusion was set for 30s.

2.2.6 Data analysis

All .raw files were converted to mzML by MSconvert tool and then analyzed with the TopFD and TopPIC pipeline [32] with an estimated 1% FDR at the spectrum level and 5% FDR at the proteoform level. Cysteine carbamidomethylation was set as a fixed modification. The maximum number of unexpected modifications was 2. The precursor and fragment mass error tolerances were 15 ppm. The maximum mass shift of unknown modifications was 500 Da.

2.3 Results

2.3.1 Top-Down Proteomics Workflow of Leaf and Chloroplast Samples

The general workflow for capillary zone electrophoresis separation coupled with electrospray ionization-tandem mass spectrometry (CZE-ESI-MS/MS) is shown in **Figure 2.1**, along with a representative electropherogram. A sample of total leaf tissue was prepared from wild-type A. thaliana leaves late in the vegetative growth stage using 2% SDS in the presence of a protease inhibitor cocktail. After sample collection and purification, the sample was separated by SEC into six fractions, followed by a 100-minute-CZE-ESI-MS/MS online separation of each fraction. Using an estimated 5% FDR at the proteoform level and 1% FDR at the MS/MS spectrum

level by the TopPIC suite [32], we identified 3198 unique proteoforms from 458 proteins across the six SEC fractions of the total leaf sample with an average of 20.8 matched fragment ions per proteoform.

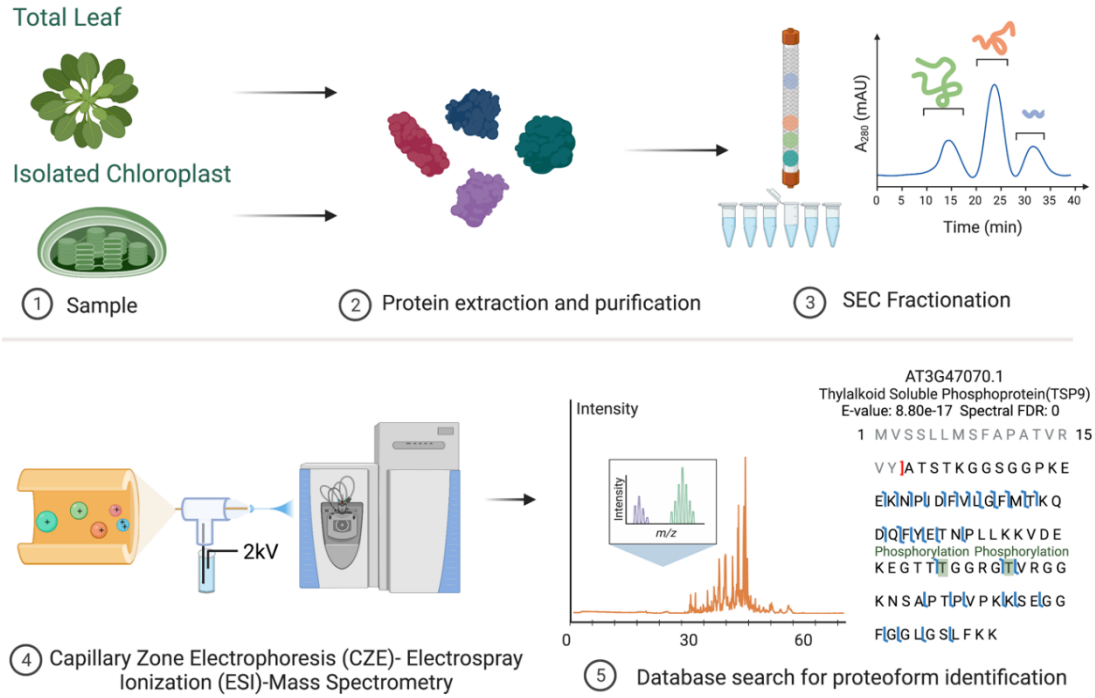


Figure 2.1. The top-down proteomics workflow of total leaf and isolated chloroplast samples. After isolation of total leaf and chloroplast samples, proteins were extracted in 2% (w/v) sodium dodecyl sulfate detergent, precipitated by acetone, resuspended with 8 M Urea in 100 mM ammonium bicarbonate (ABC), and buffer exchanged into 100 mM ABC. After separation into six size-exclusion chromatography (SEC) fractions, they were characterized by CZE-ESI-MS/MS and identified with the TopPIC MS/MS analysis suite using the TAIR10 *A. thaliana* protein sequence database. An example of an identified double phosphorylation on the Thylakoid Soluble Protein 9 (TSP9) is shown with its fragmentation pattern. Figure was made in BioRender.

According to the 5-level classification system established by Smith et al. [20] that describes the level of ambiguity within a proteoform identification, 47.9% of proteoform identifications are categorized as Level 1, indicating no ambiguity at all, in which PTMs are both characterized and well assigned (**Table 2.1**). To provide a targeted survey of chloroplast proteoforms, we also investigated whole chloroplast samples isolated from *A. thaliana* leaf tissue during the middle of vegetative growth. The same pipeline as for total leaf was used for the subsequent proteomics analysis using nine SEC fractions with a 120-minute CZE-ESI-MS/MS online separation for each fraction.

Table 2.1. Five-Level Classification of Identified Proteoforms

Level:		1	2A	2B	3	4	5	Total
Total Leaf	No	1233	0	0	0	0	0	1233
	1 Mod	272	42	177	1101	0	0	1592
	2 Mods	26	0	22	267	0	0	315
	3 Mods	1	0	0	57	0	0	56
	Total	1532	42	199	1425	0	0	3198
Chloroplast	No	636	0	0	0	0	0	636
	1 Mod	101	105	119	741	0	0	1272
	2 Mods	10	7	6	111	0	0	134
	3 Mods	0	0	0	0	0	0	0
	Total	747	112	125	852	0	0	1836

In total, 1836 proteoforms from 200 proteins were identified in the chloroplast sample, with 40.7% of proteoform identifications categorized at Level 1. The proteoforms were identified with an average of 20.7 fragment ions per proteoform. We illustrate MS/MS spectra and electropherogram of one representative proteoform in **Figure 2.2**.

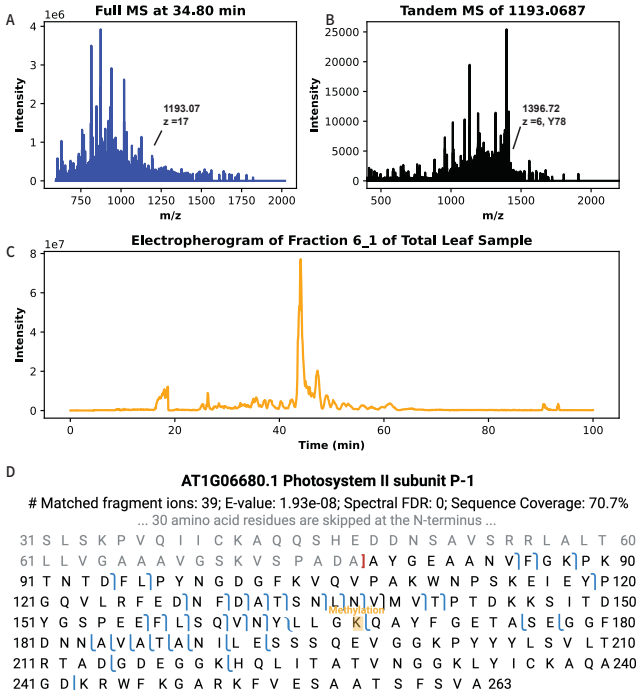


Figure 2.2. The electropherogram, mass spectrum, and fragmentation map for proteoform #1685 [AT1G06680] in the total leaf sample. (A), (B), (C), and (D) shows the full mass spectrum, the tandem mass spectrum, the electropherogram, and the fragmentation map of the proteoform, respectively. The sequence coverage is compared to the full length of the sequence.

Comparing the total leaf and chloroplast samples, identifications of 242 proteoforms from 99 proteins are present in both experiments (**Figure 2.3**). As an example output, the fragmentation pattern of a double phosphorylated proteoform of Thylakoid Soluble Phosphoprotein 9 (at3g47070) is shown in **Figure 2.1**. Residue-level assignment of the two phosphorylation sites is possible due to the fragmentation within the consecutive Thr residues.

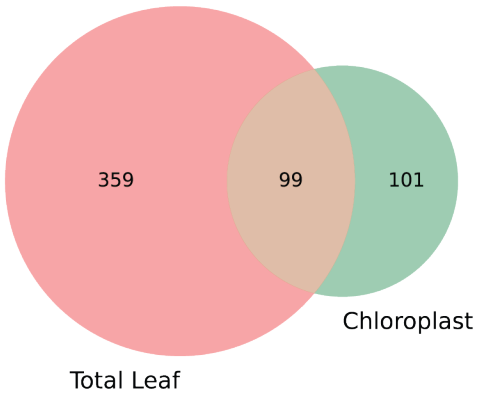


Figure 2.3. Venn diagram of proteins identified in total leaf sample and chloroplast sample.

2.3.2 Proteoform Mass Shifts and Post-translational Modifications

Across the total leaf samples, a total of 2390 mass shifts were identified. In fact, over 61% of identified proteoforms in our datasets contained at least one mass shift. We generated a histogram of these shifts to identify those most frequently represented within our datasets (**Figure 2.4**). The most prevalent mass shifts match with common PTMs, such as acetylation (460 proteoforms), N-terminal Met excision (357 proteoforms), oxidation (179 proteoforms), and methylation (28 proteoforms). Acetylation was the most abundant PTM identified in the total leaf proteoforms, 87% of which was found to occur on the N-terminus, generally accompanied by Met excision. Oxidation was found on multiple amino acids, most frequently on Lys. Twenty-eight proteoforms were methylated which localized on Lys, Gln, Asp, and Glu. Sixteen proteoforms were found to be phosphorylated despite the absence of special enrichment of phosphoproteins or the use of phosphatase inhibitors. Numerous other mass shifts were found which could not be readily assigned to a specific PTM. Possible interpretations of these unassigned mass shifts, based on the Unimod database [38], are indicated in **Figure 2.4**.

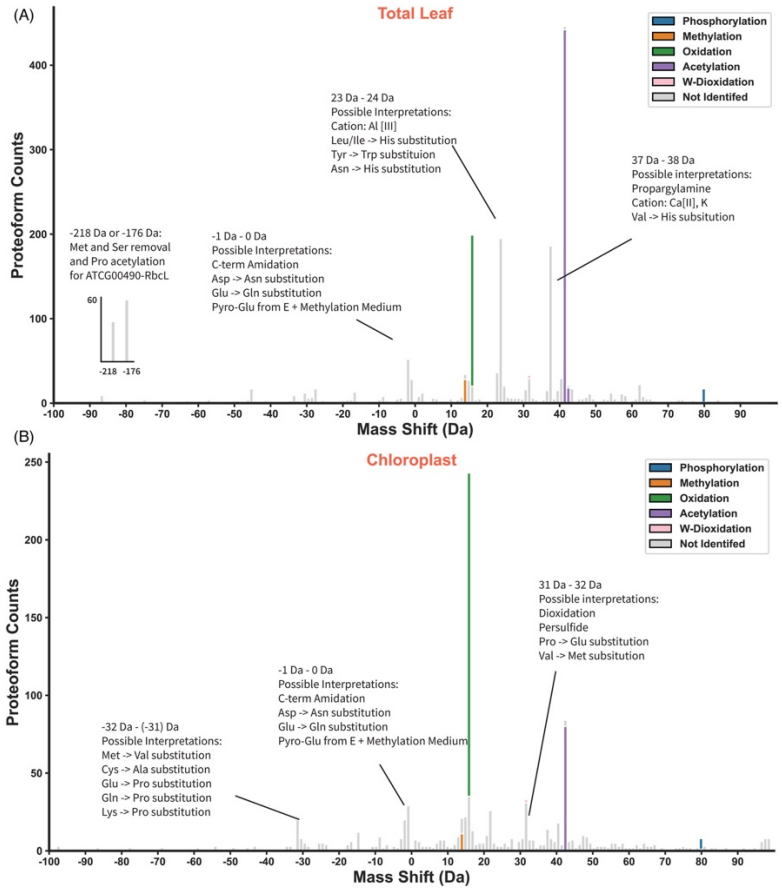


Figure 2.4. Proteoform mass shift distribution from –100 Da to +100 Da. Common PTMs are marked according to the color code indicated in the legend. (A). A histogram of mass shifts among proteoforms identified in the total leaf sample. Two bins containing prevalent mass shifts which are out of the indicated range are included as subsets in the bottom left, both of which are specific to RbcL proteoforms. Because the underlying predicted N-terminus of the proteoforms are different, both mass shifts result in RbcL proteoforms that initiate with an N-terminally acetylated Pro3 residue, as described in the text. (B). A histogram of mass shifts among proteoforms identified in the chloroplast sample. The bin size of all histograms is 1 Da. Possible interpretations of the three most prevalent unidentified mass shifts are proposed based on the Unimod database.

While almost all mass shifts fell within the –100 to +100 Da range, we did identify two peaks of – 176.1 Da and – 218.1 Da found on 59 and 49 proteoforms, respectively, of the Rubisco large subunit (RbcL, tcg00490). Manual interpretation indicated that software incorrectly predicted the N-terminal residue to be the initiating Met1. The mass shift of - 176.1 Da was found on proteoforms lacking a predicted acetylation and was consistent with removal of Met1 and Ser2 and inclusion of an acetylation. Meanwhile, the – 218.1 Da mass shift was found on proteoforms with a predicted acetylation and was consistent with removal of Met1 and Ser2. Similarly, 14 RbcL proteoforms, all with a predicted removal of Met1, were found with a mass

shift of – 45.1 Da, consistent with removal of Ser2 and an acetylation. Thus, we manually corrected 122 RbcL proteoforms, all resulting in Pro3 as the first residue and with the presence of an acetylation. In total, 144 proteoforms of RbcL were found with Pro3 as first residue and an acetyl PTM. This assignment is consistent with previous studies [39] and with their electrophoretic mobility in CZE, as described below in **Section 2.3.3**.

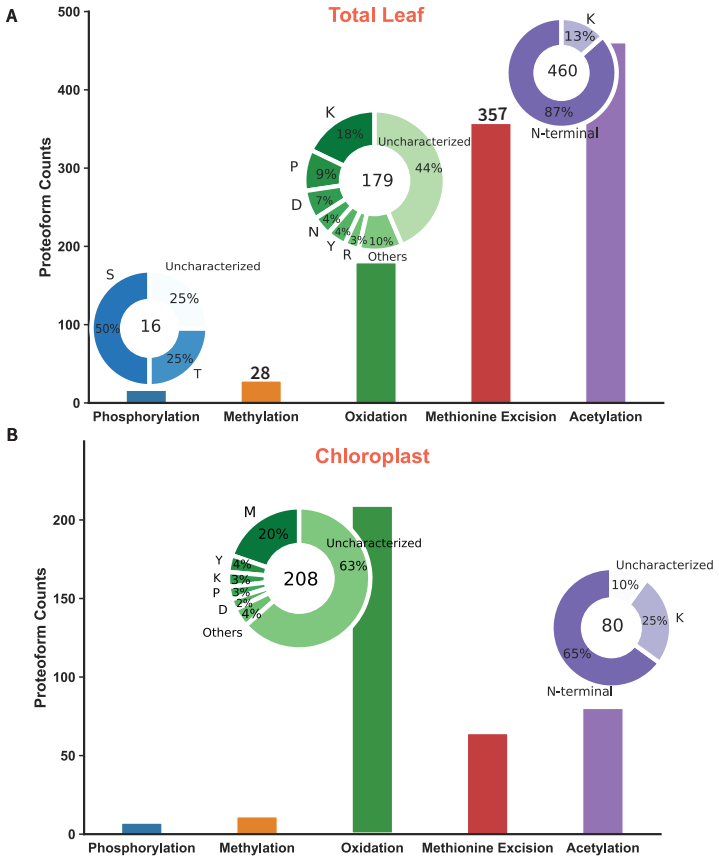


Figure 2.5. Prevalence of common PTMs identified on proteoforms in total leaf (A) and in chloroplast (B) samples. The modified residues receiving a given PTM are shown in the corresponding pie charts.

It is notable that very few proteoforms were found to start with Ser2, which amount to less than 4% of total RbcL proteoform feature intensity. Although the mature form of RbcL is recognized to begin with an N-acetylated Pro3, it remains unclear whether processing of the initiating Met1 and Ser2 occurs in a stepwise fashion or as a single cleavage event between Ser2 and Pro3. Our proteoform identifications uncover little evidence of an intermediate state in which only the Met1 is removed. This strongly suggests that N-terminal processing of RbcL occurs in a single step from an unknown dipeptidase, as suggested previously [40].

We identified 16 proteoforms with phosphorylation in the total leaf sample, including five different phosphorylated proteoforms of Plastocyanin-1 and -2 (at1g76100, at1g20340), as shown in **Figure 2.5 A**. Although twelve of the sixteen proteoforms were chloroplast-localized, a non-overlapping set of phosphorylated proteoforms were identified in the chloroplast sample. This included three proteoforms of TSP9. As noted above and observed in previous BUP experiments [41, 42], double phosphorylation of Thr66 and Thr71 was observed. Phosphorylation on Thr64 was also observed in two other proteoforms, however never shared with phosphorylation of Thr66 or Thr71. This suggests that phosphorylation of the Thr66/Thr71 pair and phosphorylation of Thr64 may be mutually exclusive. Curiously, oxidation of Met42 on TSP9 was also observed in high abundance in the chloroplast sample, although never in combination with phosphorylation of Thr64 or Thr66/Thr71.

Across the chloroplast sample, we found a total of 1540 mass shifts (Table S2). Over 76.6% of identified proteoforms include at least one mass shift. As with total leaf sample, the most prevalent mass shifts match with common PTMs such as oxidation (208 proteoforms), acetylation (80 proteoforms), Met excision (64 proteoforms), and methylation (11 proteoforms) (**Figures 2.4 B** and **2.5 B**). In addition, we also found that the relative ratio of oxidation is substantially higher in the chloroplast sample (5.6% of proteoforms vs. 13.5% of proteoforms, respectively), which we suggest to be physiologically relevant as a reflection of the high oxidative pressure found in the chloroplast [43-45]. In support of physiologically relevant oxidation in the chloroplast, we identified (di)oxidation of tryptophan (+ 15.99 and + 31.99 Da) on six different chloroplast proteins: Photosystem I reaction center subunit N (PsaN; at5g64040), CP12 (at2g47400), CP12-like (at3g62410), Photosystem II light harvesting complex protein (LHCII-1.5; at2g34420), Chlorophylll a/b binding protein 3 (CAB3; at1g29910), and RbcL. Trp dioxidation of Executer 1 and Executer 2 of the chloroplast thylakoid has previously been shown to function as a specific sensor of oxidative stress through reaction with singlet oxygen, triggering retrograde signaling [46]. Our identification of Trp dioxidation modifications on several additional proteins may indicate a broader suite of singlet oxygen sensors than was previously recognized. Curiously, in addition to the six chloroplast proteins, a single non-chloroplast protein, Pathogenesis-related 5 (PR5; at1g75040), also had (di)oxidation found on Trp 37 in the total leaf sample.

2.3.3 Predicting electrophoretic mobility with CZE-MS/MS

As an open tubular configuration, CZE outperforms RPLC in the accurate prediction of proteoform separation times based on electrophoretic mobility (μ_{ef}) [47]. The semi-empirical prediction model of protein μ_{ef} has been modified and evaluated for the large-scale CZE-MS/MS-based proteomics and has been discussed at length previously [29].

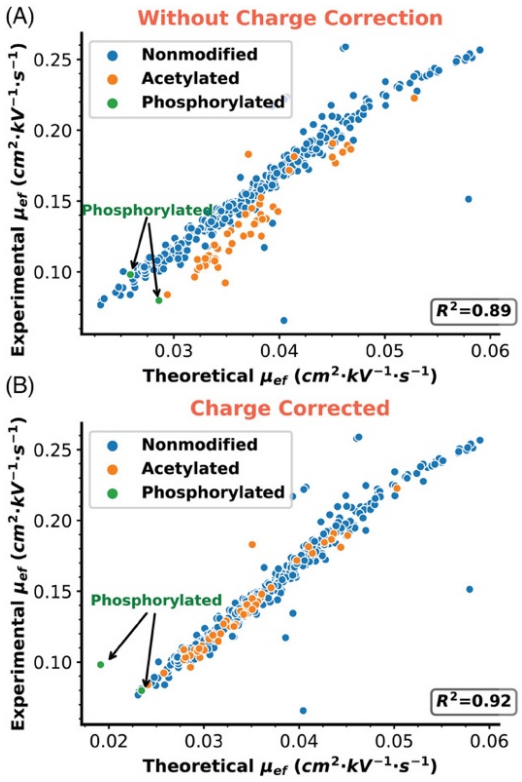


Figure 2.6. Proteoform mass shift distribution from –100 Da to +100 Da. Common PTMs are marked according to the color code indicated in the legend. (A). A histogram of mass shifts among proteoforms identified in the total leaf sample. Two bins containing prevalent mass shifts which are out of the indicated range are included as subsets in the bottom left, both of which are specific to RbcL proteoforms. Because the underlying predicted N-terminus of the proteoforms are different, both mass shifts result in RbcL proteoforms that initiate with an N-terminally acetylated Pro3 residue, as described in the text. (B). A histogram of mass shifts among proteoforms identified in the chloroplast sample. The bin size of all histograms is 1 Da. Possible interpretations of the three most prevalent unidentified mass shifts are proposed based on the Unimod database.

Accurate prediction of retention/migration times can assist in correctly identifying proteoforms and corresponding mass shifts. To explore the prediction of proteoform mobility within the context of our total leaf and chloroplast samples, we applied our prediction model to proteoforms without modifications, or with only one acetylation or phosphorylation, using

proteoform identifications from the second run of total leaf fraction 6, which has the highest number of proteoforms.

While predicted and experimentally determined μ_{ef} values of unmodified proteoforms aligned excellently, the phosphorylated and single N-terminal/lysine-acetylated proteoforms deviated from expectation, as seen in **Figure 2.6 A**. Acetylation removes the positive charge on the N-terminus or on the lysine side chain, while phosphorylation adds a single negative charge. After accounting for the (–1) charge reduction of these PTMs, we found that most corrected proteoforms aligned well with the trend line (**Figure 2.6 B**), and the R^2 increased to 0.92 from 0.89. The R^2 value for non-modified proteoforms alone is 0.91, showing that the modified charge proteoforms match the linear correlation well. The several remaining outliers may represent incomplete unfolding in the 10% acetic acid or incorrect proteoform IDs. The well improved linear correlation between experimental and predicted μef of proteoforms after charge correction highlights the value of CZE-MS/MS for confident proteoform identification and accurate characterization.

We further looked specifically at the identified RbcL proteoforms. The predicted and experimental μ_{ef} of all 43 proteoforms of RbcL with mass shifts are shown in **Figure 2.7 A**. There are five non-modified proteoforms, 22 proteoforms with single acetylation, 13 proteoforms with a – 176 Da mass shift, one proteoform with a – 45 Da mass shift, one proteoform with a + 37.9 Da mass shift, and one proteoforms with a – 2 Da mass shift. The linear correlation between experimental and predicted μef is poor ($R^2 = 0.69$) due to the PTMs of proteoforms, which relate to the mass shifts. The mass shifts (i.e., – 176 Da and – 45 Da) are difficult to explain. However, after –1 and –2 charge corrections for RbcL proteoforms with mass shifts, as highlighted in **Figure 2.7 B**, the linear correlation was drastically improved ($R^2 = 0.99$). The data suggest that those mass shifts reduced the positive charges of proteoforms significantly during CZE separation. Considering the positive charge reduction from the – 176 Da mass shift, we attributed the mass shift to the loss of Met1 (– 131 Da) and Ser2 (– 87 Da) amino acid residues plus an N-terminal acetylation on Pro3 (+ 42 Da). Similarly, we speculated that the – 45 Da mass shift was due to the removal of Ser1 residue and N-terminal acetylation on Pro2 residue. Four proteoforms with –2 charge reduction most likely had a combination of multiple PTMs that reduced the charge of proteoforms. Three of these four proteoforms had a – 175 Da mass shift and one N-terminal acetylation. We expect that those proteoforms have loss of the first

two amino acid residues as mentioned before (Met1, – 131 Da, and Ser2, – 87 Da), and two acetylation sites including the identified N-terminal acetylation. The results further demonstrate that CZE-MS/MS has the capability for accurate characterization of proteoforms with PTMs.

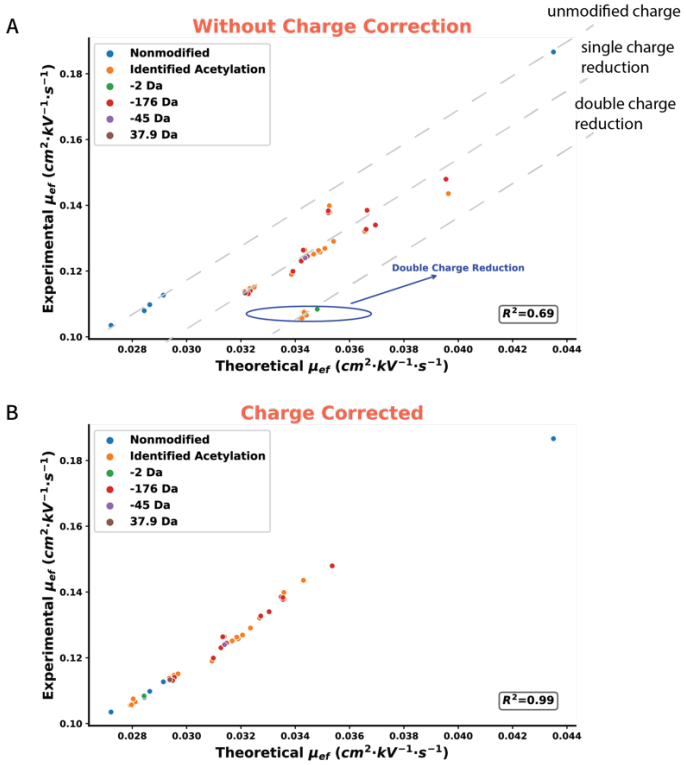


Figure 2.7. The linear correlation between predicted electrophoretic mobility (μ_{ef}) and experimental electrophoretic mobility (μ_{ef}) of rubisco large subunit, RbcL [ATCG00490] from the second run of fraction 6 in the total leaf sample, before charge correction (A) and after charge correction (B). Nonmodified proteoforms with zero mass shifts were labeled in blue and the proteoforms with only one acetylation were labeled in orange. Proteoforms with – 176 Da mass shift are labeled in black. (A), the theoretical electrophoretic mobility is calculated by the positive charges (counts of positively charged residue K, R, H and N-term) and mass of each proteoform; (B), the charge was adjusted by applying (-1) or (-2) to the charge state of each proteoform that deviated from expectation of an unmodified proteoform, depending on the extent of deviation. Note that experimental proteoform migration deviated further from expectation when two charge reduction PTMs were present. The required charge reduction implies the presence of one or two acetylation or phosphorylation PTMs.

2.3.4 Identification of processed protein sequences and truncation pattern

The identification of intact proteoforms offers a prime opportunity to establish the processed N-termini of mature protein sequences, including putative cleavage sites of sub-cellular localization signals such as cTPs or mTPs. Taking advantage of the proteoform data generated from our total leaf and chloroplast samples, we proposed mature N-termini for 343

proteins (proteoforms of an additional 216 proteins were clearly limited to internal fragments of the full protein and hence N-termini could not be determined). Determination was performed manually, relying on frequency of N-termini among proteoforms, relative abundance, and coincident N-terminal acetylation. Of the 343 proteins for which mature N-termini were proposed, 253 were consistent with the prediction from TargetP 2.0. including predicted cleavage sites of 65 cTPs, 9 mTPs, 20 SPs, and 16 luTPs. Of those that were inconsistent, most were cTPs (40), or SPs (30). Significantly, the confidence values of TargetP predictions that were inconsistent with the experimental evidence were, on average, almost as high as those that matched with experimental results (88% vs. 94%, respectively). This indicates that the measure of confidence of TargetP 2.0 may not provide a reliable indication of incorrect predictions. Below, we consider results from each class of sub-cellular localization signal separately, highlighting representative proteins in each case.

2.3.4.1 Chloroplast transit peptides

Our proteoform identifications are rich in nuclear-encoded chloroplast proteins and allow us to propose cTP cleavage sites for 105 proteins. Localization of experimental and predicted cTP sites were consistent in a majority of cases, with 45.4% of proteoforms precisely matched with the predicted cleavage site of cTPs (**Figure 2.8**). In one case, that of the cold-regulated protein 15a (AtCor15a; at2g42540), we observed proteoforms starting sequentially from residue 38 to residue 43 in both total leaf and chloroplast samples (**Figure 2.9**). While the N-termini of the most abundant proteoform was consistent with the predicted cleavage site (i.e., residue 38), the sequential coverage of five residues likely indicates imprecise cleavage of this cTP.

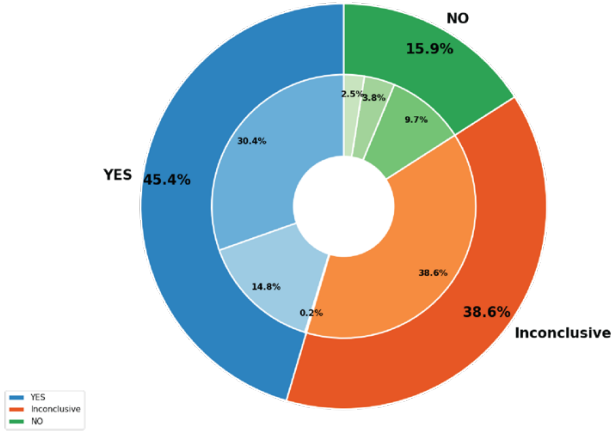


Figure 2.8. Consistency between experimental and predicted cleavage sites for sub-cellular localization signals.

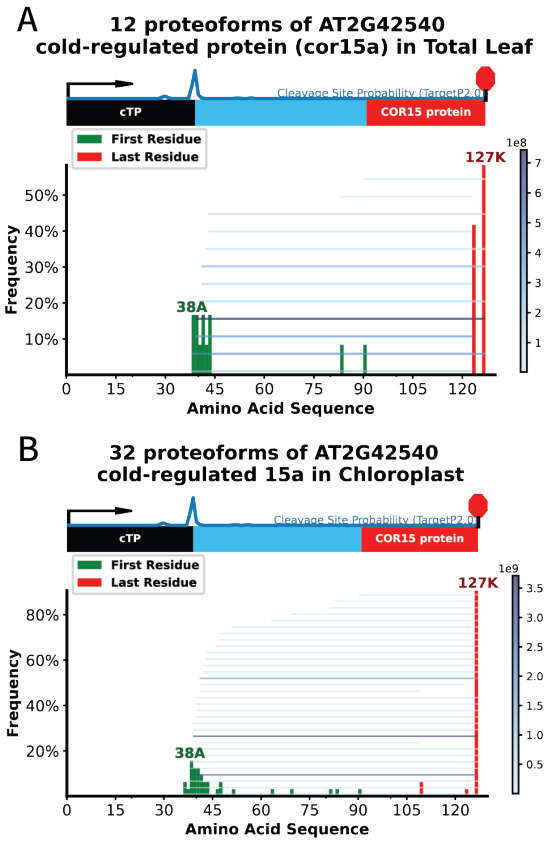


Figure 2.9. Evidence of imprecise cTP cleavage. Multiple initiating residues are found in both total leaf and chloroplast samples surrounding the predicted cleavage site. The frequency of initiating residues is indicated with green bars. The blue curve on top of the protein cartoon is the predicted probability of the TargetP 2.0 cleavage site, lying at 38A - 39A. The regions of predicted chloroplast transit peptide and the functional domain (predicted by HMMPFAM) is highlighted in the sequence. The gradient color in the histogram is normalized based on feature intensity of each proteoform, as indicated in the scale to the right.

In 11 cases we suggest a corrected cTP cleavage site. For example, LHCII-CP26 (at4g10340) was predicted to have a 25% possibility of cleaving between 50K-51A and a 21% possibility between 36V-37A by TargetP 2.0 (**Figure 2.10**). In contrast, we identified 11 out of 15 proteoforms starting from residue 38L in total leaf sample (CP26 was not identified in chloroplast), comprising over 90% of total proteoform intensity. Based on these experimental results we suggest the cTP cleavage site of CP26 is, in fact, 37A-38L. Likewise, a cTP cleavage site is predicted for Rubredoxin A (RubA; at1g54500) at residue 59. However, proteoform evidence from both total leaf and chloroplast samples suggest the mature protein sequence begins at residue 55. This is supported by the identification of multiple abundant proteoforms

beginning at residue 55 in both total leaf (with N-acetylation) and chloroplast samples, as well as the absence of any proteoform beginning at residue 59.

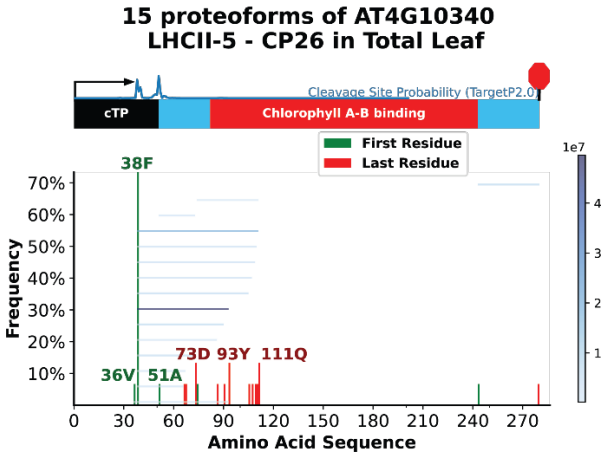


Figure 2.10. Proteoform alignment corrects the cTP cleavage site of LHCII-CP26. Experimental results suggest the 37A - 38F as cTP cleavage site, in contrast to the relatively weak and conflicting predicted sites at 36V-37A and 50K - 51A. The blue curve on top of the protein cartoon diagram is the TargetP 2.0 predicted probability of cleavage site. The regions of predicted chloroplast transit peptide and functional domain (predicted by HMMPFAM) are highlighted in the sequence. The gradient color in the histogram is normalized based on the feature intensity of each proteoform, as indicated in the scale to the right.

Several of our corrected cTP cleavage sites are consistent with results from other studies that rely on orthogonal (i.e., non-TDP-based) experimental methods. Based on proteoform identifications we determined cleavage sites for CP29 (at5g01530) and PsbS (at1g44575) at 31T-32A and 53L–54F, respectively. These two results are consistent with a previous TDP study [30]. Similarly, Heat Shock Protein of 70 kDa (HSP70; at4g24280) was predicted to start at 93A in TargetP 1.0 and updated to 69T in TargetP2.0. In total leaf, we identified three proteoforms, all of which began at residue 78E, indicating the cTP cleavage site lies at 77N-78E, as previously reported from the TAILS experiment [7].

Proteoform patterns of the two chloroplast ferredoxin isoforms (Ferredoxin-1; at1g10960, and Ferrredoxin-2; at1g60950) represent unusual cases (**Figure 2.11**). Both proteins hold a predicted cTP cleavage site at residue 52 M-53A (the same residue and position in both isoforms). Consistent with this, proteoforms identified from the chloroplast sample routinely begin at residue 53A. But surprisingly, these proteoforms represent a miniscule proportion of all proteoforms identified from the total leaf sample, even though the chloroplast-localized, and

hence processed, proteins should be highly represented in these samples. Instead, a majority of proteoforms, both in prevalence and in relative abundance, start at residue 78D (the same residue and position in both isoforms). This places the starting residue well within the 2Fe-2S Ferredoxin-type iron-sulfur binding domain, indicating that a substantial proportion of the domain is not present within the proteoforms and that they are not functional. As the chloroplast sample is dominated with proteoforms initiating at 53A, consistent with the TargetP 2.0 prediction, we conclude that the true cTP cleavage site of both Ferredoxin isoforms is 53A, however it remains unclear why a majority of both isoforms are processed precisely to the 78D residue specifically in the total leaf sample.

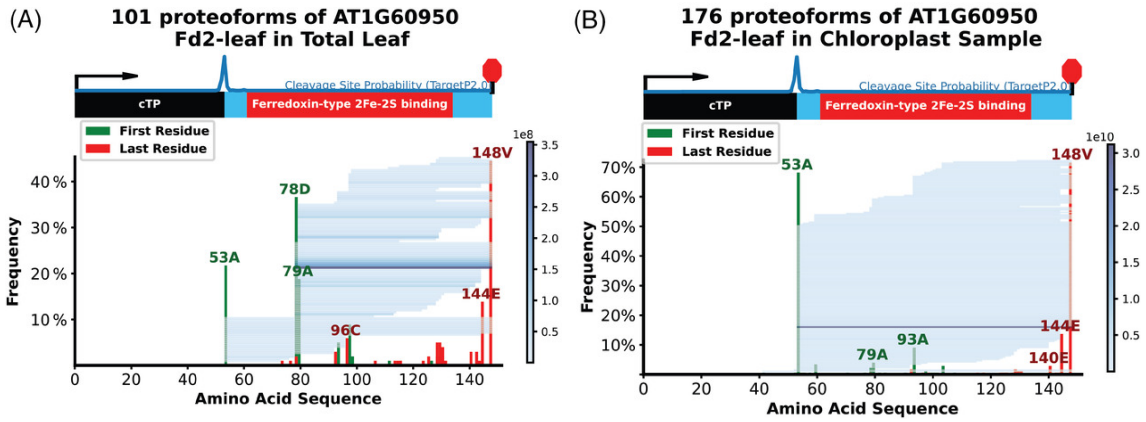


Figure 2.11. Proteoform identifications of Ferredoxin-2 in total leaf (A) and chloroplast (B) samples. Proteoforms identified in the chloroplast sample support cTP cleavage at residues 52–53, while an additional, frequent cleavage site at residues 77–78 is seen specifically in the total leaf sample. A cartoon diagram of the protein sequence domains overlays, and is aligned to, the proteoform identifications indicated as blue horizontal lines shaded according to their estimated abundance (based on ion intensity). The blue trace above the cartoon represents the TargetP 2.0 predicted probability of the cTP cleavage site for each peptide bond. A histogram of green and red bins, overlaying the identified proteoforms, indicates the frequency with which each residue represents either the N-terminal residue (green) or the C-terminal residue (red) among all identified proteoforms. In total leaf sample, the dominant proteoform is 79A-148S, with a mass shift of +23.9 Da. This proteoform is not identified in the chloroplast sample. In the chloroplast sample, the dominant proteoform is 53A-148S with a mass shift of –39.4 Da. This proteoform is identified in the total leaf sample as well but is only 0.28% the intensity of the highest abundant proteoform in the total leaf sample.

2.3.4.2 Lumenal transit peptides

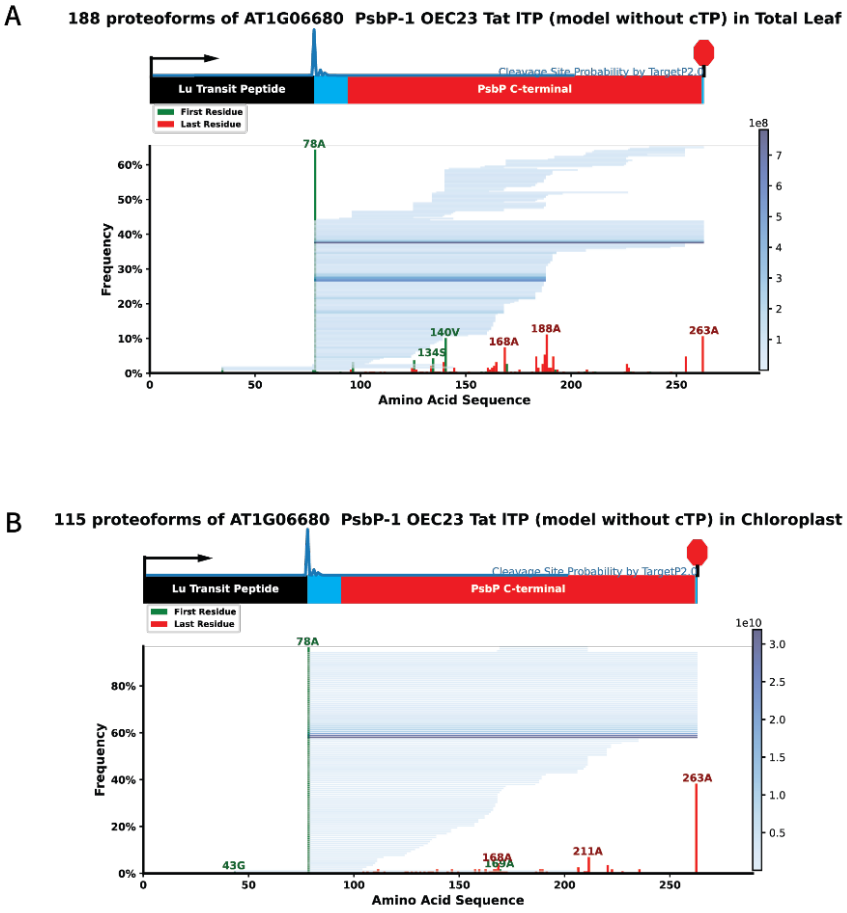


Figure 2.12. Proteoform identification supports lumenal transit peptide (luTP) cleavage site prediction and identifies possible intermediates with cleaved chloroplast transit peptide (cTP). All identified proteoforms of PsbP-1 OEC23 [AT1G06680] in total leaf (A) and chloroplast (B) and the predicted cleavage site from TargetP2.0. Experimental results strongly support the predicted luTP cleavage at 77A - 78A. Identification of three proteoforms initiating at 34T in total leaf sample may represent protein with cTP removed but luTP still intact. The blue curve on top of the protein sequence structure is the predicted TargetP 2.0 probability of luTP cleavage site. And the regions of predicted lumenal transit peptide and function domain (predicted by HMMPFAM) are highlighted in the sequence. The gradient color in the histogram is normalized based on feature intensity of each proteoform, as indicated in the scale to the right.

Remarkably, among predicted luTPs, all were consistent with our experimental determinations. For example, proteoforms of PsbP-1 predominantly began at 78A, consistent with the predicted luTP cleavage site at 77A-78A (**Figure 2.12**). Interestingly, three lower abundant proteoforms of PsbP-1 were identified that start from 34T, which may represent proteins that have had their cTP processed but still await transport into the lumen and subsequent removal of the luTP.

2.3.4.3 Signal peptides

Predicted SP cleavage sites were consistent with our experimental determinations for 40.0% of proteins. Remarkably, 14 SP predicted proteins appeared not to have processed sub-cellular localization signals of any sort, instead accumulating proteoforms starting at Met1 or residue 2. In fact, eight of the 14 proteins were found acetylated on the N-terminus. For example, the fragmentation pattern of a single proteoform of Transmembrane Protein (at1g11850) was found with almost full sequence and initiating from the Met1 residue (**Figure 2.13**).

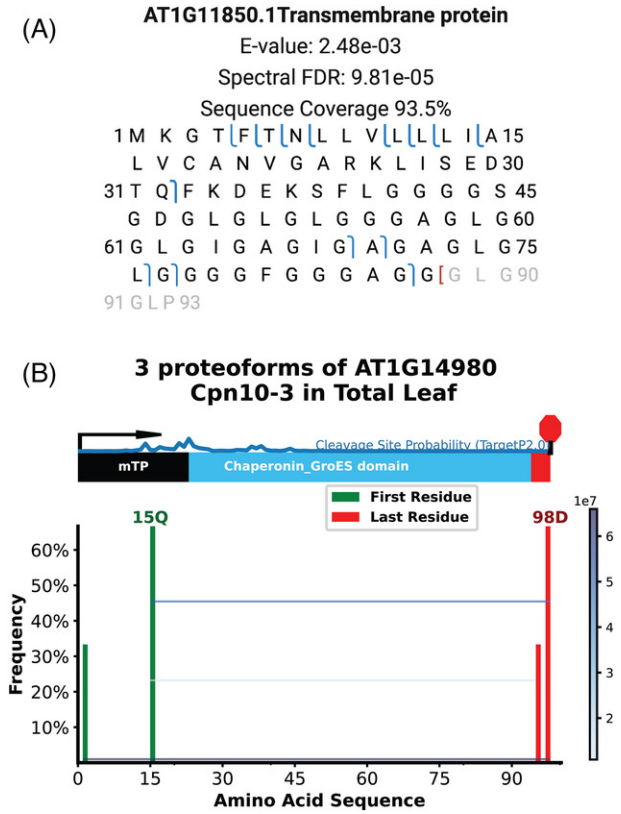


Figure 2.13. Suggested corrections to the predicted sub-cellular localization signals of two proteins. (A) Transmembrane Protein (at1g11850) is predicted to use a Signal Peptide cleaved at 23A-24R. However, the only proteoform identified from this protein reveals a sequence initiating with Met1, indicating the protein does not harbor a cleavable Signal Peptide. The fragmentation pattern from the identified proteoform, along with e-value and FDR are indicated. (B) Chaperonin 10-3 (Cpn10-3; at1g14980) is predicted to use a mitochondrial targeting peptide cleaved at 22K-23T. However, proteoform identifications indicate the protein cleaves at 14V-15Q. A cartoon diagram of the protein sequence domains overlays, and is aligned to, the proteoform identifications indicated as blue horizontal lines shaded according to their estimated abundance (based on ion intensity). The blue trace above the cartoon represents the TargetP 2.0 predicted probability of the cTP cleavage site for each peptide bond. A histogram of green and red bins, overlaying the identified proteoforms, indicates the frequency with which each residue represents either the N-terminal residue (green) or the C-terminal residue (red) among all identified proteoforms.

In four cases we could propose a corrected SP cleavage site. The SP of a protein of unknown function (at3g07470) is predicted to cleave at 24A-25I. However, a single proteoform, beginning with residue 14 V and continuing to the final encoded residue, was identified. Similarly, the SP of the TSK-associating protein (at1g52410) is predicted to cleave at 29C-30Q. In contrast, the most abundant proteoforms were found to begin at 21L. Furthermore, only a single proteoform at much lower abundance was found to begin at 30Q.

2.3.4.4 Mitochondrial transit peptides

Our experimentally concluded mTP cleavage sites coincided with prediction in 9 out of 18 (50%) proteins. We identified three proteoforms of Cpn10-3 (at1g14980) in the total leaf, which all conflicted with the predicted cleavage site at 22K-23T (**Figure 2.13B**). Our results suggest that the mTP cleavage site for this protein is 14V-15Q. While the highest abundant proteoform begins at residue Met1 (and is evidently not imported into the mitochondria), the two remaining proteoforms both initiate at residue 15Q.

Significantly, we identified likely mTP cleavage sites on three other proteins that are not currently predicted to contain mTPs (or any other sub-cellular localization signal) in TargetP 2.0. To conclude a mitochondrial localization for these proteins we relied on the SUBA4 database, which compiles a consensus localization based on disparate experimental and predictive datasets, including MS-based proteomics, fluorescent protein tagging experiments, co-expression data, and 22 computational prediction algorithms [48]. According to SUBA4, the three proteins (Voltage Dependent Ion Channel 3 [VDAC3, at5g15090], Caspase 6 [CASP6, at2g15000], and D-Tyr-tRNA Deacylase family protein [YtDA, at4g18460]) are all strongly expected to localize in the mitochondria. Consistent with this notion, a single proteoform was identified from each of the proteins, each consistent with a cleavage site ranging from residue 33 to 65. Significantly, no proteoform was identified initiating at residue Met1, as would be expected based on the TargetP 2.0 prediction. The identified proteoforms from VDAC3, CASP6, and YtDA began at residues 35S, 65P, and 65D, respectively, directly presenting putative mTP cleavage sites.

2.3.5 Residue frequency of cleavage sites

We plotted residue site occupancy around the updated cleavage sites for cTP, luTP, mTP, and SP sequences. WebLogos representing the absolute frequency of residues at each position relative to the cleavage site reveal a weak preference for Ala in the –1 position (relative to the cleavage site) in cTP, SP, and luTP sequences (**Figure 2.14 A-D**). Conversely, mTPs displayed a

somewhat stronger preference for Phe in the –1 position, as well as Ser in the +1 position and a clear preference for Arg in the –3 position. These patterns were drawn out more clearly when presented as an iceLogo which calculates a residue probability in each position by normalizing the absolute frequency of a residue by its frequency throughout all A. thaliana protein sequences **(Figure 2.14 E-H)**. Site occupancies of the sub-cellular localization signals are consistent with those reported previously [49, 7].

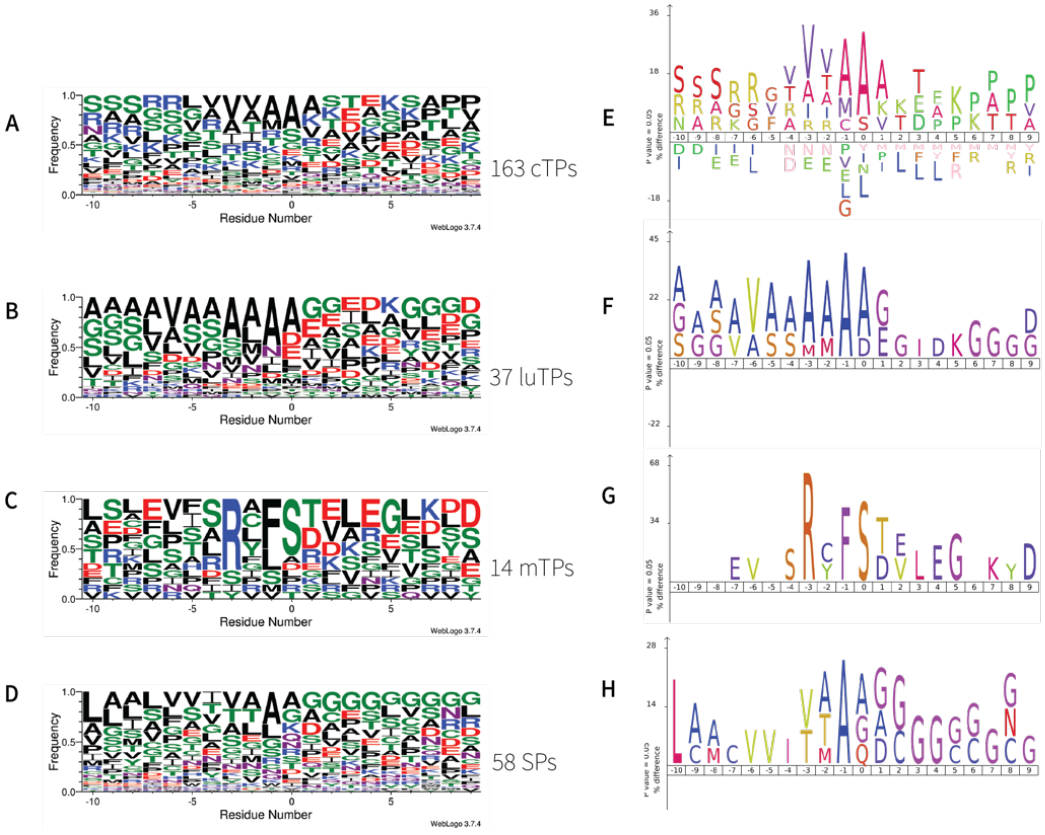


Figure 2.14. Sequence logo comparing the amino acid frequency of cleavage sites for chloroplast transit peptide (cTP), lumenal transit peptide (luTP), mitochondrial transit peptide (mTP), and signal peptide (SP), as determined from proteoform identifications in the total leaf sample. In the middle, the number of unique proteoform identifications used to generate the iceLogo is indicated. The WebLogo on the left (A), (B), (C), and (D) compares the frequency percentage of an amino acid at a certain location in the multiple sequence alignment. The iceLogo on the right (E), (F), (G), (H) compares the probability value of an amino acid at a certain location by normalizing for residue frequency in the A. thaliana proteome. The sample sequences were selected from the total leaf sample with the first residue starting at 15-87 position, and any duplicated sequences were removed.

2.4 Discussion

Methodological and technical advancements in the past 10 years have greatly expanded the capabilities of TDP, including more powerful MS/MS search algorithms and increased

resolution of mass analyzers [30, 32, 50]. However, these advancements have not, thus far, been applied to large-scale studies in plants. It was the objective of this study to exploit and exhibit the capabilities of TDP in characterizing the proteoforms of A. thaliana leaf tissue, with a particular focus on the chloroplast. Using CZE-MS/MS analysis and offline pre-fractionation by SEC, we identified over 4700 unique proteoforms across total leaf and chloroplast samples. This included a substantial number of proteoforms in each sample that contain mass shifts, arising from PTMs or sequence differences relative to the reference protein database. While most mass shifts could not be confidently associated with common PTMs, 683 and 306 mass shifts in total leaf and chloroplast samples were assigned to common PTMs, such as phosphorylation, oxidation, and acetylation (**Figure 2.4)**. Some assignments were based on manual curation of the data, comparing mass shift values with monoisotopic masses and literature reports. Identification of these common PTMs was based on the MIscore [32, 51] as well as electrophoretic mobility, providing robust confidence in the identifications. Among the identified PTMs was Trp oxidation (+15.99 Da) or dioxidation (+31.99 Da) on seven proteins, six of which are chloroplast localized. This somewhat lesser known PTM has recently been found to arise through reaction of singlet oxygen with the Trp side chain and is a crucial component of singlet oxygen retrograde signaling [52, 46]. Trp oxidation of multiple proteins in ROS-producing mitochondria has similarly been reported to function in retrograde signaling [53]. The functional role (if any) of the Trp (di)oxidation identified in this study is not clear, however the oxidation disrupts the indole ring of the Trp side chain affecting the physico-chemical properties.

A primary goal of this study was to identify mature N-termini of proteins and, by extension, propose cleavage sites of sub-cellular localization signals. In 35 cases, we could confidently propose cleavage sites inconsistent with prediction from the TargetP 2.0 algorithm. Alignment of cleavage sites, as determined from our datasets, produced residue frequency plots (i.e., WebLogo and iceLogo) that were largely consistent with those reported in other studies and with other experimental methods [49, 7]. Significantly, we did not observe free cTPs. This is consistent with observations from others that turnover of cTPs occurs rapidly following their cleavage after chloroplast import [54, 7].

Among the interesting observations we report from our study, we found that many proteins accumulated mature sequences with multiple N-termini, often varying by a single residue. It is unclear whether this holds functional significance for a given protein, though it

seems likely that it would influence stability of at least some proteins. The multiple N-termini may arise due to imprecise cleavage of processing peptidases that recognize sub-cellular localization signals. Alternatively, and not mutually exclusive, the multiple N-termini may represent evidence of additional processing following cleavage of a sub-cellular localization signal. Indeed, Rowland, et al. conclude from their chloroplast N-terminome study that additional, and yet to be identified, peptidases further process the N-terminus of cTP-cleaved proteins to arrive at a limited set of N-terminal residues in mature protein sequences [7]. Importantly, our proteoform identifications are limited to those less than ca. 30 kD. Identification of larger proteoforms is a well-known challenge of TDP studies that is attributed to the negative effects of larger molecular species on the signal/noise ratio [55-57]. As a molecular species gets larger, its number of possible charge states increases, leading to a dilution of ion intensity across an increasingly larger number of charge state molecules. Identification of larger proteoforms is however possible, and has been accomplished, but generally requires simpler protein mixtures [55]. The development of strategies to handle the mass problem represents one of the greatest opportunities for future improvements in TDP analysis.

2.5 Acknowledgements

This work was supported by a grant from the National Science Foundation to P.K.L (MCB-2034631). L.S. thanks the support from the National Institutes of Health through Grants R01GM125991, R01GM118470, and R01CA247863, and from the National Science Foundation through Grant DBI-1846913 (CAREER Award). We thank Dr. Daoyang Chen (Merck) for assistance with analysis of the electrophoretic mobility shifts, Dr. Xiaojing Shen (Protein Simple) and Tian Xu (MSU) for advice with chromatography methods, and Dr. Carrie Hiser (MSU) for assistance with sample preparation.

REFERENCES

[1] Smith, L. M., Kelleher, N. L., & Consortium for Top Down, P. Proteoform: a single term describing protein complexity. Nat Methods. 2013; 10(3):186-187.

[2] Longoni, P., Douchi, D., Cariti, F., Fucile, G., & Goldschmidt-Clermont, M. Phosphorylation of the Light-Harvesting Complex II Isoform Lhcb2 Is Central to State Transitions. Plant Physiology. 2015. 169(4):2874-2883.

[3] Bouchnak, I., & van Wijk, K. J. N-Degron Pathways in Plastids. Trends Plant Sci. 2019; 24(10):917-926.

[4] Tasaki, T., Sriram, S. M., Park, K. S., & Kwon, Y. T. The N-end rule pathway. Annu Rev Biochem. 2012; 81:261-289.

[5] Almagro Armenteros, J. J., Salvatore, M., Emanuelsson, O., Winther, O., von Heijne, G., Elofsson, A., & Nielsen, H. Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance. 2019; 2(5): e201900429.

[6] Christian, R. W., Hewitt, S. L., Nelson, G., Roalson, E. H., & Dhingra, A. Plastid transit peptides-where do they come from and where do they all belong? Multi-genome and pan-genomic assessment of chloroplast transit peptide evolution. PeerJ. 2020; 8: e9772.

[7] Rowland, E., Kim, J., Bhuiyan, N. H., & van Wijk, K. J. The Arabidopsis Chloroplast Stromal N-Terminome: Complexities of Amino-Terminal Protein Maturation and Stability. Plant Physiology. 2015; 169(3):1881-1896.

[8] Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., & Nakai, K. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007; 35(Web Server issue): W585-587.

[9] Emanuelsson, O., Brunak, S., von Heijne, G., & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2007; 2(4):953-971.

[10] Emanuelsson, O., Nielsen, H., Brunak, S., & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 2000; 300(4):1005-1016.

[11] Emanuelsson, O., & von Heijne, G. Prediction of organellar targeting signals. Biochim Biophys Acta. 2001; 1541(1-2):114-119.

[12] Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., & van Wijk, K. J. Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. PLoS One. 2008; 3(4), e1994.

[13] Gomez, S. M., Nishio, J. N., Faull, K. F., & Whitelegge, J. P. The chloroplast grana proteome defined by intact mass measurements from liquid chromatography mass spectrometry. Mol Cell Proteomics. 2002; 1(1):46-59.

[14] Ryan, C. M., Souda, P., Bassilian, S., Ujwal, R., Zhang, J., Abramson, J., . . . Whitelegge, J. P. Post-translational modifications of integral membrane proteins resolved by top-down Fourier transform mass spectrometry with collisionally activated dissociation. Mol Cell Proteomics. 2010; 9(5):791-803.

[15] Whitelegge, J. P., Zhang, H., Aguilera, R., Taylor, R. M., & Cramer, W. A. Full subunit coverage liquid chromatography electrospray ionization mass spectrometry (LCMS+) of an

oligomeric membrane protein: cytochrome b(6)f complex from spinach and the cyanobacterium Mastigocladus laminosus. Mol Cell Proteomics. 2002; 1(10):816-827.

[16] Granvogl, B., Zoryan, M., Ploscher, M., & Eichacker, L. A. Localization of 13 one-helix integral membrane proteins in photosystem II subcomplexes. Anal Biochem. 2008; 383(2):279-288.

[17] Russell, J. D., Scalf, M., Book, A. J., Ladror, D. T., Vierstra, R. D., Smith, L. M., & Coon, J. J. Characterization and quantification of intact 26S proteasome proteins by real-time measurement of intrinsic fluorescence prior to top-down mass spectrometry. PLoS One. 2013 8(3): e58157.

[18] Lambertz, J., Liauw, P., Whitelegge, J. P., & Nowaczyk, M. M. Mass spectrometry analysis of the photosystem II assembly factor Psb27 revealed variations in its lipid modification. Photosynth Res. 2021;152(3):305-316.

[19] Gomez, S. M., Bil, K. Y., Aguilera, R., Nishio, J. N., Faull, K. F., & Whitelegge, J. P. Transit peptide cleavage sites of integral thylakoid membrane proteins. Mol Cell Proteomics. 2003; 2(10):1068-1085.

[20] Smith, L. M., Thomas, P. M., Shortreed, M. R., Schaffer, L. V., Fellers, R. T., LeDuc, R. D., . . . Kelleher, N. L. A five-level classification system for proteoform identifications. Nat Methods. 2019; 16(10):939-940.

[21] Zabrouskov, V., Giacomelli, L., van Wijk, K. J., & McLafferty, F. W. A new approach for plant proteomics: characterization of chloroplast proteins of Arabidopsis thaliana by top-down mass spectrometry. Mol Cell Proteomics. 2003; 2(12): 1253-1260.

[22] Chen, D., McCool, E. N., Yang, Z., Shen, X., Lubeckyj, R. A., Xu, T., . . . Sun, L. Recent advances (2019-2021) of capillary electrophoresis-mass spectrometry for multilevel proteomics. Mass Spectrom Rev. 2023;42(2):617-642.

[23] Shen, X., Yang, Z., McCool, E. N., Lubeckyj, R. A., Chen, D., & Sun, L. Capillary zone electrophoresis-mass spectrometry for top-down proteomics. Trends Analyt Chem. 2019; 120, 115644-115644.

[24] Sun, L., Zhu, G., Zhang, Z., Mou, S., & Dovichi, N. J. Third-generation electrokinetically pumped sheath-flow nanospray interface with improved stability and sensitivity for automated capillary zone electrophoresis-mass spectrometry analysis of complex proteome digests. J Proteome Res. 2015; 14(5):2312-2321.

[25] Wojcik, R., Dada, O. O., Sadilek, M., & Dovichi, N. J. Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. Rapid Commun Mass Spectrom. 2010; 24(17):2554-2560.

[26] Zhu, G., Sun, L., & Dovichi, N. J. (2016). Thermally-initiated free radical polymerization for reproducible production of stable linear polyacrylamide coated capillaries, and their application to proteomic analysis using capillary zone electrophoresis-mass spectrometry. Talanta. 2016; 146, 839-843.

[27] Lubeckyj, R. A., McCool, E. N., Shen, X., Kou, Q., Liu, X., & Sun, L. Single-Shot Top-Down Proteomics with Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry for Identification of Nearly 600 Escherichia coli Proteoforms. Anal Chem. 2017; 89(22):12059-12067.

[28] McCool, E. N., Lubeckyj, R. A., Shen, X., Chen, D., Kou, Q., Liu, X., & Sun, L. Deep Top-Down Proteomics Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry: Identification of 5700 Proteoforms from the Escherichia coli Proteome. Anal Chem. 2018; 90(9):5529-5533.

[29] Chen, D., Lubeckyj, R. A., Yang, Z., McCool, E. N., Shen, X., Wang, Q., . . . Sun, L. Predicting Electrophoretic Mobility of Proteoforms for Large-Scale Top-Down Proteomics. Anal Chem. 2020; 92(5):3503-3507.

[30] Fellers, R. T., Greer, J. B., Early, B. P., Yu, X., LeDuc, R. D., Kelleher, N. L., & Thomas, P. M. (2015). ProSight Lite: graphical software to analyze top-down mass spectrometry data. Proteomics, 15(7), 1235-1238. doi: 10.1002/pmic.201570050

[31] Greer, J. B., Early, B. P., Durbin, K. R., Patrie, S. M., Thomas, P. M., Kelleher, N. L., . . . Fellers, R. T. ProSight Annotator: Complete control and customization of protein entries in UniProt XML files. Proteomics. 2022; (11-12): e2100209.

[32] Kou, Q., Xun, L., & Liu, X. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics. 2016. 32(22):3495-3497

[33] Smith, L. M., Agar, J. N., Chamot-Rooke, J., Danis, P. O., Ge, Y., Loo, J. A., . . . Consortium for Top-Down, P. The Human Proteoform Project: Defining the human proteome. Sci Adv. 2021; 7(46): eabk0734.

[34] Pfitzner, J. Poiseuille and his law. Anaesthesia. 1976; 31(2):273-275.

[35] Britz-McKibbin, P., & Chen, D. D. Y. Selective Focusing of Catecholamines and Weakly Acidic Compounds by Capillary Electrophoresis Using a Dynamic pH Junction. Analytical Chemistry. 2000; 72(6):1242-1252.

[36] Zhu, G., Sun, L., Yan, X., & Dovichi, N. J. Bottom-Up Proteomics of Escherichia coli Using Dynamic pH Junction Preconcentration and Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry. Analytical Chemistry. 2014; 86(13):6331-6336.

[37] 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. 2019; 30(8):1435-1445.

[38] Creasy, D. M., & Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. Proteomics. 2004; 4(6):1534-1536.

[39] Houtz, R. L., Stults, J. T., Mulligan, R. M., & Tolbert, N. E. Post-translational modifications in the large subunit of ribulose bisphosphate carboxylase/oxygenase. Proc Natl Acad Sci U S A. 1989; 86(6):1855-1859.

[40] Houtz, R. L., Magnani, R., Nayak, N. R., & Dirk, L. M. Co- and post-translational modifications in Rubisco: unanswered questions. J Exp Bot. 2008; 59(7):1635-1645.

[41] Al-Momani, S., Qi, D., Ren, Z., & Jones, A. R. Comparative qualitative phosphoproteomics analysis identifies shared phosphorylation motifs and associated biological processes in evolutionary divergent plants. J Proteomics. 2018; 181:152-159.

[42] Carlberg, I., Hansson, M., Kieselbach, T., Schroder, W. P., Andersson, B., & Vener, A. V. A novel plant protein undergoing light-induced phosphorylation and release from the photosynthetic thylakoid membranes. Proc Natl Acad Sci U S A. 2003; 100(2):757-762.

[43] Pinnola, A., & Bassi, R. Molecular mechanisms involved in plant photoprotection. Biochem Soc Trans. 2018; 46(2):467-482.

[44] Niyogi, K. K. Safety valves for photosynthesis. Curr Opin Plant Biol. 2000; 3(6):455-460.

[45] Li, Z., Wakao, S., Fischer, B. B., & Niyogi, K. K. Sensing and responding to excess light. Annu Rev Plant Biol. 2009; 60(1):239-260.

[46] Dogra, V., Li, M., Singh, S., Li, M., & Kim, C. Oxidative post-translational modification of EXECUTER1 is required for singlet oxygen sensing in plastids. Nature Communications. 2019; 10(1):2834.

[47] Cifuentes, A., & Poppe, H. Simulation and optimization of peptide separation by capillary electrophoresis. J Chromatogr A. 1994; 680(1):321-340

[48] Heazlewood, J. L., Verboom, R. E., Tonti-Filippini, J., Small, I., & Millar, A. H. SUBA: the Arabidopsis Subcellular Database. Nucleic Acids Res. 2007; 35(Database issue): D213-218.

[49] Huang, S., Taylor, N. L., Whelan, J., & Millar, A. H. Refining the Definition of Plant Mitochondrial Presequences through Analysis of Sorting Signals, N-Terminal Modifications, and Cleavage Motifs. Plant Physio. 2009; 150(3):1272-1285.

[50] Wu, Z., Roberts, D. S., Melby, J. A., Wenger, K., Wetzel, M., Gu, Y., . . . Ge, Y. MASH Explorer: A Universal Software Environment for Top-Down Proteomics. J Proteome Res. 2020; 19(9):3867-3876.

[51] Kou, Q., Zhu, B., Wu, S., Ansong, C., Tolic, N., Pasa-Tolic, L., & Liu, X. Characterization of Proteoforms with Unknown Post-translational Modifications Using the MIScore. J Proteome Res. 2016; 15(8):2422-2432.

[52] Dogra, V., & Kim, C. Chloroplast protein homeostasis is coupled with retrograde signaling. Plant Signal Behav. 2019; 14(11):1656037.

[53] Taylor, S. W., Fahy, E., Murray, J., Capaldi, R. A., & Ghosh, S. S. Oxidative post-translational modification of tryptophan residues in cardiac mitochondrial proteins. J Biol Chem. 2003; 278(22):19587-19590.

[54] Richter, S., & Lamppa, G. K. Stromal processing peptidase binds transit peptides and initiates their ATP-dependent turnover in chloroplasts. J Cell Biol. 1999; 147(1):33-44.

[55] Cai, W., Tucholski, T., Chen, B., Alpert, A. J., McIlwain, S., Kohmoto, T., . . . Ge, Y. Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion Chromatography Strategy. Anal Chem. 2017; 89(10):5467-5475.

[56] Chen, B., Brown, K. A., Lin, Z., & Ge, Y. Top-Down Proteomics: Ready for Prime Time? Anal Chem. 2018; 90(1):10-127.

[57] Melby, J. A., Roberts, D. S., Larson, E. J., Brown, K. A., Bayne, E. F., Jin, S., & Ge, Y. Novel Strategies to Address the Challenges in Top-Down Proteomics. J Am Soc Mass Spectrom. 2021; 32(6), 1278-1294.

CHAPTER 3.* Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Top-Down Proteomics of Mouse Brain Integral Membrane Proteins

3.1 Introduction

Integral membrane proteins (IMPs) embedded within the lipid bilayer membranes are fundamental for essential cellular functions as transporters, receptors, channels, and enzymes, and they are also important drug targets [1,2]. While membrane protein-coding genes with transmembrane domains (TMDs) were predicted to make up 26%–36% of the human protein-coding genes, IMPs are underrepresented in most proteomic studies due to their relatively low abundance and high hydrophobicity [3−5]. Although bottom-up proteomics (BUP) has great success in the identification of IMPs by their peptides, their intact pictures are lost [6−10]. Top-down proteomics (TDP) will provide us with a bird's eye view of intact proteoforms from the same gene. Proteoforms from the same gene due to genetic variations, alternative splicing, and post-translational modifications (PTMs) can have different biological functions [11−15]. PTMs of IMPs are related to their stability and functional regulation as well as apoptosis [16,17]. Therefore, TDP analysis of IMPs in a proteoform-specific manner will offer a much better understanding of their biological functions.

Top-down MS characterization of IMPs is a hot research topic, and great effort has been made to study well-purified specific IMPs or IMP complexes [18−25]. Some studies have been performed on coupling offline or online liquid-phase separation methods (e.g., liquid chromatography (LC)) to MS/MS for proteome-scale studies of IMPs [26−30]. The Kelleher group employed a sodium dodecyl sulfate (SDS) buffer for IMP solubilization, followed by gel-eluted liquid fraction entrapment electrophoresis (GELFrEE) fractionation and LC-MS/MS for the identification of IMP proteoforms from human cell lysates [27,28]. Whitelegge et al. utilized high-concentration formic acid (FA) to dissolve IMPs, followed by size-exclusion chromatography (SEC) and reversed-phase LC separations before MS and MS/MS [29]. The Ge group also used a similar protein solubilization and separation procedure for TDP of IMPs purified from human embryonic kidney cells and cardiac tissue lysates, identifying up to nearly 200 IMPs [30].

---

* *This chapter is partially adapted with permission from Wang, Q., Xu, T., Fang, F., Wang, Q., Lundquist, PK., Sun, L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Top-Down Proteomics of Mouse Brain Integral Membrane Proteins. Anal Chem. 2023; 95(34):12590-12594.*

Capillary zone electrophoresis (CZE)-MS/MS has been proven as a valuable tool for TDP of complex biological samples due to its high separation efficiency and high sensitivity for proteoform separation and detection [31−34]. For example, recently our group reported the identification of over 23,000 proteoforms from human cancer cell lysates using CZE-MS/MS, advancing the TDP for global proteoform identifications substantially [33]. CZE-MS/MS for TDP analysis of single human cells has also been reported [34]. CZE for membrane protein analysis started in the 1990s using background electrolytes containing a high concentration of salts or detergents [35,36]. CZE-MS methods have also been developed for BUP of hydrophobic peptides [37,38]. However, to the best of our knowledge, there is no report on coupling CZE to MS/MS online for TDP of IMPs.

Herein, we demonstrate the first example of CZE-MS/MS for the TDP of IMP proteoforms purified from mouse brains. Briefly, the IMPs were extracted from the mouse brain and purified by an alkaline and urea wash following procedures in the literature [39,40]. After chloroform–methanol precipitation, 120 μg of IMPs was dissolved in a buffer containing 30% (v/v) FA and 60% (v/v) methanol, followed by SEC fractionation using a mobile phase (5% FA, 60% methanol, v/v). Each SEC fraction was analyzed by CZE-MS/MS with a BGE containing 30% (v/v) acetic acid (AA) and 30% (v/v) methanol. The proteoform identification was performed via database search against a UniProt mouse database (UP000000589, 55315 entries) using TopPIC software (TOP-down mass spectrometry-based proteoform identification and characterization, versions 1.6.0 and 1.6.2) [41]. The proteoform identification was filtered by a proteoform-level false discovery rate (FDR) of less than 5%. The number of transmembrane domains (TMDs) of identified proteoforms was predicted by Deep TMHMM, which is a deep learning model for transmembrane topology prediction [28]. TMHMM has been widely used to predict the transmembrane helices with a high accuracy since the first version in 2001 [42].

3.2 Experimental section

3.2.1 Materials and chemicals

MS-grade water, methanol (MeOH), formic acid (FA), and acetic acid (AA), sodium phosphate monobasic monohydrate were purchased from Fisher Chemical (Hampton, NH). A.C.S. Grade sodium phosphate dibasic was bought from Jade Scientific, and sodium chloride was bought from ChemPure Brand. HPLC-Grade chloroform, sodium dodecyl sulfate, 3-(Trimethoxysilyl) propyl methacrylate, acrylamide, ammonium persulfate, and ammonium

bicarbonate (NH4AC) were purchased from Sigma-Aldrich (St. Louis, MO). 48-51% hydrofluoric acid (HF) was bought from ACROS Organics (NJ, USA). Bare fused silica capillaries (50-μm i.d., 360-μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ). cOmplete, Mini protease inhibitor cocktail (EASYpacks) was from Roche (Indianapolis, IN).

3.2.2 Preparation of integral membrane proteins (IMPs) from mouse brain

A 2-month-old male mouse (strain BL-6, wild type), which is kindly provided by Professor Yuan Wang's group at the Department of Animal Science, Michigan State University, was used for the collection of the brain. The procedure of IMPs extraction was performed following the previous publication with minor adjustments [39,40]. Briefly, the tissue (~ 0.5 g) was washed with cold phosphate-buffered saline (PBS) three times and further homogenized in 5 mL of high salt buffer (2M NaCl, 1×PBS, pH 7.4 with protease inhibitor cocktail) using FisherbrandTM 150 Homogenizer for 2 minutes and repeated three times. The samples were sonicated on ice for five minutes with a Branson Sonifier 250 (VWR Scientific, Batavia, IL) using 7 output control and 50% duty cycle and repeated three times. Then the lysate was centrifuged at 1500×g for 10 minutes at 4 °C to collect the supernatant. To purify the integral membrane proteins, the solution was centrifuged at 29,700×g for 60 minutes at 4 °C. The pellet was resuspended in 2.5 mL high pH buffer (0.1M Na_2CO_3, pH 11.3 with protease inhibitor cocktail) and incubated on ice for 60 minutes to remove peripheral membrane proteins. The centrifugation was performed under the same conditions. After a repeated wash in 2.5 mL high pH buffer, the pellet was further washed twice with the urea buffer (4 M urea with 1×PBS with protease inhibitor cocktail) and centrifuged at 29,700×g for 60 minutes at 12 °C to diminish the soluble proteins. The pellet was washed with 1×PBS buffer twice, followed by a 10-minute 10000×g centrifugation at 4 °C to remove the urea. The membrane protein pellet was recovered in 2% SDS (1×PBS with protease inhibitor cocktail), aliquoted, and quantified by bicinchoninic acid (BCA) assay. Around 2 mL solution was collected at a concentration of 0.92 mg/mL. Then the aliquot (120 ug per tube) was precipitated with methanol/chloroform precipitation to remove lipids and stored at -80 °C before use.

3.2.3 Size exclusion chromatography (SEC) Separation

The size exclusion chromatography column (4.6 x 300 mm, 3 μm particles, 500 Å pores) from Agilent was used for separation on a 1260 Infinity II HPLC system from Agilent (Santa

Clara, CA). Detection was performed using an ultraviolet-visible detector at a wavelength of 254 nm. After resuspending 120 μg proteins in 20 μL prechilled 90% formic acid (FA), the solution was diluted into 60 μL 3/6/1 FA/methanol/water (v/v) and ultrasonicated on ice for five minutes. The separation was carried out with the 5% FA and 60% methanol (v/v, pH 2.09) mobile phase with 0.25 mL/min under 30 °C column temperature. Four fractions were collected from 8 min to 16 min and dried by speed vacuum.

3.2.4 CZE-MS/MS

A CESI 8000 Plus (Beckman Coulter) was coupled to the Q-Exactive HF Orbitrap (Thermo Fisher Scientific) by EMASS-II (CMP Scientific) interface for CZE-ESI-MS/MS [43]. A 2.2~2.4 kV spray voltage was applied on the 25~30 μm glass spray emitter containing 0.2% (v/v) formic acid and 10% (v/v) methanol sheath liquid. The capillary (1-meter-long, 50 μm i.d, 360 μm o.d) was coated with linear polyacrylamide (LPA) based on the previous protocol for 55 minutes to increase endurance [44−46].The tip of the capillary was etched by HF (48%-51% solution in water) for 90 minutes to decrease the distance between the capillary outlet to the emitter orifice to around 0.4 mm. The BGE is 30% acetic acid (AA) and 30% methanol in water (v/v). For the single-shot sample, one aliquot of the protein precipitation (120 μg) was resolubilized in cold 80% FA and diluted into 40 μL 3/6/1 FA/Methanol/water (v/v). SEC fractions were also resolubilized in cold 80% FA and diluted into 20 μL 3/6/1 FA/Methanol/water (v/v). After applying 5 psi for 29 s, 150 nL samples (7.5% of column volume, average protein concentration 1.4 mg/mL for SEC fractions, and 3 mg/mL for single-shot sample) were loaded into the capillary. Under a 30 kV separation voltage, the CZE separation was carried out under low pressure (0.1 psi for the first 80 minutes, 0.5 psi for 80-100 minutes, 2 psi for 100-110 min, and 10 psi for the last 10 minutes of flushing). The separation current is around 2.5-4.3 μA. After duplicate runs of each fraction, the capillary was cleaned by flushing with 0.5% ammonium hydroxide at 10 psi for 10 minutes, then flushed with water and BGE at the same condition.

For the mass spectrometer parameters, we set up the temperature of the ion transfer tube at 320 °C, the S lens radio frequency (RF) level at 60% and turned on the intact protein mode. The MS/MS experiments were performed using data-dependent acquisition (DDA). MS1 spectra were collected with the following parameters: an m/z range of 600-2500, orbitrap resolution of 120,000 (at m/z of 200), an automatic gain control (AGC) target value of 3e6, a maximum injection time of 100 ms, and a microscan number of 1. The top five most abundant precursor

ions (charge state higher than 5 and intensity threshold 1e4) in full MS spectra were isolated with a window of 2 m/z and fragmented using HCD with Normalized collision energy (NCE) of 20%. The settings for MS2 spectra were a resolution of 60,000 (at m/z 200), an m/z range of 200-2000, a microscan number of 3, an AGC target value of 1e6, and a maximum injection time of 200 ms. The dynamic exclusion was applied with a duration of 30 s, and the exclusion of isotopes was enabled.

3.2.5 Data analysis

The raw files were converted into mzML using MSConvert with peak picking algorithm and then deconvoluted by TOP-down mass spectrometry feature detection (TopFD) and searched by TOP-down mass spectrometry-based proteoform identification and characterization (TopPIC) (https://www.toppic.org/software/toppic/publications.html) [47]. The UniProt database of mouse (UP000000589, 55315 entries) was accessed on Sep 20th, 2022. TopFD was performed in the default settings to generate the deconvoluted msalign files, and TopPIC was performed using the target-decoy approach with a spectrum-level FDR of 1% and a proteoform-level FDR of 5%. For the combined database search of SEC-CZE-MS/MS data, TopPIC version 1.6.2 was used. For others, TopPIC version 1.6.0 was used. The maximum number of mass shifts is two, and the max variable PTM number is three. The mass error tolerance is 10 ppm with a 1.2 Da PrSM cluster error tolerance. For the single-shot experiment, the mass error tolerance is 15 ppm with a 1.2 Da PrSM cluster error tolerance.

3.3 Results and discussions

Single-shot CZE-MS/MS analysis identified 21 proteins and 65 proteoforms from the mouse brain. Of the proteoforms identified, 51 (79%) have at least one TMD predicted by Deep TMHMM. The data demonstrate that our CZE-MS/MS method is efficient for IMP proteoform identification. We also checked the mass shifts of identified proteoforms for potential formylation modifications [+28 Da] due to the high concentration of FA that was used for the solubilization of IMPs before CZE-MS/MS. We did not see any 28-Da mass shifts in the identified proteoforms. We suspect the lack of artifactual formylation is due both to our storage of the FA solution at −20 °C before use and to the rapid dilution after solubilization [48]. We noted that all the identified IMP proteoforms were smaller than 18 kDa. As shown in **Figure 3.1**, the mouse brain IMP proteoforms migrated through the CZE capillary within the time window of 40–90 min. The earlier peaks during 40–50 min correspond to proteoforms smaller than 18 kDa.

The later peaks during 50–90 min are most likely relatively large proteoforms, and there are no clear charge state distributions in those mass spectra under our instrument conditions.

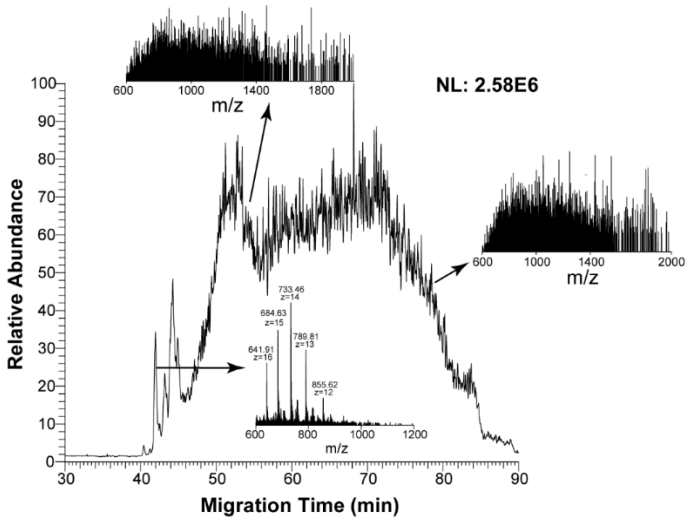


Figure 3.1. Base peak electropherogram of mouse brain integral membrane protein fraction after CZE-MS/MS analysis. Insert figures are mass spectra at three selected migration times.

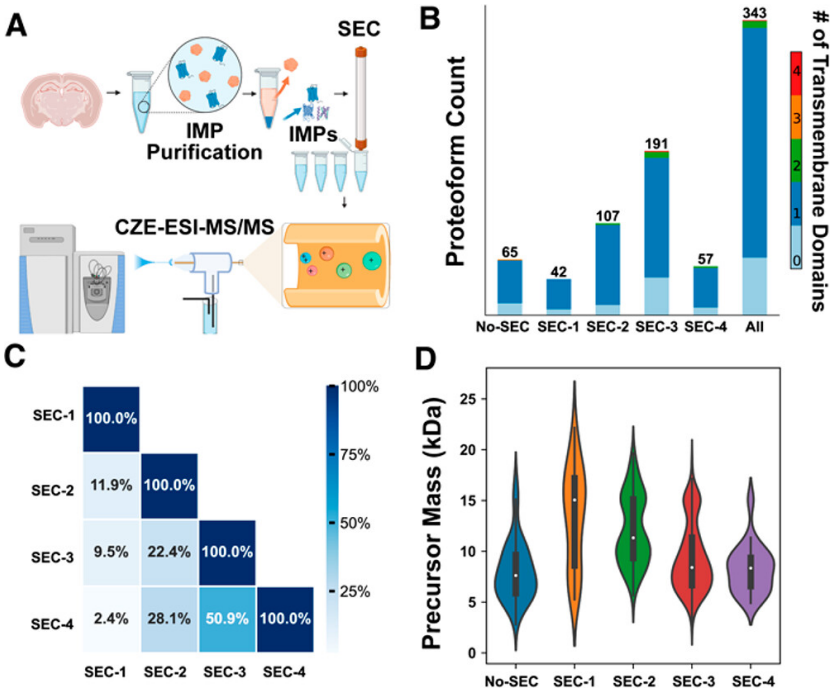


Figure 3.2. Intact integral membrane protein (IMP) analysis by CZE-MS/MS. (A) Flowchart of SEC-CZE-ESI-MS/MS for profiling IMPs of mouse brains, created with Biorender. (B) The identification number of proteoforms from each SEC fraction and the combined result from the four fractions. Different colors represent the number of predicted transmembrane domains (TMDs) based on the proteoform sequence by DeepTMHMM. (C) Heatmap of proteoform overlaps between any two SEC fractions. (D) Violin plots of mass distributions of identified proteoforms from the four SEC fractions and CZE-MS/MS alone (No-SEC).

To improve the identification of large proteoforms, we further fractionated the mouse brain IMP fraction dissolved in 30% (v/v) FA and 60% (v/v) methanol by SEC into four fractions to reduce the sample complexity. The mobile phase of SEC was kept at a high concentration of FA (5%) and methanol (60%) to maintain the solubility of IMPs. Each SEC fraction was analyzed by CZE-MS/MS in a technical duplicate. We identified 42, 107, 191, and 57 proteoforms from SEC fractions 1, 2, 3, and 4, respectively, **Figure 3.2 B**. In total, 65 proteins and 343 proteoforms were identified from the mouse brain sample using SEC-CZE-MS/MS. Out of the identified proteoforms, 276 proteoforms are IMP proteoforms and had 1–4 TMDs, including proteoforms from enzymes (e.g., ATP synthase, Cytochrome c oxidase, and NADH dehydrogenase), channels (i.e., voltage-dependent anion-selective channel protein 1), and receptors (i.e., mitochondrial import receptor). The number of IMP proteoforms from SEC-CZE-MS/MS is improved by more than 5-fold compared to that from single-shot CZE-MS/MS (276 vs 51). We need to highlight that SEC-CZE-MS/MS identified 13 proteoforms higher than 18 kDa and that single-shot CZE-MS/MS did not identify any proteoforms larger than 18 kDa. The data clearly indicate that our SEC-CZE-MS/MS technique could be a useful tool for the proteome-scale characterization of IMP proteoforms. The MS raw files have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD042298 [49]. We noted that most of the identified IMP proteoforms had one TMD, which is consistent across all of the SEC fractions, **Figure 3.2 B**. The phenomenon is most likely due to the relatively small proteoforms identified in this study. As shown in **Figure 3.3**, CZE-MS/MS alone identified proteoforms in the mass range of 3–17 kDa, with the majority in the range of 5–10 kDa. SEC-CZE-MS/MS identified proteoforms in a mass range of 3–22 kDa, producing much more proteoforms larger than 10 kDa compared to CZE-MS/MS alone. Identification of large proteoforms (i.e., >30 kDa) from complex proteomes by TDP is challenging due to their low signal-to-noise ratios caused by wide charge state distributions and due to the limited mass resolution of commonly used mass spectrometers in TDP studies [30,32,33].

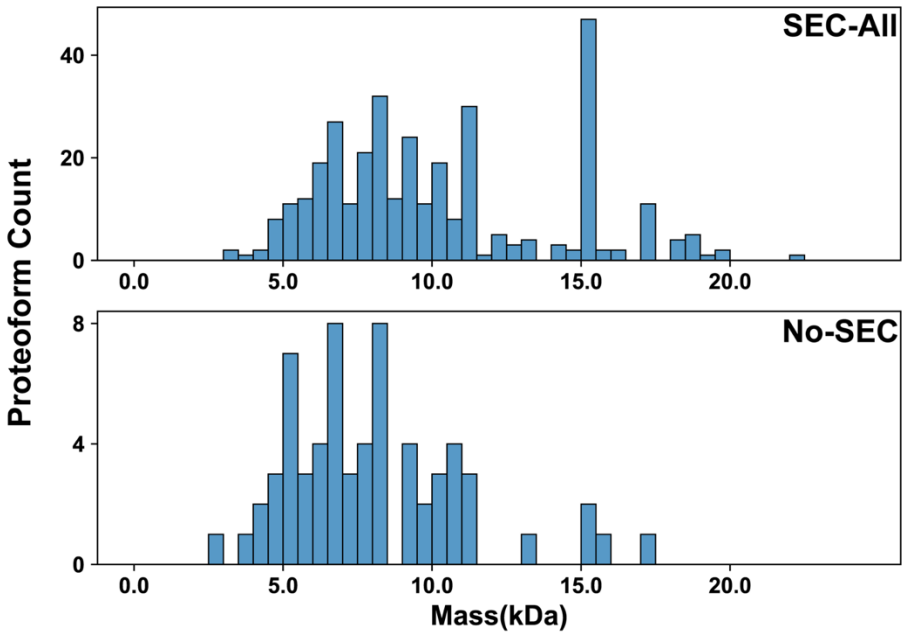


Figure 3.3 Mass distribution of the identified proteoforms from the mouse brain sample using single-shot CZE-MS (No-SEC) and SEC-CZE-MS/MS (SEC-All).

Figure 3.4 shows a Venn diagram about the proteoforms identified by CZE-MS/MS and SEC-CZE-MS/MS. Interestingly, although SEC-CZE-MS/MS identified many more proteoforms than CZE-MS/MS alone (343 vs 65), only less than 50% of proteoforms from CZE-MS/MS alone are covered by SEC-CZE-MS/MS.

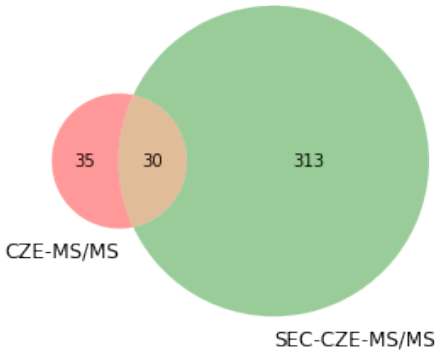


Figure 3.4 The Venn diagram of total identified unique proteoforms from CZE-MS/MS (Single-shot) and SEC-CZE-MS/MS (SEC-All). The percentage of proteoforms containing 1-4 TMDs are 77.1% in the CZE-MS/MS only, 80.0% in the overlapped identifications, and 79.9% in the SEC-CZE-MS/MS only.

We further studied the proteoform overlaps between any two SEC fractions, as shown in **Figure 3.2 C**. The proteoform overlap ranges from 2% to 51% and becomes smaller when the

two fractions have a bigger elution time difference. This suggests that SEC can fractionate the mouse brain IMP proteoforms efficiently. **Figure 3.2 D** shows the violin plots of mass distributions of identified proteoforms from CZE-MS/MS alone (no-SEC) and the four SEC fractions. It is clear that the median proteoform mass gradually decreases from Fraction 1 (SEC-1, ~15 kDa) to fraction 3 (SEC-3, ~8 kDa), indicating a reasonably good separation of SEC for the proteoforms by their size. Fractions 3 and 4 (SEC-3 and SEC-4) have comparable mass distributions, although substantially more proteoforms were identified in F3 compared to F4 (191 vs 57). SEC-CZE-MS/MS improved the identification of relatively large proteoforms compared to that of CZE-MS/MS alone (no-SEC).

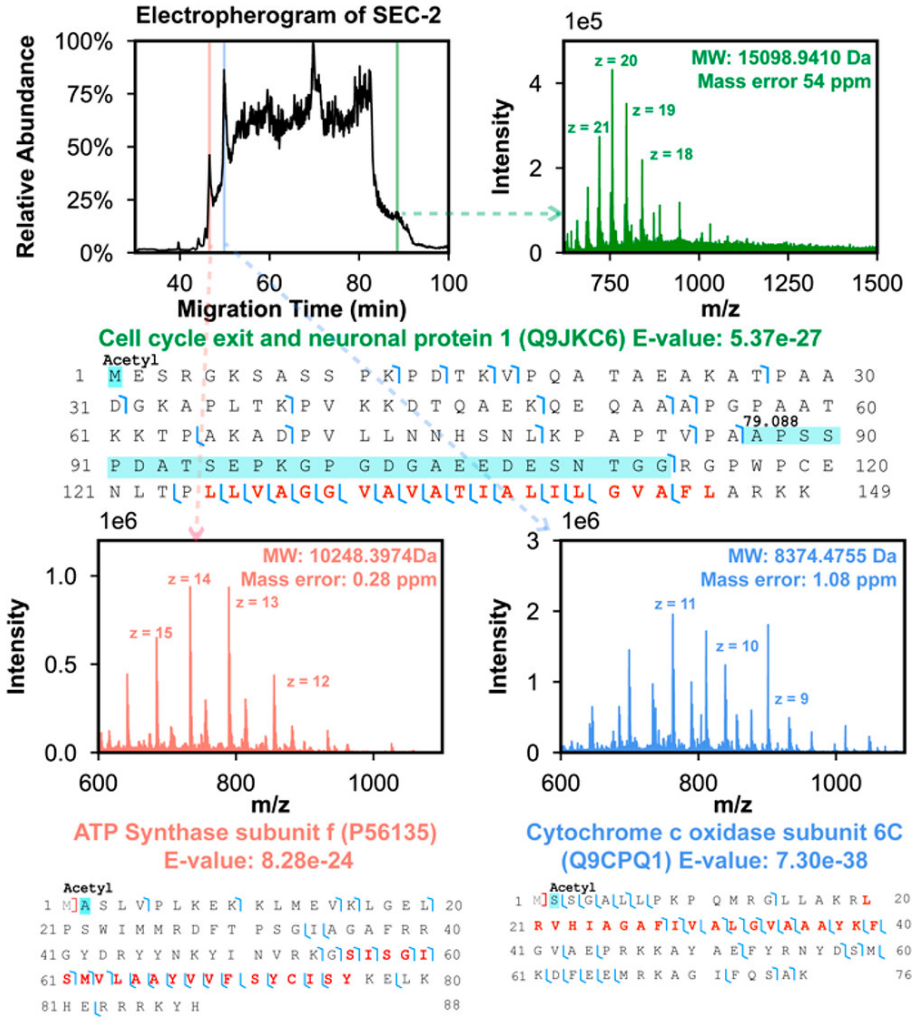


Figure 3.5. Mouse brain IMP data of SEC fraction 2 by CZE-MS/MS. The electropherogram of SEC fraction 2 and mass spectra of three IMP proteoforms identified at three different migration times are presented. The proteoform sequences and fragmentation patterns of the three IMP proteoforms are also shown. The TMD of proteoforms are marked in red in the proteoform sequences. The detected mass shifts and PTMs are labeled.

Figure 3.5 shows the representative electropherogram of CZE-MS/MS analysis of SEC fraction 2 including three examples of IMP proteoforms identified in the CZE-MS/MS run at different migration times. These three proteoforms were identified with low E-values ranging from 8.3×10^{-24} to 7.3×10^{-38}, indicating high-confidence identifications. These three proteoforms all contain one TMD, and good backbone cleavage coverages were achieved for the TMDs using the higher energy collision dissociation (HCD) method. Electropherograms of SEC fractions 1 and 3 as well as some example proteoforms identified in those fractions are shown in **Figures 3.6**.

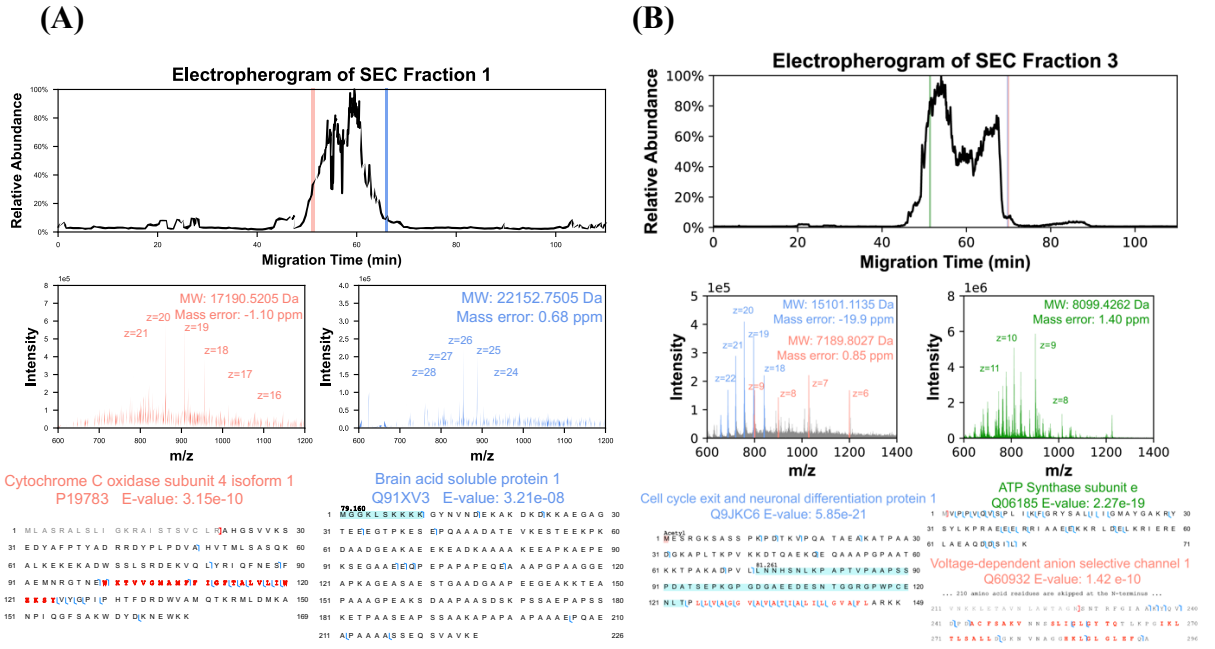


Figure 3.6. The electropherogram of SEC Fraction 1(A) and Fraction 3 (B) of the mouse brain IMPs and two representative proteoforms. The mass spectra and fragmentation patterns of the two proteoforms are shown. The predicted transmembrane domain (TMD) is marked in red in the sequence and the detected mass shift is labeled.

We determined N-terminal methionine removal (ATP synthase subunit e, Figure S5), N-terminal truncation, mitochondrion transit peptide cleavage (Cytochrome c oxidase subunit 4 isoform 1, mitochondrial, Cox4i1, **Figure 3.6 A**), N-terminal acetylation, phosphorylation (cell cycle exit and neuronal differentiation protein 1, Cend1, **Figure 3.6 B**), and myristoylation on the N-terminal glycine residue (Brain acid soluble protein 1, Basp1, **Figure 3.6 A**) on those proteoforms. The Basp1 proteoform has a mass shift of 79.160 Da on the first ten amino acid residues, which matches with the methionine removal and myristoylation on N-terminal glycine [210.198 Da–131.040 Da = 79.159 Da]. The PTM information matches well with that in the

UniProt database (https://www.uniprot.org/uniprotkb/Q91XV3/entry). The N-myristoylation is required for Basp1 as a transcriptional corepressor to remove histone modifications H3K9ac and H3K4me3 [50]. We identified over 60 Cend1 proteoforms, and the majority of them have N-terminal acetylation. We also detected various mass shifts (e.g., 78–81 Da, 122 Da, 159–161, 238, and 478 Da) on the Cend1 proteoforms. Those mass shifts could be explained as single phosphorylation (80 Da), multiple phosphorylation (i.e., 160, 240, and 480 Da), or a combination of phosphorylation and acetylation (i.e., 122 Da). Cend1 has multiple phosphorylation sites according to the UniProt database and PhosphoSitePlus database (https://www.phosphosite.org). Cend1 is a neuronal protein highly expressed in the postnatal mouse brain and modulates cell cycle exit and neuronal differentiation [51]. However, the functions of specific phosphorylated Cend1 proteoforms are still not known. Extracted ion electropherograms of some of the example proteoforms are shown in **Figures 3.7**.

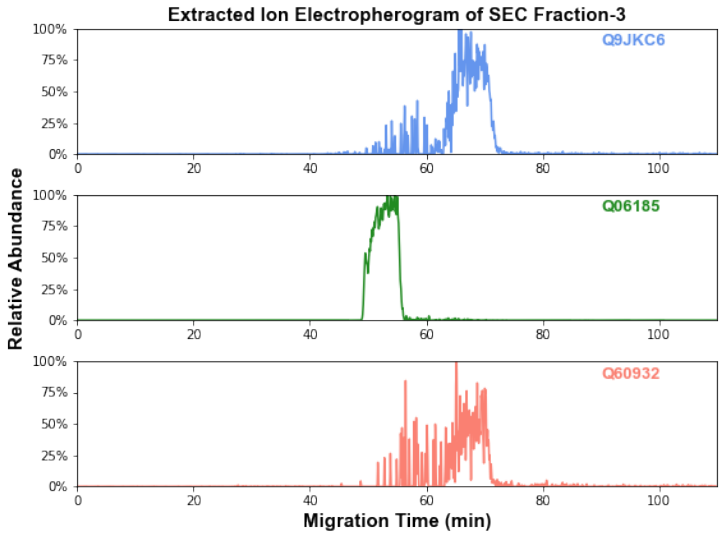


Figure 3.7. The extracted ion electropherograms of the three example proteoforms in SEC Fraction-3 in **Figure 3.6 B**. For peak extraction, m/z 796.26 (+19) was used for the proteoform of Q9CPQ1, m/z 901.50 (+9) was used for the proteoform of Q06185, and m/z 1028.69 (+7) was used for the proteoform of Q60932. The mass tolerance was 5 ppm for peak extraction and Gaussian smoothing (7 points) was applied.

3.4 Conclusions

In summary, for the first time, we developed a CZE-MS/MS technique for TDP of IMPs, and by employing SEC-CZE-MS/MS, we achieved the identification of hundreds of IMP proteoforms from mouse brains. We expect that coupling SEC fractionation with CZE-MS/MS will be a useful tool for large-scale TDP of IMPs. The current limitation of the technique is the

identification of large IMP proteoforms due to the limited mass resolution of the mass spectrometer used in this study. We believe that coupling our SEC-CZE separations to a high-end orbitrap [52], FT ion cyclotron resonance (ICR) [53], or time-of-flight (TOF) [54] mass spectrometers will improve the characterization of large IMP proteoforms substantially.

3.5 Acknowledgments

We thank Prof. Chen Chen's group at the Department of Animal Science, Michigan State University for kindly providing the mouse samples for our experiment. We also thank Prof. Julian Whitelegge at University of California, Los Angeles for all the valuable discussions. The work was funded by the National Cancer Institute (NCI) through the grant R01CA247863 (to LS) and the National Science Foundation through grant MCB-2034631 (to PKL). We thank the support from the National Institute of General Medical Sciences (NIGMS) through grants R01GM125991 and R01GM118470. We also thank the support from the National Science Foundation (CAREER Award, grant DBI1846913).

REFERENCES

[1] Kar UK, Simonian M, Whitelegge JP. Integral membrane proteins: bottom-up, top-down and structural proteomics. Expert Rev Proteomics. 2017; 14:715–723.

[2] Wu CC, Yates JR III. The application of mass spectrometry to membrane proteomics. Nat Biotechnology. 2003; 21:262–267.

[3] Sinitcyn P, Richards AL, Weatheritt RJ, et al. Global detection of human variants and isoforms by deep proteome sequencing. Nat Biotechnology. 2023; 41:1776–1786.

[4] Fagerberg L, Jonasson K, von Heijne G, et al. Prediction of the human membrane proteome. Proteomics. 2010; 10:1141–1149.

[5] Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. Science. 2015;347.

[6] Chu H, Zhao Q, Liu J, et al. Ionic liquid-based extraction system for in-depth analysis of membrane protein complexes. Anal Chem. 2022; 94:758–767.

[7] Zhao Q, Fang F, Shan Y, et al. In-depth proteome coverage by improving efficiency for membrane proteome analysis. Anal Chem. 2017; 89:5179–5185.

[8] Fischer F, Wolters D, Rögner M, et al. Toward the complete membrane proteome. Mol Cell Proteomics. 2006; 5:444–453.

[9] Lee HC, Carroll A, Crossett B, et al. Improving the identification and coverage of plant transmembrane proteins in Medicago using bottom–up proteomics. Front Plant Sci. 2020; 11.

[10] Chen EI, McClatchy D, Park SK, et al. Comparisons of mass spectrometry compatible surfactants for global analysis of the mammalian brain proteome. Anal Chem. 2008; 80:8694–8701.

[11] Smith LM, The Consortium for Top Down Proteomics, Kelleher NL. Proteoform: a single term describing protein complexity. Nat Methods. 2013; 10:186–187.

[12] Smith LM, Kelleher NL. Proteoforms as the next proteomics currency. Science. 2018; 359:1106–1107.

[13] Costa HA, Leitner MG, Sos ML, et al. Discovery and functional characterization of a neomorphic PTEN mutation. Proc Natl Acad Sci U S A. 2015; 112:13976–13981.

[14] Yang X, Coulombe-Huntington J, Kang S, et al. Widespread expansion of protein interaction capabilities by alternative splicing. Cell. 2016; 164:805–817.

[15] Jenuwein T, Allis CD. Translating the histone code. Science. 2001; 293: 1074–1080.

[16] Kerner J, Lee K, Tandler B, et al. VDAC proteomics: Post-translation modifications. Biochim Biophys Acta Biomembr. 2012; 1818:1520–1525.

[17] Kerner J, Lee K, Hoppel CL. Post-translational modifications of mitochondrial outer membrane proteins. Free Radic Res. 2011; 45:16–28.

[18] Keener JE, Zhang G, Marty MT. Native mass spectrometry of membrane proteins. Anal Chem. 2021; 93:583–597.

[19] Barrera NP, Isaacson SC, Zhou M, et al. Mass spectrometry of membrane transporters reveals subunit stoichiometry and interactions. Nat Methods. 2009; 6:585–587.

[20] Yen H-Y, Liko I, Song W, et al. Mass spectrometry captures biased signalling and allosteric modulation of a G-protein-coupled receptor. Nat Chem. 2022; 14:1375–1382.

[21] Keener JE, Zambrano DE, Zhang G, et al. Chemical additives enable native mass spectrometry measurement of membrane protein oligomeric state within intact nanodiscs. J Am Chem Soc. 2019; 141:1054–1061.

[22] Susa AC, Lippens JL, Xia Z, et al. Submicrometer emitter ESI tips for native mass spectrometry of membrane proteins in ionic and nonionic detergents. J Am Soc Mass Spectrom. 2018; 29:203–206.

[23] Harvey SR, O'Neale C, Schey KL, et al. Native mass spectrometry and surface induced dissociation provide insight into the post-translational modifications of tetrameric AQP0 isolated from bovine eye lens. Anal Chem. 2022; 94:1515–1519.

[24] Ro SY, Schachner LF, Koo CW, et al. Native top-down mass spectrometry provides insights into the copper centers of membrane-bound methane monooxygenase. Nat Commun. 2019; 10.

[25] Fantin SM, Parson KF, Niu S, et al. Collision induced unfolding classifies ligands bound to the integral membrane translocator protein. Anal Chem. 2019; 91:15469–15476.

[26] Donnelly DP, Rawlins CM, DeHart CJ, et al. Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. Nat Methods. 2019; 16:587–594.

[27] Catherman AD, Li M, Tran JC, et al. Top down proteomics of human membrane proteins from enriched mitochondrial fractions. Anal Chem. 2013; 85:1880–1888.

[28] Catherman AD, Durbin KR, Ahlf DR, et al. Large-scale top-down proteomics of the human proteome: Membrane proteins, mitochondria, and senescence. Mol Cell Proteomics. 2013; 12:3465–3473.

[29] Whitelegge JP, Zhang H, Aguilera R, et al. Full subunit coverage liquid chromatography electrospray ionization mass spectrometry (LCMS+) of an oligomeric membrane protein. Mol Cell Proteomics. 2002; 1:816–827.

[30] Brown KA, Tucholski T, Alpert AJ, et al. Top-down proteomics of endogenous membrane proteins enabled by cloud point enrichment and multidimensional liquid chromatography–mass spectrometry. Anal Chem. 2020; 92:15726–15735.

[31] Han X, Wang Y, Aslanian A, et al. In-line separation by capillary electrophoresis prior to analysis by top-down mass spectrometry enables sensitive characterization of protein complexes. J Proteome Res. 2014; 13:6078–6086.

[32] Chen D, McCool EN, Yang Z, et al. Recent advances (2019–2021) of capillary electrophoresis-mass spectrometry for multilevel proteomics. Mass Spectrom Rev. 2023; 42:617–642.

[33] McCool EN, Xu T, Chen W, et al. Deep top-down proteomics revealed significant proteoform-level differences between metastatic and nonmetastatic colorectal cancer cells. Sci Adv. 2022; 8.

[34] Johnson KR, Gao Y, Greguš M, et al. On-capillary cell lysis enables top-down proteomic analysis of single mammalian cells by CE-MS/MS. Anal Chem. 2022; 94:14358–14367.

[35] Josić D, Zeilinger K, Reutter W, et al. High-performance capillary electrophoresis of hydrophobic membrane proteins. J Chromatogr A. 1990; 516:89–98.

[36] Nelson BC, Malik S, Seeley SK, et al. Characterization of the transmembrane serine receptor by capillary zone electrophoresis. Chromatographia. 1999; 49:28–34.

[37] Cheng J, Morin GB, Chen DDY. Bottom-up proteomics of envelope proteins extracted from spinach chloroplast via high organic content CE-MS. Electrophoresis. 2020; 41:370–378.

[38] Cheng J, Chen DDY. Nonaqueous capillary electrophoresis mass spectrometry method for determining highly hydrophobic peptides. Electrophoresis. 2018;39:1216–1221.

[39] Lu A, Wiśniewski JR, Mann M. Comparative proteomic profiling of membrane proteins in rat cerebellum, spinal cord, and sciatic nerve. J Proteome Res. 2009; 8:2418–2425.

[40] Zhao Q, Sun L, Liang Y, et al. Prefractionation and separation by C8 stationary phase: Effective strategies for integral membrane proteins analysis. Talanta. 2012; 88:567–572.

[41] Kou Q, Xun L, Liu X. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics. 2016; 32:3495–3497.

[42] Krogh A, Larsson B, von Heijne G, et al. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol. 2001; 305:567–580.

[43] Sun L, Zhu G, Zhang Z, et al. A Third-Generation Electro- Kinetically Pumped Sheath Flow Nanospray Interface with Improved Stability and Sensitivity for Automated Capillary Zone Electrophoresis-Mass Spectrometry analysis of complex proteome digest. J Proteome Res. 2015; 14(5):2312-21.

[44] Zhu G, Sun L, Dovichi NJ. Thermally-initiated free radical polymerization for reproducible production of stable linear polyacrylamide coated capillaries, and their application to proteomic analysis using capillary zone electrophoresis-mass spectrometry. Talanta . 2016; 146:839–843.

[45] Mccool EN, Lubeckyj R, Shen X, et al. Large-Scale Top- down Proteomics Using Capillary Zone Electrophoresis Tandem Mass Spectrometry. J Vis Exp. 2018.

[46] Xu T, Han L, George Thompson AM, et al. An improved capillary isoelectric focusing-mass spectrometry method for high-resolution characterization of monoclonal antibody charge variants. Anal Methods. 2022; 14:383–393.

[47] Kou Q, Xun L, Liu X. TopPIC: A Software Tool for Top-down Mass Spectrometry- Based Proteoform Identification and Characterization. Bioinformatics. 2016;32.

[48] Doucette AA, Vieira DB, Orton DJ, et al. Resolubilization of precipitated intact membrane proteins with cold formic acid for analysis by mass spectrometry. J Proteome Res. 2014; 13:6001–6012.

[49] Perez-Riverol Y, Csordas A, Bai J, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019; 47:D442–D450.

[50] Moorhouse AJ, Loats AE, Medler KF, et al. The BASP1 transcriptional corepressor modifies chromatin through lipid-dependent and lipid-independent mechanisms. iScience. 2022; 25:104796.

[51] Segklia K, Stamatakis A, Stylianopoulou F, et al. Increased anxiety-related behavior, impaired cognitive function and cellular alterations in the brain of Cend1-deficient mice. Front Cell Neurosci. 2019;12.

[52] Wang C, Liang Y, Zhao B, et al. Ethane-bridged hybrid monolithic column with large mesopores for boosting top-down proteomic analysis. Anal Chem. 2022; 94:6172–6179.

[53] Tucholski T, Knott SJ, Chen B, et al. A top-down proteomics platform coupling serial size exclusion chromatography and Fourier transform ion cyclotron resonance mass spectrometry. Anal Chem. 2019; 91:3835–3844.

[54] Fornelli L, Parra J, Hartmer R, et al. Top-down analysis of 30–80 kDa proteins by electron transfer dissociation time-of-flight mass spectrometry. Anal Bioanal Chem. 2013; 405:8505–8514.

CHAPTER 4. A simple and efficient approach for preparing cationic coating with tunable electroosmotic flow for capillary zone electrophoresis-mass spectrometry-based top-down proteomics

4.1 Introduction

Top-down proteomics (TDP) aims to characterize diverse intact proteoforms resulting from alternative splicing, post-translational modifications (PTMs) and proteolytic cleavages in complex biological samples and discover proteoform biomarkers for various disease [1–4].

It has been noted that the effective separation of proteoforms prior to mass spectrometry (MS) is critical to reduce the complexity and capture the lower abundant species [5,6]. Capillary zone electrophoresis-tandem mass spectrometry (CZE-MS/MS) has been more and more recognized as a valuable method for TDP as its low sample consumption, high peak capacity and high sensitivity for various proteoform samples [4,7–10].

Coatings on the capillary inner wall are crucial for CZE-MS/MS-based TDP to maintain high separation efficiency and wide separation windows. Neutral coatings (e.g., linear polyacrylamide, LPA) and positively charged coatings (e.g., polyethyleneimine, PEI) have been utilized in CZE-MS-based analyses of large biomolecules (i.e., proteoforms) [10-14]. The silanol group on the inner wall of the bare fused silica capillary carries negative charge at pH above 2, causing a flat electroosmotic flow (EOF) driven by the electrical double layer forming on the surface [15]. Commonly used linear polyacrylamide (LPA) neutral and hydrophilic coating minimizes the EOF and reduces the static adsorption, thereby enabling a high separation efficiency and wide separation windows [10, 16-18]. However, in our most recent study, we discovered that LPA-coating still has significant protein adsorption, which leads to the reproducibility issue in long-term TDP measurements [19]. There is an urgent need for investigating new capillary coatings to improve the separation reproducibility and separation resolution of proteoforms by CZE. The positively charged coating could generate a countercurrent EOF while analytes are separated by their charge-to-size ratios under a strong electric field [20]. The cationic coating can benefit the separation resolution and prevent the non-specific adsorption of positively charged proteoforms by Coulombic repulsion force [21]. Therefore, positively charged coatings can be valuable alternatives to the LPA coating to further advance CZE-MS/MS-based TDP.

Positively charged coatings can be dynamic or covalent [21-23]. Dynamic coating is a quick implementation and easy to remove, however, the additives in the background electrolyte (BGE) usually has a lower compatibility coupling with MS [23-25]. Permanent coatings and semi-permanent, such us polyethyleneimine (PEI), maintain reproducibility of migration time and a stable EOF [26-29]. However, very few MS-based intact protein studies have employed positively charged coating-based CZE-MS [30-33]. Here, we systematically studied the performance of one positively charged coating for CZE-MS/MS-based TDP. The positive charge coating is based on one monomer carrying a permanent positive charge, i.e., (3-acrylamidopropyl) trimethylammonium chloride (APTAC). The monomer APTAC has been used in one recent study to prepare positive charge coating for CZE-UV and CZE-MS analyses of basic small molecular drugs [34]. In that study, it has been demonstrated that co-polymerization of acrylamide (AM) and APTAC to make can make a tunable EOF for CZE separation and the EOF is stable in pH 2~10. In this study, we refined the procedure of preparing the co-polymer of poly(acrylamide-co-(3-acrylamidopropyl)trimethylammonium chloride) (PAMAPTAC) and implemented it in CZE-MS/MS-based TDP. The stability and reproducibility were well-evaluated in the study. The enhanced separation resolution of the capillary indicates its strong potential for the identification of low-abundance, large, and hydrophobic proteoforms. The variable charge density of the cationic coating provides the possibility for broad biological applications.

4.2 Experimental section

4.2.1 Materials and Chemicals

Bare fused silica capillaries (50-μm i.d., 360-μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ). Ammonium persulfate, ammonium bicarbonate (NH_4HCO_3), Dulbecco's phosphate buffered saline, 3-(Trimethoxysilyl)propyl methacrylate and Amicon Ultra centrifugal filter units (0.5 mL, 30 kDa cut-off) were purchased from Sigma-Aldrich (St. Louis, MO). Hydrofluoric acid, acrylamide, and LC/MS grade water, methanol, acetic acid, and formic acid were purchased from Fisher Scientific (Pittsburgh, PA). Urea was bought from Alfa Aesar. Protease inhibitor (cOmplete ULTRA Tables, Roche) and phosphatase inhibitor (PhosSTOP, Roche) was bought through Fisher Scientific.

4.2.2 Sample preparation

A standard protein mixture (accumulative concentration 1.25 mg/mL: 0.05 mg/mL ubiquitin, 0.2 mg/mL myoglobin, 0.3 mg/mL carbonic anhydrous, and 0.7 mg/mL bovine serum albumin) was prepared in the BGE solution of 5% acetic acid (pH 2.4).

Around 50 g of baker's yeast (Saccharomyces cerevisiae, strain ATCC 204508/S288c) was added in 1 L of yeast extract peptone dextrose (YPD) medium (autoclaved) and cultured at 37 °C (300 rpm shaking) overnight in an incubator shaker (Thermo Scientific MaxQ 4000). The yeast was harvested by centrifugation (3,000 g, 5 min), followed by washing with Dulbecco's phosphate buffered saline (dPBS) for three times. Two grams of the yeast pellet was resuspending in 10 mL of lysis buffer containing 8 M urea, 100 mM ammonium bicarbonate (ABC, pH 8.0), complete protease inhibitor and phosphatase. The cells were lysed (3 min, 3 times) using a homogenizer 150 for (Fisher Scientific) and sonicated on ice (3 min, 5 times) with Branson Sonifier 250 (VWR Scienfic). The concentration was determined using a bicinchoninic acid (BCA) kit.

Around 200 μg of yeast lysates was centrifuged on an Amicon 30 kDa molecular cut-off filter at 14, 000 g for 20 min. The proteins retain inside of the filter was washed two times with 100 mM ABC, two times with water, and two times with 5% AA. A small amount of precipitation was noticed during acidifying. Around 30 μL solution was collected after buffer exchange (1.4 mg/mL). The protein concentration was measured by Bradford assay. Prior to CE separation, the sample was centrifuged at 10,000 g for 3 min to remove potential precipitates.

4.2.3 Capillary preparation

To prepare the capillary for coating, the bare fused silica capillaries (1 m length, 50 μm i.d., 360 μm o.d.) was flushed with 300 μL of 1 M sodium hydroxide (NaOH), water, 1 M hydrochloric acid (HCl), water and methanol, successively. After degassing with nitrogen, 50% (v/v) 3-(trimethoxysiyl) propyl methacrylate in methanol was introduced into the capillary and incubated for three days grafting double bonds. The capillaries were rinsed by methanol and dried under nitrogen.

The linear poly acrylamide (LPA) coating was prepared based on our previous procedure with minor modifications [17]. Briefly, 500 μL of 4% (w/v) acrylamide was mixed with 3.5 μL of 5% ammonium persulfate (APS). The solution was degassed with nitrogen for 15 minutes and infused into the capillary, then incubated in 50 °C water bath for 1 hour. Three 50%

PAMAPTAC-coated capillaries, the experimental condition is based on the literature while initiator was changed to APS [34]. A 500 µL of 5% acrylamide was mixed with 500 µL of 5% APTAC and 7 µL of 5% (w/v) APS, degassed for 15 minutes, and infused into the capillary. The incubation is under 66 °C water bath for 1.5 hours. To note: APS generates the radical under high temperature and basic condition [35,36]. Given the weak basic of the ammonium ion hydrolysis (pH 8.6 for the coating solution), it is suggested to conduct the coating step in a time sensitive manner and a completed degas condition.

One end of the coated capillary was etched with hydrofluoric acid (HF) for 85 minutes to reduce the outer diameter to around 70 µm. Detailed video procedure is accessed from our published procedure [37].

4.2.4 CZE-ESI-MS/MS

A Beckman CESI 8000 capillary electrophoresis autosampler (Sciex) was coupled to the mass spectrometer by an in-house-constructed electrokinetically pumped sheath-flow CE-MS nanospray interface [38,39]. The electrospray emitter was a borosilicate glass capillary (1.0 mm o.d., 0.75 mm i.d.) pulled with a Sutter instrument P-1000 flaming/brown micropipette puller with an orifice size 20 ~30 µm and a length of 5 cm. The sheath liquid consists of 0.2% (v/v) formic acid and 10% (v/v) methanol. The spray voltage is around + 2.0 kV and the distance between the spray emitter orifice and the mass spectrometer entrance was ~2.8 mm.

The inlet of the capillary was inserted to the background electrolyte of CZE, 5% (v/v) acetic acid. For all experiment, a 50 nL injection was carried by applying 5 psi pressure for 9.5 s, based on the Poiseuille's law. For the separation of standard protein mixtures, + 30 kV was applied at the injection end of the LPA coated capillary for 30 minutes. And – 30 kV was applied at the injection end of PAMAPTAC coated capillary for 40 minutes. A 10-psi pressure was also applied during the last ten minutes of the runs. For the yeast lysate, the separation time was extended to 60 minutes. A 10-minute flush with 10 psi and +/- 30 kV was used to clean up the capillary.

Most of the experiments were conducted using an Orbitrap Exploris 480 mass spectrometer except the cross-capillary of yeast lysate, which was conducted on the Q-Exactive HF mass spectrometer. The parameters of the Exploris 480 mass spectrometer are listed below. The ion transfer tube temperature was 320 °C. The intact protein mode was turned on and the low-pressure mode was selected. Full MS scan was performed with the following parameters:

Orbitrap resolution of 480,000 (at m/z of 200), m/z range 500-3200, normalized automatic gain control (AGC) target of 300%, maximum injection time mode of auto, and microscan of 1. The six most intense precursor ions with charge state in the range of 5-60 and intensity higher than the threshold of 1E4 were isolated with 2 m/z window, followed by fragmentation energy high energy collisional dissociation (HCD) normalized collision energy as 25%. The orbitrap resolution is 120,000 (at m/z 200) for MS/MS with the microscan of 3. The normalized AGC target value for MS/MS is 100% and the maximum ion injection time mode is auto. Dynamic exclusion was enabled with the following settings: repeat count as 3, exclusion duration as 30 s and the exclusion of isotopes was enabled. For Low resolution MS, the full MS resolution is 7,500 with a microscan of 10. Source fragmentation energy was enabled for 5 V. The parameters of MS/MS stayed the same, other than the precursor ions with undetermined charge states was considered for isolation and fragmentation with a step normalized HCD collision energies of 25%, 35%, 45% and the microscan was 2. For the cross-capillary experiment using Q-Exactive HF mass spectrometer, full MS scan was performed with the following parameters: Orbitrap resolution of 120,000 (at m/z of 200), m/z range 600-2000, AGC target of 3E6, maximum injection time mode of 100 ms, and microscan of 3. The five most intense precursor ions were isolated with 2 m/z window, followed by fragmentation energy HCD normalized collision energy of 20%. The intensity threshold is 1E4 and the charge state of 1-5 was excluded for fragmentation. Product ions were detected with a resolution of 60,000 (at m/z 200), AGC target of 1E6 and the maximum injection time is 200 ms. The dynamic exclusion was enabled with a duration of 30 s and the exclusion of isotopes were enabled.

4.2.5 Data Analysis

The standard protein mixture was analyzed using Xcalibur software (Thermo Fisher Scientific) for the intensity, migration time, and full width at half maximum of proteoforms. For complex sample analyzed by low-high mode, the RAW data was deconvoluted using UniDec to detect the large proteoforms [40].

For complex sample analyzed by high-high mode, proteoform identification and quantification from yeast lysate was performed using TopPIC (Top-down mass spectrometry-based Proteoform Identification and Characterization) pipelines [41]. Briefly, the raw file was converted to mzML files by MSConvert using peak-picking algorithm [42]. Then, the spectral deconvolution was performed using Top-down mass spectrometry Feature Detection (TopFD,

version 1.7.0) with default parameter [43]. The database search was performed using TopPIC (1.7.0) against UniProt proteome database of Yeast (UP000002311, 6735 entries, accessed on 03/24/2023). The parameters were set as follows: mass error tolerance of matching masses of 25 ppm, maximum number of variable modifcations (including acetylation, phosphorylation, oxidation and methylation) of 3, maximum number of mass shifts of unknown modification of 2 ranging from -500 Da to 500 Da. The false discovery rates (FDRs) were estimated using the target-decoy approach. The spectrum level FDR cutoff was 1%, and the proteoform level FDR cutoff was 5%. The Top-down mass spectrometry-based identification of Differentially expressed proteoforms (TopDiff) was used to perform label-free quantification of the proteoforms. The lists of identified proteoforms from the yeast replicates are shown in Supporting Information I. The MS RAW files were deposited to the ProteomeXchange Consortium via the PRIDE [44] partner repository with the dataset identifier of PXD051862.

4.3 Results and discussions

4.3.1 Comparison of LPA and PAMAPTAC coatings for CZE-MS-based protein measurement

We first compared LPA-coated and PAMAPTAC-coated capillaries in terms of proteoform separations using a standard protein mixture as the sample. The standard protein mixture contained four proteins, 0.05 mg/mL ubiquitin, 0.2 mg/mL myoglobin, 0.3 mg/mL carbonic anhydrous, and 0.7 mg/mL bovine serum albumin, in a mass range of 8.6 - 66.7 kDa. For the LPA-coated capillary, the separation was conducted via applying a positive voltage of 30 kV at the sample injection end and about 2 kV at the CE-MS interface for ESI. The BGE was 5% AA (pH 2.4). For the PAMAPTAC-coated capillary, we applied a voltage of -30 kV at the sample injection end and ~2 kV at the spray emitter for separation. The BGE was the same as the LPA-coated capillary. The sample injection volume for both conditions was 50 nL. As depicted in **Figure 4.1**, CZE-MS using LPA and PAMAPTAC-coated capillaries produced reproducible separation and detection of the standard protein mixture. The PAMAPTAC-coated capillary provided a much wider separation window compared to the LPA-coated capillary (10 vs. 5 minutes). The neutral LPA coating minimized the EOF, and the migration of positively charged proteoforms were due to their electrophoretic mobilities under a strong electric field created by the +28-kV voltage applied across the capillary. Conversely, in the PAMAPTAC-coated capillary and under a -32 kV voltage condition, proteoform migration resulted from a combination of

forward EOF and backward electrophoresis, producing a complete inversion in the proteoform order compared to the LPA-coated capillary. We measured the EOF mobility of one bare fused silica capillary, one LPA-coated capillary, and one PAMAPTAC-coated capillary using a neutral marker benzyl alcohol. We determined the EOF mobilities as 2.01E-8, 2.27E-9, and 4.88E-8 $m^2V^{-1}s^{-1}$, respectively. The EOF mobility of PAMAPTAC-coated capillary agreed with the literature for poly(diallyldimethylammonium chloride), poly(allylamine hydrochloride), and PEI [45, 46].

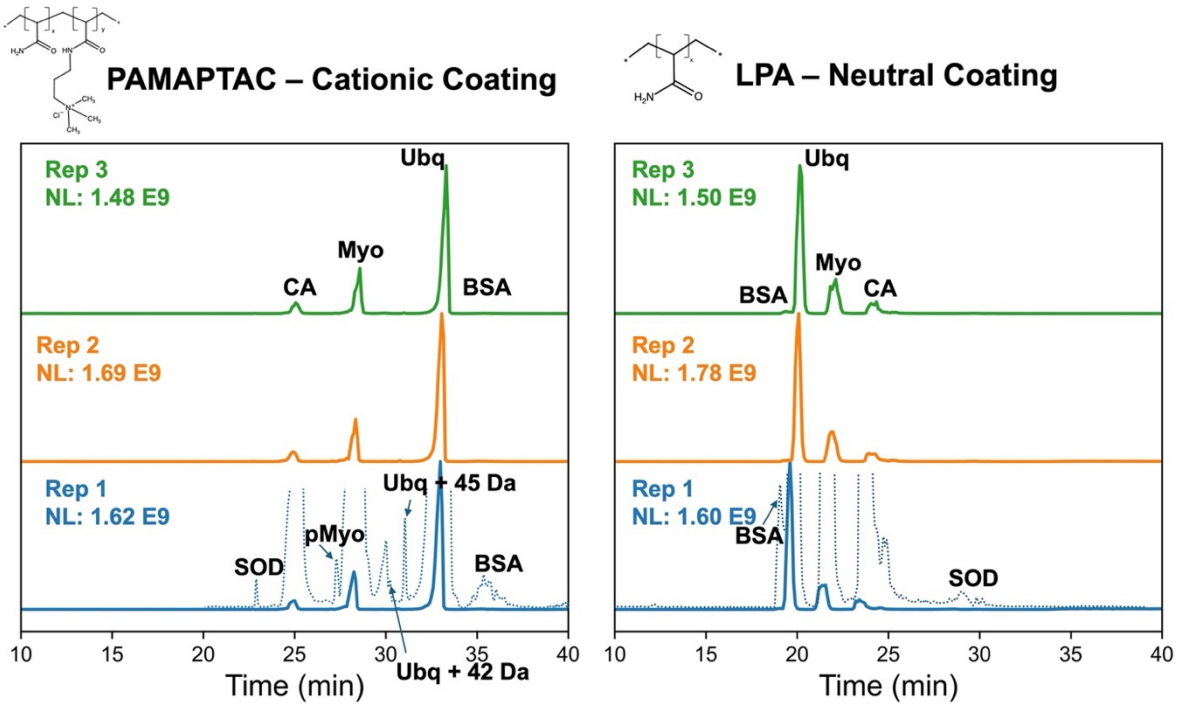


Figure 4.1. Electropherograms of a standard protein mixture analyzed by CZE-MS with a PAMAPTAC-coated capillary (A) and an LPA-coated capillary (B). Fifty nanoliter of the sample was injected for CZE-MS analysis in one run. The sample contained four standard proteins (carbonic anhydrous, CA; myoglobin, Myo; Ubiquitin, Ubq; bovine albumin serum, BSA) and the impurity superoxide dismutase (SOD) and it was dissolved in the background electrolyte (BGE, 5% acetic acid) of CZE. The dashed lines in (A) and (B) show zoomed-in electropherograms (~100x)).

We further compared the separation resolution of the standard proteins by LPA-coated and PAMAPTAC-coated capillaries, as shown in **Table 4.1**. CZE-MS using the PAMAPTAC-coated capillary improved the separation resolution between the adjacent protein peaks, on average, by 107%, compared to that using the LPA-coated capillary. Additionally, the PAMAPTAC coating provided much higher separation efficiency compared to the LPA coating,

evidenced by the much better average number of theoretical plates from the PAMAPTAC coating (35,183 vs. 12,648). Interestingly, two proteoforms of myoglobin (with or without a phosphorylation) and three proteoforms of ubiquitin (+0, +42 and +45 Da) were separated using the PAMAPTAC-coated capillary but not the LPA-coated capillary. Modifications such as phosphorylation and acetylation altered the net charge carried by the proteoform, thereby affecting the charge-to-size ratio and separation resolution. The data highlight the advantages of PAMAPTAC coating for bettering proteoform separation compared to the LPA coating. We need to point out that the preparation of the PAMAPTAC coating is as simple as the LPA coating. The results demonstrate that the PAMAPTAC coating could be a valuable alternative to the commonly used LPA-coating for further advancing TDP.

Table 4.1. Summary of the separation resolution between adjacent protein peaks of a standard protein mixture measured by CZE-MS using an LPA-coated capillary and a PAMAPTA-coated capillary

Run	BSA - Ubq	Ubq - Myo	Myo - CA	CA - SOD1
LPA-1	0.82	2.66	1.96	6.2
LPA-2	0.84	2.4	2.08	5.52
LPA-3	1.12	2.42	1.97	5.54
Mean±SD	**0.93 ± 0.17**	**2.49 ± 0.14**	**2.00 ± 0.07**	**5.75 ± 0.38**
PAMAPTA -1	2.21	7.26	5.31	4.8
PAMAPTA -2	1.49	7.16	5.17	3.88
PAMAPTA -3	1.83	7.25	5.47	4.35
Mean±SD	**1.84 ± 0.36**	**7.22 ± 0.06**	**5.32 ± 0.15**	**4.34 ± 0.46**

4.3.2 CZE-MS/MS using PAMAPTAC-coated capillaries for TDP of a complex yeast cell lysate

We further evaluated the performance of PAMAPTAC coating for CZE-MS/MS characterization of a complex sample, i.e., a yeast cell lysate. The yeast lysate extracted by 8M urea in 100 mM ammonium bicarbonate (pH 8.0) was buffer exchanged to 5% (v/v) acetic acid (pH 2.4). A fifty nanoliter sample (~70 ng) was injected to the same 1-meter capillary under

identical separation conditions as the standard proteins. The duration was extended to 70 minutes and repeated for six runs, **Figure 4.2 A**.

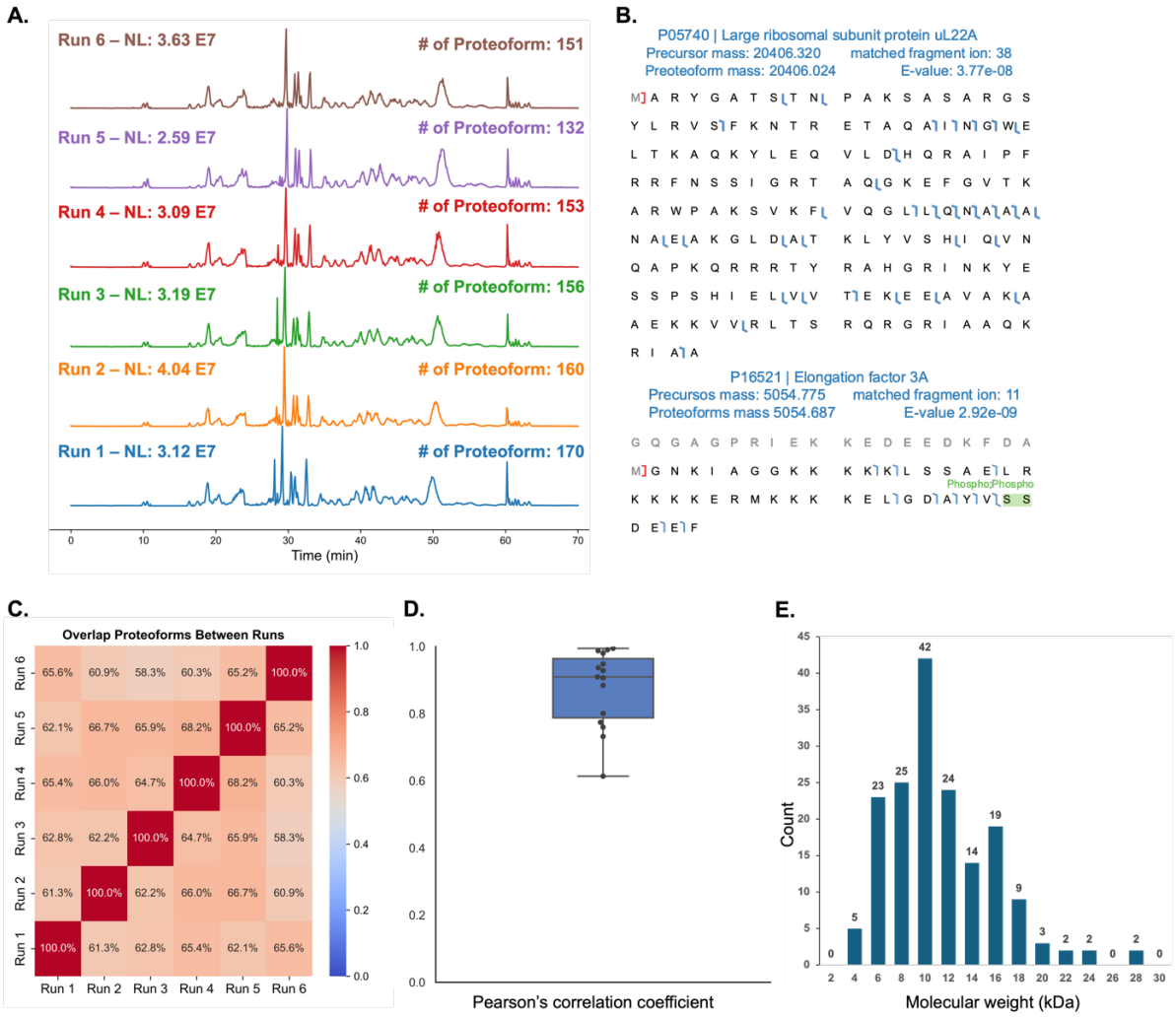


Figure 4.2. Summary of the yeast cell lysate data from CZE-MS/MS using a PAMAPTAC-coated capillary. (A). The successive six electropherograms of CZE-MS/MS analyses of a yeast cell lysate. The normalized level (NL) signal intensity and the number of identified proteoforms are labeled for each run. (B). Two examples of proteoforms identified across six runs. (C). Proteoform overlap between runs. (D). The box plot of the Pearson's correlation coefficients of proteoform intensity between any two runs. (E). The histogram of molecular weight distribution of the proteoforms identified in Run 1.

On average, approximately 154 proteoforms were identified per run, with 69 proteoforms consistently identified across all runs. Although the total number of identified proteoforms is limited, the ratio of identified proteoforms larger than 10 kDa is 43.70 ± 1.92 %. This ratio is higher than those reported in previous studies [19,47]. A histogram of molecular weight distribution is shown in **Figure 4.2 E**. The observed distribution shift can be attributed to a

different sample preparation technique. In this study, a 30 kDa MWCO filter was used with 6 washes, rather than the 10 kDa MWCO filter with three washes. Efficient washing in sample preparation resulted in fewer identifications and relatively larger proteofroms. Another reason of the limited numbers of proteoform identification was the smaller sample injection amount due to the lack of a concentration method [48,49]. **Figure 4.2 B** presents two proteoform examples. The large ribosomal subunit protein uL22A (RPL17A) is the largest proteoform consistently identified in each run. The intact proteoform from the initiator methionine removal to the end was identified with 38 matched fragment ions and a low E-value of 3.77E-08. The average migration time of RPL17A is 43.36 min with a remarkably low relative standard deviation of 1.24%, underscoring the high reproducibility of our CZE-MS/MS system. The second proteoform, the elongation factor 3A (YEF3) with dual phosphorylation sites at pSer 1039 and pSer 1040 was also identified in all six runs. This elongation factor is required for the yeast ribosome and plays a role as a negative regulator of nonderepressible 2 (GCN2) kinase activity [50,51]. The two phosphorylation sites were reported earlier in large-scale phosphorylation analysis [52,53]. Additionally, two proteoforms of nascent polypeptide-associated complex subunit alpha (EGD2) covering Gly89 - Lys174 were identified. One proteoform has a phosphorylation (+80.0010 Da) on Ser 101 and the other doesn't. The migration time of the phosphorylated form and unphosphorylated form was 17.74 ± 0.09 and 16.61 ± 0.04 min, respectively. The single phosphorylation altered the net charge of the proteoform, thereby altering its electrophoretic mobility. The distinct difference in migration time further demonstrates the high resolution of our CZE-MS method for proteoform separation.

We also assessed the proteoform identification between every two runs, noting around 60% proteoforms overlapped between pairs of runs, **Figure 4.2 C**. The box plot of Pearson's correlation coefficients of the intensity of overlapped proteoforms, **Figure 4.2 D**, indicates that the CZE-MS system is quantitatively reproducible with a mean and a medium value of 0.88 and 0.91, respectively.

We further evaluated the capillary-to-capillary reproducibility of the PAMAPTAC coating. We prepared another two PAMAPTAC-capillaries (capillary 2 and 3) following the procedure. Both standard protein mixtures and complex yeast samples were separated under identical separation conditions as used in PAMAPTAC capillary 1. As illustrated in **Figure 4.3 A** and **B**, the separation profiles of the standard protein mixture remained consistent with the

previous capillary, which suggests the high capillary-to-capillary reproducibility. Moreover, **Figure 4.3 C** shows the separation of the same yeast sample using three PAMAPTAC capillaries on a QE-HF instrument. Despite minor shift in migration times, the separation was well aligned, further confirming the high reproducibility across capillaries for complex samples.

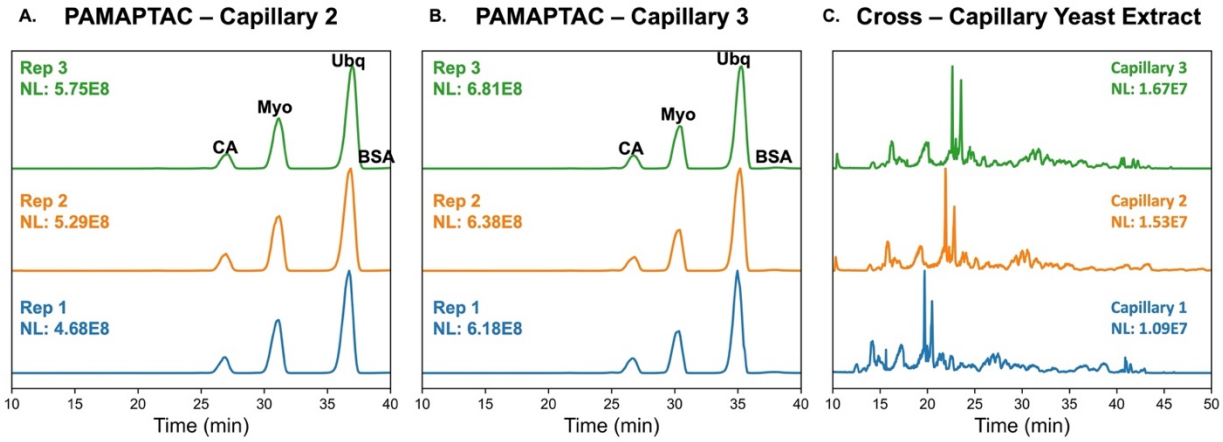


Figure 4.3. The reproducibility across capillaries. (A) and (B) are the electropherograms of standard protein mixture using different PAMAPTAC capillaries. (C). The electropherograms of the same yeast lysate using three different PAMAPTAC capillaries.

The identified proteoforms predominantly range between 5-20 kDa, as shown in **Figure 4.2 E.** Large proteoform identification from complex cell lysates is usually restricted in global TDP due to wide charge state distributions, resulting in low sensitivity [54]. In a typical TDP study, the mass of identified proteoforms usually is lower than 30 kDa [4,19,47]. Here we investigated the potential of our CZE-MS/MS technique with PAMAPTAC coating for measuring large proteoforms in a yeast cell lysate. To detect the large proteoforms in the yeast lysate, we applied low-resolution MS1 (resolution 7,500 at m/z of 200) and deconvoluted based on charge-state distribution by UniDec [40]. The CZE separation condition remains the same.

Without any size-based prefractionation method, two large proteoforms, 44.95 kDa and 45.14 kDa co-migrated, **Figure 4.4 B**, and they shared similar charge state distributions. **Figure 4.4 C and D** illustrate the detection of additional large proteoforms in the 25-35 kDa range. The data demonstrate that the cationic coating-based CZE-MS has a high potential to advance TDP toward the identification of large proteoforms. It is important to note that the identification of the large proteoforms was hindered by the limited fragment ions available. This insufficient gas-phase fragmentation of large proteoforms is a critical challenge in the field of TDP.

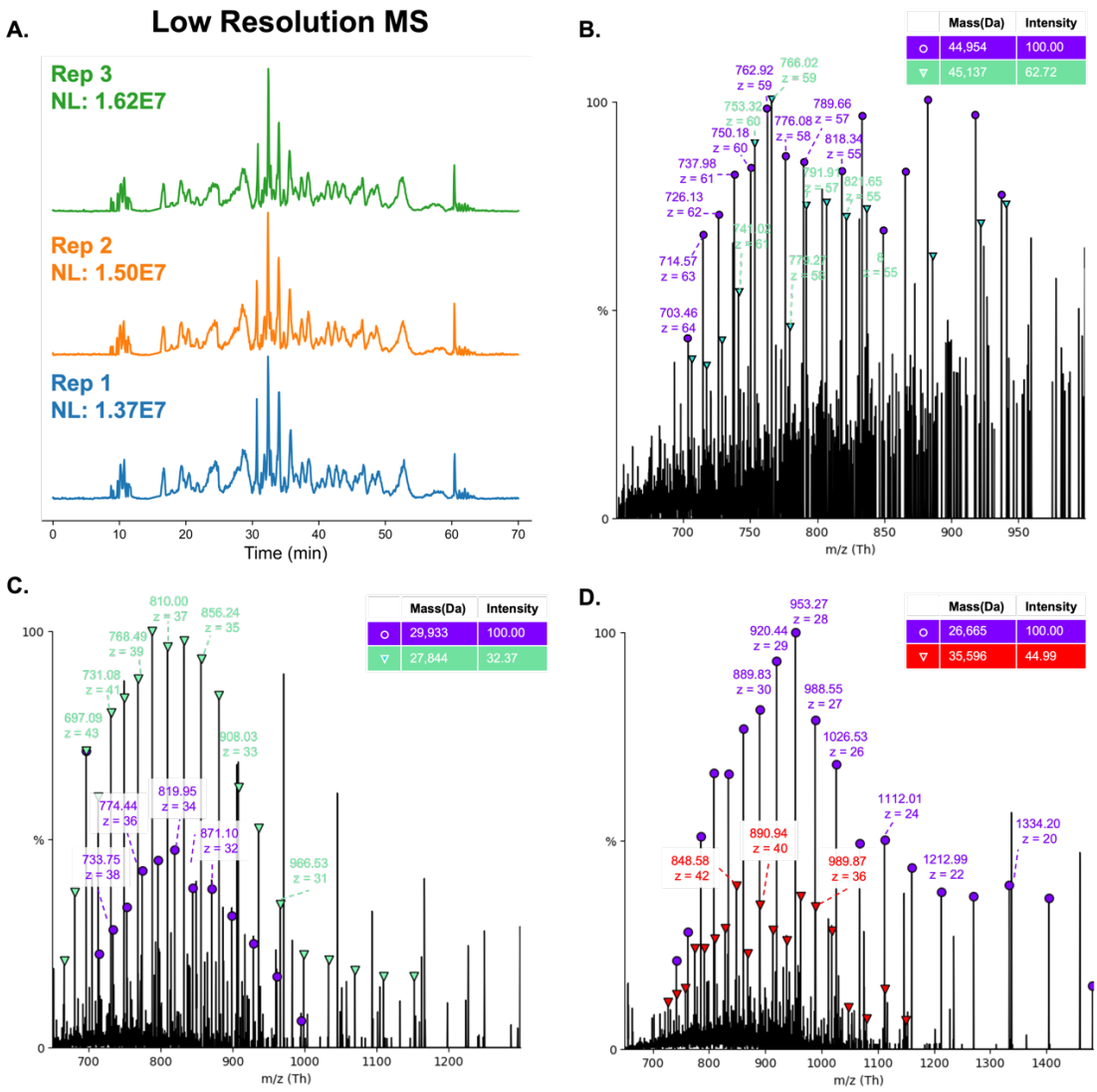


Figure 4.4. Low resolution MS1 of the yeast cell lysate. (A). The successive three electropherograms of the yeast lysate at the low-resolution MS1. (B), (C), (D) are the mass spectra and deconvoluted masses of large proteoform examples detected. The deconvolution is performed using UniDec software [40].

4.3.3 Proteoforms' electrophoretic mobility prediction from CZE-MS measurement using PAMAPTAC-coated capillary

Previous studies have demonstrated that the tunable EOF rate of PAMAPTAC capillaries, adjustable from 0 to 4E-8 $m^2V^{-1}s^{-1}$, depending on the charged monomer ratio, could enhance the electrophoretic separation resolution [34,55]. In this study, we examined the separation of standard protein mixtures using a PAMAPTAC capillary with 50% APTAC, observing a significantly higher separation resolution compared to an LPA capillary. We also tested the same

standard protein mixture using a PAMAPTAC capillary with 25% APTAC, As shown in **Figure 4.5 A**. The lower charge density of the monomer led to a reduced migration velocity, increasing the average separation resolution between CA and myoglobin from 5.32 ± 0.15 to 6.38 ± 0.66. Consequently, this adjustment delayed the migration of ubiquitin and BSA, requiring the pressure application after 60 minutes of separation. These findings state the impact of charged monomer concentration on both migration time and separation resolution. Therefore, optimizing migration time through adjustments in APTAC composition allows for the customization of separation conditions to meet the specific analytical requirements of a variety of proteins.

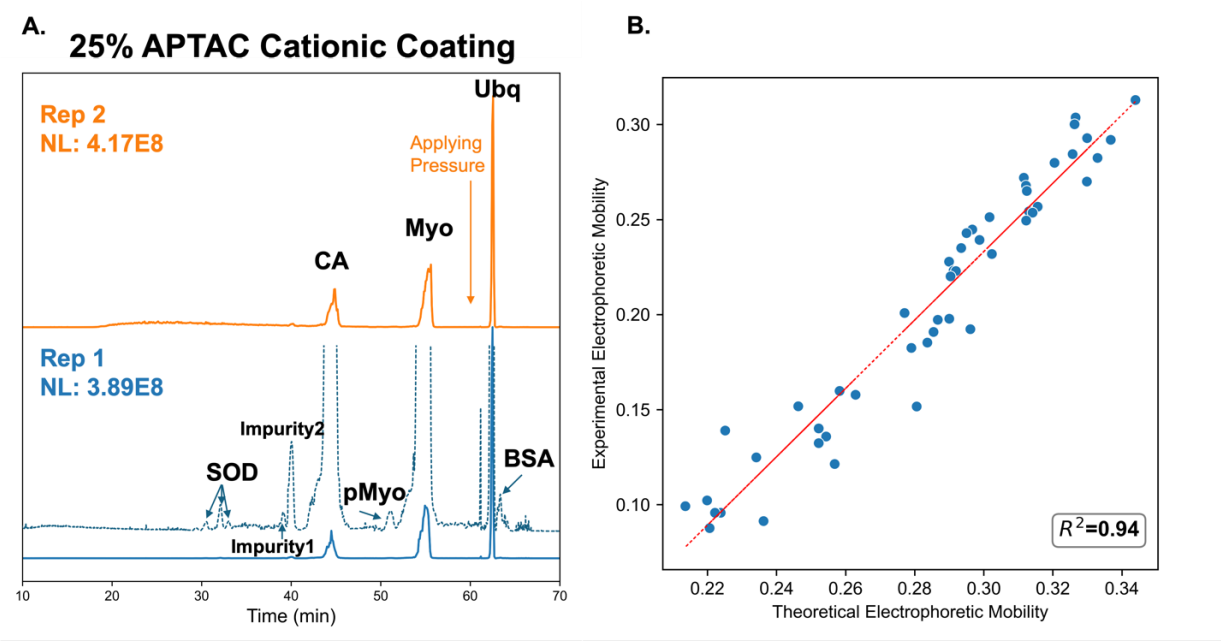


Figure 4.5. The predictability of cationic coated capillary electrophoresis. A. The electropherogram of standard protein mixture by 25% charged monomer. The dashed line shows a magnified view of the data (~100x). B. The predicted and experimental electrophoretic mobility of the unmodified proteoforms in the yeast lysate using 50% PAMAPTAC capillary.

To predict the migration time of proteoforms under EOF and electrophoretic mobility, we adopted a semiempirical model of the peptides and proteoforms in the open-tubular CZE separation [56-58]. Briefly, the theoretical electrophoretic mobility μ is calculated based on the net charge of protein, Q, and the proteoforms mass, M. The net charge is determined by accounting for positively charged residues (Ks, Rs, Hs, and the N-terminal amino group) in an acidic background electrolyte, and the proteoform mass is represented by the precursor mass. We proposed the predicted μ, combining the forward EOF with reverse electrophoresis, and an optimized coefficient:

$$Predicted\ \mu = 0.488 - 5 * \frac{\ln(1+0.35Q)}{M^{0.411}} \quad (1)$$

The experimental mobility (μ) is calculated by dividing the experimental velocity of the solute (v) by the applied electrical field (E):

$$Experimental\ \mu = \frac{v_{ep}}{E} = \frac{\frac{L}{t_m}}{\frac{V}{L}} = \frac{L^2}{V * t_m} \quad (2)$$

where L is the capillary length in meters, t_m is the migration time in seconds, and V is the applied potential in volts.

Validation using unmodified proteoforms identified in yeast lysate showed a good linearity (R^2=0.94), as depicted in **Figure 4.5 B**. Based on the correlation, the migration time of proteoforms can be predicted using this model. And integrating this predictive model with the correlation between EOF rate and charged monomer percentage, enables a robust framework for optimizing separation conditions tailored to diverse complex samples.

4.4 Conclusions

In conclusion, we developed a simple and efficient approach for preparing cationic capillary coating for CE-MS-based TDP. The cationic coating, PAMAPTAC, significantly enhanced the separation resolution, with its reproducibility confirmed through intra- and inter-capillary replicates. Importantly, our approach facilitated the detection of large proteoforms (greater than 30 kDa) without prefractionation, demonstrating its strong potential in identifying large and hydrophobic proteoforms while minimizing absorption. Additionally, this tunable EOF rate and predictable mobility offer promising potential for broader biological applications.

The current method prepared samples in the BGE solution, limiting the sample injection amount as well as the number of identified proteoforms. Future studies should focus on optimizing the online concentration method and reducing the sample complexity to enhance proteoform identification. Field-amplified sample stacking could be beneficial while maintaining proteins in a positively charged form [59]. Further investigations are required to explore the long-term durability of the coating.

4.5 Acknowledgments

We thank the support from the National Institute of General Medical Sciences for this project through grant R35GM153479. We also thank the support from the National Cancer Institute (NCI) through the grant R01CA247863.

REFERENCES

[1] Roberts, D. S.; Mann, M.; Melby, J. A.; Larson, E. J.; Zhu, Y.; Brasier, A. R.; Jin, S.; Ge, Y. Structural O-Glycoform Heterogeneity of the SARS-CoV-2 Spike Protein Receptor-Binding Domain Revealed by Top-Down Mass Spectrometry. J. Am. Chem. Soc. 2021; 143 (31):12014–12024.

[2] Smith, L. M.; Kelleher, N. L. Proteoforms as the next Proteomics Currency. Science 2018; 359 (6380):1106–1107.

[3] Adams, L. M.; DeHart, C. J.; Drown, B. S.; Anderson, L. C.; Bocik, W.; Boja, E. S.; Hiltke, T. M.; Hendrickson, C. L.; Rodriguez, H.; Caldwell, M.; Vafabakhsh, R.; Kelleher, N. L. Mapping the KRAS Proteoform Landscape in Colorectal Cancer Identifies Truncated KRAS4B That Decreases MAPK Signaling. J. Biol. Chem. 2023; 299 (1):102768.

[4] McCool, E. N.; Xu, T.; Chen, W.; Beller, N. C.; Nolan, S. M.; Hummon, A. B.; Liu, X.; Sun, L. Deep Top-down Proteomics Revealed Significant Proteoform-Level Differences between Metastatic and Nonmetastatic Colorectal Cancer Cells. Sci. Adv. 2022; 8 (51): eabq6348.

[5] Guo, Y.; Cupp-Sutton, K. A.; Zhao, Z.; Anjum, S.; Wu, S. Multidimensional Separations in Top–down Proteomics. Anal. Sci. Adv. 2023; 4 (5–6):181–203.

[6] Po, A.; Eyers, C. E. Top-Down Proteomics and the Challenges of True Proteoform Characterization. J. Proteome Res. 2023; 22 (12):3663–3675.

[7] Johnson, K. R.; Gao, Y.; Greguš, M.; Ivanov, A. R. On-Capillary Cell Lysis Enables Top-down Proteomic Analysis of Single Mammalian Cells by CE-MS/MS. Anal. Chem. 2022; 94 (41):14358–14367.

[8] Schlecht, J.; Jooß, K.; Moritz, B.; Kiessig, S.; Neusüß, C. Two-Dimensional Capillary Zone Electrophoresis-Mass Spectrometry: Intact mAb Charge Variant Separation Followed by Peptide Level Analysis Using In-Capillary Digestion. Anal. Chem. 2023; 95 (8):4059–4066.

[9] Wang, Q.; Xu, T.; Fang, F.; Wang, Q.; Lundquist, P.; Sun, L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Top-Down Proteomics of Mouse Brain Integral Membrane Proteins. Anal. Chem. 2023; 95 (34):12590–12594.

[10] Chen, D.; McCool, E. N.; Yang, Z.; Shen, X.; Lubeckyj, R. A.; Xu, T.; Wang, Q.; Sun, L. Recent Advances (2019–2021) of Capillary Electrophoresis-Mass Spectrometry for Multilevel Proteomics. Mass Spectrom. Rev. 2023; 42 (2):617-642.

[11] Shen, X.; Yang, Z.; McCool, E. N.; Lubeckyj, R. A.; Chen, D.; Sun, L. Capillary Zone Electrophoresis-Mass Spectrometry for Top-down Proteomics. Trends Anal. Chem. TRAC 2019; 120:115644.

[12] Gomes, F. P.; Yates III, J. R. Recent Trends of Capillary Electrophoresis-Mass Spectrometry in Proteomics Research. Mass Spectrom. Rev. 2019; 38 (6):445–460.

[13] Zhao, Y.; Sun, L.; Knierman, M. D.; Dovichi, N. J. Fast Separation and Analysis of Reduced Monoclonal Antibodies with Capillary Zone Electrophoresis Coupled to Mass Spectrometry. Talanta 2016; 148:529–533.

[14] Schwenzer, A.-K.; Kruse, L.; Jooß, K.; Neusüß, C. Capillary Electrophoresis-Mass Spectrometry for Protein Analyses under Native Conditions: Current Progress and Perspectives. PROTEOMICS 2024; 24 (3–4):2300135.

[15] Lowe, B. M.; Skylaris, C.-K.; Green, N. G. Acid-Base Dissociation Mechanisms and Energetics at the Silica–Water Interface: An Activationless Process. J. Colloid Interface Sci. 2015; 451:231–244.

[16] Nagy, C.; Andrási, M.; Hamidli, N.; Gyémánt, G.; Gáspár, A. Top-down Proteomic Analysis of Monoclonal Antibodies by Capillary Zone Electrophoresis-Mass Spectrometry. J. Chromatogr. Open 2022; 2:100024.

[17] Zhu, G.; Sun, L.; Dovichi, N. J. Thermally-Initiated Free Radical Polymerization for Reproducible Production of Stable Linear Polyacrylamide Coated Capillaries, and Their Application to Proteomic Analysis Using Capillary Zone Electrophoresis-Mass Spectrometry. Talanta 2016; 146:839–843.

[18] Hamidli, N.; Andrasi, M.; Nagy, C.; Gaspar, A. Analysis of Intact Proteins with Capillary Zone Electrophoresis Coupled to Mass Spectromery Using Uncoated and Coated Capillaries. J. Chromatogr. A 2021; 1654:462448.

[19] Sadeghi, S. A.; Chen, W.; Wang, Q.; Wang, Q.; Fang, F.; Liu, X.; Sun, L. Pilot Evaluation of the Long-Term Reproducibility of Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Top-Down Proteomics of a Complex Proteome Sample. J. Proteome Res. 2024; 23 (4):1399–1407.

[20] Kar, S.; Dasgupta, P. K. Improving Resolution in Capillary Zone Electrophoresis through Bulk Flow Control. Microchem. J. 1999; 62 (1):128–137.

[21] Meagher, R. J.; Seong, J.; Laibinis, P. E.; Barron, A. E. A Very Thin Coating for Capillary Zone Electrophoresis of Proteins Based on a Tri(Ethylene Glycol)-terminated Alkyltrichlorosilane. ELECTROPHORESIS 2004; 25 (3):405–414.

[22] Towns, J. K.; Regnier, F. E. Polyethyleneimine-Bonded Phases in the Separation of Proteins by Capillary Electrophoresis. J. Chromatogr. A 1990; 516 (1):69–78.

[23] Znaleziona, J.; Drahoňovský, D.; Drahoš, B.; Ševčík, J.; Maier, V. Novel Cationic Coating Agent for Protein Separation by Capillary Electrophoresis†. J. Sep. Sci. 2016; 39 (12):2406–2412.

[24] Mező, E.; Páger, C.; Makszin, L.; Kilár, F. Capillary Zone Electrophoresis of Proteins Applying Ionic Liquids for Dynamic Coating and as Background Electrolyte Component. ELECTROPHORESIS 2020; 41 (24):2083–2091.

[25] Córdova, E.; Gao, J.; Whitesides, G. M. Noncovalent Polycationic Coatings for Capillaries in Capillary Electrophoresis of Proteins. Anal. Chem. 1997; 69 (7):1370–1379.

[26] Giorgetti, J.; Lechner, A.; Del Nero, E.; Beck, A.; François, Y.-N.; Leize-Wagner, E. Intact Monoclonal Antibodies Separation and Analysis by Sheathless Capillary Electrophoresis-Mass Spectrometry. Eur. J. Mass Spectrom. 2019; 25 (3):324–332.

[27] Katayama, H.; Ishihama, Y.; Asakawa, N. Stable Cationic Capillary Coating with Successive Multiple Ionic Polymer Layers for Capillary Electrophoresis. Anal. Chem. 1998; 70 (24):5272–5277.

[28] Crihfield, C. L.; Kristoff, C. J.; Veltri, L. M.; Penny, W. M.; Holland, L. A. Semi-Permanent Cationic Coating for Protein Separations. J. Chromatogr. A 2019; 1607:460397.

[29] Giorgetti, J.; Beck, A.; Leize-Wagner, E.; François, Y.-N. Combination of Intact, Middle-up and Bottom-up Levels to Characterize 7 Therapeutic Monoclonal Antibodies by Capillary Electrophoresis - Mass Spectrometry. J. Pharm. Biomed. Anal. 2020; 182:113107.

[30] Bush, D. R.; Zang, L.; Belov, A. M.; Ivanov, A. R.; Karger, B. L. High Resolution CZE-MS Quantitative Characterization of Intact Biopharmaceutical Proteins: Proteoforms of Interferon-B1. Anal. Chem. 2016; 88 (2):1138–1146.

[31] Han, X.; Wang, Y.; Aslanian, A.; Bern, M.; Lavallée-Adam, M.; Yates, J. R. I. Sheathless Capillary Electrophoresis-Tandem Mass Spectrometry for Top-Down Characterization of Pyrococcus Furiosus Proteins on a Proteome Scale. Anal. Chem. 2014; 86 (22):11006–11012.

[32] Haselberg, R.; Ratnayake, C. K.; de Jong, G. J.; Somsen, G. W. Performance of a Sheathless Porous Tip Sprayer for Capillary Electrophoresis–Electrospray Ionization-Mass Spectrometry of Intact Proteins. J. Chromatogr. A 2010; 1217 (48):7605–7611.

[33] Li, Y.; Compton, P. D.; Tran, J. C.; Ntai, I.; Kelleher, N. L. Optimizing Capillary Electrophoresis for Top-down Proteomics of 30–80 kDa Proteins. PROTEOMICS 2014; 14 (10):1158–1164.

[34] Konášová, R.; Butnariu, M.; Šolínová, V.; Kašička, V.; Koval, D. Covalent Cationic Copolymer Coatings Allowing Tunable Electroosmotic Flow for Optimization of Capillary Electrophoretic Separations. Anal. Chim. Acta 2021; 1178:338789.

[35] Furman, O. S.; Teel, A. L.; Ahmad, M.; Merker, M. C.; Watts, R. J. Effect of Basicity on Persulfate Reactivity. J. Environ. Eng. 2011; 137 (4):241–247.

[36] Ahmed, M.; Erdőssy, J.; Horvath, V. The Role of the Initiator System in the Synthesis of Acidic Multifunctional Nanoparticles Designed for Molecular Imprinting of Proteins. Period. Polytech. Chem. Eng. 2020; 65.

[37] McCool, E. N.; Lubeckyj, R.; Shen, X.; Kou, Q.; Liu, X.; Sun, L. Large-Scale Top-down Proteomics Using Capillary Zone Electrophoresis Tandem Mass Spectrometry. J. Vis. Exp. JoVE 2018; No. 140.

[38] Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J. Simplified Capillary Electrophoresis Nanospray Sheath-Flow Interface for High Efficiency and Sensitive Peptide Analysis. Rapid Commun. Mass Spectrom. RCM 2010; 24 (17):2554–2560.

[39] Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J. A Third-Generation Electro-Kinetically Pumped Sheath Flow Nanospray Interface with Improved Stability and Sensitivity for Automated Capillary Zone Electrophoresis-Mass Spectrometry Analysis of Complex Proteome Digests. J. Proteome Res. 2015; 14 (5):2312–2321.

[40] Marty, M. T.; Baldwin, A. J.; Marklund, E. G.; Hochberg, G. K. A.; Benesch, J. L. P.; Robinson, C. V. Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. Anal. Chem. 2015; 87 (8):4370–4376.

[41] Kou, Q.; Xun, L.; Liu, X. TopPIC: A Software Tool for Top-down Mass Spectrometry-Based Proteoform Identification and Characterization. Bioinformatics 2016; 32 (22):3495.

[42] Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. Bioinformatics 2008; 24 (21):2534–2536.

[43] Basharat, A. R.; Zang, Y.; Sun, L.; Liu, X. TopFD: A Proteoform Feature Detection Tool for Top–Down Proteomics. Anal. Chem. 2023; 95 (21):8189–8196.

[44] Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Pérez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaíno, J. A. The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data. Nucleic Acids Res. 2019; 47 (D1): D442–D450.

[45] Roca, S.; Dhellemmes, L.; Leclercq, L.; Cottet, H. Polyelectrolyte Multilayers in Capillary Electrophoresis. ChemPlusChem 2022; 87 (4): e202200028.

[46] Roca, S.; Leclercq, L.; Gonzalez, P.; Dhellemmes, L.; Boiteau, L.; Rydzek, G.; Cottet, H. Modifying Last Layer in Polyelectrolyte Multilayer Coatings for Capillary Electrophoresis of Proteins. J. Chromatogr. A 2023; 1692:463837.

[47] Xu, T.; Wang, Q; Wang, Q.; Sun, L. Coupling High-Field Asymmetric Waveform Ion Mobility Spectrometry with Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Top-Down Proteomics. Anal. Chem. 2023; 95, (25):9497-9504.

[48] Tůma, P.; Hložek, T.; Sommerová, B.; Koval, D. Large Volume Sample Stacking of Antiepileptic Drugs in Counter Current Electrophoresis Performed in PAMAPTAC Coated Capillary. Talanta 2021; 221:121626.

[49] Lubeckyj, R. A.; McCool, E. N.; Shen, X.; Kou, Q.; Liu, X.; Sun, L. Single-Shot Top-Down Proteomics with Capillary Zone Electrophoresis-Electrospray Ionization-Tandem Mass Spectrometry for Identification of Nearly 600 Escherichia Coli Proteoforms. Anal. Chem. 2017; 89 (22):12059–12067.

[50] Dasmahapatra, B.; Chakraburtty, K. Protein Synthesis in Yeast. I. Purification and Properties of Elongation Factor 3 from Saccharomyces Cerevisiae. J. Biol. Chem. 1981; 256 (19):9999–10004.

[51] Visweswaraiah, J.; Lee, S. J.; Hinnebusch, A. G.; Sattlegger, E. Overexpression of Eukaryotic Translation Elongation Factor 3 Impairs Gcn2 Protein Activation. J. Biol. Chem. 2012; 287 (45):37757–37768.

[52] Holt, L. J.; Tuch, B. B.; Villén, J.; Johnson, A. D.; Gygi, S. P.; Morgan, D. O. Global Analysis of Cdk1 Substrate Phosphorylation Sites Provides Insights into Evolution. Science 2009; 325 (5948):1682.

[53] Li, X.; Gerber, S. A.; Rudner, A. D.; Beausoleil, S. A.; Haas, W.; Villén, J.; Elias, J. E.; Gygi, S. P. Large-Scale Phosphorylation Analysis of α-Factor-Arrested Saccharomyces Cerevisiae. J. Proteome Res. 2007; 6 (3):1190–1197.

[54] Compton PD, Zamdborg L, Thomas PM, et al. On the Scalability and Requirements of Whole Protein Mass Spectrometry. Anal Chem. 2011; 83:6868–6874.

[55] P. Tůma, D. Koval, B. Sommerová, Š. Vaculín, Separation of anaesthetic ketamine and its derivates in PAMAPTAC coated capillaries with tuneable counter-current electroosmotic flow. Talanta 2020; 217:121094.

[56] Chen, W.; McCool, E. N.; Sun, L.; Zang, Y.; Ning, X.; Liu, X. Evaluation of Machine Learning Models for Proteoform Retention and Migration Time Prediction in Top-Down Mass Spectrometry. J. Proteome Res. 2022; 21 (7):1736–1747.

[57] Chen, D.; Lubeckyj, R. A.; Yang, Z.; McCool, E. N.; Shen, X.; Wang, Q.; Xu, T.; Sun, L. Predicting Electrophoretic Mobility of Proteoforms for Large-Scale Top-Down Proteomics. Anal. Chem. 2020; 92 (5):3503–3507.

[58] Cifuentes, A.; Poppe, H. Simulation and Optimization of Peptide Separation by Capillary Electrophoresis. J. Chromatogr. A 1994; 680 (1):321–340.

[59] Simpson, S. L.; Quirino, J. P.; Terabe, S. On-Line Sample Preconcentration in Capillary Electrophoresis. Fundamentals and Applications. J. Chromatogr. A 2008; 1184 (1–2):504–541.

CHAPTER 5.* Native Proteomics by Capillary Zone Electrophoresis-Mass Spectrometry

5.1 Introduction

Proteins regulate cellular processes by their diverse proteoforms [1,2] and the various protein complexes via non-covalent protein-protein interactions, protein-ligand bindings, and protein-DNA/RNA interactions [3]. Native mass spectrometry (nMS) provides critical insights into the structures, functions, and dynamics of proteoforms and protein complexes near physiological conditions [4–7]. Native MS has been widely employed to study well-purified proteoforms and protein complexes with low complexity through either direct infusion [8-13] or coupling with online/offline native separation methods, including size-exclusion chromatography (SEC) [14-18], ion-exchange chromatography (IEX) [19,20], hydrophobic interaction chromatography (HIC) [21,22], and capillary zone electrophoresis (CZE) [23].

Native proteomics aims to measure endogenous proteoforms and protein complexes under a near physiological condition on a proteome scale and it requires highly efficient separation techniques for protein complexes prior to nMS [24]. The first native proteomics study coupled off-line IEX or native gel-eluted liquid fractionation with direct infusion nMS for the characterization of protein complexes in mouse heart and human cancer cell lines, identifying 125 endogenous complexes from about 600 fractions [25]. More recently, direct infusion nMS was employed to measure protein complexes from a human heart tissue lysate using a Fourier-transform ion cyclotron resonance (FTICR) mass spectrometer with the identification of a handful of protein complexes about 30 kDa or smaller [26]. Native CZE-MS (nCZE-MS) has high separation efficiency and high detection sensitivity for protein complexes and has been applied to analyzing low-complexity protein samples, i.e., monoclonal antibodies [27], large protein complexes like GroEL (near 1MDa) [28-30], ribosomes [31], and nucleosomes [32]. Native SEC fractionation and online nCZE-MS analysis of an *E. coli* cell lysate identified 23 protein complexes smaller than 30 kDa, representing the first native proteomics study of a complex proteome using online liquid-phase separation-MS [33]. However, those native proteomics studies are either too time and labor-consuming or only able to detect small proteoforms/protein complexes from complex proteomes.

---

* *This chapter is partially adapted with permission from Wang, Q., Wang, Q., Qi, Z., Moeller, W., Wysocki, VH., Sun, L. Native proteomics by capillary zone electrophoresis-mass spectrometry. bioRxiv, 2024.04. 24.590970*

In this study, we developed a high-throughput nCZE-MS technique for native proteomics measurement of large proteoforms and protein complexes up to 400 kDa from complex samples, i.e., an *E. coli* cell lysate. The nCZE-MS technique is based on the online coupling of nCZE to an ultra-high mass range (UHMR) Orbitrap mass spectrometer. We first evaluated the nCZE-MS technique using a standard protein complex mixture. Then, we employed the technique to measure endogenous proteoforms and protein complexes in *E. coli* cells. We also compared our nCZE-MS data with mass photometry results in terms of mass distribution of *E. coli* proteoforms and protein complexes [34].

5.2 Experimental section

5.2.1 Material and chemicals

Bare fused silica capillaries (50-μm i.d., 360-μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ). 3-(Trimethoxysilyl) propyl methacrylate, ammonium persulfate, ammonium acetate (AmAc), Dulbecco's phosphate-buffered saline (dPBS), bovine serum albumin (BSA) and carbonic anhydrase (CA) from bovine erythrocytes, Cytochrome C (Cyt C), myoglobin from equine (Myo), C-reactive protein (CRP), glutamate dehydrogenase (GDH) were purchased from Sigma-Aldrich (St. Louis, MO). Hydrofluoric acid (HF), streptavidin (SA) and LC/MS grade water were purchased from Fisher Scientific (Pittsburgh, PA). Acrylamide were purchased from Acros Organics (NJ, USA). Micro Bio-SpinTM 6 kDa gel-filtration column units for buffer exchange was purchased from Bio-Rad. Protease inhibitors (cOmplete ULTRA Tables) and phosphatase inhibitors (PhosSTOP) were from Roche.

5.2.2 Sample preparation

A mixture of standard protein complexes containing Cyt C (0.7 μM), Myo (0.5 μM), CA (3 μM), SA (1.1 μM), BSA (0.7 μM), CRP (1 μM) and GDH (6 μM) was prepared in 20 mM AmAc (pH ~ 7.0).

E. coli (strain Top10) was cultured in Terrific Broth (TB) medium at 37 °C until OD_{600} reached 0.7. After washed with dPBS three times, 2 g pellet was suspended in 5 mL dPBS buffer plus complete protease inhibitors and phosphatase inhibitors and homogenized for 30 s, followed by sonication with a Branson Sonifier 250 (VWR Scientific, Batavia, IL) on ice for 2 minutes, 3 times. After centrifugation at 10,000 g for 10 minutes, the supernatant containing the extracted proteins was collected. A small aliquot of the diluted sample was used for the bicinchoninic acid (BCA) assay to determine the protein concentration (~2 mg/mL). One aliquot of the *E. coli*

lysate was diluted 8,000 times (~2.5 nM assuming an average molecular weight of 80 kDa) by 20 mM AmAc and directly measured using mass photometry.

Another aliquoted *E. coli* lysate was buffer exchanged to 20 mM AmAc by Bio-Spin 6 kDa gel-filtration column. The column which was washed with 20 mM AmAc and centrifuged at 1,000 x (g) for 2 minutes and repeated for 3 times. A 50 µL (100 µg protein) cell lysate was loaded on a 6 kDa gel-filtration and centrifuged for 4 minutes at 1,000 x (g). And the step was repeated with another pre-washed gel-filtration column to ensure the depletion of dPBS.

5.2.3 Mass photometry

Mass photometry experiments were conducted on a TwoMP instrument (Refeyn Inc.) Glass coverslips and silicone gaskets used in this measurement were cleaned with ultrapure water and isopropanol (IPA) sequentially in order of water - IPA - water - IPA - water, then dried by pure nitrogen. The oil immersion objective was covered with a clean coverslip, and a 6-well silicone gasket was placed on the top of the coverslip.

Calibration was carried out using a mixture of 10 nM thyroglobulin (TG) and beta-amylase (BAM). Four peaks corresponding to the monomer, dimer tetramer of BAM, together with the dimer of TG, have been detected. These contrasts and corresponding masses generated a calibration curve with an R square value of 0.99999, and the calibration was used to identify the rough mass of individual proteins or protein complexes existing in the cell lysate.

5.2.4 Preparation of LPA-coated separation capillary

The inner wall of the separation capillary (50-µm i.d., 360-µm o.d.) was coated with linear polyacrylamide (LPA) based on the protocol described in previous references [35,36]. Briefly, a bare fused silica capillary was successively flushed with 1 M sodium hydroxide, water, 1 M hydrochloric acid, water, and methanol, followed by treatment with 3-(trimethoxysilyl) propyl methacrylate for at least 24 hours to introduce carbon-carbon double bonds on the inner wall of the capillary. The treated capillary was filled with degassed acrylamide solution in water (4%) containing ammonium persulfate, followed by incubation at 50 °C water bath for 55 min with both ends sealed by silica rubber. After that, the capillary was flushed with water to remove the unreacted reagents. Then one end of the LPA-coated capillary was etched with HF based on the protocol in reference [37] for 85 minutes to reduce its outer diameter to around 70 µm.

5.2.4 Native CZE-ESI-MS

A Beckman CESI8000 Plus capillary electrophoresis autosampler was used for the automated operation of capillary zone electrophoresis (CZE). A commercialized electrokinetically pumped sheath flow interface (CMP Scientific) was used to couple CZE to mass spectrometer [38,39]. A Q-Exactive UHMR mass spectrometer (Thermo Fisher Scientific) were used for the experiments. The interface was directly attached to mass spectrometer. The ESI emitters of the interface were pulled from borosilicate glass capillaries (1.0 mm o.d., 0.75 mm i.d.) with a Sutter P-1000 flaming/brown micropipet puller with an orifice size ~25 μm. The sheath liquid contains 10 mM AmAc. Voltage for ESI was ~2 kV. A 1-meter LPA coated capillary (50-μm i.d. and 360-μm o.d.) was used for the CZE. The background electrolyte (BGE) for CZE was 25 mM AmAc (pH ~ 7.0).

The transfer capillary temperature was 250 oC, and the S-lens RF level was 200. The number of microscans was 5 for MS and the in-source trapping (IST) desolvation voltage was -30V. The resolution for MS was 6250 (m/z 200). The AGC target was 1E6 for MS. The maximum injection time was 200 ms for MS. The mass range for MS scans was 1000-10000 m/z. The *E. coli* sample was injected into the separation capillary for CZE-MS/MS with 5-psi pressure for 9.5 s (50 nL, 2.5% of capillary volume). A 70-minute CZE separation with 30 kV applied at the BGE end and 1 psi was applied at the mean time. For standard protein mixture, the separation is 45 min, and the separation is under 1.5 psi. The mass spectrometer setting is the same other than the -50V IST desolvation voltage.

5.2.5 Data analysis

All the mass spectra were averaged by a time window of every 30 s, followed by manual deconvolution. The deconvolution data was further checked by either UniDec or ESIprot [40, 41]

5.3 Results and discussions

5.3.1 High sensitivity CZE-ESI-UHMR for standard protein complex

Figure 5.1 shows the workflow of native proteomics analysis of an *E. coli* cell lysate using our nCZE-UHMR Orbitrap platform. Briefly, the cultured *E. coli* cells (Top10 strain) were lysed in a Dulbecco's phosphate-buffered saline (dPBS) buffer containing complete protease inhibitors and phosphatase inhibitors. The cell lysate was then buffer-exchanged on a spin column (Biorad P6) to a buffer containing 20 mM ammonium acetate (AmAc, pH ~7.0) by gel filtration, followed by nCZE-MS analysis. The online nCZE-MS was assembled by coupling a

Sciex CESI-8000 Plus CE autosampler to a Thermo Fisher Scientific Q-Exactive UHMR mass spectrometer through a commercialized electrokinetically pumped sheath flow CE-MS interface (EMASS-II, CMP Scientific) [38,39]. A 1-meter-long linear polyacrylamide (LPA) coated capillary (50-µm i.d., 360-µm o.d.) was used for the CZE separation and the LPA coating was employed to reduce the protein non-specific adsorption onto the capillary inner wall. The background electrolyte (BGE) for CZE was 25 mM AmAc (pH ~7.0), and the sheath buffer for electrospray ionization (ESI) was 10 mM AmAc (pH ~7.0). Only roughly a 50-ng aliquot of the *E. coli* sample was consumed in a single nCZE-MS run. Raw MS data were averaged every 30 seconds manually, followed by manual mass deconvolution and check using either UniDec or ESIprot [40, 41].

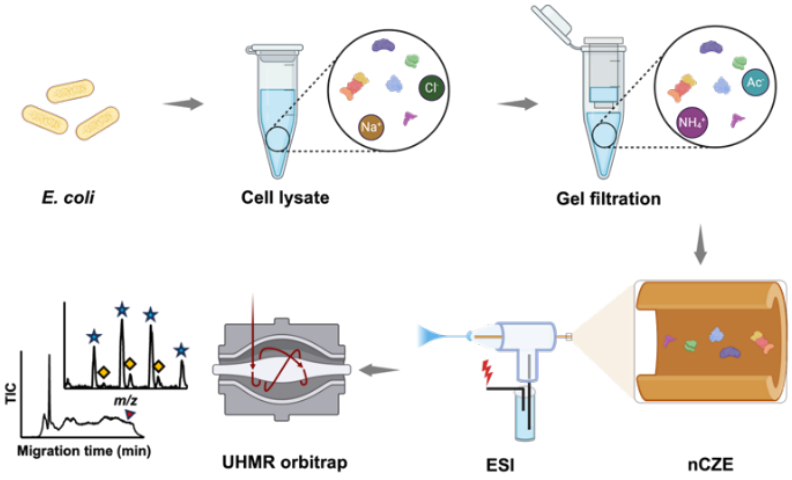


Figure 5.1. Flow chart of nCZE-ESI-MS for native proteomics of an *E. coli* cell lysate. The figure is created using the BioRender and is used here with permission.

We investigated the sensitivity of the nCZE-ESI-UHMR platform for measuring protein complexes using a mixture of standard proteins and protein complexes, as shown in **Figure 5.2**. Strong signals were observed for streptavidin (SA, 53 kDa), carbonic anhydrase (CA, 29 kDa), C-reactive protein (CRP, 115 kDa), and bovine serum albumin (BSA, 66 kDa) in the original sample with a total protein injection of about 600 femtomoles. After sample dilution by a factor of 50, a clear CRP peak was still observed, even though only 100 pg of the protein complex (~800 attomole) was loaded, indicating the high sensitivity of the technique.

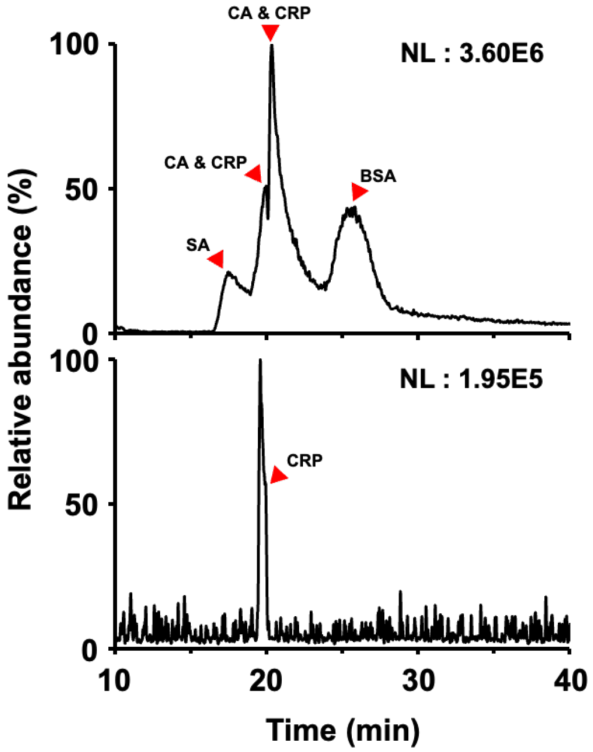
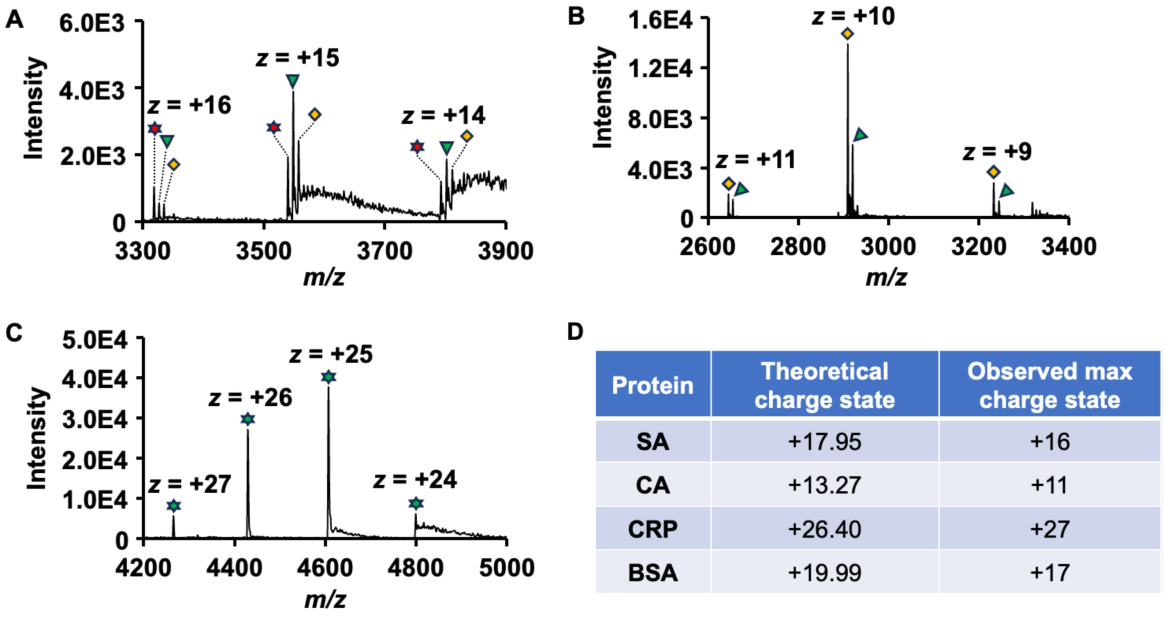


Figure 5.2. The data of a mixture of standard protein complexes analyzed by CZE-ESI-UHMR, original concentration (13 μM, top) and 50-time dilution (bottom). SA: Streptavidin; CA: carbonic anhydrase; CRP: C-reactive protein; BSA: bovine serum albumin. Cyt C, Myo, and GDH were not detected in the runs.

Figure 5.3 A shows one mass spectrum of three SA tetramers with masses as 53084.67 Da, 53216.07 Da, and 53347.97 Da. A 131-Da mass difference was observed between neighboring SA complexes, corresponding to N-terminal methionine variation on SA, which is consistent with the literature [42]. **Figure 5.3 B** shows a mass spectrum of CA-Zn(II) complex (29088.10 Da) and another CA complex (29194.01 Da) with an additional 107-Da mass shift compared to the CA-Zn(II) complex [42,43]. **Figures 5.3 C** shows the mass spectra of pentameric CRP complex in the original sample. Based on De La Mora's prediction of the maximum (Rayleigh) charge 'Z_R' of a native protein during the ESI process (Z_R = 0.0778*M0.5), the max charge of CRP is around 26.4 [44,45]. The max charge states of CRP observed in the original and 50-time diluted samples are 27 and 26, matching well with the Z_R of native CRP. We observed slightly lower max charge states compared to the theoretical charge states for the SA tetramer, CA-Zn (II) complex, and BSA, **Figure 5.3 D**. The data clearly demonstrate that intact protein complexes are maintained in native-like states during nCZE-ESI-UHMR measurements.



Protein	Theoretical charge state	Observed max charge state
SA	+17.95	+16
CA	+13.27	+11
CRP	+26.40	+27
BSA	+19.99	+17

Figure 5.3. The mass spectra of some standard protein complexes from **Figure 5.2.** (A) Tetramer of SA (red star: 53084.49 Da; green inverted triangle: 53215.89 Da; yellow diamond: 53348.18 Da). (B) CA (yellow diamond: 29087.73 Da; green inverted triangle: 29194.01 Da). (C) CRP (green star: 115148.23 Da). (D) Summary of theoretical charge states and observed max charge states of 4 standard proteins or protein complexes.

5.3.2 Detection of proteoforms or protein complexes from an *E. coli* cell lysate

The high sensitivity of nCZE-UHMR for the standard protein complexes motivated us to analyze an *E. coli* cell lysate. **Figure 5.4 A** shows an example electropherogram of the sample from nCZE-MS. The proteoforms or protein complexes migrated out of the capillary in a time range of 20-65 minutes, allowing the mass spectrometer sufficient time for data acquisition (i.e., acquiring mass spectra and tandem mass spectra). In total, we detected 104 proteoforms or protein complexes in a mass range of 10-400 kDa after manual spectrum averaging and deconvolution. Information on the detected 104 proteoforms or protein complexes is listed in Table S1. **Figures 5.4 B-G** show the mass spectra of some examples larger than 40 kDa, i.e., ~41, 139, 143, 146, 318, 340, and 387 kDa. Those proteoforms or protein complexes show native-like and clear mass spectra. For example, **Figure 5.4 F** shows two co-migrating proteoforms or protein complexes with masses ~318 and ~340 kDa. Their most-abundance charge states are +34 and +36, respectively.

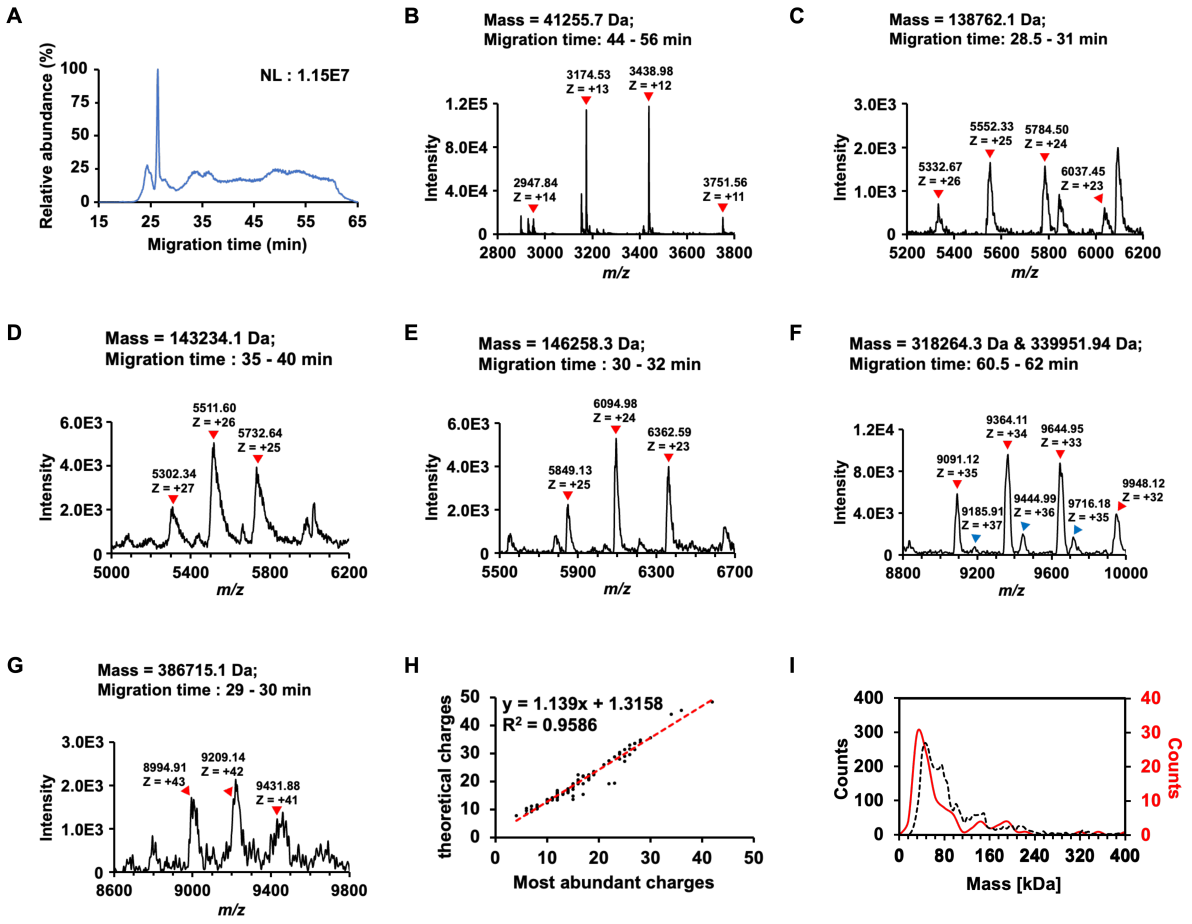


Figure 5.4. Summary of detected proteoforms or protein complexes from an *E. coli* cell lysate using nCZE-ESI-UHMR. (A) Representative electropherogram of nCZE-ESI-UHMR analyses of the *E. coli* cell lysate. (B)-(G) Mass spectra of six examples of large proteoforms/protein complexes detected. The charge states and deconvolved mass of each proteoform/protein complex are labelled. (H) Linear correlation between the most abundant charges and theoretical Rayleigh charges (Z_R) of all proteoforms/protein complexes detected in single-shot nCZE-UHMR. (I) Alignment of the mass distribution of proteoforms/protein complexes in the *E. coli* cell lysate from mass photometry (black dash line) and nCZE-UHMR (red line) analyses.

The largest proteoform or protein complex detected in this study is ~387 kDa, carrying around 42 charges (**Figure 5.4G**). Some additional examples are shown in **Figure 5.5**.

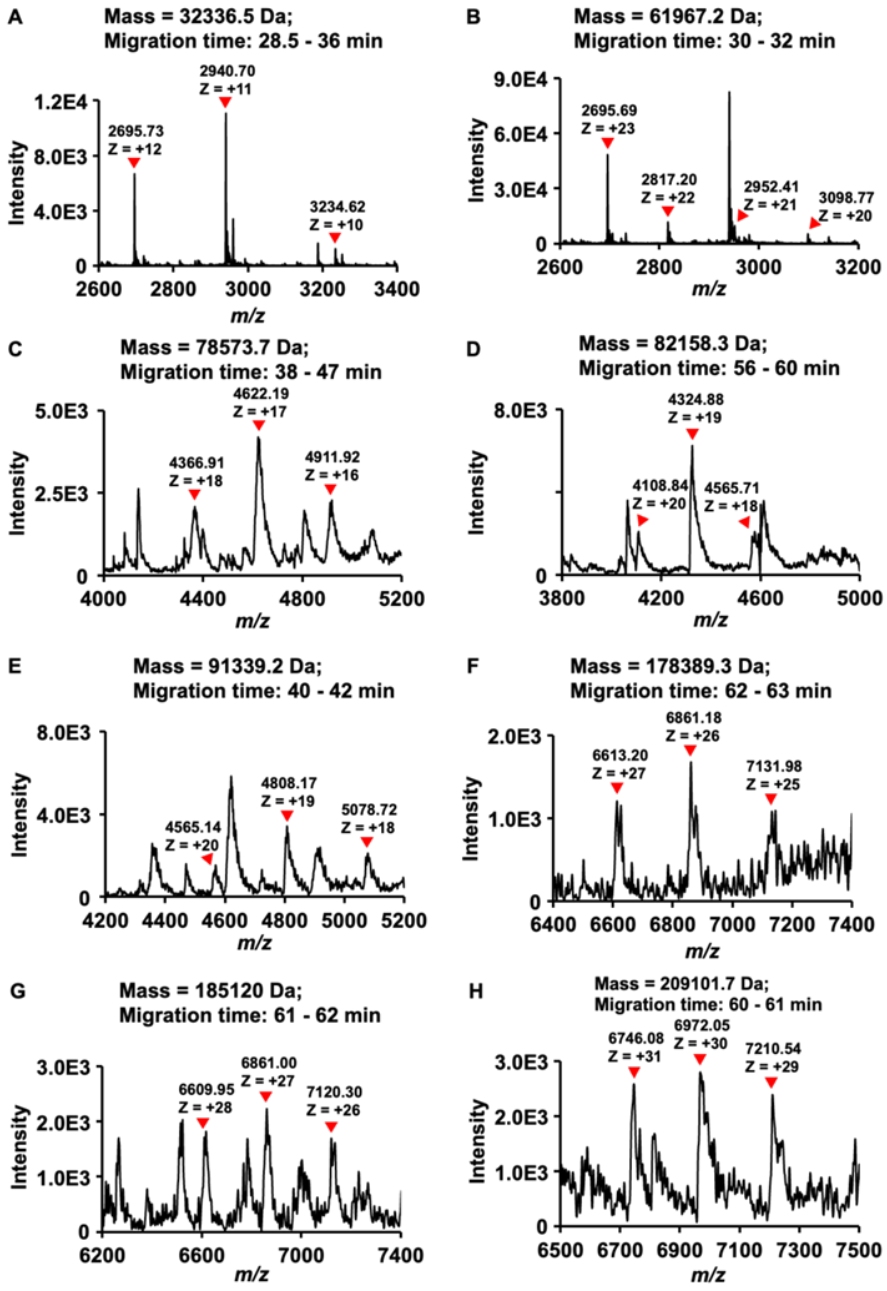


Figure 5.5. Representative mass spectra of proteoforms/protein complexes detected from the *E. coli* sample.

We further examined the correlation between the predicted Rayleigh charge (Z_R) from De La Mora's theory and the experimental maximum charge state of detected proteoforms or protein complexes, **Figure 5.4 H** [44,45]. We used the most abundant charge state instead of the highest charge state for each proteoform/protein complex here to avoid potential variations introduced during manual determination of the highest charge state. We observed a strong linear correlation

(R^2 = 0.96, slope of 1.14) between the experimental and predicted charge states. The slope indicates that the theoretical charges are slightly higher than the most abundant charges, suggesting the preservation of native states of the proteoforms or protein complexes in this experiment. We further employed mass photometry (MP) to measure the individual mass of proteoforms/protein complexes and their counts in the same *E. coli* cell lysate in a nearly physiological solution based on the quantification by light scattering [34,46,47].

The masses of proteoforms/protein complexes range from 10 kDa to 400 kDa according to the MP data, **Figure 5.4 I** (black dashed line). Nearly 70% of the molecule counts (2366 of 3555) from the MP analysis are in the mass range of 30 to 95 kDa. Interestingly, the molecular mass distributions from the MP and nCZE-MS analysis agree reasonably well, **Figure 5.4 I**, considering the low mass cutoff of MP. For example, the largest proteoform or protein complex detected by nCZE-MS is close to 400 kDa and 70% (73 out of 104) of the proteoforms/protein complexes from nCZE-MS are in the mass range of 32 to 96 kDa. It has been demonstrated that nMS and MP can produce reasonably consistent mass assessment of large proteins or protein complexes and offer complementary information about the analytes [48].

Our native proteomics study here is important because, for the first time, we can achieve a proteome-scale measurement of endogenous proteoforms and protein complexes in a complex biological sample under near physiological conditions by nMS with relatively high throughput. Over 100 endogenous intact proteoforms and protein complexes up to 400 kDa were detected from an *E. coli* cell lysate by online nCZE-MS in roughly 1-hour measurements with the consumption of 50-ng protein material. nCZE-MS can maintain the protein molecules from a complex cell lysate in close-to-native states during the measurement, evidenced by the strong linear correlation between the predicted Rayleigh charge (Z_R) and experimental most-abundance charge state of detected proteoforms or protein complexes, as well as the strong agreement in molecular mass distributions between the nCZE-MS and MP data.

The current study has two limitations. Firstly, we did not generate high-quality MS/MS data for the large proteoforms or protein complexes during the nCZE-MS run, impeding the identification of each protein. We will solve this issue by optimizing surface-induced dissociation or higher energy collisional dissociation (HCD) to achieve better fragmentation of large proteoforms or protein complexes in our future study. Second, the separations of large protein complexes by nCZE needs to be further improved regarding separation peak capacity and

reproducibility. **Figure 5.6** shows the electropherograms of the *E. coli* cell lysate after triplicate nCZE-MS measurements. The separation peak capacity is limited, which may be due to the non-specific adsorption of proteins onto the capillary inner wall. The separation profiles have some significant changes after 45 min in the second and third runs compared to the first run, most likely due to changes at the capillary inner wall after the first run of the *E. coli* sample. We need to develop procedures to clean up the capillary inner wall between nCZE-MS runs [49] and improve the capillary inner wall coating through different chemistries, e.g., carbohydrate-based neutral coating [27], to reduce protein adsorption for better separation peak capacity and reproducibility.

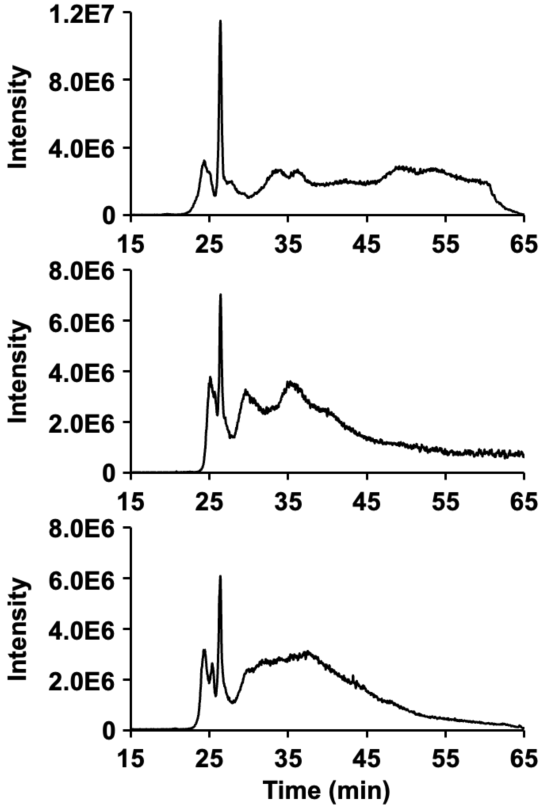


Figure 5.6. Electropherograms of an *E. coli* cell lysate after triplicate analyses by nCZE-ESI-UHMR. The electropherograms were aligned according to the most abundant peak.

5.4 Conclusions

In summary, we have demonstrated, for the first time, that nCZE coupled to an Orbitrap UHMR mass spectrometer is an effective and sensitive platform to measure large proteoforms or protein complexes up to 400 kDa from a complex proteome sample. This nCZE-MS technique enabled highly sensitive detection of standard protein complexes via consuming only pg amounts

of protein material. The technique successfully detected over one hundred proteoforms or protein complexes from an *E. coli* cell lysate in a mass range of 10-400 kDa. With further improvements in gas-phase fragmentation and nCZE separation peak capacity and reproducibility, we envision that nCZE-orbitrap UHMR will become a powerful tool in native proteomics of complex proteome samples.

5.5 Acknowledgments

The authors thank the support from the National Cancer Institute (NCI) through grant R01CA247863 (Sun), the National Institute of General Medical Sciences (NIGMS), through grants R01GM125991 (Sun) and R01GM118470 (Sun), and the National Science Foundation through the grant DBI1846913 (CAREER Award, Sun). This research was supported by NIH Native Mass Spectrometry-Guided Structural Biology Center (RM1GM149374 to V.H.W.)

REFERENCES

[1] Smith, L. M.; The Consortium for Top Down Proteomics; Kelleher, N. L. Proteoform: A Single Term Describing Protein Complexity. Nat. Methods 2013; 10 (3):186–187.

[2] Smith, L. M.; Agar, J. N.; Chamot-Rooke, J.; Danis, P. O.; Ge, Y.; Loo, J. A.; Paša-Tolić, L.; Tsybin, Y. O.; Kelleher, N. L.; The Consortium for Top-Down Proteomics. The Human Proteoform Project: Defining the Human Proteome. Sci. Adv. 2021; 7 (46).

[3] Alberts, B. The Cell as a Collection of Protein Machines: Preparing the next Generation of Molecular Biologists. Cell 1998; 92 (3):291–294.

[4] Liu, R.; Xia, S.; Li, H. Native Top-down Mass Spectrometry for Higher-Order Structural Characterization of Proteins and Complexes. Mass Spectrom. Rev. 2023; 42 (5):1876–1926.

[5] Tamara, S.; den Boer, M. A.; Heck, A. J. R. High-Resolution Native Mass Spectrometry. Chem. Rev. 2022; 122 (8):7269–7326.

[6] Karch, K. R.; Snyder, D. T.; Harvey, S. R.; Wysocki, V. H. Native Mass Spectrometry: Recent Progress and Remaining Challenges. Annu. Rev. Biophys. 2022; 51:157–179.

[7] Snyder, D. T.; Harvey, S. R.; Wysocki, V. H. Surface-Induced Dissociation Mass Spectrometry as a Structural Biology Tool. Chem. Rev. 2022; 122 (8):7442–7487.

[8] van de Waterbeemd, M.; Fort, K. L.; Boll, D.; Reinhardt-Szyba, M.; Routh, A.; Makarov, A.; Heck, A. J. R. High-Fidelity Mass Analysis Unveils Heterogeneity in Intact Ribosomal Particles. Nat. Methods 2017; 14 (3):283–286.

[9] Li, H.; Nguyen, H. H.; Ogorzalek Loo, R. R.; Campuzano, I. D. G.; Loo, J. A. An Integrated Native Mass Spectrometry and Top-down Proteomics Method That Connects Sequence to Structure and Function of Macromolecular Complexes. Nat. Chem. 2018; 10 (2):139–148.

[10] Fantin, S. M.; Parson, K. F.; Yadav, P.; Juliano, B.; Li, G. C.; Sanders, C. R.; Ohi, M. D.; Ruotolo, B. T. Ion Mobility-Mass Spectrometry Reveals the Role of Peripheral Myelin Protein Dimers in Peripheral Neuropathy. Proc. Natl. Acad. Sci. U. S. A. 2021; 118 (17).

[11] Keener, J. E.; Zambrano, D. E.; Zhang, G.; Zak, C. K.; Reid, D. J.; Deodhar, B. S.; Pemberton, J. E.; Prell, J. S.; Marty, M. T. Chemical Additives Enable Native Mass Spectrometry Measurement of Membrane Protein Oligomeric State within Intact Nanodiscs. J. Am. Chem. Soc. 2019; 141 (2):1054–1061.

[12] Vimer, S.; Ben-Nissan, G.; Morgenstern, D.; Kumar-Deshmukh, F.; Polkinghorn, C.; Quintyn, R. S.; Vasil'ev, Y. V.; Beckman, J. S.; Elad, N.; Wysocki, V. H.; Sharon, M. Comparative Structural Analysis of 20S Proteasome Ortholog Protein Complexes by Native Mass Spectrometry. ACS Cent. Sci. 2020; 6 (4):573–588.

[13] Gault, J.; Liko, I.; Landreh, M.; Shutin, D.; Bolla, J. R.; Jefferies, D.; Agasid, M.; Yen, H.-Y.; Ladds, M. J. G. W.; Lane, D. P.; Khalid, S.; Mullen, C.; Remes, P. M.; Huguet, R.; McAlister, G.; Goodwin, M.; Viner, R.; Syka, J. E. P.; Robinson, C. V. Combining Native and "omics" Mass Spectrometry to Identify Endogenous Ligands Bound to Membrane Proteins. Nat. Methods 2020; 17 (5):505–508.

[14] VanAernum, Z. L.; Busch, F.; Jones, B. J.; Jia, M.; Chen, Z.; Boyken, S. E.; Sahasrabuddhe, A.; Baker, D.; Wysocki, V. H. Rapid Online Buffer Exchange for Screening of Proteins, Protein Complexes and Cell Lysates by Native Mass Spectrometry. Nat. Protoc. 2020; 15 (3):1132–1157.

[15] Ventouri, I. K.; Veelders, S.; Passamonti, M.; Endres, P.; Roemling, R.; Schoenmakers, P. J.; Somsen, G. W.; Haselberg, R.; Gargano, A. F. G. Micro-Flow Size-Exclusion Chromatography for Enhanced Native Mass Spectrometry of Proteins and Protein Complexes. Anal. Chim. Acta 2023; 1266: 341324.

[16] Sahasrabuddhe, A.; Hsia, Y.; Busch, F.; Sheffler, W.; King, N. P.; Baker, D.; Wysocki, V. H. Confirmation of Intersubunit Connectivity and Topology of Designed Protein Complexes by Native MS. Proc. Natl. Acad. Sci. U. S. A. 2018; 115 (6):1268–1273.

[17] Ren, C.; Bailey, A. O.; VanderPorten, E.; Oh, A.; Phung, W.; Mulvihill, M. M.; Harris, S. F.; Liu, Y.; Han, G.; Sandoval, W. Quantitative Determination of Protein–Ligand Affinity by Size Exclusion Chromatography Directly Coupled to High-Resolution Native Mass Spectrometry. Anal. Chem. 2019; 91 (1):903–911.

[18] Busch, F.; VanAernum, Z. L.; Lai, S. M.; Gopalan, V.; Wysocki, V. H. Analysis of Tagged Proteins Using Tandem Affinity-Buffer Exchange Chromatography Online with Native Mass Spectrometry. Biochemistry 2021; 60 (24):1876–1884.

[19] Muneeruddin, K.; Nazzaro, M.; Kaltashov, I. A. Characterization of Intact Protein Conjugates and Biopharmaceuticals Using Ion-Exchange Chromatography with Online Detection by Native Electrospray Ionization Mass Spectrometry and Top-down Tandem Mass Spectrometry. Anal. Chem. 2015; 87 (19):10138–10145.

[20] Yan, Y.; Liu, A. P.; Wang, S.; Daly, T. J.; Li, N. Ultrasensitive Characterization of Charge Heterogeneity of Therapeutic Monoclonal Antibodies Using Strong Cation Exchange Chromatography Coupled to Native Mass Spectrometry. Anal. Chem. 2018; 90 (21):13013–13020.

[21] Debaene, F.; Bœuf, A.; Wagner-Rousset, E.; Colas, O.; Ayoub, D.; Corvaïa, N.; Van Dorsselaer, A.; Beck, A.; Cianférani, S. Innovative Native MS Methodologies for Antibody Drug Conjugate Characterization: High Resolution Native MS and IM-MS for Average DAR and DAR Distribution Assessment. Anal. Chem. 2014; 86 (21):10674–10683.

[22] Yan, Y.; Xing, T.; Wang, S.; Daly, T. J.; Li, N. Online Coupling of Analytical Hydrophobic Interaction Chromatography with Native Mass Spectrometry for the Characterization of Monoclonal Antibodies and Related Products. J. Pharm. Biomed. Anal. 2020; 186:113313.

[23] Chen, D.; McCool, E. N.; Yang, Z.; Shen, X.; Lubeckyj, R. A.; Xu, T.; Wang, Q.; Sun, L. Recent Advances (2019-2021) of Capillary Electrophoresis-Mass Spectrometry for Multilevel Proteomics. Mass Spectrom. Rev. 2023; 42 (2):617–642.

[24] Jooß, K.; McGee, J. P.; Kelleher, N. L. Native Mass Spectrometry at the Convergence of Structural Biology and Compositional Proteomics. Acc. Chem. Res. 2022; 55 (14):1928–1937.

[25] Skinner, O. S.; Haverland, N. A.; Fornelli, L.; Melani, R. D.; Do Vale, L. H. F.; Seckler, H. S.; Doubleday, P. F.; Schachner, L. F.; Srzentić, K.; Kelleher, N. L.; Compton, P. D. Top-down Characterization of Endogenous Protein Complexes with Native Proteomics. Nat. Chem. Biol. 2018; 14 (1):36–41.

[26] Chapman, E. A.; Li, B. H.; Krichel, B.; Chan, H.-J.; Buck, K. M.; Roberts, D. S.; Ge, Y. Native Top-down Mass Spectrometry for Characterizing Sarcomeric Proteins Directly from Cardiac Tissue Lysate. J. Am. Soc. Mass Spectrom. 2024; 35 (4):738–745.

[27] Shen, X.; Liang, Z.; Xu, T.; Yang, Z.; Wang, Q.; Chen, D.; Pham, L.; Du, W.; Sun, L. Investigating Native Capillary Zone Electrophoresis-Mass Spectrometry on a High-End Quadrupole-Time-of-Flight Mass Spectrometer for the Characterization of Monoclonal Antibodies. Int. J. Mass Spectrom. 2021; 462:116541.

[28] Marie, A.-L.; Georgescauld, F.; Johnson, K. R.; Ray, S.; Engen, J. R.; Ivanov, A. R. Native Capillary Electrophoresis–Mass Spectrometry of near 1 MDa Non-covalent GroEL/GroES/Substrate Protein Complexes. Adv. Sci. (Weinh.) 2024; 11.

[29] Jooß, K.; McGee, J. P.; Melani, R. D.; Kelleher, N. L. Standard Procedures for Native CZE-MS of Proteins and Protein Complexes up to 800 KDa. Electrophoresis 2021; 42 (9–10):1050–1059.

[30] Belov, A. M.; Viner, R.; Santos, M. R.; Horn, D. M.; Bern, M.; Karger, B. L.; Ivanov, A. R. Analysis of Proteins, Protein Complexes, and Organellar Proteomes Using Sheathless Capillary Zone Electrophoresis - Native Mass Spectrometry. J. Am. Soc. Mass Spectrom. 2017; 28 (12):2614–2634.

[31] Mehaffey, M. R.; Xia, Q.; Brodbelt, J. S. Uniting Native Capillary Electrophoresis and Multistage Ultraviolet Photodissociation Mass Spectrometry for Online Separation and Characterization of Escherichia Coli Ribosomal Proteins and Protein Complexes. Anal. Chem. 2020; 92 (22), 15202–15211.

[32] Jooß, K.; Schachner, L. F.; Watson, R.; Gillespie, Z. B.; Howard, S. A.; Cheek, M. A.; Meiners, M. J.; Sobh, A.; Licht, J. D.; Keogh, M.-C.; Kelleher, N. L. Separation and Characterization of Endogenous Nucleosomes by Native Capillary Zone Electrophoresis–Top-down Mass Spectrometry. Anal. Chem. 2021; 93 (12):5151–5160.

[33] Shen, X.; Kou, Q.; Guo, R.; Yang, Z.; Chen, D.; Liu, X.; Hong, H.; Sun, L. Native Proteomics in Discovery Mode Using Size-Exclusion Chromatography–Capillary Zone Electrophoresis–Tandem Mass Spectrometry. Anal. Chem. 2018; 90 (17):10095–10099.

[34] Wu, D.; Piszczek, G. Standard Protocol for Mass Photometry Experiments. Eur. Biophys. J. 2021; 50 (3–4):403–409.

[35] McCool, E. N.; Lubeckyj, R.; Shen, X.; Kou, Q.; Liu, X.; Sun, L. Large-Scale Top-down Proteomics Using Capillary Zone Electrophoresis Tandem Mass Spectrometry. J. Vis. Exp. 2018; No. 140.

[36] Zhu, G.; Sun, L.; Dovichi, N. J. Thermally-Initiated Free Radical Polymerization for Reproducible Production of Stable Linear Polyacrylamide Coated Capillaries, and Their Application to Proteomic Analysis Using Capillary Zone Electrophoresis–Mass Spectrometry. Talanta 2016; 146:839–843.

[37] Sun, L.; Zhu, G.; Zhao, Y.; Yan, X.; Mou, S.; Dovichi, N. J. Ultrasensitive and Fast Bottom-up Analysis of Femtogram Amounts of Complex Proteome Digests. Angew. Chem. Int. Ed Engl. 2013; 52 (51):13661–13664.

[38] Wojcik, R.; Dada, O. O.; Sadilek, M.; Dovichi, N. J. Simplified Capillary Electrophoresis Nanospray Sheath-flow Interface for High Efficiency and Sensitive Peptide Analysis. Rapid Commun. Mass Spectrom. 2010; 24 (17):2554–2560.

[39] Sun, L.; Zhu, G.; Zhang, Z.; Mou, S.; Dovichi, N. J. Third-Generation Electrokinetically Pumped Sheath-Flow Nanospray Interface with Improved Stability and Sensitivity for

Automated Capillary Zone Electrophoresis–Mass Spectrometry Analysis of Complex Proteome Digests. J. Proteome Res. 2015; 14 (5):2312–2321.

[40] Marty, M. T.; Baldwin, A. J.; Marklund, E. G.; Hochberg, G. K. A.; Benesch, J. L. P.; Robinson, C. V. Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. Anal. Chem. 2015; 87 (8):4370–4376.

[41] Winkler, R. ESIprot: A Universal Tool for Charge State Determination and Molecular Weight Calculation of Proteins from Electrospray Ionization Mass Spectrometry Data. Rapid Commun. Mass Spectrom. 2010; 24 (3):285–294.

[42] Xu, T.; Han, L.; Sun, L. Automated Capillary Isoelectric Focusing-Mass Spectrometry with Ultrahigh Resolution for Characterizing Microheterogeneity and Isoelectric Points of Intact Protein Complexes. Anal. Chem. 2022; 94 (27):9674–9682.

[43] Schachner, L. F.; Ives, A. N.; McGee, J. P.; Melani, R. D.; Kafader, J. O.; Compton, P. D.; Patrie, S. M.; Kelleher, N. L. Standard Proteoforms and Their Complexes for Native Mass Spectrometry. J. Am. Soc. Mass Spectrom. 2019; 30 (7):1190–1198.

[44] Fernandez de la Mora, J. Electrospray Ionization of Large Multiply Charged Species Proceeds via Dole's Charged Residue Mechanism. Anal. Chim. Acta 2000; 406 (1):93–104.

[45] Heck, A. J. R.; van den Heuvel, R. H. H. Investigation of Intact Protein Complexes by Mass Spectrometry. Mass Spectrom. Rev. 2004; 23 (5):368–389.

[46] Cole, D.; Young, G.; Weigel, A.; Sebesta, A.; Kukura, P. Label-Free Single-Molecule Imaging with Numerical-Aperture-Shaped Interferometric Scattering Microscopy. ACS Photonics 2017; 4 (2):211–216.

[47] Young, G.; Hundt, N.; Cole, D.; Fineberg, A.; Andrecka, J.; Tyler, A.; Olerinyova, A.; Ansari, A.; Marklund, E. G.; Collier, M. P.; Chandler, S. A.; Tkachenko, O.; Allen, J.; Crispin, M.; Billington, N.; Takagi, Y.; Sellers, J. R.; Eichmann, C.; Selenko, P.; Frey, L.; Riek, R.; Galpin, M. R.; Struwe, W. B.; Benesch, J. L. P.; Kukura, P. Quantitative Mass Imaging of Single Biological Macromolecules. Science 2018; 360 (6387):423–427.

[48] den Boer, M. A.; Lai, S.-H.; Xue, X.; van Kampen, M. D.; Bleijlevens, B.; Heck, A. J. R. Comparative Analysis of Antibodies and Heavily Glycosylated Macromolecular Immune Complexes by Size-Exclusion Chromatography Multi-Angle Light Scattering, Native Charge Detection Mass Spectrometry, and Mass Photometry. Anal. Chem. 2022; 94 (2):892–900.

[49] Sadeghi, S. A.; Chen, W.; Wang, Q.; Wang, Q.; Fang, F.; Liu, X.; Sun, L. Pilot Evaluation of the Long-Term Reproducibility of Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Top-Down Proteomics of a Complex Proteome Sample. J Proteome Res 2024; 23 (1):1399–1407.

CHAPTER 6. Conclusion and future directions

6.1 Conclusion of current challenges and conclusion

Proteomics study to interrogate sequence and PTM features of proteoforms is crucial for understanding their roles in complex biological and disease events. Compared to conventional BUP, which acquires regional information about proteoforms by mapping peptides, TDP deciphers the combinatorial PTMs and amino acid variations in individual proteoforms, enabling a comprehensive insight into the proteoform landscape of biological subjects [1].

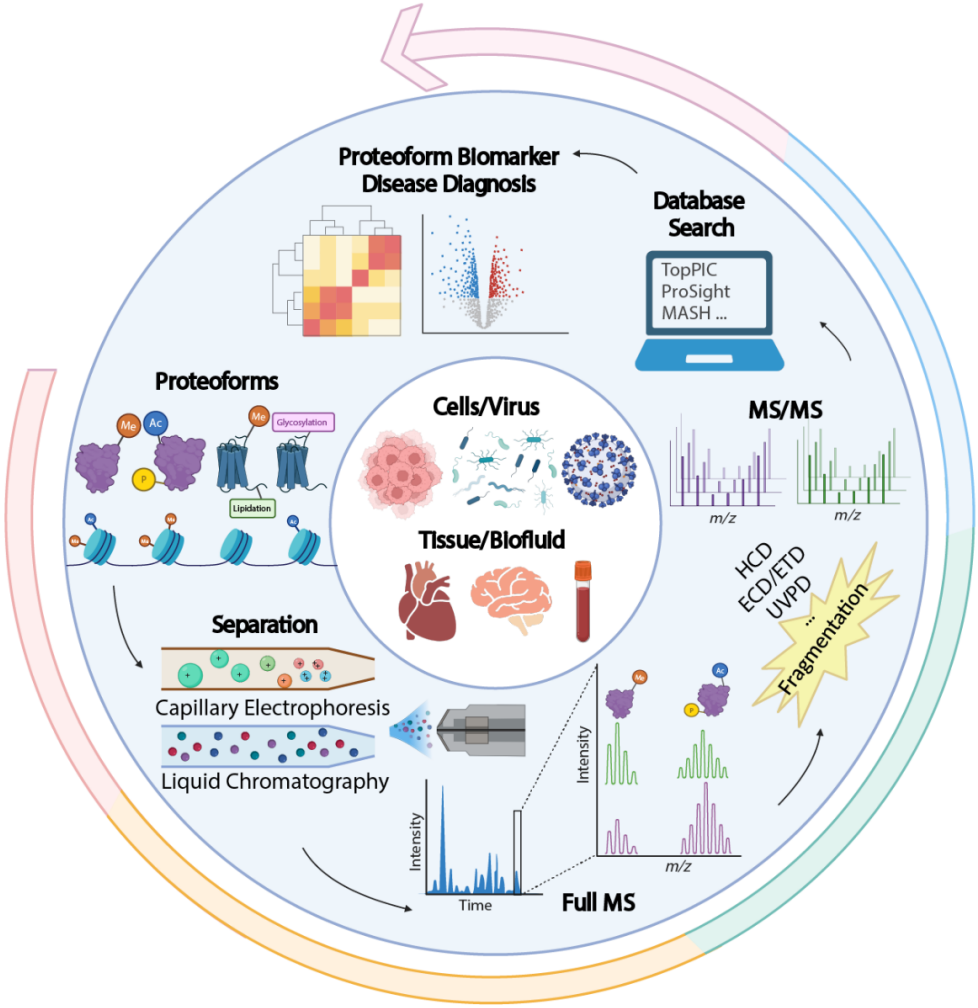


Figure 6.1. Schematic of a typical top-down proteomics (TDP) workflow, including proteoform extraction from various biological samples, separation by liquid chromatography (LC) or capillary electrophoresis (CE), mass spectrometry (MS) and tandem MS (MS/MS) via various gas-phase fragmentation techniques, and database search using bioinformatics tools for proteoform identification and quantification to discover new proteoform biomarkers of diseases. The figure is created using the BioRender and is used here with permission.

A universal TDP workflow includes protein extraction from biological samples, separation of proteoforms by liquid chromatography (LC) or capillary electrophoresis (CE), proteoform mass measurement and fragmentation by state-of-the-art mass spectrometers, and interpretation of proteomics data by advanced bioinformatics tools, as shown in **Figure 6.1**.

This dissertation starts with the first large-scale TDP analysis on *Arabidopsis thaliana* leaf and chloroplast samples. We identified 3198 and 1836 proteoforms identified from leaf and chloroplast, respectively. Proteoforms with phosphorylation and acetylation were validated by their electrophoretic mobility shifts [2]. Additionally, the direct evidence of N- and C- terminal sequencing allowed precise delineation true transit peptide cleavage site, providing important information to plant biologists.

However, this large-scale TDP study mainly focused on hydrophilic proteoforms which are lower than 30 kDa. MS-based TDP was considered challenging in multiple aspects. First, complete characterization of large proteoforms is difficult due to coelution in separation, low sensitivity in MS detection, and insufficient gas-phase fragmentation. Second, MS-intensive TDP of membrane proteins is challenging due to their low solubility in MS-compatible buffers. Third, delineating heavily modified proteoforms (e.g., histones) requires much more effort to improve separations, fragmentations, and bioinformatics tools for accurate PTM determination and localization. Fourth, achieving the identification (ID) of proteoforms of thousands of protein-coding genes from human cells in a single study is challenging for MS-intensive TDP due to the extremely high sample complexity.

Addressing the outlined challenges, this dissertation focused on the advancements in the front-end separation using the high-sensitivity CZE. We established a methodology for the separation of intact integral membrane proteins using CZE-MS/MS, identifying 343 proteoforms with 80% featuring at least one transmembrane domain [3]. We also developed a simple and efficient approach for creating cationic coatings with tunable electroosmotic flow in CZE/MS/MS, which minimized protein adsorption and enabled the detection of large proteoforms up to 45 kDa. Furthermore, we applied native CZE-ESI-UHMR to detect large proteins and protein complexes up to 318 kDa [4], demonstrating CZE's adaptability of various protein samples under multiple separation conditions. Notability, these advancements did not utilize online concentration methods for CZE, limiting the loading capacity. Therefore, future

work focusing on enhancing sample loading capacity to expand the scope of these techniques is crucial.

6.2 Future direction – Reproducibility and Robustness of CZE-MS

CZE emerged in the 1980s as a key technique for separating large biomolecules, especially proteoforms, driven by differences in their electrophoretic mobility (μ_{ep}), which is proportional to the ratio of a proteoform's net charge to its hydrodynamic radius [5, 6]. With lower diffusion coefficients than peptides and metabolites, proteoforms achieve high separation efficiency in CZE, which is particularly effective for those with charge inducing PTMs [6]. CZE has minimal sample consumption, low flow rates, and high peak capacity, reaching nearly one million theoretical plates in proteoform separation [7]. Recent characterization of three human issues demonstrating the complementary separation of CZE and conventional RPLC (charge-to-size ratio vs. hydrophobicity), with 28% of proteins and 56% of proteoforms were uniquely identified by CZE-MS, highlighting the gain and complementarity of CZE-MS in TDP [8].

Despite its advantages, CZE-MS integration in TDP has been limited until recently due to concerns about its robustness, reproducibility, and the technical challenges of CZE-MS interfacing, such as the necessity to apply an electric field across the separation capillary. Significant advancements in industrial and academic settings over recent decades have enhanced the sensitivity and robustness of CZE-MS interfacing. Innovations such as nano sheath-flow and sheathless interfaces have made CZE-MS more user-friendly and accessible [9-16]. In addition, different types of CE platforms, including capillary- and microfluidic chip-based, have been established over the past years [10-13]. Advances in CE has been applied in diverse fields such as biopharmaceutical characterization [17-20], global proteoforms profiling [21-23], single cells analysis [11, 24-29], and disease-related protein studies [30,31].

Thus, reevaluating the performance of CZE-MS for TDP is a vital future direction that promises to draw increased attention to and further development of CE-MS-based techniques. This effort will be supported by an extensive study of identical samples from a global cross-laboratory study, which spans a comprehensive range of commercially available CE systems for TDP applications.

6.3 Future direction – Combining BUP and TDP

PGs have been suggested to serve as regulatory and enzymatic hubs of chloroplast membrane remodeling and adaptation based on mutant phenotypes and gene co-expression

patterns [32,33]. Understanding how proteins are localized on the PG and dynamic recruited to PG are important for facilitating plant resilience under conditions of stress and senescence. It is known that protein targeting to cytosolic lipid droplets is mediated either by a hydrophobic hairpin loop or a n amphipathic helix (AH) [34-36]. A recent study showing that PG-localized Fibrillins harbor an amphipathic helix at the lip of their β-barrel that is necessary for proper PG association, as shown in **Figure 6.2 A** [37]. Structural comparison of the PG-localized AtFBN1a and the thylakoid-localized AtFBN3a is shown in **Figure 6.2 B**.

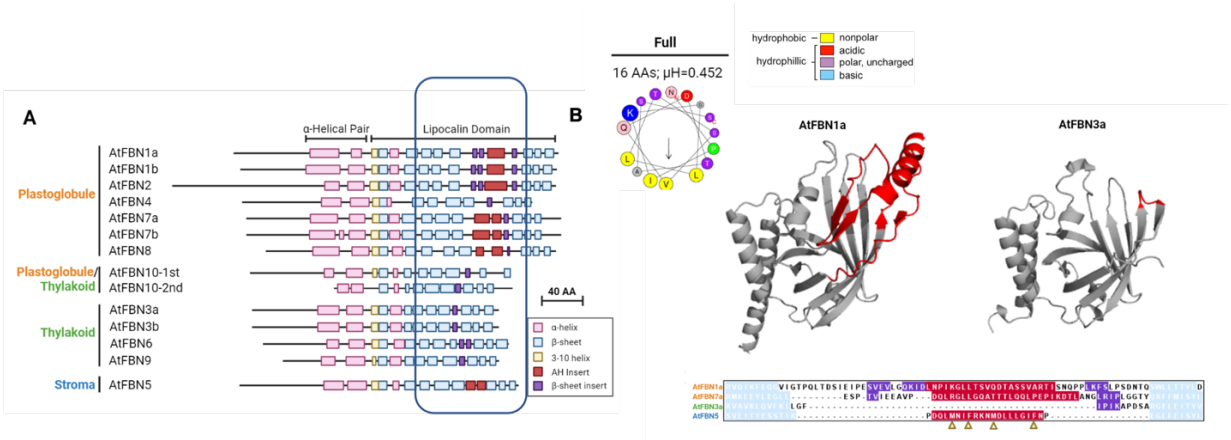


Figure 6.2. (A). The Schematic of secondary structures of all thirteen *A. thaliana* FBN proteins, as predicted by AlphaFold. The schematic is grouped by sub-plastidic localization, and each sequence is drawn to scale and aligned based on the first β-sheet of the lipocalin domain which contains the absolutely conserved GxW motif. The 3-11 helix is presumed based on the helical structure of the AlphaFold model and the solved structures of other lipocalin domains. Note that AtFBN10 contains two repeated lipocalin domains which are illustrated on separate lines in the schematic; where the first sequence drops off, the sequence is then directly continued on the second sequence without un-captured gaps. (B). Structural comparison of the plastoglobule-localized AtFBN1a and the thylakoid-localized AtFBN3a. Sequence highlighted in red indicates the 45-residue sequence surrounding the predicted AH in AtFBN1a and the corresponding sequence region in AtFBN3a which lacks any amphipathic helical insert. Sequence alignment of the amphipathic helical insert region from selected FBN proteins. The alignment highlights sequence differences, in particular the absence of large hydrophobic side chains from AtFBN1a and AtFBN7a, which are present in the non-plastoglobule-localized AtFBN5 AH and noted with gold arrowheads under the alignment. The figure is adapted from reference [37].

We hypothesized that the C-terminal amphipathic helix plays a crucial role in anchoring FBN proteins to the surface of PGs. As illustrated in **Figure 6.2**, we proposed a combined approach of top-down and bottom-up proteomics to investigate this hypothesis. This procedure starts by incubating PGs with trypsin overnight, allowing for selective proteolytic digestion of the accessible region. Subsequently, the flow-through peptides are collected for analysis. The

remaining PGs could then be processed to extract the membrane embedded proteins, potentially anchor segment of the PG-proteins. Comprehensive quantitative BUP utilizing both CE-MS/MS and LC-MS/MS, analyzes peptide abundance in both flow-through and retained fractions, pinpointing the regions critical for localization on PGs. TDP will depict a complete picture of these critical regions with their PTM states, revealing how these modifications influence protein localization. The combined BUP and TDP to study PG-localized protein under different condition will provide more information of the dynamic recruitment and interaction of these proteins.

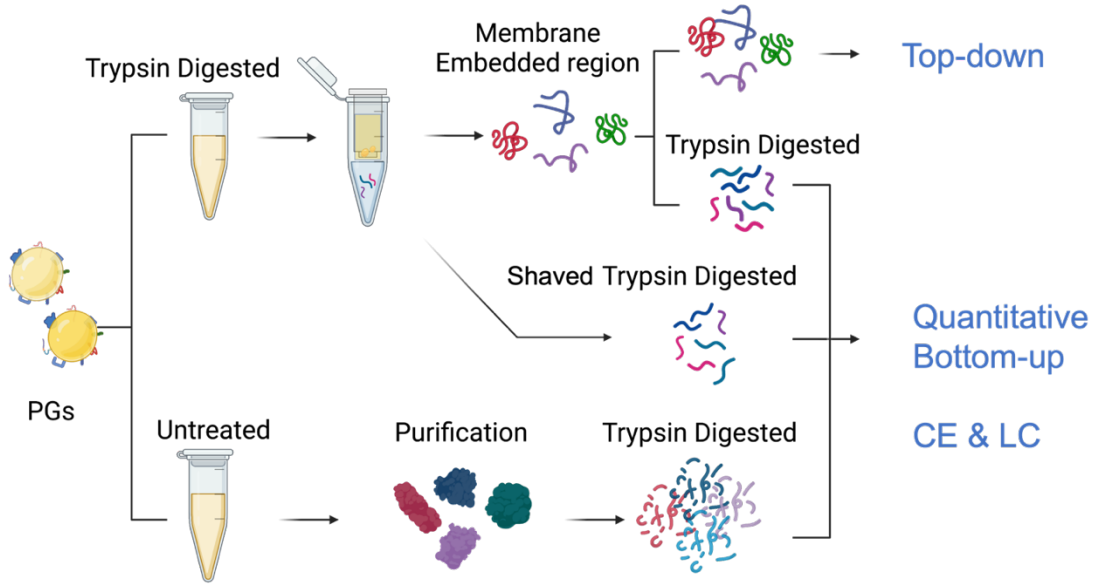


Figure 6.2. Schematic of the proposed PG shaving experiment combining bottom-up proteomics (BUP) and top-down proteomics (TDP) workflow. The process begins with overnight trypsin treatment of PGs to release accessible peptides, followed by quantitative BUP analysis using capillary electrophoresis and liquid chromatography. TDP of the membrane embedded region validate the membrane embedded region with relative post-translational modifications.

Looking ahead, the future direction of integrating BUP and TDP will facilitate a more detailed understanding of protein modifications, interactions, and functions at a molecular level.

REFERENCES

[1] Schaffer, LV., Millikin, RJ., Miller, RM., et al. Identification and Quantification of Proteoforms by Mass Spectrometry. Proteomics. 2019; 19(10):e1800361.

[2] Wang, Q., Lundquist, PK., Sun, L. Large-scale top-down proteomics of the Arabidopsis thaliana leaf and chloroplast proteomes. Proteomics. 2022; 23(3-4):2100277.

[3] Wang, Q., Xu, T., Fang, F., Wang, Q., Lundquist, PK., Sun, L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry for Top-Down Proteomics of Mouse Brain Integral Membrane Proteins. Anal Chem. 2023; 95(34):12590-12594.

[4] Wang, Q., Wang, Q., Qi, Z., Moeller, W., Wysocki, VH., Sun, L. Native proteomics by capillary zone electrophoresis-mass spectrometry. bioRxiv, 2024.04. 24.590970

[5] Chen, D., et al., Recent advances (2019-2021) of capillary electrophoresis-mass spectrometry for multilevel proteomics. Mass Spectrom Rev, 2023. 42(2): p. 617-642.

[6] Jorgenson, J.W. and K.D. Lukacs, Capillary Zone Electrophoresis. Science, 1983. 222(4621): p. 266-272.

[7] Lubeckyj, R.A., et al., 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, 2019. 30(8): p. 1435-1445.

[8] Drown, B.S., et al., Mapping the Proteoform Landscape of Five Human Tissues. Journal of Proteome Research, 2022. 21(5): p. 1299-1310.

[9] Wojcik, R., et al, Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. Rapid Commun Mass Spectrom, 2010. 24(17): p. 2554-60.

[10] Sun, L., et al., Third-generation electrokinetically pumped sheath-flow nanospray interface with improved stability and sensitivity for automated capillary zone electrophoresis-mass spectrometry analysis of complex proteome digests. J Proteome Res, 2015. 14(5): p. 2312-21.

[11] Choi, S.B., A.M. Polter, and P. Nemes, Patch-Clamp Proteomics of Single Neurons in Tissue Using Electrophysiology and Subcellular Capillary Electrophoresis Mass Spectrometry. Anal Chem, 2022. 94(3): p. 1637-1644.

[12] Schlecht, J., et al., nanoCEasy: An Easy, Flexible, and Robust Nanoflow Sheath Liquid Capillary Electrophoresis-Mass Spectrometry Interface Based on 3D Printed Parts. Anal Chem, 2021. 93(44): p. 14593-14598.

[13] Carillo, S., C. Jakes, and J. Bones, In-depth analysis of monoclonal antibodies using microfluidic capillary electrophoresis and native mass spectrometry. J Pharm Biomed Anal, 2020. 185: p. 113218.

[14] Maxwell, E.J. and D.D. Chen, Twenty years of interface development for capillary electrophoresis-electrospray ionization-mass spectrometry. Anal Chim Acta, 2008. 627(1): p. 25-33.

[15] Moini, M., Simplifying CE-MS operation. 2. Interfacing low-flow separation techniques to mass spectrometry using a porous tip. Anal Chem, 2007. 79(11): p. 4241-6.

[16] Zhong, X., et al., Flow-through microvial facilitating interface of capillary isoelectric focusing and electrospray ionization mass spectrometry. Anal Chem, 2011. 83(22): p. 8748-55.

[17] Belov, A.M., et al., Complementary middle-down and intact monoclonal antibody proteoform characterization by capillary zone electrophoresis - mass spectrometry. Electrophoresis, 2018. 39(16): p. 2069-2082.

[18] Bush, D.R., et al., High Resolution CZE-MS Quantitative Characterization of Intact Biopharmaceutical Proteins: Proteoforms of Interferon-β1. Anal Chem, 2016. 88(2): p. 1138-46.

[19] Khawli, L.A., et al., Charge variants in IgG1: Isolation, characterization, in vitro binding properties and pharmacokinetics in rats. MAbs, 2010. 2(6): p. 613-24.

[20] Shah, A., et al., Characterization of bispecific antigen-binding biotherapeutic fragmentation sites using microfluidic capillary electrophoresis coupled to mass spectrometry (mCZE-MS). Analyst, 2023. 148(3): p. 665-674.

[21] McCool, E.N., et al., Deep top-down proteomics revealed significant proteoform-level differences between metastatic and nonmetastatic colorectal cancer cells. Sci Adv, 2022. 8(51): p. eabq6348.

[22] McCool, E.N., et al., Deep Top-Down Proteomics Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry: Identification of 5700 Proteoforms from the Escherichia coli Proteome. Anal Chem, 2018. 90(9): p. 5529-5533.

[23] Gomes, F.P., et al., EThcD and 213 nm UVPD for Top-Down Analysis of Bovine Seminal Plasma Proteoforms on Electrophoretic and Chromatographic Time Frames. Anal Chem, 2020. 92(4): p. 2979-2987.

[24] Bagwe, K., et al., Single-cell omic molecular profiling using capillary electrophoresis-mass spectrometry. Trends Analyt Chem, 2023. 165.

[25] Johnson, K.R., et al., On-capillary Cell Lysis Enables Top-down Proteomic Analysis of Single Mammalian Cells by CE-MS/MS. Analytical Chemistry, 2022. 94(41): p. 14358-14367.

[26] Lombard-Banek, C., S.A. Moody, and P. Nemes, Single-Cell Mass Spectrometry for Discovery Proteomics: Quantifying Translational Cell Heterogeneity in the 16-Cell Frog (Xenopus) Embryo. Angew Chem Int Ed Engl, 2016. 55(7): p. 2454-8.

[27] Melby, J.A., et al., High sensitivity top-down proteomics captures single muscle cell heterogeneity in large proteoforms. Proc Natl Acad Sci U S A, 2023. 120(19): p. e2222081120.

[28] Mellors, J.S., et al., Integrated microfluidic device for automated single cell analysis using electrophoretic separation and electrospray ionization mass spectrometry. Anal Chem, 2010. 82(3): p. 967-73.

[29] Vaclavek, T. and F. Foret, Microfluidic device integrating single-cell extraction and electrical lysis for mass spectrometry detection of intracellular compounds. Electrophoresis, 2022.

[30] Mao, P. and D. Wang, Top-down proteomics of a drop of blood for diabetes monitoring. J Proteome Res, 2014. 13(3): p. 1560-9.

[31] Wei, L., et al., Novel Sarcopenia-related Alterations in Sarcomeric Protein Post-translational Modifications (PTMs) in Skeletal Muscles Identified by Top-down Proteomics. Mol Cell Proteomics, 2018. 17(1): p. 134-145.

[32] Lundquist PK, Poliakov A, Bhuiyan NH, et al. The functional network of the Arabidopsis plastoglobule proteome based on quantitative proteomics and genome-wide coexpression analysis. Plant Physiol. 2012;158:1172–1192.

[33] Lundquist PK, Poliakov A, Giacomelli L, et al. Loss of Plastoglobule Kinases ABC1K1 and ABC1K3 Causes Conditional Degreening, Modified Prenyl-Lipids, and Recruitment of the Jasmonic Acid Pathway. The Plant Cell. 2013;25:1818–1839.

[34]. Kory, N., Farese, R. V., Jr., and Walther, T. C. (2016) Targeting Fat: Mechanisms of Protein Localization to Lipid Droplets Trends Cell Biol 26, 535–546

[35]. Dhiman, R., Caesar, S., Thiam, A. R., and Schrul, B. (2020) Mechanisms of protein targeting to lipid droplets: A unified cell biological and biophysical perspective Semin Cell Dev Biol 108, 4–13

[36]. Olarte, M. J., Swanson, J. M. J., Walther, T. C., and Farese, R. V., Jr.. (2022) The CYTOLD and ERTOLD pathways for lipid droplet-protein targeting Trends Biochem Sci 47, 39–51

[37] Shivaiah, KK., Boren, DM., Tequia-Herrera, A., Vermaas, J., Lundquist, PK. An amphipathic helix drives interaction of Fibrillins with plastoglobuli lipid droplets. Biorxiv, 2023.09.28.559984.