ADVANCING ANALYTICAL TECHNIQUES FOR MASS SPECTROMETRY BASED-MULTI-LEVEL PROTEOMICS

By

Qianyi Wang

A DISSERTATION

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

Chemistry – Doctor of Philosophy

2024

ABSTRACT

Proteomics, the intricate study of the proteome, has evolved significantly with advancements in mass spectrometry (MS)-based techniques. While bottom-up proteomics (BUP), relying on proteolytic peptides, offers extensive protein identification, it grapples with challenges like limited sequence coverage and ambiguity in proteoform differentiation. Conversely, top-down proteomics (TDP), targeting intact proteoforms, provides a comprehensive view, capturing post-translational modifications (PTMs) but is constrained by low sensitivity and complex fragmentation. Analytical techniques like capillary zone electrophoresis (CZE) and liquid chromatography (LC) play pivotal roles in enhancing separation efficiency, while ion mobility spectrometry (IMS) complements mass spectrometry by offering an additional dimension of gas-phase separation. The unity of multi-dimensional separations and cutting-edge bioinformatics tools has expanded the horizons of proteomics, yet challenges remain.

In chapter 2, we engineered a high-throughput BUP workflow for plasma and serum analysis by integrating nanoparticle (NP) protein corona formation, rapid on-bead tryptic digestion, and CZE-tandem mass spectrometry (CZE-MS/MS). Four distinct magnetic NPs with varied functional groups on their surfaces using SDS-PAGE and CZE-MS/MS on healthy human plasma were firstly evaluated. The optimized workflow with amine-terminated and carboxylate-terminated NPs to analyze serum samples from both healthy and NUT cancer-afflicted mice was applied. This approach facilitated the identification of hundreds of proteins from plasma and serum samples, achieving high throughput within only 3.5-hours from sample to data. Leveraging the NP protein corona, rapid digestion, and CZE-MS/MS, we unveiled potential cancer biomarkers through quantitative proteomics.

In chapter 3, we explored magnetic NP-based immobilized metal affinity chromatography (IMAC) using Ti^{4+} and Fe^{3+} for the selective enrichment of phosphoproteoforms from a standard protein mixture and yeast cell lysate. This method demonstrated reproducible, high-efficiency enrichment, outperforming a commercial phosphoprotein enrichment kit in terms of capture efficiency and recovery. Reversed-phase LC (RPLC)-MS/MS analysis of yeast cell lysates post-IMAC enrichment yielded nearly 100% more phosphoproteoform identifications than without enrichment. Intriguingly, phosphoproteoforms identified after Ti^{4+}-IMAC or Fe^{3+}-IMAC enrichment corresponded to proteins of significantly lower abundance and distinct pools, suggesting their combination could enhance phosphoproteome coverage. These findings

underscore the potential of our magnetic NP-based Ti^{4+}-IMAC and Fe^{3+}-IMAC in advancing top-down MS characterization of phosphoproteoforms in complex biological systems.

In chapter 4, we introduced the first integration of CZE, IMS, and MS for online multi-dimensional separations of histone proteoforms. This innovative CZE-high-field asymmetric waveform IMS (FAIMS)-MS/MS platform enabled the identification of 366 histone proteoforms (ProSight PD) and 602 (TopPIC) from a commercial calf histone sample, using only a low microgram starting material. Remarkably, CZE-FAIMS-MS/MS achieved a threefold increase in histone proteoform identifications compared to CZE-MS/MS alone. These findings suggested that CZE-FAIMS-MS/MS holds significant potential for the comprehensive and highly sensitive characterization of histone proteoforms, paving the way for deeper insights into their complex roles in epigenetic control.

In chapter 5, we advanced native proteomics by analyzing large proteoforms and protein complexes, up to 400 kDa, from complex proteomes using native CZE (nCZE) coupled with an ultra-high mass range (UHMR) Orbitrap mass spectrometer. Our nCZE-MS technique successfully measured a 115-kDa standard protein complex with minimal sample consumption of only 0.1 ng. When applied to an *E. coli* cell lysate, nCZE-MS detected 72 proteoforms and complexes in the 30-400 kDa range from only 50 ng of material in a single run. Notably, the mass distribution of detected proteoforms and complexes was consistent with mass photometry measurements, marking a technical leap in native proteomics for complex proteome analysis.

This thesis is dedicated to my parents.
Thank you for always supporting me and believing in me.

ACKNOWLEDGEMENTS

Principally, I would like to express my deepest gratitude to my advisor Dr. Liangliang Sun for his invaluable guidance, constant support, and unwavering patience throughout my graduate study. His dedication, thoughtful critique, and encouragement are essential to my academic growth and the successful completion of this thesis. Discussing both research and professional advice with him is always a stress-free experience, and he is always available to guide students with hands-on support. He encourages us a lot to engage in outdoor activities rather than marinating at home or spending the entire week in the lab. He also consistently provides us with opportunities to attend academic conferences, research symposiums, collaborative projects, and internships, helping us broaden our perspective on cutting-edge fields and build valuable professional connections. I feel so lucky to have had the chance to work with him and learn from him.

I want to show my sincerest thanks to Dr. Dana Spence, Dr. Gary Blanchard, and Dr. Jetze Tepe as my research committee members. I greatly appreciate their questions and suggestions during my first committee meeting and the comprehensive exam. These interactions made me realize that there is always more to learn, and they reinforced the importance of continually asking 'why' throughout my scientific journey.

I would like to extend my appreciation to my internship mentor Weiwen Sun and my manager Kévin Contrepois for their amazing mentorship for this challenging project at AstraZeneca. I am so grateful for the opportunity to expand my skills in industry-related aspects of science and to gain a deeper understanding of how everything operates in the biopharmaceuticals. I would also like to thank the entire OMICS LC/MS group for their kind support and valuable suggestions. This experience has further solidified my determination to pursue a career in the biopharmaceutical industry.

I am thankful to my collaborators, Zihao, William and Dr. Vicki Wysocki for their commitment to the native CZE-SID project. I was greatly impressed by their insightful thinking, efficient work, and comprehensive understanding of mass spectrometers. It was a truly remarkable experience working with them to overcome all the challenges of this project.

I also want to give credit to our research group members. Xiaojing, Daoyang, and Zhichang taught me extensive training in fundamental experimental skills when I first joined the group. Xiaojing provided hands-on instruction in using Q-TOF for native proteomics, while

Daoyang guided me through using CE and the data analysis of histone proteins. They were strict with me regarding all the experimental details and independent thinking. Thank Zhichang for instructing me in mass spectrometer cleaning and providing valuable career advice. Thank Eli and Rachele for their support and encouragement. I appreciate that Tian instructed me in doing cIEF and shared her experiences on finding internships. I want to give special thanks to Qianjie for being not only an excellent teammate but also a true friend in daily life. I also really thank Fei for her patience in teaching me massive experimental skills and data analyses and for her advice in many aspects. I'm also grateful to have Guangyao as my friend, although he just joined the group for one year. We have known each other since day one at MSU, and he has always been a trustworthy friend with much support. Thank Amir for many collaborations in the research projects and for helping me with a lot of experiments. I also would like to extend thanks to other current members, Dr. Zhu, Jorge, Olivia, Mehrdad, Maryam, Bahar, and Lance. Working with everyone in the lab has always been inspiring, thanks to their kindness. Beyond our time in the lab, I have many fond memories of spring trips to Mackinac Island, weekends at state parks, and many other group activities, all of which I will cherish forever.

A heartfelt thanks will be given to my advisor Dr. Jimin Zheng when I was a graduate student at Beijing Normal University. I still remember him emphasizing the challenges of pursuing a Ph.D. in another country, while also wishing me the best of luck and encouraging me to do everything with dedication and hard work before my last day at BNU.

I'm extremely grateful to own so many sincere friends who offered me joy and support not just during the happy moments but also through my lowest points over the past five years, including but not limited to Zhili Guo, Yi Huang, Ziqi Lyu, Qishuo Tan, Ziting Gao, Ruiyue Tan, Grace Ren, Andrew Yang, Zeng Jin, Junyan Yang, Yuhan Jiang, Kunli Liu, Bowen Shen, Zhitao Zhao, Keyi Zhu, Haowei An, Hanqing Guo, Yu Mei, Dong Hae, Chenning Li, Peikai Qi, Shitan Xu.

Lastly, I would like to express my deepest gratitude to my parents. I always feel I'm the happiest person in the world to have them, surrounded by their unconditional love and support. They have always been my greatest motivation to strive to become a better person. It's a big regret that I haven't had the opportunity to visit them in the past five years. I wish to have more time to accompany them in the future.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

2D	Two-dimensional
ABC	Ammonium bicarbonate
ACN	Acetonitrile
AD	Alzheimer’s disease
AI-ETD	Activated ion electron transfer dissociation
APS	Ammonium persulfate
ATNP	Amine-terminated nanoparticles
BCA	Bicinchoninic acid
BGE	Background electrolyte
BSA	Bovine serum albumin
BUP	Bottom-up proteomics
CA	Carbonic anhydrase
CE	Capillary electrophoresis
CID	Collision-induced dissociation
CRP	C-reactive protein
CV	Compensation voltage
Cyt C	Cytochrome C
CZE	Capillary zone electrophoresis
D	Diffusion coefficient
DC	Direct current
DDA	Data-dependent acquisition
DPBS	Dulbecco's Phosphate-Buffered Saline
DSB	Double-strand break
DTIMS	Drift tube ion mobility spectrometry
DTT	Dithiothreitol
ECD	Electron capture dissociation
EDS	Energy Dispersive X-ray Microanalysis
EOF	Electroosmotic flow
EPF	Electrophoretic flow
ESI	Electrospray ionization

ETD	Electron transfer dissociation
ETnoD	Nondissociative electron transfer dissociation
FA	Formic acid
FACS	Fluorescence-activated cell sorting
FAIMS	Field asymmetric waveform ion mobility spectrometry
FDR	False discovery rate
FT	Fourier transform
FTICR	Fourier-transform ion cyclotron resonance
GDH	Glutamate dehydrogenase
HCD	Higher-energy collisional dissociation
HCl	Hydrochloric acid
HETP	Height equivalent to a theoretical plate
HF	Hydrofluoric acid
HILIC	Hydrophilic interaction liquid chromatography
iBAQ	Intensity-based absolute quantification
ID	Identification
IEX	Ion exchange chromatography
IMAC	Immobilized metal affinity chromatography
IMS	Ion mobility spectrometry
IPA	Ingenuity pathway analysis
IST	In-source trapping
iTRAQ	Isobaric tags for relative and absolute quantification
L	Column length
LC	Liquid chromatography
LDV	Laser Doppler Velocimetry
LFQ	Label-free quantification
LPA	Linear polyacrylamide
m/z	Mass-to-charge ratio
MALDI	Matrix-assisted laser desorption ionization
MD	Multi-dimensional
MDP	Middle-down proteomics

MP	Mass photometry
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
Myo	Myoglobin
N	Number of theoretical plates
Na_2HPO_4	Sodium phosphate dibasic
$NaBH_3CN$	Sodium cyanoborohydride
NaCl	Sodium chloride
NaH_2PO_4	Sodium phosphate monobasic
nanoESI	Nano-electrospray ionization
NaOH	Sodium hydroxide
NCE	Normalized collision energy
nCZE	Native capillary zone electrophoresis
NH_4OAc	Ammonium acetate
nMS	Native mass spectrometry
NP	Nanoparticles
nTDP	Native top-down proteomics
PrSM	Proteoform spectrum match
PTCR	Proton transfer charge reduction
PTM	Post-translational modification
Q	Charge
QqQ	Triple quadrupole
RF	Radiofrequency
RPLC	Reverse-phase liquid chromatography
RT	Room temperature
SA	Streptavidin
SDS	Sodium dodecyl sulfate
SEC	Size exclusion chromatography
SID	Surface-induced dissociation
SP3	Single-pot solid-phase-enhanced sample preparations
TB	Terrific Broth

TDP	Top-down proteomics
TEM	Transmission electron microscopy
TGA	Thermogravimetric analysis
TIMS	Trapped ion mobility spectrometry
TMT	Tandem mass tags
TOF	Time-of-flight
TWIMS	Travelling wave ion mobility spectrometry
u	Flow rate
UHMR	Ultra-high mass range
UVPD	Ultraviolet photodissociation
V	Voltage
WCX	Weak cation exchange
μ_{all}	Overall electrophoretic mobility
μ_{eo}	Electroosmotic mobility
μ_{ep}	Electrophoretic mobility

CHAPTER 1. Introduction

1.1 Multi-level mass spectrometry-based proteomics

1.1.1 Overview of proteomics

Proteins are one of the most critical components at the molecular level in regulating biological functions within cells, including signal transduction, cell proliferation, cell death (apoptosis, autophagy, necrosis), cell division, and many others [1-5]. The central dogma provides an outline where the genetic information in DNA is transferred to RNA by transcription, followed by RNA translation into proteins as the final target. The proteome refers to the entire set of protein products from an organism's genome. The proteome has a significantly higher heterogeneity than the genome due to multiple biological processes like RNA alternative splicing, genetic variants, and protein post-translational modifications (PTMs) (**Figure 1.1**) [6]. Proteoforms, as the basic units in the proteome, are used to represent all the protein forms arising from the same gene due to, e.g., amino acid sequence variations and PTMs [7]. For example, it is estimated that 20,000 encoding genes in the human genome can produce approximately over 1 million proteoforms [7].

Proteomics is defined as the study of the proteome, aiming to identify, quantify, and analyze the structures, functions, and interactions of all the proteoforms within a biological system [8, 9]. Mass spectrometry (MS)-based proteomics has become critical for measuring proteins to better understand the molecular mechanisms within diverse cellular processes and disease developments [10]. This was achieved by the development of two advanced ionization techniques of biological macromolecules for mass spectrometric analysis: electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI), which earned the Chemistry Nobel Prize in 2002. A general workflow for MS-based proteomics includes protein extraction from cells or tissues, followed by either enzymatic digestion into peptides or remaining intact proteoforms. Then, the peptides or proteoforms will experience online or offline liquid phase separation coupled with ionization steps and the downstream MS analysis. Next, the mass spectra of precursor ions and fragment ions are both acquired to collect enough information for protein identifications from database searching. There are two main strategies, bottom-up proteomics (BUP) and top-down proteomics (TDP), decided by the inclusion of enzymatic digestion or not during the protein preparation (**Figure 1.2**) [11].

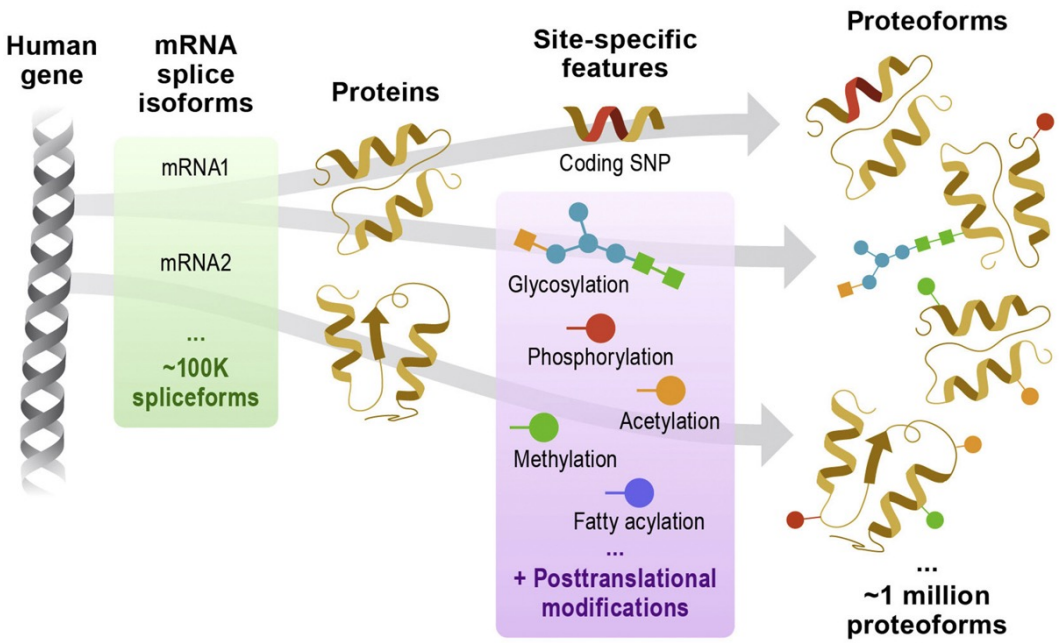


Figure 1.1. From human genes (far left), diverse forms of mature endogenous protein molecules are expressed (far right). The figure is reprinted with permission from reference [6].

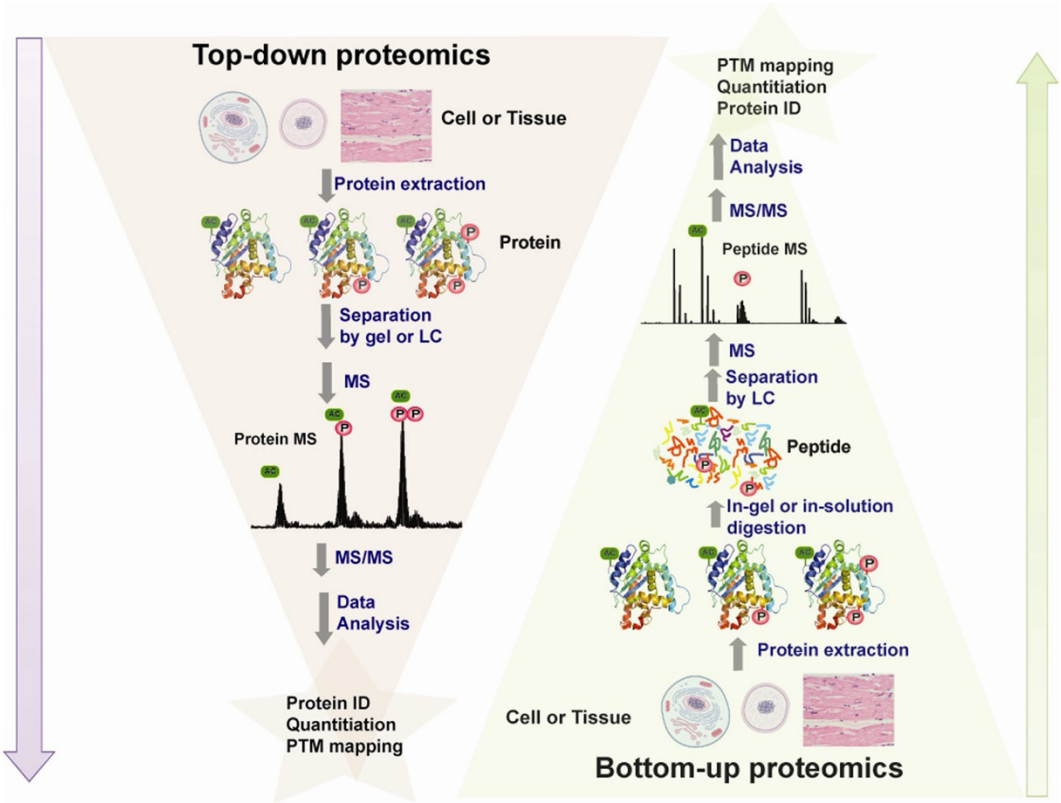


Figure 1.2. Schematic illustration of the general workflows of top-down and bottom-up proteomics. The figure is reprinted with permission from reference [11].

1.1.2 Bottom-up proteomics (BUP)

BUP achieves the protein information by analyzing the proteolytic digested peptides from the corresponding proteins. **Figure 1.2** shows a typical BUP workflow where the proteins are first extracted from the cells or tissues. Then, the proteins are either in-gel digested or in-solution digested by specific enzymes based on the needs. This enzyme digestion usually requires several steps to assure the digestion efficiency including the denaturation of proteins, reduction of disulfide bonds, and protective alkylation of proteins. The most commonly used enzyme is serine protease trypsin with high specificity and high proteolytic activity. Trypsin could specifically cleave the proteins at the carboxyl side of lysine and arginine to generate an average range of peptides at 10-20 amino acids. The product peptides will carry at least two protonation sites: the amine group at the N-terminus and the C-terminus lysine or arginine. With such an averaged size of peptides and the charges, the tryptic-digested peptides are ideal for ionization and fragmentation during the MS analysis. Next, the peptides are typically under liquid chromatography (LC) separation and followed by downstream ESI-tandem MS (ESI-MS/MS) analysis. At last, the protein identifications are attained by matching the identified peptide sequence to the protein sequence [12]. The peptide identifications apply the precursor ion and the corresponding fragmentation ion information against the database from the in-silico digestion of the protein sequences derived from the known genome. During this process, one peptide could belong to not only one protein but multiple proteins, designated as one protein group.

Quantitative proteomics measures the abundance differences of proteins across samples to provide crucial insights into the biological states of cells and tissues, significantly advancing our understanding of key biological processes [13]. Two techniques are commonly used in quantitative BUP: isotopic labeling quantification strategy and label-free quantification (LFQ) strategy [14-19].

Tandem mass tags (TMT) [16] and isobaric tags for relative and absolute quantification (iTRAQ) [17] are the two most used multi-channel labeling methods for isotopic labeling quantification. For example, TMT (**Figure 1.3**) contains one mass reporter to show the relative abundance of labeled peptides across different samples based on the relative intensity of the reporter ions, one mass normalizer to balance the overall mass of each tag equally and one amine-reactive group to attach to the N-terminus or the side chain of lysine residue of peptides. The general TMT workflow first labels the peptides across different samples with different channels of

isobaric mass tags by the amine-reactive group. Then, the peptide samples are mixed as one sample for LC-MS analysis. The same peptides from different samples having isobaric mass tags will be co-eluted during LC separation because of the same hydrophobicity and mass. Furthermore, they will be co-isolated for fragmentation to get the relative abundance information from the cleavable report ions. This analysis happens within the same LC-MS run to measure multi-channeled samples, eliminating the random quantification errors from the measurement of different samples by different LC-MS runs. The multiplexity of TMT labeling could measure up to 18 channels [20] of samples, enabling the high throughput of BUP analysis.

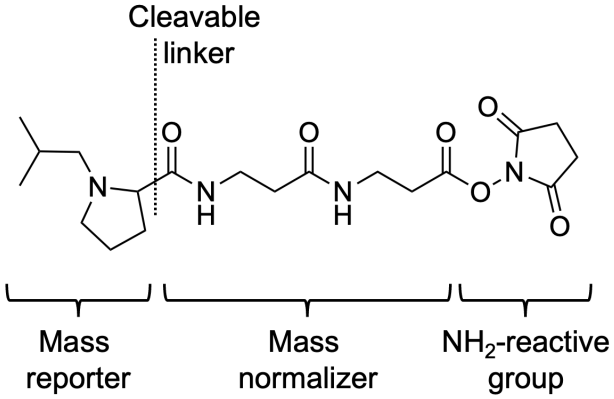


Figure 1.3. TMT reagent structure including functional regions and higher-energy collisional dissociation (HCD) fragmentation site.

In the LFQ approach, isotopic labeling is no longer required, simplifying sample preparation and eliminating the limitation of sample multiplicity for quantification. LFQ methods rely on comparing precursor ions' intensities from the extracted ion chromatograms between independent LC-MS measurements. Signal intensities of the peptide ions from ESI are directly correlated with the peptide concentrations [21, 22]. Therefore, the extracted peak areas corresponding to specific peptide ions from chromatograms in LC-MS measurements can be utilized for the relative quantification of particular peptides and proteins across different samples.

LFQ employs the integration of all the peak areas under the curve of peptides from the separation profile to estimate the corresponding protein abundance [23]. Several factors are crucial to affect the accuracy of LFQ analysis, including the reproducibility across different LC separation runs, the stability of the ESI source, the retention time alignment, the computational algorithms for abundance comparison, and the statistical evaluation of multiple LC-MS datasets [24].

Although BUP has become the gold standard approach in MS-based proteomics studies, several intrinsic drawbacks are still impeding the development of this technique. The bottom-up approach often struggles to differentiate proteoforms from the same gene because the intact proteoform information is lost during the proteolytic process. Furthermore, the sequence coverage of proteins from the bottom-up method is limited because only a subset of peptides from a protein is typically detected. Moreover, peptides may be shared between different proteins, leading to ambiguity in protein identification in complex biological samples.

1.1.3 Top-down proteomics (TDP)

Unlike BUP, TDP is a proteoform-centric approach, employing MS and MS/MS to characterize intact proteoforms without proteolytic cleavage. Most steps in the typical TDP workflow (**Figure 1.2**) are similar to BUP except enzymatic digestion, including protein extraction from cells or tissues, one or multiple-dimensional liquid-phase separations, soft ionization of intact proteins, MS and MS/MS acquisition, and data analysis. One significant advantage of TDP is that it could provide a bird's eye view of all proteoforms in a sample while fully preserving the PTM information in the proteoform sequence (**Figure 1.4A**). It could prevent false identifications from the proteoforms with the high similarities of the sequence compared to BUP. However, two main challenges in TDP are the identification of the proteoforms in low abundance and the precise localization of PTMs on each proteoform. The larger intact proteoforms carrying more charges than peptides cause substantial MS signal dilution, inducing the overall lower sensitivity for MS detection (**Figure 1.4B**). Because intact proteoforms are considerably larger than peptides, the backbone cleavage is more strenuous for proteoform fragmentation than BUP. Furthermore, the denaturation step in traditional TDP impedes the detection of protein complexes due to the disassembly of the protein complex. Native TDP (native proteomics) offers the most precise profiling of proteome samples by analyzing them under near-physiological conditions and directly assessing protein complexes. Nevertheless, this approach faces several technical hurdles. Firstly, native proteomics often struggles with detecting low-abundance protein complexes within complex biological samples due to its inherent low sensitivity. Consequently, this technique is most effective for analyzing highly abundant protein complexes or purified samples [25-27], which limits its broader applicability in proteomic studies. Secondly, analyzing large protein complexes and identifying individual proteoforms within them necessitates the use of advanced and specialized mass spectrometers. Techniques such as MS/MS

(MS^2) or MS^3 analysis are essential, adding complexity to the experimental setup and requiring substantial expertise and resources [28].

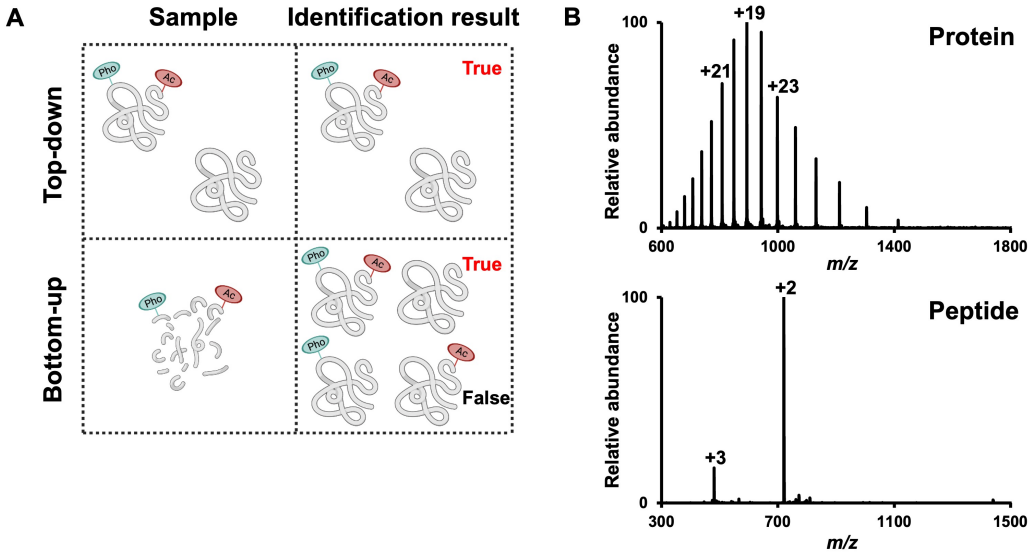


Figure 1.4. TDP and BUP. (A). Determination of proteoforms and PTMs from TDP and BUP. Pho: phosphorylation. Ac: acetylation. (B). Charge state distribution of proteins and peptides.

Despite the challenges and limitations, TDP has made substantial technological advancements in the past decades involving the characterization of large proteoforms, heavily modified proteoforms, global proteoform profiling, and bioinformatics tools [29-39].

Currently, most large-scale TDP studies have concentrated on proteoforms smaller than 30 kDa, primarily due to the ion suppression of coelution of high and low abundant proteoforms, low sensitivity from broad charge state distribution, and low backbone cleavage coverage by conventional collision-based fragmentation [40, 41]. Size exclusion chromatography (SEC) has been recognized as an effective method to fractionate the proteoforms by size to lower the complexity of the sample, improving the detection of larger proteoforms. The Smith research group coupled SEC with reverse-phase LC (RPLC)-MS/MS to achieve the identification of a 140-kDa protein (endogenous human cardiac myosin binding protein C) from human heart samples [29]. The Zhang group applied a novel monolithic reverse-phase capillary column for an SEC-RPLC-MS platform to identify 347 proteoforms over 30 kDa from *E. coli* lysate in a single RPLC run [30]. To improve the sensitivity of detecting large proteoforms, the Sun group coupled high field asymmetric waveform ion mobility spectrometry (FAIMS) to capillary zone electrophoresis-MS/MS (CZE-MS/MS) as an online two-dimensional separation to boost the number of proteoform identifications by 6-fold in the mass range of 20-45 kDa [31]. For better backbone

cleavage coverage, electron or photon-based fragmentation techniques have shown significantly superior performance compared to collision-based techniques in fragmenting large proteoforms [42-44]. The internal fragment ions are another huge part typically neglected by the bioinformatics tools because they don't contain any N-terminal or C-terminal fragments of the proteoforms [45, 46]. The Loo group achieved the highest sequence coverage (75%) of an intact mAb by combining electron capture dissociation (ECD), higher-energy collisional dissociation (HCD), and internal fragment ion assignments for top-down analysis of the intact NIST mAb [32].

The delineation of proteoforms carrying many PTMs is challenging due to the high heterogeneity of proteoforms and the potential loss of labile PTMs during collision-based fragmentation. Take histone, a heavily modified protein containing methylation, acetylation, phosphorylation, and citrullination, and many others, as an example [47, 48]. High-resolution separation for the isobaric and isomeric histone proteoforms and the effective gas-phase fragmentation for the labile PTMs are crucial. The Sun group developed SEC-CZE-MS/MS and CZE-FAIMS-MS/MS platforms for high-capacity separation and highly sensitive TDP analysis of histone proteoforms, resulting in the identification of nearly 400 and 600 histone proteoforms, respectively, from a commercial calf thymus sample [33, 34]. To precisely localize PTMs on histone proteoforms, the Brodbelt group combined ultraviolet photodissociation (UVPD) with gas-phase proton transfer charge reduction (PTCR) to achieve a drastic improvement of sequence coverage of histone proteoforms (e.g., 73% for H2A and 91% for H4) with a substantially accurate characterization of PTMs [35].

Global TDP profiling of complex samples is challenged by their immense complexity, yet multi-dimensional separations prior to MS and MS/MS have been employed to improve proteome coverage in MS-based TDP. The Kelleher group applied immunomagnetic enrichment and fluorescence-activated cell sorting (FACS) to selectively enrich specific cell types from human blood and bone marrow, followed by RPLC-MS/MS-based proteoform profiling [36]. This approach led to the identification of nearly 30,000 unique proteoforms originating from 1,690 human genes across 21 different human hematopoietic cell types and plasma. The Sun group combined an offline LC fractionation with the downstream CZE-MS/MS for TDP of two colorectal cancer cell lines (SW480 and SW620), resulting in the identification of 23,622 proteoforms from 2,332 proteins [37]. This improvement enhanced the understanding of

colorectal cancer metastasis at the proteoform level through large-scale proteomics, and previously unknown protein biomarkers for cancer diagnosis and drug development were discovered.

Sophisticated algorithms for bioinformatics tools are also critical to boost the proteoform characterization and quantification in MS-based TDP. In the last decades, many bioinformatics tools emerged and are well established, such as ProSight [49], TopPIC [50], Metamorpheus [51], MASH [52], ClipMS [53], Informed-Proteomics [54] and many others. Current bioinformatics tools for MS-based TDP struggle with the identification of large proteoforms (over 30 kDa) because they generally depend on resolved isotopic peaks in each charge state to accurately determine the monoisotopic mass of proteoforms or their fragment ions for subsequent database searching, but the resolution for typical mass spectrometers is not high enough for large proteoforms. To improve this situation, the Kohlbacher group introduced FLASHDeconv, an ultrafast deconvolution tool for TDP that can analyze both isotopically resolved and unresolved peaks across a wide charge and mass range in MS spectra, making it particularly effective for identifying large proteoforms [38]. For the intact proteoform quantification, the Petyuk group designed a companion R package (TopPICR) for TopPIC to enhance cross-data set quantification based on LFQ [39]. The TopPICR tool added a critical step of clustering features across datasets in LC-MS space to make the transformation of LC and MS dimensions into Z-scores to be statistically interpretable.

1.1.4 Mass spectrometry (MS)

MS has progressed from merely cataloging proteins in biological systems to evaluating protein properties and their functional modulation at multiple levels, such as protein identification, quantification, PTM, protein dynamics, and many others [55]. The basic principle of MS is to measure the charged molecules by their mass-to-charge ratios (*m/z*) in the gas phase. Proteins or peptides are typically separated by liquid-phase separation before MS analysis. Owing to the invention and development of ESI, protein and peptide ions can gently transform from liquid phase into gas phase while keeping their integrity. Unlike MALDI commonly forms singly charged analytes, ESI can generate multiple charging to benefit not only the detection of the large biomolecules on mass spectrometers with limited *m/z* range but also the fragmentation in tandem MS (MS/MS). In the positive ion mode of the ESI process (**Figure 1.5**), a high voltage is applied at the capillary end at an electric potential of several kV [56, 57]. The capillary is filled with an

analyte solution carrying typically an acidic buffer to provide protonation for the analytes. The Taylor cone is first formed due to the electric field and the surface tension, followed by the initial micron-sized droplet generation into the gas phase. The initial droplets shrink into the final nanometer-sized droplets and form naked charged analytes caused by the evaporation of the solvent. Ultimately, the charged analytes move into the mass spectrometer by the electric field for MS analysis.

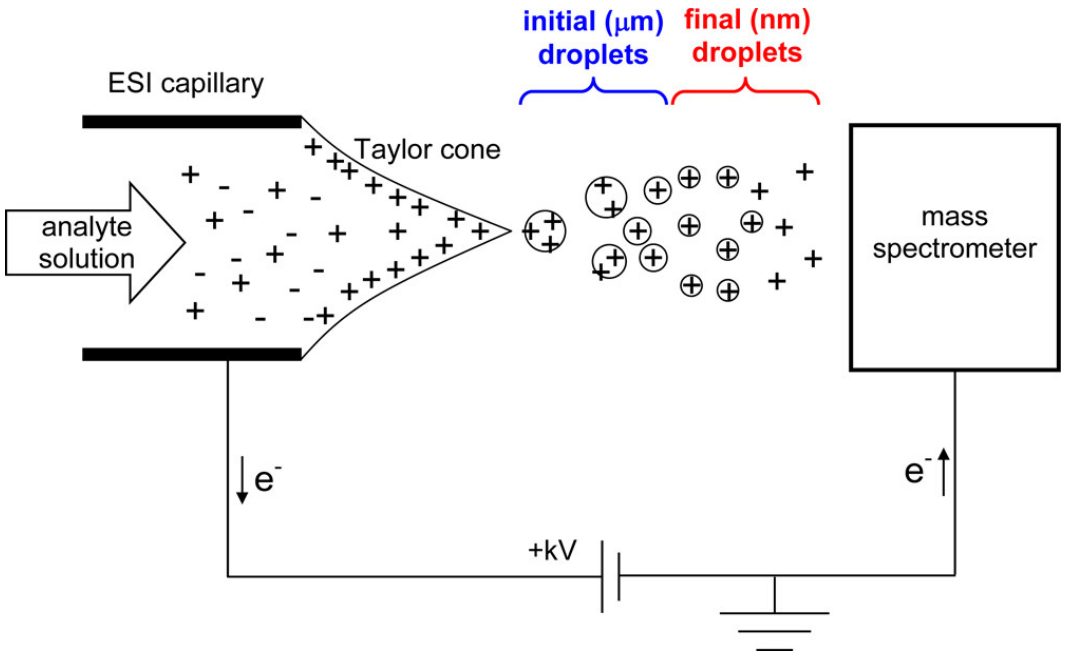


Figure 1.5. Schematic depiction of the ESI process operated in positive ion mode. The figure is reprinted with permission from reference [56].

To measure the protein or peptide ions, the mass analyzer inside the mass spectrometer is the key unit to separate ions of different *m/z*. Filtering mass analyzer, ion trapping mass analyzer, and time-of-flight (TOF) are the three most common types of mass analyzers. Quadrupole is the most representative and frequently used filtering-type mass analyzer, invented by Wolfgang Paul Helmut Steinwedel in the 1950s. A quadrupole consists of four parallel metal rods arranged in a square configuration, which create an oscillating electric field that acts as a mass filter (**Figure 1.6**) [58]. This field allows only ions with a specific *m/z* ratio to pass through while deflecting others. The operation of a quadrupole mass analyzer involves applying both direct current (DC) and radiofrequency (RF) voltages to the rods. By adjusting these voltages, the analyzer can selectively stabilize the trajectory of ions with a particular *m/z* ratio, enabling them to reach the detector, while ions with different *m/z* ratios are deflected or filtered out. Quadrupoles offer

several critical advantages that make them essential tools in the MS field, including but not limited to high selectivity and sensitivity, rapid mass scanning speed, and robustness and durability [59-63]. Triple quadrupole (QqQ) is designed by aligning three quadrupoles in series to filter not only the precursor ions but also the fragment ions to selectively monitor and quantify analyte ions with high specificity and high sensitivity [64]. However, the limitations of quadrupoles are the low mass resolution and mass accuracy and the limited scanning mass range (typically <3000 *m/z*) [62, 65].

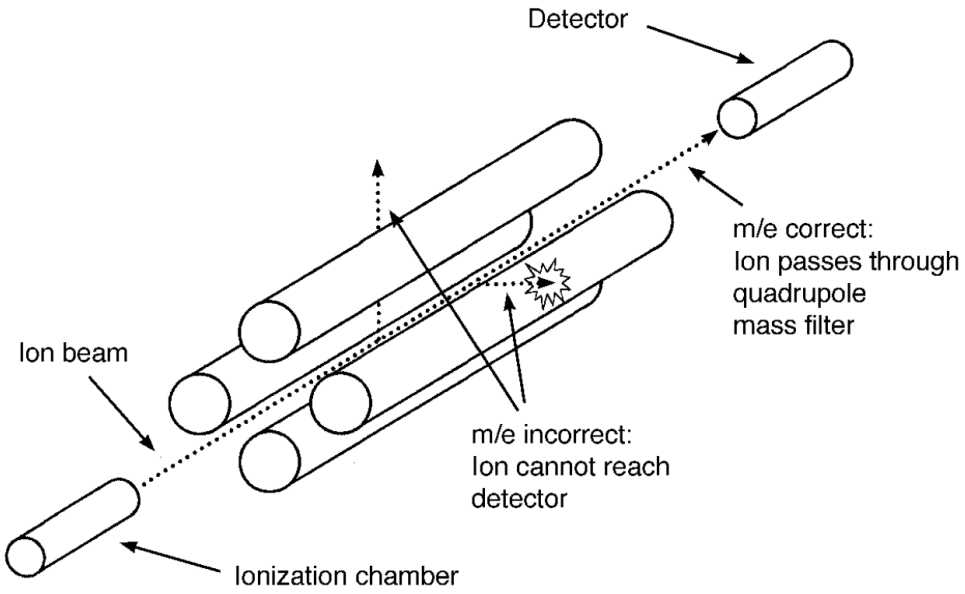


Figure 1.6. Schematic of quadrupole. m: mass of the ion; e: charge of the ion. The figure is reprinted with permission from reference [58].

The Orbitrap is an ion-trapping type mass analyzer based on Fourier transform (FT) and was first invented by Dr. Makarov in 1999 [66]. The orbitrap mass analyzer contains a spindle-shaped rod as the central electrode and a barrel-like outer electrode (**Figure 1.7**) [67]. The analyte ions have the trajectories of integrating rotation around the central electrode and oscillations along the *z*-axis to form a three-dimensional spiral motion. This intricate motion enables the ions to stay trapped within the orbitrap mass analyzer long enough for accurate measurement. The image current is generated from the analyte harmonic oscillations and detected on the split outer electrodes to provide the *m/z* information of the analytes from FT. The Orbitrap mass analyzer has been extensively utilized in MS-based proteomics for its high resolution (up to 1 million FWHM at *m/z* 200) and high mass accuracy (up to sub-1 ppm). Conventional Orbitrap mass analyzer

could only process mass range lower than 8,000 *m/z*, but the release of ultra-high mass range (UHMR) Orbitrap boosts the mass range up to 80,000 *m/z* to benefit the detection and the characterization of large biomolecules like native protein complexes [68]. The Ivanov group successfully coupled the native CZE to UHMR orbitrap to characterize a near-1 MDa GroEL protein complex with the binding molecules to show its conformational changes [69]. However, the Orbitrap mass analyzer suffers from a low scanning speed (typically lower than 40 Hz), limiting its applications for high-throughput screening and real-time analysis.

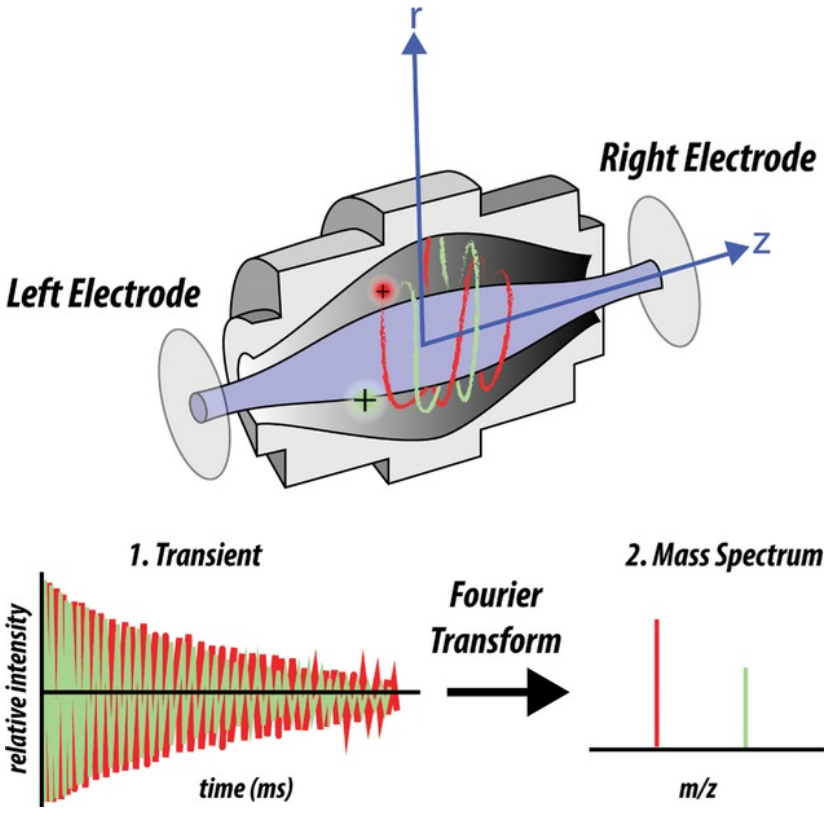


Figure 1.7. The Orbitrap mass analyzer. The figure is reprinted with permission from reference [67].

Alternative to filtering-type and ion-trapping mass analyzers, the TOF mass analyzer is a foundational technique in MS to determine the *m/z* ratio of ions based on their flight time through a field-free drift region. Originating in the 1940s, TOF operates under the principle that ions, once accelerated to uniform kinetic energy, will traverse the drift region at velocities inversely proportional to their *m/z* ratio (**Figure 1.8**) [61, 62, 70]. Consequently, ions with lower *m/z* ratios arrive at the detector more rapidly than those with higher *m/z* values. The precise measurement of

these time differences facilitates the accurate determination of the ions' *m/z*. TOF mass analyzers are particularly noted for their unlimited mass range and rapid data acquisition capabilities, which render them exceptionally suited for the analysis of large biomolecules and complex matrices [71]. TOF mass analyzers often incorporate reflectron technology to enhance the mass resolution by correcting kinetic energy differences among ions with the same *m/z*, causing them to converge in time at the detector and resulting in high-resolution peaks in the mass spectrum [72]. Although TOF mass analyzers generally exhibit lower resolving power compared to high-resolution counterparts such as Orbitrap mass analyzers, their capacity for high-speed scanning is unparalleled, making them indispensable for high-throughput screening applications [73].

Recent technological advancements have further expanded TOF's utility, notably through its integration with ion mobility spectrometry (IMS), which significantly improves its ability to resolve complex mixtures by adding an additional dimension of separation such as trapped IMS-TOF (TIMS-TOF) [74]. However, the performance of TOF mass analyzers can be subject to matrix effects, particularly in MALDI-TOF applications, where the choice of matrix and the co-crystallization process can critically influence spectral quality [75].

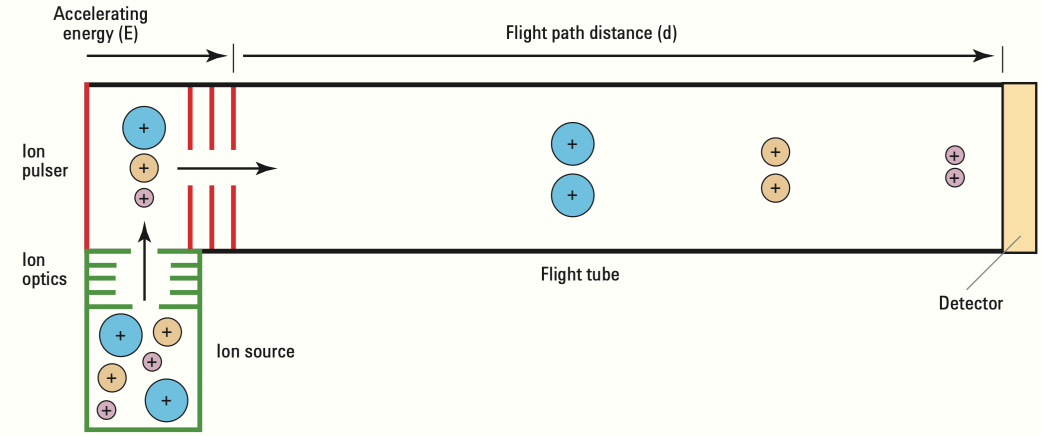


Figure 1.8. Schematic of a linear TOF analysis of singly charged ions. The figure is reprinted from reference [70].

1.1.5 Tandem mass spectrometry (MS/MS)

In MS-based proteomics, tandem MS (MS/MS) is a powerful tool for dissecting the intricate details of proteins beyond their intact masses, such as their sequences, and PTMs. The process begins with the mass spectrometer scanning the full mass spectrum (MS1) and isolating ions based on their *m/z* at a narrow isolation window. These ions, known as precursor ions, are then fragmented into smaller pieces called fragment ions using gas-phase techniques, which are

subsequently analyzed by a second mass analyzer to obtain the tandem mass spectrum scan (MS2). The mass analyzers precisely measure the molecular mass of gas-phase ions by analyzing their motions in magnetic or electric fields [62]. Both *m/z* and charge states, which are determined from isotopic patterns, play a crucial role in calculating the ions' mass with high accuracy. Identifying proteins or peptides in complex biological samples based solely on intact molecular mass from MS1 can be a daunting task because of the presence of isomeric and isobaric species. Therefore, MS2 is necessary to provide the fragment ions as a fingerprint of the specific protein or peptide isolated after MS1 for identification. Protein or peptide identification becomes more accurate as these empirical spectra are matched against theoretical ones from vast databases. Depending on the location of cleavage at the protein sequence, different types of fragment ions are generated (**Figure 1.9**). Numerous gas-phase fragmentation techniques with distinct mechanisms have been developed to enhance fragmentation efficiency and offer complementary insights into proteins and peptides.

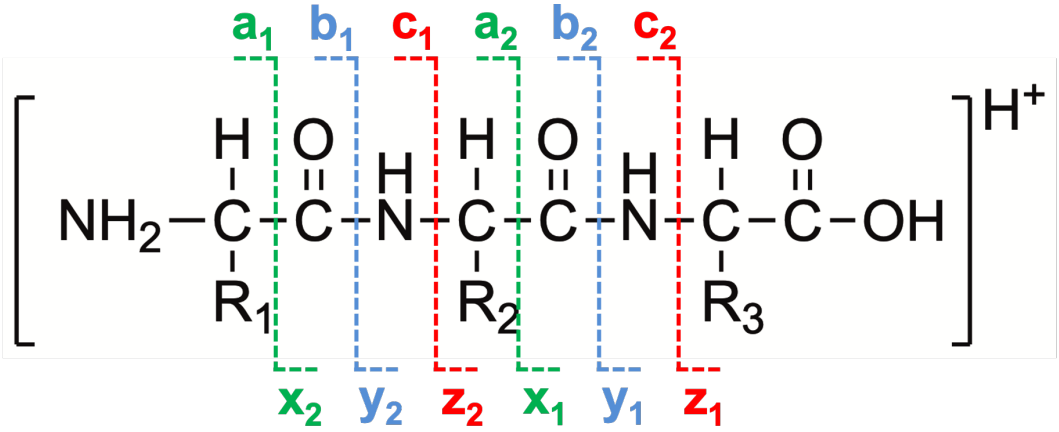


Figure 1.9. An example of peptide fragmentation nomenclature.

Collision-induced dissociation (CID) stands as the most widely employed method for generating fragment ions from peptides and proteins in MS-based proteomics. In CID, protein or peptide ions are accelerated and collide with neutral gas molecules, such as nitrogen, helium, or argon, within an ion trap (**Figure 1.10**) [76]. The kinetic energy from these collisions is deposited into the internal energy of ions. When this energy exceeds the threshold required to break chemical bonds, b-type and y-type ions are predominantly produced from the fragmentation at the peptide bond [77]. One deficiency of conventional CID is that it happens in an ion trap where the low *m/z* ions cannot be retained effectively. To advance CID, HCD follows a similar collisional

mechanism but takes place in a multipole collision cell (**Figure 1.10**) [76, 78]. This improvement allows for the inclusion of low-mass ion detection and enables applications like isobaric tag-based labeling quantification [79]. Despite these advancements, both CID and HCD tend to preferentially cleave the most labile bonds of proteins or peptides, bringing the risk of breaking labile PTMs [80]. Thus, applying CID or HCD limits sequence coverage and hinders the accurate localization of labile PTMs in the proteoforms, making PTM mapping particularly challenging [81, 82].

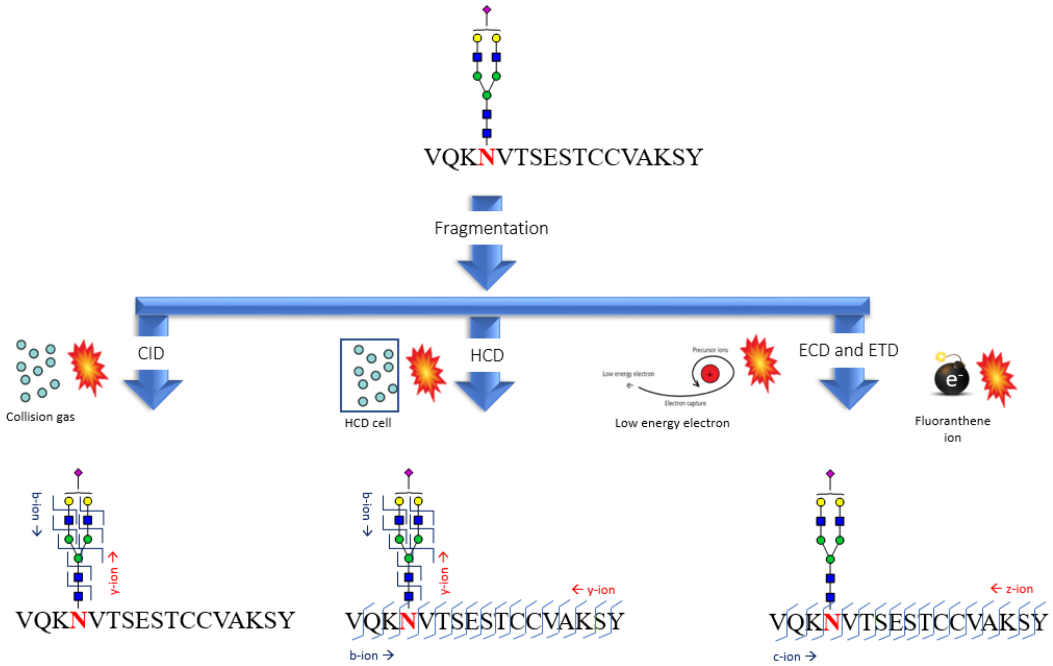


Figure 1.10. Mechanism of CID, high-energy collision dissociation (HCD), ECD, and electron transfer dissociation (ETD) fragmentation. The figure is reprinted with permission from reference [76].

ECD and electron transfer dissociation (ETD) are widely employed electron-based activation methods as the alternative gas-phase fragmentation to collision-based dissociation for proteins and peptides. The mechanisms for ECD and ETD are similar except for the source of electrons. In ECD, the free electrons are generated typically from a heated filament and captured by the multiply charged analyte cations to form highly reactive cation radicals, followed by the bond cleavage at N-C α bonds along the protein or peptide backbone, resulting in the formation of c-type and z-type fragment ions (**Figure 1.10**) [76, 83]. ETD also produces the same types of fragment ions, but the reactive cation radicals originate from the transfer of the anion radicals (formed from the reagents like azulene and fluoranthene) to the multiply charged analyte cations

[84, 85]. ETD/ECD acts as a robust approach for characterizing PTMs for large peptides or entire proteins while preserving labile modifications at modified residues, making it central to PTM analysis [86-88]. This is because bond dissociation happens immediately following electron transfer or capture, before any energy redistribution, which protects labile modifications [89]. However, ETD/ECD fragmentation efficiency is highly dependent on precursor charge density, with higher charge densities leading to more extensive precursors, while lower charge densities make the precursors more compact [86, 90, 91]. During the ETD/ECD process, the non-covalent interactions may still hold the fragment ions together within the low charge density precursors, known as nondissociative ETD (ETnoD) [92]. Supplemental activation for ETD/ECD has been recognized as an efficient approach to minimize ETnoD such as activated ion ETD (AI-ETD) and electron-transfer/higher-energy collision dissociation (EThcD) [93-96]. AI-ETD uses infrared photoactivation during ETD to disrupt noncovalent binding, while EThcD activates all ETD product ions with HCD energy, both generating additional b-type and y-type ions that enhance complementary fragment series and improve protein backbone coverage [97].

UVPD is another gas-phase fragmentation method that couples well with MS, and it utilizes UV lasers (commonly at 157, 193, 213, or 266 nm wavelength) to activate the precursor ion by the adsorption of one or more high-energy UV photons [98]. This process deposits internal energy into the analyte ions, exciting them to electronic states where dissociation occurs as soon as they gain enough energy to surpass the dissociation barrier [99, 100]. Unlike collisional-based and electron-based dissociation, UVPD can yield all types of fragment ions (a-, x-, b-, y-, c-, z-type ions), generating a series of complementary fragments for more comprehensive sequence coverage. Also, UVPD achieves higher sequence coverage compared to collisional-based dissociation and to preserve labile PTMs for protein and peptide characterization, while performing little dependence on the charge states of the analyte ions [101-105].

1.2 Separation techniques in MS-based proteomics

1.2.1 Capillary zone electrophoresis (CZE)

CZE is a highly efficient open tubular liquid-phase separation technique based on the electrophoretic mobility (μ_{ep}) of analytes, and it has been applied broadly in proteomics [106-110]. CZE separation requires a fused silica capillary (typically 10-75 μm inner diameter and 20 to 100 cm length), background electrolyte (BGE) buffer filled in the capillary for conductivity, and a high voltage source to provide a strong electric field. In separation science, the number of

theoretical plates (N) and the height equivalent to a theoretical plate (HETP) are commonly used to evaluate the column separation efficiency. A theoretical plate is a hypothetical concept and represents a single equilibrium step of the solute reaching between the stational phase and the mobile phase. More theoretical plate numbers of a column reflect the better separation efficiency. HETP is defined as the length of one theoretical plate within the separation column [111]. The relationship of the column length (L), N, and HETP are expressed in **Equation 1.1**:

$$N = \frac{L}{HETP} \quad \textbf{Equation 1.1}$$

HETP depends on various factors during the separation and can be explained by the van Deemter equation (**Equation 1.2**) [112].

$$HETP = A + \frac{B}{u} + Cu \quad \textbf{Equation 1.2}$$

In the van Deemter equation, the A-term is the eddy diffusion parameter that arises from the multiple flow paths that molecules can take through the column due to variations in the packing of particles, B-term is the longitudinal diffusion parameter that occurs as molecules disperse along the column axis and is inversely proportional to the flow rate (u), C-term is the mass transfer parameter that is associated with resistance to mass transfer between the mobile and stationary phases to reflect the delay in equilibrium between the phases [113]. In CZE, there is no stationary phase contributing to the A-term and C-term, leading to a much lower HETP and higher separation efficiency compared to other packed chromatographic separations. CZE also shows great potential in separating such large biomolecules as intact proteins. Another way of describing the CZE separation efficiency is shown in **Equation 1.3**.

$$N = \frac{\mu_{all} V}{2D} \quad \textbf{Equation 1.3}$$

In this equation, N is related to overall electrophoretic mobility (μ_{all}), voltage applied at the capillary (V), and diffusion coefficient of the analytes (D). A higher voltage is commonly applied to achieve a higher N for the separation of intact proteins owing to the low D of these proteins. For example, one work from our group showed the CZE could reach nearly 1 million theoretical plates for the separation of myoglobin proteoforms [114].

The μ_{all} of analytes in the CZE process is the sum of two contribution factors: μ_{ep} and electroosmotic mobility (μ_{eo}), as shown in **Equation 1.4**.

$$\mu_{all} = \mu_{ep} + \mu_{eo} \quad \textbf{Equation 1.4}$$

The μ_{ep} and μ_{eo} can be described in **Equation 1.5** and **Equation 1.6**:

$$\mu_{ep} = \frac{q}{6\pi\eta r}$$ **Equation 1.5**

$$\mu_{eo} = \frac{\varepsilon\zeta}{4\pi\eta}$$ **Equation 1.6**

Where q is the charge of the molecule, η is the viscosity of the BGE, r is the radius of the molecule, ε is the dielectric constant of the BGE, and ζ is the zeta potential. In the typical denatured CZE separation of proteins (**Figure 1.11**), the protein cations (assume the pH of BGE is lower than the pIs of all proteins) move towards the cathode by electrophoretic flow (EPF) and electroosmotic flow (EOF). EOF arises from the zeta potential of the double layer at the capillary inner wall and drives the whole bulk solution along with all the molecules. Only charged molecules have EPF, while the neutral molecules don't and move solely by EOF. EOF is considered an acceleration force for rapid CZE separation in most cases of proteomics studies. However, the presence of EOF can also be a limitation, as it reduces the separation window, potentially providing insufficient time for mass spectrometer acquisition in CZE-MS analysis. The common way of controlling EOF's effect is by either adjusting the pH of BGE or various types of coatings at the capillary inner wall. The advantages of capillary coating are not only helping manage the effects of EOF, but also reducing or eliminating the non-specific binding of proteins to better the separation efficiency for proteomics studies. One example of such a capillary coating is applying neutral and hydrophilic linear polyacrylamide (LPA) to eliminate EOF in the capillary for a wide separation window and minimize the interaction between the proteins or peptides with the capillary inner wall [115-117].

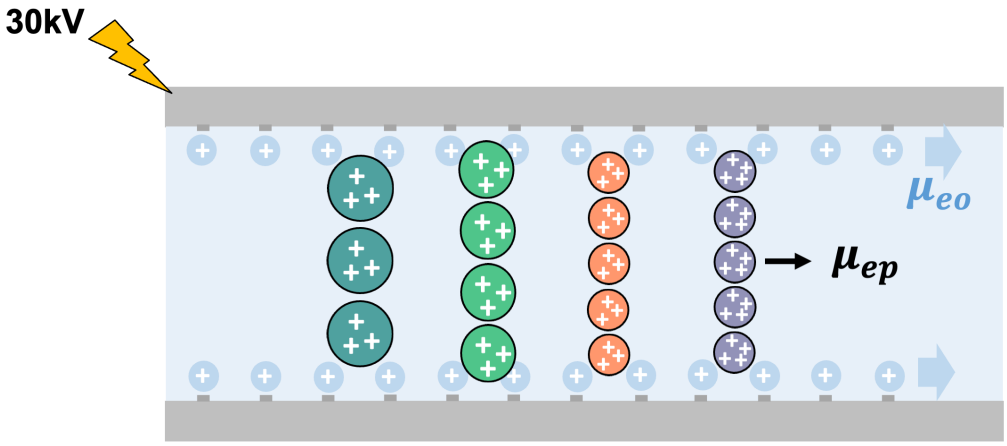


Figure 1.11. The mechanism of CZE separation. Molecules with more charges and smaller radii move faster.

To online couple CZE to ESI-MS for proteomics, the interface used for completing the electrical circuits of CZE separation and providing voltage for ESI is critical. A series of advanced CZE-MS interfaces have been developed with the improvement of sensitivity, stability, and easy operation [118-123]. Our group uses the electrokinetically pumped sheath flow nanoelectrospray interface (**Figure 1.12**), first reported by the Dovichi group in 2010 and subsequently upgraded in 2013 and 2015 [121-123]. A separation capillary is inserted through a junction and enters a glass ESI emitter, where it connects via a side arm to a sheath electrolyte reservoir linked to a power supply. The application of voltage to the sheath electrolyte induces EOF within the emitter, allowing for precise control of sheath fluid pumping at nL/min flow rates of spraying. With the development from the first to the third generation, the larger size orifice of the ESI glass emitter and the closer distance of the capillary tip towards the emitter orifice escalate the robustness without the loss of sensitivity [122].

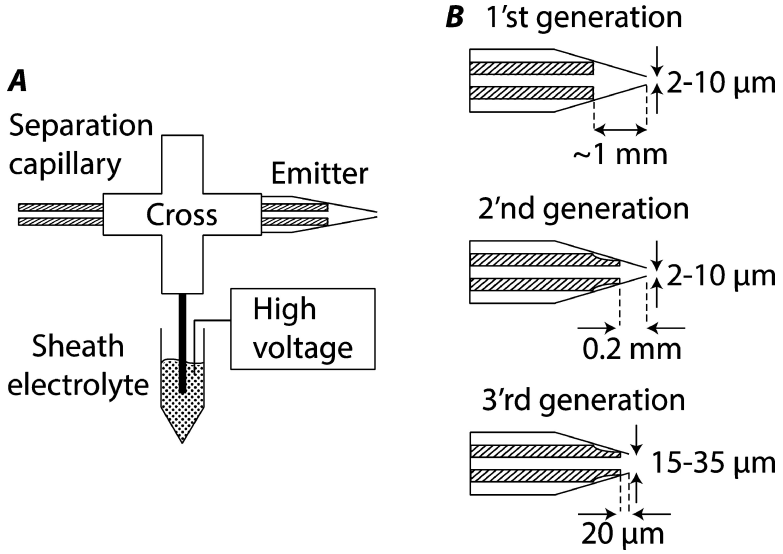


Figure 1.12. Diagrams of the basic design of the electrokinetically pumped sheath flow nanoelectrospray CE-MS interface (A) and its three different generations (B). The figure is reprinted with permission from reference [122].

CZE-MS has been extensively applied in both denatured and native TDP (nTDP) and offers high efficiency of separation and high sensitivity of detection. Our group applied the single-shot CZE-tandem MS (CZE-MS/MS) run of *E. coli* cell lysate and attained a peak capacity of 300 and proteoform identification of 600 under a 90-min separation window [124]. The Yates group demonstrated that CZE-MS offers significantly higher sensitivity for proteoforms compared to RPLC-MS, delivering comparable signal-to-noise ratios of protein targets while using 100-fold

less sample [125]. Our group developed a novel CZE-MS platform by coupling native CZE (nCZE) to a UHMR Orbitrap mass spectrometer to separate the whole *E. coli* lysate and detect nearly 100 protein/protein complexes ranging up to 400 kDa with only 50ng of sample consumption [126]. The Kelleher group reported a nCZE-top-down MS (nCZE-TDMS) system to obtain the attomole level of nucleosome characterization, with the detection of histone PTM profile changes [127]. However, a long-existing limitation of CZE-MS is the low sample loading capacity. In the CZE-MS analysis, typically, only 1% of the total capillary volume equivalent sample is injected to maintain the high separation efficiency. This limitation is a concern for TDP due to the intrinsic much lower sensitivity for intact protein measurement than peptides. Online capillary stacking methods are commonly used to increase the loading capacity, such as dynamic pH junction [128]. The dynamic pH junction method is based on applying the pH differences between the sample buffer and the BGE to concentrate the proteins at the pH boundary to achieve up to 50% capillary [124, 129].

1.2.2 Liquid chromatography (LC)

LC is one of the most fundamental analytical separation techniques via the liquid-phase sample interacting with the stationary phase and the mobile phase when flowing through the column packed with solid particles. LC offers a series of separation choices from distinct principles such as the analyte's hydrophobicity, size, and ionic strength. Reverse phase LC (RPLC) is the most dominant LC applied in proteomics studies based on the hydrophobicity of proteins or peptides.

RPLC contains a non-polar stationary phase and a polar mobile phase for separation and the retention of proteins or peptides on the stationary phase is directly correlated to their hydrophobicity. Isocratic elution and gradient elution by the mobile phase are the two common strategies when eluting the biomolecules from the stationary phase. Isocratic elution refers to the constant composition ratio and flow rate of the mobile phase, while gradient elution applies an increasing concentration of organic solvent in the mobile phase during analysis. The gradient elution is preferable when coupling to MS for proteomics because it provides improved resolution and sensitivity, enhanced peak capacity, and reduced analysis time over the isocratic elution. The gradual concentration change of the mobile phase enhances the resolution between closely eluting biomolecules, leading to sharper and more concentrated peaks which are beneficial for the sensitivity of detecting low-abundance biomolecules. Also, the gradient flow allows a broader

range of biomolecule elution within a shorter analysis time of a single run, increasing the number of biomolecules that can be effectively separated and improving the overall peak capacity. Acetonitrile (ACN) is frequently as the organic solvent in the mobile phase because of its strong solvent strength to effectively elute a wide range of biomolecules, low viscosity to significantly reduce the backpressure of the system and support high flow rates for fast analyses, and good volatility to be easily removed by evaporation and minimize the solvent contamination for downstream MS analysis [130].

The stationary phase of RPLC generally consists of porous silica particles covalently bonded with the alkyl chains. The selection of particle size, pore size, and alkyl chain can significantly affect the separation efficiency for RPLC. Applying smaller particles benefits the separation by decreasing eddy diffusion and resistance to mass transfer (**Equation 1.2**) to lower the height equivalent to a theoretical plate [131]. The pore size of the particles must keep a good balance between surface area and pore volume, allowing for efficient interactions between the analytes and the stationary phase. Larger pore size (typically 300 Å) gives better access to the interior of the silica particles for large biomolecules like intact proteins, while smaller pore size (typically 100 Å) is optimal for separating smaller biomolecules like peptides [132]. The alkyl chains provide the hydrophobicity for the stationary phase to interact and separate the analytes. C18 is composed of a linear arrangement of 18 carbon atoms and is most widely utilized for separating small molecules like peptides because of its strong hydrophobicity [133]. Shorter alkyl chains like C1-C4 are commonly for intact protein separations because of their low hydrophobicity to reduce the inevitable sample loss [132].

Microflow and nanoflow RPLC-MS are the two frequently used methods in proteomics studies. Microflow RPLC-MS is termed by the 0.5-1 mm i.d. analytical column with a flow rate of 10-200 μL/min, while nanoflow RPLC-MS refers to the columns with an i.d. <100 μm used at a flow rate <1 μL/min. Microflow RPLC-MS shows great capabilities in high-throughput analyses with excellent reproducibility in both retention time and protein quantification [134]. For example, the Kuster group demonstrated that microflow RPLC-MS could conduct up to 1,500 analyses in a month and process over 14,000 samples on a single column while maintaining consistent chromatographic performance [135]. However, the large sample consumption and sample loss from the high flow rate of microflow RPLC-MS are not ideal for limited sample analysis. The high flow rate also causes lower sensitivity for the downstream detection due to the

sample dilution in the column and the insufficient ionization efficiency in the electrospray. On the contrary, nanoflow RPLC-MS was developed to improve the sensitivity with lower sample consumption for the proteomics workflow [136-139]. The Olsen lab introduced an optimized fast and sensitive data-dependent acquisition method for nanoflow RPLC-MS by comparing the isotopically labeled yeast proteome [138]. With less than 125 ng of sample loading, the optimized nanoflow RPLC-MS workflow could identify and quantify above 2500 proteins from yeast proteome with 1 h of analysis time. The Mann group reported the high sensitivity with the single run nanoflow RPLC-MS, identifying approximately 1000 of a total of 5000 proteins from a human embryonic kidney cell line (HEK293) at an estimated detection limit of fewer than 100 attomoles per protein [139].

1.2.3 Ion mobility spectrometry (IMS)

IMS is a gas-phase separation technique measuring ions' trajectories under the influence of carrier gas and electric field, based on the sizes, charges, and shapes of ions. Owing to the intrinsic principle of IMS and MS both analyzing gas-phase charged ions, coupling IMS to MS shows an easy compatibility and can be firstly traced back to the 1970s [140]. IMS-MS has greatly improved the resolution of the analysis with complementary separations in mobility and mass dimensions, providing outstanding selectivity and sensitivity [141-143]. Revealing the structural information of analytes is a significant advantage of IMS-MS, typically achieved through the conversion of measured mobility into the calculated collision cross-section value (CCS). Several types of IMS are developed to create distinct instrument platforms coupled to MS, such as drift tube IMS (DTIMS), traveling wave IMS (TWIMS), TIMS, and field asymmetric IMS (FAIMS).

FAIMS commonly operates as an online mobility filter, positioned directly after the ion source and prior to the entrance region of the mass spectrometer. The gas-phase ions are transported under a parallel carrier gas flow between two electrodes where the asymmetric waveform electric field oscillates in the high and low field with opposite polarities [144-146]. Typically, the magnitude of the high field (reported as dispersion voltage, (DV)) is twice the low field, but the duration is half of the low field (**Figure 1.13A**) [144, 146]. The filtration of the ions is achieved by applying a small compensation voltage (CV) at the inner electrode to select ions with specific sizes, charges, and shapes passing through. The ion trajectories are generalized into three types (**Figure 1.13B**): ions with increasing mobility at high field (type A) such as

declustering from adducts with decreasing CCS, ions with initially increasing mobility but followed by decreasing mobility because of the high-energy collisions (type B), ions with decreasing mobility (type C) like protein unfolding with increasing CCS [144, 146]. The analysis of short-chain glycine-based peptides by ESI-FAIMS-MS demonstrated the transformation from type A to type C with the increasing peptide length due to the increasing CCS [147]. For proteomics studies, FAIMS tuned to transmit longer peptides will effectively exclude short peptides and small molecules, offering a cleaner background of the mass spectrum with higher sensitivity compared to no-FAIMS. FAIMS is typically added as an online additional dimension of separation for liquid-phase chromatography-MS (LC-MS) for both peptide and intact protein analyses [148, 149]. The Ivanov lab reported an optimized ultralow flow LC-FAIMS-MS platform to achieve up to 131% more protein identifications by applying four CVs within a single shot run from 1 ng of Hela digest, showing the high sensitivity of FAIMS [148]. The Petyuk group applied an Alzheimer's disease (AD) brain tissue sample by LC-FAIMS-MS to attain the double identification numbers of unique proteoforms with the external CV stepping compared to no-FAIMS [149]. Therefore, FAIMS has the potential to be coupled with different separation techniques to form online multi-dimensional separations to enhance the proteome coverage for MS-based proteomics studies.

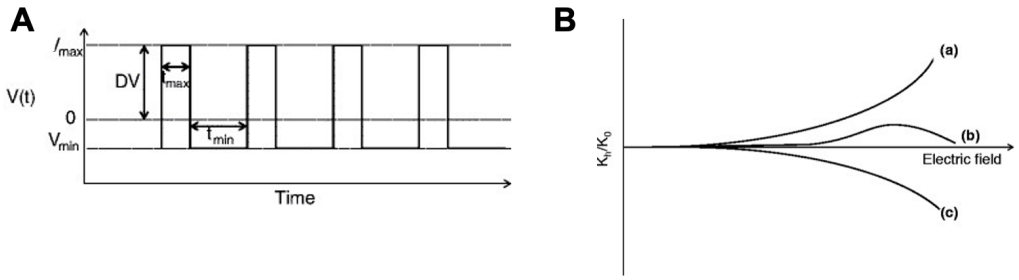


Figure 1.13. Separation principles of FAIMS. (A) asymmetric waveform on FAIMS electrodes; (B) three types of ion mobility in FAIMS. The figure is reprinted with permission from reference [146].

1.3 Summary

This chapter introduced multi-level MS-based multi-level and well-developed separation techniques for proteins and peptides. BUP and TDP are two popular complementary strategies in the characterization of peptides and intact proteoforms with their advantages and limitations. BUP suffers from limited sequence coverage and difficulties in differentiating proteoforms, while TDP

faces challenges with sensitivity and the identification of low-abundance proteoforms. However, it can be foreseen that coupling these two strategies (named multi-level proteomics) is a bright direction for better delineation of proteins with their PTMs to better the understanding of molecular mechanisms in cellular processes and diseases. CZE and RPLC are two liquid-phase separation methods presenting great potential and applications in both BUP and TDP. IMS is a gas-phase separation technique often coupled with MS to improve resolution and sensitivity by providing additional dimensions of separation. More advanced analytical techniques and workflows must be further developed to improve MS-based proteomics. The following chapters will build on these concepts to explore advanced methodologies and applications in proteomics research.

REFERENCES

[1] Hunter T. Signaling--2000 and beyond. Cell. 2000 Jan;100(1):113-27.

[2] Ruvolo PP, Deng X, May WS. Phosphorylation of Bcl2 and regulation of apoptosis. Leukemia. 2001 Apr;15(4):515-22.

[3] Wang C, Wang H, Zhang D, Luo W, Liu R, Xu D, Diao L, Liao L, Liu Z. Phosphorylation of ULK1 affects autophagosome fusion and links chaperone-mediated autophagy to macroautophagy. Nat Commun. 2018 Aug;9(1):3492.

[4] Gramaglia D, Gentile A, Battaglia M, Ranzato L, Petronilli V, Fassetta M, Bernardi P, Rasola A. Apoptosis to necrosis switching downstream of apoptosome formation requires inhibition of both glycolysis and oxidative phosphorylation in a BCL-X(L)- and PKB/AKT-independent fashion. Cell Death Differ. 2004 Mar;11(3):342-53.

[5] Hans F, Dimitrov S. Histone H3 phosphorylation and cell division. Oncogene. 2001 May;20(24):3021-7.

[6] Burnum-Johnson KE, Conrads TP, Drake RR, Herr AE, Iyengar R, Kelly RT, Lundberg E, MacCoss MJ, Naba A, Nolan GP, Pevzner PA, Rodland KD, Sechi S, Slavov N, Spraggins JM, Van Eyk JE, Vidal M, Vogel C, Walt DR, Kelleher NL. New Views of Old Proteins: Clarifying the Enigmatic Proteome. Mol Cell Proteomics. 2022 Jul;21(7):100254.

[7] Smith LM, Kelleher NL; Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. Nat Methods. 2013 Mar;10(3):186-7.

[8] Tyers M, Mann M. From genomics to proteomics. Nature. 2003 Mar;422(6928):193-7.

[9] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003 Mar;422(6928):198-207.

[10] Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature. 2016 Sep;537(7620):347-55.

[11] Gregorich ZR, Chang YH, Ge Y. Proteomics in heart failure: top-down or bottom-up? Pflugers Arch. 2014 Jun;466(6):1199-209.

[12] Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR 3rd. Protein analysis by shotgun/bottom-up proteomics. Chem Rev. 2013 Apr;113(4):2343-94.

[13] Schubert OT, Röst HL, Collins BC, Rosenberger G, Aebersold R. Quantitative proteomics: challenges and opportunities in basic and applied research. Nat Protoc. 2017 Jul;12(7):1289-1294.

[14] Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol. 1999 Oct;17(10):994-9.

[15] Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002 May;1(5):376-86.

[16] Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem. 2003 Apr;75(8):1895-904.

[17] Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics. 2004 Dec;3(12):1154-69.

[18] Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci U S A. 2003 Jun;100(12):6940-5.

[19] Michna T, Tenzer S. Quantitative Proteome and Phosphoproteome Profiling in Magnaporthe oryzae. Methods Mol Biol. 2021;2356:109-119.

[20] Li J, Cai Z, Bomgarden RD, Pike I, Kuhn K, Rogers JC, Roberts TM, Gygi SP, Paulo JA. TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. J Proteome Res. 2021 May;20(5):2964-2972.

[21] Voyksner RD, Lee H. Investigating the use of an octupole ion guide for ion storage and high-pass mass filtering to improve the quantitative performance of electrospray ion trap mass spectrometry. Rapid Commun Mass Spectrom. 1999;13(14):1427-37.

[22] Wiener MC, Sachs JR, Deyanova EG, Yates NA. Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. Anal Chem. 2004 Oct;76(20):6085-96.

[23] Bondarenko PV, Chelius D, Shaler TA. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. Anal Chem. 2002 Sep;74(18):4741-9.

[24] Rozanova S, Barkovits K, Nikolov M, Schmidt C, Urlaub H, Marcus K. Quantitative Mass Spectrometry-Based Proteomics: An Overview. Methods Mol Biol. 2021;2228:85-116.

[25] Li H, Nguyen HH, Ogorzalek Loo RR, Campuzano IDG, Loo JA. An integrated native mass spectrometry and top-down proteomics method that connects sequence to structure and function of macromolecular complexes. Nat Chem. 2018 Feb;10(2):139-148.

[26] Keener JE, Zambrano DE, Zhang G, Zak CK, Reid DJ, Deodhar BS, Pemberton JE, Prell JS, Marty MT. Chemical Additives Enable Native Mass Spectrometry Measurement of Membrane Protein Oligomeric State within Intact Nanodiscs. J Am Chem Soc. 2019 Jan;141(2):1054-1061.

[27] Wörner TP, Snijder J, Bennett A, Agbandje-McKenna M, Makarov AA, Heck AJR. Resolving heterogeneous macromolecular assemblies by Orbitrap-based single-particle charge detection mass spectrometry. Nat Methods. 2020 Apr;17(4):395-398.

[28] Skinner OS, Haverland NA, Fornelli L, Melani RD, Do Vale LHF, Seckler HS, Doubleday PF, Schachner LF, Srzentić K, Kelleher NL, Compton PD. Top-down characterization of endogenous protein complexes with native proteomics. Nat Chem Biol. 2018 Jan;14(1):36-41.

[29] Schaffer LV, Tucholski T, Shortreed MR, et al. Intact-Mass Analysis Facilitating the Identification of Large Human Heart Proteoforms. Anal Chem. 2019; 91(17):10937-10942.

[30] 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(16):6172-6179.

[31] Xu T, Wang Q, Wang Q, et al. 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.

[32] 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(24):9347-9356.

[33] Chen D, Yang Z, Shen X, et al. Capillary Zone Electrophoresis-Tandem Mass Spectrometry As an Alternative to Liquid Chromatography-Tandem Mass Spectrometry for Top-down Proteomics of Histones. Anal Chem. 2021; 93(10):4417-4424.

[34] Wang Q, Fang F, Wang Q, et al. Capillary zone electrophoresis-high field asymmetric ion mobility spectrometry-tandem mass spectrometry for top-down characterization of histone proteoforms. Proteomics. 2024 Feb;24(3-4):e2200389.

[35] Walker JN, Lam R, Brodbelt JS. Enhanced Characterization of Histones Using 193 nm Ultraviolet Photodissociation and Proton Transfer Charge Reduction. Anal Chem. 2023; 95(14):5985-5993.

[36] Melani RD, Gerbasi VR, Anderson LC, et al. The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. Science. 2022; 375(6579):411-418.

[37] McCool EN, Xu T, Chen W, Beller NC, Nolan SM, Hummon AB, Liu X, Sun L. Deep top-down proteomics revealed significant proteoform-level differences between metastatic and nonmetastatic colorectal cancer cells. Sci Adv. 2022 Dec;8(51):eabq6348.

[38] Jeong K, Kim J, Gaikwad M, et al. FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics. Cell Systems. 2020;10:213-218.e6.

[39] Martin EA, Fulcher JM, Zhou M, et al. TopPICR: A Companion R Package for Top-Down Proteomics Data Analysis. J Proteome Res. 2023;22:399–409.

[40] Tiambeng TN, Wu Z, Melby JA, et al. Size Exclusion Chromatography Strategies and MASH Explorer for Large Proteoform Characterization. Methods Mol Biol. 2022; 2500: 15-30.

[41] Fornelli L, Toby TK. Characterization of large intact protein ions by mass spectrometry: What directions should we follow?. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics. 2022;1870(4):140758.

[42] Riley NM, Westphall MS, Coon JJ. Sequencing Larger Intact Proteins (30-70 kDa) with Activated Ion Electron Transfer Dissociation. J Am Soc Mass Spectrom. 2018; 29(1):140-149.

[43] Greisch JF, den Boer MA, Lai SH, et al. Extending Native Top-Down Electron Capture Dissociation to MDa Immunoglobulin Complexes Provides Useful Sequence Tags Covering Their Critical Variable Complementarity-Determining Regions. Anal Chem. 2021; 93(48):16068-16075.

[44] Shaw JB, Li W, Holden DD, et al. Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation. J Am Chem Soc. 2013; 135(34):12646-51.

[45] Harvey SR, Porrini M, Konijnenberg A, et al. Dissecting the dynamic conformations of the metamorphic protein lymphotactin. J Phys Chem B. 2014;118(43):12348-59.

[46] Durbin KR, Skinner OS, Fellers RT, et al. Analyzing internal fragmentation of electrosprayed ubiquitin ions during beam-type collisional dissociation. J Am Soc Mass Spectrom. 2015 May;26(5):782-7.

[47] Jenuwein T, Allis CD. Translating the histone code. Science. 2001; 293(5532):1074-80.

[48] Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nat Rev Genet. 2016; 17(8):487-500.

[49] LeDuc RD, Taylor GK, Kim YB, et al. ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. Nucleic Acids Res. 2004; 32(Web Server issue):W340-5.

[50] 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.

[51] Stefan K Solntsev, Michael R Shortreed, Brian L Frey, Lloyd M Smith. Enhanced Global Post-translational Modification Discovery with MetaMorpheus. J Proteome Res. 2018 May;17(5):1844-1851.

[52] Cai W, Guner H, Gregorich ZR, et al. MASH Suite Pro: A Comprehensive Software Tool for Top-Down Proteomics. Mol Cell Proteomics. 2016; 15(2):703-14.

[53] Lantz C, Zenaidee MA, Wei B, et al. ClipsMS: An Algorithm for Analyzing Internal Fragments Resulting from Top-Down Mass Spectrometry. J Proteome Res. 2021; 20(4):1928-1935.

[54] Park J, Piehowski PD, Wilkins C, et al. Informed-Proteomics: Open Source Software Package for Top-down Proteomics. Nat Methods. 2017;14:909–914.

[55] Meissner F, Geddes-McAlister J, Mann M, Bantscheff M. The emerging role of mass spectrometry-based proteomics in drug discovery. Nat Rev Drug Discov. 2022 Sep;21(9):637-654.

[56] Konermann L, Ahadi E, Rodriguez AD, Vahidi S. Unraveling the mechanism of electrospray ionization. Anal Chem. 2013 Jan 2;85(1):2-9.

[57] Ho CS, Lam CW, Chan MH, Cheung RC, Law LK, Lit LC, Ng KF, Suen MW, Tai HL. Electrospray ionisation mass spectrometry: principles and clinical applications. Clin Biochem Rev. 2003;24(1):3-12.

[58] Honour JW. Benchtop mass spectrometry in clinical biochemistry. Ann Clin Biochem. 2003 Nov;40(Pt 6):628-38.

[59] Chernushevich IV, Loboda AV, Thomson BA. An introduction to quadrupole-time-of-flight mass spectrometry. J Mass Spectrom. 2001 Aug;36(8):849-65.

[60] Hopfgartner G, Varesio E, Tschäppät V, Grivet C, Bourgogne E, Leuthold LA. Triple quadrupole linear ion trap mass spectrometer for the analysis of small molecules and macromolecules. J Mass Spectrom. 2004 Aug;39(8):845-55.

[61] Glish GL, Burinsky DJ. Hybrid mass spectrometers for tandem mass spectrometry. J Am Soc Mass Spectrom. 2008 Feb;19(2):161-72.

[62] Glish GL, Vachet RW. The basics of mass spectrometry in the twenty-first century. Nat Rev Drug Discov. 2003 Feb;2(2):140-50.

[63] 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.

[64] Yost RA, Enke CG. Triple quadrupole mass spectrometry for direct mixture analysis and structure elucidation. Anal chem. 1979 Oct;51(12):1251-64.

[65] Douglas DJ. Linear quadrupoles in mass spectrometry. Mass Spectrom Rev. 2009 Nov-Dec;28(6):937-60.

[66] Makarov A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem. 2000 Mar;72(6):1156-62.

[67] Savaryn JP, Toby TK, Kelleher NL. A researcher's guide to mass spectrometry-based proteomics. Proteomics. 2016 Sep;16(18):2435-43.

[68] van de Waterbeemd M, Fort KL, Boll D, Reinhardt-Szyba M, Routh A, Makarov A, Heck AJ. High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. Nat Methods. 2017 Mar;14(3):283-286.

[69] Marie AL, Georgescauld F, Johnson KR, Ray S, Engen JR, Ivanov AR. Native Capillary Electrophoresis-Mass Spectrometry of Near 1 MDa Non-Covalent GroEL/GroES/Substrate Protein Complexes. Adv Sci (Weinh). 2024 Mar;11(11):e2306824.

[70] Fjeldsted J. Time-of-flight mass spectrometry. Technical overview. Agilent Technologies. 2003:22.

[71] Radionova A, Filippov I, Derrick PJ. In pursuit of resolution in time-of-flight mass spectrometry: A historical perspective. Mass Spectrom Rev. 2016 Oct;35(6):738-757.

[72] Cotter RJ, Griffith W, Jelinek C. Tandem time-of-flight (TOF/TOF) mass spectrometry and the curved-field reflectron. J Chromatogr B Analyt Technol Biomed Life Sci. 2007 Aug;855(1):2-13.

[73] Delannoy CP, Heuson E, Herledan A, Oger F, Thiroux B, Chevalier M, Gromada X, Rolland L, Froguel P, Deprez B, Paul S, Annicotte JS. High-Throughput Quantitative Screening of Glucose-Stimulated Insulin Secretion and Insulin Content Using Automated MALDI-TOF Mass Spectrometry. Cells. 2023 Mar;12(6):849.

[74] Meier F, Park MA, Mann M. Trapped Ion Mobility Spectrometry and Parallel Accumulation-Serial Fragmentation in Proteomics. Mol Cell Proteomics. 2021;20:100138.

[75] Kafka AP, Kleffmann T, Rades T, McDowell A. The application of MALDI TOF MS in biopharmaceutical research. Int J Pharm. 2011 Sep;417(1-2):70-82.

[76] Illiano A, Pinto G, Melchiorre C, Carpentieri A, Faraco V, Amoresano A. Protein Glycosylation Investigated by Mass Spectrometry: An Overview. Cells. 2020 Aug;9(9):1986.

[77] Huang Y, Pasa-Tolić L, Guan S, Marshall AG. Collision-induced dissociation for mass spectrometric analysis of biopolymers: high-resolution Fourier transform ion cyclotron resonance MS4. Anal Chem. 1994 Dec;66(24):4385-9.

[78] Olsen JV, Macek B, Lange O, Makarov A, Horning S, Mann M. Higher-energy C-trap dissociation for peptide modification analysis. Nat Methods. 2007 Sep;4(9):709-12.

[79] McAlister GC, Phanstiel DH, Brumbaugh J, Westphall MS, Coon JJ. Higher-energy collision-activated dissociation without a dedicated collision cell. Mol Cell Proteomics. 2011 May;10(5):O111.009456.

[80] Huang Y, Triscari JM, Tseng GC, Pasa-Tolic L, Lipton MS, Smith RD, Wysocki VH. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. Anal Chem. 2005 Sep;77(18):5800-13.

[81] Scigelova M, Hornshaw M, Giannakopulos A, Makarov A. Fourier transform mass spectrometry. Mol Cell Proteomics. 2011 Jul;10(7):M111.009431.

[82] Mikesh LM, Ueberheide B, Chi A, Coon JJ, Syka JE, Shabanowitz J, Hunt DF. The utility of ETD mass spectrometry in proteomic analysis. Biochim Biophys Acta. 2006 Dec;1764(12):1811-22.

[83] Zubarev RA, Kelleher NL, McLafferty FW. Electron capture dissociation of multiply charged protein cations. A nonergodic process. J Am Chem Soc. 1998 Apr;120(13):3265-6.

[84] Compton PD, Strukl JV, Bai DL, Shabanowitz J, Hunt DF. Optimization of electron transfer dissociation via informed selection of reagents and operating parameters. Anal Chem. 2012 Feb;84(3):1781-5.

[85] Kim MS, Pandey A. Electron transfer dissociation mass spectrometry in proteomics. Proteomics. 2012 Feb;12(4-5):530-42.

[86] Good DM, Wirtala M, McAlister GC, Coon JJ. Performance characteristics of electron transfer dissociation mass spectrometry. Mol Cell Proteomics. 2007 Nov;6(11):1942-51.

[87] Coon JJ, Ueberheide B, Syka JE, Dryhurst DD, Ausio J, Shabanowitz J, Hunt DF. Protein identification using sequential ion/ion reactions and tandem mass spectrometry. Proc Natl Acad Sci U S A. 2005 Jul;102(27):9463-8.

[88] Hogan JM, Pitteri SJ, Chrisman PA, McLuckey SA. Complementary structural information from a tryptic N-linked glycopeptide via electron transfer ion/ion reactions and collision-induced dissociation. J Proteome Res. 2005 Mar-Apr;4(2):628-32.

[89] Zhang Q, Frolov A, Tang N, Hoffmann R, van de Goor T, Metz TO, Smith RD. Application of electron transfer dissociation mass spectrometry in analyses of non-enzymatically glycated peptides. Rapid Commun Mass Spectrom. 2007;21(5):661-6.

[90] Liu J, McLuckey SA. Electron Transfer Dissociation: Effects of Cation Charge State on Product Partitioning in Ion/Ion Electron Transfer to Multiply Protonated Polypeptides. Int J Mass Spectrom. 2012 Dec;330-332:174-181.

[91] Pitteri SJ, Chrisman PA, Hogan JM, McLuckey SA. Electron transfer ion/ion reactions in a three-dimensional quadrupole ion trap: reactions of doubly and triply protonated peptides with SO2*-. Anal Chem. 2005 Mar;77(6):1831-9.

[92] Pitteri SJ, Chrisman PA, McLuckey SA. Electron-transfer ion/ion reactions of doubly protonated peptides: effect of elevated bath gas temperature. Anal Chem. 2005 Sep;77(17):5662-9.

[93] Horn DM, Ge Y, McLafferty FW. Activated ion electron capture dissociation for mass spectral sequencing of larger (42 kDa) proteins. Anal Chem. 2000 Oct;72(20):4778-84.

[94] Ge Y, Lawhorn BG, ElNaggar M, Strauss E, Park JH, Begley TP, McLafferty FW. Top down characterization of larger proteins (45 kDa) by electron capture dissociation mass spectrometry. J Am Chem Soc. 2002 Jan;124(4):672-8.

[95] Yu Q, Wang B, Chen Z, Urabe G, Glover MS, Shi X, Guo LW, Kent KC, Li L. Electron-Transfer/Higher-Energy Collision Dissociation (EThcD)-Enabled Intact Glycopeptide/Glycoproteome Characterization. J Am Soc Mass Spectrom. 2017 Sep;28(9):1751-1764.

[96] Frese CK, Altelaar AF, van den Toorn H, Nolting D, Griep-Raming J, Heck AJ, Mohammed S. Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. Anal Chem. 2012 Nov;84(22):9668-73.

[97] Riley NM, Coon JJ. The Role of Electron Transfer Dissociation in Modern Proteomics. Anal Chem. 2018 Jan;90(1):40-64.

[98] Brodbelt JS, Morrison LJ, Santos I. Ultraviolet Photodissociation Mass Spectrometry for Analysis of Biological Molecules. Chem Rev. 2020 Apr;120(7):3328-3380.

[99] Bowers WD, Delbert SS, Hunter RL, McIver Jr RT. Fragmentation of oligopeptide ions using ultraviolet laser radiation and Fourier transform mass spectrometry. J Am Chem Soc. 1984 Nov;106(23):7288-9.

[100] Toby TK, Fornelli L, Kelleher NL. Progress in Top-Down Proteomics and the Analysis of Proteoforms. Annu Rev Anal Chem (Palo Alto Calif). 2016 Jun;9(1):499-519.

[101] Cleland TP, DeHart CJ, Fellers RT, VanNispen AJ, Greer JB, LeDuc RD, Parker WR, Thomas PM, Kelleher NL, Brodbelt JS. High-Throughput Analysis of Intact Human Proteins Using UVPD and HCD on an Orbitrap Mass Spectrometer. J Proteome Res. 2017 May;16(5):2072-2079.

[102] Fort KL, Dyachenko A, Potel CM, Corradini E, Marino F, Barendregt A, Makarov AA, Scheltema RA, Heck AJ. Implementation of Ultraviolet Photodissociation on a Benchtop Q Exactive Mass Spectrometer and Its Application to Phosphoproteomics. Anal Chem. 2016 Feb;88(4):2303-10.

[103] Madsen JA, Cheng RR, Kaoud TS, Dalby KN, Makarov DE, Brodbelt JS. Charge-site-dependent dissociation of hydrogen-rich radical peptide cations upon vacuum UV photoexcitation. Chemistry. 2012 Apr;18(17):5374-83.

[104] Madsen JA, Kaoud TS, Dalby KN, Brodbelt JS. 193-nm photodissociation of singly and multiply charged peptide anions for acidic proteome characterization. Proteomics. 2011 Apr;11(7):1329-34.

[105] Greer SM, Holden DD, Fellers R, Kelleher NL, Brodbelt JS. Modulation of Protein Fragmentation Through Carbamylation of Primary Amines. J Am Soc Mass Spectrom. 2017 Aug;28(8):1587-1599.

[106] Li Y, Champion MM, Sun L, Champion PA, Wojcik R, Dovichi NJ. Capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry as an alternative proteomics platform to ultraperformance liquid chromatography-electrospray ionization-tandem mass spectrometry for samples of intermediate complexity. Anal Chem. 2012 Feb;84(3):1617-22.

[107] Faserl K, Sarg B, Kremser L, Lindner H. Optimization and evaluation of a sheathless capillary electrophoresis-electrospray ionization mass spectrometry platform for peptide analysis: comparison to liquid chromatography-electrospray ionization mass spectrometry. Anal Chem. 2011 Oct;83(19):7297-305.

[108] Zhu G, Sun L, Yan X, Dovichi NJ. Single-shot proteomics using capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry with production of more than 1250 Escherichia coli peptide identifications in a 50 min separation. Anal Chem. 2013 Mar;85(5):2569-73.

[109] Ramautar R, Heemskerk AA, Hensbergen PJ, Deelder AM, Busnel JM, Mayboroda OA. CE-MS for proteomics: Advances in interface development and application. J Proteomics. 2012 Jul;75(13):3814-28.

[110] Wang Y, Fonslow BR, Wong CC, Nakorchevsky A, Yates JR 3rd. Improving the comprehensiveness and sensitivity of sheathless capillary electrophoresis-tandem mass spectrometry for proteomic analysis. Anal Chem. 2012 Oct;84(20):8505-13.

[111] Colmsjö AL, Ericsson MW. Assessment of the height equivalent to a theoretical plate in liquid chromatography. J Chromatogr A. 1987 Jan 1;398:63-71.

[112] van Deemter JJ, Zuiderweg FJ, Klinkenberg AV. Longitudinal diffusion and resistance to mass transfer as causes of nonideality in chromatography. Chem Eng. Sci. 1956 Sep;5(6):271-89.

[113] Billen J, Desmet G. Understanding and design of existing and future chromatographic support formats. J Chromatogr A. 2007 Oct;1168(1-2):73-99; discussion 71-2.

[114] Lubeckyj RA, Basharat AR, 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 Aug;30(8):1435-1445.

[115] 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 Jan;146:839-43.

[116] Haselberg R, de Jong GJ, Somsen GW. Low-flow sheathless capillary electrophoresis-mass spectrometry for sensitive glycoform profiling of intact pharmaceutical proteins. Anal Chem. 2013 Feb;85(4):2289-96.

[117] Busnel JM, Schoenmaker B, Ramautar R, Carrasco-Pancorbo A, Ratnayake C, Feitelson JS, Chapman JD, Deelder AM, Mayboroda OA. High capacity capillary electrophoresis-electrospray ionization mass spectrometry: coupling a porous sheathless interface with transient-isotachophoresis. Anal Chem. 2010 Nov;82(22):9476-83.

[118] Mellors JS, Gorbounov V, Ramsey RS, Ramsey JM. Fully integrated glass microfluidic device for performing high-efficiency capillary electrophoresis and electrospray ionization mass spectrometry. Anal Chem. 2008 Sep;80(18):6881-7.

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

[120] Maxwell EJ, Zhong X, Zhang H, van Zeijl N, Chen DDY. Decoupling CE and ESI for a more robust interface with MS. Electrophoresis. 2010 Apr;31(7):1130-1137.

[121] Wojcik R, Dada OO, Sadilek M, Dovichi NJ. Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. Rapid Commun Mass Spectrom. 2010 Sep;24(17):2554-60.

[122] Sun L, Zhu G, Zhang Z, Mou S, Dovichi NJ. 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 May;14(5):2312-21.

[123] Sun L, Zhu G, Zhao Y, Yan X, Mou S, Dovichi NJ. Ultrasensitive and fast bottom-up analysis of femtogram amounts of complex proteome digests. Angew Chem Int Ed Engl. 2013 Dec;52(51):13661-4.

[124] Lubeckyj RA, McCool EN, 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 Nov;89(22):12059-12067.

[125] Han X, Wang Y, Aslanian A, Fonslow B, Graczyk B, Davis TN, Yates JR 3rd. In-line separation by capillary electrophoresis prior to analysis by top-down mass spectrometry enables sensitive characterization of protein complexes. J Proteome Res. 2014 Dec;13(12):6078-86.

[126] Wang Q, Wang Q, Qi Z, Moeller W, Wysocki VH, Sun L. Native Proteomics by Capillary Zone Electrophoresis-Mass Spectrometry. bioRxiv [Preprint]. 2024 Jul:2024.04.24.590970.

[127] Jooß K, Schachner LF, Watson R, Gillespie ZB, Howard SA, Cheek MA, Meiners MJ, Sobh A, Licht JD, Keogh MC, Kelleher NL. Separation and Characterization of Endogenous Nucleosomes by Native Capillary Zone Electrophoresis-Top-Down Mass Spectrometry. Anal Chem. 2021 Mar;93(12):5151-5160.

[128] Zhu G, Sun L, Yan X, Dovichi NJ. Bottom-up proteomics of Escherichia coli using dynamic pH junction preconcentration and capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry. Anal Chem. 2014 Jul;86(13):6331-6.

[129] Shen X, Yang Z, McCool EN, Lubeckyj RA, Chen D, Sun L. Capillary zone electrophoresis-mass spectrometry for top-down proteomics. Trends Analyt Chem. 2019 Nov;120:115644.

[130] Rafferty JL, Siepmann JI, Schure MR. Mobile phase effects in reversed-phase liquid chromatography: a comparison of acetonitrile/water and methanol/water solvents as studied by molecular simulation. J Chromatogr A. 2011 Apr;1218(16):2203-13.

[131] Wang Z, Ma H, Smith K, Wu S. Two-Dimensional Separation Using High-pH and Low-pH Reversed Phase Liquid Chromatography for Top-down Proteomics. Int J Mass Spectrom. 2018 Apr;427:43-51.

[132] Shen Y, Tolić N, Piehowski PD, Shukla AK, Kim S, Zhao R, Qu Y, Robinson E, Smith RD, Paša-Tolić L. High-resolution ultrahigh-pressure long column reversed-phase liquid chromatography for top-down proteomics. J Chromatogr A. 2017 May;1498:99-110.

[133] Ishihama Y, Rappsilber J, Andersen JS, Mann M. Microcolumns with self-assembled particle frits for proteomics. J Chromatogr A. 2002 Dec;979(1-2):233-9.

[134] Bian Y, Zheng R, Bayer FP, Wong C, Chang YC, Meng C, Zolg DP, Reinecke M, Zecha J, Wiechmann S, Heinzlmeir S, Scherr J, Hemmer B, Baynham M, Gingras AC, Boychenko O, Kuster B. Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC-MS/MS. Nat Commun. 2020 Jan;11(1):157.

[135] Bian Y, Bayer FP, Chang YC, Meng C, Hoefer S, Deng N, Zheng R, Boychenko O, Kuster B. Robust Microflow LC-MS/MS for Proteome Analysis: 38 000 Runs and Counting. Anal Chem. 2021 Mar;93(8):3686-3690.

[136] Shen Y, Tolić N, Masselon C, Pasa-Tolić L, Camp DG 2nd, Hixson KK, Zhao R, Anderson GA, Smith RD. Ultrasensitive proteomics using high-efficiency on-line micro-SPE-nanoLC-nanoESI MS and MS/MS. Anal Chem. 2004 Jan;76(1):144-54.

[137] Smith RD, Shen Y, Tang K. Ultrasensitive and quantitative analyses from combined separations-mass spectrometry for the characterization of proteomes. Acc Chem Res. 2004 Apr;37(4):269-78.

[138] Kelstrup CD, Young C, Lavallee R, Nielsen ML, Olsen JV. Optimized fast and sensitive acquisition methods for shotgun proteomics on a quadrupole orbitrap mass spectrometer. J Proteome Res. 2012 Jun;11(6):3487-97.

[139] Thakur SS, Geiger T, Chatterjee B, Bandilla P, Fröhlich F, Cox J, Mann M. Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. Mol Cell Proteomics. 2011 Aug;10(8):M110.003699.

[140] Cohen MJ, Karasek FW. Plasma chromatography™—a new dimension for gas chromatography and mass spectrometry. J Chromatogr Sci. 1970 Jun;8(6):330-7.

[141] Kanu AB, Dwivedi P, Tam M, Matz L, Hill HH Jr. Ion mobility-mass spectrometry. J Mass Spectrom. 2008 Jan;43(1):1-22.

[142] McLean JA, Ruotolo BT, Gillig KJ, Russell DH. Ion mobility–mass spectrometry: a new paradigm for proteomics. Int J Mass Spectrom. 2005 Feb;240(3):301-15.

[143] Kliman M, May JC, McLean JA. Lipid analysis and lipidomics by structurally selective ion mobility-mass spectrometry. Biochim Biophys Acta. 2011 Nov;1811(11):935-45.

[144] Swearingen KE, Moritz RL. High-field asymmetric waveform ion mobility spectrometry for mass spectrometry-based proteomics. Expert Rev Proteomics. 2012 Oct;9(5):505-17.

[145] Dodds JN, Baker ES. Ion Mobility Spectrometry: Fundamental Concepts, Instrumentation, Applications, and the Road Ahead. J Am Soc Mass Spectrom. 2019 Nov;30(11):2185-2195.

[146] Cooper HJ. To What Extent is FAIMS Beneficial in the Analysis of Proteins? J Am Soc Mass Spectrom. 2016 Apr;27(4):566-77.

[147] Purves RW, Guevremont R. Electrospray ionization high-field asymmetric waveform ion mobility spectrometry-mass spectrometry. Anal Chem. 1999 Jul;71(13):2346-57.

[148] Greguš M, Kostas JC, Ray S, Abbatiello SE, Ivanov AR. Improved Sensitivity of Ultralow Flow LC-MS-Based Proteomic Profiling of Limited Samples Using Monolithic Capillary Columns and FAIMS Technology. Anal Chem. 2020 Nov;92(21):14702-14712.

[149] Fulcher JM, Makaju A, Moore RJ, Zhou M, Bennett DA, De Jager PL, Qian WJ, Paša-Tolić L, Petyuk VA. Enhancing Top-Down Proteomics of Brain Tissue with FAIMS. J Proteome Res. 2021 May;20(5):2780-2795.

CHAPTER 2. High-throughput bottom-up proteomics of human plasma enabled by advanced CZE-MS/MS and nanoparticle protein corona

2.1 Introduction

Nanomedicine has gained significant interest in pharmaceutical research due to the promising capability of drug targeting and drug delivery [1-6]. The therapeutic efficacy, targeting ability, toxicity, cellular interactions, and biodistribution of nanoparticle-bound medicine are heavily influenced by the formation of the biomolecule corona, i.e., protein corona [7-10]. Nanoparticle protein corona refers to a layer of proteins that are naturally attached to the surface of nanoparticles (NPs) when they enter a biological environment, such as blood or other body fluids [11-13]. It has been well established that the distinct profiles of nanoparticle protein corona layers can not only reflect the complex thermodynamics, kinetics, and biological interactions of NPs but also provide a snapshot of the proteome information [13-18]. Therefore, studying the composition of the protein corona has the potential to reveal the biological identity of NPs and provide insights into the proteome of the surrounding biological environment.

Blood plasma plays a central and integrative role in human physiology, acting as a universal reflection of an individual's state or phenotype for disease diagnosis and therapeutic monitoring [19]. However, plasma proteomics is challenging as the broad dynamic range of protein abundance in plasma remains the major difficulty [20]. 22 proteins compose 99% of plasma proteins by mass with albumin alone contributing 55% [21-22]. The presence of highly abundant proteins dominates the mass spectra during mass spectrometry (MS) analysis, hindering the in-depth analysis of plasma proteome and comprehensive proteome coverage. To improve this, one of the emerging applications of nanoparticle protein corona is to reduce blood plasma proteome complexity and protein concentration dynamic range, facilitating the detection and identification of low-abundance disease-associated biomarkers [23-26].

Quantitative proteomics offers valuable insights into the biological states of relevant cells or tissues and has significantly advanced both biological and clinically focused research [27]. Label-free quantification (LFQ) and isobaric tagging strategies (i.e., TMT and iTRAQ) are commonly used in quantitative proteomics to identify the differentially expressed proteins for understanding the dynamics of protein-protein interactions across distinct cellular states and uncovering disease-related molecular mechanisms for better diagnosis [28-32]. MS-based bottom-up proteomics (BUP) is broadly recognized as an effective approach for characterizing the protein

corona, allowing for the precise identification and quantification of proteins adsorbed on nanoparticle surfaces [23-25, 33]. The throughput remains a significant challenge in plasma studies using BUP for achieving rapid clinical diagnostics of biomarkers because the enzymatic digestion step in typical BUP workflow is time-consuming (4-18 h). Shen, et al. demonstrated comparable numbers of protein identifications (IDs) from 15 min of immobilized trypsin digestion and 12 h of free trypsin digestion of mouse brain tissue sample, proving the application of rapid tryptic digestion for high-throughput BUP by NPs [34]. Capillary zone electrophoresis (CZE)-tandem MS (MS/MS) has been widely acknowledged as a valuable technique for BUP such as highly sensitive analysis of mass-limited biological samples, analysis of disease-related biomarkers, large-scale quantitative analysis, and many others [35-40]. CZE is a high-efficiency separation method based on an analyte's electrophoretic mobility. The use of shorter capillaries can accelerate CZE separations, making it a promising approach for high-throughput proteomics.

In this work, we developed a high-throughput BUP workflow for plasma/serum analysis by coupling nanoparticle protein corona, rapid on-bead tryptic digestion, and CZE-MS/MS. We first compared the workflow using 4 types of magnetic NPs with distinct functional groups on the nanoparticle surface and healthy human plasma with SDS-PAGE and CZE-MS/MS. Next, we applied the optimized workflow containing amine-terminated and carboxylate-terminated NPs to a pair of mouse serum samples (healthy and NUT cancer). We identified hundreds of proteins from plasma/serum samples with high throughput in 3.5 hours of total analysis time using nanoparticle protein corona, fast protein digestion, and CZE-MS/MS. Overall, we discovered potential cancer biomarkers by a quantitative proteomics analysis of the pair of mouse serum samples.

2.2 Experimental section

2.2.1 Ethical statement

All animal work was carried out under PROTO202000143 and PROTO 202300127, approved by the Michigan State University (MSU) Campus Animal Resources (CAR) and Institutional Animal Care and Use Committee (IACUC) in AAALAC credited facilities.

2.2.2 Materials and reagents

Amine-terminated NPs (Catalog #BP617) and carboxylate-terminated NPs (Catalog #BP618) were purchased from Bangs Laboratories, Inc. (Fishers, IN). Single-pot solid-phase-enhanced sample preparations (SP3) hydrophilic NPs (Catalog # 45152105050250) and SP3

hydrophobic NPs (Catalog # 65152105050250) were obtained from Cytiva (Marlborough, MA). Dulbecco's Phosphate-Buffered Saline (DPBS, 1X), Sodium dodecyl sulfate (SDS), ammonium bicarbonate (ABC), and dithiothreitol (DTT) were from Sigma-Aldrich (St. Louis, MO). Plain polystyrene NPs were obtained from Polysciences (www.polysciences.com). Trypsin (Bovine pancreas TPCK-treated), formic acid (FA), acetonitrile (ACN), methanol, LC/MS grade water, and bicinchoninic acid (BCA) assay kit were purchased from Fisher Scientific (Pittsburgh, PA). The protein LoBind tube was from Eppendorf (Enfield, CT). Healthy human plasma protein was purchased from Innovative Research (www.innov-research.com) and diluted to 55% using 1X DPBS. Biological triplicates of healthy and NUT cancer mouse serum (healthy: BN010, BN012, and BN110; NUT cancer: KBN002, KBN010 and KBN111) [41]. *Krt14Cre*-driven NUT carcinoma mouse models were generated as described by crossing the *Krt14Cre* mouse line with the NUT Carcinoma Translocator mouse line (MMRRC 071753-MU) [41]. Serum was sampled from end-stage tumor-bearing mice and healthy controls following the retro-orbital blood collection procedure as previously described [42].

2.2.3 Formation of nanoparticle protein corona

For human plasma nanoparticle protein corona, 1.25 mg amine-terminated magnetic NPs, carboxylate-terminated magnetic NPs, SP3 hydrophilic magnetic NPs, SP3 hydrophobic magnetic NPs or polystyrene NPs were individually washed with 200 μL water twice, and then incubated with 1 mL 55% human plasma at 37 °C for 1 h with constant stirring at 350 rpm to form nanoparticle protein corona. A magnet rack was used to separate the solution for four magnetic NPs, while centrifugation at 14,000 g for 20 min was applied to remove the solution for polystyrene NPs. Next, the nanoparticle protein coronas were washed with 500 μL ice-cold DPBS twice, followed by 500 μL ice-cold water twice. The resulting nanoparticle protein coronas in the solid phase were collected for further BUP sample preparation. For SDS-PAGE evaluation of the performances by different nanoparticle protein coronas, the proteins from nanoparticle protein coronas were eluted with 100 μL 0.2% SDS in water. The protein concentrations in the eluates were measured by BCA assay. 15 μg eluted protein was loaded into each lane of SDS-PAGE gel for analysis.

For mouse serum nanoparticle protein corona, 200 μg amine-terminated magnetic NPs, and carboxylate-terminated magnetic NPs were individually washed with 200 μL water twice and then incubated with 40 μL 55% mouse serum at 37 °C for 1 h with constant stirring at 350 rpm to

form nanoparticle protein corona. A magnet rack was used to separate the solution for four magnetic NPs, while centrifugation at 14,000 g for 20 min was applied to remove the solution for polystyrene NPs. Next, the nanoparticle protein coronas were washed with 20 μL ice-cold DPBS twice, followed by 20 μL ice-cold water twice. The resulting nanoparticle protein coronas in the solid phase were collected for further BUP sample preparation.

2.2.4 Sample preparation for BUP

The resulting nanoparticle protein coronas in the "Formation of nanoparticle protein corona" section (~15 μg of total proteins) was dispersed in 15 μL of 100 mM ABC buffer (pH 8.0) containing 5 mM DTT and led the protein corona to be fully denatured and reduced at 90 °C for 15 min. Then, the protein corona was cooled down to room temperature, followed by trypsin (3 μg) digestion at 37 °C for 1 h. The digestion was finally terminated by adding formic acid (0.6 % (v/v) final concentration), and the supernatants were collected in the LoBind tubes. To fully elute the peptides, 10 μL 20% ACN in 100 mM ABC was incubated with the nanoparticle at 37 °C for 10 min and combined with the supernatant from the previous portion for downstream CZE-MS/MS analysis.

2.2.5 CZE-MS/MS analysis

Linear polyacrylamide (LPA)-coated fused silica capillaries (50 μm i.d., 360 μm o.d.) were prepared according to our previous studies [43, 44]. The CZE-MS/MS system configuration involved the integration of a CESI 8000 Plus CE system (Beckman Coulter) with an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific), employing an in-house-built electrokinetically pumped sheath-flow CE-MS nanospray interface [45, 46]. The interface featured a glass spray emitter pulled using a Sutter P-1000 flaming/brown micropipette puller to achieve an orifice size of 30−35 μm, filled with sheath buffer composed of 0.2% (v/v) formic acid and 10% (v/v) methanol. The spray voltage was about 2 kV. The length of the LPA-coated CZE capillary was 80 cm. The capillary inlet was securely affixed within the cartridge of the CE system, while its outlet was inserted into the emitter of the interface. The capillary outlet to emitter orifice distance was maintained at approximately 0.5 mm. 50 nL (~30 ng) of each corona peptide sample was loaded for CZE-MS/MS. Following this, the capillary inlet was filled with the background electrolyte (BGE, 5% (v/v) acetic acid), initiating the CZE separation process under a separation voltage of 30 kV, and the separation time was 50 min. After the separation, 30 kV voltage and 15 psi pressure were applied for 10 min to clean up the capillary.

For the mass spectrometer, all experiments were conducted using an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific) in Data-dependent acquisition (DDA) mode. Ion transfer tube temperature was set at 320 °C and peptide mode was enabled. Pressure mode was set as standard. Full MS scans were acquired in the Orbitrap mass analyzer over the *m/z* 300-1500 range with a resolution of 60,000 (at 200 *m/z*). Normalized AGC targets for MS and MS/MS were set at 300% and 150%, respectively. Only precursor ions with an intensity exceeding 1E5 and a charge state between 2 and 7 were fragmented in the higher-energy collisional dissociation (HCD) cell and analyzed by the Orbitrap mass analyzer with a resolution of 7,500 (at 200 *m/z*). The numbers of dependent scans for human plasma and mouse serum were at 15 and 25, respectively. Monoisotopic peak determination was set to peptide and the option "Relax restrictions when too few precursors are found" was checked. One microscan was used for both MS and MS/MS. The normalized collision energy was set at 30%. The maximum ion injection times for MS and MS/MS spectra acquisition were both set as auto. The precursor isolation width was 1.4 *m/z*. The first mass for MS/MS was set at *m/z* 100. The dynamic exclusion was applied with a duration of 15 s, and the exclusion of isotopes was enabled.

2.2.6 Database search

For human plasma nanoparticle protein corona, database searching of the raw files was performed in Proteome Discoverer 2.2 with SEQUEST HT search engine against the UniProt proteome database of human (UP000005640, 82697 entries, version 12/2023). Database searching of the reversed database was also performed to evaluate the false discovery rate (FDR). Search parameters included full tryptic digestion, allowing up to 2 missed cleavages, with precursor mass tolerance set to 20 ppm and fragment mass tolerance to 0.05 Da. Acetyl (protein N-term) and phospho (S, T, Y) were set as variable modifications. Data was filtered with a peptide-level FDR of 1%, and protein grouping was applied.

For mouse serum nanoparticle protein corona, raw data files were analyzed using MaxQuant software (Version 2.1.2.0) with the Andromeda search engine against the UniProt proteome database of mouse (UP000000589, 54707 entries, version 02/2024) [47]. The peptide mass tolerances for the initial and main searches were set to 20 ppm and 4.5 ppm, respectively, with a fragment ion mass tolerance of 20 ppm. Trypsin was specified as the protease, and dynamic modifications included oxidation on methionine and acetylation at the protein N-terminus. LFQ option was checked and the LFQ minimum ratio count was at 2. Intensity-based

absolute quantification (iBAQ) was checked to report the measure of protein abundance. The minimum peptide length was set to 7 amino acids, and FDRs were controlled at 1% for both peptides and proteins.

2.2.7 Statistical analysis

For LFQ of amine-terminated nanoparticle and carboxylate-terminated nanoparticle treated mouse serum (healthy and nut cancer), each sample contains technical duplicate CZE-MS/MS runs. 12 MS raw files for each nanoparticle were applied for statistical analysis, and iBAQ values were used as the absolute abundance. The quantitative results were further analyzed using Perseus software [48]. The intensities of each protein were log2 transformed, and the significantly differentially expressed proteins were determined by performing t-test analysis using the Perseus software to generate '-log (P-value)' and 'log 2 (fold change of KBN to BN)'. P-value at 0.05 and fold change at 2 were used for making volcano plots by DataGraph software (Version 5.3).

2.3 Results and discussion

2.3.1 A high-throughput BUP workflow of CZE-MS/MS using human plasma

To develop an effective BUP workflow for high-throughput analysis of human plasma, we tested four types of magnetic NPs including amine-terminated NPs, carboxylate-terminated NPs, SP3 hydrophilic NPs, and SP3 hydrophobic NPs. We used a commercialized healthy human plasma and CZE-MS/MS for peptide separation, detection, and identification. **Figure 2.1** describes the detailed BUP workflow, which takes approximately 3.5 hours from sample preparation to generate MS raw files. The full procedures are outlined in the experimental section. Briefly, 55% human plasma was first incubated with magnetic NPs for 1 hour to form a nanoparticle protein corona, followed by washing steps to remove unbound proteins. The nanoparticle protein corona underwent rapid on-bead tryptic digestion for 1 hour, with hydrophilic peptides remaining in the aqueous phase. To ensure complete peptide elution, the NPs were incubated with an elution buffer containing 20% ACN. Finally, the second eluted peptides were combined with the initial aqueous phase and directly subjected to a 1-hour CZE-MS/MS analysis without any additional processing.

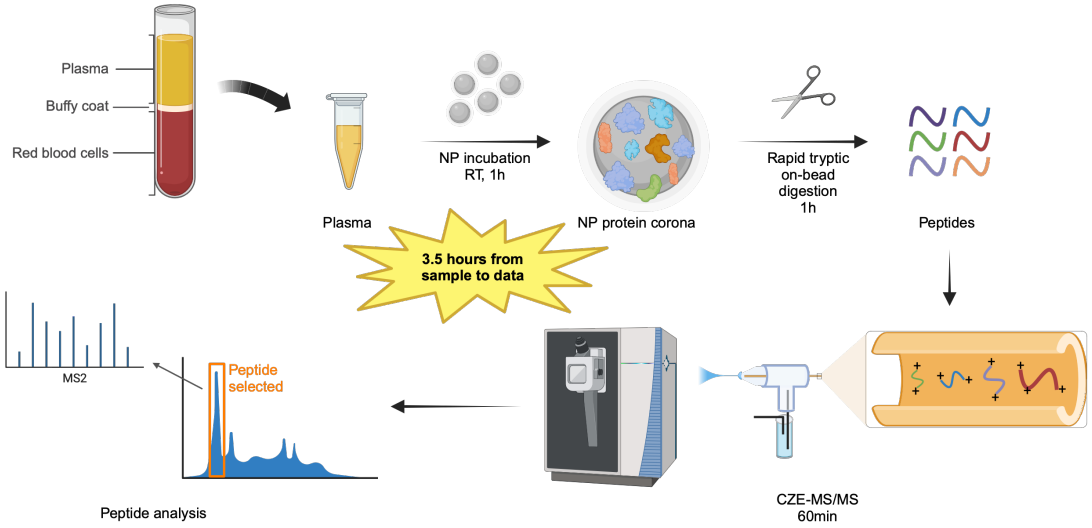


Figure 2.1. Schematic of the MS-based BUP workflow for magnetic nanoparticle protein corona using amine-terminated NPs, carboxylate-terminated NPs, SP3 hydrophilic NPs, SP3 hydrophobic NPs, a human plasma sample, and CZE-MS/MS.

2.3.2 Validation of the high-throughput BUP workflow using human plasma

Prior to CZE-MS/MS, we employed a preliminary SDS-PAGE analysis (**Figure 2.2A**) to compare the profiles of the untreated human plasma, four different magnetic nanoparticle protein coronas, and a non-magnetic polystyrene NPs, which has demonstrated the robustness for forming protein corona of human plasma [33]. Among four magnetic NPs, SP3 hydrophilic NPs and SP3 hydrophobic NPs are well developed for rapid, robust, and efficient protein sample processing for BUP [49]. Instead of the tryptic cleavage, the intact proteins were eluted from the nanoparticle surface by an elution buffer containing SDS. The SDS-PAGE result showed that the major protein bands (37 kDa - 150 kDa) from carboxylate-terminated NPs are relatively consistent with those from SP3 hydrophilic NPs and SP3 hydrophobic NPs. It must be noted that both SP3 NPs contain a carboxyl group covalently bound on the nanoparticle surface, leading to a similar surface chemistry to carboxylate-terminated NPs of forming protein corona. However, the protein profiles from carboxyl group-based NPs, amine-terminated NPs, and polystyrene NPs are all significantly different from each other, representing the distinct pools of protein coronas by different NPs. Furthermore, all NPs performed obvious drops at the intense bands compared to the untreated human plasma, such as near-150 kDa (possibly IgGs) and near-50 kDa (possibly α-1-antitrypsin, haptoglobin or plasminogen) bands, showing the effective depletion of high-abundance proteins

in human plasma. Next, CZE-MS/MS was applied to deeply explore the protein corona profiles by different NPs. **Figure 2.2B** depicts the protein and peptide identification numbers (IDs) from untreated human plasma and four magnetic nanoparticle protein coronas. Amine-terminated NPs achieved slightly higher protein IDs (191 IDs) than untreated human plasma (180 IDs) with fewer peptide IDs (2000 IDs vs 2172 IDs), while all carboxyl group-based NPs showed lower protein IDs (153, 104, and 140 IDs for carboxylate-terminated NPs, SP3 hydrophilic NPs and SP3 hydrophobic NPs, respectively) and peptide IDs (1134, 703 and 1117 IDs for carboxylate-terminated NPs, SP3 hydrophilic NPs and SP3 hydrophobic NPs, respectively). This revealed individual NPs didn't perform a significant increase in the protein or peptide IDs compared to untreated human plasma. However, by combining four different NPs, we could attain much higher protein IDs (253 IDs) with comparable peptide IDs (2272 IDs) to untreated human plasma, showing the potential of coupling distinct NPs for protein corona to reach higher protein IDs.

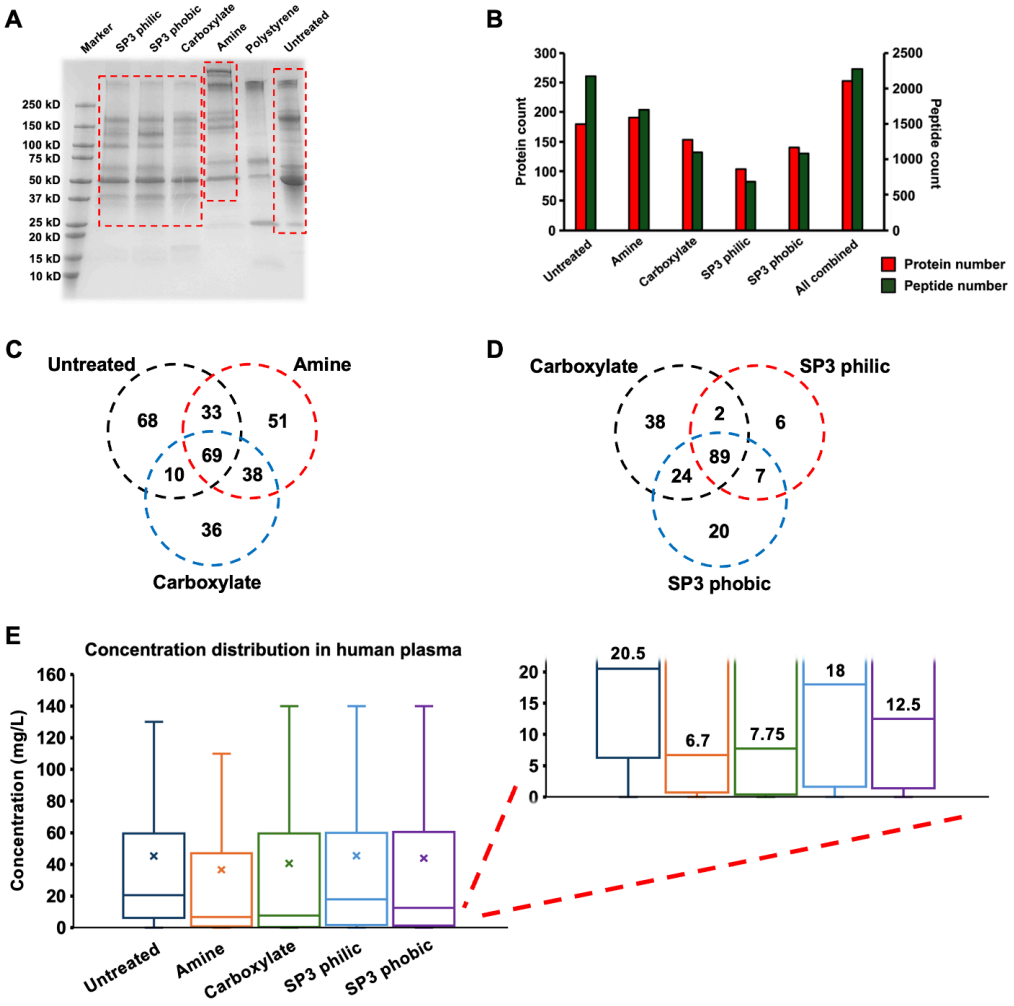


Figure 2.2. Different protein profiles revealed by amine-terminated NPs (Amine), carboxylate-terminated NPs (Carboxylate), SP3 hydrophilic NPs (SP3 philic), SP3 hydrophobic NPs (SP3 phobic), polystyrene NPs (Polystyrene) and untreated human plasma. (A). SDS-PAGE data of nanoparticle protein coronas and untreated human plasma. (B). The protein and peptide identification counts by different nanoparticle protein coronas and untreated human plasma. (C). Protein overlaps among amine-terminated NPs, carboxylate-terminated NPs, and untreated human plasma. (D). Protein overlaps among carboxylate-terminated NPs, SP3 hydrophilic NPs and SP3 hydrophobic NPs. (E). Protein concentration distribution by different nanoparticle protein coronas and untreated human plasma, and all the numbers are at the unit of mg/L. The protein abundance information was obtained from the Human Protein Atlas database.

Interestingly, the untreated human plasma, amine-terminated NPs, and carboxylate-terminated NPs produced different protein profiles, evidenced by the low protein identification overlap among them (**Figure 2.2C**). This can be explained by the distinct functional groups on the nanoparticle surface to selectively capture different proteins. On the contrary, a high protein identification overlap was found within three carboxyl group-based NPs (**Figure 2.2D**), which is

consistent with the similar protein band distributions in the previous SDS-PAGE result. We then speculated that the proteins identified in the untreated human plasma contained many high-abundance proteins such as albumin and immunoglobulin, which are also identified from each magnetic NPs as the overlap. However, for the proteins identified from each nanoparticle protein corona, there are still a good amount of non-overlapped proteins identified and we assumed the nanoparticle protein corona could significantly decrease the concentration dynamic range of human plasma. To prove this hypothesis, we checked the identified proteins' concentrations by untreated human plasma and four nanoparticle protein coronas from the blood protein section in the Human Protein Atlas database [50]. As shown in **Figure 2.2E**, the protein concentration distribution data demonstrated the benefits of nanoparticle protein corona for reducing the concentration dynamic range to detect more low-abundance proteins in human plasma, verified by the lower median value from different NPs (6.7, 7.75, 18, and 12.5 mg/L) compared to untreated human plasma (20.5 mg/L). Considering the low protein overlaps from NPs with different functional groups and the capability of detecting more low-abundance proteins, we decided to apply only amine-terminated NPs and carboxylate-terminated NPs as the representative NPs for further experiments.

2.3.3 Application of the high-throughput BUP workflow CZE-MS/MS-based BUP using a pair of mouse serums (healthy and NUT cancer)

NUT carcinoma (NC) is an aggressive cancer characterized by chromosomal rearrangements, typically involving the fusion of the NUTM1 gene with genes like BRD4, leading to uncontrolled cellular growth and blocked differentiation [51]. This rare carcinoma occurs primarily in midline structures such as the head, neck, and mediastinum, affecting both children and adults, with a poor prognosis despite intense treatment [52]. Research is actively exploring the origins and epigenetic mechanisms of NC, aiming to develop more effective therapeutic strategies and early-stage detection. Mouse serum proteomics can be used to study disease models to identify biomarkers, and a genetically engineered NC mouse model was developed to replicate NC oncogenesis in a controlled experimental setting with the precise regulation of key parameters [41]. Therefore, we decided to apply the serum samples from this NC mouse model to further validate the performance of our high-throughput BUP CZE-MS/MS workflow.

Three biological triplicates of mouse serum from the healthy mouse model and NC mouse model were used to analyze the proteins recovered by amine-terminated and carboxylate-terminated nanoparticle protein coronas. **Figure 2.3A** and **2.3B** displayed that amine-terminated nanoparticle protein corona outperformed carboxylate-terminated nanoparticles in terms of protein (average 426 IDs vs 274 IDs) and peptide IDs (average 3081 IDs vs 2853 IDs), agreeing well with the findings from previous human plasma experiments. Besides, amine-terminated nanoparticle protein corona showed consistent IDs at the protein level (4 % relative standard deviation), while carboxylate-terminated nanoparticles reached 9 % relative standard deviation for protein IDs, both presenting the credible stability of protein corona formation by each NPs. To further investigate the protein identified in each nanoparticle protein corona, we studied the protein overlaps between the biological triplicates of each mouse model by each NPs (**Figure 2.3C and 2.3D**). Good reproducibility was found for both NPs between each mouse serum sample, ranging from 0.75 to 0.89 for amine-terminated NPs and from 0.65 to 0.89 for carboxylate-terminated NPs. The data suggests that two NPs can both perform nanoparticle protein corona efficiently with good reproducibility, allowing for the confidently quantitative analysis of proteins.

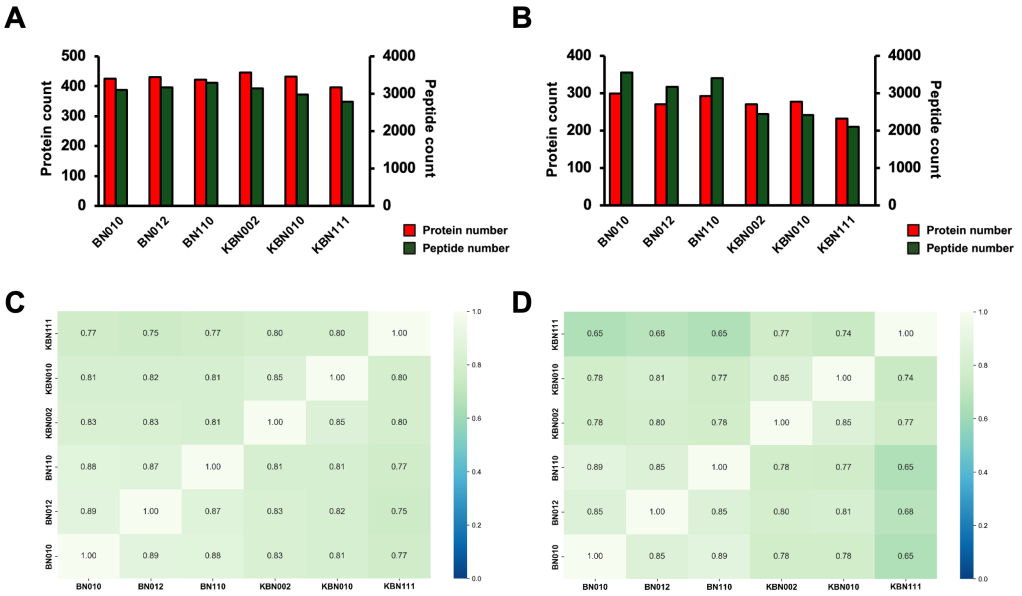


Figure 2.3. Protein and peptide identification counts and protein overlaps by amine-terminated NPs (A and C), carboxylate-terminated NPs (B and D) using serum from healthy mouse model (BN010, BN012 and BN110) and NC mouse model (KBN002, KBN010 and KBN111). (A). The protein and peptide identification counts by amine-terminated NPs. (B). The protein and peptide identification counts by carboxylate-terminated NPs. (C). Protein overlaps by amine-terminated NPs. (D). Protein overlaps by carboxylate-terminated NPs.

Next, LFQ was applied for both amine-terminated nanoparticle-treated and carboxylate-terminated nanoparticle-treated mouse serum. The volcano plot in **Figure 2.4A** and **2.4B** showed the up-regulated (red) and down-regulated (blue) proteins corresponding to genes by an abundance ratio cutoff at 2 (KBN/BN) and P-value at 0.05 in the amine-terminated nanoparticle-treated and carboxylate-terminated nanoparticle-treated NC mouse serum model, respectively. Overall, 116 proteins of 489 protein IDs and 111 proteins from 332 protein IDs had statistical differences in abundance between the healthy and NC mouse serum by amine-terminated NPs and carboxylate-terminated NPs, respectively. Interestingly, more up-regulated proteins were found in NC mouse serum by amine-terminated NPs, and more down-regulated proteins were found in NC mouse serum by carboxylate-terminated NPs. To be specific, 67 proteins were up-regulated and 49 proteins down-regulated in NC mouse serum by amine-terminated NPs, while 31 proteins were up-regulated and 80 proteins down-regulated in NC mouse serum by carboxylate-terminated NPs. This further showed the advantage of coupling different types of nanoparticle protein corona to achieve more protein IDs with significant abundance differences. Furthermore, we found 15 genes (up-regulated : down-regulated = 11 : 4 in NC mouse serum) from amine-terminated NPs and 8

genes (up-regulated : down-regulated = 1 : 7 in NC mouse serum) from carboxylate-terminated NPs reported to be candidate biomarkers of negative survival in the tumors by comparing the proteomic data to the transcriptomic data. In total, 19 genes (shown in **Table 2.1**) were revealed to be differentially expressed by combining two NPs' data, suggesting that different NPs could be used as complementary tools to identify more biomarkers from serum samples. Among the 19 genes, Spp1 is a confident gene associated with aggressive cancers, produced in various organs and found in body fluids such as serum and urine. SPP1 is expressed in specific cell types, including osteoblasts, macrophages, and immune cells, and is also present in cancer cells, with elevated levels of SPP1 correlating with poor prognosis in several cancers [53-55]. SPP1 promotes cancer cell growth and resistance to chemoradiotherapy through the induction of epithelial-mesenchymal transition, autophagy, and metabolic alterations, primarily via activation of the PI3K/Akt and MAPK pathways [56].

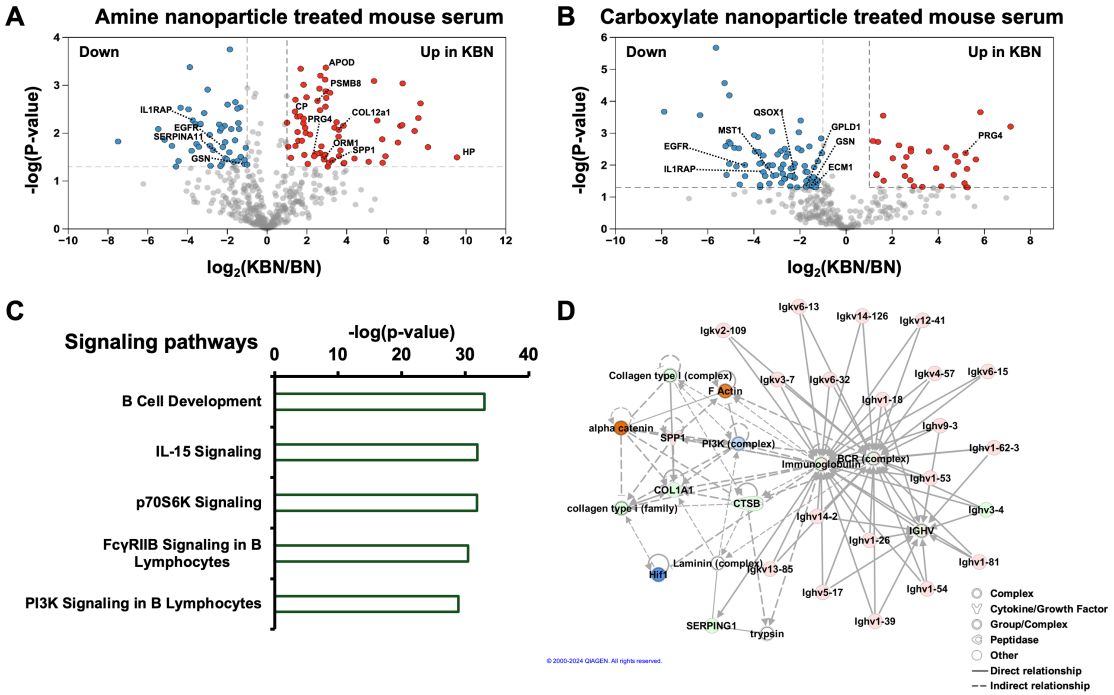


Figure 2.4. LFQ of healthy mouse serum model (BN) and NC mouse serum model (KBN). Volcano plot of LFQ of amine-terminated nanoparticle-treated mouse serum (A) and carboxylate-terminated nanoparticle-treated mouse serum (B). Up-regulated biomarkers in the NC mouse serum model are labeled in red, while down-regulated ones are marked in blue. Examples of some genes related to candidate biomarkers of negative survival in the tumors are marked. (C). An ingenuity pathway analysis reported some cancer, organismal injury, and abnormalities diseases that are related to the differentially expressed genes in the two mouse serum models by amine-terminated NPs. (D). Proteins with significant abundance differences within two mouse serum models correspond to genes that are involved in cancer-related networks with high scores. Those genes are highlighted in red (increased), green (decreased), orange (predicted activation), and blue (predicted inhibition). Copyright permission has been granted by QIAGEN for using the network data.

Table 2.1. Summary of cancer-related protein biomarkers identified by nanoparticle protein corona.

Gene	Protein name	Amine-terminated nanoparticle protein corona	Carboxylate-terminated nanoparticle protein corona	Up or down regulated in NC mouse model serum
Spp1	Secreted phosphoprotein 1	x		Up
Col12a1	Collagen, type XII, alpha 1	x		Up
Hp	Haptoglobin	x		Up
Apod	Apolipoprotein D	x		Up
Orm1	Alpha-1-acid glycoprotein 1	x		Up
Psmb8	Proteasome subunit beta type-8	x		Up
Prg4	Proteoglycan 4	x	x	Up
Lrg1	Leucine-rich HEV glycoprotein	x		Up
Cp	ferroxidase	x		Up
Ctla2a	Protein CTLA-2-alpha	x		Up
Lbp	Lipopolysaccharide-binding protein	x		Up
Gsn	Gelsolin	x	x	Down
Serpina11	Serpin A11	x		Down
Egfr	Epidermal growth factor receptor	x	x	Down
Il1rap	Interleukin-1 receptor accessory protein	x	x	Down
Ecm1	Extracellular matrix protein 1		x	Down
Gpld1	Phosphatidylinositol-glycan-specific phospholipase D		x	Down
Qsox1	Sulfhydryl oxidase 1		x	Down
Mst1	Macrophage stimulating 1 (hepatocyte growth factor-like)		x	Down

We then performed an ingenuity pathway analysis (IPA) of the genes with differential abundance between healthy and NC mouse serum models using amine-terminated NPs as an example. The genes are involved in cancer-related pathways such as immune-related pathways (B cell development, IL-15 signaling, and FcγRIIB signaling), cell growth pathways (p70S6K

signaling), and proliferation and migration pathways (PI3K signaling), as shown in **Figure 2.4C**. IPA network analysis revealed that 20 proteins (highlighted in red) showed higher abundance and 4 proteins (highlighted in green) showed lower abundance in the NC mouse serum model to healthy mouse serum model involving cancer, organismal injury, and abnormality-related network (score, 57) (**Figure 2.4D**). Those proteins belong to several groups such as growth factor (e.g., SPP1), peptidase (e.g., CTSB), complex (e.g., Collagen type i or I) and others, and the proteins all have direct (solid line) and indirect (dotted line) interactions with one another. The abundance changes of proteins in the serum could potentially reflect the activities of proteins in the cell, e.g., PI3K complex. PI3K (phosphatidylinositol 3-kinase, located in the cytoplasm) intracellular pathway is critical in regulating cell growth, survival, and metabolism, and the PI3K mutations activate the downstream AKT/mTOR signaling cascade, which promotes tumorigenesis by enabling cells to evade apoptosis, enhance proliferation, and develop resistance to conventional therapies [57]. In the IPA network, we noticed that up-regulated proteins like SPP1 and down-regulated proteins like collagen type I, collagen type i, and COL1A1 in the NC mouse cancer model have indirect interaction (dotted line) with PI3K, which is predicted to indirectly activate proteins located in cytoplasms like alpha-catenin and F Actin. We also noted that PI3K also has indirect interaction (dotted line) with immunoglobulin BCR complex which directly interacted with many up-regulated immunoglobulin variables found from our experimental result. All the differentially expressed proteins associated with PI3K and PI3K-affected proteins could be used as potential biomarkers for cancer, organismal injury, and abnormality.

2.4 Conclusion

In this study, we have developed a novel high-throughput nanoparticle protein corona workflow coupled with CZE-MS/MS for proteomic analysis of plasma and serum samples, achieving 3.5 hours from sample to data. Our results demonstrated the effectiveness of this approach in reducing sample complexity and increasing the identification of low-abundance proteins, addressing the inherent challenges in plasma proteomics. By utilizing distinct functionalized nanoparticles, we were able to selectively enrich different subsets of proteins, further enhancing proteomic coverage. Moreover, the application of rapid on-bead digestion significantly reduced the overall experimental time without compromising data quality. Our workflow's application to NC mouse models highlighted its utility in discovering potential cancer biomarkers. Comparative analysis of serum samples between healthy and NC mouse models

revealed significant differences in protein expression, with potential implications for understanding cancer progression and therapeutic response. However, some limitations need to be addressed for future work. Firstly, the current workflow could identify ~200 protein groups from human plasma, requiring some strategies to increase the proteome coverage, such as applying other types of NPs with distinct functional groups. Secondly, the current capillary length used for CZE-MS/MS is 80 cm, and the separation could be faster with shorter capillaries to improve the throughput. Thirdly, the scanning rate of orbitrap-based mass spectrometers is limited, causing a relatively lower number of protein IDs from fewer MS/MS spectra compared to other fast mass analyzer-based mass spectrometers like time-of-flight. Overall, this high-throughput workflow offers a robust platform for large-scale proteomics studies and could be further applied to various clinical and research settings, including biomarker discovery and disease monitoring.

2.5 Acknowledgment

This work was funded in part by the National Institute of General Medical Sciences through the grant R35GM153479 and the National Cancer Institute (NCI) through the grant R01CA247863.

REFERENCES

[1] Riehemann K, Schneider SW, Luger TA, Godin B, Ferrari M, Fuchs H. Nanomedicine--challenge and perspectives. Angew Chem Int Ed Engl. 2009;48(5):872-97.
[2] Bhatia SN, Chen X, Dobrovolskaia MA, Lammers T. Cancer nanomedicine. Nat Rev Cancer. 2022 Oct;22(10):550-556.

[3] Hajipour MJ, Fromm KM, Ashkarran AA, Jimenez de Aberasturi D, de Larramendi IR, Rojo T, Serpooshan V, Parak WJ, Mahmoudi M. Antibacterial properties of NPs. Trends Biotechnol. 2012 Oct;30(10):499-511.

[4] Patra JK, Das G, Fraceto LF, Campos EVR, Rodriguez-Torres MDP, Acosta-Torres LS, Diaz-Torres LA, Grillo R, Swamy MK, Sharma S, Habtemariam S, Shin HS. Nano based drug delivery systems: recent developments and future prospects. J Nanobiotechnology. 2018 Sep 19;16(1):71.

[5] Mitchell MJ, Billingsley MM, Haley RM, Wechsler ME, Peppas NA, Langer R. Engineering precision NPs for drug delivery. Nat Rev Drug Discov. 2021 Feb;20(2):101-124.

[6] Attia MF, Anton N, Wallyn J, Omran Z, Vandamme TF. An overview of active and passive targeting strategies to improve the nanocarriers efficiency to tumour sites. J Pharm Pharmacol. 2019 Aug;71(8):1185-1198.

[7] Foroozandeh P, Aziz AA. Merging worlds of nanomaterials and biological environment: factors governing protein corona formation on NPs and its biological consequences. Nanoscale Res Lett. 2015 May 16;10:221.

[8] Gunawan C , Lim M , Marquis CP , Amal R . Nanoparticle-protein corona complexes govern the biological fates and functions of NPs. J Mater Chem B. 2014 Apr 21;2(15):2060-2083.

[9] Karmali PP, Simberg D. Interactions of NPs with plasma proteins: implication on clearance and toxicity of drug delivery systems. Expert Opin Drug Deliv. 2011 Mar;8(3):343-57.

[10] Kumar A, Bicer EM, Morgan AB, Pfeffer PE, Monopoli M, Dawson KA, Eriksson J, Edwards K, Lynham S, Arno M, Behndig AF, Blomberg A, Somers G, Hassall D, Dailey LA, Forbes B, Mudway IS. Enrichment of immunoregulatory proteins in the biomolecular corona of NPs within human respiratory tract lining fluid. Nanomedicine. 2016 May;12(4):1033-1043.

[11] Cedervall T, Lynch I, Lindman S, Berggård T, Thulin E, Nilsson H, Dawson KA, Linse S. Understanding the nanoparticle-protein corona using methods to quantify exchange rates and affinities of proteins for NPs. Proc Natl Acad Sci U S A. 2007 Feb 13;104(7):2050-5.

[12] Tenzer S, Docter D, Rosfa S, Wlodarski A, Kuharev J, Rekik A, Knauer SK, Bantz C, Nawroth T, Bier C, Sirirattanapan J, Mann W, Treuel L, Zellner R, Maskos M, Schild H, Stauber RH. Nanoparticle size is a critical physicochemical determinant of the human blood plasma corona: a comprehensive quantitative proteomic analysis. ACS Nano. 2011 Sep 27;5(9):7155-67.

[13] Monopoli MP, Aberg C, Salvati A, Dawson KA. Biomolecular coronas provide the biological identity of nanosized materials. Nat Nanotechnol. 2012 Dec;7(12):779-86.

[14] Walczyk D, Bombelli FB, Monopoli MP, Lynch I, Dawson KA. What the cell "sees" in bionanoscience. J Am Chem Soc. 2010 Apr 28;132(16):5761-8.

[15] Casals E, Pfaller T, Duschl A, Oostingh GJ, Puntes V. Time evolution of the nanoparticle protein corona. ACS Nano. 2010 Jul 27;4(7):3623-32.

[16] Lundqvist M, Stigler J, Elia G, Lynch I, Cedervall T, Dawson KA. Nanoparticle size and surface properties determine the protein corona with possible implications for biological impacts. Proc Natl Acad Sci U S A. 2008 Sep 23;105(38):14265-70.

[17] Monopoli MP, Walczyk D, Campbell A, Elia G, Lynch I, Bombelli FB, Dawson KA. Physical-chemical aspects of protein corona: relevance to in vitro and in vivo biological impacts of NPs. J Am Chem Soc. 2011 Mar 2;133(8):2525-34.

[18] Milani S, Bombelli FB, Pitek AS, Dawson KA, Rädler J. Reversible versus irreversible binding of transferrin to polystyrene NPs: soft and hard corona. ACS Nano. 2012 Mar 27;6(3):2532-41.

[19] Geyer PE, Holdt LM, Teupser D, Mann M. Revisiting biomarker discovery by plasma proteomics. Mol Syst Biol. 2017 Sep 26;13(9):942.

[20] Zhang Q, Faca V, Hanash S. Mining the plasma proteome for disease applications across seven logs of protein abundance. J Proteome Res. 2011 Jan 7;10(1):46-50.

[21] Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics. 2002 Nov;1(11):845-67.

[22] Pernemalm M, Sandberg A, Zhu Y, Boekel J, Tamburro D, Schwenk JM, Björk A, Wahren-Herlenius M, Åmark H, Östenson CG, Westgren M, Lehtiö J. In-depth human plasma proteome analysis captures tissue proteins and transfer of protein variants across the placenta. Elife. 2019 Apr 8;8:e41608.

[23]Trinh DN, Gardner RA, Franciosi AN, McCarthy C, Keane MP, Soliman MG, O'Donnell JS, Meleady P, Spencer DIR, Monopoli MP. Nanoparticle Biomolecular Corona-Based Enrichment of Plasma Glycoproteins for N-Glycan Profiling and Application in Biomarker Discovery. ACS Nano. 2022 Apr 26;16(4):5463-5475.

[24] Blume JE, Manning WC, Troiano G, Hornburg D, Figa M, Hesterberg L, Platt TL, Zhao X, Cuaresma RA, Everley PA, Ko M, Liou H, Mahoney M, Ferdosi S, Elgierari EM, Stolarczyk C, Tangeysh B, Xia H, Benz R, Siddiqui A, Carr SA, Ma P, Langer R, Farias V, Farokhzad OC. Rapid, deep and precise profiling of the plasma proteome with multi-nanoparticle protein corona. Nat Commun. 2020 Jul 22;11(1):3662.

[25] Corbo C, Li AA, Poustchi H, Lee GY, Stacks S, Molinaro R, Ma P, Platt T, Behzadi S, Langer R, Farias V, Farokhzad OC. Analysis of the Human Plasma Proteome Using Multi-Nanoparticle Protein Corona for Detection of Alzheimer's Disease. Adv Healthc Mater. 2021 Jan;10(2):e2000948.

[26] Mahmoudi M, Landry MP, Moore A, Coreas R. The protein corona from nanomedicine to environmental science. Nat Rev Mater. 2023 Mar 24:1-17.

[27] Schubert OT, Röst HL, Collins BC, Rosenberger G, Aebersold R. Quantitative proteomics: challenges and opportunities in basic and applied research. Nat Protoc. 2017 Jul;12(7):1289-1294.

[28] Megger DA, Bracht T, Meyer HE, Sitek B. Label-free quantification in clinical proteomics. Biochim Biophys Acta. 2013 Aug;1834(8):1581-90.

[29] Thompson A, Wölmer N, Koncarevic S, Selzer S, Böhm G, Legner H, Schmid P, Kienle S, Penning P, Höhle C, Berfelde A, Martinez-Pinna R, Farztdinov V, Jung S, Kuhn K, Pike I.

TMTpro: Design, Synthesis, and Initial Evaluation of a Proline-Based Isobaric 16-Plex Tandem Mass Tag Reagent Set. Anal Chem. 2019 Dec 17;91(24):15941-15950.

[30] Nusinow DP, Szpyt J, Ghandi M, Rose CM, McDonald ER 3rd, Kalocsay M, Jané-Valbuena J, Gelfand E, Schweppe DK, Jedrychowski M, Golji J, Porter DA, Rejtar T, Wang YK, Kryukov GV, Stegmeier F, Erickson BK, Garraway LA, Sellers WR, Gygi SP. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. Cell. 2020 Jan 23;180(2):387-402.e16.

[31] Trinh HV, Grossmann J, Gehrig P, Roschitzki B, Schlapbach R, Greber UF, Hemmi S. iTRAQ-Based and Label-Free Proteomics Approaches for Studies of Human Adenovirus Infections. Int J Proteomics. 2013;2013:581862.

[32] Kapoor I, Pal P, Lochab S, Kanaujiya JK, Trivedi AK. Proteomics approaches for myeloid leukemia drug discovery. Expert Opin Drug Discov. 2012 Dec;7(12):1165-75.

[33] Ashkarran AA, Gharibi H, Voke E, Landry MP, Saei AA, Mahmoudi M. Measurements of heterogeneity in proteomics analysis of the nanoparticle protein corona across core facilities. Nat Commun. 2022 Nov 3;13(1):6610.

[34] Shen X, Sun L. Systematic Evaluation of Immobilized Trypsin-Based Fast Protein Digestion for Deep and High-Throughput Bottom-Up Proteomics. Proteomics. 2018 May;18(9):e1700432.

[35] Lombard-Banek C, Moody SA, Manzini MC, Nemes P. Microsampling Capillary Electrophoresis Mass Spectrometry Enables Single-Cell Proteomics in Complex Tissues: Developing Cell Clones in Live Xenopus laevis and Zebrafish Embryos. Anal Chem. 2019 Apr 2;91(7):4797-4805.

[36] Yang Z, Shen X, Chen D, Sun L. Improved Nanoflow RPLC-CZE-MS/MS System with High Peak Capacity and Sensitivity for Nanogram Bottom-up Proteomics. J Proteome Res. 2019 Nov 1;18(11):4046-4054.

[37] Frantzi M, Gomez Gomez E, Blanca Pedregosa A, Valero Rosa J, Latosinska A, Culig Z, Merseburger AS, Luque RM, Requena Tapia MJ, Mischak H, Carrasco Valiente J. CE-MS-based urinary biomarkers to distinguish non-significant from significant prostate cancer. Br J Cancer. 2019 Jun;120(12):1120-1128.

[38] Pelander L, Brunchault V, Buffin-Meyer B, Klein J, Breuil B, Zürbig P, Magalhães P, Mullen W, Elliott J, Syme H, Schanstra JP, Häggström J, Ljungvall I. Urinary peptidome analyses for the diagnosis of chronic kidney disease in dogs. Vet J. 2019 Jul;249:73-79.

[39] Yan X, Sun L, Dovichi NJ, Champion MM. Minimal deuterium isotope effects in quantitation of dimethyl-labeled complex proteomes analyzed with capillary zone electrophoresis/mass spectrometry. Electrophoresis. 2020 Aug;41(15):1374-1378.

[40] Faserl K, Chetwynd AJ, Lynch I, Thorn JA, Lindner HH. Corona Isolation Method Matters: Capillary Electrophoresis Mass Spectrometry Based Comparison of Protein Corona Compositions Following On-Particle versus In-Solution or In-Gel Digestion. Nanomaterials (Basel). 2019 Jun 20;9(6):898.

[41] Zheng D, Elnegiry AA, Luo C, Bendahou MA, Xie L, Bell D, Takahashi Y, Hanna E, Mias GI, Tsoi MF, Gu B. Brd4::Nutm1 fusion gene initiates NUT carcinoma in vivo. Life Sci Alliance. 2024 May 9;7(7):e202402602.

[42] Greenfield EA. Sampling and Preparation of Mouse and Rat Serum. Cold Spring Harb Protoc. 2017 Nov 1;2017(11):pdb.prot100271.

[43] 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 Jan 1;146:839-43.

[44] 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 Jun 21;142(12):2118-2127.

[45] Wojcik R, Dada OO, Sadilek M, Dovichi NJ. Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. Rapid Commun Mass Spectrom. 2010 Sep 15;24(17):2554-60.

[46] Sun L, Zhu G, Zhang Z, Mou S, Dovichi NJ. 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 May 1;14(5):2312-21.

[47] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008 Dec;26(12):1367-72.

[48] Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016 Sep;13(9):731-40.

[49] Hughes CS, Moggridge S, Müller T, Sorensen PH, Morin GB, Krijgsveld J. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. Nat Protoc. 2019 Jan;14(1):68-85.

[50] Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Björk L, Breckels LM, Bäckström A, Danielsson F, Fagerberg L, Fall J, Gatto L, Gnann C, Hober S, Hjelmare M, Johansson F, Lee S, Lindskog C, Mulder J, Mulvey CM, Nilsson P, Oksvold P, Rockberg J, Schutten R, Schwenk JM, Sivertsson Å, Sjöstedt E, Skogs M, Stadler C, Sullivan DP, Tegel H, Winsnes C, Zhang C, Zwahlen M, Mardinoglu A, Pontén F, von Feilitzen K, Lilley KS, Uhlén M, Lundberg E. A subcellular map of the human proteome. Science. 2017 May 26;356(6340):eaal3321.

[51] Hakun MC, Gu B. Challenges and Opportunities in NUT Carcinoma Research. Genes (Basel). 2021 Feb 5;12(2):235.

[52] French CA. NUT midline carcinoma. Cancer Genet Cytogenet. 2010 Nov;203(1):16-20.

[53] Kariya Y, Kariya Y. Osteopontin in cancer: mechanisms and therapeutic targets. Int J of Transl Med. 2022 Aug 19;2(3):419-47.

[54] Zhao H, Chen Q, Alam A, Cui J, Suen KC, Soo AP, Eguchi S, Gu J, Ma D. The role of osteopontin in the progression of solid organ tumour. Cell Death Dis. 2018 Mar 2;9(3):356.

[55] Shi L, Wang X. Role of osteopontin in lung cancer evolution and heterogeneity. Semin Cell Dev Biol. 2017 Apr;64:40-47.

[56] Hao C, Lane J, Jiang WG. Osteopontin and Cancer: Insights into Its Role in Drug Resistance. Biomedicines. 2023 Jan 12;11(1):197.

[57] Rascio F, Spadaccino F, Rocchetti MT, Castellano G, Stallone G, Netti GS, Ranieri E. The Pathogenic Role of PI3K/AKT Pathway in Cancer Onset and Drug Resistance: An Updated Review. Cancers (Basel). 2021 Aug 5;13(16):3949.

CHAPTER 3.* Pilot investigation of magnetic nanoparticle-based immobilized metal affinity chromatography for efficient enrichment of phosphoproteoforms for mass spectrometry-based top-down proteomics

3.1 Introduction

Protein phosphorylation is a vital and common post-translational modification (PTM), modulating various biological processes and diseases [1-5]. Mass spectrometry (MS)-based phosphoproteomics has been widely deployed for large-scale characterization of protein phosphorylation in cells and tissues across various biological conditions [6]. However, almost all the phosphoproteomics studies were performed using bottom-up proteomics (BUP) [6], which cannot provide clear knowledge of phosphoprotein proteoforms due to enzymatic digestion. Proteoforms are a group of protein molecules that derive from the same gene due to RNA alternative splicing and protein PTMs [7]. Strong pieces of evidence suggest that proteoforms from the same gene can have drastically different biological functions [8-11]. Therefore, the characterization of phosphoproteins in a proteoform-specific manner (i.e., phosphoproteoform) is critical for accurate understanding of phosphoproteins' biological function in biological processes and diseases [12].

Top-down proteomics (TDP), unlike BUP, characterizes intact proteoforms using mass spectrometry (MS) and tandem mass spectrometry (MS/MS), providing rich information on PTMs and their combinations on proteins [13]. TDP has been employed for large-scale identification and quantification of proteoforms across cells and tissues to better our understanding of fundamental biological processes and to determine disease-related proteoform biomarkers [12-19]. However, the TDP study of phosphoproteoforms has largely lagged because of the relatively low abundance of phosphoproteoforms in complex proteomes. It is also more challenging to enrich phosphoproteoforms compared to phosphopeptides due to their much larger size and much more complex structure than phosphopeptides. To advance the TDP of phosphoproteoforms, highly efficient and selective enrichment technologies for phosphoproteoforms are crucial.

---

* *This Chapter is partially adapted with permission from Wang Q, Fang F, Sun L. Pilot investigation of magnetic nanoparticle-based immobilized metal affinity chromatography for efficient enrichment of phosphoproteoforms for mass spectrometry-based top-down proteomics. Anal Bioanal Chem. 2023 Jul;415(18):4521-4531.*

A variety of techniques have been developed for the enrichment of phosphopeptides from complex peptide mixtures for BUP-based phosphoproteomics [6], including the most widely used immobilized metal affinity chromatography (IMAC, e.g., Ti^{4+} and Fe^{3+}) [20-23], antibody [24-26], ion exchange chromatography (IEX) [27, 28], and affinity tag [29, 30]. However, only a few studies investigated techniques for highly selective phosphoproteoform isolation for TDP. The Ge group synthesized several different novel nanoparticles with affinity tags for the enrichment of phosphoproteoforms from complex samples, followed by liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based TDP analysis [31-33]. The Yu group synthesized a Ti^{4+}-IMAC material based on polyoxometalate/polydopamine composite microspheres for selective isolation of phosphoproteins, and SDS-PAGE was utilized to evaluate the performance of the IMAC for standard phosphoproteins and low-complexity samples [34]. Those studies have demonstrated the potential of large-scale top-down characterization of phosphoproteoforms with the assistance of selective enrichment techniques. However, much more efforts need to be made to achieve comprehensive TDP of phosphoproteome in a proteoform-specific manner regarding phosphoproteoform enrichment, separation, MS/MS, and identification through bioinformatic tools.

Herein, we investigated the magnetic nanoparticle-based IMAC materials with Ti^{4+} and Fe^{3+} for highly specific phosphoproteoform enrichment from a complex cell lysate for TDP for the first time. The IMAC procedure is similar to the typical one using IMAC for phosphopeptides but with substantially different buffers. The Ti^{4+}-IMAC and Fe^{3+}-IMAC magnetic materials were prepared using a well-established procedure with a relatively new linker for the immobilization of Ti^{4+} and Fe^{3+}. We systematically characterized the magnetic IMAC nanomaterials and evaluated their performance for phosphoproteoform enrichment using SDS-PAGE and LC-MS/MS. We compared the Ti^{4+}-IMAC and Fe^{3+}-IMAC methods regarding the identified phosphoproteoforms.

3.2 Experimental section

3.2.1 Materials and reagents

All materials are purchased from Sigma-Aldrich (St. Louis, MO) unless otherwise stated. Amine-terminated nanoparticles were purchased from Polysciences (Warrington, PA). Sodium phosphate monobasic (NaH_2PO_4) and sodium cyanoborohydride ($NaBH_3CN$) were ordered from Fisher Scientific (Pittsburgh, PA). Pierce™ Phosphoprotein Enrichment Kit, Coomassie Brilliant Blue G-250, Pro-Q™ Diamond Phosphoprotein Gel Stain, and SYPRO™ Ruby Protein Gel Stain

were purchased from Thermo Scientific (Waltham, MA). Mini-PROTEAN Precast Mini PAGE Gel was from Bio-rad (Hercules, CA). Tris, HEPES, sodium phosphate dibasic (Na_2HPO_4), and sodium chloride (NaCl) were ordered from Invitrogen (Waltham, MA), GoldBio (St. Louis, MO), Jade Scientific (Westland, MI), and ChemPure Brand (Westland, MI), respectively.

3.2.2 Preparation of Ti^{4+}-IMAC magnetic nanoparticles

The Ti^{4+}-IMAC nanoparticles were synthesized using commercially available reagents following the schematics shown in **Figure 3.1A**. (1) Amine-terminated magnetic nanoparticles (10 mg) were suspended in 200 μL of deionized water. Water was removed by a magnet for isolation. Then, 600 μL of 10% (v/v) glutaraldehyde in 100 mM phosphate aqueous buffer (93.5 mM Na_2HPO_4, 6.5 mM NaH_2PO_4, pH 8.0) was added and mixed well with the nanoparticles via vortex for 10 s. Then, the mixture was kept at room temperature for 6 h under a gentle mixing condition with a tube revolver rotator (Thermo Scientific) at 15 rpm. After the reaction, the aldehyde-functionalized nanoparticles (Intermediate 1) were washed three times with 100 mM phosphate aqueous buffer (600 μL each time). (2) 600 μL of AMPA (amino-methylphosphonic acid, 2 mg/mL) and $NaBH_3CN$ (10 mg/ mL) in 100 mM phosphate aqueous buffer was added to the Intermediate 1 for reaction at room temperature for 6 h under the same mixing condition as step 1. The Intermediate 2 was washed three times with LC-MS water. (3) 1.8 mL of 100 mM $TiCl_4$ in LC-MS water (note: precipitates were removed from the solution after mixing $TiCl_4$ with the water) was added to Intermediate 2 for reaction at room temperature for 6 h under the same mixing condition as the previous steps to produce the final product, Ti^{4+}-IMAC magnetic nanoparticles. After washing with LC-MS water three times, the final product was kept in 200 μL of LC-MS water at 4 °C before use.

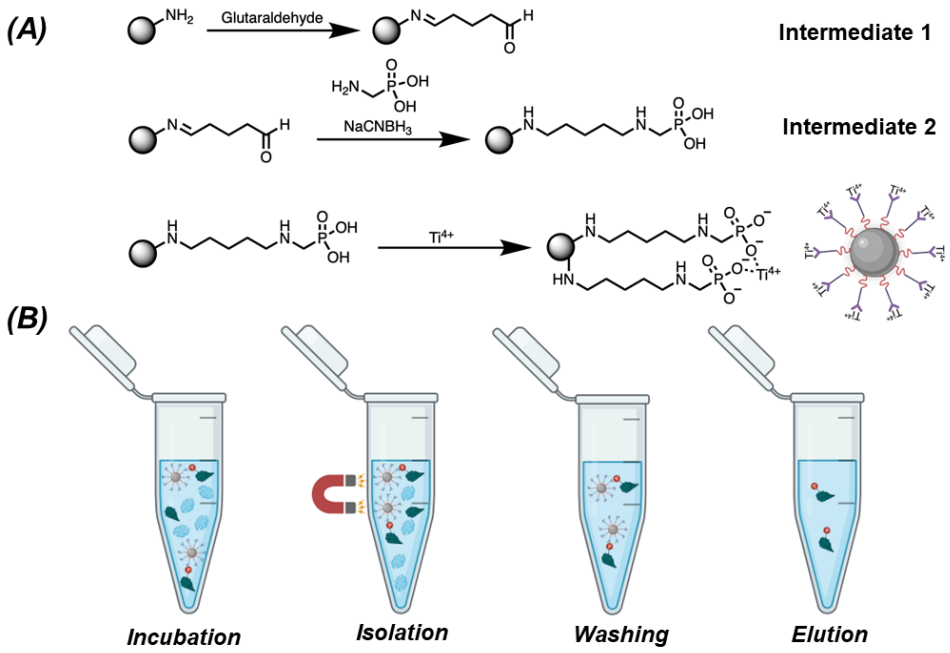


Figure 3.1. (A). Schematic of the synthesis process of Ti^{4+}-IMAC magnetic nanoparticles. The Ti^{4+} chelated on the linker has strong binding to the phosphate groups. The cartoon at the bottom right shows the schematic diagram of the final product of Ti^{4+}-IMAC magnetic nanoparticles. (B). Schematic of the phosphoproteoform enrichment process from protein mixtures using magnetic nanoparticle-based Ti^{4+}-IMAC.

3.2.3 Characterization of Ti^{4+}-IMAC magnetic nanoparticles

Transmission electron microscopy (TEM) was conducted on a JEOL JEM-1400 Flash instrument operated at 80 kV. Energy Dispersive X-ray Microanalysis (EDS) was carried out at 10kV by an ultra-high resolution JEOL 7500F scanning electron microscope equipped with Oxford EDS systems for elemental analysis. Zeta potential was measured using the Zetasizer Nano instrument (Malvern Panalytical) at 377.6 kcps count rates, 12 zeta runs, 2.00 mm measurement position, and 5 attenuator. Thermogravimetric analysis (TGA) was conducted using a TGA Q500 thermal analysis system (Waters Corporation) under a N_2 atmosphere at a constant heating rate of 10 °C/min from 100 °C to 600 °C. All samples were first heated to 100 °C and held at that temperature for 2 min to fully dry down the particles.

3.2.4 Phosphoprotein enrichment using Ti^{4+}-IMAC magnetic nanoparticles and the commercial kit: standard proteins

The procedure for phosphoprotein enrichment with the Ti^{4+}-IMAC is shown in **Figure 3.1B**. A mixture of bovine serum albumin (BSA, unphosphorylated protein) and β-casein

(phosphorylated protein) was dissolved in the loading buffer (50 mM HEPES-NaOH, 200 mM NaCl, pH 7.0). Then, the sample (4.5 mL) was mixed with Ti^{4+}-IMAC nanoparticles (10 mg) and kept at room temperature for 2 hours. Next, the nanoparticles were washed twice with the loading buffer (4.5 mL) to remove non-specific binding proteins. At last, the phosphoproteins bound to the nanoparticles were eluted twice with about 2 mL of elution buffer (200 mM Na_2HPO_4, 200 mM NaCl, pH 7.0). The two eluates were combined. Protein concentration in the loading mixture (LM), flow-through (FT), and elution (E) samples were measured by the Bicinchoninic Acid (BCA) assay. These samples were mixed with a 4X SDS-PAGE sample buffer (1.0 M Tris pH 6.8, 2 mL; SDS, 0.8g; Bromophenol Blue, 0.04g; Glycerol, 4mL; LC/MS grade water, 4mL; 1M DTT, 1mL), and boiled at 94 °C for 10min. Then a Mini-PROTEAN Precast Mini PAGE Gel (Bio-rad) was used for gel electrophoresis. The parameters for SDS PAGE were set at 150 V for 50min. After Coomassie Blue staining, the recovery rate was calculated using the ImageJ software (*%Recovery = amount of β-casein after enrichment / amount of β-casein before enrichment ×100%*).

For the phosphoprotein enrichment using the commercial kit, the Pierce™ phosphoprotein enrichment kit was used following the manufacturer's protocol. First, the storage solution was removed, and 5 mL of the loading buffer was applied to equilibrate the commercial column via centrifugation at 1000 × g for 1 minute at 4 °C. Second, 4 mg of the standard protein mixture (BSA, 10 μM; β-casein, 10 μM) in the loading buffer (LM) was incubated with the column on a tube revolver rotator for 30 minutes at 4°C. Third, the flow-through (FT) was collected by centrifugation at 1,000 × g for 1 minute at 4 °C. Fourth, the column was washed three times with 4.5 mL loading buffer in total via centrifugation at 1,000 × g for 1 minute at 4°C to remove the non-specific protein binding (W1, W2, and W3). Finally, the phosphoproteins were eluted four times using a total volume of 4.5 mL elution buffer (E) via centrifugation at 1,000 × g for 1 minute at 4°C. The eluted sample was desalted and concentrated to 1 mL for SDS-PAGE analysis as described above.

3.2.5 Preparation of yeast cell lysate, phosphoproteoform enrichment, and SDS-PAGE analysis

The baker's yeast (YSC1 purchased from Sigma Aldrich) was cultured in YPD broth (Sigma Aldrich) according to the general yeast growth protocol. After the harvest, the yeast was washed 3 times by the loading buffer and well dispersed in the loading buffer supplemented with

1× cOmplete protease inhibitor and 1× PhosSTOP phosphatase inhibitor. The yeast was lysed for 2 minutes using a homogenizer 150 (Fisher Scientific) and then sonicated on ice for 10 minutes by a Branson Sonifier 250 (VWR Scientific). Next, the yeast lysate was centrifuged at 14,000 g for 10 min under 4°C. The supernatant was kept, and the protein concentration was measured by the BCA assay. The extracted proteins were stored at -80 °C before use.

The phosphoproteoform enrichment procedure was the same as the standard protein mixture experiment (BSA and β-casein) with some modifications. 3 mg yeast proteins in 1 mL loading buffer were mixed with 10 mg of IMAC magnetic nanoparticles. 1mL of loading buffer was used for washing the beads twice to remove non-specific binding. 200 µL of elution buffer was used to elute phosphoproteoform from beads twice. After enrichment, loading mixture (LM), flow-through (FT), and elution (E) were desalted with a 10-kDa molecular weight cut-off membrane (Millipore Sigma, Inc) before analysis.

15 µg yeast proteins of LM, FT, and E were separated by SDS-PAGE. Precast Mini PAGE Gel (4-20%, Bio-rad) was used for gel electrophoresis (150 V, 50 min). The gel was stained with Pro-Q™ Diamond Phosphoprotein Gel Stain followed by SYPRO™ Ruby Protein Gel Stain according to the manufacturer's protocols. To visualize phosphoproteins and all the proteins, the gel was imaged using a ChemiDoc MP system (Bio-rad) with built-in settings for Pro-Q Diamond and SYPRO Ruby fluorescent dye separately.

3.2.6 Phosphoproteoform enrichment from yeast cell lysate and reversed-phase LC (RPLC)-MS/MS

The phosphoproteoform enrichment procedure was the same as mentioned before except that 5 mg of yeast proteins were used as the starting material. About 200 µg of proteins were recovered in the eluate (E) after the enrichment. The protein sample was dissolved in 400 µL of 0.1% (v/v) formic acid (0.5 mg/mL).

For RPLC-MS/MS, an EASY-RPLC™ 1200 system and a Q-Exactive HF mass spectrometer (Thermo Fisher Scientific) (Thermo Fisher Scientific) were used. The yeast sample was dissolved in 0.1% (v/v) FA. 1 µL of the sample corresponding to 0.5 µg proteins was separated on a home-packed C4 separation column (100-µm i.d. × 30 cm, 3 µm particles, 300 Å, Sepax Technologies, Inc.) at a flow rate of 500 nL/min. Mobile phase A was 5% (v/v) ACN in water containing 0.1% (v/v) FA), and mobile phase B was 80% (v/v) ACN and 0.1% (v/v) FA.

For separation, a 105-min gradient was used: 0-85 min, 5-70% B; 85-90 min, 70-100% B; 90-105 min, 100% B.

The electrospray voltage was set to 1.8 kV. A Top5 DDA method was used. The mass resolution was set to 120,000 (at *m/z* 200) for full MS scans and 60,000 (at *m/z* 200) for MS/MS scans. For full MS scans and MS/MS scans, the target values were 3E6 and 1E6, and the maximum injection time was 100 ms and 200 ms, respectively. The scan range was 600 to 2000 *m/z* for full MS scans. For MS/MS scans, the isolation window was 4 *m/z*. Fragmentation in the HCD cell was performed with a normalized collision energy of 20%. The fixed first mass was set to 100 *m/z* for MS/MS. Dynamic exclusion was applied and it was set to 30 s. Ions with charge states from +1 to +5 were not considered for fragmentation.

3.2.7 Data analysis

All the RAW files were analyzed with the TopPIC (Top-down mass spectrometry-based proteoform identification and characterization) software (version 1.5.2) [35].

The RAW files were first converted to mzML files with the MsConvert software [36], and spectral deconvolution was performed with the TopFD (Top-down mass spectrometry feature detection) software, generating msalign files, which were used as the input for database searching using TopPIC. The spectra were searched against a yeast database (downloaded from Swiss-Uniprot, September 2021). False discovery rates (FDRs) were estimated using the target-decoy approach [37, 38]. A 1% proteoform spectrum match (PrSM)-level FDR and a 5% proteoform-level FDR were employed to filter the identifications. The mass error tolerance was 15 ppm. The mass error tolerance was 1.2 Da for identifying PrSM clusters. The maximum mass shift was 500 Da. The maximum number of mass shift was set to 2.

For phosphoproteoform determination, we used multiple strategies. First, we manually checked the reported mass shifts from TopPIC, considering single phosphorylation (around 80-Da mass shift), multiple phosphorylation (i.e., around 160-Da and 240-Da mass shifts), and combinations of phosphorylation and other common PTMs (e.g., methylation and acetylation). Second, we confirmed those PTMs according to the information on the UniProt database (https://www.uniprot.org) and YAAM database (http://yaam.ifc.unam.mx). Third, we manually checked the MS/MS spectra of some identified phosphoproteoforms for the neural loss of phosphorylation (80-Da or 98-Da) caused by HCD fragmentation.

3.3 Results and discussion

3.3.1 Characterization of Ti^{4+}-IMAC magnetic nanoparticles

The synthesis of Ti^{4+}-IMAC nanoparticles is shown in **Figure 3.1A**. First, the amine-terminated nanoparticles (ATNPs) were mixed with glutaraldehyde to bring an aldehyde group to the NPs. Then, this intermediate **1** reacted with sodium cyanoborohydride and AMPA to generate stable phosphate groups to the NPs. Finally, the Ti^{4+} was immobilized on the surface of intermediate **2** based on the chelating interaction between the phosphate group and Ti^{4+}.

The TEM results (**Figure 3.2A**) revealed the size of the Ti^{4+}-IMAC magnetic nanoparticles is smaller than 20 nm, and the functionalization didn't increase the particle size compared to the initial ATNPs, most likely because the reactions only added short carbon chains to the particle surface. **Figure 3.2B** shows the elemental composition analysis data of ATNPs and Ti^{4+}-IMAC nanoparticles from EDS. Ti^{4+}-IMAC nanoparticles had substantially higher amounts of Ti^{4+} and P compared to ATNPs (Ti^{4+}, 8.5% vs. 0%; P, 4.7% vs. 0%), indicating the successful functionalization of the magnetic nanoparticles with phosphate groups and Ti^{4+} ions. Additionally, we further characterized the zeta potentials of ATNPs and Ti^{4+}-IMAC nanoparticles by Laser Doppler Velocimetry (LDV), **Figure 3.2C**. The zeta potential of the ATNPs and Ti^{4+}-IMAC nanoparticles were 19.8 mV and -29.6 mV. This negative shift is due to the replacement of the amine group with the phosphate group on the nanoparticle surface. Finally, the TGA analysis results of ATNPs and Ti^{4+}-IMAC nanoparticles demonstrated a much more significant weight loss of Ti^{4+}-IMAC compared to the original ATNPs, **Figure 3.2D**. The phenomenon is due to several more chemical modifications of the Ti^{4+}-IMAC nanoparticles compared to the original ATNPs. **Figure 3.2E** shows that the Ti^{4+}-IMAC magnetic particles can be well dispersed in water and easily separated from water by a magnet, which guarantees efficient interactions between nanoparticles and phosphoproteins as well as easy operations.

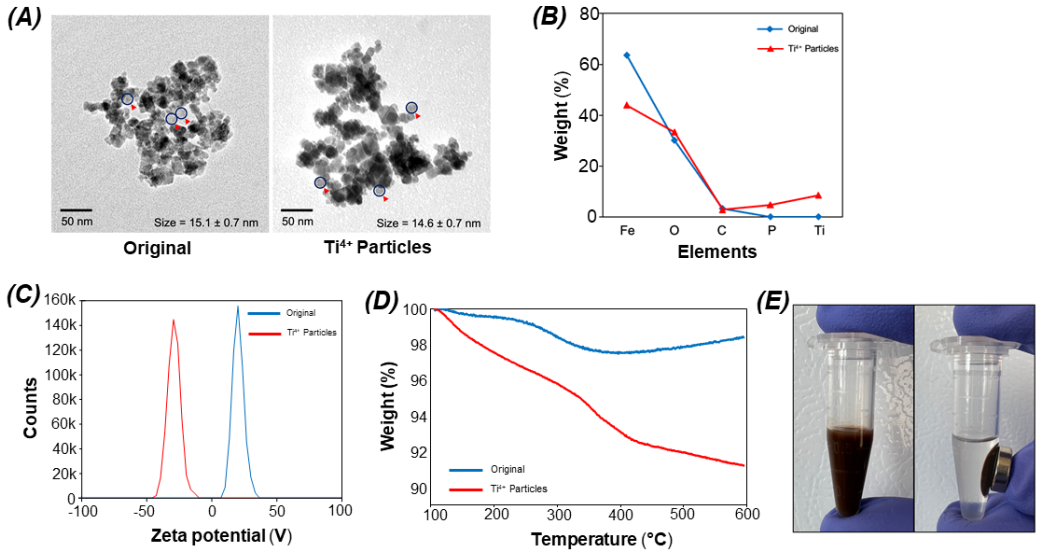


Figure 3.2. Characterization results of magnetic nanoparticle-based Ti^{4+}-IMAC material, including TEM (A), elemental composition analysis (B), zeta potential analysis (C), weight loss analysis (D), and water dispersion and magnetic separation tests (E). Each red arrow in (A) is pointing at a single magnetic nanoparticle and those nanoparticles were used to estimate the means and standard deviations of the size of original ATNPs and Ti^{4+}-IMAC nanoparticles.

3.3.2 Phosphoprotein enrichment by the Ti^{4+}-IMAC nanoparticles and SDS-PAGE analysis: standard proteins

We used a standard protein mixture containing β-casein (β, a phosphoprotein) and bovine serum albumin (BSA, a non-phosphoprotein) to evaluate the performance of Ti^{4+}-IMAC nanoparticles for selectively isolating phosphoproteins. The experimental procedure is shown in **Figure 3.1B**, including (1) mixing and incubating the protein mixture (loading mixture, LM) with the Ti^{4+}-IMAC, (2) selectively isolating the phosphoproteins by the Ti^{4+}-IMAC and removing the non-phosphoproteins in the solution (flow-through, FT), (3) washing away the non-phosphoproteins efficiently by a couple of washing steps (Wash 1 and 2, W1 and W2), and (4) eluting phosphoproteins from the Ti^{4+}-IMAC (Elution, E). We chose the salt concentrations in the loading buffer, washing buffer, and elution buffer according to one previous report. [31] We optimized the pH of the loading buffer, washing buffer, and elution buffer using the standard protein mixture (BSA and β-casein molar ratio as 10:1, 100 μM:10 μM), shown in **Figure 3.3**. We used SDS-PAGE to evaluate the enrichment efficiency of phosphoproteins, and the gel was stained with Coomassie blue dye to observe the phosphoproteins and non-phosphoproteins. After considering both non-phosphoprotein removal and phosphoprotein recovery, we decided to choose the pH 7.0 buffers for all the following experiments. The loading and washing buffer

contained 50 mM HEPES-NaOH and 200 mM NaCl (pH 7.0). The elution buffer contained 200 mM NaCl and 200 mM Na_2HPO_4 (pH 7.0).

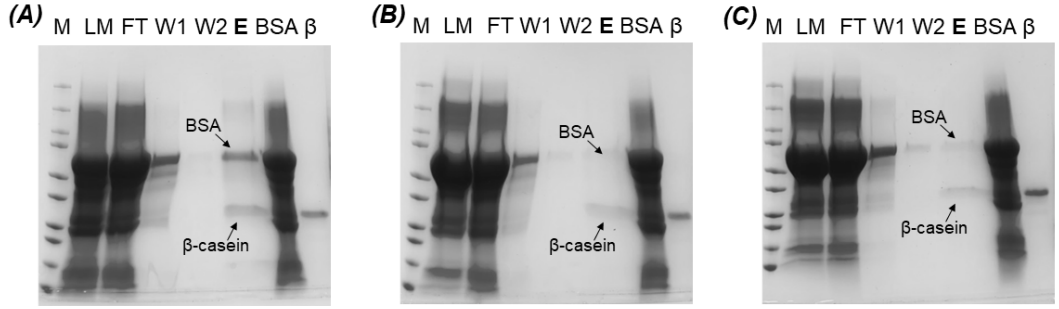


Figure 3.3. Optimization of the pH of the loading buffer, washing buffer, and elution buffer. Three different pH were evaluated, pH 6.0 (A), pH 7.0 (B), and pH 8.0 (C). The composition of the loading buffer and washing buffer is the same, containing 50 mM HEPES and 200 mM NaCl. The elution buffer contained 200 mM Na_2HPO_4 and 200 mM NaCl. The loading mixture (LM) contained a mixture of BSA and β-casein with a molar ratio of 10:1 (100 μM:10 μM). M: protein molecular weight marker; LM: loading mixture (the standard protein mixture before IMAC enrichment); FT: flow-through; W1 and W2: the first and second wash; E: eluate from the Ti^{4+}-IMAC magnetic nanoparticles after enrichment; β: β-casein standard; BSA: BSA standard.

Figure 3.4 shows the SDS-PAGE data of the standard protein mixture (1:1 molar ratio of BSA: β-casein, 10 μM:10 μM) after treatment by the Ti^{4+}-IMAC (A) and the commercial kit (B) as well as the standard protein mixture (10:1 molar ratio of BSA: β-casein, 100 μM:10 μM) after treatment by the Ti^{4+}-IMAC (C). It is clear that Ti^{4+}-IMAC can selectively capture the phosphoprotein (β-casein) and efficiently remove the non-phosphoprotein (BSA) even when BSA has a 10-fold higher concentration than β-casein, **Figure 3.4A** and **3.4C**. Compared to the commercial kit (**Figure 3.4B**), the Ti^{4+}-IMAC had a better performance regarding the capture efficiency for phosphoproteins. As marked by the red ovals, clear β-casein bands were observed in the flow-through (FT) and Wash (W1) samples from the commercial kit; no obvious signals were obtained in FT and W1 samples from the Ti^{4+}-IMAC. The β-casein recovery from the Ti^{4+}-IMAC is much higher than that from the commercial kit (46% vs. 37%). We further tested the reproducibility of the Ti^{4+}-IMAC for phosphoprotein enrichment using the standard protein mixture (BSA:β-casein, 100 μM:10 μM). A reproducible β-casein recovery (48±8%) was produced from quadruplicate preparations. We want to highlight that another important advantage of our Ti^{4+}-IMAC magnetic particles compared to the commercial kit is its easy operations via a magnet without the need for centrifugation. Additionally, the Ti^{4+}-IMAC method could be used for a variety of initial amounts of protein materials via a simple adjustment

of the mass of magnetic particles depending on the availability of the biological samples. We noted that there is still a visible BSA band in the eluates of the Ti^{4+}-IMAC (**Figures 3.4A** and **3.4C**) and there is no clear BSA signal in the elution sample of the commercial kit (**Figure 3.4B**). Some further improvement of the surface chemistry of Ti^{4+}-IMAC magnetic particles could be done to reduce the non-specific binding of non-phosphoproteins and will be investigated in our future study.

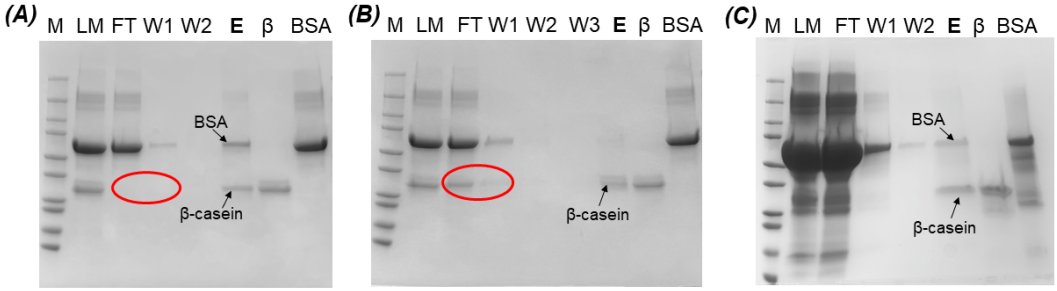


Figure 3.4. SDS-PAGE data of a standard protein mixture (BSA and β-casein) after selective isolation of phosphoprotein β-casein with magnetic nanoparticle-based Ti^{4+}-IMAC (A and C) and the commercial phosphoprotein enrichment kit (B). For A and B, the concentration of BSA and β-casein in the sample was both 10 μM. For C, the concentration of BSA was 10 times higher than β-casein (100 μM vs. 10 μM). M: protein molecular weight marker; LM: loading mixture (the standard protein mixture before IMAC enrichment); FT: flow-through; W1, W2, and W3: the first, second, and third wash; E: eluate from the Ti^{4+}-IMAC magnetic nanoparticles after enrichment; β: β-casein standard; BSA: BSA standard.

3.3.3 Phosphoproteoform enrichment by Ti^{4+}-IMAC nanoparticles and SDS-PAGE analysis: a yeast cell lysate

We further validated the performance of the Ti^{4+}-IMAC magnetic nanoparticles for phosphoproteoform enrichment from a complex sample, a yeast cell lysate. 3 mg of yeast proteins and 10 mg of Ti^{4+}-IMAC magnetic nanoparticles were used. The experiment was performed in triplicate. We loaded an equal amount of proteins for loading mixture (LM), flow-through (FT), and elution (E) into each lane of SDS-PAGE gel for analysis. We first stained the gel using Pro-Q Diamond to detect phosphoproteoforms specifically. Then, we de-stained the gel and re-stained it with SYPRO Ruby to detect total proteins in the samples.

As shown in **Figure 3.5A**, much more visible phosphoproteoform bands were observed in the eluates (E1, E2, and E3) compared to the loading mixture (LM) in the triplicate preparations. The Ti^{4+}-IMAC method has nice reproducibility according to the phosphoproteoform profiles in the three eluates (E1, E2, and E3). **Figure 3.5B** further shows the

nice reproducibility of the technique at the total proteoform level (E1, E2, and E3). By comparing the total proteoform and phosphoproteoform profiles in the eluates, we observed that the major proteoform bands (≤75 kDa) are relatively consistent, indicating the reasonably high specificity of the technique for phosphoproteoform enrichment from complex samples. We noted that many visible phosphoproteoform bands exist for the flow-through sample (FT) and some bands even have a higher intensity than that in the eluates. To get a better understanding of this phenomenon, we determined the loading capacity of the Ti^{4+}-IMAC nanoparticles using β-casein as the sample, shown in **Figure 3.6A**. The loading capacity is about 140 μg phosphoproteins/mg nanoparticles for Ti^{4+}-IMAC and the enrichment process could be done within one hour. The 10 mg of Ti^{4+}-IMAC magnetic nanoparticles used in the experiment could capture more than 1 mg of phosphoproteoforms. Interestingly, only about 200 μg of proteins were recovered in the eluate © after the enrichment. The results suggest that the Ti^{4+}-IMAC cannot capture all the phosphoproteoforms in the cell lysate, probably due to the three-dimensional structure of intact phosphoproteoforms and the selectivity of Ti^{4+}.

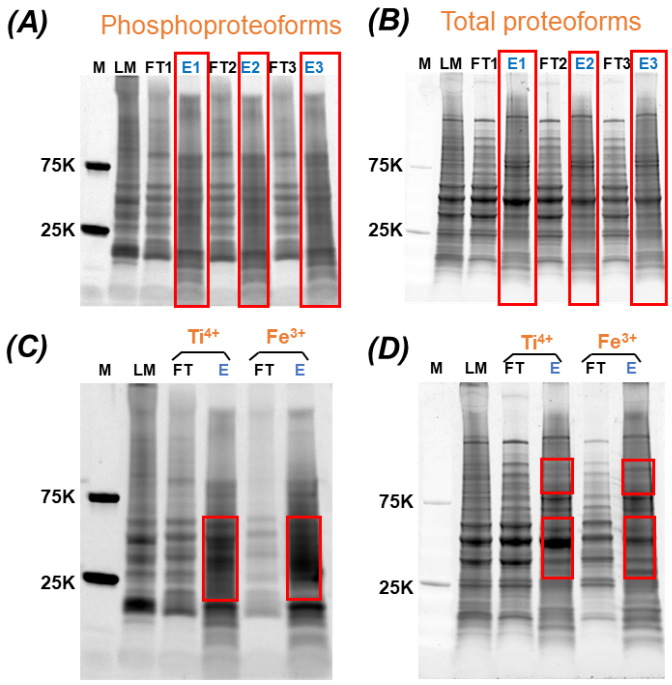


Figure 3.5. SDS-PAGE data of a yeast cell lysate. Visualization of phosphoproteoforms by the Pro-Q Diamond staining (A) and total proteoforms by SYPRO Ruby staining (B) after phosphoproteoform enrichment by magnetic nanoparticle-based Ti^{4+}-IMAC in triplicate experiments. M: protein molecular weight marker; LM: loading mixture (the yeast cell lysate before IMAC enrichment); FT1, FT2, and FT3: flow-through from the first, second, and third experiment; E1, E2, and E3: eluate from the Ti^{4+}-IMAC magnetic nanoparticles after enrichment in the first, second, and third experiment. Direct comparisons of Ti^{4+}-IMAC and Fe^{3+}-IMAC regarding the profile of phosphoproteoforms (C) and total proteoforms (D) isolated from the yeast cell lysate.

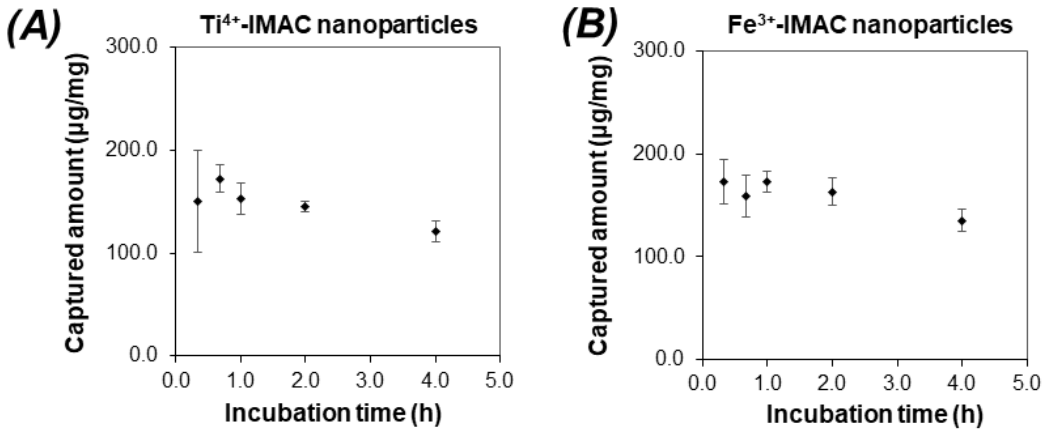


Figure 3.6. Loading capacity measurement of Ti^{4+}-IMAC nanoparticles (A) and Fe^{3+}-IMAC nanoparticles (B). 1 mg nanoparticles were incubated with 1000 µL of 0.6 mg/mL of β-casein solution in the loading buffer. After different incubation time periods (20-min, 40-min, 1 hour, 2 hours, and 4 hours), aliquots of the protein solution were collected for protein concentration measurement using the BCA assay. According to the protein concentration difference between the original β-casein solution and the solution after incubation with IMAC magnetic nanoparticles, we determined the captured protein amount. The error bars show the standard deviations of captured phosphoprotein amount from triplicate measurements.

Considering that IMAC with different metal ions (e.g., Ti^{4+} and Fe^{3+}) could enrich different pools of phosphopeptides from complex proteome samples [39, 40], we compared Ti^{4+}-IMAC and Fe^{3+}-IMAC for phosphoproteoform enrichment for the first time here. The Fe^{3+}-IMAC magnetic nanoparticles were prepared using the same procedure as the Ti^{4+}-IMAC material, and the salt $FeCl_3$ was used as the source of the Fe^{3+}. We employed the same protocol for the phosphoproteoform enrichment from the yeast cell lysate using the Ti^{4+}-IMAC and Fe^{3+}-IMAC magnetic nanoparticles. The loading capacity is about 160 µg phosphoproteins/mg nanoparticles for Fe^{3+}-IMAC nanoparticles, see **Figure 3.6B**. The Ti^{4+}-IMAC and Fe^{3+}-IMAC produced substantially different profiles of phosphoproteoforms, as evidenced by the SDS-PAGE data in **Figure 3.5C**. The total proteoform data in **Figure 3.5D** also indicates the distinguishable differences between Ti^{4+}-IMAC and Fe^{3+}-IMAC eluates. The data indicate that Ti^{4+}-IMAC and Fe^{3+}-IMAC are complementary for phosphoproteoform enrichment from complex proteomes and a combination of the two methods will be useful for improving the phosphoproteoform coverage.

3.3.4 RPLC-MS/MS-based top-down proteomics of yeast phosphoproteoforms enriched by Ti^{4+}-IMAC and Fe^{3+}-IMAC

We further enriched phosphoproteoforms from the yeast cell lysate using both Ti^{4+}-IMAC and Fe^{3+}-IMAC and analyzed the loading mixture (LM) and eluates (E) by RPLC-MS/MS. After database search by the TopPIC, 15, 28, and 32 phosphoproteoforms were identified from the LM, E of Ti^{4+}-IMAC, and E of Fe^{3+}-IMAC, respectively, with a 5% proteoform-level FDR.

The IMAC technique yielded about 100% more phosphoproteoform identifications compared to a direct RPLC-MS/MS analysis of the LM without enrichment (about 30 vs. 15). Interestingly, Ti^{4+}-IMAC and Fe^{3+}-IMAC produced different phosphoproteoform profiles, evidenced by the low proteoform-level overlap between the two methods, **Figure 3.7A**. In total, 48 phosphoproteoforms were identified by the two IMAC methods and only 12 of them were identified by both methods. The data agrees well with the data in **Figure 3.5C**. We noted that only 3 out of 15 phosphoproteoforms identified in the LM sample were also identified in the E of Ti^{4+}-IMAC or Fe^{3+}-IMAC, indicating that some phosphoproteoforms cannot be captured by the IMAC materials during the enrichment step, which agrees well with the SDS-PAGE data in **Figure 3.5**.

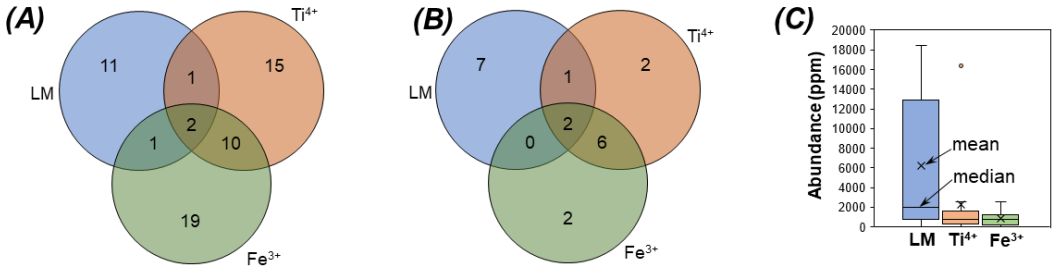


Figure 3.7. Phosphoproteoform identification results from the yeast cell lysate by RPLC-MS/MS. (A) Proteoform-level overlap among phosphoproteoforms identified from the yeast sample before IMAC enrichment (loading mixture, LM), after Ti^{4+}-IMAC enrichment (Ti^{4+}), and after Fe^{3+}-IMAC enrichment (Fe^{3+}). (B) Protein-level overlap among LM, Ti^{4+}, and Fe^{3+} for the identified phosphoproteoforms. (C) Boxplots of abundance (ppm) of proteins corresponding to the identified phosphoproteoforms from LM, Ti^{4+}, and Fe^{3+}. The protein abundance information was obtained from the Protein Abundance Database (PAXdb, version 4.2, https://pax-db.org/species/4932).

We speculated that proteins corresponding to the phosphoproteoforms identified in the LM had a relatively high abundance in the yeast cells, and the phosphoproteoforms can be identified directly by RPLC-MS/MS without the need for IMAC enrichment. However, for the phosphoproteoforms identified in the E after Ti^{4+}-IMAC or Fe^{3+}-IMAC, the corresponding proteins have relatively low abundance in the yeast cells and IMAC enrichment is critical for the characterization of those phosphoproteoforms. To prove this hypothesis, we first checked the protein-level overlaps among LM, E of Ti^{4+}-IMAC, and E of Fe^{3+}-IMAC for the identified phosphoproteoforms, followed by the investigation of protein relative abundance according to the Protein Abundance Database (PAXdb, version 4.2, https://pax-db.org/species/4932). As shown in **Figure 3.7B**, the protein-level overlaps between LM and E of Ti^{4+}-IMAC or LM and E of Fe^{3+}-IMAC are low. The protein abundance data in **Figure 3.7C** clearly indicate that phosphoproteins identified in the Es after IMAC enrichment have much lower abundance compared to that identified in the LM. The data clearly demonstrate the benefits of Ti^{4+}-IMAC and Fe^{3+}-IMAC enrichment for top-down proteomics of phosphoproteoforms with low abundance.

We noted that the number of phosphoproteoforms identified from IMAC eluates here is small compared to the total number of proteoform identifications (~30 *vs.* ~600). Those about 600 proteoforms correspond to roughly 200 proteins and the approximate 30 phosphoproteoforms derive from about 10 proteins. We further manually checked the identified total proteins from the Ti^{4+}-IMAC and Fe^{3+}-IMAC eluates in terms of phosphorylation through the online protein phosphorylation database PhosphoGRID (https://phosphogrid.org/). We found that at least more than 50% of those proteins have been reported as phosphorylated proteins. The reason why we only identified roughly 30 phosphoproteoforms of 10 proteins from each of the IMAC eluates by RPLC-MS/MS might be due to the phosphate group loss during sample processing and storage because of their dynamic features. This is because the buffer exchange steps prior to MS analysis removed phosphatase inhibitors in the solution, and enzymatic activities might happen. However, it is hard to make a solid conclusion about this point here. We will study the sample processing procedure in more detail and more samples to achieve a better understanding of this phenomenon in our future work.

Figure 3.8 shows two examples of identified phosphoproteoforms from the yeast cell lysate by Ti^{4+}-IMAC or Fe^{3+}-IMAC enrichment and RPLC-MS/MS. One phosphoproteoform of

gene *HYP2* (eukaryotic translation initiation factor 5A-1) shows a strong signal after Ti^{4+}-IMAC enrichment, but without a visible signal before enrichment, **Figure 3.8A**. In another example shown in **Figure 3.8B**, one phosphoproteoform of gene *STF2* (ATPase-stabilizing factor 15 kDa protein) has a drastically better signal after Fe^{3+}-IMAC enrichment compared to before enrichment. The data further demonstrate the highly efficient phosphoproteoform enrichment from complex samples by the Ti^{4+}-IMAC and Fe^{3+}-IMAC. Interestingly, we detected clear signals of the corresponding un-phosphoproteoforms of the genes *HYP2* and *STF2* not only before enrichment but also after IMAC enrichment, which might be due to either the dynamic nature of protein phosphorylation (loss of phosphate groups during the steps after enrichment) or the non-specific binding of un-phosphoproteoforms on the IMAC magnetic nanoparticles. **Figures 3.8C** and **3.8D** show the sequences and fragmentation patterns of two identified phosphoproteoforms. The two phosphoproteoforms were identified with high confidence and were characterized reasonably well. We noted that the proteoform shown in **Figure 3.8C** has both acetylation and phosphorylation close to its N-terminus. Although the database search software assigned the acetylation to the S2 residue and the phosphorylation to the S8 residue, there are still uncertainties in the PTM localization because of the lack of fragment ions from the first 10 amino acid residues. The data indicate a general challenge in top-down proteomics for accurate PTM localization.

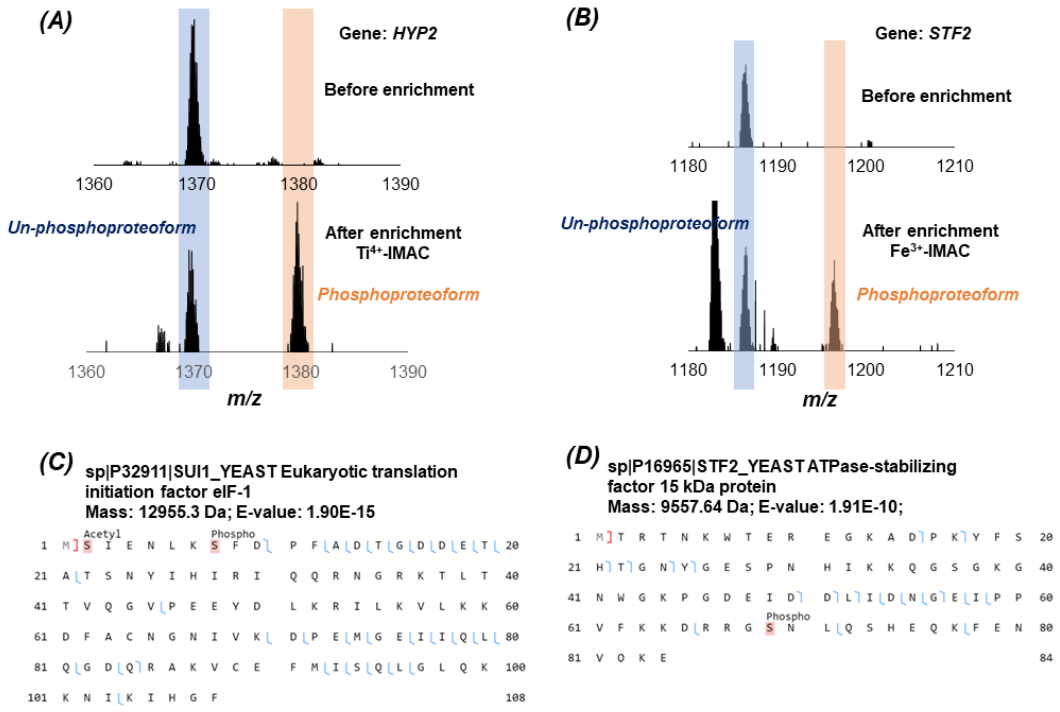


Figure 3.8. (A). Mass spectra of one *HYP2* phosphoproteoform before and after Ti^{4+}-IMAC enrichment. (B). Mass spectra of one *STF2* phosphoproteoform before and after Fe^{3+}-IMAC enrichment. (C, D). Sequences and fragmentation patterns of two example phosphoproteoforms.

Figure 3.9 in shows the sequences and fragmentation patterns of four example phosphoproteoforms with the combinations of multiple PTMs. Three phosphoproteoforms of Eukaryotic translation initiation factor 5A-1 (IF5A1) were identified with two phosphorylation sites (A), combinations of phosphorylation, hypusination, and acetylation (B), and combinations of phosphorylation, hypusination, and truncation (C). We identified over 10 different phosphopoteoforms of IF5A1 by Ti^{4+}-IMAC and Fe^{3+}-IMAC enrichment, suggesting the huge potential heterogeneity of phosphoproteoforms from the same gene. It is impossible to reveal the proteoform-level heterogeneity using the traditional BUP strategy. IF5A is a translation factor, and it has crucial functions in modulating cancer and brain aging. [41,42] However, the detailed functions of IF5A phosphorylation and hypusination in those processes are not clear. The capability of delineating various IF5A phosphoproteoforms with or without hypusination using TDP will establish the foundation for further elucidating their functions in cancer and brain aging. The data here highlight the significance of TDP for protein characterization in a proteoform-specific manner.

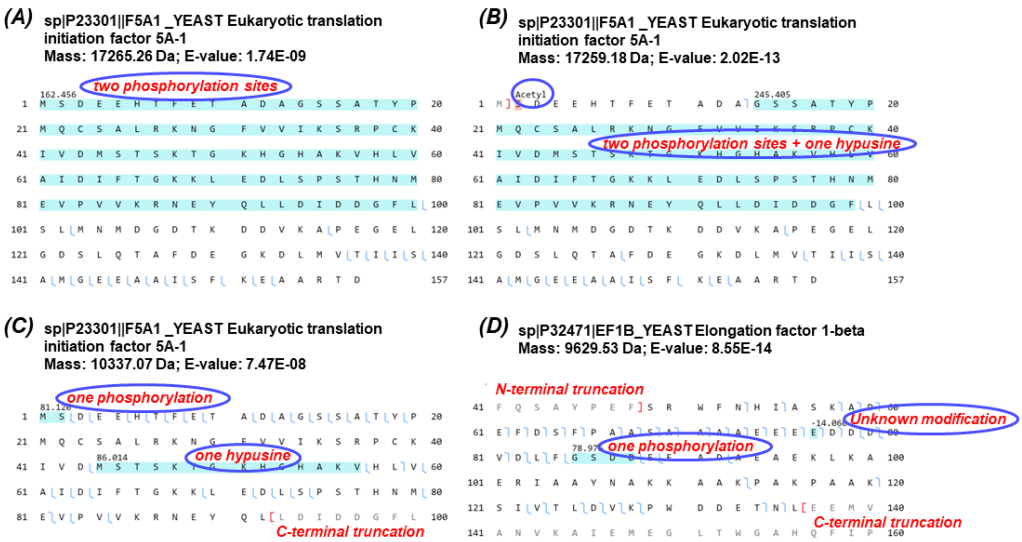


Figure 3.9. Sequences and fragmentation patterns of example phosphoproteoforms identified from the IMAC eluates by RPLC-MS/MS. (A). one phosphoproteoform with two phosphorylation sites; (B). one phosphoproteoform with N-terminal methionine removal, N-terminal acetylation, two phosphorylation sites, and one hypusine; (C). one phosphoproteoform with one phosphorylation, one hypusine, and C-terminal truncation; (D). one phosphoproteoform with one phosphorylation, one unknown modification, and terminal truncations.

3.4 Conclusion

In this pilot study, we investigated magnetic nanoparticles-based IMAC (Ti^{4+} and Fe^{3+}) for the enrichment of phosphoproteoforms from simple and complex protein mixtures for MS-based top-down proteomics. The IMAC methods achieved highly efficient and reproducible enrichment of intact phosphoproteoforms from a standard protein mixture and a yeast cell lysate. Substantially more phosphoproteoforms were identified from the yeast cell lysate after IMAC enrichment with Ti^{4+} or Fe^{3+} compared to that without enrichment. Interestingly, we documented that Ti^{4+}-IMAC and Fe^{3+}-IMAC tended to isolate different pools of phosphoproteoforms from a complex proteome.

We note that some improvements need to be made to achieve global top-down proteomics of phosphoproteoforms from complex proteomes. First, the surface chemistry of Ti^{4+} and Fe^{3+}-IMAC magnetic particles could be improved to reduce the non-specific binding of non-phosphoproteins and further boost the phosphoproteoform recovery. Second, the mass of all the identified phosphoproteoforms by RPLC-MS/MS is smaller than 20 kDa in this work due to the low sensitivity of top-down proteomics for the characterization of large proteoforms. Improvement of MS-based top-down proteomics technique for the identification of large

phosphoproteoforms will be an important topic in our future studies. Third, the full characterization of phosphoproteoforms is hampered by the unsatisfying backbone cleavage coverage of proteoforms from typical LC-MS/MS techniques with collision-based gas-phase fragmentation. We expect that coupling our IMAC techniques to LC-MS/MS equipped with collision, electron, and photon-based gas-phase fragmentation methods will advance the top-down proteomics of phosphoproteoforms drastically.

3.5 Acknowledgments

We thank the support from National Cancer Institute through Grant R01CA247863. We also thank the support from the National Institute of General Medical Sciences (NIGMS) through Grants 2R01GM118470 and R01GM125991 and the National Science Foundation through Grant DBI1846913 (CAREER Award).

REFERENCES

[1] Graves JD, Krebs EG. Protein phosphorylation and signal transduction. Pharmacol Ther. 1999; 82:111–121.

[2] Pawson T, Scott JD. Protein phosphorylation in signaling--50 years and counting. Trends Biochem. Sci. 30:286–290.

[3] Tarrant MK, Cole PA. The Chemical Biology of Protein Phosphorylation. Annu Rev Biochem. 2009; 78:797–825.

[4] Lemeer S, Heck AJR. The phosphoproteomics data explosion. Curr Opin Chem Biol. 2009; 13:414–420.

[5] Ardito F, Giuliani M, Perrone D, Troiano G, Lo Muzio L. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). Int J Mol Med. 2017; 40:271–280.

[6] Wang F, Song C, Cheng K, Jiang X, Ye M, Zou H. Perspectives of comprehensive phosphoproteome analysis using shotgun strategy. Anal Chem. 2011; 83:8078–8085.

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

[8] Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, Sun S, Yang F, Shen YA, Murray RR, Spirohn K, Begg BE, Duran-Frigola M, MacWilliams A, Pevzner SJ, Zhong Q, Trigg SA, Tam S, Ghamsari L, Sahni N, Yi S, Rodriguez MD, Balcha D, Tan G, Costanzo M, Andrews B, Boone C, Zhou XJ, Salehi-Ashtiani K, Charloteaux B, Chen AA, Calderwood MA, Aloy P, Roth FP, Hill DE, Iakoucheva LM, Xia Y, Vidal M. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell. 2016; 164:805–817.

[9] Smith LM, Agar JN, Chamot-Rooke J, Danis PO, Ge Y, Loo JA, Paša-Tolić L, Tsybin YO, Kelleher NL. The Human Proteoform Project: Defining the human proteome. Sci Adv. 2021; 7:eabk0734.

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

[11] Wang T, Holt M V, Young NL. The histone H4 proteoform dynamics in response to SUV4-20 inhibition reveals single molecule mechanisms of inhibitor resistance. Epigenetics Chromatin. 2018; 11:29.

[12] Tucholski T, Cai W, Gregorich ZR, Bayne EF, Mitchell SD, McIlwain SJ, de Lange WJ, Wrobbel M, Karp H, Hite Z, Vikhorev PG, Marston SB, Lal S, Li A, Dos Remedios C, Kohmoto T, Hermsen J, Ralphe JC, Kamp TJ, Moss RL, Ge Y. Distinct hypertrophic cardiomyopathy genotypes result in convergent sarcomeric proteoform profiles revealed by top-down proteomics. Proc Natl Acad Sci U S A. 2020; 117:24691–24700.

[13] Toby TK, Fornelli L, Kelleher NL. Progress in Top-Down Proteomics and the Analysis of Proteoforms. Annu Rev Anal Chem (Palo Alto Calif). 2016; 9:499–519.

[14] Chen B, Brown KA, Lin Z, Ge Y. Top-Down Proteomics: Ready for Prime Time? Anal Chem. 2018; 90:110–127.

[15] Wang Q, Sun L, Lundquist PK. Large-scale top-down proteomics of the Arabidopsis thaliana leaf and chloroplast proteomes. Proteomics. 2022; e2100377.

[16] Ansong C, Wu S, Meng D, Liu X, Brewer HM, Deatherage Kaiser BL, Nakayasu ES, Cort JR, Pevzner P, Smith RD, Heffron F, Adkins JN, Pasa-Tolic L. Top-down proteomics reveals a unique protein S-thiolation switch in Salmonella Typhimurium in response to infection-like conditions. Proc Natl Acad Sci U S A. 2013; 110:10153–10158.

[17] Ntai I, Fornelli L, DeHart CJ, Hutton JE, Doubleday PF, LeDuc RD, van Nispen AJ, Fellers RT, Whiteley G, Boja ES, Rodriguez H, Kelleher NL. Precise characterization of KRAS4b proteoforms in human colorectal cells and tumors reveals mutation/modification cross-talk. Proc Natl Acad Sci U S A. 2018; 115:4140–4145.

[18] Melani RD, Gerbasi VR, Anderson LC, Sikora JW, Toby TK, Hutton JE, Butcher DS, Negrão F, Seckler HS, Srzentić K, Fornelli L, Camarillo JM, LeDuc RD, Cesnik AJ, Lundberg E, Greer JB, Fellers RT, Robey MT, DeHart CJ, Forte E, Hendrickson CL, Abbatiello SE, Thomas PM, Kokaji AI, Levitsky J, Kelleher NL. The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells. Science. 2022; 375:411–418.

[19] Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, Tipton JD, Vellaichamy A, Kellie JF, Li M, Wu C, Sweet SMM, Early BP, Siuti N, LeDuc RD, Compton PD, Thomas PM, Kelleher NL. Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature. 2011; 480:254–258.

[20] Villen J, Gygi SP. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. Nat Protoc. 2008; 3:1630–1638.

[21] Zhou H, Ye M, Dong J, Han G, Jiang X, Wu R, Zou H. Specific phosphopeptide enrichment with immobilized titanium ion affinity chromatography adsorbent for phosphoproteome analysis. J Proteome Res. 2008; 7:3957–3967.

[22] Gruhler A, Olsen J V, Mohammed S, Mortensen P, Faergeman NJ, Mann M, Jensen ON. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. Mol Cell Proteomics. 2005; 4:310–327.

[23] Andersson L, Porath J. Isolation of phosphoproteins by immobilized metal (Fe3+) affinity chromatography. Anal Biochem. 1986; 154:250–254.

[24] Steen H, Kuster B, Fernandez M, Pandey A, Mann M. Tyrosine phosphorylation mapping of the epidermal growth factor receptor signaling pathway. J Biol Chem. 2002; 277:1031–1039.

[25] Grønborg M, Kristiansen TZ, Stensballe A, Andersen JS, Ohara O, Mann M, Jensen ON, Pandey A. A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies: identification of a novel protein, Frigg, as a protein kinase A substrate. Mol Cell Proteomics. 2002; 1:517–527.

[26] Pandey A, Podtelejnikov A V, Blagoev B, Bustelo XR, Mann M, Lodish HF. Analysis of receptor signaling pathways by mass spectrometry: identification of vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. Proc Natl Acad Sci U S A. 2000; 97:179–184.

[27] Schmidt SR, Schweikart F, Andersson ME. Current methods for phosphoprotein isolation and enrichment. J Chromatogr B. 2007; 849:154–162.

[28] Alpert AJ, Hudecz O, Mechtler K. Anion-exchange chromatography of phosphopeptides: weak anion exchange versus strong anion exchange and anion-exchange chromatography versus electrostatic repulsion-hydrophilic interaction chromatography. Anal Chem. 2015; 87:4704–4711.

[29] Li Y, Wang Y, Dong M, Zou H, Ye M. Sensitive Approaches for the Assay of the Global Protein Tyrosine Phosphorylation in Complex Samples Using a Mutated SH2 Domain. Anal Chem. 2017; 89:2304–2311.

[30] Bian Y, Li L, Dong M, Liu X, Kaneko T, Cheng K, Liu H, Voss C, Cao X, Wang Y, Litchfield D, Ye M, Li SS-C, Zou H. Ultra-deep tyrosine phosphoproteomics enabled by a phosphotyrosine superbinder. Nat Chem Biol. 2016; 12:959–966.

[31] Hwang L, Ayaz-Guner S, Gregorich ZR, Cai W, Valeja SG, Jin S, Ge Y. Specific enrichment of phosphoproteins using functionalized multivalent nanoparticles. J Am Chem Soc. 2015; 137:2432–2435.

[32] Roberts DS, Chen B, Tiambeng TN, Wu Z, Ge Y, Jin S. Reproducible Large-Scale Synthesis of Surface Silanized Nanoparticles as an Enabling Nanoproteomics Platform: Enrichment of the Human Heart Phosphoproteome. Nano Res. 2019; 12:1473–1481.

[33] Chen B, Hwang L, Ochowicz W, Lin Z, Guardado-Alvarez TM, Cai W, Xiu L, Dani K, Colah C, Jin S, Ge Y. Coupling functionalized cobalt ferrite nanoparticle enrichment with online LC/MS/MS for top-down phosphoproteomics. Chem Sci. 2017; 8:4306–4311.

[34] Wang M-M, Chen S, Yu Y-L, Wang J-H. Novel Ti(4+)-Chelated Polyoxometalate/Polydopamine Composite Microspheres for Highly Selective Isolation and Enrichment of Phosphoproteins. ACS Appl Mater Interfaces. 2019; 11:37471–37478.

[35] 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.

[36] Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics. 2008; 24:2534–2536.

[37] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002; 74:5383–5392.

[38] Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007; 4:207–214.

[39] Yue X, Schunter A, Hummon AB. Comparing multistep immobilized metal affinity chromatography and multistep TiO2 methods for phosphopeptide enrichment. Anal Chem. 2015; 87:8837–8844.

[40] Tsai C-F, Hsu C-C, Hung J-N, Wang Y-T, Choong W-K, Zeng M-Y, Lin P-Y, Hong R-W, Sung T-Y, Chen Y-J. Sequential phosphoproteomic enrichment through complementary metal-directed immobilized metal ion affinity chromatography. Anal Chem. 2014; 86:685–693.

[41] Mathews MB, Hershey JWB. The translation factor eIF5A and human cancer. Biochim Biophys Acta. 2015; 1849: 836–844.

[42] Liang Y, Piao C, Beuschel CB, Toppe D, Kollipara L, Bogdanow B, Maglione M, Lützkendorf J, See JCK, Huang S, Conrad TOF, Kintscher U, Madeo F, Liu F, Sickmann A, Sigrist SJ. eIF5A hypusination, boosted by dietary spermidine, protects from premature brain aging and mitochondrial dysfunction. Cell Rep. 2021; 35:108941.

CHAPTER 4.* Capillary Zone Electrophoresis-High Field Asymmetric Ion Mobility Spectrometry-Tandem Mass Spectrometry for Top-down Characterization of Histone Proteoforms

4.1 Introduction

Histones, serving as the backbone of the nucleosome where the genomic DNA is packed along, are essential for the epigenetic gene regulation and the structural stability of the chromatin [1-6]. Histones consist of two parts: the core histones (H2A, H2B, H3, and H4) as the key components of the nucleosome and the linker histone (H1) as the linkage between nucleosomes. The long N-terminal tails of all four core histones, which protrude from the nucleosome, have a high diversity of post-translational modifications (PTMs) such as methylation, acetylation, phosphorylation, and citrullination [2, 7]. "Histone code" was proposed to state that the PTMs of histone directly contribute to the alternation of chromatin structure and thereby influence DNA transcription [6]. It is crucial to characterize histone PTMs for a better understanding of epigenetic gene regulation. With the diverse PTMs and high heterogeneity of histones, fully deciphering the "histone code" using proteomics is super challenging [8, 9].

Top-down proteomics (TDP), middle-down proteomics (MDP), and bottom-up proteomics (BUP) are three popular ways applied to the decryption of histone code [10-15]. Unlike partial or full digestion of protein into peptides by MDP or BUP, TDP targets intact proteins for the utmost preservation of the PTM information of proteoforms via mass spectrometry (MS) and tandem mass spectrometry (MS/MS) [13, 15, 16]. To delineate rich PTMs of the histone, TDP is definitely the ideal choice among these three methods. In the meantime, TDP requires sufficient separation techniques to improve proteoform detection and identification, especially for complex proteoform mixtures. Reversed-phase liquid chromatography (RPLC)-MS has been frequently used for TDP of histones [17-21]. To advance the separation and throughput of histone variants for TDP, Tian *et al.* reported a weak cation exchange-hydrophilic interaction liquid chromatography (WCX-HILIC) platform for two-dimensional (2D) separations prior to MS analysis, leading to over 700 histone proteoform identifications [22]. Besides LC-MS, capillary zone electrophoresis (CZE)-MS has also been

* *This Chapter is partially adapted with permission from Wang Q, Fang F, Wang Q, Sun L. Capillary zone electrophoresis-high field asymmetric ion mobility spectrometry-tandem mass spectrometry for top-down characterization of histone proteoforms. Proteomics. 2024 Feb;24(3-4):e2200389.*

developed for TDP of histone proteoforms by our lab [11]. CZE provides efficient separation of proteoforms due to PTMs by their charge-to-size ratios and CZE-MS has been proven as a powerful analytical technique for large-scale delineation of proteoforms of histones, bacteria, brains, and human cancer cells [11, 23-27]. Usually, offline LC fractionation is coupled to CZE-MS/MS for TDP of a complex mixture (e.g., histones) to achieve an in-depth proteoform characterization [11]. Development of an online multi-dimensional separation technique involving CZE-MS/MS will be invaluable for TDP of histone proteoforms and complex proteomes in general because of potentially much higher throughput and much less sample loss during sample transfer.

High-field asymmetric waveform ion mobility spectrometry (FAIMS) is drawing more attention as a highly efficient gas-phase online separation technique, which could significantly reduce the background chemical noise, improve sensitivity, and fractionate ions based on their mobility differences under the asymmetric oscillation between high and low electric fields [28-31]. In the FAIMS device, a compensation voltage (CV) is uniquely applied to the inner electrode and amends the trajectory of passing ions based on their mass, charge, and shape. FAIMS has been online coupled with RPLC-MS/MS for TDP to efficiently fractionate proteoform ions based on their masses by applying different CVs [32-35]. Also, Pham *et al.* reported customized tandem nonlinear and linear IMS (FAIMS-TIMS) coupled to MS for bettering the characterization of human histone H2A and H4 proteoforms [36, 37]. Therefore, we expect that coupling FAIMS to CZE-MS/MS to build an online 2D platform will be useful for further advancing the TDP of complex samples, for example, histones.

Here, we present the first example of 2D-CE-FAIMS-MS/MS for TDP of histones. We optimized the CZE separation of histone proteoforms by adjusting the pH of the background electrolyte (BGE) and sample buffer. We achieved nearly 400 and 600 histone proteoform identifications by CZE-FAIMS-MS/MS analyses of a commercial calf histone sample with nine different FAIMS CVs by using two different data analysis tools (ProSight PD and TopPIC Suite).

4.2 Experimental section

4.2.1 Materials and reagents

Histone extract (from calf thymus, Product No. 10223565001) and all chemicals were obtained from Sigma-Aldrich (St. Louis, MO) unless stated otherwise. Acetic acid, formic acid,

methanol, LC/MS grade water, and ammonium hydroxide were ordered from Fisher Chemical (Hampton, New Hampshire). Ammonium acetate (NH_4OAc) was purchased from Invitrogen (Waltham, MA). Acrylamide was purchased from Acros Organics (NJ, USA). Fused silica capillaries (50 μm i.d./360 μm o.d.) were purchased from Polymicro Technologies (Phoenix, AZ).

4.2.2 Capillary coating

The linear polyacrylamide (LPA) coated capillary was prepared according to the previous publications [38, 39]. In brief, a one-meter-long fused silica capillary (50 μm i.d. 360 μm o.d.) was successively flushed by 1 M sodium hydroxide (NaOH), water, 1 M hydrochloric acid (HCl), water, and methanol, followed by overnight nitrogen flow. Then, the capillary was treated with 50% (v/v) 3-(trimethoxysilyl) propyl methacrylate in methanol at room temperature (RT) for 24 hours. Next, the capillary was flushed with methanol and dried under overnight nitrogen flow. For coating the inner wall of the capillary, 500 μL of 4% (w/v) acrylamide in water was mixed with 3.5 μL of 5% (w/v) ammonium persulfate (APS) in water, followed by a 15 min degassing procedure using nitrogen flow. This mixture was then introduced to the capillary using a vacuum suction. Both ends of the capillary were sealed before the incubation in a 50 °C water bath for 1 hour. At last, the capillary was flushed with water to remove any residue reactants and kept at RT before use.

4.2.3 Sample preparation

The histone extract was dissolved in 50 mM NH_4OAc (pH 6.5 or pH 9.0) to prepare 2 mg/mL histone samples for CZE-MS/MS analysis.

4.2.4 Optimization of CZE conditions for CZE-MS/MS of histones

CZE-MS/MS platform was built up by connecting a CESI 8000 Plus CE system (Sciex) to an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific) with an in-house constructed electrokinetically pumped sheath-flow nano-electrospray ionization (nanoESI) interface [40, 41]. A glass electrospray emitter with orifice size ranging from 30-35 μm was pulled using a Sutter P-1000 flaming/brown micropipette puller, and the sheath liquid contained 0.2% (v/v) formic acid and 10% (v/v) methanol in water. The electrospray voltage was set at 2.2-2.4 kV to the sheath liquid reservoir for ionization. The inlet of the capillary was installed in the cartridge of the CE system and the outlet was fit into the glass electrospray emitter (around 0.5 mm to the emitter tip), following the previous procedures [41].

For the mass spectrometer settings, the ion transfer tube temperature was set to 320 °C, and the RF lens was 60%. The application mode was set to intact protein mode with low C-trap pressure. The isolation window was 0.7 *m/z* for isolating parent ions for high-energy collision dissociation (HCD). The normalized collision energy (NCE) of HCD was 25%. The MS/MS experiments were performed using data-dependent acquisition (DDA). Full MS scan was performed with the following parameters: orbitrap resolution of 480,000 (at *m/z* of 200), *m/z* range of 300-2000, normalized AGC target of 300%, microscans of 1. The top 6 most intense precursors with charge states in the range of 5-60 in full MS spectra were isolated and fragmented, and the threshold of precursors was set at 10,000. Other parameters for MS/MS include the resolution of 120,000 (at *m/z* 200), *m/z* range of 100-1500, microscans of 3, normalized AGC target of 100%, auto maximum injection time and dynamic exclusion of 30 s.

To achieve a better histone proteoform separation for more proteoform identifications, we optimized the BGE of CZE. Two BGEs were evaluated, and they were 5% (v/v) acetic acid (pH 2.4) and 20 mM NH_4OAc (pH 5.0). For the BGE 5% (v/v) acetic acid (pH 2.4), we used a sample buffer of 50 mM NH_4OAc (pH 6.5) for dynamic pH junction sample stacking [39]. For the 20 mM NH_4OAc BGEs (pH 5.0 achieved by adding acetic acid), we chose the 50 mM NH_4OAc (pH 9) as the sample buffer to maintain the pH difference between sample buffer and BGE for efficient dynamic pH sample stacking. A 1-meter-long LPA-coated separation capillary (50 µm i.d./360 µm o.d.) was used for the project. The histone sample was injected under 5 psi for 5s to introduce about 25 nL for each run (50 ng loaded). Then, a 30 kV voltage was applied for separation. After the separation, 30 kV voltage and 15 psi pressure were applied for 10 min to clean up the capillary.

4.2.5 CZE-FAIMS-MS/MS

All CZE separation conditions of histones and mass spectrometer settings were the same as those mentioned above unless stated otherwise. For the CZE separation, the BGE was 20 mM NH_4OAc (pH 5.0). The sample injection volume was about 25 nL and about 50-ng histone was loaded for each run. The separation voltage was 30 kV, and the separation time was 50 min.

For the FAIMS fractionation, the FAIMS Pro Duo interface (Thermo Fisher Scientific) was installed prior to the mass spectrometer. After the auto DV tune, the FAIMS Pro Duo interface was set to standard resolution and the nitrogen carrier gas was set as default (4.6 L/min). Different CV voltages (-60V, -50V, -40V, -30V, -20V, -10V, +10V, +20V, and +30V)

were individually tested for triplicate CZE-FAIMS-MS/MS runs to examine the fractionation performance of the FAIMS.

For the mass spectrometer settings, intact protein mode, low C-trap pressure, and isolation window 0.7 *m/z* for MS/MS were employed. 28% HCD energy was applied for FAIMS CV ranging from -60 V to -40 V, and 30% HCD energy was applied for FAIMS CV ranging from -30 V to -10 V, and 35% HCD energy was applied for FAIMS CV ranging from +10 V to +30 V. The top 6 most intense precursors with charge states in the range of 5-60 in full MS spectra were isolated and fragmented for FAIMS CV ranging from -50 V to +30 V, and precursor charge states in the range of 3-60 were selected for FAIMS CV at -60 V.

4.2.6 Data analysis

Proteome Discoverer 2.2 software (Thermo Fisher Scientific) with the ProSightPD 1_1 node for TDP was used for database search [42]. The detailed database searching setup was the same as our previous work [11]. Briefly, the MS1 spectra were first averaged using the cRAWler algorithm in Proteome Discoverer. The precursor *m/z* tolerance was set to 0.2 *m/z*. For both precursor and fragmentation Xtract parameters, the signal-to-noise ratio threshold, the lowest and the highest *m/z* were set to 3, 200, and 4000, respectively. Then deconvolution was performed by the Xtract algorithm followed by database searching against a Bos taurus database (downloaded from http://proteinaceous.net/-database-warehouse-legacy/ in April 2018). A three-prone database searching was performed: (1) a search was performed with a 2-Da and 10-ppm mass tolerance of absolute mass for MS1 and MS2, respectively; (2) a subsequent biomarker search was performed to find unreported truncated proteoforms with 10 ppm tolerance for both MS1 and MS2; (3) the last search was performed with a 1000-Da mass tolerance for MS1 and a 10-ppm mass tolerance for MS2 for matching unexpected PTMs. The target-decoy strategy was exploited for evaluating the false discovery rates (FDRs) [43, 44]. FDR estimation was performed for each of the three search strategies. The identified proteoform-spectrum matches (PrSMs) and proteoforms were filtered using a 1% FDR.

For the CZE-FAIMS-MS/MS data, the raw files for FAIMS CV ranging from -60 V to -10 V were searched by ProSightPD, and the raw files from CV ranging from +10 V to +30 V were analyzed with the TopPIC (Top-down mass spectrometry-based proteoform identification and characterization) software (version 1.6.2) [45]. The raw files were firstly converted to mzML files with the MsConvert software [46], and spectral deconvolution was performed with the

TopFD (Top-down mass spectrometry feature detection) software, generating msalign files, which were used as the input for database searching using TopPIC. The spectra were searched against a Bos taurus database (downloaded from Swiss-Uniprot, March 2022). FDRs were estimated using the target-decoy approach [43, 44]. A 1% PrSM-level FDR and a 5% proteoform-level FDR were employed to filter the identifications. The mass error tolerance was 15 ppm. The mass error tolerance was 1.2 Da for identifying PrSM clusters. The maximum mass shift was 500 Da. The maximum number of mass shift was set to 2.

All the CZE-MS/MS and CZE-FAIMS-MS/MS data were further analyzed by the TopPIC software (version 1.6.2) [45]. The parameters were the same as previously described except for several differences. A 1% PrSM-level FDR and a 1% proteoform-level FDR were employed to filter the identifications. The identified proteoforms were further filtered by the E value lower than 0.001. The maximum variable PTM number was set to 5.

4.2.7 Data analysis

We followed the procedure in our previous work for calculating the experimental and predicted μ_{ef} [11]. The experimental μ_{ef} was calculated by **eq 1**,

$$\text{experimental } \mu_{ef} = L / ((30 - 2) / L \times t_M) (\text{unit of } cm^2\ kV^{-1}s^{-1}) \quad (1)$$

where L is the capillary length in cm, 30 and 2 are the separation voltage, and electrospray voltage in kV. The **eq 1** is obtained from the literature [47, 48]. The predicted μ_{ef} was calculated by **eq 2**,

$$\text{predicted } \mu_{ef} = \ln(1 + 0.350 \times Q) / M^{0.411} \quad (2)$$

where Q is the number of charges of the proteoform in the BGE by counting the number of positively charged amino acid residues in the proteoform sequence (K, R, H, and N-terminus). M is the molecular mass obtained by MS measurement in Da. The **eq 2** is obtained based on previous publications [47-49].

4.3 Results and discussion

4.3.1 Optimization of CZE for better separation and identification of histone proteoforms

CZE separates histone proteoforms according to their electrophoretic mobilities, which relate to their charge-to-size ratios. Histones are super basic and are highly positively charged under our typical CZE BGE condition (i.e., 5% (v/v) acetic acid (~pH 2.4)) for TDP [50]. The pH of BGE will influence the charge of histone proteoforms and impact the CZE separations. A BGE with a higher pH value decreases the charge of histone proteoforms, resulting in potentially

bigger differences in charge-to-size ratios of histone proteoforms, which eventually leads to better separation resolution and more histone proteoform identifications. To test our hypothesis, we studied two different BGEs: 5% (v/v) acetic acid (~pH 2.4) and 20 mM NH_4OAc (pH 5.0). We maintained the same sample injection volume and protein injection amount for the two conditions (25 nL and 50 ng). To maintain a sufficient pH difference between the sample buffer and BGE for dynamic pH junction sample stacking, we employed 50 mM NH_4OAc (pH 9) as the sample buffer for the BGE of 20 mM NH_4OAc (pH 5.0). For the 5% (v/v) acetic acid (~pH 2.4) BGE, the sample buffer was 50 mM NH_4OAc (pH 6.5).

As shown in **Figure 4.1**, CZE-MS/MS using a BGE pH 5.0 produced a substantially wider separation window and better resolution for histone proteoforms compared to a BGE pH 2.4. The separation window of histone proteoforms was about 3 minutes for the BGE pH 2.4 and about 9 minutes for the BGE pH 5.0. Histone H2A and H2B co-migrated under the pH 2.4 BGE condition, agreeing well with our previous data [11]. Interestingly, Histone H2A and H2B were well separated using a BGE pH of 5.0. Due to the much better separation and dramatically wider separation window, CZE-MS/MS using a BGE pH 5.0 identified 60% (TopPIC) or 85% (ProSight PD) more proteoforms that are larger than 10 kDa than that using a BGE pH 2.4 in triplicate runs, **Figure 4.1**. CZE-MS/MS with a BGE pH 5.0 produced reproducible separations of histone proteoforms in terms of separation profiles and proteoform intensity, **Figure 4.2**. Considering the overall number of large intact histone proteoforms (over 10 kDa), the separation profiles, and proteoform intensity, the BGE pH 5.0 was used for all the following experiments. We observed that the histone proteoforms from pH 2.4 and 5 were substantially different, and only 16% of identified histone proteoforms were shared, **Figure 4.3.** We need to point out that a decrease in separation voltage is another potential way to increase the separation window for histone proteoforms, but this approach will most likely reduce the separation efficiency.

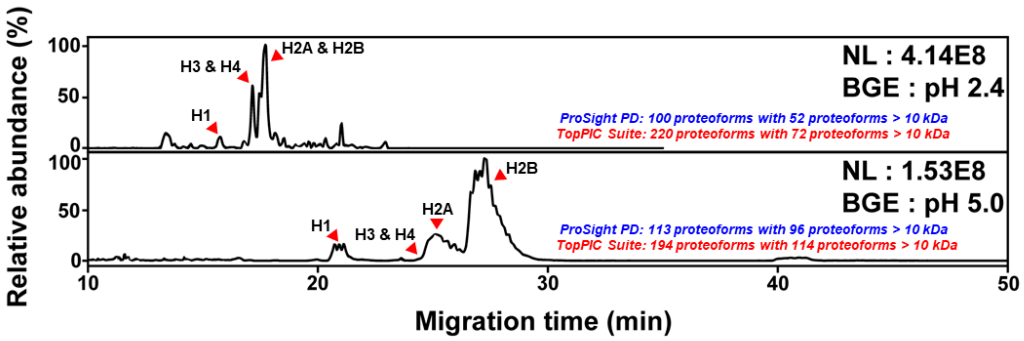


Figure 4.1. Electropherograms of CZE-MS/MS analysis of histone proteoforms under different BGE conditions. Two different BGEs with pH 2.4 (5% (v/v) acetic acid) and pH 5.0 (20 mM NH_4OAc by adding acetic acid to achieve the pH) were studied. The peaks of H1, H2A, H2B, H3, and H4 are marked with red arrows. The total number of proteoform identifications from ProSightPD and TopPIC Suite, and the number of proteoforms larger than 10 kDa from triplicate analyses are labeled.

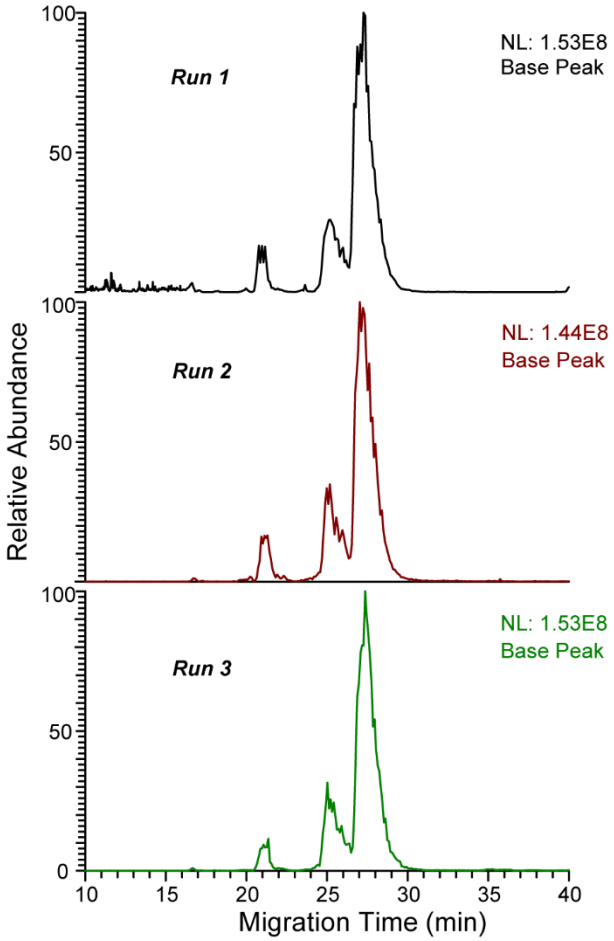


Figure 4.2. Base peak electropherograms of the calf histone sample were analyzed by CZE-MS/MS in triplicate runs using the BGE pH 5.0 (20 mM NH4OAc by adding acetic acid to achieve the pH).

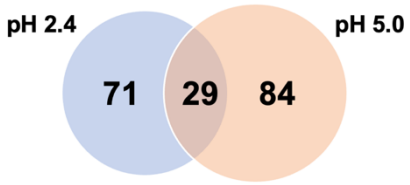


Figure 4.3. The overlap of identified histone proteoforms between two different BGEs: pH 2.4 (5% (v/v) acetic acid) and pH 5.0 (20 mM NH_4OAc by adding acetic acid to achieve the pH). The histone proteoform data here is from the analysis by ProSightPD.

4.3.2 CZE-FAIMS-MS/MS as an online two-dimensional technique for the characterization of histone proteoforms

Because of the extreme complexity of histone proteoforms, multi-dimensional (MD) separations are crucial for delineating histone proteoforms. Here, we integrated FAIMS into the CZE-MS/MS system to carry out additional proteoform separations in the gas phase based on the ion mobility principle between liquid-phase CZE and gas-phase MS separations. After initial liquid phase CZE separation, histone proteoforms are further online fractionated in the gas phase by FAIMS based on their charges and sizes prior to MS and MS/MS. For FAIMS fractionation, nine different CVs ranging from -60 V to +30 V with 10-V increments were studied (triplicate runs for each CV).

As shown in **Figure 4.4A**, each CV displays its unique histone separation profile. Main peaks of H2A and H2B were detected from -40 V to -10 V and from -50 V to -20 V CVs, respectively. Interestingly, the main peak of H1 emerged at CV of -10 V and became the only one at CV of +20 V and +30 V. This is consistent with the database search result using the TopPIC. Almost all the identified proteoforms at CV of +20 V and +30 V were from histone H1. Because the size of H1 (> 20 kDa) is larger than that of H2A and H2B, this phenomenon is in agreement with the literature that protein ions are fractionated by FAIMS according to their masses [34, 35].

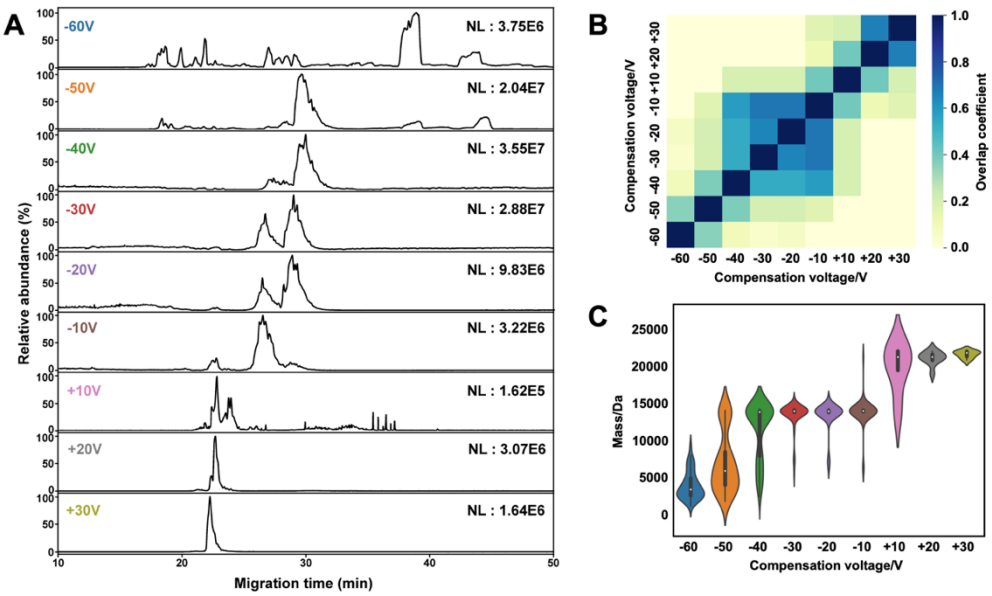


Figure 4.4. (A). Electropherograms of CZE-FAIMS-MS/MS analysis of histone proteoforms under different CVs. CV values from -60 V to +30 V are listed from top to bottom. (B). Overlap of identified proteoforms of histones between FAIMS CVs. (C). Violin plots of mass distributions of identified histone proteoforms by different FAIMS CVs.

Totally, we identified 366 (from ProSight PD) and 602 (from TopPIC Suite) histone proteoforms with the combination of 9 CVs by CZE-FAIMS-MS/MS, and the number of histone proteoforms is improved by about 3 folds compared to that from CZE-MS/MS alone (without FAIMS) (366 *vs.* 113 proteoforms from ProSightPD, 602 *vs.* 194 proteoforms from TopPIC Suite). We previously coupled size-exclusion chromatography (SEC) to CZE-MS/MS for TDP of histones with the identification of about 400 histone proteoforms from the same calf histone sample [11]. Both offline 2D-SEC-CZE-MS/MS and online 2D-CZE-FAIMS-MS/MS are efficient for histone proteoform characterization. The unique advantage of online 2D-CZE-FAIMS-MS/MS is the much lower requirement for initial histone material compared to offline 2D platforms. The offline 2D-SEC-CZE-MS/MS used hundreds of micrograms of protein material to start the analysis and online 2D-CZE-FAIMS-MS/MS only required less than 10 µg of histone material to initiate and complete the analyses because it avoided any potential sample loss due to, e.g., LC fraction collection and sample transfers. We expect the online 2D-CZE-FAIMS-MS/MS will be a powerful tool for TDP of mass-limited biological samples.

To further investigate the histone proteoform fractionation performance of FAIMS, we studied the histone proteoform overlaps between different CVs, **Figure 4.4B**. The proteoform overlap coefficients between any two different CVs became smaller when the CV difference

increased. For example, the proteoform overlap coefficient was close to 0.4 between -60 V and -50 V CVs; the coefficient was reduced to lower than 0.2 between -60 V and -40 V CVs. The data suggests that FAIMS can perform proteoform fractionation efficiently in the gas phase. Next, the violin plots in **Figure 4.4C** show the mass distributions of histone proteoforms identified under each CV condition. It is clear that CZE-FAIMS-MS/MS with a larger CV tends to identify histone proteoforms with higher masses. The median mass of identified histone proteoforms increased from 3.4 kDa to 21.3 kDa when the CV was enlarged from -60 to +30 V. The results indicate that the CV value of FAIMS and proteoform mass have a clear correlation in our CZE-FAIMS-MS/MS condition. The histone proteoform overlaps between different CVs and proteoform mass distributions across different CVs from TopPIC Suite were in consistent with that from ProSightPD, **Figure 4.5**.

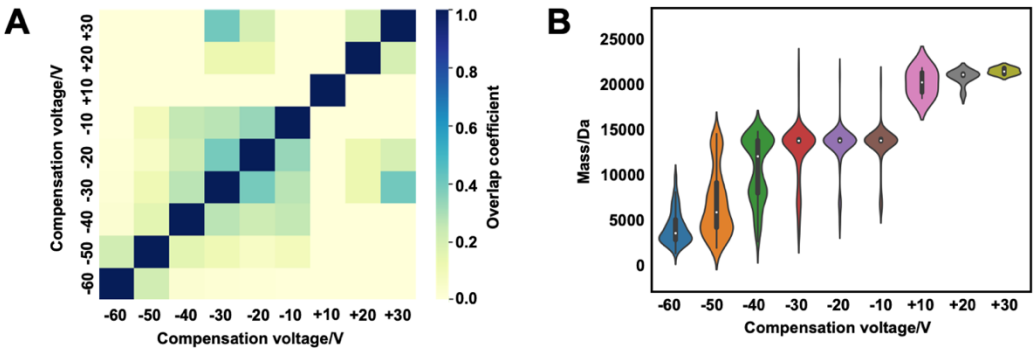


Figure 4.5. (A). Overlap of identified proteoforms of histones between FAIMS CVs. (B). Violin plots of mass distributions of identified histone proteoforms by different FAIMS CVs. The histone proteoform data here is from the analysis by TopPIC Suite.

Besides global analyses of the histone proteoform data from CZE-FAIMS-MS/MS, we also tried to investigate the performance of FAIMS for separations of near isobaric histone proteoforms. For example, we found that the H2B type 1-N (Protein Accession: Q32L48) had two nearly isobaric proteoforms well separated by FAIMS. The one with two acetylation (Theo. MH+: 13868.4992 Da) was only identified by -40-V CV, while the one with one phosphorylation (Theo. MH+: 13864.4444 Da) was only identified by -20-V CV. The data suggests that CZE-FAIMS-MS/MS is promising for the characterization of isobaric or nearly isobaric histone proteoforms, which will be further systematically studied in our future work.

4.3.3 Electrophoretic mobility prediction of histone proteoforms

Recently, our lab published the first examples of accurately predicting proteoforms' μ_{ef} by optimized semiempirical models using large-scale CZE−MS/MS datasets of *E. coli*, zebrafish

brain, plant leaf, and calf histone samples [11, 47, 51]. This new approach could help validate the confidence of proteoform identifications by examining the correlation between predicted and experimental μ_{ef} of proteoforms. The previous works employed 5% (v/v) acetic acid (~pH 2.4) as the BGE in CZE−MS/MS analysis for denaturing histone proteoforms. Here, we employed an optimized BGE (20 mM NH_4OAc, pH 5.0) to better separations of histone proteoforms. Under the pH 5.0 condition, the histone proteoforms most likely tend to carry fewer positive charges and unfold less compared to the pH 2.4 condition, which could substantially influence the μ_{ef} prediction of histone proteoforms by the semi-empirical model used in our previous studies [11, 47]. Here we further studied the μ_{ef} prediction of histone proteoforms under the two BGE conditions, pH 2.4 and pH 5.0. The details of calculating the predicted and experimental μ_{ef} of the proteoforms were described in the ***Data analysis*** part. Only histone proteoforms without unknown modifications were used in this study. Both CZE-MS/MS datasets from BGE 2.4 and 5.0 without FAIMS were utilized.

For the CZE-MS/MS dataset from the BGE pH 2.4, we first optimized the prefactor of 'Q' and the power factor of 'M' in ***eq 2*** using proteoforms without any PTMs by independent adjustment of the factors. We found that 0.218 as the prefactor of 'Q' and 0.411 as the power factor of 'M' produced the best linear correlation coefficient (R^2) as 0.9824, **Figure 4.6A**. Then, the optimized μ_{ef} prediction equation was used for histone proteoforms with PTMs (i.e., acetylation and phosphorylation). Proteoform acetylation and phosphorylation were reported to reduce the charge (Q) by roughly one unit according to our previous studies [47]. Some of the histone proteoforms with acetylation and/or phosphorylation were clearly off the trendline without charge Q corrections, **Figure 4.6B**. After we applied charge Q reduction for those acetylated and/or phosphorylated proteoforms, the linear correlation coefficient between predicted and experimental μ_{ef} increased from 0.9058 to 0.9548, **Figure 4.6C**. The nice linear correlations between experimental and theoretical μ_{ef} of identified histone proteoforms suggest high-confidence proteoform identifications in this study, which agrees with the P-Score distribution of proteoforms, **Figure 4.6D.** The histone proteoforms identified by CZE-MS/MS have P-scores centering around 10^{-20}. The low P-score value indicates confident proteoform identifications [50].

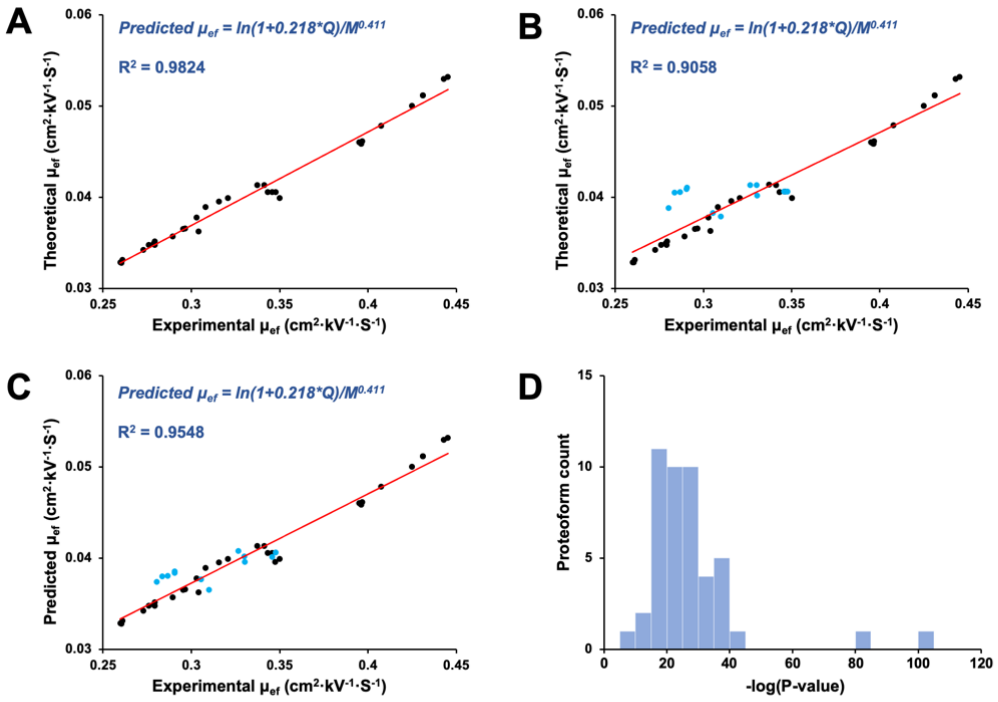


Figure 4.6. Linear correlations between theoretical and experimental μ_{ef} of unmodified histone proteoforms (A) and unmodified plus phosphorylated and acetylated proteoforms (B) identified in a single CZE-MS/MS analysis under the BGE of 5% (v/v) acetic acid (pH 2.4). The black dots represent proteoforms without PTMs and the light blue dots represent proteoforms with phosphorylation and/or acetylation. (C). Linear correlations between theoretical and experimental μ_{ef} of unmodified histone proteoforms plus phosphorylated and acetylated proteoforms after charge Q corrections due to PTMs. (D). Distribution of the −log(P-score) of the identified histone proteoforms by CZE-MS/MS.

We further studied the CZE-MS/MS dataset from the BGE pH 5.0, **Figure 4.7**. After optimizations, the best linear correlation coefficient (R^2=0.9417) for histone proteoforms without PTMs was obtained using a prefactor of 'Q' as 0.277, **Figure 4.7**, which is substantially different from that for the BGE pH 2.4. The best linear correlation coefficient under BGE pH 5.0 is significantly lower than that under BGE 2.4 (0.9417 vs. 0.9824). We also explored the CZE-MS/MS dataset by TopPIC Suite from both BGE pH 2.4 and pH 5.0, **Figures 4.8** and **4.9**. After applying charge Q reduction for acetylated and/or phosphorylated proteoforms, the linear correlation coefficient between predicted and experimental μ_{ef} increased from 0.8538 to 0.8963 at pH 2.4, while the linear correlation coefficient at pH 5.0 almost had no changes (from 0.8774 to 0.8792). We suspected that the rise of BGE pH to 5.0 led to a more folded condition of histone proteoforms and made accurate charge and size calculations more difficult, resulting in a lower correlation coefficient.

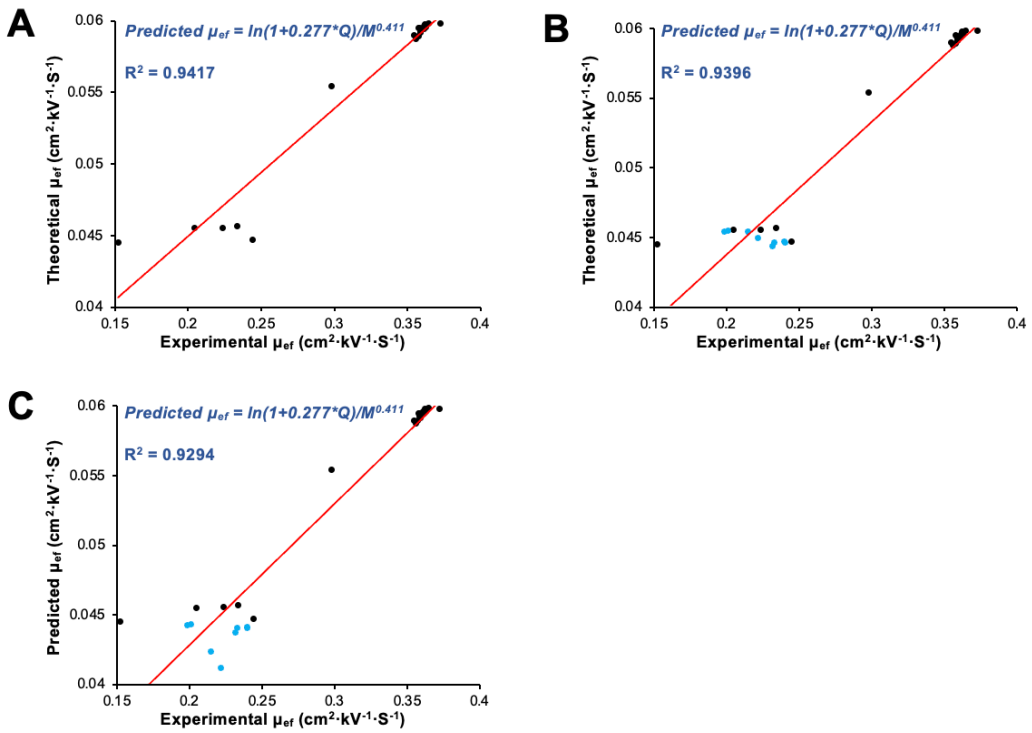


Figure 4.7. Linear correlations between theoretical and experimental μ_{ef} of unmodified histone proteoforms (A) and unmodified plus phosphorylated and acetylated proteoforms (B) identified in a single CZE-MS/MS analysis under the BGE of 20mM NH_4OAc (pH 5.0 achieved by adding acetic acid). The black dots represent proteoforms without PTMs and the light blue dots represent proteoforms with phosphorylation and/or acetylation. (C). Linear correlations between theoretical and experimental μ_{ef} of unmodified histone proteoforms plus phosphorylated and acetylated proteoforms after charge Q corrections due to PTMs. The histone proteoform data used here is from ProSightPD.

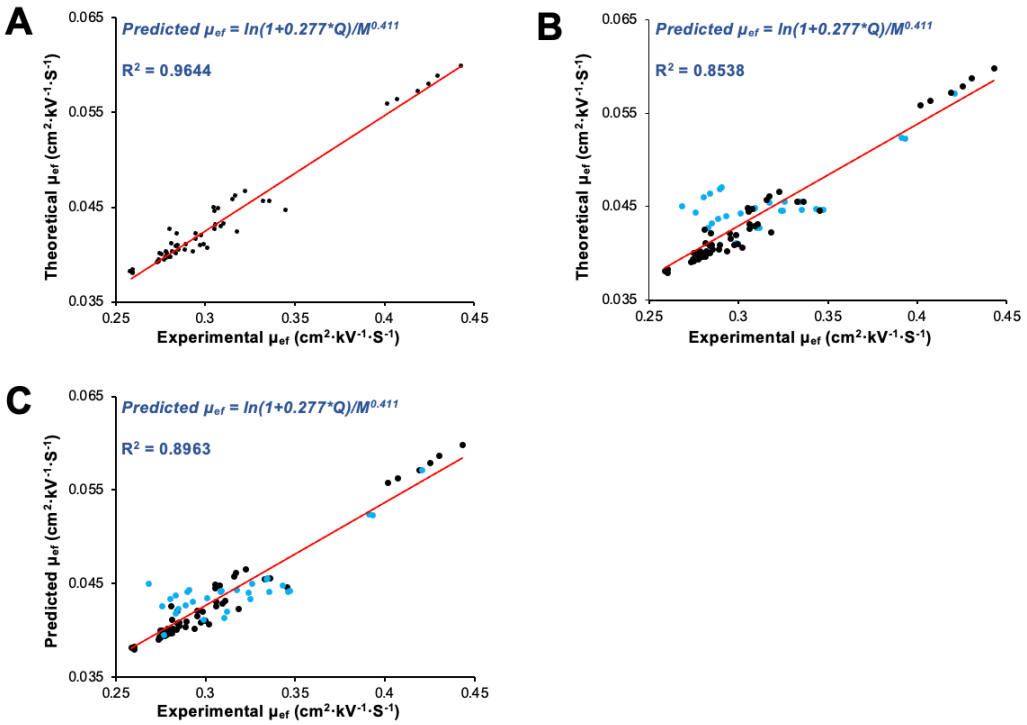


Figure 4.8. Linear correlations between theoretical and experimental μ_{ef} of unmodified histone proteoforms (A) and unmodified plus phosphorylated and acetylated proteoforms (B) identified in a single CZE-MS/MS analysis under the BGE of 5% (v/v) acetic acid (pH 2.4). The black dots represent proteoforms without PTMs and the light blue dots represent proteoforms with phosphorylation and/or acetylation. (C). Linear correlations between theoretical and experimental μ_{ef} of unmodified histone proteoforms plus phosphorylated and acetylated proteoforms after charge Q corrections due to PTMs. The histone proteoform data here is from the analysis by TopPIC Suite.

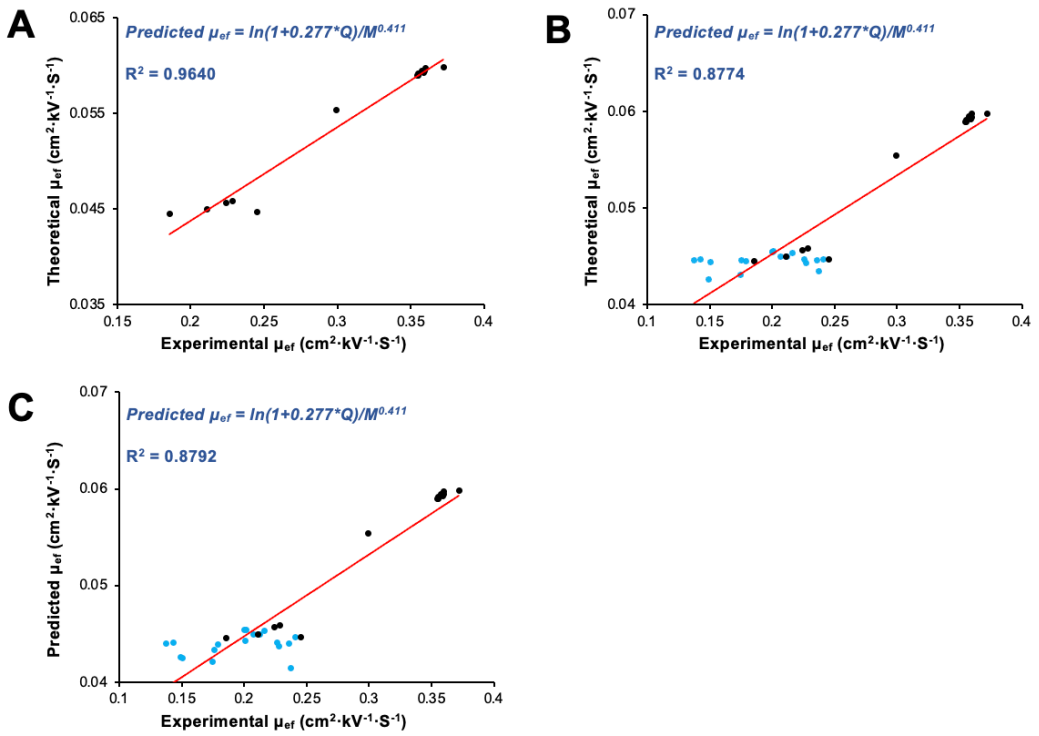


Figure 4.9. Linear correlations between theoretical and experimental μ_{ef} of unmodified histone proteoforms (A) and unmodified plus phosphorylated and acetylated proteoforms (B) identified in a single CZE-MS/MS analysis under the BGE of 20mM NH_4OAc (pH 5.0 achieved by adding acetic acid). The black dots represent proteoforms without PTMs and the light blue dots represent proteoforms with phosphorylation and/or acetylation. (C). Linear correlations between theoretical and experimental μ_{ef} of unmodified histone proteoforms plus phosphorylated and acetylated proteoforms after charge Q corrections due to PTMs. The histone proteoform data here is from the analysis by TopPIC Suite.

The results here demonstrate that the pH of BGE can strongly affect the prediction of histone proteoforms' μ_{ef}. To achieve better μ_{ef} prediction of histone proteoforms under BGE pH 5.0, more efforts need to be made regarding the collection of much larger histone proteoform datasets and more systematic investigations of factors that potentially influence the charge and size of histone proteoforms.

4.3.4 Histone PTMs

The individual PTMs and the combination of diverse PTMs located on histone proteoforms are critical for the epigenetic control of gene expression. Some examples of histone proteoforms with PTMs identified in our CZE-FAIMS-MS/MS study by the TopPIC software are shown in **Figure 4.10**. All four examples of histone proteoforms were identified with multi-PTMs under high confidence, less than 10-ppm mass errors, and better than 1×10^{-13} E-values.

Figure 4.10A and **4.10B** show proteoforms of H2B type 1K and H2B type 1, while **Figure 4.10C** and **4.10D** display two proteoforms from H2A type 1. The proteoform in **Figure 4.10A** has four PTMs at S6 (phospho), K15 (methyl), R86 to R92 (citrullination), L101 to A110 (methyl). The phosphorylation of H2B at S6 was reported to occur during the early mitotic phases and may prevent chromosomal instability and aneuploidy [52]. The proteoform in **Figure 4.10B** has one phosphorylation at the S14 and one citrullination at Q22. The phosphorylation of H2B at S14 was associated with the apoptotic chromatin condensation pathway and regulation of monoubiquitination of H2B [53, 54]. The proteoform in **Figure 4.10C** has N-terminal acetylation at S1, one citrullination at R3, and another citrullination between N68 and N89. The citrullination of histones by PADs (protein arginine deiminases) was reported to be correlated with both transcriptional activation and repression [55]. The proteoform in **Figure 4.10D** carries N-terminal acetylation and one phosphorylation at T101. Phosphorylation of H2A at T101 may play a crucial role in creating distinctive binding sites for DNA double-strand break (DSB) response proteins, and lead to alterations in the local chromatin structure [56]. We noted that ProSight PD also identified similar histone proteoforms to that shown in **Figures 4.10B** and **4.10C**. ProSight PD did not identify the proteoforms shown in **Figures 4.10A** and **4.10D**, which is most likely due to the fact that TopPIC and ProSight PD employ drastically different database search strategies.

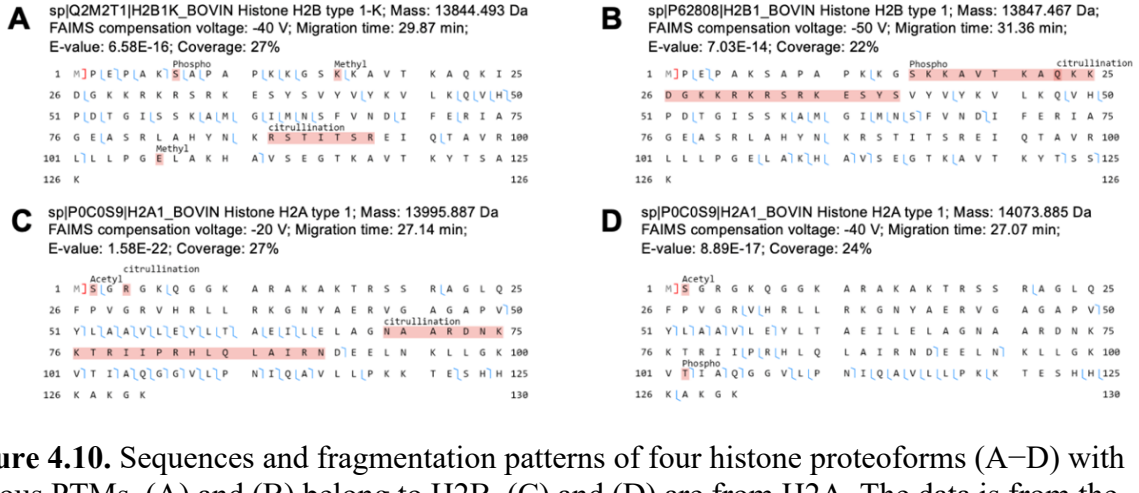


Figure 4.10. Sequences and fragmentation patterns of four histone proteoforms (A−D) with various PTMs. (A) and (B) belong to H2B. (C) and (D) are from H2A. The data is from the TopPIC Suite.

We further compared some identified histone proteoforms carrying PTMs in this work with that identified in one recent study from the Brodbelt group, which employed ultraviolet

photodissociation (UVPD) for TDP of the same calf histone sample as our study [57]. Several histone proteoforms with PTMs were identified with high fragmentation coverage by ProSight PD and highlighted in the Brodbelt group's study. The first one was histone H4 carrying N-terminal acetylation, R3 dimethylation, and K12 acetylation (acH4R3me2K12ac). The second one was histone H4 with N-terminal acetylation and R3 dimethylation (acH4R3me2). The third one was histone H2A with N-terminal acetylation (acH2A). In our study, we also identified similar histone proteoforms by ProSight PD, **Figures 4.11-4.13**. We identified three possible H4 proteoforms, acH4R3me2K5ac, acH4R3me2K8ac, and acH4R3me2K12ac, **Figure 4.11**. Due to the limited backbone cleavage at the N-terminus, we can't distinguish those three H4 proteoforms. **Figure 4.12** shows another possible H4 proteoform, acH4R3me2, identified in this study. We noted that there are other possible explanations regarding the PTMs in **Figure 4.12** due to the limited fragmentation coverage at the N-terminus. **Figure 4.13** shows a high-confidence identification of one H2A proteoform, acH2A, in this work. The data demonstrate that our histone proteoform data and the Brodbelt group's data agree reasonably well regarding the specific histone H4 and H2A proteoforms discussed above.

Figure 4.11. Sequences and fragmentation patterns of three possible histone H4 proteoforms identified by ProSight PD from our CZE-FAIMS-MS/MS data. (A). acH4R3me2K5ac; (B). acH4R3me2K8ac; (C). acH4R3me2K12ac. PCS (Proteoform characterization score) and P-Scores are labeled.

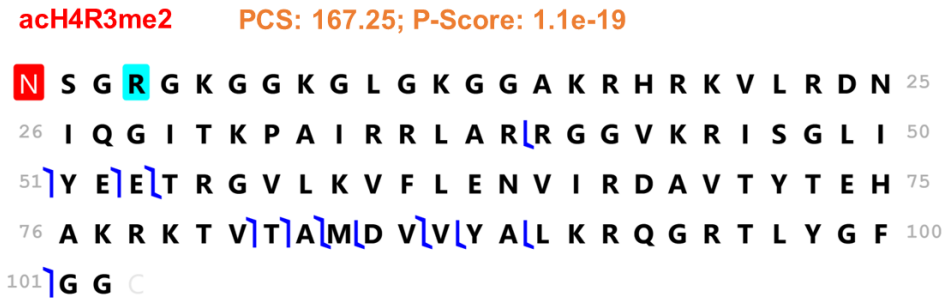


Figure 4.12. Sequence and fragmentation pattern of one possible histone H4 proteoform (acH4R3me2) identified by ProSight PD from our CZE-FAIMS-MS/MS data. PCS (Proteoform characterization score) and P-Score are labeled.

Figure 4.13. Sequence and fragmentation pattern of one possible histone H2A proteoform (acH2A) identified by ProSight PD from our CZE-FAIMS-MS/MS data. PCS (Proteoform characterization score) and P-Score are labeled.

4.4 Conclusion

We presented the first example of coupling CZE, IMS, and MS as a multi-dimensional platform for characterization of histone proteoforms with the identification of 366 histone proteoforms (from ProSightPD) and 602 histone proteoforms (from TopPIC Suite) using a low microgram amount of histone sample as the starting material. We revealed that the pH of BGE could affect the CZE separation and μ_{ef} prediction of histone proteoforms substantially. We documented that FAIMS is an efficient gas-phase separation method for histone proteoforms and can fractionate histone proteoforms according to their masses.

One limitation of our current CZE-FAIMS-MS/MS platform is the low backbone cleavage coverage for histone proteoforms with HCD fragmentation, which impedes accurate localizations of PTMs on histone proteoforms. We expect that the integration of alternative gas-phase fragmentation techniques like UVPD [57], electron capture dissociation [58-61], and electron transfer dissociation [62, 63] will drastically benefit the characterization of histone proteoforms.

4.5 Acknowledgment

We thank the support from the National Cancer Institute through Grant R01CA247863, and National Institute of General Medical Sciences (NIGMS) through Grants R01GM125991 and 2R01GM118470. We also thank the support from the National Science Foundation through Grant DBI1846913 (CAREER Award).

REFERENCES

[1] Badeaux AI, Shi Y. Emerging roles for chromatin as a signal integration and storage platform. Nat Rev Mol Cell Biol. 2013 Apr;14(4):211-24.

[2] Kimura H. Histone modifications for human epigenome analysis. J Hum Genet. 2013 Jul;58(7):439-45.

[3] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74.

[4] Woodcock CL, Skoultchi AI, Fan Y. Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. Chromosome Res. 2006;14(1):17-25.

[5] Kimura H. Histone dynamics in living cells revealed by photobleaching. DNA Repair (Amst). 2005 Jul 28;4(8):939-50.

[6] Jenuwein T, Allis CD. Translating the histone code. Science. 2001 Aug 10;293(5532):1074-80.

[7] Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nat Rev Genet. 2016 Aug;17(8):487-500.

[8] Zhao Y, Garcia BA. Comprehensive Catalog of Currently Documented Histone Modifications. Cold Spring Harb Perspect Biol. 2015 Sep 1;7(9):a025064.

[9] Brown DT. Histone variants: are they functionally heterogeneous? Genome Biol. 2001;2(7):REVIEWS0006.

[10] Moradian A, Kalli A, Sweredoski MJ, Hess S. The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications. Proteomics. 2014 Mar;14(4-5):489-97.

[11] Chen D, Yang Z, Shen X, Sun L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry As an Alternative to Liquid Chromatography-Tandem Mass Spectrometry for Top-down Proteomics of Histones. Anal Chem. 2021 Mar 16;93(10):4417-4424.

[12] Zheng Y, Fornelli L, Compton PD, Sharma S, Canterbury J, Mullen C, Zabrouskov V, Fellers RT, Thomas PM, Licht JD, Senko MW, Kelleher NL. Unabridged Analysis of Human Histone H3 by Differential Top-Down Mass Spectrometry Reveals Hypermethylated Proteoforms from MMSET/NSD2 Overexpression. Mol Cell Proteomics. 2016 Mar;15(3):776-90.

[13] Shliaha PV, Gorshkov V, Kovalchuk SI, Schwämmle V, Baird MA, Shvartsburg AA, Jensen ON. Middle-Down Proteomic Analyses with Ion Mobility Separations of Endogenous Isomeric Proteoforms. Anal Chem. 2020 Feb 4;92(3):2364-2368.

[14] Sidoli S, Lin S, Karch KR, Garcia BA. Bottom-up and middle-down proteomics have comparable accuracies in defining histone post-translational modification relative abundance and stoichiometry. Anal Chem. 2015 Mar 17;87(6):3129-33.

[15] Zheng Y, Huang X, Kelleher NL. Epiproteomics: quantitative analysis of histone marks and codes by mass spectrometry. Curr Opin Chem Biol. 2016 Aug;33:142-50.

[16] Toby TK, Fornelli L, Kelleher NL. Progress in Top-Down Proteomics and the Analysis of Proteoforms. Annu Rev Anal Chem (Palo Alto Calif). 2016 Jun 12;9(1):499-519.

[17] Su X, Jacob NK, Amunugama R, Lucas DM, Knapp AR, Ren C, Davis ME, Marcucci G, Parthun MR, Byrd JC, Fishel R, Freitas MA. Liquid chromatography mass spectrometry profiling of histones. J Chromatogr B Analyt Technol Biomed Life Sci. 2007 May 1;850(1-2):440-54.

[18] Zhou Y, Zhang X, Fornelli L, Compton PD, Kelleher N, Wirth MJ. Chromatographic efficiency and selectivity in top-down proteomics of histones. J Chromatogr B Analyt Technol Biomed Life Sci. 2017 Feb 15;1044-1045:47-53.

[19] Pesavento JJ, Bullock CR, LeDuc RD, Mizzen CA, Kelleher NL. Combinatorial modification of human histone H4 quantitated by two-dimensional liquid chromatography coupled with top down mass spectrometry. J Biol Chem. 2008 May 30;283(22):14927-37.

[20] Contrepois K, Ezan E, Mann C, Fenaille F. Ultra-high performance liquid chromatography-mass spectrometry for the fast profiling of histone post-translational modifications. J Proteome Res. 2010 Oct 1;9(10):5501-9.

[21] Holt MV, Wang T, Young NL. High-Throughput Quantitative Top-Down Proteomics: Histone H4. J Am Soc Mass Spectrom. 2019 Dec;30(12):2548-2560.

[22] Tian Z, Tolić N, Zhao R, Moore RJ, Hengel SM, Robinson EW, Stenoien DL, Wu S, Smith RD, Paša-Tolić L. Enhanced top-down characterization of histone post-translational modifications. Genome Biol. 2012 Oct 3;13(10):R86.

[23] McCool EN, Xu T, Chen W, Beller NC, Nolan SM, Hummon AB, Liu X, Sun L. Deep top-down proteomics revealed significant proteoform-level differences between metastatic and nonmetastatic colorectal cancer cells. Sci Adv. 2022 Dec 21;8(51):eabq6348.

[24] Chen D, McCool EN, Yang Z, Shen X, Lubeckyj RA, Xu T, Wang Q, Sun L. Recent advances (2019-2021) of capillary electrophoresis-mass spectrometry for multilevel proteomics. Mass Spectrom Rev. 2023 Mar;42(2):617-642.

[25] Lubeckyj RA, Sun L. Laser capture microdissection-capillary zone electrophoresis-tandem mass spectrometry (LCM-CZE-MS/MS) for spatially resolved top-down proteomics: a pilot study of zebrafish brain. Mol Omics. 2022 Feb 21;18(2):112-122.

[26] Xu T, Shen X, Yang Z, Chen D, Lubeckyj RA, McCool EN, Sun L. Automated Capillary Isoelectric Focusing-Tandem Mass Spectrometry for Qualitative and Quantitative Top-Down Proteomics. Anal Chem. 2020 Dec 15;92(24):15890-15898.

[27] McCool EN, Lubeckyj RA, 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 May 1;90(9):5529-5533.

[28] Barnett DA, Ells B, Guevremont R, Purves RW. Application of ESI-FAIMS-MS to the analysis of tryptic peptides. J Am Soc Mass Spectrom. 2002 Nov;13(11):1282-91.

[29] Saba J, Bonneil E, Pomiès C, Eng K, Thibault P. Enhanced sensitivity in proteomics experiments using FAIMS coupled with a hybrid linear ion trap/Orbitrap mass spectrometer. J Proteome Res. 2009 Jul;8(7):3355-66.

[30] Hebert AS, Prasad S, Belford MW, Bailey DJ, McAlister GC, Abbatiello SE, Huguet R, Wouters ER, Dunyach JJ, Brademan DR, Westphall MS, Coon JJ. Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. Anal Chem. 2018 Aug 7;90(15):9529-9537.

[31] Bekker-Jensen DB, Martínez-Val A, Steigerwald S, Rüther P, Fort KL, Arrey TN, Harder A, Makarov A, Olsen JV. A Compact Quadrupole-Orbitrap Mass Spectrometer with FAIMS Interface Improves Proteome Coverage in Short LC Gradients. Mol Cell Proteomics. 2020 Apr;19(4):716-729.

[32] Kaulich PT, Cassidy L, Winkels K, Tholey A. Improved Identification of Proteoforms in Top-Down Proteomics Using FAIMS with Internal CV Stepping. Anal Chem. 2022 Mar 1;94(8):3600-3607.

[33] Takemori A, Kaulich PT, Cassidy L, Takemori N, Tholey A. Size-Based Proteome Fractionation through Polyacrylamide Gel Electrophoresis Combined with LC-FAIMS-MS for In-Depth Top-Down Proteomics. Anal Chem. 2022 Sep 20;94(37):12815-12821.

[34] Fulcher JM, Makaju A, Moore RJ, Zhou M, Bennett DA, De Jager PL, Qian WJ, Paša-Tolić L, Petyuk VA. Enhancing Top-Down Proteomics of Brain Tissue with FAIMS. J Proteome Res. 2021 May 7;20(5):2780-2795.

[35] Gerbasi VR, Melani RD, Abbatiello SE, Belford MW, Huguet R, McGee JP, Dayhoff D, Thomas PM, Kelleher NL. Deeper Protein Identification Using Field Asymmetric Ion Mobility Spectrometry in Top-Down Proteomics. Anal Chem. 2021 Apr 27;93(16):6323-6328.

[36] Pham KN, Fernandez-Lima F. Structural Characterization of Human Histone H4.1 by Tandem Nonlinear and Linear Ion Mobility Spectrometry Complemented with Molecular Dynamics Simulations. ACS Omega. 2021 Oct 27;6(44):29567-29576.

[37] Pham KN, Mamun Y, Fernandez-Lima F. Structural Heterogeneity of Human Histone H2A.1. J Phys Chem B. 2021 May 20;125(19):4977-4986.

[38] 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 Jan 1;146:839-43.

[39] 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 Jun 21;142(12):2118-2127.

[40] Wojcik R, Dada OO, Sadilek M, Dovichi NJ. Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. Rapid Commun Mass Spectrom. 2010 Sep 15;24(17):2554-60.

[41] Sun L, Zhu G, Zhang Z, Mou S, Dovichi NJ. 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 May 1;14(5):2312-21.

[42] Zamdborg L, LeDuc RD, Glowacz KJ, Kim YB, Viswanathan V, Spaulding IT, Early BP, Bluhm EJ, Babai S, Kelleher NL. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W701-6.

[43] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002 Oct 15;74(20):5383-92.

[44] Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007 Mar;4(3):207-14.

[45] Kou Q, Xun L, Liu X. TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics. 2016 Nov 15;32(22):3495-3497.

[46] Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics. 2008 Nov 1;24(21):2534-6.

[47] Chen D, Lubeckyj RA, Yang Z, McCool EN, Shen X, Wang Q, Xu T, Sun L. Predicting Electrophoretic Mobility of Proteoforms for Large-Scale Top-Down Proteomics. Anal Chem. 2020 Mar 3;92(5):3503-3507.

[48] Krokhin OV, Anderson G, Spicer V, Sun L, Dovichi NJ. Predicting Electrophoretic Mobility of Tryptic Peptides for High-Throughput CZE-MS Analysis. Anal Chem. 2017 Feb 7;89(3):2000-2008.

[49] Cifuentes A, Poppe H. Simulation and optimization of peptide separation by capillary electrophoresis. J Chromatogr A. 1994 Sep 30;680(1):321-40.

[50] Lubeckyj RA, McCool EN, 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 Nov 21;89(22):12059-12067.

[51] Wang Q, Sun L, Lundquist PK. Large-scale top-down proteomics of the Arabidopsis thaliana leaf and chloroplast proteomes. Proteomics. 2023 Feb;23(3-4):e2100377.

[52] Seibert M, Krüger M, Watson NA, Sen O, Daum JR, Slotman JA, Braun T, Houtsmuller AB, Gorbsky GJ, Jacob R, Kracht M, Higgins JMG, Schmitz ML. CDK1-mediated phosphorylation at H2B serine 6 is required for mitotic chromosome segregation. J Cell Biol. 2019 Apr 1;218(4):1164-1181.

[53] Suraweera A, Gandhi NS, Beard S, Burgess JT, Croft LV, Bolderson E, Naqi A, Ashton NW, Adams MN, Savage KI, Zhang SD, O'Byrne KJ, Richard DJ. COMMD4 functions with the histone H2A-H2B dimer for the timely repair of DNA double-strand breaks. Commun Biol. 2021 Apr 19;4(1):484.

[54] Cheung WL, Ajiro K, Samejima K, Kloc M, Cheung P, Mizzen CA, Beeser A, Etkin LD, Chernoff J, Earnshaw WC, Allis CD. Apoptotic phosphorylation of histone H2B is mediated by mammalian sterile twenty kinase. Cell. 2003 May 16;113(4):507-17.

[55] Fuhrmann J, Thompson PR. Protein Arginine Methylation and Citrullination in Epigenetic Regulation. ACS Chem Biol. 2016 Mar 18;11(3):654-68.

[56] DeMicco A, Bassing CH. Deciphering the DNA damage histone code. Cell Cycle. 2010 Oct 1;9(19):3845.

[57] Walker JN, Lam R, Brodbelt JS. Enhanced Characterization of Histones Using 193 nm Ultraviolet Photodissociation and Proton Transfer Charge Reduction. Anal Chem. 2023 Apr 11;95(14):5985-5993.

[58] Zubarev RA, Horn DM, Fridriksson EK, Kelleher NL, Kruger NA, Lewis MA, Carpenter BK, McLafferty FW. Electron capture dissociation for structural characterization of multiply charged protein cations. Anal Chem. 2000 Feb 1;72(3):563-73.

[59] Shen X, Xu T, Hakkila B, Hare M, Wang Q, Wang Q, Beckman JS, Sun L. Capillary Zone Electrophoresis-Electron-Capture Collision-Induced Dissociation on a Quadrupole Time-of-Flight Mass Spectrometer for Top-Down Characterization of Intact Proteins. J Am Soc Mass Spectrom. 2021 Jun 2;32(6):1361-1369.

[60] Chen B, Guo X, Tucholski T, Lin Z, McIlwain S, Ge Y. The Impact of Phosphorylation on Electron Capture Dissociation of Proteins: A Top-Down Perspective. J Am Soc Mass Spectrom. 2017 Sep;28(9):1805-1814.

[61] Jeanne Dit Fouque K, Miller SA, Pham K, Bhanu NV, Cintron-Diaz YL, Leyva D, Kaplan D, Voinov VG, Ridgeway ME, Park MA, Garcia BA, Fernandez-Lima F. Top-"Double-Down" Mass Spectrometry of Histone H4 Proteoforms: Tandem Ultraviolet-Photon and Mobility/Mass-Selected Electron Capture Dissociations. Anal Chem. 2022 Nov 8;94(44):15377-15385.

[62] Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci U S A. 2004 Jun 29;101(26):9528-33.

[63] Riley NM, Coon JJ. The Role of Electron Transfer Dissociation in Modern Proteomics. Anal Chem. 2018 Jan 2;90(1):40-64.

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 essential insights into the structures, functions, and dynamics of proteoforms and protein complexes near physiological conditions [4-7]. nMS 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

---

* *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. Angew Chem Int Ed Engl. 2024 Aug 28:e202408370, and Wang Q. Exploring the Function of Plastoglobules Using Top-down and Bottom-up Proteomics by Capillary Zone Electrophoresis–Mass Spectrometry (Doctoral dissertation, Michigan State University).*

and labor-consuming or only able to detect small proteoforms/protein complexes from complex proteomes.

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 analyze endogenous proteoforms and protein complexes in *E. coli* cells. We also compared our nCZE-MS data with mass photometry results in terms of the mass distribution of *E. coli* proteoforms and protein complexes [34].

5.2 Experimental section

5.2.1 Materials and reagents

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 (NH_4OAc), 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 was purchased from Acros Organics (NJ, USA). Micro Bio-SpinTM 6 kDa gel-filtration column units for buffer exchange were 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 NH_4OAc (pH ~ 7.0).

E. coli (strain Top10) was cultured in Terrific Broth (TB) medium at 37 °C until OD600 reached 0.7. After being washed with DPBS three times, a 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 NH_4OAc and directly measured using mass photometry.

Another aliquoted *E. coli* lysate was buffer exchanged to 20 mM NH_4OAc by Bio-Spin 6 kDa gel-filtration column. The column was washed with 20 mM NH_4OAc and centrifuged at 1,000 x (g) for 2 minutes and repeated 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). 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 sequentially in order of water - isopropanol - water - isopropanol - 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 and beta-amylase. 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.

The parameters related to the mass photometry are as follows. The mass precision was 2%. The mass error was 5%. The resolution (defined as FWHM) was 25 kDa @ 66 kDa and 60 kDa @ 660 kDa, respectively.

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.

5.2.5 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 a mass spectrometer [38, 39]. A Q-Exactive UHMR mass spectrometer (Thermo Fisher Scientific) was used for the experiments. The interface was directly attached to the 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 micropipette puller with an orifice size ~25 μm. The sheath liquid contains 10 mM NH_4OAc. 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 NH_4OAc (pH ~ 7.0).

The transfer capillary temperature was 250 °C, and the S-lens RF level was 200. The number of micro scans was 5 for MS, and the in-source trapping (IST) desolvation voltage was -30V. The trapping gas flow was at 5 (UHV readback showed ~1E-11 mbar). 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, ~50 ng). A 70-minute CZE separation with 30 kV was applied at the BGE end and 1 psi was applied in the meantime. For a standard protein mixture, the separation is 45 min, and the separation is under 1.5 psi.

5.2.6 Data analysis

All the mass spectra were firstly averaged by a time window of every 30 s, followed by inputting the data into UniDec. Only peaks with S/N better than 10 were analyzed. Most of the settings of UniDec [40] analysis was at default except applying the 'Automatic *m/z* Peak Width' and the 'Suppress Artifacts' with 'Some' or 'Lots'. Next, the successive charge state distribution of proteoforms/protein complexes was manually checked to ensure the correct distribution. At last, we calculated the mass and the standard deviation of the proteoform/protein complex by ESIProt [41] based on the *m/z* of the successive charge states from UniDec.

5.3 Results and discussion

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

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 (Bio-Rad P6) to a buffer containing 20 mM ammonium acetate (NH_4OAc, pH ~7.0) by gel filtration, followed by nCZE-MS analysis. The online nCZE-MS was assembled by coupling a Sciex CESI-8000 Plus capillary electrophoresis (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 NH_4OAc (pH ~7.0), and the sheath buffer for electrospray ionization (ESI) was 10 mM NH_4OAc (pH ~7.0). Only roughly 50 ng of the *E. coli* sample was consumed in a single nCZE-MS run. Raw MS data were averaged every 30 seconds, followed by mass deconvolution and check using UniDec and ESIprot [40, 41].

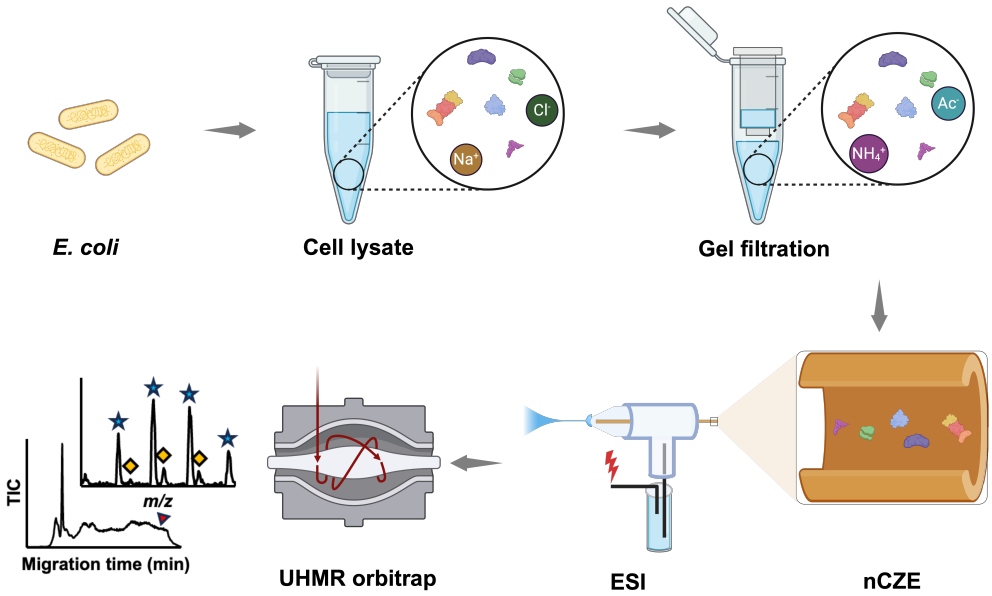


Figure 5.1. Flow chart of nCZE-ESI-MS for native proteomics of an *E. coli* cell lysate. The figure is created using the BioRender and used here with permission.

We investigated the sensitivity of the nCZE-ESI-UHMR platform for measuring protein complexes using a mixture of standard proteins and protein complexes, **Figure 5.2**. High intensity was 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 via consuming only about 15 ng of those proteins. After sample dilution by a factor of 50, a clear CRP peak was still observed, even though only 100 pg of the protein complex was loaded, indicating the high sensitivity of the technique. **Figure 5.3A** shows one mass spectrum of three SA tetramers with masses of 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 5.3B** shows a mass spectrum of the 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]. **Figure 5.3C** shows the mass spectrum of the 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 ($Z_R = 0.0778*M^{0.5}$), 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 5.3D**. The data demonstrate that intact protein complexes are maintained in native-like states during nCZE-ESI-UHMR measurements.

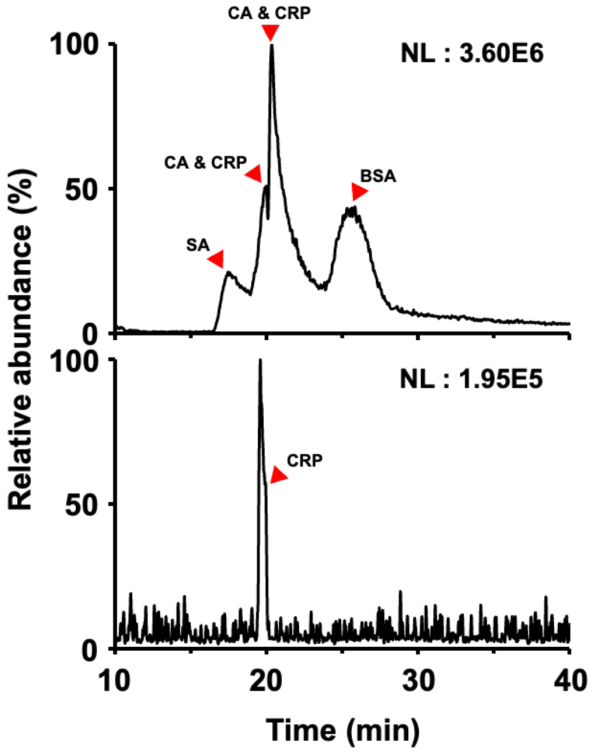
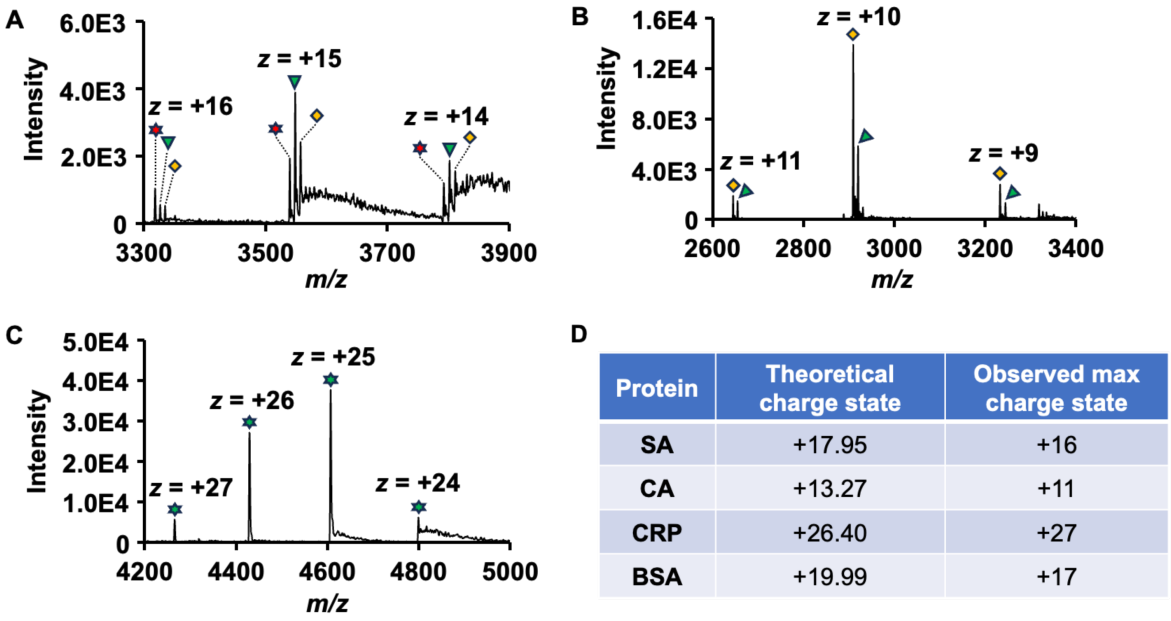


Figure 5.2. The data of a mixture of standard protein complexes was analyzed by CZE-ESI-UHMR, original concentration (13 μM, top), and 50-time dilution (bottom). SA: Streptavidin; CA: carbonic anhydrase; CRP: C-reactive protein; BSA: bovine serum albumin. Cyt C, Myo, and GDH were not detected in the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



D

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 are from Figure S1. (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 (calculated by Rayleigh charge 'Z_R' = 0.0778*$M^{0.5}$) and observed max charge states of 4 standard proteins or protein complexes.

5.3.2 Detection of proteoforms or protein complexes from *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 5.4A** 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 99 proteoforms or protein complexes in a mass range of 10-400 kDa after spectrum averaging and mass deconvolution. **Figures 5.4B-5.4F** show the mass spectra of some examples larger than 40 kDa, i.e., ~41, 139, 146, 318, 340, and 387 kDa. Those proteoforms or protein complexes show native-like and clear mass spectra. For example, **Figure 5.4E** shows two co-migrating proteoforms or protein complexes with masses ~318 and ~340 kDa. Their most-abundance charge states are +34 and +36, respectively. The largest proteoform or protein complex detected in this study is ~387 kDa, carrying around 42 charges (**Figure 5.4F**). Some additional examples are shown in **Figure 5.5**.

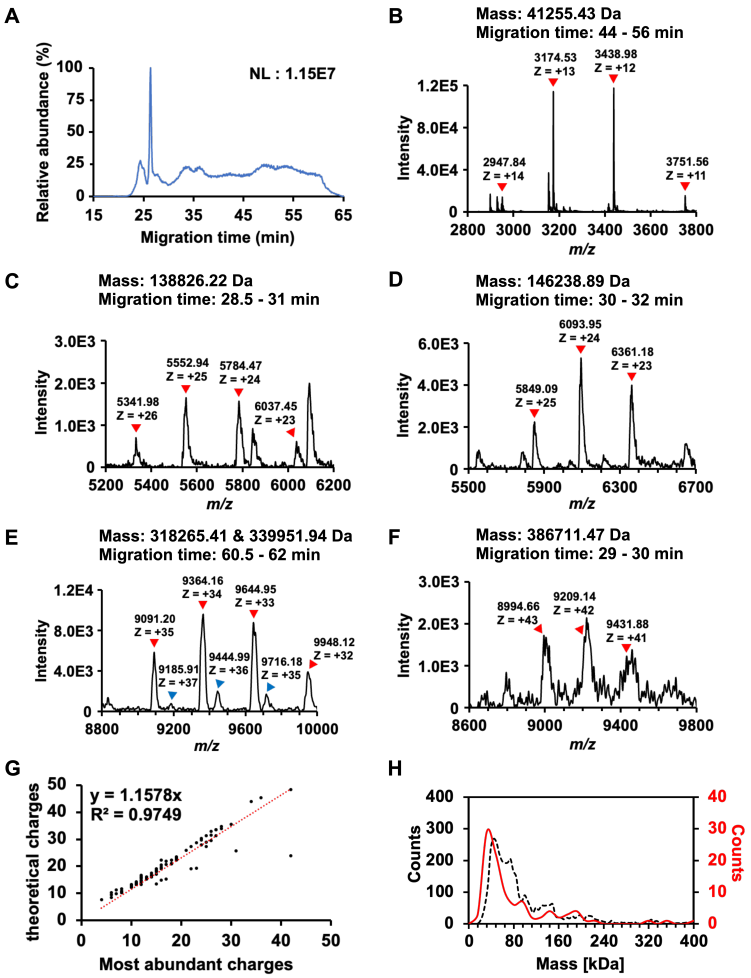


Figure 5.4. Summary of detected proteoforms or protein complexes from an *E. coli* cell lysate using nCZE-ESI-UHMR. (A). Representative electropherogram of nCZE-ESI-UHMR analyses of the *E. coli* cell lysate. (B)-(F). Mass spectra of five examples of large proteoforms/protein complexes were detected. The charge states and deconvolved mass of each proteoform/protein complex is labeled. (G). Linear correlation between the most abundant charges and theoretical Rayleigh charges (Z_R) of all proteoforms/protein complexes detected in single-shot nCZE-UHMR. (H). 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.

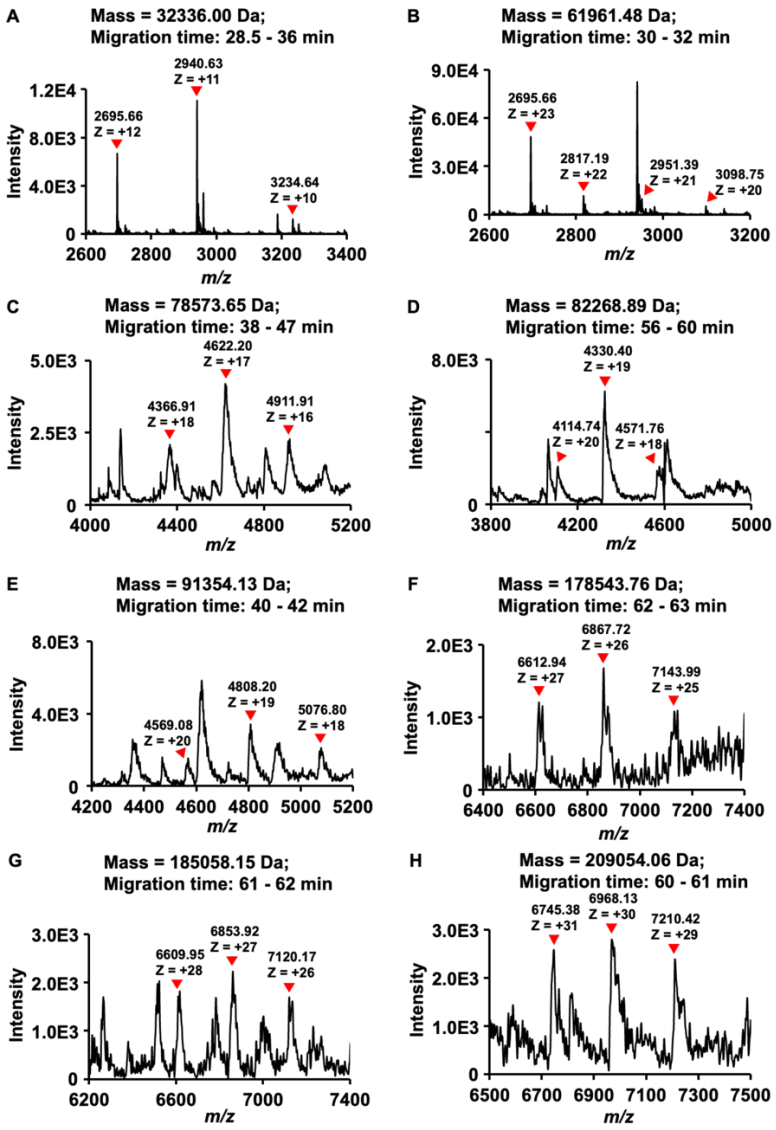


Figure 5.5. Representative mass spectra of proteoforms/protein complexes detected from the *E. coli* sample.

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 5.4G** [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 the manual determination of the highest charge state. We observed a strong linear correlation ($R^2 = 0.97$, slope of 1.16) 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 5.4H** (black dashed line). About 72% of the molecule counts (2558 of 3555) from the MP analysis are smaller than 100 kDa. Interestingly, the molecular mass distributions from the MP and nCZE-MS analysis agree reasonably well, **Figure 5.4H,** 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 78% (77 out of 99) of the proteoforms/protein complexes from nCZE-MS are smaller than 100 kDa. It has been demonstrated that nMS and MP can produce reasonably consistent mass assessments 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. Nearly 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.

Compared with native SEC-MS as another well-recognized technique for native proteomics, nCZE-MS has better sensitivity due to higher separation resolution and a much lower flow rate for ESI. However, native SEC-MS is robust and has high throughput [15, 49]. We expect that coupling native SEC fractionation with nCZE-MS will be helpful for further boosting the proteome coverage of native proteomics because the two separation techniques offer orthogonal separations of protein complexes.

The current study still has several limitations. Firstly, we only observed the mass information of proteoforms or protein complexes and did not generate high-quality MS/MS data during the nCZE-MS run, impeding the accurate identification of each protein. Those detected

proteoforms belong to level 5 identifications [50]. We will solve this issue by optimizing surface-induced dissociation (SID) or higher-energy collisional dissociation (HCD) to achieve better fragmentation of large proteoforms or protein complexes in our future study. Second, the sample loading capacity of nCZE is low, impeding the detection of low-abundance proteoforms or protein complexes and reducing the quality of acquired MS/MS spectra. We will enhance the overall sample loading capacity of nCZE by some online stacking techniques (e.g., capillary isoelectric focusing [27]) or offline fractionation techniques (e.g., SEC [33]). Third, the separations of large protein complexes by nCZE need to be further improved regarding separation peak capacity and reproducibility. **Figure 5.6** shows the electropherograms of triplicate nCZE-MS measurements of the *E. coli* cell lysate. **Figure 5.7** shows the extracted ion electropherograms of two example proteoforms/protein complexes. The peaks are much wider than that in denaturing CZE. The roughly estimated peak capacity of the nCZE separation is 15 based on the separation window and the average full peak width at half maximum of the two examples in **Figure 5.7**. The relatively low peak capacity is possibly due to the protein dispersion under the applied pressure and non-specific protein adsorption on 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 [51] 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. Lastly, the bioinformatics tool for data analysis needs to be improved. We employed mass deconvolution using UniDec [40] and ESIprot [41] for each averaged mass spectrum across the whole run. This approach was tedious and could be problematic for low-abundance proteoforms or protein complexes. More efforts are needed to build streamlined bioinformatic tools for large-scale native proteomics using, e.g., nCZE-MS.

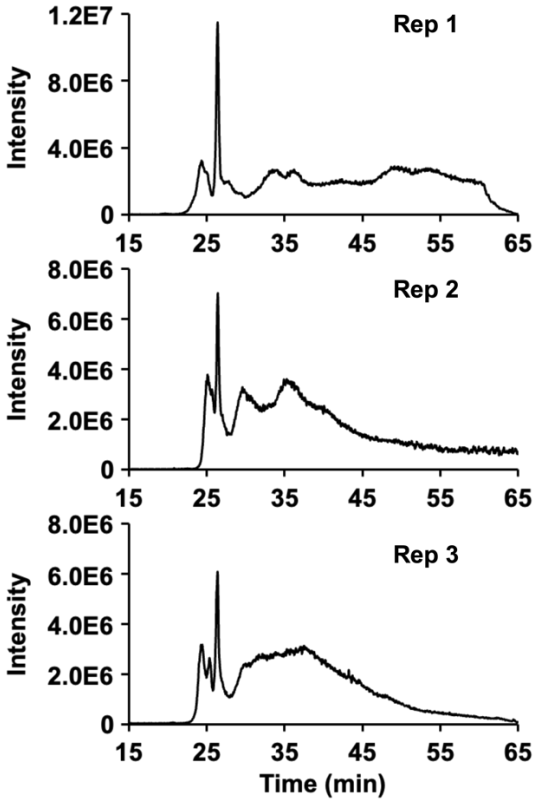


Figure 5.6. Electropherograms of triplicate analyses of an *E. coli* cell lysate by nCZE-ESI-UHMR. The electropherograms were aligned according to the most abundant peak.

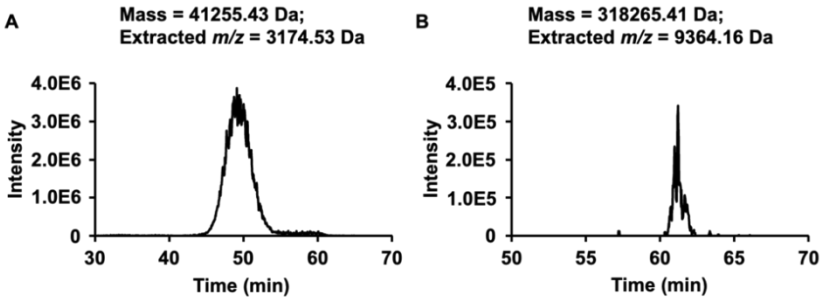


Figure 5.7. Extracted ion electropherograms of two example proteoform/protein complexes. The mass tolerance is set to 500 ppm, and Gaussian smoothing was enabled at 5 points.

5.4 Conclusion

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 nearly 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 Acknowledgment

We 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 LM, Kelleher NL; Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. Nat Methods. 2013 Mar;10(3):186-7.

[2] Smith LM, Agar JN, Chamot-Rooke J, Danis PO, Ge Y, Loo JA, Paša-Tolić L, Tsybin YO, Kelleher NL; Consortium for Top-Down Proteomics. The Human Proteoform Project: Defining the human proteome. Sci Adv. 2021 Nov 12;7(46):eabk0734.

[3] Zoll J, Heus HA, van Kuppeveld FJ, Melchers WJ. The structure-function relationship of the enterovirus 3'-UTR. Virus Res. 2009 Feb;139(2):209-16.

[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 Sep-Oct;42(5):1876-1926.

[5] Tamara S, den Boer MA, Heck AJR. High-Resolution Native Mass Spectrometry. Chem Rev. 2022 Apr 27;122(8):7269-7326.

[6] Karch KR, Snyder DT, Harvey SR, Wysocki VH. Native Mass Spectrometry: Recent Progress and Remaining Challenges. Annu Rev Biophys. 2022 May 9;51:157-179.

[7] Snyder DT, Harvey SR, Wysocki VH. Surface-induced Dissociation Mass Spectrometry as a Structural Biology Tool. Chem Rev. 2022 Apr 27;122(8):7442-7487.

[8] van de Waterbeemd M, Fort KL, Boll D, Reinhardt-Szyba M, Routh A, Makarov A, Heck AJ. High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. Nat Methods. 2017 Mar;14(3):283-286.

[9] Li H, Nguyen HH, Ogorzalek Loo RR, Campuzano IDG, Loo JA. An integrated native mass spectrometry and top-down proteomics method that connects sequence to structure and function of macromolecular complexes. Nat Chem. 2018 Feb;10(2):139-148.

[10] Fantin SM, Parson KF, Yadav P, Juliano B, Li GC, Sanders CR, Ohi MD, Ruotolo BT. Ion mobility-mass spectrometry reveals the role of peripheral myelin protein dimers in peripheral neuropathy. Proc Natl Acad Sci U S A. 2021 Apr 27;118(17):e2015331118.

[11] Keener JE, Zambrano DE, Zhang G, Zak CK, Reid DJ, Deodhar BS, Pemberton JE, Prell JS, Marty MT. Chemical Additives Enable Native Mass Spectrometry Measurement of Membrane Protein Oligomeric State within Intact Nanodiscs. J Am Chem Soc. 2019 Jan 16;141(2):1054-1061.

[12] Vimer S, Ben-Nissan G, Morgenstern D, Kumar-Deshmukh F, Polkinghorn C, Quintyn RS, Vasil'ev YV, Beckman JS, Elad N, Wysocki VH, Sharon M. Comparative Structural Analysis of 20S Proteasome Ortholog Protein Complexes by Native Mass Spectrometry. ACS Cent Sci. 2020 Apr 22;6(4):573-588.

[13] Gault J, Liko I, Landreh M, Shutin D, Bolla JR, Jefferies D, Agasid M, Yen HY, Ladds MJGW, Lane DP, Khalid S, Mullen C, Remes PM, Huguet R, McAlister G, Goodwin M, Viner R, Syka JEP, Robinson CV. Combining native and 'omics' mass spectrometry to identify endogenous ligands bound to membrane proteins. Nat Methods. 2020 May;17(5):505-508.

[14] VanAernum ZL, Busch F, Jones BJ, Jia M, Chen Z, Boyken SE, Sahasrabuddhe A, Baker D, Wysocki VH. Rapid online buffer exchange for screening of proteins, protein complexes and cell lysates by native mass spectrometry. Nat Protoc. 2020 Mar;15(3):1132-1157.

[15] Ventouri IK, Veelders S, Passamonti M, Endres P, Roemling R, Schoenmakers PJ, Somsen GW, Haselberg R, Gargano AFG. Micro-flow size-exclusion chromatography for enhanced native mass spectrometry of proteins and protein complexes. Anal Chim Acta. 2023 Jul 25;1266:341324.

[16] Sahasrabuddhe A, Hsia Y, Busch F, Sheffler W, King NP, Baker D, Wysocki VH. Confirmation of intersubunit connectivity and topology of designed protein complexes by native MS. Proc Natl Acad Sci U S A. 2018 Feb 6;115(6):1268-1273.

[17] Ren C, Bailey AO, VanderPorten E, Oh A, Phung W, Mulvihill MM, Harris SF, 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 Jan 2;91(1):903-911.

[18] Busch F, VanAernum ZL, Lai SM, Gopalan V, Wysocki VH. Analysis of Tagged Proteins Using Tandem Affinity-Buffer Exchange Chromatography Online with Native Mass Spectrometry. Biochemistry. 2021 Jun 22;60(24):1876-1884.

[19] Muneeruddin K, Nazzaro M, Kaltashov IA. 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 Oct 6;87(19):10138-45.

[20] Yan Y, Liu AP, Wang S, Daly TJ, Li N. Ultrasensitive Characterization of Charge Heterogeneity of Therapeutic Monoclonal Antibodies Using Strong Cation Exchange Chromatography Coupled to Native Mass Spectrometry. Anal Chem. 2018 Nov 6;90(21):13013-13020.

[21] Debaene F, Boeuf 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 Nov 4;86(21):10674-83.

[22] Yan Y, Xing T, Wang S, Daly TJ, 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 Jul 15;186:113313.

[23] Chen D, McCool EN, Yang Z, Shen X, Lubeckyj RA, Xu T, Wang Q, Sun L. Recent advances (2019-2021) of capillary electrophoresis-mass spectrometry for multilevel proteomics. Mass Spectrom Rev. 2023 Mar;42(2):617-642.

[24] Jooß K, McGee JP, Kelleher NL. Native Mass Spectrometry at the Convergence of Structural Biology and Compositional Proteomics. Acc Chem Res. 2022 Jul 19;55(14):1928-1937.

[25] Skinner OS, Haverland NA, Fornelli L, Melani RD, Do Vale LHF, Seckler HS, Doubleday PF, Schachner LF, Srzentić K, Kelleher NL, Compton PD. Top-down characterization of endogenous protein complexes with native proteomics. Nat Chem Biol. 2018 Jan;14(1):36-41.

[26] Chapman EA, Li BH, Krichel B, Chan HJ, Buck KM, Roberts DS, Ge Y. Native Top-Down Mass Spectrometry for Characterizing Sarcomeric Proteins Directly from Cardiac Tissue Lysate. J Am Soc Mass Spectrom. 2024 Apr 3;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 Apr;462:116541.

[28] Marie AL, Georgescauld F, Johnson KR, Ray S, Engen JR, Ivanov AR. Native Capillary Electrophoresis-Mass Spectrometry of Near 1 MDa Non-Covalent GroEL/GroES/Substrate Protein Complexes. Adv Sci (Weinh). 2024 Mar;11(11):e2306824.

[29] Jooß K, McGee JP, Melani RD, Kelleher NL. Standard procedures for native CZE-MS of proteins and protein complexes up to 800 kDa. Electrophoresis. 2021 May;42(9-10):1050-1059.

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

[31] 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 Nov 17;92(22):15202-15211.

[32] Jooß K, Schachner LF, Watson R, Gillespie ZB, Howard SA, Cheek MA, Meiners MJ, Sobh A, Licht JD, Keogh MC, Kelleher NL. Separation and Characterization of Endogenous Nucleosomes by Native Capillary Zone Electrophoresis-Top-Down Mass Spectrometry. Anal Chem. 2021 Mar 30;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 Sep 4;90(17):10095-10099.

[34] Wu D, Piszczek G. Standard protocol for mass photometry experiments. Eur Biophys J. 2021 May;50(3-4):403-409.

[35] Wojcik R, Dada OO, Sadilek M, Dovichi NJ. Simplified capillary electrophoresis nanospray sheath-flow interface for high efficiency and sensitive peptide analysis. Rapid Commun Mass Spectrom. 2010 Sep 15;24(17):2554-60.

[36] Sun L, Zhu G, Zhang Z, Mou S, Dovichi NJ. 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 May 1;14(5):2312-21.

[37] Marty MT, Baldwin AJ, Marklund EG, Hochberg GK, Benesch JL, Robinson CV. Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. Anal Chem. 2015 Apr 21;87(8):4370-6.

[38] 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 Feb;24(3):285-94.

[39] 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 Jul 12;94(27):9674-9682.

[40] Schachner LF, Ives AN, McGee JP, Melani RD, Kafader JO, Compton PD, Patrie SM, Kelleher NL. Standard Proteoforms and Their Complexes for Native Mass Spectrometry. J Am Soc Mass Spectrom. 2019 Jul;30(7):1190-1198.

[41] De La Mora JF. Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. Anal Chim Acta. 2000 Feb;406(1), 93-104.

[42] Heck AJ, Van Den Heuvel RH. Investigation of intact protein complexes by mass spectrometry. Mass Spectrom Rev. 2004 Sep-Oct;23(5):368-89.

[43] 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 Feb 15;4(2):211-216.

[44] Young G, Hundt N, Cole D, Fineberg A, Andrecka J, Tyler A, Olerinyova A, Ansari A, Marklund EG, Collier MP, Chandler SA, Tkachenko O, Allen J, Crispin M, Billington N, Takagi Y, Sellers JR, Eichmann C, Selenko P, Frey L, Riek R, Galpin MR, Struwe WB, Benesch JLP, Kukura P. Quantitative mass imaging of single biological macromolecules. Science. 2018 Apr 27;360(6387):423-427.

[45] den Boer MA, Lai SH, Xue X, van Kampen MD, Bleijlevens B, Heck AJR. 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 Jan 18;94(2):892-900.

[46] Deslignière E, Ley M, Bourguet M, Ehkirch A, Botzanowski T, Erb S, Hernandez-Alba O, Cianférani S. Pushing the limits of native MS: Online SEC-native MS for structural biology applications. Int J Mass Spectrom. 2021 Mar;461:116502.

[47] Smith LM, Thomas PM, Shortreed MR, Schaffer LV, Fellers RT, LeDuc RD, Tucholski T, Ge Y, Agar JN, Anderson LC, Chamot-Rooke J, Gault J, Loo JA, Paša-Tolić L, Robinson CV, Schlüter H, Tsybin YO, Vilaseca M, Vizcaíno JA, Danis PO, Kelleher NL. A five-level classification system for proteoform identifications. Nat Methods. 2019 Oct;16(10):939-940.

[48] Sadeghi SA, 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 Apr 5;23(4):1399-1407.

[49] McCool EN, 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 Oct 24;(140):58644.

[50] 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 Jan 1;146:839-43.

[51] Sun L, Zhu G, Zhao Y, Yan X, Mou S, Dovichi NJ. Ultrasensitive and fast bottom-up analysis of femtogram amounts of complex proteome digests. Angew Chem Int Ed Engl. 2013 Dec 16;52(51):13661-4.

CHAPTER 6. Conclusion and future directions

This dissertation aims to advance analytical techniques for MS-based multi-level proteomics. First, we introduced a novel high-throughput BUP workflow for plasma and serum samples by coupling magnetic nanoparticle protein corona and CZE-MS/MS, allowing for a rapid analysis for discovering potential cancer biomarkers. Second, a magnetic nanoparticle-based IMAC (Ti^{4+} and Fe^{3+}) for the enrichment of intact phosphoproteoforms from simple and complex protein mixtures for MS-based TDP was investigated, showing a high efficiency and good reproducibility. Third, we presented the first TDP example of coupling CZE, IMS, and MS as a multi-dimensional platform for the characterization of histone proteoforms with the 3-fold increasing identification numbers compared to no-IMS using a low microgram histone as the starting material. Fourth, we demonstrated that nCZE coupled to an Orbitrap UHMR mass spectrometer is a potent and sensitive platform for native proteomics to measure large proteoforms or protein complexes up to 400 kDa from a complex proteome sample.

Previous chapters mainly cover the sample preparation and separation improvement by reducing the sample complexity by nanoparticle-based strategies, enhancing the separation efficiency by the additional dimension of separation, and boosting the native protein separation and detection by creating a platform of coupling nCZE to a mass spectrometer with UHMR. However, extensive fragmentation for characterizing intact proteoforms remains highly demanded. Conventional collision-based fragmentation methods (CID or HCD) often struggle with comprehensive backbone cleavage of intact proteoforms, leading to challenges in precise sequence identification and accurate localization of PTMs [1, 2]. Electron or photon-based fragmentation techniques (ETD, ECD, or UVPD) significantly outperform HCD in fragmenting intact proteoforms, and it could be further enhanced by extra collision energy to electron-based fragmentation, such as ECciD (ECD followed by CID) [3, 4]. Recently, we evaluated the fragmentation performance of EThcD (ETD followed by HCD) compared to HCD-only for an intact standard protein mixture by coupling CZE to an Orbitrap Ascend Tribrid mass spectrometer. **Figure 6** illustrates how carbonic anhydrase (~29 kDa) was fragmented, achieving only 24% sequence coverage with HCD but an impressive 60% with EThcD. This underscores that the combination of electron-based and collision-based fragmentation can significantly enhance backbone cleavage, making it a powerful approach for the detailed characterization of intact proteins.

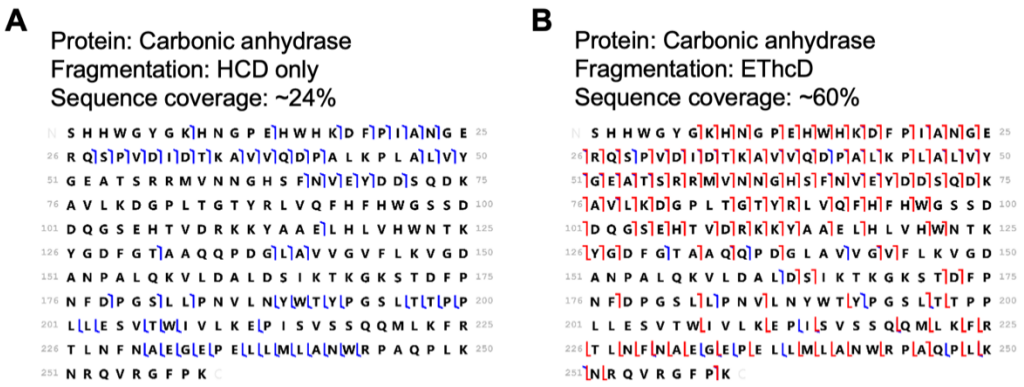


Figure 6. Fragmentation pattern of carbonic anhydrase (~29 kDa) by HCD (**A**) and EThcD (**B**). Blue: b/y ions. Red: c/z ions.

Despite significant advancements in analytical techniques, spanning sample preparation, separation methods, gas-phase fragmentation, and mass spectrometry for multi-level proteomics, there remains a need for further innovation. Future efforts should focus on refining mass-limited sample preparation with high recovery rates, like in single-cell proteomics, enhancing high-capacity, high-resolution separation through multi-dimensional approaches, and advancing the characterization of large biopharmaceuticals, including monoclonal antibodies and antibody-drug conjugates. Moreover, achieving global and in-depth analyses of native proteins and protein complexes remains a crucial frontier to explore.

REFERENCES

[1] Cai W, Tucholski T, Chen B, Alpert AJ, McIlwain S, Kohmoto T, Jin S, Ge Y. Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion Chromatography Strategy. Anal Chem. 2017 May 16;89(10):5467-5475.

[2] Schaffer LV, Tucholski T, Shortreed MR, Ge Y, Smith LM. Intact-Mass Analysis Facilitating the Identification of Large Human Heart Proteoforms. Anal Chem. 2019 Sep 3;91(17):10937-10942.

[3] Riley NM, Westphall MS, Coon JJ. Sequencing Larger Intact Proteins (30-70 kDa) with Activated Ion Electron Transfer Dissociation. J Am Soc Mass Spectrom. 2018 Jan;29(1):140-149.

[4] Shen X, Xu T, Hakkila B, Hare M, Wang Q, Wang Q, Beckman JS, Sun L. Capillary Zone Electrophoresis-Electron-Capture Collision-Induced Dissociation on a Quadrupole Time-of-Flight Mass Spectrometer for Top-Down Characterization of Intact Proteins. J Am Soc Mass Spectrom. 2021 Jun 2;32(6):1361-1369.