UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

QUANTIFYING THE INTACT PROTEOME: OPTIMIZATION OF ANALYTICAL METHODS FOR HIGH-THROUGHPUT TOP-DOWN PROTEOMICS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

YANTING GUO
Norman, Oklahoma
2023

QUANTIFYING THE INTACT PROTEOME: OPTIMIZATION OF

ANALYTICAL METHODS FOR HIGH-THROUGHPUT TOP-DOWN

PROTEOMICS

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY THE COMMITTEE CONSISTING OF

Dr. Si Wu, Chair

Dr. Christina Bourne

Dr. Rakhi Rajan

Dr. Robert L. White

Dr. Yuchen Qiu

Acknowledgements

Firstly, I would like to express my deepest gratitude to my major advisor, Dr. Si Wu, for her tremendous support, continuous guidance and encouragement, and professional assistance during the five-year academic journey. Completion of this undertaking would not have been possible without her strong support and assistance. I am also extremely grateful to all the help I received from my committee members, Dr. Christina Bourne, Dr. Rakhi Rajan, Dr. Robert L. White, and Dr. Yuchen Qiu through the years, especially on my preliminary exam and general exam.

Secondly, special thanks to all Wu lab members during the last five years: Dr. Dahang Yu for mentoring me for all the benchwork, LC-MS/MS data collection, and data analysis in the lab. Dr. Kellye A. Cupp-Sutton for her tremendous help not only in research but also in Ph.D. life as a good friend. Dr. Zhe Wang for supporting me during my early days in lab on LC-MS data collection for my first project. Mulin Fang, one of my best friends, for all the care and help. Dr. Lushuang Huang, Ji Kang, Jiaxue Li, Zhitao Zhao, Trishika Chowdhury, Anju Teresa Sunny, Samin Anjum, Joel B. Langford, Walter P. Galie, Meena Seshadri, all undergraduate students I mentored, for all the help and backup.

Lastly, but most importantly, I would like to express my deepest appreciation to my family, my dad, mom, and my brother for their endless love and strong support. Especially, I would like to thank my husband, Shengfeng and our puppy, Beta, for their immense love and tremendous support during the most difficult time. I won’t be able to do it without them.

Table of Contents

List of Tables

List of Figures

Key to Abbreviations

MS	mass spectrometry
2D PAGE	two-dimensional polyacrylamide gel electrophoresis
TDP	top-down proteomics
BUP	bottom-up proteomics
PTMs	post-translational modifications
LC	liquid chromatography
1D	1 dimensional
MD	Multidimensional
RPLC	reversed phase liquid chromatography
HIC	hydrophobic interaction chromatography
HILIC	hydrophilic interaction chromatography
IEX	ion-exchange chromatography
SEC	size-exclusion chromatography
CE	capillary electrophoresis
CZE	capillary zone electrophoresis
BGE	background electrolyte
EOF	electroosmotic flow
cIEF	capillary isoelectric focusing
pI	isoelectric point
ALS-WPAGE	acid-labile surfactant-polyacrylamide gel electrophoresis

GELFrEE	Gel-Eluted Liquid Fraction Entrapment Electrophoresis
PEPPI	Passively Eluting Proteins from Polyacrylamide gels as Intact species
SILAC	stable isotope labeling by amino acids in cell culture
NeuCode SILAC	neutron encoding (NeuCode) SILAC
TIPMI	tunable intact protein mass increases
DiLeu	N, N-dimethylleucine
iTRAQ	isobaric tagging for relative and absolute quantification
TMT	tandem mass tag
iodoTMT	iodoacetyl tandem mass tag
m/z	mass-to-charge ratio
MPA	mobile phase A
MPB	mobile phase B
MW	molecular weight
MWCO	molecular weight cutoff
PMSF	phenylmethylsulfonyl fluoride
ACN	acetonitrile
IPA	isopropanol
HOAc	acetic acid
AF	ammonium formate
ABC	ammonium bicarbonate

PBS	phosphate-buffered saline
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
TFA	trifluoroacetic acid
TCEP	tris (2-carboxyethyl) phosphine hydrochloride
IAA	iodoacetamide
TEAB	triethylammonium bicarbonate
DMEM	Dulbecco's modified Eagle medium
FBS	fetal bovine serum
NaF	sodium fluoride
E. coli	*Escherichia coli*
ESI	electrospray ionization
HCD	high-energy collision dissociation
EThcD	electron transfer/higher-energy collisional dissociation
ETD	electron transfer dissociation

Abstract

Proteins play an indispensable cellular role within organisms and participate in a majority of biological functions, including enzymatic catalysis, immune reaction, and signaling. Multiple forms of a protein (*i.e.*, proteoforms) can be generated from one single gene due to genetic variation, alternative RNA splicing, and post-translational modifications (PTMs). It has been reported that there are about ~20,000 human protein-coding genes, ~100,000 RNA splice variants, and an estimated 1 million proteoforms in complex biological systems. Different proteoforms from the same protein may have different functions and abundances; therefore, studying proteoforms is essential to understanding biological functions and the mechanisms of different biological processes.

Top-down mass spectrometry (MS)-based proteomics (TDP) techniques analyze intact proteoforms for high throughput identification, quantification, and characterization of intact proteoforms. Recent advancements in commercially available, high resolution MS instrumentation, such as improved MS resolution and scan speed, have made MS-based proteomics more accessible to the proteomics community. However, due to the extremely high proteome complexity and wide dynamic range of proteoform concentrations in complex biological samples, TDP generally suffers from low sensitivity and low proteome coverage. As such, advancements in technology to quantify and characterize intact proteoforms using top-down proteomics are imperative.

Currently, label-free quantitation is the most common quantitative approach in TDP due to the overall simplicity of application and sample handling; however,

TDP suffers from run-to-run instrument variation, no multiplexing, and difficulty in implementing multidimensional (MD) separation. Application of isobaric chemical tag labeling (*e.g.*, tandem mass tag, TMT), which is the gold-standard for bottom-up quantitation, has been limited in application to quantitative analysis of intact proteoforms to pure proteins and simple protein mixtures. Further application of isobaric chemical tag labeling to complex biological samples (*e.g.*, cell lysate) remains challenging for three reasons: (1) protein precipitation under labeling conditions due to introduction of organic solvent; (2) production of side products including underlabeled (incompletely labeled) and overlabeled (labeling of unintended residues) species; (3) inability of MS fragmentation to simultaneously achieve adequate quantitation and identification. In this dissertation, I will present the development and optimization of sample preparation and MS methods to enable the application of TMT labeling to intact, complex biological samples.

Initially, we found that large molecular weight proteoforms tended to precipitate and "crash out" of solution under TMT labeling conditions. To minimize protein precipitation under labeling conditions, we developed a "filter-SEC" technique that couples 100 kDa MWCO filtration and size exclusion chromatography to enrich small molecular weight proteoforms (Yu D. *et al.* 2021). By removing larger proteoforms, we were able to accurately quantify and characterize smaller proteoforms (<35 kDa) in complex biological samples using TMT labeling. However, the production of side products, including both underlabeled and overlabeled species increased the sample complexity and decreased the signal-to-noise ratios which impeded the accurate quantification and

characterization of intact proteoforms, particularly low abundance proteoforms, in complex biological samples. We systematically optimized the intact protein-level TMT labeling conditions and achieved >90% labeling efficiency for TMT-labeled *E. coli* cell lysate and ~86% labeling efficiency for TMT-labeled *HeLa* cell lysate using the optimized conditions. Finally, we evaluated the higher-energy collisional dissociation (HCD) using an Orbitrap Exploris 240 mass spectrometer for TMT-labeled complex protein mixtures. We found that high HCD energies resulted in high-intensity reporter ion peaks for accurate quantification and relatively lower HCD energies resulted in production of adequate proteoform backbone fragment ions for confident identification. However, single HCD energies were not adequate to provide accurate quantification and confident characterization simultaneously. Therefore, we evaluated stepped normalized collision energy (SNCE) schemes between 30% to 50% and found that these stepped schemes provided balance between accurate quantification and confident identification.

The innovation of an intact protein TMT labeling platform for proteoform quantitation further allowed the application of multidimensional (MD) separation to quantitative top-down proteomics to improve proteome coverage. MD separation couples two or more orthogonal separation approaches to improve separation to increase detection of low abundance proteoforms. Previously, MD separation has not been used for quantitative TDP because MD separation is not compatible with label-free quantitation due to the inability to accurately quantify proteoforms that elute in multiple first dimension fractions. We successfully integrated an automated, online 2-dimensional (2D) high-pH/low-pH reversed phase liquid chromatography

(RPLC)-MS separation system with intact protein-level TMT labeling for deep proteome profiling and quantitative analysis of complex biological samples such as the *HeLa* proteome.

In summary, this dissertation presents the optimization of intact protein-level TMT labeling conditions to decrease the production of side products, the optimization of MS2 fragmentation energies to achieve a balance between accurate quantification and confident identification, and the coupling of automated online 2D RPLC-MS platform with protein-level TMT labeling for deep proteoform characterization and quantification in TDP. I believe that the isobaric labeling-based quantitative TDP platform demonstrated here holds great potential for the quantitative analysis of intact proteoforms in real biological samples. This will have significant impacts on many areas of proteomics research including understanding of disease state, disease progression, and biomarker discovery.

Chapter 1 Introduction

1.1 BACKGROUND

1.1.1 Proteome and proteomics

Proteins are macromolecules made up of amino acids through peptide bonds that play an indispensable cellular role within organisms and participate in majority of the biological reactions such as enzymatic catalysis, immune reaction, and signaling.[1] The central dogma explains the transition of genetic information within a biological system from DNA to RNA through transcription and from RNA to proteins through translation, where each single step undergoes different types of variations that contribute to the biological system complexity (***Figure 1-1***).[2–4] As such, multiple forms of a protein can be generated from one single gene and these variants are called "proteoforms".[3] The entire set of proteoforms expressed in an organism was termed the "proteome".[5] It has been reported that there are about ~20,000 human protein-coding genes, ~100,000 RNA splice variants, and an estimated 1 million proteoforms resulting in extremely high proteome complexity in complex biological systems.[3,6]

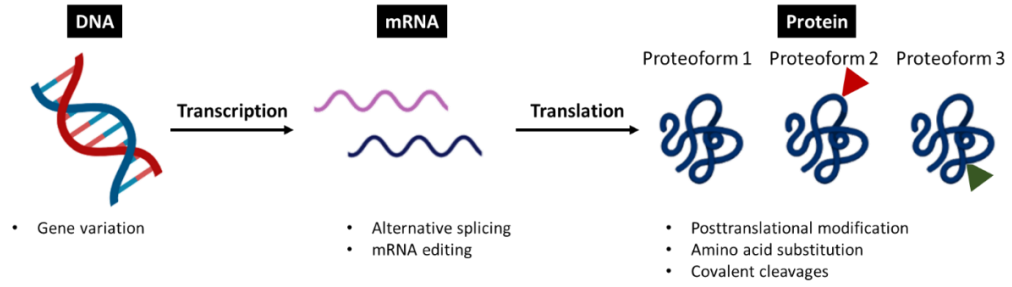


Figure 1-1 Central dogma of the flow of genetic information and generation of various proteoforms through variation at each step. Figure was created with BioRender.com.

Proteins are the key intermediaries connecting genotype and phenotype and it is impossible to understand the function of the biological systems without knowing what type of proteoforms are present, the abundance of different proteoforms, the location within cells or tissues, and the interactions with other proteoforms.[2] Proteomics, the large-scale study of proteomes of a cell, tissue, or organism, enables the investigation of proteoform expression, proteoform abundance, protein-protein interactions, and the understanding of biological mechanisms and pathways in complex biological systems (*e.g.*, cell lysate or biological fluids).[7] Mass spectrometry (MS)-based proteomics has emerged as an important technology to identity, quantify, and characterize proteoforms in a large-scale and high-throughput manner.[8] Recent advancements in commercially available, high resolution MS instrumentation, such as improved MS resolution and scan speed, have made MS-based proteomics more accessible to the proteomics community.[9,10] However, due to the extremely high proteome complexity (millions of proteoforms from approximately 20,000 human genes[2]) and wide dynamic range (>5-10 orders of magnitude[8,11]) of the protein mixture obtained from the complex biological sample, front-end separation methods are often applied to reduce sample complexity before MS analysis. Two-dimensional polyacrylamide gel electrophoresis (2D PAGE) was developed and has been traditionally applied as a standard procedure for proteomic research. 2D PAGE separates proteins based on molecular weight (MW) and isoelectric point (pI) with subsequent mass spectrometry analysis of protein spots isolated from the gel.[12,13] This process is typically labor-intensive and time-consuming and is not suitable for high

throughput proteomics analysis. Alternatively, front-end separation methods such as liquid chromatography (LC) and capillary electrophoresis (CE) have been developed and applied to MS-based proteomics to separate proteins and reduce sample complexity. These separation method can also facilitate online coupling to MS for direct analysis.[11] MS-based, large-scale proteomics studies utilizing LC or CE as front-end separation techniques are typically called high-throughput proteomics.

1.1.2 High-throughput proteomics

Two major high-throughput proteomics approaches have been developed and applied, bottom-up and top-down, ***Figure 1-2***. Bottom-up proteomics (BUP, peptide-level) is the most common strategy to study complex samples. In BUP, proteins are enzymatically digested into peptides, the peptide mixture is separated by 1-dimensional (1D) or multidimensional (MD) separation, and peptides are analyzed by MS.

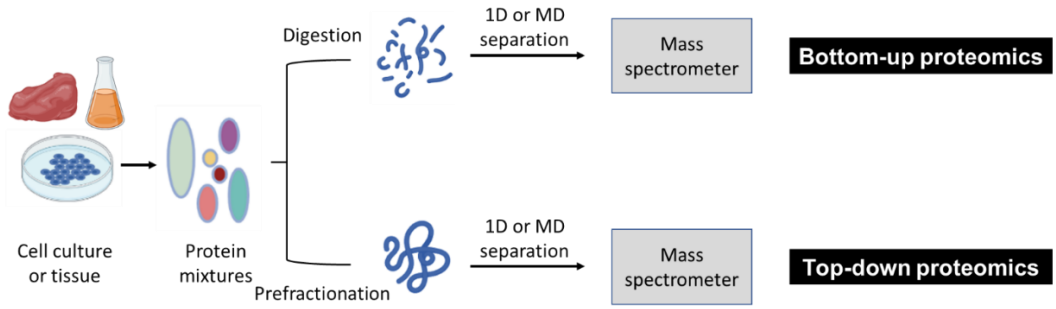


Figure 1-2 Typical workflow for high-throughput proteomics. Figure was created with BioRender.com.

Bottom-up proteomics is a robust and sensitive approach, which provides an indirect analysis of proteins through detection of peptides generated from proteins. Digestion of proteins into small peptides makes the sample more

homogeneous and limits the number of charge states which increases separation efficiency and benefits the MS detection, respectively.[14] Moreover, since proteins ware digested into peptides, there is no size limitation for proteins that can be analyzed by bottom-up proteomics. However, a key limitation to bottom-up proteomics is the loss of information regarding the intact proteoform such as the location of PTMs, the number of PTMs, and endogenous proteolysis due to incomplete sequence coverage and coexistence of multiple PTMs on the same protein, ***Figure 1-3***.[8,15]

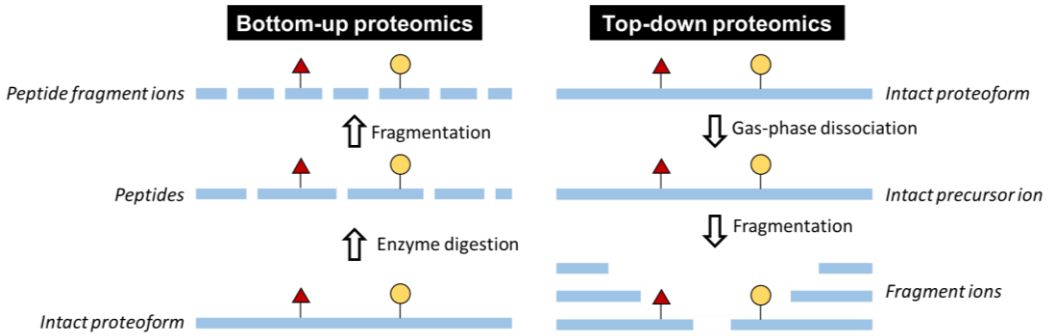


Figure 1-3 Bottom-up and top-down proteomics.

Unlike bottom-up proteomics, top-down proteomics (TDP, intact proteoform-level proteomics) allows for the identification and characterization of intact proteoforms arising from genetic variation, alternative RNA splicing or mutation, and post-translational modifications (PTMs).[3,11,15,16] Digestion is not required for TDP, so the location and identity of PTMs can be characterized, ***Figure 1-3***. PTMs have been found to play critical roles in development and progression of many human diseases including cardiovascular diseases, cancer, infectious diseases, neurodegenerative diseases, and immunobiology.[5] Moreover, different proteoforms from the same protein may have different functions and abundances in

difference cell types and as a result of diseases state. Therefore, quantitative and qualitative study of proteoforms is essential to understand biological functions and the mechanisms of different biological processes.

High-resolution mass spectrometers such as Fourier transform (FT) MS instruments are commonly required for high-throughput top-down proteomics.[17] Moreover, because of the high sample complexity, wide dynamic range, and heterogeneity of the intact proteoform mixture, efficient separation techniques such as LC and CE are required for the application of top-down proteomics in complex biological samples.[8,11] Both LC and CE have been commonly applied in TDP;[11,18–24] however, 1 dimensional (1D) separation may not provide enough separation resolution for high-throughput analysis of complex intact proteoform mixtures. Alternatively, multidimensional (MD) separation that couples two or more orthogonal separation techniques is often applied for in-depth proteome characterizations in TDP.

1.2 SEPARATION TECHNIQUES IN TOP-DOWN PROTEOMICS

Multidimensional (MD) separation couples two or more orthogonal separation techniques and is often applied for in-depth proteome characterizations in TDP. Orthogonal separation techniques are regarded as those having differing selectivities, and coupling orthogonal methods tends to result in increased peak capacities and minimized ionization suppression from coeluted species which is especially important for the complex spectra produced in TDP.[25,26] For MD separation, fractions from the first dimension are collected, either online or offline, and transferred to the second dimension for further separation and online MS

detection. Offline MD separation is flexible and can be applied to various first-dimension separation systems; however, additional sample handling steps in offline MD separations, such as buffer exchange and sample concentration, can be labor-intensive and may introduce sample loss, contamination, and degradation.[27] Alternatively, in online MD separation, eluent from the first-dimension system is either directly injected into the second dimension system or concentrated in a trapping column before injection into the second dimension system for separation, which eliminates many offline sample handling steps. Moreover, the sample required for online MD separation is typically much lower than the sample consumption in offline MD separation. On the other hand, online MD separation often has less flexibility due to issues with coupling various types of chromatography (*e.g.*, buffer compatibility).[8]

Recent advancements in TDP, including improved separation technologies, have been reviewed by various research groups.[11,15,26,28–45] Here, we discuss the development and applications of MD separation techniques to TDP reported within the last two decades. Experimental details, including separation column/capillary, sample type and loading amount, online/offline, and pros and cons of each approach, are summarized for each reviewed method. We also discuss perspectives and potential future applications of novel methods to MD separation**.**

1.2.1 Current state of separation techniques in top-down proteomics

Separation techniques applied to TDP have focused primarily on liquid chromatography (LC) and capillary electrophoresis (CE). As applied to TDP, LC

methods encompass diverse separation strategies with high separation efficiency to characterize intact proteoforms in complex samples.[11,46] These LC-based methods include (1) polarity-based separation: reverse phase chromatography (RPLC or RPC)[8,20], hydrophobic interaction chromatography (HIC)[47,48], and hydrophilic interaction chromatography (HILIC)[49,50]; (2) ionic strength-based separation: ion-exchange chromatography (IEX or IEC)[51]; (3) size-based separation: size-exclusion chromatography (SEC)[52,53].

The application of CE separation for TDP has been popularized more recently because it can be coupled directly to MS for data collection and features high separation efficiency, short separation time, low-volume sample requirements (nL to μL level), and ultra-low detection limits (zmol).[54–57] CE has several varieties that have been applied to TDP, including (1) capillary zone electrophoresis (CZE): separates proteins based on the differences in electrophoretic mobility[55]; and (2) capillary isoelectric focusing (cIEF): separates proteins based on isoelectric point (pI).[29,58]

Apart from LC and CE separation techniques, gel electrophoresis-based techniques have also been applied as a dimension of separation for TDP. These prefractionation methods separate intact proteins based on size and include Acid-Labile Surfactant–Polyacrylamide Gel Electrophoresis (ALS)-PAGE[59,60]**,** Gel-Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE)[46,61–70]**,** and Passively Eluting Proteins from Polyacrylamide gels as Intact species for MS (PEPPI-MS)[71,72]**.**

Together, these methods encompass all published separation strategies coupled for MD separation of intact proteins for MS-based TDP. Researchers have coupled these in a variety of ways to present unique and innovative platforms, as demonstrated in **Table *1-2* and *Table 1-2***. Here, we discuss the development and applications of these MD separation platforms as well as the advantages and disadvantages as they apply to the separation of intact proteins for TDP.

1.2.2 Multidimensional separations in top-down proteomics

Researchers have coupled different separation approaches in TDP in a variety of ways to present unique and innovative platforms, including RPLC-RPLC, IEC-RPLC, HIC-RPLC, RPLC-HILIC, SEC-RPLC, gel electrophoresis-RPLC, IEC-HIC, RPLC-CE, SEC-CE, gel electrophoresis-CE, and CE-CE, as demonstrated in ***Figure 1-4,*****Figure *1-4* Schematic of MD separation in top-down proteomics with** (A) RPLC-RPLC; (B) HIC-RPLC; (C) RPLC-HILIC; (D) IEX-RPLC; (E) SEC-RPLC; (F) Gel electrophoresis-RPLC; (G) CE-MD separations. ***Table 1-2, and Table 1-2***. Here, the development and application of these MD separation platforms applied in TDP were discussed.

RPLC-based MD separations: Reversed-phase liquid chromatography (RPLC) is the most commonly used separation approach in high-throughput TDP due to its MS-compatible buffers and high separation resolution.[73] RPLC separates proteins based on the strength of hydrophobic (*i.e.*, nonpolar) interactions between proteins and the nonpolar stationary phase (***Figure 1-4A***). Buffer gradients that move from polar (aqueous) to nonpolar (organic) are used to elute proteins based on differences in hydrophobicity. Coupling multiple RPLC methods for MD

separation may seem counterintuitive as a lack of orthogonality would be expected from methods with similar selectivities. It is possible, however, to couple RPLC techniques via a change in the buffer system; specifically, combination of two RPLC separations with high- (first dimension) and low- (second dimension) pH mobile phases is feasible due to variance in selectivity arising from differentially charged amino acids at different pH.[74] As positive ion mode is typically used for MS detection of intact proteins, all iterations of pH-based RPLC-RPLC MD separation for TDP feature low-pH RPLC as the second dimension directly coupled with MS for analysis.

In bottom-up proteomics, the success of 2D high-pH/low-pH RPLC separation has been limited by the average peptide length as differential charge distributions are limited with shorter peptides[74]. In contrast, intact proteins are much longer than peptides and contain more groups that maintain different charges at different pHs. As such, orthogonality between high-pH and low-pH RPLC was found to be much higher for intact proteins, and there was efficient utilization of the separation space without fraction concatenation.[18] The intrinsic orthogonality makes high-pH/low-pH RPLC an attractive MD method for both online and offline TDP analysis. Moreover, the MS-compatible mobile phases utilized in both dimensions require no desalting steps and further simplify sample processing.

The Wu lab developed and applied an offline high-pH/low-pH RPLC platform for the separation of intact cell lysate proteins.[75] As shown in ***Table 1-2***, a normal-flow Waters XBridge BEH C4 column (300 Å, 3.5 μm, 2.1 mm × 250 mm) was used for the 1st dimension high-pH RPLC separation, and an in-house

packed capillary C5 column (5 μm, 75 μm × 75 cm) was used for 2nd dimension low-pH RPLC separation. Different protein elution profiles were observed, demonstrating the orthogonality between high-pH RPLC and low-pH RPLC. 365 proteins and 886 proteoforms were identified from 500 μg intact *E. coli* lysate, compared to 163 proteins and 328 proteoforms identified using a 1D RPLC (low-pH)–MS approach (in-house packed capillary C5 column) from 10 μg intact *E. coli* lysate. This platform was also applied to intact *HeLa* cell lysate with a successful characterization of 20 different types of PTMs, including phosphorylation, acetylation, glutathionylation, and octanoylation.[76]

Sample handling steps used in offline MDLC commonly include buffer exchange and sample concentration steps that introduce sample loss, are low throughput, and may also cause sample degradation or addition of artificial PTMs such as oxidation.[8,11] Online MDLC platforms have been developed to overcome such challenges. For example, the Schug group reported the first top-down, online high-pH/low-pH RPLC platform with two sample storage loops for online coupling and a flow splitter for MS detection.[73] Microflow RPLC columns were used for both dimensions and a steep gradient (<1 min) was utilized in the 2nd low-pH RPLC separation to overcome the mismatch between mobile phases with different pH and organic solvent content. However, the overall separation resolution was relatively low due to the short second dimension gradient (*e.g.*, peak capacity of 19), resulting in limited improvement in proteome coverage. To improve the separation resolution, the Wu group developed an alternative online 2D high-pH/low-pH RPLC system with a short nanoRPLC column (*i.e.*, 150 μm i.d., 10 cm length) for

1D high-pH separation and a long nanoRPLC capillary column (*i.e.*, 75 μm i.d., 100 cm length) for 2D low-pH RPLC separation.[18] Using this online system, comparable identification numbers were observed with 100-times less initial sample injection amount (*e.g.*, 5 μg) compared with the offline high-pH/low-pH RPLC platform.[75]

Very recently, the Tholey group has offered a novel coupling method for RPLC-RPLC analysis for TDP. This group utilized low-pH RPLC in both dimensions and employed different ion pairing reagents and stationary phases to change selectivity between the two RPLC dimensions. In addition to concatenation to improve sample throughput, this group demonstrated the first implementation of low-pH/low-pH RPLC for MD separation and TDP.[77]

In summary, good orthogonality, improved separation efficiency, and more comprehensive protein characterization has been achieved using high-pH/low-pH RPLC platforms compared with 1D RPLC for TDP. MD RPLC takes advantage of the high-resolution separation offered by RPLC for both dimensions and requires no desalting steps due to the application of MS-compatible buffers. As such, only simple sample dilution with compatible buffers between dimensions is required to combine these techniques offline. This also reduces sample loss and simplifies the system for online coupling. However, RPLC separates proteins in a denaturing mode such that the protein tertiary structure and bioactivity cannot be preserved limiting its application to native MS analysis.

HIC-based MD separations: Hydrophobic interaction chromatography (HIC) separates proteins based on hydrophobic interactions between the stationary

phase and proteins. The separation environment is less denaturing and more polar compared with RPLC (***Figure 1-4B***).[47,48] HIC is a valuable tool because it provides high-resolution separation of intact proteins under nondenaturing conditions so that the protein tertiary structure and bioactivity can be preserved allowing separation of intact proteins with small conformational variations.[47,78,79] A salt gradient, from high to low salt concentration, is generally required as the mobile phase for HIC separation and is the key to ensuring the retention of hydrophobic proteins and elution of hydrophilic proteins.[78] As such, proteins elute in order of increasing surface hydrophobicity.

Generally, the requirement of non-volatile salt buffers (*e.g.,* sulfate, phosphate, or citrate salts) has largely limited the application of HIC for direct MS analysis and as the terminal separation method for MD separation.[47,80] However, the Ge group evaluated the use of volatile ammonium tartrate for HIC separation and found that similar results were obtained compared with ammonium sulfate.[78] With the development of a HIC separation protocol for intact proteins that utilized MS-compatible, volatile salts, the Ge group further developed a MD separation method coupling HIC with RPLC.[78] They found that, even though both HIC and RPLC separate proteins based on hydrophobicity, orthogonality was observed between the two methods based on the change in protein folding between the conditions. Specifically, proteins maintain native state folding under HIC buffer condition and become denatured under RPLC conditions resulting in different exposure of hydrophobic regions of the proteins. Therefore, complementary selectivity was achieved between the two dimensions.

Additionally, the Ge group published a 3DLC platform coupling IEC, HIC and RPLC for intact protein separation. The MS-compatible buffer, ammonium tartrate, previously optimized for HIC separation was utilized for both IEC and HIC.[81] IEC was coupled before HIC because a salt gradient from low to high is required to elute proteins in IEC, while a higher starting salt concentration is required for HIC separation; therefore, desalting steps were not required between separation dimensions.

Overall, HIC is a promising method for TDP due to its ability to separate proteins in a nondenaturing mode to preserve the tertiary structure of proteins and biological activity.[78] This allows for separation of intact proteins with small changes in conformations, such as those induced by protein modification (*e.g.*, PTMs). Integration of HIC with RPLC improved protein characterization compared to 1D RPLC. Recently, the Ge group introduced an online HIC-MS platform that utilizes a mobile phase comprised of relatively low concentrations of ammonium acetate and organic solvent to retain the tertiary structure of proteins. This platform demonstrated the potential of applying HIC as the terminal dimension in MD TDP.[48] As a hybrid method, this innovation would allow for the implementation of HIC as a terminal separation method for MD separation and would open up a new avenue of coupling HIC with other separation methods.

HILIC-based MD separations: Hydrophilic interaction liquid chromatography (HILIC) separates intact proteins based on interactions of polar functional groups with polar stationary phases (***Figure 1-4C***). Mobile phases for HILIC generally consist of aqueous solutions containing water-miscible polar

organic solvents such as acetonitrile. Gradients begin with high concentrations of organic between 50-95% and become more aqueous over the gradient. Volatile salts with high solubility in organic solvent, such as ammonium acetate, are typically added to control interactions between proteins and the stationary phase.[82] In HILIC separation, polar proteins elute later than nonpolar proteins, roughly the reverse of the elution order in RPLC separation. Compared with RPLC, HILIC utilizes higher concentrations of organic solvent in the mobile phase, making it suitable for online coupling with MS. It has been reported that HILIC tends to result in higher sensitivity compared with RPLC, which has a more extensive range of organic solvent concentrations in the mobile phase.[83] As such, HILIC has been successfully applied to TDP to analyze intact proteins including histones, membrane proteins, and monoclonal antibodies.[84–88] Notably, HILIC was able to resolve different proteoforms resulting from acetylation and methylation of histones which may coelute in RPLC separation.[78,85] Moreover, weak cation exchange-hydrophilic interaction (WCX-HILIC), a mixed-mode chromatography, separates proteins based on both hydrophilic interaction and electrostatic interaction between the stationary phase and protein molecules.[89] Application of WCX-HILIC has resulted in successful separation of differentially methylated histone isoforms[89,90] that were not separated by RPLC[91].

The Kelleher group developed an offline 2DLC system with RPLC as the first dimension and HILIC as the second dimension to characterize histone isoforms to improve separation resolution and increase proteome coverage.[86] In this study, RPLC fractions were collected from purified H4 protein samples and analyzed by

HILIC-MS/MS. In total, 42 H4 isoforms were characterized using this system. Similarly, the Paša-Tolić group demonstrated the first online metal-free RPLC-WCX-HILIC-MS platform for the characterization of intact core histone proteins (H4, H2B, H2A, and H3).[27,92] Specifically, this group developed a salt-free pH-gradient[93] that was adapted for the second dimension WCX-HILIC separation in this study resulting in increased sensitivity and throughput. 41 H4 proteoforms were identified from ~24 μg core histones purified from human fibroblasts.[27] Additionally, this group presented a nanoflow-based online 2DLC system with WCX-HILIC as the first dimension and RPLC as the second dimension for the analysis of *HeLa* core histones.[19]

Overall, HILIC has demonstrated good orthogonality to RPLC for the TDP study of intact proteins. The high resolution of RPLC combined with the increased sensitivity of HILIC is particularly beneficial for MD analysis of hypermodified proteoforms as has been demonstrated for the analysis of histone proteoforms resulting from methylation and acetylation.

IEX-based MD separations: Ion exchange chromatography (IEX) separates protein molecules based on ionic interactions between charged protein molecules and ion exchanger stationary phases that can be positively (anion-exchange chromatography, AEX) or negatively (cation-exchange chromatography, CEX) charged based on functional groups (***Figure 1-4D***).[51] Both AEX and CEX can be further categorized into two types: weak and strong, including weak anion exchange (WAX), strong anion exchange (SAX), weak cation exchange (WCX), and strong cation exchange (SCX). Weak exchangers are only ionized within a

limited pH range (*e.g.*, pH = 5-9), and strong exchangers are ionized over the full pH range (*e.g.*, pH = 0-14).[94] IEX uses a salt concentration gradient for intact protein separation in which proteins with fewer charged groups are eluted out at lower salt concentrations and proteins with more charged groups are eluted out at higher salt concentrations. IEX cannot be directly coupled to MS without additional desalting steps and, as such, is often used as the non-terminal dimension in MD separation.

Both AEX and CEX have been coupled with RPLC for MD separations to analyze complex samples such as cell lysate. For example, the Paša-Tolić group applied WAX-RPLC separation to analyze *Shewanella oneidensis* MR-1 cell lysate. In this study, offline fractions from WAX were dialyzed to remove the salt before being subjected to RPLC-MS analysis.[95] More recently, an online 2DLC-MS platform that coupled AEX as the first dimension and RPLC as the second dimension was developed to analyze adeno-associated viruses (AAVs) with online desalting.[96] The Tholey group utilized SCX to reduce the abundance of low-charged ions prior to RPLC separation using monolithic columns and MS detection to characterize low molecular weight proteomes and short open reading frame-encoded peptides (SEP) using combined bottom-up proteomics (BUP) and TDP.[97]

A variant of IEX, chromatofocusing, utilizes a pH gradient to focus proteins based on their isoelectric point (pI) values.[98] Chromatofocusing has been reported to provide better peak shapes, compared with salt gradients, since proteins can be focused into relatively narrow bands.[98] However, generating a pH gradient that encompasses the full range of pI values for all intact proteins in complex samples

has proven challenging. To address this challenge, the Greibrokk group innovated a mixture of six amine buffers that provided a wide pH gradient descending from 10.5 to 3.5. Fractions from pH gradient separation were subjected to RPLC, followed by UV and ESI-TOF-MS analysis.[98]

IEX-RPLC is an attractive MDLC separation technique for TDP analysis due to high orthogonality between IEX and RPLC (*e.g.*, charge *vs.* hydrophobicity). Both salt-gradient-based IEX and pH-gradient-based IEX have been used in MD separation for TDP. Although salt buffers are typically required for IEX separation, with online desalting or substitution of volatile buffers, IEX-RPLC still holds great potential for deeper proteome profiling in TDP.

Size-based MD separations: Characterization of high molecular weight (MW) intact proteoforms in complex biological samples is challenging due to wide dynamic range, high sample complexity, low sensitivity, and limitations in high resolution MS technology.[8] Specifically, it has been reported that the signal-to-noise ratio (S/N) has an exponential decay for proteins as a function of increasing MW, and this problem was amplified for high MW proteoforms that coeluted with low MW proteoforms.[24,99] Enrichment of proteins with MW suited for MS-based TDP (*e.g.*, <30 kDa)[100] can lower spectral complexity and increase sensitivity for analysis of intact proteins. Thus, the inclusion of size-based separation techniques has proven valuable for MD separation. These size-based separation techniques can be coupled with other methods such as RPLC for in-depth proteome characterization (***Figure 1-4E&F***). To date, several size-based separation techniques have been developed or implemented for MDLC separation in TDP,

including size exclusion chromatography (SEC)[24,101,102], Gel-Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE)[46,61–70], Passively Eluting Proteins from Polyacrylamide gels as Intact species for MS (PEPPI-MS)[71,72], and acid-labile surfactant–polyacrylamide gel electrophoresis (ALS-PAGE)[59,60].

Size exclusion chromatography (SEC) separates proteins based on size as they pass through a packed column of porous beads (***Figure 1-4E***). Smaller proteins are able to diffuse into the pores on the beads resulting in increased retention compared with larger proteins that are excluded from the pores and tend to elute earlier.[52,53] SEC is compatible with various solvents and salt buffers, including MS-compatible salts such as ammonium acetate that preserve native protein folding.[103] However, SEC separation suffers from relatively low resolution, and proteins are highly diluted after SEC separation.[52,104] These drawbacks can be partially addressed by coupling SEC with orthogonal techniques such as RPLC. 2D SEC-RPLC separation has been previously applied to analyze complex biological samples such as intact *Shewanella oneidensis* proteome[102], endogenous membrane protein proteome profiling for HEK 293T cell lysate and cardiac tissue lysate[101], and bovine seminal plasma proteins[105]. Moreover, the Ge group developed a serial SEC (sSEC)-RPLC platform to increase SEC resolution prior to RPLC-MS analysis. This platform coupled 3 SEC columns with 1000 Å, 500 Å, and 500 Å pore size, sequentially, for high-resolution size-based fractionation. Significantly more unique proteoforms were identified using this sSEC-RPLC platform compared to 1D RPLC, including higher MW proteins up to 223 kDa.[24] Overall, coupling SEC

with RPLC for TDP has been valuable in reducing spectral complexity and is particularly useful when coupled with MS compatible buffers.

Other size-based prefractionation methods that have also been applied for MD separation and TDP include Acid-Labile Surfactant-Polyacrylamide Gel Electrophoresis (ALS-PAGE), Gel-Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE), Passively Eluting Proteins from Polyacrylamide gels as Intact species for MS (PEPPI-MS), ***Figure 1-4F***. ALS-PAGE and GELFrEE are tube gel electrophoresis methods with continuous elution, which separate proteins using polyacrylamide gel similar to the Sodium Dodecyl-Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE) technique.[106] However, unlike SDS-PAGE, in which proteins remain on the gel for staining and analysis, tube gel electrophoresis techniques elute proteins from the matrix for further TDP analysis. In ALS-PAGE, an acid-labile surfactant is used to aid in protein denaturation and charge distribution. This acid-labile surfactant can be degraded in acid before RPLC-MS analysis.[59,60] Similarly, the GELFrEE platform can be used for size-based prefractionation of proteins; however, Sodium Dodecyl-Sulfate (SDS) is used as the surfactant. As SDS is not compatible with MS detection, GELFrEE separation must be followed up by SDS removal using detergent removal techniques such as chloroform/methanol/water precipitation[107,108] or spin columns[109,110]. The original GELFrEE platform featured a single lane for protein separation; the authors further enhanced and commercialized the technique for multiple simultaneous separations in multiplexed GELFrEE (mGELFrEE).[63] GELFrEE has been successfully coupled with RPLC and/or solution isoelectric focusing (sIEF) in MD TDP for complex

sample analysis.[46,61–70] While GELFrEE is a robust and relatively high-resolution technique for size-based separations, the consumable gel cartridges for the commercial GELFrEE instrument have unfortunately become discontinued, making the method inaccessible to the general public. As such, other groups have developed gel-based fractionation techniques to apply to MD separations for TDP.

One alternative method to GELFrEE for size-based prefractionation is PEPPI. PEPPI was designed to passively extract intact proteins from precast gels as a prefractionation method in MD TDP.[71] After SDS removal by methanol/chloroform/water precipitation, proteins can be directly analyzed using RPLC-MS. In addition, anion-exchange disk-assisted sequential sample preparation (AnExSP) can also be coupled to PEPPI to remove SDS from fractions.[72] PEPPI-MS has been successfully applied to analyze complex lysates such as *E. coli*. Overall, PEPPI is an accessible alternative to continuous elution tube gel electrophoresis that can be easily implemented by many labs.

Capillary electrophoresis-coupled MD separations: Capillary electrophoresis (CE) separates proteins in open-bore submillimeter diameter capillaries using electrokinetic forces. Two commonly applied CE techniques in TDP include capillary zone electrophoresis (CZE), which separates proteins based on electrophoretic mobility in an electric field, and capillary isoelectric focusing (cIEF), which separates proteins based on their isoelectric points (pIs), ***Figure 1-4G***. Recently, CE separations have increased in popularity in TDP due to the associated high sensitivity, high-resolution separation, and small sample

mass/volume requirements. Moreover, CE can be directly coupled with MS for online analysis since MS-compatible buffers are utilized.

Various RPLC-CE separation methods (*e.g.*, CE as the 2nd dimension in coupling with MS) have been developed for TDP. For example, the Smith group offline coupled RPLC with cIEF for analysis of *Shewanella oneidensis* proteome.[102] The Sun group developed an automated online cIEF-MS/MS platform and performed the large-scale characterization of proteoforms in *E. coli* cell lysate by coupling SEC-cIEF-ESI-MS/MS.[111] The online cIEF-MS/MS platform was also successfully applied for characterizing monoclonal antibody (mAb) charge variants.[112] Additionally**,** the Yates group demonstrated the offline coupling of RPLC with CZE for analyzing *Pyrococcus furiosus* cell lysate[55] Using this platform, the Dovichi group further optimized the CZE separation and MS data collection to improve proteoform characterization.[113] Online coupling of RPLC with CE separation is challenging because separation conditions between these two formats are incompatible due to the differences in separation parameters including flow rates, solvents, and sample volumes.[114–118] To overcome these challenges, the Neusüß lab presented an online nanoRPLC-CZE-MS platform that utilized a 4-port valve interface to transfer small sample volumes from the nanoRPLC dimension to the CE capillary for online CZE-MS. This group initially used this platform for the heart-cut analysis of a simple protein mixture[116] and further applied it to the analysis of human alpha-1-acid glycoprotein and nearly doubled the number of assigned glycoforms compared with one-dimensional CZE-MS[119]. Overall, coupling RPLC

with CZE or cIEF results in good orthogonality and improved proteoform characterization compared with 1D separation techniques.

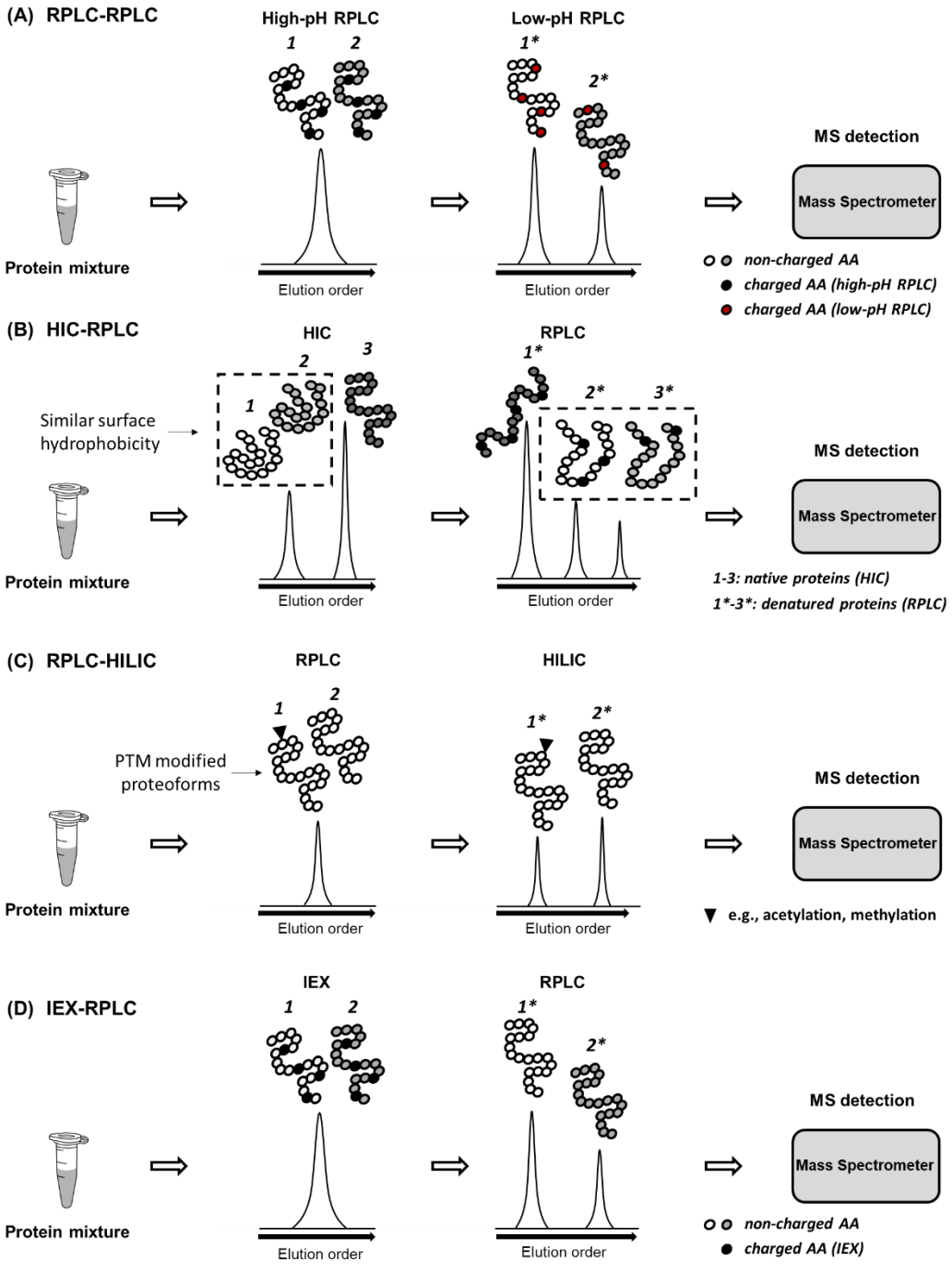
CE-based methods, cIEF and CZE, can be coupled in MD TDP because the separation principles between the techniques are orthogonal (*e.g.*, pI for cIEF and electrophoretic mobility for CZE). However, direct coupling of these two methods is complicated because of the difference in separation electrolytes and addition of ampholytes in cIEF that may affect the ESI process.[120–122] Using a similar 4-port valve coupling mechanism as discussed above, the Neusüß group developed an online cIEF-CZE-MS platform for the TDP analysis of standard protein mixtures[123] and deglycosylated model antibody (mAb) charge variants[124]. However, this platform is not readily applied to complex biological samples such as cell lysates because the heart-cut fraction selection process is incapable of consecutive analysis of multiple fractions from the first-dimension separation.

Size-based fractionation has also been commonly used as a prefractionation method for CE-based MD separations in TDP.[21,100,105,125–127] The Kelleher group used a GELFrEE-CZE-MS platform to characterize large intact proteins in the 30-80 kDa mass range from *Pseudomonas aeruginosa* whole cell lysate and found that size-based fractionation prior to CE-MS can improve identification of large proteoforms.[100] Alternatively, the Sun group developed an offline SEC-CZE-MS platform to characterize complex proteomes such as *E. coli* cell lysate. Initially, this platform used native conditions for both SEC and CZE to perform native MS on *E. coli* lysate.[127] This group further implemented SEC-CZE-MS to analyze denatured samples to improve proteoform characterization.[125,126] The Yates group recently

applied both SEC-CZE and SEC-RPLC for proteome profiling of bovine seminal plasma proteins using electron-transfer/higher-energy collisional dissociation (EThcD) and 213 nm ultraviolet photodissociation (213 nm UVPD).[105]

CE methods have also been integrated into 3D separations. For example, the Sun group demonstrated a 3D separation platform by coupling SEC, RPLC, and CZE-MS for deep intact proteome characterization of *E. coli* cell lysate.[22] This 3D platform was recently used in conjunction with CZE-MS/MS, SEC-CZE-MS/MS, and RPLC-CZE-MS/MS to analyze human colorectal cancer cell lines. Altogether, this study identified 23,622 proteoforms of 2,332 proteins from two cancer cell lines, representing one of the largest TDP datasets currently reported.[21]

CE has been demonstrated to be a powerful addition to MD separation platforms because of its high sensitivity, small sample volume, and high-resolution separation. Moreover, use of MS-compatible buffers in CE makes it capable of direct coupling with MS and ideal for coupling with RPLC and size-based separation techniques. However, while there are a few instances of online MD separations with CE, developing robust online platforms for complex sample analysis is still challenging due to the extremely low sample injection volume required for CE analysis.


(A) RPLC-RPLC
High-pH RPLC
Low-pH RPLC
1
2
1*
2*
MS detection
Mass Spectrometer
Protein mixture
Elution order
Elution order
non-charged AA
charged AA (high-pH RPLC)
charged AA (low-pH RPLC)
(B) HIC-RPLC
HIC
RPLC
Similar surface hydrophobicity
1
2
3
1*
2*
3*
MS detection
Mass Spectrometer
Protein mixture
Elution order
Elution order
1-3: native proteins (HIC)
1*-3*: denatured proteins (RPLC)
(C) RPLC-HILIC
RPLC
HILIC
PTM modified proteoforms
1
2
1*
2*
MS detection
Mass Spectrometer
Protein mixture
Elution order
Elution order
e.g., acetylation, methylation
(D) IEX-RPLC
IEX
RPLC
1
2
1*
2*
MS detection
Mass Spectrometer
Protein mixture
Elution order
Elution order
non-charged AA
charged AA (IEX)

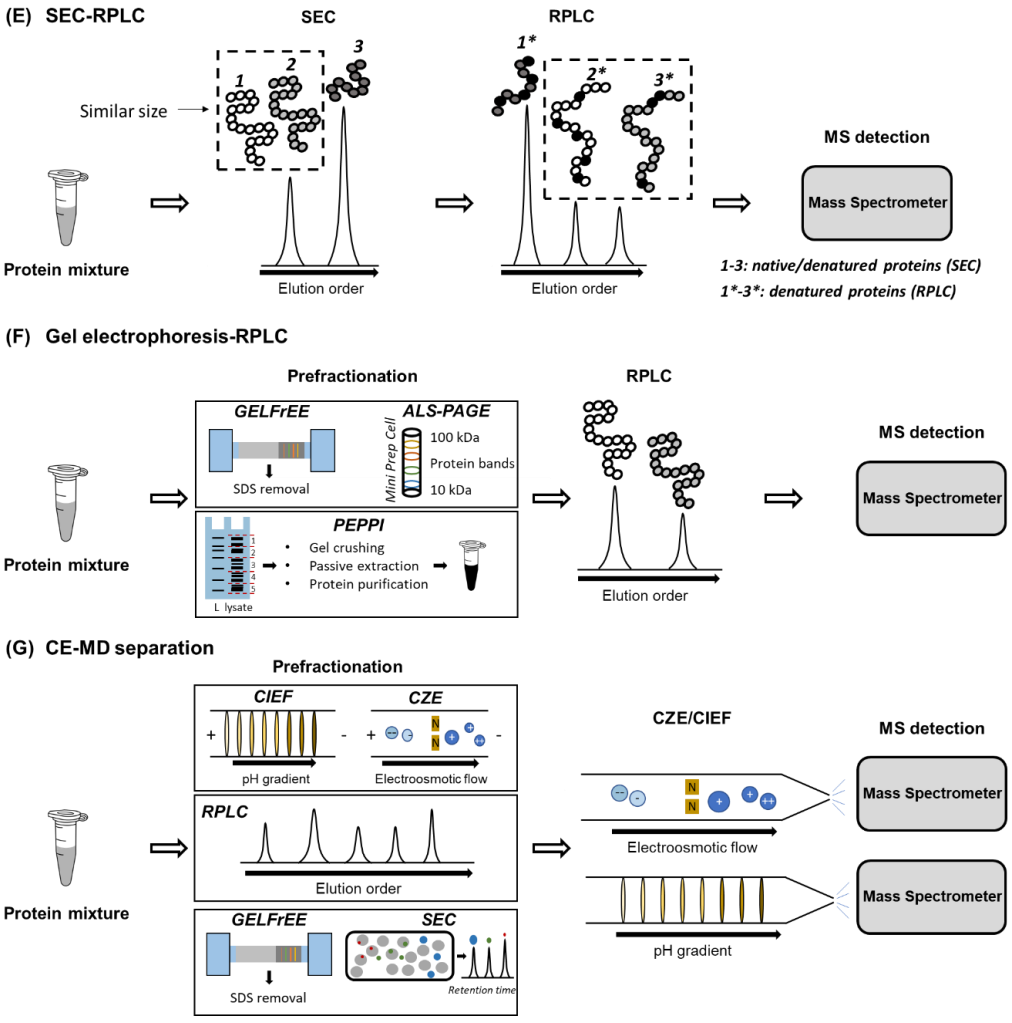


Figure 1-4 Schematic of MD separation in top-down proteomics with (A) RPLC-RPLC; (B) HIC-RPLC; (C) RPLC-HILIC; (D) IEX-RPLC; (E) SEC-RPLC; (F) Gel electrophoresis-RPLC; (G) CE-MD separations.

1.2.3 Current applications and perspectives

Since the development of top-down MD separations, the platforms discussed here have been applied to many systems such as *E. coli* cell lysate[18,22,75,78,125,127,128], *Shewanella oneidensis* cell lysate[55,95,102], *Pyrococcus furiosus* cell lysate[55], yeast cell lysate[113], *Pseudomonas aeruginosa PA01* cell lysate[100], zebrafish brain extract[126], *HeLa* cell lysate[46,76], *HEK 293* cell lysate[81,101], cardiac protein extract[24], bovine seminal plasma proteins[105], and human plasma

proteins[98]. Apart from these complex biological samples, other systems have also been studied, including the blood proteome in human hematopoietic cells, histone proteoforms, charge variants of mAbs, and other proteins with PTMs.[19,27,70,86,92,129] These are all impressive implementations of MD separation technologies; however, some additional developments such as improved quantitation methods, increased throughput, and improved MS data acquisition and data analysis would improve the robustness of these methods and make them more accessible to researchers.

TDP quantitation has been primarily limited to label-free quantitation, which is challenging to implement for MD separation due to the inability to accurately quantify proteoforms if separated into multiple first-dimension fractions.[15] Recently, intact protein-level tandem mass tag (TMT) labeling[130] and thiol-directed TMT labeling[131] have been optimized with respect to sample preparation[130,131], labeling parameters[132], and fragmentation[131,133] for labeling of intact, complex biological samples. Implementing isobaric chemical tag labeling also increases the throughput of quantitative TDP experiments. The throughput of MD TDP may also be improved by shortening cycle time with optimized short-gradient separations or through the use of LC systems with multiple columns.[134–140] Methods to increase the throughput of CE-MS for proteomics analysis have also been developed, including multisegmented injection CE-MS[141] which injects multiple samples for CE-MS analysis prior to initiating separation voltage. Limitations in applicable quantitative methods and low throughput are primary drawbacks to MD separation for TDP, methods such as these to quantify intact

proteoforms and increase throughput will greatly increase the feasibility of applying these methods to relevant biochemical research.

In summary, with the improved proteome coverage and separation efficiency acquired using MD separations in TDP, we believe that TDP holds great potential to address biological questions in real complex biological samples.

Table 1-1 Summary of multidimensional separation in TDP. Grey highlights represent reported combinations of two separation techniques with related references noted in the brackets. RPC: reverse phase chromatography; HIC: hydrophobic interaction chromatography; HILIC: hydrophilic interaction chromatography; (2) IEC: ion-exchange chromatography; SEC: size-exclusion chromatography; GE: gel electrophoresis. CE: capillary electrophoresis.

	1D						
2D	**RPC**	**IEC**	**HIC**	**HILIC**	**SEC**	**GE**	**CE**
RPC	[18,73,75–77]	[19,27,70,81,95–98,128]	[78,81]	[19,27]	[24,101,102,105]	[46,59–72]	
IEC							
HIC		[81]					
HILIC	[86,92]						
SEC							
GE							
CE	[21,22,55,102,113,116,119]				[21,105,111,125–127]	[64,100]	[123,124,129]

Table 1-2 Summary of multidimensional separation in TDP.

MDLC	Coupling strategy	1st separation technique	2nd separation technique	1st column	2nd column	Separation sample type	Sample starting amount	MS	Proteoform/protein count
RPLC-RPLC	Offline	High-pH RPLC	Low-pH RPLC	Waters XBridge BEH C4 (25 cm × 2.1 mm, 3.5 μm particles, 300 Å)	In-house packed capillary C5 (75 cm × 75 μm, 5 μm)	*E. coli* cell lysate	500 μg	LTQ Orbitrap Elite MS	886 proteoforms; 365 proteins[75]
	Offline	High-pH RPLC	Low-pH RPLC	Waters XBridge BEH C4 (25 cm × 2.1 mm, 3.5 μm particles, 300 Å)	In-house packed capillary C5 (75 cm × 75 μm, 5 μm)	*HeLa* cell lysate	1 mg	LTQ Orbitrap Elite MS	2778 proteoforms; 628 proteins[76]
	Online	High-pH RPLC	Low-pH RPLC	Waters Acquity UPLC BEH C4 column (150 mm × 1 mm, 1.7 μm, 300 Å)	Halo protein C4 column (30 mm × 3 mm, 2.7 μm, 1000 Å)	*E. coli* cell lysate	90 μg	Shimadzu LCMS-8050 triple quadrupole MS	N/A[73]
	Online	High-pH RPLC	Low-pH RPLC	In-house packed Waters XBridge BEH C4 (5 cm × 150 μm, 3.5 μm, 300 Å)	In-house packed C5, 75 cm × 75 μm, 5 μm)	*E. coli* cell lysate	5 μg	LTQ Orbitrap Elite MS	1507 proteoforms; 308 proteins[18]
	Offline	Low-pH RPLC	Low-pH RPLC	Agilent PLRP-S column (150 mm × 2.1 mm, 8 μm, 1,000 Å) with Trifluoroacetic acid (TFA) as the ion-pair reagent	Thermo Fisher Scientific C4 analytical column (Accucore™, 50 cm × 75 μm, 2.6 μm, 150 Å) with formic acid (FA) as ion-pair reagent	Caco-2 proteome	500 μg	FAIMS Orbitrap Tribrid MS	4,900 proteoforms; 1,250 protein[77]
HIC-RPLC	Offline	HIC	RPLC	PolyLC PolyPROPYL A column (100 mm × 4.6 mm, 3 μm, 1500 Å)	New Objective PicoFrit PLRP-S column (100 mm × 100 μm i.d., 5 μm, 1000 Å;	*E. coli* BL21 cell lysate	200 mg of *E. coli* cells	Q Exactive benchtop Orbitrap MS	20 proteins in HIC fraction 2[78]
	Offline	IEC and HIC	RPLC	PolyCAT WAX A column (200 mm × 4.6 mm i.d., 5 μm, 1000 Å) for	PLRP-S column (100 mm × 100 μm i.d., 5 μm, 1000 Å)	HEK 293 cell lysate	100 μL of HEK 293 cell lysate	Q Exactive benchtop Orbitrap MS	640 proteins in IEC fraction 3[81]

				IEC; PolyPROPYL A column (100 mm × 4.6 mm i.d., 3 µm, 1500 Å) for HIC					
RPLC-HILIC	Offline	RPLC	WCX-HILIC	Ultrapore RPSC C3 column (250 × 10 mm, 5 µm particle pore size, 30 nm pore size	PolyCAT A column (4.6 mm x 150 mm; 3 µm, 1000 Å)	*HeLa* cell histone H4	150 µg	Custom 8.5 T Quadrupole-FTMS hybrid	42 H4 proteoforms[86]
	Online	RPLC	WCX-HILIC	C5 (600 mm × 200 µm, 5 µm, 300 Å)	PolyCAT A (50 cm × 100 µm, 5 µm, 1000 Å)	*HeLa* core histones	7.5 µg	LTQ Orbitrap MS	708 proteoforms[92]
	Online	WCX-HILIC	RPLC	PolyCAT A column (150 mm × 75 µm, 5 µm particles, 1000 Å pore size)	In-house packed C18 column (100 mm × 50 µm, 3 µm, 300 Å, Phenomenex Jupiter particles)	*HeLa* core histones	1.5 µg	Q Exactive HF Orbitrap MS modified to incorporate 193 nm ultraviolet photodissociation (UVPD)	395±20 proteoforms [19]
	Online	WCX-HILIC	RPLC	PolyCAT A (800 mm × 200 µm, 5 µm; 1000 Å pore size)	Phenomenex Jupiter C5 (600 mm × 75 µm, 5 µm; 300 Å)	Histone H4	24 µg	LTQ Orbitrap MS	41 H4 proteoforms [27]
IEX-RPLC	Offline	IEX	RPLC	PL-SAX column (25 cm × 0.32 mm) with a wide pH range (10.5-3.5)	PS-DVB column (10 cm × 0.3 mm, 4000 Å)	plasma samples	10 µL	TOF MS	+200 protein peaks[98]
	Offline	WAX	RPLC	PolyWAX LP column (100 mm × 4.6 mm, 5 µm, 1000 Å)	In-house packed C5 column (80 cm × 75 µm, 5 µm, 300 Å, Phenomenex Jupiter particles)	Shewanella oneidensis cell lysate	4.35 mg	12 T FTICR	715 proteins [95]
	Online	SAX	RPLC	Thermo ProPac SAX-10 column (250 mm × 2 mm, 10 µm)	ACQUITY UPLC Protein BEH C4 column (150 mm × 2.1 mm, 1.7 µm, 300 Å)	Adeno-associated viruses (AAVs)	1 µL	Thermo Scientific Orbitrap Exploris 480 MS	13 proteoforms for AAV8 and AAV1 samples[96]

	Offline	SCX	RPLC	PolyLC PolySulfoethyl A^{tm} column (200 × 2.1 mm, 5 µm, 300 Å)	Thermo monolithic ProSwift RP-4H, 500 × 0.1 mm)	*Methanosarcina mazei Gö1* cell lysate	41.6±1.5 mg wet weight pellet (~3.44±0.48mg proteins)	Orbitrap Fusion Lumos mass spectrometer	1356 proteoforms; 353 proteins[97]
	Offline	sSEC	RPLC	PolyHEA column (200 mm × 9.4 mm, 3 µm, 1000-500-500 Å)	PLRP column (250 mm × 0.5 mm, 10 µm, 1000 Å)	Cardiac protein extract	200 µg	MaXis II Q-TOF	5360 proteoforms[24]
	Offline	SEC	RPLC	PolyHEA column (200 × 4.6 mm, 3 µm, 1000 Å)	PLRP capillary (250 mm × 0.250 mm, 5 µm, 1000 Å)	HEK293T cells and cardiac tissues	150 µg-360 µg	MaXis II Q-TOF	HEK293T: 1688 proteoforms, + 188 proteins; cardiac tissues: 1781,124 proteins[101]
	Offline	SEC	RPLC	BioSep-SEC-S2000, 600 × 21.2 mm column	In-house packed C5 column (80 cm × 150 µm, 5 µm, 300 Å, Phenomenex Jupiter particles)	*Shewanella oneidensis* cell lysate	22.5 mg	FTICR	297 proteins[102]
size-RPLC	Offline	sIEF and GELFrEE	RPLC	sIEF buffer: 8M urea, 2M thiourea, 50mM DTT, 1% w/v Biolyte 3/10 carrier ampholytes GELFrEE: 12% T resolving	New Objective PLRP-S column (10 cm × 75 µm, 5 µm, 1000 Å)	*HeLa* S3 cells	0.5-1 mg	12 T LTQ FT Ultra MS	+3000 proteoforms; 1043 proteins[46]
	Offline	GELFrEE	RPLC	15% T resolving	New Objective PLRP-S column (10 cm × 75 µm, 5 µm, 1000 Å)	nuclear and cytosolic extracts from Hela S3 cells	~100 µL	12T or 14.5T LTQ-FT-Ultra MS	N/A[61]

Offline	GELFrEE	RPLC	12% or 10% T resolving	New Objective PLRP-S column (10 cm × 75 µm, 5 µm, 1000 Å)	Peripheral blood mononuclear cells (PBMCs)	300 µg	Velos Orbitrap Elite MS	2905 proteoforms; 344 proteins[62]
Offline	GELFrEE	RPLC	15% T resolving	home-packed C4 column (5 cm × 0.5 mm, 5 µm, 300 Å)	Six protein standards	60 µg	MicroTOF MS	428 proteins[63]
Offline	GELFrEE	RPLC and CZE	10% T resolving	PLRP-S column (25 cm × 75 µm, 5 µm, 1000 Å) or 90 cm, 30 µm i.d., CESI 800 Plus (Sciex), neutrally coated	Human tissue samples	300 µg	Orbitrap Eclipse Tribrid MS or a Fusion Lumos Orbitrap Tribrid MS	8784 proteoforms; 343 proteins[64]
Offline	GELFrEE	RPLC	10% T resolving	Home-packed PLRP-S column (5 µm, 1000 Å, Phenomenex media)	Nuclear and cytosolic extracts of E. granulosus protoscolece s	250 µg	Orbitrap Elite	207 proteoforms; 186 proteins [65]

	Offline	GELFrEE and sIEF	RPLC	GELFrEE: 10% T resolving. sIEF buffer: 7M urea,2M thiourea, 30 mM DTT, and 2% Bio-Lyte 3/10 carrier ampholytes, pH range 3-10	Home-packed PLRP-S column (20 cm × 75 µm, 5 µm, 1000 Å, Phenomenex media)	venom of Ophiophagus hannah (king cobra)	300 µg	Orbitrap Elite and Q-Exactive HF	184 proteoforms; 131 proteins[66]
	Offline	GELFrEE	RPLC	10% T resolving or 8% T resolving	Home-packed PLRP-S column (10 cm × 75 µm, 5 µm, 1000 Å, Phenomenex media) or monolithic Thermo Dionex RP-4H analytical columns (50 cm × 100 µm)	IMR90 human fibroblasts	30 µg	Orbitrap Elite	1577 proteoforms; 1180 proteins[67]
	Offline	sIEF and GELFrEE	RPLC	sIEF buffer: 4M urea,2M thiourea, 50 mM DTT, and 1% Bio-Lyte 3/10 carrier ampholytes GELFrEE: 12% or 10% T resolving	New Objective PLRP-S column (10 cm × 75 µm, 5 µm, 1000 Å)	Human cell line H1299	500 µg-3 mg	Orbitrap Elite MS	1220 proteins[68]
	Offline	sIEF and GELFrEE	RPLC	sIEF buffer: 4 M urea, 2 M thiourea, 50 mm DTT, 1% (w/v) Biolyte 3/10 carrier ampholytes GELFrEE: 15% T resolving	New Objective PLRP, 100 mm × 75 µm, 5 µm, 300 Å)	*Saccharomyces cerevisiae* cell lysate	3 mg	12T LTQ-FT-Ultra MS	10-60 proteins (70-80 kDa) per LC run[69]

	Offline	IEX and GELFrEE	RPLC	IEX: Pierce strong anion exchange spin columns (Thermo Fisher Scientific) GELFrEE: 8% or 12% T resolving	ProSwift RP-4H column (50 cm × 100 μm) or PLRP-S (20 cm × 75 μm, 5 μm, 1000 Å)	human blood or bone marrow	0.4-100 million cells	Orbitrap Elite or Modified Q-Exactive HF or Orbitrap Fusion Lumos or 21 Tesla FT-ICR	17103 proteoforms; 1690 proteins[70]
	Offline	PEPPI	RPLC	Thermo Fisher Scientific precast NuPAGE 4−12% Bis−Tris gradient gels	Agilent PLRP-S (17.5 cm × 150 μm, 5 μm, 1000 Å)	*Escherichia coli* strain MG1655	∼80−110 μg	21 T FTICR MS	+1000 proteoforms (≤50 kDa)[71]
	Offline	PEPPI-AnExSP/PEPPI-MCW	RPLC	Thermo Fisher Scientific precast NuPAGE 4−12% Bis−Tris gradient gels	Thermo Fisher Scientific C4 column (50 cm × 75 μm, 2.6 μm, 150 Å)	human protein extract (Promega)	40 μg per well	Fusion Lumos Tribrid MS equipped with the FAIMS Pro Interface	PEPPI-AnExSP: 6728 proteoforms; PEPPI-MCW: 3088 proteoforms[72]
	Offline	ALS-PAGE	RPLC	Bio-Rad model 491 Prep Cell with 12% T resolving gel and 0.1% ALS	Waters C4 Exterra RPLC column (100 mm × 4.6) or Eichrom Technologies NPS column (14 × 4.6 mm)	Saccharomyces cerevisiae cell lysate	3 mL	Quadrupole-FTMS Hybrid	N/A[59]
	Offline	ALS-PAGE	RPLC	Bio-Rad Mini Prep Cell with 12% T resolving gel and 0.1% ALS	Vydac C4 column (25 cm × 320 μm) and New Objective C4 column (10 cm × 75 μm)	*Saccharomyces cerevisiae* (S288C) cell lysate	300 μL	Custom Q-FTMS	37 yeast protein forms between 6 and 30 kDa[60]
CIEF/CZE-CZE	Online	CIEF	CZE	50 cm, 50 μm i.d., 363 μm o.d. , Ultratrol™ dynamic pre-coat LN coated capillary	87 cm, 50 μm i.d., 363 μm o.d. , Ultratrol™ dynamic pre-coat LN coated capillary	hemoglobin and glycated hemoglobin	N/A	micrOTOF-Q instrument	N/A[123]

	Online	CIEF	CZE	60 cm 50 μm i.d., 360 μm o.d. PVA capillary	80 cm 50 μm i.d., 360 μm o.d. PVA capillary	charge variants of mAb	18.75 or 75μg	compactquadrupole–quadrupole time of flight	N/A[124]
	Online	CZE	CZE	60cm, 50 μm i.d., 360 μm o.d., bare	80cm, 50 μm i.d., 360 μm o.d., PVA coated	de/glycosylation of trastuzumab	18 mg/mL	compact™ quadrupole–quadrupole time-of-flight	N/A[129]
RPLC-CZE/CIEF	Offline	RPLC	CIEF	C4 column (250 × 10 mm, 10 mm, 300 Å, Phenomenex Jupiter)	80 cm × 50 μm i.d., 190 μm o.d., HPC coated	*Shewanella oneidensis* proteome	4 mg	FTICR	166 proteins[102]
	Offline	RPLC	CZE	C4 column (8 cm × 250 μm, 5 μm, 300 Å)	90 cm, 30 μm i.d., 150 μm o.d., PEI coated	*Pyrococcus furiosus* cell lysate	270 ng	Orbitrap Elite	291 proteoforms; 134 proteins[55]
	Offline	RPLC	CZE	C5 column (250 × 4.6 mm, 5 μm, 300 Å)	60 or 80 cm, 50 μm i.d., 150 μm o.d., LPA coated	Red Star Active Dry Yeast	320 μg	Q Exactive HF	580 proteoforms; 180 proteins[113]
	Offline	SEC, RPLC, SEC-RPLC	CZE	SEC: 300 × 4.6 mm, 500 Å RPLC: C4 column (250 × 2.1 mm)	100 or 70 cm, 50 μm i.d., 360 μm o.d., LPA coated	SW480 and SW620	100 μg-2 mg	Q-Exactive HF	23622 proteoforms; 2332 proteins[21]
	Online	RPLC	CZE	ZORBAX 300SB C8 (50 ×0.075 mm, 3.5 μm)	10cm, 50 μm i.d., PVA coated	Standard protein mixture	50 ng	QTOF–MS	N/A[116]
	Online	RPLC	CZE	PLRP-S column (250 mm × 75 μm, 5 μm, 1000 Å)	70 cm, 50 μm i.d., <150 μm o.d., PDADMAC-PMA or DEAEDq-PMA-coated	intact human alpha-1-acid glycoprotein	50 μg/mL	QTOF-MS	368 glycoforms[119]
size-CZE	Offline	SEC and RPLC	CZE	SEC: 300 × 4.6 mm, 3 μm, 300 Å RPLC: C4 column (250 × 2.1 mm, 3.0 μm, 300 Å)	1m, 50 μm i.d., 360 μm o.d., LPA coated	*E. coli* cell lysate	1 mg	Q-Exactive HF	5705 proteoforms; 850 proteins[22]

Offline	SEC	CZE	300 × 4.6 mm, 3 µm, 300 Å	1m, 55 µm i.d., 360 µm o.d., LPA coated	*E. coli* cell lysate	663 µg	Orbitrap Fusion Lumos	957 proteoforms; 253 proteins[125]
Offline	SEC	CZE	300 × 4.6 mm, 3 µm, 300 Å	1m, 50 µm i.d., 360 µm o.d., LPA coated	zebrafish brain	700 µg	Orbitrap Fusion Lumos	600 proteoform; 369 proteins[126]
Offline	SEC	CZE	300 × 4.6 mm, 2.7 µm, 300 Å	1m, 50 µm i.d., 360 µm o.d. LPA coated	*E. coli* cell lysate	600 µg	Q-Exactive HF	672 proteoform; 144 proteins[127]
Offline	SEC	cIEF	Agilent, 300 × 4.6 mm, 3 µm, 150 Å	1m, 50 µm i.d., 360 µm o.d. LPA coated	*E. coli* cell lysate and zebrafish brain lysate	200 µg for *E. coli* and 500 µg for zebrafish brain proteins	Q-Exactive HF	*E. coli* cell lysate:1896 proteoforms; 365 proteins among all SEC fractions Zebrafish brain lysate:171-1260 proteoforms and 51-216 proteins from different SEC fractions[111]
Offline	GELFrEE	CZE	8% cartridge	30 cm, 50 µm i.d., 150 µm o.d., PA coated	*Pseudomonas aeruginosa PA01* cell lysate	400 µg	Q-Exactive	30 proteins (30-80kDa)[100]
Offline	GELFrEE	CZE and RPLC	10% T resolving	90 cm, 30 µm i.d., CESI 800 Plus (Sciex), neutrally coated or PLRP-S column (25 cm × 75 µm, 5 µm, 1000 Å)	Human tissue samples	300 µg	Orbitrap Eclipse Tribrid MS or a Fusion Lumos Orbitrap Tribrid MS	8784 proteoforms; 343 proteins[64]

1.3 QUANTITATIVE TOP-DOWN PROTEOMICS

Quantitative top-down proteomics (TDP) enables the quantification of intact proteins at the proteoform level and allows for the characterization of physiological differences between biological samples. Quantitation of intact proteoforms in top-down proteomics can elucidate proteoforms that are involved in the biological pathways, determine the effect of proteoform expression on disease states, and potentially be used for biomarker discovery and drug target identification. There are three major approaches that have been used for quantitative TDP including label-free quantitation, metabolic label-based quantitation, and isobaric chemical tag-based quantitation.[15]

1.1.1 Label-free quantitation (LFQ)

Label-free quantitation (LFQ) quantifies intact proteoforms in different biological samples by direct comparison of proteoform intensity between individual LC-MS runs. LFQ relatively quantifies intact proteoforms at the MS1 level using the extracted ion chromatogram (EIC) peak area or MS spectra abundance, ***Figure 1-5***.[142] There are generally 4 steps required for LFQ[15]: (1) individual LC-MS runs will be collected; (2) quantification of the mass features and identification of the proteoforms using bioinformatic tools such as Biopharma Finder, Proteome Discoverer, TopPIC Suite, *etc.*; (3) relative quantitation using the intensities of mass features/proteoforms. Statistical analysis such as *t-test* will also be applied at this step to evaluate the statistical significance of changes in proteoform abundance between experimental conditions.

It has been reported that the proteoform intensities among different LC-MS runs vary with a coefficient of variation ranging from 8-20% for a simple mixture of standard proteins.[143] The variation might become larger for complex biological sample because of the increased sample complexity and wider dynamic range. To decrease the run-to-run variation, intensities from total ion chromatograms (TICs) or based peak chromatograms (BPCs) among different samples are commonly normalized for relative quantitation. Alternatively, standard protein can be added into the samples prior to analysis and the intensity of the standard protein can be utilized for normalization among different samples. On the other hand, if expression of different proteoforms within the same run can be compared between experimental conditions, there is no need for mass feature matching or normalization between runs.[144]

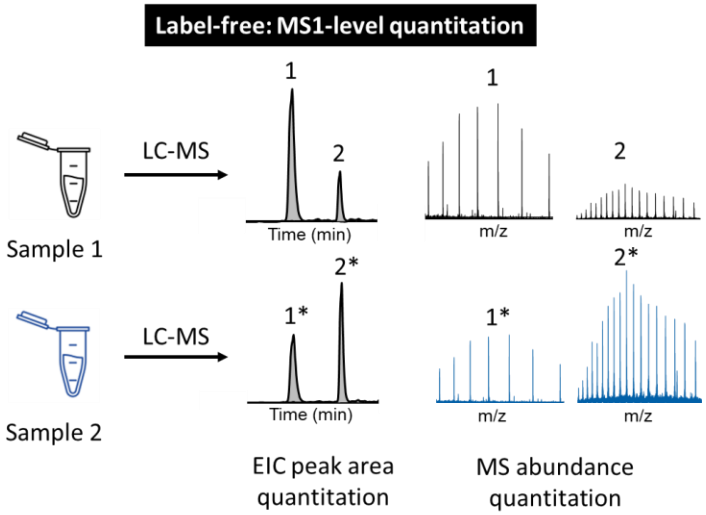


Figure 1-5 Label-free quantitation in top-down proteomics.

Overall, since LFQ does not require labeling and can be applied to any protein solution or complex biological sample, LFQ is the simplest quantification approach and has been commonly utilized in TDP. However, label-free quantitation

is sensitive to run-to-run instrument variation and because individual LC-MS runs must be collected for each sample. This process is generally time-consuming and especially challenging when dealing with a large number of runs. Additionally, inconsistent sample handling will further decrease the reproducibility of the quantitation. Moreover, LFQ cannot typically be coupled with multidimensional (MD) separation approaches for relative quantitation since the proteoforms of interested may be eluted in multiple first dimension fractions.

1.1.2 Metabolic labeling-based quantitation

In the metabolic labeling-based quantitation approach, cells are cultured with isotopically labeled media (*i.e.*, "heavy" culture) or typical media (*i.e.*, "light" culture) so that differentially isotopically labeled proteoforms are expressed under different experimental conditions. The isotopically labeled sample and unlabeled sample will be mixed, and one single LC-MS run will be performed. The relative quantitation will be completed at the MS1 level using the intensity ratio of labeled and unlabeled proteoforms, ***Figure 1-6***.[38] Several metabolic labeling techniques have been developed and applied in TDP, such as stable isotope labeling by amino acids in cell culture (SILAC)[145–148], neutron encoding (NeuCode) SILAC[149–151], and tunable intact protein mass increases (TIPMI)[152]. Other metabolic labeling approaches that culture cells by using heavy isotopes such as labeled nitrogen sources have also been applied in TDP.[153]

Specifically, in SILAC-based quantitation, cells are cultured in the media containing isotopically labeled amino acids such as ^{13}C or ^{15}N labeled lysine, leucine, and tryptophan[154]; the control cells are cultured in the normal media

without isotopically labeled amino acids. Several generations of cell cultures are needed to ensure the efficient labeling.[147] After protein extraction, "heavy" proteoforms (labeled proteoforms) and "light" proteoforms (unlabeled proteoforms) will be combined and analyzed for LC-MS. The intensity ratio at the MS1 level of the "heavy" proteoforms and "light" proteoforms will be extracted and used for the relative quantitation. SILAC has been commonly utilized for peptide quantitation in bottom-up proteomics studies, and it has also been successfully extended to the relative quantitation of intact proteins using top-down protoemics.[146,155]

However, SILAC is commonly used to compare two conditions and may lack multiplexing. Additionally, the incorporation of isotopes on amino acids complicates the isotopic distribution and further challenge data analysis.[156] NeuCode SILAC was developed using different isotopologues of the same amino acid for cell culture to provide more multiplexing.[151] However, the mass difference is relatively small (~ mDa scale), and high-resolution MS instruments are required for NeuCode SILAC-based quantitation approach to be able to resolve the subtle mass difference. TIPMI can be applied to the study of animal models since the isotopically labeled sugar or deuterated water can be added into the media or feedstock so that proteoforms can be differently labeled.[152]

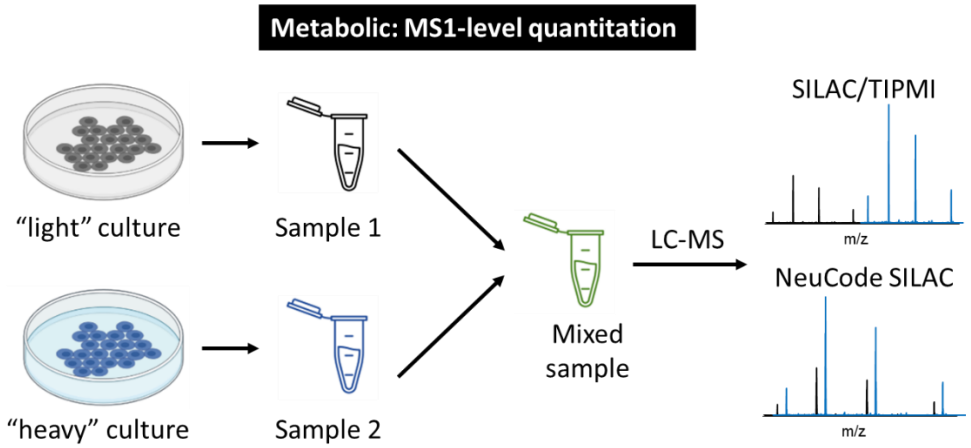


Figure 1-6 Metabolic labeling-based quantitation in top-down proteomics. Figure was created with BioRender.com.

Overall, metabolic labeling-based quantitation enables the characterization of the proteoform in *vivo* for relative quantitation between different conditions. However, metabolic labeling is restricted to the samples that can be cultured or grown in the laboratory, which limits its application to clinical samples that can not be metabolically labeled. Also, MS spectra will be much more complicated because of the introduction of light/heavy isotope labeling; therefore, it relies on advanced MS instruments with high resolving power. The complicated spectra could further complicate the data analysis process to the relative quantitation application of intact proteins using top-down proteomics.

1.1.3 Isobaric chemical tag-based quantitation

Another alternative quantitation approach that has been applied in top-down proteomics (TDP) is isobaric chemical tag labeling, which covalently labels proteins at specific amino acids.[15] All isobaric tags share similar physical and chemical properties as well as the same total masses, with different distributions of isotopes in each tag channel, ***Figure 1-7***.[157] Many isobaric chemical tag labeling

techniques have been developed for relative and absolute quantitation (iTRAQ)[158], tandem mass tag (TMT)[159,160], cysteine-reactive tandem mass tag (iodoTMT)[131,161], and *N*, *N*-dimethylleucine (DiLeu)[162–164]. A standard isobaric tag design commonly includes three parts: (1) a reactive group that targets specific amino acid residues such as lysine and N-terminus, (2) a reporter group that contains a different number of isotopes for different tag channels, and (3) a balance group that links the reactive group and reporter group to ensure the identical total mass for different tag channels. Different samples are labeled with different chemical tags, and the labeled samples are mixed into one single tube for LC-MS analysis. The bond between the reporter group and the balance group will be cleaved during MS2 fragmentation and the ratio of the reporter ion intensities will be used for relative quantitation between different samples. Since isobaric chemical tag labeling quantifies proteoforms at the MS2 level, multiple proteoform masses will not complicate MS1 spectra, which increases quantitation sensitivity compared with metabolic labeling techniques. These isobaric chemical tags enable the quantification of multiple samples (*e.g.*, up to 18 samples with TMTpro 18-plex and 21 samples using 21-plex DiLeu[163]) in a single run, allowing for simultaneous quantification and identification at the MS2 level.

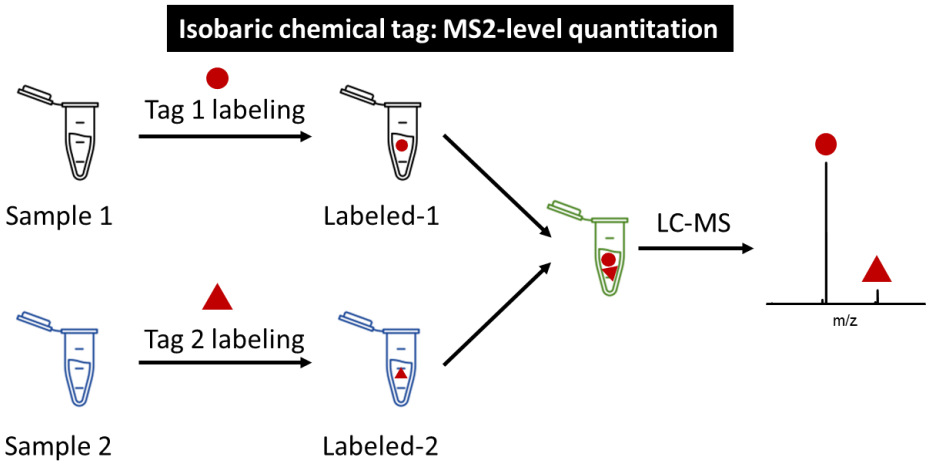


Figure 1-7 Isobaric chemical tag-based quantitation in top-down proteomics.

Isobaric chemical tag labeling has been widely applied to bottom-up proteomics (BUP) for the relative quantification of peptides in complex biological samples.[157,165,166] The application of isobaric chemical tag labeling in intact proteoforms has been limited majorly due to the nature of intact proteins that (1) protein tend to precipitate out under the chemical labeling condition since organic solvent is commonly required to avoid the isobaric chemical tag hydrolysis; (2) the production of incomplete or labeling of unintended residues.[157,167] Even with these challenges, protein-level labeling is still beneficial over peptide-level labeling to some extent. For example, the variation from downstream sample preparation steps in peptide-level labeling such as protein digestion and desalting can be reduced; protein-level labeling also enables the application of intact protein separation approaches such as MD separations for intact proteins to improve the proteome coverage.

iTRAQ was first applied to simple intact protein mixtures for relative quantitation[168], then TMT was also applied to protein-level labeling in standard protein or simple protein mixtures for the quantitative analysis.[168–171] Only a few

publications have applied isobaric chemical tag labeling for quantitative analysis in complex biological samples using TDP because of the increased sample complexity, protein precipitation, and production of side products.[130–133,172] To tackle the protein precipitation issue, our group has recently developed a "filter-SEC" technique that encompasses 100 kDa molecular weight cutoff spin filtration (MWCO) and size exclusion chromatography (SEC) for the removal of larger proteoforms (> 30 kDa), which allows for the relative quantitation of smaller intact proteoforms in complex biological samples using protein-level TMT labeling.[130] Later, Winkels and the coworkers developed a labeling strategy using thiol-directed TMT (*e.g.,* cysteine labeling) for the quantitation of cysteine-containing proteoforms (< 30 kDa) after prefractionation using GELFrEE or high-pH RPLC, followed by low pH RPLC-MS/MS analysis.[131]

Overall, isobaric chemical tag labeling-based quantitation allows for simultaneous identification and quantification at the MS2 level without complicating the MS1 level. The successful application of protein-level labeling in standard proteins, simple protein mixtures, and complex biological samples also enables the protein-level quantification in TDP. However, there are still some challenges existing that need to be addressed, such as the formation of side products that increase the sample complexity and decrease the signal-to-noise ratios and insufficient separation due to the high complexity and wide dynamic range of proteoforms in complex samples.

1.4 DISSERTATION SYNOPSIS

This dissertation presents the development of an intact protein-level TMT labeling platform for high-throughput and deep quantitative top-down proteomics. Some challenges in protein-level TMT labeling in complex biological samples include production of side products, inefficient quantification and identification, and insufficient separation for TMT-labeled complex samples.

First, to reduce the production of side products and maximize the percentage of completely labeled proteoforms, a systematic optimization of protein-level TMT labeling conditions was performed (***Chapter 2***)[132], and a benchmarking protocol for protein-level TMT labeling was demonstrated (***Appendix A – A Benchmarking Protocol for Intact Protein-Level Tandem Mass Tag (TMT) Labeling for Quantitative Top-Down Proteomics***)[172]. Parameters that were optimized include TMT-to-protein mass ratio, pH/concentration of quenching buffer, protein concentration, reaction time, and reaction buffer. Method optimization improved the labeling efficiency to 90% for TMT-labeled *E. coli* cell lysate and ~86% for TMT-labeled *HeLa* cell lysate and improved proteoform quantitation.

Isobaric chemical tag-based protein quantitation relies on the efficient and accurate measurement of low-mass reporter ions generated in MS2 fragmentation; therefore, optimal fragmentation energy for TMT-labeled complex protein mixture is essential to generate high-intensity reporter ion peaks for accurate quantification, as well as adequate protein fragment ions for confident identification.[133] We have optimized the fragmentation energies (higher-energy collisional dissociation (HCD) on Orbitrap Exploris 240 mass spectrometer) for TMT-labeled complex intact

proteoforms to achieve a balance between accurate quantitation and confident identification (***Chapter 3***).[133] We found that stepped normalized collision energy (SNCE) schemes with NCEs between 30% and 50% provided a balance between high reporter ion intensity that was sufficient for accurate and sensitive quantification and adequate fragment ions for confident identification.

Finally, we applied MD separation techniques for deep proteome quantitation of TMT-labeled intact protein lysate. Due to the high complexity and wide dynamic range of the TMT-labeled complex sample, one dimensional (1D) separation using LC or capillary electrophoresis **(**CE) may not provide sufficient proteome coverage. Therefore, multidimensional (MD) separation platforms that couple different orthogonal separation methods are commonly utilized to improve the separation resolution and identification (***1.2.2 Multidimensional separations in top-down proteomics***).

Successful application of protein-level TMT labeling in complex samples not only allows identification and quantification of intact proteoforms with high multiplexing capacity and eliminates variability from downstream sample preparation (*e.g.,* peptide digestion) but also enables protein-based fractionation such as multidimensional separation.[130,173] We have demonstrated the deep quantitative analysis of human cell lysate using top-down TMT labeling[130] and online 2D high-pH/low-pH LC separation[18] (***Chapter 4***). In summary, we observed ~5600 quantifiably proteoforms from 20 µg TMT-labeled intact protein lysate. This analysis demonstrated the first application of isobaric chemical tag labeling for

deep, quantitative top-down proteome profiling using MD separation with low microgram sample amounts.

Section 1.2 in Chapter 1 of "Separation techniques in top-down proteomics" are adapted from:

Yanting Guo[*], Kellye A. Cupp-Sutton[*], Zhitao Zhao, Samin Anjum, Si Wu[*] "Multidimensional Separations in Top-Down Proteomics." Analytical Science Advances (***manuscript accepted***)"

[*]Authors contributed equally

Author contributions: Yanting Guo: Methodology, Literature summary, validation, Writing - original draft. Kellye A. Cupp-Sutton: Methodology, Literature summary, validation, Writing - review & editing. Zhitao Zhao: Literature summary for CZE-based separation. Samin Anjum: Literature summary for part of LC-based separation. Paper Si Wu: Conceptualization, Validation, Supervision, Funding acquisition, Writing – review & editing.

Copyright information for Section 1.2 in Chapter 1:

This chapter is unpublished research from the Wu lab at the University of Oklahoma Department of Chemistry and Biochemistry. All the figures and tables were made by Yanting Guo.

Chapter 2 Optimization of protein-level tandem mass tag (TMT) labeling conditions in complex samples with top-down proteomics

2.1 ABSTRACT

Isobaric chemical tag labels (*e.g.*, iTRAQ and TMT) have been extensively utilized as a standard quantification approach in bottom-up proteomics, which provides high multiplexing capacity and enables MS2-level quantification while not complicating the MS1 scans. We recently demonstrated the feasibility of intact protein TMT labeling for the identification and quantification with top-down proteomics of smaller intact proteoforms (<35 kDa) in complex biological samples through the removal of large proteins prior to labeling. Still, the production of side products during TMT labeling (*i.e.*, incomplete labeling or labeling of unintended residues) complicated the analysis of complex protein samples. In this study, we systematically evaluated the protein-level TMT labeling reaction parameters, including TMT-to-protein mass ratio, pH/concentration of quenching buffer, protein concentration, reaction time, and reaction buffer. Our results indicated that: (1) high TMT-to-protein mass ratio (*e.g.*, 8:1, 4:1), (2) high pH/concentration of quenching buffer (pH > 9.1, final hydroxylamine concentration >0.3%), and (3) high protein concentration (*e.g.*, > 1.0 μg/μL) resulted in optimal labeling efficiency and minimized production of over/underlabeled side products. >90% labeling efficiency was achieved for *E. coli* cell lysate after optimization of protein-level TMT labeling conditions. In addition, a double labeling approach was developed for efficiently labeling limited biological samples with low concentrations. This research provides practical guidance for efficient TMT

labeling of complex intact protein samples, which can be readily adopted in the high-throughput quantification top-down proteomics.

2.2 INTRODUCTION

The ability to identify and relatively quantify intact proteoforms using quantitative top-down proteomics techniques has enabled the study of diverse biological processes that are mediated by post-translational modifications (PTMs).[11,16] Various quantitative techniques have been adapted and applied to top-down proteomics, including label-free quantitation, metabolic labeling, and isobaric chemical labeling. However, isobaric chemical tag labeling methods including relative and absolute quantitation (iTRAQ)[158], tandem mass tags (TMT)[159,160], and N, N-dimethylleucine (DiLeu) isobaric tags[162,174] have been widely applied to the quantification of peptides and proteins in bottom-up proteomics studies, the application to top-down proteomics has been limited.

Previously, intact protein-level isobaric tag labeling has been coupled with bottom-up proteomics methods to improve proteome coverage or characterize protein structures. iTRAQ was first applied to protein-level labeling using standard proteins by Wiese *et al.* to measure the labeling efficiency of intact proteins using matrix-assisted laser desorption ionization-time-of-flight (MALDI-TOF) mass spectrometry analysis.[175] Sinclair *et al.* further applied protein-level TMT labeling to human serum samples; the TMT labeled proteins were separated using strong anion exchange (SAX) chromatography and digested to achieve a higher proteome coverage.[169] Nie *et al.* applied isobaric chemical tag labeling for serum glycoprotein quantification and evaluated the effect of various digestion enzymes, isobaric

chemical tags, and organic solvents on the identification and quantification of peptides.[167] TMT was also applied to whole cells for protein quantification with combination of enrichment using anti-TMT antibody.[176] Kaiwen et al. also developed a native protein TMT method to profile Lys accessibility and applied it to investigate structural change in Alzheimer's disease brain specimens.[177]

Top-down proteomics on protein-level isobaric chemical tag labeled standard proteins has been previously performed by Hung *et al;*[171] however, the application of protein-level isobaric tag labeling to complex samples has been limited. One major challenge has been protein precipitation during labeling due to the presence of organic solvent which limits labeling efficiency. We previously demonstrated that the removal of large proteins by coupling molecular weight cutoff (MWCO) and size exclusion chromatography prevents intact proteins from precipitating under labeling conditions and allows for the identification and quantification of smaller proteoforms (≤ 35 kDa) in *E. coli* cell lysate.[130] Moreover, the application of thiol-directed isobaric labeling for quantification and identification using top-down proteomics has also been demonstrated for *E. coli* and combined *E. coli*/yeast samples.[173]

Another challenge for protein-level TMT labeling is the production of side products (*i.e.*, overlabeled and underlabeled species) during the labeling reaction. Overlabeling (labeling of residues other than lysine or the N-terminal) may occur on residues with side chains containing hydroxyl groups (Ser, Thr, Tyr). Underlabeling (missing labels) may occur on the N-terminal or lysine residues. Over and underlabeling does not have a significant impact on quantification results;

however, the side products will decrease the signal-to-noise ratio and increase the sample complexity.[171,175]

Recently, the TMT labeling condition for peptides was optimized by Zecha *et al.* and the impact of labeling reaction parameters was evaluated to achieve >99% labeling efficiency.[175] Considering protein-level labeling is more complicated than peptide-level labeling, it is necessary to optimize protein-level labeling conditions to increase the labeling efficiency. In this study, we systematically optimized the parameters of the protein-level TMT labeling reaction using to maximize the production of the completely labeled species. Overall, we evaluated the effect of TMT-to-protein ratio (w/w), labeling reaction conditions (*i.e.*, quenching buffer, reaction time, and reaction buffer), and protein concentration on protein-level TMT labeling efficiency. Our results demonstrate an optimized condition for intact protein-level TMT labeling of complex samples. We also innovated a double labeling strategy as an alternative approach to achieve sufficient labeling efficiency for mass limited samples that did not have adequate concentrations to implement the optimized labeling conditions (*e.g.*, clinical samples).

2.3 EXPERIMENTAL

2.3.1 Chemicals and materials.

All reagents were purchased from Sigma-Aldrich (St. Louis, MO, USA) unless noted otherwise. Acetonitrile (ACN), isopropanol (IPA), trifluoroacetic acid (TFA), and water used for LC mobile phases were LC-MS grade. Tris (2-carboxyethyl) phosphine hydrochloride (TCEP), Pierce™ BCA Protein Assay Kit, molecular weight cutoff spin filters (10K and 100K MWCO), TMT 6-plex Isobaric

Label Reagent set (90061) and TMT zero Isobaric Label Reagent set (90067) were obtained from Thermo Fisher (Waltham, MA, USA).

2.3.2 *E. coli* cell lysate preparation.

E. coli K12 cells were inoculated in 250 mL sterilized LB medium at 37 ^{o}C, 250 rpm for 8 hours. Then 250 mL *E. coli* was transferred into 2 L LB solution and incubated at 37 ^{o}C, 250 rpm for 12 hours. The cells were collected and pelleted through centrifugation at 10,000 × g, 4 ^{o}C for 20 minutes. Cell pellets were resuspended in 25 mM ammonium bicarbonate (ABC) buffer at the ratio of 5:1 (w/w) with addition of 0.1% (v/v) PMSF, as described previously.[178] Cells were lysed using high pressure Avestin C3 EmulsiFlex homogenizer (ATA Scientific Instruments, Australia) with the operating pressure between 1500 psi and 2000 psi. The lysate was cooled on ice to offset the slow heat buildup during homogenization. After homogenization, the cell lysate was centrifuged at 10,000 × g, 4 ^{o}C for 40 minutes to remove the insoluble debris. Protein concentration was measured using Pierce™ BCA Protein Assay Kit. The *E. coli* lysate was aliquoted and stored at -80 °C.

2.3.3 *HeLa* cell culture and cell lysate preparation.

HeLa cells were grown and lysed as previously described.[76] Briefly, *HeLa* cells were cultured in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 2% penicillin-streptomycin under 5% CO_2 at 37 °C. *Hela* cells were washed with ice-cold PBS buffer and scratched off, followed by centrifugation at 2, 500×g, 4 °C for 20 minutes, resuspended into lysis buffer (1 μM PMSF and 20 mM NaF in PBS, pH = 7.4), and lysed using a sonicator.

Cell lysate was centrifuged at 15,000 ×g, 4 °C for 30 min and the supernatant was collected. Protein concentration was measured by Pierce™ BCA Protein Assay Kit. The *HeLa* cell lysate was then aliquoted and stored at -80 °C.

2.3.4 Protein-level TMT labeling.

Spin filters were washed using the labeling buffer (100 mM TEAB, pH=8.5 or 100 mM HEPES, pH=8.5) prior to protein filtration. Cell lysates were centrifuged at 12,000 × g and 4 ^{o}C for 20 minutes filtered in 100 kDa MWCO spin filters to remove high molecular weight proteins. The low MW filtrate was then concentrated using a pre-washed 10 kDa MWCO spin filter for the removal of small proteins and protein concentration by centrifugation at 12,000 × g and 4 ^{o}C for 15 minutes. Right before labeling, the protein sample was diluted to 1 μg/μL using 6 M fresh urea in the labeling buffer to the final urea concentration of 3 M. 800 μg proteins were reduced by adding 160 μL 0.5 M TCEP (incubating for 15 minutes at room temperature), and then alkylated by adding 215 μL of 375 mM IAA (incubating at room temperature at dark for 30 minutes). The reduced and alkylated proteins were then buffer-exchanged and concentrated with labeling buffer using a 10 kDa MWCO spin filter to the appropriate concentration for TMT labeling. TMTsixplex reagent or TMTzero reagent was used for the optimization of intact protein-level labeling. Reactant volumes, concentrations, and reaction conditions for each experimental condition are specified in ***Table 1***. The labeled sample was centrifuged at 12,000 × g and 4 ^{o}C for 30 minutes before LC-MS/MS analysis to remove any precipitation.

For TMT labeling *of HeLa* cell lysate, 100 kDa MWCO filtered *HeLa* cell lysate was reduced and alkylated as described above. This samples were then buffer exchanged into 100 mM TEAB buffer, pH 8.5 using a 10 kDa MWCO filter. Protein concentration was diluted to 1.5 µg/µL before TMT labeling. *HeLa* cell lysate was aliquoted into 6 tubes and labeled with TMTsixplex individually at the optimal conditions: TMT reagent and proteins were incubated with a 4:1 TMT-to-protein mass ratio at room temperature for 1 hour, then the same amount of TMT reagent was added again, and the sample was incubated for an additional hour to complete double labeling. The TMT labeling reaction was quenched using hydroxylamine to a final concentration of 1.2% (pH 9.5). Six TMT-labeled samples were mixed with a mass ratio of 5:5:2:2:1:1 for LC-MS/MS analysis.

Table 2-1 Experimental design of protein-level TMT labeling optimization

Group No.	Description	TMT reagent	TMT-to-protein ratio (wt/wt)	Initial protein concentration (g/L)	Protein quantities (μg)	TMT quantities (μg)	Reaction volume (μL)	Protein conc. during labeling (g/L)	TMT conc. during labeling (mM)	Varying parameters
0	Reproducibility	TMTzero	4:1	1.5	50	200	43.3	1.2	13.6	N/A
1	TMT-to-protein ratio	TMTsixplex	1:1	1.2	45	45	55.5	0.8	2.3	TMT final conc. = 2.3 mM
			2:1			90			4.5	TMT final conc. = 4.5 mM
			4:1			180			9.1	TMT final conc. = 9.1 mM
			8:1			360			18.2	TMT final conc. = 18.2 mM (Control*)
2	Quenching buffer	TMTzero	4:1	1.5	50	200	43.3	1.2	13.6	0.3% final conc. of hydroxylamine
										0.6% final conc. of hydroxylamine
										1.2% final conc. of hydroxylamine (Control*)
										50 mM final conc. of Tris
3	Protein conc. (single label)	TMTzero	4:1	0.5	50	200	110	0.5	5.4	Final reaction volume = 110 μL
				1.0			60	0.8	9.8	Final reaction volume = 60 μL
				1.5			43.3	1.2	13.6	Final reaction volume = 43.3 μL
				2.0			35	1.4	16.9	Final reaction volume = 35 μL (Control*)
4	Protein conc. (double label)	TMTzero	4:1[#] & 8:1[*]	0.5	50	200[#] & 400[*]	120	0.5[#] & 0.4[*]	5.4[#] & 9.8[*]	Final reaction volume = 120 μL
				1.0			70.0	0.8[#] & 0.7[*]	9.8[#] & 16.9[*]	Final reaction volume = 70 μL
				1.5			53.3	1.2[#] & 0.9[*]	13.6[#] & 22.1[*]	Final reaction volume = 53.3 μL
				2.0			45.0	1.4[#] & 1.1[*]	16.9[#] & 26.2[*]	Final reaction volume = 45 μL (Control*)
5	Labeling reactions	TMTsixplex	4:1	1.4	50	200	46.0	1.1	12.2	TEAB, 60 min, hydroxylamine quenching (Control*)
										30-min labeling
										HEPES reaction buffer (pH 8.5, 100 mM)
										TEAB, 60 min, hydroxylamine quenching (Control*)
6	*HeLa* cell lysate	TMTsixplex	4:1[#] & 8:1[*]	1.5	50	200[#] & 400[*]	43.3	1.2[#] & 0.9[*]	13.6[#] & 22.1[*]	N/A

2.3.5 Top-down RPLC-MS/MS analysis.

10 μg of TMT-labeled protein sample was loaded onto a trapping column (150 μm i.d., 5 cm length, Jupiter particles, 5 μm diameter, 300 Å pore size) and separated using a home-packed C5 RPLC capillary column (75 μm i.d., 70 cm length, Jupiter particles, 5 μm diameter, 300 Å pore size) on a modified Thermo Scientific (Waltham, MA, USA.) Accela LC system.[179,180] Mobile phase A was 0.01% TFA, 0.585% acetic acid, 2.5% isopropanol, and 5% acetonitrile in water. Mobile phase B was 0.01% TFA, 0.585% acetic acid, 45% isopropanol, and 45% acetonitrile in water. A 200-min gradient from 10% to 70% of MPB was applied with a flow rate of 400 nL/min.[130,179–181] A customized nano-ESI interface[14] was coupled to an LTQ Orbitrap Elite mass spectrometer using positive ion mode (Thermo Fisher Scientific, Hanover Park, IL, Bremen, Germany, USA). The temperature of the inlet capillary was set to 275 °C and the spray voltage was 2.6 kV. Full MS scans were collected with a resolving power setting of 240,000 (at m/z 400) with two micro scans. The top 6 most abundant precursor ions in the full MS scans were selected for MS/MS fragmentation. The MS/MS data were obtained using a resolving power setting of 60,000 (at m/z 400) with two microscans, an isolation window of 5.0 m/z and 60 s dynamic exclusion. Ions with charge state <4 were rejected for MS/MS fragmentation. Collision-induced dissociation (CID) with a normalized collision energy of 35 eV was applied. The maximum injection time was 1000 ms for full mass scans and 800 ms for MS/MS scans. The AGC target was 1×10^6 for full mass scans and 6×10^5 for MS/MS scans.

For TMT-labeled *HeLa* samples, a 60-min gradient from 10% to 70% MPB was applied, and the other LC parameters were consistent with the previously stated parameters. MS detection was performed using and Orbitrap Exploris 240 mass spectrometer with positive mode (Thermo Fisher Scientific, Hanover Park, IL, Bremen, Germany, USA). Full MS resolution was set at 120,000 with three microscans. The resolution of MS/MS was 120,000 with 2 microscans. The AGC target was set to 3×10^6 for MS and 1×10^6 for MS/MS. Maximum injection time was 1000 ms for MS and 500 ms for MS/MS. Top 6 most abundant precursor ion peaks were selected for MS/MS fragmentation, and 35% of normalized HCD collision energy was used for MS/MS. The isolation window was set to 2 m/z. Dynamic exclusion was 20 s. Charge 4-50 ions will be included for MS/MS fragmentation.

2.3.6 Data analysis.

All top-down proteomics datasets were deconvoluted using Biopharma Finder (Thermo Fisher Scientific) and were searched against the annotated *E. coli* protein database (UniProt 2019-10-13, 4519 species) or Human protein database (UniProt 2020-07-11, 20368 species) using TopPIC Suite.[182,183] The TopPIC parameters were MS1 signal-to-noise ratio was 3, the precursor window size was 5.0 *m/z* and the maximum mass was 50,000 Daltons. Decoy database searching was used, and the FDR cutoff was set to 0.01 for the spectrum and proteoform levels. The maximum number of mass shifts was 2 and the mass shift range was ± 500 Daltons. TMT labeling at the N-terminal and lysine residues was set as a fixed PTM and the variable PTM list is given in ***Table 2-2***. All other parameters were set as

default values. MASH Suite[184] and ProSight Lite[185] were used for manual interpretation and spectrum presentation.

Table 2-2. Post-translational modifications (PTMs) in TopPIC Suite searching.

Description	Monoisotopic mass	Residues	UnimodID	Status
Acetylation	42.0106	K	1	Variable
Carboxymethyl	58.0055	C	6	Variable
Phosphorylation	79.9663	STY	21	Variable
Oxidation	15.9949	MCPKDNRY	35	Variable
Methylation	14.0157	CKRHDENQ, N-termini	34	Variable
TMTsixplex	229.1629	K, N-termini	737	Fixed
TMTzero	224.2994	K, N-termini	739	Fixed

2.3.7 Labeling efficiency evaluation.

The labeling efficiency was evaluated and calculated by the relative intensity of each labeled product (*e.g.*, completely labeled, under-labeled, or over-labeled) under individual labeling conditions using an in-house python package. First, proteoforms with mass shifts of ± 229.163 Da for TMTsixplex or ± 224.152 Da for TMTzero were grouped together in the same "protein group" with the cutoff of mass shift less than 200 ppm and retention time shift less than 6 min. Here, "protein group" means the same proteoform with different TMT labeling status. The proteoforms with masses that are decreased by the mass of one or more TMT labels compared to reference proteoforms were considered to be under labeled and proteoforms with masses increased by the mass of one or more TMT labels compared to the reference proteoforms were considered to be over labeled. The most abundant species in the protein group was used as reference to calculate the relative intensities for each labeled product.

Second, the identification results were matched with protein groups in deconvolution results based on monoisotopic mass and retention time to generate the identification heatmap (Example in ***Figure 2-2C*****).** In the identification heatmap, 0 was used to represent the completely labeled proteoforms, negative integers were used to represent under labeling (*e.g.*, -1, -2), and positive integers were used to represent over labeling (*e.g.*, 1, 2). Protein group number is used as the y-axis ranking from high molecular weight to low molecular weight of the completely labeled proteoforms and the rows in the heatmap represent the normalized intensity of each proteoform.

We also evaluated the average TMT labeling status (weighted average) under each condition using ***Equation 1*** shown below (Example in **Figure *2-2D***). Labeling status is the integer we assigned for each labeled product in the same "protein group", such as -1, -2, -3 for under labeling, 0 for complete labeling, 1, 2, 3 for over labeling. The intensity for each labeled product was also extracted for calculation. The mean of weighted average values for all protein groups of each sample was calculated demonstrating the average labeling status. Theoretically, the closer the weighted average is to 0, the more completely labeled the species; the more negative the value, the more under labeling is present; the more positive the value, the more over labeling is present.

$$Weighted\ average = \frac{\sum labeling\ status * intensity}{\sum intensity} \qquad \textbf{Equation 1}$$

2.4 RESULTS AND DISCUSSION

2.4.1 Removal of high molecular weight proteoforms

We previously found that removal of large molecular weight proteins (>100 kDa) decreased sample precipitation and increased labeling efficiency for protein-level TMT labeling,[130] so we evaluated and adapted this sample preparation process for the removal of large proteins prior to TMT labeling (*e.g.*, molecular weight cutoff (MWCO) filtration step to enrich proteins <100 kDa). We examined the reproducibility of protein-level TMT labeling of complex samples using *E. coli* cell lysate. We labeled these samples using TMTzero reagent as described before with minor modifications.[130] In short, we reduced, alkylated, and buffer-exchanged the lysate and labeled with TMTzero reagent with a TMT-to-protein mass ratio of 1:4 at room temperature for 1 hour. The labeling reaction was quenched with 5% hydroxylamine to a final concentration of 0.6% and incubated for 15 minutes. Lysate was labeled in duplicate to evaluate the labeling reproducibility. No significant precipitation was observed during the labeling process, indicating that the 100 kDa molecular weight cutoff approach can efficiently remove large proteins for protein-level TMT labeling. In total, 352 unique proteoforms were identified from these two samples. Base peak chromatogram (BPC) shown in ***Figure 2-1A*** demonstrated high reproducibility between the two replicate samples. An identification heatmap (***Figure 2-1B***) was utilized to compare the overall labeling efficiency between two replicates. Each row in heatmap represents a "protein group" which contains all TMT labeled products from one proteoform, including completed labeled, underlabeled, and overlabeled products. The greyscale indicates the relative intensity of each labeled product. The heatmap demonstrated high similarity of labeling efficiency between two replicates.

Overall, we found that the protein-level TMT labeling is reproducible and 100 kDa MWCO filtration removes large proteins to avoid protein precipitation during the TMT labeling reaction. Using this MWCO filtering method to prepare intact protein samples for TMT labeling, we systematically evaluated protein-level TMT labeling reaction parameters, including TMT-to-protein ratios, pH/concentration of the quenching buffer, initial protein concentration, reaction time, and reaction buffer to minimize the under-/over- labeling commonly observed in protein-level TMT labeling.[167,171]

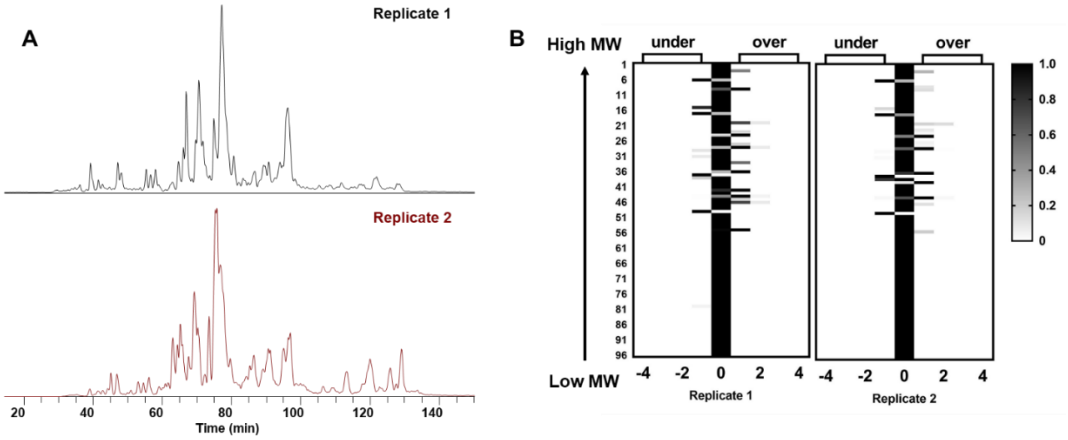


Figure 2-1 Simple sample preparation and reproducibility evaluation. (A) The base peak chromatogram (BPC) for two replicates. **(B)** Identification heatmap of group 0 for protein-level TMT labeling reproducibility test. The protein groups were found in both replicates.

2.4.2 Optimization of TMT-to-protein mass ratios in the labeling reaction

The TMT-to-protein mass ratio recommended by the manufacturer is 8:1, which is an approximately 20-fold molar excess of the labeling reagent.[175] However, a recently published bottom-up study determined that a TMT-to-peptide mass ratio of 1:1 demonstrated a high labeling efficiency with high reproducibility. Here, to minimize under-labeling and over-labeling products in the protein-level TMT

labeling process, TMT-to-protein mass ratio (1:1, 2:1, 4:1 to 8:1) was evaluated (***Figure 2-2A***). Briefly, 45 µg of *E. coli* proteins were labeled with TMTsixplex reagent at different ratios. The protein reaction concentration after mixing was 0.8 µg/µL (*e.g.*, the final TMT concentration varied from 2.3 mM to 18.2 mM).

In total, 747 unique proteoforms were identified from these 4 samples using top-down LC-MS/MS analysis. The labeling efficiency for identified proteoforms in each sample is demonstrated in an identification heatmap (***Figure 2-2C***). We found that 31.00% (31 out of 100), 34.76% (65 out of 187), 52.69% (98 out of 186), and 67.39% (124 out of 184) of the protein groups from the samples labeled using 1:1, 2:1, 4:1 and 8:1 TMT-to-protein ratios, respectively, were found to only include the completely labeled products; improperly labeled side products were not detected. 67.00%, 64.71%, 36.02% and 12.50% of protein groups, respectively, were found to include underlabeled products; 4.00%, 6.42%, 23.66%, and 25.54% of protein groups were found to include overlabeled products. The percentage of protein groups that contained a completely labeled product with >80% intensity (*e.g.*, completely labeled product was the dominant protein peak) was found to be 35.00%, 36.36%, 59.68%, and 80.98%, respectively.

We also evaluated the average labeling status using a violin plot (***Figure 2-2D***). The mean of the weighted averages for the 1:1, 2:1, 4:1 and 8:1 TMT-to-protein labeled samples were -1.19 ± 1.22, -0.56 ± 0.53, -0.09 ± 0.36 and 0.07 ± 0.27. This suggests that as TMT-to-protein mass ratio increases, the amount of completely labeled protein groups increases as demonstrated by the mean value approaching 0. Additionally, at labeling ratios of 4:1 and 8:1, the dominant

distributions are localized around 0, indicating a majority of protein groups primarily contain completely labeled proteins. However, in the sample labeled at 1:1 and 2:1 ratio, a large portion of protein groups were underlabeled. Also, in the samples labeled with a 1:1 TMT-to-protein ratio, multiple underlabeled products were observed. In the sample labeled with a 2:1 ratio, the dominant product is underlabeled by 1 TMT tag (*e.g.*, a dominant distribution at -1 on the violin plot).

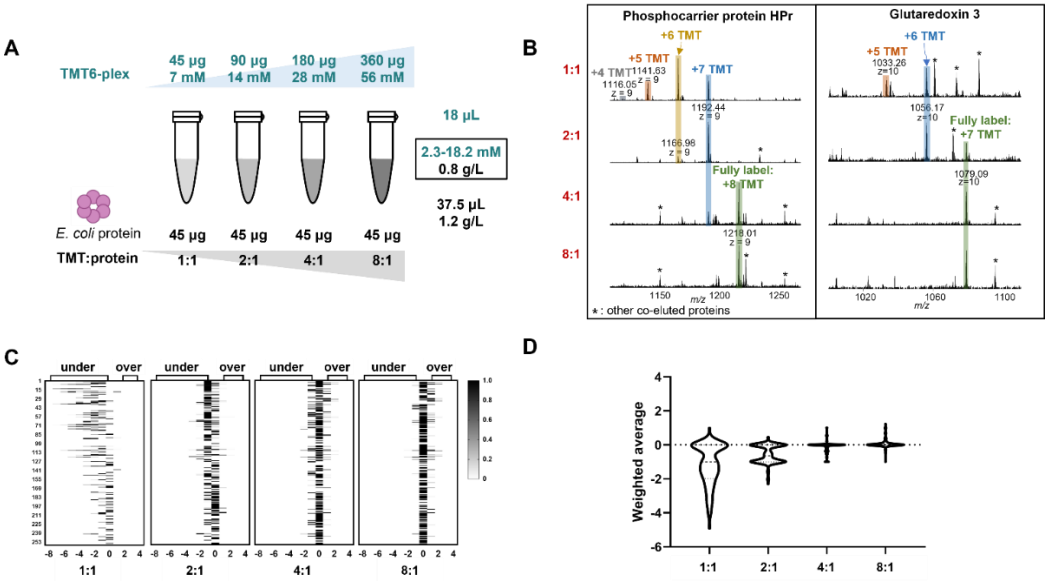


Figure 2-2 Optimization of TMT-to-protein mass ratio. (A) Quantities and concentrations of TMT reagent (blue) and intact E. coli protein lysate (grey). The reaction volume and protein quantity were kept constant at 55.5 µL and 45 µg for all 4 conditions. The mass of the TMT reagent was increased from 45 to 360 µg in a constant volume of ACN yielding an increasing TMT-to-protein mass ratio from 1:1 to 8:1. **(B)** MS of 2 identified proteins labeled using the different TMT-to-protein ratios: Phosphocarrier protein HPr, P0AA04 and Glutaredoxin 3, P0AC63 (asterisk (*) represents coeluted species). **(C)** Identification heatmap demonstrating all labeled species observed under each TMT-to-protein ratio condition. Y-axis shows the number of "protein groups" ranked from high to low MW of the completely labeled proteoform. The labeling status is shown for each intact proteoform as x-axis with under-/over-labeled status indicated by the brackets. Grey scale (0-1) represents the relative intensity of each proteoform. **(D)** Violin plots of average labeling status to show overall TMT labeling efficiency.

Two examples of identified proteoforms are shown in ***Figure 2-2B***, Phosphocarrier protein HPr (Uniprot: P0AA04) and Glutaredoxin 3 (Uniprot: P0AC62), which showed multiple labeled products in the MS spectra. For Phosphocarrier protein HPr protein, the dominant peak that represented the completely labeled proteoform was observed at 8:1 and 4:1 TMT-to-protein labeling ratios. However, in the sample with a 4:1 ratio, a small peak that represented the underlabeled product (underlabeled by 1 tag) was also observed. In the samples labeled with 1:1 and 2:1 labeling ratios, there were significant underlabeled products detected. Interestingly, the sample with a 1:1 ratio showed a normal distribution of underlabeled peaks with multiple peaks representing underlabeled products. However, the sample with a 2:1 ratio only contains one major underlabeled product (underlabeled by 1 tag) which is consistent with the violin plot (***Figure 2-2D***). The MS/MS fragmentation confirmed that the missing label was located on the *N*-terminus (***Figure 2-3A-C***), which is consistent with our previously reported results that the decreased reactivity of the *N*-terminus compared with lysine (pKa of N-terminal is 7.7 ± 0.5 and lysine is 10.5 ± 0.1) causes underlabeling to occur more frequently on the N-terminus.[130,186] Similar results were observed for P0AC62 with the exception that the sample labeling with a 2:1 TMT-to-protein ratio had no observable underlabeling. Overall, analysis of MS/MS spectra for identified proteoforms confirmed that the missing labels in the 2:1 TMT-to-protein ratio sample was primarily localized to the N-terminus (***Figure 2-3D-E***).[130]

Overall, higher TMT-to-protein ratio (w/w: 8:1, 4:1) reaction conditions resulted in higher ratios of completely labeled products compared with underlabeled products. However, some protein groups labeled using 8:1 and 4:1 TMT-to-protein ratios showed overlabeled products, which suggested that overlabeling also needs to be minimized by optimization.

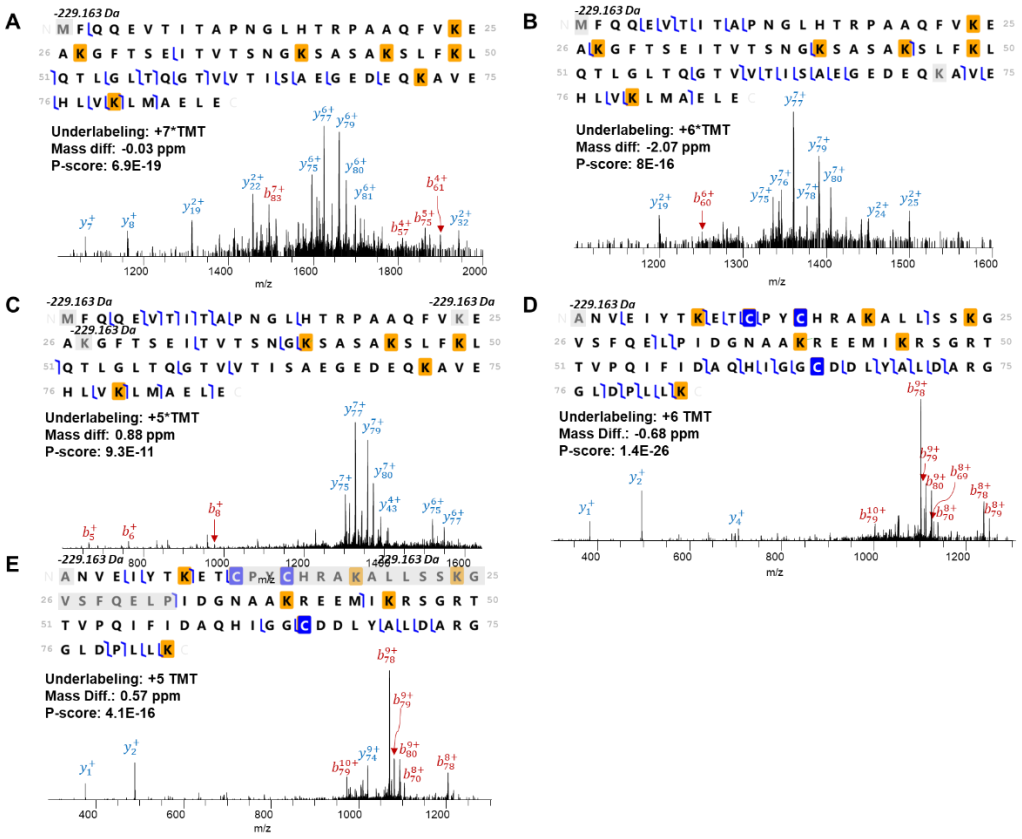


Figure 2-3 Fragmentation examples from optimization of TMT-to-protein mass ratio group. Phosphocarrier protein HPr (P0AA04) underlabeled species with **(A)** 7 TMT; **(B)** 6 TMT; **(C)** 5 TMT labels. Glutaredoxin 3 (P0AC63) underlabeled species with **(D)** 6 TMT; **(E)** 5 TMT labels. Amino acids highlighted in yellow indicate TMT labeling on N-terminus or K; blue highlight represents the carboxymethyl modification on C. Missing labels were indicated by grey highlight. All fragmentations were collected from 1:1 TMT-to-protein ratio sample.

2.4.3 Optimization of quenching buffers in the labeling reaction

For bottom-up TMT labeling, 5% hydroxylamine (0.3% final concentration) was recommended as the quenching buffer by the manufacturer. However, Tris buffer (50 mM at pH 8.0) has also been utilized as the quenching buffer in previous literature during peptide-level TMT labeling.[175] In this study, we evaluated different quenching buffers for intact protein-level TMT labeling (*e.g.,* hydroxylamine buffer with a final concentration of 1.2%, 0.6%, or 0.3%, and Tris buffer with a final concentration of 50 mM ***Figure 2-4A***). We also measured final pH values of the reaction solution after quenching: 9.1 (0.3% hydroxylamine), 9.3 (0.6% hydroxylamine), 9.5 (1.2% hydroxylamine), and 8.0 (50 mM Tris).

A total of 610 unique proteoforms were identified from 4 samples with different quenching buffers using LC-MS/MS analysis. An identification heatmap was generated to evaluate the labeling efficiency for each identified protein (***Figure 2-4C***). In the solution quenched by Tris buffer (pH 8.0), many overlabeled products were observed with relatively high intensity. With the increased pH/final concentration of hydroxylamine, the relative intensity of overlabeled products decreased and more completely labeled product were detected. The percentage of protein groups that only included the completely labeled product (no improperly labeled products) were 89.24% (141 out of 158), 77.86% (109 out of 140), 80.16% (101 out of 126), and 60.13% (92 out of 153) for the samples quenched at pH 9.5, pH 9.3, pH 9.1 and pH 8.0, respectively. 3.16%, 15.00%, 15.08%, and 35.29% of protein groups, respectively, included overlabeled products, which suggested that higher pH after quenching led to a decrease in overlabeled products. Additionally,

7.59%, 7.86%, 4.76%, and 8.50% of the protein groups were found to include underlabeled products. The percentage of protein groups that contained a completely labeled product with >80% intensity (*e.g.*, completely labeled product was the dominant peak) was found to be 92.41%, 85.00%, 84.92%, and 67.97%, respectively. Mean values of the weighted averages for the four samples were calculated to be -0.03 ± 0.20, 0.02 ± 0.30, 0.06 ± 0.33 and 0.26 ± 0.54 for pH 9.5, pH 9.3, pH 9.1, and pH 8.0, respectively (***Figure 2-4D***), which indicated that quenching at pH > 9.1 decreases overlabeling.

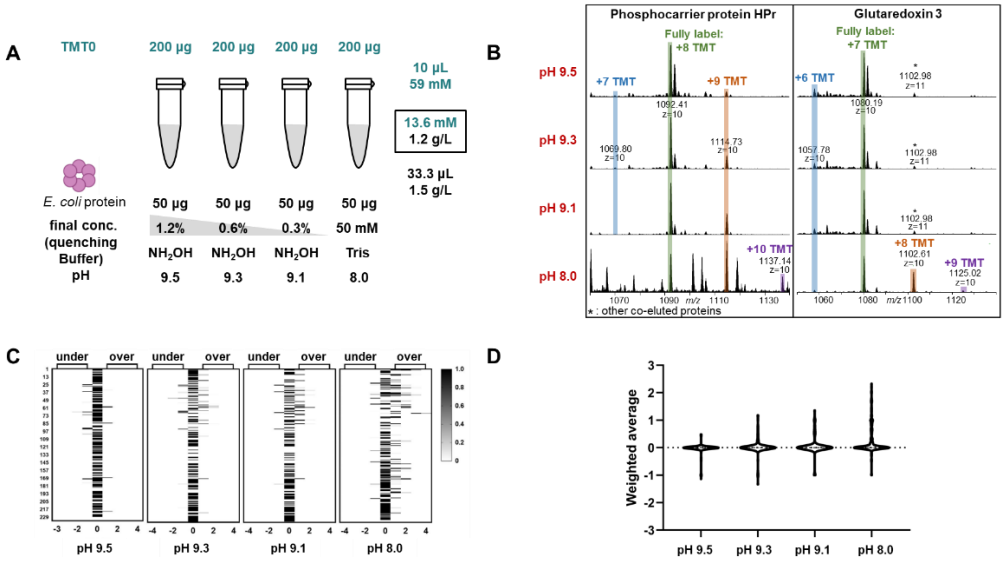


Figure 2-4 Optimization of quenching buffer in protein-level TMT labeling. **(A)** Quantities and concentrations of TMT reagent (blue) and intact E. coli protein (grey). The reaction parameters were kept the same for all four experiments, but the final concentration of the quenching buffer was varied: 1.2%, 0.6%, and 0.3% hydroxylamine or 50 mM Tris which resulted in final pH of 9.5, 9.3, 9.1, and 8.0, respectively. **(B)** MS of 2 identified proteins under different quenching conditions: Phosphocarrier protein HPr, P0AA04 and Glutaredoxin 3, P0AC63 (asterisk (*) represents coeluted species). **(C)** Identification heatmap of quenching buffer optimization group. Y-axis shows the number of "protein groups" ranked from high to low MW of the completely labeled proteoform. The labeling status is shown for each intact proteoform as x-axis with under-/over-labeled status indicated by the brackets. Grey scale (0-1) represents the relative intensity of each proteoform. **(D)** Violin plots demonstrating the average labeling status to show overall TMT labeling efficiency.

As shown in ***Figure 2-4B,*** for Phosphocarrier protein HPr (P0AA04), only a very small portion of the overlabeled product was observed in the reaction quenched at high pH (hydroxylamine 1.2%, pH 9.5). Generally, as the quenching pH increased, overlabeling decreased. In the solution quenched with Tris buffer (pH=8.0), the dominant peak represented the product overlabeled by one TMT label. Similar results were observed for Glutaredoxin 3 (P0AC62). MS/MS fragmentation (***Figure 2-5A*****)** for Phosphocarrier protein HPr protein localized the overlabeling site at T_{30}; ***Figure 2-5B*** for Glutaredoxin 3 indicated the overlabeling site was within a range with several STY residues, which is consistent with previous reported results that TMT overlabeling tends to occur on STY amino acids.[167,175]

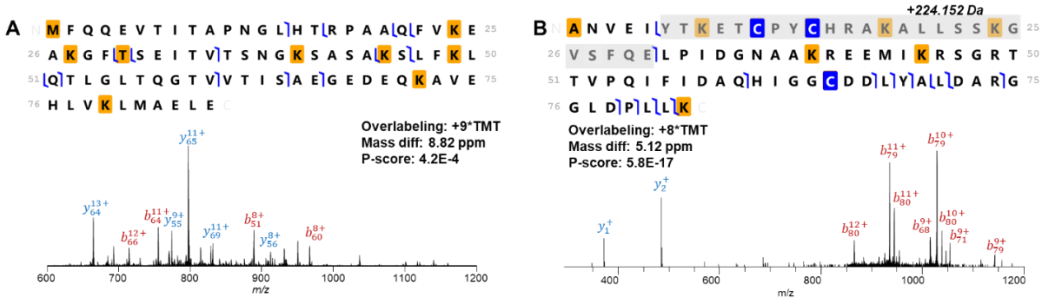


Figure 2-5 Fragmentation examples from optimization of quenching buffer group. (A) Phosphocarrier protein HPr (P0AA04) overlabeled species with 9 TMT labels. **(B)** Glutaredoxin 3 (P0AC63) overlabeled species with 8 TMT labels. Amino acids highlighted in yellow indicate TMT labeling on N-terminus, K, and T; blue highlight represents the carboxymethyl modification on C. All fragmentation examples were taken from Tris sample.

The side chains of lysine and tyrosine have similar pKa values (Lys: 10.5 ± 1.1; Tyr: 10.3 ± 1.2);[186] therefore, it is expected that tyrosine would also be labeled by TMT at basic pH. It has also been reported that O-acylation can occur on serine and threonine because of the nucleophilicity of the hydroxyl oxygen due to the hydrogen bonding of hydroxyl hydrogen to proton acceptors such as histidyl

imidazole groups.[187,188] Hydroxyl-containing amino acids also show reactivity towards NHS esters when separated by one or two residues from a histidyl residue to form an H-X-[STY] or [STY]-X-H motif.[175,188] O-acylation was found to hydrolyze within a few minutes at basic pH values,[175,189,190] which explains why hydroxylamine quenching buffer can be used to reverse the overlabeling by increasing the pH to ≥ 9.0. [187,189,189,191]

Overall, our results demonstrate that a hydroxylamine quenching buffer (final concentration of 0.3% or higher) resulting in >pH 9.1 after quenching can reverse unwanted O-acylation reactions and reduce the production of overlabeled products.

2.4.4 Optimization of the initial protein sample concentration in the labeling reaction

In the TMT and protein labeling reaction, higher molarity of reactants (TMT and protein) increases the rate of labeling and reduces the hydrolysis of the TMT reagent which may reduce the production of underlabeled species.[175] Hence, to fully understand the effect of reactant concentration on intact protein-level TMT labeling, the initial concentration of protein was varied (2 μg/μL, 1.5 μg/μL, 1.0 μg/μL, and 0.5 μg/μL) while maintaining the same TMT-to-protein mass ratio of 4:1, ***Figure 2-6A***. Specifically, 200 μg TMT in 10 μL of ACN solvent was reacted with 50 μg protein in different buffer volumes from 25 μL to 100 μL, which varied the final reactant concentrations of TMT and proteins from 5.4 to 16.9 mM and 0.5 to 1.4 μg/μL, respectively.

LC-MS/MS analysis identified 559 unique proteoforms among 4 samples. In the identification heatmap (***Figure 2-6C*****)**, we found that 61.22% (90 out of 147), 75.76% (125 out of 165), 79.61% (121 out of 152), and 75.89% (107 out of 141) of the protein groups for samples labeled with protein concentration from 0.5 μg/μL to 2 μg/μL, respectively, which suggested underlabeled products could be reduced by increasing protein initial concentration. The percentage of protein groups that contained underlabeled products were 37.41%, 20.00%, 8.55% and 10.64% as protein concentration increased; the percentage of protein groups that include overlabeled products decreased slightly as concentration increased (3.40%, 5.45%, 13.82% and 15.60%). In addition, the means of weighted averages (***Figure 2-6D*****)** from were calculated to be -0.16 ± 0.38, -0.08 ± 0.27, 0.02 ± 0.25 and 0.03 ± 0.28 from 0.5 μg/μL to 2 μg/μL. This trend further demonstrates that under labeling was reduced as reactant concentration increased. The percentage of protein groups that contained a completely labeled product peak with >80% intensity (*e.g.*, completely labeled product was dominant) was found to be 69.39%, 86.06%, 88.16%, and 83.69% for samples with protein concentrations from 0.5 μg/μL to 2 μg/μL, respectively.

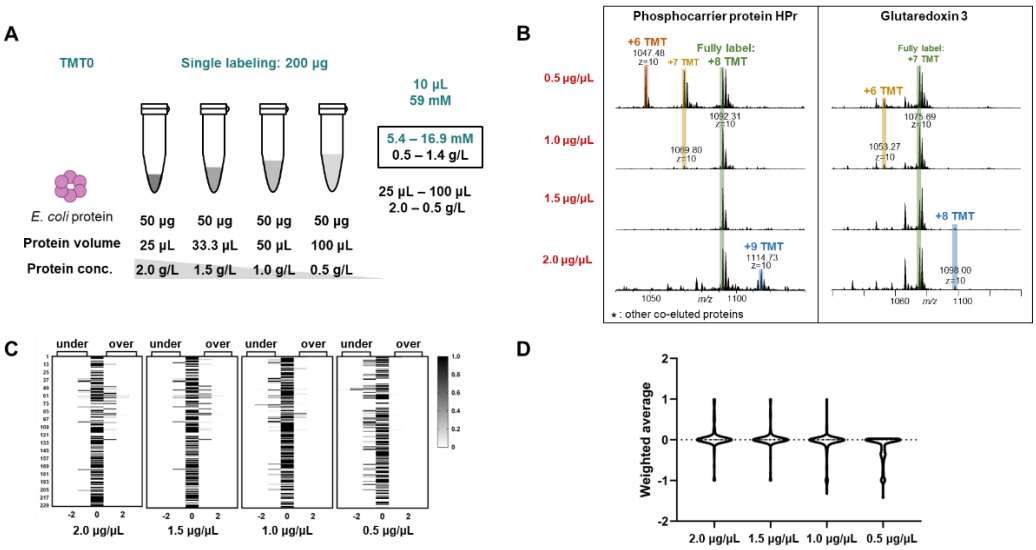


Figure 2-6 Optimization of initial protein concentration in the labeling reaction. (A) Quantities and concentrations of TMT reagent (blue) and intact E. coli protein (grey). The reaction TMT and protein quantity were kept constant at 200 µg and 50 µg for all 4 conditions and the total reaction volume was varied from 35 µL to 110 µL by changing initial protein concentration from 2.0 µg/µL to 0.5 µg/µL. **(B)** MS of 2 identified protein examples under different reaction conditions: Phosphocarrier protein HPr, P0AA04 and Glutaredoxin 3, P0AC63 (asterisk (*) represents coeluted species). **(C)** Identification heatmap of all four samples under initial protein concentration optimization group. Y-axis shows the number of "protein groups" ranked from high to low MW of the completely labeled proteoform. The labeling status is shown for each intact proteoform as x-axis with under-/over-labeled status indicated by the brackets. Grey scale (0-1) represents the relative intensity of each proteoform. **(D)** Violin plots demonstrating the average labeling status to show overall TMT labeling efficiency.

To further demonstrate the influence of reactant concentration on labeling efficiency, MS data for two identified proteins are shown in ***Figure 2-6B***. The MS spectra of Phosphocarrier protein HPr (P0AA04) showed that the dominant peak that represented the completely labeled product was observed in all samples. However, the sample with an initial protein concentration of 0.5 µg/µL showed 2 underlabeled peaks with similar intensities as the completely labeled peak (***Figure 2-6B***). Another example, Glutaredoxin 3 (P0AC62), also showed the completely labeled product peak as the highest peak in all samples; however, a small peak that

represents the underlabeled product was detected in the samples with initial protein concentration of 1.0 μg/μL and 0.5μg/μL (***Figure 2-6B***). Fragmentation and MS/MS spectra for the two examples in **Figure *2-7*** suggested that proteoforms that were under labeled by one TMT tag were missing labels on the N-terminal which is consistent with previous observations.[14]

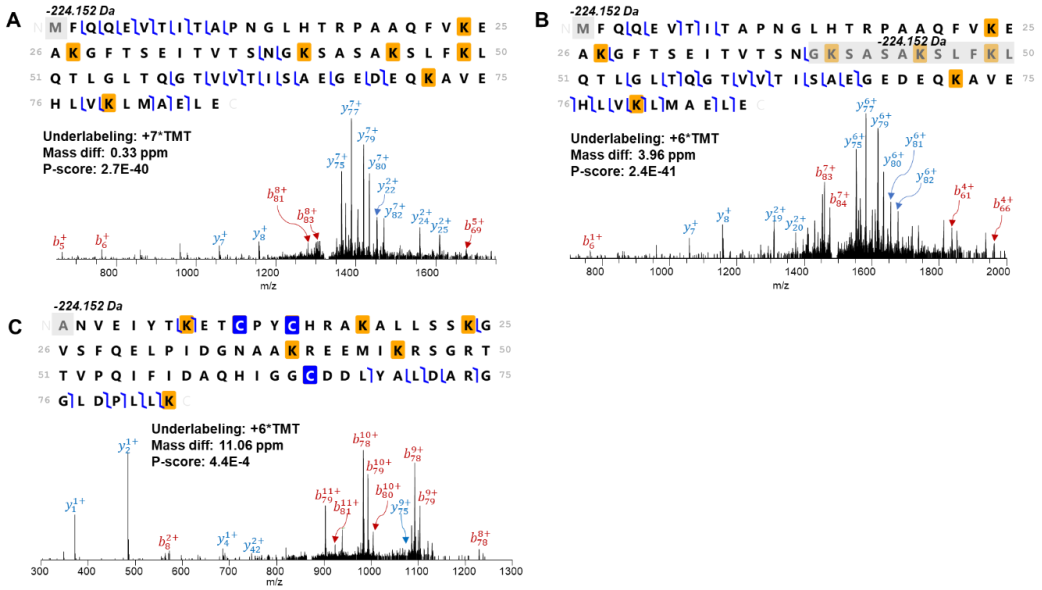


Figure 2-7 Fragmentation examples of initial protein concentration with single labeling optimization group. Phosphocarrier protein HPr (P0AA04) underlabeled species with **(A)** 7 TMT labels; **(B)** 6 TMT labels. **(C)** Glutaredoxin 3 (P0AC63) underlabeled species with 7 TMT labels. Amino acids highlighted in yellow indicated TMT labeling on N-terminus or K; blue highlight represents the carboxymethyl modification on C. The range highlighted in light grey represents the range that the mass shift lies within. All fragmentations were taken from 0.5 μg/μL sample.

Overall, our results indicated that initial protein concentration >1.0 μg/μL with a TMT-to-protein ratio of 4:1 (wt/wt) or higher was best for intact protein TMT labeling. These conditions minimized underlabeling and resulted in high labeling efficiency to produce primarily completely labeled species.

2.4.5 Investigation of TMT labeling reaction conditions influence in labeling efficiency.

Here, we evaluated other reaction parameters that might affect the efficiency of protein-level TMT labeling, including (1) reaction time (60 minutes and 30 minutes); (2) pH 8.5 reaction buffer (100 mM TEAB and 100 mM HEPES) **(*Figure 2-8A*)**. LC-MS/MS analysis identified 367 proteoforms among 3 samples.

Reaction time evaluation: The manufacturer protocol recommended a 60-min reaction time for TMT labeling. To evaluate the effect of reduced labeling time on labeling efficiency, we changed the labeling time to 30 minutes and kept the other parameters the same (***Table 2-1***). Two protein examples: Phosphocarrier protein HPr (P0AA04) and Glutaredoxin 3 (P0AC63), ***Figure 2-8B,*** demonstrated similar labeling for the samples incubated for 30 and 60 minutes with the exception that the sample incubated for 60 minutes generated a higher intensity of over labeled peak in P0AA04. In the identification heatmaps **(*Figure 2-8C*)**, 77.08% (148 out of 192) of protein groups from the 60-min incubation sample and 75.47% (160 out of 212) of protein groups from the 30-min incubation sample demonstrated only the completely labeled products. 11.46% and 12.26% of protein groups, respectively, were found to include under labeled products and 12.50% and 15.09% of protein groups were found to include over labeled products. The percentage of protein groups that contained a completely labeled product with >80% intensity (*e.g.*, completely labeled product was dominant) was 80.21% and 80.19% for 60-min and 30-min incubations, respectively. There is no significant difference between these two samples with respect to labeling efficiency. Additionally, the

weighted averages shown in the violin plot **(*Figure 2-8D*)** indicate similar overall TMT labeling with mean values of 0.02 ± 0.36 for 60-min incubation and 0.05 ± 0.39 for 30-min incubation. These results indicate that a 30-minute incubation time is sufficient for high protein-level TMT labeling efficiency.

Reaction buffer evaluation: To determine if different reaction buffers will affect the labeling efficiency, we tested the protein-level TMT labeling in 100 mM TEAB (pH=8.5) and 100 mM HEPES (pH=8.5), the other parameters were shown in ***Table 2-1***. The examples (Phosphocarrier protein HPr (P0AA04) and Glutaredoxin 3 (P0AC63)) shown in ***Figure 2-8B*** suggest a similar labeling efficiency between two samples except that the HEPES sample had a higher ratio of over labeled product peaks compared to TEAB buffer for P0AA04. According to the identification heatmap in ***Figure 2-8C***, the labeling efficiency was similar between the HEPES and TEAB buffers. 77.08% (148 out of 192) and 78.95% (165 out of 209) of the protein groups in TEAB and HEPES, respectively, were found to only produce the completely labeled products. 11.46% and 9.09% of protein groups were found to contain under labeled products; 12.50% and 12.44% of protein groups included over labeled products. The percentage of protein groups that contained a completely labeled product with >80% intensity (*e.g.*, completely labeled product was dominant in majority of protein groups) was 80.21% and 82.78% for TEAB and HEPES, respectively. The weighted average distribution (***Figure 2-8D***) for the two samples was also quite similar with mean values of 0.02 ± 0.36 and 0.03 ± 0.33, respectively, which was consistent with the heatmap, suggesting similar labeling efficiency between labeling buffers.

In conclusion, we evaluated the reaction conditions for protein-level TMT labeling and concluded that the reaction buffer and time did not significantly affect labeling efficiency. The result indicated that the labeling reaction time can be reduced and HEPES may be used as alternative reaction buffer.

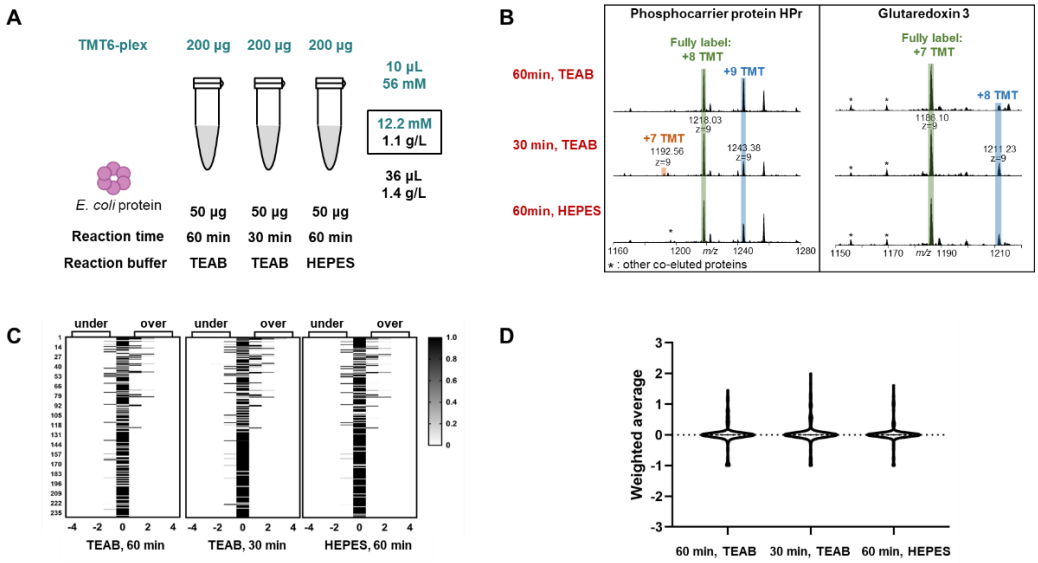


Figure 2-8 Investigation of TMT labeling reaction conditions influence in labeling efficiency. (A) Quantities and concentrations of TMT reagent (blue) and intact E. coli protein (grey). The reaction TMT and protein quantity were kept constant at 200 µg and 50 µg for all 4 conditions. Reaction buffer and time are indicated in the figure. **(B)** MS of 2 identified protein examples with labeling status at different reaction conditions: Phosphocarrier protein HPr, P0AA04 and Glutaredoxin 3, P0AC63 (asterisk (*) represents coeluted species). **(C)** Identification heatmap for reaction optimization group. Y-axis shows the number of "protein groups" ranked from high to low MW of the completely labeled proteoform. The labeling status is shown for each intact proteoform as x-axis with under-/over-labeled status indicated by the brackets. Grey scale (0-1) represents the relative intensity of each proteoform. **(D)** Violin plots of average labeling status to show overall TMT labeling efficiency.

2.4.6 Double labeling: to achieve complete labeling on low concentration samples.

Our results indicate that high protein and TMT reagent concentrations are optimum for intact protein labeling. However, some samples (*e.g.*, clinical samples)

may be limited and may not be sufficient to produce high concentration protein solutions (*e.g.*, >1.0 μg/μL). For efficient labeling of low concentration protein samples, we designed and optimized a double labeling technique using various initial protein concentrations (2 μg/μL, 1.5 μg/μL, 1.0 μg/μL, and 0.5 μg/μL), ***Figure 2-9A*.** Briefly, TMT labeling was conducted twice before quenching; 200 ug of TMT in 10 μL ACN was incubated with the protein solutions in 100 mM TEAB buffer (pH 8.5) for 1 hour followed by a second addition of TMT reagent and incubation for an additional hour. TMT-to-protein ratio remained constant and the final reactant concentrations were varied from 10.8 to 33.8 mM (TMT) and 0.5 to 1.4 μg/μL (protein).

The LC-MS/MS analysis identified 619 unique proteoforms among the 4 samples. The heatmap in ***Figure 2-9C*** suggested a similar labeling efficiency for all 4 conditions. The percentage of protein groups that only included completely labeled products was found to be 87.73% (193 out of 220), 88.54% (170 out of 192), 90.78% (187 out of 206), and 89.15% (189 out of 212) from 0.5 μg/μL to 2 μg/μL, respectively. 10.45%, 7.81%, 5.83% and 6.60% of protein groups, respectively, were found to include under labeled products and 2.27%, 4.17%, 3.88% and 4.72% of protein groups were found to include over labeled products. The percentage of protein groups that contained a completely labeled product with >80% intensity (*e.g.*, majority of protein product was completely labeled) was found to be 93.64%, 92.71%, 93.69%, and 93.87% for initial protein concentration samples from 0.5 μg/μL to 2 μg/μL, respectively. As shown in the violin plot (***Figure 2-9D***), there is no significant difference among the mean values of the weighted averages (-0.04 ±

0.20, -0.03 ± 0.22, -0.03 ± 0.20 and -0.02 ± 0.18). Overall, we found that there was no significant difference in TMT labeling efficiency among 4 samples with different initial protein concentrations when the double labeling technique was used. Furthermore, the MS spectra of both Phosphocarrier protein HPr (P0AA04) and Glutaredoxin 3 (P0AC62) proteins showed that the completely labeled product was the only observable proteoform peak in all samples after double labeling (***Figure 2-9B***).

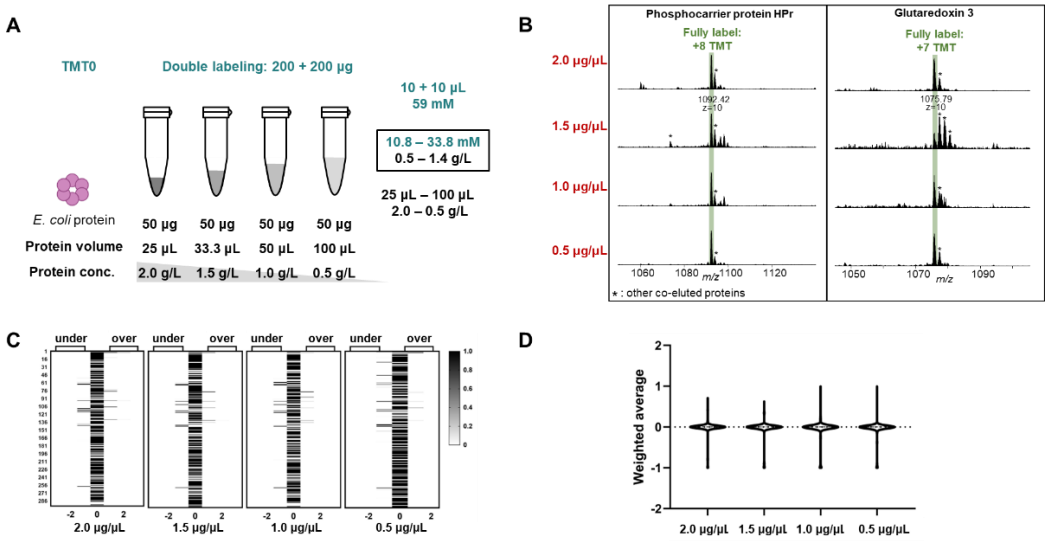


Figure 2-9 Double labeling: to achieve complete labeling on low-concentration samples. (A) Quantities and concentrations of TMT reagent (blue) and intact E. coli protein (grey). The reaction TMT and protein quantity were kept constant at 200 µg and 50 µg for all 4 experiments but the final reaction volume after double labeling was varied from 45 µL to 120 µL by changing initial protein concentration from 2.0 µg/µL to 0.5 µg/µL. **(B)** MS of 2 identified protein examples under different reaction conditions: Phosphocarrier protein HPr, P0AA04 and Glutaredoxin 3, P0AC63 (asterisk (*) represents coeluted species). **(C)** Identification heatmap of all four samples under initial protein concentration optimization group. Y-axis shows the number of "protein groups" ranked from high to low MW of the completely labeled proteoform. The labeling status is shown for each intact proteoform as x-axis with under-/over-labeled status indicated by the brackets. Grey scale (0-1) represents the relative intensity of each proteoform. **(D)** Violin plots demonstrating the average labeling status to show overall TMT labeling efficiency.

Our results indicated that the double labeling can be an alternative approach to achieve high labeling efficiency for limited samples with low protein concentration (*i.e.*, ≥0.5 μg/μL initial protein concentration).

2.4.7 Application of optimized protein-level TMT labeling conditions on *HeLa* cell lysate

We applied the optimized conditions to the TMT labeling of *HeLa* cell lysate to evaluate the TMT labeling efficiency and quantification accuracy. Briefly, *HeLa* proteins (1.5 μg/μL) were aliquoted into 6 tubes and labeled with TMTsixplex under the optimized conditions: TMT-to-protein mass ratio of 8:1 (double labeling) and hydroxylamine quench (final concentration of 1.2%, pH 9.5). Then, the six TMT-labeled *HeLa* protein samples were mixed to a mass ratio of 5:5:2:2:1:1 and analyzed using LC-MS/MS.

179 proteoforms were identified from the LC-MS/MS analysis. MS spectra of four protein examples are given in ***Figure 2-10A***. All four proteins were found to only show the completely labeled proteoforms (indicated by *) and the dashed lines labeled with -1 and +1 represent the theoretical location of proteoforms under labeled and over labeled by one TMT tag, respectively. An identification heatmap with 133 clustered protein groups (***Figure 2-10B*)** was generated to illustrate the overall labeling efficiency. 86.47% (115 out of 133) protein groups were found to only include completely labeled products without any improperly labeled side products, 10.53% protein groups contained under labeled products and 6.77% protein groups included over labeled products. 91.73% of protein groups that contained a completely labeled product with >80% intensity (*e.g.*, completely

labeled product was dominant in majority of protein groups). The mean value of the weighted average was -0.05 ± 0.21, which confirmed that most of the protein groups showed primarily completely labeled products. In ***Figure 2-10C***, the Jupiter microtubule associated homolog 1 protein (Q9UK76) was found to include only the completely labeled proteoform. Additionally, the intensities of all reporter ions were extracted and the ratios were calculated as previously described.[130] The average reporter ion ratios were 5.20 ± 1.50, 5.00 ± 1.37, 1.97 ± 0.76, 1.98 ± 0.67, 1.05 ± 0.34 (theoretical ratios as 5:5:2:2:1), which represents a sizeable improvement compared with our previously published data before reaction optimization.[130]

Overall, the intact protein-level TMT labeling under these optimized conditions not only increased the labeling efficiency, but also improved the quantification accuracy.

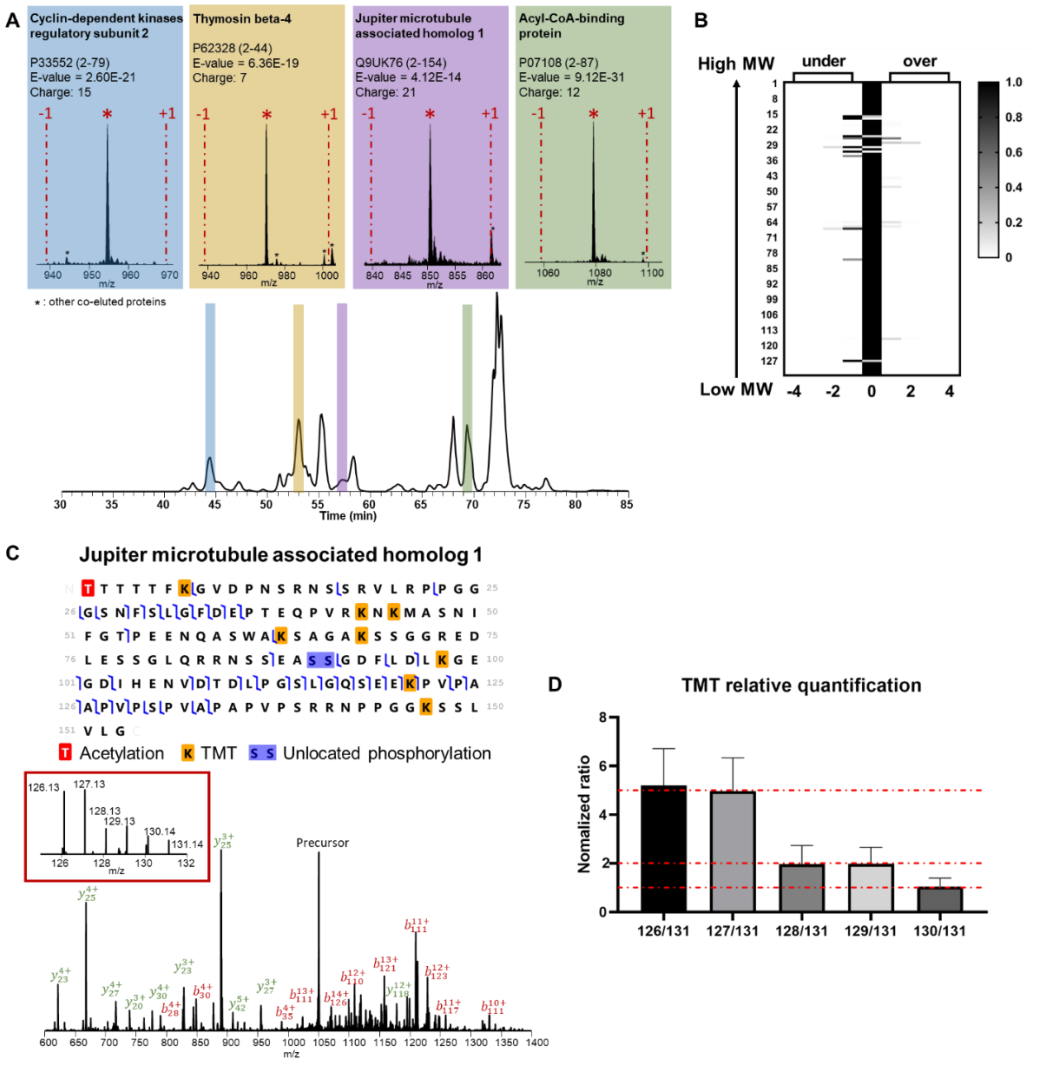


Figure 2-10 Application of optimized protein-level TMT labeling conditions on *HeLa* cell lysate. (A) MS of 4 identified proteoforms. The completely labeled proteoform is indicated by an asterisk and the -1 and +1 represent the theorical location of proteoform peaks under labeled and over labeled by one TMT tag. **(B)** Identification heatmap of TMT-labeled HeLa proteoforms. Y-axis shows the number of "protein groups" ranked from high to low MW of the completely labeled proteoform. The labeling status is shown for each intact proteoform as x-axis with under-/over-labeled status indicated by the brackets. Grey scale (0-1) represents the relative intensity of each proteoform. **(C)** MS/MS spectra of the completely labeled Jupiter microtubule associated homolog 1 protein. **(D)** Normalized intensity ratios (126/131, 127/131, 128/131, 129/131, and 130/131) of TMT-labeled HeLa proteins. The error bars represent the standard deviation, and the red dashed lines represent the theoretical ratio (5:5:2:2:1).

2.5 CONCLUSION

We have determined that the optimized conditions for protein-level TMT labeling in complex samples, such as *E. coli* cell lysate and *HeLa* cell lysate. A protein solution with initial concentration > 1.0 μg/μL labeled with TMT as 4:1 or higher TMT-to-protein mass ratio, then quenched by hydroxylamine to a final concentration as > 0.3% (final pH > 9.1) resulted in optimal labeling efficiency and minimized production of over/underlabeled side products). Double labeling can also be applied for low concentration samples for better labeling efficiency. With optimal protein-level TMT labeling conditions, >90% labeling efficiency (percentage of protein groups that only contained completely labeled product) was achieved. When the optimal conditions were applied to *HeLa* cell lysate, 86.47% labeling efficiency was achieved and 91.73% of protein groups contained a completely labeled product with >80% intensity (*e.g.*, completely labeled product was dominant). These optimized conditions can potentially be adapted to the other isobaric chemical tag labeling approaches (*e.g.*, iTRAQ, DiLeu).

In previous work, optimization of peptide-level TMT labeling suggested that a 1:1 TMT-to-peptide mass ratio was sufficient for labeling;[175] however, our results indicated a higher TMT-to-protein mass ratio (4:1 or 8:1) is required for efficient top-down TMT labeling. The higher TMT-to-protein mass ratio required for sufficient labeling may be due to structural hindrance caused by the secondary structure that may be present in intact proteins even after urea denaturation, reduction, and alkylation.[192] Higher organic concentration and/or higher labeling temperature may improve intact protein denaturation by reducing the remaining

secondary structure, which can potentially enhance labeling efficiency at a lower TMT-to-protein mass ratio.[193] This may be determined by further optimization of organic content and temperature during TMT labeling.

We also demonstrated that the enrichment of low molecular weight proteins using 100 kDa MWCO only was sufficient for protein-level TMT labeling of complex samples to limit precipitation under the labeling conditions. Simplified sample preparation reduced the sample processing time and may prevent sampling bias. Other molecular weight strategies such as Gel-Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE)[194] can also be utilized in the future for enhanced enrichment of low molecular weight proteoforms. The presented protocol was optimized for low molecular weight proteoforms quantification using protein-level TMT labeling, which may not be readily applicable to larger proteins such as IgG (~150 kDa). In the future, MS compatible detergent such as 4-hexylphenylazosulfonate (Azo)[11,195], will be evaluated to optimize TMT labeling for high molecular weight proteins.

In addition, we will optimize the fragmentation parameters in the future to increase the number of proteoform identifications and increase proteome coverage. Also, fractionation methods such as high-pH and low-pH reversed phase chromatography[18,75,76] can also be applied to improve intact proteoform identification and characterization.

*The materials in Chapter 2 are adapted from:

Yanting Guo[*], Dahang Yu[*], Kellye A. Cupp-Sutton, Xiaowen Liu, and Si Wu. "Optimization of protein-level tandem mass tag (TMT) labeling conditions in complex samples with top-down proteomics." Analytica Chimica Acta (2022): 340037."

[*]Authors contributed equally

Author contributions: Yanting Guo: Methodology, Data curation, Formal analysis, validation, Writing - original draft. Dahang Yu: Methodology, Data curation, Formal analysis, Validation, Writing - original draft. Kellye A. Cupp-Sutton: Validation, Writing - review & editing. Xiaowen Liu: Software, Funding acquisition. Si Wu: Conceptualization, Validation, Supervision, Funding acquisition, Writing – review & editing.

Copyright information for Chapter 2:

Chapter 3 Optimization Higher-Energy Collisional Dissociation Fragmentation Energy for Intact Protein-level Tandem Mass Tag Labeling

3.1 ABSTRACT

Isobaric chemical tag labeling (*e.g.,* TMT) is a commonly used approach in quantitative proteomics and quantification is enabled through detection of low-mass reporter ions generated after MS2 fragmentation. Recently, we have introduced and optimized an intact protein-level TMT labeling platform that demonstrated >90% labeling efficiency in complex samples with top-down proteomics. Higher-energy collisional dissociation (HCD) is commonly utilized for isobaric tag labeled-peptides fragmentation because it produces accurate reporter ion intensities and avoids loss of low mass ions. HCD energies have been optimized for isobaric tag labeled-peptides but have not been systematically evaluated for isobaric tag-labeled intact proteins. In this study, we report a systematic evaluation of normalized HCD fragmentation energies (NCEs) on TMT-labeled *HeLa* cell lysate using top-down proteomics. Our results suggested that reporter ions often result in higher ion intensities at higher NCEs. Optimal fragmentation of intact proteins for identification, however, required relatively lower NCE. We further demonstrated that a stepped NCE scheme with energies from 30 to 50% resulted in optimal quantification and identification of TMT-labeled *HeLa* proteins. These parameters resulted in an average reporter ion intensity of ~4E4 and average proteoform spectrum matches (PrSMs) of >1000 per RPLC-MS/MS run with a 1% false discovery rate (FDR) cutoff.

3.2 INTRODUCTION

Top-down mass spectrometry (MS)-based proteomics analyzes intact proteins in complex samples, allowing for the characterization of intact proteoforms arising from genetic variation, alternative RNA splicing, and post-translational modifications (PTMs).[11,16,196] Quantitative top-down proteomics enables measurement of variation in proteoform-level expression resulting from different biological conditions to better understand protein function, cellular mechanism, disease state, and to aid in biomarker discovery.[11,15] Isobaric chemical tag labeling, such as isobaric tag for relative and absolute quantitation (iTRAQ)[158], tandem mass tag (TMT)[159,160]**,** and *N*, *N*-dimethylleucine (DiLeu)[162–164] has become one of the most popular approaches in quantitative bottom-up proteomics.[165,166] These isobaric chemical tags may be used to quantify multiple samples (*e.g.*, up to 18 samples with TMTpro 18-plex and 21 samples using 21-plex DiLeu[163]) in a single run allowing for simultaneous quantification and identification at the MS2 level. Isobaric chemical tag-based quantification has been extended to top-down proteomics for the analysis of standard proteins or simple protein mixtures[168–171] but has remained limited with respect to complex samples due to the presence of protein precipitation and side products during the labeling.[130,132] To overcome these challenges, our group recently developed and optimized an intact protein-level TMT labeling platform for the quantification of low molecular weight proteoforms (<35 kDa).[17] Moreover, a systematic evaluation of protein labeling conditions was performed using top-down proteomics and >90% labeling efficiency was achieved in complex samples.[130,132] Konrad *et al.* also applied thiol-directed isobaric labeling

(*e.g.*, cysteine labeling) for the quantification of *E. coli* and combined *E. coli*/yeast cell lysate using top-down proteomics.[173]

Protein quantification using isobaric chemical tag-based labeling relies on the efficient and accurate measurement of low-mass reporter ions generated in MS2 fragmentation; therefore, the selection and optimization of fragmentation techniques is essential to generate reporter ions and protein fragment ions simultaneously. Higher-energy collisional dissociation (HCD) fragments ions in a collision cell instead of an ion trap and operates at higher collisional energies than collision-induced dissociation (CID)[197]. HCD has been commonly utilized for isobaric-based quantification because HCD fragmentation resulted in high quality MS2 spectra by providing high resolution ion detection, no low-mass cutoff, and high intensity low-mass reporter ions that improve the accuracy and precision of isobaric-based quantification (*e.g.*, iTRAQ and TMT).[198,199]

It has been reported that quantification precision and accuracy decreased for isobaric tag-labeled precursors if reporter ion signals are low due to low HCD fragmentation energy; however, over-fragmentation issues due to high HCD fragmentation energy can result in less confident identification.[200] Therefore**,** a fine tuning of the collision energy for MS2 fragmentation is required for a balance between confident identification and highly sensitive and accurate quantification. Some groups have studied the influence of fragmentation energy on the reporter ion intensity for isobaric tag-labeled peptides, achieving a compromise between peptide identification and quantification.[200–205] A normalized collision energy (NCE) of 40-45% is used in many laboratories for routine analysis of TMT-labeled

peptides.[203,206–208] There have also been reports for evaluation of MS2 fragmentation on isobaric tag-labeled intact protein using solutions of standard protein. Chien-Wen *et al.* evaluated reporter ion intensities from intact TMT-labeled myoglobin using different HCD energies and found that reporter ion signals increased as NCE increased and reached a maximum at 30% NCE using a LTQ-Orbitrap Velos mass spectrometer.[171] Konrad and coworkers recently applied a combined fragmentation scheme for iodoTMTzero-labeled lysozyme (*e.g.*, cysteine labeling) using an Orbitrap Fusion Lumos Tribrid mass spectrometer. This study found that different fragmentation energies were required for efficient backbone fragmentation and cleavage of reporter ions.[173] Thus, they utilized a sequential scan approach with 80% HCD used for reporter ion quantification and electron-transfer dissociation (ETD) fragmentation or collision induced dissociation (CID) for intact protein fragmentation. To date, no systematic evaluation of HCD fragmentation energy on TMT-labeled intact proteins in complex samples such as cell lysate has been reported.

In this study, we systematically evaluated the effect of HCD normalized energy on fragmentation efficiency and reporter ion intensities of TMTzero-labeled *HeLa* cell lysate using an Orbitrap Exploris 240 mass spectrometer. Our results indicated that reporter ion intensities increased with increasing NCE, which provided improved quantification sensitivity and accuracy; however, high NCE (*e.g.*, 50% or higher) often introduced over-fragmentation of intact proteoforms, resulting in poor identification confidence. Based on these observations, we proposed a stepped HCD strategy with three normalized HCD energies from 30 to

50%. With the optimized HCD strategy, we confidently identified >1000 PrSMs with an average reporter ion intensity of ~4E4 in TMT-labeled intact proteins from *HeLa* cell lysate with a single LC-MS/MS run.

3.3. EXPERIMENTAL

3.3.1 Chemicals and materials.

Tris (2-carboxyethyl) phosphine hydrochloride (TCEP), Pierce™ BCA Protein Assay Kit, molecular weight cutoff filters (10K and 100K MWCO), and TMT isobaric label reagents (both TMTzero and TMT10plex) were obtained from Thermo Fisher (Waltham, MA, USA). LC-MS grade acetonitrile (ACN), isopropanol (IPA), trifluoroacetic acid (TFA), water used for LC mobile phases, and other chemicals were purchased from Sigma-Aldrich (St. Louis, MO, USA) unless noted otherwise.

3.3.2 *HeLa* cell lysate preparation.

HeLa cells were grown in Dulbecco's modified Eagle medium (DMEM), with 10% fetal bovine serum (FBS) and 2% penicillin-streptomycin under 5% CO_2 at 37 °C until 80-90% density was achieved. Then *HeLa* cells were collected and cell lysate were prepared as previously reported.[209] Protein concentration was measured by Pierce™ BCA Protein Assay Kit (Thermo Fisher). *HeLa* cell lysate was aliquoted and stored at -80 °C.

3.3.3 Protein-level TMT labeling.

Optimized protein-level TMT labeling protocol was reported here (*Chapter 617810432**Appendix A – A Benchmarking Protocol for Intact Protein-Level Tandem Mass Tag (TMT) Labeling for Quantitative Top-Down Proteomics***).[172]

Briefly, low molecular weight (<100 kDa) *HeLa* cell lysate proteins were enriched using 100 kDa MWCO spin filters and concentrated using 10 kDa MWCO spin filters. Low M.W. *HeLa* proteins were denatured by urea, reduced by TCEP, alkylated by IAA, buffer-exchanged to 100 mM TEAB buffer (pH 8.5), and concentrated to > 1 μg/μL using 10 kDa MWCO spin filters. TMTzero reagent was then added to the protein solution with a mass ratio of 8:1 (TMT-to-protein) and incubated for 1 hour at room temperature. Then, the same amount of TMTzero reagent was added for double labeling at room temperature for an additional hour, followed by quenching by hydroxylamine to a final concentration of 1.2% for 15 min. The TMT-labeled *HeLa* lysate was centrifuged at 12,000 × g and 4 ^{o}C for 30 minutes to remove any precipitation before LC-MS/MS analysis. Moreover, three aliquots of *HeLa* cell lysate proteins were individually labeled with TMT10plex tags (129C, 130C, 131) using the same conditions to evaluate the quantification. These TMT-labeled *HeLa* samples were then mixed with a mass ratio of 1:4:2 (129C:130C:131), centrifuged at 12, 000 × g and 4 ^{o}C for 30 minutes, and analyzed by nano-RPLC-MS/MS.

3.3.4 Top-down RPLC-MS/MS analysis.

All RPLC separations were performed on a modified Thermo Scientific (Waltham, MA, USA.) Accela LC system.[130,179–181] Mobile phase A (MPA) consisted of 0.01% TFA, 0.585% acetic acid, 2.5% isopropanol, and 5% acetonitrile in water. Mobile phase B (MPB) consisted of 0.01% TFA, 0.585% acetic acid, 45% isopropanol, and 45% acetonitrile in water. 5 μg of TMT-labeled *HeLa* lysate was loaded onto a trapping column (150 μm i.d., 5 cm length, Jupiter particles, 5 μm

diameter, 300 Å pore size), and then separated by a C2 capillary from CoAnn Technologies (100 μm i.d., 30 cm length, 3 μm diameter, 300 Å pore size) with a flow rate of 400 nL/min. The LC eluent was analyzed using an Orbitrap Exploris 240 mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a customized nano-ESI interface.[14] A 60-min gradient from 10% to 70% of MPB was applied. MS parameters were set as follows: inlet capillary temperature was 275 °C; spray voltage was 3.0 kV; resolution for MS1 and MS2 was set to 120, 000 and 60, 000, respectively; AGC target was 1×10^6 with 2 microscans for both MS1 and MS2 scans. Maximum injection time was 500 ms for MS1 and 300 ms for MS2 scans; the isolation window was set as 2 m/z. Dynamic exclusion was 90 s; top 6 most abundant precursor ion peaks (charge 4-50) in MS1 scans were selected for MS2 fragmentation. Normalized HCD energy was varied for MS2 fragmentation: single energies of 25%, 30%, 35%, 40%, 45%, 50% and 80%; stepped energies of 30/40/50% and 35/45/50%. TMT-labeled *HeLa* lysate was analyzed in triplicate for each selected energy scheme.

3.3.5 Data analysis.

An in-house python package (https://github.com/OUWuLab/Yanting_TMT-extraction.git) was utilized to extract reporter ion peaks. TopPIC Suite was used for proteoform search against human protein database (UniProt, 2020-07-11, 20368 species).[182,183] Decoy database searching was used and the FDR cutoff was set as 0.01 for both spectrum and proteoform levels. The maximum number of mass shifts was 2 with TMTzero or TMT10plex at both *N*-terminal and lysine residues as the fixed modification. All

the other parameters were set as default. MASH Suite[184] and ProSight Lite[185] were used for manual interpretation and spectrum presentation.

3.4 RESULTS AND DISCUSSION

The fragmentation energy used for multistage fragmentation has an important impact on the quality of fragmentation spectra. In top-down MS, isobaric chemical tag-based quantification normally requires higher fragmentation energy to produce intense reporter ion peaks for accurate and sensitive quantification.[203] However, high fragmentation energy may also introduce over-fragmentation of intact proteoforms, resulting in low-quality MS2 spectra that do not result in confident proteoform identification.[28] This is especially true for low abundance isobaric tag-labeled species in complex samples. Therefore, a balanced fragmentation energy is required to obtain both confident proteoform identification and highly sensitive and accurate quantification. Here, we systematically evaluated different normalized HCD energies for optimal proteoform identification and reporter ion intensity using an Orbitrap Exploris 240 mass spectrometer for TMT-labeled intact proteoforms in *HeLa* cell lysate.

3.4.1 Evaluation of normalized HCD energy for optimal reporter ion intensities

An in-house python package was developed to extract TMTzero reporter ion peaks from mzML files (converted from RAW using MSConvert[210]). The reporter ion intensities from all MS2 scans collected using various normalized HCD energies (NCEs) were plotted using a box plot, not including the scans that reporter ions were not detected, as shown in ***Figure 3-1A***. As HCD fragmentation energy

increased (25-80%), average reporter ion intensity increased from 4.6E3 at NCE=25% to 1.8E5 at NCE=80% (~38-fold increase). We also plotted the reporter ion intensities for all proteoform spectrum matches (PrSMs) (*i.e.*, identified MS2[35] spectra, not including the scans that reporter ions were not detected) with different HCD energies in ***Figure 3-1B***. The overall trend was consistent with ***Figure 3-1A***, but there was less variation observed in HCD 45-80% after matching to PrSMs. In addition, the percentage of PrSMs without reporter ion detected in each NCE scheme was also evaluated. As shown in ***Figure 3-1C***, the percentage of PrSMs without reporter ion detected decreased as NCE increased from 25% to 80% as expected. 47.84±5.20% PrSMs did not have reporter ions detected at NCE = 25%, while it decreased to 0% at NCE = 80% and close to 0% (0.47±0.20%) at NCE = 50%. In summary, our results suggested that higher HCD energies generally produced higher reporter ion intensities, which can be beneficial for TMT quantification.[28]

The relationship between precursor intensity and reporter ion intensity in PrSMs was also explored by the scatter plots in ***Figure 3-1D,*** not including the scans that reporter ions were not detected. A slight trend was observed that indicated precursor intensity was proportional to reporter ion intensity. Overall, higher HCD energies produced reporter ions with higher intensities for mass features with similar intensities. For example, for features with intensities from 7.9E7 to1.3E8 (log value 7.9-8.1), average reporter ion intensities varied from 1.33E3 to 2.63E5 as NCE increased from 20% to 80% (~200-fold increase). However, even though 80% NCE generated the highest average reporter ion

intensities (~2.63E5), a very small number of PrSMs were found. This is likely due to over-fragmentation of proteoforms caused by high fragmentation energy resulting in fewer identifiable terminal fragment ions. On the other hand, we noticed some PrSMs with relatively high reporter ion intensities at low NCEs (e.g., NCE <35%). Notably, most of these PrSMs are proteoforms with low MW (< 5 kDa) that are less likely to be identified using higher NCEs due to over-fragmentation.

The intact parathymosin proteoform (Uniprot# P20962, ***Figure 3-2*****)** demonstrated how the reporter ion intensity changed across different HCD energies. The observed monoisotopic mass was 14348.13 Da with 3.62 ppm difference to the theoretical monoisotopic mass of 14348.19 Da. As shown in ***Figure 3-2***, for parathymosin with charge state = 15, no reporter ion was detected at NCE = 25% and the precursor ion peak (957.03 m/z, z = 15) was still the most abundant peak in MS2 spectra. The reporter ion intensities increased from 8.80E2 at NCE = 30% to 2.95E5 at NCE = 80%, suggesting a ~335-fold improvement. However, as mentioned above, when the normalized HCD energy is too high (*e.g.*, NCE = 80%), over-fragmentation may be observed, which can be seen from the poor MS2 spectra quality at NCE = 80% where very few identifiable peaks were observed. In fact, when the NCE = 50%, we observed fewer fragment peaks, particularly for large MW fragment ions.

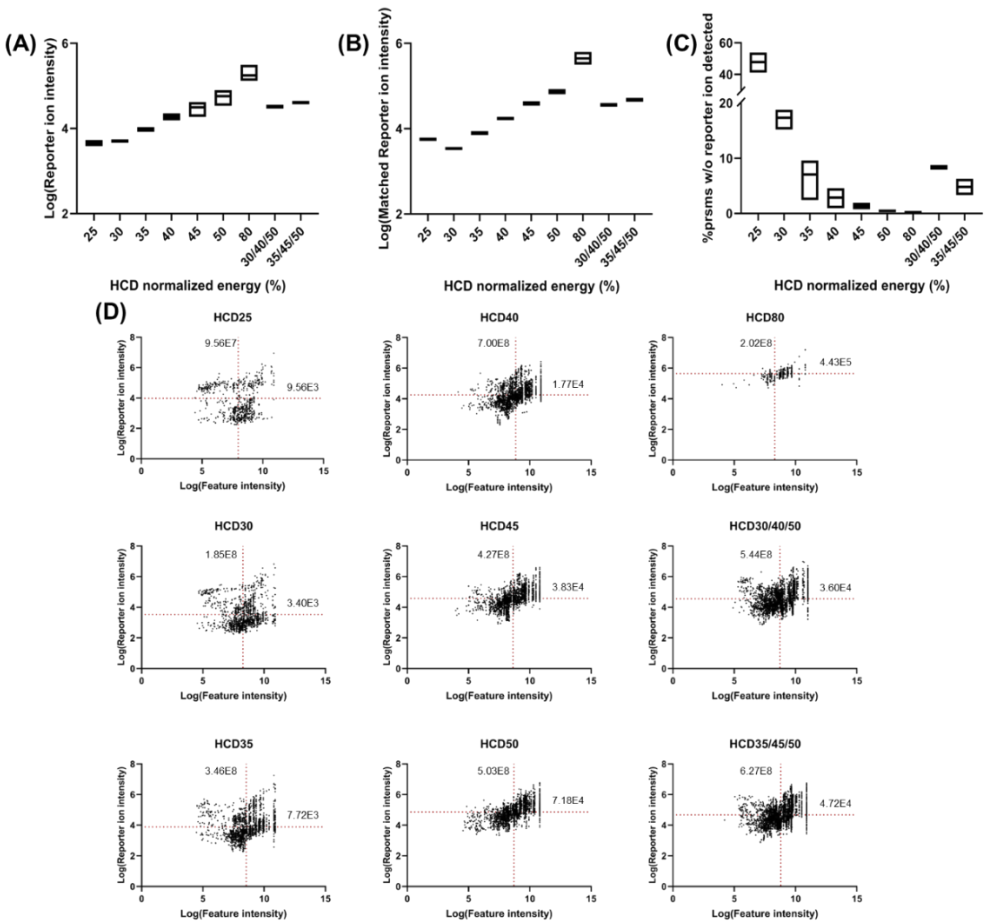


Figure 3-1 Evaluation of reporter ion intensities. Box plot representation of **(A)** average reporter ion intensities for all MS2 scans (excluding scans with no reporter ions detected); **(B)** average reporter ion intensities for all PrSMs (excluding scans with no reporter ions detected); **(C)** percentage of PrSMs with no reporter ions detected; (D) scatter plots of precursor ion intensities versus reporter ion intensities in PrSMs (excluding scans with no reporter ions detected). The red dotted lines represent the mean values of log (reporter ion intensity) and log(feature intensity) for different NCEs.

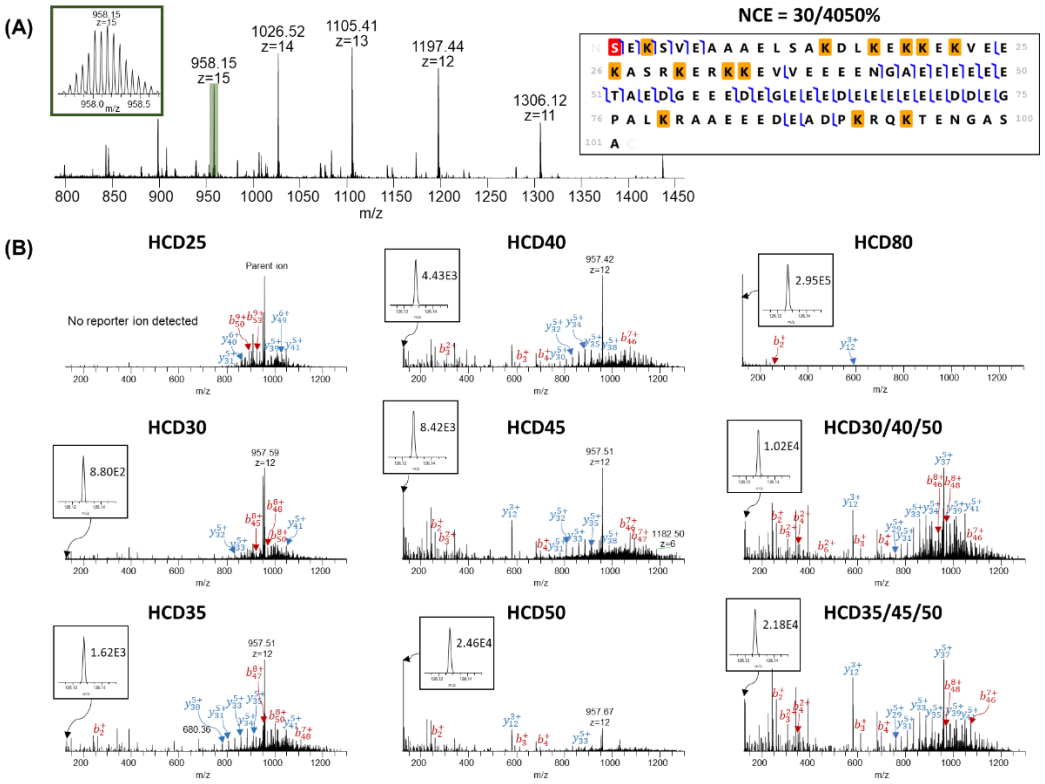


Figure 3-2 Parathymosin protein (Uniprot#P20962). (A) MS spectrum (inset: isotopic distribution of charge state 15) and representative protein sequence fragmented using SNCE = 30/40/50%. Red highlight represents acetylation and yellow highlight represent TMT tags. **(B)** MS/MS identification spectra for charge state = 15 using single HCD energies and SNCE schemes. Reporter ion peak and intensity are shown in the inset for each energy in which a reporter ion was detected.

Overall, higher normalized HCD fragmentation energy generated higher reporter ion intensities, which can be beneficial for improved accuracy and sensitivity of the quantification. However, high normalized HCD fragmentation energy may also lead to over-fragmentation, resulting in poor sequence coverage and identification. Therefore, for balanced identification and quantification, the effect of normalized HCD fragmentation energies on intact proteoform identification should also be evaluated.

3.4.2 Evaluation of the effect of normalized HCD fragmentation energy on intact proteoform identification

To determine the optimal normalized HCD energy for proteoform identification, the number of PrSMs matched and unique proteoforms identified using each energy by TopPIC Suite were evaluated.[183] As shown in ***Figure 3-3A***, the number of PrSMs increased as NCE increased from 25-40% and remained approximately constant for NCE = 40-50% with approximately 800 PrSMs. However, PrSMs dramatically decreased to ~100 when NCE was increased to 80%. ***Figure 3-3B*** similarly demonstrated that for NCE <50% the number of unique proteoforms that were identified had a similar trend to the PrSMs and again the number of unique proteoforms identified with NCE = 80% decreased (~20 proteoforms) due to over-fragmentation.

Identified PrSMs and proteoforms alone cannot entirely gauge the quality of MS2; these spectra can be further evaluated using the matched fragment peaks (*i.e.*, fragment peaks matched in MS2 spectra, one fragment ion may have multiple fragment peaks due to charge distribution). We plotted the average number of matched peaks for all PrSMs (***Figure 3-3C***) against different normalized HCD energies and found that the number of matched peaks increased as NCE increased from 25% to 45% but dramatically decreased for 50% and 80%. A similar trend was observed for the matched unique fragment ions (*i.e.*, matched fragment ions counted only once regardless of the number of charge states matched) as shown in ***Figure 3-3D***. Additionally, we examined the *E*-value (expected value, representing a nonlinear transformation of the number of matched peaks when searching the

spectrum against a protein database, the lower the E-value, the more significant the database match is)[125,211] for the PrSMs to determine the confidence of the identifications. The average -log(*E*-value) for all PrSMs were plotted against different normalized HCD energies, as demonstrated in ***Figure 3-3E***, identification confidence increased as NCE increased from 25% to 45% but decreased at 50% and 80%. Overall, the results suggested that as NCE increased from 25% to 45%, the identification confidence level increased, but decreased when NCE = 50% and 80%.

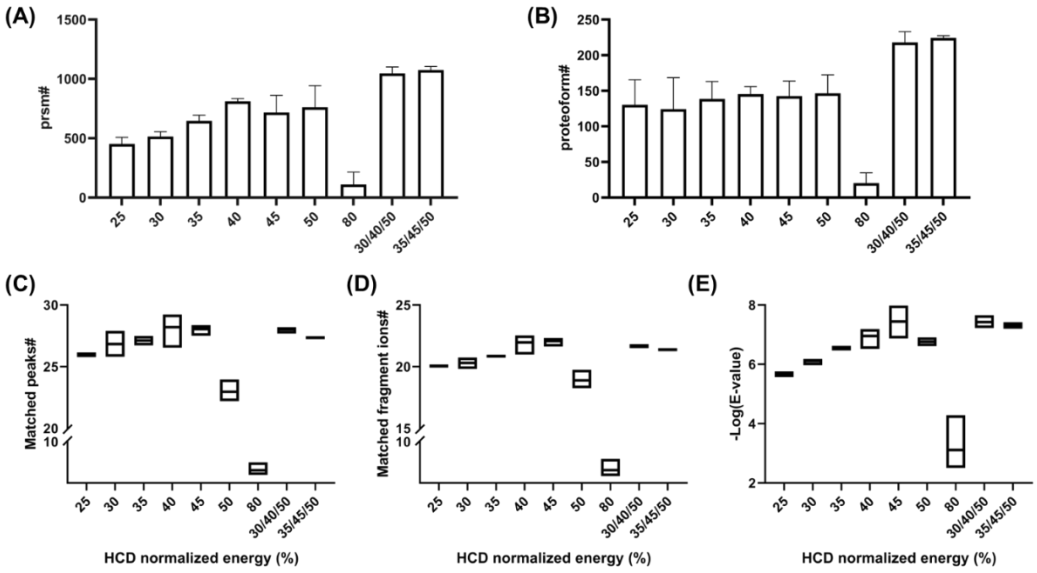


Figure 3-3 Identification of PrSMs and unique proteoforms using single NCEs and SNCE schemes. Bar graph representation of **(A)** matched PrSMs; **(B)** uniquely identified proteoforms. Box plot representation of the **(C)** average number of matched peaks for matched PrSMs; **(D)** average number of uniquely matched fragment ions for matched PrSMs; **(E)** average E-values for matched PrSMs. *Error bars represent the standard deviations.

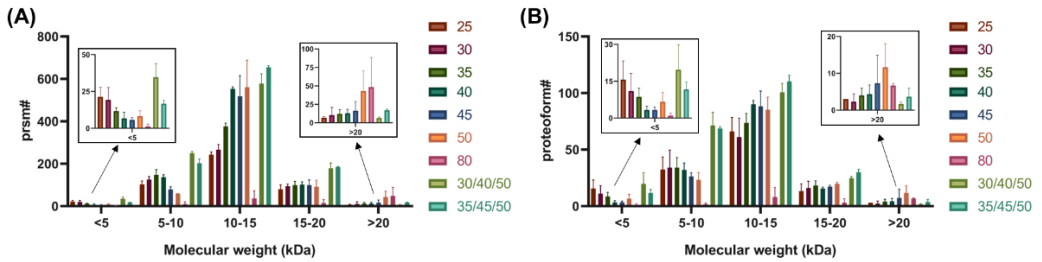


Figure 3-4 Bar graph depicting the number of (A) matched PrSMs and **(B)** proteoforms identified as a function of molecular weight for each single NCE and SNCE scheme. The legends show the NCE used for fragmentation and the error bars represent standard deviation. Inserts show PrSMs and proteoforms when molecular weight is below 5 kDa and above 20 kDa.

We further evaluated the fragmentation efficiency using different normalized HCD energies as a function of proteoform molecular weight (***Figure 3-4***). For moderately sized proteoforms (10-20 kDa), the number of PrSMs generated and proteoforms identified increased with increasing NCE up to 45-50%. These proteoforms represented approximately 70% of all identified proteoforms. Interestingly, for higher molecular weight proteoforms (>20 kDa), higher normalized HCD energies (50% and 80%) generated more PrSMs than the lower energies. This suggested that larger proteoforms may need a higher energy for efficient backbone fragmentation.[212] A 22 kDa intact proteoform of Stathmin (STMN1, Uniprot#P16949) was shown as an example in ***Figure 3-5***. The observed monoisotopic mass was 22358.55 Da with a mass measurement accuracy of 5.96 ppm compared with the theoretical mass of 22358.42 Da. The reporter ion intensities increased from 5.12E2 at NCE = 30% to 4.62E5 at NCE = 80%. No protein fragments were observed at NCE = 25% and 30% (no reporter ion was detected at NCE = 25%). The highest sequence coverage was observed at NCE = 50%, which yielded a 13% sequence coverage with the reporter ion intensity of

2.14E5 (***Figure 3-5B***). NCE = 80%, on the other hand, resulted in the highest reporter ion intensity (4.62E5), but no fragment ions were detected. We also noticed that the peak with the highest intensity at NCE= 50% was the precursor ion, which suggested that the optimal HCD NCE would be between 50% and 80%.

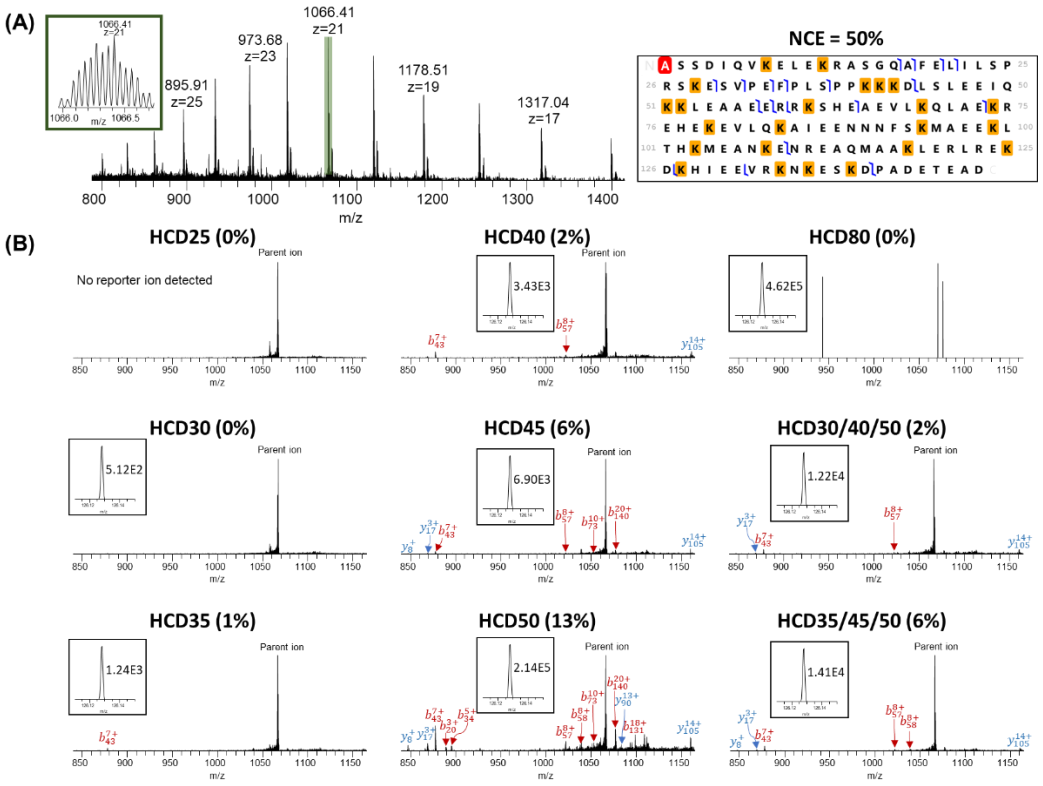


Figure 3-5 HCD analysis of Stathmin (STMN1, Uniprot#P16949) with different NCEs. (A) MS spectrum and representative fragmentation map using NCE = 50%. Red highlight represents acetylation and yellow highlights represent TMT tags; **(B)** MS/MS spectra for +21 charge state precursor ions using single NCEs and SNCE schemes. Reporter ion peaks and the intensities are shown in the inset.

Another 20 kDa intact proteoform of endothelial differentiation-related factor 1 (EDF1, Uniprot#O60869) was shown in ***Figure 3-6***. The observed monoisotopic mass was 20752.89 Da with the mass measurement accuracy of 10.55 ppm to its theoretical monoisotopic mass. For this proteoform, no fragments (either protein fragments or reporter ions) were detected at the NCE = 25% and 30%. The

reporter ion intensities increased from 1.00E3 at NCE = 35% to 9.54E4 at NCE = 80% (~95-fold). The precursor ion peak (903.80 m/z, z = 23) was the dominant peak when the NCE was below 45% with only a few fragment ions detected. The MS2 spectrum at NCE = 50% showed the highest number of detected fragment peaks and relatively high reporter ion intensity (1.37E4). NCE = 80% was too high for efficient terminal fragment detection, but this energy resulted in the highest intensity of reporter ion (9.54E4).

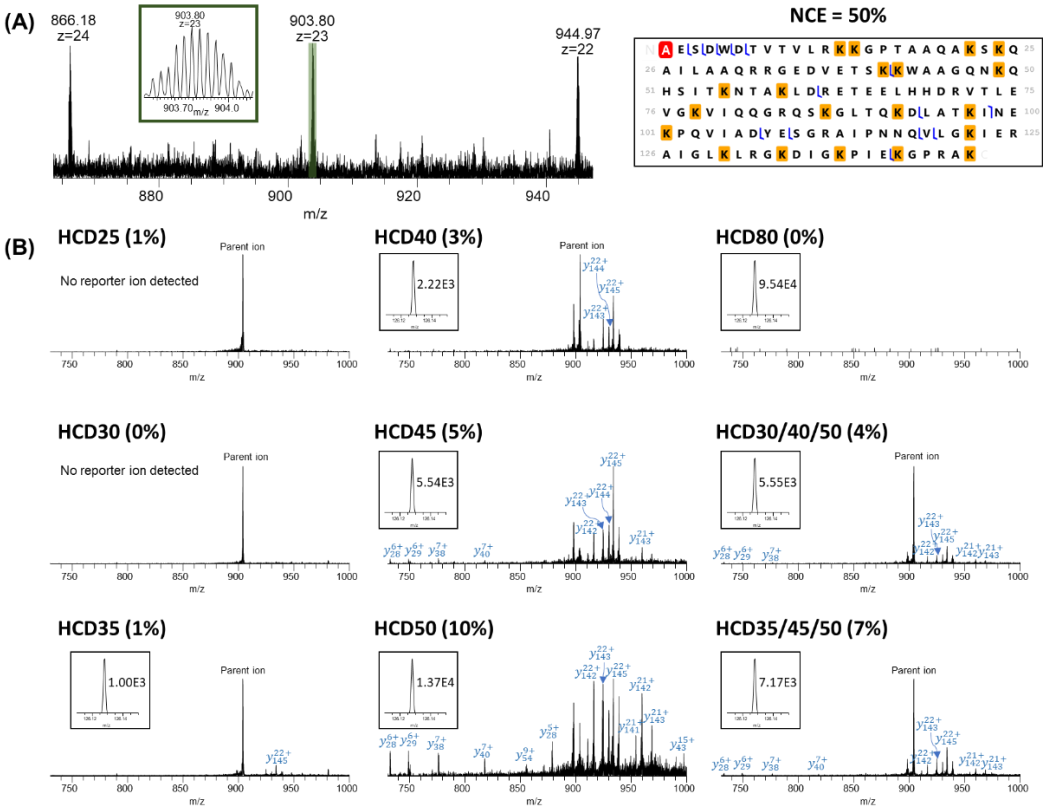


Figure 3-6 HCD analysis of Endothelial differentiation-related factor 1 (EDF1, Uniprot#O60869) with different NCEs. (A) MS spectrum and representative fragmentation map using NCE = 50%. Red highlight represents acetylation and yellow highlights represent TMT tags; **(B)** MS/MS spectra for +23 charge state precursor ions using single NCEs and SNCE schemes. Reporter ion peaks and the intensities are shown in the inset.

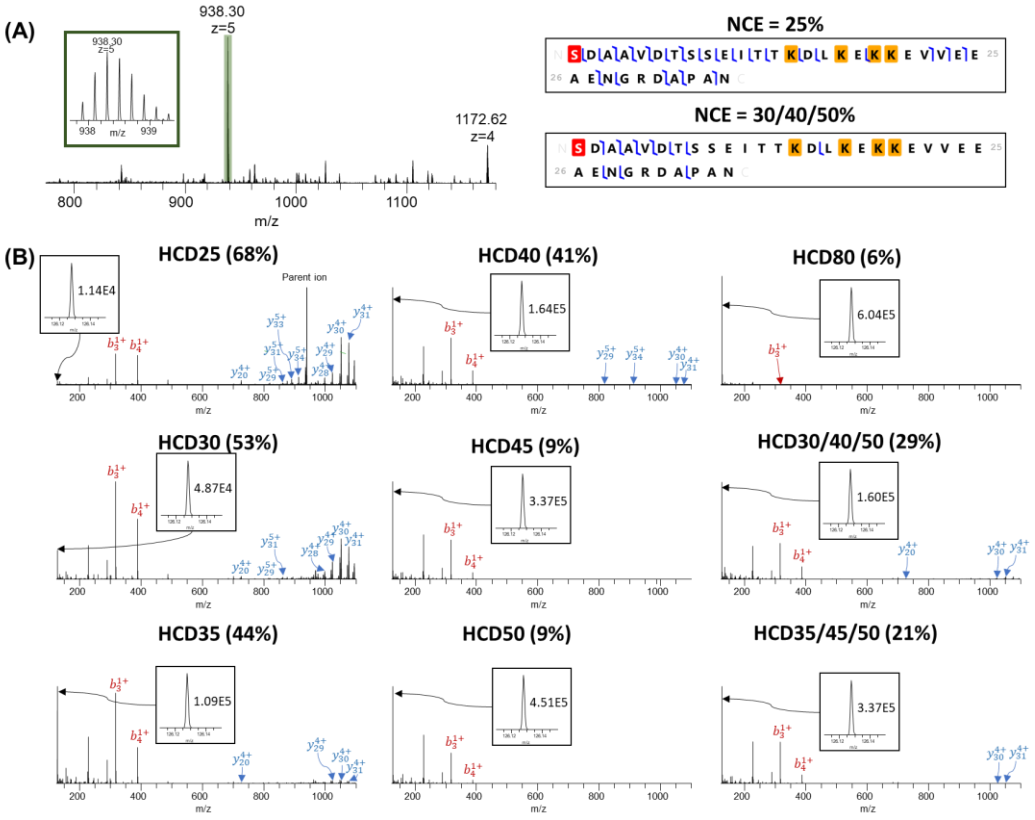


Figure 3-7 HCD analysis of Prothymosin alpha (PTMA, Uniprot#P06454) with different NCEs. (A) MS spectrum and representative fragmentation map using NCE = 25% and SNCE = 30/40/50%. Red highlight represents acetylation and yellow highlights represent TMT tags; **(B)** MS/MS spectra for +5 charge state precursor ions using single NCEs and SNCE schemes. Reporter ion peaks and the intensities are shown in the inset.

The trend was opposite for small proteoforms with M.W. < 5 kDa compared with large M.W. proteoforms. For these smaller proteoforms, the NCE = 25% identified the highest number of PrSMs. For example, MS2 spectra of a small intact protein prothymosin alpha (Uniprot#P06454, observed monoisotopic mass: 4684.43 Da; theoretical monoisotopic mass: 4684.46; mass measurement accuracy: 5.08 ppm, 938.30 m/z, z = 5) was evaluated as shown in ***Figure 3-7***. As expected, reporter ion intensities increased as NCE increased from 1.14E4 at 25% to 6.04E5 at 80% (~53-fold). However, sequence coverage decreased from 68% at NCE = 25% to 3% at NCE = 80%, suggesting smaller proteoforms (<5 kDa) may need a

lower fragmentation energy for better protein backbone fragmentation as opposed to larger proteoforms (>20 kDa).[212] For most proteoforms with M.W. >5 kDa, NCE = 25% may not be sufficient for complete protein fragmentation while NCE = 50% or 80% may start to cause over-fragmentation. For example, as shown in ***Figure 3-8A*****,** a fragmentation heatmap of parathymosin (Uniprot#P20962) at differing normalized HCD energies depicts all N- and C-terminal fragments (*i.e.*, internal fragments that contain neither N- or C-termini were not considered[213,214]). Most of the fragments of parathymosin were located in the *N*-terminal and the middle of the sequence. ~30% sequence coverage was achieved when NCE was below 40%. Fewer large fragment ions were detected when higher NCEs were applied (*e.g.*, NCE ≥ 40%). Bias against longer terminal fragment ions was previously reported to be linked with repeated backbone cleavage for internal fragment production, which increased with increased fragmentation energy.[215]

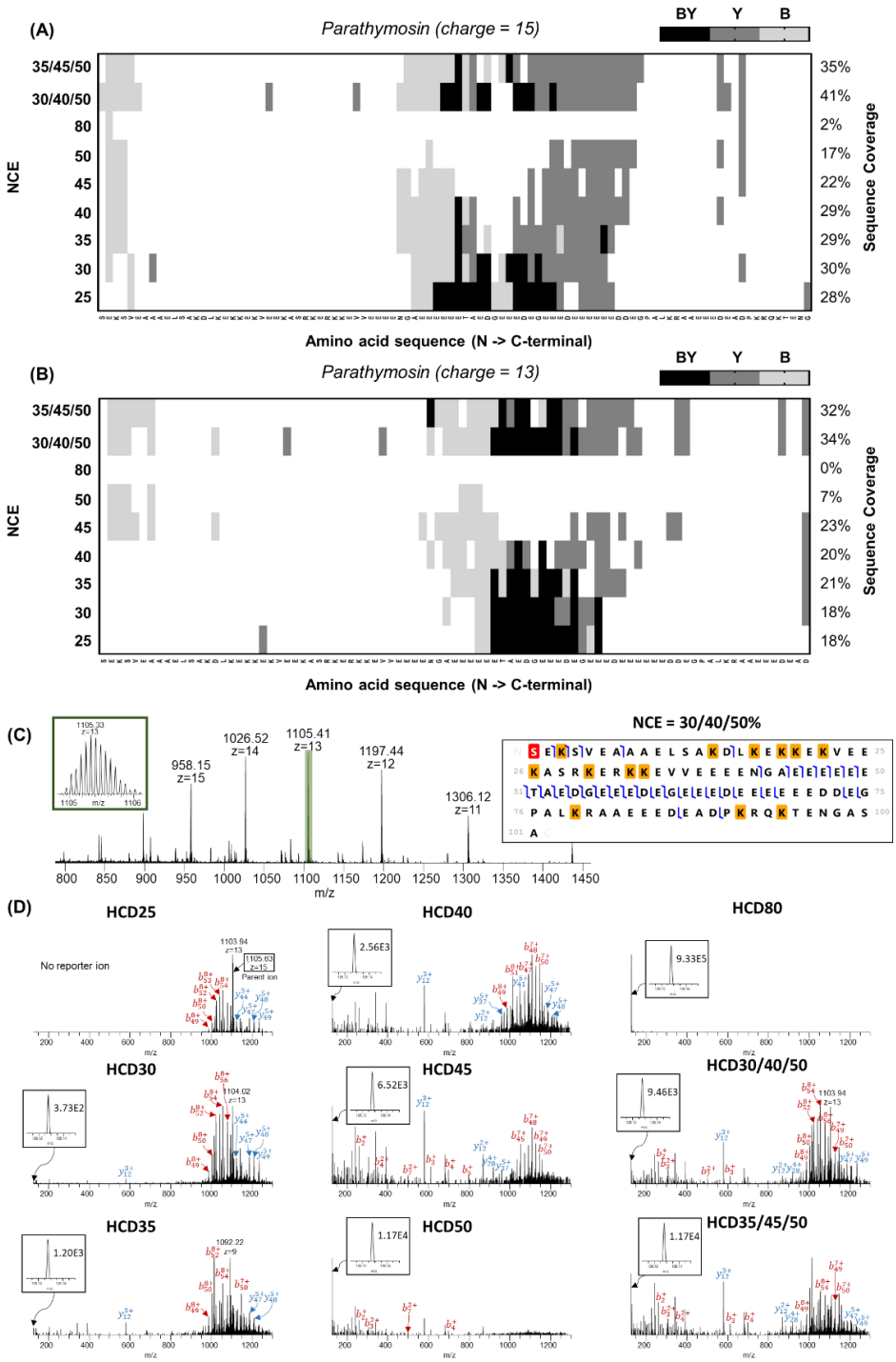


Figure 3-8 Fragment ion heatmaps of parathymosin protein (Uniprot#P20962) with (A) +15 charge state and **(B)** +13 charge state using different NCEs. Color scale represents the detection of b, y, and b/y ions; **(C)** MS/MS spectra of parathymosin with +13 charge state using different NCEs; **(D)** representative fragment maps at different NCEs. Red highlight represents acetylation, and yellow highlights represent TMT tags.

3.4.3 Optimization of stepped normalized HCD fragmentation scheme.

Stepped normalized collisional energies (SNCE) fragments precursor ions at different energies (*e.g.*, low, medium, and high) and then all fragment ions are combined and detected simultaneously.[197] The use of SNCE is particularly helpful for TMT-labeled intact proteoforms in complex sample due to the high sample complexity and large range of proteoform size. Our results suggested that the SNCE fragmentation schemes with NCEs between 30% and 50% can efficiently fragment intact proteoforms with an extensive range of molecular weights (e.g., lower than 30 kDa) with reasonable reporter ion intensities. We selected 3 energies in this range for the SNCE schemes (NCE = 30/40/50% and 35/45/50%). The NCE = 80% was not included in the initial test because very few fragment ions were detected for protein identification. NCE = 25% was not included because of low reporter ion intensities and the presence of intense peaks representing unfragmented precursor ions. Use of SNCE aims to improve both broad proteoform identification and TMT reporter ion quantification.

Overall, both fragmentation schemes generated relatively high reporter ion intensities, an average intensity of 3.24E4 for SNCE = 30/40/50% and an average intensity of 4.03E4 for SNCE = 35/45/50%, respectively (***Figure 3-1***). These results were similar to the average reporter ion intensity from the NCE = 45% run but with lower variation. No significant difference was observed between these two schemes but the average reporter ion intensity with the 35/45/50% scheme was slightly higher since relatively higher average NCE was used. As shown in ***Figure 3-1D,*** many low abundance proteoforms showed high reporter ion intensities with

stepped HCD energies. For example, for features with intensities from 7.9E7 to1.3E8 (log value 7.9-8.1, ~10E3 fold lower than the highest proteoforms intensity), average reporter ion intensities were 3.14E4 at SNCE = 30/40/50% and 2.87E4 at SNCE = 35/45/50%.

Also, the fragmentation schemes with stepped HCD energies (30/40/50% and 35/45/50%) in general identified more PrSMs and unique proteoforms compared to other energies (***Figure 3-1A&B***). In 30/40/50% scheme, 1047±44 PrSMs and 218±12 unique proteoforms were confidently identified with an average E-value of 3.8E-8. In 35/45/50% scheme, 1076±24 PrSMs and 224±3 unique proteoforms were confidently identified with an average E-value of 5.0E-8. For experiments using single NCE values, the NCE = 40% provided the highest number of PrSMs (811±19), NCE = 50% provided the highest number of unique proteoforms (146±21), and the NCE = 45% demonstrated the lowest average E-value as 3.6E-8. Overall, the stepped schemes allowed for improved proteoform identification with higher confidence. In addition, SNCE = 35/45/50% provided slightly higher reporter ion intensities and identified more PrSMs and proteoforms compared to SNCE = 30/40/50%, but the average E-value was slightly higher in NCE 35/45/50%, suggesting slightly better identification confidence in NCE = 30/40/50%. The SNCE schemes significantly improved the identification of proteoforms with molecular weight <20 kDa and especially for proteoforms from 5-10 kDa, there was a >2-fold improvement of the number of identified PrSMs and proteoforms compared with all schemes using single NCE value. For high M.W. proteoforms (>20 kDa), higher NCE energy (*e.g.*, larger than 50%) resulted in

improved fragmentation, so the current stepped scheme may not outperform a scheme using a single NCE value (50% or 80%).

As shown in ***Figure 3-2,*** parathymosin (charge state = 15) showed reporter ion intensities of 1.02E4 at SNCE = 30/40/50% and 2.18E4 at SNCE = 35/45/50%. Although these reporter ion intensities were similar to the reporter ion intensities at NCE = 45% (8.42E3) and 50% (2.46E4), there were more matched fragment peaks in the MS2 spectra (142 for NCE = 30/40/50% and 101 for 35/45/50%) when compared with NCE = 45% and 50% (35 and 33, respectively). A higher number of matched fragment peaks result in improved MS2 spectra quality using SCNE schemes. A similar trend was observed for charge state = 13 (1105.41 m/z) (***Figure 3-8 B-D***), higher sequence coverage was achieved in the SNCE runs (34% and 32% for 30/40/50% and 35/45/50%, respectively) compared to runs using single NCEs. Reporter ion intensities increased from 0 to 9.33E5 as NCE increased from 25% to 80%, similar to those detected for charge state = 15.

In addition, the largest fragment was 10333.81 Da at SNCE = 30/40/50%, and 8742.64 Da at SNCE = 35/45/50% compared with 7882.33 Da at NCE = 45% and 7495.21 Da at NCE = 50%, suggesting SNCE performed better to maintain large fragments compared with the schemes using single NCE values (figure). We also evaluated the largest fragment detected using different NCEs as well as SNCEs in other protein examples (

Table 3-1). Our results suggested that SNCEs were able to maintain large fragments for a variety of proteoforms with different molecular weights (*e.g.*, 5-22 kDa). As shown in ***Figure 3-8A***, higher sequence coverage was observed in the SNCE runs

(41% and 35% for 30/40/50% and 35/45/50%, respectively for parathymosin (charge state = 15, 958.09 m/z). Sequence coverage decreased to 22% for NCE = 45% and 17% for NCE = 50% due to relatively poor MS2 spectra. Although NCE = 80% provided the highest reporter ion intensity (2.95E5), over-fragmentation resulted in only 2 matched fragment peaks and 2% sequence coverage. When NCE < 45%, the reporter ion intensity was < 5E3 (no reporter ion peak was detected at NCE = 25%) and fewer fragment peaks (<100) were matched, resulting in approximatedly30% sequence coverage.

Overall, the results suggested the SNCE scheme outperformed schemes using single HCD energy to balance accurate quantification and confident identification.

Table 3-1 Largest fragment mass for example proteins at different NCEs.

Parathymosin (PTMS, Uniprot# P20962)				
	Charge = 15		Charge = 13	
HCD	Largest fragment observed mass (Da)	Ion type	Largest fragment observed mass (Da)	Ion type
25	9430.85	B63	14100.91	B98
30	9430.93	B63	13280.59	Y93
35	9115.75	B60	10320.22	B70
40	8613.56	B56	8112.40	B51
45	8011.36	B50	7882.32	B49
50	7753.30	B48	7495.16	B46
80	N/A	N/A	1733.97	Y12
30/40/50	9115.77	B60	10333.76	Y77
35/45/50	8871.72	B58	8742.70	B57
Stathmin (STMN1, Uniprot#P16949)			***Endothelial differentiation-related factor 1 (EDF1, Uniprot#O60869)***	
HCD	Largest fragment	Ion type	Largest fragment	Ion type

	observed mass (Da)		observed mass (Da)	
25	N/A	N/A	15168.74	Y106
30	N/A	N/A	N/A	N/A
35	6134.48	B43	20639.90	Y146
40	16233.94	Y105	20382.74	B146
45	21512.10	B140	20423.80	Y144
50	21512.10	B140	20510.83	Y145
80	N/A	N/A	N/A	N/A
30/40/50	16233.94	Y105	20308.84	Y143
35/45/50	16233.94	Y105	20639.93	Y146
Prothymosin alpha (PTMA, Uniprot#P06454)				
HCD	Largest fragment observed mass (Da)		Ion type	
25	4555.39		Y34	
30	4440.36		Y33	
35	4369.33		Y32	
40	4555.38		Y34	
45	386.14		B4	
50	386.14		B4	
80	386.14		B4	
30/40/50	4369.33		Y32	
35/45/50	4298.30		Y31	

3.4.4 Evaluation of the effect of normalized HCD fragmentation energy on TMT-based quantification

To verify the labeling accuracy, the intensity ratios among each pair of samples were compared (130C/129C, 131/129C) for all the scans in which all reporter ions were detected (***Figure 3-9A***). In all single NCE runs (*e.g.*, from 25% to 80%), the average reporter ion ratios were similar to the theoretical ratios for both sample ratios. For example, the average reporter ion ratios between 130C/129C were between 3.37 ± 1.41 (NCE =25%) and 3.93 ± 2.13 (NCE=50%), which is consistent with the theoretical ratio of 4. In both SNCE schemes, the average reporter ion ratios were also close to the theoretical ratio (*e.g.*, 3.94 ± 2.31 for SNCE 30/40/50% and 3.89 ± 2.21 for SNCE 35/45/50%). However, the average

reporter ion intensities were much lower (3.24E3 for 130C tag) in NCE of 25% compared with higher NCE values (*e.g.*, 8.02E4 for 130C tag at NCE = 50%). A similar trend was observed when the average reporter ion ratios were calculated for only scans that produced PrSMs, as shown in ***Figure 3-9B***. Our results suggested that MS/MS scans in which all report ions were detected provided good quantification accuracy, even with low reporter ion intensities. However, when NCE value was too low, many MS/MS scans contained missing reporter ions (shown in ***Figure 3-9C***). ≥ 50% of the scans showed missing reporter ions using NCE ≤ 30%, while most MS/MS scans at NCE ≥ 50% had no missing reporter ions (e.g., <1 % in NCE =50%). Overall, our results indicated that when single NCEs are used, higher NCE (*i.e.*, ≥ 40%) was required for a relatively accurate quantification with no missing reporter ions. The SNCE schemes (30/40/50% and 35/45/50%) provided good quantification accuracy compared to single NCEs.

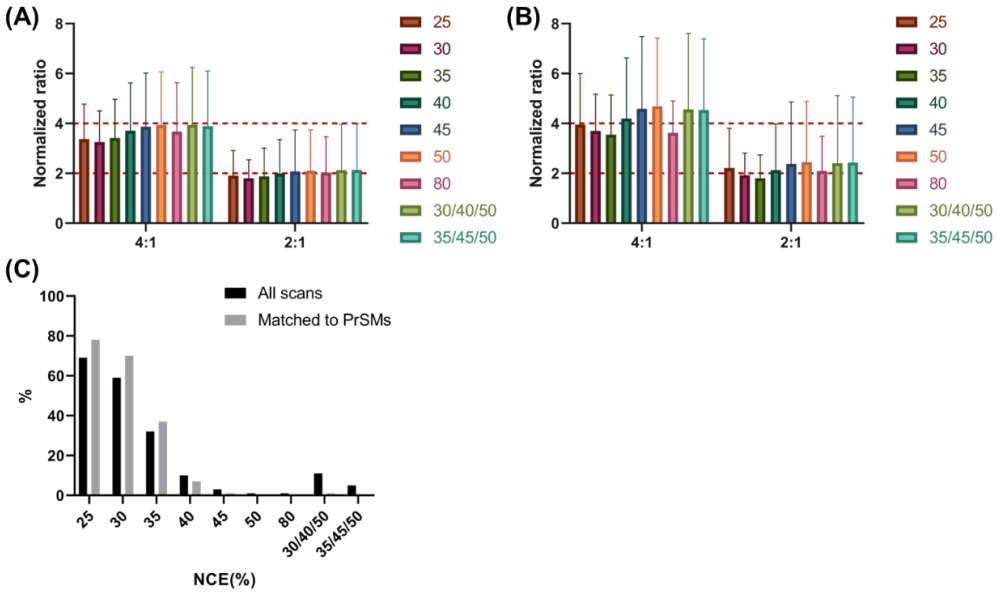


Figure 3-9 Evaluation of quantification accuracy using different NCEs. Bar graph representations of the normalized intensity ratio (130C/129C, 131/129C) of **(A)** all MS/MS scans and **(B)** all PrSMs. In these graphs, scans with any missing

tags were not included. Error bars represent the standard deviation of reporter ion ratios from all scans. Red lines represent theoretical normalized ratios, 4:1 and 2:1. **(C)** Percentage of MS/MS scans (black) or PrSMs (grey) with missing detection of reporter ions at different NCEs.

We further evaluated the identification and quantification of multiple proteoforms of the same protein. ***Figure 3-10*** shows two fully labeled proteoforms of heterogeneous nuclear ribonucleoprotein U (HNRNPU, Uniprot#Q00839) using SNCE = 30/40/50%. The initiator methionine was removed for both proteoforms. Proteoform 1 contained an acetylation at the N-terminal serine, and proteoform 2 included the addition of an unlocalized phosphorylation. Our results demonstrated similar fragmentation patterns between these two proteoforms using the SNCE schemes. No neutral loss peaks were detected for the phosphorylated proteoform, consistent with previous reports.[216] Quantification accuracy was evaluated by calculating normalized ratios of reporter ion intensities (130C/129C, 131/129C) for all the scans that matched to PrSMs of these HNRNPU proteoforms. The quantification accuracy between the two proteoforms was similar, with normalized ratios of 3.53 ± 0.20 and 1.95 ± 0.13 for proteoform 1 and 3.73 ± 0.28 and 2.04 ± 0.12 for proteoform 2. Similar results were observed for two Calmodulin-1 (CALM1, Uniprot#P0DP23) proteoforms using SNCE = 30/40/50% (***Figure 3-11***). The initiator methionine was removed for both proteoforms. Proteoform 1 had 1 acetylation at the N-terminal alanine and 1 trimethylation at K116; proteoform 2 had 1 acetylation and 1 unlocated phosphorylation. Similar quantification accuracy was achieved for the two proteoforms. In summary, similar quantification accuracy and fragmentation patterns were observed for the same protein with different PTMs.

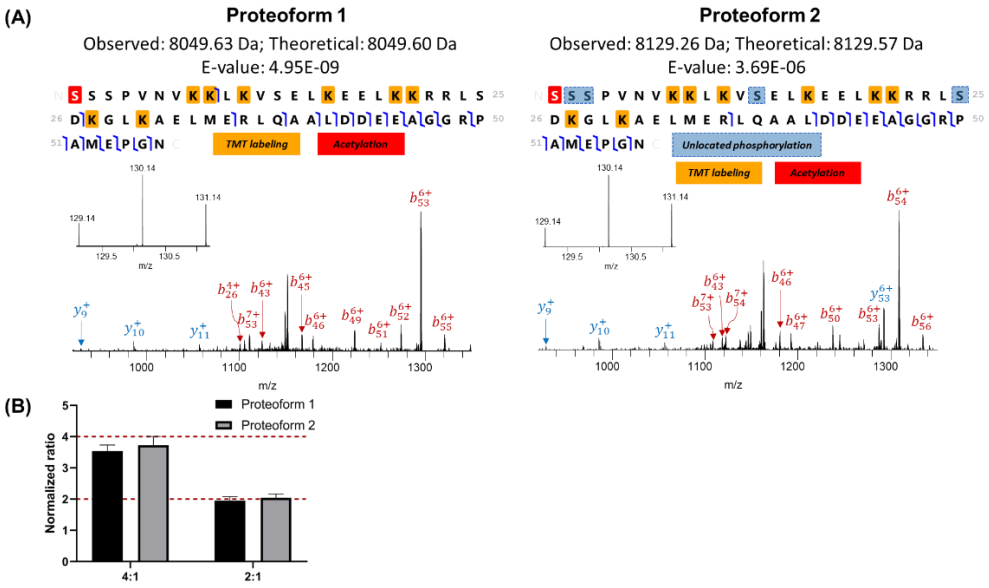


Figure 3-10 Evaluation of the effect of PTMs on quantification and identification. (A) Sequence and MS/MS spectra for two proteoforms of heterogeneous nuclear ribonucleoprotein U (HNRNPU, Uniprot#Q00839) at SNCE = 30/40/50%. **(B)** Bar graph showing quantification of the two proteoforms. Error bars shows the standard deviation of reporter ion intensities among all identified PrSMs for HNRNPU.

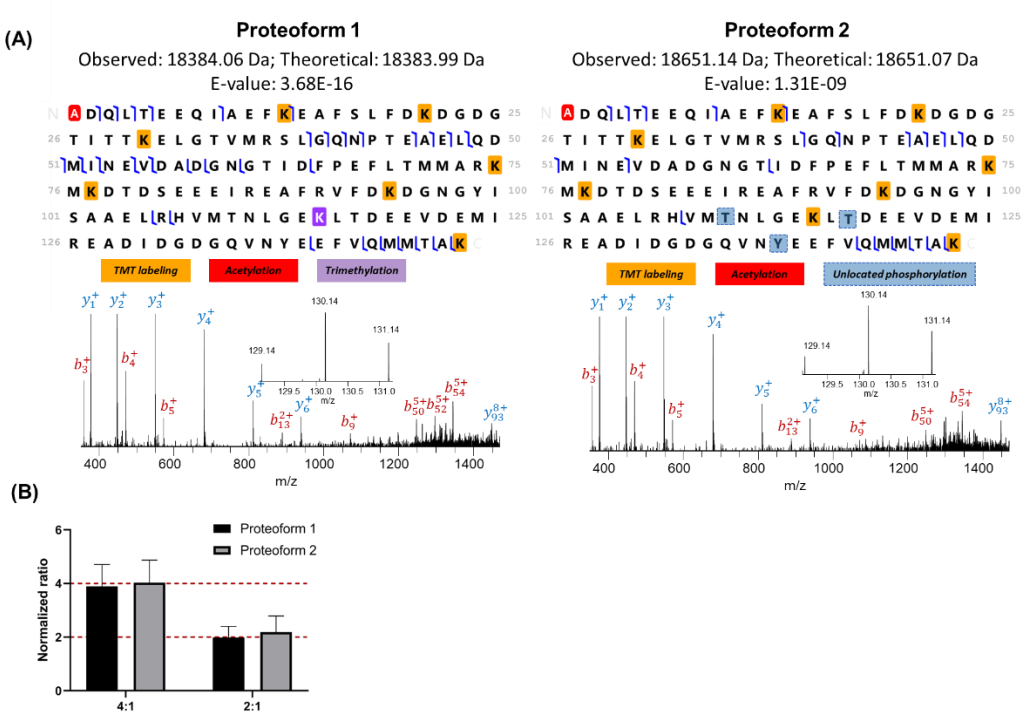


Figure 3-11 Evaluation of the effect of PTMs on quantification and identification. (A) Sequence and MS/MS spectra for two proteoforms of Calmodulin-1 (CALM1, Uniprot#P0DP23) at SNCE = 30/40/50%. **(B)** Bar graph

showing quantification of the two proteoforms. Error bars shows the standard deviation of reporter ion intensities among all identified PrSMs for CALM1.

3.5 CONCLUSION

Here, we have explored the effect of variation in HCD fragmentation energies on reporter ion intensity and proteoform identification for TMT-labeled intact proteoforms in *HeLa* cell lysate. Our results indicated that although SNCE energies did not provide the highest reporter ion intensity, SNCEs outperformed single NCEs by providing a balance between high reporter ion intensity that was sufficient for accurate and sensitive quantification and confident identification leading to more PrSMs and identified proteoforms More matched fragment peaks (including some large fragments) and relatively low E-values were also observed using SNCEs. In this study, we evaluated two SNCE schemes with NCEs between 30% and 50%. The development and optimization of new stepped schemes may be required for different biological samples with different applications. For example, single cell proteomics (SCP) may require the inclusion of higher NCE (e.g., 80%) to generate sufficient reporter ion intensities in each cell channel. Higher NCE may also be beneficial for high MW proteins. On the other hand, the inclusion of lower NCE, such as 25%, may increase the sequence coverage for identification of site-specific modifications.[203]

Quantification accuracy was also evaluated at different NCEs, and the results suggested that when lower NCEs were applied, quantification accuracy was significantly skewed due to the missing detection of reporter ions. NCE $\geq$ 40% was required to obtain good quantification accuracy. SNCEs provided good overall quantification accuracy. The effect of PTMs on identification and quantification of

the same protein was also evaluated and similar quantification accuracy was observed. Future work should be done to systematically evaluate the effect of PTMs on identification and quantification in a proteome-wide range.

We also observed that optimal NCE was dependent on the proteoform size and larger proteoforms (>20 kDa) require higher NCE (*e.g.*, NCE = 50%) for better quantification and identification. SNCE schemes, however, significantly improved the identification for proteoforms with molecular weight <20 kDa.[212] Thus, molecular weight-based fractionation methods such as Gel-Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE)[63,194], Passively Eluting Proteins from Polyacrylamide gels as Intact species for MS (PEPPI-MS)[71,72], and size exclusion chromatography (SEC)[24,52,217] may be considered for prefractionation so that single NCEs or more targeted SNCEs may be used for improved proteoform identification in a narrow M.W. range for a better balance between quantification sensitivity and identification confidence. As mentioned previously, the TMT labeling protocol was optimized for quantification of low molecular weight proteoforms[132], thus sample preparation to observe larger proteins such as IgG (~150 kDa) need to be optimized in the future using MS compatible detergent (*e.g.*, Azo)[11,132,195]. In this case, the optimization of NCE used for TMT-labeled larger intact proteoforms may also need to be optimized. In addition, some multidimensional separation approaches (*e.g.*, high-pH/low-pH RPLC)[18,75,76,209] can also be implemented to reduce sample complexity for improved identification while maintaining high quantification sensitivity and accuracy. Furthermore, the optimal collision energy may vary depending on the instruments and samples, therefore, the optimization of

fragmentation energy may be is necessary before the MS analysis. Other instrument parameters such as maximum injection time and auto gain control (AGC) can also be optimized. Some other combined fragmentation approaches may also be considered, such as EThcD (combined electron-transfer dissociation, ETD and HCD) for a balance between identification and quantification.[173]

*The materials in Chapter 3 are adapted with permission from:

Yanting Guo, Trishika Chowdhury, Meena Seshadri, Kellye A. Cupp-Sutton, Qingyu Wang, Dahang Yu, and Si Wu. “Optimization of Higher-Energy Collisional Dissociation Fragmentation Energy for Intact Protein-Level Tandem Mass Tag Labeling." Journal of Proteome Research Article (2023) DOI: 10.1021/acs.jproteome.2c00549. Copyright (2023) American Chemical Society

Author contributions: Yanting Guo: Methodology, Data curation, Formal analysis, validation, Writing - original draft. Trishika Chowdhury: Sample preparation. Meena Seshadri: Data analysis. Kellye A. Cupp-Sutton: Validation, Writing - review & editing. Qingyu Wang: Software. Dahang Yu: Validation. Si Wu: Conceptualization, Validation, Supervision, Funding acquisition, Writing – review & editing.

Copyright information for Chapter 3:

This introductory chapter was originally published with American Chemical Society publisher in Journal of Proteome Research on January 1st , 2023. This article, copyright © 2023 Guo, Chowdhury, Seshadri, Cupp-Sutton, Wang, Yu, and Wu is under the Copyright license (**Appendix C – Copyright licence for Chapter 3**) which states that the article contents can be shared (copied or redistributed) and adapted.

Chapter 4 Deep Quantification Of Intact Proteoforms Using Tandem Mass Tag Labeling And Online 2D High-pH/Low-pH-RPLC Top-Down Proteomics

4.1 ABSTRACT

Top-down proteomics has become a promising approach to characterize intact proteoforms at the proteoform level. Isobaric chemical tag labels (e.g., iTRAQ, TMT, and DiLeu) have been most frequently utilized as a standard quantification approach in bottom-up proteomics. We recently demonstrated the feasibility of intact protein-level TMT labeling for the identification and quantification of small intact proteoforms (< 35 kDa) in complex biological samples through the removal of large proteins prior to labeling with top-down proteomics; more than 90% labeling efficiency was achieved with optimized protein-level TMT labeling conditions in complex cell lysate sample. Furthermore, a stepped higher energy collisional dissociation (HCD) fragmentation scheme was optimized for a balance between accurate quantification and confident identification. The success of top-down TMT labeling allows identification and quantification of intact proteoforms with high multiplexing capacity, eliminates variability from downstream sample preparation (*e.g.*, peptide digestion), and enables protein-based multidimensional separation. To further improve the throughput and sensitivity, we integrated an online 2D high-pH/low-pH-RPLC system and TMT labeling for deep intact proteoform quantification in complex *Hela* cell lysate. In total, ~5600 unique proteoforms were detected from 20 μg TMT-labeled *HeLa* proteins. This research holds the potential for deep proteome

characterization and quantification with low microgram sample amounts and top-down TMT labeling approach.

4.2 INTRODUCTION

Top-down proteomics has become a promising approach to analyze intact proteoforms resulting from genetic variations, alternatively RNA splicing and post-translational modifications (PTMs).[3,16] Three quantitative methods have been successfully applied in quantitative top-down proteomics, including label-free quantitation, metabolic label-based quantitation, and isobaric chemical tag-based quantitation.[15] Label-free quantitation is mostly applied up to date due to its ease of operation as it doesn't require any labeling steps. However, it suffers from limited multiplexing capacity, sensitivity to run-to-run instrument variation, as well as difficulty in combination with multidimensional separation. In metabolic labeling-based quantitation techniques, cells are cultured in the media with and without isotopically labeled compounds, and samples are relatively quantified based on the ratio of labeled and unlabeled proteins. Metabolic label-based quantitation directly probes the proteome in *vivo* and may also be able to be applied in animal models for proteomics analysis; however, the complexity of MS1 spectra is significantly increased due to the introduction of light/heavy isotope-labeled proteins, which further complicated the data analysis process.

Isobaric chemical labeling such as relative and absolute quantitation (iTRAQ)[158], tandem mass tags (TMT)[159,160], and N, N-dimethylleucine (DiLeu) isobaric tags[162–164] allow quantification on the MS2 level while not complicating

the MS1-level spectra. Such isobaric tags may be used to quantify up to 18 samples with TMTpro 18-plex and 21 samples with 21-plex DiLeu[163], which has been considered as a "gold standard" quantitation approach in bottom-up proteomics. Later, isobaric chemical tag-based quantification study was applied to pure protein or simple protein mixtures[168–171], but its application in complex samples (*e.g.*, cell lysate) is still challenging mainly due to protein precipitation introduced from organic solvent used during TMT labeling. Another issue is the formation of side products (*i.e.*, incomplete labeling or labeling of unintended residues) during TMT labeling.[130] To overcome these challenges, our group recently developed protein-level TMT labeling and quantification of intact proteoforms under 35 kDa in *E. coli* cell lysate by using a "filter-SEC" technique that combines a molecular weight cutoff filtration with size exclusion chromatography (SEC) to remove large proteoforms that tend to "crash" out of solution during TMT labeling condition.[130] Winkels *et al.* also further adapted top-down thiol-directed isobaric labeling (*i.e.*, cysteine labeling) to quantify and identify proteoforms in *E. coli* cell lysate.[173] Later, we optimized the protein-level TMT labeling conditions and achieved >90% labeling efficiency in *E. coli* cell lysate and 86.47% labeling efficiency in *HeLa* cell lysate. With optimized protein-level TMT conditions, side products can be limited, further decreasing the sample complexity and increasing signal-to-noise ratio.[132,218] In addition, MS2 fragmentation was also optimized for TMT-labeled intact proteins in complex biological samples and a stepped higher energy collisional dissociation (HCD) fragmentation was suggested for a balance between accurate quantification and confident identification.[133]

Successful application of protein-level TMT labeling in complex samples allows identification and quantification of intact proteoforms with high multiplexing capacity, eliminates variability from downstream sample preparation (e.g., peptide digestion), and enables protein-based fractionation, such as multidimensional separation.[130,173] However, due to the high complexity and wide dynamic range of complex samples (*e.g.*, cell lysate), one dimensional (1D) separation by using LC or capillary electrophoresis (CE) may not provide sufficient identifications. Therefore, many multidimensional (MD) separation platforms by coupling different separation methods, such as reversed-phase chromatography (RPLC), ion exchange chromatography (IEX), hydrophilic interaction chromatography (HILIC), hydrophobic interaction chromatography (HIC), size exclusion chromatography (SEC), have been applied to improve the separation resolution and identification based on separation orthogonality.[20,75,76,219–221] RPLC has become one of the most popular chromatographic separation approaches in high-throughput top-down proteomics due to the high separation resolution and MS-compatible mobile phases used.[73] Offline MD separation is commonly utilized because of the simple operation and flexibility of generally being able to couple any separation techniques. However, offline 2D separation generally requires additional sample handling steps, which introduces more sample variation and degradation products as well as reduces the throughput and sample recovery. Moreover, starting protein material for offline 2D separation is relatively high, ranging from hundreds of micrograms to milligrams.[68,75,76,86] However, such high starting material is not practical for TMT-labeled samples due to sample limitation.

Therefore, the online MDLC system was commonly used to overcome these challenges. In online MDLC systems, only few micrograms of proteins are needed for the deep proteome profiling.[18,19,27,92] Our lab has recently developed a 2D high-pH/low-pH RPLC-MS/MS system that integrated an automatic microtrap column-based platform by coupling high-pH RPLC and low-pH RPLC separation, enabling the characterization of >1000 intact proteoforms from only 5 µg *E. coli* cell lysate in 10 online-collected fractions.[18]

Here we demonstrated the deep quantitative analysis of human cell lysate using top-down TMT labeling[130] and the automatic online 2D high-pH/low-pH RPLC separation[18]. Microgram-level TMT-labeled protein samples were subjected to the online 2D system for high-pH separation, followed by online microtrap column dilution, low-pH RPLC separation by an ultrahigh-pressure long capillary column, and detection by a mass spectrometer. Overall, we successfully quantified ~5600 proteoforms from 20 µg TMT-labeled intact proteins, demonstrating the deep proteome profiling and quantitation with low microgram sample amounts and top-down TMT labeling approach.

4.3. EXPERIMENTAL

4.3.1 Chemicals and materials.

LC/MS grade acetonitrile (ACN), isopropanol (IPA), water, analytical reagent (AR) grade ammonium formate (AF) and acetic acid (Hac) were purchased from Sigma-Aldrich (St. Louis, MO, USA) unless noted otherwise. Trifluoroacetic acid (TFA), Tris (2-carboxyethyl) phosphine hydrochloride (TCEP), 1M

Triethylammonium bicarbonate (TEAB), Pierce™ BCA Protein Assay Kit, molecular weight cutoff spin filters (10K and 100K MWCO), TMT 10-plex Isobaric Label Reagent set (90061) was obtained from Thermo Fisher Scientific (Waltham, MA, USA).

4.3.2 TMT-labeled *HeLa* proteins reparation.

HeLa cells were grown in Dulbecco's modified Eagle medium (DMEM), with 10% fetal bovine serum (FBS) and 2% penicillin-streptomycin under 5% CO_2 at 37 °C until 80-90% density was achieved. *HeLa* cells were then collected and lysed via a sonication-based approach, as previously described.[76] 100 kDa molecular weight cutoff (MWCO) filters were used to remove larger molecular weight proteoforms through centrifugation at 12,000 × g and 4 °C for 20 minutes. Then 3 kDa MWCO filters were used to concentrate the sample down to > 1 μg/μL in 100 mM TEAB buffer (pH 8.5). The filtered protein sample was aliquoted into 6 tubes and labeled with 6 different TMT reagents (126, 127N, 127C, 128C, 129N, 129C) individually at the optimal condition as discussed previously.[132,218] The labeled sample was centrifuged at 12,000 × g and 4 °C for 30 minutes before LC-MS/MS analysis to remove any precipitation. Then six TMT-labeled *HeLa* cell lysate samples were mixed with a mass ratio of 4:4:2:2:1:1 for online 2D RPLC-MS/MS analysis.

4.3.4 Online 2D high-pH/low-pH RPLC-MS system setup.

The online 2D-RPLC system was built as previously described.[18] TMT-labeled proteins were separated by first dimensional (1D) home-packed high-pH

C4 column (10 cm, 150 μm I.D. × 360 μm O.D., Waters BEH C4 3.5 μm particles, 300 Å pore size), and further eluted onto a trapping column (150 μm i.d., 5 cm length, Jupiter particles, 5 μm diameter, 300 Å pore size), then followed by the separation of a second dimensional (2D) low-pH C4 RPLC capillary column (CoAnn Technologies, 100 μm i.d., 60 cm length, 3.4 μm diameter, 300 Å pore size). 20 mM ammonium formate in water (pH = 10) and 20 mM ammonium formate in acetonitrile (pH = 10) were prepared as mobile phase A and mobile phase B, respectively, for 1D high-pH RPLC separation.[18,209] Mobile phase A (0.01% TFA, 0.585% acetic acid, 2.5% isopropanol, and 5% acetonitrile in water) and mobile phase B (0.01% TFA, 0.585% acetic acid, 45% isopropanol, and 45% acetonitrile in water) were used for second dimensional low-pH RPCL separation. Detailed 2D LC gradient schedule is shown in ***Table 4-1***.

Table 4-1 The online high-pH/low-pH RPLC gradient schedule. The (*) marked B% can be changed for different online fractions. The gradient time used in the paper was 100 minutes.

Time	1D high-pH RPLC		2D low-pH RPLC	
	B%	Flowrate (μL/min)	B%	Flowrate (μL/min)
0	0	5	0	25
5	30*	5	0	25
55	30*	5	0	25
56	0	5	0	25
60	0	5	10	25
160	0	5	70	95
161	0	5	90	95
171	0	5	90	95
172	0	5	0	95
212	0	5	0	95
213	0	5	0	25
215	0	5	0	25

4.3.5 Top-down MS analysis.

Proteins separated by 2D RPLC were further analyzed by an Orbitrap Exploris 240 mass spectrometer with positive mode (Thermo Fisher Scientific, Bremen, Germany) through a customized nano-ESI interface.[14] MS parameters were listed below: The temperature of the inlet capillary was set to 275 °C and the spray voltage was 3.0 kV. Full MS was acquired from 550-1500 m/z with mass resolution of 120,000 at m/z=200 with two microscans. The resolution of MS/MS was 60,000 at m/z=200 with 100-1800 m/z and 4 microscans. The AGC target was set to 1E6 for both MS and MS/MS. Maximum injection time was 500 ms for MS and 200 ms for MS/MS. Top 6 most abundant precursor ion peaks (charge 4-50) were selected for MS/MS fragmentation. The isolation window was set to 2 m/z with 90 s as the dynamic exclusion. Higher-energy C-trap dissociation (HCD) with a stepped normalized collision energy of 35/45/50% was applied for fragmentation.

4.3.6 Data analysis.

Deconvolution for all datasets were performed using Biopharma Finder (Thermo Fisher Scientific). Datasets were searched against the annotated Human protein database (UniProt 2020-07-11, 20368 species) using TopPIC Suite (Version 1.4.10) for identification.[182,183] The TopPIC parameters were set as shown below: decoy database searching was used, and the FDR cutoff was set as 0.01 for the spectrum level and 0.05 for the proteoform level. The maximum number of mass shifts was 2, and the mass shift range was ± 500 Daltons. TMT labeling at the N-terminal and lysine residues was set as the fixed PTMs; acetylation, carboxymethyl,

phosphorylation, oxidation and methylation were set as the variable PTMs.[132] All other parameters were set as default values. ProSight Lite[185] was used for manual interpretation and spectrum presentation. An in-house Python package was utilized for reporter ion peaks extraction.[133]

4.4 RESULTS AND DISCUSSION

The performance of the online 2D system was evaluated using denatured *HeLa* cell lysate samples. Briefly, after reduction and alkylation (before TMT labeling), a 10 µg of *HeLa* cell lysate were injected into a BEH C4 capillary column for 1st dimension high-pH RPLC separation (pH = 10), then separated by a C2 capillary column for 2nd dimension low-pH separation (pH = 2), followed by the MS analysis. Two 2-fraction runs (the first fraction is flow through with 0% mobile phase B from 1D high-pH RPLC; the second fraction is the proteins eluted using 97% mobile phase B from 1D high-pH RPLC) were collected. As shown in ***Figure 4-1***, the majority of proteins eluted out in the 97% fraction, suggesting efficient binding and elution from the 1st dimension high-pH columns. In addition, the two 2-fraction runs showed similar elution profiles, suggesting the reproducibility of the online 2D RPLC system.

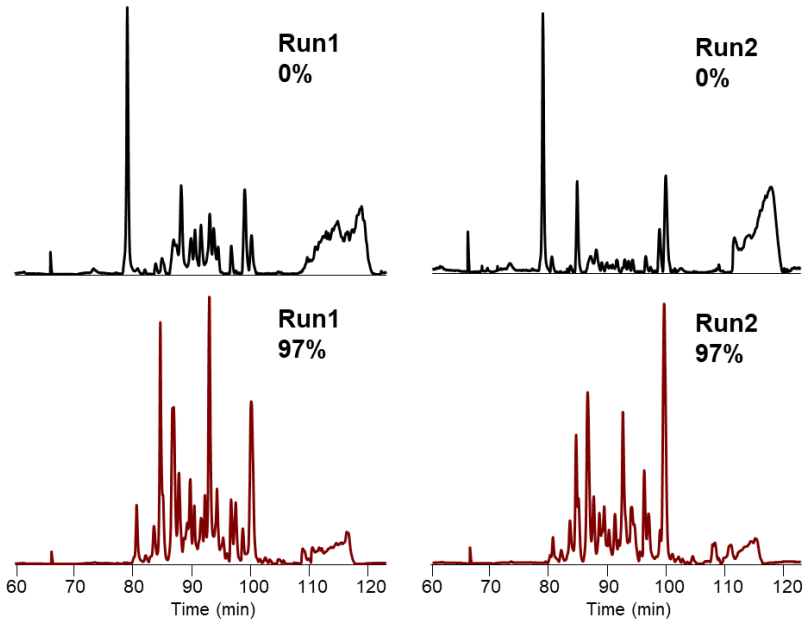


Figure 4-1 Two 2-fraction runs testing the reproducibility of online 2D system by using 10 µg of reduced *HeLa* cell lysate.

Next, we applied the 2-fraction 2D separation (0% and 97%) for the analysis of TMT-labeled *Hela* protein samples. 5 µg of the TMT-labeled *HeLa* proteins with a mass ratio of 1:1:2:2:4:4 was loaded onto the online high-pH/low-pH RPLC-MS/MS system. Similarly, most TMT-labeled proteoforms were detected in the 97% fraction, as shown in ***Figure 4-2A***. However, we noticed significant pattern differences between the labeled and unlabeled samples, suggesting that the first dimension fraction collection steps should be further optimized compared to the unlabeled intact protein samples. Firstly, a 2D separation with 4 fractions was applied (***Figure 4-2B*****),** including 0%, 30%, 60%, and 97% fractions. Both 60% and 97% fractions contain more proteoforms compared to the 0% fraction and 30% fraction, which can be further split to improve proteome coverage. A 7-fraction 2D separation was then evaluated (0%, 30%, 45%, 60%, 70%, 80%, 97%) as shown in ***Figure 4-2C***. 251 unique proteoforms were identified in the 2-fraction runs, 328

unique proteoforms were identified in the 4-fraction runs, and 412 unique proteoforms were identified in the 7-fraction runs, suggesting a significant increase in proteoform detection and quantification with increased 1st dimensional fraction numbers.

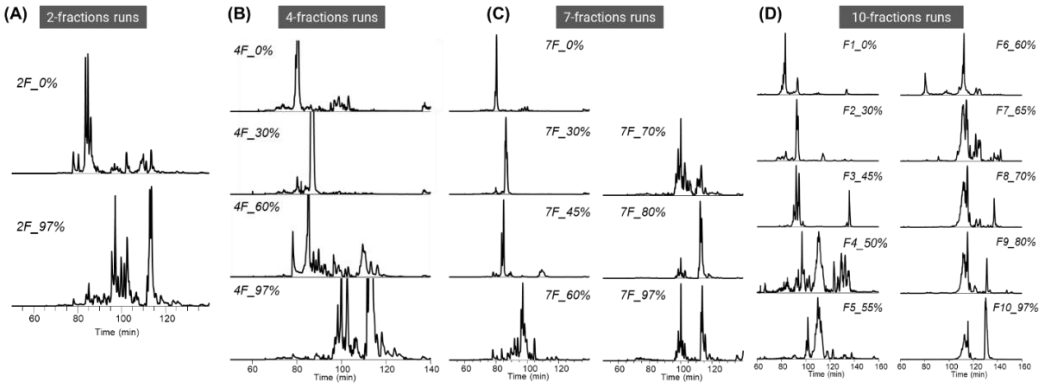


Figure 4-2 Total ion chromograms (TICs) for optimizing elution steps of 1D high-pH RPLC separation. (A) 2-fraction runs; (B) 4-fraction runs; (C) 7-fraction runs; and (D) 10-fraction runs. A 60-min 2nd dimension RPLC gradient was applied with the injection of 5 µg of TMT-labeled *HeLa* proteins for 2-fraction, 4-fraction, and 7-fraction runs. A100-min 2nd dimension RPLC gradient was applied with the injection of 20 µg of TMT-labeled *HeLa* proteins for the 10-fraction runs.

Based on the optimization results, we applied a 10-fraction scheme (***Figure 4-2D***) for the online analysis of 20 µg of the TMT-labeled *HeLa* proteins with a mass ratio of 1:1:2:2:4:4. To further improve the proteoform identification and quantification number, we also extended the 2nd dimension separation time from 60 minutes to 100 minutes. In total, 5628 unique proteoforms were detected in a single 10-fraction 2D analysis. Among them, 1022 unique proteoforms from 231 unique proteins were identified, similar to our previously reported 2D analysis of unlabeled

samples (***Figure 4-4A***). Total ion chromatograms (TICs) of the 10 online collected fractions suggested different elution profiles in different fractions, ***Figure 4-2D***.

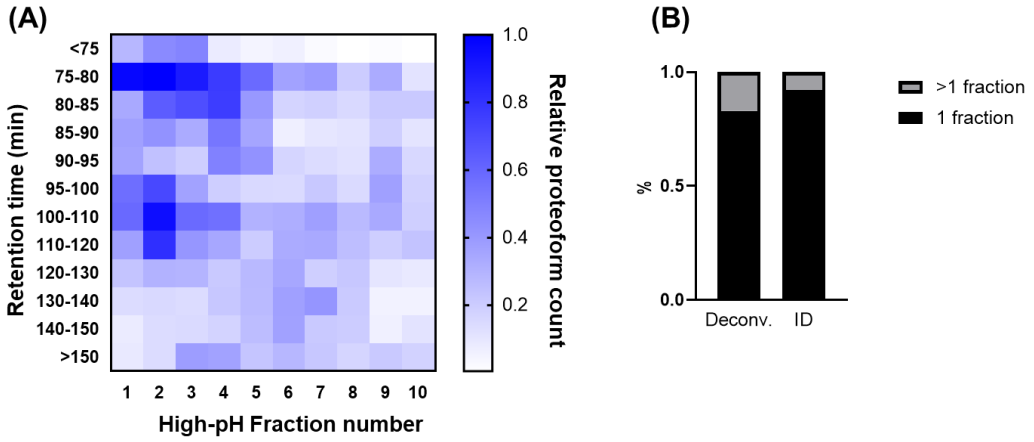


Figure 4-3 Orthogonality evaluation of the online collected 10 fractions for TMT-labeled HeLa proteins. (A) Heatmap demonstrating proteoform elution patterns for different fractions. Each column bin in the heatmap represents the relative number of uniquely identified proteoforms in the elution window in the low pH RPLC separation. Relative proteoform count was normalized to the maximum proteoform count in all bins. **(B)** Percentage of mass features or proteoforms were detected or identified in different fractions. 1 fraction means the mass feature/proteoforms were only eluted in 1 fraction among all 10 online-collected fractions. >1 fraction means the mass feature/proteoforms were eluted in multiple fractions among all 10 online-collected fractions.

In addition, to evaluate the separation orthogonality between high-pH and low-pH RPLC separation, the relative number of uniquely detected proteoforms in each elution window from low-pH RPLC were plotted against the high-pH RPLC fraction number in a heatmap as shown in ***Figure 4-3A***. The results demonstrated that TMT-labeled intact human proteoforms were separated orthogonally by using the online 2D system, which is compatible with the results from our previously reported online 2D system for *E. coli* cell lysate.[18] Also, 83% of the uniquely detected mass features were eluted in 1 fraction and 17% of them were eluted in

multiple fractions; In comparison, 94% of the uniquely identified proteoforms were eluted in 1 fraction and 6% of them were eluted in multiple fractions, which further confirmed the orthogonality between high-pH RPLC and low-pH RPLC for TMT-labeled intact proteoforms ***(Figure 4-3B)***.

The quantification accuracy of TMT-based intact proteoform in complex samples was also evaluated. Reporter ion peak intensities for all scans among all 10 online fractions were extracted using the in-house Python package.[133] Only scans with all six reporter ions present were considered. The intensity of reporter ions was distributed from ~ 1.2E2 to ~ 2.5E7, suggesting a good dynamic range of relative quantification for intact proteoforms in complex samples with top-down proteomics. The reporter ratio was normalized against the intensity from 127N reporter channel (*i.e.,* 129C/127N, 129N/127N, 128C/127N, 127C/127N, 126/127N) as shown in **Error! Reference source not found.** that demonstrated the m easured reporter ion ratios on all the scans extracted from raw file (**All scans**) as well as the scans that matched to proteoform spectrum matches (**PrSMs)**. All measured reporter ion ratios have the maximum close to the expected values, suggesting relatively high quantification accuracy. The observed average reporter ion ratios were 4.24±2.74, 3.79±2.11, 2.32±1.08, 2.32±1.73, and 1.31±1.14 for all scans. For all identified PrSMs, the average reporter ion ratios were 4.34±2.48, 3.95±2.14, 2.42±1.42, 2.36±2.31, 1.22±1.41, which suggested similar quantification results to all scans.

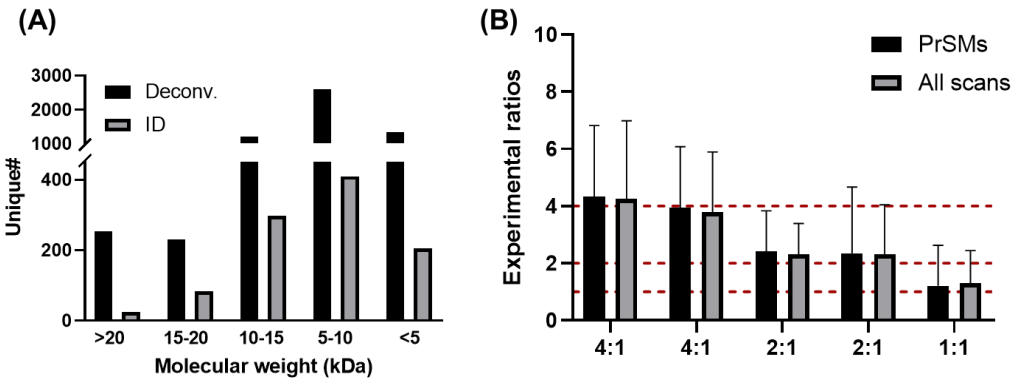


Figure 4-4 Evaluation of online 2DLC-MS/MS system with 10 fractions. (A) Detected and identified unique proteoforms from 20 µg TMT-labeled *HeLa* proteins with 10 online collected fractions. (B) Distribution of measured ratios for quantification accuracy evaluation with all reporter ions extracted from raw file (All scans) and the reporter ions after matching to PrSMs with the corresponding quantifications scans (PrSMs). Dashed red lines represent the theoretical values.

Figure 4-5A showed the total ion chromatogram (TIC) of low-pH RPLC separation for fraction 4 with the highlighted peak for Jupiter microtubule associated homolog 1 (JPT1, Uniprot# Q9UK76) as an example. Parent ion peak of 1050.37 m/z with charge state as 17 was chosen for the HCD fragmentation to generate the MS2 spectra as shown in ***Figure 4-5B***. The observed monoisotopic mass for JPT1 proteoform was 17829.21 Da with mass measurement accuracy (MMA) of 3.05 ppm to its theoretical monoisotopic mass. For this proteoform, initial methionine was removed and the second amino acid, threonine (T_2) was modified by acetylation. All 8 lysine residues were labeled with TMT tags, and unlocated phosphorylation was identified at serine residue at S_{87} or S_{88}. The overall sequence coverage was 23%, with a p-score of 3.5E-14 from ProSight Lite[185]. For quantification, there are 4 PrSMs identified in total for this proteoform and the average observed reporter ion ratios among all 4 PrSMs were 3.66±0.69, 3.15±0.73,

1.94±0.34, 1.98±0.46, 1.02±0.32, which was close to the expected ratio as 4:4:2:2:1 (129C/127N, 129N/127N, 128C/127N, 127C/127N, 126/127N).

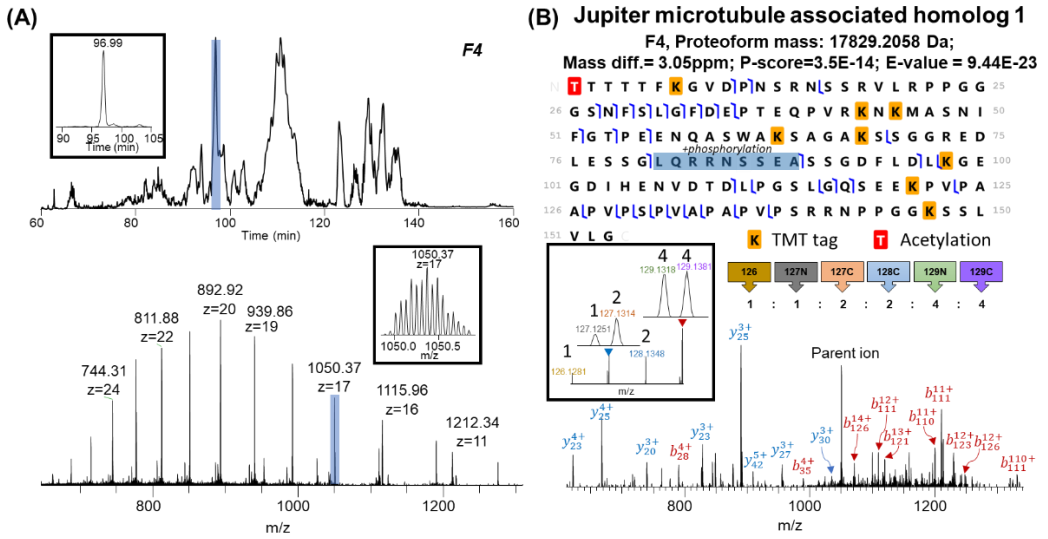


Figure 4-5 Quantification evaluation with Jupiter microtubule associated homolog 1 as the example. (A) Total ion chromatogram (TIC) of low-pH RPLC separation for fraction 4 (50%) and the MS1 spectra for (JPT1, Uniprot# Q9UK76). (B) Representative MS/MS spectra and sequence of JUP1 proteoform identified in fraction 4 with its corresponding reporter ions for quantification. Theoretical reporter ion ratios should be 1:1:2:2:4:4.

4.5 CONCLUSION

In summary, we have demonstrated the deep quantitative proteome profiling of human cell lysate by coupling an automated online high-pH/low-pH RPLC-MS/MS platform and intact protein-level TMT labeling that is highly promising for the deep quantitative analysis of intact proteoforms in complex biological samples using top-down proteomics. ~5600 proteoforms were detected and quantified from only 20 µg TMT-labeled intact proteoforms. The observed TMT reporter ion ratios also strongly resemble the theoretical ones, which demonstrated good and accurate quantification. With significantly improved proteome coverage and accurate quantification, this platform coupling protein-level

TMT labeling and the automated online high-pH/low-pH RPLC-MS/MS holds great potential for deep proteome profiling and quantitative analysis in real biological samples.

Authors for Chapter 4:

Yanting Guo, Trishika Chowdhury, Walter P. Galie, Kellye A. Cupp-Sutton, Zhitao Zhao, Dahang Yu, and Si Wu. “Deep intact proteoform quantification of microgram-scale proteome by tandem mass tag labeling and online 2D high-pH/low-pH-RPLC top-down analysis." (***Manuscript in preparation***)

Author contributions: Yanting Guo: Methodology, Data curation, Formal analysis, validation, Writing - original draft. Trishika Chowdhury: Sample preparation. Walter P. Galie: Sample preparation. Zhitao Zhao: Data analysis. Kellye A. Cupp-Sutton: Validation, Writing - review & editing. Dahang Yu: Methodology. Si Wu: Conceptualization, Validation, Supervision, Funding acquisition, Writing – review & editing.

Copyright information for Chapter 4

This chapter is unpublished research from the Wu lab at the University of Oklahoma Department of Chemistry and Biochemistry. All the figures were made by Yanting Guo.

Chapter 5 Overall summary and future perspectives

5.1 OVERALL SUMMARY

Overall, the work presented here demonstrated a series of optimizations for analytical methods including tandem mass tag (TMT) labeling and multidimensional separation for deep quantitative analysis of intact proteoforms using top-down proteomics, which holds great potential for the identification and quantitation of intact proteoforms, understating of proteoforms functions, characterization of the effects of intact proteoforms in disease state and progression, and help with biomarker discovery.

Isobaric tag labeling-based quantitation has been commonly utilized in bottom-up proteomics for quantitative analysis in complex biological sample digest.[157] Although a few publications have demonstrated the application of isobaric tag labeling in pure intact proteins or simple protein mixtures168–171, it has been limited to the application of complex biological sample for high throughput quantitative analysis using top-down proteomics. One of the major challenges is the protein precipitation during the labeling reaction because of the use of organic solvent, which largely affected the quantification of intact proteoforms in complex biological sample. To tackle such challenge, we previously demonstrated the quantification of smaller proteoforms (<35 kDa) in complex biological using intact protein-level TMT labeling and top-down proteomics by removing large proteoforms with molecular weight cutoff (MWCO) spin filters and size exclusion chromatography.[130] However, another challenge that limits the

application of protein-level TMT labeling in complex biological samples is the production of side products including both underlabeled and overlabeled products, which significantly increases sample complexity and decrease the signal-to-noise ratio.

To limit the side products generated during protein-level TMT labeling process and maximize the completely labeled products, we have performed a systematic optimization of intact protein-level TMT labeling conditions including TMT-to-protein mass ratio, protein initial concentration, quenching buffer pH, reaction buffer, reaction time (***Chapter 2***). We concluded that a complex protein solution with initial concentration >1.0 μg/μL labeled with TMT as 4:1 or higher TMT-to-protein mass ratio, then quenched by hydroxylamine to a final concentration as > 0.3% (final pH > 9.1) resulted in optimal labeling efficiency and minimized production of side products. A double labeling approach was also innovated to overcome the low concentration of limited samples such as clinical samples. Moreover, a benchmarking protocol for intact protein-level TMT labeling in complex biological sample was also summarized in ***Appendix A – A Benchmarking Protocol for Intact Protein-Level Tandem Mass Tag (TMT) Labeling for Quantitative Top-Down Proteomics***.

Moreover, protein quantification using isobaric chemical tag-based labeling relies on the efficient and accurate measurement of low-mass reporter ions generated in MS2 fragmentation; therefore, higher-energy collisional dissociation (HCD) fragmentation energy for TMT-labeled intact proteoforms in complex biological sample was further optimized to better generate reporter ions and protein

fragment ions simultaneously (***Chapter 3***). Both quantification accuracy and identification confidence were evaluated among different HCD energies. Our results indicated that higher normalized collision energies (NCEs) such as NCE = 50% and 80% resulted in higher reporter ion intensities for accurate quantitation. However, only a few fragments were generated at high NCEs for identification. We further demonstrated that a stepped normalized collision (SNCE) with energies from 30% to 50% resulted in optimal identification and quantification of TMT-labeled intact proteoforms.

Another challenge that occurs in high throughput top-down proteomics is the high sample complexity and wide dynamic range. To improve the proteome coverage, coupling 2 or more orthogonal separation approaches has been applied to further separate the coeluted species.[8,196] Here, intact protein-level TMT labeling and online high-pH/low-pH RPLC-MS/MS platform was coupled to perform the deep quantitative analysis for complex proteome using top-down proteomics (***Chapter 4***). The automated, online, 2D nano-high-pH/low-pH RPLC platform enables the characterization of intact proteoforms with small quantity sample consumption and our results indicated that the platform successfully characterized and quantified over 5600 intact mass features from only 20 μg TMT-labeled intact proteoforms using top-down proteomics. Overall, the online 2D-TMT platform holds great potential for deep proteome quantitative analysis of mass-limited samples, which allows for the characterization and quantification of thousands of intact proteoforms from microgram-level complex biological samples.

In summary, the optimization of intact protein-level TMT labeling reactions simplifies the sample complexity by decreasing the production of side products which lays the foundation for the quantitative analysis of intact proteomics in complex biological sample using isobaric tag labeling in top-down proteomics. The optimization of MS2 fragmentation energies for TMT-labeled intact proteoforms in complex biological samples further allows for a balance between accurate quantification and confident identification. Moreover, the successful coupling between online 2D RPLC platform with intact protein-level TMT labeling allows for deep proteome quantitative analysis using top-down proteomics, which paves the way for future applications of quantitative analysis in real biological samples such as cancer cells or human serum proteins.

5.2 FUTURE PERSPECTIVES

Characterization of intact proteoforms has become increasingly important since many proteoforms with different PTMs have been identified and linked to the progression of human diseases.[5] Therefore, simultaneous identification and quantification of intact proteoforms in complex biological samples using quantitative top-down proteomics (TDP) are essential to characterize proteoform functions and understand the effect of PTMs on disease state and progression.

Label-free quantitation approach is the most popular strategy used in quantitative TDP, and it has been applied to many complex biological samples for quantitative analysis and potential biomarker discovery. However, label-free quantitation naturally suffers from run-to-run variability that decreases the label-free quantitation accuracy and low multiplexing that needs a lot of instrument time.

The quantitation accuracy can be further decreased because of the increasing number of LC-MS runs. Moreover, label-free quantitation is not suitable for quantitative analysis in multidimensional separation techniques because the proteoforms may be eluted in multiple fractions. Therefore, the isobaric chemical tag labeling-based quantitative approach in TDP has been developed and further extended to the applications of complex biological samples. This dissertation has demonstrated a series of systematic optimizations of protein-level labeling in TDP to limit the formation of side products, increase quantitation accuracy, decrease the separation difficulty, and improve proteome coverage. With the optimized intact protein-level TMT labeling, we believe that quantitative TDP holds great potential for deep proteome profiling to address biological questions in real complex biological samples. Additionally, the multiplexing ability of intact protein-level TMT labeling holds the potential for the top-down proteomics investigation of cellular heterogeneity in mass-limited samples or at the single-cell level.[54] The intact protein-level TMT labeling for top-down proteomics may also be applied to study protein functions and protein interactions with functional top-down proteomics such as stability proteomics.[15,222]

Further improvements are still required in quantitative TDP using isobaric chemical tag labeling. Advancing instrumentation on high-resolution MS instruments and separation techniques will benefit the TMT-based quantitative top-down proteomics analysis of real biological samples. Developing new isobaric chemical tags that are more soluble in aqueous solution will solve the protein precipitation, simplify the sample handling process, and decrease the potential bias

from sample preparation. Moreover, current data analysis still relies on manual data interpretation; therefore, advanced bioinformatic tools that can provide robust and accurate data interpretation are urgently needed.

References:

(1) Widłak, W. Protein Structure and Function. In *Molecular Biology: Not Only for Bioinformaticians*; Widłak, W., Ed.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2013; pp 15–29. https://doi.org/10.1007/978-3-642-45361-8_2.

(2) Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Ogorzalek Loo, R. R.; Lundberg, E.; MacCoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schlüter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlén, M.; Van Eyk, J. E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.; Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B. How Many Human Proteoforms Are There? *Nat Chem Biol* **2018**, *14* (3), 206–214. https://doi.org/10.1038/nchembio.2576.

(3) Smith, L. M.; Kelleher, N. L. Proteoform: A Single Term Describing Protein Complexity. *Nat Methods* **2013**, *10* (3), 186–187. https://doi.org/10.1038/nmeth.2369.

(4) Wang, E. T.; Sandberg, R.; Luo, S.; Khrebtukova, I.; Zhang, L.; Mayr, C.; Kingsmore, S. F.; Schroth, G. P.; Burge, C. B. Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature* **2008**, *456* (7221), 470–476. https://doi.org/10.1038/nature07509.

(5) Smith, L.; Agar, J.; Chamot-Rooke, J.; Danis, P.; Ge, Y.; Loo, J.; Pasa-Tolic, L.; Tsybin, Y.; Kelleher, N. The Human Proteoform Project: A Plan to Define the Human Proteome. **2020**. https://doi.org/10.20944/preprints202010.0368.v1.

(6) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. NCBI Reference Sequences (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins. *Nucleic Acids Res* **2007**, *35* (Database issue), D61–D65. https://doi.org/10.1093/nar/gkl842.

(7) Tyers, M.; Mann, M. From Genomics to Proteomics. *Nature* **2003**, *422* (6928), 193–197. https://doi.org/10.1038/nature01510.

(8) Zhang, Z.; Wu, S.; Stenoien, D. L.; Paša-Tolić, L. High-Throughput Proteomics. *Annual Review of Analytical Chemistry* **2014**, *7* (1), 427–454. https://doi.org/10.1146/annurev-anchem-071213-020216.

(9) Yuan, H.; Jiang, B.; Zhao, B.; Zhang, L.; Zhang, Y. Recent Advances in Multidimensional Separation for Proteome Analysis. *Anal. Chem.* **2019**, *91* (1), 264–276. https://doi.org/10.1021/acs.analchem.8b04894.

(10) Yates, J. R.; Ruse, C. I.; Nakorchevsky, A. Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. *Annual Review of Biomedical Engineering* **2009**, *11* (1), 49–79. https://doi.org/10.1146/annurev-bioeng-061008-124934.

(11) Melby, J. A.; Roberts, D. S.; Larson, E. J.; Brown, K. A.; Bayne, E. F.; Jin, S.; Ge, Y. Novel Strategies to Address the Challenges in Top-Down Proteomics. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (6), 1278–1294. https://doi.org/10.1021/jasms.1c00099.
(12) Gygi, S. P.; Corthals, G. L.; Zhang, Y.; Rochon, Y.; Aebersold, R. Evaluation of Two-Dimensional Gel Electrophoresis-Based Proteome Analysis Technology. *Proceedings of the National Academy of Sciences* **2000**, *97* (17), 9390–9395. https://doi.org/10.1073/pnas.160270797.
(13) Monteoliva, L.; Albar, J. P. Differential Proteomics: An Overview of Gel and Non-Gel Based Approaches. *Briefings in Functional Genomics* **2004**, *3* (3), 220–239. https://doi.org/10.1093/bfgp/3.3.220.
(14) Zhang, Y.; Fonslow, B. R.; Shan, B.; Baek, M.-C.; Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem Rev* **2013**, *113* (4), 2343–2394. https://doi.org/10.1021/cr3003533.
(15) A. Cupp-Sutton, K.; Wu, S. High-Throughput Quantitative Top-down Proteomics. *Molecular Omics* **2020**, *16* (2), 91–99. https://doi.org/10.1039/C9MO00154A.
(16) Smith, L. M.; Kelleher, N. L. Proteoforms as the next Proteomics Currency. *Science* **2018**, *359* (6380), 1106–1107. https://doi.org/10.1126/science.aat1884.
(17) Kelleher, N. L. Top-down Proteomics. *Industrial And Engineering Chemistry Analytical Edition* **2004**, *76* (11).
(18) Wang, Z.; Yu, D.; Cupp-Sutton, K. A.; Liu, X.; Smith, K.; Wu, S. Development of an Online 2D Ultrahigh-Pressure Nano-LC System for High-PH and Low-PH Reversed Phase Separation in Top-Down Proteomics. *Anal. Chem.* **2020**, *92* (19), 12774–12777. https://doi.org/10.1021/acs.analchem.0c03395.
(19) Gargano, A. F. G.; Shaw, J. B.; Zhou, M.; Wilkins, C. S.; Fillmore, T. L.; Moore, R. J.; Somsen, G. W.; Paša-Tolić, L. Increasing the Separation Capacity of Intact Histone Proteoforms Chromatography Coupling Online Weak Cation Exchange-HILIC to Reversed Phase LC UVPD-HRMS. *J. Proteome Res.* **2018**, *17* (11), 3791–3800. https://doi.org/10.1021/acs.jproteome.8b00458.
(20) Shen, Y.; Tolić, N.; Piehowski, P. D.; Shukla, A. K.; Kim, S.; Zhao, R.; Qu, Y.; Robinson, E.; Smith, R. D.; Paša-Tolić, L. High-Resolution Ultrahigh-Pressure Long Column Reversed-Phase Liquid Chromatography for Top-down Proteomics. *Journal of Chromatography A* **2017**, *1498*, 99–110. https://doi.org/10.1016/j.chroma.2017.01.008.
(21) McCool, E. N.; Xu, T.; Chen, W.; Beller, N. C.; Nolan, S. M.; Hummon, A. B.; Liu, X.; Sun, L. Deep Top-down Proteomics Revealed Significant Proteoform-Level Differences between Metastatic and Nonmetastatic Colorectal Cancer Cells. *Science Advances* **2022**, *8* (51), eabq6348. https://doi.org/10.1126/sciadv.abq6348.
(22) McCool, E. N.; Lubeckyj, R. A.; Shen, X.; Chen, D.; Kou, Q.; Liu, X.; Sun, L. Deep Top-Down Proteomics Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry: Identification of 5700 Proteoforms from the

Escherichia Coli Proteome. *Anal. Chem.* **2018**, *90* (9), 5529–5533. https://doi.org/10.1021/acs.analchem.8b00693.
(23) Shen, X.; Yang, Z.; McCool, E. N.; Lubeckyj, R. A.; Chen, D.; Sun, L. Capillary Zone Electrophoresis-Mass Spectrometry for Top-down Proteomics. *Trends Analyt Chem* **2019**, *120*, 115644. https://doi.org/10.1016/j.trac.2019.115644.
(24) Cai, W.; Tucholski, T.; Chen, B.; Alpert, A. J.; 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**, *89* (10), 5467–5475. https://doi.org/10.1021/acs.analchem.7b00380.
(25) François, I.; Sandra, K.; Sandra, P. Comprehensive Liquid Chromatography: Fundamental Aspects and Practical Considerations—A Review. *Analytica Chimica Acta* **2009**, *641* (1), 14–31. https://doi.org/10.1016/j.aca.2009.03.041.
(26) Pirok, B. W. J.; Stoll, D. R.; Schoenmakers, P. J. Recent Developments in Two-Dimensional Liquid Chromatography: Fundamental Improvements for Practical Applications. *Anal. Chem.* **2019**, *91* (1), 240–263. https://doi.org/10.1021/acs.analchem.8b04841.
(27) Tian, Z.; Zhao, R.; Tolić, N.; Moore, R. J.; Stenoien, D. L.; Robinson, E. W.; Smith, R. D.; Paša-Tolić, L. Two-Dimensional Liquid Chromatography System for Online Top-down Mass Spectrometry. *Proteomics* **2010**, *10* (20), 3610–3620. https://doi.org/10.1002/pmic.201000367.
(28) Stolz, A.; Jooß, K.; Höcker, O.; Römer, J.; Schlecht, J.; Neusüß, C. Recent Advances in Capillary Electrophoresis-Mass Spectrometry: Instrumentation, Methodology and Applications. *ELECTROPHORESIS* **2019**, *40* (1), 79–112. https://doi.org/10.1002/elps.201800331.
(29) Gomes, F. P.; Yates, J. R. Recent Trends of Capillary Electrophoresis-Mass Spectrometry in Proteomics Research. *Mass Spectrom Rev* **2019**, *38* (6), 445–460. https://doi.org/10.1002/mas.21599.
(30) Kohl, F. J.; Sánchez-Hernández, L.; Neusüß, C. Capillary Electrophoresis in Two-Dimensional Separation Systems: Techniques and Applications. *ELECTROPHORESIS* **2015**, *36* (1), 144–158. https://doi.org/10.1002/elps.201400368.
(31) Kler, P. A.; Sydes, D.; Huhn, C. Column–Coupling Strategies for Multidimensional Electrophoretic Separation Techniques. *Anal Bioanal Chem* **2015**, *407* (1), 119–138. https://doi.org/10.1007/s00216-014-8099-7.
(32) Siuti, N.; Kelleher, N. L. Decoding Protein Modifications Using Top-down Mass Spectrometry. *Nat Methods* **2007**, *4* (10), 817–821. https://doi.org/10.1038/nmeth1097.
(33) Toby, T. K.; Fornelli, L.; Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu Rev Anal Chem (Palo Alto Calif)* **2016**, *9* (1), 499–519. https://doi.org/10.1146/annurev-anchem-071015-041550.
(34) Brown, K. A.; Melby, J. A.; Roberts, D. S.; Ge, Y. Top-down Proteomics: Challenges, Innovations, and Applications in Basic and Clinical Research. *Expert Review of Proteomics* **2020**, *17* (10), 719–733. https://doi.org/10.1080/14789450.2020.1855982.

(35) Schaffer, L. V.; Millikin, R. J.; Miller, R. M.; Anderson, L. C.; Fellers, R. T.; Ge, Y.; Kelleher, N. L.; LeDuc, R. D.; Liu, X.; Payne, S. H.; Sun, L.; Thomas, P. M.; Tucholski, T.; Wang, Z.; Wu, S.; Wu, Z.; Yu, D.; Shortreed, M. R.; Smith, L. M. Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **2019**, *19* (10), e1800361. https://doi.org/10.1002/pmic.201800361.
(36) Cui, W.; W. Rohrs, H.; L. Gross, M. Top-down Mass Spectrometry : Recent Developments, Applications and Perspectives. *Analyst* **2011**, *136* (19), 3854–3864. https://doi.org/10.1039/C1AN15286F.
(37) Armirotti, A.; Damonte, G. Achievements and Perspectives of Top-down Proteomics. *PROTEOMICS* **2010**, *10* (20), 3566–3576. https://doi.org/10.1002/pmic.201000245.
(38) Gregorich, Z. R.; Ge, Y. Top-down Proteomics in Health and Disease: Challenges and Opportunities. *PROTEOMICS* **2014**, *14* (10), 1195–1210. https://doi.org/10.1002/pmic.201300432.
(39) Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top Down Proteomics: Facts and Perspectives. *Biochemical and Biophysical Research Communications* **2014**, *445* (4), 683–693. https://doi.org/10.1016/j.bbrc.2014.02.041.
(40) Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y. Top-Down Proteomics: Ready for Prime Time? *Anal. Chem.* **2018**, *90* (1), 110–127. https://doi.org/10.1021/acs.analchem.7b04747.
(41) Kelleher, N. L. Peer Reviewed: Top-down Proteomics. *Analytical chemistry* **2004**, *76* (11), 196 A-203 A.
(42) Cai, W.; Tucholski, T. M.; Gregorich, Z. R.; Ge, Y. Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. *Expert Review of Proteomics* **2016**, *13* (8), 717–730. https://doi.org/10.1080/14789450.2016.1209414.
(43) Kelleher, N. L.; Thomas, P. M.; Ntai, I.; Compton, P. D.; LeDuc, R. D. Deep and Quantitative Top-down Proteomics in Clinical and Translational Research. *Expert Review of Proteomics* **2014**, *11* (6), 649–651. https://doi.org/10.1586/14789450.2014.976559.
(44) Savaryn, J. P.; Catherman, A. D.; Thomas, P. M.; Abecassis, M. M.; Kelleher, N. L. The Emergence of Top-down Proteomics in Clinical Research. *Genome Medicine* **2013**, *5* (6), 53. https://doi.org/10.1186/gm457.
(45) Reid, G. E.; McLuckey, S. A. 'Top down' Protein Characterization via Tandem Mass Spectrometry. *Journal of Mass Spectrometry* **2002**, *37* (7), 663–675. https://doi.org/10.1002/jms.346.
(46) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. Mapping Intact Protein Isoforms in Discovery Mode Using Top-down Proteomics. *Nature* **2011**, *480* (7376), 254–258. https://doi.org/10.1038/nature10575.

(47) Queiroz, J. A.; Tomaz, C. T.; Cabral, J. M. S. Hydrophobic Interaction Chromatography of Proteins. *Journal of Biotechnology* **2001**, *87* (2), 143–159. https://doi.org/10.1016/S0168-1656(01)00237-1.
(48) Chen, B.; Peng, Y.; Valeja, S. G.; Xiu, L.; Alpert, A. J.; Ge, Y. Online Hydrophobic Interaction Chromatography–Mass Spectrometry for Top-Down Proteomics. *Anal. Chem.* **2016**, *88* (3), 1885–1891. https://doi.org/10.1021/acs.analchem.5b04285.
(49) Hemström, P.; Irgum, K. Hydrophilic Interaction Chromatography. *Journal of Separation Science* **2006**, *29* (12), 1784–1821. https://doi.org/10.1002/jssc.200600199.
(50) Gama, M. R.; da Costa Silva, R. G.; Collins, C. H.; Bottoli, C. B. G. Hydrophilic Interaction Chromatography. *TrAC Trends in Analytical Chemistry* **2012**, *37*, 48–60. https://doi.org/10.1016/j.trac.2012.03.009.
(51) Acikara, Ö. B.; Acikara, Ö. B. *Ion-Exchange Chromatography and Its Applications*; IntechOpen, 2013. https://doi.org/10.5772/55744.
(52) Chen, X.; Ge, Y. Ultrahigh Pressure Fast Size Exclusion Chromatography for Top-down Proteomics. *PROTEOMICS* **2013**, *13* (17), 2563–2566. https://doi.org/10.1002/pmic.201200594.
(53) Barth, H. G.; Boyes, B. E.; Jackson, C. Size Exclusion Chromatography. *Anal. Chem.* **1996**, *68* (12), 445–466. https://doi.org/10.1021/a19600193.
(54) Cupp-Sutton, K. A.; Fang, M.; Wu, S. Separation Methods in Single-Cell Proteomics: RPLC or CE? *Int J Mass Spectrom* **2022**, *481*, 116920. https://doi.org/10.1016/j.ijms.2022.116920.
(55) Han, X.; Wang, Y.; Aslanian, A.; Bern, M.; Lavallée-Adam, M.; Yates, J. R. I. Sheathless Capillary Electrophoresis-Tandem Mass Spectrometry for Top-Down Characterization of Pyrococcus Furiosus Proteins on a Proteome Scale. *Anal. Chem.* **2014**, *86* (22), 11006–11012. https://doi.org/10.1021/ac503439n.
(56) Huang, L.; Fang, M.; Cupp-Sutton, K. A.; Wang, Z.; Smith, K.; Wu, S. Spray-Capillary-Based Capillary Electrophoresis Mass Spectrometry for Metabolite Analysis in Single Cells. *Anal. Chem.* **2021**, *93* (10), 4479–4487. https://doi.org/10.1021/acs.analchem.0c04624.
(57) Huang, L.; Wang, Z.; Cupp-Sutton, K. A.; Smith, K.; Wu, S. Spray-Capillary: An Electrospray-Assisted Device for Quantitative Ultralow-Volume Sample Handling. *Anal. Chem.* **2020**, *92* (1), 640–646. https://doi.org/10.1021/acs.analchem.9b04131.
(58) Kahle, J.; Stein, M.; Wätzig, H. Design of Experiments as a Valuable Tool for Biopharmaceutical Analysis with (Imaged) Capillary Isoelectric Focusing. *ELECTROPHORESIS* **2019**, *40* (18–19), 2382–2389. https://doi.org/10.1002/elps.201900162.
(59) Meng, F.; Cargile, B. J.; Patrie, S. M.; Johnson, J. R.; McLoughlin, S. M.; Kelleher, N. L. Processing Complex Mixtures of Intact Proteins for Direct Analysis by Mass Spectrometry. *Anal. Chem.* **2002**, *74* (13), 2923–2929. https://doi.org/10.1021/ac020049i.
(60) Du, Y.; Meng, F.; Patrie, S. M.; Miller, L. M.; Kelleher, N. L. Improved Molecular Weight-Based Processing of Intact Proteins for Interrogation by

Quadrupole-Enhanced FT MS/MS. *J Proteome Res* **2004**, *3* (4), 801–806. https://doi.org/10.1021/pr0499489.

(61) Lee, J. E.; Kellie, J. F.; Tran, J. C.; Tipton, J. D.; Catherman, A. D.; Thomas, H. M.; Ahlf, D. R.; Durbin, K. R.; Vellaichamy, A.; Ntai, I.; Marshall, A. G.; Kelleher, N. L. A Robust Two-Dimensional Separation for Top-Down Tandem Mass Spectrometry of the Low-Mass Proteome. *J Am Soc Mass Spectrom* **2009**, *20* (12), 2183–2191. https://doi.org/10.1016/j.jasms.2009.08.001.

(62) Savaryn, J. P.; Toby, T. K.; Catherman, A. D.; Fellers, R. T.; LeDuc, R. D.; Thomas, P. M.; Friedewald, J. J.; Salomon, D. R.; Abecassis, M. M.; Kelleher, N. L. Comparative Top down Proteomics of Peripheral Blood Mononuclear Cells from Kidney Transplant Recipients with Normal Kidney Biopsies or Acute Rejection. *PROTEOMICS* **2016**, *16* (14), 2048–2058. https://doi.org/10.1002/pmic.201600008.

(63) Tran, J. C.; Doucette, A. A. Multiplexed Size Separation of Intact Proteins in Solution Phase for Mass Spectrometry. *Anal. Chem.* **2009**, *81* (15), 6201–6209. https://doi.org/10.1021/ac900729r.

(64) Drown, B. S.; Jooß, K.; Melani, R. D.; Lloyd-Jones, C.; Camarillo, J. M.; Kelleher, N. L. Mapping the Proteoform Landscape of Five Human Tissues. *J. Proteome Res.* **2022**, *21* (5), 1299–1310. https://doi.org/10.1021/acs.jproteome.2c00034.

(65) Lorenzatto, K. R.; Kim, K.; Ntai, I.; Paludo, G. P.; de Lima, J. C.; Thomas, P. M.; Kelleher, N. L.; Ferreira, H. B. Top Down Proteomics Reveals Mature Proteoforms Expressed in Subcellular Fractions of the Echinococcus Granulosus Preadult Stage. *J Proteome Res* **2015**, *14* (11), 4805–4814. https://doi.org/10.1021/acs.jproteome.5b00642.

(66) Melani, R. D.; Skinner, O. S.; Fornelli, L.; Domont, G. B.; Compton, P. D.; Kelleher, N. L. Mapping Proteoforms and Protein Complexes From King Cobra Venom Using Both Denaturing and Native Top-down Proteomics. *Mol Cell Proteomics* **2016**, *15* (7), 2423–2434. https://doi.org/10.1074/mcp.M115.056523.

(67) Durbin, K. R.; Fornelli, L.; Fellers, R. T.; Doubleday, P. F.; Narita, M.; Kelleher, N. L. Quantitation and Identification of Thousands of Human Proteoforms below 30 KDa. *J Proteome Res* **2016**, *15* (3), 976–982. https://doi.org/10.1021/acs.jproteome.5b00997.

(68) Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L. Large-Scale Top-down Proteomics of the Human Proteome: Membrane Proteins, Mitochondria, and Senescence. *Mol Cell Proteomics* **2013**, *12* (12), 3465–3473. https://doi.org/10.1074/mcp.M113.030114.

(69) Vellaichamy, A.; Tran, J. C.; Catherman, A. D.; Lee, J. E.; Kellie, J. F.; Sweet, S. M. M.; Zamdborg, L.; Thomas, P. M.; Ahlf, D. R.; Durbin, K. R.; Valaskovic, G. A.; Kelleher, N. L. Size-Sorting Combined with Improved Nanocapillary Liquid Chromatography−Mass Spectrometry for Identification of Intact Proteins up to 80 KDa. *Anal. Chem.* **2010**, *82* (4), 1234–1244. https://doi.org/10.1021/ac9021083.

(70) Melani, R. D.; Gerbasi, V. R.; Anderson, L. C.; Sikora, J. W.; Toby, T. K.; Hutton, J. E.; Butcher, D. S.; Negrão, F.; Seckler, H. S.; Srzentić, K.; Fornelli, L.; Camarillo, J. M.; LeDuc, R. D.; Cesnik, A. J.; Lundberg, E.; Greer, J. B.; Fellers, R. T.; Robey, M. T.; DeHart, C. J.; Forte, E.; Hendrickson, C. L.; Abbatiello, S. E.; Thomas, P. M.; Kokaji, A. I.; Levitsky, J.; Kelleher, N. L. The Blood Proteoform Atlas: A Reference Map of Proteoforms in Human Hematopoietic Cells. *Science* **2022**, *375* (6579), 411–418. https://doi.org/10.1126/science.aaz5284.

(71) Takemori, A.; Butcher, D. S.; Harman, V. M.; Brownridge, P.; Shima, K.; Higo, D.; Ishizaki, J.; Hasegawa, H.; Suzuki, J.; Yamashita, M.; Loo, J. A.; Loo, R. R. O.; Beynon, R. J.; Anderson, L. C.; Takemori, N. PEPPI-MS: Polyacrylamide-Gel-Based Prefractionation for Analysis of Intact Proteoforms and Protein Complexes by Mass Spectrometry. *J. Proteome Res.* **2020**, *19* (9), 3779–3791. https://doi.org/10.1021/acs.jproteome.0c00303.

(72) Takemori, A.; Kaulich, P. T.; 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**. https://doi.org/10.1021/acs.analchem.2c02777.

(73) Baghdady, Y. Z.; Schug, K. A. Online Comprehensive High PH Reversed Phase × Low PH Reversed Phase Approach for Two-Dimensional Separations of Intact Proteins in Top-Down Proteomics. *Anal. Chem.* **2019**, *91* (17), 11085–11091. https://doi.org/10.1021/acs.analchem.9b01665.

(74) Wang, Y.; Yang, F.; Gritsenko, M. A.; Wang, Y.; Clauss, T.; Liu, T.; Shen, Y.; Monroe, M. E.; Lopez-Ferrer, D.; Reno, T.; Moore, R. J.; Klemke, R. L.; Camp, D. G.; Smith, R. D. Reversed-Phase Chromatography with Multiple Fraction Concatenation Strategy for Proteome Profiling of Human MCF10A Cells. *Proteomics* **2011**, *11* (10), 2019–2026. https://doi.org/10.1002/pmic.201000722.

(75) 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. *International Journal of Mass Spectrometry* **2018**, *427*, 43–51. https://doi.org/10.1016/j.ijms.2017.09.001.

(76) Yu, D.; Wang, Z.; Cupp-Sutton, K. A.; Liu, X.; Wu, S. Deep Intact Proteoform Characterization in Human Cell Lysate Using High-PH and Low-PH Reversed-Phase Liquid Chromatography. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (12), 2502–2513. https://doi.org/10.1007/s13361-019-02315-2.

(77) Kaulich, P. T.; Cassidy, L.; Tholey, A. Identification of Proteoforms by Top-down Proteomics Using Two-Dimensional Low/Low PH Reversed-Phase Liquid Chromatography-Mass Spectrometry. *PROTEOMICS n/a* (n/a), 2200542. https://doi.org/10.1002/pmic.202200542.

(78) Xiu, L.; Valeja, S. G.; Alpert, A. J.; Jin, S.; Ge, Y. Effective Protein Separation by Coupling Hydrophobic Interaction and Reverse Phase Chromatography for Top-down Proteomics. *Anal. Chem.* **2014**, *86* (15), 7899–7906. https://doi.org/10.1021/ac501836k.

(79) Hjertén, S. Some General Aspects of Hydrophobic Interaction Chromatography. *Journal of Chromatography A* **1973**, *87* (2), 325–331. https://doi.org/10.1016/S0021-9673(01)91733-9.
(80) Melander, W.; Horváth, C. Salt Effects on Hydrophobic Interactions in Precipitation and Chromatography of Proteins: An Interpretation of the Lyotropic Series. *Archives of Biochemistry and Biophysics* **1977**, *183* (1), 200–215. https://doi.org/10.1016/0003-9861(77)90434-9.
(81) Valeja, S. G.; Xiu, L.; Gregorich, Z. R.; Guner, H.; Jin, S.; Ge, Y. Three Dimensional Liquid Chromatography Coupling Ion Exchange Chromatography/Hydrophobic Interaction Chromatography/Reverse Phase Chromatography for Effective Protein Separation in Top-Down Proteomics. *Anal. Chem.* **2015**, *87* (10), 5363–5371. https://doi.org/10.1021/acs.analchem.5b00657.
(82) Buszewski, B.; Noga, S. Hydrophilic Interaction Liquid Chromatography (HILIC)—a Powerful Separation Technique. *Anal Bioanal Chem* **2012**, *402* (1), 231–247. https://doi.org/10.1007/s00216-011-5308-5.
(83) Nguyen, H. P.; Schug, K. A. The Advantages of ESI-MS Detection in Conjunction with HILIC Mode Separations: Fundamentals and Applications. *Journal of Separation Science* **2008**, *31* (9), 1465–1480. https://doi.org/10.1002/jssc.200700630.
(84) Lindner, H.; Sarg, B.; Helliger, W. Application of Hydrophilic-Interaction Liquid Chromatography to the Separation of Phosphorylated H1 Histones. *Journal of Chromatography A* **1997**, *782* (1), 55–62. https://doi.org/10.1016/S0021-9673(97)00468-8.
(85) Garcia, B. A.; Pesavento, J. J.; Mizzen, C. A.; Kelleher, N. L. Pervasive Combinatorial Modification of Histone H3 in Human Cells. *Nat Methods* **2007**, *4* (6), 487–489. https://doi.org/10.1038/nmeth1052.
(86) Pesavento, J. J.; Bullock, C. R.; LeDuc, R. D.; Mizzen, C. A.; Kelleher, N. L. Combinatorial Modification of Human Histone H4 Quantitated by Two-Dimensional Liquid Chromatography Coupled with Top down Mass Spectrometry. *J Biol Chem* **2008**, *283* (22), 14927–14937. https://doi.org/10.1074/jbc.M709796200.
(87) Carroll, J.; Fearnley, I. M.; Walker, J. E. Definition of the Mitochondrial Proteome by Measurement of Molecular Masses of Membrane Proteins. *Proceedings of the National Academy of Sciences* **2006**, *103* (44), 16170–16175. https://doi.org/10.1073/pnas.0607719103.
(88) Domínguez-Vega, E.; Tengattini, S.; Peintner, C.; van Angeren, J.; Temporini, C.; Haselberg, R.; Massolini, G.; Somsen, G. W. High-Resolution Glycoform Profiling of Intact Therapeutic Proteins by Hydrophilic Interaction Chromatography-Mass Spectrometry. *Talanta* **2018**, *184*, 375–381. https://doi.org/10.1016/j.talanta.2018.03.015.
(89) Mant, C. T.; Hodges, R. S. Mixed-Mode Hydrophilic Interaction/Cation-Exchange Chromatography (HILIC/CEX) of Peptides and Proteins. *J Sep Sci* **2008**, *31* (15), 2754–2773. https://doi.org/10.1002/jssc.200800243.

(90) Alpert, A. J. Hydrophilic-Interaction Chromatography for the Separation of Peptides, Nucleic Acids and Other Polar Compounds. *J Chromatogr* **1990**, *499*, 177–196. https://doi.org/10.1016/s0021-9673(00)96972-3.
(91) Lindner, H.; Wesierska-Gadek, J.; Helliger, W.; Puschendorf, B.; Sauermann, G. Identification of ADP-Ribosylated Histones by the Combined Use of High-Performance Liquid Chromatography and Electrophoresis. *J Chromatogr* **1989**, *472* (1), 243–249. https://doi.org/10.1016/s0021-9673(00)94110-4.
(92) Tian, Z.; Tolić, N.; Zhao, R.; Moore, R. J.; Hengel, S. M.; Robinson, E. W.; Stenoien, D. L.; Wu, S.; Smith, R. D.; Paša-Tolić, L. Enhanced Top-down Characterization of Histone Post-Translational Modifications. *Genome Biology* **2012**, *13* (10), R86. https://doi.org/10.1186/gb-2012-13-10-r86.
(93) Young, N. L.; DiMaggio, P. A.; Plazas-Mayorca, M. D.; Baliban, R. C.; Floudas, C. A.; Garcia, B. A. High Throughput Characterization of Combinatorial Histone Codes. *Mol Cell Proteomics* **2009**, *8* (10), 2266–2284. https://doi.org/10.1074/mcp.M900238-MCP200.
(94) Jungbauer, A.; Hahn, R. Chapter 22 Ion-Exchange Chromatography. In *Methods in Enzymology*; Burgess, R. R., Deutscher, M. P., Eds.; Guide to Protein Purification, 2nd Edition; Academic Press, 2009; Vol. 463, pp 349–371. https://doi.org/10.1016/S0076-6879(09)63022-6.
(95) Sharma, S.; Simpson, D. C.; Tolić, N.; Jaitly, N.; Mayampurath, A. M.; Smith, R. D.; Paša-Tolić, L. Proteomic Profiling of Intact Proteins Using WAX-RPLC 2-D Separations and FTICR Mass Spectrometry. *J. Proteome Res.* **2007**, *6* (2), 602–610. https://doi.org/10.1021/pr060354a.
(96) Wu, Z.; Wang, H.; Tustian, A.; Qiu, H.; Li, N. Development of a Two-Dimensional Liquid Chromatography-Mass Spectrometry Platform for Simultaneous Multi-Attribute Characterization of Adeno-Associated Viruses. *Anal. Chem.* **2022**, *94* (7), 3219–3226. https://doi.org/10.1021/acs.analchem.1c04873.
(97) Cassidy, L.; Helbig, A. O.; Kaulich, P. T.; Weidenbach, K.; Schmitz, R. A.; Tholey, A. Multidimensional Separation Schemes Enhance the Identification and Molecular Characterization of Low Molecular Weight Proteomes and Short Open Reading Frame-Encoded Peptides in Top-down Proteomics. *Journal of Proteomics* **2021**, *230*, 103988. https://doi.org/10.1016/j.jprot.2020.103988.
(98) Pepaj, M.; Wilson, S. R.; Novotna, K.; Lundanes, E.; Greibrokk, T. Two-Dimensional Capillary Liquid Chromatography: PH Gradient Ion Exchange and Reversed Phase Chromatography for Rapid Separation of Proteins. *Journal of Chromatography A* **2006**, *1120* (1), 132–141. https://doi.org/10.1016/j.chroma.2006.02.031.
(99) Compton, P. D.; Zamdborg, L.; Thomas, P. M.; Kelleher, N. L. On the Scalability and Requirements of Whole Protein Mass Spectrometry. *Anal Chem* **2011**, *83* (17), 6868–6874. https://doi.org/10.1021/ac2010795.
(100) Li, Y.; Compton, P. D.; Tran, J. C.; Ntai, I.; Kelleher, N. L. Optimizing Capillary Electrophoresis for Top-down Proteomics of 30-80 KDa Proteins. *Proteomics* **2014**, *14* (10), 1158–1164. https://doi.org/10.1002/pmic.201300381.

(101) Brown, K. A.; Tucholski, T.; Alpert, A. J.; Eken, C.; Wesemann, L.; Kyrvasilis, A.; Jin, S.; Ge, Y. Top-Down Proteomics of Endogenous Membrane Proteins Enabled by Cloud Point Enrichment and Multidimensional Liquid Chromatography–Mass Spectrometry. *Anal. Chem.* **2020**, *92* (24), 15726–15735. https://doi.org/10.1021/acs.analchem.0c02533.

(102) Simpson, D. C.; Ahn, S.; Pasa-Tolic, L.; Bogdanov, B.; Mottaz, H. M.; Vilkov, A. N.; Anderson, G. A.; Lipton, M. S.; Smith, R. D. Using Size Exclusion Chromatography-RPLC and RPLC-CIEF as Two-Dimensional Separation Strategies for Protein Profiling. *ELECTROPHORESIS* **2006**, *27* (13), 2722–2733. https://doi.org/10.1002/elps.200600037.

(103) Ventouri, I. K.; Malheiro, D. B. A.; Voeten, R. L. C.; Kok, S.; Honing, M.; Somsen, G. W.; Haselberg, R. Probing Protein Denaturation during Size-Exclusion Chromatography Using Native Mass Spectrometry. *Analytical Chemistry* **2020**. https://doi.org/10.1021/acs.analchem.9b04961.

(104) Burgess, R. R. A Brief Practical Review of Size Exclusion Chromatography: Rules of Thumb, Limitations, and Troubleshooting. *Protein Expression and Purification* **2018**, *150*, 81–85. https://doi.org/10.1016/j.pep.2018.05.007.

(105) Gomes, F. P.; Diedrich, J. K.; Saviola, A. J.; Memili, E.; Moura, A. A.; Yates, J. R. I. EThcD and 213 Nm UVPD for Top-Down Analysis of Bovine Seminal Plasma Proteoforms on Electrophoretic and Chromatographic Time Frames. *Anal. Chem.* **2020**, *92* (4), 2979–2987. https://doi.org/10.1021/acs.analchem.9b03856.

(106) Tran, J. C.; Doucette, A. A. Gel-Eluted Liquid Fraction Entrapment Electrophoresis: An Electrophoretic Method for Broad Molecular Weight Range Proteome Separation. *Anal. Chem.* **2008**, *80* (5), 1568–1573. https://doi.org/10.1021/ac702197w.

(107) Botelho, D.; Wall, M. J.; Vieira, D. B.; Fitzsimmons, S.; Liu, F.; Doucette, A. Top-Down and Bottom-Up Proteomics of SDS-Containing Solutions Following Mass-Based Separation. *J. Proteome Res.* **2010**, *9* (6), 2863–2870. https://doi.org/10.1021/pr900949p.

(108) Wessel, D.; Flügge, U. I. A Method for the Quantitative Recovery of Protein in Dilute Solution in the Presence of Detergents and Lipids. *Anal Biochem* **1984**, *138* (1), 141–143. https://doi.org/10.1016/0003-2697(84)90782-6.

(109) Antharavally, B. S.; Mallia, K. A.; Rosenblatt, M. M.; Salunkhe, A. M.; Rogers, J. C.; Haney, P.; Haghdoost, N. Efficient Removal of Detergents from Proteins and Peptides in a Spin Column Format. *Analytical Biochemistry* **2011**, *416* (1), 39–44. https://doi.org/10.1016/j.ab.2011.05.013.

(110) Hengel, S. M.; Floyd, E.; Baker, E. S.; Zhao, R.; Wu, S.; Paša-Tolić, L. Evaluation of SDS Depletion Using an Affinity Spin Column and IMS-MS Detection. *PROTEOMICS* **2012**, *12* (21), 3138–3142. https://doi.org/10.1002/pmic.201200168.

(111) Xu, T.; Shen, X.; Yang, Z.; Chen, D.; Lubeckyj, R. A.; McCool, E. N.; Sun, L. Automated Capillary Isoelectric Focusing-Tandem Mass Spectrometry for Qualitative and Quantitative Top-Down Proteomics. *Anal Chem* **2020**, *92* (24), 15890–15898. https://doi.org/10.1021/acs.analchem.0c03266.

(112) Xu, T.; Han, L.; George Thompson, A. M.; Sun, L. An Improved Capillary Isoelectric Focusing-Mass Spectrometry Method for High-Resolution Characterization of Monoclonal Antibody Charge Variants. *Anal Methods* **2022**, *14* (4), 383–393. https://doi.org/10.1039/d1ay01556g.

(113) Zhao, Y.; Sun, L.; Zhu, G.; Dovichi, N. J. Coupling Capillary Zone Electrophoresis to a Q Exactive HF Mass Spectrometer for Top-down Proteomics: 580 Proteoform Identifications from Yeast. *J. Proteome Res.* **2016**, *15* (10), 3679–3685. https://doi.org/10.1021/acs.jproteome.6b00493.

(114) Yamamoto, H.; Manabe, T.; Okuyama, T. Gel Permeation Chromatography Combined with Capillary Electrophoresis for Microanalysis of Proteins. *Journal of Chromatography A* **1989**, *480*, 277–283. https://doi.org/10.1016/S0021-9673(01)84295-3.

(115) Bushey, M. M.; Jorgenson, J. W. Automated Instrumentation for Comprehensive Two-Dimensional High-Performance Liquid Chromatography of Proteins. *Anal. Chem.* **1990**, *62* (2), 161–167. https://doi.org/10.1021/ac00201a015.

(116) Jooß, K.; Scholz, N.; Meixner, J.; Neusüß, C. Heart-Cut Nano-LC–CZE–MS for the Characterization of Proteins on the Intact Level. *ELECTROPHORESIS* **2019**, *40* (7), 1061–1065. https://doi.org/10.1002/elps.201800411.

(117) Mellors, J. S.; Black, W. A.; Chambers, A. G.; Starkey, J. A.; Lacher, N. A.; Ramsey, J. M. Hybrid Capillary/Microfluidic System for Comprehensive Online Liquid Chromatography-Capillary Electrophoresis-Electrospray Ionization-Mass Spectrometry. *Anal. Chem.* **2013**, *85* (8), 4100–4106. https://doi.org/10.1021/ac400205a.

(118) Chambers, A. G.; Mellors, J. S.; Henley, W. H.; Ramsey, J. M. Monolithic Integration of Two-Dimensional Liquid Chromatography-Capillary Electrophoresis and Electrospray Ionization on a Microfluidic Device. *Anal Chem* **2011**, *83* (3), 842–849. https://doi.org/10.1021/ac102437z.

(119) Stolz, A.; Neusüß, C. Characterisation of a New Online NanoLC-CZE-MS Platform and Application for the Glycosylation Profiling of Alpha-1-Acid Glycoprotein. *Anal Bioanal Chem* **2022**, *414* (5), 1745–1757. https://doi.org/10.1007/s00216-021-03814-6.

(120) Tang, Qing.; Harrata, A. Kamel.; Lee, C. S. Capillary Isoelectric Focusing-Electrospray Mass Spectrometry for Protein Analysis. *Anal. Chem.* **1995**, *67* (19), 3515–3519. https://doi.org/10.1021/ac00115a021.

(121) Mohan, D.; Lee, C. S. On-Line Coupling of Capillary Isoelectric Focusing with Transient Isotachophoresis-Zone Electrophoresis: A Two-Dimensional Separation System for Proteomics. *Electrophoresis* **2002**, *23* (18), 3160–3167. https://doi.org/10.1002/1522-2683(200209)23:18<3160::AID-ELPS3160>3.0.CO;2-L.

(122) Chen, J.; Balgley, B. M.; DeVoe, D. L.; Lee, C. S. Capillary Isoelectric Focusing-Based Multidimensional Concentration/Separation Platform for Proteome Analysis. *Anal Chem* **2003**, *75* (13), 3145–3152. https://doi.org/10.1021/ac034014+.

(123) Hühner, J.; Neusüß, C. CIEF-CZE-MS Applying a Mechanical Valve. *Anal Bioanal Chem* **2016**, *408* (15), 4055–4061. https://doi.org/10.1007/s00216-016-9498-8.
(124) Hühner, J.; Jooß, K.; Neusüß, C. Interference-Free Mass Spectrometric Detection of Capillary Isoelectric Focused Proteins, Including Charge Variants of a Model Monoclonal Antibody. *Electrophoresis* **2017**, *38* (6), 914–921. https://doi.org/10.1002/elps.201600457.
(125) McCool, E. N.; Lodge, J. M.; Basharat, A. R.; Liu, X.; Coon, J. J.; Sun, L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry with Activated Ion Electron Transfer Dissociation for Large-Scale Top-down Proteomics. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (12), 2470–2479. https://doi.org/10.1007/s13361-019-02206-6.
(126) McCool, E. N.; Chen, D.; Li, W.; Liu, Y.; Sun, L. Capillary Zone Electrophoresis-Tandem Mass Spectrometry with Ultraviolet Photodissociation (213 Nm) for Large-Scale Top–down Proteomics. *Anal. Methods* **2019**, *11* (22), 2855–2861. https://doi.org/10.1039/C9AY00585D.
(127) Shen, X.; Kou, Q.; Guo, R.; Yang, Z.; Chen, D.; Liu, X.; Hong, H.; Sun, L. Native Proteomics in Discovery Mode Using Size-Exclusion Chromatography–Capillary Zone Electrophoresis–Tandem Mass Spectrometry. *Anal. Chem.* **2018**, *90* (17), 10095–10099. https://doi.org/10.1021/acs.analchem.8b02725.
(128) Ren, J.; Beckner, M. A.; Lynch, K. B.; Chen, H.; Zhu, Z.; Yang, Y.; Chen, A.; Qiao, Z.; Liu, S.; Lu, J. J. Two-Dimensional Liquid Chromatography Consisting of Twelve Second-Dimension Columns for Comprehensive Analysis of Intact Proteins. *Talanta* **2018**, *182*, 225–229. https://doi.org/10.1016/j.talanta.2018.01.072.
(129) Jooß, K.; Hühner, J.; Kiessig, S.; Moritz, B.; Neusüß, C. Two-Dimensional Capillary Zone Electrophoresis-Mass Spectrometry for the Characterization of Intact Monoclonal Antibody Charge Variants, Including Deamidation Products. *Anal Bioanal Chem* **2017**, *409* (26), 6057–6067. https://doi.org/10.1007/s00216-017-0542-0.
(130) Yu, D.; Wang, Z.; Cupp-Sutton, K. A.; Guo, Y.; Kou, Q.; Smith, K.; Liu, X.; Wu, S. Quantitative Top-Down Proteomics in Complex Samples Using Protein-Level Tandem Mass Tag Labeling. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (6), 1336–1344. https://doi.org/10.1021/jasms.0c00464.
(131) Winkels, K.; Koudelka, T.; Tholey, A. Quantitative Top-Down Proteomics by Isobaric Labeling with Thiol-Directed Tandem Mass Tags. *J. Proteome Res.* **2021**, *20* (9), 4495–4506. https://doi.org/10.1021/acs.jproteome.1c00460.
(132) Guo, Y.; Yu, D.; Cupp-Sutton, K. A.; Liu, X.; Wu, S. Optimization of Protein-Level Tandem Mass Tag (TMT) Labeling Conditions in Complex Samples with Top-down Proteomics. *Analytica Chimica Acta* **2022**, *1221*, 340037. https://doi.org/10.1016/j.aca.2022.340037.
(133) Guo, Y.; Chowdhury, T.; Seshadri, M.; Cupp-Sutton, K. A.; Wang, Q.; Yu, D.; Wu, S. Optimization of Higher-Energy Collisional Dissociation Fragmentation Energy for Intact Protein-Level Tandem Mass Tag Labeling. *J. Proteome Res.* **2023**. https://doi.org/10.1021/acs.jproteome.2c00549.

(134) Webber, K. G. I.; Truong, T.; Johnston, S. M.; Zapata, S. E.; Liang, Y.; Davis, J. M.; Buttars, A. D.; Smith, F. B.; Jones, H. E.; Mahoney, A. C.; Carson, R. H.; Nwosu, A. J.; Heninger, J. L.; Liyu, A. V.; Nordin, G. P.; Zhu, Y.; Kelly, R. T. Label-Free Profiling of up to 200 Single-Cell Proteomes per Day Using a Dual-Column Nanoflow Liquid Chromatography Platform. *Anal. Chem.* **2022**, *94* (15), 6017–6025. https://doi.org/10.1021/acs.analchem.2c00646.

(135) Wang, H.; Hanash, S. M. Increased Throughput and Reduced Carryover of Mass Spectrometry-Based Proteomics Using a High-Efficiency Nonsplit Nanoflow Parallel Dual-Column Capillary HPLC System. *J. Proteome Res.* **2008**, *7* (7), 2743–2755. https://doi.org/10.1021/pr700876g.

(136) Livesay, E. A.; Tang, K.; Taylor, B. K.; Buschbach, M. A.; Hopkins, D. F.; LaMarche, B. L.; Zhao, R.; Shen, Y.; Orton, D. J.; Moore, R. J.; Kelly, R. T.; Udseth, H. R.; Smith, R. D. Fully Automated Four-Column Capillary LC−MS System for Maximizing Throughput in Proteomic Analyses. *Anal. Chem.* **2008**, *80* (1), 294–302. https://doi.org/10.1021/ac701727r.

(137) Orton, D. J.; Wall, M. J.; Doucette, A. A. Dual LC–MS Platform for High-Throughput Proteome Analysis. *J. Proteome Res.* **2013**, *12* (12), 5963–5970. https://doi.org/10.1021/pr400738a.

(138) Lee, H.; Lee, J. H.; Kim, H.; Kim, S.-J.; Bae, J.; Kim, H. K.; Lee, S.-W. A Fully Automated Dual-Online Multifunctional Ultrahigh Pressure Liquid Chromatography System for High-Throughput Proteomics Analysis. *Journal of Chromatography A* **2014**, *1329*, 83–89. https://doi.org/10.1016/j.chroma.2013.12.084.

(139) Lee, H.; Mun, D.-G.; Bae, J.; Kim, H.; Yeon Oh, S.; Soo Park, Y.; Lee, J.-H.; Lee, S.-W. A Simple Dual Online Ultra-High Pressure Liquid Chromatography System (SDO-UHPLC) for High Throughput Proteome Analysis. *Analyst* **2015**, *140* (16), 5700–5706. https://doi.org/10.1039/C5AN00639B.

(140) Hosp, F.; Scheltema, R. A.; Eberl, H. C.; Kulak, N. A.; Keilhauer, E. C.; Mayr, K.; Mann, M. A Double-Barrel Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS) System to Quantify 96 Interactomes per Day*. *Molecular & Cellular Proteomics* **2015**, *14* (7), 2030–2041. https://doi.org/10.1074/mcp.O115.049460.

(141) Kuehnbaum, N. L.; Kormendi, A.; Britz-McKibbin, P. Multisegment Injection-Capillary Electrophoresis-Mass Spectrometry: A High-Throughput Platform for Metabolomics with High Data Fidelity. *Anal. Chem.* **2013**, *85* (22), 10664–10669. https://doi.org/10.1021/ac403171u.

(142) Wu, S.; Brown, J. N.; Tolić, N.; Meng, D.; Liu, X.; Zhang, H.; Zhao, R.; Moore, R. J.; Pevzner, P.; Smith, R. D.; Paša-Tolić, L. Quantitative Analysis of Human Salivary Gland-Derived Intact Proteome Using Top-down Mass Spectrometry. *Proteomics* **2014**, *14* (10), 1211–1222. https://doi.org/10.1002/pmic.201300378.

(143) Ntai, I.; Kim, K.; Fellers, R. T.; Skinner, O. S.; Smith, A. D.; Early, B. P.; Savaryn, J. P.; LeDuc, R. D.; Thomas, P. M.; Kelleher, N. L. Applying Label-

Free Quantitation to Top down Proteomics. *Anal Chem* **2014**, *86* (10), 4961–4968. https://doi.org/10.1021/ac500395k.
(144) Zhang, J.; Guy, M. J.; Norman, H. S.; Chen, Y.-C.; Xu, Q.; Dong, X.; Guner, H.; Wang, S.; Kohmoto, T.; Young, K. H.; Moss, R. L.; Ge, Y. Top-down Quantitative Proteomics Identified Phosphorylation of Cardiac Troponin I as a Candidate Biomarker for Chronic Heart Failure. *J Proteome Res* **2011**, *10* (9), 4054–4065. https://doi.org/10.1021/pr200258m.
(145) Collier, T. S.; Sarkar, P.; Rao, B.; Muddiman, D. C. Quantitative Top-down Proteomics of SILAC Labeled Human Embryonic Stem Cells. *J Am Soc Mass Spectrom* **2010**, *21* (6), 879–889. https://doi.org/10.1016/j.jasms.2010.01.031.
(146) Waanders, L. F.; Hanke, S.; Mann, M. Top-down Quantitation and Characterization of SILAC-Labeled Proteins. *J Am Soc Mass Spectrom* **2007**, *18* (11), 2058–2064. https://doi.org/10.1016/j.jasms.2007.09.001.
(147) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; 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**, *1* (5), 376–386. https://doi.org/10.1074/mcp.m200025-mcp200.
(148) Collier, T. S.; Hawkridge, A. M.; Georgianna, D. R.; Payne, G. A.; Muddiman, D. C. Top-down Identification and Quantification of Stable Isotope Labeled Proteins from Aspergillus Flavus Using Online Nano-Flow Reversed-Phase Liquid Chromatography Coupled to a LTQ-FTICR Mass Spectrometer. *Anal Chem* **2008**, *80* (13), 4994–5001. https://doi.org/10.1021/ac800254z.
(149) Rhoads, T. W.; Rose, C. M.; Bailey, D. J.; Riley, N. M.; Molden, R. C.; Nestler, A. J.; Merrill, A. E.; Smith, L. M.; Hebert, A. S.; Westphall, M. S.; Pagliarini, D. J.; Garcia, B. A.; Coon, J. J. Neutron-Encoded Mass Signatures for Quantitative Top-Down Proteomics. *Anal. Chem.* **2014**, *86* (5), 2314–2319. https://doi.org/10.1021/ac403579s.
(150) Shortreed, M. R.; Frey, B. L.; Scalf, M.; Knoener, R. A.; Cesnik, A. J.; Smith, L. M. Elucidating Proteoform Families from Proteoform Intact-Mass and Lysine-Count Measurements. *J Proteome Res* **2016**, *15* (4), 1213–1221. https://doi.org/10.1021/acs.jproteome.5b01090.
(151) Merrill, A. E.; Hebert, A. S.; MacGilvray, M. E.; Rose, C. M.; Bailey, D. J.; Bradley, J. C.; Wood, W. W.; El Masri, M.; Westphall, M. S.; Gasch, A. P.; Coon, J. J. NeuCode Labels for Relative Protein Quantification. *Mol Cell Proteomics* **2014**, *13* (9), 2503–2512. https://doi.org/10.1074/mcp.M114.040287.
(152) Quijada, J. V.; Schmitt, N. D.; Salisbury, J. P.; Auclair, J. R.; Agar, J. N. Heavy Sugar and Heavy Water Create Tunable Intact Protein Mass Increases for Quantitative Mass Spectrometry in Any Feed and Organism. *Anal Chem* **2016**, *88* (22), 11139–11146. https://doi.org/10.1021/acs.analchem.6b03234.
(153) Du, Y.; Parks, B. A.; Sohn, S.; Kwast, K. E.; Kelleher, N. L. Top-down Approaches for Measuring Expression Ratios of Intact Yeast Proteins Using Fourier Transform Mass Spectrometry. *Anal Chem* **2006**, *78* (3), 686–694. https://doi.org/10.1021/ac050993p.

(154) Chen, X.; Wei, S.; Ji, Y.; Guo, X.; Yang, F. Quantitative Proteomics Using SILAC: Principles, Applications, and Developments. *Proteomics* **2015**, *15* (18), 3175–3192. https://doi.org/10.1002/pmic.201500108.
(155) Veenstra, T. D.; Martinović, S.; Anderson, G. A.; Pasa-Tolić, L.; Smith, R. D. Proteome Analysis Using Selective Incorporation of Isotopically Labeled Amino Acids. *J Am Soc Mass Spectrom* **2000**, *11* (1), 78–82. https://doi.org/10.1016/S1044-0305(99)00120-8.
(156) Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J. Neutron-Encoded Mass Signatures for Multiplexed Proteome Quantification. *Nat Methods* **2013**, *10* (4), 332–334. https://doi.org/10.1038/nmeth.2378.
(157) Sivanich, M. K.; Gu, T.-J.; Tabang, D. N.; Li, L. Recent Advances in Isobaric Labeling and Applications in Quantitative Proteomics. *PROTEOMICS* **2022**, *22* (19–20), 2100256. https://doi.org/10.1002/pmic.202100256.
(158) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; 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, D. J. Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents *. *Molecular & Cellular Proteomics* **2004**, *3* (12), 1154–1169. https://doi.org/10.1074/mcp.M400129-MCP200.
(159) Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Anal. Chem.* **2003**, *75* (8), 1895–1904. https://doi.org/10.1021/ac0262560.
(160) Li, J.; Van Vranken, J. G.; Pontano Vaites, L.; Schweppe, D. K.; Huttlin, E. L.; Etienne, C.; Nandhikonda, P.; Viner, R.; Robitaille, A. M.; Thompson, A. H.; Kuhn, K.; Pike, I.; Bomgarden, R. D.; Rogers, J. C.; Gygi, S. P.; Paulo, J. A. TMTpro Reagents: A Set of Isobaric Labeling Mass Tags Enables Simultaneous Proteome-Wide Measurements across 16 Samples. *Nat Methods* **2020**, *17* (4), 399–404. https://doi.org/10.1038/s41592-020-0781-4.
(161) Qu, Z.; Meng, F.; Bomgarden, R. D.; Viner, R. I.; Li, J.; Rogers, J. C.; Cheng, J.; Greenlief, C. M.; Cui, J.; Lubahn, D. B.; Sun, G. Y.; Gu, Z. Proteomic Quantification and Site-Mapping of S-Nitrosylated Proteins Using Isobaric IodoTMT Reagents. *J. Proteome Res.* **2014**, *13* (7), 3200–3211. https://doi.org/10.1021/pr401179v.
(162) Xiang, F.; Ye, H.; Chen, R.; Fu, Q.; Li, L. N,N-Dimethyl Leucines as Novel Isobaric Tandem Mass Tags for Quantitative Proteomics and Peptidomics. *Anal. Chem.* **2010**, *82* (7), 2817–2825. https://doi.org/10.1021/ac902778d.
(163) Frost, D. C.; Feng, Y.; Li, L. 21-Plex DiLeu Isobaric Tags for High-Throughput Quantitative Proteomics. *Anal. Chem.* **2020**, *92* (12), 8228–8234. https://doi.org/10.1021/acs.analchem.0c00473.
(164) Wang, D.; Ma, M.; Huang, J.; Gu, T.-J.; Cui, Y.; Li, M.; Wang, Z.; Zetterberg, H.; Li, L. Boost-DiLeu: Enhanced Isobaric N,N-Dimethyl Leucine Tagging Strategy for a Comprehensive Quantitative Glycoproteomic Analysis.

Anal. Chem. **2022**, *94* (34), 11773–11782. https://doi.org/10.1021/acs.analchem.2c01773.
(165) Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B. Quantitative Mass Spectrometry in Proteomics: A Critical Review. *Anal Bioanal Chem* **2007**, *389* (4), 1017–1031. https://doi.org/10.1007/s00216-007-1486-6.
(166) Chen, X.; Sun, Y.; Zhang, T.; Shu, L.; Roepstorff, P.; Yang, F. Quantitative Proteomics Using Isobaric Labeling: A Practical Guide. *Genomics, Proteomics & Bioinformatics* **2021**, *19* (5), 689–706. https://doi.org/10.1016/j.gpb.2021.08.012.
(167) Nie, S.; Lo, A.; Zhu, J.; Wu, J.; Ruffin, M.; Lubman, D. Isobaric Protein-Level Labeling Strategy for Serum Glycoprotein Quantification Analysis by Liquid Chromatography-Tandem Mass Spectrometry. *Analytical chemistry* **2013**. https://doi.org/10.1021/ac400838s.
(168) Wiese, S.; Reidegeld, K. A.; Meyer, H. E.; Warscheid, B. Protein Labeling by ITRAQ: A New Tool for Quantitative Mass Spectrometry in Proteome Research. *Proteomics* **2007**, *7* (3), 340–350. https://doi.org/10.1002/pmic.200600422.
(169) Sinclair, J.; Timms, J. F. Quantitative Profiling of Serum Samples Using TMT Protein Labelling, Fractionation and LC-MS/MS. *Methods* **2011**, *54* (4), 361–369. https://doi.org/10.1016/j.ymeth.2011.03.004.
(170) Prudova, A.; Keller, U. auf dem; Butler, G. S.; Overall, C. M. Multiplex N-Terminome Analysis of MMP-2 and MMP-9 Substrate Degradomes by ITRAQ-TAILS Quantitative Proteomics *. *Molecular & Cellular Proteomics* **2010**, *9* (5), 894–911. https://doi.org/10.1074/mcp.M000050-MCP201.
(171) Hung, C.-W.; Tholey, A. Tandem Mass Tag Protein Labeling for Top-Down Identification and Quantification. *Anal. Chem.* **2012**, *84* (1), 161–170. https://doi.org/10.1021/ac202243r.
(172) Guo, Y.; Yu, D.; Cupp-Sutton, K. A.; Liu, X.; Wu, S. A Benchmarking Protocol for Intact Protein-Level Tandem Mass Tag (TMT) Labeling for Quantitative Top-Down Proteomics. *MethodsX* **2022**, 101873. https://doi.org/10.1016/j.mex.2022.101873.
(173) Winkels, K.; Koudelka, T.; Tholey, A. Quantitative Top-Down Proteomics by Isobaric Labeling with Thiol-Directed Tandem Mass Tags. *J. Proteome Res.* **2021**, *20* (9), 4495–4506. https://doi.org/10.1021/acs.jproteome.1c00460.
(174) *21-plex DiLeu Isobaric Tags for High-Throughput Quantitative Proteomics | Analytical Chemistry*. https://pubs.acs.org/doi/abs/10.1021/acs.analchem.0c00473 (accessed 2021-12-23).
(175) Zecha, J.; Satpathy, S.; Kanashova, T.; Avanessian, S. C.; Kane, M. H.; Clauser, K. R.; Mertins, P.; Carr, S. A.; Kuster, B. TMT Labeling for the Masses: A Robust and Cost-Efficient, In-Solution Labeling Approach. *Mol Cell Proteomics* **2019**, *18* (7), 1468–1478. https://doi.org/10.1074/mcp.TIR119.001385.
(176) Megger, D. A.; Pott, L. L.; Rosowski, K.; Zülch, B.; Tautges, S.; Bracht, T.; Sitek, B. NHS-Based Tandem Mass Tagging of Proteins at the Level of Whole Cells: A Critical Evaluation in Comparison to Conventional TMT-Labeling

Approaches for Quantitative Proteome Analysis. *Anal. Sci.* **2017**, *33* (12), 1387–1394. https://doi.org/10.2116/analsci.33.1387.

(177) Yu, K.; Niu, M.; Wang, H.; Li, Y.; Wu, Z.; Zhang, B.; Haroutunian, V.; Peng, J. Global Profiling of Lysine Accessibility to Evaluate Protein Structure Changes in Alzheimer's Disease. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (4), 936–945. https://doi.org/10.1021/jasms.0c00450.

(178) Fang, M.; Wang, Z.; Cupp-Sutton, K. A.; Welborn, T.; Smith, K.; Wu, S. High-Throughput Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS) Coupled with Subzero-Temperature Ultrahigh Pressure Liquid Chromatography (UPLC) Separation for Complex Sample Analysis. *Analytica Chimica Acta* **2021**, *1143*, 65–72. https://doi.org/10.1016/j.aca.2020.11.022.

(179) Wu, S.; Brown, R. N.; Payne, S. H.; Meng, D.; Zhao, R.; Tolić, N.; Cao, L.; Shukla, A.; Monroe, M. E.; Moore, R. J.; Lipton, M. S.; Paša-Tolić, L. Top-Down Characterization of the Post-Translationally Modified Intact Periplasmic Proteome from the Bacterium Novosphingobium Aromaticivorans. *Int J Proteomics* **2013**, *2013*, 279590. https://doi.org/10.1155/2013/279590.

(180) Ansong, C.; Wu, S.; Meng, D.; Liu, X.; Brewer, H. M.; Kaiser, B. L. D.; Nakayasu, E. S.; Cort, J. R.; Pevzner, P.; Smith, R. D.; Heffron, F.; Adkins, J. N.; Paša-Tolić, L. Top-down Proteomics Reveals a Unique Protein S-Thiolation Switch in Salmonella Typhimurium in Response to Infection-like Conditions. *PNAS* **2013**, *110* (25), 10153–10158. https://doi.org/10.1073/pnas.1221210110.

(181) Wu, S.; Yang, F.; Zhao, R.; Tolić, N.; Robinson, E. W.; Camp, D. G.; Smith, R. D.; Paša-Tolić, L. Integrated Workflow for Characterizing Intact Phosphoproteins from Complex Mixtures. *Anal. Chem.* **2009**, *81* (11), 4210–4219. https://doi.org/10.1021/ac802487q.

(182) Liu, Z.; Wang, R.; Liu, J.; Sun, R.; Wang, F. Global Quantification of Intact Proteins via Chemical Isotope Labeling and Mass Spectrometry. *J. Proteome Res.* **2019**, *18* (5), 2185–2194. https://doi.org/10.1021/acs.jproteome.9b00071.

(183) 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. https://doi.org/10.1093/bioinformatics/btw398.

(184) Guner, H.; Close, P. L.; Cai, W.; Zhang, H.; Peng, Y.; Gregorich, Z. R.; Ge, Y. MASH Suite: A User-Friendly and Versatile Software Interface for High-Resolution Mass Spectrometry Data Interpretation and Visualization. *J. Am. Soc. Mass Spectrom.* **2014**, *25* (3), 464–470. https://doi.org/10.1007/s13361-013-0789-4.

(185) Fellers, R. T.; Greer, J. B.; Early, B. P.; Yu, X.; LeDuc, R. D.; Kelleher, N. L.; Thomas, P. M. ProSight Lite: Graphical Software to Analyze Top-down Mass Spectrometry Data. *PROTEOMICS* **2015**, *15* (7), 1235–1238. https://doi.org/10.1002/pmic.201400313.

(186) Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. A Summary of the Measured PK Values of the Ionizable Groups in Folded Proteins. *Protein Sci* **2009**, *18* (1), 247–251. https://doi.org/10.1002/pro.19.
(187) Miller, B. T.; Collins, T. J.; Nagle, G. T.; Kurosky, A. The Occurrence of O-Acylation during Biotinylation of Gonadotropin-Releasing Hormone and Analogs. Evidence for a Reactive Serine. *J Biol Chem* **1992**, *267* (8), 5060–5069.
(188) Wiktorowicz, J. E.; English, R. D.; Wu, Z.; Kurosky, A. Model Studies on ITRAQ Modification of Peptides: Sequence-Dependent Reaction Specificity. *J. Proteome Res.* **2012**, *11* (3), 1512–1520. https://doi.org/10.1021/pr2003165.
(189) Yang, W.-C.; Mirzaei, H.; Liu, X.; Regnier, F. E. Enhancement of Amino Acid Detection and Quantification by Electrospray Ionization Mass Spectrometry. *Anal. Chem.* **2006**, *78* (13), 4702–4708. https://doi.org/10.1021/ac0600510.
(190) Wang, S.; Regnier, F. E. Proteomics Based on Selecting and Quantifying Cysteine Containing Peptides by Covalent Chromatography. *J Chromatogr A* **2001**, *924* (1–2), 345–357. https://doi.org/10.1016/s0021-9673(01)00961-x.
(191) Regnier, F. E.; Julka, S. Primary Amine Coding as a Path to Comparative Proteomics. *Proteomics* **2006**, *6* (14), 3968–3979. https://doi.org/10.1002/pmic.200500553.
(192) Nick Pace, C.; Huyghues-Despointes, B. M. P.; Fu, H.; Takano, K.; Scholtz, J. M.; Grimsley, G. R. Urea Denatured State Ensembles Contain Extensive Secondary Structure That Is Increased in Hydrophobic Proteins. *Protein Sci* **2010**, *19* (5), 929–943. https://doi.org/10.1002/pro.370.
(193) Griebenow, K.; Klibanov, A. M. On Protein Denaturation in Aqueous−Organic Mixtures but Not in Pure Organic Solvents. *J. Am. Chem. Soc.* **1996**, *118* (47), 11695–11700. https://doi.org/10.1021/ja961869d.
(194) Witkowski, C.; Harkins, J. Using the GELFREE 8100 Fractionation System for Molecular Weight-Based Fractionation with Liquid Phase Recovery. *J Vis Exp* **2009**, No. 34, 1842. https://doi.org/10.3791/1842.
(195) Brown, K. A.; Chen, B.; Guardado-Alvarez, T. M.; Lin, Z.; Hwang, L.; Ayaz-Guner, S.; Jin, S.; Ge, Y. A Photocleavable Surfactant for Top-down Proteomics. *Nat Methods* **2019**, *16* (5), 417–420. https://doi.org/10.1038/s41592-019-0391-1.
(196) Melby, J. A.; Roberts, D. S.; Larson, E. J.; Brown, K. A.; Bayne, E. F.; Jin, S.; Ge, Y. Novel Strategies to Address the Challenges in Top-Down Proteomics. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (6), 1278–1294. https://doi.org/10.1021/jasms.1c00099.
(197) Diedrich, J. K.; Pinto, A. F. M.; Yates, J. R. Energy Dependence of HCD on Peptide Fragmentation: Stepped Collisional Energy Finds the Sweet Spot. *J Am Soc Mass Spectrom* **2013**, *24* (11), 10.1007/s13361-013-0709–7. https://doi.org/10.1007/s13361-013-0709-7.
(198) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-Energy C-Trap Dissociation for Peptide Modification Analysis. *Nat Methods* **2007**, *4* (9), 709–712. https://doi.org/10.1038/nmeth1060.

(199) Jedrychowski, M. P.; Huttlin, E. L.; Haas, W.; Sowa, M. E.; Rad, R.; Gygi, S. P. Evaluation of HCD- and CID-Type Fragmentation Within Their Respective Detection Platforms For Murine Phosphoproteomics. *Mol Cell Proteomics* **2011**, *10* (12), M111.009910. https://doi.org/10.1074/mcp.M111.009910.

(200) Köcher, T.; Pichler, P.; Schutzbier, M.; Stingl, C.; Kaul, A.; Teucher, N.; Hasenfuss, G.; Penninger, J. M.; Mechtler, K. High Precision Quantitative Proteomics Using ITRAQ on an LTQ Orbitrap: A New Mass Spectrometric Method Combining the Benefits of All. *J. Proteome Res.* **2009**, *8* (10), 4743–4752. https://doi.org/10.1021/pr900451u.

(201) Zhang, Y.; Ficarro, S. B.; Li, S.; Marto, J. A. Optimized Orbitrap HCD for Quantitative Analysis of Phosphopeptides. *Journal of the American Society for Mass Spectrometry* **2009**, *20* (8), 1425–1434. https://doi.org/10.1016/j.jasms.2009.03.019.

(202) Pichler, P.; Köcher, T.; Holzmann, J.; Möhring, T.; Ammerer, G.; Mechtler, K. Improved Precision of ITRAQ and TMT Quantification by an Axial Extraction Field in an Orbitrap HCD Cell. *Anal. Chem.* **2011**, *83* (4), 1469–1474. https://doi.org/10.1021/ac102265w.

(203) Chiva, C.; Sabidó, E. HCD-Only Fragmentation Method Balances Peptide Identification and Quantitation of TMT-Labeled Samples in Hybrid Linear Ion Trap/Orbitrap Mass Spectrometers. *Journal of Proteomics* **2014**, *96*, 263–270. https://doi.org/10.1016/j.jprot.2013.11.013.

(204) Dayon, L.; Pasquarello, C.; Hoogland, C.; Sanchez, J.-C.; Scherl, A. Combining Low- and High-Energy Tandem Mass Spectra for Optimized Peptide Quantification with Isobaric Tags. *Journal of Proteomics* **2010**, *73* (4), 769–777. https://doi.org/10.1016/j.jprot.2009.10.015.

(205) Wang, Y.; Tian, Z. New Energy Setup Strategy for Intact N-Glycopeptides Characterization Using Higher-Energy Collisional Dissociation. *J. Am. Soc. Mass Spectrom.* **2020**, *31* (3), 651–657. https://doi.org/10.1021/jasms.9b00089.

(206) Altelaar, A. F. M.; Frese, C. K.; Preisinger, C.; Hennrich, M. L.; Schram, A. W.; Timmers, H. Th. M.; Heck, A. J. R.; Mohammed, S. Benchmarking Stable Isotope Labeling Based Quantitative Proteomics. *Journal of Proteomics* **2013**, *88*, 14–26. https://doi.org/10.1016/j.jprot.2012.10.009.

(207) Ting, L.; Rad, R.; Gygi, S. P.; Haas, W. MS3 Eliminates Ratio Distortion in Isobaric Multiplexed Quantitative Proteomics. *Nat Methods* **2011**, *8* (11), 937–940. https://doi.org/10.1038/nmeth.1714.

(208) Wang, H.; Alvarez, S.; Hicks, L. M. Comprehensive Comparison of ITRAQ and Label-Free LC-Based Quantitative Proteomics Approaches Using Two Chlamydomonas Reinhardtii Strains of Interest for Biofuels Engineering. *J. Proteome Res.* **2012**, *11* (1), 487–501. https://doi.org/10.1021/pr2008225.

(209) Cupp-Sutton, K. A.; Wang, Z.; Yu, D.; Wu, S. RPLC-RPLC-MS/MS for Proteoform Identification. In *Proteoform Identification: Methods and Protocols*; Sun, L., Liu, X., Eds.; Methods in Molecular Biology; Springer US: New York, NY, 2022; pp 31–42. https://doi.org/10.1007/978-1-0716-2325-1_4.

(210) Holman, J. D.; Tabb, D. L.; Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. *Curr Protoc Bioinformatics* **2014**, *46*, 13.24.1-13.24.9. https://doi.org/10.1002/0471250953.bi1324s46.
(211) Liu, X.; Segar, M. W.; Li, S. C.; Kim, S. Spectral Probabilities of Top-down Tandem Mass Spectra. *BMC Genomics* **2014**, *15* (1), S9. https://doi.org/10.1186/1471-2164-15-S1-S9.
(212) Révész, Á.; Hevér, H.; Steckel, A.; Schlosser, G.; Szabó, D.; Vékey, K.; Drahos, L. Collision Energies: Optimization Strategies for Bottom-up Proteomics. *Mass Spectrometry Reviews n/a* (n/a), e21763. https://doi.org/10.1002/mas.21763.
(213) Lantz, C.; Zenaidee, M. A.; Wei, B.; Hemminger, Z.; Ogorzalek Loo, R. R.; Loo, J. A. ClipsMS: An Algorithm for Analyzing Internal Fragments Resulting from Top-Down Mass Spectrometry. *J. Proteome Res.* **2021**, *20* (4), 1928–1935. https://doi.org/10.1021/acs.jproteome.0c00952.
(214) Durbin, K. R.; Skinner, O. S.; Fellers, R. T.; Kelleher, N. L. Analyzing Internal Fragmentation of Electrosprayed Ubiquitin Ions During Beam-Type Collisional Dissociation. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (5), 782–787. https://doi.org/10.1007/s13361-015-1078-1.
(215) Lyon, Y. A.; Riggs, D.; Fornelli, L.; Compton, P. D.; Julian, R. R. The Ups and Downs of Repeated Cleavage and Internal Fragment Production in Top-Down Proteomics. *J. Am. Soc. Mass Spectrom.* **2018**, *29* (1), 150–157. https://doi.org/10.1007/s13361-017-1823-8.
(216) Everley, R. A.; Huttlin, E. L.; Erickson, A. R.; Beausoleil, S. A.; Gygi, S. P. Neutral Loss Is a Very Common Occurrence in Phosphotyrosine-Containing Peptides Labeled with Isobaric Tags. *J Proteome Res* **2017**, *16* (2), 1069–1076. https://doi.org/10.1021/acs.jproteome.6b00487.
(217) Tucholski, T.; Knott, S. J.; Chen, B.; Pistono, P.; Lin, Z.; Ge, Y. A Top-Down Proteomics Platform Coupling Serial Size Exclusion Chromatography and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem.* **2019**, *91* (6), 3835–3844. https://doi.org/10.1021/acs.analchem.8b04082.
(218) Guo, Y.; Yu, D.; Cupp-Sutton, K. A.; Liu, X.; Wu, S. A Benchmarking Protocol for Intact Protein-Level Tandem Mass Tag (TMT) Labeling for Quantitative Top-down Proteomics. *MethodsX* **2022**, *9*, 101873. https://doi.org/10.1016/j.mex.2022.101873.
(219) Wang, Z.; Liu, X.; Muther, J.; James, J. A.; Smith, K.; Wu, S. Top-down Mass Spectrometry Analysis of Human Serum Autoantibody Antigen-Binding Fragments. *Sci Rep* **2019**, *9* (1), 2345. https://doi.org/10.1038/s41598-018-38380-y.
(220) McCool, E. N.; Lubeckyj, R. A.; Shen, X.; Chen, D.; Kou, Q.; Liu, X.; Sun, L. Deep Top-Down Proteomics Using Capillary Zone Electrophoresis-Tandem Mass Spectrometry: Identification of 5700 Proteoforms from the Escherichia Coli Proteome. *Anal. Chem.* **2018**, *90* (9), 5529–5533. https://doi.org/10.1021/acs.analchem.8b00693.
(221) Choi, S. B.; Zamarbide, M.; Manzini, M. C.; Nemes, P. Tapered-Tip Capillary Electrophoresis Nano-Electrospray Ionization Mass Spectrometry

for Ultrasensitive Proteomics: The Mouse Cortex. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (4), 597–607. https://doi.org/10.1007/s13361-016-1532-8.

(222) Savitski, M. M.; Reinhard, F. B. M.; Franken, H.; Werner, T.; Savitski, M. F.; Eberhard, D.; Martinez Molina, D.; Jafari, R.; Dovega, R. B.; Klaeger, S.; Kuster, B.; Nordlund, P.; Bantscheff, M.; Drewes, G. Tracking Cancer Drugs in Living Cells by Thermal Profiling of the Proteome. *Science* **2014**, *346* (6205), 1255784. https://doi.org/10.1126/science.1255784.

(223) Kellye. A. Cupp-Sutton, Zhe Wang, Dahang Yu, Si Wu. *Proteoform Identification Methods and Protocols Chapter 4: RPLC-RPLC-MS/MS for Proteoform Identification*, 1st ed.; 1064-3745.

(224) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat Biotechnol* **2012**, *30* (10), 918–920. https://doi.org/10.1038/nbt.2377.

(225) Fellers, R. T.; Greer, J. B.; Early, B. P.; Yu, X.; LeDuc, R. D.; Kelleher, N. L.; Thomas, P. M. ProSight Lite: Graphical Software to Analyze Top-down Mass Spectrometry Data. *PROTEOMICS* **2015**, *15* (7), 1235–1238. https://doi.org/10.1002/pmic.201400313.

Appendices

Appendix A – A Benchmarking Protocol for Intact Protein-Level Tandem Mass Tag (TMT) Labeling for Quantitative Top-Down Proteomics

ABSTRACT

Isobaric chemical tag labeling for quantification of intact proteins in complex samples is limited due to the tendency of intact proteins precipitate under labeling conditions and increased sample complexity as a result of side products (*i.e.*, incomplete labeling or labeling of unintended residues). To reduce precipitation under labeling conditions, we developed a technique to remove large proteoforms that allowed for the labeling and characterization of small proteoforms (< 35 kDa) using top-down proteomics. We also systematically optimized protein-level tandem mass tag (TMT) labeling conditions to obtain optimal labeling parameters for complex samples. Here, we present a benchmarking protocol for protein-level TMT labeling for quantitative top-down proteomics, including complex intact protein sample preparation, protein-level TMT labeling, top-down LC/MS analysis, and TMT reporter ion quantification.

- An optimized protocol for protein-level TMT labeling in complex sample
- Limits production of incorrectly labeled side products for minimization of spectral complexity.
- A guideline for isobaric chemical tag quantification in top-down proteomics

METHOD DETAILS

HeLa cell culture and cell lysate preparation

Supplies

1. Nunc™ EasYDish™ Dishes (Thermo Fisher, 150466)
2. Sterile pipettes

3. 50 mL falcon tubes (Corning, 430828)
4. Eppendorf Centrifuge 5804 R with rotors S-4-72 and FA-45-45-11

Chemicals

1. *HeLa* cell stock (stored at -80°C)
2. Eagle's minimal essential medium- high glucose (DEME) (Sigma-Aldrich: D6429)
3. Penicillin-streptomycin (Sigma-Aldrich: P4333)
4. Fetal bovine serum (FBS) (Fisher Scientific, SH3008704HI)
5. TrypLE™ Express Enzyme (1X), phenol red (Thermo Fisher Scientific, 12605-010)
6. Dulbecco's Phosphate buffered saline, 10X (Sigma-Aldrich, D1408)
7. 75% ethanol
8. Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA, 23225)

Procedures

HeLa cells were grown and lysed as described previously.[76] *HeLa* cells must be cultured in a biosafety cabinet under sterile conditions. Proper PPE should be worn.

Step 1: *HeLa* cell recovery

1. Mix DMEM with pen-strep and FBS to final concentration of 10% FBS and 2% penicillin-streptomycin. Then warm the DMEM mix in a water bath at 37 °C for around 1 hour. Thaw the frozen *HeLa* cells in a 37 °C water bath quickly.

2. Add 8 mL of the DMEM mixture into the cell culture plate, transfer the *HeLa* cells into the plate. Shake the plate gently on the bench.

3. Incubate the cells at 37 °C with 5% CO_2. Change the media every two days until the cell density is around 80-90% confluent under the microscope.

Step 2: HeLa cell passage

1. Mix DMEM with pen-strep and FBS to final concentration of 10% FBS and 2% penicillin-streptomycin. Then warm the DMEM mix, 1X PBS buffer, and TrypIE at 37 °C for around 1 hour.

2. Remove the old media from the plates. Add 5 mL 1X PBS buffer to the plate from the edge and shake gently on the bench, then remove the PBS using pipette.

3. Add 2.5 mL of TrypIE and incubate at 37 °C for around 10 minutes to dissociate the cells from the plate floor.

4. After 10-min TrypIE incubation, quench the reaction by adding 7.5 mL DMEM mix by moving the plate gently on the bench.

5. To passage 1 plate to 3 plates, prepare the new plates by adding 6.7 mL DMEM to each plate. Write down the cell line name, the passage number, operator name, and date on the plates for record.

6. Add 3.3 mL of parent cells to each new plate. Mix gently on the bench, then incubate at 37 °C and 5% CO_2 for 2 days.

7. Change the medium every two days by following steps 2-1 and 2-2, then add 10 mL warm DMEM mix to the plate.

Step 3: HeLa cell harvest

1. Remove old media from the plates with a pipette from the edge of the plate. Add 5 mL 1X PBS buffer (ice cold, pH = 7.4) to the plate from the edge and

shake gently on the bench. Remove the PBS using pipette. Repeat this step 2 additional times.

2. Add 2 mL 1X PBS buffer (ice cold, pH = 7.4) to each plate and scratch the cells off the plates into 50 mL falcon tubes.

3. Pellet and collect the cells through centrifugation at 4,200 rpm and 4 °C for 20 minutes.

4. Add 20 mL 1X PBS buffer (ice cold, pH = 7.4) to wash the cells again. Repeat this step 2 additional times.

5. Centrifuge the cells down at 2,000 rpm and room temperature for 5 minutes. Remove the supernatant and store the cells at -80 °C for future recovery or lysis.

Step 4: HeLa cell lysis

1. Resuspend *HeLa* cells into the lysis buffer (1 μM PMSF and 20 mM NaF in PBS, pH = 7.4).

2. Homogenize cells in a sonicator for 5 minutes and allow them to rest for 5 minutes, repeat this step two additional times.

3. Centrifuge the cell lysate at 15, 000 ×g, 4 °C for 30 minutes.

4. Evaluate protein concentration and patter using Pierce™ BCA Protein Assay Kit and SDS-PAGE. Aliquot *HeLa* protein samples and store at -80 °C for future use.

Protein-level TMT labeling

1. **Materials**
   1.1. Sartorius Vivaspin™ 20 Centrifugal Concentrator (100 kDa MWCO spin filter, Fisher Scientific, 14558509)

   1.2. Sartorius Vivaspin™ Turbo 15 PES Centrifugal Concentrator (10 kDa MWCO spin filter, Fisher Scientific, 14558651)

   1.3. Eppendorf Centrifuge 5804 R with S-4-72 rotor

2. **Chemicals**
   2.1. Triethylammonium bicarbonate (TEAB) buffer (pH = 8.5, Sigma-Aldrich, T7408)

   2.2. TMTsixplex isobaric label reagent (Thermo Fisher Scientific, 90061)

   2.3. Urea (Sigma-Aldrich, 208884)

   2.4. Iodoacetamide (IAA) (Sigma-Aldrich, I1149)

   2.5. 0.5M Tris(2-carboxyethyl) phosphine (TCEP) (Thermo Fisher Scientific, 77720)

   2.6. 50% hydroxylamine solution (Sigma-Aldrich, 467804)

   2.7. Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific, 23225)

3. **Procedures**

   3.1. **Sample preparation prior to TMT labeling**

      3.1.1. Rinse spin filters using labeling buffer (100 mM TEAB, pH=8.5).

      3.1.2. Remove large proteins (> 100 kDa) in *HeLa* cell lysate using 100 kDa MWCO filtration through centrifugation at 4200 RPM (3234 ×g) and 4 ºC for 20 minutes. Remove the filtrate and add ~

5 mL of the labeling buffer to wash the proteins and centrifuge at 3234 ×g and 4 ºC for 20 minutes. Remove the filtrate again and combine with previous filtrate. Repeat rinsing 3-5 times.

3.1.3. Combine all the filtrate from Step 3.1.2 and concentrate using a 10 kDa MWCO spin filter through centrifugation at 3234 ×g RPM and 4 ºC until the protein concentration is ~ 2 μg/μL. Measure protein concentration using BCA assay.

3.1.4. Mix the concentrated protein sample (800 μg) with 6 M urea at equal volume for protein denaturation. Final concentration of urea is ~3 M and protein concentration is ~1 μg/μL.

3.1.4.1. Note: Protein mass (800 μg) may be adjusted based on experimental design, but concentration should remain consistent for all samples.

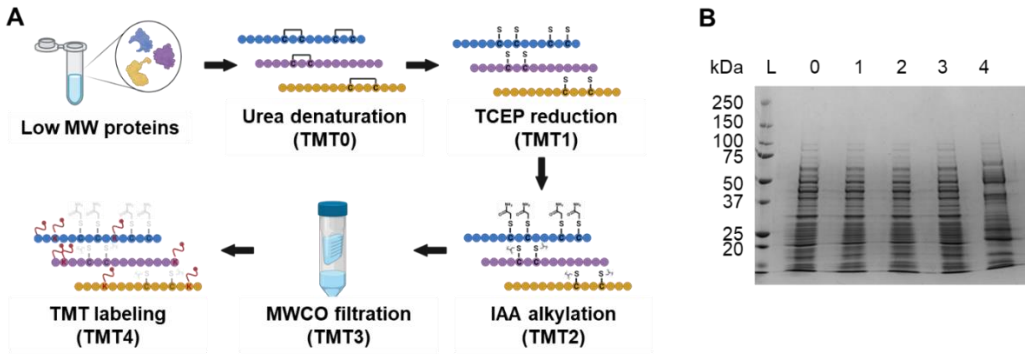
3.1.4.2. Note: Urea should be prepared fresh.

3.1.5. Add 160 μL of 0.5 M TCEP to the denatured protein solution (800 μg) to reduce disulfide bonds. Incubate for 15 minutes at room temperature and then add 215 μL (375 mM) IAA at room temperature. Incubate this solution in the dark for 30 minutes.

3.1.5.1. Note: IAA is light sensitive and should be prepared in labeling buffer and wrapped with aluminum foil to prevent decomposition.

3.1.6. Buffer exchanges the alkylated protein solution to 100 mM TEAB buffer (pH=8.5) and concentrates to ≥ 1 µg/µL using a 10 kDa MWCO spin filter.

3.1.6.1. Note: A small portion (~10 µg) of proteins after each step (urea denaturation, TCEP reduction, IAA alkylation, MWCO buffer exchange) may be removed and analyzed using SDS-PAGE to observe sample consistency (***Figure S1B: Sample 1-4***).



Supplementary Figure 1. Protein-level TMT labeling sample preparation. (A) Overall complex protein sample preparation workflow. (B) SDS-PAGE analysis of protein samples taken from each step (L: standard protein ladder); TMT0: urea denaturation; TMT1: TCEP reduction; TMT2: IAA alkylation; TMT3: MWCO buffer exchange; TMT4: TMT labeling. Figure was created in BioRender.com.

3.2. **TMT labeling of intact complex protein mixture.**

3.2.1. Centrifuge TMT reagent tubes for 1 minute with a benchtop centrifuge, add 41 µL of ACN to each tube, and vortex the tubes briefly to dissolve TMT. Then centrifuge the tubes briefly using a benchtop centrifuge to bring solution down to the bottom of the tube.

3.2.2. Mix 50 μg of protein sample produced in Step 3.1.6 (concentration ≥ 1 μg/μL) with 200 μg TMT reagent in 10 μL ACN. Incubate the sample at room temperature for 1 hour, then quench the reaction by addition of 5% hydroxylamine to a final concentration of > 0.3% (final pH > 9.1) for 15 minutes at room temperature.

3.2.2.1. Note: Protein mass may be adjusted based on experimental design; however, it is recommended that the protein amount be no less than 20 μg for good labeling efficiency.

3.2.2.2. Note: If the protein concentration is lower than 1 μg/μL, double labeling can be applied. After the first 1-hour reaction, add the same amount of TMT reagent (200 μg in 10 μL) to the solution and incubate for 1 additional hour, then quench the reaction using hydroxylamine to a final concentration of > 0.3% (final pH > 9.1).

3.2.3. Analyze TMT-labeled samples by SDS-PAGE to make sure there is no obvious sample loss after TMT labeling (***Figure 1B: Sample 4***).

3.2.4. Centrifuge the TMT-labeled sample at 12,000 × g and 4 ºC for 30 minutes before LC-MS/MS analysis to remove any possible precipitation.

4. Top-down RPLC-MS/MS analysis

4.1. **Supplies**

4.1.1. Home-packed C5 trapping column (150 μm i.d., 5 cm length, Jupiter particles, 5 μm diameter, 300 Å pore size)[75,219]

4.1.2. Home-packed C5 RPLC capillary column (75 μm i.d., 70 cm length, Jupiter particles, 5 μm diameter, 300 Å pore size)[75,219]

4.1.3. Modified Thermo Scientific (Waltham, MA, USA.) Accela LC system[179,180]

4.1.4. Thermo Orbitrap Exploris 240 mass spectrometer (Thermo Fisher Scientific, Bremen, Germany)

4.2. **Chemicals**

4.2.1. LC-MS grade acetonitrile (ACN) (Sigma-Aldrich, 34998)

4.2.2. LC-MS grade iso-propanol (IPA) (Sigma-Aldrich, 34863)

4.2.3. Pierce™ Trifluoroacetic Acid (TFA), LC-MS Grade (Thermo Scientific, 85183)

4.2.4. Acetic acid, glacial, ACS reagent, ≥99.7% (Sigma-Aldrich, 695092)

4.2.5. LC-MS grade HPLC water (Sigma-Aldrich, 270733)

4.3. **Procedures**[130,132]

4.3.1. Mobile phase A (MPA) is made up of 0.01% TFA, 0.585% acetic acid, 2.5% isopropanol, and 5% acetonitrile in water. To prepare 1 L of MPA, add 50 mL ACN, 25 mL IPA, and 920 mL HPLC water to a 1000 mL media storage bottle. Then, add 5.85 mL acetic acid and 0.1 mL TFA to the solution using a glass

pipette. Sonicate the buffer for 15 minutes before use and stored at room temperature.[223]

4.3.1.1. Note: MPA should be prepared in the hood. Gloves, goggles, and lab coat should be worn while preparing buffer.

4.3.1.2. Note: Trifluoroacetic acid (TFA) may react with oxygen containing plastics and should only be transferred using glass pipettes. Additionally, TFA is a corrosive chemical that can irritate the throat, nose, and lungs and may also burn the skin and eyes. Therefore, proper PPE should be worn while handling TFA.

4.3.2. Mobile phase B is made up of 0.01% TFA, 0.585% acetic acid, 45% IPA, and 45% ACN in water. To prepare 1 L of MPB, add 450 mL ACN, 450 mL IPA, and 95 mL HPLC water to a 1000 mL media storage bottle. Then, add 5.85 mL acetic acid and 0.1 mL TFA to the solution using a glass pipette. Sonicate the buffer for 15 minutes before use and store at room temperature.

4.3.2.1. Note: MPB should be prepared in the hood. Gloves, goggles, and lab coat should be worn while preparing buffer.

4.3.2.2. Note: Trifluoroacetic acid (TFA) may react with oxygen containing plastics and should only be transferred using glass pipettes. Additionally, TFA is a corrosive chemical that can irritate the throat, nose, and lungs and may

also burn the skin and eyes. Therefore, proper PPE should be worn while handling TFA.

4.3.3. Dilute the TMT-labeled protein sample with HPLC water to a final organic percentage below 10%. Then load 10 μg of the diluted TMT-labeled protein sample onto a home-packed trapping column with a flow rate of 5~8 μL/min for 20 minutes.

4.3.4. Proteins bound on trapping column are then eluted onto a C5 RPLC capillary column using a modified Thermo Scientific (Waltham, MA, USA.) Accela LC system for seperation.[179,181] A 200-min gradient from 10% to 70% of MPB is applied for protein separation at a flow rate of 400 nL/min. The LC eluent is analyzed by an Orbitrap Exploris 240 mass spectrometer in positive mode (Thermo Fisher Scientific, Bremen, Germany) using a customized nano-ESI interface.[219]

4.3.5. Gradient time can be adjusted according to sample requirements.

4.3.6. Parameters used for Orbitrap Exploris 240 mass spectrometer have been previously reported.[130] Briefly, the temperature of the inlet capillary is set to 275 °C and the spray voltage is 2.6 kV. The resolution of full MS scans (500-2000 m/z) is set to 120,000 with three micro scans. The top 6 most abundant precursor ions (charge 4-50) are selected for MS/MS fragmentation. The resolution of MS/MS scans (100-1600 m/z) is

set to 120,000 with two microscans, 2.0 m/z as the isolation window, and 20 s as the dynamic exclusion. Higher-energy collisional dissociation (HCD) with a normalized energy as 35% is performed for MS/MS fragmentation. The maximum injection time is set to 1000 ms for full mass scans and 500 ms for MS/MS scans. The AGC target is set to 3E6 for full mass scans and 1E6 for MS/MS scans.

5. Data analysis

5.1. **Software**

5.1.1. MSconvert[224]

5.1.2. TopPIC Suite[182] (includes TopFD and TopPIC)

5.1.3. Python 3.8

5.1.4. ProSight Lite[225]

5.2. **Procedures**

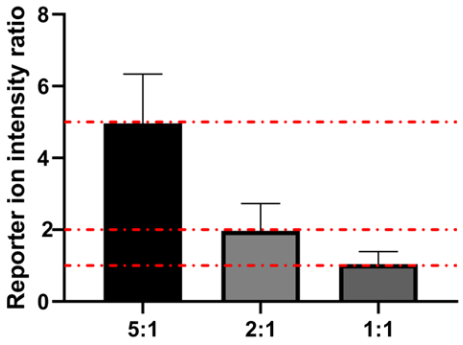
5.2.1. **Identification:** Convert all MS raw files to mzML format using MSconvert.[224]

5.2.1.1. Set the TopFD[182] deconvolution with parameters as follows: MS1 signal-to-noise ratio as 3, MS2 signal-to-noise as 1, the precursor window size as 3.0 *m/z,* maximum mass as 50,000 Dalton, maximum charge as 30, and m/z error as 0.02.

5.2.1.2. TopPIC[182] is utilized for identification against the annotated Human protein database (UniProt 2021-05-14,

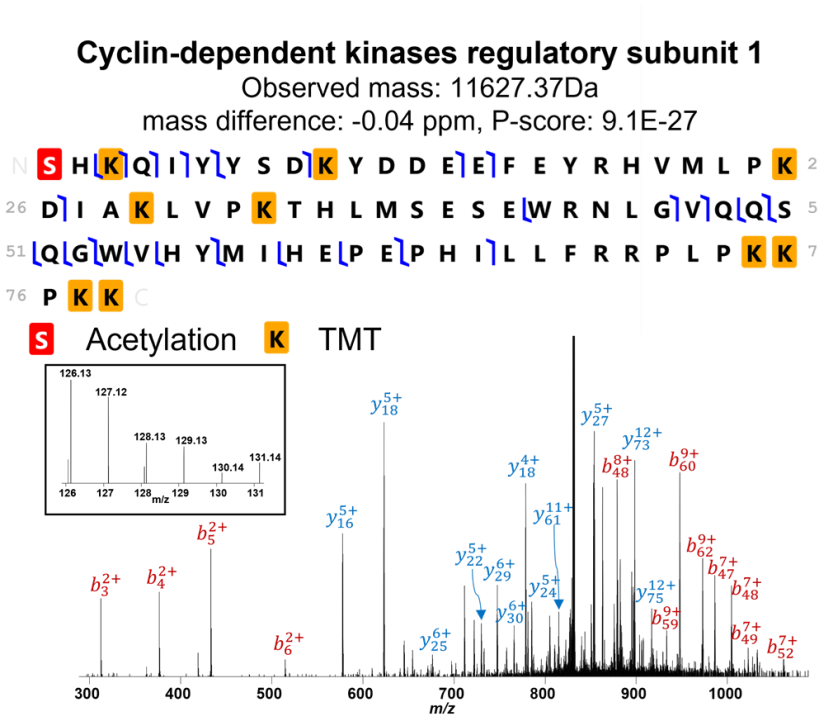
20380 species). Parameters for TopPIC identification are decoy database searching with FDR = 0.01 for spectrum and proteoform level. The maximum number of mass shifts is 2 and the mass shift range is ± 500 Dalton. TMT at both N-terminal and lysine residues were set as fixed modifications. Acetylation, phosphorylation, oxidation, methylation and carbamidomethyl were set at dynamic PTMs. Parameters not given here are default.

5.2.2. **TMT Quantitation:** To evaluate the protein-level TMT labeling quantification, separate the buffer-exchanged protein sample Step 3.1.6 to three aliquots and label individually with four tags from TMTsixplex (127, 128, 130, and 131) (, then mixed with the mass ratio as 5:2:1:1 for LC-MS/MS analysis. Extract the intensities of the reporter ions for each scan from converted mzML files using the in-house python coding described previously.[130] Calculate and compare the intensity ratios (127/131, 128/131, 130/131) among samples with the theoretical ratios (5:2:1) (Example in **Figure S2**). Here, the observed ratio is5.00 ± 1.37, 1.97 ± 0.76, 1.05 ± 0. 34, which is close to theoretical ratio as 5:2:1))[132], suggesting accurate quantification as expected.



Supplementary Figure 2. The bar graph shows the normalized intensity ratio of TMT-labeled *HeLa* protein after optimization (127/131, 128/131 and 130/131). The error bars were the calculated standard deviation, and the red dashed lines represent the theoretical ratio as 5:2:1.

5.2.3. **Data visualization:** Protein identification can be confirmed and visualized using ProSight Lite.[225] Load the protein mass list, precursor monoisotopic mass, and sequence obtained for the protein of interest from TopPIC into ProSight to visualize fragmentation. Appropriate PTMs or TMT tags can be manually placed on the amino acid residues to obtain the best fragmentation patterns. As shown in ***Figure 3***, cyclin-dependent kinases regulatory subunit 1 (Uniprot ID: P61024) from TMT-labeled *HeLa* cell lysate demonstrates there is an acetylation at the N-terminal and a TMT tag on each lysine residues. Quantification results can be obtained from the relative intensity of the reporter ions.



Supplementary Figure 3. Fragmentation and post-translational modifications observed on cyclin-dependent kinases regulatory subunit 1 (Uniprot ID: P61024) created by ProSight Lite. Red highlight represents acetylation; yellow highlights represent TMT tag labeling.

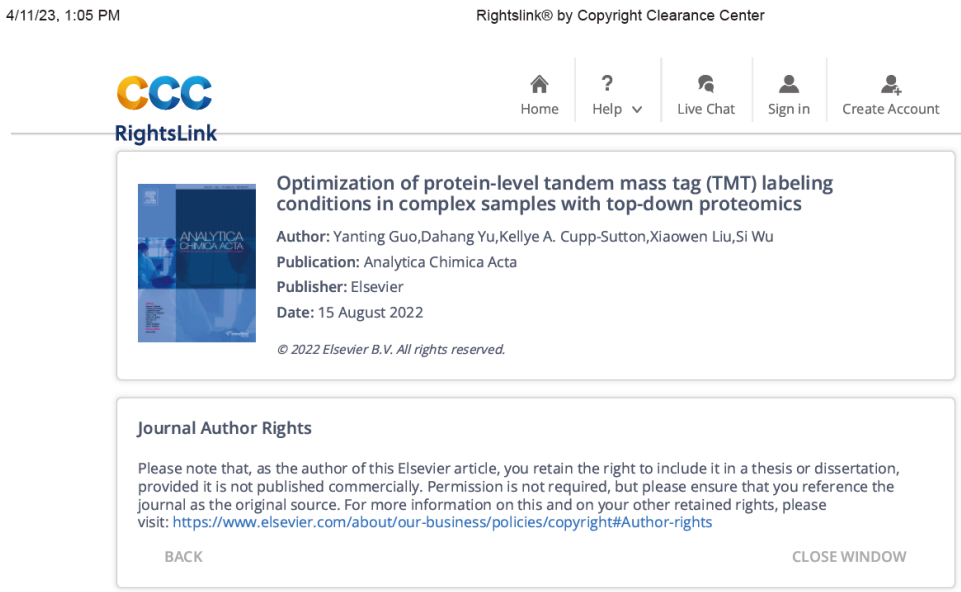
*The materials in ***Appendix A*** are adapted from:

Yanting Guo, Dahang Yu, Kellye A. Cupp-Sutton, Xiaowen Liu, and Si Wu. "A benchmarking protocol for intact protein-level Tandem Mass Tag (TMT) labeling for quantitative top-down proteomics." MethodsX 9 (2022): 101873.

Author contributions: Yanting Guo: Methodology, Data curation, Formal analysis, validation, Writing - original draft. Dahang Yu: Methodology, validation. Kellye A. Cupp-Sutton: Validation, Writing - review & editing. Xiaowen Liu: Software, Validation. Si Wu: Conceptualization, Validation, Supervision, Funding acquisition, Writing – review & editing.

Copyright information for Appendix A:

Appendix B – Copyright licence for Chapter 2



Appendix C – Copyright licence for Chapter 3

4/11/23, 1:01 PM Rightslink® by Copyright Clearance Center

RightsLink

Home | Help | Live Chat | Sign in | Create Account

ACS Publications
Most Trusted. Most Cited. Most Read.

Optimization of Higher-Energy Collisional Dissociation Fragmentation Energy for Intact Protein-Level Tandem Mass Tag Labeling

Author: Yanting Guo, Trishika Chowdhury, Meena Seshadri, et al
Publication: Journal of Proteome Research
Publisher: American Chemical Society
Date: Jan 1, 2023

Copyright © 2023, American Chemical Society

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

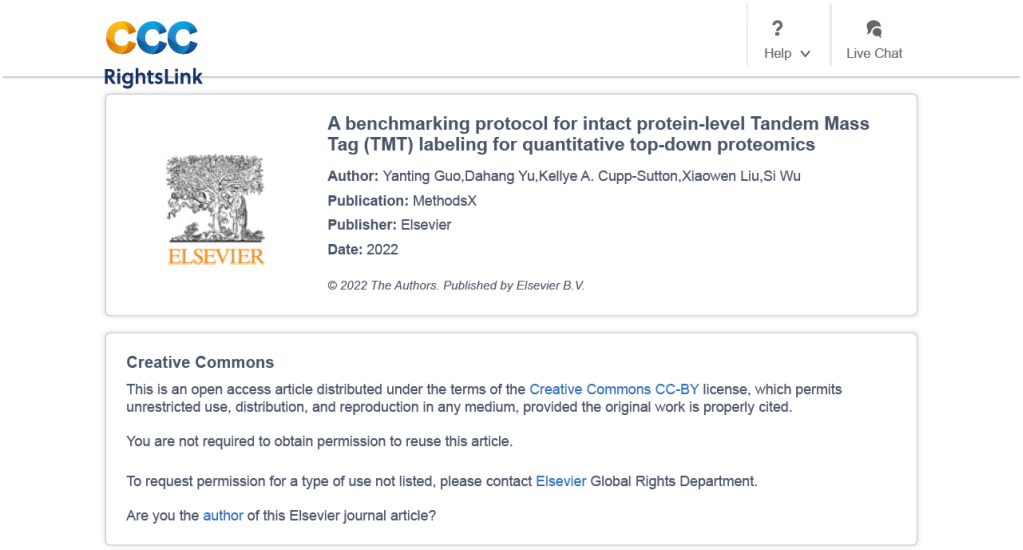
This type of permission/license, instead of the standard Terms and Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from {COMPLETE REFERENCE CITATION}. Copyright {YEAR} American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your RightsLink request. No additional uses are granted (such as derivative works or other editions). For any uses, please submit a new request.

If credit is given to another source for the material you requested from RightsLink, permission must be obtained from that source.

BACK CLOSE WINDOW

Appendix D – Copyright licence for Appendix A



RightsLink

Help Live Chat

A benchmarking protocol for intact protein-level Tandem Mass Tag (TMT) labeling for quantitative top-down proteomics

Author: Yanting Guo,Dahang Yu,Kellye A. Cupp-Sutton,Xiaowen Liu,Si Wu

Publication: MethodsX

Publisher: Elsevier

Date: 2022