UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DEVELOPMENTS OF HIGH-THROUGHPUT QUANTITATIVE TOP-DOWN PROTEOMICS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

DAHANG YU
Norman, Oklahoma
2021

DEVELOPMENTS OF HIGH-THROUGHPUT QUANTITATIVE TOP-DOWN PROTEOMICS

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY THE COMMITTEE CONSISTING OF

Dr. Si Wu, Chair

Dr. Richard Broughton

Dr. Christina Bourne

Dr. Rakhi Rajan

Acknowledgments

First and foremost, I sincerely appreciate my advisor, Dr. Si Wu. She not only taught me the knowledge about proteomics and MS but also showed me how to be a rigorous and thoughtful scientist. Her care and patience deeply affected my teaching style to undergraduates and junior lab members. I am so proud to be a member of this group, the time spent in the Wu Lab is very meaningful to me, I will miss it in the days to come.

I would like to appreciate my committee members, Dr. Rakhi Rajan, Dr. Christina Bourne, and Dr. Richard Broughton. Thank you very much for your support and guidance over years.

I am grateful to all the lab members. Zhe Wang taught me how to build up my first LC-MS system and how to maintain it hand in hand. Dr. Kellye Cupp-Sutton gave me lots of useful advice on both manuscript writing and rebuttal arguing. Lushuang Huang and Mulin Fang were helping with the MS cleaning and instrument setup. Yanting Guo worked with me together on the TMT project in Chapter 4. I also thank Daniel G. Dalefield, Toni Woodard, Bo Zhao, Siyu Wu, Ji Kang, Josh Li, Trishika Chowdhury, and Anju Teresa Sunny.

Finally, I would thank my family for their love and support. I am glad they understand my choice to study abroad.

Table of Contents

List of Figures

Key to abbreviations

2D-PAGE	two-dimensional polyacrylamide gel electrophoresis
CE	capillary electrophoresis
CID	collision induced dissociation
cTnl	cardiac troponin I
cyro-EM	cryoelectron microscopy
CZE	capillary zone electrophoresis
DiLeu	N, N-dimethylleucine
DMEM	Dulbecco's modified Eagle medium
E. coli	*Escherichia coli*
ECD	electron capture dissociation
EOF	electroosmotic flow
ESI	electrospray ionization
ETD	electron transfer dissociation
EThcD	electron transfer/higher-energy collisional dissociation
FT-ICR	Fourier-transform ion cyclotron resonance
GELFrEE	gel-eluted liquid fraction entrapment

HCD	high-energy collision dissociation
HILIC	hydrophilic interaction chromatography
HRMS	high-resolution mass spectrometry
IEC	ion-exchange chromatography
IMAC	immobilized metal affinity chromatography
iTRAQ	isobaric tags for relative and absolute quantification
LC	liquid chromatography
LFQ	label-free quantification
m/z	mass-to-charge ratio
MPA	mobile phase A
MPB	mobile phase B
MS	mass spectrometry
MW	molecular weight
NeuCode SILAC	neutron encoding stable isotope labeling in cell culture
pI	isoelectric points
PMSF	phenylmethylsulfonyl fluoride
PTM	post-translational modification

Q-TOF	quadrupole-time-of-flight
RPLC	reversed-phase liquid chromatography
SEC	size exclusion chromatography
sIEF	solution isoelectric focusing
SILAC	stable isotope labeling in cell culture
STM	*Salmonella Typhimurium*
TIPMI	tunable intact mass protein mass increases
TMT	tandem mass tags
TOF	time-of-flight
UPLC	ultra-high-pressure liquid chromatography
UVPD	ultraviolet photodissociation
WCX	weak cation exchange

Abstract

The LC-MS techniques are commonly used to analyze intact proteoforms for top-down proteomics. To deepen the coverage of the intact human proteome, multidimensional separations are often applied prior to the MS analysis. However, the most common quantitative top-down approach, label-free quantification, cannot be applied to multidimensional separations because proteins that elute in multiple fractions cannot be quantified. The focus of this dissertation is the development of high-throughput quantitative top-down proteomics techniques for the analysis of complex biological samples (*e.g.*, bacteria and human cell lysate) using isobaric chemical tag labeling for application to multidimensional separation and quantitative top-down proteomics.

To improve the intact proteome coverage of human protein samples, we applied a 2D pH RP/RPLC-MS platform to characterize intact human proteoforms. The high resolution of RPLC and orthogonality between high-pH and low-pH RPLC separations provide an aerial view of the intact human proteome. Our results demonstrated that the 2DLC platform dramatically improves the identification of intact proteins and proteoforms, and largely enhances the detection of low abundant proteoforms and PTMs.

The isobaric chemical tag labeling quantification has been widely applied in bottom-up proteomics. However, protein-level labeling is challenging due to the tendency of intact proteins to aggregate under labeling conditions. We developed a filter-SEC approach to enrich low MW proteins from the complex sample, these proteins maintain solubility in protein-level TMT labeling for quantitative top-down MS analysis.

Once the protein aggregation issue had been addressed, however, we found that protein-level TMT labeling under standard labeling conditions resulted in unwanted side products (underlabeled or overlabeled species). This issue is more serious for intact proteoform labeling because a higher number of residues are potentially labeled when compared with shorter peptides. We comprehensively optimized and evaluated the protein-level TMT labeling conditions including TMT to protein ratio, final pH (quenching solution concentration), reaction concentration (starting sample volume), reaction time, and reaction buffers. After optimization of parameters, we found that 85% of the proteoforms were completely labeled after the reaction; this improvement in labeling efficiency allows protein-level isobaric chemical tag labeling to be used for real applications.

Overall, our results demonstrate the potential of quantitative top-down proteomics using TMT and multidimensional separation. High throughput quantitative top-down proteomics techniques using protein-level TMT labeling hold great potential for the relative quantification of intact proteoforms. We hope that our work can push the isobaric chemical tag labeling quantification for intact proteoforms into practical applications.

Chapter 1 Introduction

1.1 Proteomics

Proteins are macromolecules composed of amino acid residues. They participate in every process in cells, and they work as essential elements in biological systems. The functions of proteins are diverse due to the many unique structures of proteins, thus, studying proteins is a key point to understand the functions of biological systems. The proteome, which includes all proteins expressed by the biological system, is extremely complex due to genetic variation, alternative splicing, and post-translational modifications (PTMs).[1] (**Figure 1.1**) In fact, many proteoforms, different forms of a protein expressed from the same gene, may be generated by a single cell which further complicates the proteome.[2]

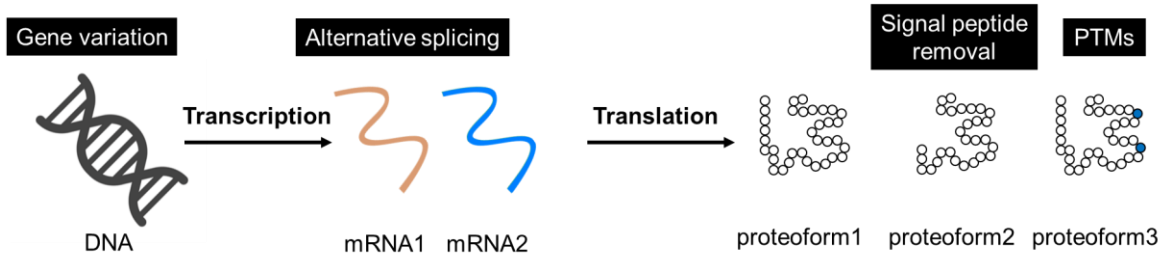


Figure 1.1 Various proteoforms generated from 1 protein-coding gene

Proteomics, the large-scale study of all expressed proteins, has enabled investigation into the proteins expressed in a complex system.[3] There are various biochemical techniques to study proteins and proteoforms, but most of these techniques require relatively large amounts of pure protein and cannot analyze multiple proteins in a complex mixture. For example, western blot has been used to detect specific proteins

in a complex biological sample, but it requires known protein-antibody binding.[4] X-ray crystallography can provide detailed protein crystal structure information but purified protein and high-quality protein crystals are necessary.[5] These challenges are partially addressed by the development of cryoelectron microscopy (cyro-EM).[6] Compared to the methods mentioned above, proteomics can study complex protein mixtures and provide the analysis of multiple proteins within the mixture. The central dogma indicates that DNA encodes the instructions for transcription into RNA, and this then holds the information for translation to proteins. However, the proteome is dynamic and responds to changes in the biological environment. This results in a proteome that has a higher complexity than the genome. To effectively study the proteome, a two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) was applied to separate proteins based on molecular weight (MW) and isoelectric points (pI); then the proteins were analyzed using mass spectrometry.[7] The separation prior to MS analysis reduced the complexity of the protein sample, which enhanced the analysis of individual protein species by MS. This technique provided a proof of concept for the characterization of multiple proteins in complex samples, despite the time-consuming and low throughput nature of 2D-PAGE. Subsequently, high-throughput approaches to analyze complex protein samples have been extensively studied.

1.2 High-throughput top-down proteomics

Different approaches based on upstream sample preparation, bottom-up and top-down proteomics, are applied to study the protein samples (**Figure 1.2**). For bottom-up proteomics, protein samples are digested to peptides for MS analysis. Trypsin is commonly utilized in bottom-up proteomics for two reasons:[8] (1) tryptic digested

peptides have an average molecular weight (MW) of 600-1000 Da and these shorter peptides are easier to analyze using MS compared to intact protein molecules. (2) These peptides are detected using MS under positive ion mode because they have more than 2 positive charges attached to amine groups, including N-terminus, Lys, and Arg residue. These multiply charged peptides can be selected for fragmentation specifically by the tandem MS. However, the bottom-up sample preparation is tedious and the limited sequence coverage from peptides cannot always provide comprehensive information regarding intact proteoforms.

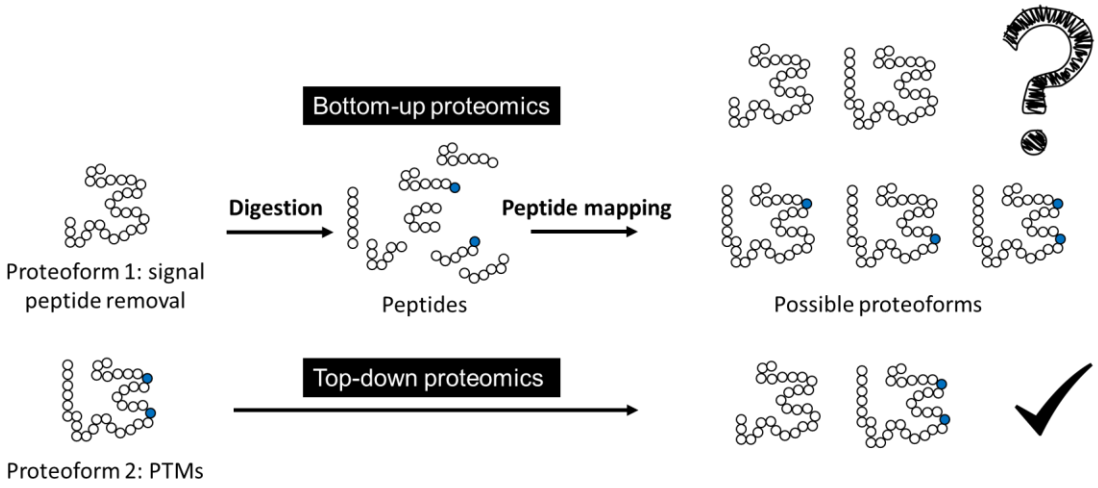


Figure 1.2 Bottom-up vs. top-down proteomics.

In bottom-up proteomics, the proteins are digested to peptides and analyzed using MS. In top-down proteomics, the proteins are directly analyzed using MS.

The top-down proteomics approach surpasses the bottom-up proteomics in terms of analyzing intact proteoforms, though the MW of the detected proteoform is limited by current MS technology. The workflow of top-down proteomics does not require a digestion step, the intact proteins are directly ionized and introduced into the MS for intact mass measurement. Sequence information is acquired through the activation and fragmentation of intact precursor ions using MS/MS techniques. (**Figure 1.3**) Therefore,

top-down MS can analyze unique proteoforms that have altered primary sequence and/or PTMs.

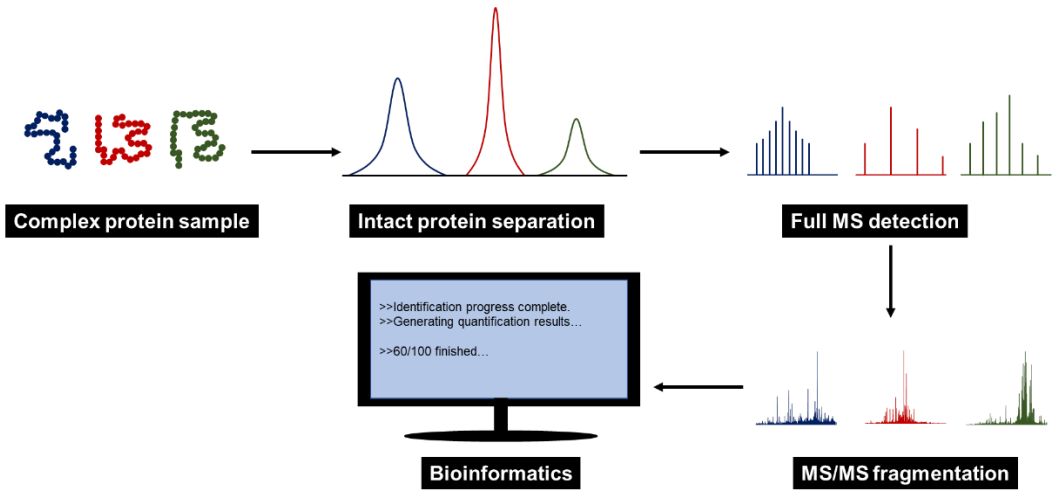


Figure 1.3 Typical top-down proteomics workflow.

1.2.1 Separation methods for top-down proteomics

Top-down proteomics is a standard approach to analyze intact proteoforms however, improvements are still needed to address the most severe challenges to the technique. These challenges include high complexity and a large dynamic range of complex biological protein samples. For example, it has been estimated that the human proteome contains over 20,000 protein-coding genes and more than 1 million proteoforms [1], and the human plasma contains an enormous dynamic range for a protein concentration of 11 orders of magnitude.[9] In complex sample analysis, high sample complexity increases the complexity of the mass spectra which complicates data analysis. Furthermore, the signal from high abundant proteins can suppress low abundance proteoform signals limiting the proteome coverage. To tackle these challenges, powerful protein enrichment and separation techniques are required before MS analysis of complex samples. Several options for protein enrichment exist; for example,

phosphoproteins can be enriched using immunoprecipitation and immobilized metal affinity chromatography (IMAC),[10] and antibodies can be enriched from serum using ammonium sulfate precipitation.[11] Alternatively, samples can be fractionated so that the sample complexity of each fraction is reduced. One example of a fractionation method is gel-eluted liquid fraction entrapment (GELFrEE), a technique to separate the proteins based on MW. GELFrEE has been utilized extensively in top-down proteomics since it can enrich the low MW proteins suitable for intact protein MS analysis.[12-15] Solution isoelectric focusing (sIEF) is another fractionation technique that separates the proteins based on pI.[13, 14]

Different LC methods have been utilized to fraction complex protein samples based on specific binding mechanisms, including reversed-phase liquid chromatography (RPLC) [16-18], size-exclusion chromatography (SEC) [19-21], hydrophilic interaction chromatography (HILIC)[22], and ion-exchange chromatography (IEC)[23]. Since the development of electrospray ionization (ESI), some of these LC approaches can be coupled online with MS. Further, the development of the autosampler has allowed the sample injection, separation, and detection to be automated for convenient LC-MS analysis. Among these methods, RPLC is the most commonly applied method for intact protein separation with MS detection.[24] RPLC separation achieves high peak capacities for faster chromatographic partitioning [25] and low salt buffers provide clean ions for MS.[26, 27] RPLC-MS has been previously utilized to analyze complex protein samples and has been reported to identify 1,665 proteoforms of 563 proteins from the *Salmonella Typhimurium* (STM) proteome using an ultra-high pressure RPLC-MS.[28] Capillary electrophoresis (CE) has also been utilized as a separation method for intact proteins, Llubeckyj et al. identified about 600 proteoforms of 200 proteins using capillary zone electrophoresis (CZE) from *E. coli* cell lysate.[29] CE provides high theoretical plate

numbers and efficient separation, however, CE application in top-down proteomics has previously been limited by the small sample loading amount, limited separation window, and poor reproducibility.[29, 30]

These LC approaches, however, offer limited coverage of biological proteomes. (**Figure 1.4**) To increase the number of proteins and proteoforms identified, various separation techniques have been combined for multidimensional separation. Separation techniques that utilize different properties of proteins (*e.g.*, MW, charge, hydrophobicity, affinity, etc.) are utilized as separate dimensions for separation.[31] These combined orthogonal separations can improve total peak capacity to improve protein identification.[32] The size-based separations, SEC and GELFrEE, are popular as first-dimension since they can enrich low MW proteins suitable for top-down MS detection or remove the low MW molecules to achieve high MW protein analysis. GELFrEE has a higher resolution separation than SEC, but detergent in the GELFrEE buffer makes the sample handling strenuous.[33] The Kelleher group has coupled CZE with a Q-Exactive mass spectrometer to analyze GELFrEE fractions from *P. aeruginosa* PA01 cell lysate to identify 30 proteins between 30-80 kDa.[12] Further, this same group developed native GELFrEE using gentle detergents to maintain natural protein interactions during separation.[33] SEC can be performed using native separation conditions but has low resolution compared with some other methods. The Ge group developed a serial-SEC (sSEC) fractionation technique to improve SEC resolution and they detected high MW proteins up to 223 kDa with a quadrupole-time-of-flight (Q-TOF) mass spectrometer.[20]

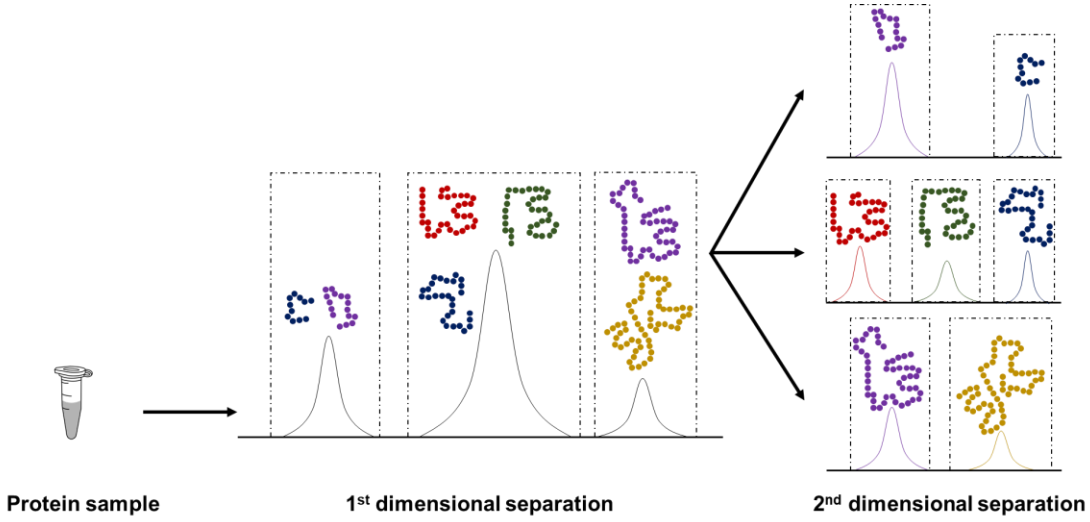


Figure 1.4 Co-eluted proteins on 1st dimension are separated on 2nd dimension

RPLC is popularly used for top-down, MS-based proteomics because it has great separation power and uses MS-friendly buffers; therefore, RPLC is widely used for second-dimensional separation.[14, 16, 26, 34] For example, Cleland et al. coupled GELFrEE and RPLC to separate intact human proteins for MS analysis.[15] Coupling high-resolution separation techniques with lower resolution techniques, however, limits the overall separation efficiency of a 2D separation. Thus, the change in hydrophobicity of intact proteins under different pH conditions has been utilized to combine RPLC for both dimensions of a 2D separation. The hybridization of high pH and low pH RPLC allows comprehensive intact proteoform coverage. Our group developed a 2D high-pH and low-pH RPLC/RPLC-MS approach for top-down proteomics, which not only enhanced the low-abundant proteoform identification but also detected rare PTMs from *E. coli* and *HeLa* cell lysate.[16, 18, 26] Three-dimensional separations are also used to further improve the coverage for complex samples. The Sun group developed a 3-dimensional platform, combining SEC, RPLC, and CZE that allowed 5,700 proteoforms identified from the *E. coli* proteome.[35] Furthermore, the Ge group combined IEC, HIC, and RPLC into a 3-

dimensional platform for intact protein analysis.[34]

Although the offline platforms have a larger sample capacity, they suffer from labor-intensive handling and possible sample loss and contamination.[36] The two-dimensional separations are designed for an automated online approach to reduce the total sample amount. The Paša-Tolić group presented an online weak cation exchange (WCX)-HILIC to RPLC approach which identified hundreds of *HeLa* core histone proteoforms starting from 1.5 μg material.[23] Our group demonstrated that the automatic, online 2D pH RP/RPLC system can reduce sample amounts from 500 μg to 5μg and identify more than 1000 intact proteoforms.[18]

1.2.2 MS methods for top-down proteomics

Another challenge for top-down proteomics is the detection and identification of intact proteoforms on a proteome-wide level.[24] High-resolution and mass accuracy are important to study the intact proteoforms for small, PTM induced mass shifts.[37] High sensitivity is also critical in top-down proteomics due to the high molecular mass of intact proteins and the resultant signal distribution into more isotopic peaks and charge states.[38] The commercial high-resolution mass spectrometers, such as Orbitrap, time-of-flight (TOF), and Fourier transform ion cyclotron resonance (FTICR), make it feasible to analyze intact proteoforms and PTMs in top-down proteomics. The Orbitrap and FTICR have both high sensitivity and high mass accuracy, which enables the detection of intact proteoforms and more confident identification with mass fragmentation. The TOF instruments, however, provide a higher m/z range and are commonly used to characterize high MW proteins.[20]

Various dissociation methods are applied to aid in the identification of intact proteoforms with specific fragmentation mechanisms resulting in unique types of

peptide fragmentation. (**Figure 1.5**) Collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD) are commonly utilized in top-down proteomics. In these collision-based methods, ions are accelerated by electrical potential and collide with gas molecules, which result in peptide bonds being cleaved and the production of b- and y- ions. (**Figure 1.5**) HCD occurs in the multiple collision cell and is capable of low-mass fragmentation; thus, HCD can be used in quantitative proteomics with isobaric chemical tag labeling.[39] However, collision-based methods cleave the most labile bonds first so the sequence coverage and PTM localization are limited.[40] Electron capture dissociation (ECD) and electron transfer dissociation (ETD) are other fragmentation methods applied to top-down proteomics. In ECD, the multiply-charged analyte cations (precursor ions) react with low-energy electrons to cleave the peptide bonds.[41] The cation radical is generated by an anion radical transferring to the analyte cation in ETD.[40] Electron-based methods are suitable for highly positively-charged precursors; however, the dissociation efficiency is suppressed by inducing negative charges.[42] Electron transfer/higher-energy collisional dissociation (EThcD) is a sequential combination of ETD and HCD that improves sequence coverage because complementary fragment ions are generated.[43] Activated-ion ETD (AI-ETD) generates ions with infrared photons during the reaction and has been reported to improve sequence coverage of intact proteoforms with PTMs.[44] Ultraviolet photodissociation (UVPD) utilizes high-energy UV photons to activate and dissociate ions and exhibits impressive performance for protein sequence coverage and PTMs characterization.[15]

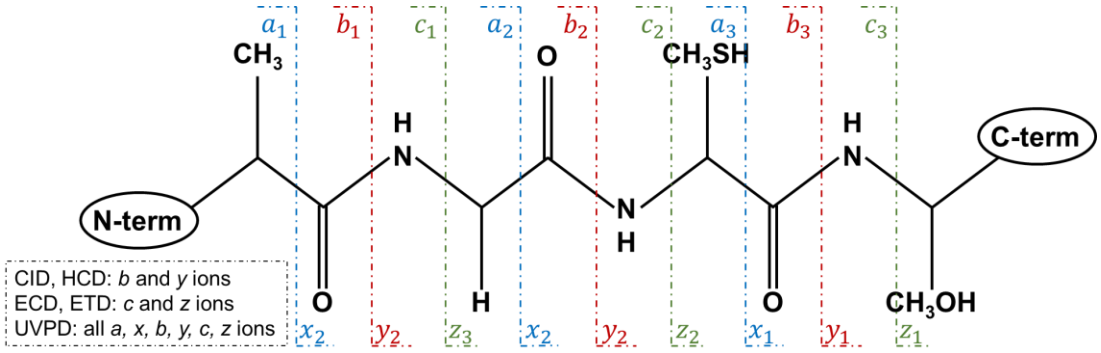


Figure 1.5 Various dissociation techniques result in different ions

1.3 Quantitative top-down proteomics

1.3.1 Label-free quantification

The quantitative study in top-down proteomics is critical to further investigate intact proteoform variation as a function of the cell cycle, drug effect, or pathological change. The quantification of intact proteoforms and PTMs is gaining importance as it is known that some specific proteoforms and PTMs are functional or indicate the change in biological systems. Therefore, quantification of proteins may not be sufficient to monitor these changes, quantification of unique proteoforms of the same protein among different samples may be required.

The simplest method for top-down quantitation is label-free quantification, which directly compares intensity ratios of proteoforms for relative quantitation (**Figure 1.6**). The MS spectra comparison is powerful to study proteoforms rich in PTMs like histone or the phosphoproteome. Quantification accuracy is affected by the differing ionization efficiencies of unique proteoforms; however, the ionization efficiency of intact proteoforms is less affected by PTMs than smaller peptides.[45] The Ge group applied this method to relatively quantify the phospho-proteoforms of cardiac troponin

I (cTnl) from different biological samples, including heart tissues[46] and serum samples[47]. The Young group relatively quantified the histone proteoforms by determining relative ion abundance.[48]

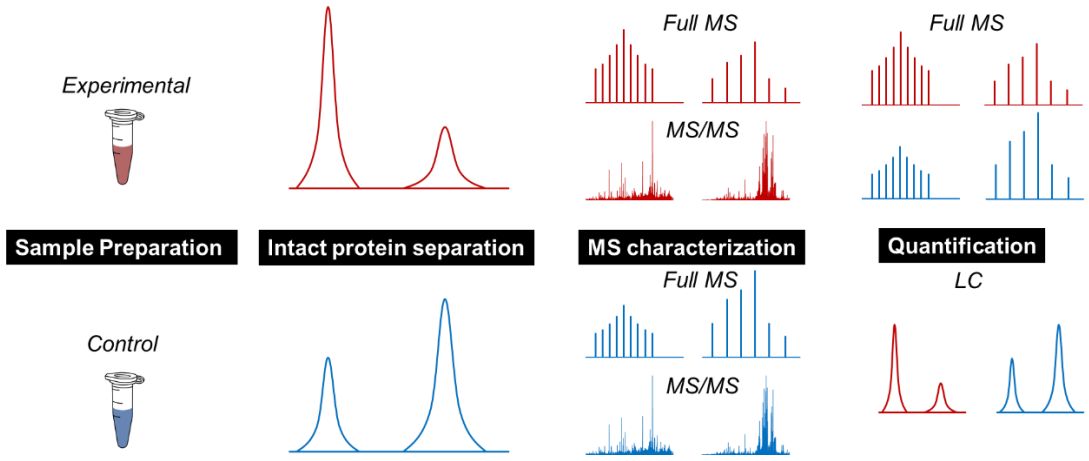


Figure 1.6 Schematic of label-free quantification in top-down proteomics.

The label-free quantification in top-down proteomics is still the most popular approach because it is applicable for all different samples and the sample handling is simple. Usually, various samples are analyzed by LC-MS individually, the proteins are quantified based on spectral count[49], proteoform intensity[50], or the extracted ion chromatography of the same proteoform[51] among different samples. However, simple experimental procedures and data processing sacrifice quantitative accuracy. The accuracy and robustness of label-free quantification rely on the reproducibility of LC-MS instruments.[52] The low throughput also hinders the analysis of large batches of samples. Additionally, label-free quantification in multi-dimensional separation is challenging due to the difficulty to sum the quantification from fractions.[53]

1.3.2 Isobaric metabolic labeling quantification

Since the label-free quantification is limited by low sample throughput and experimental results are affected by the state of the instrument, isotopic labeling methods have been developed to merge different samples in one, single analysis. (**Figure 1.7**) One method of isotopic labeling applied to proteomics is stable isotope labeling in cell culture (SILAC).[54, 55] SILAC incorporates isotopically labeled amino acids into media during cell culture and normal protein synthesis. Using full MS detection, the proteins from "heavy-labeled" samples are detected separately from proteins in normal samples.[54] Additionally, a triple labeling approach using isotopic lysine and arginine in the medium was applied by Olsen et al. to compare phosphorylation dynamics among 3 conditions.[56]

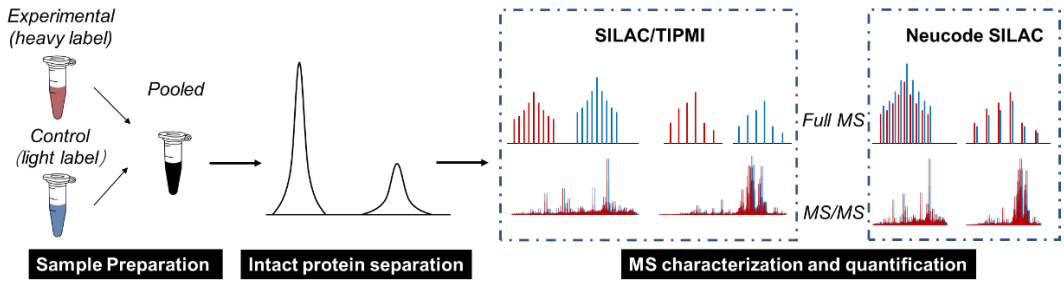


Figure 1.7 Schematic of isobaric metabolic labeling quantification in top-down proteomics.

SILAC demonstrates advantages over label-free quantification for lab-cultured cells for the study of proteoforms in top-down proteomics.[57] The Smith group reported a modified SILAC quantification approach using isotopic leucine (Leu-D10) in top-down proteomics.[58] The Mann group extended SILAC quantification to characterize recombinant proteins using isotopic arginine and lysine.[57] However, SILAC suffers from limited throughput, the current technology only allows 2 or 3 samples to be analyzed in

a single run; additionally, incorporation of isotopes is restricted by the amino acid structure spectral complexity is increased resulting in multiple isotopic clusters which complicate data analysis.[59] To tackle these issues, neutron encoding (NeuCode) SILAC in top-down proteomics was developed by the Coon group.[60] NeuCode SILAC cleverly applied 6 isotopologues of lysine which, which relies on the advanced resolving power of the modern FT-MS instruments, can be isotopically resolved using high-resolution MS/MS for 6-plex SILAC analysis. Another metabolic labeling strategy for top-down proteomics, tunable intact mass protein mass increases (TIPMI), utilizes spiking of ^{13}C-sugar or ^{2}H-water into cell cultures for intact protein quantification.[61] Compared to SILAC quantification, TIPMI has more symmetric and predictable mass peak pairs. However, TIPMI still introduces multiple labeled products from one protein which complicates full MS spectral, and there's no available software for TIPMI quantification. Also, the most serious issue for all metabolic labeling approaches is that they are only applicable for lab cultured cells or animal models, so these approaches are limited for clinical samples.[52]

1.3.3 Isobaric chemical tag labeling quantification

Isobaric chemical tag labeling, including tandem mass tags (TMT)[62] and isobaric tags for relative and absolute quantification (iTRAQ)[63], has been widely used as the gold standard in quantitative bottom-up proteomics. These chemical tags are popular because they can label all different kinds of protein samples with high sample throughput, up to 21-plex using N, N-dimethylleucine (DiLeu) tags developed by the Li group [64]. The Peng group hybridized TMT 11-plex and 16-plex to achieve a 27-plex assay, which combines MS1 and MS2 quantification.[65]

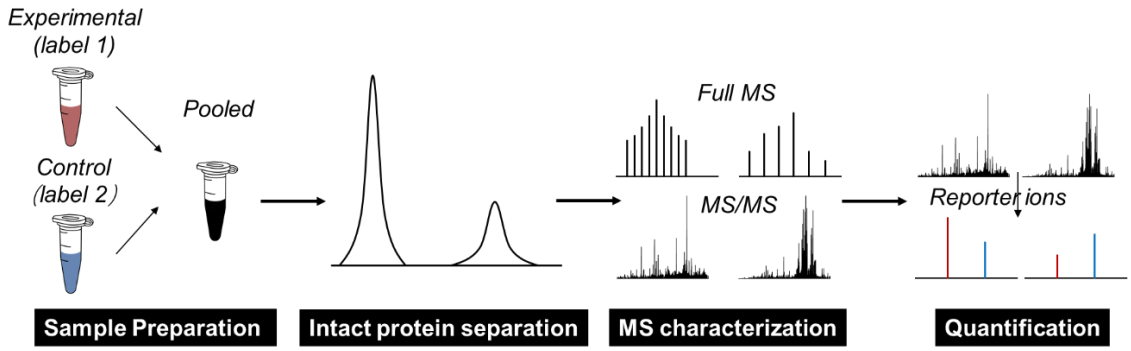


Figure 1.8 Schematic of isobaric chemical labeling quantification in top-down proteomics.

Isobaric chemical tag labeling quantification has been widely applied to different biological sample studies in bottom-up proteomics.[66-68] Compared to peptide-level labeling, protein-level labeling has more potential benefits.[69] For instance, the labeled protein samples can be pooled before sample preparation steps, which reduces the variation because these samples are treated together. Also, protein-level prefractionation can be utilized to improve proteome coverage. However, the application of isobaric chemical tags in top-down proteomics has been limited. Hung et al. reported the use of isobaric chemical tags for intact standard protein quantification using protein-level TMT labeling.[70] Two major issues blocked the path to the successful application of isobaric chemical tags to intact level protein labeling. First, the organic solvent that is required to dissolve the labeling reagent-induced protein precipitation.[71] Second, tryptic peptides usually only contain N-terminal and one Lys residue to react with labeling reagent, however, the intact proteoform contains multiple Lys residues, which may result in several side products.

To utilize the superiority of protein-level labeling and to avoid shortcomings, researchers digest the intact labeled proteins followed by analysis using bottom-up proteomics. Nie et al. used this method to study protein-level labeling for purified

glycoprotein quantification in serum samples.[69] The protein-level isobaric chemical labeling coupled to digestion is also applied to quantify protein structural changes. Hill et al. quantified the site occupancy of individual Lys residues in the monoclonal antibody using protein-level TMT labeling.[72] Yu et al. applied protein-level TMT quantification to evaluate protein structural changes in Alzheimer's Disease.[73]

Recently, our group developed an approach using protein-level TMT labeling to accurately characterize and quantify intact proteoforms in biological samples. (Chapter 3)[71] We also optimized the TMT labeling parameters of protein-level TMT labeling to facilitate the complete labeling of intact proteins for top-down proteomics. (Chapter 4)

1.4 Dissertation synopsis

This dissertation presents the development and evaluation of a high-throughput, quantitative top-down proteomics approach, which achieves an aerial view of the identification and quantification of intact proteoforms in complex protein samples. We describe in subsequent chapters how we established a protein-level TMT labeling method and 2D RPLC separation analysis platform for quantitative study of soluble proteomes, such as cell lysates.

It has been estimated that about one-million proteoforms exist in intact human samples and these proteoforms are expressed in a large dynamic range. To provide a comprehensive view of the intact proteome and allow for the identification of novel proteoforms, we applied the 2D pH RP/RPLC-MS platform to characterize intact human proteoforms. Compared to the 1DLC platform, the identification of proteins and proteoforms was largely improved and low abundant proteoforms with rare PTMs were identified. (**Chapter 2**)

TMT quantification has been widely applied in bottom-up proteomics, it has advantages in eliminating instrument variation, high-throughput capability, and can be applied to lab-cultured and clinical samples. However, TMT quantification is limited in top-down proteomics applications. The major challenge is that the proteins will aggregate under TMT labeling conditions. We developed a filter-SEC technique, which combines MW cut-off filtering and SEC separation, to enrich low MW proteins. The intact proteoforms and PTMs are identified and quantified using our protein-level TMT labeling sample preparation platform. (**Chapter 3**) The second challenge that limits TMT application to top-down proteomics is the creation of side products after labeling, which reduces the signal-to-noise ratio and complicates data processing. We systematically optimized and evaluated the parameters of protein-level TMT labeling, which include TMT-to-protein ratios, final pH/concentration of quenching buffer, initial protein concentration, reaction time, and reaction buffer. We achieved an average of 85% complete labeling by adjusting the parameters, which simplifies MS spectra and subsequent data analysis (**Chapter 4**).

Chapter 2 Deep intact proteoform characterization in human cell lysate using high-pH and low-pH reversed-phase liquid chromatography

*The materials in Chapter 2 are adapted from

Yu, D., Wang, Z., Cupp-Sutton, K.A., Liu, X. and Wu, S., 2019. Deep intact proteoform characterization in human cell lysate using high-pH and low-pH reversed-phase liquid chromatography. *Journal of the American Society for Mass Spectrometry*, *30*(12), pp.2502-2513.

2.1 Abstract

Post-translational modifications (PTMs) play critical roles in biological processes and have significant effects on the structures and dynamics of proteins. Top-down proteomics methods were developed for and applied to the study of intact proteins and their PTMs in human samples. However, the large dynamic range and complexity of human samples make the study of human proteins challenging. To address these challenges, we developed a 2D pH RP/RPLC-MS/MS technique that fuses high-resolution separation and intact protein characterization to study the human proteins in *HeLa* cell lysate. Our results provide deep coverage of soluble proteins in human cancer cells. Compared to 225 proteoforms from 124 proteins identified when 1D separation was used, 2778 proteoforms from 628 proteins were detected and characterized using our 2D separation method. Many proteoforms with critically functional PTMs including phosphorylation were characterized. Additionally, we present the first detection of intact human GcvH proteoforms with rare modifications such as octanoylation and lipoylation. Overall, the increase in the number of proteoforms identified using 2DLC separation is largely due to the reduction in sample complexity through improved separation resolution, which enables the detection of low-abundance PTM-modified proteoforms. We demonstrate here that 2D pH RP/RPLC is an effective technique to analyze complex protein samples using top-down proteomics.

2.2 Introduction

The ability to identify functional proteins with post-translational modifications (PTMs) is a crucial element in understanding cellular function. For example, acetylation of acidic residues can impact protein function and acetylation of protein N-terminal

amines can protect proteins from degradation [74]. Furthermore, the acetylation of primary amines on histones often regulates protein-DNA interactions [75-78]. Protein PTMs have emerged in the post-genomic era as critical features in regulating and diversifying the biological activity of proteins. However, identifying specific proteoforms with PTMs and understanding their function on individual proteins is currently limited by the lack of effective and accessible analytical methods. Therefore, there is a need to develop a high-throughput approach to characterize intact proteins and their modified proteoforms at the systems level.

Proteomics is a high-throughput bio-analytical approach to study proteins in a complex sample using mass spectrometry (MS)-based methods. Currently, there are two fundamental approaches to MS-based protein identification and characterization: bottom-up and top-down proteomics. Recent developments in bottom-up proteomics significantly expanded the depth of proteomic analysis, both qualitatively and quantitatively [79, 80]. Two independent bottom-up-based human proteomics studies confirmed ~ 20,000 protein-coding genes from human samples, which provided direct evidence for the actual translation of over 90% of putative protein-coding genes [81, 82]. The throughput of top-down proteomics has also improved significantly during the last decade due to the advancements in chromatographic separation and high-resolution MS instrumentation [14, 35, 42, 48, 83-86].

Reversed-phase liquid chromatography (RPLC) is the most prevalent separation technique in top-down proteomics. Long capillary column (*i.e.*, 1 m length, 75 μm i.d.) ultra-high-pressure liquid chromatography has been coupled to high-resolution mass spectrometry (UPLC-HRMS) for the analysis of intact proteins in complex mixtures [17, 87]. However, in cases of high complexity and large dynamic range, such as in human cell lysate, 1D-RPLC may not provide sufficient proteome coverage due to a limited

sample loading capacity. To overcome this challenge, different LC separation formats are often combined to increase the separation power based on the orthogonality. LC techniques often used in proteomics include reversed-phase chromatography (RPLC), ion exchange chromatography (IEX), hydrophilic interaction chromatography (HILIC), hydrophobic interaction chromatography (HIC), size exclusion chromatography (SEC), etc. [26, 88-92]. Several electrophoresis-based separation methods such as gel-eluted liquid fraction entrapment electrophoresis (GELFrEE) [93] and solution isoelectric focusing (sIEF) [13] can be applied as pre-fractionation methods for RPLC; however, these methods cannot be directly coupled to MS for detection. Capillary electrophoresis (CE) is a very promising electrophoresis-based separation technique that can be coupled after RPLC separation due to the small sample loading amount. CE can be directly coupled with MS either by the sheathless interface [94] or by a sheath-flow-based interface [95]. McCool et al. reported a three-dimensional platform, SEC-RPLC-CZE-MS/MS, that was applied to study intact proteins in *E. coli* lysate. A total of 5705 proteoforms from 850 proteins were identified using this method [35].

Recently, our lab developed a two-dimensional separation platform using high-pH and low-pH RPLC for top-down proteomics [26]. This method was applied to analyze intact proteins and proteoforms in *E. coli* cell lysate, and our results demonstrated that the platform provided high-resolution protein separation due to the comparatively good separation resolution and orthogonality between the two RPLC formats. Another advantage of this approach is the simple sample handling procedure which utilizes MS-compatible buffers. In this study, we further applied 2D pH RP/RPLC coupled with top-down MS to characterize intact proteins and proteoforms in human cell lysate (*e.g.*, *HeLa* cell lysate). A fraction-to-fraction orthogonal selectivity analysis between high-pH and low-pH separations was performed. In addition, we compared different

cell lysis sample preparation methods (*i.e.*, with and without the removal of small proteins with a 10 kDa MW cutoff filter). In total, 2778 proteoforms from 628 proteins were identified and manually confirmed. A total of 20 different types of PTMs were characterized, including intact proteoforms with phosphorylation, lipoylation, and glutathionylation. Overall, our results suggested that the two-dimensional high-pH and low-pH RPLC-MS can be easily adapted to complex protein samples for deep proteoform characterization.

2.3 Materials and Methods

2.3.1 Chemicals and Reagents

All chemicals including LC-MS grade water and organic solvents were purchased from Sigma-Aldrich (Milwaukee, WI) unless noted otherwise. Jupiter particles used as column packing material were purchased from Phenomenex (Torrance, CA).

2.3.2 *HeLa* Cell Culture and Cell Lysate Preparation

The *HeLa* cells were incubated at 37 °C in a humidified atmosphere containing 5% CO2 in Dulbecco's modified Eagle medium (DMEM) with 10% fetal bovine serum (FBS) and 2% penicillin-streptomycin. The growth status of the cells was monitored using a microscope. When the cells were 80–90% confluent, they were washed with ice-cold PBS buffer and scratched off the petri dish. The cells were centrifuged at 2500×g, 4 °C for 30 min. Ten plates (10 cm diameter) of *HeLa* cells were resuspended to 10 mL in lysis buffer (1 μM PMSF and 20 mM NaF in PBS, pH = 7.4). The cells were broken

using a sonication homogenizer on ice for 30 min, and insoluble matter was removed by centrifugation at 15,000×g, 4 °C for 30 min. The lysate (unfiltered sample) was aliquoted and stored at − 80 °C. For a portion of the lysate sample, small proteins or degradation products were removed using a 10 kDa MWCO spin filter at 4 °C (two times with 25 mM ammonium bicarbonate), referred to as "filtered sample."

2.3.3 High-pH RPLC Fractionation

The intact proteins from *HeLa* cell lysate were separated and analyzed using the 2D pH RP/RPLC-MS/MS platform as previously described [26]. Briefly, 1 mg of proteins was fractionated by an offline high-pH RPLC separation (pH = 10) using an XBridge Protein BEH C4 column (300 Å, 3.5 μm, 2.1 mm × 250 mm) from Waters, Inc. (Milford, MA). The mobile phase A (MPA) was 20 mM ammonium formate in water (pH = 10), and mobile phase B (MPB) was 20 mM ammonium formate in acetonitrile (pH = 10), the pH was adjusted to 10 using ammonium hydroxide. The unfiltered sample was separated using a gradient from 3 to 90% MPB over 60 min. The eluted proteins were fractionated into 24 2.5-min fractions starting from 16 min after sample injection (*i.e.*, the first detected UV peak). Based on the results from the unfiltered sample, we fractionated the filtered sample into nine fractions. All the fractions were vacuum dried and resuspended to a final volume of 50 μL using 25 mM ammonium bicarbonate before injecting onto the second-dimension column.

2.3.4 Low-pH RPLC-MS Analysis

Ten micrograms of protein (filtered, unfiltered, or fractions from high-pH RPLC separation) from the high-pH fractions was loaded onto a trapping column (150 μm i.d., 10 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 [17]. Specifically, MPA was 0.01% TFA, 0.585% acetic acid, 2.5% isopropanol, and 5% acetonitrile in water. MPB was 0.01% TFA, 0.585% acetic acid, 45% isopropanol, and 45% acetonitrile in water [28, 87, 96]. A 200-min gradient from 10 to 65% of MPB was applied with a flow rate of 400 nL/min. The LC elution was sprayed from an etched capillary tip through a customized nano-ESI interface to an LTQ Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific, Hanover Park, IL, Bremen, Germany, USA) [97]. The temperature of the inlet capillary was set to 275 °C and the spray voltage was 2.6 kV. Full MS scans used the resolving power setting of 100,000 (at m/z 400) with two micro scans. Collision-induced dissociation (CID) with a normalized collision energy of 35 eV was applied. The top 6 most abundant precursor ions in the full MS scans were selected for MS/MS fragmentation. The MS/MS data were obtained at a resolving power setting of 60,000 (at m/z 400) with two micro scans and an isolation window of 3.0 m/z. The AGC target was set as 5 × 105 for full mass scans and 3 × 105 for MS/MS scans. All the data was collected using Xcalibur 3.0 software (Thermo Fisher Scientific, Bremen, Germany).

2.3.5 Data Analysis

Raw data were deconvoluted and searched against the Swiss-Prot reviewed Homo sapiens protein database (published on May 2018, 20,355 proteins) using TopPIC Suite [98]. All parameters were set as follows: the error tolerance was 15 ppm; the maximum value for the unexpected mass shift was 500 Da; the maximum number of the unexpected mass shifts was 1. The decoy database was utilized to calculate PSM-level FDRs and proteoform-level FDRs in TopPIC. The E-value cutoff value was 8.81E−8

for PSM-level FDRs < 1% and 5.31E−9 for proteoform-level FDRs < 1%. For each dataset, we only considered proteoforms with E-values less than the proteoform-level FDR cutoff values. Proteoform identifications in different fractions were aggregated based on their retention time (~ 3 min normalized retention time window) and detected masses (15 ppm). No protein level FDR was calculated. All proteins with proteoforms that passed the E-value cutoff were reported. All identified proteoforms were also manually evaluated. MASH Suite [99] and ProSight Lite [100] were used for manual interpretation and spectrum presentation.

2.4 Results and Discussion

2.4.1 2D pH RP/RPLC Top-Down MS on *HeLa* Cell Lysate

In this study, we applied the previously developed 2D pH RP/RPLC separation and top-down MS to characterize intact proteins and proteoforms in human *HeLa* cell lysate. A total of 24 first-dimension fractions were analyzed with long-column high-pressure RPLC with top-down MS [26] and by SDS-PAGE (**Figure 2.1 a, b**). In total, 2778 unique proteoforms from 628 proteins were identified and manually confirmed. Seventy-seven percent of the proteoforms were detected in one fraction, 14% of the proteoforms were detected in two fractions, and 9% of the proteoforms were detected in three or more fractions. Our results suggest that high-pH RPLC provides good separation power as the first-dimensional separation.

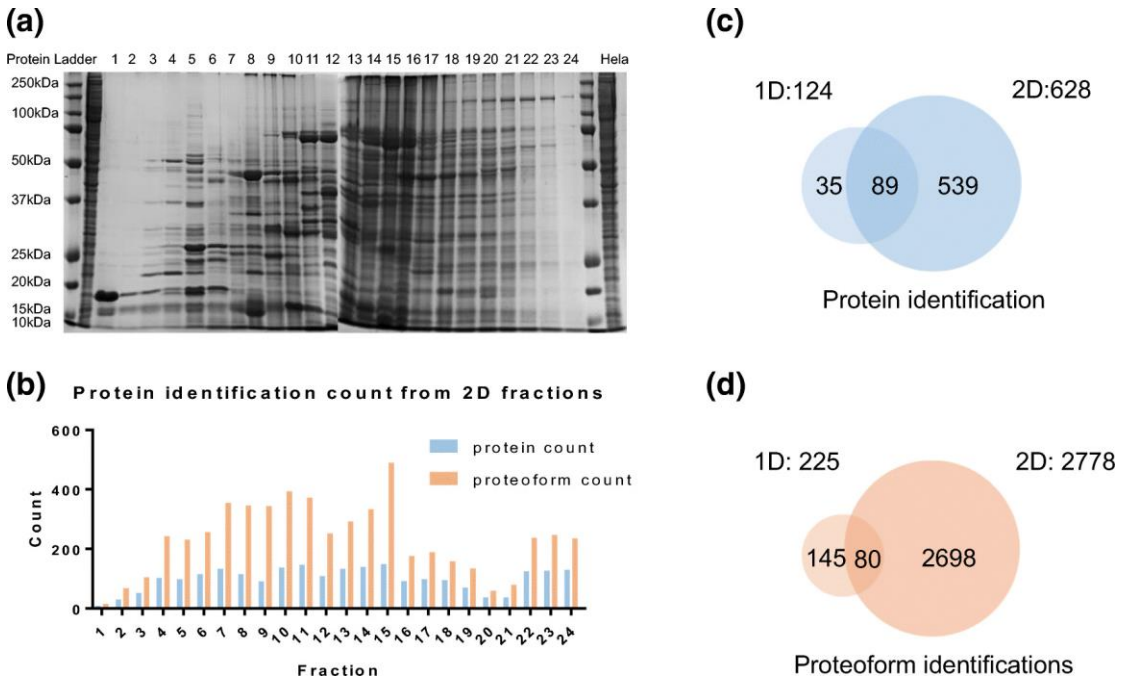


Figure 2.1 Protein and proteoform identification count of *HeLa* cell lysate.

(a) First-dimension fractions performed by SDS-PAGE. (b) Protein and proteoform identification count in each fraction under high pH conditions. (c) Venn diagram of proteins identified using the 1D and 2D platforms. (d) Venn diagram of proteoforms identified using the 1D and 2D platforms

The same sample was also analyzed using 1D low-pH RPLC top-down MS without pre-fractionation; 124 intact proteins and 225 intact proteoforms were identified. We compared the results of 1D and 2D separation of the same sample (**Figure 2.1 c, d**). We observed an increase in the number of proteoform identifications when 2D separation was used, and there are two possible reasons for this increase: (1) there is a higher chance that MS/MS analysis will be performed on low-abundance proteoforms in the 2D analysis due to extended MS analysis time and (2) improved separation performance in 2D analysis reduces sample complexity and allows for the detection of proteoforms that cannot be detected in 1D analysis because many co-eluted species split the ion current. To evaluate the effects of improved separation, we performed two 1D RPLC experiments using *E. coli* cell lysate: one with a 70-min gradient (lower peak

capacity) and one with a 200-min gradient (higher peak capacity). All the detected mass features were deconvoluted with a 1-min moving average window (*i.e.*, all spectra in a 1-min window were summed and peak deconvolution was performed). We compared the detected mass features in different MW ranges (**Figure 2.2**). In total, we detected 137 mass features > 5 kDa from the 70-min run and 367 mass features > 5 kDa from the 200-min run. Our results suggested better separation facilitated the detection of more species and provided higher chances to identify them using MS/MS analysis. Previously, we demonstrated that 2D pH RP/RPLC separation of *E. coli* cell lysate has high peak capacity due to good separation power on both high-pH RPLC and low-pH RPLC. Here, we applied 2D pH RP/RPLC to human cell lysate resulting in significantly improved detection of intact mass features (2637 mass features > 5 kDa) compared to the 1D 200-min RPLC run.

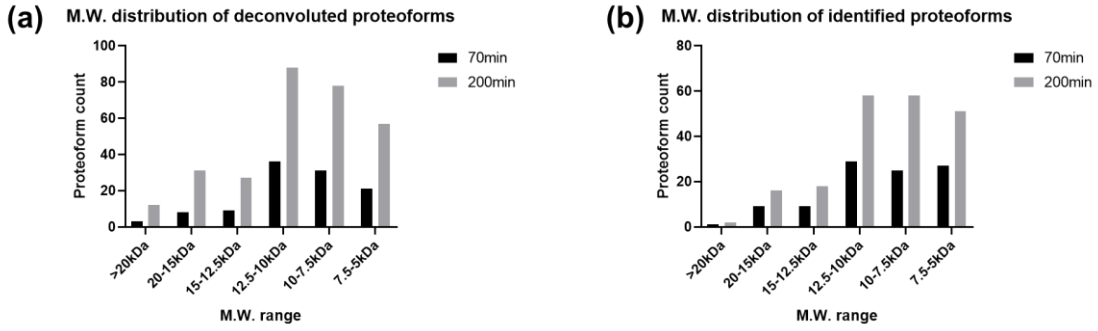


Figure 2.2. The gradient of 1st-dimensional high-pH separation was evaluated using a 70-min range and 200-min range.

Both the deconvoluted species count (a) and identified proteoforms count (b) from 2 sets of data are compared.

We further evaluated two cell lysate preparation approaches for 2D pH RP/RPLC top-down MS analysis (*i.e.*, with and without the removal of small proteins

using a 10-kDa MW cutoff filter). Based on the 24-fraction identification results from the unfiltered sample, the filtered samples were fractionated into nine fractions (**Figure 2.3 a**) to reduce the MS operation time. A total of 1710 unique mass features > 5 kDa were detected in nine fractions vs. 2637 unique mass features > 5 kDa in 24 fractions. Using the same MS conditions, fewer mass features were detected when nine fractions were collected, but much less MS analysis time was required. We also compared the masses of detected proteoforms in different MW ranges (**Figure 2.3 b**). The MW distribution is similar between these two fractionation schemes, and both approaches have similar percentages of detected mass features less than 10 kDa. One possible reason for the discrepancy in the percentage of proteoforms is that some proteins may degrade during the concentration process (*i.e.*, speedVac) before MS analysis. If protein degradation during concentration is an issue, filter-based concentrators or an online 2D system can be adapted in the future to reduce sample degradation.

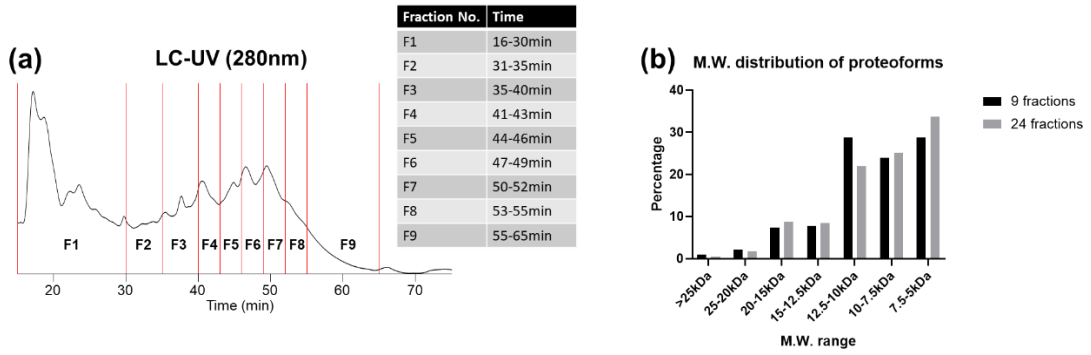


Figure 2.3. The 1st dimensional, high pH fractions were combined into 9 fractions based on the protein content of each fraction.

(a) LC-UV and fractionation for high pH separation. (b) The percentage of deconvoluted proteoforms within different MW ranges in 9 fractions vs. 24 fractions.

In general, current commercially available MS instruments are limited in their

ability to detect large proteins (MW > 30 kDa) with low abundance due to the low S/N ratios inherent to mass measurement. In our study, only a very small portion of proteins > 20 kDa were detected even though many large proteins were observed in the same fractions by SDS-PAGE. Gel-eluted liquid fraction entrapment electrophoresis (GELFrEE) has been coupled with RPLC for deep intact human proteoform characterization [13, 101], and different fragmentation approaches such as UVPD [15] and Ai-ETD [42] have been coupled with GELFrEE for better sequence coverage of PTM localization in human samples. In addition, many proteoforms > 30 kDa were successfully detected in GELFrEE fractions that contained enriched large proteoforms using a 21T FTICR-MS [85]. By removing co-eluting small proteoforms, GELFrEE allows for the detection of large proteoforms with low S/N ratios in complex samples. However, GELFrEE often requires MS-incompatible detergents such as SDS as well as specific instrumentation. Therefore, GELFrEE is challenging to use online with MS analysis. SEC is another size-based separation that may enhance MS detection of large proteoforms and has been coupled online with MS analysis [20, 34]. However, SEC in general does not provide good separation resolution in the full MW range because the separation selectivity of SEC depends on the resin pore size. Our results suggested that both high-pH RPLC and low-pH RPLC provide good separation power, which allows for deep proteoform detection in complex samples. With an additional dilution step, our approach can be easily adapted into an online platform to improve the robustness and sensitivity of the 2D pH RP/RPLC. Additionally, larger proteins can be enriched prior to 2D pH RP/RPLC separation using size-based approaches such as GELFrEE or SEC for better detection of larger proteins.

2.4.2 Orthogonality Between High-pH RPLC and Low-pH RPLC for Intact Protein Separation

In bottom-up proteomics, it has been reported that high-pH RPLC only has semi-orthogonality with low-pH RPLC for peptide analysis [102, 103]. For example, more hydrophobic peptides tend to elute late regardless of high or low pH. Therefore, the separation window of the secondary dimension cannot be fully utilized, which limits the usage of online 2D pH RP/RPLC analysis in bottom-up proteomics. Our previous study in top-down proteomics suggested that high-pH RPLC had good orthogonality against low-pH RPLC [26]. To further assess if these two separation forms have semi-orthogonality or full orthogonality for intact proteoform analysis, we evaluated proteoform elution patterns in high-pH fractions (**Figure 2.4**). For each fraction (each column in **Figure 2.4**), a heatmap was generated using the relative number of uniquely identified proteoforms in each bin (10-min windows). It has been demonstrated that in an ideally orthogonal system, area coverage > 63% represents full orthogonality [102]. Our results suggested that proteoforms were distributed across the entire efficient elution window of the second dimension in most of the fractions (81% bins containing 5 or more mass features > 2.5 kDa). The good orthogonality between RPLC performed under different pH conditions can be explained by the change in charge distribution caused by the change in pH of the mobile phase [102, 104, 105]. Compared to peptide chains, intact proteins are more complicated with more dynamic and versatile changes in charge distribution under different pH conditions. These changes in charge distribution can significantly affect protein retention and elution orders [26, 106].

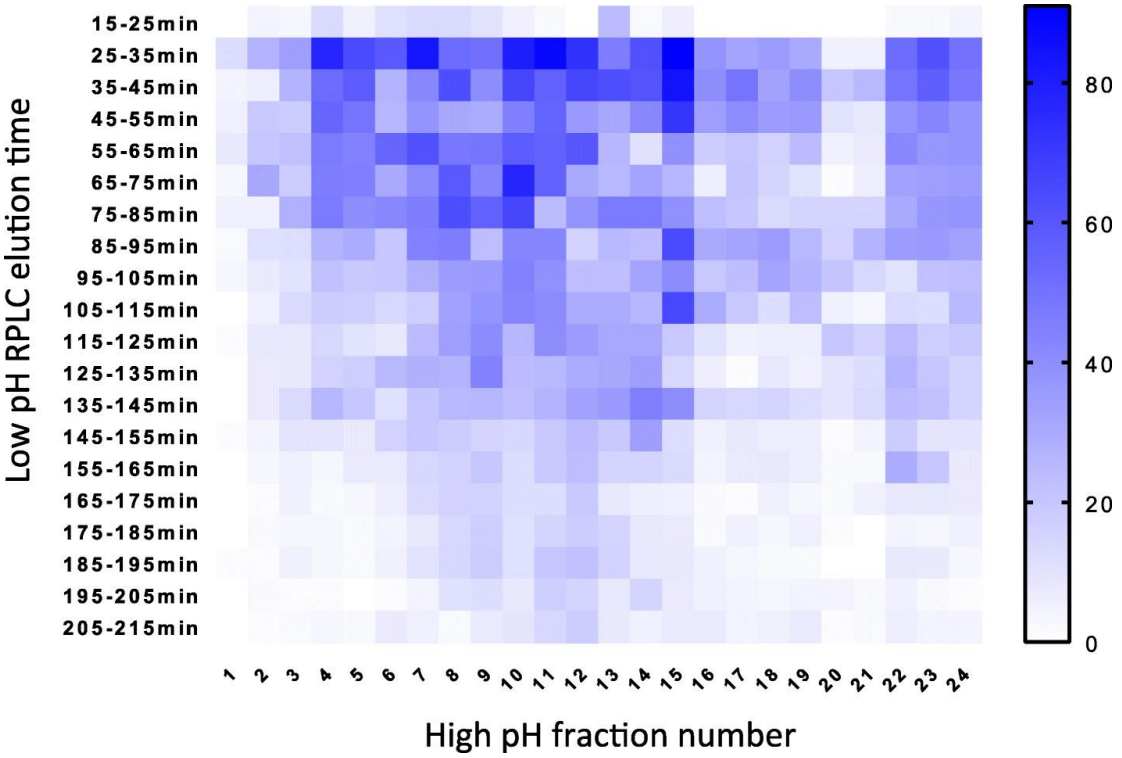


Figure 2.4 The proteoform elution patterns in high-pH fractions.

Each bin in the heatmap represents the number of detected mass features > 2.5 kDa in each 10-min elution window in the low-pH RPLC separation

To further evaluate the effects of pI values on intact protein retention behavior in high-pH RPLC separation, we filtered all identified intact proteoforms from the nine-fraction samples. The pI values were calculated for each filtered proteoform, and the average pI values and their standard deviations were plotted for each fraction (**Figure 2.5**). The average pI values were between 4 and 7 for most fractions, which suggests that the pI value alone does not alter the retention behavior. Still, the average pI value (4.48) in fraction 1 was significantly lower than other fractions indicating that pI may have some effect on separation. At pH 10, most proteins with pI values around 4 carry multiple negative charges, which reduces their retention times. In general, we did not detect many proteins with pI values larger than 8. One possible reason is that proteins with high pI values do not carry many charges and may not be easily eluted from the

high-pH RPLC column.

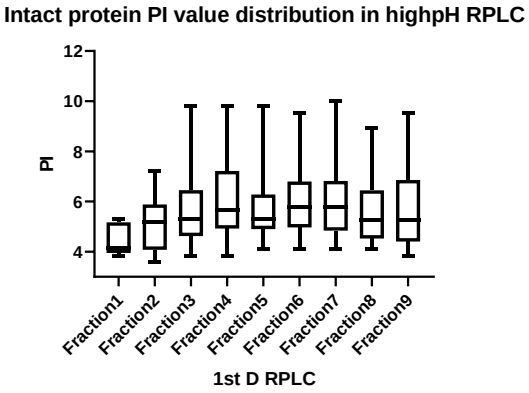


Figure 2.5 The PI value distribution of intact proteins in each fraction after 1 dimensional, high pH separation.

2.4.3 Identification and Characterization of Intact Proteins

Overall, a total of 20 different types of PTMs were characterized in the proteoforms detected using the 2D analysis. Only 6 types of PTMs were detected using 1D analysis. Several commonly observed PTMs were detected including phosphorylation, acetylation, and glutathionylation. We were able to identify 96 intact proteoforms with phosphorylation and 771 intact proteoforms with acetylation. Several less common functional PTMs were also detected using the 2D platform including lipoylation and octanoylation. To increase the number of proteoforms detected using 1D RPLC separation, a combined top-down and intact mass analysis, such as that developed by Smith's group, may be utilized [107]. The Smith group proposed a strategy to combine both top-down MS and intact mass determination, which addresses the issue that some proteoforms may be detected, but not identified, using MS/MS. The mass shift between experimentally detected species in the same run was calculated, and experimental delta

mass histograms were plotted (**Figure 2.6**). Mass shifts (16 Da and 80 Da) were often detected in pairs in both nine-fraction analysis and 24-fraction analysis, which may correspond to oxidation and phosphorylation. Interestingly, although we identified 771 intact proteoforms with acetylation, we did not observe many proteoform pairs with 42 Da shifts. This may be due to the fact that many acetylated species we detected have static N-term acetylation.

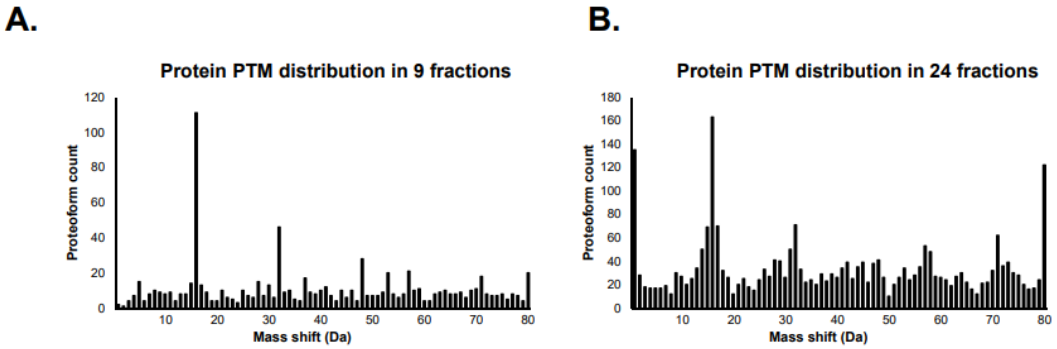


Figure 2.6 Mass shift histograms of experimental mass figures.
(A) 9-fraction and (B) 24-fraction separation.

We also conducted a comparison of the proteoforms detected using both the 1D and 2D platforms (**Figure 2.7**). We observed significant improvements in the S/N ratios with higher signal and lower background noise when 2DLC separation methods were used, which is consistent with previous literature [108]. For example, calmodulin-1 (Swiss-Prot, P0DP23) was detected in both 1D and 2D results, but there was a much higher S/N ratio in 2D (S/N = 647) than in 1D (S/N = 37) (**Figure 2.7**). Additionally, a proteoform of calmodulin-1 which lost three amino acids (*i.e.*, TyrAlaLys) at the C-terminus was detected using 2D (S/N = 23) but not 1D separation. The C-terminus deletion mutants of

calmodulin-1 were reported to reduce the Ca2+ binding ability because it had minor conformers [109]. This suggests that the C-terminus of calmodulin-1 is active in inter-domain communication while Ca2+ induces structural transition so the truncated proteoform detected is functionally relevant. To summarize, our results suggest that the 2D pH RP/RPLC separation and top-down MS can be applied to complex protein samples for deep proteoform characterization, especially to detect and characterize low-abundance proteoforms and PTMs. Additionally, the application of 2D LC separation improves the S/N ratio, which increases the limit of detection of the analytes [108].

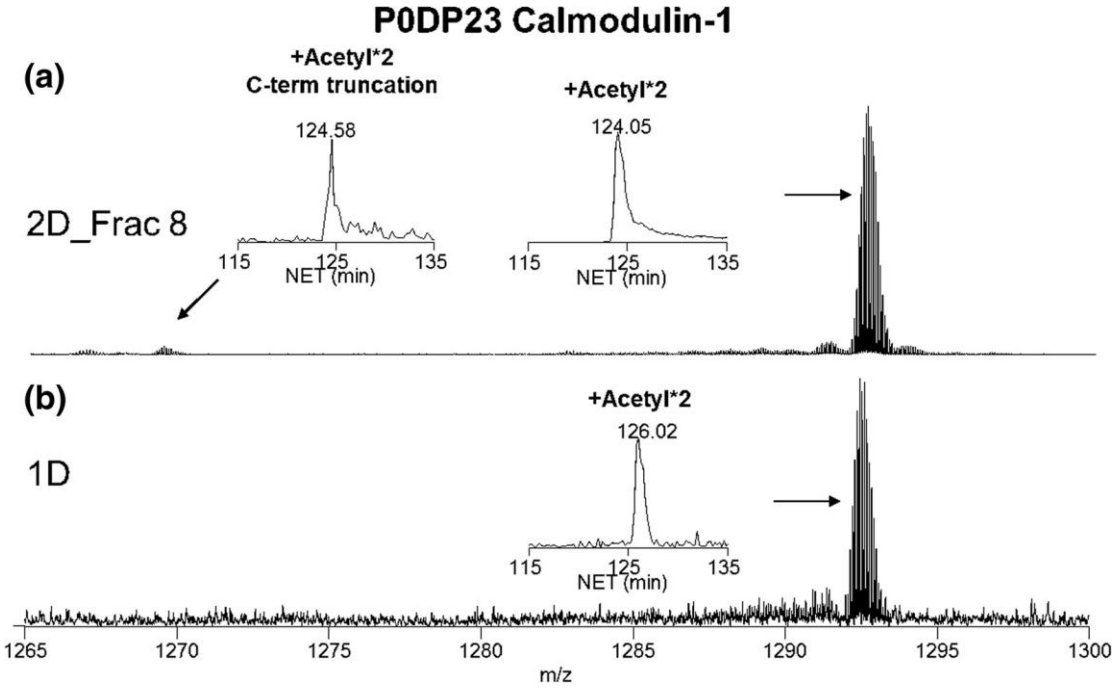


Figure 2.7 MS detection of calmodulin-1 proteoforms.

(a) Full MS spectrum and extracted ion chromatograms (EICs) of proteoforms after 2D separation. (b) Full MS spectrum and extracted ion chromatograms (EICs) of proteoforms after 1D separation.

2.4.4 Characterization of Intact Proteoforms with Phosphorylation

The 2D LC technique applied in our study increases the detection of phosphorylated proteoforms. Many low-abundance PTM-modified proteoforms were detected using the 2D platform with good S/N ratios but were not observed in the 1D analysis. Two proteoforms of parathymosin (Swiss-Prot, P20962), with and without phosphorylation, were characterized in fraction 5 during the 2D analysis (**Figure 2.8 a**). The MS/MS spectra elucidated mass differences between b51 and b53 ions indicating the possible locations of the phosphorylation (**Figure 2.8 b**). We observed a decrease in the elution time of the phosphorylated proteoforms in both 1D and 2D results because the phosphorylation added an additional charge to the protein that may have decreased the hydrophobicity. Decreased hydrophobicity of the phosphorylated proteins weakened the binding of these proteins to the stationary phase in RPLC, so they often eluted earlier compared to non-phosphorylated proteins [110]. Additionally, some Fe[III] adducts were detected from phosphorylated parathymosin proteins (* in **Figure 2.8 a**). The Fe[III] adducts may come from the stainless-steel parts of the high-pressure system, such as the frit and the union. To avoid the production of Fe[III] adducts and increase the sensitivity for phosphoprotein detection, a "metal-free" RPLC-ESI-MS platform could be adapted to our 2D pH RP/RPLC platform [87].

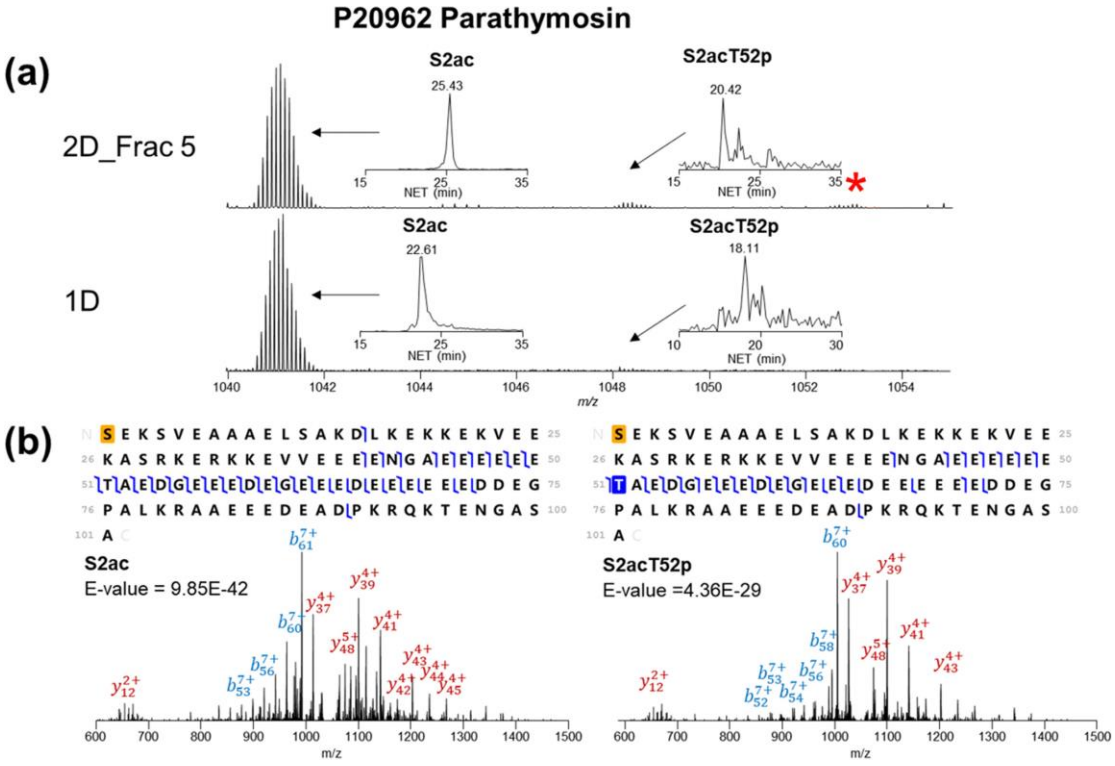


Figure 2.8 MS detection and MS/MS identification of parathymosin proteoforms.

(a) Full MS spectrum of proteoforms. The asterisk (*) shows the Fe[III] adduct. The EIC of each proteoform is shown. (b) MS/MS identification of proteoforms. The orange serine residue on the N-terminus is acetylated. The tyrosine residue highlighted in blue is phosphorylated.

Ribosomal proteins are involved in protein synthesis by providing control mechanisms for transcription and translation [111]. The 60S acidic ribosomal protein P2 (Swiss-prot, P05387, RPLP2) is coded by the gene RPLP2 and participates in the elongation step in protein synthesis and GTPase activation [112, 113]. Phosphorylation has been proven to stimulate interaction with eukaryotic translation elongation factor 2 (eEF-2) [113]. Phosphorylation of serine residues located near the C-terminus increases the affinity of RPLP2 for eEF-2. The phosphorylation of these residues also causes changes in eEF-2 activity, which suggests the C-terminus is involved in this functionality. The phosphorylation on the C-terminus of RPLP2 is also related to autoimmune disease. Autoantibodies against the C-terminus peptide of 60S acidic ribosomal proteins exist in

15% of systemic lupus erythematosus patients [114].

We observed a total of ten proteoforms of RPLP2 using 2D methods with modifications including phosphorylations, oxidations, and a sequence variation, Tyr->Gln (**Figure 2.9**). **Figure 2.10 a** shows the MS spectrum of 60S acidic ribosomal protein P2 proteoforms in fraction 14 from the 2D separation. The 60S acidic ribosomal protein P2 variations include the mature proteoform RPLP2_P0, and the proteoforms RPLP2_P1 and RPLP2_P2, which have 1 and 2 phosphorylations, respectively. **Figure 2.10 b** illustrates the identification and PTM characterization of RPLP2_P0, P1, and P2. The MS/MS spectra elucidated the mass differences between b96 and b107 ions of the three proteoforms indicating the possible locations of the phosphorylations. However, there was no fragmentation between Ser 102 and Ser 105, so the location of the phosphorylation of RPLP2_P1 could not be determined. The NET of proteoforms shifted from 119 to 126 min and the extracted ion chromatogram (EIC) of RPLP2_P1 was split into two peaks (*i.e.*, 120.42 and 120.87). This splitting might suggest the existence of two proteoforms with one phosphorylation on Ser102 or Ser105. To determine the location of the phosphorylation, the fragmentation method should be optimized to enhance the sequence coverage using electron transfer dissociation (ETD) and/or UVPD in the future [115]. We also found that phosphorylated proteoforms were eluted later than non-phosphorylated proteoforms, which disagrees with previous observations [110]. This may be due to the fact that both phosphorylated sites were located on the highly flexible and acidic C-terminus of the protein.

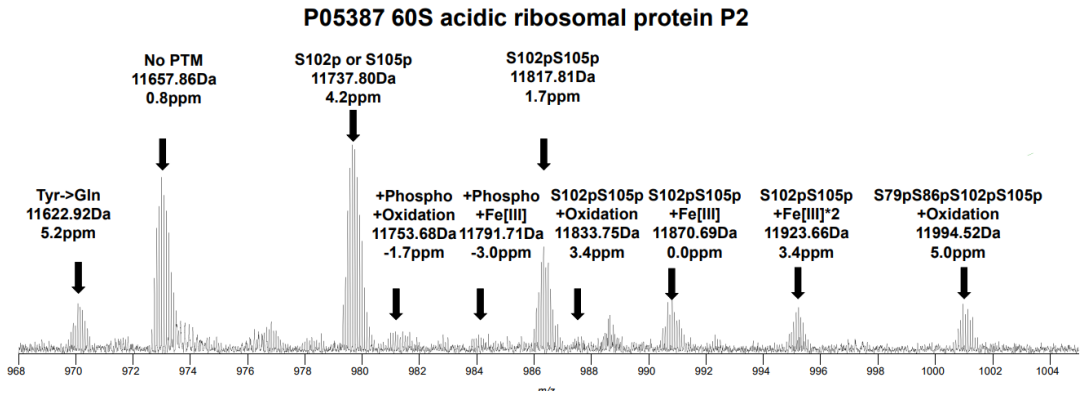


Figure 2.9 The MS spectra of all RPLP2 proteoforms identified.

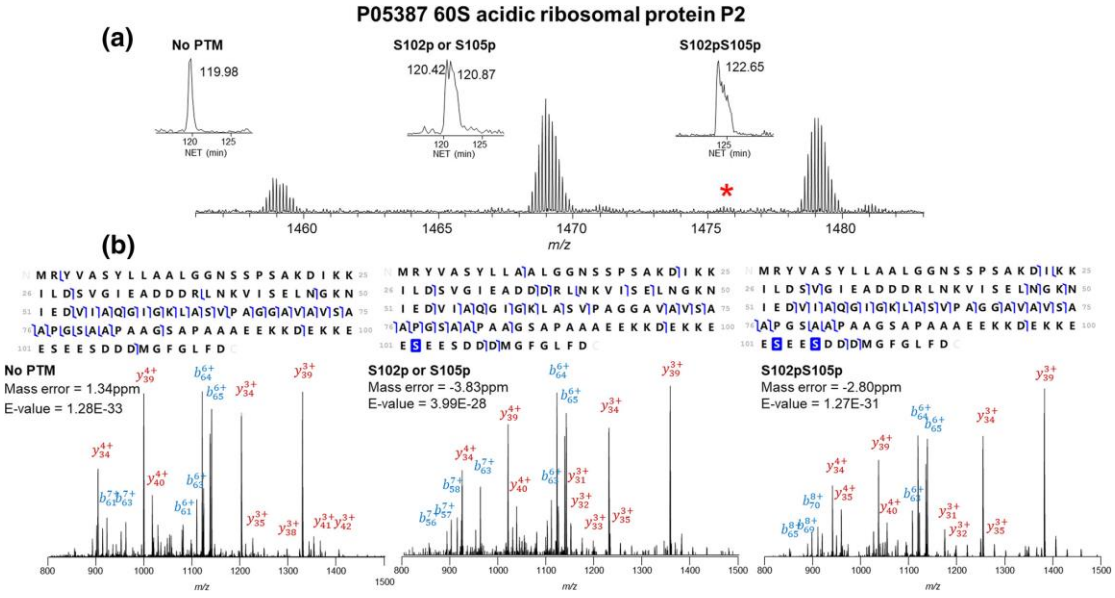


Figure 2.10 MS detection and MS/MS identification of RPLP2 proteoforms.

(a) Full MS spectrum of proteoforms: RPLP_P0, P1, P2. The asterisk (*) indicates the Fe[III] adduct. The EIC of each proteoform is shown. (b) MS/MS identification of RPLP_P0, P1, P2. The serine residues highlighted in blue are phosphorylated

2.4.5 Identification of Intact Proteoforms with Functional Modifications

The 2DLC technique studied here allowed for the identification of functional PTMs that have not been commonly studied. For example, octanoylation and lipoylation

were characterized on the glycine cleavage system H protein (Swiss-Prot, P23434, GcvH). The GcvH protein was reported as a subunit of the glycine cleavage system (GCV). GCV is active in glycine and serine catabolism [116]. The GcvH protein shuttles the methylamine group of glycine from the P protein to the T protein [117]. We confidently identified two GcvH proteoforms with octanoylation and lipoylation (**Figure 2.11**), which have never been reported in previous top-down studies. The transit peptides of both proteoforms were cleaved suggesting that these proteoforms were mature and transferred to the mitochondrion. The difference between the y66 and y67 ions, observed via MS/MS fragmentation, suggests two different PTMs (octanoylation and lipoylation) are located at the same amino acid, Lys50, of the proteoforms. Octanoylation and lipoylation are rare modifications characterized by the addition of a heptane chain and a heptane chain ending in a 5-membered ring containing two sulfur atoms, respectively. Lipoylation plays a critical role in the function of GcvH. The oxidized form of the lipoylation serves as the acceptor of the intermediary methylamine group derived from glycine as the two sulfur atoms bind to the glycine residue [117, 118]. Lipoylation was also reported to bind with the substrates of other enzymes (*e.g.*, pyruvate dehydrogenase and alpha-ketoglutarate dehydrogenase) [119]. Octanoylation is an intermediate of lipoylation. Jordan et al. published a mechanism proposing that octanoyl-ACP is converted to lipoyl-ACP in *E. coli*. In this mechanism, LipA lipoate synthase inserts two sulfur atoms at C6 and C8 on the octanyl chain [120, 121]. Our results demonstrated the first detection of octanoylation on GcvH on human proteins, which suggests that the lipoyl-GcvH was synthesized through a similar pathway in human biological systems.

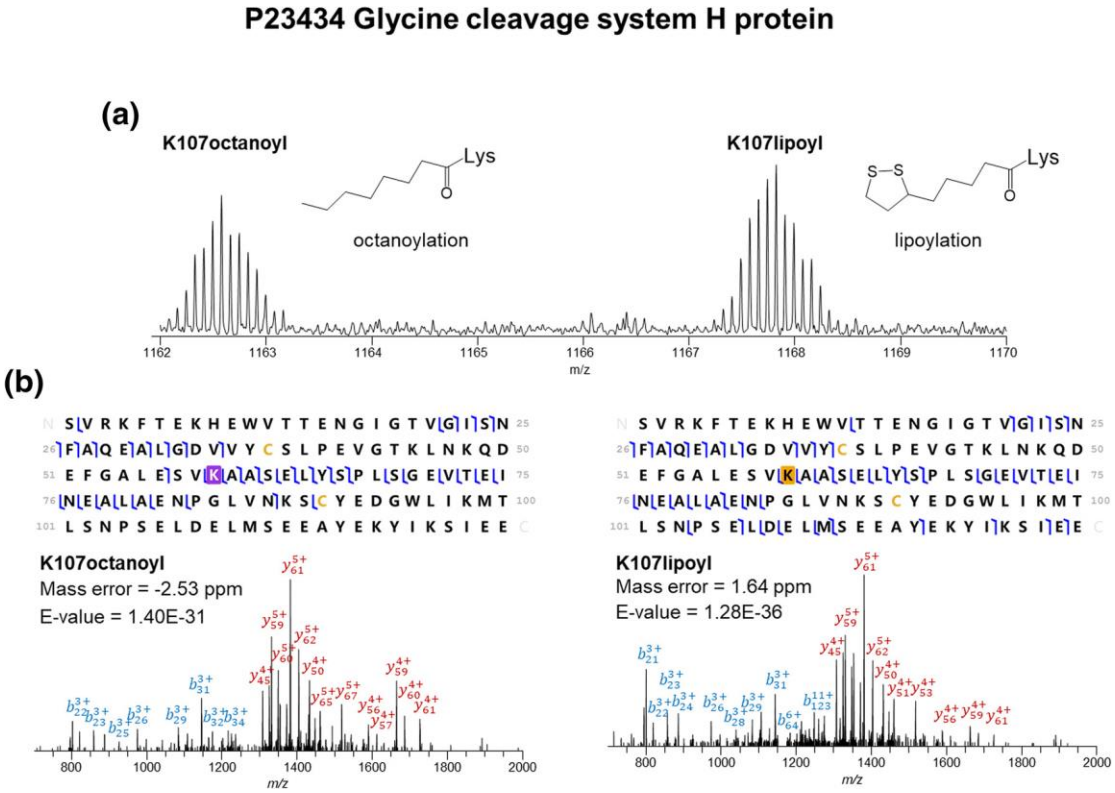


Figure 2.11 MS detection and MS/MS identification of GcvH proteoforms.

(a) Full MS spectrum of GcvH proteoforms with octanoylation and lipoylation. (b) MS/MS identification of the GcvH with octanoylation and lipoylation. The lysine residue highlighted in orange is lipoylated. The yellow cysteine residues do not participate in disulfide bonds

Another example of a rare PTM identified by our method is hypusine on the eukaryotic translation initiation factor 5A-1 (Swiss-Prot, P63241, eIF5A-1). eIF5A-1 is an mRNA-binding protein and is involved in translation elongation. It plays such an essential role in eukaryotic cell growth and animal development that its inactivation causes death in yeast [122, 123]. We characterized two proteoforms of eIF5A-1 containing a hypusine residue and acetylation at Lys47 (**Figure 2.12**). The hypusine was found on both proteoforms using a y110 and y98 ion with + 87.08 Da. Hypusine, a unique PTM only reported on eIF5A, is derived from polyamine spermidine [122]. Interestingly, the hypusine residue is highly conserved from yeast to mammals [124]. Hypusine has a longer side chain than lysine and promotes the activity of eIF5A by enhancing translation

termination through stimulating peptide release from peptide bond formation at polyproline stretches and ribosome pausing sites [122]. It was reported that the basic charge from hypusine at Lys50 is important for acetylation at Lys47 [123]. The acetylation on Lys 47 was characterized by the mass difference of a y110 and y98 ion (+ 42.01 Da) from two proteoforms and has been previously reported in the literature [124]. The acetylation at Lys47 also regulates the activity of eIF5A. Previous research suggests a basic charge at Lys47 is critical for eIF5A activity [124]. Overall, the deep proteoform characterization made possible by 2DLC separation allows for the detection and quantification of functional PTMs and proteoforms, which are not commonly detected using other methods.

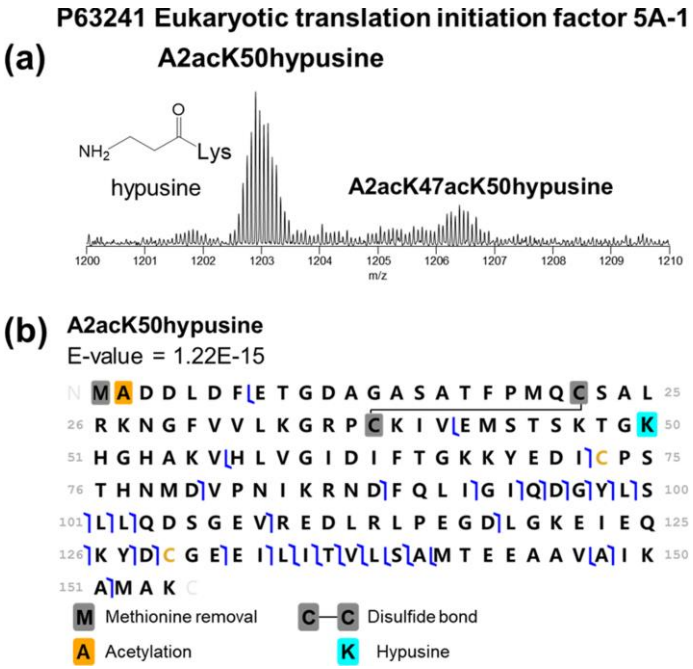


Figure 2.12 MS detection and MS/MS identification of eIF5A-1 proteoforms.

(a) Full MS spectrum of eIF5A-1 proteoforms with hypusine and 1 or 2 acetylations. (b) MS/MS identification of the eIF5A-1 proteoform with 1 acetylation on N-terminus. The cysteine residues highlighted in gray participate in a disulfide bond. The methionine residue highlighted in gray is removed. The yellow cysteine residues do not participate in disulfide bonds. The orange alanine residue on the N-terminus is acetylated. The Lys50 highlighted in blue is modified with a Hypusine.

2.5 Conclusion

In our study, the 2D pH RP/RPLC-MS/MS platform was applied to study the human proteins and PTMs in *HeLa* cell lysate. The protein, proteoform, and PTM identifications were enhanced using the 2D platform compared to 1D. The 2D separation implemented here made possible the characterization of proteoforms with complicated PTM distribution, such as multiple phosphorylations. In addition, we presented the first detection of intact human GcvH proteoforms with rare modifications such as octanoylation and lipoylation. The 2D platform is "salt-free" with good orthogonality, making the platform easily optimized to online 2D methods. Size-based separation approaches such as GELFrEE or SEC can be used before the 2D pH RP/RPLC for large protein fractionation and characterization. Different buffer conditions and column materials can be evaluated to improve the elution of hydrophobic proteins with high pI values. To characterize specific proteins, the 2D platform may be coupled with other sample preparation methods, such as immobilized metal affinity chromatography for phosphoprotein study[10]. In summary, our application provides a comprehensive aerial view of proteins in human cancer cells. This technique may also be used with isobaric-labeling-based quantification (TMT or iTRAQ) and applied to other human proteomics studies.

Chapter 3 Quantitative top-down proteomics in complex samples using protein-level tandem mass tag labeling

*The materials in Chapter 3 are adapted from

Yu, D., Wang, Z., Cupp-Sutton, K.A., Guo, Y., Kou, Q., Smith, K., Liu, X. and Wu, S., 2021. Quantitative Top-Down Proteomics in Complex Samples Using Protein-Level Tandem Mass Tag Labeling. *Journal of the American Society for Mass Spectrometry*, *32*(6), pp.1336-1344.

3.1 Abstract

Labeling approaches using isobaric chemical tags (*e.g.*, isobaric tagging for relative and absolute quantification, iTRAQ and tandem mass tag, TMT) have been widely applied for the quantification of peptides and proteins in bottom-up MS. However, until recently, successful applications of these approaches to top-down proteomics have been limited because proteins tend to precipitate and "crash" out of solution during TMT labeling of complex samples making the quantification of such samples difficult. In this study, we report a top-down TMT MS platform for confidently identifying and quantifying low molecular weight intact proteoforms in complex biological samples. To reduce the sample complexity and remove large proteins from complex samples, we developed a filter-SEC technique that combines a molecular weight cutoff filtration step with high-performance size exclusion chromatography (SEC) separation. No protein precipitation was observed in filtered samples under the intact protein-level TMT labeling conditions. The proposed top-down TMT MS platform enables high-throughput analysis of intact proteoforms, allowing for the identification and quantification of hundreds of intact proteoforms from Escherichia coli cell lysates. To our knowledge, this represents the first high-throughput TMT labeling-based, quantitative, top-down MS analysis suitable for complex biological samples.

3.2 Introduction

Quantitative top-down MS analysis of intact proteoforms with different post-translational modifications (PTMs) that arise from one gene is important because PTMs often cause the proteoforms to exhibit differing functions in biological processes [28, 125, 126]. Currently, peak intensity-based, label-free quantitation is the most commonly used

quantitation approach in top-down proteomics. It has been successfully applied to the quantitation of proteoforms in biological samples such as tumor samples [127], animal tissues [125, 128], human saliva [51], murine mitochondria [107], and the yeast proteome [129]. Moreover, label-free, top-down MS quantitation has been successfully used to quantify 1000+ human proteoforms below 30 kDa based on a pregenerated protein identification database [130]. However, one of the major challenges of label-free quantitation is run-to-run LC-MS instrument variation, which may result in inconsistent analysis among multiple runs. In addition, multidimensional separation techniques, which have been applied for increasing the throughput of top-down MS, may split a single proteoform into different fractions and make accurate, label-free quantitation difficult.

Various isotope-labeling techniques, both *in vivo* and *in vitro*, have also been applied to quantitative, top-down proteomics. Stable isotope labeling by amino acids in cell culture (SILAC) is an *in vivo* metabolic labeling technology that relies on the incorporation of isotopically labeled amino acids in the growth medium during cell growth [54, 131]. In top-down SILAC MS, peak detection and spectral deconvolution are challenging because multiple isotopic distributions in parent scans are often partially overlapped [59]. The spectral complexity can be further increased by even a small amount of incomplete labeling. More recently, a neutron coding (NeuCode) SILAC method has been introduced to incorporate isotope-labeled amino acids with subtle mass differences. High-resolution MS instruments (*i.e.*, mass resolution >120 K) are coupled with NeuCode SILAC and can alleviate some of the drawbacks of the traditional SILAC approach for the quantitation of intact proteoforms [83]. However, like other *in vivo* isotope-labeling approaches, the application of NeuCode SILAC is largely limited to lab-cultured samples. Alternatively, *in vitro* isotope-labeling approaches can be

applied to lab-cultured or clinical samples. For example, the pseudo isobaric dimethyl labeling (pIDL) approach utilizes isotopically labeled formaldehyde to label lysine residues, and protein quantification is based on the fragment ion pairs [132, 133].

Isobaric labeling, such as isobaric tagging for relative and absolute quantification (iTRAQ) [63], tandem mass tags (TMT) [62], and *N*,*N*-dimethyl leucine (DiLeu) [64, 134], has become one of the most applied quantitative approaches in bottom-up proteomics [135, 136]. Currently, isobaric labeling approaches can simultaneously quantify up to 21 biological samples in a single run, which largely increases the throughput and reduces the need to control run-to-run instrument variation [64]. With isobaric chemical tags, signals across different fractions from multidimensional separation can be normalized, which increases the accuracy of quantification and enables deeper identification and quantification than can be achieved using one-dimensional separation. Coupling isobaric labeling and multidimensional separation could improve protein quantification in complex biological samples such as cell lysates; however, under typical labeling conditions, intact proteins are challenging to label using isobaric chemical tags. In fact, isobaric chemical tag labeling of intact proteins has only been achieved using pure proteins or simple protein mixtures [70, 137-139]. Previously, TMT label-based protein quantification of complex protein mixtures such as cell lysates has not been achieved, and one major challenge with intact complex sample labeling is that proteins tend to precipitate and "crash" out of solution due to the organic solvents required for chemical labeling.

We applied a protein-level TMT labeling platform to quantify intact proteins and proteoforms in complex protein samples, such as *Escherichia coli* cell lysates, using top-down proteomics. To reduce sample complexity and to remove large proteins in the complex sample, we developed a filter-SEC approach that combines a molecular weight

cutoff (MWCO) step with high-performance size exclusion chromatography (SEC) separation. Our results indicated that the filter-SEC approach efficiently enriched proteins with molecular weight (MW) less than 30 kDa and, most importantly, avoided protein precipitation for efficient labeling. We further combined filter-SEC enrichment, intact protein-level TMT labeling, and top-down MS to identify and quantify low MW proteins and proteoforms (*i.e.*, < 30 kDa) in an *E. coli* cell lysate. A total of 411 intact proteoforms of 95 proteins were identified and quantified from two LC-MS/MS runs after manual evaluation. Our results demonstrate that the proposed protein-level TMT labeling approach can efficiently label and quantify proteoforms in complex protein samples.

3.3 Materials and Methods

3.3.1 Chemicals and Reagents

Acetonitrile (ACN), trifluoroacetic acid (TFA), formic acid, water, triethylammonium bicarbonate (TEAB), 50% (w/w) hydroxylamine, iodoacetamide (IAA), phenylmethanesulfonylfluoride (PMSF), and ammonium bicarbonate were purchased from Sigma-Aldrich (St. Louis, MO, USA). ACN, TFA, and water used for LC mobile phases were LC-MS grade. Tris (2-carboxyethyl) phosphine hydrochloride (TCEP), BCA assay, Bradford assay, and the TMT 6-plex isobaric label reagent set were obtained from Thermo Fisher (Waltham, MA, USA). Molecular weight cutoff spin filters were purchased from Sartorius (Gottingen, Germany).

3.3.2 *E. coli* Cell Lysate Preparation

E. coli K12 cells were incubated in LB medium at 37 °C for 12 h. The cells were collected and pelleted through centrifugation (10000*g*) at 4 °C for 20 min. Cell lysates were prepared using the bead-beating approach as previously described [140]. Cell lysis buffer (1 mM PMSF and 25 mM NH_4HCO_3), cell pellets, and 0.1 mm zirconia/silica beads were mixed at a ratio of 2:1:1 (v/v/v), and the cells were lysed in a bullet blender (Next Advance, Atkinson, NH) at 4 °C for 3 min. After bead-beating, the cell debris was removed by centrifugation at 20000*g* and 4 °C for 40 min. Protein concentration of collected cell lysates was measured by a BCA assay before storage at −80 °C.

3.3.3 Filter-SEC Separation

Large proteins in cell lysates were removed using a 100 kDa MWCO spin filter at 4 °C. After filtration, the low MW protein fraction was concentrated to 8 μg/μL using a 3 kDa MWCO filter (low MWCO sample) or a 10 kDa MWCO filter (high MWCO sample). 100 μL of each sample was injected onto a BioSuite UHR column for high-pressure SEC separation (UHR SEC 250, 4.6 mm i.d., 30 cm length, 4 μm particle size). The mobile phase was 150 mM ammonium acetate (pH 7.4), and the flow rate was 200 μL/min. Thirty (30) 1 min fractions were collected. These fractions were analyzed using SDS-PAGE before TMT labeling and top-down MS analysis.

3.3.4 Protein-level TMT Labeling

SEC fractions were combined into two samples based on the molecular weight range determined by SDS-PAGE analysis. For the protein-level labeling, the protein concentration of samples was adjusted to 1 μg/μL. This is comparable to previous literature which reports protein concentrations between 1–5 μg/μL before reduction and

alkylation [70, 138, 139]. 300 μg of the proteins were diluted to 300 μL using 150 mM ammonium acetate, reduced with 10 μL of 200 mM TCEP for 1 h at 55 °C, and alkylated with 20 μL of 375 mM iodoacetamide (IAA) at room temperature in the dark for 30 min. The samples were then buffer exchanged with 100 mM TEAB (pH 8.5). A total of 200 μg of each protein sample was diluted to 300 μL (protein concentration: 0.67 μg/μL) using 100 mM TEAB (pH 8.5) and separated into four aliquots with a volume ratio of 5:2:2:1. A total of 41 μL of ACN was added to each TMT reagent vial. Each aliquot of protein sample was mixed at room temperature with the corresponding TMT reagent at a protein-to-TMT reagent ratio of 1:4 (w/w). The reaction was quenched after 1 h by incubation with 8 μL of 5% hydroxylamine at room temperature for 15 min.

3.3.5 Top-Down LC-MS

Labeled proteins were injected onto a home-packed C5 RPLC capillary column (75 μm i.d., 70 cm length, Jupiter particles, 5 μm diameter, 300 Å pore size) using a Thermo Scientific (Waltham, MA, USA.) Accela LC System, as previously described [17, 26]. The packing materials for the C5 column (Jupiter particles, 5 μm diameter, 300 Å pore size) were purchased from Phenomenex (Torrance, CA). Mobile phase A was 0.01% TFA, 0.585% acetic acid, 2.5% 2-propanol, and 5% acetonitrile in water. Mobile phase B was 0.01% TFA, 0.585% acetic acid, 45% 2-propanol, and 45% ACN in water. A 200 min gradient from 10% to 65% mobile phase B was applied at a flow rate of 400 nL/min. The LC was coupled online to an LTQ Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) through a customized nano-ESI interface. The temperature of the inlet capillary was 300 °C, and the electrospray voltage was 2.6 kV. Full MS scans used a resolving power setting of 100 000 (at *m*/*z* 400) with two micro scans. Higher energy collisional induced dissociation (HCD) with a normalized collision

energy of 30% was applied. The MS/MS data were collected using data-dependent acquisition and the top six most abundant precursor ions were selected. The MS/MS data were obtained at a resolving power setting of 60 000 (at *m*/*z* 400) with two micro scans and an isolation window of 3.0 *m*/*z* and 60 s dynamic exclusion. Electron transfer dissociation (ETD) was applied for the MS/MS acquisition as well. The activation time used for ETD was 20 ms. The ETD MS/MS data were obtained at a resolving power setting of 100 000 (at *m*/*z* 400). All other parameters were the same as used for HCD. All the data were collected with the Xcalibur 3.0 software (Thermo Fisher Scientific, Bremen, Germany).

3.3.6 Data Analysis

The top-down proteomics raw data were converted to mzXML files using MSConvert [141], deconvoluted using Biopharma Finder (Thermo Fisher Scientific), and searched against the annotated *E. coli* protein database (UniProt 2017-03-18) using TopPIC Suite [98]. TMT on lysine residues and the N-terminus were set as fixed PTMs, as well as carbamidomethylation of cysteine residues. The error tolerance was set to 15 ppm, the maximum number of unexpected mass shifts in a proteoform spectrum-match was 1, and the maximum mass shift was 500 Da. Detailed parameters for TopPIC can be found in. MASH Suite [99] and ProSight Lite [100] were used for manual interpretation and spectrum presentation. A python program was developed in-house to extract the ion intensities of individual TMT reporter ions from the corresponding MS/MS spectrum in the converted mzML files [142, 143].

The labeling efficiency was calculated from the deconvolution and identification results in top-down analysis using an in-house python package. Proteoforms with mass shifts of ±229.163 Da (TMT6-plex monoisotopic mass) were grouped together as a

"protein group" if the mass shift was less than 200 ppm and the retention time shift was less than 6 min. Here, a protein group indicates the same proteoform with different numbers of TMT labels. The identified proteoforms from TopPIC Suite were utilized to assign the TMT labeling status of each proteoform in a protein group. The labeling efficiency was calculated using the following equation (*I*c: intensity of completely labeled proteoform; *I*o: intensity of overlabeled proteoforms; *I*u: intensity of underlabeled proteoforms):

$$labeling\ efficiency = \frac{ic}{ic + iu + io}$$

3.4 Results and Discussion

3.4.1 Low Molecular Weight Protein Enrichment Using the Filter-SEC

To enrich low molecular weight proteins, a 100 kDa MWCO spin filter was used to remove high MW proteins, and the remaining low MW fraction was concentrated using a 3 kDa MWCO spin filter. A sample of low molecular weight enriched proteins prepared with the MWCO filters and an analogous unfiltered sample was injected individually onto the SEC column to compare the separation efficiency (**Figure 3.1 a, b**). The MWCO filtering step significantly improved the SEC separation of low MW proteins in this complex sample and decreased the background and peak broadening caused by the high abundance of large MW proteins in the unfiltered samples. In addition, the filter-SEC allowed for the enrichment of low MW proteins resulting in the injection of more low MW proteins in a single LC-MS run. After the filter-SEC step, low MW fractions (fraction 19–26) were combined into two samples based on the molecular weight as determined by SDS-PAGE analysis. Sample 1 contained fractions

19–21 (<66 kDa), and Sample 2 contained fractions 22–26 (<35 kDa). These two samples were used for the subsequent intact protein TMT labeling. In addition, several replicate experiments were performed to demonstrate the reproducibility of the filter-SEC (**Figure 3.1 c**).

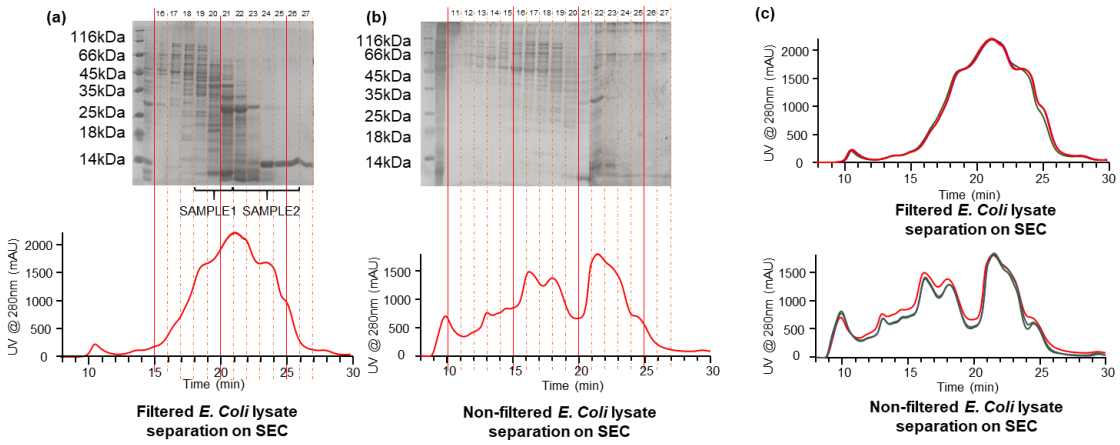


Figure 3.1 SEC separation performance evaluation.

The SEC separations are compared between (a) filtered *E. coli* proteins and (b) non-filtered *E. coli* proteins. (c) Multiple filter-SEC runs were compared to determine reproducibility. The same amount of filtered and non-filtered *E. coli* protein was separated by an SEC column and the fractions (from red traces) were evaluated using SDS-PAGE.

3.4.2 Intact Protein-level TMT Labeling for the Low MW Proteins Enriched by Filter-SEC

A common problem for intact protein TMT labeling is the reduction in solubility of the intact proteins after reduction and alkylation of the disulfide bonds. However, the cleavage of disulfide bonds is important for efficient protein-level TMT labeling, so we examined the effect of the cleavage of disulfide bonds on protein solubility under the labeling conditions (**Figure 3.2 a**). The protein recovery after reduction and alkylation

of disulfide bonds for a nonfiltered sample of *E. coli* cell lysate was approximately 60%, but there was no significant loss of protein for the filter-SEC prepared low MW samples. There was also visible protein precipitation in the nonfiltered sample after the reduction and alkylation, while no visible precipitation formed in the filter-SEC prepared low MW samples (fractions 19–26). These results indicate that the filter-SEC enrichment of low MW proteins prevents protein precipitation under labeling conditions.

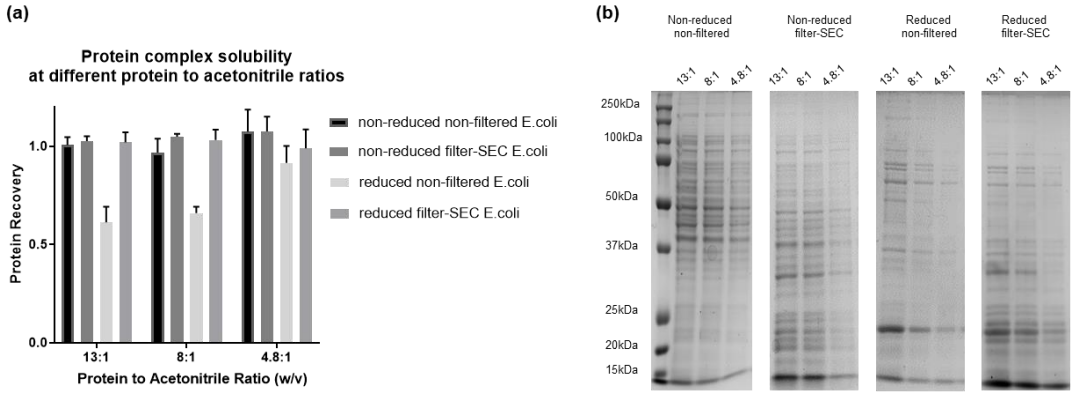


Figure 3.2 Evaluation of protein solubility under different TMT labeling conditions.

Reduced (reduced by TCEP and alkylated by iodoacetamide) and nonreduced *E. coli* proteins were mixed with acetonitrile at different ratios. (a) Protein concentrations of the supernatants were measured using the BCA assay. (b) The protein content of each supernatant was evaluated using SDS-PAGE.

We then evaluated the ratio of the labeling reagent to protein for TMT labeling. Since the organic solvent (ACN) in the reaction solution may cause the proteins to precipitate, we evaluated the protein solubility in ACN with no TMT reagent added at varying protein concentrations. According to the manufacturer instructions, TMT reagent should be dissolved in ACN at a ratio of 800 μg of TMT to 41 μL of ACN. A sample-to-reagent ratio between 1:1 and 1:8 (w/w) is suggested for peptide labeling, which corresponds to a peptide-to-ACN ratio from 19.2:1 to 2.4:1 (w/v) [144]. We

analyzed the percent recovery of samples with protein-to-ACN ratios of 13:1, 8:1, and 4.8:1 (w/v), which is in the middle of the range of recommended ratios for peptide labeling. In the filter-SEC low MW samples, an average protein recovery of 98.9% was observed with no visible precipitation.

The protein patterns of the soluble proteins were analyzed by SDS-PAGE before and after reduction and alkylation for the filter-SEC and nonfiltered samples, and no significant changes were observed for the filter-SEC samples under all three protein concentration conditions (**Figure 3.2 b**). The filter-SEC sample also maintained similar protein patterns before and after the cleavage of disulfide bonds as observed using SDS-PAGE; a significant loss of large MW proteins was observed on the SDS-PAGE after the reduction and alkylation of the nonfiltered samples. Our results indicate that the cleavage of disulfide bonds can significantly reduce the solubility of large MW proteins due to the exposure of hydrophobic groups as suggested in previous studies [145]. The filter-SEC procedure removed most of the large MW proteins, and the remaining low MW proteins were soluble before and after the reduction and alkylation of cysteine residues. Moreover, these filter-SEC enriched low MW proteins remained soluble under various protein-to-ACN ratio conditions. In summary, our filter-SEC prevented protein precipitation and preserved similar protein patterns under denaturing TMT labeling conditions.

3.4.3 Intact Protein and Proteoform Identification after TMT Labeling

Top-down LC-MS/MS was performed for the identification of labeled proteoforms in the filter-SEC low MW samples (*e.g.*, Sample 1 and Sample 2, **Figure 3.3 a**). There are 411 proteoforms of 95 proteins identified from these 2 samples: 309 proteoforms from 72 proteins identified in Sample 1 and 222 proteoforms from 60

proteins identified in sample 2 after TMT labeling. The labeling efficiency for identified proteoforms was listed in a heatmap (**Figure 3.3 b**). Each row in the heatmap represents a "protein group", which contains all TMT labeled products from 1 proteoform (*i.e.*, completely/under-/overlabeled products), and the relative intensity of each labeled product is indicated in greyscale. In sample 1, 175 out of 211 protein groups display only the completely labeled products with no detection of improperly labeled side products (**Figure 3.3 b**). 17.1% (36 of 211) of the protein groups demonstrated overlabeled products, and 10.4% (22 of 211) of the protein groups demonstrated underlabeled products. In addition, 88.6% (187 of 211) of the protein groups show the completely labeled product with >80% intensity. In sample 2, 114 of 183 protein groups showed only the completely labeled species, 31.1% (57 of 183) of the protein groups included overlabeled products, and 20.8% (38 of 183) of the protein groups included underlabeled products. 74.9% (137 of 183) of the protein groups show a completely labeled product with >80% intensity. The average labeling efficiency is 87.7% ± 5.8% and 81.3% ± 7.6% in sample 1 and sample 2, respectively, where the average labeling efficiency is the mean of the labeling efficiency of all the protein groups.

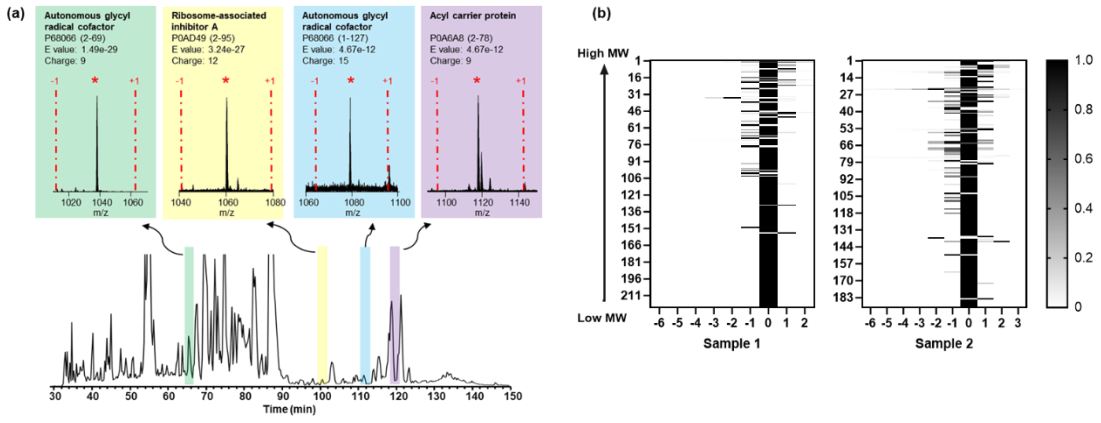


Figure 3.3 Intact protein labeling and identification.

(a) Examples of labeled proteoforms identified in one LC-MS run. The completely labeled forms are indicated by an asterisk and the −1 and +1 represent proteoforms that are underlabeled or overlabeled by one tag. (b) The identification heatmap of samples 1 and 2. The labeling status is indicated for each intact proteoform after labeling.

Overlabeling was observed in these two samples when the NHS-ester reactive group on the TMT labels reacted with amino acid residues other than K. For example, the pK_a of Y (10.3 ± 1.2) is similar to K (10.5 ± 1.1) [146], which suggests that Y may react with TMT under similar pH conditions [144]. The pK_a of H (6.6 ± 1.0) is lower than the N-terminus (7.7 ± 0.5) [146] and, accordingly, the N-imidazole of H is more deprotonated than N-terminus, which promotes TMT labeling on H. Moreover, the hydrogen bonds within the secondary structure increase the nucleophilicity of the hydroxyl-group, leading to the O-acylation of S, T, and Y residues [147, 148]. As suggested in the previous bottom-up study, the optimization of the quenching solution may help to reduce the overlabeled products [144].

We also found that a significantly lower number of proteins were identified with MW larger than 15 kDa when compared with the number of proteins identified smaller than 15 kDa. To evaluate the ability of TMT to efficiently label proteins >15 kDa, we compared the filter-SEC samples before and after TMT labeling (high MWCO sample) using SDS-PAGE. The SDS results suggested that there was no significant loss of larger proteins (*i.e.*, between 20 kDa and 50 kDa) before and after labeling. However, it is still possible that large proteins precipitated after labeling in organic condition and were reconstituted in when SDS was added prior to SDS-PAGE analysis. To remove any proteins that may have precipitated due to addition of organic solvents, the samples were centrifuged at 13 000 rpm for 30 min before SDS-PAGE or LC-MS. Removal of the large proteins was not observed in the centrifuged samples, and, furthermore, continuous injection of these samples did not clog the LC column. This suggests that proteins were

not precipitated under labeling conditions or LC buffer conditions. The low identification number of larger proteins may be related to the MS instrument limitations (*e.g.*, LTQ-Orbitrap). Because of the limited charge capacity of MS instruments, it is a challenge to identify proteoforms >30 kDa as has been discussed previously [53]. Kelleher and co-workers also discussed that protein identifications are dominated by species under 30 kDa due to both inefficient front-end separations and increased difficulty in detection of high mass proteins [12]. Yang et al. reported identified proteoforms that were smaller than 35 kDa from both *E. coli* and HepG2 cells using an CZE separation coupled to Orbitrap QE-HF [149]. Another possible reason for these issues may be the heterogeneity caused by unwanted species such as underlabeled or overlabeled products. These species are often coeluted with each other and/or with other protein species. In ion trap instruments with limited charge capacity, these complications can significantly reduce the signal-to-noise ratio (S/N). With improved MS instrumentation (*e.g.*, ion–ion proton transfer and parallel ion parking [150]) and efficient enrichment/separation approaches for large proteins, we expect that the proposed TMT labeling approach may be applied to proteins with relatively high MW (up to 50 kDa).

Additionally, we noticed that a significant number of proteoform features cannot be identified, even when good quality MS/MS spectra were collected. This may be due to limitations associated with the identification software (TopPIC) which limits the number of dynamic modifications (up to two) without significantly increasing the FDR. The current software considers any N-terminal protein processing (*e.g.*, N-terminal methionine cleavage, signal peptide removal, and N-terminal acetylation) as well as other known or unknown modifications as dynamic modifications. In addition, because we defined TMT as a fixed modification on the lysine residues and N-terminus, any underlabeling or overlabeling events are considered as dynamic modifications. If more

than two underlabeling and/or PTM events are present, our customized TopPIC software cannot automatically identify the detected proteoforms. Therefore, we need to improve the current bioinformatic tools for TMT top-down MS to allow the characterization of unexpected modifications as well as underlabeled or overlabeled proteoforms.

3.4.4 Quantification of Intact Proteins and Proteoforms Using TMT Labeling

To evaluate the quality of protein-level TMT labeling and its quantification accuracy and reproducibility, the filter-SEC sample was separated to four aliquots with a 5:2:2:1 volume ratio, and the aliquots were labeled independently with TMT reagent 128, 129, 130, and 131. The intensity of reporter ions was distributed from 1e2 to 1e6, which suggests a potentially large dynamic range of relative quantification with top-down proteomics. To verify the labeling accuracy, the intensity ratios among each pair of samples were compared (128:131, 129:131, and 130:131) for all the scans in which all reporter ions were detected (**Figure 3.4 a**). The average reporter ion ratios were 6.22 ± 2.17, 2.42 ± 0.97, and 2.60 ± 0.99 for Sample 1 and 5.12 ± 3.33, 2.20 ± 0.88, and 2.40 ± 0.90 for Sample 2. To evaluate the reproducibility of the method, we tested two replicate samples that were individually labeled with TMT 129 and 130. The reporter ion intensities in all the identified MS/MS scans were plotted between these two samples (**Figure 3.4 b, c**), and good linear correlations were observed for both Sample 1 and Sample 2 (R^2 = 0.9922 and 0.9844, respectively). We also examined the quantification results of underlabeled and overlabeled proteoforms and found that the results were similar to completely labeled proteoforms. However, the over/underlabeled species tended to have lower intensities, which may result from lower parent ion intensities of the side products. These results indicate that it is possible to quantify complex protein mixtures with intact protein labeling and top-down proteomics.

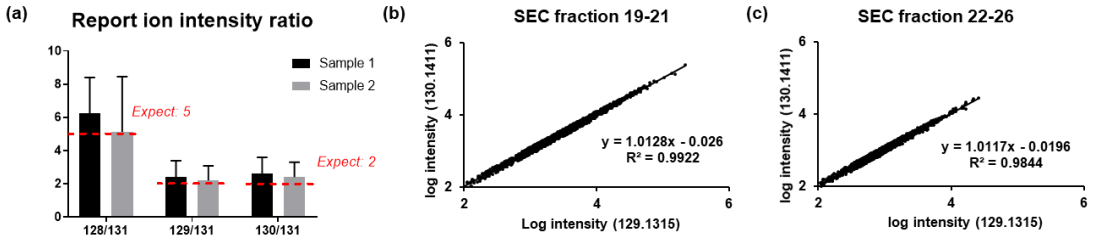


Figure 3.4 Evaluation of protein-level TMT quantification.

(a) The normalized intensity ratio (128/131, 129/131, and 130/131) of the two samples. The error bars represent the calculated standard deviation of reporter ion ratios from all MS/MS scans. The intensity of reporter ions 129 and 130 are correlated to evaluate the reproducibility of the quantification: (b) Sample 1 (fractions 19–21) and (c) Sample 2 (fractions 22–26).

3.4.5 Examples of Protein-level TMT Labeling

Two 50S ribosomal protein L6 (P0AG55) proteoforms were observed (**Figure 3.5**). 50S ribosomal protein L6 contains six lysine residues and an amine group on the N-terminus. As such, the completely labeled proteoform should contain 7 TMT tags. In the labeled sample, one peak that represented the completely labeled proteoform and a small shoulder peak that represented the underlabeled proteoform containing six labels were observed. The HCD MS/MS spectrum confirmed that the proteoform with 7 TMT tags was the completely labeled ribosomal L6 protein. The reporter ions were confidently detected in the HCD spectrum, and the ratio correlated well with the theoretical ratio (5:2:2:1). Another interesting observation is that the charge state distributions were similar between the unlabeled proteoforms and the labeled proteoforms. This observation indicates that the proton affinity on the labeled group is similar to that of the lysine amino group, which is consistent with previous reports [151, 152]. We also identified the proteoform with 6 TMT tags and confirmed that the TMT

underlabeling was either on the N-terminus or Lys-5. Interestingly, reporter ion intensity ratios among different samples correlated well with theoretical ratios suggesting that underlabeling may not affect the quantification results. It has been reported that ETD can also be used to identify and quantify TMT-labeled peptides [153], so we performed an ETD analysis on the same sample. However, the overall fragmentation efficiency was low, and no observable reporter ions were detected in the ETD spectra (**Figure 3.5 d**). Additional optimization or evaluation should be performed for ETD-based top-down TMT experiments.

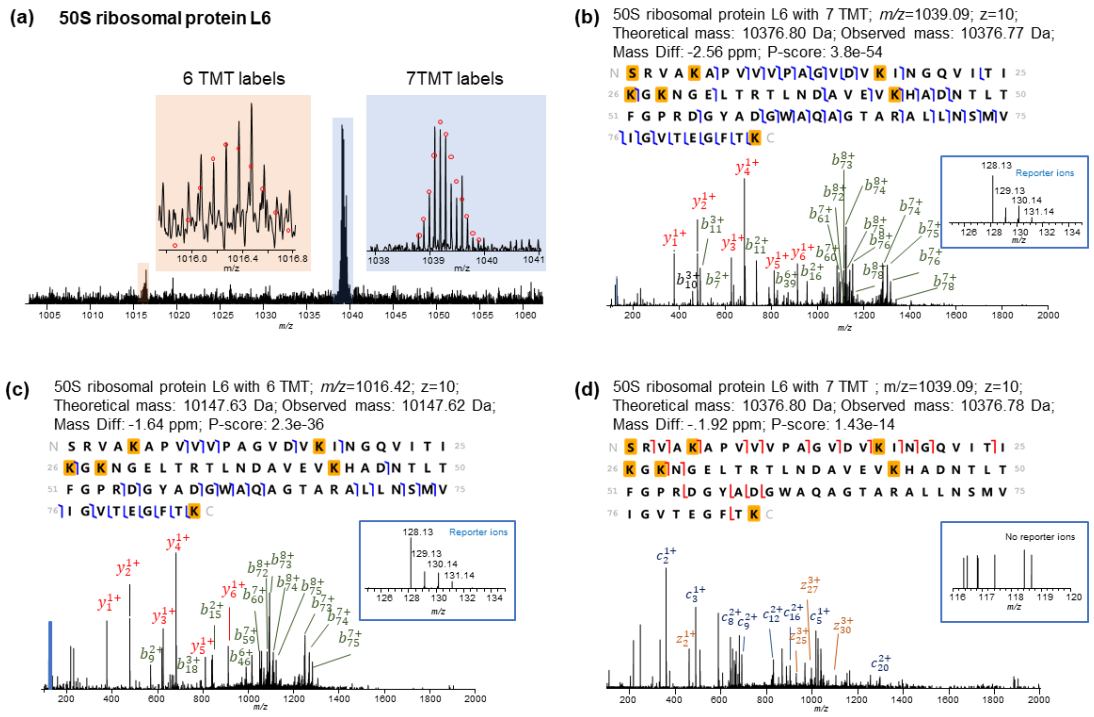


Figure 3.5 Identification of 50S ribosomal protein L6 with 6 and 7 TMT labels.

(a) Full MS spectrum of the protein with 6 and 7 TMT labels. The HCD MS/MS spectra and identification of (b) protein with 7 TMT labels and (c) protein with 6 TMT labels. (d) ETD MS/MS spectrum of protein with 7 TMT labels.

The thiol peroxidase (P0A862) proteoforms were studied to evaluate the quantitative accuracy of overlabeled species (**Figure 3.6 a**). Thiol peroxidase contains

seven lysine residues and an amine group on the N-terminus. In the labeled sample, three peaks indicated an underlabeled proteoform (7 TMT), completely labeled proteoform (8 TMT; **Figure 3.6 b**), and overlabeled proteoform (9 TMT; **Figure 3.6 c**) were observed. The completely labeled and overlabeled proteoforms were confirmed with the HCD MS/MS spectrum. In the overlabeled proteoform, the excess label was located between Ser-1 and His-5 as determined using continuous b ions from MS/MS fragmentation. It has been reported that overlabeling may occur on S/T/Y residues separated by 1 residue from H [144, 146, 148]. Therefore, the overlabeled residue may be Thr-3. The reporter ions were confidently detected in the HCD spectrum for the overlabeled species and the experimental ratio (5.28:2.73:2.79:1) of the overlabeled species correlated well with the experimental ratio of the completely labeled species (5.04:2.43:2.88:1) and the theoretical ratio (5:2:2:1). It is apparent that over/underlabeled side products provide accurate quantitation results and may be used as replicates for quantification.

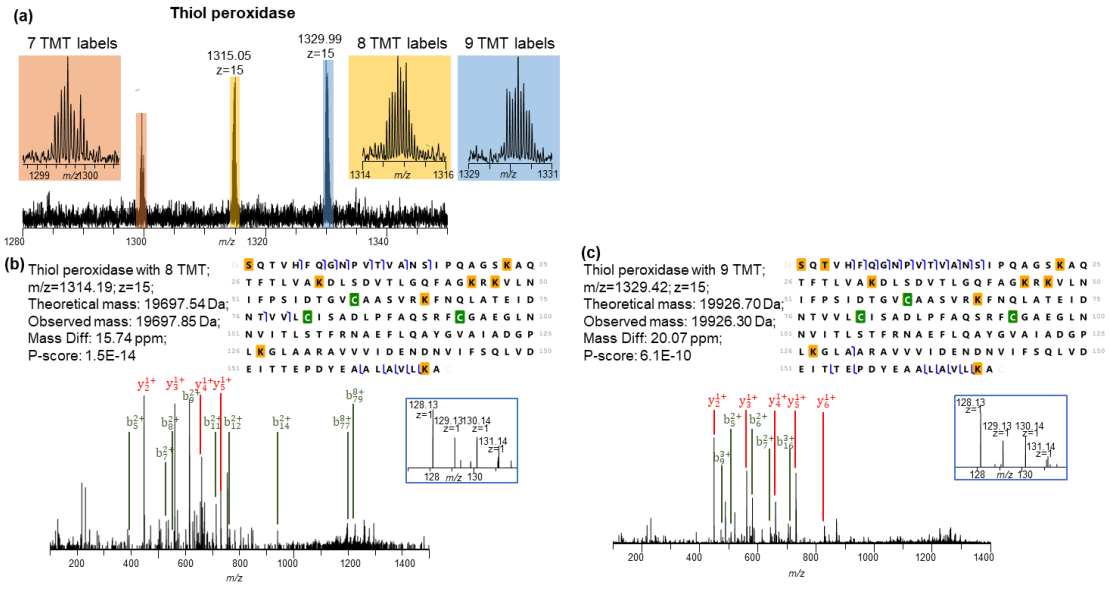


Figure 3.6 Identification of Thiol peroxidase with 9, 8, and 7 TMT labels.

(a) Full MS spectrum of Thiol peroxidase. The HCD MS/MS spectra and identification of (b) protein with 8 TMT labels and (c) protein with 9 TMT labels.

We also evaluated the labeling efficiency on large proteins. The completely labeled ribose import binding protein (RbsB) was detected with all 24 lysine residues and the N-terminus labeled in the MS spectrum (**Figure 3.7 a**) and was confirmed by the MS/MS spectrum (**Figure 3.7 b**). The reporter ions were also detected in the HCD spectrum with good intensities, and the ratio correlated well with the theoretical ratio. However, the underlabeled proteoform that lacked TMT labeling on the N-terminus had a relatively high abundance. Underlabeling of the proteoform indicates the ratio between labeling reagent and protein needs to be optimized for larger proteins (*i.e.*, larger than 20 kDa).

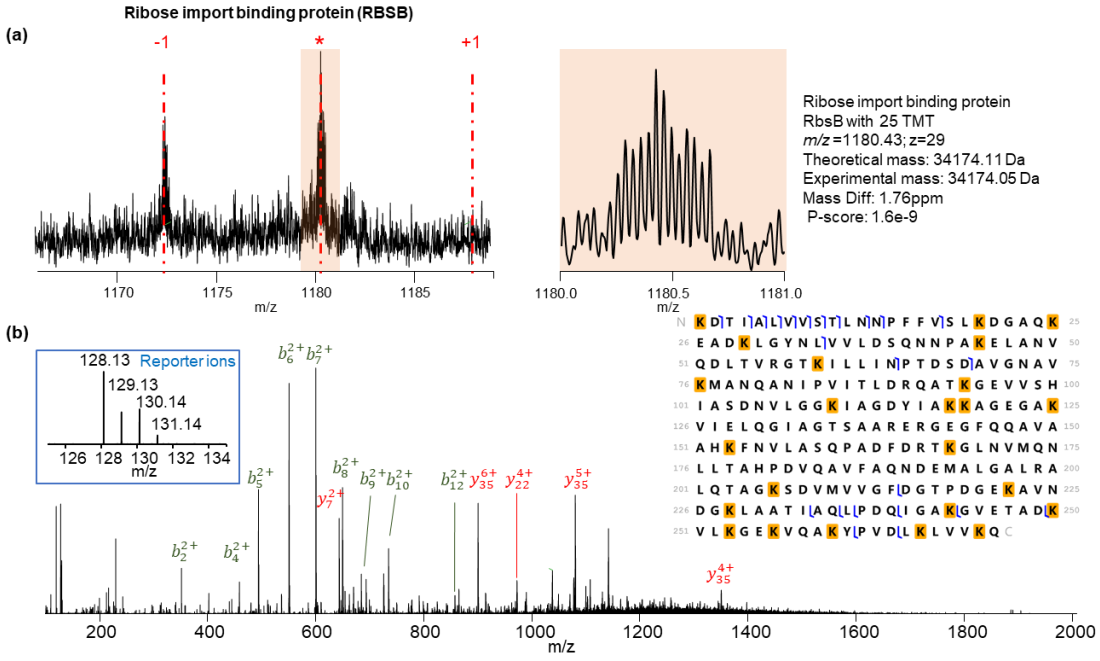


Figure 3.7 Identification of completely labeled ribose import binding protein RbsB.

(a) MS spectrum of the completely labeled proteoform (25 TMT labels) and one underlabeled proteoform (24 TMT labels). (b) HCD MS/MS spectrum and identification of the completely labeled RbsB. The lysine on the N-terminus is labeled on both its side chain and N-terminus primary amine.

3.5 Conclusion

We have developed a method to identify and quantify intact TMT-labeled proteins using top-down proteomics. Our work provides a tool to enrich proteins smaller than 30 kDa, a workflow for intact protein-level TMT labeling, and quantitative top-down analysis of TMT-labeled proteins with high-resolution FTMS instruments (*e.g.*, orbitrap or FTICR). We have demonstrated that most identified proteins were efficiently labeled (*e.g.*, >80% labeling efficiency). Additionally, our quantification results with proteins smaller than 30 kDa were highly replicable, even for proteins that were not completely labeled. Further studies will be focused on the optimization of the protein-level TMT labeling conditions to improve the intact protein labeling efficiency such as TMT reagent to protein ratio, TMT reagent and protein concentrations, quenching solution concentration and pH, labeling time, and reaction buffers. In addition, other separation methods that select for different molecule sizes may be used, such as GELFrEE [13] or serial size exclusion chromatography [20] to enrich low molecular weight proteins. Our filter-SEC method enriches proteins smaller than 30 kDa from the *E. coli* lysate, which excludes some larger proteins in the *E. coli* proteome, so complete proteome analysis cannot be achieved. For larger proteins, different organic solvents can be evaluated to improve sample solubility. In summary, the intact TMT labeling technique extends isobaric labeling quantification to a wide variety of applications in top-down proteomics and could be easily incorporated into many areas of quantitative proteomics research including the discovery of therapeutic proteins and disease biomarkers.

Chapter 4 Optimization of protein-level tandem mass tag (TMT) labeling in the complex sample with top-down proteomics

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

D.Y. and Y.G. contributed equally.

Author contributions: D.Y., Y.G., and S.W. designed research; D.Y., Y.G., and S.W. performed research; D.Y., Y.G., and X.L. contributed data analysis software; D.Y. and Y.G. analyzed data; and D.Y., Y.G., K.A.C. and S.W. wrote the paper.

4.1 Abstract

Isobaric chemical tag labels have been widely applied for the quantification of peptides and proteins in bottom-up proteomic studies. However, until recently, intact protein quantification using isobaric chemical tag labeling methods has been limited. A major issue is the presence of side products (*i.e.*, incomplete labeling or labeling of unintended residues) generated from the labeling reaction. These side products result in complicated MS spectra and challenging data analysis. In this study, we systematically evaluated the TMT labeling reaction parameters, including TMT-to-protein ratios, and pH/concentration of quenching buffer. Our results indicated that: (1) high TMT-to-protein ratio (w/w, 8:1) and (2) high pH/concentration of quenching buffer (pH = 9.5, final concentration 1.2%) were the optimal conditions for intact protein labeling. We also developed software to calculate the relative abundance of labeled products, which provided a method to evaluate the labeling efficiency of intact proteins. This research provides practical guidance for efficient TMT labeling on complex intact protein samples, which can be readily adopted in high-throughput quantification top-down proteomics.

4.2 Introduction

Isobaric chemical tag labeling techniques have been considered to be the "gold standard" in peptide-based protein quantification studies.[154] Various isobaric labeling approaches have been developed including iTRAQ,[138] TMT,[62] and N, N-dimethylleucine (DiLeu).[64] These approaches enable relative quantification using the MS2 spectrum without an increase in MS1 complexity.[154] Meanwhile, protein-level labeling shows advantages over peptide-level labeling, including the potential to utilize

intact protein separation after labeling, and elimination of variability from digestion.[69] Nie et al. applied isobaric tag labeling for serum glycoprotein quantification and evaluated different digestion enzymes, isobaric tags, and organic solvents to optimize the identification and quantification of peptides.[69] TMT labeling was also effectively applied in cells to compare different approaches (protein-level labeling and peptide-level labeling).[155]

Isobaric labeling of intact proteins with top-down proteomics has also been studied for intact proteoform quantification. Intact labeled standard proteins have been previously analyzed using top-down proteomics by Hung et al;[70] however, the applications of protein-level isobaric tag labeling to top-down analysis are limited. First, the organic solvent required for labeling causes protein precipitation which limits labeling efficiency.[69] We have demonstrated a filtering technique in our reported top-down TMT MS platform[71] that prevented protein precipitation induced by labeling by the removal of large proteins in complex samples. The second issue with intact-protein TMT labeling is incorrectly labeled side products are generated during the labeling reaction, including overlabeling (unexpected labeling) and underlabeling (Lys and N-term are not labeled). The overlabeling occurs on residues with a hydroxyl group (Ser, Thr, Tyr) and underlabeling occurs on the N-terminal or lysine residues. We have demonstrated that overlabeling and underlabeling will not affect quantification results; however, the underlabeled or overlabeled products will decrease the signal-to-noise ratio and increase the sample complexity[70, 144]. Winkels et al. reported a strategy using iodoacetyl TMT (iodoTMT) a thiol-reactive reagent that targets less common thiol groups to reduce side reactions, however, iodoTMT sacrificed the coverage of proteins without Cys.[154]

The TMT labeling condition for peptides was optimized by Zecha et al. and the

impact of labeling reaction parameters was optimized to achieve >99% labeling efficiency[144]. Considering protein-level TMT labeling is more complicated than peptide-level labeling, it is necessary to optimize protein-level labeling to achieve more complete labeling. In this study, we systematically optimized the parameters of the protein-level TMT labeling reaction using top-down proteomics to maximize complete labeling. Here, we demonstrated the effect of TMT-to-protein ratios (w/w) and quenching buffer concentration on protein-level TMT labeling efficiency. Our results demonstrate an optimized condition for protein-level TMT labeling using top-down proteomics.

4.3 Materials and Methods

4.3.1 Chemicals and materials

Acetonitrile (ACN), ammonium bicarbonate (ABC), 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), triethylammonium bicarbonate (TEAB, pH 8.5), 50% (w/w) hydroxylamine, Tris hydrochloride (Tris-HCl), dithiothreitol (DTT), iodoacetamide (IAA), phenylmethanesulfonylfluoride (PMSF), trifluoroacetic acid (TFA), formic acid (FA), water and other reagents were purchased from Sigma-Aldrich (St. Louis, MO, USA) unless otherwise noted. ACN, 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).

4.3.2 Protein sample preparation before labeling

The *E. coli* cell lysate was prepared as previously described.[156] Both 100 K and 10 K MWCO were pre-washed using MiliQ water 3 times, then washed with 100 mM TEAB buffer (or 100 mM HEPES at pH 8.5) 3 times through centrifugation at 12,000 × g and 4 °C. The pre-washed 100 K MWCO spin filter was utilized to first remove the large proteins in *E. coli* cell lysate through centrifugation at 12,000 × g and 4 °C for 20 minutes. Then the filtrate was collected into a pre-washed 10 K MWCO spin filter for the removal of small proteins and to concentrate protein by centrifugation at 12,000 × g and 4 °C for 15 minutes. The concentration of the filtered protein sample was determined by Pierce™ BCA Protein Assay Kit (Thermo) and the proteins were aliquoted and stored at -80 °C for future use. Right before labeling, the protein sample was thawed on ice for 20 minutes and diluted to 1 μg/μL using 6 M fresh urea in TEAB (pH=8.5) to the final urea concentration of 3 M. Then 800 μg proteins were reduced by adding 160 μL 0.5 M TCEP (40 mmol), incubating for 15 minutes at room temperature to break disulfide bonds, alkylated by adding 215 μL of 375 mM IAA (80 mmol) incubating at room temperature at dark for 30 minutes. Then the alkylated proteins were buffer-exchanged and concentrated with 100 mM TEAB (pH = 8.5) to the appropriate concentration for TMT labeling (e.g, 1.5 μg/μL) The prepared samples were stored at -20 °C until future use.

4.3.3 Protein-level TMT labeling.

TMT6-plex reagent or TMTzero (Thermo Fisher) was dissolved in 100% anhydrous acetonitrile, then transferred to protein sample tubes. The respective volumes, concentrations, and reaction conditions are specified in the results section. The protein-TMT mixture was incubated for 1 hour at 25 °C and 400 rpm on a thermomixer. The

labeling reaction was quenched at 25 ºC and 400 rpm by adding 5% hydroxylamine to the final concentration of 0.6% (or 1.2% and 0.3% for quenching buffer optimization) or 1 M Tris (pH = 8) to the final concentration of 50 mM for 15 minutes. Then the sample was centrifuged at 12,000 × g and 4 ºC for 30 minutes to remove any precipitation before LC-MS/MS analysis. Also, after labeling, part of the sample was digested for bottom-up analysis.

4.3.4 Top-down RPLC-MS/MS

The TMT-labeled protein samples were first diluted to 0.4 μg/μL for sample loading. Individually, 10 μg of TMT-labeled protein sample from each optimization condition 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. [28, 96] Specifically, the mobile phase A was 0.01% TFA, 0.585% acetic acid, 2.5% isopropanol, and 5% acetonitrile in water. The 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% MPB was applied with a flow rate of 400 nL/min. The LC elution was sprayed from an etched capillary tip through a customized nano-ESI interface to an LTQ Orbitrap Elite mass spectrometer under positive 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 used the 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 at a resolving power setting of 60,000 (at m/z 400) with two micro scans and an isolation

window of 5.0 m/z. Ions with less than 4 charges were rejected for the selection of MS/MS scans. Collision-induced dissociation (CID) with a normalized collision energy of 35 eV was applied. The maximum injection time was set as 1000 ms for full mass scans and 800 ms for MS/MS scans. The AGC target was set as 1×106 for full mass scans and 6×105 for MS/MS scans. All the data was collected using Xcalibur 3.0 software (Thermo Fisher Scientific, Bremen, Germany).

4.3.5 Data analysis

The protein species in the raw data were deconvoluted using Biopharma Finder (Thermo) using the sliding window mode with parameters shown below: max RT gap was 5 minutes, target average spectrum width was 1 minute and target average spectrum offset was 25%, and merge tolerance was 50 ppm. All the other parameters were set as default. For identification, the top-down proteomics raw data were searched against the annotated *E. coli.* protein database (UniProt 2019-10-13, 4519 species) using TopPIC Suite.[133] Below are the parameters used for TopFD: MS1 signal-to-noise ratio was 3, the precursor window size was 5.0 *m/z* and the maximum mass was 50,000 Dalton, decoy database searching was applied and FDR = 0.01 was set as the cutoff value in both spectrum and proteoform level. The maximum number of mass shifts was 2 and the mass shift range was ± 500 Dalton. The TMT labeling at N-terminal and lysine residue was set as fixed PTM. All the other parameters were default.

To estimate the relative change in the labeling efficiency between different samples for each set of optimization parameters, an in-house python package was developed. Identification heatmap was generated for all identified species with actual labeling status to evaluate the labeling efficiency of different samples in each optimization group. First, the Biopharma Finder (Thermo Fisher Scientific)

deconvolution results for samples in the same optimization group were combined if the mass shift was less than 35 ppm and the retention time shift was less than 6.5 min. Second, proteoforms with a mass shift of ± 1.007825 Da (H^1 monoisotopic mass) were combined and filtered if mass shifts were less than 35 ppm and retention time shifts were less than 6.5 min. Third, proteoforms with mass shifts of ± 229.163 Da for TMT6-plex or ± 224.152 Da for TMTzero were grouped in the same "protein group" with the cutoffs 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 statuses. Fourth, the most abundant species in the control sample was considered as a reference to evaluate the relative labeling efficiency shift among different samples. The proteoforms with masses that are decreased by the mass of one or more TMT labels (229.163 Da for TMT 6-plex or 224.152 Da for TMT-zero) compared to reference proteoforms were considered to be underlabeling and proteoforms with masses that are increased by the mass of one or more TMT labels compared to the reference proteoforms were considered to be overlabeling.

Also, MS raw data were converted to mzML files by using MSConvert.[141] TopPIC Suite was then utilized for mass spectral identification and the actual TMT labeling status was checked manually. Later, the identification results were matched with protein groups in deconvolution results based on monoisotopic mass and retention time to generate the identification heatmap. Here, 0 was used to represent complete labeling in the heatmap, negative integers were used to represent underlabeling (*e.g.,* -1, -2), and positive integers were used to represent overlabeling (*e.g.,* 1, 2). The protein group number is used as the y-axis aligning 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.

4.4 Results and Discussion

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

First, the TMT-to-protein mass ratio was varied (1:1, 2:1, 4:1, and 8:1) to evaluate the effect on intact protein labeling efficiency (8:1 is the ratio recommended by the manufacturer for peptide level labeling). Briefly, 45 µg of *E. coli* proteins were labeled with 45 µg, 90 µg, 180 µg, and 360 µg of TMT6-plex reagent. The final protein concentration was 0.8 µg/µL and the final TMT concentration was varied from 2.3 mM to 18.2 mM.

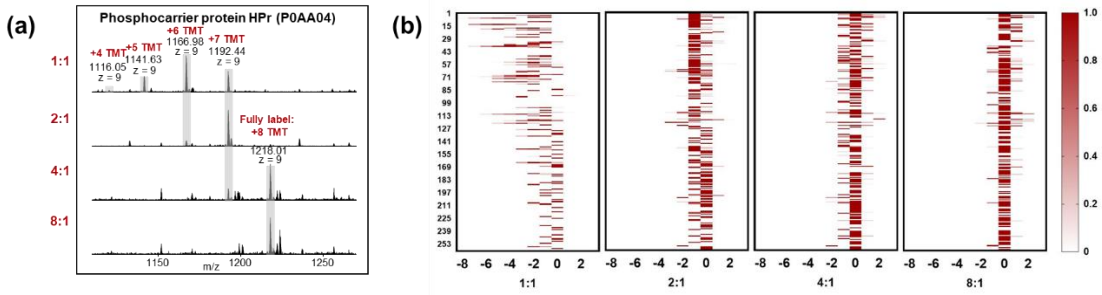


Figure 4.1 Optimization of TMT-to-protein mass ratio with fixed protein amount.

(a) Phosphocarrier protein HPr labeled using different TMT-to-protein ratios. (b) Identification heatmap demonstrating all labeled species observed under each TMT-to-protein ratio condition (aligned with MW).

The MS spectrum example of phosphocarrier protein HPr (Uniprot: P0AA04) showed multiple labeled products under different conditions, which demonstrated the effect of different TMT-to-protein ratios on the labeling status. We observed that the number of TMT labels on the protein increased from 4 to 8 (completely labeled) as the concentration of TMT reagent increased. The samples labeled at 8:1 primarily showed the completely labeled products, which suggested the 8:1 ratio limited under labeling.

To statistically evaluate the labeling efficiency for all identified proteoforms, a heatmap was generated to show the production and identification of the

completely/over-/underlabeled products. LC-MS/MS analysis of the variable TMT-to-protein ratio samples identified 747 unique proteoforms of 99 proteins among the 4 samples. In total, 253 protein groups were clustered (**Figure 4.1 b**). Our results indicate increasing the TMT-to-protein ratio increases the percentage of complexly labeled products.

Furthermore, to verify our labeling efficiency evaluation, the intact labeled proteins were digested and analyzed using bottom-up proteomics methods. The bottom-up data confirmed the trend observed in the top-down evaluation: as the TMT-to-protein ratio increased from 1:1 to 8:1, the percentage of peptides that exhibited an overlabeled product increased (0.0%, 0.9%, 6.1%, and 7.1%, respectively). Further, the percentage of peptides that exhibited an underlabeled product decreased (18.2%, 8.7%, 4.3%, and 0.8%, respectively).[144] We also used the bottom-up data to further confirm the location of under labeling. There were a total of 29 unique peptides that were underlabeled, in which 12 out of 29 (41.4%) were underlabeled at the N-terminus. In the sample labeled with a 2:1 ratio, 10 out of 14 (71.4%) unique peptides were under labeled at the N-terminus. In the samples labeled with a 4:1 ratio, 5 out of 7 (71.4%) unique peptides were under labeled at the N-terminus. In the sample labeled with an 8:1 ratio, 0 out of 1 (0.0%) unique peptides were under labeled at the N-terminus. Therefore, we concluded that underlabeling by 1 tag tend to occur on the N-terminal, which may be due to the decreased reactivity of the N-terminal amino group compared with the lysine sidechain due to the difference in pKa values (pKa of N-terminal is 7.7 ± 0.5 and lysine is 10.5 ± 0.1[146]).

In summary, our results suggested that a TMT-to-protein ratio of 8:1 can dramatically increase the intensity of the completely labeled product peak and decrease the overall sample complexity when compared with lower TMT-to-protein

concentration ratios.

4.4.2 Optimization of final quenching pH in the labeling reaction

Quenching with 5% hydroxylamine (final concentration 0.3%) was recommended to stop the TMT labeling reaction in the commercial protocol. Tris-HCl buffer (50 mM, pH 8.0) was also used for quenching TMT reactions in previous literature to evaluate peptide-level labeling.[144] Here, we optimized the final concentration of hydroxylamine (1.2%, 0.6%, and 0.3%) which resulted in varying final pH (9.5, 9.3, and 9.1, respectively). We further evaluated labeled products to determine the effect on labeling efficiency.

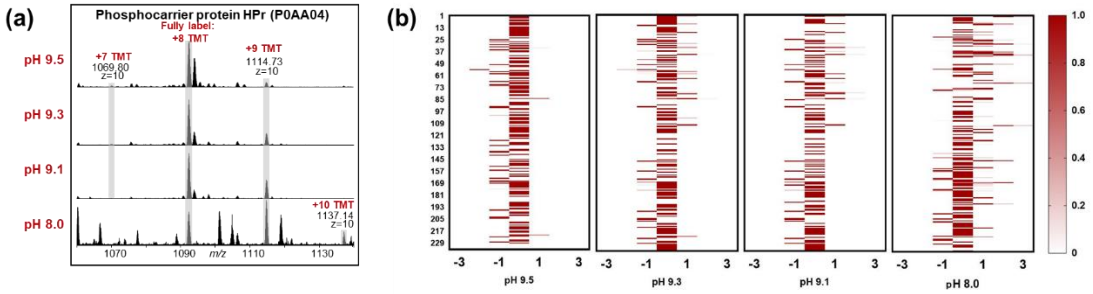


Figure 4.2 Quenching buffer optimization in protein-level TMT labeling.

(a) Phosphocarrier protein HPr labeling products under varying quenching pH conditions. (b) Identification heatmap demonstrating all labeled species observed under each quenching pH condition (aligned with MW).

LC-MS/MS analysis of the different pH samples identified 610 unique proteoforms of 114 proteins among the 4 samples. The MS spectra of Phosphocarrier protein HPr (P0AA04) showed that the completely labeled product was the dominant peak with a small portion of the overlabeled product under pH 9.5, while the percentage of overlabeling increased in other samples. This same protein, when Tris was used as the quenching buffer, demonstrated 2 overlabeled products, and the product that was

over labeled by one tag was the dominant peak. (**Figure 4.2 a**).

To evaluate the labeling efficiency for each identified protein, 235 protein groups were clustered in the identification heatmap (**Figure 4.2 b**). According to the heatmap, a high level of overlabeling occurred at pH 8.0. According to the literature, increasing the pH of the quenching solution may destabilize the O-acylation products on STY resulting in less overlabeled products.[144] From the heatmap, we observed that pH 9.1 to 9.5 resulted in less overlabeling compared with pH 8.0 quenching. This demonstrated that a quenching pH of 9.1 or higher was able to reverse unwanted overlabeling.

It has previously been reported that TMT overlabeling tends to occur on STY amino acids[69, 144]. The side chains of lysine and tyrosine have similar pKa values (Lys: 10.5 ± 1.1; Tyr: 10.3 ± 1.2); [146] therefore, it is expected that tyrosine would also be labeled by TMT at basic pH. However, some studies determined that the O-acylation of tyrosine is favored at acidic pH (pH 6.0), whereas the N-terminal and lysine are prone to react under alkaline conditions (pH 8.4).[144, 157, 158] This phenomenon occurs because the O-acylation of tyrosine product hydrolyzes within a few minutes at basic pH values.[158, 159] Therefore, hydroxylamine quenching buffer can be used to reverse the overlabeling of tyrosine by increasing the pH to ≥9.0 upon quenching.[144, 158, 160, 161]

In conclusion, more proteoforms tend to be overlabeled when the TMT labeling reaction was quenched using Tris-HCl (pH 8.0) when compared with hydroxylamine quenching with pH greater than 9.1. Our results show that adding hydroxylamine could reverse the overlabeling of intact proteins and higher concentrations of hydroxylamine (higher post quenching pH) resulted in fewer overlabeled products.

4.4.3 Application of optimized TMT labeling on intact human proteins

The optimized TMT-to-protein ratio (8:1) and quenching buffer concentration (final 1.2%) were applied to the labeling of intact human protein lysate. 200 μg of *HeLa* lysate proteins were aliquoted and labeled with individual TMT reagents and mixed at a 5:5:2:2:1:1 ratio.

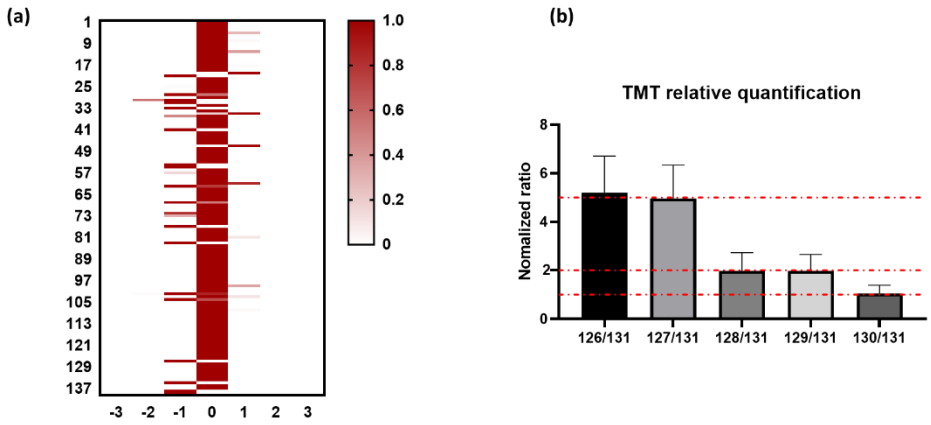


Figure 4.3 Evaluation of optimized protein-level labeling on intact human protein sample.

(a) Identification heatmap demonstrating all labeled species observed (aligned with MW). (b) The normalized intensity ratio (normalized to TMT-131) of the labeled samples. The error bars represent the calculated standard deviation of reporter ion ratios from all MS/MS scans.

LC-MS/MS analysis identified 161 unique proteoforms of 66 proteins from one intact human protein sample. 139 protein groups were clustered in the identification heatmap (**Figure 4.3 a**). According to the heatmap, we observed that most of the protein groups (117 of 139) only have one product detected. This result indicated that our optimized condition dramatically reduced side products in the labeling reaction. The reporter ion intensities were also displayed in the bar graph (**Figure 4.3 b**). The average reporter ion intensities were 5.20 ± 1.50, 4.97 ± 1.37, 1.97 ± 0.76, 1.98 ± 0.67, and 1.05 ± 0.34. This result showed that our method has the potential to quantify complex protein

mixtures with top-down proteomics.

Overall, our optimized protein-level TMT labeling method effectively labeled the *HeLa* lysate protein sample for accurate relative quantitation.

4.5 Conclusion

The protein-level isobaric chemical tag labeling has shown an advantage over the peptide-level labeling. We have previously published the high-throughput quantitative top-down proteomics study using protein-level TMT labeling. Here, we systematically optimized the TMT labeling reaction conditions to achieve better complete labeling of intact proteins. Our work illustrated that a higher TMT-to-protein mass ratio (*i.e,* 8:1) improves labeling efficiency. Also, a high concentration of the quenching solution (*i.e.*, 1.2% of hydroxylamine, final pH 9.5) reverses the unwanted overlabeling. This optimized protein-level TMT labeling method can be widely applied to different intact protein samples for various experimental purposes, including protein biomarker discovery for early diagnosis and therapeutic protein studies.

Chapter 5 Overall summary and future directions

5.1 Overall summary

Overall, the work presented here regarding the high-efficiency separation and protein-level quantitative chemical labeling of complex intact protein samples has contributed to deepen the coverage of intact proteomes and to the progress of quantitative top-down proteomics approaches. The application of 2D pH RP/RPLC-MS to human cell lysate demonstrated the orthogonality of protein separation between high-pH and low-pH RPLC and revealed multiple low abundant proteoforms and PTMs that could not be detected using 1D methods. This displayed the utility of this separation method for deep proteome analysis for further implementation to study biologically relevant and sample-limited systems. To broaden the utility of top-down proteomics, protein-level proteoform quantification with TMT labeling combines advantages from isobaric tag quantification and top-down proteomics. However, the challenges presented by labeling intact proteins including protein precipitation and side reactions had previously limited the application of TMT labeling to intact proteins. The reaction optimization demonstrated here has enabled the application of TMT labeling for top-down proteomics to allow relative intact proteoform quantification of complex protein samples. It has also opened the way for top-down quantitative analysis using multidimensional separation for deeper quantitative proteome coverage. These approaches are powerful tools to investigate the quantification of intact human proteoforms.

The quantitative top-down proteomics approaches discussed here have improved proteome coverage and intact proteoform quantification using isobaric chemical tag

labeling for application to systems biology and functional protein analysis. We hope this approach can be adapted for broad application to enhance functional proteomics methods, such as thermal proteome profiling (TPP)[162] and target identification using drug affinity responsive target stability (DARTS)[163]. The sensitivity of the approach can also be improved with other advanced techniques, for example, the broad-spectrum optimization of selective triggering (BOOST). [138]

5.2 Future directions

Quantitative top-down proteomics has been used to reveal the antigen-binding fragment related to systemic lupus erythematosus[17], and the molecular characterization of hypertrophic cardiomyopathy phenotype[46], but these are label-free quantification applications. Modern high-resolution instruments also allow for isotopic quantification approaches including metabolic and chemical tag labeling. These quantitation techniques improve sample consistency to resolve instrument variation and sample throughput issues to increase the analysis speed and make the results more credible. We believe our work can contribute to the intact proteoform quantification applications with isotopic chemical tag labeling.

With the development of multidimensional LC/MS and intact isobaric chemical tag labeling techniques, we have a set of tools to comprehensively study the human proteome and quantitively analyze changes of intact proteoform abundance. For future applications of the research presented here, we propose the combined application of protein-level TMT labeling and online 2D pH RP/RPLC-MS to study the intact human proteome. The online 2DLC platform greatly reduces the starting sample amount and increases separation efficiency for deeper proteome coverage of limited samples such as

clinical samples. Furthermore, the protein-level TMT quantification can evaluate biological replicates for increased throughput and accuracy. As a result, we hope to quantitatively analyze the intact proteoform changes in relevant biological systems.

Overall, the recent advances in the top-down proteomics field indicate that the intact proteoform analysis is ready for prime time. More and more research groups and laboratories have started to utilize quantitative top-down approaches to investigate the important roles of proteoforms and PTMs in biological regulations, disease mechanisms, and other biochemistry fields. Our presented projects contributed to the quantitative top-down proteomics area, especially the intact human proteoform study and the protein-level quantitative proteomics using TMT.

Still, quantitative top-down proteomics requires further research development. There is still room to improve the components of quantitative top-down proteomics, for example, the improvement of isobaric chemical tags that are water soluble can help simplify the sample preparation. The developments of advanced HRMS instruments provide a better chance for quantitative protein analysis, and new data analysis software also helps to explain the MS data. Also, more extensive application scenarios pose challenges to the development of quantitative top-down proteomics. For instance, limited sample loading for single-cell proteomics, data analysis progress for multi-omics, the salty neutral pH ionization condition for native MS, interlaboratory reproducibility, and processing of large numbers of the dataset.

References

1. 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., How many human proteoforms are there? *Nature chemical biology* **2018,** *14* (3), 206.
2. Smith, L. M.; Kelleher, N. L., Proteoform: a single term describing protein complexity. *Nature methods* **2013,** *10* (3), 186-187.
3. Tyers, M.; Mann, M., From genomics to proteomics. *Nature* **2003,** *422* (6928), 193-197.
4. Burnette, W. N., "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Analytical biochemistry* **1981,** *112* (2), 195-203.
5. Wang, D.; Huang, S.; Liu, P.; Liu, X.; He, Y.; Chen, W.; Hu, Q.; Wei, T.; Gan, J.; Ma, J., Structural analysis of the Hg (II)-regulatory protein Tn501 MerR from Pseudomonas aeruginosa. *Scientific reports* **2016,** *6* (1), 1-9.
6. Wang, H. W.; Wang, J. W., How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Science* **2017,** *26* (1), 32-39.
7. Cordwell, S. J.; Wilkins, M. R.; Cerpa-Poljak, A.; Gooley, A. A.; Duncan, M.; Williams, K. L.; Humphery-Smith, I., Cross-species identification of proteins separated by two-dimensional gel electrophoresis using matrix-assisted laser desorption ionisation/time-of-flight mass spectrometry and amino acid composition. *Electrophoresis* **1995,** *16* (1), 438-443.
8. Laskay, U. n. A.; Lobas, A. A.; Srzentić, K.; Gorshkov, M. V.; Tsybin, Y. O., Proteome digestion specificity analysis for rational design of extended bottom-up and middle-down proteomics experiments. *Journal of proteome research* **2013,** *12* (12), 5558-5569.
9. Liu, T.; Qian, W.-J.; Gritsenko, M. A.; Xiao, W.; Moldawer, L. L.; Kaushal, A.; Monroe, M. E.; Varnum, S. M.; Moore, R. J.; Purvine, S. O., High dynamic range characterization of the trauma patient plasma proteome. *Molecular & Cellular Proteomics* **2006,** *5* (10), 1899-1913.
10. Qu, Y.; Wu, S.; Zhao, R.; Zink, E.; Orton, D. J.; Moore, R. J.; Meng, D.; Clauss, T. R.; Aldrich, J. T.; Lipton, M. S., Automated immobilized metal affinity chromatography system for enrichment of Escherichia coli phosphoproteome. *Electrophoresis* **2013,** *34* (11), 1619-1626.
11. Pringels, L.; Broeckx, V.; Boonen, K.; Landuyt, B.; Schoofs, L., Abundant plasma protein depletion using ammonium sulfate precipitation and protein A affinity chromatography. *Journal of Chromatography B* **2018,** *1089*, 43-59.
12. 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.
13. 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., Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011,** *480* (7376), 254-258.
14. 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. *Molecular & Cellular*

Proteomics **2016,** *15* (7), 2423-2434.
15. Cleland, T. P.; DeHart, C. J.; Fellers, R. T.; VanNispen, A. J.; Greer, J. B.; LeDuc, R. D.; Parker, W. R.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S., High-throughput analysis of intact human proteins using UVPD and HCD on an orbitrap mass spectrometer. *Journal of proteome research* **2017,** *16* (5), 2072-2079.
16. 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. *Journal of The American Society for Mass Spectrometry* **2019,** *30* (12), 2502-2513.
17. 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. *Scientific reports* **2019,** *9* (1), 1-9.
18. Wang, Z.; Yu, D.; Cupp-Sutton, K. A.; Liu, X.; Smith, K.; Wu, S., Development of an Online 2D Ultra-High-Pressure Nano-LC System for High-pH and Low-pH Reversed Phase Separation in Top-Down Proteomics. *Analytical Chemistry* **2020**.
19. Chen, X.; Ge, Y., Ultrahigh pressure fast size exclusion chromatography for top-down proteomics. *Proteomics* **2013,** *13* (17), 2563-2566.
20. 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. *Analytical chemistry* **2017,** *89* (10), 5467-5475.
21. 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. *Analytical chemistry* **2019,** *91* (6), 3835-3844.
22. Gargano, A. F.; Roca, L. S.; Fellers, R. T.; Bocxe, M.; Domínguez-Vega, E.; Somsen, G. W., Capillary HILIC-MS: a new tool for sensitive top-down proteomics. *Analytical chemistry* **2018,** *90* (11), 6601-6609.
23. Gargano, A. F.; 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. *Journal of proteome research* **2018,** *17* (11), 3791-3800.
24. 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.
25. Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C., Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. *Journal of separation science* **2005,** *28* (14), 1694-1703.
26. 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.
27. Yang, F.; Shen, Y.; Camp, D. n.; Smith, R., High pH reversed-phase chromatography with fraction concatenation as an alternative to strong-cation exchange chromatography for two-dimensional proteomic analysis. *Expert Rev. Proteomics* **2012,** *9* (2), 129-134.
28. 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., Top-down proteomics reveals

a unique protein S-thiolation switch in Salmonella Typhimurium in response to infection-like conditions. *Proceedings of the National Academy of Sciences* **2013,** *110* (25), 10153-10158.
29. Lubeckyj, R. A.; McCool, E. N.; Shen, X.; Kou, Q.; Liu, X.; Sun, L., Single-shot top-down proteomics with capillary zone electrophoresis-electrospray ionization-tandem mass spectrometry for identification of nearly 600 Escherichia coli proteoforms. *Analytical chemistry* **2017,** *89* (22), 12059-12067.
30. Shen, X.; Yang, Z.; McCool, E. N.; Lubeckyj, R. A.; Chen, D.; Sun, L., Capillary zone electrophoresis-mass spectrometry for top-down proteomics. *TrAC Trends in Analytical Chemistry* **2019,** *120*, 115644.
31. Link, A. J., Multidimensional peptide separations in proteomics. *Trends in Biotechnology* **2002,** *20* (12), s8-s13.
32. Giddings, J., Concepts and comparisons in multidimensional separation. *Journal of High Resolution Chromatography* **1987,** *10* (5), 319-323.
33. Skinner, O. S.; Do Vale, L. H.; Catherman, A. D.; Havugimana, P. C.; Sousa, M. V. d.; Compton, P. D.; Kelleher, N. L., Native GELFrEE: a new separation technique for biomolecular assemblies. *Analytical chemistry* **2015,** *87* (5), 3032-3038.
34. 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. *Analytical chemistry* **2015,** *87* (10), 5363-5371.
35. 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. *Analytical chemistry* **2018,** *90* (9), 5529-5533.
36. 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. *Analytical chemistry* **2019,** *91* (17), 11085-11091.
37. Tipton, J. D.; Tran, J. C.; Catherman, A. D.; Ahlf, D. R.; Durbin, K. R.; Kelleher, N. L., Analysis of intact protein isoforms by mass spectrometry. *Journal of Biological Chemistry* **2011,** *286* (29), 25451-25458.
38. Compton, P. D.; Zamdborg, L.; Thomas, P. M.; Kelleher, N. L., On the scalability and requirements of whole protein mass spectrometry. *Analytical chemistry* **2011,** *83* (17), 6868-6874.
39. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., Higher-energy C-trap dissociation for peptide modification analysis. *Nature methods* **2007,** *4* (9), 709-712.
40. Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J., Performance characteristics of electron transfer dissociation mass spectrometry. *Molecular & Cellular Proteomics* **2007,** *6* (11), 1942-1951.
41. Zubarev, R. A., Electron capture dissociation LC/MS/MS for bottom-up proteomics. In *Mass Spectrometry of Proteins and Peptides*, Springer: 2009; pp 413-416.
42. Riley, N. M.; Sikora, J. W.; Seckler, H. S.; Greer, J. B.; Fellers, R. T.; LeDuc, R. D.; Westphall, M. S.; Thomas, P. M.; Kelleher, N. L.; Coon, J. J., The value of activated ion electron transfer dissociation for high-throughput top-down characterization of intact proteins. *Analytical chemistry* **2018,** *90* (14), 8553-8560.
43. Brunner, A. M.; Lössl, P.; Liu, F.; Huguet, R.; Mullen, C.; Yamashita, M.; Zabrouskov, V.; Makarov, A.; Altelaar, A. M.; Heck, A. J., Benchmarking multiple

fragmentation methods on an orbitrap fusion for top-down phospho-proteoform characterization. *Analytical chemistry* **2015,** *87* (8), 4152-4158.
44. Rush, M. J.; Riley, N. M.; Westphall, M. S.; Coon, J. J., Top-down characterization of proteins with intact disulfide bonds using activated-ion electron transfer dissociation. *Analytical chemistry* **2018,** *90* (15), 8946-8953.
45. Chen, B.; Guo, X.; Tucholski, T.; Lin, Z.; McIlwain, S.; Ge, Y., The impact of phosphorylation on electron capture dissociation of proteins: a top-down perspective. *Journal of the American Society for Mass Spectrometry* **2017,** *28* (9), 1805-1814.
46. Tucholski, T.; Cai, W.; Gregorich, Z. R.; Bayne, E. F.; Mitchell, S. D.; McIlwain, S. J.; de Lange, W. J.; Wrobbel, M.; Karp, H.; Hite, Z., Distinct hypertrophic cardiomyopathy genotypes result in convergent sarcomeric proteoform profiles revealed by top-down proteomics. *Proceedings of the National Academy of Sciences* **2020,** *117* (40), 24691-24700.
47. Tiambeng, T. N.; Roberts, D. S.; Brown, K. A.; Zhu, Y.; Chen, B.; Wu, Z.; Mitchell, S. D.; Guardado-Alvarez, T. M.; Jin, S.; Ge, Y., Nanoproteomics enables proteoform-resolved analysis of low-abundance proteins in human serum. *Nature communications* **2020,** *11* (1), 1-12.
48. Dang, X.; Singh, A.; Spetman, B. D.; Nolan, K. D.; Isaacs, J. S.; Dennis, J. H.; Dalton, S.; Marshall, A. G.; Young, N. L., Label-free relative quantitation of isobaric and isomeric human histone H2A and H2B variants by Fourier transform ion cyclotron resonance top-down MS/MS. *Journal of proteome research* **2016,** *15* (9), 3196-3203.
49. Geis-Asteggiante, L.; Ostrand-Rosenberg, S.; Fenselau, C.; Edwards, N. J., Evaluation of spectral counting for relative quantitation of proteoforms in top-down proteomics. *Analytical chemistry* **2016,** *88* (22), 10900-10907.
50. Ntai, I.; Kim, K.; Fellers, R. T.; Skinner, O. S.; Smith IV, 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. *Analytical chemistry* **2014,** *86* (10), 4961-4968.
51. Wu, S.; Brown, J. N.; Tolić, N.; Meng, D.; Liu, X.; Zhang, H.; Zhao, R.; Moore, R. J.; Pevzner, P.; Smith, R. D., Quantitative analysis of human salivary gland‐derived intact proteome using top‐down mass spectrometry. *Proteomics* **2014,** *14* (10), 1211-1222.
52. Cupp-Sutton, K. A.; Wu, S., High-throughput quantitative top-down proteomics. *Molecular Omics* **2020,** *16* (2), 91-99.
53. 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., Identification and quantification of proteoforms by mass spectrometry. *Proteomics* **2019,** *19* (10), 1800361.
54. 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. *Molecular & cellular proteomics* **2002,** *1* (5), 376-386.
55. Terzi, F.; Cambridge, S., An overview of advanced SILAC-labeling strategies for quantitative proteomics. *Methods in enzymology* **2017,** *585*, 29-47.
56. Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M., Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006,** *127* (3), 635-648.
57. Waanders, L. F.; Hanke, S.; Mann, M., Top-down quantitation and characterization of SILAC-labeled proteins. *Journal of the American Society for Mass*

Spectrometry **2007,** *18* (11), 2058-2064.
58. Veenstra, T. D.; Martinović, S.; Anderson, G. A.; Paša-Tolić, L.; Smith, R. D., Proteome analysis using selective incorporation of isotopically labeled amino acids. *Journal of the American Society for Mass Spectrometry* **2000,** *11* (1), 78-82.
59. 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. *Nature methods* **2013,** *10* (4), 332-334.
60. 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., NeuCode labels for relative protein quantification. *Molecular & Cellular Proteomics* **2014,** *13* (9), 2503-2512.
61. 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. *Analytical chemistry* **2016,** *88* (22), 11139-11146.
62. 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. *Analytical chemistry* **2003,** *75* (8), 1895-1904.
63. Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics* **2004,** *3* (12), 1154-1169.
64. Frost, D. C.; Feng, Y.; Li, L., 21-plex DiLeu isobaric tags for high-throughput quantitative proteomics. *Analytical Chemistry* **2020**.
65. Wang, Z.; Yu, K.; Tan, H.; Wu, Z.; Cho, J.-H.; Han, X.; Sun, H.; Beach, T. G.; Peng, J., 27-Plex tandem mass tag mass spectrometry for profiling brain proteome in alzheimer's disease. *Analytical chemistry* **2020,** *92* (10), 7162-7170.
66. Clark, D. J.; Fondrie, W. E.; Liao, Z.; Hanson, P. I.; Fulton, A.; Mao, L.; Yang, A. J., Redefining the breast cancer exosome proteome by tandem mass tag quantitative proteomics and multivariate cluster analysis. *Analytical chemistry* **2015,** *87* (20), 10462-10469.
67. Chen, C.-J.; Chou, C.-Y.; Shu, K.-H.; Chen, H.-C.; Wang, M.-C.; Chang, C.-C.; Hsu, B.-G.; Wu, M.-S.; Yang, Y.-L.; Liao, W.-L., Discovery of Novel Protein Biomarkers in Urine for Diagnosis of Urothelial Cancer Using iTRAQ Proteomics. *Journal of Proteome Research* **2021**.
68. Navarrete-Perea, J.; Yu, Q.; Gygi, S. P.; Paulo, J. A., Streamlined tandem mass tag (SL-TMT) protocol: an efficient strategy for quantitative (phospho) proteome profiling using tandem mass tag-synchronous precursor selection-MS3. *Journal of proteome research* **2018,** *17* (6), 2226-2236.
69. Nie, S.; Lo, A.; Zhu, J.; Wu, J.; Ruffin, M. T.; Lubman, D. M., Isobaric protein-level labeling strategy for serum glycoprotein quantification analysis by liquid chromatography–tandem mass spectrometry. *Analytical chemistry* **2013,** *85* (11), 5353-5357.
70. Hung, C.-W.; Tholey, A., Tandem mass tag protein labeling for top-down identification and quantification. *Analytical chemistry* **2012,** *84* (1), 161-170.
71. 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. *Journal of the American Society for Mass Spectrometry* **2021**.

72. Hill, J. J.; Tremblay, T.-L.; Corbeil, C. R.; Purisima, E. O.; Sulea, T., An accurate TMT-based approach to quantify and model lysine susceptibility to conjugation via N-hydroxysuccinimide esters in a monoclonal antibody. *Scientific reports* **2018,** *8* (1), 1-14.
73. 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. *Journal of the American Society for Mass Spectrometry* **2021**.
74. Hershko, A.; Heller, H.; Eytan, E.; Kaklij, G.; Rose, I. A., Role of the alpha-amino group of protein in ubiquitin-mediated protein breakdown. *Proceedings of the National Academy of Sciences* **1984,** *81* (22), 7021-7025.
75. Recht, J.; Tsubota, T.; Tanny, J.; Diaz, R.; Berger, J. M.; Zhang, X.; Garcia, B.; Shabanowitz, J.; Burlingame, A.; Hunt, D., Histone chaperone Asf1 is required for histone H3 lysine 56 acetylation, a modification associated with S phase in mitosis and meiosis. *Proceedings of the National Academy of Sciences* **2006,** *103* (18), 6988-6993.
76. Grunstein, M., Histone acetylation in chromatin structure and transcription. *Nature* **1997,** *389* (6649), 349-352.
77. Miao, J.; Lawrence, M.; Jeffers, V.; Zhao, F.; Parker, D.; Ge, Y.; Sullivan Jr, W. J.; Cui, L., Extensive lysine acetylation occurs in evolutionarily conserved metabolic pathways and parasite-specific functions during P lasmodium falciparum intraerythrocytic development. *Molecular microbiology* **2013,** *89* (4), 660-675.
78. Phanstiel, D.; Brumbaugh, J.; Berggren, W. T.; Conard, K.; Feng, X.; Levenstein, M. E.; McAlister, G. C.; Thomson, J. A.; Coon, J. J., Mass spectrometry identifies and quantifies 74 unique histone H4 isoforms in differentiating human embryonic stem cells. *Proceedings of the National Academy of Sciences* **2008,** *105* (11), 4093-4098.
79. Shen, X.; Sun, L., Systematic evaluation of immobilized trypsin-based fast protein digestion for deep and high-throughput bottom-up proteomics. *Proteomics* **2018,** *18* (9), 1700432.
80. Gautier, V.; Boumeester, A. J.; Lössl, P.; Heck, A. J., Lysine conjugation properties in human IgGs studied by integrating high-resolution native mass spectrometry and bottom-up proteomics. *Proteomics* **2015,** *15* (16), 2756-2765.
81. 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 research* **2007,** *35* (suppl_1), D61-D65.
82. Hubbard, T. J.; Aken, B. L.; Beal, K.; Ballester, B.; Cáccamo, M.; Chen, Y.; Clarke, L.; Coates, G.; Cunningham, F.; Cutts, T., Ensembl 2007. *Nucleic acids research* **2007,** *35* (suppl_1), D610-D617.
83. 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., Neutron-encoded mass signatures for quantitative top-down proteomics. *Analytical chemistry* **2014,** *86* (5), 2314-2319.
84. Peng, Y.; Gregorich, Z. R.; Valeja, S. G.; Zhang, H.; Cai, W.; Chen, Y.-C.; Guner, H.; Chen, A. J.; Schwahn, D. J.; Hacker, T. A., Top-down proteomics reveals concerted reductions in myofilament and Z-disc protein phosphorylation after acute myocardial infarction. *Molecular & Cellular Proteomics* **2014,** *13* (10), 2752-2764.
85. Anderson, L. C.; DeHart, C. J.; Kaiser, N. K.; Fellers, R. T.; Smith, D. F.; Greer, J. B.; LeDuc, R. D.; Blakney, G. T.; Thomas, P. M.; Kelleher, N. L.,

Identification and characterization of human proteoforms by top-down LC-21 tesla FT-ICR mass spectrometry. *Journal of proteome research* **2017,** *16* (2), 1087-1096.
86. Karch, K. R.; Coradin, M.; Zandarashvili, L.; Kan, Z.-Y.; Gerace, M.; Englander, S. W.; Black, B. E.; Garcia, B. A., Hydrogen-deuterium exchange coupled to top-and middle-down mass spectrometry reveals histone tail dynamics before and after nucleosome assembly. *Structure* **2018,** *26* (12), 1651-1663. e3.
87. Wu, S.; Yang, F.; Zhao, R.; Tolic, N.; Robinson, E. W.; Camp, D. G.; Smith, R. D.; Pasa-Tolic, L., Integrated workflow for characterizing intact phosphoproteins from complex mixtures. *Analytical chemistry* **2009,** *81* (11), 4210-4219.
88. Bloh, A. M.; Campos, J. M.; Alpert, G.; Plotkin, S. A., Determination of N-formimidoylthienamycin concentration in sera from pediatric patients by high-performance liquid chromatography. *Journal of Chromatography B: Biomedical Sciences and Applications* **1986,** *375*, 444-450.
89. Yang, Y.; Gu, D.; Aisa, H. A.; Ito, Y., Studies on the effect of column angle in figure-8 centrifugal counter-current chromatography. *Journal of Chromatography A* **2011,** *1218* (36), 6128-6134.
90. Chen, B.; Peng, Y.; Valeja, S. G.; Xiu, L.; Alpert, A. J.; Ge, Y., Online hydrophobic interaction chromatography–mass spectrometry for top-down proteomics. *Analytical chemistry* **2016,** *88* (3), 1885-1891.
91. 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. *Analytical chemistry* **2014,** *86* (15), 7899-7906.
92. Zhou, M.; Wu, S.; Stenoien, D. L.; Zhang, Z.; Connolly, L.; Freitag, M.; Paša-Tolić, L., Profiling changes in histone post-translational modifications by top-down mass spectrometry. In *Eukaryotic Transcriptional and Post-Transcriptional Gene Expression Regulation*, Springer: 2017; pp 153-168.
93. Vellaichamy, A.; Tran, J. C.; Catherman, A. D.; Lee, J. E.; Kellie, J. F.; Sweet, S. M.; Zamdborg, L.; Thomas, P. M.; Ahlf, D. R.; Durbin, K. R., Size-sorting combined with improved nanocapillary liquid chromatography− mass spectrometry for identification of intact proteins up to 80 kDa. *Analytical chemistry* **2010,** *82* (4), 1234-1244.
94. Haselberg, R.; de Jong, G. J.; Somsen, G. W., Low-flow sheathless capillary electrophoresis–mass spectrometry for sensitive glycoform profiling of intact pharmaceutical proteins. *Analytical chemistry* **2013,** *85* (4), 2289-2296.
95. Zhao, Y.; Sun, L.; Knierman, M. D.; Dovichi, N. J., Fast separation and analysis of reduced monoclonal antibodies with capillary zone electrophoresis coupled to mass spectrometry. *Talanta* **2016,** *148*, 529-533.
96. Wu, S.; Brown, R. N.; Payne, S. H.; Meng, D.; Zhao, R.; Tolić, N.; Cao, L.; Shukla, A.; Monroe, M. E.; Moore, R. J., Top-down characterization of the post-translationally modified intact periplasmic proteome from the bacterium Novosphingobium aromaticivorans. *International journal of proteomics* **2013,** *2013*.
97. Kelly, R. T.; Page, J. S.; Tang, K.; Smith, R. D., Array of chemically etched fused-silica emitters for improving the sensitivity and quantitation of electrospray ionization mass spectrometry. *Analytical chemistry* **2007,** *79* (11), 4192-4198.
98. 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.
99. 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. *Journal of The American Society for Mass Spectrometry* **2014,** *25* (3), 464-470.
100. 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.
101. 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. *Journal of proteome research* **2016,** *15* (3), 976-982.
102. Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C., Orthogonality of separation in two-dimensional liquid chromatography. *Analytical chemistry* **2005,** *77* (19), 6426-6434.
103. Dwivedi, R. C.; Spicer, V.; Harder, M.; Antonovici, M.; Ens, W.; Standing, K. G.; Wilkins, J. A.; Krokhin, O. V., Practical implementation of 2D HPLC scheme with accurate peptide retention prediction in both dimensions for high-throughput bottom-up proteomics. *Analytical chemistry* **2008,** *80* (18), 7036-7042.
104. Wang, Y.; Yang, F.; Gritsenko, M. A.; Wang, Y.; Clauss, T.; Liu, T.; Shen, Y.; Monroe, M. E.; Lopez-Ferrer, D.; Reno, T., Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **2011,** *11* (10), 2019-2026.
105. Delmotte, N.; Lasaosa, M.; Tholey, A.; Heinzle, E.; Huber, C. G., Two-dimensional reversed-phase× ion-pair reversed-phase HPLC: an alternative approach to high-resolution peptide separation for shotgun proteome analysis. *Journal of proteome research* **2007,** *6* (11), 4363-4373.
106. Baghdady, Y. Z.; Schug, K. A., Qualitative evaluation of high pH mass spectrometry-compatible reversed phase liquid chromatography for altered selectivity in separations of intact proteins. *Journal of Chromatography A* **2019,** *1599*, 108-114.
107. Schaffer, L. V.; Rensvold, J. W.; Shortreed, M. R.; Cesnik, A. J.; Jochem, A.; Scalf, M.; Frey, B. L.; Pagliarini, D. J.; Smith, L. M., Identification and quantification of murine mitochondrial proteoforms using an integrated top-down and intact-mass strategy. *Journal of proteome research* **2018,** *17* (10), 3526-3536.
108. Choi, B.; Hercules, D.; Gusev, A., LC-MS/MS signal suppression effects in the analysis of pesticides in complex environmental matrices. *Fresenius' journal of analytical chemistry* **2001,** *369* (3), 370-377.
109. Pagano, A.; Rovelli, G.; Mosbacher, J.; Lohmann, T.; Duthey, B.; Stauffer, D.; Ristig, D.; Schuler, V.; Meigel, I.; Lampert, C., C-terminal interaction is essential for surface trafficking but not for heteromeric assembly of GABAB receptors. *Journal of Neuroscience* **2001,** *21* (4), 1189-1202.
110. Kim, J.; Petritis, K.; Shen, Y.; Camp II, D. G.; Moore, R. J.; Smith, R. D., Phosphopeptide elution times in reversed-phase liquid chromatography. *Journal of Chromatography A* **2007,** *1172* (1), 9-18.
111. Uechi, T.; Tanaka, T.; Kenmochi, N., A complete map of the human ribosomal protein genes: assignment of 80 genes to the cytogenetic map and implications for human disorders. *Genomics* **2001,** *72* (3), 223-230.
112. Rich, B. E.; Steitz, J. A., Human acidic ribosomal phosphoproteins P0, P1, and P2: analysis of cDNA clones, in vitro synthesis, and assembly. *Molecular and cellular biology* **1987,** *7* (11), 4065-4074.
113. Bargis-Surgey, P.; Lavergne, J. P.; Gonzalo, P.; Vard, C.; Filhol-Cochet, O.; Reboud, J. P., Interaction of elongation factor eEF-2 with ribosomal P proteins. *European journal of biochemistry* **1999,** *262* (2), 606-611.

114. Artero-Castro, A.; Perez-Alea, M.; Feliciano, A.; Leal, J. A.; Genestar, M.; Castellvi, J.; Peg, V.; Ramon y Cajal, S.; LLeonart, M. E., Disruption of the ribosomal P complex leads to stress-induced autophagy. *Autophagy* **2015,** *11* (9), 1499-1519.
115. Sidoli, S.; Lin, S.; Xiong, L.; Bhanu, N. V.; Karch, K. R.; Johansen, E.; Hunter, C.; Mollah, S.; Garcia, B. A., Sequential window acquisition of all theoretical mass spectra (SWATH) analysis for characterization and quantification of histone post-translational modifications. *Molecular & Cellular Proteomics* **2015,** *14* (9), 2420-2428.
116. Kikuchi, G., The glycine cleavage system: composition, reaction mechanism, and physiological significance. *Molecular and cellular biochemistry* **1973,** *1* (2), 169-187.
117. Koyata, H.; Hiraga, K., The glycine cleavage system: structure of a cDNA encoding human H-protein, and partial characterization of its gene in patients with hyperglycinemias. *American journal of human genetics* **1991,** *48* (2), 351.
118. Fujiwara, K.; Okamura-Ikeda, K.; Motokawa, Y., Chicken liver H-protein, a component of the glycine cleavage system. Amino acid sequence and identification of the N epsilon-lipoyllysine residue. *Journal of Biological Chemistry* **1986,** *261* (19), 8836-8841.
119. Zhou, Z. H.; McCarthy, D. B.; O'Connor, C. M.; Reed, L. J.; Stoops, J. K., The remarkable structural and functional organization of the eukaryotic pyruvate dehydrogenase complexes. *Proceedings of the National Academy of Sciences* **2001,** *98* (26), 14802-14807.
120. Hermes, F. A.; Cronan, J. E., Scavenging of cytosolic octanoic acid by mutant LplA lipoate ligases allows growth of Escherichia coli strains lacking the LipB octanoyltransferase of lipoic acid synthesis. *Journal of bacteriology* **2009,** *191* (22), 6796-6803.
121. Jordan, S. W.; Cronan Jr, J. E., The Escherichia coli lipB gene encodes lipoyl (octanoyl)-acyl carrier protein: protein transferase. *Journal of bacteriology* **2003,** *185* (5), 1582-1589.
122. Park, M. H.; Wolff, E. C., Hypusine, a polyamine-derived amino acid critical for eukaryotic translation. *Journal of Biological Chemistry* **2018,** *293* (48), 18710-18718.
123. Saini, P.; Eyler, D. E.; Green, R.; Dever, T. E., Hypusine-containing protein eIF5A promotes translation elongation. *Nature* **2009,** *459* (7243), 118-121.
124. Cano, V. S.; Jeon, G. A.; Johansson, H. E.; Henderson, C. A.; Park, J. H.; Valentini, S. R.; Hershey, J. W.; Park, M. H., Mutational analyses of human eIF5A-1–identification of amino acid residues critical for eIF5A activity and hypusine modification. *The FEBS journal* **2008,** *275* (1), 44-58.
125. Zhang, J.; Guy, M. J.; Norman, H. S.; Chen, Y.-C.; Xu, Q.; Dong, X.; Guner, H.; Wang, S.; Kohmoto, T.; Young, K. H., Top-down quantitative proteomics identified phosphorylation of cardiac troponin I as a candidate biomarker for chronic heart failure. *Journal of proteome research* **2011,** *10* (9), 4054-4065.
126. Chamot-Rooke, J.; Mikaty, G.; Malosse, C.; Soyer, M.; Dumont, A.; Gault, J.; Imhaus, A.-F.; Martin, P.; Trellet, M.; Clary, G., Posttranslational modification of pili upon cell contact triggers N. meningitidis dissemination. *Science* **2011,** *331* (6018), 778-782.
127. Ntai, I.; LeDuc, R. D.; Fellers, R. T.; Erdmann-Gilmore, P.; Davies, S. R.; Rumsey, J.; Early, B. P.; Thomas, P. M.; Li, S.; Compton, P. D., Integrated bottom-

up and top-down proteomics of patient-derived breast tumor xenografts. *Molecular & Cellular Proteomics* **2016,** *15* (1), 45-56.
128. Wei, L.; Gregorich, Z. R.; Lin, Z.; Cai, W.; Jin, Y.; McKiernan, S. H.; McIlwain, S.; Aiken, J. M.; Moss, R. L.; Diffee, G. M., Novel sarcopenia-related alterations in sarcomeric protein post-translational modifications (PTMs) in skeletal muscles identified by top-down proteomics. *Molecular & Cellular Proteomics* **2018,** *17* (1), 134-145.
129. 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. *Analytical chemistry* **2006,** *78* (3), 686-694.
130. Fornelli, L.; Durbin, K. R.; Fellers, R. T.; Early, B. P.; Greer, J. B.; LeDuc, R. D.; Compton, P. D.; Kelleher, N. L., Advancing top-down analysis of the human proteome using a benchtop quadrupole-orbitrap mass spectrometer. *Journal of proteome research* **2017,** *16* (2), 609-618.
131. 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. *Analytical chemistry* **2008,** *80* (13), 4994-5001.
132. Fang, H.; Xiao, K.; Li, Y.; Yu, F.; Liu, Y.; Xue, B.; Tian, Z., Intact protein quantitation using pseudoisobaric dimethyl labeling. *Analytical chemistry* **2016,** *88* (14), 7198-7205.
133. Liu, Z.; Wang, R.; Liu, J.; Sun, R.; Wang, F., Global quantification of intact proteins via chemical isotope labeling and mass spectrometry. *Journal of proteome research* **2019,** *18* (5), 2185-2194.
134. 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. *Analytical chemistry* **2010,** *82* (7), 2817-2825.
135. Cox, J.; Mann, M., Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry* **2011,** *80*, 273-299.
136. Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry* **2007,** *389* (4), 1017-1031.
137. 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.
138. 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.
139. Prudova, A.; auf dem Keller, U.; 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.
140. Chowdhury, S. M.; Du, X.; Tolic, N.; Wu, S.; Moore, R. J.; Mayer, M. U.; Smith, R. D.; Adkins, J. N., Identification of cross-linked peptides after click-based enrichment using sequential collision-induced dissociation and electron transfer dissociation tandem mass spectrometry. *Analytical chemistry* **2009,** *81* (13), 5524-5532.
141. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J., A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology* **2012,** *30* (10), 918-920.
142. Levitsky, L. I.; Klein, J. A.; Ivanov, M. V.; Gorshkov, M. V., Pyteomics 4.0:

five years of development of a Python proteomics framework. *Journal of proteome research* **2018,** *18* (2), 709-714.
143. Goloborodko, A. A.; Levitsky, L. I.; Ivanov, M. V.; Gorshkov, M. V., Pyteomics—a Python framework for exploratory data analysis and rapid software prototyping in proteomics. *Journal of The American Society for Mass Spectrometry* **2013,** *24* (2), 301-304.
144. 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. *Molecular & Cellular Proteomics* **2019,** *18* (7), 1468-1478.
145. Rossky, P. J., Protein denaturation by urea: Slash and bond. *Proceedings of the National Academy of Sciences* **2008,** *105* (44), 16825-16826.
146. Grimsley, G. R.; Scholtz, J. M.; Pace, C. N., A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Science* **2009,** *18* (1), 247-251.
147. Uchida, K.; Kawakishi, S., Sequence-dependent reactivity of histidine-containing peptides with copper (II)/ascorbate. *Journal of Agricultural and Food Chemistry* **1992,** *40* (1), 13-16.
148. Miller, B. T.; Kurosky, A., Elevated intrinsic reactivity of seryl hydroxyl groups within the linear peptide triads His-Xaa-Ser or Ser-Xaa-His. *Biochemical and biophysical research communications* **1993,** *196* (1), 461-467.
149. Yang, Z.; Shen, X.; Chen, D.; Sun, L., Towards a universal sample preparation method for denaturing top-down proteomics of complex proteomes. *Journal of Proteome Research* **2020**.
150. Ugrin, S. A.; English, A. M.; Syka, J. E.; Bai, D. L.; Anderson, L. C.; Shabanowitz, J.; Hunt, D. F., Ion-Ion Proton Transfer and Parallel Ion Parking for the Analysis of Mixtures of Intact Proteins on a Modified Orbitrap Mass Analyzer. *Journal of the American Society for Mass Spectrometry* **2019,** *30* (10), 2163-2173.
151. Bush, M. F.; Oomens, J.; Williams, E. R., Proton Affinity and Zwitterion Stability: New Results from Infrared Spectroscopy and Theory of Cationized Lysine and Analogues in the Gas Phase. *The Journal of Physical Chemistry A* **2009,** *113* (2), 431-438.
152. Thingholm, T. E.; Palmisano, G.; Kjeldsen, F.; Larsen, M. R., Undesirable Charge-Enhancement of Isobaric Tagged Phosphopeptides Leads to Reduced Identification Efficiency. *Journal of Proteome Research* **2010,** *9* (8), 4045-4052.
153. Phanstiel, D.; Unwin, R.; McAlister, G. C.; Coon, J. J., Peptide Quantification Using 8-Plex Isobaric Tags and Electron Transfer Dissociation Tandem Mass Spectrometry. *Analytical Chemistry* **2009,** *81* (4), 1693-1698.
154. Winkels, K.; Koudelka, T.; Tholey, A., Quantitative Top-Down Proteomics by Isobaric Labeling with Thiol-Directed Tandem Mass Tags. *Journal of Proteome Research* **2021,** *20* (9), 4495-4506.
155. Megger, D. A.; Pott, L. L.; Rosowski, K.; Zuelch, 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. *Analytical Sciences* **2017,** *33* (12), 1387-1394.
156. 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* **2020,** *1143*, 65-72.

157. Mädler, S.; Bich, C.; Touboul, D.; Zenobi, R., Chemical cross-linking with NHS esters: a systematic study on amino acid reactivities. *Journal of mass spectrometry* **2009,** *44* (5), 694-706.
158. Yang, W.-C.; Mirzaei, H.; Liu, X.; Regnier, F. E., Enhancement of amino acid detection and quantification by electrospray ionization mass spectrometry. *Analytical Chemistry* **2006,** *78* (13), 4702-4708.
159. Wang, S.; Regnier, F. E., Proteomics based on selecting and quantifying cysteine containing peptides by covalent chromatography. *Journal of Chromatography A* **2001,** *924* (1-2), 345-357.
160. 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. *Journal of Biological Chemistry* **1992,** *267* (8), 5060-5069.
161. Regnier, F. E.; Julka, S., Primary amine coding as a path to comparative proteomics. *Proteomics* **2006,** *6* (14), 3968-3979.
162. Savitski, M. M.; Reinhard, F. B.; Franken, H.; Werner, T.; Savitski, M. F.; Eberhard, D.; Molina, D. M.; Jafari, R.; Dovega, R. B.; Klaeger, S., Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **2014,** *346* (6205).
163. Lomenick, B.; Hao, R.; Jonai, N.; Chin, R. M.; Aghajan, M.; Warburton, S.; Wang, J.; Wu, R. P.; Gomez, F.; Loo, J. A., Target identification using drug affinity responsive target stability (DARTS). *Proceedings of the National Academy of Sciences* **2009,** *106* (51), 21984-21989.

Appendix

Figure S1 The MS/MS spectra of fragment ions used to locate the phosphorylation on parathymosin.

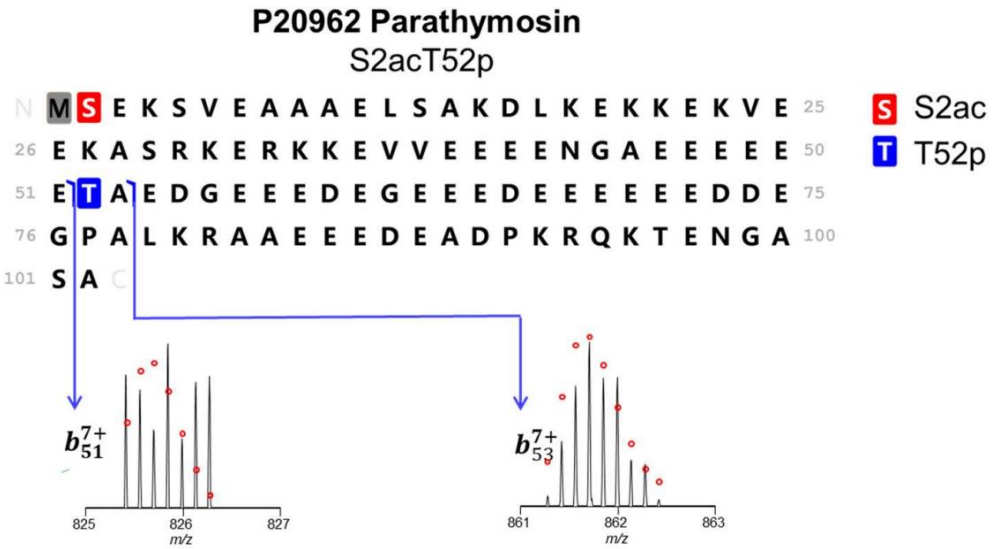


Figure S2 The MS/MS spectra of fragment ions used to locate the phosphorylation on 60S acidic ribosomal protein P2.

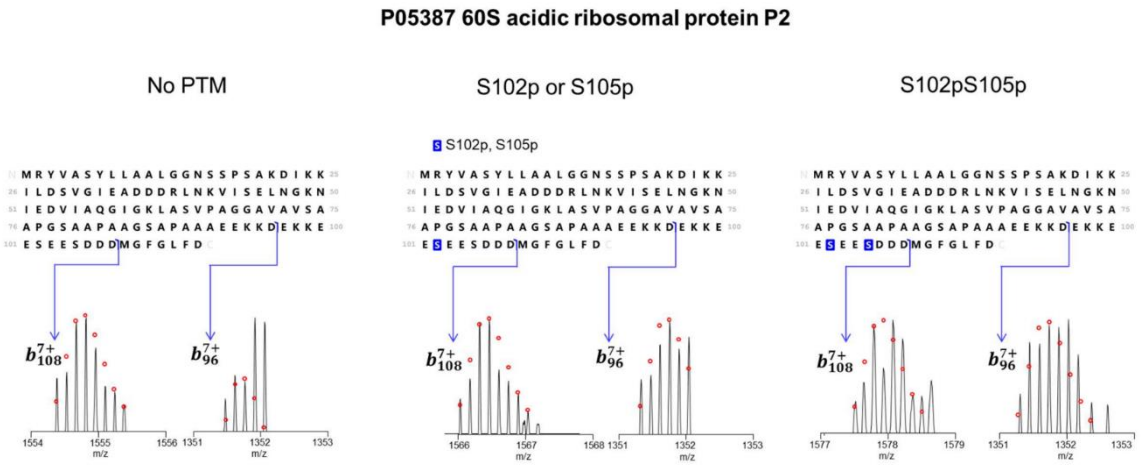


Figure S3 The MS/MS spectra of fragment ions used to locate the lipoylation and octanoylation on glycine cleavage system H protein

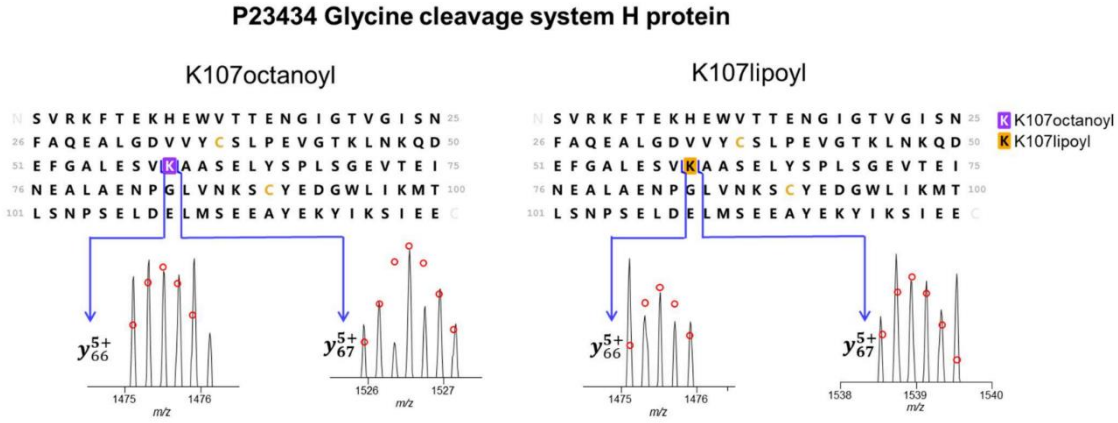


Figure S4 The MS/MS spectra of fragment ions used to locate the hypusine on eukaryotic translation initiation factor 5A-1

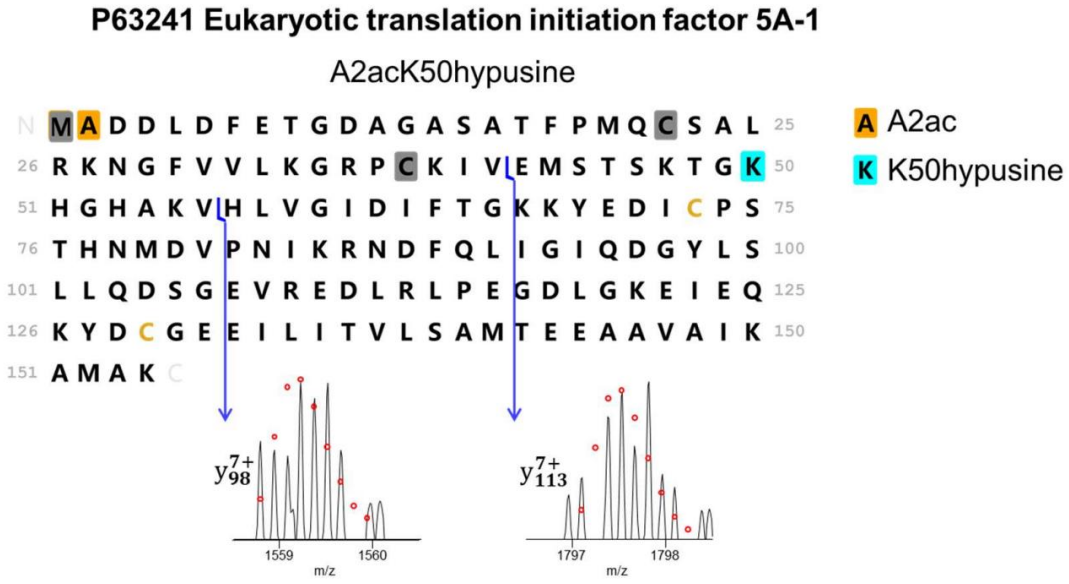


Figure S5 Correlation of intensity of reporter ions from underlabeled proteoforms and overlabeled proteoforms

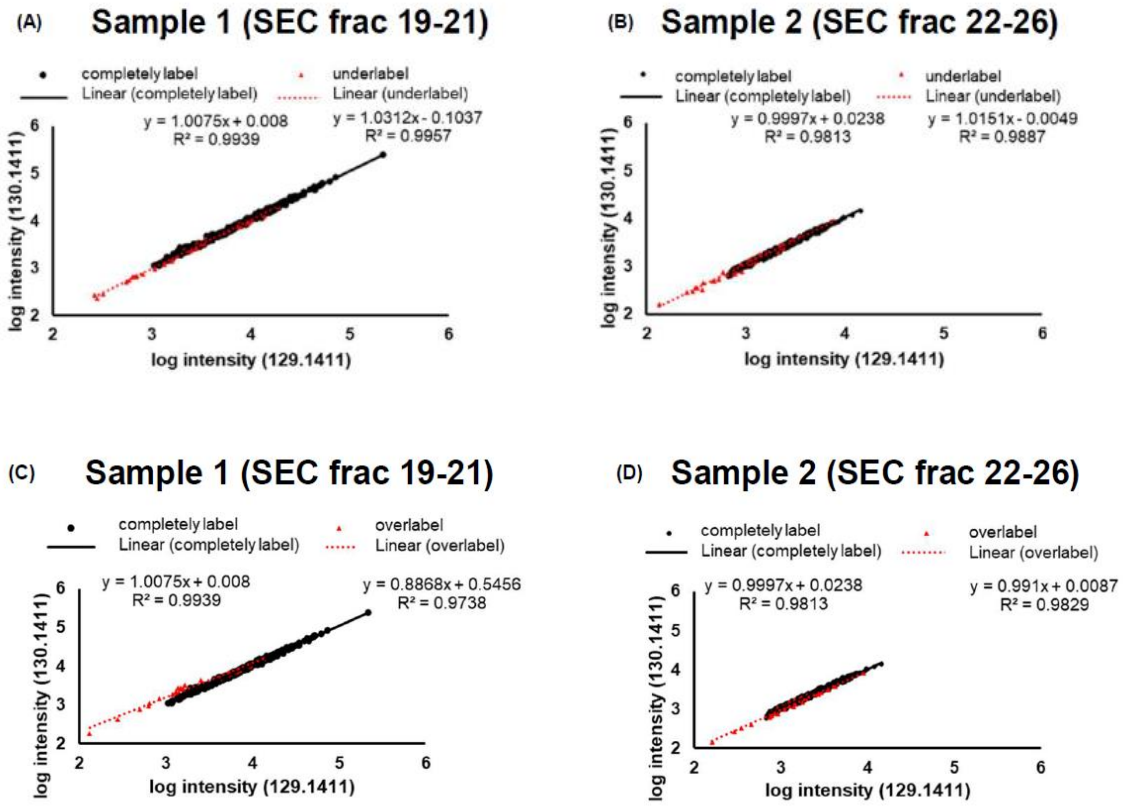
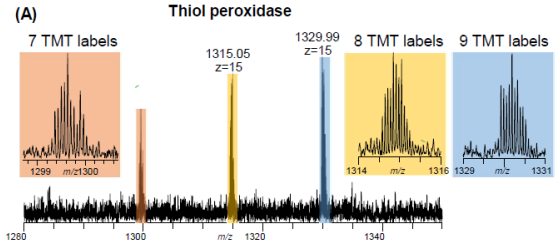
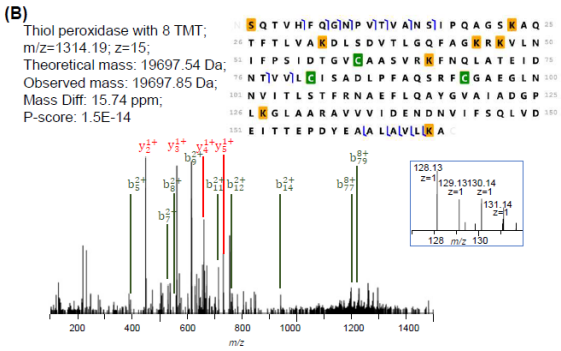


Figure S6 Identification of Thiol peroxidase with 9, 8 and 7 TMT labels.

(A) Full MS spectrum of Thiol peroxidase. The HCD MS/MS spectra and identification of (B) protein with 8 TMT labels and (C) protein with 9 TMT labels.





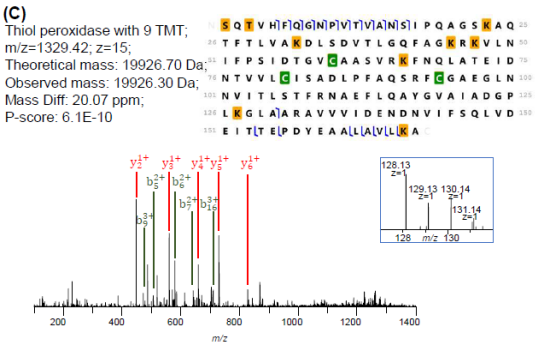


Table S1 PTMs identified using 1D and 2D platform

PTM #	Name	Description	Monoisotopic mass (Da)	proteoform count in 1D	proteoform count in 2D
1	Acetyl	Acetylation	42.011	117	771
2	Phospho	Phosphorylation	79.966	3	96
3	Oxidation	Oxidation	15.995	3	56
4	Disulfide	Disulfide bond	-2.016	18	51
5	Glutathione	Glutathionation	305.068	0	9
6	Pyro-glu	Gln->pyro-Glu	-17.027	0	7
7	Glu-loss	Loss of Glu	-129.043	0	11
8	Cation:Fe[III]	Replacement of 3 protons by iron	52.911	0	5
9	Glu	Monoglutamyl	129.043	0	3
10	Methyl	Methylation	14.016	0	4
11	Dehydrated	Loss of H2O	-18.011	1	2
12	Ethyl	Ethylation	28.031	0	2
13	Cysteinyl	Cysteinylation	119.004	0	1
14	Gly	Addtion of Glycine	57.021	1	1
15	Hypusine	Hypusine	87.068	0	1
16	Lipoy	Lipoy	188.033	0	1
17	Octanoyl	Octanoylation	126.196	0	1
18	Phosphocytidine	Cytidine monophosphate	305.041	0	1
19	Phosphopropargyl	Phospho-propargylamine	116.998	0	1
20	Sulfide	Persulfide	31.972	0	1

Table S2 Experimental design of protein-level TMT labeling optimization

Experiment series	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
TMT-to-protein ratio (vary TMT)	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 (Contorl*)
Quenching buffer	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

Table S3 Post-translational modifications (PTMs) in TopPIC Suite searching (Top-down)

Description	Monoisotopic mass (Da)	Residues	UnimodID	Status
Acetyllation	42.0106	K	1	Variable
Carboxymethyl	58.0055	C	6	Variable
Phosphoration	79.9663	STY	21	Variable
Oxidation	15.9949	MCPKDNRY	35	Variable
Methylation	14.0157	CKRHDENQ, N-termimi	34	Variable
TMT6plex	229.1629	K, N-termimi	737	Fixed
TMTzero	224.2994	K, N-termimi	739	Fixed

Table S4 Post-translational modifications (PTMs) in MS-GF+ searching (Bottom-up)

Search mode	Interim name	Monoisotopic mass (Da)	Residues	UnimodID	Status
Under-labeling	Oxidation	15.9949	K	1	Variable
	Carboxymethyl	58.0055	C	6	Fixed
	Acetylation	42.0106	Protein N-terminal	21	Variable
	TMT6plex	229.1629	K, peptide N-termimi	737	Variable
	TMT zero	224.2994	K, peptide N-termimi	739	Variable
Over-labeling	Oxidation	15.9949	K	1	Variable
	Carboxymethyl	58.0055	C	6	Fixed
	Acetylation	42.0106	Protein N-terminal	21	Variable
	TMT6plex	229.1629	K, protein N-termimi	737	Fixed
	TMT zero	224.1524	K, protein N-termimi	739	Fixed
	TMT6plex	229.1629	STYH	737	Variable
	TMT zero	224.1524	STYH	739	Variable

Reuse permission for Chapter 2

Deep Intact Proteoform Characterization in Human Cell Lysate Using High-pH and Low-pH Reversed-Phase Liquid Chromatography

ACS Publications
Most Trusted. Most Cited. Most Read.

Author: Dahang Yu, Zhe Wang, Kellye A. Cupp-Sutton, et al
Publication: Journal of the American Society for Mass Spectrometry
Publisher: American Chemical Society
Date: Dec 1, 2019

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

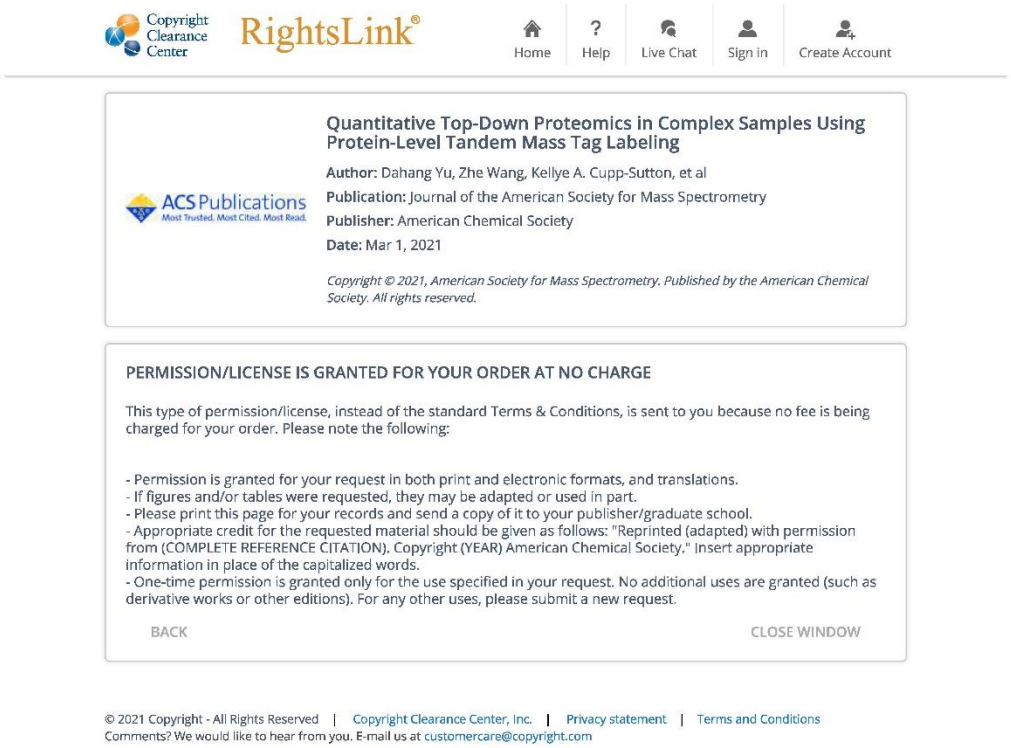
This type of permission/license, instead of the standard Terms and Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- 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

Reuse permission for Chapter 3



Copyright Clearance Center RightsLink®

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

ACS Publications
Most Trusted. Most Cited. Most Read.

Quantitative Top-Down Proteomics in Complex Samples Using Protein-Level Tandem Mass Tag Labeling

Author: Dahang Yu, Zhe Wang, Kellye A. Cupp-Sutton, et al
Publication: Journal of the American Society for Mass Spectrometry
Publisher: American Chemical Society
Date: Mar 1, 2021

Copyright © 2021, American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & 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 request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK CLOSE WINDOW