INVESTIGATIONS ON PROTEIN PHOSPHORYLATION AND ITS REGULATORS BY TOP-DOWN MASS SPECTROMETRY

By

Zhijie Wu

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

(Chemistry)

At the

UNIVERSITY OF WISCONSIN-MADISON

2020

Date of final oral examination: August 13th, 2020

This dissertation is approved by the following members of the Final Oral Exam Committee:

Ying Ge, PhD, Professor, Chemistry and Cell and Regenerative Biology

Lingjun Li, PhD, Professor, Chemistry and Pharmaceutical Sciences

Song Jin, PhD, Professor, Chemistry

Samuel H. Gellman, PhD, Professor, Chemistry

Acknowledgements

I would like to first thank Dr. Ying Ge, who is my advisor for five years, for the mentorship during my doctoral career. Dr. Ying Ge has provided me with tremendous amount of resources for the past five years, in particular for me not to get involved in teaching duties such that I can focus on doing research. She has constantly inspired me to push forward in science and has taught me many life lessons on how to handle issues more intelligently. She is a kind-hearted person and helped me to go through the PhD together, especially during my third year which is the lowest point of my doctoral journal coming from different stresses and anxiety. The work present in this dissertation would not have been possible without her guidance and support. I will miss her unconditional support after graduating from this lab.

I would also offer my appreciation to my current and past colleague in this lab. Tim Tiambeng, who is my roommate for the past two years, has been an excellent companion during works and outside of work. We had too much fun together in both our professional and leisure life. It is also a pleasure to work with David Roberts, who is never running out of energy, ideas, and excitement in science. Dr. Bifan Chen and Dr. Yutong Jin have been instrumental in training me in top-down mass spectrometry. To my peers, Kyle Brown and Trisha Tucholski, who is or soon to obtain their PhD, I am extremely honored to work with these two incredible scientists. Their work ethics and passion in science have and will inspire me to pursue the best science. I am honored to have Dr. Ziqing Lin and Dr. Yanlong Zhu during my time to ask mass spectrometry-related questions. They always have the answers for me. I would like to thank Dr. Wenxuan Cai and Dr. Zachery Gregorich for their help in training my biochemical skills and assisting in my first first-author publication. To the rest of the current lab members, Samantha Knott, Elizabeth Bayne,

Eli Larson, Jake Melby, Benjamin Wancewicz, Timothy Aballo, and Emily Reasoner, it is my pleasure to work with all of you to make this experience unforgettable. I would like to thank my committee members, Prof. Song Jin, Prof. Lingjun Li, and Prof. Samual Gellman for their feedback and discussions on my thesis background oral, original research proposal and this dissertation.

I would like to express my gratitude to my collaborators. It is my honor to work together with Dr. Sean McIlwain, Dr. Irene Ong and Kent Wenger, Molly Wetzel, Yiwen Gu, Sudharshana Govindaraj Ramanathan, and Yiran Yan in the MASH team. This project would not be half as successful without the work from all of you.

Finally, I would like to thank my family and friends. My mom has never ceased her support during my doctoral program, and helped me go through all my difficult time in graduate school. My grandmother never stops caring about me whenever I get a chance to return to China to visit her. I am blessed to have such loving family in my life. Dr. Rudi Fasan and Dr. John Frost have trained me in organic and biological chemistry during my undergraduate career. This training has benefited me immensely in graduate school and will continue to do so in my future career.

Table of Contents

Acknowledgements i

Table of Contents iii

List of Tables vi

List of Figures vii

List of Abbreviations and Acronyms xi

Abstract xiv

Chapter 1 Introduction 1

Introduction to Protein 2

Protein Reversible Phosphorylation 4

Mass Spectrometry-based Proteomics 8

References 25

Chapter 2 Impact of Phosphorylation on the Mass Spectrometry Quantification of Intact Phosphoproteins 37

Abstract 38

Introduction 39

Experimental Section 41

Results and Discussion 47

Conclusion 54

Acknowledgement 54

References 55

Supplemental Information 61

Chapter 3 Comprehensive Characterization of the Recombinant Catalytic Subunit of cAMP-Dependent Protein Kinase by Top-Down Mass Spectrometry 74
Abstract 75
Introduction 76
Experimental Section 77
Results and Discussion 81
Conclusion 89
Acknowledgement 90
References 91
Supplemental Information 101
Chapter 4 Enriching Kinases by Functionalized Nanoparticles: A Pilot Study 108
Abstract 109
Introduction 110
Experimental Section 112
Results and Discussion 124
Conclusion 129
Acknowledgement 130
Reference 131
Supplemental Information 138
Chapter 5 MASH Explorer: A Universal Software Environment for Top-Down Proteomics... 153
Abstract 154
Introduction 155
Experimental Section 156

Results and Discussion 159

Conclusion 169

Acknowledgement 170

References 171

Supplemental Information 181

Chapter 6 Enhancing Top-Down Proteomics Data Analysis by Combining Deconvolution Results through a Machine Learning Strategy 189

Abstract 190

Introduction 191

Experimental Section 193

Results and Discussion 199

Conclusion 209

Acknowledgement 209

References 210

Supplemental Information 220

Chapter 7 Conclusion and Future Directions 238

References 245

List of Tables

Table S2.1 61

Table S2.2 62

Table S5.1 181

Table 6.1 213

Table S6.1 220

Table S6.2 222

Table S6.3 223

Table S6.4 224

Table S6.5 225

Table S6.6 228

List of Figures

Figure 2.1 58
Figure 2.2 59
Figure 2.3 60
Figure S2.1 63
Figure S2.2 64
Figure S2.3 65
Figure S2.4 66
Figure S2.5 67
Figure S2.6 68
Figure S2.7 69
Figure S2.8 70
Figure S2.9 71
Figure S2.10 72
Figure S2.11 73
Figure 3.1 95
Figure 3.2 96
Figure 3.3 97
Figure 3.4 98
Figure 3.5 99
Figure 3.6 100
Figure S3.1 101
Figure S3.2 102
Figure S3.3 103

Figure S3.4 104

Figure S3.5 105

Figure S3.6 106

Figure S3.7 107

Figure 4.1 134

Figure 4.2 135

Figure 4.3 136

Figure 4.4 137

Figure S4.1 138

Figure S4.2 139

Figure S4.3 140

Figure S4.4 141

Figure S4.5 142

Figure S4.6 143

Figure S4.7 144

Figure S4.8 145

Figure S4.9 146

Figure S4.10 147

Figure S4.11 148

Figure S4.12 149

Figure S4.13 150

Figure S4.14 151

Figure S4.15 152

Figure 5.1 174

Figure 5.2 175

Figure 5.3 176

Figure 5.4 177

Figure 5.5 178

Figure 5.6 179

Figure 5.7 180

Figure S5.1 182

Figure S5.2 183

Figure S5.3 184

Figure S5.4 185

Figure S5.5 186

Figure S5.6 187

Figure S5.7 188

Figure 6.1 214

Figure 6.2 215

Figure 6.3 216

Figure 6.4 217

Figure 6.5 218

Figure 6.6 219

Figure S6.1 231

Figure S6.2 232

Figure S6.3 233

Figure S6.4 234

Figure S6.5 235

Figure S6.6 236

Figure S6.7 237

List of Abbreiations and Acronyms

ACN	Acetonitrile
ADC	Antibody-drug conjugate
ATP	Adenosine triphosphate
BCA	Bradford colorimetric assay
CAD	Collisionally activated dissociation
C-subunit	Catalytic subunit
CID	Collision-induced dissociation
CV	Coefficient of variation
Da	Dalton, unit of mass measurement (1 Da= 1g/mol)
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
ECD	Electron capture dissociation
EDTA	Ethylenediaminetetracetic acid
EGTA	Ethylene-bis(oxyethylenenitrilo)tetraacetic acid
ESI	Electrospray ionization
ETD	Electron transfer dissociation
EThcD	Electron transfer/higher-energy collision dissociation
EtOH	Ethanol
FA	Formic acid
FT-ICR	Fourier Transform Ion Cyclotron Resonance
GUI	Graphical user interface

H_2O	Water
HPLC	High-performance liquid chromatography
isCID	In-source collision-induced dissociation
IPTG	Isopropyl β-d-1-thiogalactopyranoside
λPP	Lambda protein phosphatase
LC	Liquid chromatography
m/z	Mass-to-charge ratio
MALDI	Matrix-assisted laser desorption ionization
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MW	Molecular weight
MWCO	Molecular weight cutoff
PKA	Protein kinase A/cAMP-dependent protein kinase A
PMSF	Phenylmethylsulfonyl fluoride
ppm	Parts per million
PTM	Post-translational modification
Q-TOF	Quadrupole time-of-flight
RF	Radio frequency
RPC	Reversed-phase chromatography
S/N	Signal-to-noise ratio
SEC	Size-exclusion chromatography
TCEP	Tris(2-carboxyethyl)phosphine

TFA	Trifluoroacetic acid
UVPD	Ultraviolet photodissociation

Abstract

The reversible phosphorylation of proteins is central to the regulation of most aspects of cell function, including cell cycle control, receptor-mediated signal transduction, cell differentiation and proliferation, and metabolism. Characterization of phosphorylation sites on phosphorylated proteins is crucial in signaling pathways, and quantification of phosphorylation level is shown to be relevant in understanding disease pathology. In the phosphorylation event, protein kinases are responsible for transporting phosphate groups to their substrate, and these proteins are expressed in relatively low abundance but in large varieties. In recent years, top-down proteomics has emerged as a powerful technology capable of characterizing and quantifying proteoforms arising from post-translational modifications, alternative splicing, and sequence variations at the intact protein level. In this dissertation, Chapter 1 investigated the impact of phosphorylation on phosphoprotein quantification by top-down mass spectrometry. Chapter 2 described a method to comprehensively characterize an intat phosphoprotein which is heavily phosphorylated. Chapter 3 is a pilot study on the development of a nanoproteomics platform to enable intact protein kinase enrichment for top-down mass spectrometry analysis. In Chapter 4, a universal and user-friendly software environment was developed for top-down proteomics community. Finally, Chapter 5 showed the development of a machine learning strategy for spectral deconvolution to aid top-down data analysis. I envision the development described in this dissertation will enable comprehensive characterization of phosphoproteins with accurate phosphoprotein quantification, enrichment of low-abundance proteins for mass spectrometry analysis, and high-throughput data analysis for top-down proteomics.

Chapter 1

Introduction

Introduction to Protein

Protein Structure

The central dogma of molecular biology describes a process of the flow of genetic information in the biological system.[1] DNA molecules containing genetic information are transcribed into RNA molecules, which are then translated into proteins. When compared to DNA and RNA, proteins are much more diverse. Protein structures are categorized in four levels: primary structure, secondary structure, tertiary structure, and quaternary structure.

The primary structure of proteins focuses on amino acid residues. Twenty proteinogenic amino acids are used in biosynthesis of proteins, and the amino acids in the polypeptide chain are connected by amide bonds. All of the amino acids residues, except for proline and glycine, have a primary amino group and a carboxyl group; where they differ is in their side chains, which can fall under the classifications of positively charged, negatively charged, polar and uncharged, and hydrophobic. Proline is the only amino acid containing a secondary amine, which directly connects with the main chain of the amino acid. Glycine does not have a side chain. For the hydrophobic side chains, structural motifs can range from a simple alkane chain with or without branching to ring systems containing phenyl and indole groups.

The pKa of the amino group, carboxyl group, and the side chains determines the reactivity of the amino acid. At physiological pH (~ 7.4), the amino acid is shown as a zwitterionic species in which the amino group is protonated and the carboxyl group is deprotonated. Furthermore, at physiological pH, the carboxylic side chains of both aspartic acid (pKa = 3.71) and glutamic acid (pKa = 4.15) are deprotonated, whereas the amino group on the side chain of lysine (pKa = 10.67) is protonated. Finally, the side chains of amino acids such as cysteine (pKa = 8.14) and tyrosine (pKa = 10.10) are protonated but uncharged at physiological pH; however, deprotonation will take place upon elevating solution pH above the side chain pKa level. Being able to discern features of amino acids, such as the variations in side chain structure and pKa values, is crucial to understanding biological processes. Moreover, researchers harness this knowledge in order to design unnatural amino acids, perform biochemical reactions, and create new digestive enzymes.[2-5]

Protein secondary structure is characterized by the local folded structures that form within the primary sequence of the polypeptide as a result of the interactions between atoms of the backbone. Two main structures include the α-helix and the β-pleated sheet. Hydrogen bonding between the amide proton from one amino acid and the carbonyl oxygen from another amino acid allows these secondary structures to hold their shape, but the hydrogen bonding patterns within these two structures are different. In an α-helix, the helical structure resembles a curled ribbon, with each turn containing 3.6 amino acids. The side chains of the amino acids stick outward from the helices, allowing side chain interactions with the outer environment. In a β-pleated sheet, segments of polypeptide chains align with each other and are held together by a hydrogen bonding network between sheets. Two types of β-pleated sheet include parallel and anti-parallel, which differ in their hydrogen bonding patterns. Other secondary structures include turns, loops, and paperclips.

Tertiary structure describes the overall three-dimensional structure of polypeptides. The main player in tertiary structure is the interaction between the side chains of the polypeptide sequence. Interactions such as hydrogen bonding, ionic bonding, dipole-dipole interactions, and London dispersion forces contribute to the tertiary structure. In addition, interactions with the outer environment determine the folding of polypeptide. For instance, in an aqueous environment, amino acids containing hydrophobic side chains favor the inside of a polypeptide, while those with hydrophilic residues tend to interact favorably with the outside aqueous environment. Disulfide bonds also shape the tertiary structure by linking two cysteine residues together. Finally, quaternary protein structures can be found in proteins that contain several subunits, which can be viewed as assembly of several tertiary structures.

Protein Reversible Phosphorylation

Phosphorylation

Phosphorylation, which takes place at serine, threonine, and tyrosine, is a ubiquitous and important PTM in mammalian cells.[6-7] Reversible phosphorylation, which is constituted by phosphorylation and dephosphorylation, plays an integral role in regulating the biological activity of proteins, and is thus involved in modulating numerous cellular processes such as cell cycle control, cell growth, apoptosis, and signaling transduction pathways.[8] Specifically, protein phosphorylation can alter the conformation of a protein between active and inactive states, and it can allow for proteins to bind to downstream partners. The structural change arising from phosphorylation at Thr-197 of the PKA C-subunit has been demonstrated to be relevant to the activation of the kinase activity.[9] The epidermal growth factor receptors are phosphorylated at different tyrosine residues, which initiates recruitment of various phosphotyrosine-containing motifs such as Grb2-Sos complex, Class I phosphatidylinositol 3-kinases, and phospholipase Cg for different downstream signaling pathways.[10] Moreover, protein phosphorylation can alter the protein turnover number and its activity. For example, phosphorylation at PEST sequence modulates the proteosomal-medidate rapid turnover of proteins.[11] In the case of 5-hydroxyconiferaldehyde O-methyltransferase, the enzymatic activity is controlled by phosphorylation for poplar monolignol biosynthesis.[12] Lastly, protein phosphorylation induces changes in protein localization across cellular compartments as well as PTM crosstalk with other partners. In the case of Cucumber Mosaic Virus 2b protein, phosphorylation allows it to shuttle between the cytoplasm and nucleus for protein function.[13] Phosphorylation at certain motifs inhibits O-GlcNacylation through PTM crosstalk mechanism.[14]

Because of the importance of phosphorylation, altered phosphorylation levels have been associated with the progression of diseases such as cancer, cardiovascular disease, and neurodegenerative disease. Cancer is often characterized by its aberrant signaling pathways. This includes mis-regulated expression resulting in changes in turnover of kinases that is responsible for protein phosphorylation, as well as abnormal phosphorylation such as up/down-regulation of phosphorylation levels.[15-16] Myofilaments and Z-disc proteins showed reduction in phosphorylation in acute myocardial infarction.[17] Phosphorylation of Tau protein is also relevant to the disease progression of Alzheimer’s disease.[18] As a result, protein phosphorylation may be useful as potential disease biomarkers.[19-22]

Protein Kinases

Kinases are enzymes that catalyze the reaction that transfers γ-phosphate groups of nucleotide triphosphates to their substrates. Due to its crucial role in regulating protein phosphorylation, phosphorylation by kinases needs to be highly specific.[23] As a result, while protein kinases are normally expressed in low abundance, they have a large variety.[24] For instance, protein kinase C (PKC) and cyclin-dependent kinase (CDK) have multiple isoforms in their respective class. PKC has isoforms such as PKCα, PKCβ, PKCγ and others, whereas CDK is encoded with isoforms including CDK1, CDK2, CDK3, and up to CDK21.[25-26] Dysregulation in the expression levels of kinases has significant consequences. For instance, in lung and breast cancers, epidermal growth factor receptor, which is a kinase, is found to be overexpressed in cells.[27]

Manning *et al.* classified protein kinases into eight groups including TK (tyrosine kinase), TKL (tyrosine kinase-like), STE (homologs of the yeast STE7, STE11, AND STE20 genes), CK1 (Casein Kinase 1), AGC (protein kinase A, G, and C families), CAMK (calmodulin/calcium regulated kinase families), CMGC (cyclin-dependent kinases, mitogen-activated protein kinase, glycogen synthase kinase, and dual specificity protein kinase CLK1), and RGC (Receptor guanylate cyclases).[28]

Protein kinases have a highly conserved catalytic core.[29] The N-lobe is a conserved region responsible for ATP binding, and this region consists of glycine-rich residues near a lysine residue. Kinase-dead variants in studying the signaling pathways of kinases are often made by mutating the catalytic lysine residue to other amino acid residues.[30-31] The C-lobe is responsible for binding to the peptide and directing catalysis. A conserved aspartic acid can be found in the C-lobe that is significant for the catalytic activity of the kinase. Structural insights of kinases are partly revealed in PKA C-subunit, which is one of the most studied kinases that participates in numerous biological processes.[32] The PKA has a heterotetrameric structure, which consists of two C-subunits, and two different regulatory subunits. While kinases phosphorylate their substrates, they themselves are phosphoproteins that can be autophosphorylated. In the case of PKA C-subunit, phosphorylation at Thr197 changes the PKA C-subunit from its inactive form to its active form.[9] Other phosphorylation sites in PKA C-subunit have been associated with the enzymatic activity and physiochemical properties of the protein.[33]

Kinase Inhibitors

Kinase inhibitors modulate the activities of protein kinases either reversibly or irreversibly, and these small molecules are designed based on four modes of action as defined by Zhang *et al.*[34-35] Type I inhibitors target the ATP binding site of the kinase active conformation.[34] Type II inhibitors recognize the inactive conformation of the kinase, and exploit changes in activation loop to expose additional hydrophobic binding site. Type III inhibitors modulate kinase activities by binding to an allosteric site, which demonstrates the highest degree of kinase selectivity. These allosteric sites can be either adjacent or remote to the ATP-binding pocket.[36] Type IV inhibitors covalently react with the nucleophilic cysteine residue in the kinase active site.

Kinase inhibitors can be used to control phosphorylation events by kinases and are thus useful in studying these signaling pathways. For instance, bisindolylmaleimide I is commonly used as a reagent to inhibit the activities of protein kinase C isoforms.[37-38] More importantly, kinase inhibitors have been developed as cancer therapy to target specific kinases, as kinases play an important role in signaling pathways. Since the approval of the first kinase inhibitor drug, imatinib, which is used as a chemotherapy agent for treating some types of cancer including chronic myelogenous leukemia, acute lymphocytic leukemia, and gastrointestinal stromal tumors, a total of 52 kinase inhibitors have been approved by Food and Drug Administration in 2020.[36, 39]

Investigations of these therapeutics treatments include the pharmacokinetics, the potency of inhibition, and the interactions. In particular, while kinase inhibitors are designed to be selective, they may still interact with other proteins. Researchers have designed analogs of FDA-approved kinase inhibitors and have immobilized these analogs on solid support to study the interactions. For example, AX14596 is an analog of Getifinib, a kinase inhibitor therapeutic for non–small cell lung cancer, and this analog carries an amine functional group. By reacting it with epoxy-activated Sepharose, proteins interacting with Getifinib can be evaluated by using the AX14596 analogs.[40]

Protein Phosphatases

Dephosphorylation is the counterpart of phosphorylation in the event of reversible phosphorylation.[41-42] Using a protein phosphatase, dephosphorylation catalyzes the hydrolysis of a phosphomonoester, which cleaves the phosphate group from its substrate protein. Recently, classification has been performed on protein phosphatase by studying the genomic and evolution.[43] Moreover, compared to the specificity of phosphorylation by protein kinases, dephosphorylation by protein phosphatases is less specific. Lastly, researchers have been expanding the landscape of drug development to target protein phosphatases, which allows them to approach difficult drug targets.[44-45]

Mass Spectrometry-based Proteomics

General Introduction

In the post-genome era, proteomics is the next frontier that enables in-depth understanding of the functions of cellular systems.[46-47] Proteomics is the study of proteome, which is a set of proteins expressed in a biological context such as an organism or an intracellular environment.[48-50] While the proteome reflects the transcriptome, the proteome is highly dynamic. As discussed previously, proteins in the proteome can be modified and unmodified with phosphorylation by kinases and dephosphorylation by phosphatases to activate and deactivate their activities in response to changes in the biological environment. In addition to PTMs, the location, such as where the proteins are expressed, and the abundance, including the rate of production and degradation, of proteins are also subjects of interest in the study of proteomics.[48]

MS-based proteomics is the primary tool that allows high-throughput investigation of the proteome. MS measures mass to charge ratio (*m/z*), which can be used to reflect precursor peptides or proteins, and MS/MS techniques fragment the proteins to provide information of the primary sequence of peptides or proteins. Three common proteomics techniques include bottom-up, top-down, and middle-down.

The bottom-up proteomics approach, which utilizes proteases to digest proteins into smaller peptides, is a high-throughput method for characterization of protein amino acid sequences and PTMs.[51] These digested peptides are subject to front-end separation by LC and analyzed by MS and MS/MS. By comparing the masses in the MS and MS/MS spectra to a sequence database, one or multiple identified peptides can be assembled into a protein identification. Peptides, in comparison to proteins, are small in molecular weight and are therefore less likely to be hindered by resolution, which describes the width of the peak. Using an optimal setup, a nearly complete coverage of yeast proteome can be analyzed by bottom-up proteomics.[52] This approach can also be utilized for protein structure analysis. Using a crosslinking reagent, the space in the target region of protein structure can be investigated.[53] Additionally, hydrogen-deuterium exchange enables analysis of the structure and dynamic of proteins. By studying the rate of hydrogen-deuterium exchange, researchers can identify the tightly folded and disordered regions of proteins in complement to other biophysical characterization techniques.[54]

In the bottom-up proteomics approach, proteins are inferred by identified peptides. While this strategy can provide rapid protein identifications, it suffers from this inference problem, and it is generally impossible to identify specific proteoform(s) from which the peptides originated.[55]

Additionally, despite that PTMs can be identified, bottom-up proteomics only offers a partial coverage of the identified modification, leading to loss of information. Top-down proteomics has emerged as a powerful technology capable of identifying proteoforms arising from PTMs, alternative splicing, and sequence variations at the intact protein level.[56-59] In comparison to bottom-up approach, top-down proteomics analyzes intact proteins, thus providing a “bird’s eye” view of the proteome. In recent years, top-down proteomics has drastically improved in its ability for global profiling, which facilitates identification of thousands of proteoforms in a single MS experiment. In addition to global profiling, top-down proteomics strives in targeted protein analysis as complex sequence information is preserved during MS/MS analysis. This approach is highly effective in identifying novel proteoforms and locating PTM sites by in-depth sequence characterization with the development of new fragmentation techniques. Moreover, changes in disease-associated PTMs can be quantified using top-down proteomics to reveal disease mechanisms for biomarker discovery. Finally, native MS has been continuously evolved to support structural elucidation of proteins and protein complexes.

Despite the promises offered by top-down proteomics, some proteins are difficult to study directly by this approach. For instance, it is difficult to generate enough informative fragment ions for large proteins.[60] Additionally, proteoforms with complex modifications, such as those in histones, are challenging for direct top-down analysis. A combinatorial approach known as middle-down has since been developed and involves using specialized proteases, such as Asp-N, Glu-C and Lys-C, to perform digestion to yield large polypeptides, and this digestion is then followed by MS/MS analysis.[61] For large proteins, such as those greater than 100 kDa, middle-down digestion can result in large polypeptides and subunits, which can be suitable for intact analysis.[62-63] Moreover, middle-down analysis coupled with effective LC separation enables high-throughput analysis of complex proteoforms.[64-65]

Although this thesis focuses on development of top-down proteomics, this particular proteomic technology is still under development. In the light of this, other proteomics techniques such as bottom-up and middle-down will also be included to overview the transition of specific techniques to top-down proteomics.

Chromatography-based Methods for Peptide/Protein Separation

Front-end separation prior to MS analysis is an integral component in proteomic research. Effective separation allows for increased confidence in peptide/protein identification, detection of minor variants with sequence variations and PTMs, and accurate quantification. Reverse-phase chromatography (RPC), which uses a hydrophobic stationary phase, is the most common mode of chromatography. For the stationary phase, silica particles, which are commonly used for chromatography, with different length of alkyl groups such as n-butyl (C4), octyl (C8), and octadecyl (C18), and functional groups including cyano and phenyl groups, are used for separating peptides and proteins with different retention characteristics.[66] The solvent gradient runs from a polar aqueous mobile phase to a nonpolar organic solvent mobile phase such as acetonitrile, methanol, ethanol or isopropanol. As the polarity of the mobile phase decreases, the hydrophobic interactions between the molecules and the stationary phase weaken. As a result, hydrophilic species are eluted first, followed by hydrophobic molecules. For positive mode MS, acids such as formic acid and trifluoroacetic acid are added to the mobile phase to add charges on the molecules of interest as well as provide charges on the silica surface to improve separation performance. Conversely, bases such as ammonium hydroxide and piperidine are utilized for negative mode

MS.[67] Despite silica particles being the material of choice for most chromatographic experiments, other materials, such as monolith, have emerged owing to their high permeability, low backpressure, and fast mass transfer.[68-70]

Other methods of chromatographic separation are also utilized in proteomic research, including hydrophobic interaction chromatography (HIC), hydrophilic interaction chromatography (HILIC), size-exclusion chromatography (SEC), and ion exchange chromatography (IEX). Since this thesis has not extensively studied these methods of chromatography, the mechanism and applications of these methods are briefly discussed.

Similar to RPC, HIC also separates molecules based on their hydrophobicity. Compared to RPC, which runs mostly in the denatured mode, HIC allows separation of molecules while maintaining biological activities. This mode of chromatography utilizes MS compatible salts, which reduce the solvation of the molecules, enabling analytes to interact with the stationary phase. More recently, HIC has recently been employed for online and intact MS analysis of monoclonal antibody and antibody-drug conjugate.[71-72] HILIC is an alternative form of normal phase separation.[73] This separation technique uses solvents similar to RPC; however, the analytes elute in the order of increased polarity using a gradient from a nonpolar organic mobile phase to a polar aqueous mobile phase. Applications of HILIC include analysis of N-glycan, where HILIC can separate N-glycan effectively, and membrane proteins, which have poor separation in RPC.[74-76]

SEC separates proteins by size. The porous silica materials allow smaller analytes to enter the pores, resulting retardation in retention, while the bigger analytes that do not fit in the pores elute first. Top-down proteomics has benefited from developments in SEC. In top-down analysis, the S/N ratio decreases logarithmically upon increase in molecular weight, necessitating the separation between large and small proteins.[77] Effective separation using SEC allows detection and characterization of high molecular weight proteins up to 223 kDa.[78] Sample preparation from fractionation by SEC is also MS-compatible, enabling robust analysis of large proteins.[62, 79]

IEX separates molecules by charges, and it includes anion exchange chromatography and cation exchange chromatography.[80] A strong ion exchanger can tolerate a wide range of pH values, as the material does not lose the charge after column equilibrium, whereas a weak ion exchanger cannot retain the charge and can only work in a narrower pH range. Negatively charged moieties interact strongly with the column material used in anion-exchange chromatography, and the same can be said for positively charged moieties when referring to cationic ion exchange chromatography. Anion-exchange chromatography can be applied to phosphopeptide separation to enhance its identification, as well as effective separation of ovalbumin.[81-82] Strong cation-exchange chromatography works similar to RPC and can work as an orthogonal method to enhance general peptide separation.[83] Weak cation-exchange chromatography allows effective separation of hemoglobin components for disease diagnosis.[84]

Ionization and Instrumentation of Mass Spectrometry

As MS measures *m/z*, proteins need to be ionized for MS analysis. Two soft ionization methods are widely used in the MS community for protein analysis. The first one is electrospray ionization (ESI).[85] In this method, the droplet prior to entering the mass spectrometer is heated, and the amount of liquid is slowly decreased. Coulombic explosion occurred to generate charged ions in the gas phase. During this process, nebulizer gas can assist in the removal of liquid for accelerating ion generation. Nanoelectrospray has been developed to enhance the sensitivity of the signal.[86] The second ionization method is matrix-assist laser desorption ionization (MALDI).[87] In this method, the protein sample is mixed with matrix molecules, which contain aromatic rings.

When the laser light hits the mixture, the matrix molecules absorb the energy and create ionized protein molecules in the hot plume of ablated gases. The analytes can then be analyzed by the mass spectrometer.

Several types of mass spectrometers are available for researchers to perform MS experiments. Linear ion trap is an ion trap mass spectrometer that confines ions in a two-dimensional radio frequency field.[88] Triple quadrupole instruments are composed of three quadrupole mass analyzers.[89-90] These instruments have two mass-resolving quadrupoles at both ends and one non-mass-resolving quadrupole in the middle to serve as collision cell. A time-of-flight (TOF) instrument measures the time of flight of ions, which is correlated to an ion's *m/z*.[91] As ions are accelerated in the electric field, these ions carry the same kinetic energy as other ions with the same charge. The velocity resulting in differences in arrival time is dependent on the *m/z*, where heavier ions have lower velocity and lighter ions have higher velocity.

Two types of Fourier transform instruments, including Orbitrap and Fourier transfer ion cyclotron resonance (FT-ICR), are available. For Orbitrap, the mass analyzer has an outer barrel-like electrode and a coaxial inner spindle-like electrode.[92-93] Ions trapped in the spindle under the electrostatic field exhibit an orbital motion, and the mass spectrum can be obtained by Fourier transform of the frequency signal. On the other hand, the ions in the ICR cell are subject to both electric field and magnetic field.[94] After excitation, the ions display cyclotron motion and result in free induction decay. The cyclotron frequency can be Fourier transform to mass spectrum.

During instrument development and analytical needs, hybrid instruments have become more popular in the mass spectrometry society. The quadrupole can be coupled with either the Orbitrap or the TOF mass analyzer.[95-96] The linear ion trap can be combined with FT-ICR due to the high injection efficiencies and high ion storage capacities of the linear ion trap. Different ionization techniques are also developed for different types of analytical workflow. ESI is the method of choice for most LC-MS/MS experiments. MALDI coupled with a TOF instrument, which has extended mass range, has been a robust analytical method for biomolecules. Recent advances have enabled MALDI as a prominent technique in imaging for biomarker discovery and disease prognosis.[97]

Fragmentation Techniques

MS has significantly improved the efficiency of sequencing protein primary structure using tandem MS (MS/MS) techniques.[98] MS/MS analysis involves isolation of precursor ions, which are subject to fragmentation. Common fragmentation methods for MS/MS analysis include collision-induced dissociation (CID), electron capture dissociation (ECD), and electron-transfer dissociation (ETD).[58]

For CID (also known as collisionally activated dissociation, CAD), the kinetic energy of selected ions is increased by accelerating the ions through an electrical potential gradient.[99] Neutral molecules such as helium, nitrogen, or argon are introduced to collide with these accelerated precursor ions. The kinetic energy is converted to internal energy, resulting in breakage of the backbone at the amide bond position and generating *b* and *y* ions. An alternative form of CID, higher-energy collisional dissociation (HCD), is commonly used for the Orbitrap mass spectrometer.[100] Due to the mechanism of CID, breakage of low energy bonds is favored. Limited sequence information can only be attained if a weak bond is present at a region of sequence.[63] Additionally, using the CID method can cause breakage of labile modifications, such as phosphorylation, making PTM site localization difficult.

Compared to CID, both ECD and ETD are two electron-based fragmentation methods, which are softer and can preserve labile PTMs, enabling PTM site characterization. Moreover, different mechanisms are employed for these two electron-based methods, making them complementary methods to CID for analyzing the primary structure of polypeptides. In ECD, the precursor ion interacts with a free electron to form an odd-electron ion. Fragmentation of the peptide backbone at the N-Cα position occurs, which generates *c* and $z^{\bullet}$ ions when the electric potential energy of the precursor ion is released.[101] While ECD is primarily used for the FT-ICR instrument, recent development has made ECD available to other instruments, such as ion mobility (IM)-TOF and Orbitrap.[102-104] In comparison, for ETD, the precursor ions react with reagents, which are radical anions, to become cation radicals. These cation radicals are unstable, leading to fragmentation of the peptide backbone at the N-Cα position.[105] Compared to ECD, ETD is relatively low cost and is conducted in radio frequency quadrupole ion trap devices. Hybrid mass spectrometry with front-end quadrupoles, such as Orbitrap and quadruple time-of-flight, can be equipped with ETD technology. Recently, an improved form of ETD called activated-ion ETD was developed. In activated-ion ETD, the precursor ions are further activated by photon bombardment, which enhances the ability of ETD to identify peptides in complex biological samples.[106-107] Additionally, ETD in tandem with HCD has been developed as a powerful for glycopeptide analysis, where ETD provides peptide sequence information and the subsequent HCD gives glycan fragmentation.[108-110]

Some other fragmentation methods, including infrared multi photon dissociation (IRMPD), ultraviolet photodissociation (UVPD), and surface-induced dissociation (SID), have also been developed. IRMPD involves precursor ions absorbing multiple infrared photons, leading to excitation to its more energetic vibrational states.[111] Cleavage of bonds results in *b* and *y* ions, similar to CID. UVPD is a recently revisited fragmentation technique, which uses 193 nm laser to excite the peptide backbone, leading to bond cleavages.[112] This technique generates an array of fragment ions including a, b, c, x, y, and $z^{\bullet}$, which provides rich sequence information.[113-114] Lastly, SID experiments are analogous to CID, but a surface is used instead of neutral gas as the collision target.[115-116] Applications of SID focus on dissociating protein complexes into subunits for analysis in a controlled manner.[117-118]

While the major application of fragmentation methods is to sequence the primary structure of peptides/proteins, tertiary and quaternary structural information can also be obtained in the application of intact protein analysis. For instance, peptide backbones within disulfide bond linkages are unlikely to be broken by fragmentation methods, which enables researchers to use fragmentation pattern to identify the region connected by disulfide bonds.[119] Additionally, the solvent-exposed region of protein structure can be observed during fragmentation using native MS.[72, 120-121]

Protein Quantification

Protein quantification provides important information regarding the protein expression level and PTM level, such as those between healthy and diseased samples, which are crucial for biomarker discovery.[122-123] Relative quantification studies the difference in protein expression among various conditions, and absolute quantification aims to obtain the expression level of a target protein. Relative quantification and absolute quantification of proteins are enabled by various quantification methods including isobaric tag, stable isotope labeling by/with amino acids in cell culture (SILAC), and label-free quantification.

Isobaric tagging is a commonly used strategy to quantify protein expression level and PTM level across different samples. Two common chemical tagging systems include tandem mass tags (TMT) and isobaric tag for relative and absolute quantitation (iTRAQ).[124-125] In both types of chemical tags, the mass of the entire moieties added to the samples are the same. These moieties contain a cleavable reporter group by fragmentation method, balance group, and amine reactive group. The reporter group and balance group uses isotopes to balance the mass. The reporter group is cleaved subject to collisional energy, resulting in isotopic peaks that can be used to infer the relative abundance of the protein of interest across multiple samples. Novel isobaric tagging, such as DiLeu, improves both in reactivity, which minimizes random error, and multiplexing, which increases the number of samples that can be simultaneously quantification.[126-128]

SILAC is another method for quantitative proteomics, which uses non-radioactive isotopic labeling to detect protein abundance differences across samples with mass spectrometry.[129-130] Both a “light” and a “heavy” version of amino acids are used, and these amino acids are incorporated into protein sequence during cell growth. The samples from different conditions can be mixed and analyzed by MS. Leucine, arginine, and lysine are commonly used for SILAC. While this technology has small error and good reproducibility, this method is only applicable for cell samples and requires an extended period for cell growth.

Label-free quantification is a method, which does not uses isotope labels during quantification. Two methods of label-free quantification are commonly utilized. Quantitative values can be derived from the area under the curve of precursor ions in the chromatogram and MS1 intensity values. Additionally, spectral counting, which measures the number of spectra matched to target proteins, can also be used to evaluate the abundance of proteins.[131]

Quantitative studies in top-down proteomics use primarily label-free, relative quantification. In bottom-up proteomics, the ionization efficiency among multiple peptides may differ significantly by amino acid composition and PTMs such as phosphorylation. Additionally, PTMs often add significant molecular weight to the peptides of interest, thus altering the physiochemical properties. In comparison, relative quantification is widely used in top-down proteomics, as it has been shown in several studies that modifications have negligible impact on the ionization efficiency among proteoforms. Recent advances have also showed success of quantification of protein expression and modifications using top-down proteomics.[132]

While both isobaric labeling and SILAC are popular in bottom-up proteomics quantification, these methods are less common in top-down proteomics. SILAC has been experimented on top-down proteomics; however, the incorporation efficiency in addition to impure “heavy” amino acids can lead to mass shift and imperfect isotopic distribution.[133-134] Isobaric labeling sees little applications in top-down proteomics, as the mass spectra can already be crowded by different proteoforms from the same proteins and requires fragmentation methods to yield reporter ions.[135]

Peptide/Protein Enrichment Methods

Proteomics offers a powerful tool in identification and characterization of proteins; however, low-abundance species are significantly underrepresented in the analysis. With protein enrichment, the percentage of proteins of interest in the mixture can be dramatically increased. In this section, enrichment of phosphoproteins/phosphopeptides, glycans, and kinases is discussed.

As mentioned in the earlier section, phosphorylation is an important PTM in biological processes. There are significant interests in understanding the functional consequences of specific phosphorylation sites. However, the phosphorylation level of specific sites may not be sufficiently abundant for MS detection and identification. Therefore, enrichment of these phosphorylated species is necessary. Enrichment of phosphorylated species has been shown at both the peptide level, phosphopeptides, and the protein level, phosphoprotein. For phosphopeptide enrichment, which takes place after proteolytic digestion, materials such as Fe-NTA, which can chelate phosphate groups, and TiO_2, which has affinity to phosphate groups, are available.[136] Additionally, nanomaterials such as magnetic iron oxide nanoparticles have also shown promise in phosphopeptide enrichment.[137] Despite their promise in phosphopeptide enrichment, these materials are less effective in phosphoprotein enrichment. Chemical moieties such as Phos-tag^{TM}, were developed to enrich intact phosphoproteins and have demonstrated utility in several biological applications including SDS-PAGE, western blotting, and protein purification.[138] An analog of Phos-tag was also incorporated onto nanoparticles for intact protein analysis.[139-141]

Dysregulation of kinase expression and activity is often associated with diseases such as cancer and neurodegeneration. As mentioned in previous sections, kinases are responsible for transporting phosphate groups to their substrates. Kinases are also phospho proteins, and phosphorylation of kinases regulates their protein structures and activities. One common feature of kinases is the relatively conserved active site, which allows small molecule inhibitors with high affinity to bind and thus can be exploited for achieving kinase enrichment. Kinase enrichment using kinase inhibitors is performed at the protein level since kinase inhibitors need to access the well-structured active site of kinases. To this end, several pan-kinase inhibitors, including those from modified cancer therapeutics, were incorporated on solid support and used for kinase enrichment.[142-143] This platform allows investigation of the interactions between cancer drugs and kinases.[40, 143]

Glycosylation is another prevalent PTM, and an estimated 50% proteins are post-translationally modified with this PTM.[144-145] The modification site of glycosylation and the glycan structure has attracted significant interest.[146-147] Glycosylation enrichment can be performed at both the peptide and protein level, as these ligands can well recognize the glycan structure. This enables capturing glycoproteins for identification as well as the subsequent enrichment step for glycopeptides for site localization. Several strategies have been developed to study this PTM. First, lectins, which are proteins that have an affinity for glycans, can be used for glycoprotein enrichment.[148] Different types of lectins can bind to specific glycan structures. Moreover, boronic acid chemistry has also been utilized for glycoprotein analysis.[148] It reacts with specific *cis*-diol groups on glycans to form a cyclic bromate ester, and this reaction can be reversible by changing the pH in the solution to release the glycopeptides. Finally, derivatives of monosaccharides with functional groups such as azide are synthesized, and these sugar subunits can be incorporated by normal biological machinery.[149-150] The click reaction, which is a highly efficient reaction between azide and alkyne, can be performed to capture modified glycan structures. Other strategies, including hydrazide chemistry and HILIC enrichment, were also employed for glycoprotein enrichment.[151-153]

Top-down MS-based Bioinformatics

Data analysis from proteomics experiments rely extensively on computational tools. Compared to the well-developed bottom-up proteomics software, tools for top-down proteomics are limited and underdeveloped. This is largely due to the complexity of high-resolution top-down proteomics datasets. A typical top-down proteomics data analysis includes data import, spectral deconvolution, and database search. Data import and data file conversion become straightforward with the development of ProteoWizard, which provides an array of tools for data manipulation prior to data analysis.[154] Additionally, vendors for MS instruments have provided dynamic link libraries for software developers. Therefore, data import will not be extensively discussed, whereas fundamentals and development for top-down spectral deconvolution and database search will be reviewed.

In top-down proteomics, spectral deconvolution is a challenging process characterized by converting complicated raw spectra to simplified mass lists.[155] In bottom-up proteomics, peptides are generally small in molecular weight (less than 2,500 Da). As a result, the most abundant mass (the *m/z* value of the most abundant peak minus the mass of the number of protons, given by the charge state of the peak) is likely to be equivalent to the monoisotopic mass (calculated from peptide sequence using masses from most abundant isotopes of each element). However, for larger molecular weights, isotopes play a much bigger role. The probability of having a peak with every element in its most abundant form decreases dramatically with increasing molecular weight. Therefore, the most abundant mass no longer equals the monoisotopic mass, which might not be visible in the spectrum. Additionally, peptides most often carry a charge state of one, two, or three. Consequently, the charge state of the small molecular weight ions can be easily deduced with a few isotopic peaks. In comparison, a group of isotopic peaks is necessary to evaluate the charge state for ions with higher molecular weight. In top-down proteomics, the averagine model is the primary tool to deconvolute the isotopic distributions.[156] This model uses the mass of an average amino acid based on natural abundance to be $C_{4.94}H_{7.76}N_{1.36}O_{1.48}S_{0.04}$, with a mass of 111.1254 Da.

Using the isotopic spacing in the isotopic distribution and the observed peak, an estimated elemental composition of the ions can be deduced.

In top-down proteomics, computational tools process two spectral deconvolution tasks for the datasets, including both MS and MS/MS spectra. In the MS level, software tools calculate the monoisotopic mass of the parent ions by decharging, which collapses all the charge states to a single deconvoluted peak to provide intact mass analysis of the proteins. The deconvoluted mass list is also termed "MS1 feature" in most software. These MS1 features provide precursor and deconvoluted masses of proteins in the LC timescale to assist the downstream database search task. In the MS/MS level, software tools handle deconvolution of isotopomers using an improved version of the averagine model, and they calculate the monoisotopic mass of fragment ions. The deconvoluted mass list for fragment ions can be used to match with a theoretically generated mass list from protein sequence during database search. Several algorithms have been developed to achieve one or both of these deconvolution tasks. MS-Deconv and TopFD from TopPIC Suite have integrated both tasks during its run, whereas ProMex from Informed-Proteomics and FLASHDeconv only performed MS1 feature detection.[157-160]

Database search algorithms in top-down proteomics are similar to that for bottom-up proteomics. The algorithms generate theoretical fragment ion mass lists from a given database and match with the input deconvoluted mass list for protein identification. In the database search, one or more PTMs can be included in the calculation; however, it adds complexity to the process. While top-down proteomics allows for identification of novel proteoforms, identification can be challenging since these proteoforms are not in the database. Algorithms can assign random modifications at certain sequence regions to maximize fragment ion matching and to enhance the probability of identifying proteoforms, but manual efforts are necessary to adjust the amino acid variations and PTM locations.[157-159] Some of these database search algorithms identify sequence tags, which is a three to four amino acid short sequence that is fragmented at each residue, to propose protein identification.[159, 161] In contrast to deconvolution that requires reading the spectrum from the entire x-axis (*m/z*) to identify isotopic distribution, database search can be executed simultaneously by dividing the database into subsets. The speed of database search can be improved by using computers with more threads in the processor in the case of TopPIC from TopPIC Suite and MSPathFinderT from Informed-Proteomics.[158-159] Although a majority of database search algorithms use fragment ions for protein identification, intact mass measurement can also be utilized for database search as demonstrated in Proteoform Suite.[162-163]

Machine learning methods are harnessed in most deconvolution and database search algorithms to enhance the efficiency and accuracy of algorithmic outputs. For instance, pTop utilizes a support vector machine to train a model and incorporate a variety of features to detect isotopic clusters and to determine their charge states.[164] While a plethora of algorithms has been developed, each algorithm has outputs with varied accuracy. Ensembles and machine learning algorithms are used to evaluate these results to output a consensus list.[165]

References

1. Crick, F., Central dogma of molecular biology. *Nature* **1970,** *227* (5258), 561-3.

2. Wang, L.; Brock, A.; Herberich, B.; Schultz, P. G., Expanding the genetic code of Escherichia coli. *Science* **2001,** *292* (5516), 498-500.

3. Chin, J. W.; Cropp, T. A.; Anderson, J. C.; Mukherji, M.; Zhang, Z.; Schultz, P. G., An expanded eukaryotic genetic code. *Science* **2003,** *301* (5635), 964-7.

4. Frost, J. R.; Smith, J. M.; Fasan, R., Design, synthesis, and diversification of ribosomally derived peptide macrocycles. *Curr. Opin. Struct. Biol.* **2013,** *23* (4), 571-80.

5. Liebscher, S.; Bordusa, F., Trypsiligase-Catalyzed Peptide and Protein Ligation. *Methods Mol. Biol.* **2019,** *2012*, 111-133.

6. Hunter, T., Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* **1995,** *80* (2), 225-36.

7. 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), 635-48.

8. Humphrey, S. J.; James, D. E.; Mann, M., Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends Endocrinol. Metab.* **2015,** *26* (12), 676-687.

9. Steichen, J. M.; Iyer, G. H.; Li, S.; Saldanha, S. A.; Deal, M. S.; Woods, V. L., Jr.; Taylor, S. S., Global consequences of activation loop phosphorylation on protein kinase A. *J. Biol. Chem.* **2010,** *285* (6), 3825-32.

10. Cohen, P., The regulation of protein function by multisite phosphorylation &#x2013; a 25 year update. *Trends in Biochem. Sci.* **2000,** *25* (12), 596-601.

11. Garcia-Alai, M. M.; Gallo, M.; Salame, M.; Wetzler, D. E.; McBride, A. A.; Paci, M.; Cicero, D. O.; de Prat-Gay, G., Molecular basis for phosphorylation-dependent, PEST-mediated protein turnover. *Structure* **2006,** *14* (2), 309-19.

12. Wang, J. P.; Chuang, L.; Loziuk, P. L.; Chen, H.; Lin, Y. C.; Shi, R.; Qu, G. Z.; Muddiman, D. C.; Sederoff, R. R.; Chiang, V. L., Phosphorylation is an on/off switch for 5-hydroxyconiferaldehyde O-methyltransferase activity in poplar monolignol biosynthesis. *P. Natl. Acad. Sci. USA* **2015,** *112* (27), 8481-8486.

13. Nemes, K.; Gellert, A.; Almasi, A.; Vagi, P.; Saray, R.; Kadar, K.; Salanki, K., Phosphorylation regulates the subcellular localization of Cucumber Mosaic Virus 2b protein. *Sci. Rep.* **2017,** *7*, 13444.

14. Leney, A. C.; El Atmioui, D.; Wu, W.; Ovaa, H.; Heck, A. J. R., Elucidating crosstalk mechanisms between phosphorylation and O-GlcNAcylation. *P. Natl. Acad. Sci. USA* **2017,** *114* (35), E7255-E7261.

15. Ruprecht, B.; Lemeer, S., Proteomic analysis of phosphorylation in cancer. *Expert Rev. Proteomic* **2014,** *11* (3), 259-267.

16. Cicenas, J.; Zalyte, E.; Bairoch, A.; Gaudet, P., Kinases and Cancer. *Cancers (Basel)* **2018,** *10* (3).

17. Peng, Y.; Gregorich, Z. R.; Valeja, S. G.; Zhang, H.; Cai, W. X.; Chen, Y. C.; Guner, H.; Chen, A. J.; Schwahn, D. J.; Hacker, T. A.; Liu, X. W.; Ge, Y., Top-down Proteomics Reveals Concerted Reductions in Myofilament and Z-disc Protein Phosphorylation after Acute Myocardial Infarction. *Mol. Cell. Proteomics* **2014,** *13* (10), 2752-2764.

18. Hanger, D. P.; Anderton, B. H.; Noble, W., Tau phosphorylation: the therapeutic challenge for neurodegenerative disease. *Trends Mol. Med.* **2009,** *15* (3), 112-119.

19. Chen, I. H.; Xue, L.; Hsu, C. C.; Paez, J. S. P.; Pan, L.; Andaluz, H.; Wendt, M. K.; Iliuk, A. B.; Zhu, J. K.; Tao, W. A., Phosphoproteins in extracellular vesicles as candidate markers for breast cancer. *P. Natl. Acad. Sci. USA* **2017,** *114* (12), 3175-3180.

20. Palmqvist, S.; Janelidze, S.; Quiroz, Y. T.; Zetterberg, H.; Lopera, F.; Stomrud, E.; Su, Y.; Chen, Y.; Serrano, G. E.; Leuzy, A.; Mattsson-Carlgren, N.; Strandberg, O.; Smith, R.; Villegas, A.; Sepulveda-Falla, D.; Chai, X.; Proctor, N. K.; Beach, T. G.; Blennow, K.; Dage, J. L.; Reiman, E. M.; Hansson, O., Discriminative Accuracy of Plasma Phospho-tau217 for Alzheimer Disease vs Other Neurodegenerative Disorders. *JAMA* **2020**.

21. Carter, A. M.; Tan, C.; Pozo, K.; Telange, R.; Molinaro, R.; Guo, A.; De Rosa, E.; Martinez, J. O.; Zhang, S.; Kumar, N.; Takahashi, M.; Wiederhold, T.; Ghayee, H. K.; Oltmann, S. C.; Pacak, K.; Woltering, E. A.; Hatanpaa, K. J.; Nwariaku, F. E.; Grubbs, E. G.; Gill, A. J.; Robinson, B.; Gillardon, F.; Reddy, S.; Jaskula-Sztul, R.; Mobley, J. A.; Mukhtar, M. S.; Tasciotti, E.; Chen, H.; Bibb, J. A., Phosphoprotein-based biomarkers as predictors for cancer therapy. *P. Natl. Acad. Sci. USA* **2020,** *117* (31), 18401-18411.

22. Zawadzka, A. M.; Schilling, B.; Cusack, M. P.; Sahu, A. K.; Drake, P.; Fisher, S. J.; Benz, C. C.; Gibson, B. W., Phosphoprotein secretome of tumor cells as a source of candidates for breast cancer biomarkers in plasma. *Mol. Cell. Proteomics* **2014,** *13* (4), 1034-1049.

23. Ubersax, J. A.; Ferrell, J. E., Jr., Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* **2007,** *8* (7), 530-41.

24. Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R., The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **2011,** *7*, 549.

25. Steinberg, S. F., Structural basis of protein kinase C isoform function. *Physiol. Rev.* **2008,** *88* (4), 1341-78.

26. Malumbres, M.; Harlow, E.; Hunt, T.; Hunter, T.; Lahti, J. M.; Manning, G.; Morgan, D. O.; Tsai, L. H.; Wolgemuth, D. J., Cyclin-dependent kinases: a family portrait. *Nat. Cell Biol.* **2009,** *11* (11), 1275-6.

27. Liu, T. C.; Jin, X.; Wang, Y.; Wang, K., Role of epidermal growth factor receptor in lung cancer and targeted therapies. *Am. J. Cancer Res.* **2017,** *7* (2), 187-202.

28. Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S., The protein kinase complement of the human genome. *Science* **2002,** *298* (5600), 1912-1934.

29. McClendon, C. L.; Kornev, A. P.; Gilson, M. K.; Taylor, S. S., Dynamic architecture of a protein kinase. *P. Natl. Acad. Sci. USA* **2014,** *111* (43), E4623-31.

30. Goshen-Lago, T.; Goldberg-Carp, A.; Melamed, D.; Darlyuk-Saadon, I.; Bai, C.; Ahn, N. G.; Admon, A.; Engelberg, D., Variants of the yeast MAPK Mpk1 are fully functional

independently of activation loop phosphorylation. *Mol. Biol. Cell* **2016,** *27* (17), 2771-83.

31. Martin-Doncel, E.; Rojas, A. M.; Cantarero, L.; Lazo, P. A., VRK1 functional insufficiency due to alterations in protein stability or kinase activity of human VRK1 pathogenic variants implicated in neuromotor syndromes. *Sci. Rep.* **2019,** *9*, 13381.

32. Taylor, S. S.; Zhang, P.; Steichen, J. M.; Keshwani, M. M.; Kornev, A. P., PKA: Lessons learned after twenty years. *Biochim. Biophys. Acta.* **2013,** *1834* (7), 1271-1278.

33. Yonemoto, W.; Garrod, S. M.; Bell, S. M.; Taylor, S. S., Identification of Phosphorylation Sites in the Recombinant Catalytic Subunit of Camp-Dependent Protein-Kinase. *J. Biol. Chem.* **1993,** *268* (25), 18626-18632.

34. Zhang, J. M.; Yang, P. L.; Gray, N. S., Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009,** *9* (1), 28-39.

35. Liu, Y.; Gray, N. S., Rational design of inhibitors that bind to inactive kinase conformations. *Nat. Chem. Biol.* **2006,** *2* (7), 358-364.

36. Wu, P.; Nielsen, T. E.; Clausen, M. H., FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol. Sci.* **2015,** *36* (7), 422-39.

37. Kawata, H.; Yoshida, K.; Kawamoto, A.; Kurioka, H.; Takase, E.; Sasaki, Y.; Hatanaka, K.; Kobayashi, M.; Ueyama, T.; Hashimoto, T.; Dohi, K., Ischemic preconditioning upregulates vascular endothelial growth factor mRNA expression and neovascularization via nuclear translocation of protein kinase C epsilon in the rat ischemic myocardium. *Circ Res* **2001,** *88* (7), 696-704.

38. Salameh, A.; Wustmann, A.; Karl, S.; Blanke, K.; Apel, D.; Rojas-Gomez, D.; Franke, H.; Mohr, F. W.; Janousek, J.; Dhein, S., Cyclic Mechanical Stretch Induces Cardiomyocyte Orientation and Polarization of the Gap Junction Protein Connexin43. *Circ. Res.* **2010,** *106* (10), 1592-1602.

39. Deininger, M. W.; Druker, B. J., Specific targeted therapy of chronic myelogenous leukemia with imatinib. *Pharmacol. Rev.* **2003,** *55* (3), 401-23.

40. Brehmer, D.; Greff, Z.; Godl, K.; Blencke, S.; Kurtenbach, A.; Weber, M.; Muller, S.; Klebl, B.; Cotten, M.; Keri, G.; Wissing, J.; Daub, H., Cellular targets of gefitinib. *Cancer Res.* **2005,** *65* (2), 379-82.

41. Krebs, E. G.; Beavo, J. A., Phosphorylation-dephosphorylation of enzymes. *Annu. Rev. Biochem.* **1979,** *48*, 923-59.

42. Denu, J. M.; Stuckey, J. A.; Saper, M. A.; Dixon, J. E., Form and function in protein dephosphorylation. *Cell* **1996,** *87* (3), 361-4.

43. Chen, M. J.; Dixon, J. E.; Manning, G., Genomics and evolution of protein phosphatases. *Sci. Signal.* **2017,** *10* (474).

44. McConnell, J. L.; Wadzinski, B. E., Targeting protein serine/threonine phosphatases for drug development. *Mol. Pharmacol.* **2009,** *75* (6), 1249-61.

45. Heneberg, P., Use of Protein Tyrosine Phosphatase Inhibitors as Promising Targeted Therapeutic Drugs. *Curr. Med. Chem.* **2009,** *16* (6), 706-733.

46. Anderson, N. L.; Anderson, N. G., Proteome and proteomics: new technologies, new

concepts, and new words. *Electrophoresis* **1998,** *19* (11), 1853-61.

47. Cox, J.; Mann, M., Is proteomics the new genomics? *Cell* **2007,** *130* (3), 395-398.

48. Altelaar, A. F.; Munoz, J.; Heck, A. J., Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* **2013,** *14* (1), 35-48.

49. Baker, M., Proteomics: The interaction map. *Nature* **2012,** *484* (7393), 271-5.

50. Bensimon, A.; Heck, A. J.; Aebersold, R., Mass spectrometry-based proteomics and network biology. *Annu. Rev. Biochem.* **2012,** *81*, 379-405.

51. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003,** *422* (6928), 198-207.

52. Nagaraj, N.; Kulak, N. A.; Cox, J.; Neuhauser, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M., System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap. *Mol. Cell. Proteomics* **2012,** *11* (3).

53. Klykov, O.; Steigenberger, B.; Pektas, S.; Fasci, D.; Heck, A. J. R.; Scheltema, R. A., Efficient and robust proteome-wide approaches for cross-linking mass spectrometry. *Nat. Protoc.* **2018,** *13* (12), 2964-2990.

54. Konermann, L.; Pan, J.; Liu, Y. H., Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem. Soc. Rev.* **2011,** *40* (3), 1224-34.

55. Smith, L. M.; Kelleher, N. L., Proteoforms as the next proteomics currency. *Science* **2018,** *359* (6380), 1106-1107.

56. Siuti, N.; Kelleher, N. L., Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* **2007,** *4* (10), 817-821.

57. Cai, W. X.; Tucholski, T. M.; Gregorich, Z. R.; Ge, Y., Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. *Expert Rev. Proteomic* **2016,** *13* (8), 717-730.

58. Chen, B. F.; Brown, K. A.; Lin, Z. Q.; Ge, Y., Top-Down Proteomics: Ready for Prime Time? *Anal. Chem.* **2018,** *90* (1), 110-127.

59. Toby, T. K.; Fornelli, L.; Kelleher, N. L., Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem.* **2016,** *9*, 499-519.

60. Chait, B. T., Mass spectrometry: Bottom-up or top-down? *Science* **2006,** *314* (5796), 65-66.

61. Cristobal, A.; Marino, F.; Post, H.; van den Toorn, H. W. P.; Mohammed, S.; Heck, A. J. R., Toward an Optimized Workflow for Middle-Down Proteomics. *Anal. Chem.* **2017,** *89* (6), 3318-3325.

62. Jin, Y.; Wei, L.; Cai, W.; Lin, Z.; Wu, Z.; Peng, Y.; Kohmoto, T.; Moss, R. L.; Ge, Y., Complete Characterization of Cardiac Myosin Heavy Chain (223 kDa) Enabled by Size-Exclusion Chromatography and Middle-Down Mass Spectrometry. *Anal. Chem.* **2017,** *89* (9), 4922-4930.

63. Chen, B.; Lin, Z.; Zhu, Y.; Jin, Y.; Larson, E.; Xu, Q.; Fu, C.; Zhang, Z.; Zhang, Q.; Pritts, W. A.; Ge, Y., Middle-Down Multi-Attribute Analysis of Antibody-Drug Conjugates with Electron Transfer Dissociation. *Anal. Chem.* **2019,** *91* (18), 11661-11669.

64. Coradin, M.; Mendoza, M. R.; Sidoli, S.; Alpert, A. J.; Lu, C.; Garcia, B. A., Bullet points

to evaluate the performance of the middle-down proteomics workflow for histone modification analysis. *Methods* **2020**.

65. Sidoli, S.; Garcia, B. A., Middle-down proteomics: a still unexploited resource for chromatin biology. *Expert Rev. Proteomic* **2017,** *14* (7), 617-626.

66. Field, J. K.; Euerby, M. R.; Petersson, P., Investigation into reversed phase chromatography peptide separation systems part III: Establishing a column characterisation database. *J. Chromatogr. A* **2020,** *1622*, 461093.

67. Riley, N. M.; Rush, M. J. P.; Rose, C. M.; Richards, A. L.; Kwiecien, N. W.; Bailey, D. J.; Hebert, A. S.; Westphall, M. S.; Coon, J. J., The Negative Mode Proteome with Activated Ion Negative Electron Transfer Dissociation (AI-NETD). *Mol. Cell. Proteomics* **2015,** *14* (10), 2644.

68. Wu, C.; Liang, Y.; Yang, K.; Min, Y.; Liang, Z.; Zhang, L.; Zhang, Y., Clickable Periodic Mesoporous Organosilica Monolith for Highly Efficient Capillary Chromatographic Separation. *Anal. Chem.* **2016,** *88* (3), 1521-1525.

69. Liang, Y.; Jin, Y.; Wu, Z.; Tucholski, T.; Brown, K. A.; Zhang, L.; Zhang, Y.; Ge, Y., Bridged Hybrid Monolithic Column Coupled to High-Resolution Mass Spectrometry for Top-Down Proteomics. *Anal. Chem.* **2019,** *91* (3), 1743-1747.

70. El Deeb, S.; Preu, L.; Wätzig, H., A strategy to develop fast RP-HPLC methods using monolithic silica columns. *J. Sep. Sci.* **2007,** *30* (13), 1993-2001.

71. Yan, Y.; Xing, T.; Wang, S.; Daly, T. J.; Li, N., Online coupling of analytical hydrophobic interaction chromatography with native mass spectrometry for the characterization of monoclonal antibodies and related products. *J. Pharmaceut. Biomed.* **2020,** *186*, 113313.

72. Chen, B.; Lin, Z.; Alpert, A. J.; Fu, C.; Zhang, Q.; Pritts, W. A.; Ge, Y., Online Hydrophobic Interaction Chromatography–Mass Spectrometry for the Analysis of Intact Monoclonal Antibodies. *Anal. Chem.* **2018,** *90* (12), 7135-7138.

73. Alpert, A. J., Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *J. Chromatogr. A* **1990,** *499*, 177-196.

74. Lauber, M. A.; Yu, Y.-Q.; Brousmiche, D. W.; Hua, Z.; Koza, S. M.; Magnelli, P.; Guthrie, E.; Taron, C. H.; Fountain, K. J., Rapid Preparation of Released N-Glycans for HILIC Analysis Using a Labeling Reagent that Facilitates Sensitive Fluorescence and ESI-MS Detection. *Anal. Chem.* **2015,** *87* (10), 5401-5409.

75. Bones, J.; Mittermayr, S.; O'Donoghue, N.; Guttman, A.; Rudd, P. M., Ultra Performance Liquid Chromatographic Profiling of Serum N-Glycans for Fast and Efficient Identification of Cancer Associated Alterations in Glycosylation. *Anal. Chem.* **2010,** *82* (24), 10208-10215.

76. Carroll, J.; Fearnley, I. M.; Walker, J. E., Definition of the mitochondrial proteome by measurement of molecular masses of membrane proteins. *P. Natl. Acad. Sci. USA* **2006,** *103* (44), 16170.

77. Compton, P. D.; Zamdborg, L.; Thomas, P. M.; Kelleher, N. L., On the Scalability and Requirements of Whole Protein Mass Spectrometry. *Anal. Chem.* **2011,** *83* (17), 6868-6874.

78. Cai, W.; Tucholski, T.; Chen, B.; Alpert, A. J.; McIlwain, S.; Kohmoto, T.; Jin, S.; Ge, Y., Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion

Chromatography Strategy. *Anal. Chem.* **2017,** *89* (10), 5467-5475.

79. Tucholski, T.; Knott, S. J.; Chen, B.; Pistono, P.; Lin, Z.; Ge, Y., A Top-Down Proteomics Platform Coupling Serial Size Exclusion Chromatography and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem.* **2019,** *91* (6), 3835-3844.

80. Cummins, P. M.; Rochfort, K. D.; O'Connor, B. F., Ion-Exchange Chromatography: Basic Principles and Application. *Protein Chromatography: Methods and Protocols, 2nd Edition* **2017,** *1485*, 209-223.

81. Dai, J.; Wang, L.-S.; Wu, Y.-B.; Sheng, Q.-H.; Wu, J.-R.; Shieh, C.-H.; Zeng, R., Fully Automatic Separation and Identification of Phosphopeptides by Continuous pH-Gradient Anion Exchange Online Coupled with Reversed-Phase Liquid Chromatography Mass Spectrometry. *J. Proteome Res.* **2009,** *8* (1), 133-141.

82. Fussl, F.; Criscuolo, A.; Cook, K.; Scheffler, K.; Bones, J., Cracking Proteoform Complexity of Ovalbumin with Anion-Exchange Chromatography-High-Resolution Mass Spectrometry under Native Conditions. *J. Proteome Res.* **2019,** *18* (10), 3689-3702.

83. Camerini, S.; Mauri, P., The role of protein and peptide separation before mass spectrometry analysis in clinical proteomics. *J. Chromatogr. A* **2015,** *1381*, 1-12.

84. Gupta, S. P.; Hanash, S. M., Separation of hemoglobin types by cation-exchange high-performance liquid chromatography. *Anal. Biochem.* **1983,** *134* (1), 117-121.

85. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989,** *246* (4926), 64-71.

86. Juraschek, R.; Dülcks, T.; Karas, M., Nanoelectrospray—more than just a minimized-flow electrospray ionization source. *J. Am. Soc. Mass Spectrom.* **1999,** *10* (4), 300-308.

87. Hillenkamp, F.; Karas, M.; Beavis, R. C.; Chait, B. T., Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.* **1991,** *63* (24), 1193A-1203A.

88. Douglas, D. J.; Frank, A. J.; Mao, D., Linear ion traps in mass spectrometry. *Mass Spectrom. Rev.* **2005,** *24* (1), 1-29.

89. Yost, R. A.; Enke, C. G., Selected Ion Fragmentation with a Tandem Quadrupole Mass-Spectrometer. *J. Am. Chem. Soc.* **1978,** *100* (7), 2274-2275.

90. Johnson, J. V.; Yost, R. A.; Kelley, P. E.; Bradford, D. C., Tandem-in-Space and Tandem-in-Time Mass-Spectrometry - Triple Quadrupoles and Quadrupole Ion Traps. *Anal. Chem.* **1990,** *62* (20), 2162-2172.

91. Wiley, W. C.; Mclaren, I. H., Time-of-Flight Mass Spectrometer with Improved Resolution. *Rev. Sci. Instrum.* **1955,** *26* (12), 1150-1157.

92. Zubarev, R. A.; Makarov, A., Orbitrap Mass Spectrometry. *Anal. Chem.* **2013,** *85* (11), 5288-5296.

93. Makarov, A. A., Orbitrap mass spectrometry in proteomics: past, present and future. *Febs J.* **2013,** *280*, 632-632.

94. Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S., Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom. Rev.* **1998,** *17* (1), 1-35.

95. Cavaliere, C.; Antonelli, M.; Capriotti, A. L.; La Barbera, G.; Montone, C. M.; Piovesana, S.; Lagana, A., A Triple Quadrupole and a Hybrid Quadrupole Orbitrap Mass Spectrometer in Comparison for Polyphenol Quantitation. *J. Agr. Food. Chem.* **2019,** *67* (17), 4885-4896.

96. Beck, S.; Michalski, A.; Raether, O.; Lubeck, M.; Kaspar, S.; Goedecke, N.; Baessmann, C.; Hornburg, D.; Meier, F.; Paron, I.; Kulak, N. A.; Cox, J.; Mann, M., The Impact II, a Very High-Resolution Quadrupole Time-of-Flight Instrument (QTOF) for Deep Shotgun Proteomics. *Mol. Cell. Proteomics* **2015,** *14* (7), 2014-2029.

97. McDonnell, L. A.; Heeren, R. M. A., Imaging mass spectrometry. *Mass Spectrom. Rev.* **2007,** *26* (4), 606-643.

98. Mclafferty, F. W., Tandem Mass-Spectrometry (Ms-Ms) - Promising New Analytical Technique for Specific Component Determination in Complex-Mixtures. *Acc. Chem. Res.* **1980,** *13* (2), 33-39.

99. Wells, J. M.; McLuckey, S. A., Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol.* **2005,** *402*, 148-85.

100. Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M., Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007,** *4* (9), 709-12.

101. Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W., Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.* **1998,** *120* (13), 3265-3266.

102. Williams, J. P.; Morrison, L. J.; Brown, J. M.; Beckman, J. S.; Voinov, V. G.; Lermyte, F., Top-Down Characterization of Denatured Proteins and Native Protein Complexes Using Electron Capture Dissociation Implemented within a Modified Ion Mobility-Mass Spectrometer. *Anal. Chem.* **2020,** *92* (5), 3674-3681.

103. Shaw, J. B.; Malhan, N.; Vasil'ev, Y. V.; Lopez, N. I.; Makarov, A.; Beckman, J. S.; Voinov, V. G., Sequencing Grade Tandem Mass Spectrometry for Top–Down Proteomics Using Hybrid Electron Capture Dissociation Methods in a Benchtop Orbitrap Mass Spectrometer. *Anal. Chem.* **2018,** *90* (18), 10819-10827.

104. Zhou, M. W.; Liu, W. J.; Shaw, J. B., Charge Movement and Structural Changes in the Gas-Phase Unfolding of Multimeric Protein Complexes Captured by Native Top-Down Mass Spectrometry. *Anal. Chem.* **2020,** *92* (2), 1788-1795.

105. Syka, J. E.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *P. Natl. Acad. Sci. USA* **2004,** *101* (26), 9528-33.

106. Ledvina, A. R.; McAlister, G. C.; Gardner, M. W.; Smith, S. I.; Madsen, J. A.; Schwartz, J. C.; Stafford Jr., G. C.; Syka, J. E. P.; Brodbelt, J. S.; Coon, J. J., Infrared Photoactivation Reduces Peptide Folding and Hydrogen-Atom Migration following ETD Tandem Mass Spectrometry. *Angew. Chem. Int. Ed.* **2009,** *48* (45), 8526-8528.

107. Ledvina, A. R.; Beauchene, N. A.; McAlister, G. C.; Syka, J. E. P.; Schwartz, J. C.; Griep-Raming, J.; Westphall, M. S.; Coon, J. J., Activated-Ion Electron Transfer Dissociation Improves the Ability of Electron Transfer Dissociation to Identify Peptides in a Complex Mixture. *Anal. Chem.* **2010,** *82* (24), 10068-10074.

108. Yu, Q.; Wang, B.; Chen, Z.; Urabe, G.; Glover, M. S.; Shi, X.; Guo, L.-W.; Kent, K. C.; Li, L., Electron-Transfer/Higher-Energy Collision Dissociation (EThcD)-Enabled Intact Glycopeptide/Glycoproteome Characterization. *J. Am. Soc. Mass Spectrom.* **2017,** *28* (9), 1751-1764.

109. Khatri, K.; Pu, Y.; Klein, J. A.; Wei, J.; Costello, C. E.; Lin, C.; Zaia, J., Comparison of Collisional and Electron-Based Dissociation Modes for Middle-Down Analysis of Multiply Glycosylated Peptides. *J. Am. Soc. Mass Spectrom.* **2018,** *29* (6), 1075-1085.

110. Mommen, G. P. M.; Frese, C. K.; Meiring, H. D.; van Gaans-van den Brink, J.; de Jong, A. P. J. M.; van Els, C. A. C. M.; Heck, A. J. R., Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *P. Natl. Acad. Sci. USA* **2014,** *111* (12), 4507.

111. Little, D. P.; Speir, J. P.; Senko, M. W.; O'Connor, P. B.; McLafferty, F. W., Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing. *Anal. Chem.* **1994,** *66* (18), 2809-15.

112. Brodbelt, J. S., Photodissociation mass spectrometry: new tools for characterization of biological molecules. *Chem. Soc. Rev.* **2014,** *43* (8), 2757-83.

113. Quick, M. M.; Crittenden, C. M.; Rosenberg, J. A.; Brodbelt, J. S., Characterization of Disulfide Linkages in Proteins by 193 nm Ultraviolet Photodissociation (UVPD) Mass Spectrometry. *Anal. Chem.* **2018,** *90* (14), 8523-8530.

114. Williams, P. E.; Klein, D. R.; Greer, S. M.; Brodbelt, J. S., Pinpointing Double Bond and sn-Positions in Glycerophospholipids via Hybrid 193 nm Ultraviolet Photodissociation (UVPD) Mass Spectrometry. *J. Am. Chem. Soc.* **2017,** *139* (44), 15681-15690.

115. Zhou, M.; Wysocki, V. H., Surface induced dissociation: dissecting noncovalent protein complexes in the gas phase. *Acc. Chem. Res.* **2014,** *47* (4), 1010-8.

116. Stiving, A. Q.; VanAernum, Z. L.; Busch, F.; Harvey, S. R.; Sarni, S. H.; Wysocki, V. H., Surface-Induced Dissociation: An Effective Method for Characterization of Protein Quaternary Structure. *Anal. Chem.* **2019,** *91* (1), 190-209.

117. Harvey, S. R.; Liu, Y.; Liu, W.; Wysocki, V. H.; Laganowsky, A., Surface induced dissociation as a tool to study membrane protein complexes. *Chem. Commun.* **2017,** *53* (21), 3106-3109.

118. Zhou, M.; Yan, J.; Romano, C. A.; Tebo, B. M.; Wysocki, V. H.; Pasa-Tolic, L., Surface Induced Dissociation Coupled with High Resolution Mass Spectrometry Unveils Heterogeneity of a 211 kDa Multicopper Oxidase Protein Complex. *J. Am. Soc. Mass. Spectrom.* **2018,** *29* (4), 723-733.

119. Jin, Y.; Lin, Z.; Xu, Q.; Fu, C.; Zhang, Z.; Zhang, Q.; Pritts, W. A.; Ge, Y., Comprehensive characterization of monoclonal antibody by Fourier transform ion cyclotron resonance mass spectrometry. *MAbs* **2019,** *11* (1), 106-115.

120. Li, H.; Sheng, Y.; McGee, W.; Cammarata, M.; Holden, D.; Loo, J. A., Structural Characterization of Native Proteins and Protein Complexes by Electron Ionization Dissociation-Mass Spectrometry. *Anal. Chem.* **2017,** *89* (5), 2731-2738.

121. Li, H. L.; Nguyen, H. H.; Loo, R. R. O.; Campuzano, I. D. G.; Loo, J. A., An integrated

native mass spectrometry and top-down proteomics method that connects sequence to structure and function of macromolecular complexes. *Nat. Chem.* **2018,** *10* (2), 139-148.

122. Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **2007,** *389* (4), 1017-1031.

123. Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B., Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **2012,** *404* (4), 939-65.

124. Ross, P. L.; Huang, Y. L. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004,** *3* (12), 1154-1169.

125. Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C., Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **2003,** *75* (8), 1895-1904.

126. Frost, D. C.; Greer, T.; Li, L. J., High-Resolution Enabled 12-Plex DiLeu Isobaric Tags for Quantitative Proteomics. *Anal. Chem.* **2015,** *87* (3), 1646-1654.

127. Frost, D. C.; Feng, Y.; Li, L. J., 21-plex DiLeu Isobaric Tags for High-Throughput Quantitative Proteomics. *Anal. Chem.* **2020,** *92* (12), 8228-8234.

128. Greer, T.; Hao, L.; Nechyporenko, A.; Lee, S.; Vezina, C. M.; Ricke, W. A.; Marker, P. C.; Bjorling, D. E.; Bushman, W.; Li, L. J., Custom 4-Plex DiLeu Isobaric Labels Enable Relative Quantification of Urinary Proteins in Men with Lower Urinary Tract Symptoms (LUTS). *Plos One* **2015,** *10* (8).

129. Ong, S. E.; Mann, M., A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* **2006,** *1* (6), 2650-2660.

130. Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M., Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **2002,** *1* (5), 376-386.

131. Asara, J. M.; Christofk, H. R.; Freimark, L. M.; Cantley, L. C., A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen. *Proteomics* **2008,** *8* (5), 994-9.

132. Lin, Z.; Wei, L.; Cai, W.; Zhu, Y.; Tucholski, T.; Mitchell, S. D.; Guo, W.; Ford, S. P.; Diffee, G. M.; Ge, Y., Simultaneous Quantification of Protein Expression and Modifications by Top-down Targeted Proteomics: A Case of the Sarcomeric Subproteome. *Mol. Cell. Proteomics* **2019,** *18* (3), 594-605.

133. Collier, T. S.; Sarkar, P.; Rao, B.; Muddiman, D. C., Quantitative top-down proteomics of SILAC labeled human embryonic stem cells. *J. Am. Soc. Mass Spectrom.* **2010,** *21* (6), 879-89.

134. Waanders, L. F.; Hanke, S.; Mann, M., Top-down quantitation and characterization of SILAC-labeled proteins. *J. Am. Soc. Mass Spectrom.* **2007,** *18* (11), 2058-64.

135. Hung, C. W.; Tholey, A., Tandem mass tag protein labeling for top-down identification and

quantification. *Anal. Chem.* **2012,** *84* (1), 161-70.

136. Paulo, J. A.; Navarrete-Perea, J.; Erickson, A. R.; Knott, J.; Gygi, S. P., An Internal Standard for Assessing Phosphopeptide Recovery from Metal Ion/Oxide Enrichment Strategies. *J. Am. Soc. Mass Spectrom.* **2018,** *29* (7), 1505-1511.

137. Chen, S. Y.; Juang, Y. M.; Chien, M. W.; Li, K. I.; Yu, C. S.; Lai, C. C., Magnetic iron oxide nanoparticle enrichment of phosphopeptides on a radiate microstructure MALDI chip. *Analyst* **2011,** *136* (21), 4454-9.

138. Kinoshita, E.; Kinoshita-Kikuta, E.; Koike, T., Separation and detection of large phosphoproteins using Phos-tag SDS-PAGE. *Nat. Protoc.* **2009,** *4* (10), 1513-1521.

139. Hwang, L.; Ayaz-Guner, S.; Gregorich, Z. R.; Cai, W. X.; Valeja, S. G.; Jin, S.; Ge, Y., Specific Enrichment of Phosphoproteins Using Functionalized Multivalent Nanoparticles. *J. Am. Chem. Soc.* **2015,** *137* (7), 2432-2435.

140. Chen, B. F.; Hwang, L.; Ochowicz, W.; Lin, Z. Q.; Guardado-Alvarez, T. M.; Cai, W. X.; Xiu, L. C.; Dani, K.; Colah, C.; Jin, S.; Ge, Y., Coupling functionalized cobalt ferrite nanoparticle enrichment with online LC/MS/MS for top-down phosphoproteomics. *Chem. Sci.* **2017,** *8* (6), 4306-4311.

141. Roberts, D. S.; Chen, B. F.; Tiambeng, T. N.; Wu, Z. J.; Ge, Y.; Jin, S., Reproducible large-scale synthesis of surface silanized nanoparticles as an enabling nanoproteomics platform: Enrichment of the human heart phosphoproteome. *Nano Res.* **2019,** *12* (6), 1473-1481.

142. Wissing, J.; Jansch, L.; Nimtz, M.; Dieterich, G.; Hornberger, R.; Keri, G.; Wehland, J.; Daub, H., Proteomics analysis of protein kinases by target class-selective prefractionation and tandem mass spectrometry. *Mol. Cell. Proteomics* **2007,** *6* (3), 537-547.

143. Bantscheff, M.; Eberhard, D.; Abraham, Y.; Bastuck, S.; Boesche, M.; Hobson, S.; Mathieson, T.; Perrin, J.; Raida, M.; Rau, C.; Reader, V.; Sweetman, G.; Bauer, A.; Bouwmeester, T.; Hopf, C.; Kruse, U.; Neubauer, G.; Ramsden, N.; Rick, J.; Kuster, B.; Drewes, G., Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat. Biotechnol.* **2007,** *25* (9), 1035-1044.

144. Goettig, P., Effects of Glycosylation on the Enzymatic Activity and Mechanisms of Proteases. *Int. J. Mol. Sci.* **2016,** *17* (12).

145. Ohtsubo, K.; Marth, J. D., Glycosylation in cellular mechanisms of health and disease. *Cell* **2006,** *126* (5), 855-867.

146. Pinho, S. S.; Reis, C. A., Glycosylation in cancer: mechanisms and clinical implications. *Nat. Rev. Cancer* **2015,** *15* (9), 540-555.

147. Reily, C.; Stewart, T. J.; Renfrow, M. B.; Novak, J., Glycosylation in health and disease. *Nat. Rev. Nephrol.* **2019,** *15* (6), 346-366.

148. Mechref, Y.; Madera, M.; Novotny, M. V., Glycoprotein enrichment through lectin affinity techniques. *Methods Mol. Biol.* **2008,** *424*, 373-96.

149. Clark, E. L.; Emmadi, M.; Krupp, K. L.; Podilapu, A. R.; Helble, J. D.; Kulkarni, S. S.; Dube, D. H., Development of Rare Bacterial Monosaccharide Analogs for Metabolic Glycan Labeling in Pathogenic Bacteria. *ACS Chem. Biol.* **2016,** *11* (12), 3365-3373.

150. Wang, S.; Yang, F.; Camp, D. G., 2nd; Rodland, K.; Qian, W. J.; Liu, T.; Smith, R. D., Proteomic approaches for site-specific O-GlcNAcylation analysis. *Bioanalysis* **2014,** *6* (19), 2571-80.

151. Huang, J.; Wan, H.; Yao, Y.; Li, J.; Cheng, K.; Mao, J.; Chen, J.; Wang, Y.; Qin, H.; Zhang, W.; Ye, M.; Zou, H., Highly Efficient Release of Glycopeptides from Hydrazide Beads by Hydroxylamine Assisted PNGase F Deglycosylation for N-Glycoproteome Analysis. *Anal. Chem.* **2015,** *87* (20), 10199-204.

152. Melmer, M.; Stangler, T.; Schiefermeier, M.; Brunner, W.; Toll, H.; Rupprechter, A.; Lindner, W.; Premstaller, A., HILIC analysis of fluorescence-labeled N-glycans from recombinant biopharmaceuticals. *Anal. Bioanal. Chem.* **2010,** *398* (2), 905-14.

153. Qing, G. Y.; Yan, J. Y.; He, X. N.; Li, X. L.; Liang, X. M., Recent advances in hydrophilic interaction liquid interaction chromatography materials for glycopeptide enrichment and glycan separation. *Trac-Trend. Anal. Chem.* **2020,** *124*.

154. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012,** *30* (10), 918-20.

155. Kou, Q.; Wu, S.; Liu, X. W., A new scoring function for top-down spectral deconvolution. *BMC Genomics* **2014,** *15*.

156. Senko, M. W.; Beu, S. C.; McLaffertycor, F. W., Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995,** *6* (4), 229-33.

157. Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A., Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Mol. Cell. Proteomics* **2010,** *9* (12), 2772-82.

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

159. Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Liu, T.; Petyuk, V. A.; Tolic, N.; Pasa-Tolic, L.; Smith, R. D.; Payne, S. H.; Kim, S., Informed-Proteomics: open-source software package for top-down proteomics. *Nat. Methods* **2017,** *14* (9), 909-914.

160. Jeong, K.; Kim, J.; Gaikwad, M.; Hidayah, S. N.; Heikaus, L.; Schluter, H.; Kohlbacher, O., FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics. *Cell Syst.* **2020,** *10* (2), 213-+.

161. Vyatkina, K.; Wu, S.; Dekker, L. J.; VanDuijn, M. M.; Liu, X.; Tolic, N.; Dvorkin, M.; Alexandrova, S.; Luider, T. M.; Pasa-Tolic, L.; Pevzner, P. A., De Novo Sequencing of Peptides from Top-Down Tandem Mass Spectra. *J. Proteome Res.* **2015,** *14* (11), 4450-62.

162. Cesnik, A. J.; Shortreed, M. R.; Schaffer, L. V.; Knoener, R. A.; Frey, B. L.; Scalf, M.;

Solntsev, S. K.; Dai, Y. X.; Gasch, A. P.; Smith, L. M., Proteoform Suite: Software for Constructing, Quantifying, and Visualizing Proteoform Families. *J. Proteome Res.* **2018,** *17* (1), 568-578.

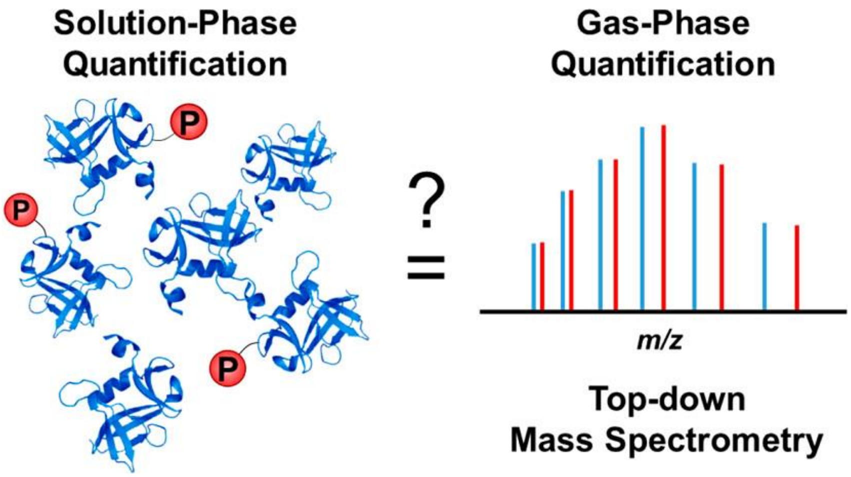
163. Schaffer, L. V.; Tucholski, T.; Shortreed, M. R.; Ge, Y.; Smith, L. M., Intact-Mass Analysis Facilitating the Identification of Large Human Heart Proteoforms. *Anal. Chem.* **2019,** *91* (17), 10937-10942.

164. Sun, R. X.; Luo, L.; Wu, L.; Wang, R. M.; Zeng, W. F.; Chi, H.; Liu, C.; He, S. M., pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification. *Anal. Chem.* **2016,** *88* (6), 3082-3090.

165. McIlwain, S. J.; Wu, Z. J.; Wetzel, M.; Belongia, D.; Jin, Y. T.; Wenger, K.; Ong, I. M.; Ge, Y., Enhancing Top-Down Proteomics Data Analysis by Combining Deconvolution Results through a Machine Learning Strategy. *J. Am. Soc. Mass Spectrom.* **2020,** *31* (5), 1104-1113.

Chapter 2

Impact of Phosphorylation on the Mass Spectrometry Quantification of Intact Phosphoproteins



Adapted from: Wu, Z.; Tiambeng, T. N.; Cai, W.; Chen, B.; Lin, Z.; Gregorich, Z. R.; Ge, Y.*, Impact of Phosphoylation on the Mass Spectrometry Quantitation of Phosphoproteins. *Anal. Chem.* **2018**, *90* (8), 4935-4939.

Abstract

Protein phosphorylation is a ubiquitous and critical PTM involved in numerous cellular processes. MS-based proteomics has emerged as the preferred technology for protein identification, characterization, and quantification. Whereas ionization/detection efficiency of peptides in ESI-MS are markedly influenced by the presence of phosphorylation, the physicochemical properties of intact proteins are assumed not to vary significantly due to the relatively smaller modification on large intact proteins. Thus the ionization/detection efficiency of intact phosphoprotein is hypothesized not to alter appreciably for subsequent MS quantification. However, this hypothesis has never been rigorously tested. Herein, we systematically investigated the impact of phosphorylation on ESI-MS quantification of mono- and multiply-phosphorylated proteins. We verified that a single phosphorylation did not appreciably affect the ESI-MS quantification of phosphoproteins as demonstrated in the enigma homolog isoform 2 (28 kDa) with mono-phosphorylation. Moreover, different ionization and desolvation parameters did not impact phosphoprotein quantification. In contrast to mono-phosphorylation, multi-phosphorylation noticeably affected ESI-MS quantification of phosphoproteins likely due to differential ionization/detection efficiency between unphosphorylated and phosphorylated proteoforms as shown in the pentakis-phosphorylated β-casein (24 kDa).

Introduction

Protein phosphorylation is an important PTM that is involved in many critical cellular processes, including cell cycle control, cell growth, apoptosis, and signaling transduction pathways.[1-2] Not surprisingly, altered phosphorylation levels have been associated with the development of diseases such as cardiovascular disease, cancer, and neurodegenerative disease.[3-6] Moreover, recent evidence indicates that protein phosphorylation may also be useful as potential disease biomarkers.[7] Therefore, accurate quantification of protein phosphorylation in different biological states not only can help elucidate intracellular signaling pathways that regulate various cellular processes, but may also be useful in understanding disease mechanism and diagnosis.[8-9]

MS-based proteomics has emerged as the preferred method for phosphoprotein identification, characterization, and quantification.[10-12] The bottom-up MS-based phosphoproteomics approach, which commonly utilizes proteases to digest phosphoproteins into smaller peptides, is a high throughput method for quantification of phosphoproteins.[13,14] However, ionization/detection efficiency of peptides in ESI-MS are markedly influenced by the presence of phosphorylation.[15] Recently, top-down MS-based proteomics has emerged as the foremost method for the identification and quantification of proteoforms, a term adopted to represent the myriad protein products of a single gene generated via sequence variations (as a consequence of mutations/polymorphisms and/or alternative splicing), as well as PTMs.[16] In top-down MS, intact proteins are analyzed, providing a “bird’s eye” view of all observed proteoforms in a given sample.[17-18] Moreover, as the physicochemical properties of intact proteins are believed to be less impacted than peptides by the addition of smaller PTMs (e.g., phosphorylation), it is hypothesized that the effect of the small modifications on the ionization/detection efficiency of intact proteins will be negligible. Thus, top-down MS has been routinely employed for the quantification of modified and un-modified protein species in biological samples.[3, 19-20]

In support of the long-held belief that the ionization/detection efficiency of intact proteins is not significantly impacted by PTMs, Kelleher and co-workers have demonstrated that the difference in ionization/detection efficiency between un-modified and acetylated intact recombinant H4 protein was minimal as observed by the small deviation (<5%) between protein ion relative ratios and solution ratio.[21] Therefore, accurate relative quantification of un-modified and acetylated proteoforms can be achieved by top-down MS analysis. On the other hand, some evidence suggests that phosphorylation can have a dramatic impact on the physicochemical properties of proteins such as hydrophobicity, viscosity, and side chain flexibility.[22,23] However, it remains unclear whether the ionization/detection efficiency of proteins will be significantly altered by phosphorylation.

Herein we systematically investigated the impact of phosphorylation on the ionization/detection efficiency of intact proteins using a mono-phosphorylated protein, enigma homolog isoform 2 (ENH2), and a multiply-phosphorylated protein, β-casein, as model phosphoproteins. Phosphorylation or dephosphorylation reaction was achieved by *in vitro* kinase or phosphatase reaction. With these model systems, we varied the solution-phase ratio of unphosphorylated and phosphorylated proteins quantified by SDS-PAGE analysis, which allowed us to evaluate the impact of phosphorylation on top-down phosphoprotein quantification analysis. We also investigated the ESI-MS response in two different types of mass spectrometers, TOF and FT-ICR, and assessed the correlation between solution-phase ratios of proteins derived from SDS-gel analysis, and their respective gas-phase ratios as measured by these two mass spectrometers.

Experimental Section

Chemicals and Reagents

All reagents were acquired from Sigma-Aldrich, Inc. (St. Louis, MO, USA), unless otherwise noted. Solvents, including HPLC grade H_2O and ACN, were purchased from Fisher Scientific (Fair Lawn, NJ, USA).

Molecular Cloning

The coding sequence of the ENH2 protein was amplified by polymerase chain reaction using 5’-GCAGCCTCGAGGTATGAGCAACTACAGTGTGTCACTGG-3’ and 5’-ATGAGAATTCGTCTGTACGTTAAGAGCACGTGCTGA-3’ as the forward and reverse primers, respectively, and the product was ligated into a pT7-flag vector (Sigma-Aldrich, Inc.) to add a C-terminal FLAG-tag to the protein. Subsequently, the sequence encoding the FLAG-tagged ENH2 protein was again amplified by polymerase chain reaction using the 5’-AGGTACCATGGGGAGCAACTACAGTGTGTCACTGGTT-3’ and 3’-GGTGGTGCTCGAGCTTGTCATCGTCGTCCTTGTAG-5’ as forward and reverse primers, respectively. The amplified sequence was ligated into the pET-28a vector (MilliporeSigma, Burlington, MA, USA), adding a C-terminal polyhistidine-tag following the FLAG-tag. The ligation product was transformed into ScarabXpress T7 *E. coli* (Scarab Genomics, Madison, WI, USA) and the recombinant DNA sequence was confirmed by sequencing.

Protein Expression and Purification

A starter culture (50 mL LB broth with 50 µg/mL kanamycin) was prepared by inoculating approximately 10 µL of the glycerol stock of the *E. coli*, for overnight at 30 ⁰C. 25 mL of the starter culture was diluted in 1 L of LB broth containing 50 µg/mL kanamycin and the culture was incubated at 37 ⁰C with 250 rpm shaking until an optical density of 0.4-0.6 was reached. Protein expression was induced by 0.1 mM IPTG and the bacteria were cultured at 37 ⁰C with 250 rpm shaking for 6 hrs. The cells were harvested by centrifugation at 4,000 rpm for 10 min, and the pellet was stored at -80 ⁰C. The *E. coli* cells were lysed by sonication in 50 mM NaH_2PO_4 pH 7.4, 250 mM NaCl (10 mL/g pellet) buffer containing 1 mM DTT, 0.25 mM PMSF and 100x protease inhibitor cocktail (Sigma-Aldrich Inc.). Sonication was performed for 5 cycles, 20 s per cycle, followed by centrifugation at 4000 rpm for 20 min, and the supernatant was collected. ENH2 protein was purified from the cell lysate by initially binding to silica-based PrepEase® Histidine-tagged High Specificity Purification Resin (Affymetrix, Santa Clara, CA, USA) for two hours before washing with 50 mM NaH_2PO_4 pH 7.4, 250 mM NaCl buffer containing 1 mM DTT and 0.25 mM PMSF, followed by 50 mM Tris pH 7.4, 50 mM NaCl buffer containing 1 mM DTT and 0.25 mM PMSF. The resin was subsequently washed with 50 mM Tris pH 7.4, 50 mM NaCl buffer containing 1 mM DTT, 100 mM imidazole, and 0.25 mM PMSF, and the protein was eluted from the resin with 50 mM Tris pH 7.4, 50 mM NaCl buffer containing 1 mM DTT, 50 mM EDTA, 0.25 mM PMSF and 100x protease inhibitor cocktail (Sigma-Aldrich Inc.). The eluted protein was concentrated and the EDTA salt was washed off by 50 mM Tris pH 7.4, 50 mM NaCl buffer using a Pierce™ Protein Concentrators PES, 10K MWCO filter (Fisher Scientific) before the phosphorylation reaction.

ENH2 Phosphorylation Reaction and β-casein Dephosphorylation Reaction

The phosphorylation reaction for purified ENH2 protein (~100 µg) was performed using ~1,250 units of PKA C-subunit (New England Biolabs Inc.), supplemented with 1X NEBuffer for Protein Kinases (PK), 200 µM ATP (New England Biolabs Inc.) incubated at 37 °C for 24 hrs to achieve complete phosphorylation. The dephosphorylation reaction for commercial β-casein (~40 µg) (Protea, Morgantown, WV, USA) was performed using ~100 units of λPP (New England Biolabs Inc.), supplemented with 1 mM $MnCl_2$ solution and 1X NEBuffer for PMP (New England Biolabs Inc.). The reaction was allowed to proceed at 30 °C for 2 hrs to achieve complete dephosphorylation of β-casein.

Solution-Phase Protein Concentration Analysis

For both ENH2 and β-casein, the protein concentrations of the stock solutions were first quantified using the Bio-Rad Protein Assay Dye Reagent (Bio-Rad Laboratories, Inc., Hercules, CA, USA) in accordance with the manufacturer's instructions. Subsequently, to establish the linear response range of SDS-PAGE for the target proteins, known amounts of ENH2 and β-casein were loaded in the polyacrylamide gel to establish a gel-based standard curve. After electrophoresis, the gel was visualized by either Coomassie Blue R-250 and destained overnight when using 12.5% Criterion™ Tris-HCl Protein Gel (Bio-Rad Laboratories, Inc.), or by Bio-Rad ChemiDoc™ MP System with Image Lab software when using the 8–16% Criterion Stain Free™ Tris-HCl Protein Gel (Bio-Rad Laboratories, Inc.). Band intensities were quantified using ImageJ[24] by integrating the area under the curve for each band. SDS-PAGE analysis showed a strong linear correlation for ENH2 and β-casein between 0.25 – 2 µg and 0.1 – 0.7 µg respectively (Figures S2.1). Thus, the amount of protein loaded in the polyacrylamide gel was within the linear range for each respective protein and the SDS-gel data was used to derive the solution-phase ratio.

Top-down Mass Spectrometry

Unless otherwise stated, all samples were diluted in 50:50 H_2O:ACN with 0.1% FA for MS and MS/MS analyses. The mixtures containing ENH2:*p*ENH2 or β-casein:5*p*β-casein were subject to MS analysis using either a Bruker maXis II Q-TOF mass spectrometer or a 12T solariX FT-ICR mass spectrometer (Bruker Daltonics, Bremen, Germany). The mass spectra obtained from the maXis II Q-TOF mass spectrometer were obtained in triplicate by direct infusion. Unless otherwise stated, mass spectra collected using the Q-TOF mass spectrometer were collected using a 4500 V spray voltage, 20 eV in-source CID (isCID) voltage, 2 L/min dry gas flow rate, 200 °C dry gas temperature, and a 1.5 bar nebulizer gas pressure. All spectra were collected for a 30 s duration. Sample were injected into the Q-TOF mass spectrometer at a sample flow rate of 2 µL/min. Mass spectra from the 12T solariX FT-ICR mass spectrometer were acquired in triplicate, using a 2 mega-words transient, 0.004 s acquisition time, and 20 eV isCID voltage. A fixed 48 scans were collected per injection. Samples were injected into the 12T solariX FT-ICR using a TriVersa NanoMate® (Advion Bioscience, Ithaca, NY, USA) with a spray voltage of 1.3 kV.

ECD was performed on a Thermo LTQ/FT Ultra 7T FT-ICR mass spectrometer (Thermo Scientific Inc., Bremen, Germany) for ENH2 and a 12T solariX FT-ICR mass spectrometer for β-casein. Mass spectra from the Thermo LTQ/FT Ultra 7T FT-ICR mass spectrometer used a resolving power of 200,000 (at 400 *m/z*). The samples were introduced into the mass spectrometer using a TriVersa NanoMate® as previously described.[3,25] The energy, delay, and duration parameters for ECD experiments were determined on a case-by-case basis to optimize protein fragmentation. On average, between 700 and 1300 scans were collected to ensure the collection of high-quality tandem mass spectra for data analysis. Mass spectra from the 12T solariX FT-ICR mass spectrometer were obtained as previously described.[26]

All reported MWs are the most abundant MWs. Data acquired on the Bruker maXis II Q-TOF mass spectrometer were analyzed using DataAnalysis 4.0 and deconvoluted using the Maximum Entropy algorithm. The relative quantification of proteoforms for deconvoluted spectra and individual charge state was determined by comparing intensities of the peaks from the SNAP algorithm results in DataAnalysis. The phosphate adducts (H_3PO_4) are non-covalent adducts to the protein molecular ions resulting in a 98 Da mass increase. Since they are non-covalently bound to the protein molecular ions, they are not considered as distinct proteoforms. For the relative quantification of either ENH2 or β-casein, proteoforms containing the phosphate adducts were added back to their parent proteoform before subsequent calculation.[27-28]

For β-casein analysis, the data acquired on the Bruker maXis II Q-TOF mass spectrometer were deconvoluted similar to ENH2 protein. For the deconvoluted spectra, the relative abundance between unphosphorylated and phosphorylated β-casein (5*p*β-casein) were calculated based on the sum of the all relative abundance of the three isoforms (A2, A1 and B) of β-casein divided by the sum of all relative abundance of the three isoforms of 5*p*β-casein using the SNAP algorithm. For justification of only using three isoforms for quantification, an additional algorithm, Sum Peak which integrates the peak area within the mass range with an increment of 1 Da, was used for comparison for the Q-TOF deconvoluted spectra (Figure S2.2). Mass range between 23550 Da and 23710 Da for β-casein and between 23950 Da and 24110 Da for 5*p*β-casein was each integrated (Figure S2.2a). These two algorithms yielded similar relative abundance quantification results for the relative percentage of β-casein using three different MS spectra of the same mixture (61% from SNAP versus 58% from Sum Peak) (Figure S2.2b). Similarly, data acquired using the 12T solariX

FT-ICR mass spectrometer were analyzed using DataAnalysis and the relative abundance was derived from the SNAP algorithm results in DataAnalysis using only three genetic variants of β-casein. ECD MS/MS spectra from both the 7T LTQ/FT Ultra mass spectrometer and the 12T solariX mass spectrometer were analyzed using MASH Suite Pro.[29-30]

Data Analysis

The CV was used to evaluate the relative variability in different datasets. Both solution-phase and gas-phase ratios were computed by dividing the numerical values (gel band intensity or MS ion abundance) of unphosphorylated proteins by those of phosphorylated proteins (i.e. unphosphorylated:phosphorylated). To calculate the percentage difference between solution-phase ratio and gas-phase ratio, normalization was performed to the higher abundance proteoform for both SDS-PAGE and mass spectra for individual mixture. For example, for a 3:1 protein mixture with a gas-phase ratio of 2.72 and a solution-phase ratio of 3.16, the percentage difference is ~5% (e.g. after normalization, the relative percentage of phosphorylated protein is 36.7% in the gas phase and 31.6% in the solution phase). A least-square linear regression was performed between two datasets as indicated in each figure.

Solution-Phase Protein Concentration Analysis and the Relationship between Solution-phase Ratio vs. Gas-phase Ratio

To determine solution-phase protein concentration, each protein solution was first evaluated by a BCA. Gel-based standard curves were established using SDS-PAGE by loading lanes with increasing protein amounts using the concentration values obtained by the BCA. The solution-phase protein ratios (e.g., 1:1) were obtained by loading lanes with the appropriate protein amounts within the linear range of the gel-based standard curves (Figure S2.1), and the ratios were confirmed by comparing the gel band intensity between unphosphorylated and phosphorylated proteins using ImageJ (Figure 2.1 and Figure 2.2). The protein mixtures of each ratio composition were then subjected to gas-phase analysis by mass spectrometry.

We used direct infusion as a simple means to introduce the samples into the gas-phase for mass spectrometry analysis. We obtained the relative ion abundance of proteoforms using SNAP algorithm (methods mentioned in the Experimental Procedure above) in the deconvoluted mass spectra, which accounts for all charge states, or by examining individual charge states. The relative abundance of proteoforms represents the gas-phase ratios, which was first compared against their respective solution-phase ratio by percentage difference (methods mentioned in the Experimental Procedure above). Furthermore, a least-square linear regression was utilized to analyze the linear relationship between the solution-phase ratios and gas-phase ratios.

Results and Discussion

Preparation and Characterization of Mono- and Multiply-Phosphorylated Proteins

Our lab previously identified the ENH2 protein, a Z-disc protein belonging to the PDZ-LIM protein family, as a phosphoprotein, and localized the sole site of phosphorylation in the endogenous protein to Ser118 using top-down MS/MS analysis. For the purposes of this study, we expressed and purified a recombinant ENH2 protein with C-terminal FLAG- and polyhistidine-tags. Purification of the recombinant ENH2 protein was confirmed by SDS-PAGE analysis, which showed a single dark band at the expected molecular weight of the recombinant protein (~28 kDa) in the elution lane (Figure S2.3). Top-down MS/MS analysis was also employed to confirm the sequence of the recombinant protein (Figure S2.4). To phosphorylate this protein, recombinant ENH2 was incubated with PKA and complete phosphorylation of the protein was verified by top-down MS analysis (Figure S2.4a). ECD-MS/MS analysis confirmed phosphorylation at a single site, Ser119, in the recombinant protein, which corresponds to Ser118 in the endogenous swine ENH2 sequence (Figure S2.4a).[3]

For the assessment of the effect of multiple phosphate groups on the ionization/detection efficiency of intact proteins, we chose the commercially available protein β-casein, which has three isoforms (termed A1, A2, and B) that exist completely in the pentakis-phosphorylated state (Figure S2.4b).[26] ECD-MS/MS localized the sites of phosphorylation in the A2 isoform to Ser15, Ser17, Ser18, Ser19, and Ser35 (Figure S2.2b), which is in agreement with the results of previous studies.[31] To generate unphosphorylated β-casein for our analysis, the commercial protein was incubated with λPP. Complete dephosphorylation of the three β-casein isoforms was verified by top-down MS analysis (Figure S2.4b).

Impact of Mono-phosphorylation on Phosphorylation Quantification

To determine whether the presence of a single phosphate moiety impacts the ionization/detection efficiency of an intact protein, stock solutions of completely unphosphorylated or phosphorylated ENH2 were mixed 1:5, 1:3, 1:1, 3:1, and 5:1 (ENH2:*p*ENH2), and EHN2 and *p*ENH2 components of these mixtures were analyzed on separate lanes by SDS-PAGE to confirm the solution-phase ratios (Figure 2.1a). Subsequently, ESI-MS analysis of the aforementioned mixtures using a maXis II Q-TOF mass spectrometer was carried out. Our results showed that gas-phase ratios of ENH2:*p*ENH2 as determined from the deconvoluted mass spectra generally correspond to their respective solution-phase ratios derived from the SDS-gel data (i.e. < 6% difference between gas phase and solution phase. Further details concerning calculation and data analysis were provided in the Supporting Information). Additionally, linear regression analysis of the solution-phase and gas-phase ratios yielded an R^2 value of 0.995, indicating good correlation between the solution-phase and gas-phase ratios when quantification of the gas-phase ENH2:*p*ENH2 ratio was determined based on the ratios in the deconvoluted mass spectra (Figure 2.1b-c). Moreover, the effect of different ionization parameters such as variations in the spray voltage, isCID voltage and solvent composition also did not affect the observed gas-phase ratio of ENH2:*p*ENH2 mixtures (Figure S2.5 and S2.6). For instance, the change in electrospray voltage from 5000 V to 3000 V has minimal impact on the phosphoprotein quantification, despite the reduction of the overall ions generated (Figure S2.5a). Additionally, phosphoprotein quantification was not affected by the variations in the desolvation processes such as changes in the nebulizer gas pressure from 0.5 to 1.5 bar (Figure S2.5c).

Nevertheless, as the relative intensities of unphosphorylated and phosphorylated ENH2 proteoforms in the deconvoluted mass spectra represent an average of the relative abundance ratios for these species across all charge states in the 500 – 3000 *m/z* range, there was the potential that the gas-phase ratios for ENH2:*p*ENH2 at individual charge states may not correlate well with the solution-phase ratios. To investigate this possibility, we also evaluated the ENH2:*p*ENH2 gas-phase ratio for three individual charge states (41^+, 40^+, and 39^+) from the raw mass spectra (Figure S2.7). The gas-phase ratio across the three selected charge states differed from the solution-phase ratio by less than 3% (Figure S2.7), which confirms that the good degree of correspondence between the solution-phase and gas-phase ratios for ENH2:*p*ENH2 was not an artifact of deconvolution. Moreover, similar results were obtained when the same mixtures were analyzed using a 12T solariX FT-ICR mass spectrometer at the most abundant charge states, 41^+ (Figure S2.8), 40^+, and 39^+ (Figure S2.9), indicating that the good correlation observed between the solution-phase and gas-phase ratios for ENH2 and *p*ENH2 was not dependent on the mass analyzer employed.

Furthermore, we have analyzed all charge states (44^+ to 25^+) and the results are summarized in Table S2.1. The data suggested that the gas-phase ratio derived based on the most abundant charge states are within 5% difference from the solution-phase ratio, which is also similar to the gas-phase ratio derived from the deconvoluted spectrum that takes into consideration of all charge states. In contrast, the gas-phase ratio derived from the highest or lowest charge states deviate significantly from the solution-phase ratio. For the spectrum obtained from 1:1 solution-phase ratio, the gas-phase ratios of the most abundant charge states are 41^+, 40^+, and 39^+ are 1.14, 1.09, and 1.09 (Table S2.1), respectively, compared to the solution-phase ratio of 0.93. In contrast, the gas-phase ratios at extreme charge states (i.e. highest or lowest) charge states at 44^+ and 25^+ are 1.37 and 0.68 (Table S2.1), respectively, which is significantly deviated from the solution-phase ratio. The gas-phase ratio obtained from deconvoluted spectrum which accounts for all charge states is 0.93, which is in close agreement with the solution-phase ratio. Collectively, these data strongly support the long-held belief in the field that the presence of a single phosphate group has a negligible impact on the ionization/detection efficiency of an intact protein.

Impact of Multiple phosphorylation on Phosphorylation Quantification

We next sought to determine the impact of multi-phosphorylation on MS quantification of phosphoproteins. Similar to the analysis for ENH2, stock solutions of completely unphosphorylated or phosphorylated β-casein were mixed 1:5, 1:3, 1:1, 3:1, and 5:1 (β-casein:5*p*β-casein), and the solution-phase ratios in the mixtures were confirmed by SDS-PAGE (Figure 2.2a). ESI-MS analysis of the above mentioned mixtures were analyzed by a maXis II Q-TOF mass spectrometer. Relative quantification of β-casein was performed using the three isoforms (A2, A1, and B), and justification was detailed in the Experimental Section (Figure S2.2). A good linear correlation between the solution-phase ratios and the gas-phase ratios was deduced from the linear regression analysis with an R^2 value of 0.999 based on the relative quantification using the deconvoluted mass spectra (Figure 2.2b-c). However, despite the good linearity between these two ratios, the gas-phase ratios for these mixtures differed from the solution-phase ratios by more than 10%. This result suggests a likelihood that the addition of five phosphate groups gives rise to differential ionization/detection efficiency between unphosphorylated and phosphorylated β-casein. Additionally, the mixtures were analyzed using a 12T solariX FT-ICR mass spectrometer at high abundance charge states, 24^+, 23^+, and 22^+ (Figure S2.10). The linear regression using a linear equation afforded an R^2 value of 0.961, whereas the gas-phase ratios for these mixtures again had more than 10% difference compared to their respective solution-phase ratios. Altogether, these results indicate that the presence of multiple phosphate groups on the intact protein may have an impact on the ionization/detection efficiency of intact proteins.

The influence of multi-phosphorylations on gas-phase ratios in MS becomes apparent when we examined the overall charge state distribution profile of the mono-phosphorylated ENH2 and the multiply-phosphorylated β-casein mixtures in 1:1 solution-phase ratio with their corresponding unphosphorylated counterparts (Figure 2.3a-b). While the mono-phosphorylated ENH2 ions have similar intensities to their unphosphorylated counterparts, the multiply-phosphorylated β-casein showed much lower ion intensities. To further investigate this discrepancy caused by multi-phosphorylation, a mixture of the unphosphorylated and the multiply-phosphorylated β-casein that led to 1:1 intensity ratio from the deconvoluted spectra was prepared (Figure S2.11). For these three mixtures (1:1 solution-phase ratio of ENH2:*p*ENH2; 1:1 solution-phase ratio of β-casein:5*p*β-casein; and 1:1 gas-phase ratio of β-casein:5*p*β-casein based on the deconvoluted spectra), gas-phase ratios of individual charge states from 44^+ to 25^+ and from 31^+ to 16^+ were analyzed for ENH2 proteoforms and β-casein proteoforms, respectively. While the CVs of all these charge states for 1:1 solution-phase ratio of both ENH2:pENH2 and β-casein:5*p*β-casein have a value of 0.24 and 0.30, respectively, the CV of 1:1 gas-phase ratio of β-casein:5*p*β-casein based on deconvoluted spectra has a greater value of 0.41 (Table S2.1 and S2.2). This result indicates that the variations of gas-phase ratios of individual charge state significantly varied even though the deconvoluted spectra implied an equal 1:1 gas-phase ratio. The high CV (0.41) in the case of 1:1 gas-phase ratio of β-casein:5*p*β-casein based on deconvoluted spectra arises from the fact that the charge state distribution of 5*p*β-casein was shifted to higher *m/z* value (Figure S2.11), inferring that the ionization/detection efficiency between β-casein and 5*p*β-casein may be affected by the negatively charged phosphate groups. Collectively, the difference in ionization/detection efficiency owing to multiple phosphate modifying groups provides a possible explanation for the discrepancy between gas-phase ratios and solution-phase ratios for β-casein analysis.

Importance of Accurate Relative Quantification of Phosphorylation Levels of Proteins

Relative quantification of phosphorylation levels of proteins provides important information which can be used to correlate with cellular processes and disease pathophysiology. Recent studies have correlated phosphorylation levels of proteins with alteration in cardiac and muscle functions.[4,7] Top-down MS is especially attractive for relative quantification of protein PTMs because it is believed that these modifications will have negligible impact on the ionization/detection efficiency of intact proteins.[17, 21] However, such assumption has not been vigorously validated. Previously, Steen *et al.* reported that the ionization/detection efficiency of phosphopeptides and their unphosphorylated cognates vary drastically.[15] They have demonstrated that more phosphopeptides show better ionization/detection efficiencies than their unphosphorylated cognates.[15] In this study, we have shown that mono-phosphorylation has minimal impact on the ionization/detection efficiency of the intact proteins as demonstrated by ESI-MS analysis of recombinant ENH2 with mono-phosphorylation using both a TOF and an FT-ICR mass spectrometers (Figure 2.1, Figure S2.5-2.9, and Table S2.1). Therefore, relative quantification using top-down proteomics analysis can be an accurate and powerful method for the relative quantification of mono-phosphorylated proteins.

In addition to mono-phosphorylation, multi-phosphorylation is also observed in biological processes.[32-33] Previous study from Medina *et al.* suggested that multiple phosphate modifying groups changed the physicochemical properties of proteins such as electrophoretic mobility and side chain flexibility of caseins.[23] Therefore, we prepared and characterized a multiply-phosphorylated protein model using β-casein, and observed discrepancy between gas-phase ratios and solution-phase ratios of β-casein:5*p*β-casein (Figure 2.2). Conceivably, our result showed that multi-phosphorylation significantly altered the electrophoretic mobility between β-casein and 5*p*β-casein (Figure 2.2a), in agreement with the previous finding from Medina *et al.*[23] Further investigation into the shift in charge state distribution between β-casein and 5*p*β-casein suggests a change in the ionization/detection efficiency resulted from possible alteration in physicochemical properties likely due to the multiple negative charges imparted by five phosphate modifying groups, and thus impacts relative quantification (Figure 2.3).

Regarding other PTMs beyond phosphorylation, previously, the Kelleher group attained quantitative information about the isomeric composition of intact histone H4 protein by monitoring the mono-, di-, tri-, and tetra-acetylation.[21] Multiple histone acetylation modifications do not appear to affect the ionization/detection efficiency of the histone H4 protein, in contrast to our results of multi-phosphorylation quantification.

Conclusion

To recapitulate, we conducted a systematic interrogation on the top-down ESI-MS-based relative quantification of phosphoproteins using a mono-phosphorylated protein model (ENH2) and a multiply-phosphorylated protein model (β-casein). Our results showed that the mono-phosphorylation does not appreciably affect ESI-MS quantification of phosphoproteins. In contrast to mono-phosphorylation, pentakis-phosphorylation noticeably influenced ESI-MS quantification of phosphoproteins, possibly due to the differential ionization/detection efficiency resulted from slightly different physicochemical properties between unphosphorylated and pentakis-phosphorylated proteoforms.

Acknowledgement

This work was supported by the NIH R01 GM117058 (to S.J. and Y.G.), and NIH F31 HL128086 (to Z.R.G.). YG would also like to acknowledge support by NIH R01 HL109810, HL096971 and S10 OD018475.

References

1. Hunter, T., Protein kinases and phosphatases: The Yin and Yang of protein phosphorylation and signaling. *Cell* **1995,** *80* (2), 225-236.

2. 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.

3. 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.; Liu, X.; Ge, Y., Top-down proteomics reveals concerted reductions in myofilament and Z-disc protein phosphorylation after acute myocardial infarction. *Mol. Cell. Proteomics* **2014,** *13* (10), 2752-64.

4. Gregorich, Z. R.; Peng, Y.; Cai, W.; Jin, Y.; Wei, L.; Chen, A. J.; McKiernan, S. H.; Aiken, J. M.; Moss, R. L.; Diffee, G. M.; Ge, Y., Top-Down Targeted Proteomics Reveals Decrease in Myosin Regulatory Light-Chain Phosphorylation That Contributes to Sarcopenic Muscle Dysfunction. *J. Proteome Res.* **2016,** *15* (8), 2706-2716.

5. Li, Y.-Y.; Popivanova, B. K.; Nagai, Y.; Ishikura, H.; Fujii, C.; Mukaida, N., Pim-3, a Proto-Oncogene with Serine/Threonine Kinase Activity, Is Aberrantly Expressed in Human Pancreatic Cancer and Phosphorylates Bad to Block Bad-Mediated Apoptosis in Human Pancreatic Cancer Cell Lines. *Cancer Res.* **2006,** *66* (13), 6741-6747.

6. Hanger, D. P.; Anderton, B. H.; Noble, W., Tau phosphorylation: the therapeutic challenge for neurodegenerative disease. *Trends Mol. Med.* **2009,** *15* (3), 112-119.

7. Zhang, J.; Guy, M. J.; Norman, H. S.; Chen, Y.-C.; Xu, Q.; Dong, X.; Guner, H.; Wang, S.; Kohmoto, T.; Young, K. H.; Moss, R. L.; Ge, Y., Top-Down Quantitative Proteomics Identified Phosphorylation of Cardiac Troponin I as a Candidate Biomarker for Chronic Heart Failure. *J. Proteome Res.* **2011,** *10* (9), 4054-4065.

8. Kosako, H.; Nagano, K., Quantitative phosphoproteomics strategies for understanding protein kinase-mediated signal transduction pathways. *Expert Rev. Proteomics* **2011,** *8* (1), 81-94.

9. Chan, C. Y. X. a.; Gritsenko, M. A.; Smith, R. D.; Qian, W.-J., The current state of the art of quantitative phosphoproteomics and its applications to diabetes research. *Expert Rev. Proteomics* **2016,** *13* (4), 421-433.

10. Taus, T.; Köcher, T.; Pichler, P.; Paschke, C.; Schmidt, A.; Henrich, C.; Mechtler, K., Universal and Confident Phosphorylation Site Localization Using phosphoRS. *J. Proteome Res.* **2011,** *10* (12), 5354-5362.

11. Ong, S.-E.; Mann, M., Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **2005,** *1* (5), 252-262.

12. Steen, H.; Jebanathirajah, J. A.; Springer, M.; Kirschner, M. W., Stable isotope-free relative and absolute quantitation of protein phosphorylation stoichiometry by MS. *Proc. Natl. Acad. Sci. U S A* **2005,** *102* (11), 3948-3953.

13. Wu, R.; Haas, W.; Dephoure, N.; Huttlin, E. L.; Zhai, B.; Sowa, M. E.; Gygi, S. P., A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nat. Method* **2011,** *8*, 677.

14. Riley, N. M.; Coon, J. J., Phosphoproteomics in the Age of Rapid and Deep Proteome Profiling. *Anal. Chem.* **2016,** *88* (1), 74-94.

15. Steen, H.; Jebanathirajah, J. A.; Rush, J.; Morrice, N.; Kirschner, M. W., Phosphorylation Analysis by Mass Spectrometry: Myths, Facts, and the Consequences for Qualitative and Quantitative Measurements. *Mol. Cel. Proteomics* **2006,** *5* (1), 172-181.

16. Smith, L. M.; Kelleher, N. L.; The Consortium for Top Down, P., Proteoform: a single term describing protein complexity. *Nat. Method* **2013,** *10* (3), 186-187.

17. Cai, W.; Tucholski, T. M.; Gregorich, Z. R.; Ge, Y., Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. *Expert Rev. Proteomics* **2016,** *13* (8), 717-730.

18. Toby, T. K.; Fornelli, L.; Kelleher, N. L., Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem. (Palo Alto Calif)* **2016,** *9* (1), 499-519.

19. Jiang, L.; Smith, J. N.; Anderson, S. L.; Ma, P.; Mizzen, C. A.; Kelleher, N. L., Global Assessment of Combinatorial Post-translational Modification of Core Histones in Yeast Using Contemporary Mass Spectrometry: LYS4 trimethylation correlates with degree of acetylation on the same H3 tail. *J. Biol. Chem.* **2007,** *282* (38), 27923-27934.

20. Chamot-Rooke, J.; Mikaty, G.; Malosse, C.; Soyer, M.; Dumont, A.; Gault, J.; Imhaus, A.-F.; Martin, P.; Trellet, M.; Clary, G.; Chafey, P.; Camoin, L.; Nilges, M.; Nassif, X.; Duménil, G., Posttranslational Modification of Pili upon Cell Contact Triggers N. meningitidis Dissemination. *Science* **2011,** *331* (6018), 778-782.

21. Pesavento, J. J.; Mizzen, C. A.; Kelleher, N. L., Quantitative Analysis of Modified Proteins and Their Positional Isomers by Tandem Mass Spectrometry: Human Histone H4. *Anal. Chem.* **2006,** *78* (13), 4271-4280.

22. Polyansky, A. A.; Zagrovic, B., Protein Electrostatic Properties Predefining the Level of Surface Hydrophobicity Change upon Phosphorylation. *J. Phys. Chem. Lett.* **2012,** *3* (8), 973-976.

23. Medina, A. L.; Colas, B.; Le Meste, M.; Renaudet, I.; Lorient, D., Physicochemical and Dynamic Properties of Caseins Modified by Chemical Phosphorylation. *J. Food Sci.* **1992,** *57* (3), 617-621.

24. Schneider, C. A.; Rasband, W. S.; Eliceiri, K. W., NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **2012,** *9* (7), 671-675.

25. Chen, Y.-C.; Ayaz-Guner, S.; Peng, Y.; Lane, N. M.; Locher, M. R.; Kohmoto, T.; Larsson, L.; Moss, R. L.; Ge, Y., Effective Top-Down LC/MS+ Method for Assessing Actin Isoforms as a Potential Cardiac Disease Marker. *Anal. Chem.* **2015,** *87* (16), 8399-8406.

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

27. Peng, Y.; Chen, X.; Zhang, H.; Xu, Q.; Hacker, T. A.; Ge, Y., Top-down Targeted Proteomics for Deep Sequencing of Tropomyosin Isoforms. *J. Proteome Res.* **2013,** *12* (1), 187-198.

28. Peng, Y.; Yu, D.; Gregorich, Z.; Chen, X.; Beyer, A. M.; Gutterman, D. D.; Ge, Y., In-depth proteomic analysis of human tropomyosin by top-down mass spectrometry. *J. Muscle Res. Cell*

Motil. **2013,** *34* (3), 199-210.

29. Cai, W.; Guner, H.; Gregorich, Z. R.; Chen, A. J.; Ayaz-Guner, S.; Peng, Y.; Valeja, S. G.; Liu, X.; Ge, Y., MASH Suite Pro: A Comprehensive Software Tool for Top-Down Proteomics. *Mol. Cel. Proteomics* **2016,** *15* (2), 703-714.

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

31. Wu, S.; Lourette, N. M.; Tolić, N.; Zhao, R.; Robinson, E. W.; Tolmachev, A. V.; Smith, R. D.; Paša-Tolić, L., An integrated top-down and bottom-up strategy for broadly characterizing protein isoforms and modifications. *J. Proteome Res.* **2009,** *8* (3), 1347-1357.

32. Ali, A. A. E.; Jukes, R. M.; Pearl, L. H.; Oliver, A. W., Specific recognition of a multiply phosphorylated motif in the DNA repair scaffold XRCC1 by the FHA domain of human PNK. *Nucleic Acids Res.* **2009,** *37* (5), 1701-1712.

33. Wang, Y.; Guan, S.; Acharya, P.; Liu, Y.; Thirumaran, R. K.; Brandman, R.; Schuetz, E. G.; Burlingame, A. L.; Correia, M. A., Multisite phosphorylation of human liver cytochrome P450 3A4 enhances Its gp78- and CHIP-mediated ubiquitination: a pivotal role of its Ser-478 residue in the gp78-catalyzed reaction. *Mol. Cell. Proteomics* **2012,** *11* (2), M111 010132.

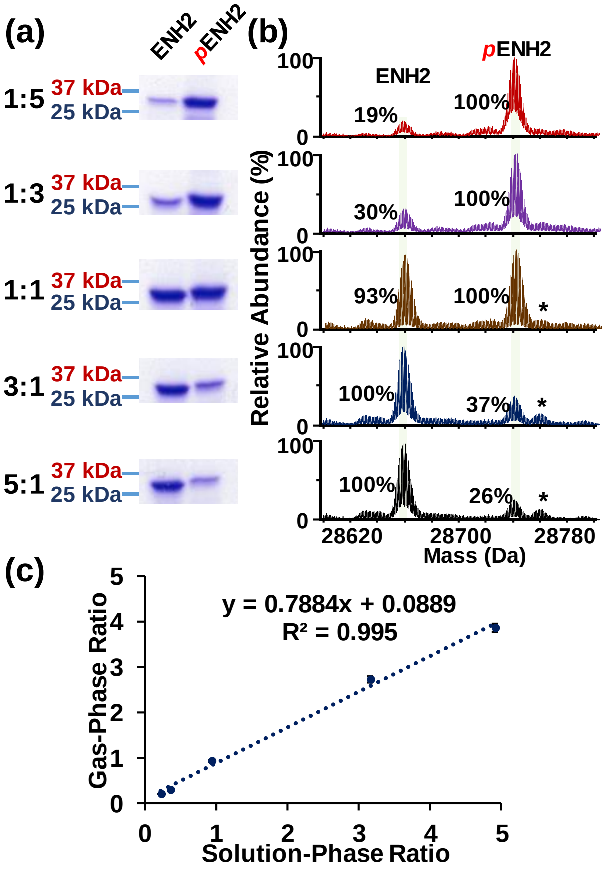


Figure 2.1. Protein quantification using SDS-PAGE gel analysis and maXis II Q-TOF MS analysis for ENH2. (a) SDS-PAGE analysis of ENH2:*p*ENH2 in five different ratios, 1:5, 1:3, 1:1, 3:1 to 5:1 (top to bottom) and (b) the corresponding Q-TOF MS deconvoluted spectra. Relative abundance is normalized to the most abundant species in each mass spectrum. (c) Correlation analysis between gas-phase ratios (derived from the Q-TOF MS data) and solution-phase ratios (derived from SDS-gel data) of ENH2:*p*ENH2 suggests a linear correspondence between these two methods. *ENH2 with non-covalent phosphate adduct (+98 Da).

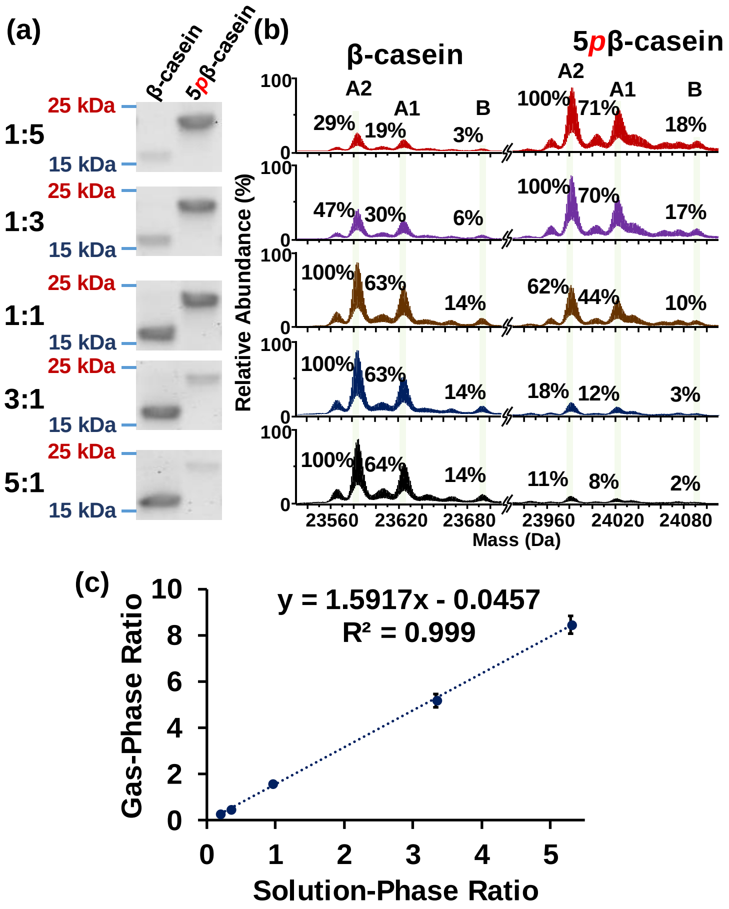


Figure 2.2. Protein quantification using SDS-PAGE gel analysis and maXis II Q-TOF MS analysis for β-casein. (a) SDS-PAGE analysis of five different β-casein:5*p*β-casein ratios from 1:5, 1:3, 1:1, 3:1 to 5:1 (top to bottom); (b) corresponding Q-TOF MS spectra (deconvoluted). In each deconvoluted spectrum, the relative abundance is normalized to the highest abundance species; (c) Correlation analysis between gas-phase ratios (derived from Q-TOF MS data) and solution-phase ratios (derived from SDS-gel data) of β-casein:5*p*β-casein suggests a linear correspondence between these two ratios.

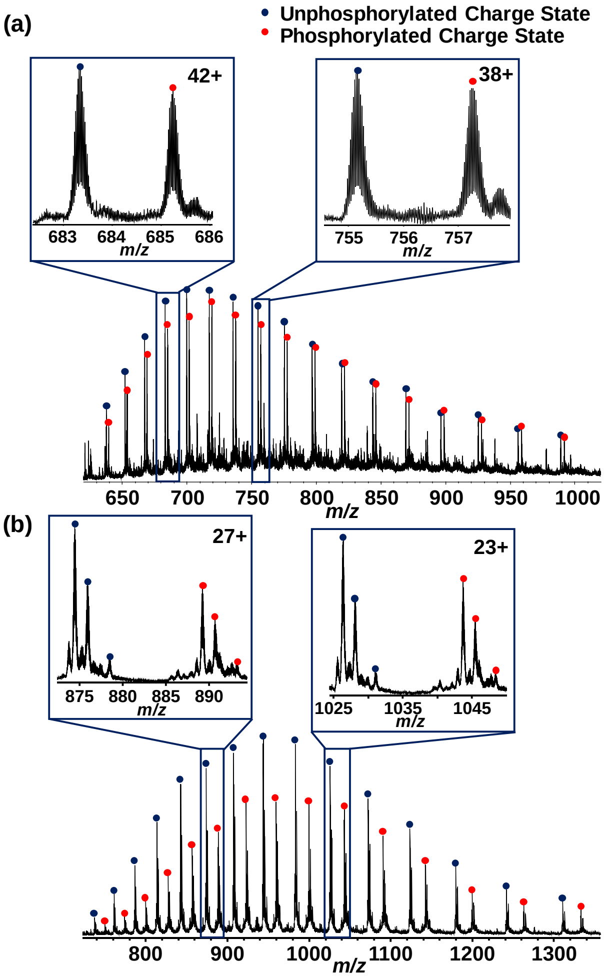


Figure 2.3. Comparison between mono-phosphorylation and multi-phosphorylation on protein quantification. Charge state distributions obtained from ESI/Q-TOF MS of 1:1 solution-phase ratio mixture of (a) ENH2:*p*ENH2 and (b) β-casein:5*p*β-casein were shown. Compared to mono-phosphorylation, multi-phosphorylation affects the phosphoprotein quantification. Insets, representative higher and lower charge state of ENH2:*p*ENH2 and β-casein:5*p*β-casein, respectively.

Supplemental Information

Table S2.1. Quantification for the equal amount of ENH2:*p*ENH2 (1:1 at all observed charge states) using Q-TOF MS.

Charge State	ENH2:*p*ENH2 gas-phase ratio
44^+	1.37
43^+	1.32
42^+	1.16
41^+	1.14
40^+	1.09
39^+	1.09
38^+	1.24
37^+	1.07
36^+	1.02
35^+	1.61
34^+	1.15
33^+	1.39
32^+	0.79
31^+	1.88
30^+	1.32
29^+	1.16
28^+	1.20
27^+	0.82
26^+	0.84
25^+	0.68
Average	1.17
Standard Deviation	0.28
Coefficient of Variation	0.24
Deconvolution*	0.93

The gas-phase ratios of ENH2:*p*ENH2 at charge state 44^+ to 25^+ were shown. The average, standard deviation and coefficient of variation of the gas-phase ratios from charge state 44^+ to 25^+ were derived. 40^+ is the most abundant charge state. *The gas-phase ratio based on deconvoluted spectrum is 0.93 as described previously in Figure 1.

Table S2.2. Quantification for 1:1 gas-phase ratio of β-casein:5*p*β-casein and 1:1 solution-phase ratio of β-casein:5*p*β-casein at all observed charge states using Q-TOF MS.

1:1 Gas-Phase Ratio of β-casein:5*p*β-casein*		1:1 Solution-Phase Ratio β-casein:5*p*β-casein	
Charge State	β-casein:5*p*β-casein gas-phase ratio	Charge State	β-casein:5*p*β-casein gas-phase ratio
31^+	1.53	31^+	2.54
30^+	1.92	30^+	1.90
29^+	1.72	29^+	1.50
28^+	1.49	28^+	2.16
27^+	1.63	27^+	1.39
26^+	1.00	26^+	1.40
25^+	0.93	25^+	1.35
24^+	0.66	24^+	0.72
23^+	0.76	23^+	0.81
22^+	0.75	22^+	1.71
21^+	1.02	21^+	1.43
20^+	0.78	20^+	1.65
19^+	1.13	19^+	1.58
18^+	0.88	18^+	1.45
17^+	0.68	17^+	1.25
16^+	0.44	16^+	1.26
Average	1.08	Average	1.51
Standard Deviation	0.44	Standard Deviation	0.45
Coefficient of Variation	0.41	Coefficient of Variation	0.30

The gas-phase ratios of β-casein:5*p*β-casein at charge state 31^+ to 16^+ were shown in both cases. The average, standard deviation and coefficient of variation of the gas-phase ratios from charge state 31^+ to 16^+ were derived. 25^+ is the most abundant charge state. *Gas-phase ratio derived based on the deconvoluted spectrum taking into account of all charge states.

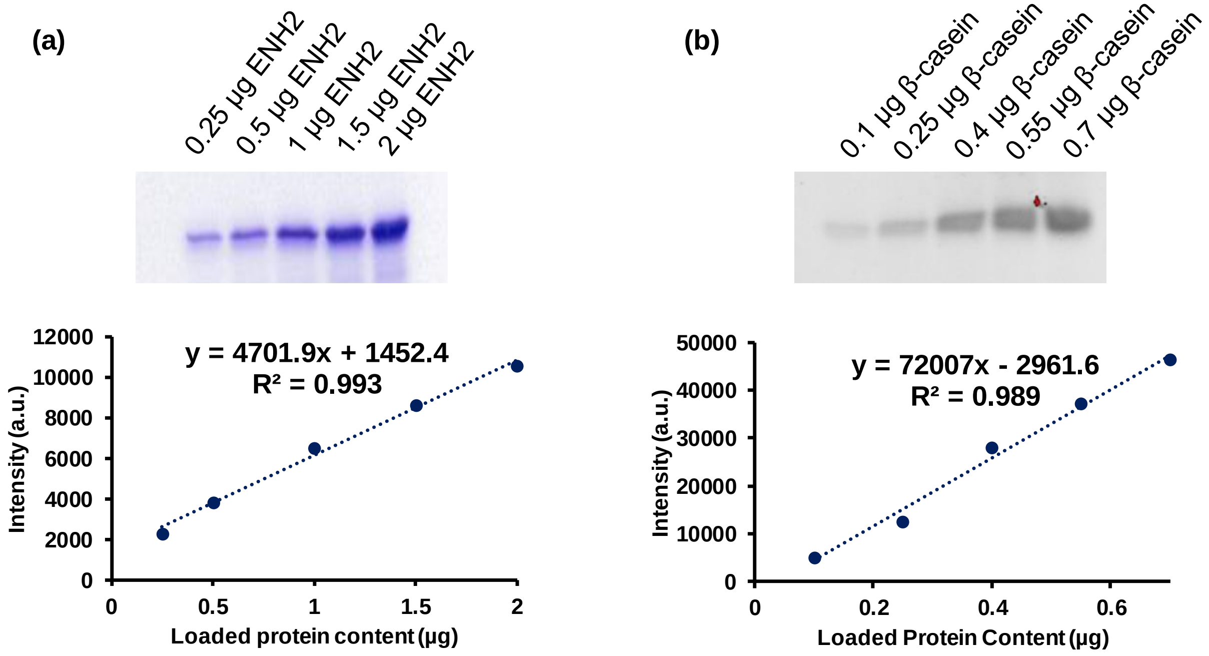


Figure S2.1. Gel-Band Quantification for ENH2 and β-casein. The band intensity from ImageJ can be correlated to loaded protein content using the gel-based standard curve. (a) ENH2 and (b) β-casein showed a respective linear range of 0.25 – 2 µg and 0.1 – 0.7 µg of protein amount.

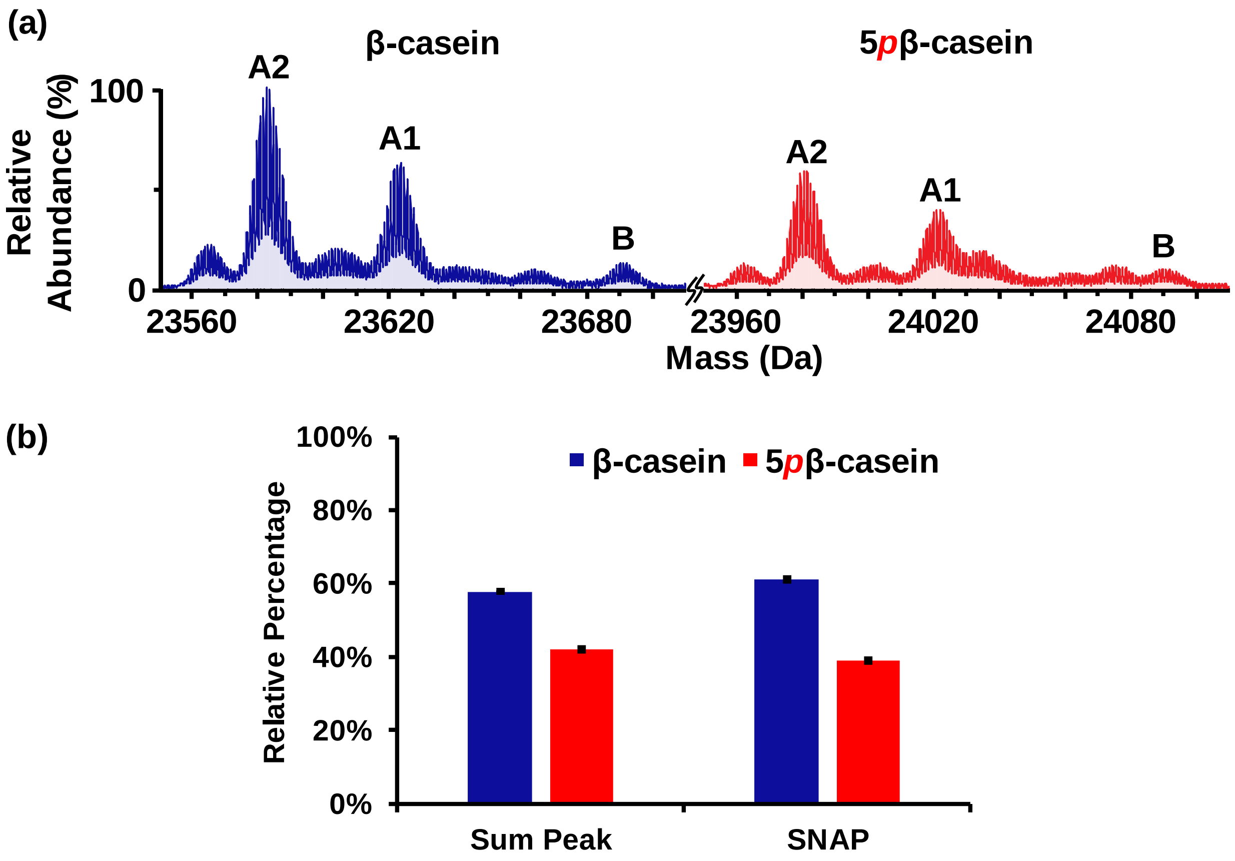


Figure S2.2. Comparison between SNAP and Sum Peak quantification method. (a) Mass range from 23550 to 23710 Da for β-casein and that from 23950 to 24110 Da for 5*p*β-casein were used for Sum Peak algorithm to calculate gas-phase ratio of β-casein:5*p*β-casein. A2, A1 and B isoforms of both β-casein and 5pβ-casein were used for SNAP algorithm to calculate Q-TOF MS ion ratio of β-casein:5*p*β-casein. (b) Relative percentage of β-casein and 5*p*β-casein was comparable using either SNAP algorithm or Sum Peak algorithm in the DataAnalysis software. Specifically, relative percentage of β-casein using Sum Peak algorithm and SNAP algorithm was 58% and 61%, respectively.

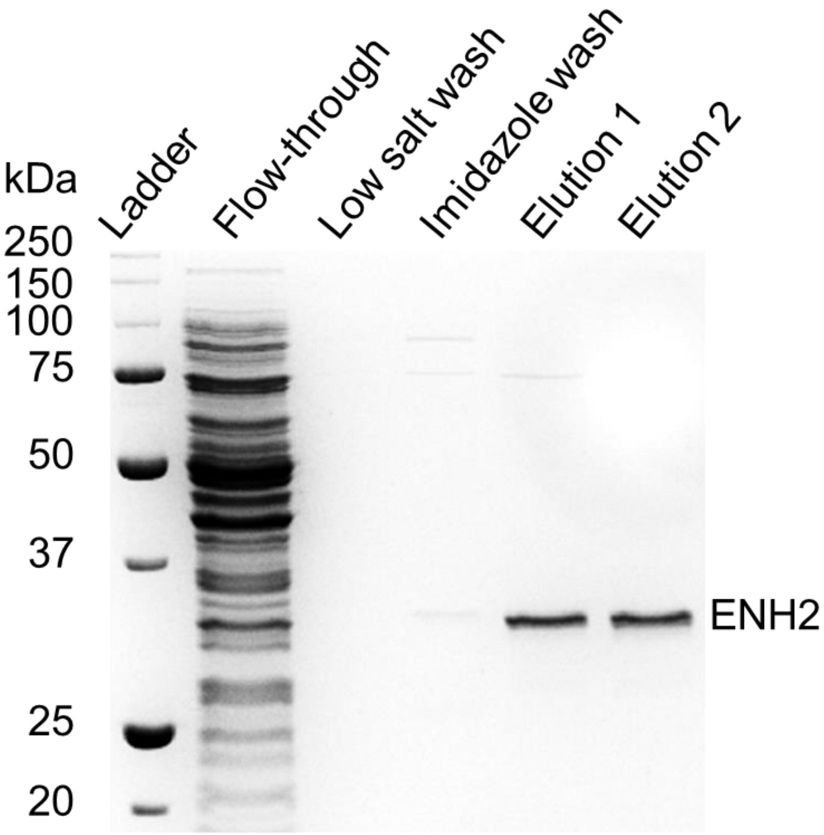


Figure S2.3. Protein purification. SDS-PAGE analysis indicating that ENH2 was purified with high purity.

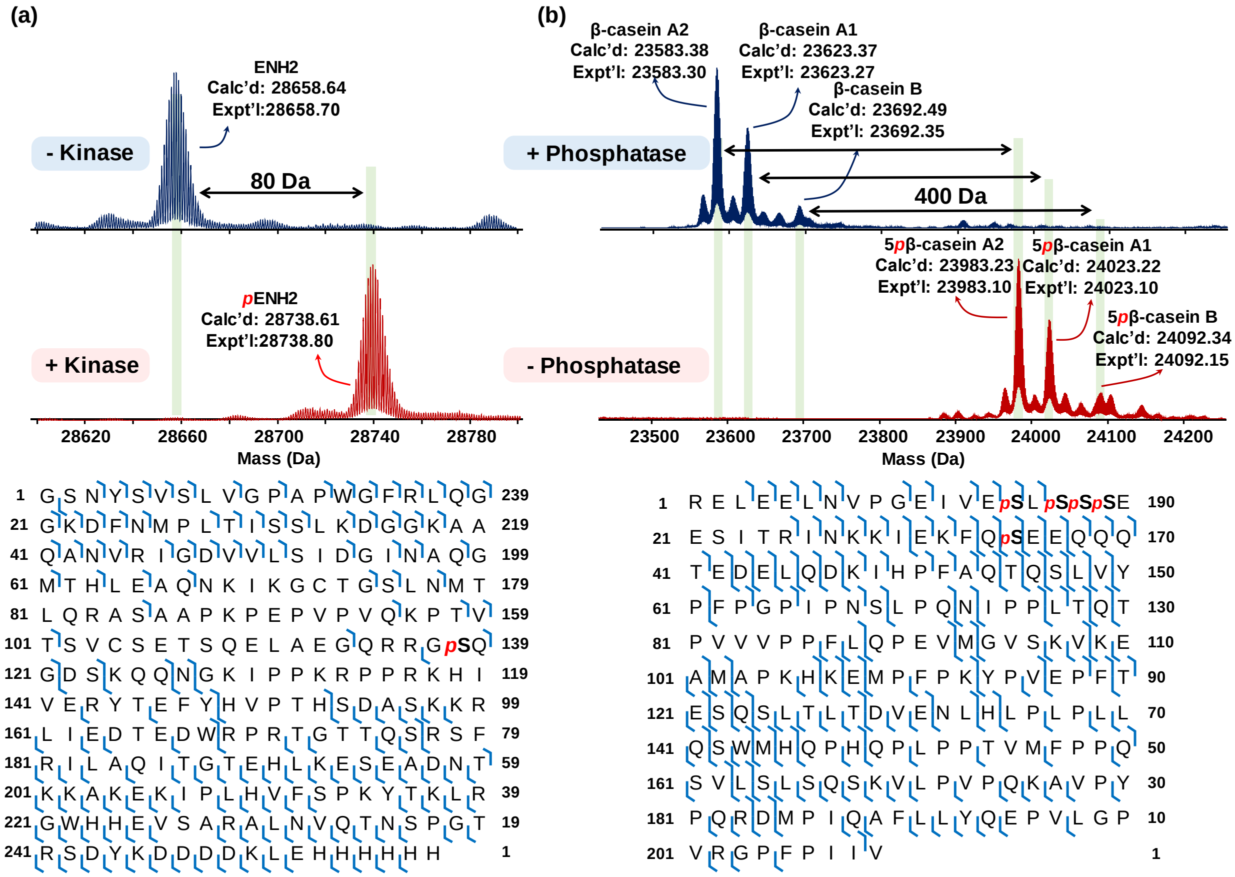


Figure S2.4. Preparation and characterization of mono-phosphorylated protein, ENH2, and multiply-phosphorylated protein, β-casein, for assessing the impact of phosphorylation on ESI-MS quantification. (a) Complete phosphorylation was achieved by incubating ENH2 in the presence of PKA. Top-down ECD experiment localized the site of phosphorylation in the recombinant ENH2 to Ser119, which corresponds to Ser118 in the endogenous protein. (b) Complete dephosphorylation of β-casein was achieved by incubation with λPP, indicated by a mass shift of 400 Da. Top-down ECD experiment confirmed the phosphorylation sites are Ser15, Ser17, Ser18, Ser19, and Ser35 in the A2 isoform β-casein. The phosphorylation sites for A1 and B isoform are identical to the A2-isoform.

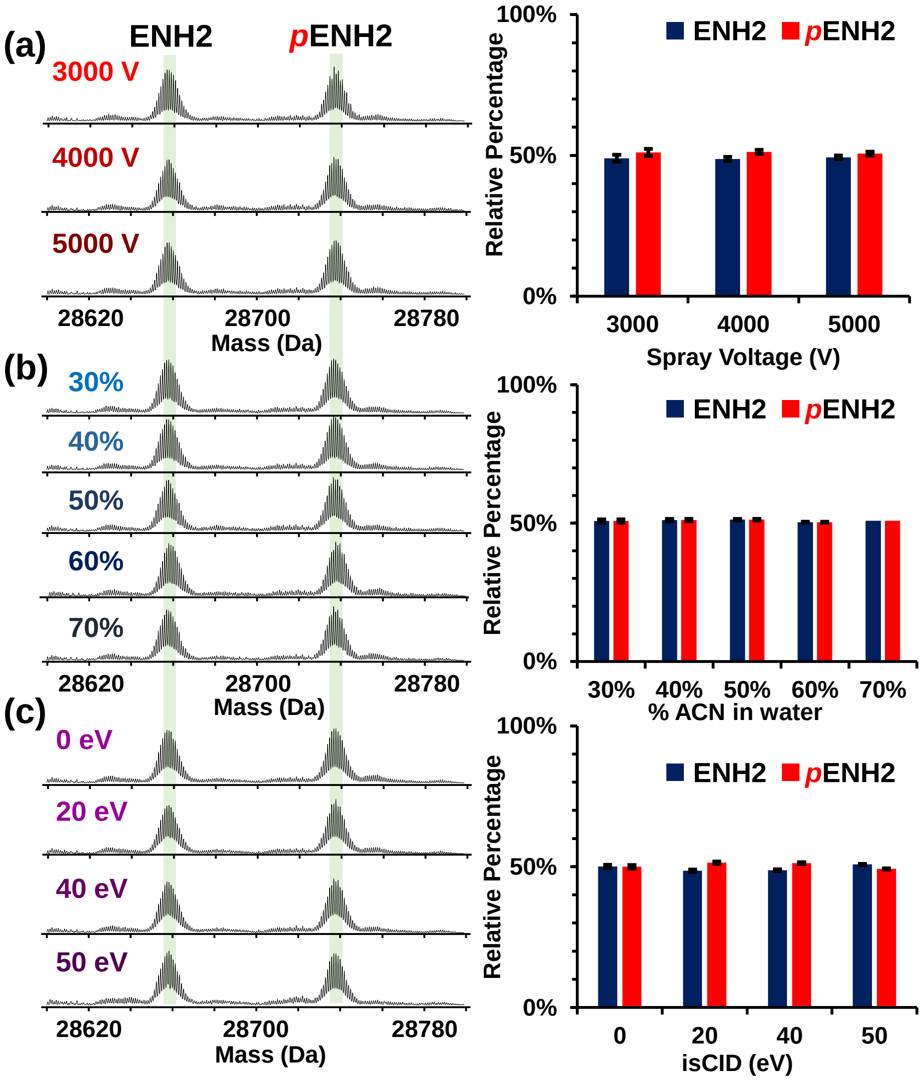


Figure S2.5. Evaluation of the impact of different ionization parameter setting on phosphoprotein quantification. Deconvoluted mass spectra and relative percentage at different conditions including (a) spray voltages, (b) solvent compositions, and (c) isCID voltages, suggests changes in ESI parameter setting do not affect the relative proteoform percentage of the 1:1 ENH2:*p*ENH2 mixture. Data were collected on a Bruker maXis II Q-TOF mass spectrometer.

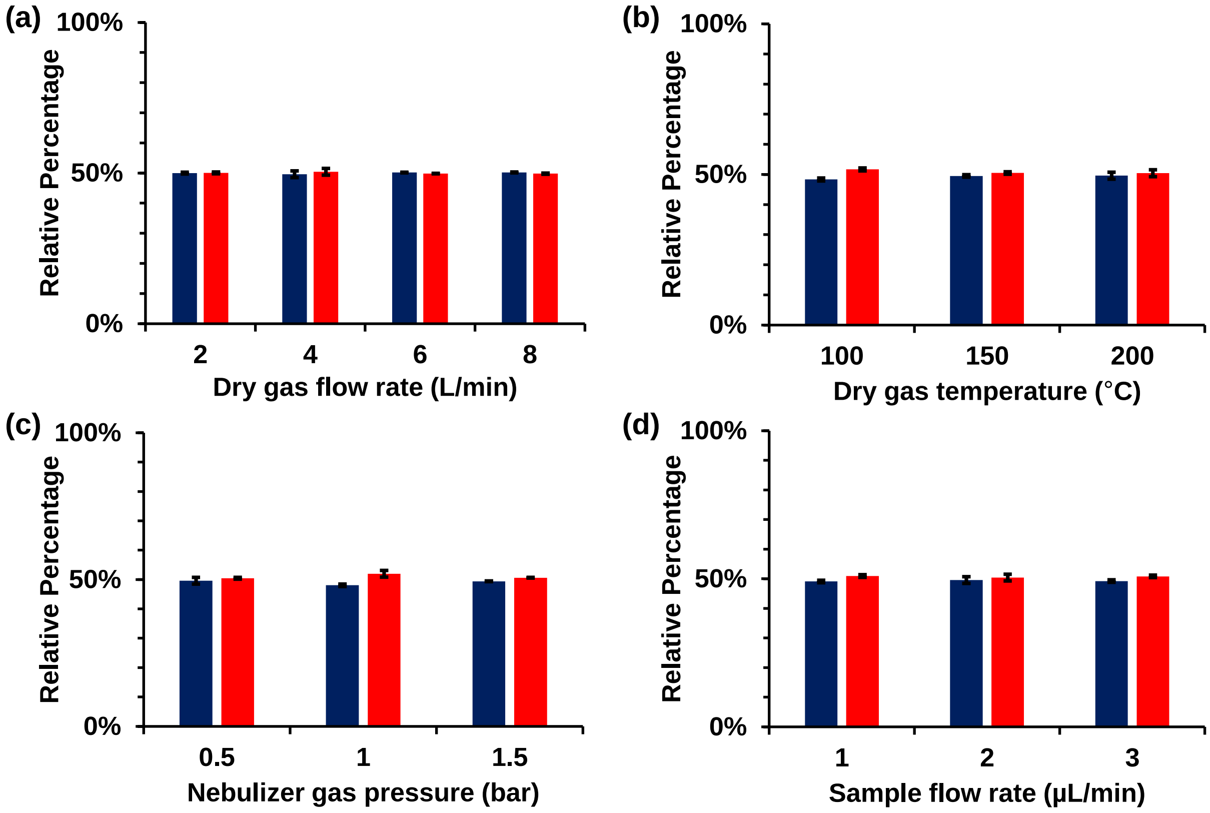


Figure S2.6. Evaluation of changes in desolvation parameters and sample flow rate on the proteoform quantification. Different conditions including (a) dry gas flow rate, (b) dry gas temperature, (c) nebulizer gas pressure, and (d) sample flow rate do not significantly vary the relative proteoform percentage. Data were collected on a Bruker maXis II Q-TOF MS.

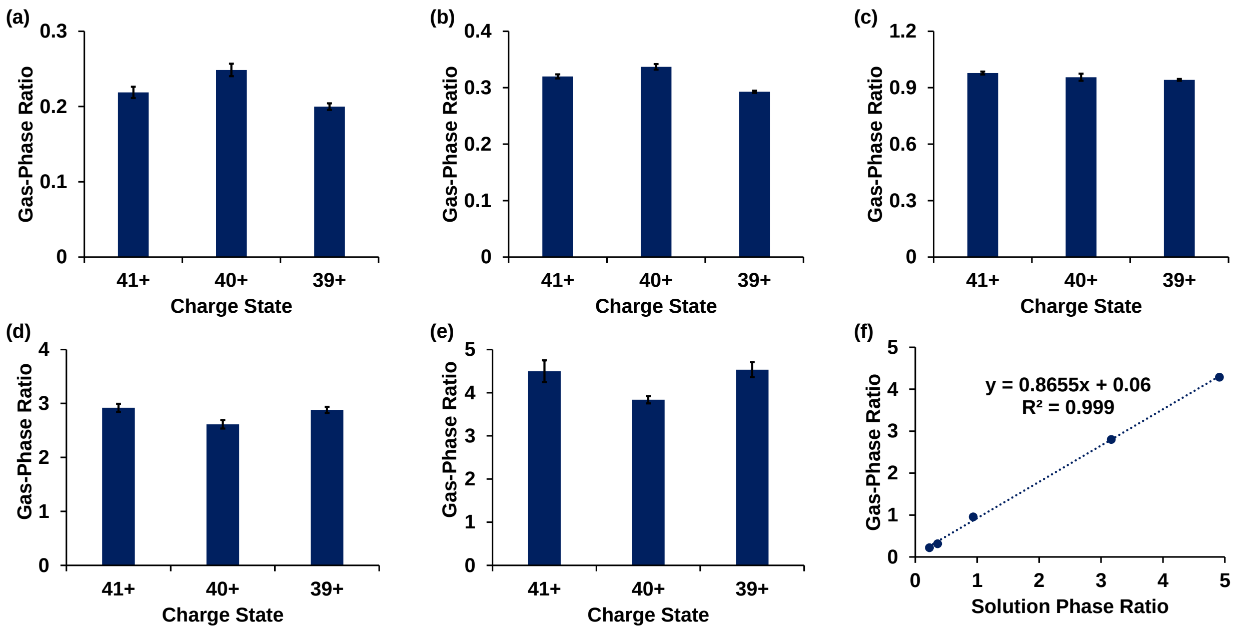


Figure 2.7. Quantification for ENH2 proteoforms at different charge states using Q-TOF mass spectrometer. Gas-phase ratios of ENH2 and pENH2 was compared with solution-phase ratios of (a) 1:5, (b) 1:3, (c) 1:1, (d) 3:1, (e) 5:1. (f) The correlation between the average of gas-phase ratios among three charge states and solution-phase ratios suggests a good linear correspondence.

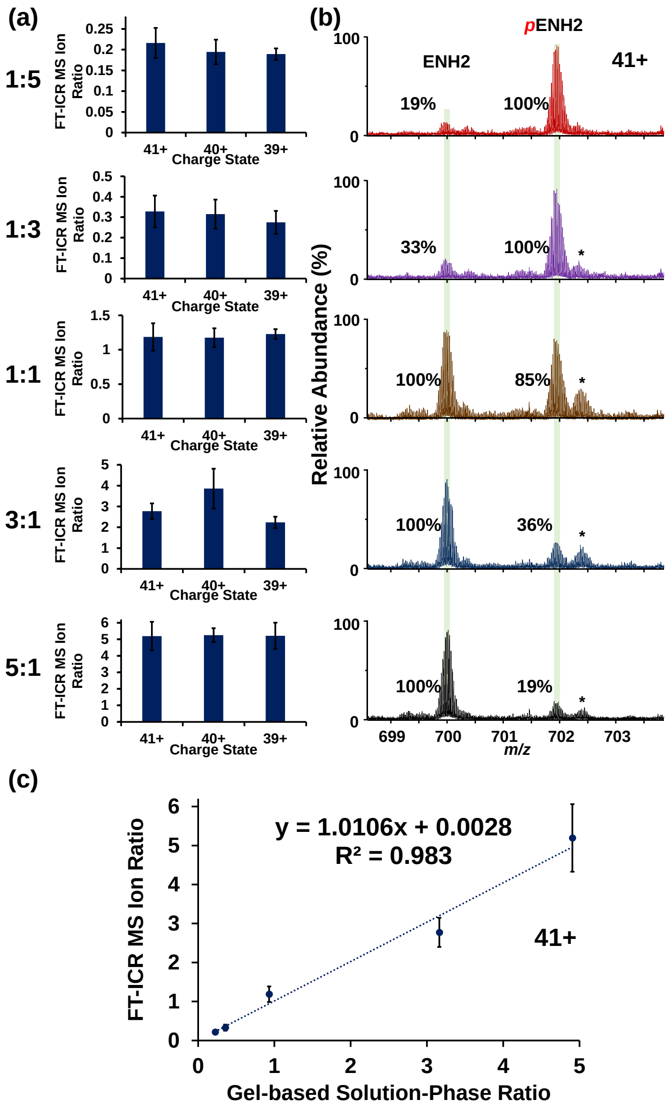


Figure S2.8. Gas-phase quantification using 12T solariX FT-ICR MS analysis for ENH2. (a) Gas-phase ratios of ENH2:*p*ENH2 at charge state 41$^+$, 40$^+$, and 39$^+$ were compared with five different solution-phase ratios from 1:5, 1:3, 1:1, 3:1 to 5:1 (top to bottom). (b) Representative FT-ICR spectra at charge state 41$^+$ of the five solution-phase ratios were shown. In each spectrum, the relative abundance is normalized to the highest abundance species. (c) Correlation analysis for charge state 41$^+$ shows a good linear correlation between the gas-phase ratios (derived from FT-ICR MS data) and solution-phase ratios (derived from SDS-gel data). *ENH2 with phosphate adduct.

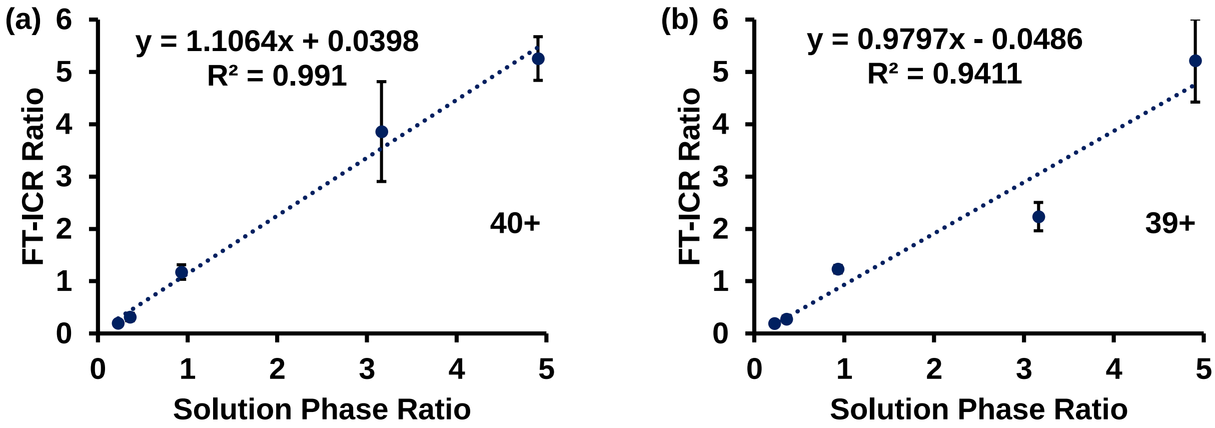


Figure S2.9. Gas-phase quantification using FT-ICR MS analysis for ENH2 at charge state 40$^+$, and 39$^+$. Correlation analysis of charge state (a) 40$^+$ and (b) 39$^+$ shows a good linear correlation between the gas-phase ratios (derived from FT-ICR MS data) and solution-phase ratios (derived from SDS-gel data).

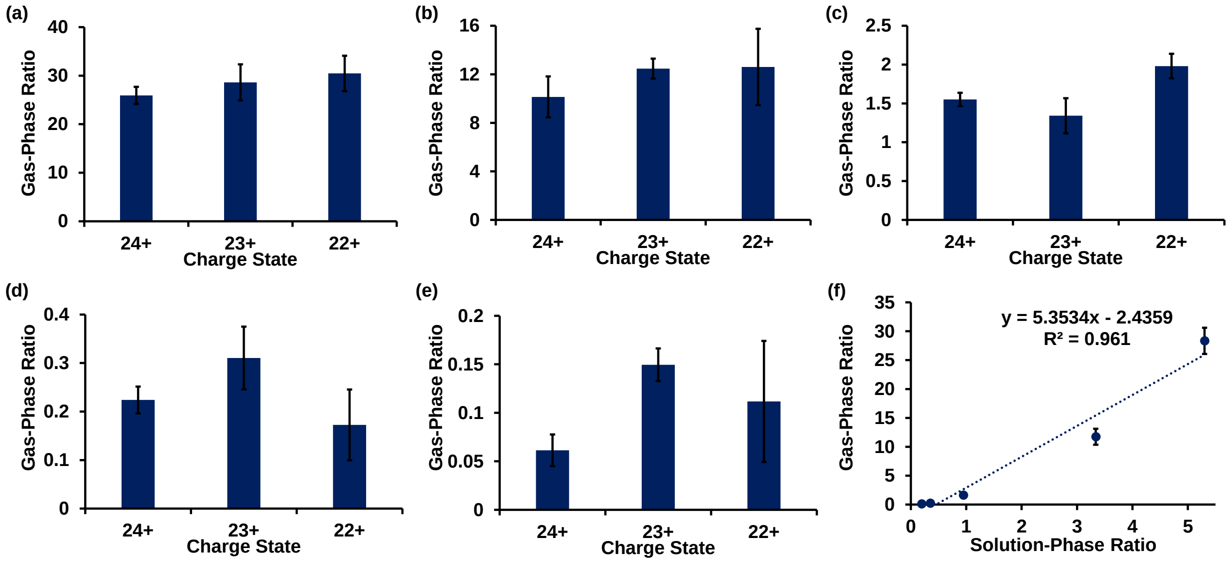


Figure S2.10. Quantification for β-casein proteoforms at different charge states using 12T solariX FT-ICR mass spectrometer. Gas-phase ratios of β-casein and 5*p*β-casein were compared with solution-phase ratios of (a) 1:5, (b) 1:3, (c) 1:1, (d) 3:1, (e) 5:1. (f) The correlation between the average of gas-phase ratios among three charge states and solution-phase ratios was shown. The data set was fit to a linear equation.

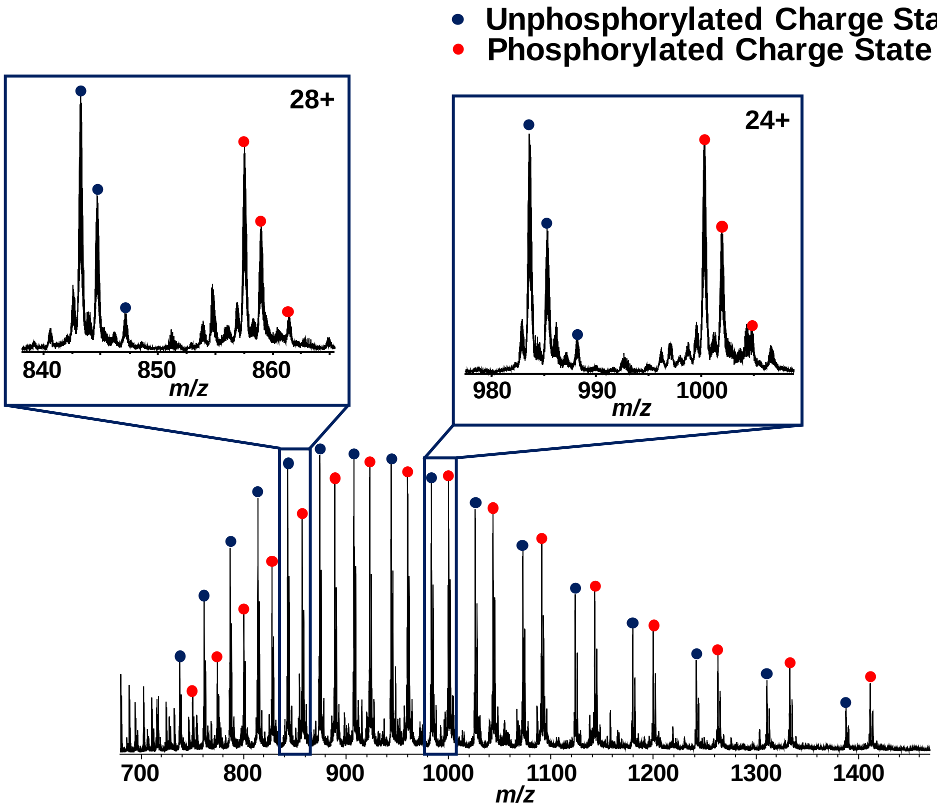
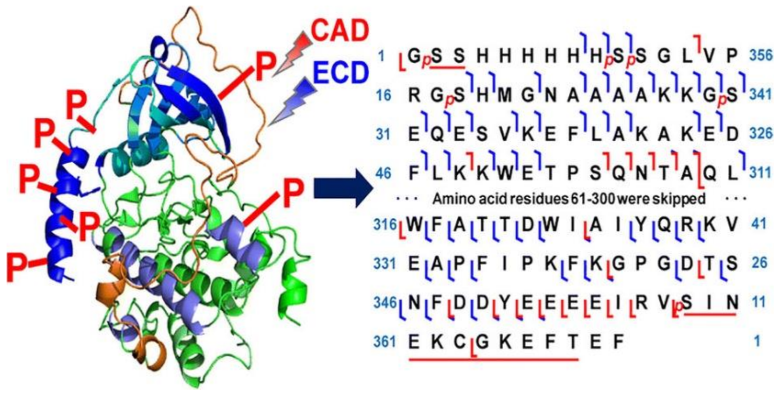


Figure S2.11. ESI/Q-TOF MS analysis of a mixture of 1:1 gas-phase ratio of β-casein:5*p*β-casein derived based on the deconvoluted spectrum in the revealed a shift in charge state distribution. Charge state distribution of solution mixture with 1:1 gas-phase ratio of β-casein:5*p*β-casein based on deconvoluted spectra is shown. Insets, representative charge states showing different gas-phase ratios.

Chapter 3

Comprehensive Characterization of the Recombinant Catalytic Subunit of cAMP-Dependent Protein Kinase by Top-Down Mass Spectrometry


CAD
ECD
P
P
P
P
P
P
P
1 GpS S H H H H H HpSpS G L V P 356
16 R GpS H M G N A A A A K K GpS 341
31 E Q E S V K E F L A K A K E D 326
46 F L K K W E T P S Q N T A Q L 311
··· Amino acid residues 61-300 were skipped ···
316 W F A T T D W I A I Y Q R K V 41
331 E A P F I P K F K G P G D T S 26
346 N F D D Y E E E E I R VpS I N 11
361 E K C G K E F T E F 1

Adapted from: Wu, Z.; Jin, Y.; Chen, B.; Gugger, M. K.; Wilkinson-Johnson, C. L.; Tiambeng, T. N.; Jin, S.; Ge, Y.*, Comprehensive Characterization of the Recombinant Catalytic Subunit of cAMP-Dependent Protein Kinase by Top-Down Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (12), 2561-2570.

Abstract

Reversible phosphorylation plays critical roles in cell growth, division, and signal transduction. Kinases which catalyze the transfer of γ-phosphate groups of nucleotide triphosphates to their substrates are central to the regulation of protein phosphorylation and are therefore important therapeutic targets. Top-down MS presents unique opportunities to study protein kinases owing to its capabilities in comprehensive characterization of proteoforms that arise from alternative splicing, sequence variations, and post-translational modifications. Here, for the first time, we developed a top-down MS method to characterize the C-subunit of an important kinase, PKA. The recombinant PKA C-subunit was expressed in E. coli and successfully purified via his-tag affinity purification. By intact mass analysis with high resolution and high accuracy, four different proteoforms of the affinity-purified PKA C-subunit were detected and the most abundant proteoform was found containing seven phosphorylations with the removal of N-terminal methionine. Subsequently, the seven phosphorylation sites of the most abundant PKA C-subunit proteoform were characterized simultaneously using tandem MS methods. Four sites were unambiguously identified as Ser10, Ser11, Ser18, and Ser30 and the remaining phosphorylation sites were localized to Ser2/Ser3, Ser358/Thr368, and Thr[215-224]Tyr in the PKA C-subunit sequence with a 20mer 6xHis-tag added at the N-terminus. Interestingly, four of these seven phosphorylation sites were located at the 6xHis-tag. Furthermore, we have performed dephosphorylation reaction by λPP, and showed that all phosphorylations of the recombinant PKA C-subunit phosphoproteoforms were removed by this phosphatase.

Introduction

Reversible phosphorylation is one of the key biological processes that govern cellular events including cell cycle control, cell growth, and signal transduction.[1-2] Aberrations in signaling events, such as up- and down-regulation of phosphorylation, are associated with the progress of human diseases.[3-8] Protein kinases are enzymes that catalyze the transfer of the γ-phosphate groups of nucleotide triphosphates to their substrates, and therefore are central to the regulation of protein phosphorylation.[9-10] Dysregulation of kinase signaling networks is increasingly recognized as an underlying mechanism that contributes to human diseases.[11-14] Consequently, numerous kinases inhibitors are currently utilized or under development for use as therapeutics.[15-16]

Protein kinases are also modulated by phosphorylation.[17] Autophosphorylation of protein kinases, or phosphorylation by other protein kinases, results in their activation or deactivation due to changes in their secondary structures.[18] Structural changes affect the binding kinetics to kinase substrates such as ATP and inhibitor peptides by altering the salt bridges and hydrogen bonding network at the active site.[18-19] One of the important protein kinases is the PKA, which partakes in many biological processes including mediating adrenergic stimulation in the heart and regulating the functions of skeletal muscle.[20-21] This protein kinase is a heterotetramer composed of two C-subunits, and two different regulatory subunits.[22-23] The PKA C-subunit has multiple phosphorylation sites displayed in the expressed proteins, which are associated with the physiochemical properties and enzymatic activity.[18-19, 24-25]

The phosphate groups at the phosphorylation sites are removed by protein phosphatase through the biological process of dephosphorylation, which together with protein phosphorylation, constitutes the reversible phosphorylation.[1] The removal of a phosphate group at the phosphorylation sites by phosphatase largely depends on the substrate specificity of the phosphatase.[26] Additionally, structural information around the phosphorylation sites can be revealed as the efficacy of dephosphorylation also relies on the accessibility, which is based on the structural environment of the phosphorylation sites under physiological conditions.[26] To better understand the function of phosphorylations on PKA C-subunit, a comprehensive characterization of the phosphorylation sites and the analysis of the dephosphorylation reaction are necessary.

Top-down MS presents unique opportunities to study protein kinases owing to its capabilities in analyzing alternative splicing, sequence variations and PTMs.[5, 27-34] Compared to bottom-up MS, which analyzes digested peptides, top-down MS analysis provides a "bird's eye" view of all proteoforms by analyzing proteins from the intact level.[35-37] In this study, we have developed a top-down MS strategy to characterize the recombinant PKA C-subunit. The affinity-purified PKA C-subunit, which was expressed in *E. coli*, was present with multiple proteoforms by intact mass analysis. The most abundant proteoform of PKA C-subunit was identified with seven phosphorylations along with the removal of N-terminal methionine. Using MS/MS techniques including CID and ECD, these seven phosphorylation sites were localized to specific amino acid residues or located to a region. Interestingly, four of these phosphorylation sites were located at the 6xHis-tag sequence. Dephosphorylation reactions using λPP suggested that all phosphorylation sites were accessible to this particular phosphatase. Taken together, we have demonstrated that top-down MS has unique advantages in comprehensively characterizing protein kinases.

Experimental Section

Chemicals and Reagents

All reagents were acquired from Sigma-Aldrich, Inc. (St. Louis, MO, USA), unless otherwise noted. Solvents, including HPLC grade H_2O, ACN and EtOH, were purchased from Fisher Scientific (Fair Lawn, NJ, USA).

Molecular Cloning

Commercial plasmid encoding the PKA C-subunit (plasmid # 14921) was purchased from Addgene (Watertown, MA, USA) in an agar gel piece.[38] A small agar piece was transferred in 5 mL TB media with 100 µg/mL ampicillin, and the mixture was allowed to grow for 9 h in a shaker. The cells were collected and the growth media was discarded. The plasmid DNA was extracted using QIAprep Spin Miniprep Kit (QIAGEN, Hilden, Germany) following the manufacturer recommended protocol. The plasmid product was transformed into ScarabXpress T7 *E. coli* cells (Scarab Genomics, Madison, WI, USA), and a glycerol stock was prepared.

Protein Expression and Purification

The protein expression and purification protocol was similar to that previously described.[39] Briefly, a starter LB broth culture with 100 µg/mL ampicillin was inoculated by glycerol stock of the *E. coli* and the starter culture was allowed to grow overnight. A small amount of starter culture was transferred to LB broth containing 100 µg/mL ampicillin. The culture was allowed to grow until the optical density of the culture reached 0.4 to 0.6. IPTG at a final concentration of 0.1 mM was introduced to induce protein expression, and the bacteria were cultured at 30 °C for 9 h. The cells were harvested by centrifugation and the cell pellets were stored at -80 °C prior to protein purification.

Unless stated otherwise, additives include 1 mM DTT and 0.25 mM PMSF. The cell pellets were lysed by sonication in 50 mM NaH_2PO_4 pH 7.4, 250 mM NaCl (10 mL/g pellet) buffer (Buffer A) with additives and protease inhibitor cocktail (Sigma-Aldrich Inc.). The cell debris were removed by centrifugation. For 2 mL of the supernatant, 250 µL of Dynabeads™ His-Tag Isolation and Pulldown (Invitrogen™, Carlsbad, CA, USA) was added, and the mixture was agitated at 4 °C for 30 min. The supernatant was removed and the Dynabeads were washed twice with Buffer A containing additives, once with 50 mM Tris pH 7.4, 50 mM NaCl buffer (Buffer B) with additives, and finally with Buffer B containing additives and 25 mM imidazole. The attached proteins were eluted with Buffer B with additives, 300 mM imidazole, and protease inhibitor cocktail, and was concentrated using a Pierce™ Protein Concentrators PES, 10K MWCO filter (Fisher Scientific). The efficacy of the protein purification was verified by SDS-PAGE analysis.

Dephosphorylation Reaction

The dephosphorylation reaction for the PKA C-subunit (~ 40 µg) was performed using ~ 150 units of λPP (New England Biolabs Inc., Ipswich, MA, USA) following the manufacturer recommended protocol. Briefly, the reaction was supplemented with 1 mM $MnCl_2$ solution and 1X NEBuffer for PMP (New England Biolabs Inc.) and allowed to proceed for 2 h at 30 °C to achieve complete dephosphorylation.

Top-down Mass Spectrometry

For online MS analysis, the PKC C-subunit samples were separated using a homemade PLRP reversed-phase column (200 mm length × 500 µm i.d., 10 µm particle size, 1,000 Å pore size). PLRP-S particles were obtained from Agilent Technologies (Santa Clara, CA, USA). Mobile phase A (MPA) contained H_2O with 0.1% FA and mobile phase B (MPB) contained 50:50 ACN:EtOH with 0.1% FA. LC was performed with a 60 min linear gradient which ran at 5% MPB from 0 to 5 min, followed by 5% to 65% MPB from 5 to 40 min, 65% to 95% MPB from 40-53 min, and back to 5% MPB at a flow rate of 12 μL/min. Five microliters (5 μL) of sample were injected for all experiments. The sample was analyzed either using a maXis II Q-TOF mass spectrometer (Bruker Daltonics, Bremen, Germany) coupled with an ACQUITY UPLC M-Class System (Waters Corporation, Milford, MA, USA), or using a 12T solariX FT-ICR mass spectrometer (Bruker Daltonics) coupled with a nanoACQUITY UPLC System (Waters Corporation). For online LC-MS/MS experiments with CID fragmentation using a maXis II Q-TOF mass spectrometer, the precursor ion was isolated and subjected to 15 - 20 eV energy for fragmentation.

For offline MS analysis, the fraction was collected using a nanoACQUITY UPLC System. The sample was introduced to a 12T solariX FT-ICR mass spectrometer using a TriVersa NanoMate® (Advion Bioscience, Ithaca, NY, USA) as previously described.[5, 40] The mass spectra were collected over a 200 to 3000 *m/z* range with 2 M transient size (1.2 s transient length) and a pulse at 28% excitation power. In MS/MS analysis, an isolation window of 1.8 – 2 *m/z* was used for the precursor ion. Mass spectra were accumulated for 500 to 750 scans. For CID experiments, an energy from 6 to 12 V was set to generate fragment ions. For ECD experiments, the parameters for ECD pulse length, ECD bias, and ECD lens were set to 0.020 s, 0.3 - 0.6 V, and 10 V, respectively.

Data Analysis

All reported masses are monoisotopic masses. For intact mass analysis, the spectra were analyzed using DataAnalysis 4.2 and deconvoluted using the Maximum Entropy deconvolution algorithm. The monoisotopic mass was calculated using the SNAP algorithm in DataAnalysis. For MS/MS analysis, the data were analyzed using MASH Suite Pro.[41] Peak extraction was performed using a signal-to-noise ratio of 3 and a minimum fit of 60%, and all peaks were subjected to manual validation. A 10-ppm mass tolerance was used to match the experimental fragment ions to the calculated fragment ions based on amino acid sequence.

Results and Discussion

We developed a top-down MS strategy for the comprehensive characterization of recombinant PKA C-subunit (Figure 3.1). The strategy started with obtaining a plasmid encoding the PKA C-subunit, and subsequently transforming the plasmid into a vector. Afterwards, the PKA C-subunit was overexpressed in *E. coli*, and the protein was purified by affinity purification. The protein was first subjected to intact mass analysis which reveals the sequence variations and PTMs by accurate mass measurements. These putative modifications were first assessed by online CID experiment for protein fragmentation analysis. Then the fraction containing the PKA C-subunit was collected after LC separation, and further subjected to offline characterization using both CID and ECD at various fragmentation settings for verification of the putative modifications.

PKA C-subunit expression and purification

The plasmid encoding the PKA C-subunit was kindly provided by Dr. Susan Taylor from UCSD through Addgene organization.[38] The plasmid includes a 20 amino acid 6xHis-tag sequence before the endogenous sequence of PKA C-subunit derived from mice [UniProtKB - P05132]. The mouse-derived C-subunit of PKA is composed of 351 amino acid residues. For overexpression of the PKA C-subunit, the plasmid was purified and transformed into the pET-28a(+) vector. To capture the 6xHis-tag on the PKA C-subunit, affinity purification using Dynabeads was employed, which is based on TALON technology (Figure S3.1a). The loading mixture, flow through, and elution fractions were evaluated by SDS-PAGE analysis (Figure S3.1b). The PKA C-subunit was determined to be successfully purified based on the presence of a dark band at around 42 kDa, which is consistent with the predicted protein mass (M_r: 42575.92 Da) from the encoding amino acid sequence. Although the PKA C-subunit was present as the most prominent band by SDS-PAGE analysis, other faint bands could also be observed in the elution lanes. In particular, some lower mass proteins might suppress the ionization and detection of the PKA C-subunit in the top-down MS analysis. Therefore, our strategy was to use RPLC methods to separate the PKA C-subunit from other proteins for both online and offline characterization.

Online LC-MS/MS Profiling of Multiple Proteoforms

The affinity-purified PKA C-subunit was subjected to RPLC separation coupled online with high-resolution MS analysis using a Q-TOF instrument. Using H_2O as MPA and 50:50 ACN:EtOH as MPB, the PKA C-subunit was separated and detected by MS with minimal impurities, and this was demonstrated by the charge state distribution envelope (Figure 3.2a and Figure S3.2). The deconvoluted spectra revealed the existence of multiple PKA C-subunit proteoforms but none of the masses of these proteoforms matched with the theoretical protein mass based on the predicted amino acid sequence (Figure 3.2a, inset). The mass difference between two neighboring peaks was 79.97 Da, indicating the occurrence of phosphorylation on these proteoforms. As it is common that the N-terminal methionine of recombinant proteins would be removed by methionyl-aminopeptidase after protein translation, the mass of methionine was first deducted from the theoretical protein mass.[42] The PKA C-subunit proteoforms were found to contain six to nine phosphorylations with N-terminal methionine removed based on the results from the deconvoluted spectra. With a mass shift of 559.49 Da, the most abundant proteoform was modified with seven phosphorylations in addition to the removal of N-terminal methionine, which accounted for ~ 45% of relative percentage of all PKA C-subunit proteoforms (Figure 3.2a, inset). Collectively, the affinity-purified PKA C-subunit was hyperphosphorylated from *E. coli* expression, which is consistent with previous studies.[25, 43]

For the initial PTM site characterization, the PKA C-subunit was subjected to online LC-MS/MS with CID fragmentation on the precursor ion corresponding to the most abundant proteoform with seven phosphorylation sites. Since phosphorylation is the only PTM being considered, the mass list was matched with the theoretical fragment ion list by adding the mass of phosphorylation modification. Asides from the precursor ion (M^{49+}), several abundant fragment ions were observed (Figure 3.2b). The masses of these abundant fragment ions were identical and could be identified as y_{114} ions at different charge states after accounting for the mass of one phosphorylation. Less abundant ions at 560 – 850 *m/z* afforded additional information regarding the phosphorylation sites (Figure 3.2b, inset). Several low mass *y* ions (y_{20}, y_{19}, andy_{15}) suggested a phosphorylation site located after Arg356 at the C-terminus. A series of *b* ions (b_{54}, b_{55}, b_{56}, and b_{57}) could also be identified with mass difference equivalent to five phosphorylations, indicating five phosphorylation sites located before Lys54 at the N-terminus. Lastly, a b_{254} ion was identified with mass difference equivalent to six phosphorylation, suggesting that a phosphorylation site was located in the middle of the recombinant PKA C-subunit sequence. Using online CID characterization, fragment ions from MS/MS spectra localized five phosphorylation site before Lys54 at the N-terminus, one phosphorylation site after Arg356 at the C-terminus, and one phosphorylation site in the middle of the sequence for the most abundant PKA C-subunit proteoform.

Characterization of PKA C-subunit phosphorylation sites by high-resolution MS/MS Analysis

We sought to localize all phosphorylation sites present in the most abundant proteoform of the PKA C-subunit using offline MS analysis combining different fragmentation methods. The fraction containing the recombinant PKA C-subunit was collected after LC separation, and the samples were analyzed on an ultrahigh-resolution FT-ICR mass spectrometer. In this study, when referring the amino acid residue in the endogenous sequence, note that the reference is to the UniProt sequence [UniProtKB - P05132] with N-terminal methionine removed to be consistent with the previous reports.[19, 24] As shown from the online CID results, there were five phosphorylation sites located near the N-terminus. As a result, ECD fragmentation method was used for site localization, which is known to preserve labile modifications such as phosphorylation.[44] ECD was able to effectively fragment most of the bonds at the N-terminus, and a plethora of fragment ions was observed in the raw spectra from the ECD experiment (Figure 3.3a and Figure S3.3).

For the first phosphorylation site at the N-terminus, both c_8 and c_9 differed from the theoretical mass by 79.97 Da, indicative of the occurrence of phosphorylation (Figure 3.3b). Ser2 and Ser3 are the only two amino acid residues that can be phosphorylated; however, the site could not be definitively localized to either Ser2 or Ser3 without additional fragment ions. Three additional phosphorylation sites were localized at Ser10, Ser11, and Ser18, which were confirmed by c_9, c_{10}, and c_{19} ions (Figure 3.3b). Ser11 and Ser18 were confirmed with only two *c* ions, c_{10}, and c_{19}, as these are the only two sites which could be phosphorylated. Intriguingly, all of these four phosphorylation sites were located at the added 6xHis-tag sequence. Hyperphosphorylation at the 6xHis-tag sequence was also observed in other case using *E. coli* for kinase expression such as that for Aurora A.[45] Iakoucheva *et al.* suggested that protein phosphorylation predominantly occurred at disordered regions [46]. The structure of the 6xHis-tag sequence along with the first 12 amino acid residues of the PKA C-subunit was found to be disordered from previous X-ray crystallography study, which supported our observation that the four phosphorylations took place at the disordered 6xHis-tag sequence.[38] This 20mer 6xHis-tag sequence (MGSSHHHHHHSSGLVPRGSH) is a common sequence added at the N-terminus due to its dual functionality.[47-48] This tag includes a 6xHis-tag for affinity purification and a thrombin cleavage site (LVPR/GS). Proteins with only Gly-Ser-His added at the N-terminus of the endogenous protein sequence could be yielded after reacting the affinity-purified protein with thrombin.[49] In the case of the PKA C-subunit, the thrombin cleavage site was not utilized as the 20mer 6xHis-tag did not affect the structure and enzymatic activity of this protein.[19, 38]

The last of the five phosphorylation sites at the N-terminus was localized at Ser30, which was confirmed by c_{18} and c_{32} ions (Figure 3.3b). This phosphorylation site is equivalent to Ser10 in the endogenous sequence. Previously, Tholey *et al.* suggested that phosphorylation at Ser10 altered the structure at the N-terminus, resulting in the amplified extent of electrostatic interaction.[50] Yonemoto *et al.* argued that Ser10 could be autophosphorylated *in vitro*, and that this site was significant for protein solubility.[19] Mutation at Ser10 significantly impaired the solubility of protein in aqueous solution. Therefore, the phosphorylation at Ser10 was shown to be important for protein structure and solubility.

Next, we sought to identify the phosphorylation site at the C-terminus. Since ECD did not generate sufficient fragment ions, the site was instead characterized primarily by CID fragment ions. The phosphorylation site was localized at Ser358 or Thr368 by y_{13} and a series of y ions from y_{15} to y_{21} (Figure 3.4a and Figure 3.4b). Although an y_7 ion without phosphorylation was observed in the CID experiment, this ion could not be used for confident identification of phosphorylation site at Ser358 due to the possibility that this ion was present after the loss of the phosphorylation at Thr368 (Figure 3.4b).[44] One of the potential phosphorylation sites, Ser358, equivalent to Ser338 in the endogenous sequence, has been reported previously.[19, 51] Yonemoto *et al.* suggested that this phosphorylation site was relevant to catalytic activity and protein stability. Mutations of recombinant PKA C-subunit with S338A or S338E either disrupted the catalytic activity or altered the binding kinetics for inhibitor peptide and ATP.[19]

The localization of phosphorylation site in the middle of the sequence required fragment ions generated from both CID and ECD fragmentation methods. The phosphorylation site was narrowed down to T[215-224]Y by b_{224} with six phosphorylations, c_{210} with five phosphorylations and $z{\bullet}_{157}$ with two phosphorylations (Figure 3.4c and Figure 3.4d). One of the potential phosphorylation sites was Thr217, which is equivalent to Thr197 in the endogenous sequence. Phosphorylation at Thr197 is crucial to catalytic activity, as it allows PKA C-subunit to change from an inactive state to an active state.[18] It does so by forming salt bridges with amino acid residues from other parts of the PKA C-subunit, such as C-helix, catalytic loop, β9, and activation loop.

In previous studies, phosphorylation sites on PKA C subunit were usually identified by bottom-up MS based on the detection of phosphopeptides, in which the identified phosphorylation sites are from a mixture of multiply-phosphorylated proteoforms.[24-25] Compared to the bottom-up MS strategy, our top-down MS strategy analyzes intact proteins, giving a bird's eye view of all proteoforms present. This approach not only shows the stoichiometry of different proteoforms in a single sample, but also provides a comprehensive analysis of all phosphorylation sites present in a single proteoform.

Top-down MS/MS Sequencing of the PKA C-subunit

By combining five CID spectra and three ECD spectra, 191 of 369 possible bonds were cleaved, providing a 52% sequence coverage for the recombinant PKA C-subunit (Figure 3.5). A series of CID fragment ions was observed at Ser[54-58]Gln, Ser[134-145]Gly, Gly[246-265]Gln, and Tyr[350-357]Arg (Figure S3.4). Interestingly, although loss of phosphorylation would sometimes occur in CID, a series of *b* ions with all phosphorylations intact was observed from Ser[54-58]Gln, Ser[134-145]Gly, and Gly[246-255]Tyr. Compared to peptide fragmentation, CID of intact proteins often could retain a portion of the labile modifications in the top-down approach [44]. Fragmentation at the amide backbone was preferred over PTM ejection, which is likely due to the higher-order structure of gas phase ions that are larger than ~8 kDa.[52] By contrast, only a few *b* ions were observed for the first 50 amino acid residues in the N-terminus, likely due to the loss of phosphate group(s) as a result of their smaller size. ECD fragmentation yielded bond cleavages unique to CID fragmentation due to the difference in the dissociation mechanism.[53] This method provided good sequence coverage at both the N- and C-terminus; however, the fragmentation efficiency was suboptimal for bonds in the middle of the sequence, despite some larger ECD fragment ions being observed (Figure S3.5 and Figure S3.6). For this study, utilizing both CID and ECD fragmentation methods, the phosphorylation sites of the PKA C-subunit proteoform with seven phosphorylations were characterized. Current development in fragmentation methods, such as UVPD, will be beneficial to achieve a higher sequence coverage due to additional generated ion species such as *a* and *x* ions, in addition to *b*, *c*, *y* and *z•* ions.[54]

Dephosphorylation of hyperphosphorylated PKA C-subunit

Dephosphorylation, which removes phosphate groups on their substrates by phosphatases, is complimentary to phosphorylation. We were interested in how the recombinant PKA C-subunit proteoforms react to a common phosphatase, λPP. The dephosphorylation reaction of the affinity-purified PKA C-subunit was performed and the reaction product was analyzed by top-down MS. A drastic shift in peaks was observed in each charge state due to the loss of multiple phosphorylations (Figure 3.6a and Figure S3.7). From the deconvoluted spectra, all phosphoproteoforms collapsed into a single unphosphorylated proteoform after the dephosphorylation reaction (Figure 3.6b). This suggested that all phosphorylation sites were accessible by λPP and were subsequently dephosphorylated. Byrne *et al.* observed that after dephosphorylation, the most abundant proteoform still possessed two phosphorylations detected by low-resolution top-down MS analysis.[25] The reaction conditions of the dephosphorylation reaction between our studies and the study done by Byrne *et al.* were different. In our protocol, we followed the manufacturer recommended conditions from the New England Biolabs and performed the dephosphorylation reaction at 30 °C. In comparison, Byrne *et al.* conducted the dephosphorylation reaction at 37 °C using bacterially expressed λPP. At elevated temperature, the phosphatase might not reach its maximum kinetics and might denature after prolonged incubation.

Conclusively, our result showed that all phosphorylations on the phosphoproteoform of the PKA C-subunit were removed by λPP. The discrepancy between our result and result from Byrne *et al.* might be due to the difference in reactions conditions.

Conclusion

For the first time, a top-down MS strategy was developed to achieve a comprehensive characterization of the recombinant PKA C-subunit. The PKA C-subunit was overexpressed in *E. coli* and the expressed protein with 6xHis-tag was successfully purified by affinity purification. The affinity-purified PKA C-subunit was subjected to intact mass analysis and proteoforms with six to nine phosphorylations were observed. The most abundant proteoform was identified with seven phosphorylations and removal of N-terminal methionine. Using CID and ECD fragmentation methods, all seven phosphorylation sites were characterized simultaneously for the first time. Four of the phosphorylation sites were unambiguously localized to Ser10, Ser11, and Ser18, which were located at the 20mer 6xHis-tag sequence, as well as Ser30, which corresponded to Ser10 in the endogenous sequence Three other phosphorylation sites were localized to Ser2/Ser3, Thr[215-224]Tyr, and Ser358/Thr368. By combining five CID and three ECD experiments, a 52% sequence coverage was achieved for the PKA C-subunit with seven phosphorylations. Finally, dephosphorylation experiments showed that all phosphorylations of the PKA C-subunit phosphoproteoform were removed by λPP.

Acknowledgement

The authors would like to thank Dr. Wenxuan Cai for making the construct for PKA C-subunit for bacterial expression. Financial support was provided by NIH R01 GM117058 (to S. J. and Y. G.) and R01 GM125085 (to Y. G.). Y. G. also would like to acknowledge the NIH grants, R01 HL096971, R01 HL109810, and S10 OD018475.

References

1. Hunter, T., Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* **1995,** *80* (2), 225-36.

2. 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), 635-48.

3. Grimes, M.; Hall, B.; Foltz, L.; Levy, T.; Rikova, K.; Gaiser, J.; Cook, W.; Smirnova, E.; Wheeler, T.; Clark, N. R.; Lachmann, A.; Zhang, B.; Hornbeck, P.; Ma'ayan, A.; Comb, M., Integration of protein phosphorylation, acetylation, and methylation data sets to outline lung cancer signaling networks. *Sci. Signal.* **2018,** *11* (531).

4. Hanger, D. P.; Anderton, B. H.; Noble, W., Tau phosphorylation: the therapeutic challenge for neurodegenerative disease. *Trends Mol. Med.* **2009,** *15* (3), 112-9.

5. 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.; Liu, X.; Ge, Y., Top-down proteomics reveals concerted reductions in myofilament and Z-disc protein phosphorylation after acute myocardial infarction. *Mol. Cell. Proteomics* **2014,** *13* (10), 2752-64.

6. Dong, X. T.; Sumandea, C. A.; Chen, Y. C.; Garcia-Cazarin, M. L.; Zhang, J.; Balke, C. W.; Sumandea, M. P.; Ge, Y., Augmented Phosphorylation of Cardiac Troponin I in Hypertensive Heart Failure. *J. Biol. Chem.* **2012,** *287* (2), 848-857.

7. Zhang, J.; Guy, M. J.; Norman, H. S.; Chen, Y. C.; Xu, Q. G.; Dong, X. T.; Guner, H.; Wang, S. J.; Kohmoto, T.; Young, K. H.; Moss, R. L.; Ge, Y., Top-Down Quantitative Proteomics Identified Phosphorylation of Cardiac Troponin I as a Candidate Biomarker for Chronic Heart Failure. *J. Proteome Res.* **2011,** *10* (9), 4054-4065.

8. Chen, I. H.; Xue, L.; Hsu, C. C.; Paez, J. S. P.; Pan, L.; Andaluz, H.; Wendt, M. K.; Iliuk, A. B.; Zhu, J. K.; Tao, W. A., Phosphoproteins in extracellular vesicles as candidate markers for breast cancer. *P. Natl. Acad. Sci. USA* **2017,** *114* (12), 3175-3180.

9. Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S., The protein kinase complement of the human genome. *Science* **2002,** *298* (5600), 1912-34.

10. Hanks, S. K.; Quinn, A. M.; Hunter, T., The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **1988,** *241* (4861), 42-52.

11. Sacco, F.; Silvestri, A.; Posca, D.; Pirro, S.; Gherardini, P. F.; Castagnoli, L.; Mann, M.; Cesareni, G., Deep Proteomics of Breast Cancer Cells Reveals that Metformin Rewires Signaling Networks Away from a Pro-growth State. *Cell Syst.* **2016,** *2* (3), 159-171.

12. Dhillon, A. S.; Hagan, S.; Rath, O.; Kolch, W., MAP kinase signalling pathways in cancer. *Oncogene* **2007,** *26* (22), 3279-90.

13. Chatterjee, K., Neurohormonal activation in congestive heart failure and the role of vasopressin. *Am. J. Cardiol.* **2005,** *95* (9a), 8b-13b.

14. Vlahos, C. J.; McDowell, S. A.; Clerk, A., Kinases as therapeutic targets for heart failure. *Nat. Rev. Drug. Discov.* **2003,** *2* (2), 99-113.

15. Zhang, J. M.; Yang, P. L.; Gray, N. S., Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009,** *9* (1), 28-39.

16. Ferguson, F. M.; Gray, N. S., Kinase inhibitors: the road ahead. *Nat. Rev. Drug. Discov.* **2018,** *17* (5), 353-376.

17. Roskoski, R., Jr., ERK1/2 MAP kinases: structure, function, and regulation. *Pharmacol. Res.* **2012,** *66* (2), 105-43.

18. Steichen, J. M.; Iyer, G. H.; Li, S.; Saldanha, S. A.; Deal, M. S.; Woods, V. L., Jr.; Taylor, S. S., Global consequences of activation loop phosphorylation on protein kinase A. *J. Biol. Chem.* **2010,** *285* (6), 3825-32.

19. Yonemoto, W.; McGlone, M. L.; Grant, B.; Taylor, S. S., Autophosphorylation of the catalytic subunit of cAMP-dependent protein kinase in Escherichia coli. *Protein Eng.* **1997,** *10* (8), 915-25.

20. Wheeler-Jones, C. P., Cell signalling in the cardiovascular system: an overview. *Heart* **2005,** *91* (10), 1366-74.

21. Ruehr, M. L.; Russell, M. A.; Ferguson, D. G.; Bhat, M.; Ma, J.; Damron, D. S.; Scott, J. D.; Bond, M., Targeting of protein kinase A by muscle A kinase-anchoring protein (mAKAP) regulates phosphorylation and function of the skeletal muscle ryanodine receptor. *J. Biol. Chem.* **2003,** *278* (27), 24831-6.

22. Bauman, A. L.; Scott, J. D., Kinase- and phosphatase-anchoring proteins: harnessing the dynamic duo. *Nat. Cell. Biol.* **2002,** *4* (8), E203-E206.

23. Taylor, S. S.; Knighton, D. R.; Zheng, J. H.; Teneyck, L. F.; Sowadski, J. M., Structural Framework for the Protein-Kinase Family. *Annu. Rev. Cell. Biol.* **1992,** *8*, 429-462.

24. Yonemoto, W.; Garrod, S. M.; Bell, S. M.; Taylor, S. S., Identification of phosphorylation sites in the recombinant catalytic subunit of cAMP-dependent protein kinase. *J. Biol. Chem.* **1993,** *268* (25), 18626-32.

25. Byrne, D. P.; Vonderach, M.; Ferries, S.; Brownridge, P. J.; Eyers, C. E.; Eyers, P. A., cAMP-dependent protein kinase (PKA) complexes probed by complementary differential scanning fluorimetry and ion mobility-mass spectrometry. *Biochem. J.* **2016,** *473* (19), 3159-75.

26. Roy, J.; Cyert, M. S., Cracking the phosphatase code: docking interactions determine substrate specificity. *Sci. Signal.* **2009,** *2* (100), re9.

27. Toby, T. K.; Fornelli, L.; Kelleher, N. L., Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem. (Palo Alto Calif)* **2016,** *9* (1), 499-519.

28. Cai, W. X.; Tucholski, T. M.; Gregorich, Z. R.; Ge, Y., Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. *Expert Rev. Proteomic* **2016,** *13* (8), 717-730.

29. Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y., Top-Down Proteomics: Ready for Prime Time? *Anal. Chem.* **2018,** *90* (1), 110-127.

30. 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. X.; Wu, C.; Sweet, S. M. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L., Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011,** *480* (7376), 254-U141.

31. Brown, K. A.; Chen, B.; Guardado-Alvarez, T. M.; Lin, Z.; Hwang, L.; Ayaz-Guner, S.; Jin, S.; Ge, Y., A photocleavable surfactant for top-down proteomics. *Nat. Methods* **2019,** *16* (5), 417-420.

32. Chen, B.; Hwang, L.; Ochowicz, W.; Lin, Z.; Guardado-Alvarez, T. M.; Cai, W.; Xiu, L.; Dani, K.; Colah, C.; Jin, S.; Ge, Y., Coupling functionalized cobalt ferrite nanoparticle enrichment with online LC/MS/MS for top-down phosphoproteomics. *Chem. Sci.* **2017,** *8* (6), 4306-4311.

33. Roberts, D. S.; Chen, B. F.; Tiambeng, T. N.; Wu, Z. J.; Ge, Y.; Jin, S., Reproducible large-scale synthesis of surface silanized nanoparticles as an enabling nanoproteomics platform: Enrichment of the human heart phosphoproteome. *Nano Res.* **2019,** *12* (6), 1473-1481.

34. Ge, Y.; Rybakova, I. N.; Xu, Q. G.; Moss, R. L., Top-down high-resolution mass spectrometry of cardiac myosin binding protein C revealed that truncation alters protein phosphorylation state. *P. Natl. Acad. Sci. USA* **2009,** *106* (31), 12658-12663.

35. Major, L. L.; Denton, H.; Smith, T. K., Coupled Enzyme Activity and Thermal Shift Screening of the Maybridge Rule of 3 Fragment Library Against Trypanosoma brucei Choline Kinase; A Genetically Validated Drug Target. In *Drug Discovery*, El-Shemy, H. A., Ed. Rijeka (HR), 2013.

36. Smith, L. M.; Kelleher, N. L., Proteoforms as the next proteomics currency. *Science* **2018,** *359* (6380), 1106-1107.

37. Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Loo, R. R. O.; Lundberg, E.; MacCoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schluter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlen, M.; Van Eyk, J. E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.; Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B., How many human proteoforms are there? *Nat. Chem. Biol.* **2018,** *14* (3), 206-214.

38. Narayana, N.; Cox, S.; Shaltiel, S.; Taylor, S. S.; Xuong, N., Crystal structure of a polyhistidine-tagged recombinant catalytic subunit of cAMP-dependent protein kinase complexed with the peptide inhibitor PKI(5-24) and adenosine. *Biochemistry* **1997,** *36* (15), 4438-48.

39. Wu, Z.; Tiambeng, T. N.; Cai, W.; Chen, B.; Lin, Z.; Gregorich, Z. R.; Ge, Y., Impact of Phosphorylation on the Mass Spectrometry Quantification of Intact Phosphoproteins. *Anal. Chem.* **2018,** *90* (8), 4935-4939.

40. Chen, Y. C.; Ayaz-Guner, S.; Peng, Y.; Lane, N. M.; Locher, M.; Kohmoto, T.; Larsson, L.; Moss, R. L.; Ge, Y., Effective top-down LC/MS+ method for assessing actin isoforms as a potential cardiac disease marker. *Anal. Chem.* **2015,** *87* (16), 8399-8406.

41. Cai, W.; Guner, H.; Gregorich, Z. R.; Chen, A. J.; Ayaz-Guner, S.; Peng, Y.; Valeja, S. G.; Liu, X.; Ge, Y., MASH Suite Pro: A Comprehensive Software Tool for Top-Down Proteomics. *Mol. Cell. Proteomics* **2016,** *15* (2), 703-14.

42. Hirel, P. H.; Schmitter, M. J.; Dessen, P.; Fayat, G.; Blanquet, S., Extent of N-terminal methionine excision from Escherichia coli proteins is governed by the side-chain length of the penultimate amino acid. *Proc. Natl. Acad. Sci. U S A* **1989,** *86* (21), 8247-51.

43. Herberg, F. W.; Bell, S. M.; Taylor, S. S., Expression of the catalytic subunit of cAMP-dependent protein kinase in Escherichia coli: multiple isozymes reflect different phosphorylation states. *Protein Eng.* **1993,** *6* (7), 771-7.

44. Siuti, N.; Kelleher, N. L., Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* **2007,** *4* (10), 817-821.

45. Haydon, C. E.; Eyers, P. A.; Aveline-Wolf, L. D.; Resing, K. A.; Maller, J. L.; Ahn, N. G., Identification of novel phosphorylation sites on Xenopus laevis Aurora A and analysis of phosphopeptide enrichment by immobilized metal-affinity chromatography. *Mol. Cell. Proteomics* **2003,** *2* (10), 1055-67.

46. Iakoucheva, L. M.; Radivojac, P.; Brown, C. J.; O'Connor, T. R.; Sikes, J. G.; Obradovic, Z.; Dunker, A. K., The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **2004,** *32* (3), 1037-1049.

47. Barraud, P.; Banerjee, S.; Mohamed, W. I.; Jantsch, M. F.; Allain, F. H., A bimodular nuclear localization signal assembled via an extended double-stranded RNA-binding domain acts as an RNA-sensing signal for transportin 1. *Proc. Natl. Acad. Sci. U S A* **2014,** *111* (18), E1852-61.

48. Comstock, M. J.; Whitley, K. D.; Jia, H.; Sokoloski, J.; Lohman, T. M.; Ha, T.; Chemla, Y. R., Protein structure. Direct observation of structure-function relationship in a nucleic acid-processing enzyme. *Science* **2015,** *348* (6232), 352-4.

49. Waugh, D. S., An overview of enzymatic reagents for the removal of affinity tags. *Protein Expres. Purif.* **2011,** *80* (2), 283-293.

50. Tholey, A.; Pipkorn, R.; Bossemeyer, D.; Kinzel, V.; Reed, J., Influence of myristoylation, phosphorylation, and deamidation on the structural behavior of the N-terminus of the catalytic subunit of cAMP-dependent protein kinase. *Biochemistry* **2001,** *40* (1), 225-31.

51. Toner-Webb, J.; van Patten, S. M.; Walsh, D. A.; Taylor, S. S., Autophosphorylation of the catalytic subunit of cAMP-dependent protein kinase. *J. Biol. Chem.* **1992,** *267* (35), 25174-80.

52. Meng, F. Y.; Cargile, B. J.; Miller, L. M.; Forbes, A. J.; Johnson, J. R.; Kelleher, N. L., Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat. Biotechnol.* **2001,** *19* (10), 952-957.

53. Zubarev, R. A.; Horn, D. M.; Fridriksson, E. K.; Kelleher, N. L.; Kruger, N. A.; Lewis, M. A.; Carpenter, B. K.; McLafferty, F. W., Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal. Chem.* **2000,** *72* (3), 563-73.

54. Holden, D. D.; Sanders, J. D.; Weisbrod, C. R.; Mullen, C.; Schwartz, J. C.; Brodbelt, J. S., Implementation of Fragment Ion Protection (FIP) during Ultraviolet Photodissociation (UVPD) Mass Spectrometry. *Anal. Chem.* **2018,** *90* (14), 8583-8591.

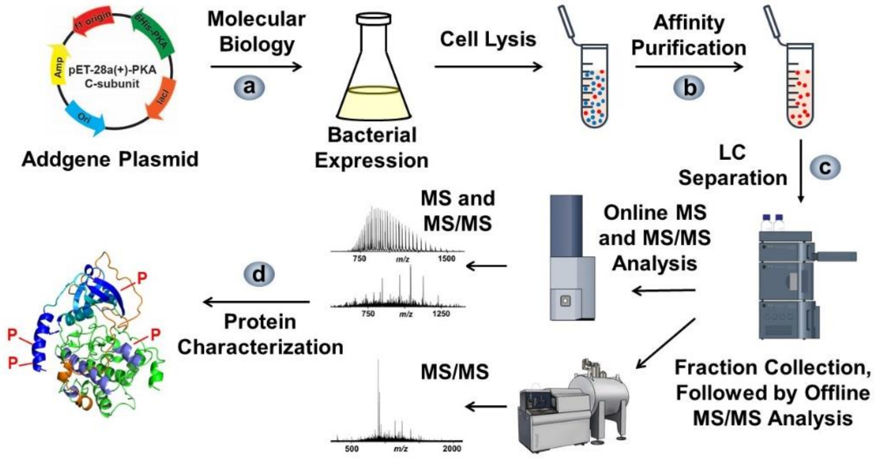


Figure 3.1. Workflow of expression, affinity purification, and top-down LC-MS analysis of the recombinant PKA C-subunit. a) The plasmid obtained from Addgene organization was transferred into a vector and expressed in *E. coli* cells. b) The PKA C-subunit was purified using affinity purification. c) The purified samples were subjected to online LC-MS and MS/MS analysis using a Q-TOF mass spectrometer and complemented with offline MS/MS analysis using a FT-ICR mass spectrometer. d) The modifications were characterized based on MS and MS/MS spectra.

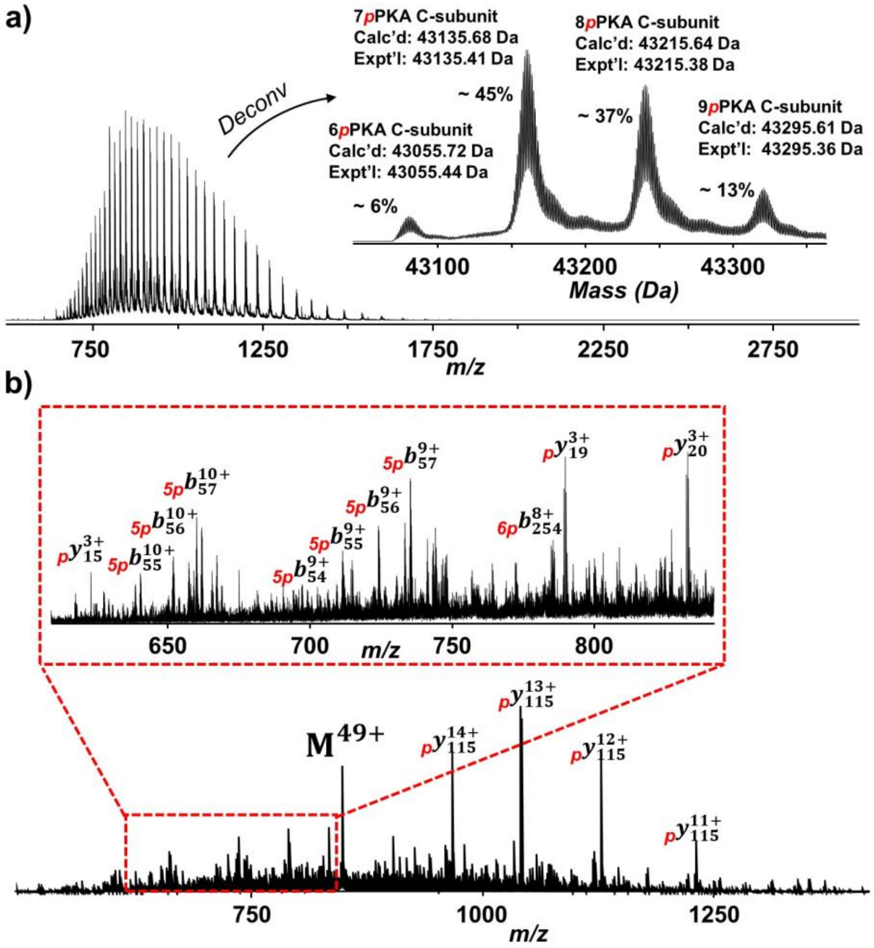


Figure 3.2. Top-down MS analysis of the PKA C-subunit. a) Mass spectra with charge state distribution and deconvoluted (inset) spectra of the PKA C-subunit. The most abundant proteoform purified from *E. coli* expression contained seven phosphorylations with removal of N-terminal methionine. b) The precursor at charge state 49^+ was subjected to online LC-MS/MS with CID fragmentation. 560 – 850 *m/z* is zoomed in to show a variety of *b* and *y* fragment ions (inset).

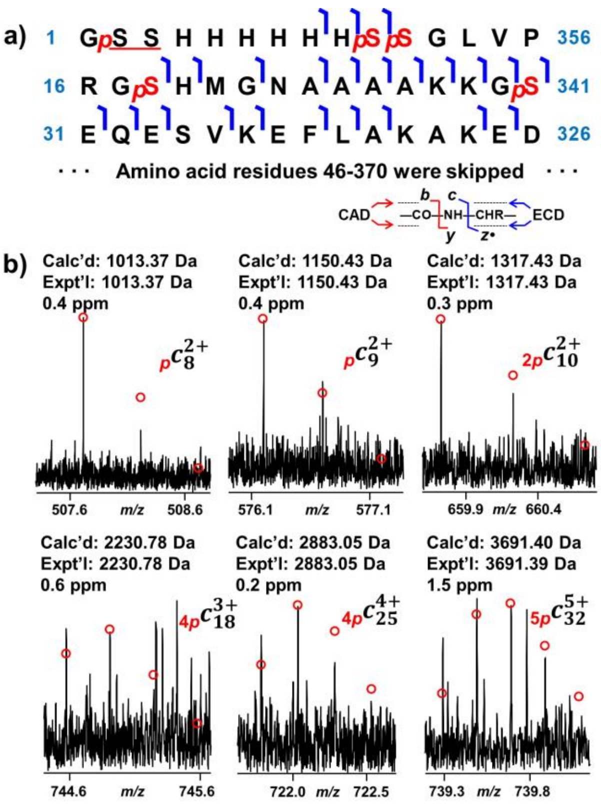


Figure 3.3. Identification of five phosphorylation sites at the N-terminal region using ECD. a) ECD fragment ion mapping for the first 45 amino acid residues. b) Representative fragment ions from ECD fragmentation. Five phosphorylation sites, Ser2/Ser3, Ser10, Ser11, Ser18, and Ser30 were confirmed by phosphorylated *c* ions.

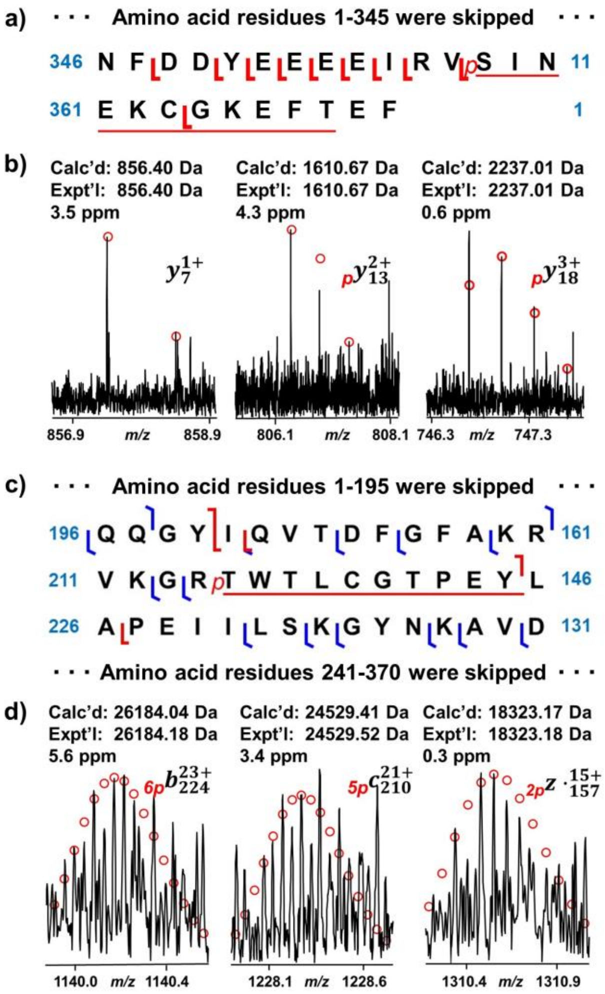
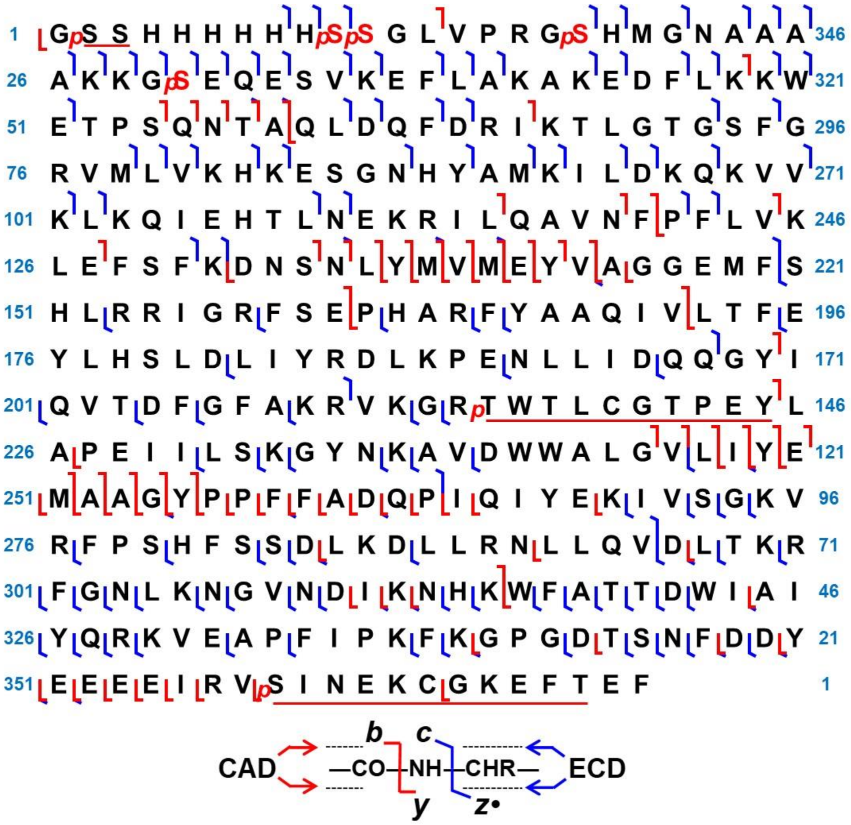


Figure 3.4. Phosphorylation site mapping at the C-terminus and in the middle region by CID and ECD. a) CID fragment ion mapping at the C-terminus. b) Representative fragment ions from CID experiment. A phosphorylation site was localized at Ser358 or Thr368. c) CID and ECD fragment ion mapping for the middle sequence of PKA C-subunit. d) Representative fragment ions from the CID and ECD experiments. A phosphorylation site was located at Thr[215-224]Tyr.



1 G pS S H H H H H H pS pS G L V P R G pS H M G N A A A 346
26 A K K G pS E Q E S V K E F L A K A K E D F L K K W 321
51 E T P S Q N T A Q L D Q F D R I K T L G T G S F G 296
76 R V M L V K H K E S G N H Y A M K I L D K Q K V V 271
101 K L K Q I E H T L N E K R I L Q A V N F P F L V K 246
126 L E F S F K D N S N L Y M V M E Y V A G G E M F S 221
151 H L R R I G R F S E P H A R F Y A A Q I V L T F E 196
176 Y L H S L D L I Y R D L K P E N L L I D Q Q G Y I 171
201 Q V T D F G F A K R V K G R pT W T L C G T P E Y L 146
226 A P E I I L S K G Y N K A V D W W A L G V L I Y E 121
251 M A A G Y P P F F A D Q P I Q I Y E K I V S G K V 96
276 R F P S H F S S D L K D L L R N L L Q V D L T K R 71
301 F G N L K N G V N D I K N H K W F A T T D W I A I 46
326 Y Q R K V E A P F I P K F K G P G D T S N F D D Y 21
351 E E E E I R V pS I N E K C G K E F T E F 1
CAD
b
c
—CO—NH—CHR—
y
z•
ECD

Figure 3.5. Top-down MS/MS sequencing of the PKA C-subunit. The map combined three ECD spectra and five CAD spectra. 191 of 369 possible bonds were cleaved, providing a 52% sequence coverage.

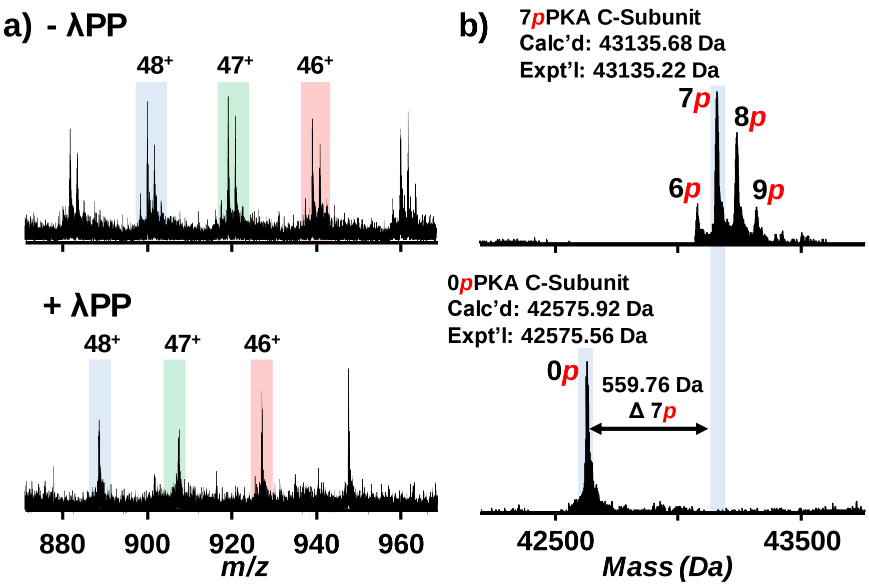


Figure 3.6. Analysis of the dephosphorylation of the PKA C-subunit using λPP. a) Charge states 48^+, 47^+, and 46^+ were shown for PKA C-subunit before (top) and after (bottom) the dephosphorylation reaction. b) Deconvoluted spectra were shown for the analysis of the dephosphorylation reaction. All phosphorylations were removed, resulting in a single unphosphorylated proteoform in the spectra.

Supplemental Information

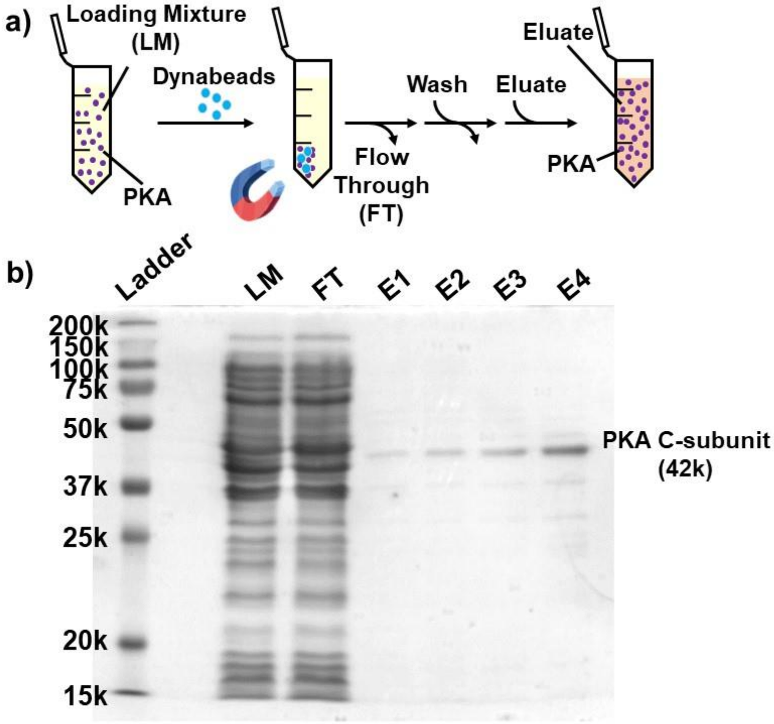


Figure S3.1. Schematic drawing and SDS-PAGE analysis of affinity purification for PKA C-subunit. a) Schematic of the affinity purification of PKA C-subunit using DynaBead for His-tag purification; b) SDS-PAGE analysis of purified PKA catalytic subunit expressed in *E. coli*.

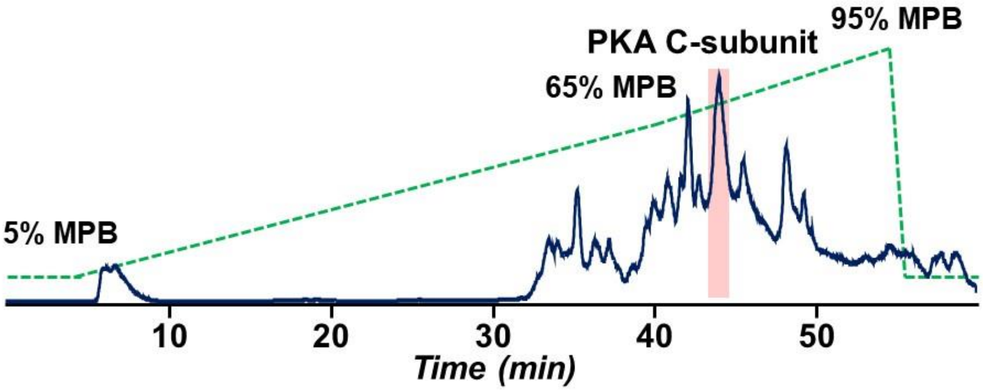


Figure S3.2. Total ion chromatogram of recombinant PKA C-subunit using liquid chromatography. The LC condition was run with H_2O with 0.1% formic acid (FA) as mobile phase A and 50:50 EtOH:ACN with 0.1% FA as mobile phase B (MPB). The gradient ran at 5% MPB for 5 min, followed by 5% to 65% MPB from 5 to 40 min, 65% to 95% MPB from 40-53 min, and back to 5% MPB. Other peaks in the chromatogram are either low mass proteins or small molecule contamination.

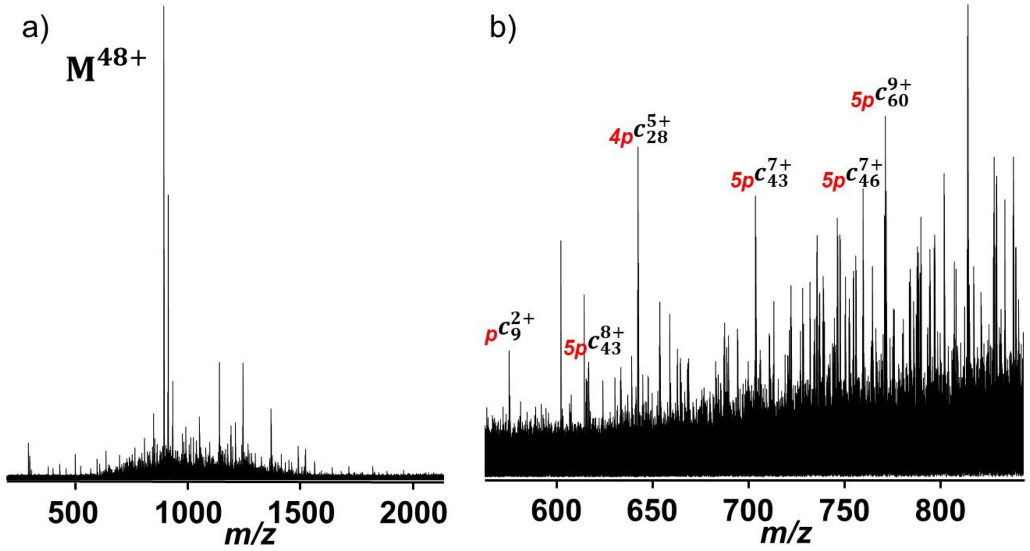


Figure S3.3. Raw spectra of ECD experiment. a) The precursor ion at charge state 48^{+} was subject to ECD fragmentation experiment, and a large number of fragment ions was yielded. b) Zoomed-in spectra from 625 – 850 *m/z* showing several *c* ions with phosphorylations intact.

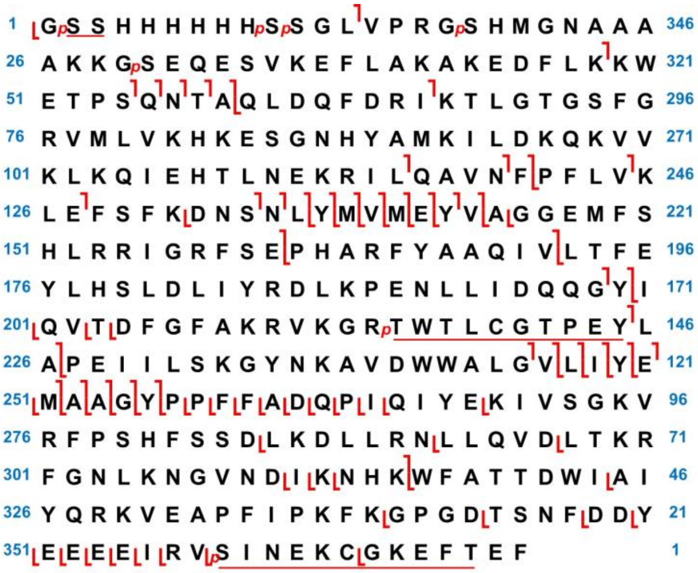

1 G p S S H H H H H H p S p S G L V P R G p S H M G N A A A 346
26 A K K G p S E Q E S V K E F L A K A K E D F L K K W 321
51 E T P S Q N T A Q L D Q F D R I K T L G T G S F G 296
76 R V M L V K H K E S G N H Y A M K I L D K Q K V V 271
101 K L K Q I E H T L N E K R I L Q A V N F P F L V K 246
126 L E F S F K D N S N L Y M V M E Y V A G G E M F S 221
151 H L R R I G R F S E P H A R F Y A A Q I V L T F E 196
176 Y L H S L D L I Y R D L K P E N L L I D Q Q G Y I 171
201 Q V T D F G F A K R V K G R p T W T L C G T P E Y L 146
226 A P E I I L S K G Y N K A V D W W A L G V L I Y E 121
251 M A A G Y P P F F A D Q P I Q I Y E K I V S G K V 96
276 R F P S H F S S D L K D L L R N L L Q V D L T K R 71
301 F G N L K N G V N D I K N H K W F A T T D W I A I 46
326 Y Q R K V E A P F I P K F K G P G D T S N F D D Y 21
351 E E E E I R V p S I N E K C G K E F T E F 1

Figure S3.4. Fragment ion mapping for CID fragmentation. A series of CID fragment ions was observed at Ser[54-58]Gln, Ser[134-145]Gly, Gly[246-265]Gln, and Tyr[350-357]Arg. *b* ions that contain a large number of phosphorylations were observed at Ser[54-58]Gln, Ser[134-145]Gly, and Gly[246-255]Tyr.

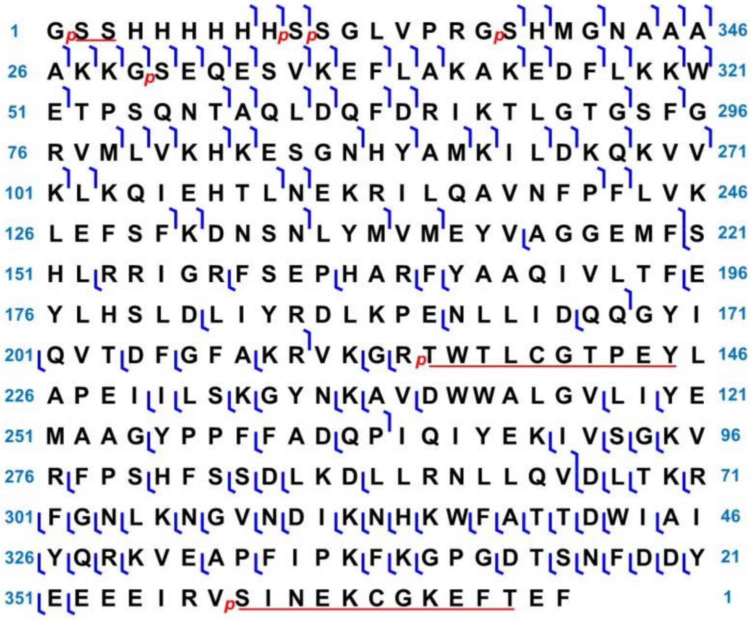

1 G S S H H H H H H S S G L V P R G S H M G N A A A 346
26 A K K G S E Q E S V K E F L A K A K E D F L K K W 321
51 E T P S Q N T A Q L D Q F D R I K T L G T G S F G 296
76 R V M L V K H K E S G N H Y A M K I L D K Q K V V 271
101 K L K Q I E H T L N E K R I L Q A V N F P F L V K 246
126 L E F S F K D N S N L Y M V M E Y V A G G E M F S 221
151 H L R R I G R F S E P H A R F Y A A Q I V L T F E 196
176 Y L H S L D L I Y R D L K P E N L L I D Q Q G Y I 171
201 Q V T D F G F A K R V K G R T W T L C G T P E Y L 146
226 A P E I I L S K G Y N K A V D W W A L G V L I Y E 121
251 M A A G Y P P F F A D Q P I Q I Y E K I V S G K V 96
276 R F P S H F S S D L K D L L R N L L Q V D L T K R 71
301 F G N L K N G V N D I K N H K W F A T T D W I A I 46
326 Y Q R K V E A P F I P K F K G P G D T S N F D D Y 21
351 E E E E I R V S I N E K C G K E F T E F 1
p

Figure S3.5. Fragment ion mapping for ECD fragmentation. Most ECD fragment ions were located at both ends of the amino acid sequence.

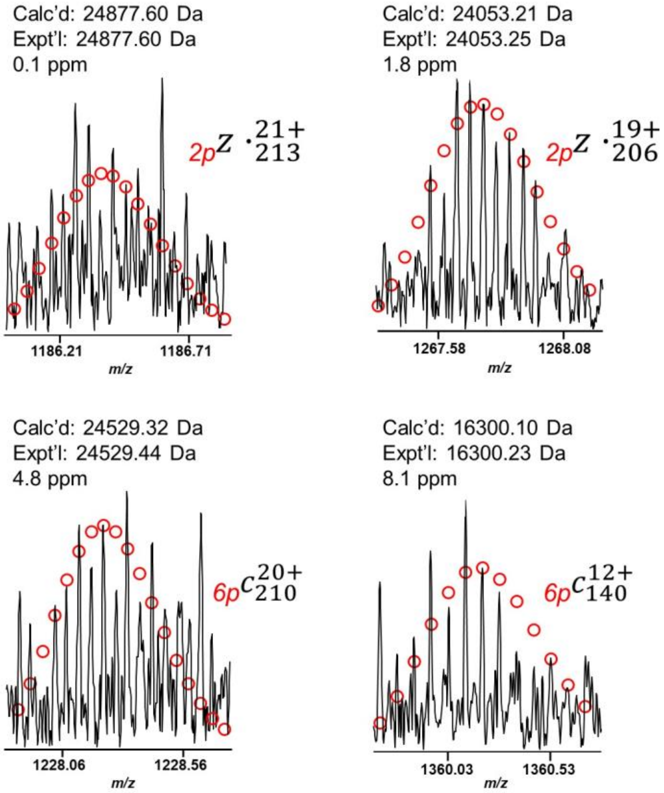


Figure S3.6. Representative fragment ions from ECD fragmentation. Four ECD fragment ions with large molecular weight were shown.

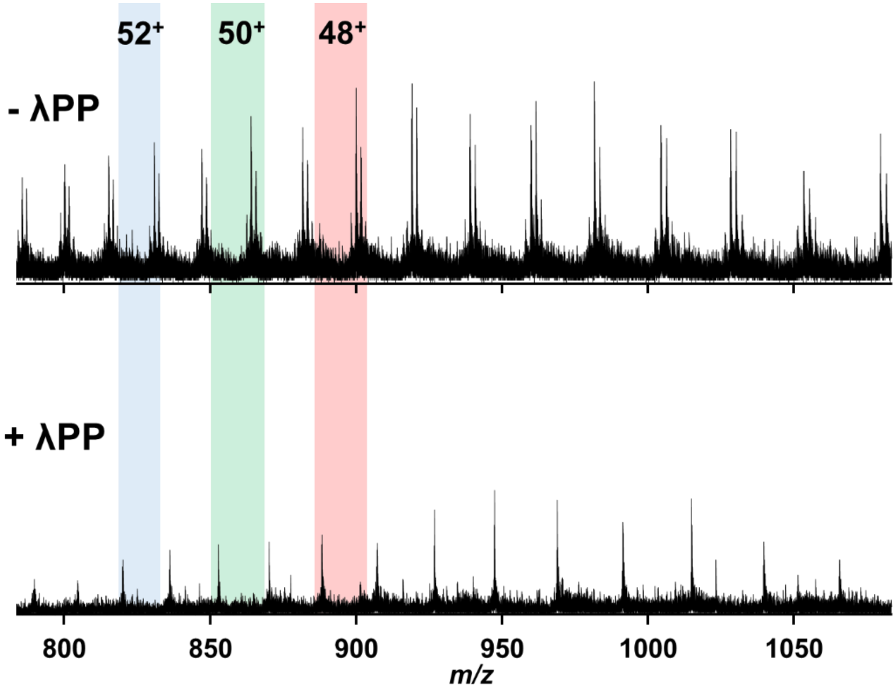


Figure S3.7. Broadband spectra for the dephosphorylation reaction. A drastic peak shift was observed for charge state 52^+, 50^+, and 48^+ due to the removal of multiple phosphorylations.

Chapter 4

Enriching Kinases by Functionalized Nanoparticles: A Pilot Study

Abstract

Reversible protein phosphorylation is crucial in cell growth, division, and signal transduction. In protein phosphorylation, protein kinases are enzymes that are responsible for transferring the phosphate group from ATP to their kinase substrates. The regulation of kinase activities is central to the regulation of protein phosphorylation, and are therefore makes kinases important therapeutic targets. Top-down MS has emerged as the method of choice to study proteoforms arising from alternative splicing, sequence variations, and post-translational modifications, and this technology is suitable for protein kinases which have a large variety. On the other hand, protein kinases are expressed in low-abundance, and thus an enrichment strategy is needed to enable top-down MS analysis. In this study, we developed a nanoproteomics platform to enrich endogenous protein kinases from samples for top-down MS analysis. Iron oxide NPs were functionalized with a pan-kinase inhibitor, which can capture a wide range of kinases. In the simple standard protein system, the functionalized NPs could effectively capture target kinases with some degrees of non-specific binding. For the kinase enrichment of complex system, an experimental workflow was established that bridged the sample preparation to downstream MS analysis. Further development in this nanoproteomics strategy could be coupled with top-down MS to allow for in-depth characterization of protein kinases.

Introduction

Reversible protein phosphorylation, which is constituted by protein phosphorylation and protein dephosphorylation is crucial to transduce external stimuli into intracellular signals in order to achieve activation, and inhibit or reverse the phosphorylation events of biological processes.[1] These phosphorylation events are important for cell cycle control, receptor-mediated signal transduction, cell differentiation and proliferation, and metabolism.[2-3] Protein kinases are enzymes that are responsible for protein phosphorylation process by transporting phosphate groups to their substrates. The set of protein kinases of an organism constitutes a specific kinome.[4] The human genome encodes for about 500 protein kinases constituting around 1.7% of the full human genome, with two major types such as tyrosine kinases and serine-threonine kinases.[5] Diseases are often stemmed from abnormal regulation of kinase signaling. For instance, cancers are observed to have aberrant kinase activity due to genetic or somatic mutations, whereas cardiovascular diseases have elevated neurohormonal systems due to enhanced stimulation of kinases.[6-7] As a result, protein kinases have emerged as major drug targets for therapeutic purposes.[8] In order to advance the understanding of the signaling networks orchestrated by protein kinases, it is important to comprehensively assess their functions by studying their own molecular events for kinase activation and inhibition, and their interactions with other key players in the signaling pathways.

The continuing development of MS with increasingly powerful instruments and novel fragmentation techniques has empowered MS-based proteomics to become the method of choice for protein identification, characterization, and quantification.[9] In particular, Top-down proteomics is a rapidly developing technique that is capable of identifying proteoforms arising from PTMs, alternative splicing and sequence variations in the intact protein level.[10-14] While the use of top-down proteomics is promising, it still faces tremendous technological challenges. This technology is particularly attractive for studying protein kinases, as these enzymes are expressed in large varieties to accurately modulate signaling events.[15] However, protein kinases are also expressed in low copies per cell.[15-16] Recent studies by Worboys *et al* has shown that the abundance of different peptides from the digestion of kinases has three order of magnitude difference in proteome dynamic range.[17] In the top-down proteomic analysis, low abundance proteins are often either overlooked by the detector because of suppression from high abundance proteins, or lack of identification due to fragmentation methods that often select several high abundance ions in the chromatogram window. This necessitates the development for kinase enrichment methods to increase the relative abundance of kinases in the sample top-down MS analysis.

Kinase inhibitors are a class of small molecules that have affinity to protein kinases.[18-19] While kinase inhibitors are developed with different modes of inhibitory actions, kinase inhibitors that are ATP-competitive should be ideal to perform kinase enrichment, which allows proteins to attach and detach from the small molecules. Several ATP-competitive kinase inhibitors are also pan-kinase inhibitors, which can capture a range of proteins.[20-22] While conventionally affinity reagents are coupled with polysaccharide beads-based platform (i.e. Sepharose) which are µm scale, recent studies has suggested that nanoparticle which are nanometer in scale may be an attractive alternative.[23-25] Development in material chemistry has enabled nanoparticles to be surface-functionalized with a variety of terminal functional groups, and have superparamagnetic properties.[26] Additionally, nanoparticles have small diameters that may facilitate diffusion through protein mixtures. As a result, nanoparticles functionalized with pan-kinase inhibitors can be an ideal platform for kinase enrichment.

In this study, we have developed a nanoproteomics platform that utilizes functionalized NPs to capture kinases for downstream proteomics. NPs were functionalized with a pan-kinase inhibitor which have affinity to a wide range of kinases. The functionalized NPs were characterized by FTIR spectroscopy. Both a simple standard protein system and a complex protein extract system were used to evaluate the performance by using the functionalized NPs. We have found success in a simple standard protein system, where the functionalized NPs can selectively capture target kinases with some degrees of non-specific binding. On the other hand, while an experimental workflow was developed for MS analysis, the functionalized NPs have not been able to capture kinases selectively and effectively.

Experimental Section

Chemicals and Supplies

All reagents were acquired from Sigma-Aldrich, Inc. (St. Louis, MO, USA) and TCI America (Portland, OR, USA), unless otherwise noted. Solvents, including HPLC grade H_2O, ACN, EtOH, and acetone, were purchased from Fisher Scientific (Fair Lawn, NJ, USA). (3-aminopropyl)triethoxysilane (APTES) was acquired from Gelest (Morrisville, PA, USA). Halt™ Protease inhibitor cocktail were purchased from ThermoFisher Scientific (Rockford, IL, USA). Protein kinase C isoform alpha (PKCα) and glycogen synthase kinase 3 isoform beta (GSK3β) were purchased from MilliporeSigma (Burlington, MA, USA). Bovine Serum Albumin (BSA) was acquired from Fisher Scientific. β-casein and carbonic anhydrase were purchased from Sigma-Aldrich, Inc.

Nuclear Magnetic Resonance and Small Molecule MS

All nuclear magnetic resonance (NMR) spectra were acquired on the Bruker Avance-500 (500 MHz, Bruker, Bremen, Germany) with the DCH cryoprobe. 16 to 32 scans were accumulated for ^{1}H spectra whereas 512 to 1024 scans were accumulated for ^{13}C spectra.

For MS analysis, small molecule samples were dissolved in 50:50:0.1 ACN:H_2O:FA. Samples were delivered to a solariX XR 12-Tesla Fourier Transform Ion Cyclotron Resonance mass spectrometer (FTICR-MS, Bruker Daltonics, Bremen, Germany) using a TriVersa Nanomate system (Advion BioSciences, Ithaca, NY, USA). Mass spectra were acquired with an acquisition size of 16M, in the mass range between 150-2000 *m/z* (with a resolution of 270,000 at 400 m/z), and around 50 scans were accumulated for each sample.

Synthesis of tert-butyl (3-bromopropyl)carbamate

3-bromopropylamine hydrobromide (4.00 g, 18.3 mmol, 1 eq.) and di-tert-butyl dicarbonate (7.99 g, 36.6 mmol, 8.4 mL, 2 eq.) were dissolved in anhydrous DCM (100 mL) under nitrogen gas flow. To the stirred solution was added DIPEA (2.61 g, 20.2 mmol, 3.5 mL, 1.1 eq.). After 12 hours the solvent was removed *in vacuo*. The residue was dissolved in ethanol (10 mL). Imidazole (1.07 g, 18.7 mmol, 1.02 eq.) was added and stirred for 30 minutes. The mixture was diluted with chloroform (100 mL) and washed with 1% HCl solution (3 x 50 mL). The organic phase was dried with sodium sulfate and evaporated to afford the product as a yellow oil (4.39 g, 17.4 mmol, 95%). ^{1}H NMR (500 MHz, DMSO-d_6): δ 6.01 (s, 1H), 3.50 (t, J = 6.5 Hz, 2H), 3.03 (q, J = 6.5 Hz, 2H), 1.90 (p, J = 6.5 Hz, 2H), 1.37 (s, 9H). ^{13}C NMR (126 MHz, DMSO-d_6): δ 156.08, 78.07, 38.94, 33.12, 32.78, 28.70. HRMS (*m/z*) calc'd for $C_8H_{16}BrNO_2$ $[M+H]^+$ 238.0437, found 238.0438.

Synthesis of tert-butyl (3-(3-(2-amino-2-oxoethyl)-1H-indol-1-yl)propyl)carbamate (1)

Indole-3-acetamide (1.00 g, 5.7 mmol, 1 eq.) was dissolved in anhydrous DMF (8.76 mL). To the stirring solution cooled in an ice bath, sodium hydride (0.3421 g, 60% in mineral oil, 8.55 mmol, 1.5 eq.) was added slowly and the mixture was allowed to stir at room temperature for 60 mins under nitrogen. Tert-butyl (3-bromopropyl)carbamate (2.28 g, 9.57 mmol, 1.7 eq.), dissolved in anhydrous DMF (4.38 mL), was added to the mixture dropwise. The reaction was allowed to proceed at room temperature overnight under nitrogen. The crude reaction mixture was diluted with ethyl acetate (30 mL) and the organic layer was washed with water (3 x 20 mL). The organic layer was dried by $MgSO_4$ and filtered. The volatile was removed under vacuum to yield a crude yellow oil. The product was purified by flash column chromatography to afford **1** (1.16g, 3.5 mmol, 61%) as a white solid using a gradient from 50% to 100% acetone in hexane. ^{1}H NMR (500 MHz, DMSO-d_6): δ 6.91 (s, 1H), 3.50 (t, J = 6.5 Hz, 2H), 3.03 (q, J = , 2H), 1.90 (p, J = 6.5 Hz, 2H), 1.37 (s, 9H). ^{13}C NMR (126 MHz, DMSO-d_6): δ156.08, 78.07, 38.94, 33.12, 32.78, 28.70. HRMS (*m/z*) calc'd for $C_{18}H_{25}N_3O_3$ $[M+H]^+$ 332.1969, found 332.1968.

Synthesis of methyl 2-(1-methyl-1H-indol-3-yl)-2-oxoacetate (2)

1-methylindole (2.00 g, 15.3 mmol, 1.90 mL, 1 equiv) was dissolved in diethyl ether (20 mL) and transferred to a 100 mL round bottom flask on an ice bath (0 °C). To the stirring solution, oxalyl chloride (1.94 g, 15.3 mmol, 1.31 mL, 1 eq.) was added dropwise, and the solution was left to stir under nitrogen at 0 °C for 30 mins. The flask was then transferred to a dry ice and acetone bath (-78 °C), to which sodium methoxide (6.62 g, 30.6 mmol, 7.0 mL at 25 wt. %, 2 eq.) was added dropwise. The solution was then promptly removed from the bath and left to slowly warm to room temperature for 60 minutes. Water (10 mL) was then added to quench the reaction, and the solid precipitate was dried and washed (first with cold water, and then with cold ethyl acetate) via vacuum filtration. The solid was then further dried to afford **2** (2.11g, 9.72 mmol, 63.7%) as a light pink solid. ^{1}H NMR (500 MHz, DMSO-d_6) δ 8.50 (s, 1H), 8.18 (d, J = 7.5 Hz, 4H), 7.62 (d, J = 8.0 Hz, 1H), 7.35 (dtd, J = 21.8, 7.3, 1.1 Hz, 3H), 3.92 (s, 3H), 3.89 (s, 3H). ^{13}C NMR (126 MHz, DMSO-d_6) δ 178.52, 164.37, 142.10, 137.91, 126.38, 124.36, 123.74, 121.72, 111.76, 111.70, 53.01, 33.98. HRMS (*m/z*) calc'd for $C_{12}H_{11}NO_3$ $[M+H]^+$ 218.0812, found 218.0812, $[M+Na]^+$, found 240.0631.

Synthesis of tert-butyl (3-(3-(4-(1-methyl-1H-indol-3-yl)-2,5-dioxo-2,5-dihydro-1H-pyrrol-3-yl)-1H-indol-1-yl)propyl)carbamate (3)

1 (0.85 g, 2.6 mmol, 1 eq.) and **2** (1.11 g, 5.13 mmol, 2 eq.) were dissolved in anhydrous THF (6 mL). To this stirring solution cooling on an ice bath (0 °C), potassium tert-butoxide (1.0M in THF, 7.70 mL, 3 eq.) was slowly added dropwise. The solution was then taken off of the ice bath after 60 seconds and allowed to gradually warm to room temperature for 2 hours under nitrogen. Subsequently, potassium tert-butoxide (1.0 M in THF, 12.83 mL, 5 eq.) was added dropwise, and the solution was left to stir for another 30 mins at room temperature. The volatile was then removed *in vacuo* to leave a crude material, which was then dissolved in ethyl acetate (175 mL) and washed with water (3 x 175 mL). For the second wash, brine (3 mL) was added to hasten separation and reduce the emulsion formed. The organic layer was then dried with magnesium sulfate and filtered. The volatile was removed to yield a crude, bright red oil which was purified by flash column chromatography. After the product band (Rf of 0.8, 50% acetone in hexane) and the impurity band (Rf of 0.7, 50% acetone in hexane) were visibly separated, the eluting solvent was changed directly from 50% to 70% acetone in hexane to afford **3** (0.77g, 1.54 mmol, 60.0%). ^{1}H NMR (500 MHz, DMSO-d$_6$) δ 10.93 (s, 1H), 7.86 (s, 1H), 7.77 (s, 1H), 7.46 (d, J = 8.2 Hz, 1H), 7.42 (d, J = 8.2 Hz, 1H), 7.07 – 6.96 (m, 3H), 6.90 (d, J = 8.0 Hz, 1H), 6.75 – 6.59 (m, 4H), 4.23 (t, J = 6.9 Hz, 2H), 3.86 (s, 3H), 2.91 (q, J = 6.5 Hz, 2H), 1.84 (p, J = 6.8 Hz, 2H), 1.39 (s, 9H). ^{13}C NMR (126 MHz, DMSO-d$_6$) δ 206.49, 172.92, 172.89, 155.61, 136.49, 135.57, 133.11, 131.97, 127.56, 126.72, 126.19, 125.57, 121.66, 121.63, 121.18, 121.15, 119.54, 119.49, 110.10, 110.05, 105.00, 104.57, 77.59, 43.46, 39.52, 37.30, 32.88, 30.71, 29.87, 28.26. HRMS (*m/z*) calc'd for $C_{29}H_{30}N_4O_4$ [M+H]$^+$ 499.2340, found 499.2340.

Synthesis of (4, Bisindolylmaleimide VIII Hydrochloride)

4.0 M HCl in dioxane (4 mL) was added to **3** (.309 g, 0.62mmol). This solution was left to stir at room temperature for 60 minutes. The volatile was then removed *in vacuo*, and methanol (1 mL) was added to solvate the remaining solid. This solution was then added to cold diethyl ether (45 mL) in a centrifuge tube. An additional aliquot of methanol (1 mL) was added to wash the remainder of the solid. The result was a bright red precipitate, which was then placed in a centrifuge (5 mins, 10 °C). The supernatant was poured off and disposed of, and the remaining solid was then dried in a desiccator under vacuum to yield **4** (.247g, 0.62 mmol, 100%). ^{1}H NMR (500 MHz, DMSO-d$_6$) δ 10.95 (s, 1H), 7.91 (s, 3H), 7.87 (s, 1H), 7.82 (s, 1H), 7.53 (d, J = 8.2 Hz, 1H), 7.42 (d, J = 8.2 Hz, 1H), 7.09 – 6.99 (m, 2H), 6.85 (d, J = 8.0 Hz, 1H), 6.73 – 6.57 (m, 3H), 4.36 (t, J = 6.9 Hz, 2H), 3.87 (s, 3H), 2.74 (q, J = 7.4, 6.5 Hz, 2H), 2.05 (p, J = 7.1 Hz, 2H). ^{13}C NMR (126 MHz, DMSO-d$_6$) δ 172.88, 172.88, 136.48, 135.49, 133.16, 131.64, 127.76, 126.46, 126.23, 125.68, 121.79, 121.67, 121.18, 120.96, 119.64, 119.55, 110.19, 110.14, 105.29, 104.55, 66.35, 42.98, 39.52, 36.43, 32.92, 27.80. HRMS (*m/z*) calc'd for $C_{24}H_{22}N_4O_2$ $[M+H]^+$ 399.1816, found 399.1819.

Synthesis of (S)-5-((1-amino-1-oxo-3-(tritylthio)propan-2-yl)amino)-5-oxopentanoic acid

The starting materials, S-trityl-L-cysteinamide (0.5 g, 1.38 mmol, 1 eq.) and glutaric anhydride (0.173 g, 1.52 mmol, 1.1 eq) were transferred to a 100 mL round bottom flask. DCM (50 mL) was added to dissolve all the solids. DIPEA (0.357 g, 2.76 mmol, 480 µL, 2 eq.) was added, and the reaction was allowed to proceed at reflux overnight. Upon completion, the volatile was removed *in vacuo* to afford a yellow solid. The solid was suspended in hexane and filtered by vacuum filtration to obtain the final product. The solid was used for further synthesis without additional purification.

Synthesis of (S)-N1-(1-amino-1-oxo-3-(tritylthio)propan-2-yl)-N5-(3-(3-(4-(1-methyl-1H-indol-3-yl)-2,5-dioxo-2,5-dihydro-1H-pyrrol-3-yl)-1H-indol-1-yl)propyl)glutaramide (5)

(S)-5-((1-amino-1-oxo-3-(tritylthio)propan-2-yl)amino)-5-oxopentanoic acid (65.7 mg, 0.138 mmol, 1.2 eq.) was incubated with HATU (55.1 mg, 0.145 mmol, 1.26 eq.) and DIPEA (26.5 mg, 0.276 mmol, 35.7 uL, 2.4 eq.) in 3 mL anhydrous DMF for 30 mins. The color changed from light yellow to slightly pink during the process. In a separate vial, **4** (50 mg, 0.115 mmol, 1 eq.) was dissolved in 3 mL anhydrous DMF, and DIPEA (13.3 mg, 0.138 mmol, 17.8 µL, 1.2 eq.). After 30 mins, the two mixtures were combined and the reaction was allowed to proceed overnight. Upon completion, the mixture was diluted with 25 mL EtOAc. The organic layer was washed with 25 mL of saturated $NaHCO_3$ solution. The aqueous layer was extracted with 25 mL EtOAc. The organic layers were combined and dried with brine and Na_2SO_4. The volatile was removed *in vacuo*. The crude mixture was diluted with 50:50 acetone:hexane. The product was purified using a gradient from 50:50 to 100:0 acetone:hexane. ^1H NMR (500 MHz, DMSO-d_6) δ 10.91 (s, 1H), 8.00 (d, J = 8.0 Hz, 1H), 7.85 (s, 2H), 7.77 (s, 1H), 7.46 (d, J = 8.0 Hz, 1H), 7.41 (d, J = 8.0 Hz, 1H), 7.35 – 7.26 (m, 15H), 7.23 (t, J = 7.1 Hz, 3H), 7.07 (s, 1H), 7.02 (q, J = 7.5 Hz, 2H), 6,89 (d, J = 8.3 Hz, 1H), 6.71 – 6.65 (m, 2H), 6.62 (t, J = 6.8 Hz, 1H), 4.32 (m, 1H), 4.23 (t, J = 6.9 Hz, 2H), 3.86 (s, 3H), 3.30 (s, 2H), 3.02 (q, J = 6.5 Hz, 2H), 2.32 (m, 2H), 2.14 (m, 3H), 1.85 (p, J = 6.9 Hz, 2H), 1.73 (p, J = 7.6 Hz, 2H).

Synthesis of (S)-N1-(1-amino-3-mercapto-1-oxopropan-2-yl)-N5-(3-(3-(4-(1-methyl-1H-indol-3-yl)-2,5-dioxo-2,5-dihydro-1H-pyrrol-3-yl)-1H-indol-1-yl)propyl)glutaramide (6, BIM-GA-Cyst)

The starting material, **5** (69 mg, 0.112 mmol) was dissolved in anhydrous DCM (5.7 mL). TIPS (150 µL, 5%) was added, followed by additional of TFA (150 µL, 5%). The reaction was allowed to proceed for an hour. The volatile was removed *in vacuo*. The residue was dissolved in minimal amount of MeOH, and transferred to a 50 mL Falcon tube. Et_2O was added to fill the rest of the volume. Precipitate was observed immediately. The red product was collected by centrifugation at 5,000 rpm for 10 mins. The supernatant was discarded to afford the red product which was dried under vacuum. ^1H NMR (500 MHz, DMSO-d_6) δ 10.91 (s, 1H), 7.95 (d, J = 8.2 Hz, 1H), 7.89 - 7.85 (m, 2H), 7.78 (s, 1H), 7.47 (d, J = 7.8 Hz, 1H), 7.41 (d, J = 8.6 Hz, 1H), 7.13 (s, 1H), 7.03 (q, J = 7.5 Hz, 2H), 6,89 (d, J = 7.9 Hz, 1H), 6.72 – 6.60 (m, 3H), 4.34 – 4.30 (m, 1H), 4.24 (t, J = 7.1 Hz, 2H), 3.86 (s, 3H), 3.18 (s, 1H), 3.02 (q, J = 4.8 Hz, 2H), 2.82 – 2.63 (m, 2H), 2.23 (t, J = 8.8 Hz, 1H), 2.18 (t, J = 7.3 Hz, 2H), 2.10 (t, J = 6.9 Hz, 2H), 1.86 (p, J = 6.9 Hz, 2H), 1.73 (p, J = 7.3 Hz, 2H). HRMS (*m/z*) calc'd for $C_{32}H_{34}N_6O_5S$ $[M+H]^+$ 615.2384, found 615.2396.

Synthesis of 3-butynoic acid

3-Butynoic acid was synthesized following a previously reported procedure,[27] with modification. To a 1 L round bottom flask, 250 mL of H_2O was added, followed by 120 µL of 70% HNO_3, $Na_2Cr_2O_7$, and $NaIO_4$ on an ice bath. The mixture was stirred for 15 mins. 3-butyn-1-ol was slowly added. The reaction was allowed to proceed overnight under nitrogen on ice and warming up to room temperature. Upon completion, the slurry was filtered by gravity filtration. The aqueous layer was extracted five times each with 80 mL Et_2O. The combined organic layer was concentrated to a brown oil. The oil was diluted with 20 mL DCM and concentrated. The procedure with additional DCM wash until the color of the product lightened to afford a yellow powder. ^1H NMR (500 MHz, $CDCl_3$-d) δ 3.38 (d, 2H, J = 2.7 Hz), 2.25 (t, 1H, J = 2.7 Hz).

Synthesis of N-(3-(triethoxysilyl)propyl)buta-2,3-dienamide (BAPTES).

To a 500 mL three-neck round bottom flask, 3-butynoic acid, 2-chloro-1-methylpyridinium iodide were dissolved in DCM (130 mL). The mixture was heated to 60 °C to reflux and stir with a condenser under nitrogen for 20 mins. In a separate flask, APTES and DIPEA (2 eq) were diluted with 130 mL DCM. The mixture was slowly transferred to the refluxing reaction using a syringe. After 30 mins, additional DIPEA (2 eq) was added to induce isomerization. The mixture was allowed to reflux overnight. Upon completion, the volatile was removed. The residual mixture was diluted with EtOAc and centrifuged at 5000 rpm for 5 mins to remove the precipitate. The supernatant was concentrated via rotary evaporation. The product was purified with flash column chromatography using a gradient of 50:50 to 100:0 ethyl acetate: hexane to yield BAPTES as a clear, orange oil (4.80 g, 67% yield). ^{1}H NMR (500 MHz, $CDCl_3$-d) δ 6.00 (s, 1H), 5.62 (t, J = 6.6 Hz, 1H), 5.20 (d, J = 6.7 Hz, 2H), 3.82 (q, J = 7.0 Hz, 6H), 3.30 (q, J = 6.7 Hz, 2H), 1.65 (m, 2H), 1.23 (t, J = 7.0 Hz, 9H), 0.65 (m, 2H).

Synthesis of Iron Oleate precursor

Sodium oleate (18.267 g, 60 mmol, 1 eq.) and iron(III) chloride hexahydrate (5.406 g, 20 mmol, 0.33 eq.) were weighed and transferred to a 500 mL round bottom flask. Nanopure H_2O (30mL), EtOH (40mL), and hexane (70mL) were added to the round bottom flask while stirring with a Teflon-coated magnetic egg-shaped stir-bar (1-1/4" x 5/8"). After all of the reagents were dissolved, the round bottom flask was attached to a condenser and put on vacuum for 5 minutes. The flask was refilled with N_2 gas and the reaction was allowed to proceed at 72 °C overnight. The product, $Fe(oleate)_3$, was then removed from heat. The organic layer was separated from the aqueous layer using a separatory funnel. The organic layer was washed with nanopure H_2O (15mL) three times. The excess hexane in $Fe(oleate)_3$ was remove *in vacuo*, and the $Fe(oleate)_3$ was stored under vacuum until use.

Synthesis of 8 nm Fe_3O_4-Oleate NP

In this order, $Fe(oleate)_3$ (10 mmol, 9.00 g, 1 eq.), oleic acid (5.5 mmol, 1.56 g, 0.55 eq.), tetradecein (50.4 mmol, 10.00 g, 5.04 eq.), and octadecein (158.4 mmol, 40.00 g, 12.84 eq.) were weighed and transferred to a 125 mL three-neck round bottom flask. The side openings were sealed by rubber stoppers. The flask was attached to vacuum and replaced with nitrogen. The reaction was allowed to proceed at 102 °C for 4 hours, and heated to 300° C over the course of two hours using a temperature controller. The ramp rate was set to 3.3° C, and temperature was held at 300° C for 30 minutes. Upon completion, the flask was cooled. For each 5 mL of Fe_3O_4-Oleate in high boiling solvent transferred to a 50 mL Falcon tube, EtOH was used to fill up to 50 mL. The solutions were sonicated and centrifuged at 5,000 rpm for 5 minutes. The supernatant was discarded, and the remaining particles were washed 3 times, each with 30 mL of EtOH. The resulting Fe_3O_4 was collected in minimal hexanes and left on vacuum to dry. The dried Fe_3O_4 was resuspended in anhydrous hexanes at a concentration of 20 mg/mL.

Synthesis of Fe_3O_4-BAPTES NP

Fe_3O_4-Oleate (80 mg) was added to 200 mL of hexane in a round-bottom flask (500 mL) to make a final concentration of 0.4 mg/mL. The mixture was stirred at 600 rpm and heated at 60° C. BAPTES was added drop-wise to the solution to make a final concentration of 0.55% (v/v) of trialkoxysilane followed by addition of acetic acid catalyst (20 uL) to make a final concentration of 0.01% (v/v) acetic acid catalyst. The capped reaction with minimal exposure to air and water was allowed to run for 24 hours. The resulting product was separated from solution through centrifugation at 5,000 rpm for 5 minutes at 10 °C. The subsequent precipitate was washed once with hexane, once with a combination of ACN, and washed once again with hexane. The final product was suspended in ACN and dried overnight to afford Fe_3O_4-BAPTES (50 mg).

Synthesis of BIM-GA-Cyst-Fe_3O_4-BAPTES NP

BIM-GA-Cyst (10 mg) was dissolved in 4 mL of 1:1 ACN:H_2O. 200 µL of 1M $(NH_3)_2CO_3$ solution was added to the BIM-GA-Cyst suspension to adjust the pH to about 8. In a separate vial, Fe_3O_4-BAPTES nanoparticle was suspended with 1 mL of ACN. The nanoparticle solution was added to the BIM-GA-Cyst solution while sonicating. The mixture was sonicated for 3 hours. The resulting nanoparticles was separated through centrifugation (13,200 rpm, 3 minutes). The nanoparticles was washed with DMSO at least six times, with a final wash with 1:1 ACN:H_2O. BIM-GA-Cyst-Fe_3O_4-BAPTES NP were subsequently isolated magnetically with a DynaMag and resuspended in 1:1 ACN:H_2O at 5 mg/mL and stored at 4 °C in the absence of light.

Transmission Fourier Transform Infrared (FTIR) Material Characterization

The samples were prepared by mixing the NP with potassium bromide to make a pellet. The sample mass loading was around 0.33 wt.%. Transmission Fourier transform infrared (FTIR) spectra were measured on a Bruker Equinox 55 FT-IR spectrometer (Bruker Optik GmbH, Ettlingen, Germany). The measurement was recorded from 4,000 cm^{-1} to 400 cm^{-1} at 2 cm^{-1} resolution.

Simple system kinase enrichment

For simple system kinase enrichment, 2 µg of PKAα and GSK3β along with 5 µg of BSA, β-casein, and carbonic anhydrase were used. The protein mixture was diluted with low-salt buffer (50 mM Tris, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1 mM PMSF, 10 mM L-methionine, 1x HALT protease inhibitor cocktail, and 1x phosphatase inhibitor cocktail A, pH 7.4). Prior to nanoparticle incubation, the protein mixture was adjusted to 1 M NaCl, 0.1% DDM, 0.02 mg/ml

DAG, and 0.1 mg/ml PS for final concentration. The protein mixture was incubated with the functionalized NPs for an hour. Upon completion, the NPs were centrifuged at 15,000 g for 5 mins at 4 °C, and collected using a DynaMag. The NPs were washed twice with high-salt buffer (50 mM Tris, 1 M NaCl, 1 mM EDTA, 1 mM EGTA, 1 mM PMSF, 10 mM L-methionine, 1x HALT protease inhibitor cocktail, and 1x phosphatase inhibitor cocktail A, pH 7.4) with 0.1% DDM, once with high-salt buffer, and once with low-salt buffer. Bound proteins were eluted with 1x Laemmli sample buffer. All fractions were concentrated using a 10k MWCO filter. The loading mixture and flow through fractions were normalized before loading on the 12.5% SDS-PAGE gel.

Complex system kinase enrichment

The protein extraction procedures were performed in a cold room (4 °C) using freshly prepared buffers. The human cardiac tissue was acquired from the University of Wisconsin Hospital and Clinic with the procedure approved by the Institutional Review Board of the University of Wisconsin-Madison. For large scale extraction, around 500 mg of tissue was homogenized in HEPES buffer (25 mM HEPES, 50 mM NaF, 5 mM DTT, 10 mM L-methonine, 5 mM EDTA, 1 mM PMSF, 1x HALT protease inhibitor cocktail, and 1x phosphatase inhibitor cocktail A, pH 7.4) using a Polytron electric homogenizer (Model PRO200; PRO Scientific, Oxford, CT, USA) on ice. The resulting homogenate was centrifuged at 10,000 g for 10 mins at 4 °C. After centrifugation, the supernatant was saved as the first HEPES protein extract. The same extraction procedure was repeated to obtain the second HEPES extract protein.

For kinase enrichment from complex system, 500 µg of protein extract was used for incubation with 1 mg of functionalized NPs. The incubation mixture was adjusted to 1 M NaCl and 0.1% DDM concentration, and the incubation was allowed to proceed at 4 °C for an hour.

Upon completion, the NPs were centrifuged at 15,000 g for 5 mins at 4 °C, and collected using a DynaMag. The NPs were washed twice with high-salt buffer with 0.1% DDM, once with high-salt buffer, and once with low-salt buffer. Bound proteins were eluted with 0.5% Azo in 65.8 mM Tris, 1% Azo in 65.8 mM Tris with 355 mM 2-mercaptoethanol, and finally with 1x Laemmli sample buffer. All fractions were concentrated using a 10k MWCO filter. The loading mixture and flow through fractions were normalized before loading on the 10% SDS-PAGE gel.

Mass Spectrometry Analysis

For bottom-up proteomics analysis, the samples were irradiated for 1-5 mins with UV lamp. 38 µL of the samples were first reduced by 2.5 µL of 100 mM dithiothreitol solution, and incubated at 37 °C for 30 mins. 7.5 µL of 100 mM iodoacetamide solution was added and incubated at room temperature under dark for 30 mins. 1 µL of 1 µg/µL trypsin was added and incubated for 16 hours at 37 °C. Upon completion, the samples were centrifuged at 15,000 g for 5 mins, and subject to LC-MS/MS analysis with a home-packed C18 column. The samples were separated using a nanoACQUITY UPLC System (Waters Corporation, Milford, MA, USA) coupled with an Impact II Q-TOF mass spectrometer (Bruker Daltonics, Bremen, Germany). Data analysis was performed using MaxQuant version 1.6.12.0 using default parameters with a human database (Uniprot-Swissprot database, released December 2019, containing 20,367 protein sequences).

Results and Discussion

Synthesis and characterization of kinase-inhibitor functionalized NPs.

We first selected a kinase inhibitor, BIM VIII, which can be easily synthesized in large-scale and have specific affinity to kinases. The class of BIM derivatives has attracted our attention as it has strong affinity for PKC isoforms and GSK isoforms, while these molecules can act as pan-kinase inhibitors to capture other kinases.[28] Out of the ten commercially available BIM derivatives, three of these molecules (BIM III, BIM VIII, and BIM X) carries an amine functionality, which enables incorporation of additional moiety on these molecules. Previous studies has examined the kinase capturing efficacy using bottom-up proteomics.[28] While BIM III captured the least amount of kinases, BIM VIII and BIM X showed similar capturing efficacy towards PKC isoforms and GSK isoforms, and could also have affinity to cyclin-dependent kinase isoform 2 (CDK2) and ribosomal S6 kinase isoform 1 (Rsk 1). Although BIM X exhibited stronger affinity to target kinases than BIM VIII, the synthesis of BIM X is more complex. Additionally, BIM X captured more proteins than BIM VIII, including non-specific binding proteins, which will put additional burden on front-end separation prior to top-down MS analysis. As a result, we have synthesized BIM VIII in a large scale following previously published protocols (Figure 4.1).[29-30]

Next, we rationally designed a strategy to incorporate BIM VIII on nanoparticles (Figure 4.2). Previous studies in our lab have developed a method to reproducibly surface silanize superparamagnetic iron-oxide (magnetite, Fe_3O_4) NPs.[25] Further investigation has shown that an organosilane link molecule, N-(3-(triethoxysilyl)propyl)buta2,3-dienamide (BAPTES), can be effectively modified on oleic acid coated Fe_3O_4 NPs.[31] This organosilane link molecule also possesses high chemoselectivity to thiol functionality due to the terminal allene carboxamide functional group, and was not prone to hydrolysis. Following the success on the surface functionalization using BAPTES functionalized nanoparticles, BIM VIII was further modified with additional moiety carrying thiol functional group *via* coupling reaction to affording the final

BIM derivative, termed BIM-GA-Cyst. This thiol-containing moiety (GA-Cyst) was synthesized by reacting glutaric anhydride with S-trityl-L-cysteinamide, which adds linker length from glutaric anhydride, allowing the molecule to reach the active site of kinases (Figure 4.1).[32]

BIM-GA-Cyst was then incorporated on NP-BAPTES (Figure 4.2a). The properties of the surface-functionalized NPs were examined at each step of the reaction by Fourier transform infrared (FTIR) spectroscopy analysis. Incorporation of BAPTES on the Fe_3O_4 showed a strong peak at 1970 and 1947 cm^{-1}, which is the characteristic peak of allene functional group (Figure 4.2b).[33] This characteristic peak disappeared after reacting with BIM-GA-Cyst. This suggests that allene was significantly depleted post-reaction, and BIM-GA-Cyst was successfully incorporated on the NP-BAPTES. When comparing the IR spectra between the fully functionalized NP and the kinase inhibitor molecule, signature peaks for functional group including carbonyl, imide, C-C aromatic alkene, and aromatic bend mode could be observed in both spectra (Figure 4.2c). Due to the high quantum yield of the BIM-GA-Cyst given its conjugated ring system, the NP-BAPTES-BIM-GA-Cyst NPs fluoresced when UV light was applied (Figure 4.2d).

Kinase enrichment with a simple standard protein system

Initial enrichment studies using the fully functionalized NPs were evaluated using a simple standard protein mixture (Figure 4.3). The protein mixture contains two protein kinases, PKCα (78 kDa) and GSK3β (46 kDa), known to be captured by BIM VIII, and three common non-kinase proteins including BSA (67 kDa), β-casein (24 kDa), and carbonic anhydrase (30 kDa). The enrichment was analyzed by SDS-PAGE loaded in different steps during the enrichment process including loading mixture (pre-enrichment sample), flow through (protein mixture of the unattached proteins), wash (washing step), and elution (eluted protein mixture) (Figure 4.3a).

For the simple system enrichment, we observed that both PKCα and GSK3β were captured, as the intensities of the bands in the flow through was weaker than those in the loading mixture (Figure 4.3b). Both of these proteins had intense bands in the elution lane when the Laemmli buffer was used for elution. In the elution lane, the majority of proteins was PKCα and GSK3β, and successfully enrichment was observed when comparing to the loading mixture lane, where these two kinases are much lower in the overall population. For two of the non-kinase proteins, BSA and carbonic anhydrase, neither of these proteins was attached on the particle. The residual protein left on the NPs could be washed off during washing step. Interesting, β-casein was observed to be strongly attached to the functionalized particles. This non-kinase protein was completed attached to the ligand after incubation and it could not be washed away during the washing step. This non-kinase protein eluted in the same conditions as protein kinases.

From the enrichment experiment using a simple standard protein mixture, we have demonstrated success in capturing protein kinases. Two of the non-kinases did not bind to the BIM ligands, while β-casein was strongly attached. This simple system experiment showcased the fundamentals of kinase enrichment. Although it would be ideal to capture only kinases, kinase inhibitors inherently had non-specific binding. In the study of identifying the cellular targets of bisindolylmaleimide class of inhibitors, Brehmer *et al* found that BIM III, which differed to BIM VIII with a methyl group, had affinity to metabolic enzymes such as glycerinaldehyde-3-phosphate dehydrogenase, glucose-6-phosphate dehydrogenase and lactate dehydrogenase B, as well as heat shock proteins (HSPs) including HSP90α, HSP70.1 and HSP73. Therefore, it was not surprising to observe binding of non-kinase proteins.[28] Additionally, Oppermann *et al* reported that the majority of proteins captured by BIM X were non-kinase proteins, and these proteins included abundant 60S and 40S ribosomal proteins.[34]

Kinase enrichment with a complex mixture

Following the investigation on a simple standard protein mixture, we conduct experiment on a complex mixture using protein extracts from heart tissue. As these samples would be subject to MS analysis, elution using Laemmli buffer, which contains a high concentration of SDS, would not be applicable as SDS is detrimental for MS experiment. Azo, which is a photo-cleavable surfactant, has been demonstrated as an alternative for intact protein analysis. As a result, the SDS component was replaced by Azo in hope for an elution condition that can work directly for MS analysis.

The experimental workflow of complex mixture analysis was similar to that for simple system. Instead of a simple Laemmli buffer elution, the attached proteins were eluted with 0.5% Azo solution, 1% Azo solution with β-mercaptoethanol, a component in the Laemmli buffer, and finally Laemmli buffer solution. As observed in Figure 4.4, the loading mixture and flow through lanes did not change significantly. 0.5% Azo solution can elute some proteins, and the overall band distribution was different to that for loading mixture and flow through. A more concentrated Azo, 1%, elution with a reducing reagent showed a distinct protein portfolio compared to loading mixture, flow through, and previous Azo elution. Interestingly, Laemmli elution at the final step did not contain a large amount of protein, which suggests that 1% Azo elution with reducing agent was a strong enough elution to detach most of the proteins that were left on the NPs.

After desalting and surfactant cleavage, both Azo elution samples were subject to bottom-up proteomics analysis. Only two kinases were identified including galactokinase in the 0.5% Azo elution sample and adenylate kinase isoenzyme 1 in both Azo elution samples. At the current stage, the kinase enrichment using the functionalized NPs have not been successful. While kinase enrichment using BIM may have non-specific binding, target kinases such as PKC isoforms, GSK-3 isoforms, and pyruvate kinase could not be identified in the elution. Extraction of kinases should first be optimized. It was unclear whether HEPES extraction could extract sufficient amounts of kinases, and whether those extracted kinases are targets of BIM. The experimental workflow may need further optimization to minimize non-specific binding such as adjusting the surfactant concentration and salt concentration. Moreover, the storage conditions may also be improved. The functionalized NPs were stored at room temperature under dark in 50:50 H_2O:ACN solution. Prolonged storage may degrade the surface silane coating to generate silanol, which contains hydrophilic groups that may introduce non-specific binding.

Conclusion

In this study, we have employed a nanoproteomics strategy to explore kinase enrichment for proteomics analysis. The NPs were synthesized and surface-functionalized with a kinase inhibitor, BIM, and characterized by FTIR spectroscopy. Using a simple standard protein mixture containing two kinases and three non-kinases, the functionalized NPs could capture and elute kinases, with some extents of non-specific binding. For the complex system using a cardiac protein extract, an experimental workflow was developed to enable kinase enrichment with MS analysis. This works presented here is a pilot study and further efforts will be needed to perform extensive evaluation of the enrichment specificity, followed by top-down LC-MS/MS analysis of the enriched kinases from complex systems. We envision ongoing efforts of this nanoproteomics strategy coupled with top-down MS will allow for in-depth characterization of protein kinases, including relative quantification of protein kinase proteoforms and localization of the phosphorylation sites of protein kinases.

Acknowledgement

The authors would like to thank National Institute of Health for financial support, R01 GM117058 (to S. J. and Y. G.) and R01 GM125085 (to Y. G.). Y. G. also would like to acknowledge the NIH grants, R01 HL096971, R01 HL109810, and S10 OD018475.

Reference

1. Kumar, R.; Singh, V. P.; Baker, K. M., Kinase inhibitors for cardiovascular disease. *J. Mol. Cell. Cardiol.* **2007,** *42* (1), 1-11.

2. Humphrey, S. J.; James, D. E.; Mann, M., Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends Endocrinol. Metab.* **2015,** *26* (12), 676-687.

3. Macek, B.; Mann, M.; Olsen, J. V., Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu. Rev. Pharmacol* **2009,** *49*, 199-221.

4. Knight, Z. A.; Lin, H.; Shokat, K. M., Targeting the cancer kinome through polypharmacology. *Nat. Rev. Cancer* **2010,** *10* (2), 130-137.

5. Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S., The Protein Kinase Complement of the Human Genome. *Science* **2002,** *298* (5600), 1912.

6. Schneider, G.; Schmid, R. M., Genetic alterations in pancreatic carcinoma. *Mol. Cancer* **2003,** *2* (1), 15.

7. Bolger, A. P.; Sharma, R.; Li, W.; Leenarts, M.; Kalra, P. R.; Kemp, M.; Coats, A. J. S.; Anker, S. D.; Gatzoulis, M. A., Neurohormonal Activation and the Chronic Heart Failure Syndrome in Adults With Congenital Heart Disease. *Circulation* **2002,** *106* (1), 92.

8. Force, T.; Kuida, K.; Namchuk, M.; Parang, K.; Kyriakis, J. M., Inhibitors of Protein Kinase Signaling Pathways. *Circulation* **2004,** *109* (10), 1196.

9. Glish, G. L.; Vachet, R. W., The basics of mass spectrometry in the twenty-first century. *Nat. Rev. Drug Discov.* **2003,** *2* (2), 140-50.

10. Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y., Top-Down Proteomics: Ready for Prime Time? *Anal. Chem.* **2018,** *90* (1), 110-127.

11. Cai, W.; Tucholski, T. M.; Gregorich, Z. R.; Ge, Y., Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. *Expert Rev. Proteomics* **2016,** *13* (8), 717-30.

12. Smith, L. M.; Kelleher, N. L.; Consortium for Top Down, P., Proteoform: a single term describing protein complexity. *Nat. Methods* **2013,** *10* (3), 186-7.

13. Siuti, N.; Kelleher, N. L., Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* **2007,** *4* (10), 817-821.

14. Smith, L. M.; Kelleher, N. L., Proteoforms as the next proteomics currency. *Science* **2018,** *359* (6380), 1106-1107.

15. Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R., The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **2011,** *7*.

16. Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M., Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Mol. Cell. Proteomics* **2012,** *11* (3).

17. Worboys, J. D.; Sinclair, J.; Yuan, Y.; Jorgensen, C., Systematic evaluation of quantotypic peptides for targeted analysis of the human kinome. *Nat. Methods* **2014,** *11* (10), 1041-1044.

18. Zhang, J. M.; Yang, P. L.; Gray, N. S., Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009,** *9* (1), 28-39.

19. Liu, Y.; Gray, N. S., Rational design of inhibitors that bind to inactive kinase conformations. *Nat. Chem. Biol.* **2006,** *2* (7), 358-364.

20. Wissing, J.; Jansch, L.; Nimtz, M.; Dieterich, G.; Hornberger, R.; Keri, G.; Wehland, J.; Daub, H., Proteomics analysis of protein kinases by target class-selective prefractionation and tandem mass spectrometry. *Mol. Cell. Proteomics* **2007,** *6* (3), 537-547.

21. Daub, H.; Olsen, J. V.; Bairlein, M.; Gnad, F.; Oppermann, F. S.; Korner, R.; Greff, Z.; Keri, G.; Stemmann, O.; Mann, M., Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle. *Mol. Cell* **2008,** *31* (3), 438-448.

22. Bantscheff, M.; Eberhard, D.; Abraham, Y.; Bastuck, S.; Boesche, M.; Hobson, S.; Mathieson, T.; Perrin, J.; Raida, M.; Rau, C.; Reader, V.; Sweetman, G.; Bauer, A.; Bouwmeester, T.; Hopf, C.; Kruse, U.; Neubauer, G.; Ramsden, N.; Rick, J.; Kuster, B.; Drewes, G., Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat. Biotechnol.* **2007,** *25* (9), 1035-1044.

23. Hwang, L.; Ayaz-Guner, S.; Gregorich, Z. R.; Cai, W. X.; Valeja, S. G.; Jin, S.; Ge, Y., Specific Enrichment of Phosphoproteins Using Functionalized Multivalent Nanoparticles. *J. Am. Chem. Soc.* **2015,** *137* (7), 2432-2435.

24. Chen, B. F.; Hwang, L.; Ochowicz, W.; Lin, Z. Q.; Guardado-Alvarez, T. M.; Cai, W. X.; Xiu, L. C.; Dani, K.; Colah, C.; Jin, S.; Ge, Y., Coupling functionalized cobalt ferrite nanoparticle enrichment with online LC/MS/MS for top-down phosphoproteomics. *Chem. Sci.* **2017,** *8* (6), 4306-4311.

25. Roberts, D. S.; Chen, B. F.; Tiambeng, T. N.; Wu, Z. J.; Ge, Y.; Jin, S., Reproducible large-scale synthesis of surface silanized nanoparticles as an enabling nanoproteomics platform: Enrichment of the human heart phosphoproteome. *Nano Res.* **2019,** *12* (6), 1473-1481.

26. Turcheniuk, K.; Tarasevych, A. V.; Kukhar, V. P.; Boukherroub, R.; Szunerits, S., Recent advances in surface chemistry strategies for the fabrication of functional iron oxide based magnetic nanoparticles. *Nanoscale* **2013,** *5* (22), 10729-10752.

27. Abbas, A.; Xing, B.; Loh, T.-P., Allenamides as Orthogonal Handles for Selective Modification of Cysteine in Peptides and Proteins. *Angew. Chem. Int. Ed.* **2014,** *53* (29), 7491-7494.

28. Brehmer, D.; Godl, K.; Zech, B.; Wissing, J.; Daub, H., Proteome-wide identification of cellular targets affected by bisindolylmaleimide-type protein kinase C inhibitors. *Mol. Cell. Proteomics* **2004,** *3* (5), 490-500.

29. Faul, M. M.; Winneroski, L. L.; Krumrich, C. A., A New, Efficient Method for the Synthesis of Bisindolylmaleimides. *J. Org. Chem.* **1998,** *63* (17), 6053-6058.

30. Remsing Rix, L. L.; Kuenzi, B. M.; Luo, Y.; Remily-Wood, E.; Kinose, F.; Wright, G.; Li, J.; Koomen, J. M.; Haura, E. B.; Lawrence, H. R.; Rix, U., GSK3 alpha and beta are new functionally relevant targets of tivantinib in lung cancer cells. *ACS Chem. Biol.* **2014,** *9* (2), 353-8.

31. 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. *Nat. Commun.* **2020,** *11* (1), 3903.

32. Grodsky, N.; Li, Y.; Bouzida, D.; Love, R.; Jensen, J.; Nodes, B.; Nonomiya, J.; Grant, S., Structure of the catalytic domain of human protein kinase C beta II complexed with a bisindolylmaleimide inhibitor. *Biochemistry* **2006,** *45* (47), 13970-13981.

33. Es-Sebbar, E.; Jolly, A.; Benilan, Y.; Farooq, A., Quantitative mid-infrared spectra of allene and propyne from room to high temperatures. *J. Mol. Spectrosc.* **2014,** *305*, 10-16.

34. Oppermann, F. S.; Gnad, F.; Olsen, J. V.; Hornberger, R.; Greff, Z.; Keri, G.; Mann, M.; Daub, H., Large-scale proteomics analysis of the human kinome. *Mol. Cell. Proteomics.* **2009,** *8* (7), 1751-64.

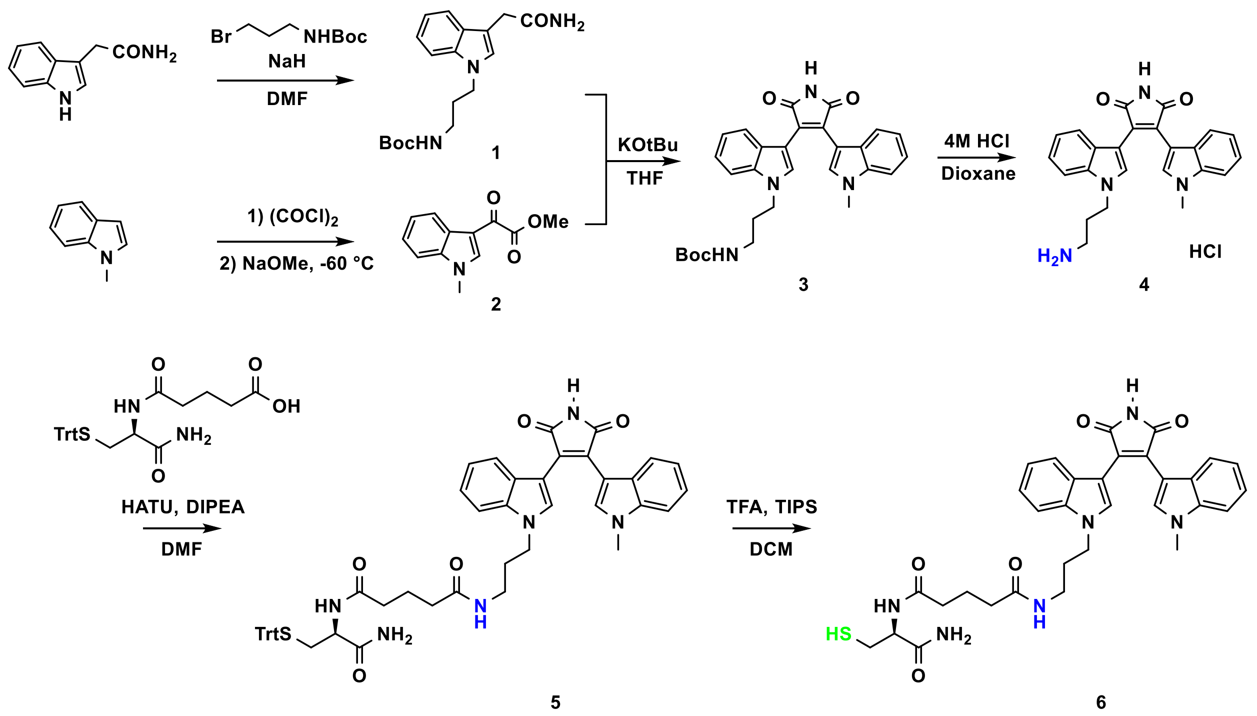


Figure 4.1. Synthesis of bisindolylmaleimide derivatives. The bisindolylmaleimide VIII (**4**) was successfully using a procedure by Faul *et al* with modifications. The molecule was further reacted with a glutaric anhydride – cysteinamide moiety to obtain a thiol functional group and increase linker length.

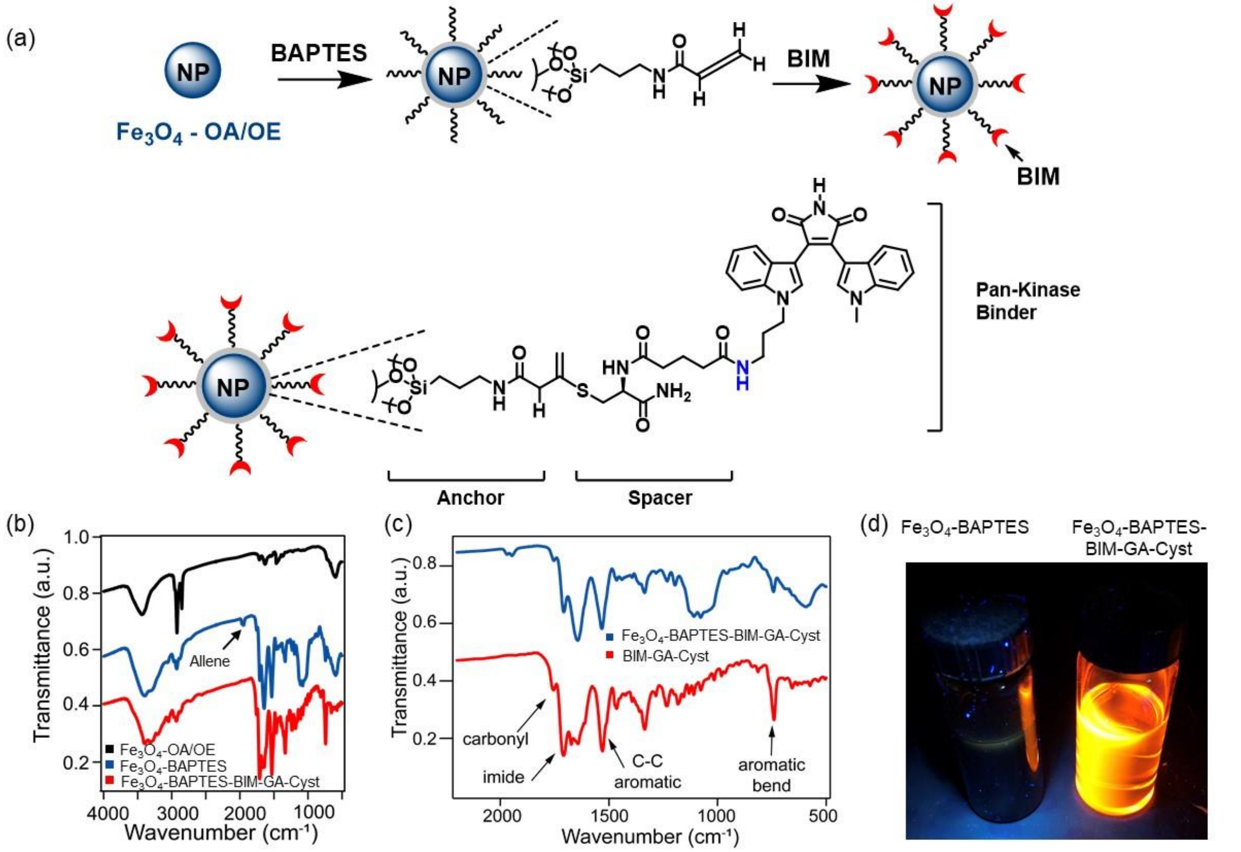


Figure 4.2. Synthesis of kinase inhibitor-functionalized Fe_3O_4 nanoparticle. (a) The functionalized NPs were synthesized by silanization of oleic acid coated Fe_3O_4 NPs. The BIM kinase inhibitor carrying a thiol functional group further reacted with the allene functional group on the BAPTES coated Fe_3O_4 NPs. (b) FTIR spectra showed the comparison among Fe_3O_4 NPs coated with oleic acid, Fe_3O_4 NPs after BAPTES silanization, and Fe_3O_4 after BIM-GA-Cyst reaction. (c) FTIR spectra comparing the small molecule by itself and the fully functionalized Fe_3O_4 NP. (d) The fully functionalized NPs fluoresced under UV light.

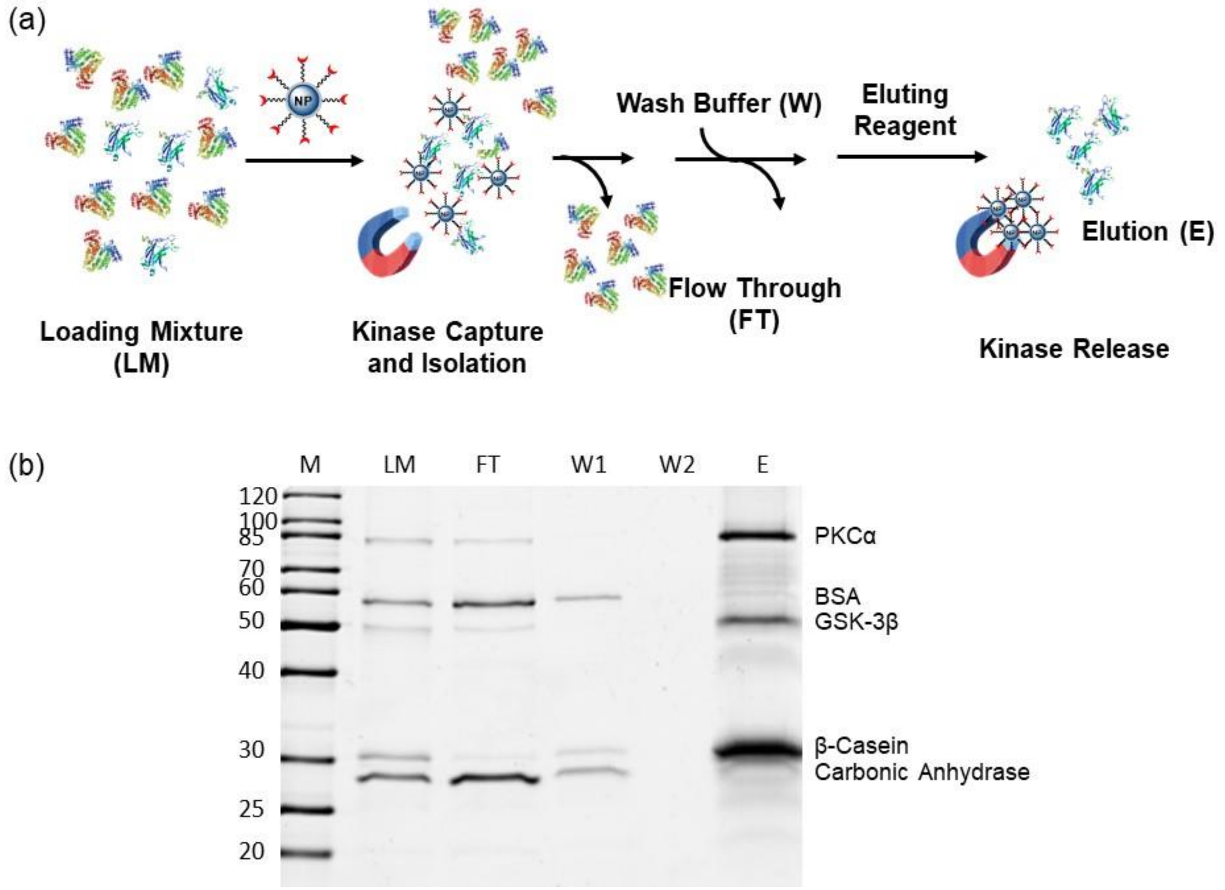


Figure 4.3. Kinase enrichment for a simple mixture. (a) Schematic workflow showing the kinase enrichment procedures. (b) SDS-PAGE analysis of the kinase enrichment for a simple standard protein mixture. Fully functionalized NPs could effective capture target kinases with some degrees of non-specific binding. M, ladder; LM, loading mixture; FT, flow through; W1, high salt wash with 0.1% DDM; W2, low salt wash; E, 1x Laemmli sample buffer elution.

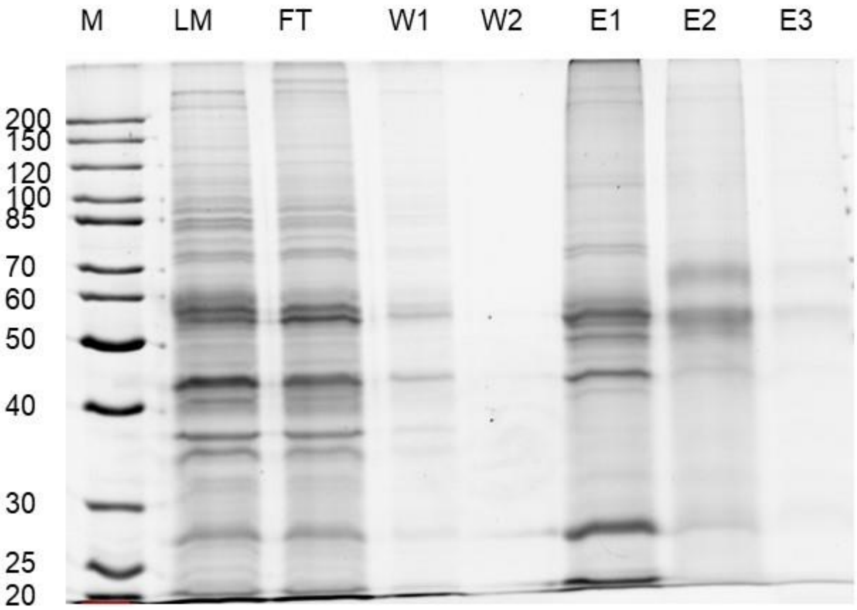


Figure 4.4. Kinase enrichment for a complex mixture. SDS-PAGE analysis of HEPES protein extract from cardiac tissue. With E1 at 0.5% Azo elution and E1 at 1% Azo elution with a reducing agent, most proteins bound on the NPs could be eluted. M, ladder; LM, loading mixture; FT, flow through; W1, high salt wash with 0.1% DDM; W2, low salt wash; E1, 0.5% Azo elution; E2, 1% Azo elution with a reducing agent; E3, 1x Laemmli sample buffer elution.

Supplemental Information

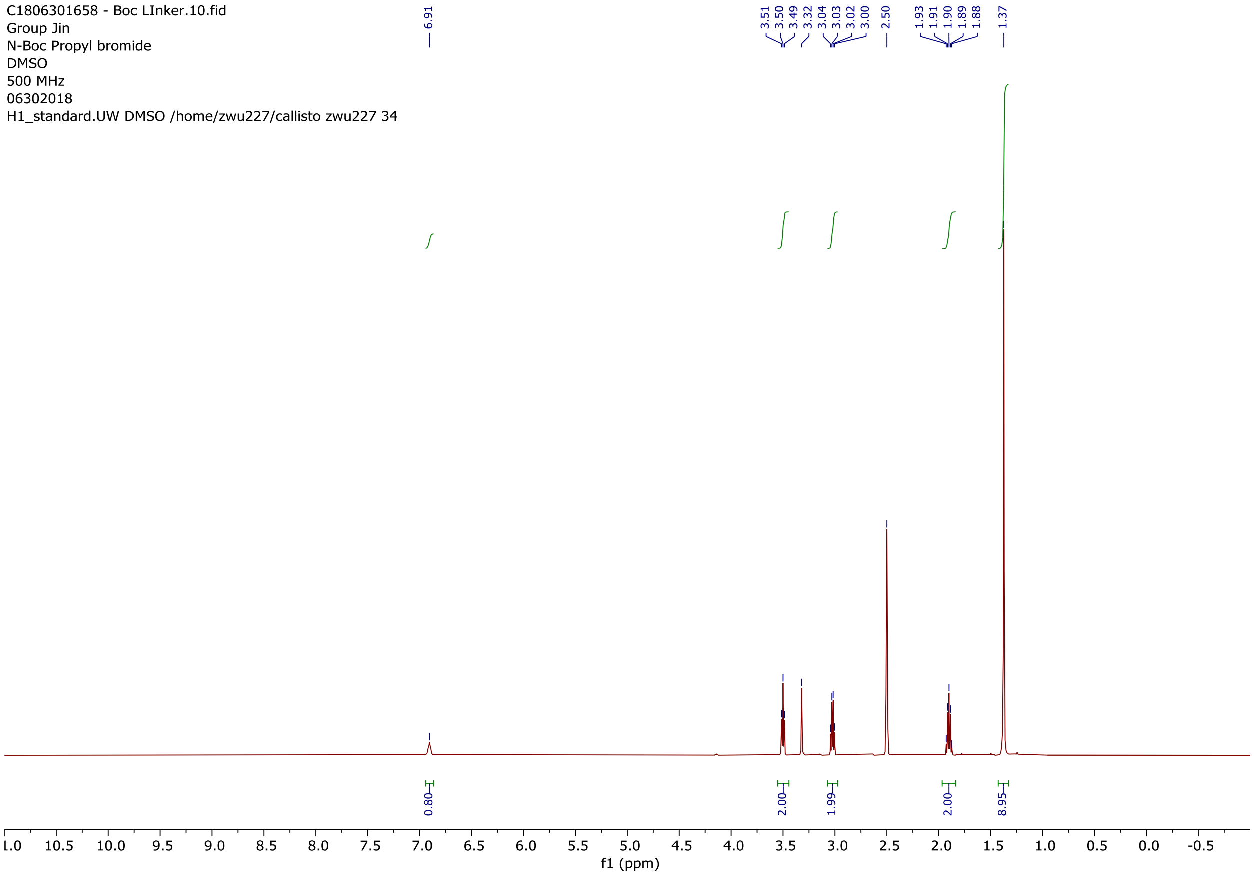


Figure S4.1. ^{1}H NMR spectrum of tert-butyl (3-bromopropyl)carbamate

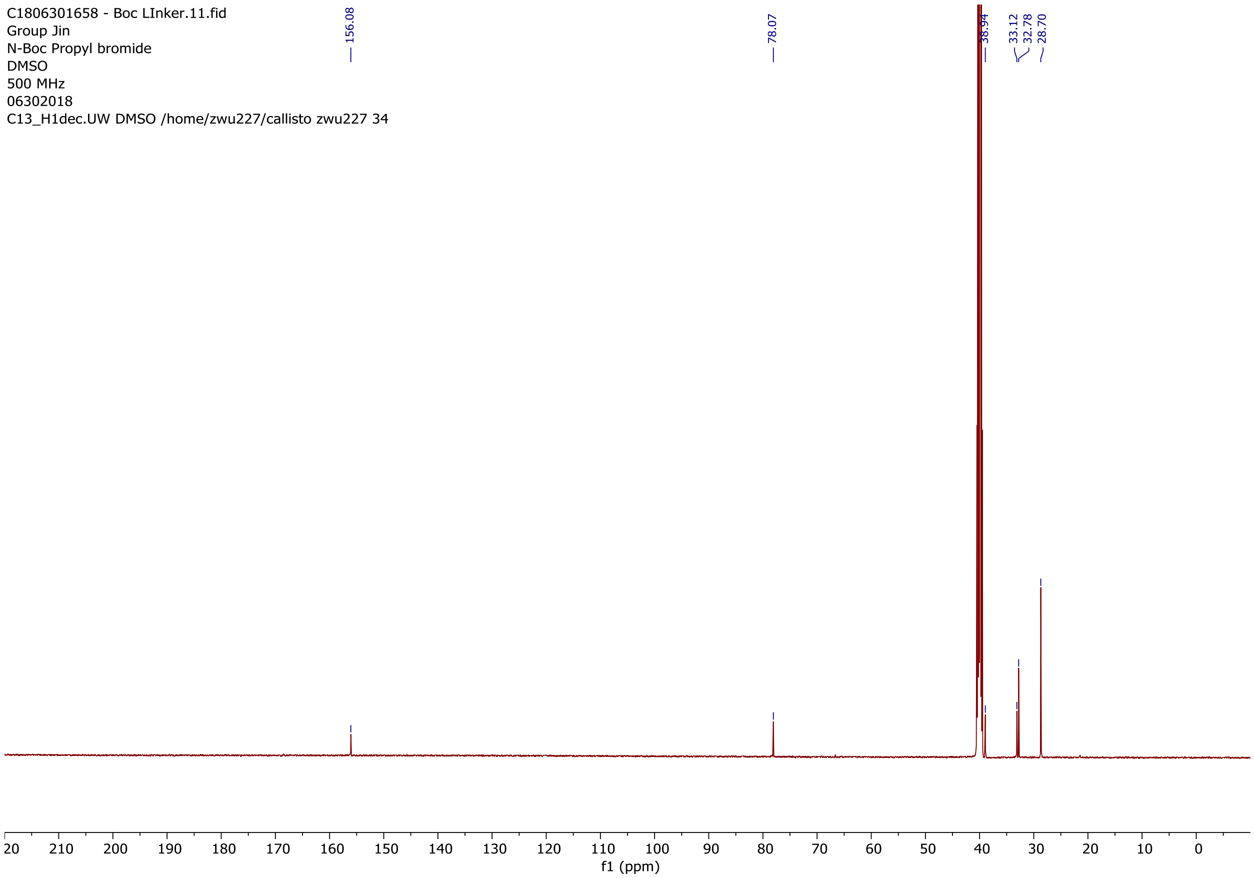


Figure S4.2. ^{13}**C NMR spectrum of tert-butyl (3-bromopropyl)carbamate**

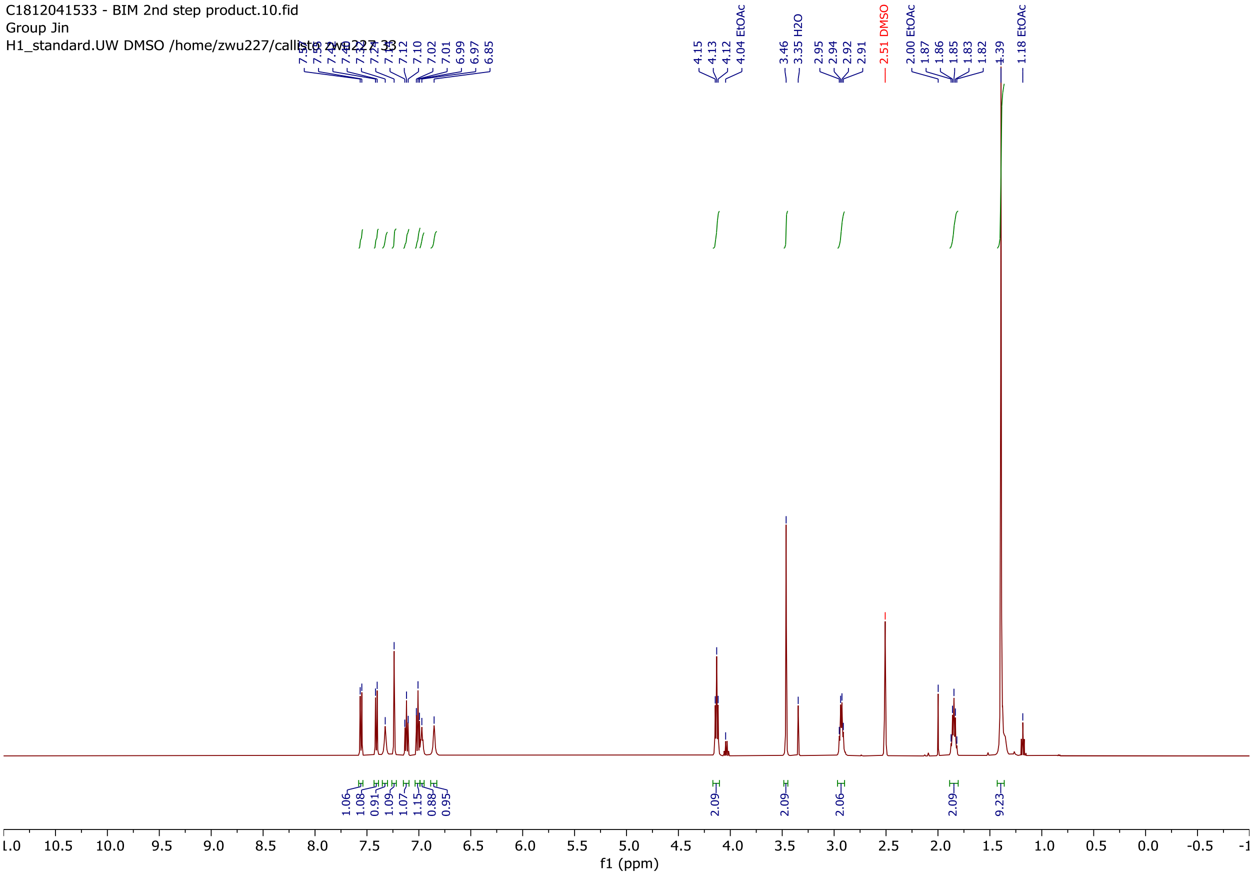


Figure S4.3. ^{1}H NMR of tert-butyl (3-(3-(2-amino-2-oxoethyl)-1H-indol-1-yl)propyl)carbamate (1)

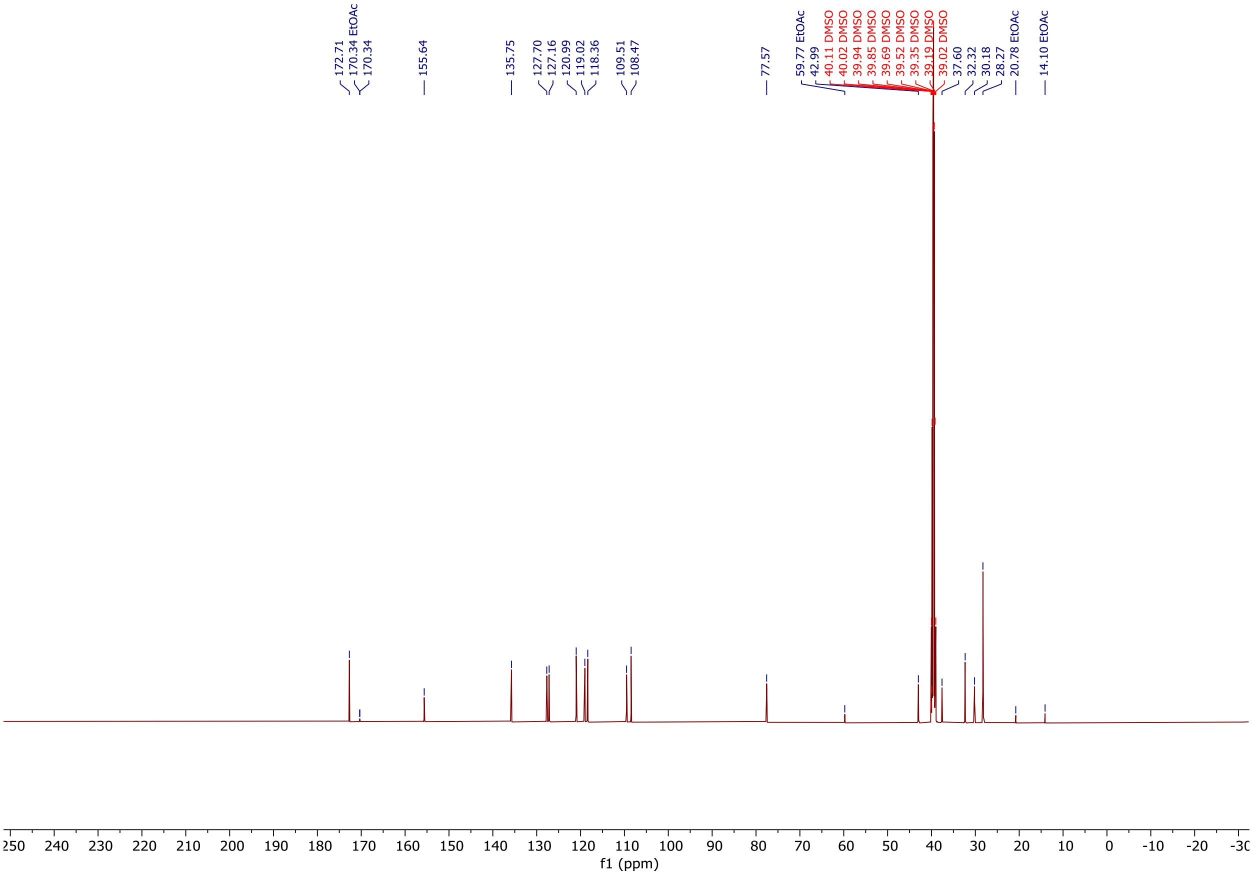


Figure S4.4. ^{13}C NMR of tert-butyl (3-(3-(2-amino-2-oxoethyl)-1H-indol-1-yl)propyl)carbamate (1)

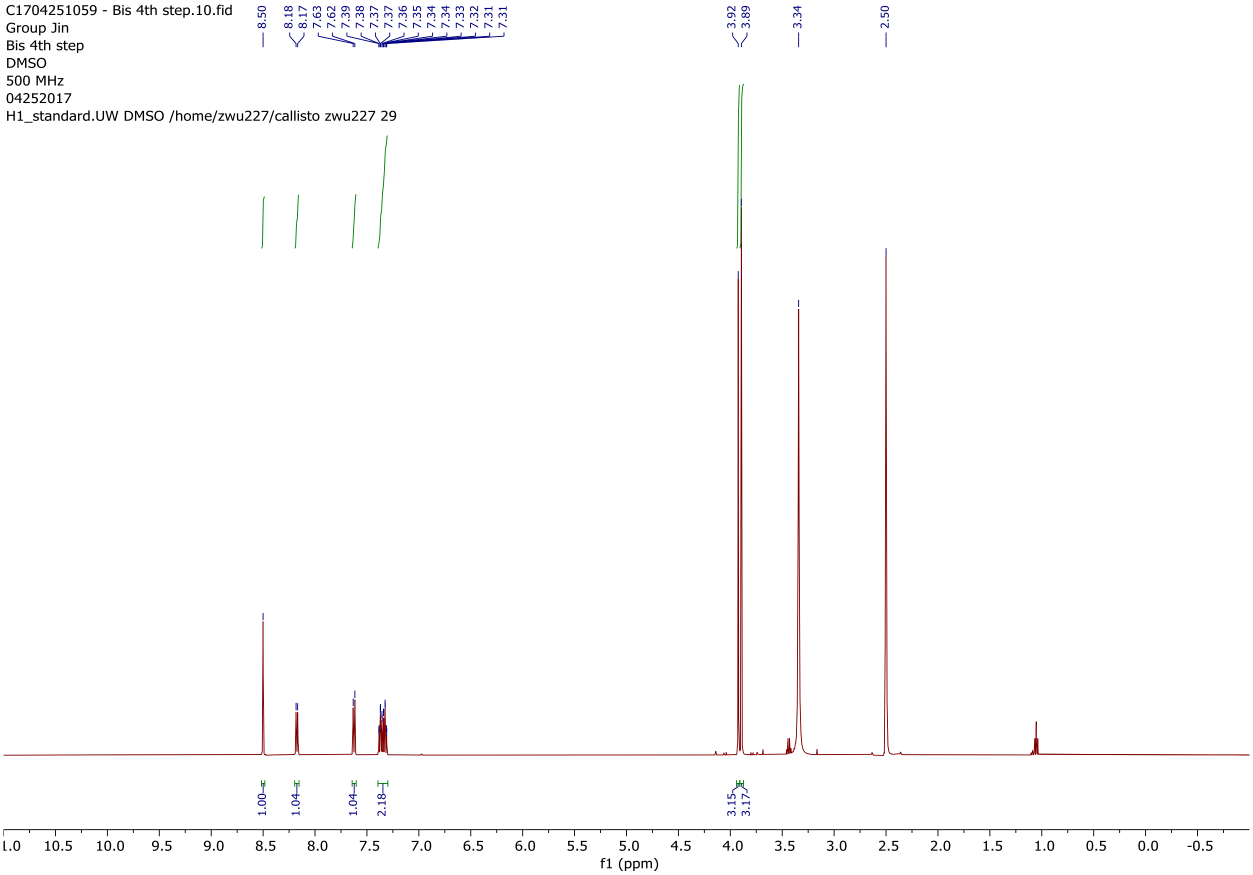


Figure S4.5. ^{1}**H NMR spectrum of methyl 2-(1-methyl-1H-indol-3-yl)-2-oxoacetate (2)**

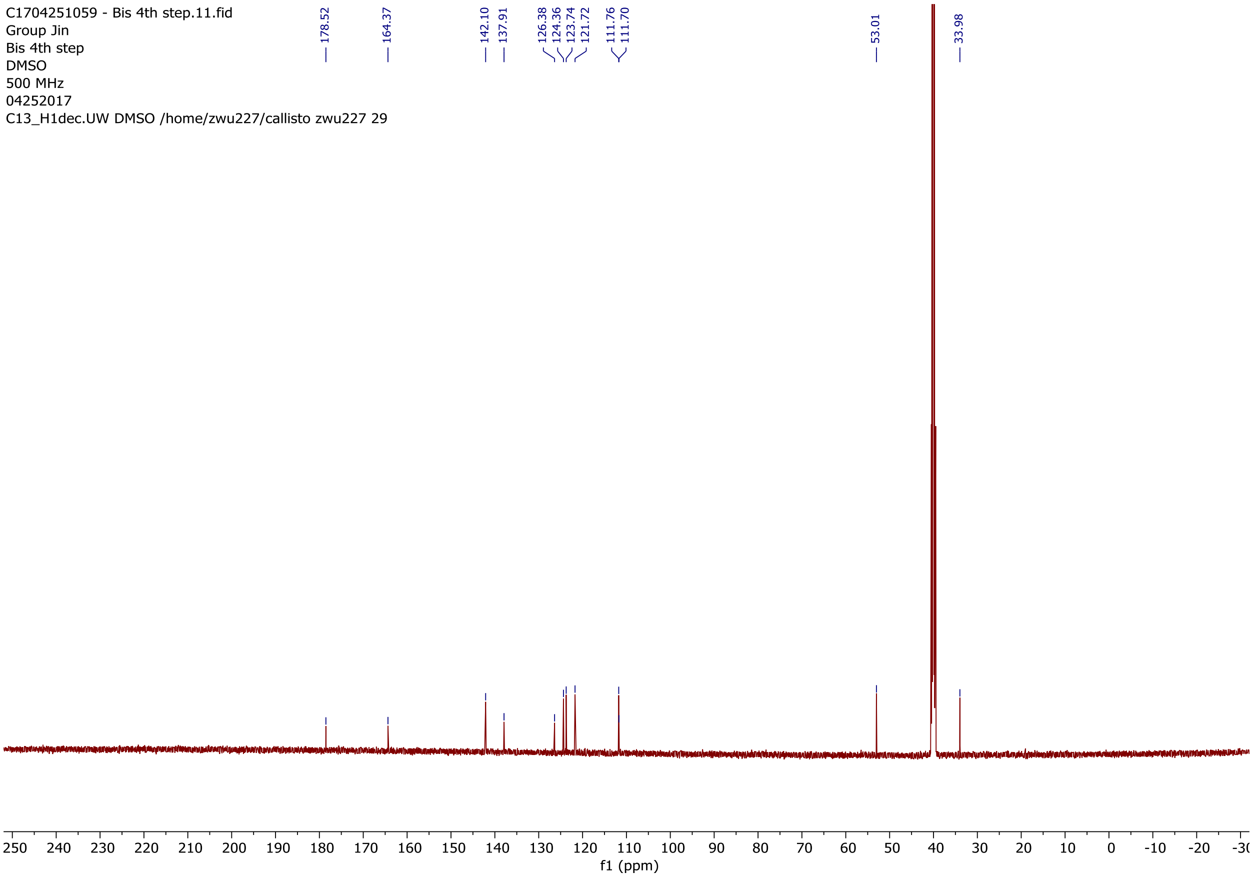


Figure S4.6. ^{13}C NMR spectrum of methyl 2-(1-methyl-1H-indol-3-yl)-2-oxoacetate (2)

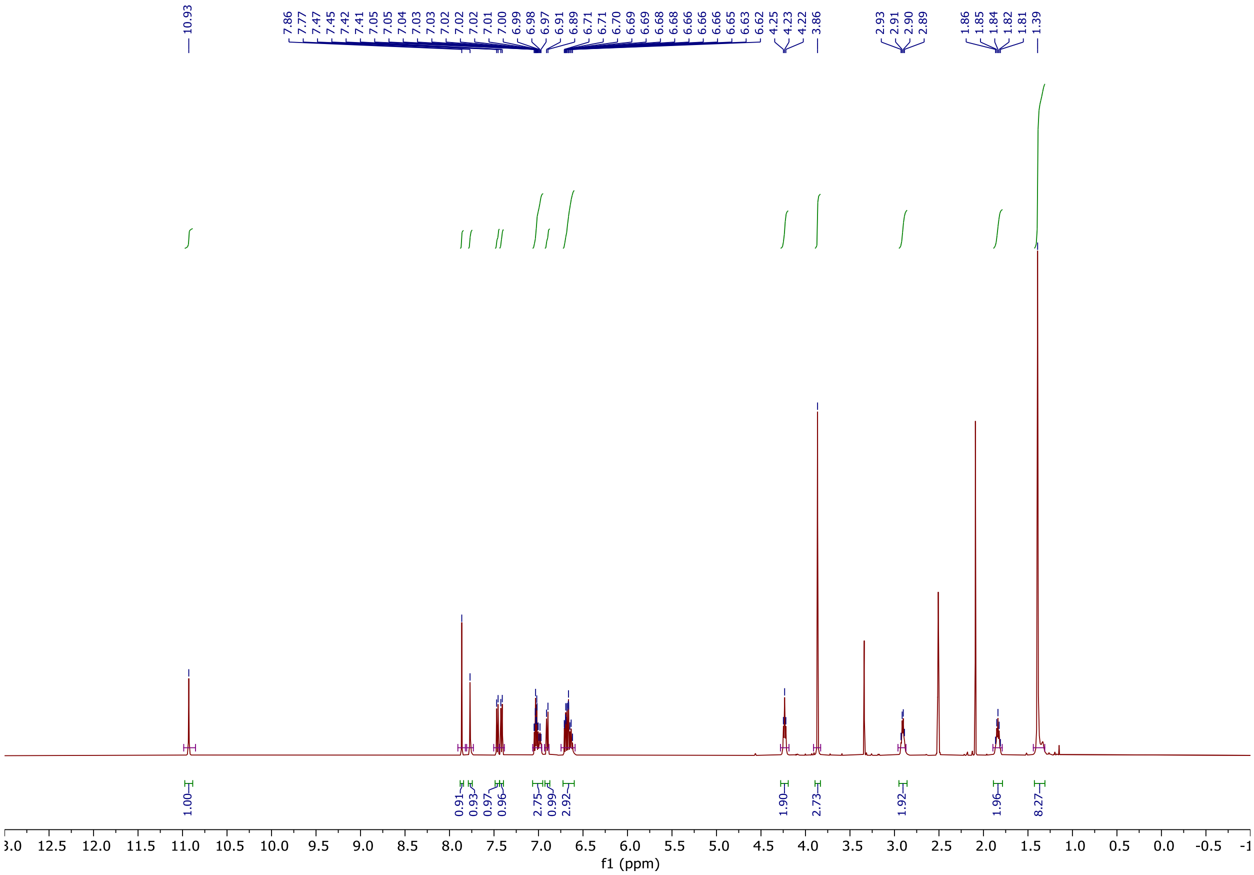


Figure S4.7. ^{1}H NMR spectrum of tert-butyl (3-(3-(4-(1-methyl-1H-indol-3-yl)-2,5-dioxo-2,5-dihydro-1H-pyrrol-3-yl)-1H-indol-1-yl)propyl)carbamate (3)

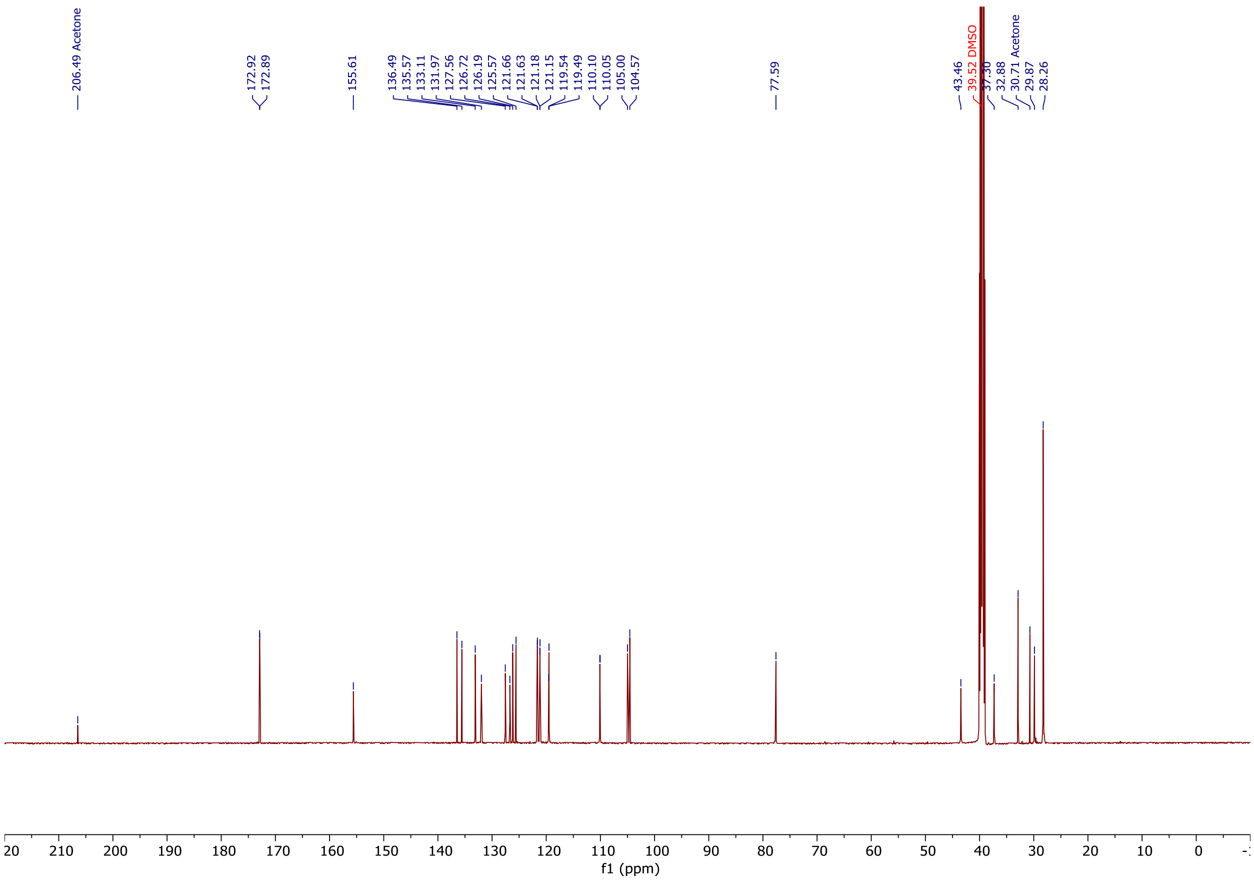


Figure S4.8. ^{13}C NMR spectrum of tert-butyl (3-(3-(4-(1-methyl-1H-indol-3-yl)-2,5-dioxo-2,5-dihydro-1H-pyrrol-3-yl)-1H-indol-1-yl)propyl)carbamate (3)

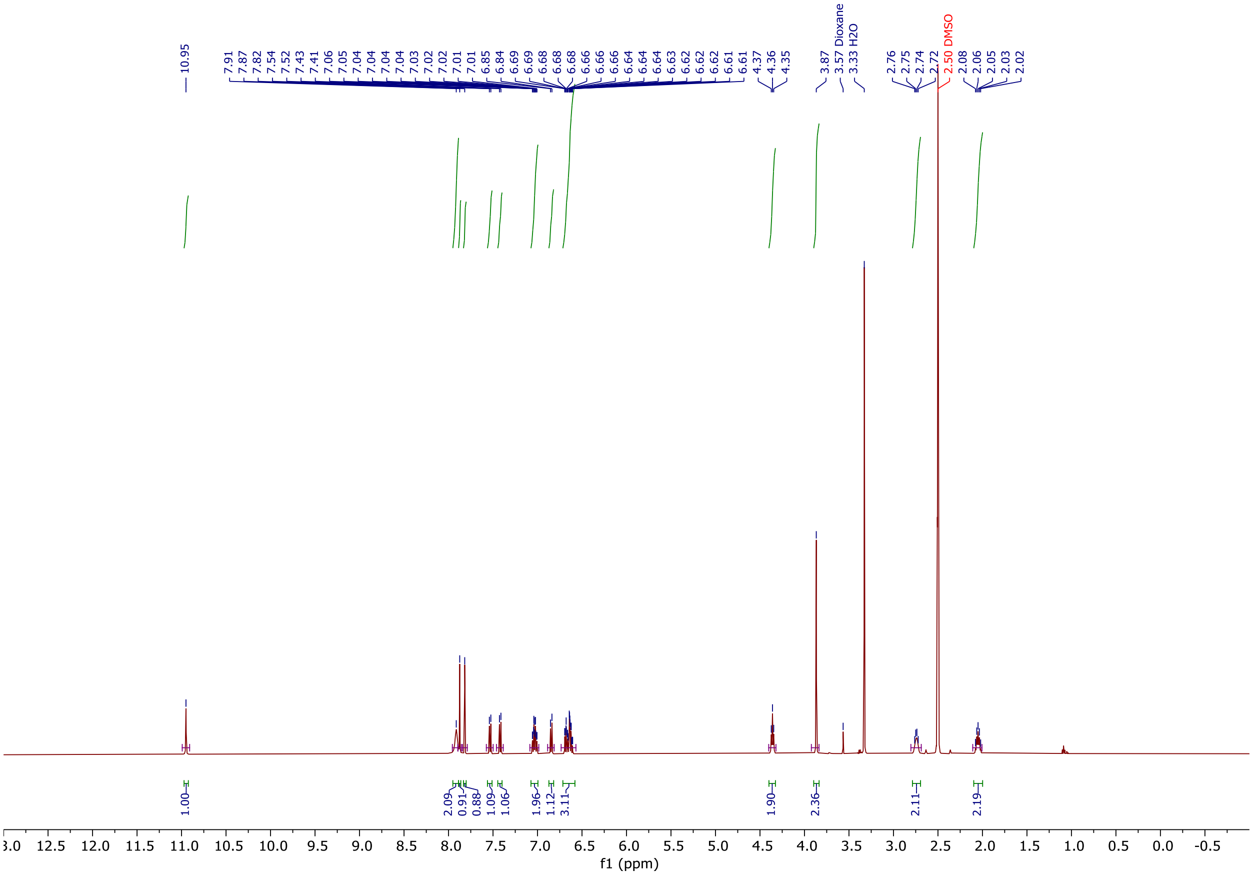


Figure S4.9. ^{1}H NMR spectrum of Bisindolylmaleimide VIII Hydrochloride (4)

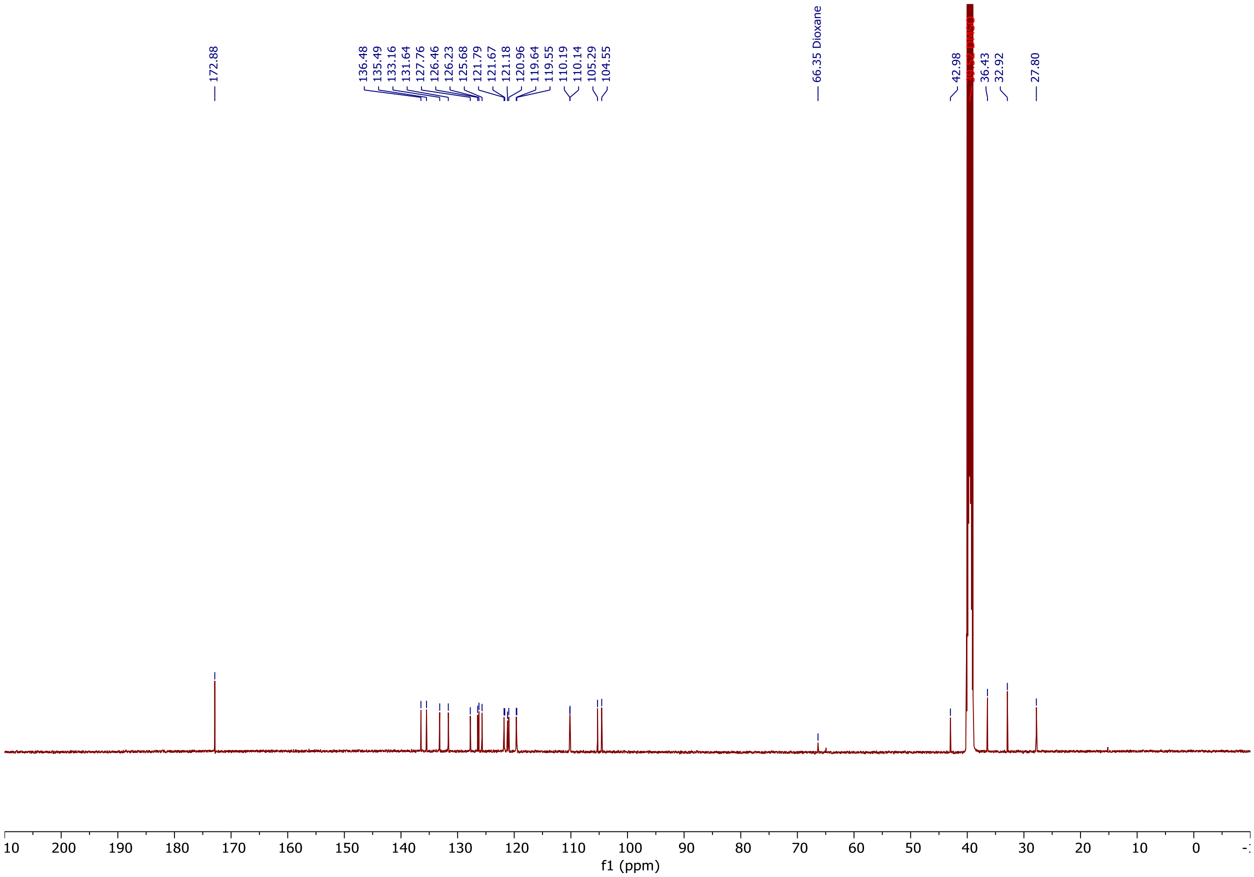


Figure S4.10. ^{13}C NMR spectrum of Bisindolylmaleimide VIII Hydrochloride (4)

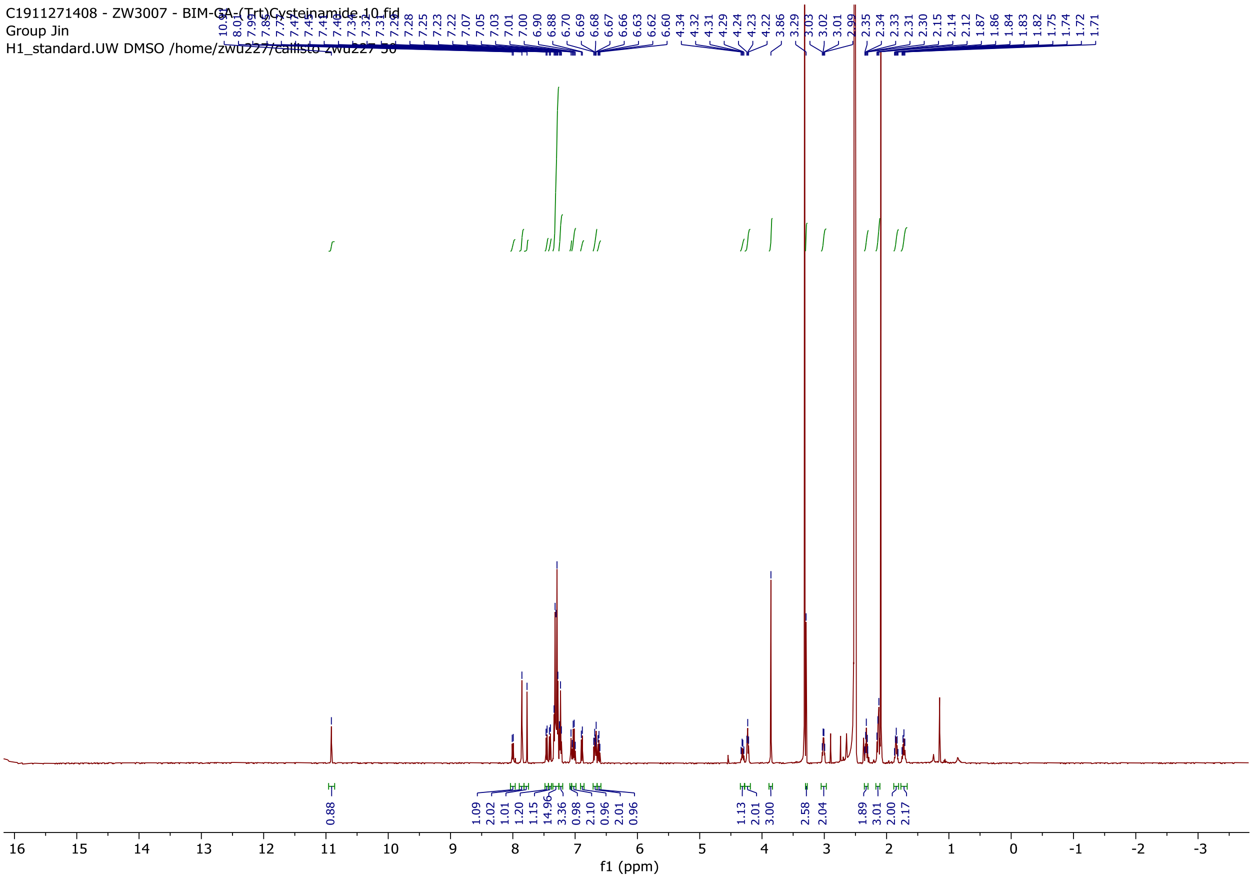


Figure S4.11. ^{1}H NMR spectrum of (S)-N1-(1-amino-1-oxo-3-(tritylthio)propan-2-yl)-N5-(3-(3-(4-(1-methyl-1H-indol-3-yl)-2,5-dioxo-2,5-dihydro-1H-pyrrol-3-yl)-1H-indol-1-yl)propyl)glutaramide (5)

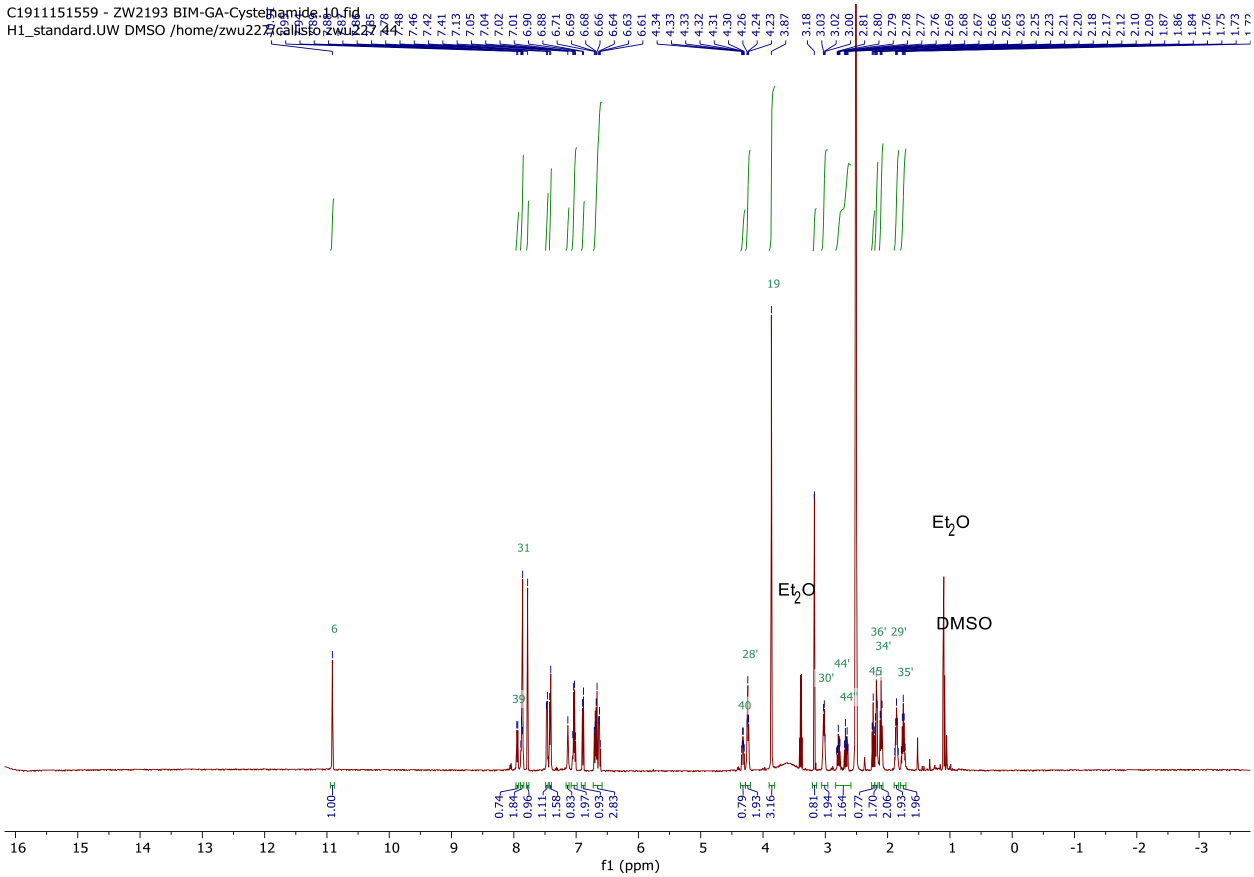


Figure S4.12. ^{1}H NMR spectrum of (S)-N1-(1-amino-3-mercapto-1-oxopropan-2-yl)-N5-(3-(3-(4-(1-methyl-1H-indol-3-yl)-2,5-dioxo-2,5-dihydro-1H-pyrrol-3-yl)-1H-indol-1-yl)propyl)glutaramide (6)

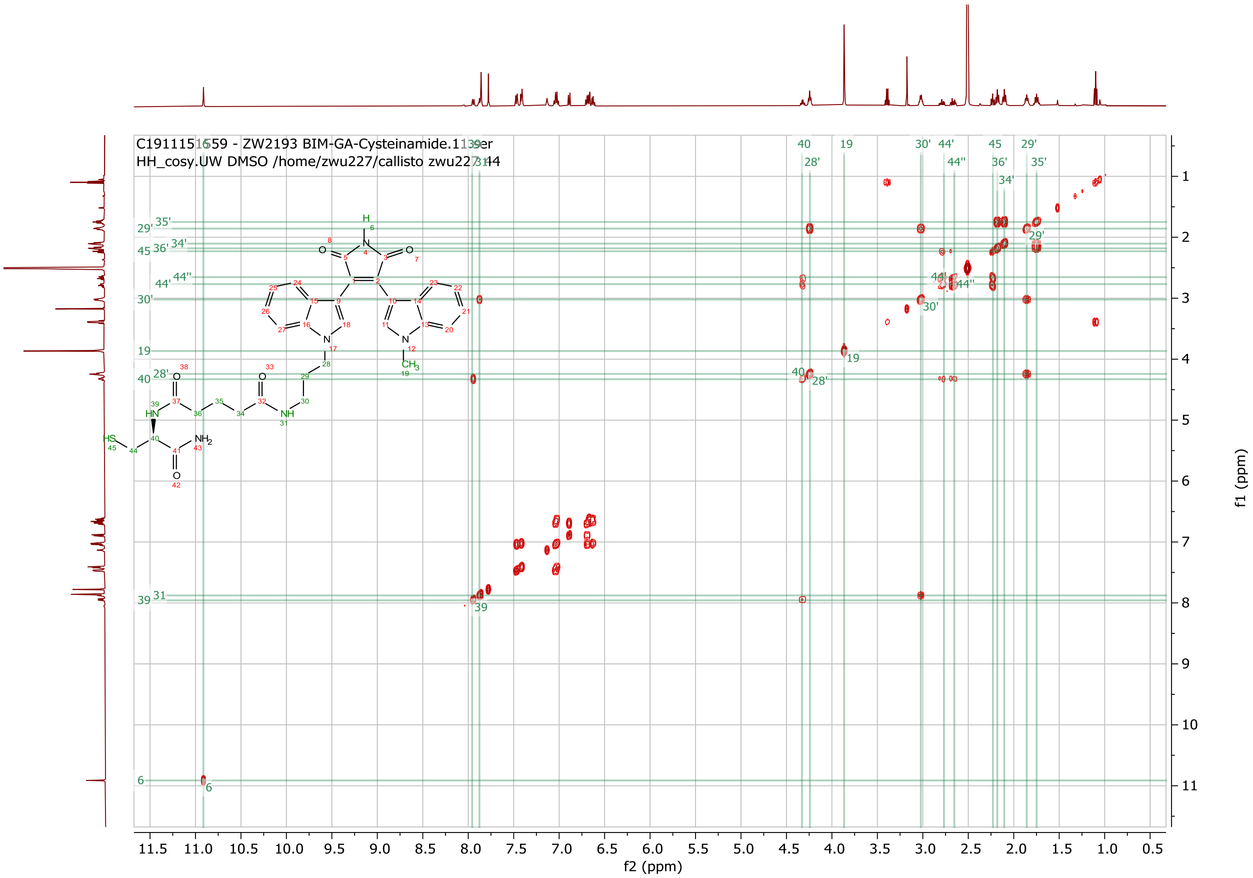


Figure S4.13. COSY NMR spectrum of (S)-N1-(1-amino-3-mercapto-1-oxopropan-2-yl)-N5-(3-(3-(4-(1-methyl-1H-indol-3-yl)-2,5-dioxo-2,5-dihydro-1H-pyrrol-3-yl)-1H-indol-1-yl)propyl)glutaramide (6)

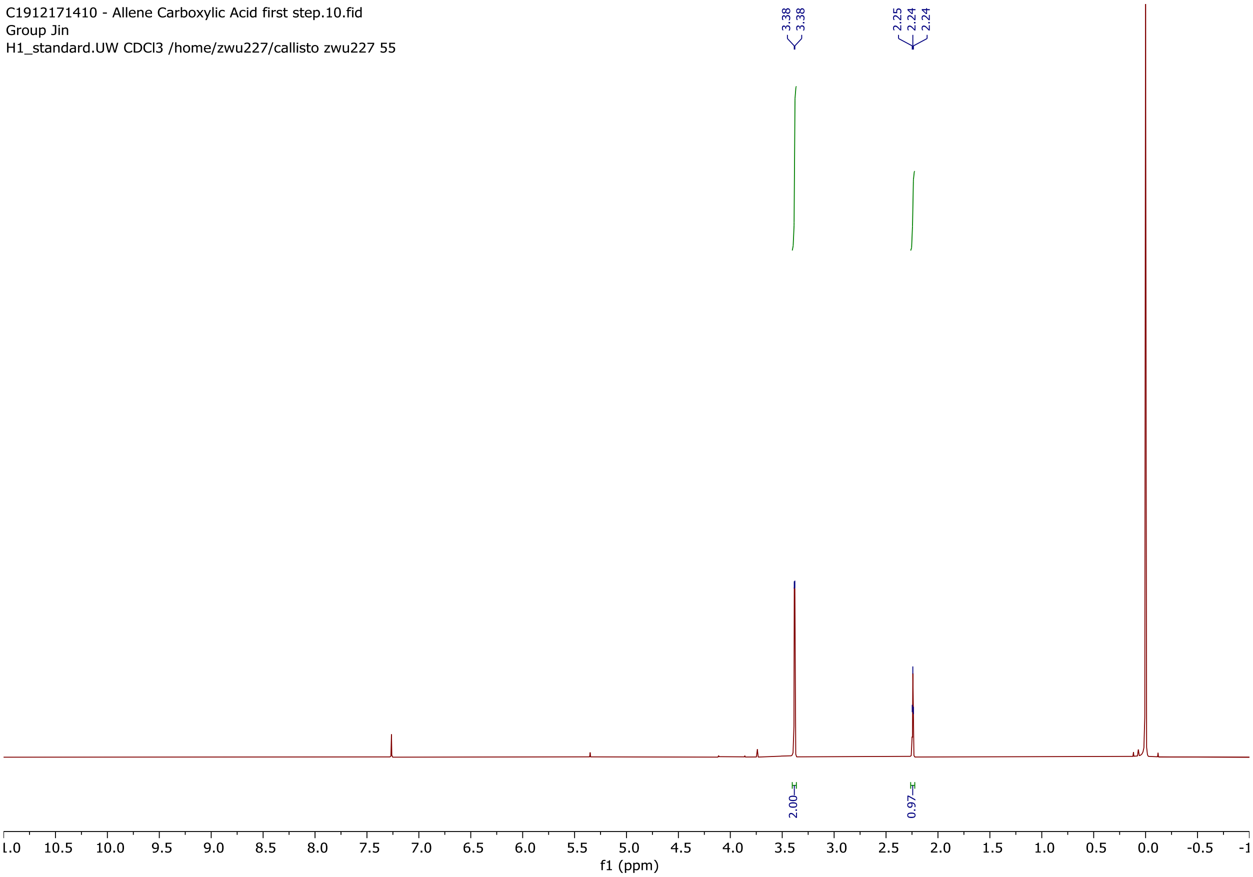


Figure S4.14. ^{1}H NMR spectrum of 3-butynoic acid.

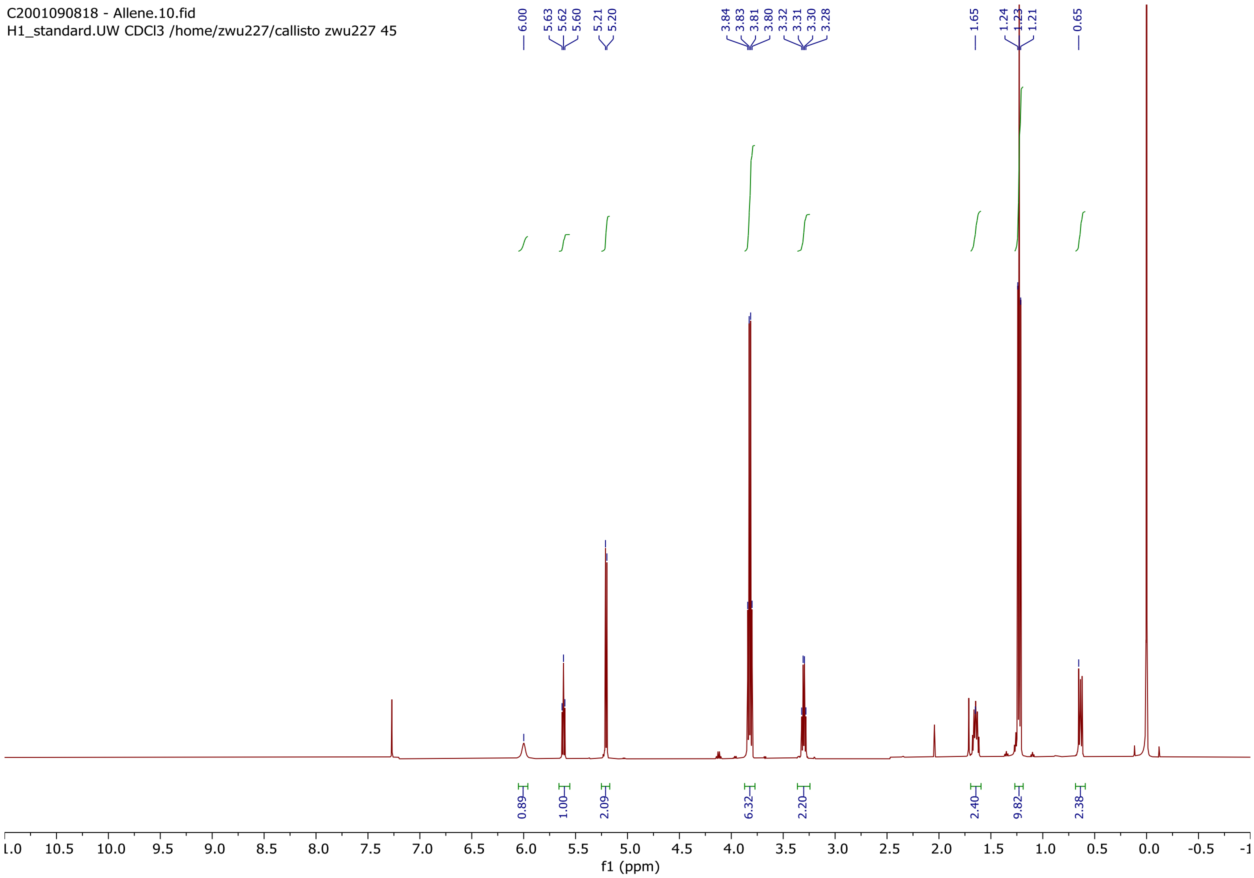
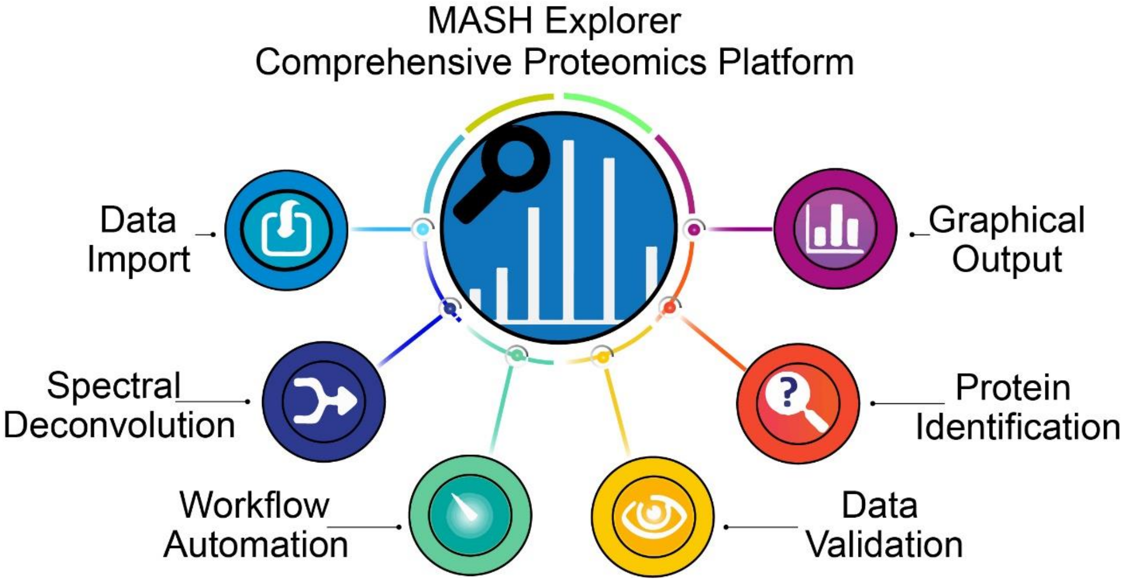


Figure S4.15. ^{1}H NMR spectrum of N-(3-(triethoxysilyl)propyl)buta-2,3-dienamide (BAPTES).

Chapter 5

MASH Explorer: A Universal Software Environment for Top-Down Proteomics



Adapted from: Wu, Z.; Roberts D. S.; Melby, J. A..; Wenger, K.; Wetzel, M.; Gu, Y.; Ramanathan, S. G..; Bayne, E. F.; Liu, X.; Sun, R.; Ong, I. M, McIlwain, S. J.; Ge, Y.*, MASH Explorer: A Universal Software Environment for Top-Down Proteomics, *J. Proteome Res., in press*.

Abstract

Top-down mass spectrometry (MS)-based proteomics enables a comprehensive analysis of proteoforms with molecular specificity to achieve a proteome-wide understanding of protein functions. However, the lack of a universal software for top-down proteomics is becoming increasingly recognized as a major barrier especially for newcomers. Here we develop MASH Explorer, a universal, comprehensive, and user-friendly software environment for top-down proteomics. MASH Explorer integrates multiple spectral deconvolution and database searching algorithms into a single, universal platform which can process top-down proteomics data from various vendor formats, for the first time. It addresses the urgent need in the rapidly growing top-down proteomics community and is freely available to all users worldwide. With the critical need and tremendous support from the community, we envision this MASH Explorer software package will play an integral role in advancing top-down proteomics to realize its full potential for biomedical research.

Introduction

Top-down mass spectrometry (MS)-based proteomics provides a comprehensive analysis of "proteoforms" — all protein products arising from post-translational modifications (PTMs), alternative splicing and genetic variations originating from a single gene — with molecular specificity to achieve a proteome-wide understanding of protein functions.[1-4] Top-down MS analyzes intact proteins without proteolytic digestion and can detect various proteoforms simultaneously in a single MS experiment, thereby enabling their comprehensive molecular characterization. Specific information about proteoforms including PTM sites and sequence variations can be further characterized by tandem MS (MS/MS).[5-7] In contrast to the well-developed software packages in the peptide-based bottom-up proteomics, the data analysis tools for protein-based top-down proteomics remain under-developed due to the major challenge in handling the enormous complexity of high-resolution intact protein mass spectra.[7-9] Particularly, the lack of a universal and user-friendly software for streamlined analysis of complex top-down proteomics data is becoming increasingly recognized as a major barrier, especially for newcomers, thus limiting the broader impact of top-down proteomics in the biomedical research communities. Additionally, the relatively high cost of commercial top-down software limits the accessibility for general users and thus necessitates a freely available academic version.

Here we develop MASH Explorer, a universal, comprehensive, user-friendly, and freely available software environment for top-down proteomics (http://ge.crb.wisc.edu/MASH_Explorer/index.htm). This software can process high-resolution MS, MS/MS data and liquid-chromatography tandem MS (LC-MS/MS) across multiple vendor-specific formats, with automated database searching for protein identification as well as user-friendly tools for proteoform characterization and data visualization/validation. MASH Explorer includes two major workflows: “Discovery Mode” for analysis of complex high-resolution LC-MS/MS data to achieve global protein identification and “Targeted Mode” for comprehensive proteoform characterization including PTMs and sequence variants, with user-friendly graphic user interface (GUI) support. Advancing on our previous generations of proteomics software, MASH Suite[10] and MASH Suite Pro,[11] MASH Explorer has many new features including: (1) development of a universal platform for streamlined data processing from various vendor formats to standardize the data analysis; (2) integration of multiple deconvolution and database search algorithms for significantly enhanced protein identifications; (3) workflow management for high-throughput data processing such as Process Wizard and Workflow Manager; (4) comprehensive proteoform characterization tools with the capability of handling highly complex data resulting from various MS/MS techniques such as collision-induced dissociation (CID), electron capture dissociation (ECD), electron transfer dissociation (ETD), and ultraviolet photodissociation (UVPD). The universal accessibility of non-proprietary, free software solutions such as MASH Explorer will significantly bolster the growth of the top-down proteomics community and welcome newcomers to employ this powerful technology to realize its impact in biomedical research.

Experimental Section

Software Design and Algorithm Support

MASH Explorer is a multithreaded Windows application implemented in C# using .NET framework within the Visual Studio Integrated Development Environment. The software visual components are provided by Microsoft Office Runtime Support. Importing data obtained from different MS instruments is supported using ProteoWizard,[12] DeconEngine,[13] and vendor provided libraries. Additionally, MASH Explorer supports multiple deconvolution and database search algorithms, including TopPIC suite,[14] pTop,[15] Informed-Proteomics,[16] MS-Deconv,[17] MS-Align+,[18] and a modified version of THRASH[19] (eTHRASH[11]). As of March 24th, 2020, the supported versions of the deconvolution and database search algorithms are summarized in Table S5.1.

Computer Setup for Data Analysis

Data analysis was performed to simulate basic research environment. This computer has Windows 10 Student Edition operating system installed. It was equipped with an Intel i5-2400 central processing unit, which has 4 cores and 4 threads for processing, 16 GB DDR3 2400 MHz random access memory, and 1 TB SATA hard drive.

Mass Spectrometry Data

Two LC-MS/MS datasets from two different mass spectrometer vendors, Thermo Scientific and Bruker Corporation (referred to as Thermo and Bruker, respectively, in this manuscript), were utilized to demonstrate the Discovery Mode workflow of the MASH Explorer. The Thermo dataset is publicly available in the MassIVE repository with identifier/username MSV000079978 (ftp://massive.ucsd.edu/MSV000079978/).[20] The dataset was acquired by extracting protein from DLD-1 parental (KRas wt/G13D) human colorectal cancer cells and using a GELFrEE system for size-based separation.[21] The MS experiment was performed using reverse-phase (RP) LC-MS/MS analysis using a 21 Tesla Fourier Transform Ion Cyclotron Resonance mass spectrometer.

The Bruker LC-MS/MS dataset used was publicly available from the PRIDE repository via ProteoXchange with identifier PXD010825.[4] Briefly, the samples from this dataset were prepared by protein extraction using a photo-cleavable surfactant, 4-hexylphenylazosulfonate (Azo), from the human embryonic kidney 293K stem cells. The samples were irradiated to cleave the Azo surfactant. RPLC-MS/MS experiment was performed on a Bruker maXis II quadrupole-time of flight (Q-TOF) mass spectrometer. For the Bruker dataset, the mass spectra were also deconvoluted using Maximum Entropy Algorithm with 80,000 resolution from 10,000 Da to 50,000 Da using Bruker DataAnalysis 4.3.

The dataset for MS/MS analysis was previously published.[22] Briefly, the samples were prepared by extracting proteins from non-human primate skeletal muscles. The dataset was published previously,[22] and is publicly available through ProteomeXchange Consortium via the PRIDE partner repository with the PXD018043 identifier.[23] Target sarcomeric proteins were fractionated using a Waters nanoAQUITY liquid chromatography system, and the fractionated samples were analyzed with a Bruker solariX 12 Tesla FT-ICR instrument using an Advion Nanomate. Specifically, beta-tropomyosin (βTpm, Uniprot-Swissprot accession number P07951) with ECD spectrum and myosin light chain 2 slow isoform (MLC-2S, Uniprot-Swissprot accession number A0A1D5RDY5) with the CID spectrum were used for demonstration of top-down protein characterization using the “Targeted Mode” of MASH Explorer.

A Bruker MS/MS dataset were used for demonstrating the functions of the Targeted Mode in MASH Explorer for characterization of the antibody-drug conjugate (ADC), Adcetris (brentuximab vedotin) subunits, were previously published.[24] Briefly, Adcetris was digested by IdeS, and the interchain disulfide bond was reduced by dithiothreitol (DTT). The subunits were analyzed by LC-MS/MS using a combination of a Waters M-Class LC system and a Bruker maXis

II Q-TOF mass spectrometer. The precursor of each subunit was subject to MS/MS experiment using both CID and ETD. The MS/MS spectra for each subunit were averaged using Bruker DataAnalysis 4.3 software and exported in .ascii format. The ions were extracted using THRASH at 60% fit, and the fragmentation ions were manually validated.

The MS/MS dataset for demonstrating ultraviolet photodissociation (UVPD) ion fragment in Figure 5.1 was previously published by the Brodbelt group and could be accessed through ProteomeXchange with the PXD009447 accession number.[25] This dataset was acquired by applying both CID and UVPD fragmentation methods on single amino acid variants of the human mitochondrial enzyme branched-chain amino acid transferase 2 using a modified prototype of Thermo Q Exactive UHMR instrument.

Algorithm Parameters and Database Search

For comparison of deconvolution and database search algorithms in this study, our analysis used the default parameters from different algorithms. Additionally, we attempted to use the same parameters to minimize runtime differences caused by parameters. For instance, all algorithms were set to 100,000 Dalton (Da) for maximum protein mass. A standard list of modifications such as N-terminal acetylation and N-terminal methionine removal was included during database search. A human database (Uniprot-Swissprot database, release December 2019, containing 20,367 protein sequences) was used for LC-MS/MS database search.

Results and Discussion

Main functions of MASH Explorer Software

MASH Explorer software is a multifaceted software, which is built upon C# programming language using Visual Studio software under .NET framework environment. The combination of C# and Visual Studio enables the development of user-friendly Windows-based graphical interface, which is very intuitive for users, especially newcomers, to learn for streamlined routine analysis. This software development environment allows high performance, low latency, and rich data interaction for high throughput data processing.

The core functions of MASH Explorer include spectral deconvolution, protein identification, proteoform characterization, graphical data output, data validation, and workflow automation (Figure 5.1). Users can choose the integrated deconvolution and database search algorithms to perform spectral deconvolution tasks, which extracts spectral features and subsequently generates a mass list from complex mass spectrum to search against a database for protein identification. Spectral deconvolution and protein identification tasks are supported by GUI tools in the MASH Explorer software for automation. The proteoform characterization function allows users to match fragment ions to protein sequence for localizing PTM sites and identifying sequence variations. MASH Explorer provides GUI to visualize experimental data for LC chromatograms, mass spectra, and fragment ion maps generated from various MS/MS experiments such as CID, ECD/ETD, and UVPD.

One unique feature of MASH Explorer is its universal data processing platform for top-down proteomics with the capability to process data from multiple vendor formats. MASH Explorer currently support specific vendor raw data format from Thermo (.raw), Bruker (.d and .ascii), and Waters (.raw) (Figure 5.1). Moreover, universal data formats such as mzXML and mgf can be imported. The data import function is supported by ProteoWizard,[12] DeconEngine,[13] and vendor provided libraries. To allow successful data import, codes in MASH Explorer are continuously updated to accommodate the latest version of ProteoWizard and vendor-specific data acquisition software.

For the first time, MASH Explorer integrates multiple deconvolution and database searching algorithms into a single platform to maximize the performance for enhanced protein identification (Figure 5.1). Currently, the software incorporates various deconvolution algorithms including MS-Deconv,[17] TopFD,[14] eTHRASH,[19] pParseTD,[15] and ProMex[26] for both MS and MS/MS deconvolution. The database searching algorithms such as MS-Align+,[18] TopPIC,[14] pTop,[15] and MSPathFinderT[26] were integrated in the software for protein identification. MASH Explorer implements the process wizard, a user-friendly GUI to allow users to easily select deconvolution and database search algorithms and to customize the parameters of the selected algorithms for data processing, which is particularly convenient for users. In contrast, some database searching algorithms, such as MS-Align+, require command line inputs using the Windows terminal, which is complicated and difficult for users with limited computational experience. The Configuration tool provides an intuitive interface for the users to find the directory of the supported deconvolution and database search algorithms (Figure S5.1).

The main interface of MASH Explorer allows users to perform data visualization, data validation, and customized output. The panels in the main interface include Workflow, Status Bar, Results View, Mass List, Logbook, and Sequence Table (Figure S5.2). In the Workflow and Parameters panel, several sections are available for users to process top-down MS data, including "Discovery Mode" for LC-MS/MS data processing, "Targeted Mode" for single protein characterization. In addition, "Data Reporting" allows users to save processed datasets in Extensible Markup Language (XML) format, which can be reopened for further analysis, and to export Microsoft object files of both mass spectra and fragment ion maps for image processing. In the Results View panel, a mass spectrum is displayed for data visualization. Users can navigate through different scans, zoom-in and zoom-out of the selected spectrum, and adjust the theoretical Gaussian distribution of the fragment ions using the buttons displayed in the panel. The Mass List panel allows users browse through deconvoluted mass list from the mass spectra for data validation. The entries in the Mass List panel interacts with the Results View and Sequence Table panels, offering users to visualize the fragment ion mapping for different types of MS/MS techniques to characterize the protein sequence. The entries in the Mass List panel can be copied to text editing software and is converted to .msalign format during data processing. In the Sequence Table panel, PTMs of the protein sequences can also be selected and analyzed. The Logbook and Status Bar panels record all data processing by the software such as the versions of the tools used for raw data import, the parameters used in deconvolution and database search tasks. Users can copy the Logbook recordings to a text editor in the event an error occurs. Moreover, the information in the Logbook recordings can help the MASH Explorer software developers troubleshoot any problems.

“Discovery Mode” for LC-MS/MS analysis

MASH Explorer features a “Discovery Mode” workflow that is useful for high-throughput data processing and proteoform identification from batch LC-MS/MS raw data files without *a priori* knowledge of specific proteins (Figure 5.2). “Discovery mode” integrates several top-down MS processing tools to centroid, deconvolute, and search databases against raw datasets for comprehensive proteoform characterization. The software environment highlights intuitive and user-friendly Process Wizard and Workflow Manager to enhance the efficiency of data processing.

MASH Explorer offers a user-friendly GUI, Process Wizard, for different deconvolution and database search algorithms (Figure S5.3). This GUI tool bundles top-down data processing steps including centroiding, deconvolution, and database search. After data import, users can choose available processing pipelines in the Process Wizard. Users can run the algorithms using default settings or change the parameters of each algorithm in the Advanced tab. Additionally, MASH Explorer implements a Workflow Manager to enhance the efficiency of processing top-down proteomics datasets (Figure S5.4). In the Workflow Manager, users can run a batch analysis of top-down proteomic datasets in sequence. The Workflow Manager achieves this function by reading the workflow log created during the algorithm process and gives instructions to wait to execute the next operation. Upon completion, the Workflow Manager automatically imports both the deconvolution and database search results into MASH Explorer for validation of identified proteins. It provides users with convenience in both automatic data file conversion and parameter input in algorithms without sacrificing the efficiency of the database search.

Incorporation of various deconvolution and database search algorithms enables MASH Explorer to improve global proteoform identification and characterization (Figure 5.3 and Figure S5.5). As an example, multiple deconvolution and database search workflows have been performed on both Thermo dataset from human colorectal cancer cell protein extracts[20] and Bruker dataset from surfactant-extracted protein mixture[4] for global proteoform identification (Figure 5.3B and Figure S5.5A) and discussed in the following sections. Identified proteoforms can be further analyzed using tools provided by MASH Explorer for comprehensive proteoform characterization (Figure 5.3C). In addition to the current list of deconvolution and database search algorithms, MASH Explorer has the capability to incorporate more algorithms, owing to the modularity of the software. The incorporation of recently developed deconvolution algorithms such as FLASHDeconv[8] and UniDec[27-28] could increase the diversity in deconvolution methods and thus enable MASH Explorer to process datasets more effectively. Moreover, the results from multiple algorithms can be used for analysis and further implementation of machine learning algorithms. Recent algorithm development in the MASH project will enable users to run a machine learning tool on deconvolution.[23] This machine learning tool used hierarchical clustering to combine deconvoluted peak lists from different algorithms, which can effectively detect true positive peaks while filtering out false positive peaks, resulting in enhanced accuracy and confidence in protein identification during database search.

“Discovery Mode” workflow for a Thermo LC-MS/MS dataset of human colorectal cancer cell protein extracts

Using a publicly available Thermo dataset, we compared the protein identifications for five workflows including MS-Deconv – TopPIC, TopFD – TopPIC, ProMex – MSPathFinderT, MS-Deconv – MS-Align+, and pParseTD – pTop (Figure 5.3B). The combination of multiple deconvolution and database search algorithms can improve global proteoform identification and characterization. MS-Deconv – TopPIC and TopFD – TopPIC workflows did not have any distinct identifications, suggesting that all identifications from these two workflows are also found in other workflows. ProMex – MSPathFinderT and pParseTD – pTop yielded unique identifications (an additional 30-50% of identifications to the 120 consensus identifications from the five workflows). MS-Deconv – MS-Align+ offered many unique identifications. While MS-Align+ can suggest unknown modifications in the proteoform to maximize the number of fragment ion matched to the sequence, this process can potentially increase false positive identifications and thus manual validation is often needed.

Proteins identified in the database search can be further validated using visual tools provided by MASH Explorer. After importing the database search results and clicking on an identified protein, MASH Explorer displays the corresponding MS/MS spectrum in the Spectrum View Panel, shows the related sequence in the Sequence Table panel, and provides the mass list of the scan in the Mass List panel. Users can evaluate the quality of the MS/MS spectrum, adjust the sequence to account for sequence variations, and view the fragment ions which matched to the sequence. ATP synthase subunit *g*, mitochondrial (Uniprot-Swissprot accession number O75964) and microsomal glutathione S-transferase 1 (Uniprot-Swissprot accession number P10620) were identified by MS-Deconv – TopPIC workflow. ATP synthase subunit *g*, mitochondrial was identified with an E-value of 9.66E-037, suggesting high-confidence identification. The protein was modified with the removal of N-terminal methionine and N-terminal acetylation, and the sequence was extensively characterized by CID fragment ions, which is indicated to be high confidence by the E-value (Figure 5.3C). In comparison, microsomal glutathione S-transferase 1 was identified with lower confidence (E-value = 2.82E-004). The N-terminal methionine of this protein was removed. While the E-value was not ideal, CID fragment ions still allowed characterization of the protein (Figure 5.3C).

“Discovery Mode” workflow for a Bruker LC-MS/MS dataset of surfactant-extracted protein mixture

The Bruker LC-MS/MS dataset was acquired in the profile mode, and thus peak picking was needed before further processing. We compared the protein identification from three available workflows for Bruker data file from a publicly available dataset,[4] including MS-Deconv – TopPIC, TopFD – TopPIC, and ProMex – MSPathFinderT (Figure S5.5).

Most protein identifications provided by TopPIC database search algorithm from either MS-Deconv or TopFD deconvolution algorithm overlapped with those from ProMex – MSPathFinderT (Figure S5.5A). As a result, the confidence of protein identification can be increased if the proteins were identified by more than one algorithm. For instance, the 60S ribosomal protein L27 (Uniprot-Swissprot accession number P61353) only has an E-value of 1.80E-004 in the TopFD - TopPIC workflow, suggesting low confidence in protein identification. However, this protein was also identified by MS-Deconv – TopPIC with an E-value of 3.06E-05, and ProMex – MSPathFinderT workflow with an E-value of 3.17E-003. All three workflows identifying the same protein provide strong evidence of a true protein identification. Indeed, the protein was detected in the mass spectrum in a combination of other proteins (Figure S5.5B). The suboptimal E-value for protein identification was most likely due to insufficient fragment ions in the MS/MS experiment.

ProMex – MSPathFinderT workflow offered several unique identifications, which was not identified by neither workflow with TopPIC database search algorithms (Figure S5.5C). While some identifications might be false positives, the intact mass of identifications with E-value >1 (indicative of low confidence in identification), could be observed in the deconvoluted mass spectra. For example, prohibitin-2 (Uniprot-Swissprot accession number Q99623) and sodium channel subunit beta-4 (Uniprot-Swissprot accession number Q8IWT1) were identified with a respective E-value of 1.51 and 10.76. The intact mass of both proteins was found in the deconvoluted mass spectra. The ability to find unique identifications for ProMex – MSPathFinderT workflow is likely the result of better performance of ProMex to extract features (i.e. MS level deconvolution).

"Targeted Mode" for MS/MS analysis

Another important feature of MASH Explorer is a complimentary "Targeted Mode" workflow that is optimized for the detailed and comprehensive characterization of individual proteins, enabling users to identify site-specific PTMs within a protein target (Figure 5.4). The "Targeted Mode" workflow was developed for comprehensive protein characterization. It includes data import, spectral deconvolution to identify and verify isotopic distributions, database search to identify target protein, and finally protein characterization by matching identified isotopic distribution to the target proteoform sequence. The "Targeted Mode" workflow aims to perform identification of fragment ions that help identify and localize PTMs of a target proteoform sequence.

In addition to the functions introduced in our previous generation software, MASH Suite Pro,[11] which provides tools for users to perform charge state and mass shift correction, the "Targeted Mode" in MASH Explorer introduces an Ion Finder Tool GUI that parses through generated ion lists from different fragmentation methods to find proteoform annotations and allow users to match theoretical and observed fragment ions (Figure S5.6). Using the Ion Finder Tool, users can input the fragment ion type and the charge state of the specific fragment ion of interest. The software will then zoom-in to the *m/z* region of targeted ion and attempt to perform fragment ion matching. The Ion Finder Tool complements the existing THRASH algorithm in MASH Explorer to provide a more comprehensive fragment ion mapping for top-down protein analysis. As an example, we have demonstrated on a previously published dataset in the characterization of cardiac sarcomeric proteins from non-human primate skeletal muscle such as βTpm, which was modified with N-terminal acetylation, and MLC-2S with N-terminal methionine removal and PTMs including N-terminal acetylation and deamidation at Asn13 (Figure S5.7).[22]

Characterization of the antibody-drug conjugate subunit using "Targeted Mode" workflow

MASH Explorer can also be extended to characterize the subunits of ADCs,[24] which combine the target specificity of monoclonal antibody and the potency of the cytotoxin drugs, gaining enormous interest in the pharmaceutical industry (Figure 5.5 and Figure 5.6). One of the analytical tasks for ADC characterization is the site localization of drug payload. The digestion of an ADC, brentuximab vedotin, with IdeS resulted (Figure 5.5A). After digstion of an ADC, bretuximab vedotin with IdeS, the resulting subunits were further reacted with DTT to reduce the inter-chain disulfide bonds and subject to LC-MS/MS analysis with both CID and ETD fragmentation (Figure 5.5A). The MS/MS spectra of each subunit were averaged and exported as separate MS file. Due to the complexity of ADC, after IdeS digestion and DTT reduction, Fd subunit with one drug payload has three isomers which can be separated by LC.

Using MASH Explorer, MS/MS spectra can be imported and performed by fragment ion mapping on specific Fd1 subunit (Figure 5.6). Spectral deconvolution was executed on the MS/MS spectra to identify isotopic distributions, which represent fragment ions of this isomer, and calculate their charge states, monoisotopic mass, most abundant mass, intensities, and other parameters. With the known sequence of the Fd subunit and the location of the disulfide bonds, identified isotopic distribution can be mapped to the sequence. Users can visually validate the ions to ensure the accurate assignments since the direct software output may contain false positives in identifying the monoisotopic peak and assigning the charge state especially for the low abundance ions. For payload localization, there are three possible sites including Cys220, Cys226, and Cys229, which are the location of inter-chain disulfide bonds. Moreover, MASH Explorer software provides an Ion Finder Tool to search for ions in the specific amino acid regions of interests. As illustrated in Figure 5.5B, $z\bullet_{15}$, $z\bullet_{16}$, $z\bullet_{23}$, and $z\bullet_{24}$ ions were visualized using the Ion Finder Tool to localize Cys220 as the site for the payload for an Fd isomer with one drug payload.

Conclusion

MASH Explorer is a non-proprietary and free software solution, providing a universal and comprehensive environment for processing top-down proteomics data. The major innovations of MASH Explorer include the integration of multiple deconvolution and search algorithms into a single, universal platform to process raw data from various vendor formats in a user-friendly interface. Since the development of the MASH project, the software has been downloaded and used by more than 600 users around the world (as of March 24th, 2020) (Figure 5.7). While the majority of users are from North America, the MASH software has continuously attracted users across the globe, including users from continents such as Europe and Asia. As the popularity of top-down MS-based proteomics grows, MASH software increasingly becomes a vital and integral tool for users to process complex high-resolution top-down LC-MS/MS data. In addition to the case studies of protein identification from human colorectal cancer cell protein extracts[20] and surfactant-extracted protein mixture,[4] as well as the characterization of ADC,[24] many other groups have used the MASH software packages in top-down proteomics projects including analysis of the light and heavy chain connectivity of a monoclonal antibody,[29] characterization of branched ubiquitin chainsm,[30-31] intact phosphoprotein characterization,[32] and localization of phosphorylation sites of a phosphatase.[33]

As the burgeoning top-down proteomics community continues its rapid growth and has gained momentum through the creation of the Consortium for Top-down Proteomics (CTDP) (http://www.topdownproteomics.org/), the need for universal, comprehensive and globally accessible top-down proteomics software increases tremendously. With the critical need and tremendous support from the community, we envision this MASH Explorer software package will serve as a powerful tool to enable top-down proteomics researchers worldwide, playing an integral role in advancing the top-down proteomics to realize its full potential for biomedical research.

Acknowledgement

This work was supported by the NIH R01 GM125085 (to Y.G). Y.G. would also like to acknowledge support by NIH R01 HL096971, GM117058 and S10 OD018475. We would like to thank Ziqing Lin, Yutong Jin, Bifan Chen, Trisha Tucholski, Kyle Brown, and Austin Carr for the helpful discussions. We also thank all the MASH user worldwide for the excellent feedback which has helped the development of the software.

References

1. Smith, L. M.; Kelleher, N. L., Proteoforms as the next proteomics currency. *Science* **2018,** *359* (6380), 1106-1107.

2. Smith, L. M.; Thomas, P. M.; Shortreed, M. R.; Schaffer, L. V.; Fellers, R. T.; LeDuc, R. D.; Tucholski, T.; Ge, Y.; Agar, J. N.; Anderson, L. C.; Chamot-Rooke, J.; Gault, J.; Loo, J. A.; Pasa-Tolic, L.; Robinson, C. V.; Schluter, H.; Tsybin, Y. O.; Vilaseca, M.; Vizcaino, J. A.; Danis, P. O.; Kelleher, N. L., A five-level classification system for proteoform identifications. *Nat. Methods* **2019**.

3. Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Loo, R. R. O.; Lundberg, E.; MacCoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schluter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlen, M.; Van Eyk, J. E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.; Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B., How many human proteoforms are there? *Nat. Chem. Biol.* **2018,** *14* (3), 206-214.

4. Brown, K. A.; Chen, B. F.; Guardado-Alvarez, T. M.; Lin, Z. Q.; Hwang, L.; Ayaz-Guner, S.; Jin, S.; Ge, Y., A photocleavable surfactant for top-down proteomics. *Nat. Methods* **2019,** *16* (5), 417-420.

5. Siuti, N.; Kelleher, N. L., Decoding protein modifications using top-down mass spectrometry. *Nat. Methods* **2007,** *4* (10), 817-21.

6. Cai, W.; Tucholski, T. M.; Gregorich, Z. R.; Ge, Y., Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. *Expert Rev. Proteomics* **2016,** *13* (8), 717-30.

7. Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y., Top-Down Proteomics: Ready for Prime Time? *Anal. Chem.* **2018,** *90* (1), 110-127.

8. Jeong, K.; Kim, J.; Gaikwad, M.; Hidayah, S. N.; Heikaus, L.; Schluter, H.; Kohlbacher, O., FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics. *Cell Syst.* **2020,** *10* (2), 213-218 e6.

9. Schaffer, L. V.; Millikin, R. J.; Miller, R. M.; Anderson, L. C.; Fellers, R. T.; Ge, Y.; Kelleher, N. L.; LeDuc, R. D.; Liu, X.; Payne, S. H.; Sun, L.; Thomas, P. M.; Tucholski, T.; Wang, Z.; Wu, S.; Wu, Z.; Yu, D.; Shortreed, M. R.; Smith, L. M., Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **2019,** *19* (10), e1800361.

10. Guner, H.; Close, P. L.; Cai, W.; Zhang, H.; Peng, Y.; Gregorich, Z. R.; Ge, Y., MASH Suite: a user-friendly and versatile software interface for high-resolution mass spectrometry data interpretation and visualization. *J. Am. Soc. Mass Spectrom.* **2014,** *25* (3), 464-70.

11. Cai, W. X.; Guner, H.; Gregorich, Z. R.; Chen, A. J.; Ayaz-Guner, S.; Peng, Y.; Valeja, S. G.; Liu, X. W.; Ge, Y., MASH Suite Pro: A Comprehensive Software Tool for Top-Down Proteomics. *Mol. Cell. Proteomics* **2016,** *15* (2), 703-714.

12. Kessner, D.; Chambers, M.; Burke, R.; Agusand, D.; Mallick, P., ProteoWizard: open

source software for rapid proteomics tools development. *Bioinformatics* **2008,** *24* (21), 2534-2536.

13. Jaitly, N.; Mayampurath, A.; Littlefield, K.; Adkins, J. N.; Anderson, G. A.; Smith, R. D., Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics* **2009,** *10*, 87.

14. 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.

15. Sun, R. X.; Luo, L.; Wu, L.; Wang, R. M.; Zeng, W. F.; Chi, H.; Liu, C.; He, S. M., pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification. *Anal. Chem.* **2016,** *88* (6), 3082-3090.

16. Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Liu, T.; Petyuk, V. A.; Tolic, N.; Pasa-Tolic, L.; Smith, R. D.; Payne, S. H.; Kim, S., Informed-Proteomics: open-source software package for top-down proteomics. *Nat. Methods* **2017,** *14* (9), 909-914.

17. Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A., Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Mol. Cell. Proteomics* **2010,** *9* (12), 2772-82.

18. Liu, X.; Sirotkin, Y.; Shen, Y.; Anderson, G.; Tsai, Y. S.; Ting, Y. S.; Goodlett, D. R.; Smith, R. D.; Bafna, V.; Pevzner, P. A., Protein Identification Using Top-Down Spectra. *Mol. Cell. Proteomics* **2012,** *11* (6).

19. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W., Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* **2000,** *11* (4), 320-332.

20. 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.; Hendrickson, C. L., Identification and Characterization of Human Proteoforms by Top-Down LC-21 Tesla FT-ICR Mass Spectrometry. *J. Proteome Res.* **2017,** *16* (2), 1087-1096.

21. Tran, J. C.; Doucette, A. A., Gel-eluted liquid fraction entrapment electrophoresis: an electrophoretic method for broad molecular weight range proteome separation. *Anal. Chem.* **2008,** *80* (5), 1568-73.

22. Jin, Y.; Diffee, G. M.; Colman, R. J.; Anderson, R. M.; Ge, Y., Top-down Mass Spectrometry of Sarcomeric Protein Post-translational Modifications from Non-human Primate Skeletal Muscle. *J. Am. Soc. Mass Spectrom.* **2019,** *30* (12), 2460-2469.

23. McIlwain, S. J.; Wu, Z.; Wetzel, M.; Belongia, D.; Jin, Y.; Wenger, K.; Ong, I. M.; Ge, Y., Enhancing Top-Down Proteomics Data Analysis by Combining Deconvolution Results through a Machine Learning Strategy. *J. Am. Soc. Mass Spectrom.* **2020,** *31* (5), 1104-1113.

24. Chen, B.; Lin, Z.; Zhu, Y.; Jin, Y.; Larson, E.; Xu, Q.; Fu, C.; Zhang, Z.; Zhang, Q.; Pritts, W. A.; Ge, Y., Middle-Down Multi-Attribute Analysis of Antibody-Drug Conjugates with Electron Transfer Dissociation. *Anal. Chem.* **2019,** *91* (18), 11661-11669.

25. Mehaffey, M. R.; Sanders, J. D.; Holden, D. D.; Nilsson, C. L.; Brodbelt, J. S., Multistage Ultraviolet Photodissociation Mass Spectrometry To Characterize Single Amino Acid Variants of Human Mitochondrial BCAT2. *Anal. Chem.* **2018,** *90* (16), 9904-9911.

26. Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Liu, T.; Petyuk, V. A.; Tolić, N.; Paša-Tolić, L.; Smith, R. D.; Payne, S. H.; Kim, S., Informed-Proteomics: open-source software package for top-down proteomics. *Nat. Methods* **2017,** *14*, 909.

27. Marty, M. T.; Baldwin, A. J.; Marklund, E. G.; Hochberg, G. K.; Benesch, J. L.; Robinson, C. V., Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Anal. Chem.* **2015,** *87* (8), 4370-6.

28. Marty, M. T., A Universal Score for Deconvolution of Intact Protein and Native Electrospray Mass Spectra. *Anal. Chem.* **2020,** *92* (6), 4395-4401.

29. Srzentic, K.; Nagornov, K. O.; Fornelli, L.; Lobas, A. A.; Ayoub, D.; Kozhinov, A. N.; Gasilova, N.; Menin, L.; Beck, A.; Gorshkov, M. V.; Aizikov, K.; Tsybin, Y. O., Multiplexed Middle-Down Mass Spectrometry as a Method for Revealing Light and Heavy Chain Connectivity in a Monoclonal Antibody. *Anal. Chem.* **2018,** *90* (21), 12527-12535.

30. Crowe, S. O.; Rana, A. S. J. B.; Deol, K. K.; Ge, Y.; Strieter, E. R., Ubiquitin Chain Enrichment Middle-Down Mass Spectrometry Enables Characterization of Branched Ubiquitin Chains in Cellulo (vol 89, pg 4428, 2017). *Anal. Chem.* **2017,** *89* (17), 9610-9610.

31. Rana, A. S. J. B.; Ge, Y.; Strieter, E. R., Ubiquitin Chain Enrichment Middle-Down Mass Spectrometry (UbiChEM-MS) Reveals Cell-Cycle Dependent Formation of Lys11/Lys48 Branched Ubiquitin Chains. *J. Proteome Res.* **2017,** *16* (9), 3363-3369.

32. Roberts, D. S.; Chen, B.; Tiambeng, T. N.; Wu, Z.; Ge, Y.; Jin, S., Reproducible large-scale synthesis of surface silanized nanoparticles as an enabling nanoproteomics platform: Enrichment of the human heart phosphoproteome. *Nano Res.* **2019,** *12* (6), 1473-1481.

33. Wu, C. G.; Chen, H.; Guo, F.; Yadav, V. K.; McIlwain, S. J.; Rowse, M.; Choudhary, A.; Lin, Z.; Li, Y.; Gu, T.; Zheng, A.; Xu, Q.; Lee, W.; Resch, E.; Johnson, B.; Day, J.; Ge, Y.; Ong, I. M.; Burkard, M. E.; Ivarsson, Y.; Xing, Y., PP2A-B' holoenzyme substrate recognition, regulation and role in cytokinesis. *Cell Discov.* **2017,** *3*, 17027.

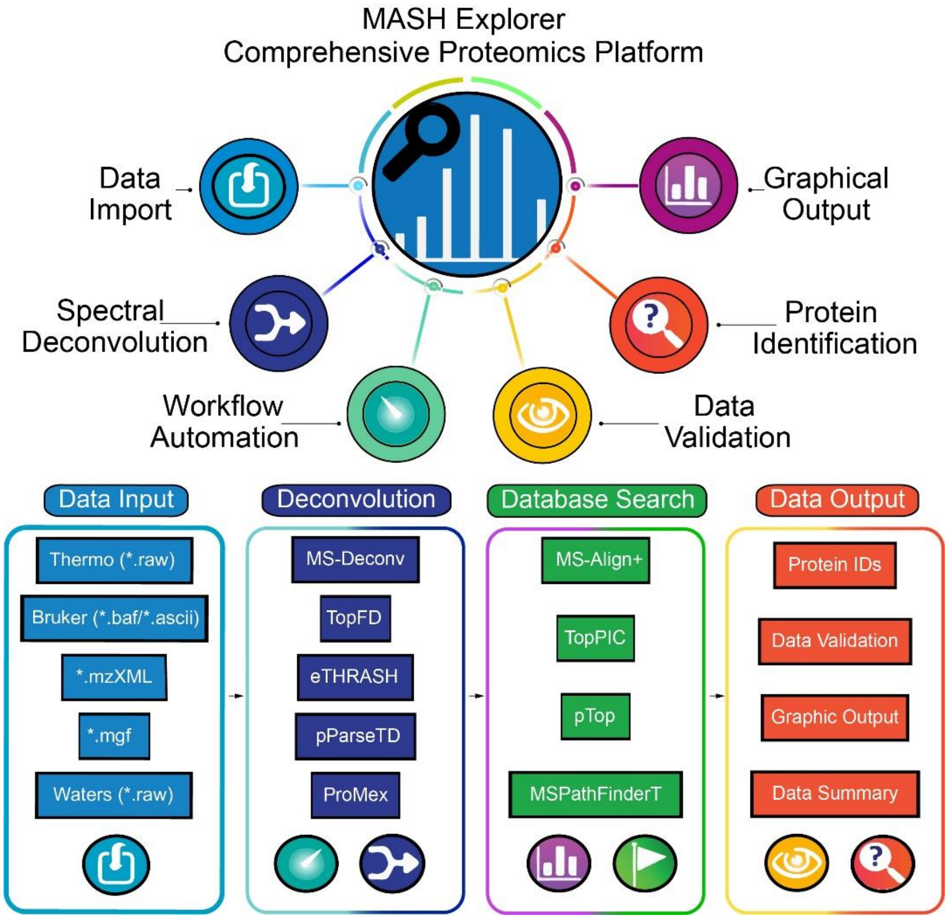


Figure 5.1. Schematic of the various MASH Explorer functions for proteomics data processing. Main functions of MASH Explorer include data import, spectral deconvolution, workflow automation, data validation, protein identification, and graphical output. MASH Explorer utilizes a new data processing module based on the ProteoWizard Library to accept various data input file formats from major instrument vendors (e.g. Thermo, Bruker, and Waters). Raw MS and MS/MS data files are then processed by deconvolution algorithms (i.e. MS-Deconv, TopFD, eTHRASH, and pParseTD), and database search algorithms (i.e. MS-Align+, TopPIC, pTop, and MSPathFinderT). MASH Explorer provides a user-friendly interface for data validation, proteoform identification, and characterization.

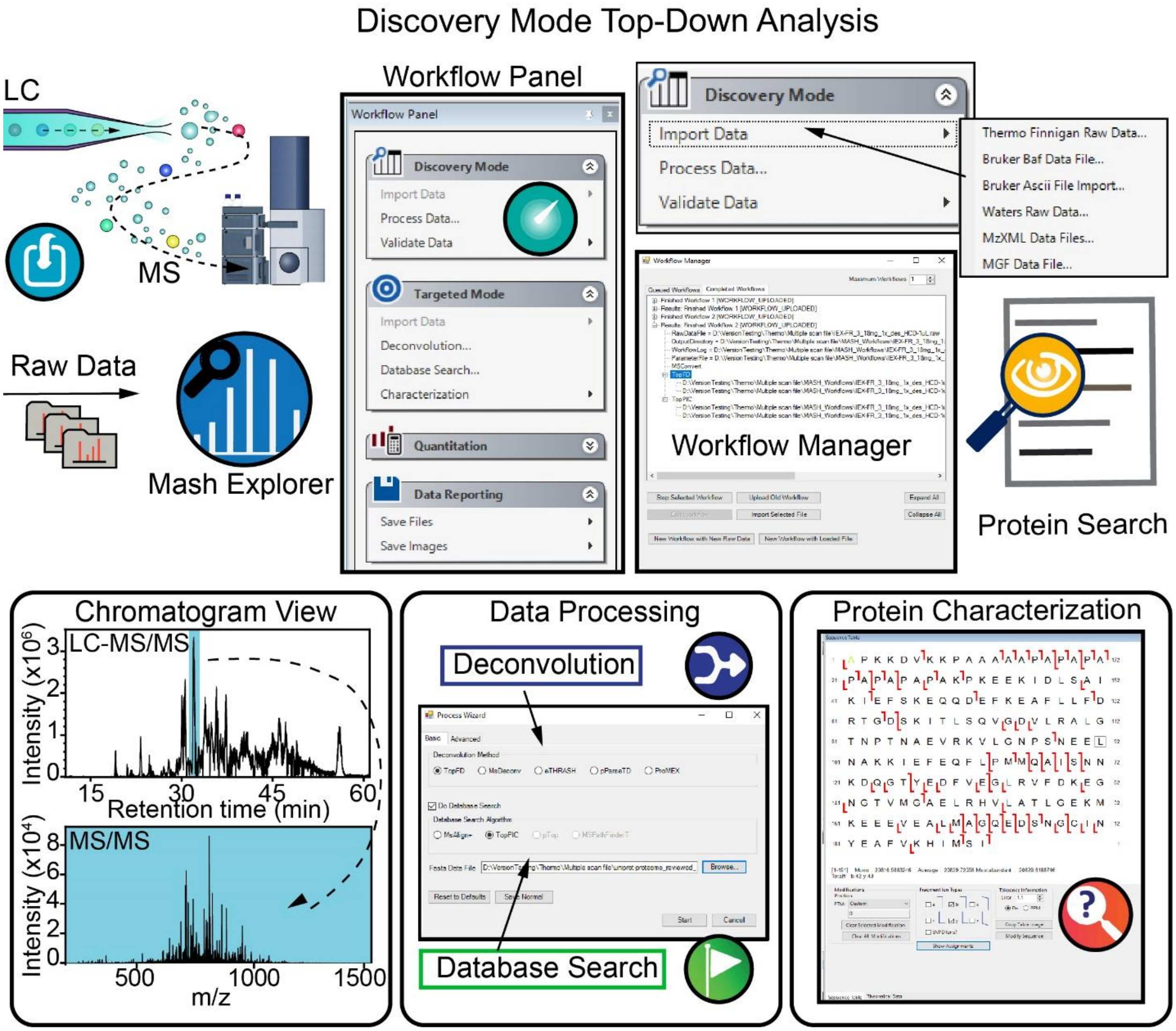


Figure 5.2. **Illustration of "Discovery Mode" for LC-MS/MS data processing.** "Discovery mode" can handle batch LC-MS/MS raw data files and includes features such as data import, data processing (deconvolution and database search), and data validation for protein identification. A simple and user-friendly Workflow Manager GUI automates the search and validation process and outputs processed data to a tabulated "Mass List" where users can view individual fragment ions and assign additional PTMs to reflect the fragment ion mapping on individual protein sequences.

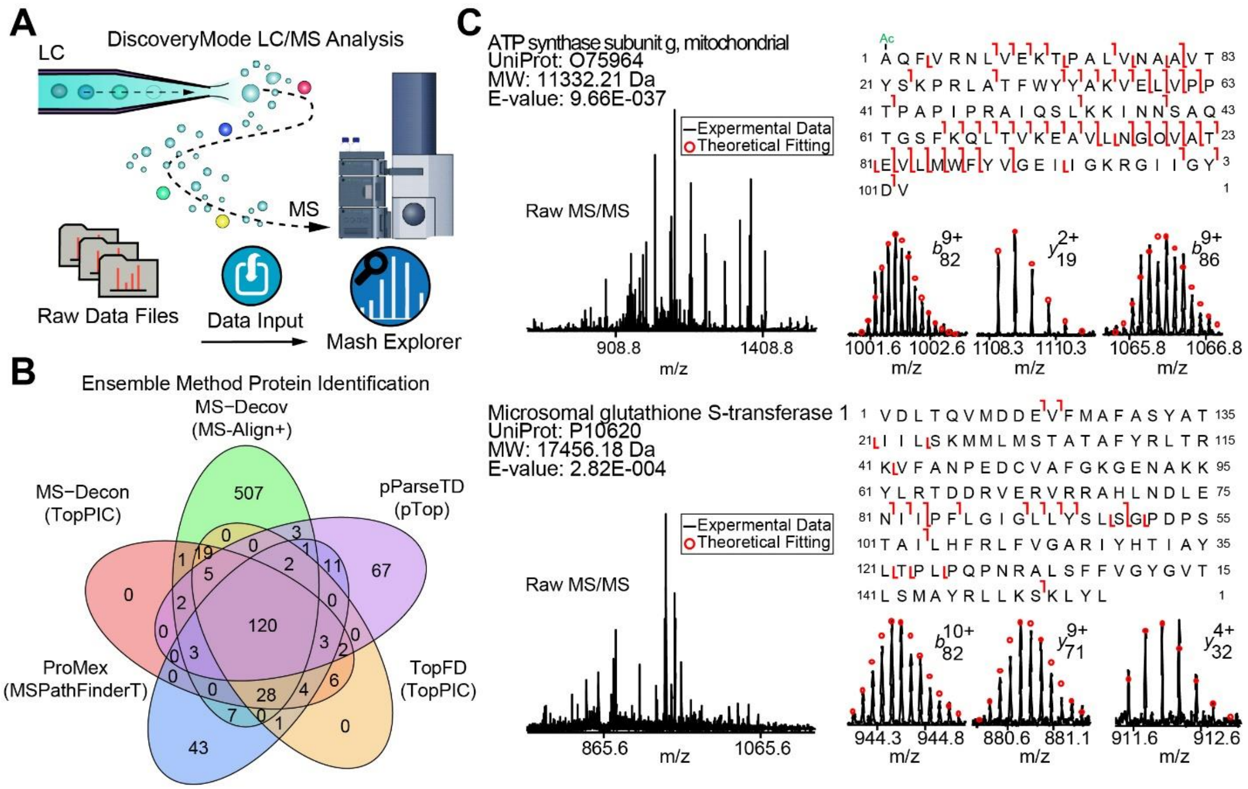


Figure 5.3. Top-down proteomics data analysis using "Discovery Mode" in MASH Explorer. **A**, Cartoon illustration of a typical "Discovery Mode" top-down LC-MS workflow. **B**, Venn diagram showing the overlap of protein identifications using an ensemble of five combined deconvolution and protein search workflows using a Thermo LC-MS/MS dataset. This combined deconvolution algorithm capability enables a deeper proteome coverage and enhanced protein identifications. **C**, Top-down MS identification and characterization using "Discovery Mode" workflow with ATP synthase subunit *g*, mitochondrial and microsomal glutathione S-transferase 1 shown as examples. The MS/MS spectra, sequence tables and fragment ions were output directly from MASH Explorer. Uniprot-Swissprot accession and protein E-value score are reported for each protein.

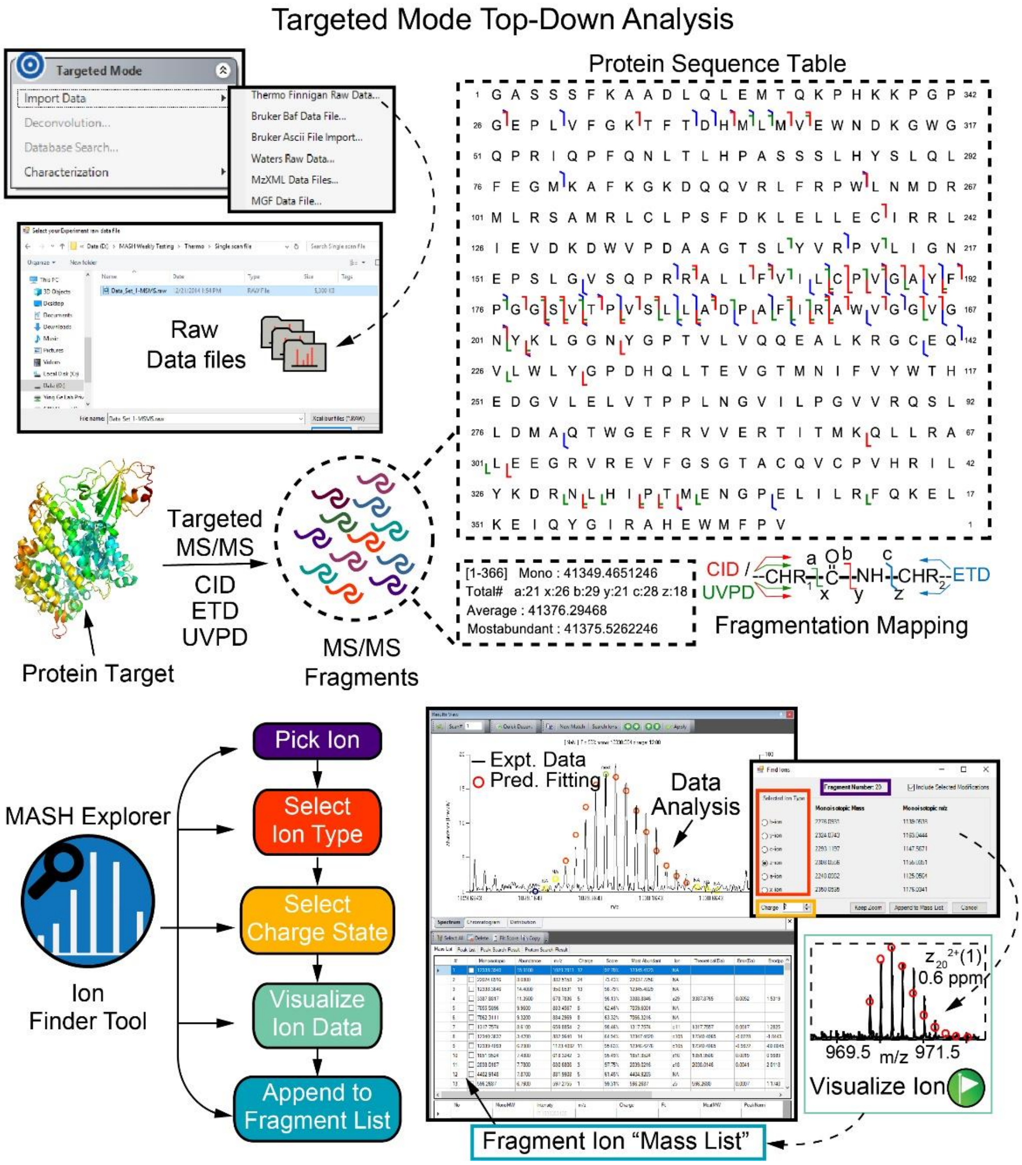

Targeted Mode Top-Down Analysis
Targeted Mode
Import Data
Deconvolution...
Database Search...
Characterization
Thermo Finnigan Raw Data...
Bruker Baf Data File...
Bruker Ascii File Import...
Waters Raw Data...
MzXML Data Files...
MGF Data File...
Raw Data files
Protein Sequence Table
1 G A S S S F K A A D L Q L E M T Q K P H K K P G P 342
26 G E P L V F G K T F T D H M L M V E W N D K G W G 317
51 Q P R I Q P F Q N L T L H P A S S S L H Y S L Q L 292
76 F E G M K A F K G K D Q Q V R L F R P W L N M D R 267
101 M L R S A M R L C L P S F D K L E L L E C I R R L 242
126 I E V D K D W V P D A A G T S L Y V R P V L I G N 217
151 E P S L G V S Q P R R A L L F V I L C P V G A Y F 192
176 P G G S V T P V S L L A D P A F I R A W V G G V G 167
201 N Y K L G G N Y G P T V L V Q Q E A L K R G C E Q 142
226 V L W L Y G P D H Q L T E V G T M N I F V Y W T H 117
251 E D G V L E L V T P P L N G V I L P G V V R Q S L 92
276 L D M A Q T W G E F R V V E R T I T M K Q L L R A 67
301 L E E G R V R E V F G S G T A C Q V C P V H R I L 42
326 Y K D R N L H I P T M E N G P E L I L R F Q K E L 17
351 K E I Q Y G I R A H E W M F P V 1
[1-366] Mono : 41349.4651246
Total# a:21 x:26 b:29 y:21 c:28 z:18
Average : 41376.29468
Mostabundant : 41375.5262246
CID / UVPD
a O b c
--CHR$_1$-C-NH-CHR$_2$--
x y z
ETD
Fragmentation Mapping
Protein Target
Targeted MS/MS
CID
ETD
UVPD
MS/MS Fragments
MASH Explorer
Ion Finder Tool
Pick Ion
Select Ion Type
Select Charge State
Visualize Ion Data
Append to Fragment List
— Expt. Data
○ Pred. Fitting
Data Analysis
Find Ions
Fragment Number: 20
Include Selected Modifications
Selected Ion Type
b-ion
y-ion
c-ion
z-ion
a-ion
x-ion
Monoisotopic Mass
Monoisotopic m/z
Charge
Keep Zoom
Append to Mass List
Cancel
$z_{20}^{2+}(1)$
0.6 ppm
969.5 m/z 971.5
Visualize Ion
Fragment Ion "Mass List"

Figure 5.4. **Illustration of "Targeted Mode" workflow for MASH Explorer.** "Targeted Mode" workflow includes data import, spectral deconvolution to identify and verify isotopic distributions, database search based on identified isotopic distributions, and proteoform characterization by matching identified isotopic distributions to the target proteoform sequence. "Targeted Mode" helps expedite PTM localization by a simple Ion finder Tool, which searches for fragment ions to confidently localize PTMs.

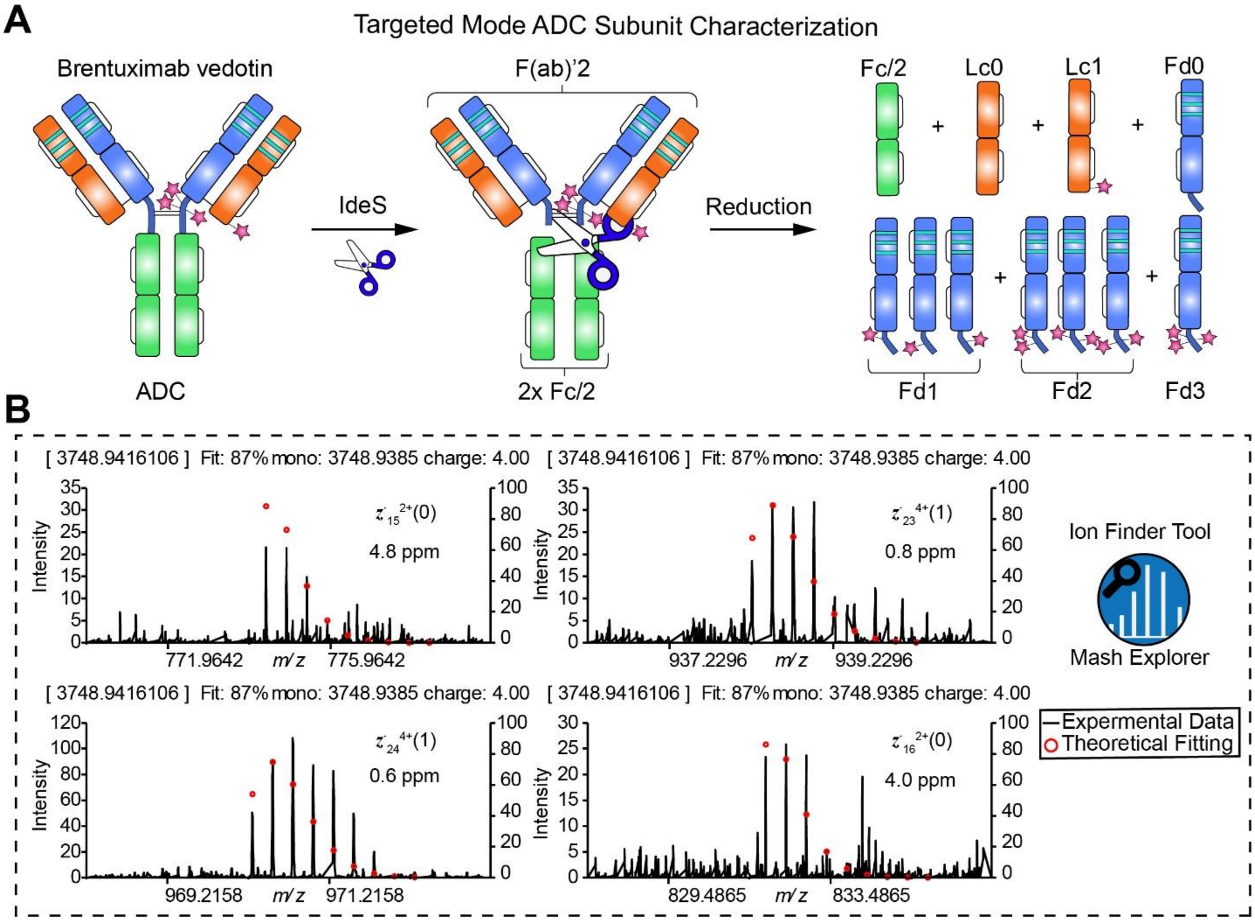


Figure 5.5. **Characterization of ADC subunits using "Targeted Mode" in MASH Explorer. A**, Intact ADC, brentuximab vedotin (Adcetris), is first subjected to IdeS digestion to cleave the hinge region and then further reduced to generate the ADC subunits (Fc/2, Lc0, Lc1, Fd0, Fd1, Fd2, and Fd3). **B**, The MASH Explorer Ion Finder Tool was used to search through candidate ions and generate fragment ion maps for the identification and localization of the site-specific drug conjugation site of a positional isomer of Fd1 subunit. The number in the parentheses represents the number of drug payloads included in the fragment ion.

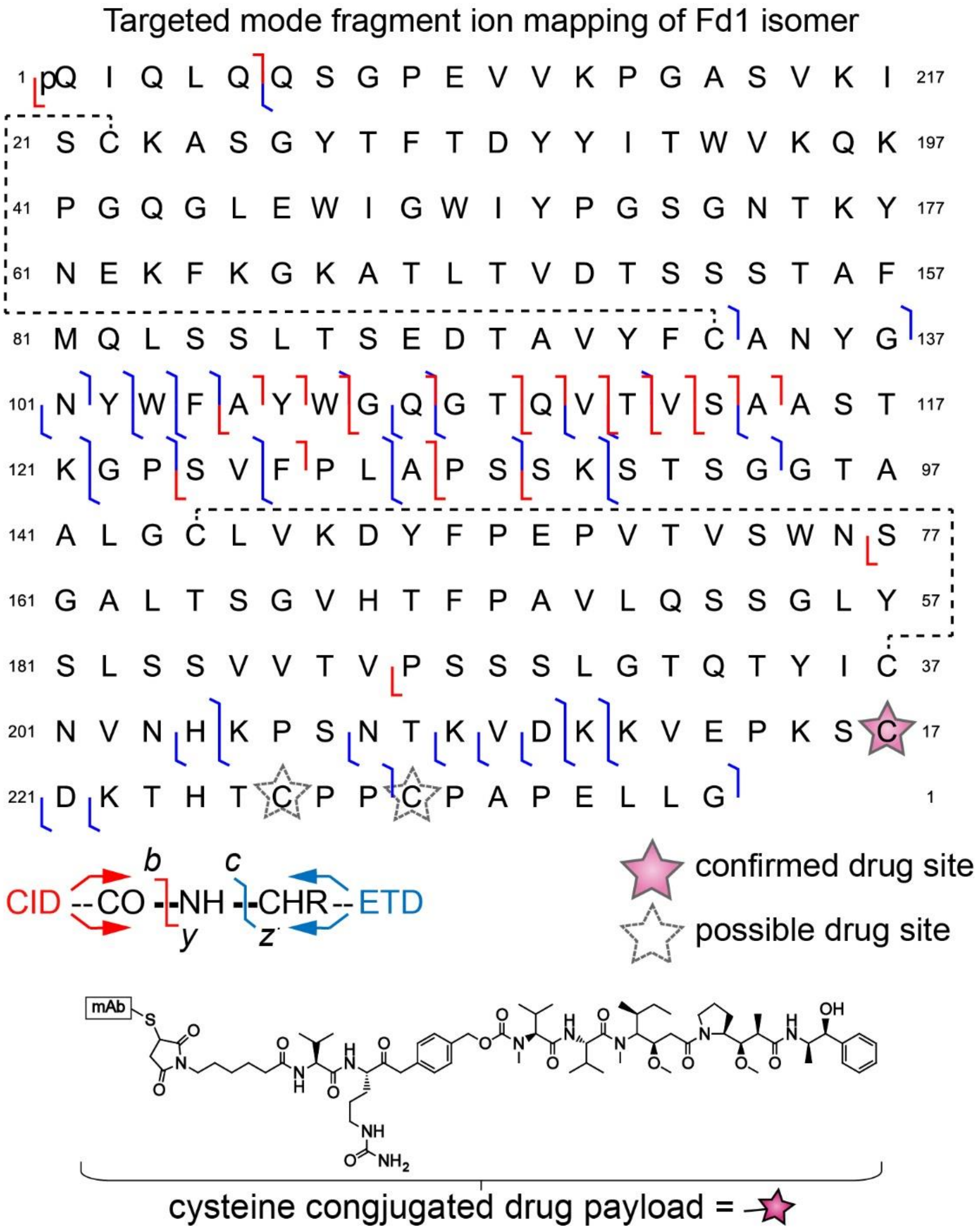

Targeted mode fragment ion mapping of Fd1 isomer
1 pQ I Q L Q Q S G P E V V K P G A S V K I 217
21 S C K A S G Y T F T D Y Y I T W V K Q K 197
41 P G Q G L E W I G W I Y P G S G N T K Y 177
61 N E K F K G K A T L T V D T S S S T A F 157
81 M Q L S S L T S E D T A V Y F C A N Y G 137
101 N Y W F A Y W G Q G T Q V T V S A A S T 117
121 K G P S V F P L A P S S K S T S G G T A 97
141 A L G C L V K D Y F P E P V T V S W N S 77
161 G A L T S G V H T F P A V L Q S S G L Y 57
181 S L S S V V T V P S S S L G T Q T Y I C 37
201 N V N H K P S N T K V D K K V E P K S C 17
221 D K T H T C P P C P A P E L L G 1
b
c
CID --CO --NH --CHR-- ETD
y
z
confirmed drug site
possible drug site
mAb
S
O
N
H
OH
NH
NH2
cysteine congjugated drug payload =

Figure 5.6. Protein sequence characterization and fragment ion mapping of Fd1 isomer from an ADC. Fragment ion map shows both CID and ETD fragment ions. Fragment ions were used to confirm the specific localization of a drug site of an Fd1 isomer. The pink star represents the cysteine-conjugated drug warhead corresponding to the Adcetris drug molecule. The data shown corresponds to the ADC fragmentation data shown in Figure 5.5.

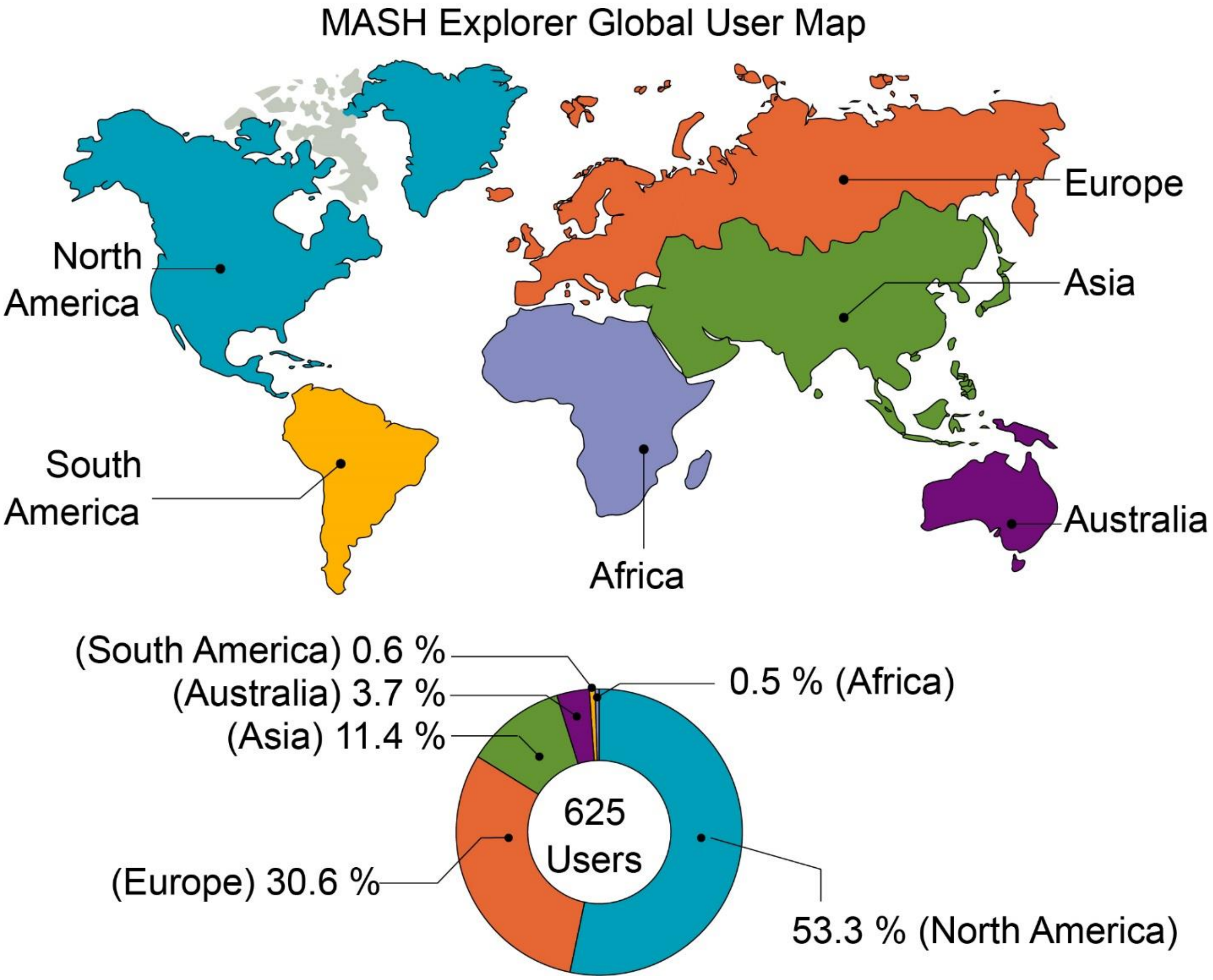
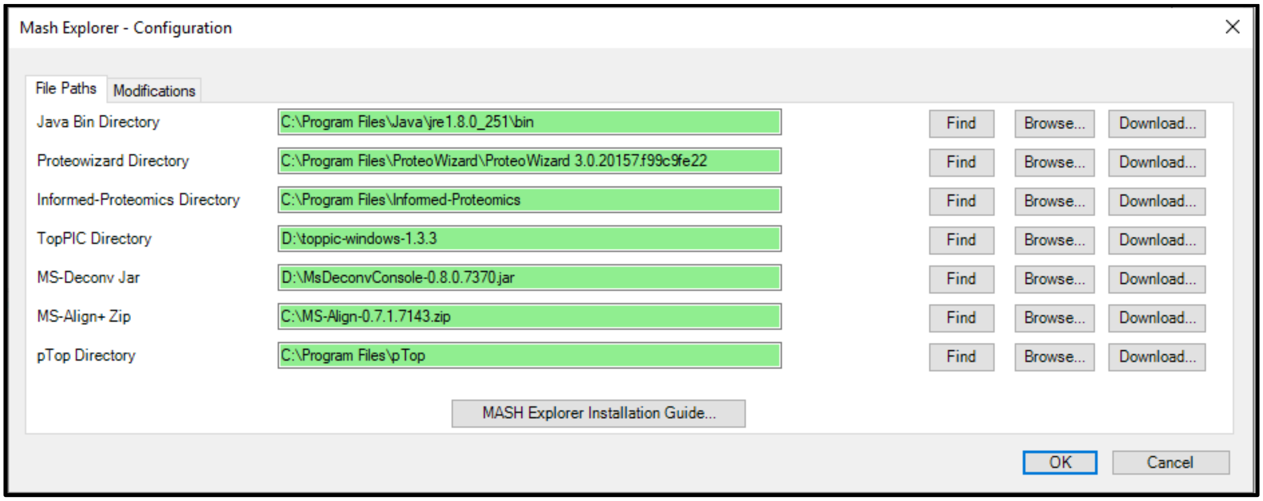


Figure 5.7. Cartoon schematic of a "world map" featuring the location distribution of MASH users across the globe. There are currently 625 active users (03/24/2020) with ~53% of users from North America, ~31% from Europe, and ~11% from Asia.

Supplemental Information

Table S5.1. Supported versions of deconvolution and database search tools.

Algorithm Tools	**Category**	**Supported Version**
TopFD[14]	Deconvolution	Up to 1.2.6
TopPIC[14]	Database Search	Up to 1.2.6
pParseTD/pTop[15]	Deconvolution and Database search	Up to 1.2
Informed Proteomics pipeline (ProMex and MSPathFinderT)[26]	Deconvolution and Database Search	Up to version 1.0.99
MS-Deconv[17]	Deconvolution	0.8.0.7370
MS-Align+[18]	Database Search	0.7.1.7143



Mash Explorer - Configuration

Java Bin Directory: C:\Program Files\Java\jre1.8.0_251\bin

Proteowizard Directory: C:\Program Files\ProteoWizard\ProteoWizard 3.0.20157.f99c9fe22

Informed-Proteomics Directory: C:\Program Files\Informed-Proteomics

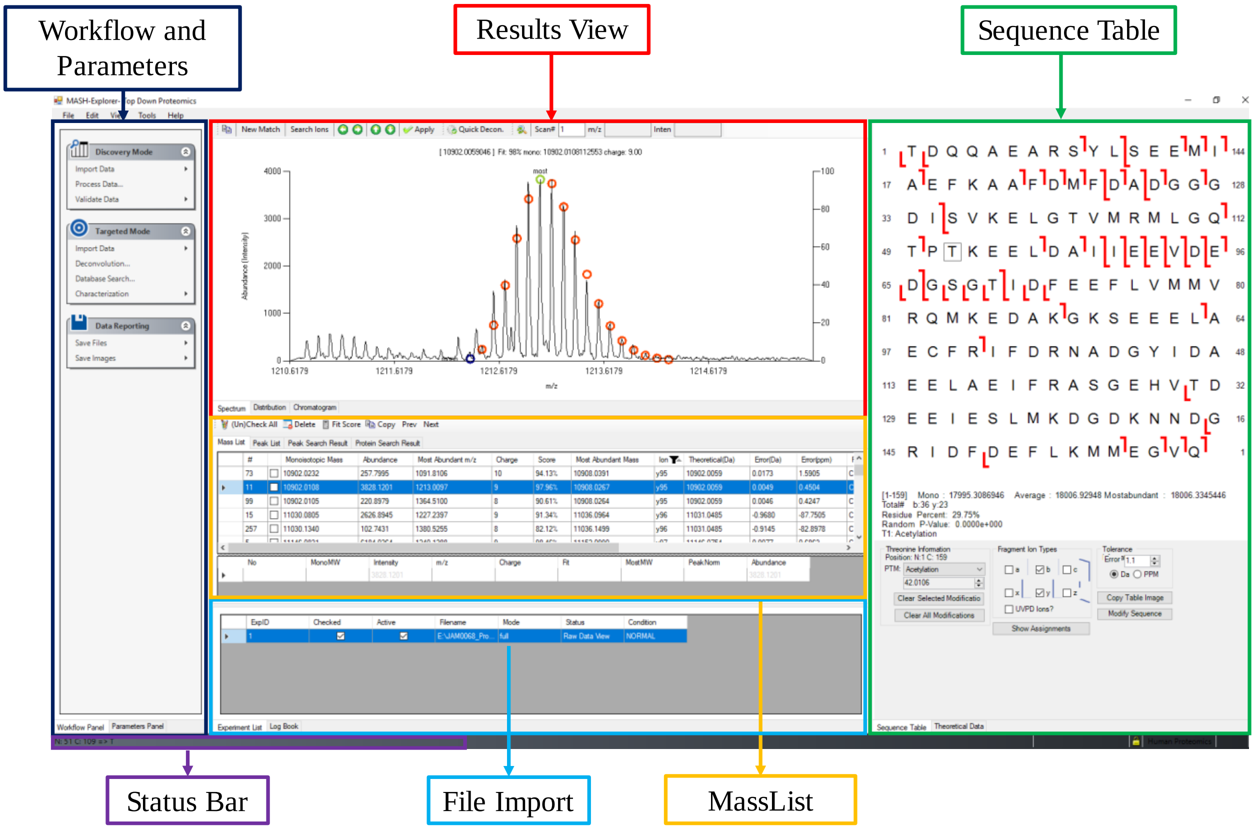
TopPIC Directory: D:\toppic-windows-1.3.3

MS-Deconv Jar: D:\MsDeconvConsole-0.8.0.7370.jar

MS-Align+ Zip: C:\MS-Align-0.7.1.7143.zip

pTop Directory: C:\Program Files\pTop

Figure S5.1. Software configuration. In the MASH Explorer, the Configuration tool provides an intuitive interface for the users to find the directory of the supported deconvolution and database search algorithms. Users can either use “Find” button to look for the default directory locations where the software was installed, or use “Browse” to manually locate the correct directory through a file browser dialog. Clicking the “Download” button will direct users to the website where the software can be downloaded. Directories found on the system are displayed with green background, while the unidentified directories are displayed in red. The Configuration tool can be found under Tools → Configuration.



[10902.0059046] Fit: 98% mono: 10902.0108112553 charge: 9.00

#		Monoisotopic Mass	Abundance	Most Abundant m/z	Charge	Score	Most Abundant Mass	Ion	Theoretical(Da)	Error(Da)	Error(ppm)
73	□	10902.0232	257.7995	1091.8106	10	94.13%	10908.0391	y95	10902.0059	0.0173	1.5905
11	□	10902.0108	3828.1201	1213.0097	9	97.96%	10908.0267	y95	10902.0059	0.0049	0.4504
99	□	10902.0105	220.8979	1364.5100	8	90.61%	10908.0264	y95	10902.0059	0.0046	0.4247
15	□	11030.0805	2626.8945	1227.2397	9	91.34%	11036.0964	y96	11031.0485	-0.9680	-87.7505
257	□	11030.1340	102.7431	1380.5255	8	82.12%	11036.1499	y96	11031.0485	-0.9145	-82.8978

No	MonoMW	Intensity	m/z	Charge	Fit	MostMW	PeakNorm	Abundance
		3828.1201						3828.1201

ExpID	Checked	Active	Filename	Mode	Status	Condition
1	☑	☑	E:\JAM0068_Pro...	full	Raw Data View	NORMAL

```
1   T D Q Q A E A R S Y L S E E M I  144
17  A E F K A A F D M F D A D G G G  128
33  D I S V K E L G T V M R M L G Q  112
49  T P T K E E L D A I I E E V D E   96
65  D G S G T I D F E E F L V M M V   80
81  R Q M K E D A K G K S E E E L A   64
97  E C F R I F D R N A D G Y I D A   48
113 E E L A E I F R A S G E H V T D   32
129 E E I E S L M K D G D K N N D G   16
145 R I D F D E F L K M M E G V Q      1
```

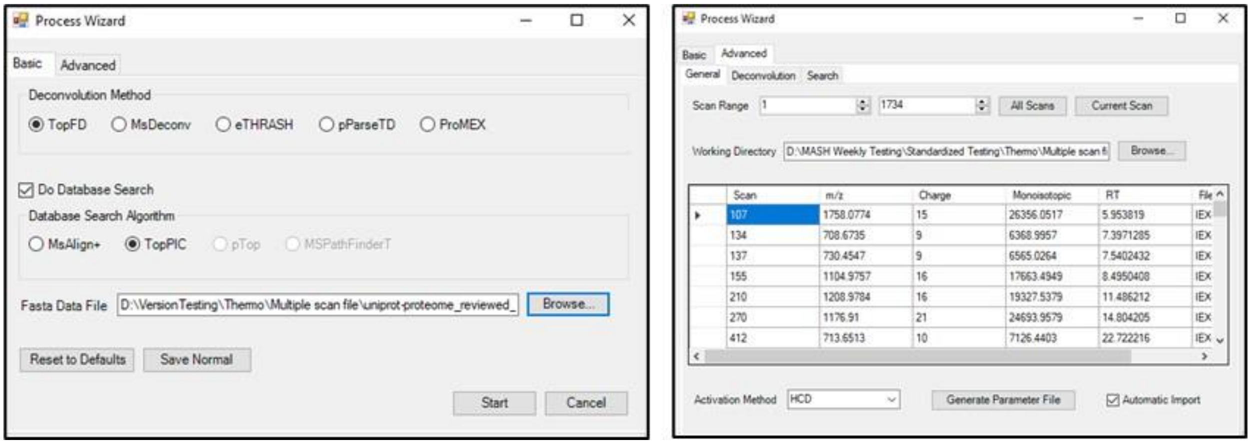
[1-159] Mono : 17995.3086946 Average : 18006.92948 Mostabundant : 18006.3345446
Total# b:36 y:23
Residue Percent: 29.75%
Random P-Value: 0.0000e+000
T1: Acetylation

Threonine Information
Position: N:1 C: 159
PTM: Acetylation
42.0106

Fragment Ion Types: □ a ☑ b □ c □ x ☑ y □ z □ UVPD Ions?

Tolerance: Error 1.1 ● Da ○ PPM

Figure S5.2. MASH Explorer main interface. Six main panels are shown. Workflow and Parameters panel handles all the core data processing. The Results View panel provides visualization of MS/MS, LC-MS, and LC-MS/MS data. The Mass List panel allows users to select deconvoluted fragment ions for manual processing. The Logbook and Status panels provide updates on the progress of data processing. The Sequence Table visualizes the fragment ions that match to the identified proteoform sequence.



Deconvolution Method

● TopFD ○ MsDeconv ○ eTHRASH ○ pParseTD ○ ProMEX

☑ Do Database Search

Database Search Algorithm

○ MsAlign+ ● TopPIC ○ pTop ○ MSPathFinderT

Fasta Data File D:\VersionTesting\Thermo\Multiple scan file\uniprot-proteome_reviewed_

Scan Range 1 1734

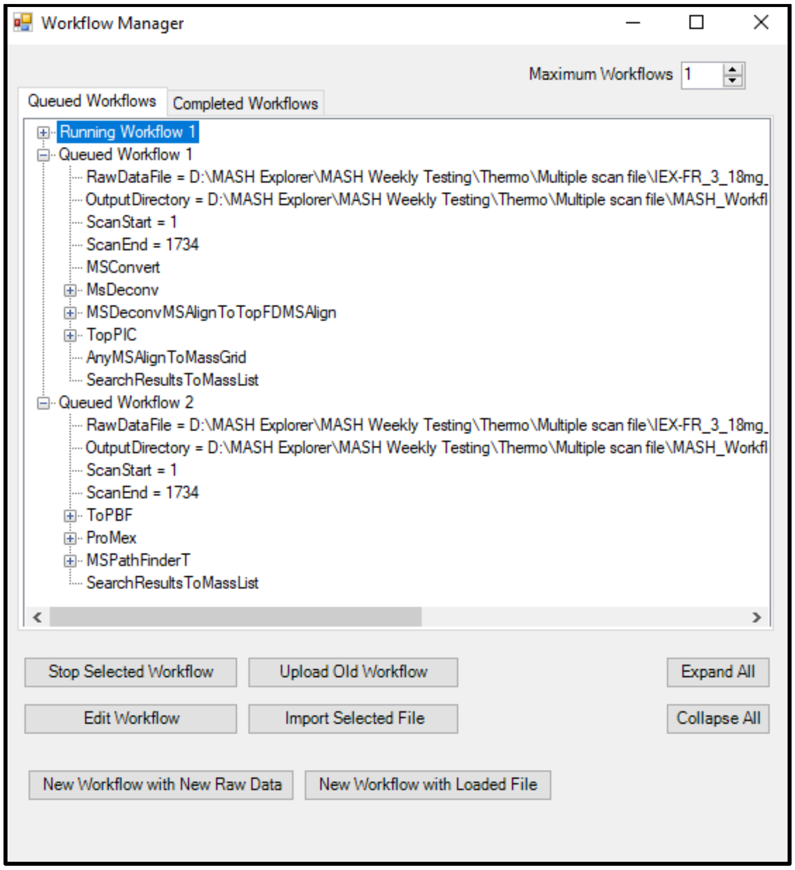
Working Directory D:\MASH Weekly Testing\Standardized Testing\Thermo\Multiple scan fi

Scan	m/z	Charge	Monoisotopic	RT	File
107	1758.0774	15	26356.0517	5.953819	IEX
134	708.6735	9	6368.9957	7.3971285	IEX
137	730.4547	9	6565.0264	7.5402432	IEX
155	1104.9757	16	17663.4949	8.4950408	IEX
210	1208.9784	16	19327.5379	11.486212	IEX
270	1176.91	21	24693.9579	14.804205	IEX
412	713.6513	10	7126.4403	22.722216	IEX

Activation Method HCD

☑ Automatic Import

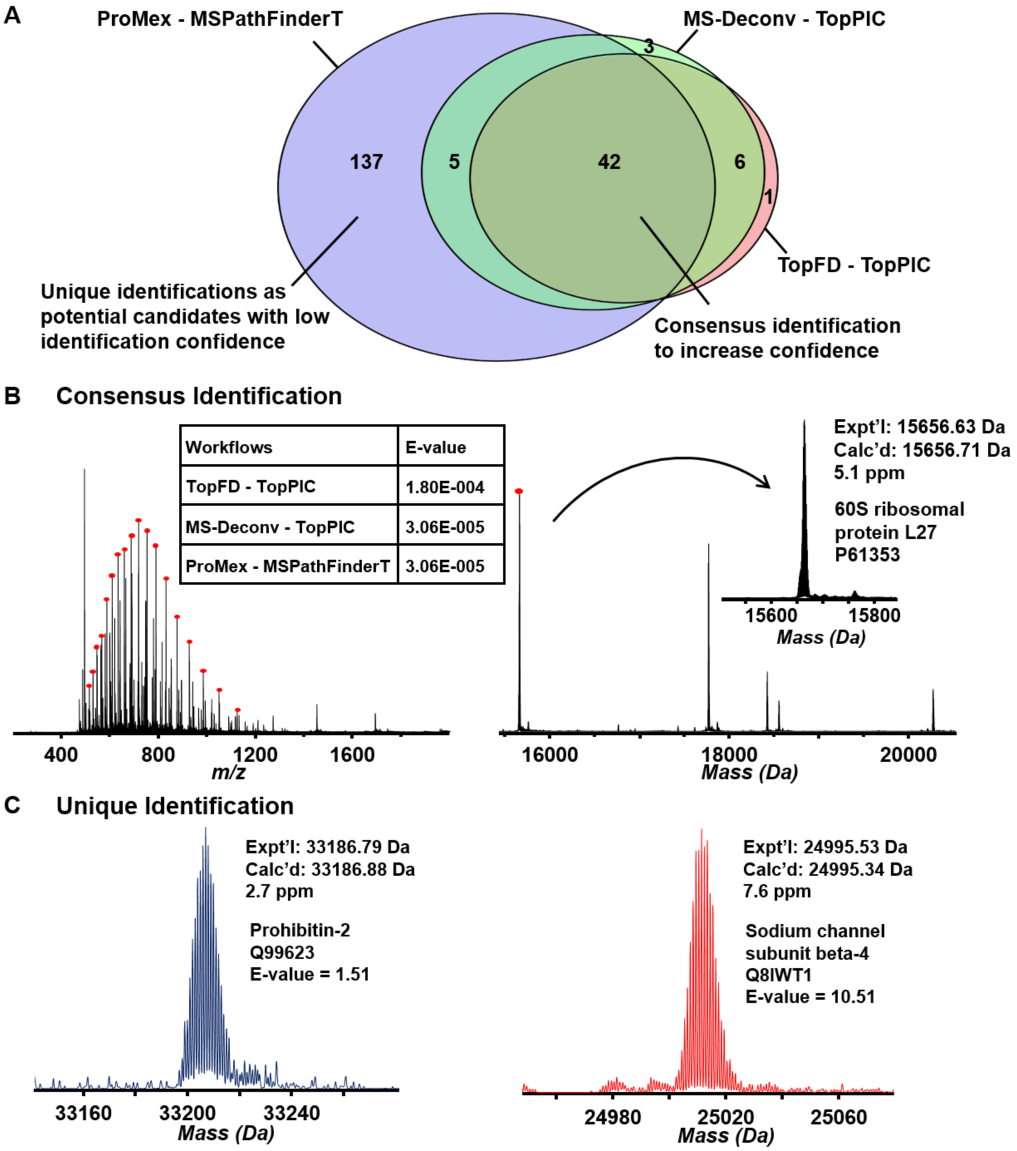
Figure S5.3. Process Wizard for top-down data processing. The Process Wizard provides an intuitive GUI for deconvolution and database search tasks in the top-down data processing workflows. Left, Basic tab of the Process Wizard. The Basic tab allows users to start deconvolution and database search tasks by selecting the radio buttons of the deconvolution and database search methods to start the data processing. Right, Advanced tab of the Process Wizard. The Advanced tab displays all the modifiable parameters of each algorithm.



Maximum Workflows 1

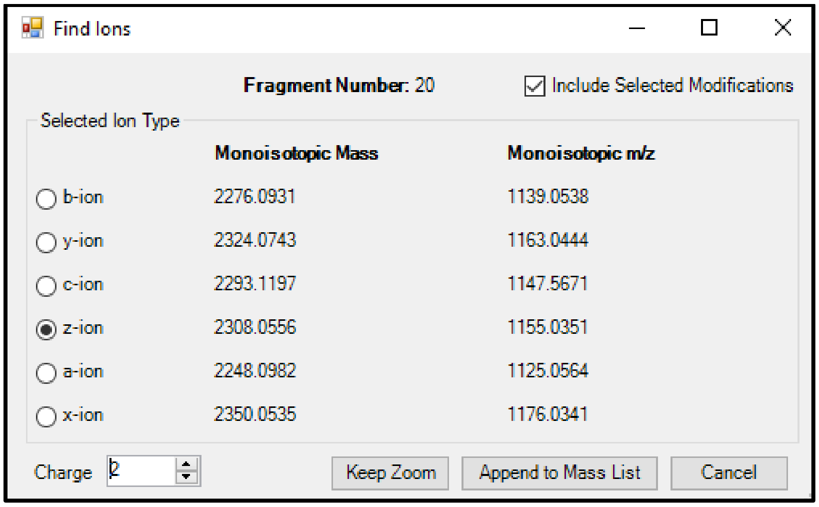
```
⊞ Running Workflow 1
⊟ Queued Workflow 1
    RawDataFile = D:\MASH Explorer\MASH Weekly Testing\Thermo\Multiple scan file\IEX-FR_3_18mg_
    OutputDirectory = D:\MASH Explorer\MASH Weekly Testing\Thermo\Multiple scan file\MASH_Workfl
    ScanStart = 1
    ScanEnd = 1734
    MSConvert
  ⊞ MsDeconv
  ⊞ MSDeconvMSAlignToTopFDMSAlign
  ⊞ TopPIC
    AnyMSAlignToMassGrid
    SearchResultsToMassList
⊟ Queued Workflow 2
    RawDataFile = D:\MASH Explorer\MASH Weekly Testing\Thermo\Multiple scan file\IEX-FR_3_18mg_
    OutputDirectory = D:\MASH Explorer\MASH Weekly Testing\Thermo\Multiple scan file\MASH_Workfl
    ScanStart = 1
    ScanEnd = 1734
  ⊞ ToPBF
  ⊞ ProMex
  ⊞ MSPathFinderT
    SearchResultsToMassList
```

Figure S5.4. Workflow Manager for batch analysis of multiple datasets. The Workflow Manager allows users to queue datasets for batch data analysis. This function enables higher throughput and efficiency for processing top-down proteomics datasets.



Workflows	E-value
TopFD - TopPIC	1.80E-004
MS-Deconv - TopPIC	3.06E-005
ProMex - MSPathFinderT	3.06E-005

Figure S5.5. "Discovery Mode" Analysis on a Bruker LC-MS/MS dataset. A. Venn diagram showed the overlap of protein identifications among three workflows. **B**. The MS1 spectrum and corresponding deconvoluted mass spectrum showed identified protein, 60S ribosomal protein L27, a consensus identification from three workflows. **C**. Deconvoluted mass spectra of prohibitin-2 and sodium channel subunit beta-4 were shown. These two proteins were uniquely identified by ProMex – MSPathFinderT algorithms with low confidence based on E-value provided by this workflow.



Fragment Number: 20

☑ Include Selected Modifications

Selected Ion Type

	Monoisotopic Mass	Monoisotopic m/z
○ b-ion	2276.0931	1139.0538
○ y-ion	2324.0743	1163.0444
○ c-ion	2293.1197	1147.5671
● z-ion	2308.0556	1155.0351
○ a-ion	2248.0982	1125.0564
○ x-ion	2350.0535	1176.0341

Charge 2

Figure S5.6. Demonstration of Ion Finder Tool. GUI of the Ion Finder Tool. The Ion Finder Tool allows users to search for a specific ion. For fragment ion number 20 shown, the user can select the six common ion types, including *a*, *b*, *c*, *x*, *y*, and *z* ions. The modifications can be included or excluded during this process.

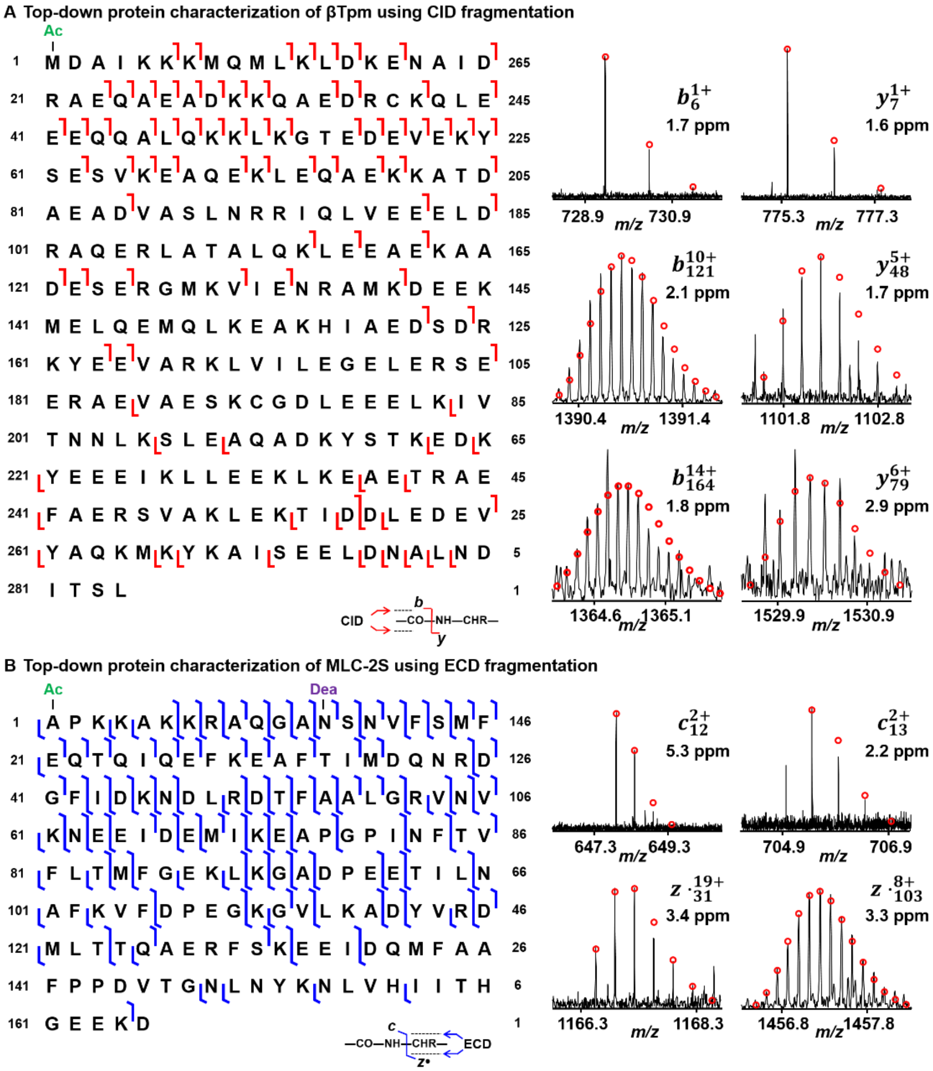
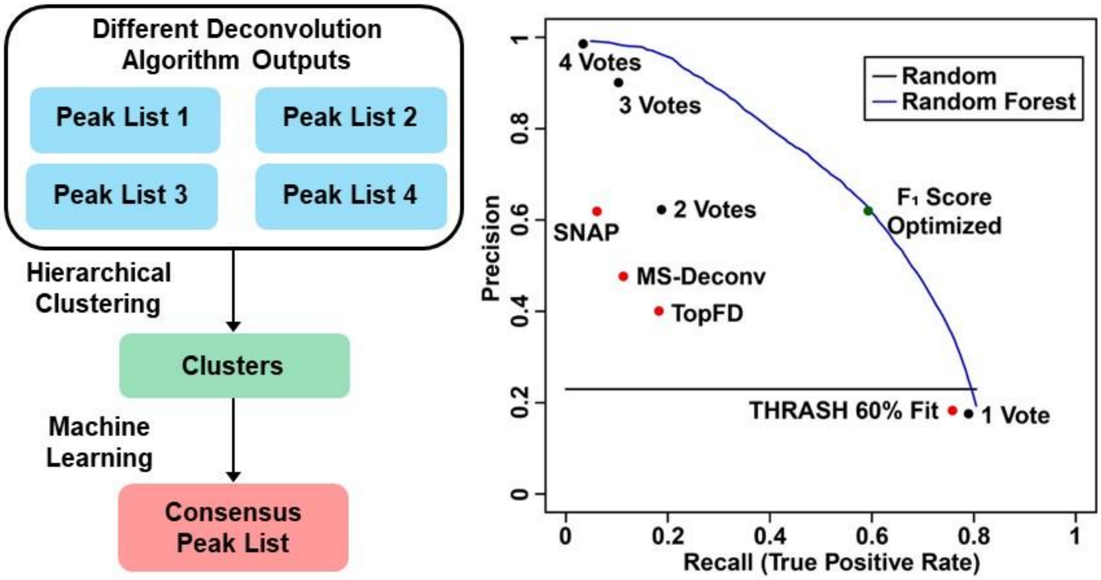


Figure S5.7. Top-down protein characterization using "Targeted Mode" workflow. A. Top-down protein characterization of beta-tropomyosin (βTpm, Uniprot-Swissprot accession number P07951) using CID fragmentation. CID fragment ions confirmed N-terminal acetylation. The MS/MS spectra, sequence tables, and fragment ions were output directly from MASH Explorer. B. Top-down protein characterization of myosin light chain isoform 2 slow isoform (MLC-2S, Uniprot-Swissprot accession number A0A1D5RDY5) using ECD fragmentation. N-terminal methionine removal, N-terminal acetylation, and deamidation at Asn13 were confirmed by ECD fragment ions.

Chapter 6

Enhancing Top-Down Proteomics Data Analysis by Combining Deconvolution Results through a Machine Learning Strategy



Adapted from: McIlwain, S. J.[#,*]; Wu, Z.[#]; Wetzel, M.; Belongia, D.; Jin, Y.; Wenger, K.; Ge, Y., Enhancing Top-Down Proteomics Data Analysis by Combining Deconvolution Results through a Machine Learning Strategy. *J. Am. Soc. Mass Spectrom,* **2020**, *31* (5), 1104-1113.

Abstract

Top-down MS is a powerful tool for identification and comprehensive characterization of proteoforms arising from alternative splicing, sequence variation, and PTMs. However, the complex dataset generated from top-down MS experiments requires multiple sequential data processing steps to successfully interpret the data for identifying and characterizing proteoforms. One critical step is the deconvolution of the complex isotopic distribution that arises from naturally occurring isotopes. Multiple algorithms are currently available to deconvolute top-down mass spectra, resulting in different deconvoluted peak lists with varied accuracy compared to true positive annotations. In this study, we have designed a machine learning strategy that can process and combine the peak lists from different deconvolution results. By optimizing clustering results, deconvolution results from THRASH, TopFD, MS-Deconv, and SNAP algorithms were combined into consensus peak lists at various thresholds using either a simple voting ensemble method or a random forest machine learning algorithm. For the random forest algorithm, which had better predictive performance, the consensus peak lists on average could achieve a recall value (true positive rate) of 0.60 and a precision value (positive predictive value) of 0.78. It outperforms the single best algorithm which only achieved a recall value of 0.47, and a precision value of 0.58. This machine learning strategy enhanced the accuracy and confidence in protein identification during database search by accelerating detection of true positive peaks while filtering out false positive peaks. Thus, this method show promise in enhancing proteoform identification and characterization for high-throughput data analysis in top-down proteomics.

Introduction

Top-down MS is a powerful tool for the identification and comprehensive characterization of proteoforms, including alternative splicing, sequence variations, and PTMs.[1-5] One of the unique advantages of top-down MS is the ability to analyze intact proteins without proteolytic cleavage to obtain the mass spectra of various proteoforms simultaneously and subsequently fragment the proteoform to locate the site(s) of modification.[6-7] A major challenge in top-down proteomics data analysis is the complexity of high-resolution top-down mass spectra.

The analysis of high-resolution top-down MS data requires several sequential processing steps, such as centroiding, deconvolution, proteoform identification, and quantification. Currently, many software tools have been developed to perform each step of the analysis process.[8] Deconvolution is a critical step early in the analysis, as the results can significantly affect the performance of the downstream methods. In addition to the first high-resolution deconvolution software THRASH,[9] other algorithms such as MS-Deconv,[10] TopFD,[11] pParseTD,[12] and UniDec[13] are also available for the deconvolution of top-down MS data. Furthermore, instrument vendors also provide deconvolution algorithms such as the SNAP algorithm[14] by the Bruker Corporation and the Xtract algorithm by Thermo Scientific within their software products.

Due to the diversity of deconvolution algorithms provided to the scientific community, one potential challenge an analyst may face is the non-standardization of their parameters. Consequently, the resulting peak list from different deconvolution algorithms cannot be directly compared. Moreover, different deconvolution algorithms performed spectral deconvolution using diverse computational methods, resulting in different peak list output. For instance, THRASH[9] is a subtractive peak finding routine that locates possible isotopic clusters in the spectrum by using least-squares fits to a theoretically derived isotopic abundance distributions. MS-Deconv[10] is a combinatorial algorithm that uses graph theory to find the heaviest path in the largest set of potential candidate envelopes. TopFD[11] is a successor to MS-Deconv which converts isotopomer envelopes to monoisotopic neutral masses after grouping top-down spectral peaks into isotopomer envelopes. The SNAP algorithm fits a function of superimposed bell curves to the peaks in order to identify the isotopic distributions. (Details regarding several common deconvolution algorithms were summarized in Table S6.1). Using a human histone dataset, Sun *et al.* showed that the peak list outputs among Xtract, MS-Deconv, and pParseTD had a maximum difference of 25% and 15% in the recalled peak rate and recalled intensity rate, respectively.[12] Finally, deconvolution algorithms may identify false positive peaks. The deconvolution results would need to be manually validated or corrected using software such as MASH Suite Pro,[15] which can be time consuming. As a consequence of all these challenges, there is a need for the standardization of different deconvolution algorithms as well as a method that analyzes and combines results from available deconvolution algorithms.

In the machine learning community, ensemble methods (e.g., simple voting) and machine learning algorithms (e.g., random forest algorithm) have been developed to enhance the predictions of multiple distinct algorithms in order to improve the overall predictive performance.[16-17] These ensemble methods and machine learning algorithms have also been employed in MS applications to improve the performance of disease diagnosis,[18] to improve target protein identification,[19] and to enhance the *de novo* peptide sequence.[20] In this study, by treating each deconvolution algorithm as a distinct algorithm, we propose that these ensemble methods and machine learning algorithms could be applied to combine different deconvolution results and obtain consensus peak lists. The resulting consensus peak lists should have higher accuracy, which will improve proteoform identification and mitigate manual validation efforts.

Herein, we report a novel use of machine learning strategy to combine the results from multiple deconvolution algorithms employed on high-resolution top-down MS to obtain consensus peak lists using an ensemble method and a machine learning algorithm. We compared and contrasted the predictive performance of our machine learning strategy against each deconvolution algorithm separately using a set of MS data that has been annotated by an expert to obtain a true positive list and showed improved performance over each individual algorithm. We demonstrated that adding more deconvolution results, even results from the same algorithm with different parameters, could further improve the predictive performance. Finally, we showed that the utility of the consensus peak list generated by our machine learning strategy could improve downstream proteoform identification using a software tool such as MS-Align+. This machine learning strategy will be integrated into our developing software, MASH Explorer,[21] a comprehensive and user-friendly tool for analyzing high-resolution top-down MS data.

Experimental Section

Skeletal Muscle Tissue Samples and Sarcomeric Protein Extraction

The collection and sarcomeric protein extraction method was previously published.[22] Briefly, biopsy samples of vastus lateralis (VL) skeletal muscle tissue were acquired from rhesus macaques at the Wisconsin National Primate Research Center using the protocols approved by the Institutional Animal Care and Use Committee of the University of Wisconsin-Madison. After dissection, the muscle tissues were immediately flash frozen and stored at −80 °C. For protein extraction, approximately 5 mg of tissue was first homogenized in 50 μL of HEPES extraction buffer to extract cytosolic proteins. The sample was centrifuged and the supernatant was removed. The pellet was subsequently resuspended and further homogenized in 50 μL TFA extraction solution to extract sarcomeric proteins. The sample was centrifuged and the resulting supernatant was used for LC-MS and MS/MS analysis.

Offline Fraction Collection and High-Resolution MS/MS for Protein Characterization

The fractions of some sarcomeric proteins from the tissue homogenate were separated using a homemade PLRP reversed-phase column (200 mm length × 500 µm i.d., 10 µm particle size, 1,000 Å pore size) with a nanoACQUITY UPLC system (Waters Corporation, Milford, MA, USA). PLRP-S particles were obtained from Agilent Technologies (Santa Clara, CA, USA). The fractions were subject to offline MS/MS to achieve a comprehensive characterization of the protein sequences and PTMs similar to the methods described in previous publications.[23-24] The collected protein fractions were analyzed by a 12-T solariX Fourier transform ion cyclotron resonance (FTICR) mass spectrometer (Bruker, Bremen, Germany) equipped with an automated chip-based nano-electrospray ionization source (Triversa NanoMate; Advion Bioscience, Ithaca, NY, USA). Targeted proteoforms were subject to both CID and ECD experiment. Typically, 100–500 transients were averaged for MS/MS experiments to ensure the collection of high-quality tandem mass spectra for protein characterization. The exact instrument and experimental settings can be found in previous publication.[22] The mass spectrometry proteomics raw data and annotations have been deposited to the ProteomeXchange Consortium via the PRIDE[25] partner repository with the dataset identifier PXD018043. The protein identification, accession number, and PTMs were provided in Table S6.2.

Peak Extraction and Expert Annotation

Deconvoluted peaks were identified by four different algorithms, including THRASH,[9] MS-Deconv,[10] TopFD,[11] and the SNAP algorithm from Bruker DataAnalysis,[14] which were available for processing the Bruker data set. The peak extraction using the MASH Explorer software was executed with the THRASH algorithm using fit parameters of 60, 70, 80, and 90%. The MS-Deconv algorithm was run using default parameters with a maximum charge of 30, a maximum mass of 50,000, an m/z error tolerance of 0.02, and an S/N ratio of 3. The TopFD deconvolution was employed using default parameters with a maximum charge of 30, an MS1 S/N ratio of 3.0, a precursor window size (m/z) of 3.0, a maximum mass (Da) of 100,000, an MS2 S/N ratio of 1.0, and an m/z error of 0.02. Using DataAnalysis available for the Bruker dataset, the deconvoluted ion list was obtained using the SNAP algorithm with a quality factor threshold of 0.1, an S/N threshold of 2, a relative intensity threshold (base peak) of 0.01%, an absolute intensity threshold of 0, and a maximum charge state of 50. All deconvolution results were output into MSAlign format, which provides information of the monoisotopic distributions including monoisotopic mass, intensity, and charge. While this manuscript focused on data acquired using Bruker instruments, this method is applicable to datasets from other vendors if the peak information was converted to MSAlign format.

Coding Environment

Python (2.7.10) was used to generate the clusters, and R (3.6.0) was used to perform the machine learning analysis and to automate the MS-Align+ searches.

Machine Learning Strategy for Combining Multiple Deconvolution Results

A general overview of the data analysis process is provided in Figure 6.1. Each MSAlign file was parsed by a Python script, and the results were concatenated into one peak list which records the monoisotopic mass, charge, and source algorithm. Peaks having the same charge and similar monoisotopic mass were clustered together as the same peak. The clusters were then filtered using either a simple voting or a machine learning methods, and the results were output into a consensus MSAlign file. Each part of the process is described in more detail below.

Hierarchical Clustering - The algorithm merges the full list of deconvoluted peaks into clusters that contain the same charge and are similar in monoisotopic mass. Inspired by Robert Tibshirani's work on 'peak probability contrasts',[26] the method uses hierarchical clustering[27] with the difference between pairs of peaks from the $\log_{10}$ transformed monoisotopic mass as the distance metric. Transforming the monoisotopic mass using log removes the linear dependence of the error with mass, so a constant cutoff can be used to determine the number of clusters. A further constraint was added to ensure that the charges are the same between peaks with the proposed clusters. Using Equation 1, a cutoff was determined using a user-defined threshold ppm error within the cluster, which ensured that the distances between the largest and smallest mass of the peaks within the cluster were not larger than the ± ppm threshold. The average of the monoisotopic mass was then used as the center of the cluster. The clustering algorithms was run on each spectrum separately.

$$\mathrm{Cutoff(ppm)} = \log_{10}(2.0\,\frac{\mathrm{ppm}}{10^6} + 1) \qquad \text{Equation 1}$$

Expert Annotation and Assignment to Clusters - The expert annotations were obtained and verified manually using the MASH software with the embedded enhanced-THRASH algorithm at 60% fit setting.[15] Some peaks were manually validated by adjusting the most abundant m/z and charge state of each monoisotopic distribution. In this study, we considered expert annotated peaks to be true positive peaks.

The identified clusters were annotated using the expert annotations by finding clusters that had the same charge and were within a ±X ppm window of the expert annotated monoisotopic mass (where X is set to the same value as used in the clustering). In cases where an expert peak could be assigned to multiple clusters, we selected the pair with the smallest distance between the monoisotopic mass with expert assignment that matched to multiple possible clusters as the true match. Clusters with assigned expert annotation were called expert matched peaks, and the unassigned clusters were labelled as unmatched expert annotated peaks.

Machine Learning Analysis - The machine learning analysis was performed using the R language. For each cluster, a feature vector was generated using the features described in Table 6.1. We set up a machine learning task to separate expert annotated matched clusters from unmatched clusters. Precision-recall curves were estimated by leave-one-spectrum-out cross-validation, where each fold estimates the probability of a true annotation for each of the clusters found in one spectrum using a machine learning model built from the other spectra feature vectors. Recall and precision are defined in Equation 2 and 3,

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \qquad \text{Equation 2}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \qquad \text{Equation 3}$$

where a true positive is a cluster peak with an expert annotation, a false negative is an unmatched expert annotation, and a false positive is a cluster peak without an expert annotation. Recall measures the percentage of expert annotations found by the algorithm, whereas precision measures the percentage of cluster peak calls that have true annotations. We compared and contrasted the predictive performance using the random forest (randomForest R package)[28] model using ntree (the number of trees used in the forest) of 100 and the remaining parameters set to their defaults.

To compare the individual algorithms on the precision-recall curves, all of the true positive, true negative (indicating cluster peaks with no expert annotations that algorithms did not call as annotations), false positive, and false negative results were aggregated before calculating precision and recall values. Deconvolution methods were also compared by calculating the F_1 score, a metric that balances precision and recall as defined in Equation 4. For the random forest, we selected the probability threshold that maximizes the F_1 score within the training dataset to make the final calls on the associated test set.

$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad \text{Equation 4}$$

Due to the high rate of false positives within the datasets, we used precision-recall curves to visualize the accuracy of the methods. Typically, precision-recall curves have a point for a recall value of 1 and a precision value of the ratio of true positives over all cluster calls. However, if we count the false negatives incurred by the upstream clustering method, then the curve will give a lower maximum achievable recall result. A superior performing classification algorithm would have a point (or curve) that is higher in precision and recall (i.e., more top-right) than for the contrasted algorithm(s). For example, the point for one deconvolution algorithm which gives a recall of 0.50 and a precision of 0.40 is outperformed by the point for the second deconvolution algorithm that gives a recall of 0.60 and a precision of 0.50.

Cluster Filtering and Consensus Deconvolution Results - To reduce false positive clusters, we explored two avenues of filtering. One was a simple voting heuristic that thresholds clusters based upon the number of deconvolution algorithms that called the peaks within the cluster.

Another route was to apply the previously described machine learning models to assign a probability of a true expert assignment to each cluster. The clusters were filtered via thresholding upon this probability. Consensus results were output as an MSAlign file and were processed using a database search algorithm.

Database Searching - All searches were performed using MS-Align+ v0.7.1.7143[29] with a fasta database file derived from a human database (Uniprot-Swissprot database, released December 2019, containing 20,367 protein sequences) for βTpm, a cynomolgus monkey database (Uniprot-Swissprot database, released January 2020, containing 77,341 protein sequences) for fsTnT5, a rat database (Uniprot-Swissprot database, release January 2020, containing 8,085 protein sequences) for αTpm, and a rhesus macaque database (Uniprot-Swissprot database, released January 2020, containing 78,285 protein sequences) for the rest of the proteins. We compared and contrasted the search results of MS-Align+ using the MSAlign file from each deconvolution algorithm, the expert annotated peaks, the simple voting method, and the random forest machine learning method.

Results and Discussion

Setting clustering ppm cutoff and clustering choices

One of the important parameters to determine for the machine learning strategy is the choice of the ppm cutoff for calling clusters. To determine the optimal cutoff for the data set presented in this work, we evaluated four parameters [the number of clusters, the percentage of peaks assigned, the cross-validated accuracy from the random forest model (the percentage of correct annotations found over the whole data set), and the random forest's F_1 score (a measure of accuracy that is the harmonic mean of precision and recall)] at multiple cutoff levels (1, 2, 5, 10, 20, 50, 100, and 200 ppm). The results in Figure 6.2 demonstrated that 10 ppm was optimal for the clustering cutoff because (1) for values greater than 10 ppm clustering cutoff, there was a noticeable drop in the number of clusters (Figure 6.2a), (2) the percent of recalled peaks was not significantly less than that from higher ppm cutoff but was greater than that from a lower ppm cutoff (Figure 6.2b), and (3) the accuracy and the overall accuracy measured by F_1 score did not differ significantly from the optimal values in both measurements (Figure 6.2c and 6.2d).

Many other clustering algorithms exist in the literature, including different linkage algorithms for hierarchical clustering.[30] In this work, we used complete hierarchical clustering, which gives tight clusters (min/max rather than the average). This is desirable for merging peaks by monoisotopic mass.

Expert annotation accuracy performance with 4-vote ensemble

After determining the optimal hierarchical clustering cutoff, the peak clusters were analyzed by ensemble/machine learning methods and individual deconvolution algorithms for comparison. In this study, we used a simple voting ensemble method which is based upon the number of unique deconvolution algorithms that called a peak within that cluster. Additionally, the random forest machine learning algorithm, which is itself an ensemble of decision trees, was utilized.[28] The random forest algorithm was shown to be able to handle large data sets and exhibit excellent performance in the classification tasks.[31] There are several other classification methods available, such as support vector machines[32] and deep learning models.[33] However, these algorithms may be difficult to tune, and deep machine learning requires numerous examples in order to learn an adequate network structure for optimizing predictive performance.

The aggregate predictive performance among individual deconvolution algorithms, the simple voting method, and the random forest machine learning algorithm are summarized in Figure 6.3. A majority vote (two or more votes, point “2 Votes”) appeared to outperform any single deconvolution method used by itself. Compared to SNAP (point “SNAP”) and TopFD (point “TopFD”) algorithms, a majority vote (two or more votes) had better recall and precision, respectively. The Venn diagram between a majority vote (two or more votes) and its overlap with expert annotation is shown in Figure S6.1. Although THRASH with 60% fit identified a total of 50,381 peaks, 41,204 of them (82%) were false positives because they were not those identified by the expert annotations. Filtering out the false positives accounted for the improved accuracy in the majority vote (two or more votes). On the other hand, this majority missed 7,181 peaks from the THRASH with 60% fit, out of 12,264 peaks (59%) that were expert annotated peaks, contributing to the low recall values. To provide a reference for the random forest method, we calculated the aggregate precision and recall score using a probability threshold cutoff that optimizes the F_1 score on the training spectra and applied it to the corresponding test set. The aggregated precision and recall value from the random forest method shown as a green point in Figure 6.3 is superior to most of the methods.

Furthermore, the results suggest that the random forest algorithm could achieve superior performance for identifying clusters which are true expert annotations. To determine average metrics (precision and recall) for random forest’s performance, the probability threshold cutoff that optimizes the F_1 score was determined in each training fold feature set. The probability threshold was then applied to the associated test fold dataset, and the resulting performance metrics were calculated. The final precision and recall were determined by averaging the results across each testing fold. Using this process, the random forest model achieved an average recall of 0.49, a precision score of 0.69, and an F_1 score of 0.55. In comparison, THRASH with 60% fit, which was the best algorithm by its F_1 score, achieved a recall of 0.76 and precision of 0.18, with an F_1 score of 0.30. Additional metrics including the median, first and third quartiles, and minimum and maximum of the F_1 score across the different deconvolution methods were compared, and the random forest model outperformed other algorithms (Figure S6.2).

A useful aspect of the random forest model is the ability to extract feature importance values. One of the metrics that the random forest can report for each feature is the mean decrease accuracy, which is an estimate of the reduction in the accuracy performance of the machine learning algorithm upon permuting the values of the current feature. The features ranked at the top of the plot reduce the accuracy of the model most significantly when permuted, and these features are considered to be the most important ones. In Figure S6.3, cluster features such as the average mass of the cluster (AvgMass), the cluster charge (charge), and the average intensity (AvgIntensity) had the most significant impact on the model, indicating that the random forest model was learning some of the spectral features such as charge and mass ranges that contribute to a true positive cluster. Features describing characteristics of the spectrum (i.e., activation type, precursor mass, and precursor charge) had a greater influence on the performance of the random forest classifier over the simple voting model. Using the vote of each deconvolution algorithm in the random forest model also provides a way to learn the confidence in each algorithm to determine an optimal score (THRASH with 60% fit, MS-Deconv, TopFD, and SNAP). While these features did not rank high in the list, the THRASH with 60% fit feature seemed to have the greatest effect on the model performance over the other deconvolution algorithms. This is possibly due to the number of proposed peaks that the THRASH with 60% fit finds in conjunction with the other features (spectrum characteristics and cluster features) to find the best scoring clusters within all of the false positive peaks (clusters with an expert annotation).

Backward selection, which iteratively removes features in model performance optimization, is an alternative route to determine feature importance. Performing a full backward selection process with leave-one-spectrum-out cross-validation and optimizing on the median F_1 score (Figure S6.4), we found that charge, precursor charge, THRASH with 60% Fit, AvgMass, and SumIntensity (five of the features shown in Table 6.1) can achieve the same performance as a model built from all of the features. Omitting one of these five features showed a significant decrease in the median F_1 score. Moreover, features such as AvgIntensity, votes, and SumIntensity are correlated by definition. Consequently, removal of two of these features would be sufficient for the discriminatory models.

Expert annotation accuracy performance with seven-vote ensemble

To test the hypothesis that more orthogonal deconvolution algorithms can further improve results, we generated the results using THRASH with different fit score parameters as separate deconvolution algorithms (Figure 6.4). Comparing the peak call results from four THRASH algorithms with different fit scores, we noticed that no result directly subsumed the peak calls of any of the others, which indicates some degree of orthogonality among the different THRASH results (Figure 6.4a). The discordant results from THRASH was not surprising since THRASH heuristically finds isotope envelopes. That is, isotopic distributions found in the beginning of the THRASH algorithm can affect the peaks found later during the algorithm process. With the additional deconvolution algorithm results added to the method, our results showed an increase in the number of assignments of clusters to annotated expert peaks and an increase in the filtering performance (Figures 6.4b and 6.4c). Additionally, other metrics using the seven-vote ensemble including the number of clusters, F_1 score, and the number of recalled peaks were also improved compared to those using the four-vote ensemble (Figure S6.5). In comparison with the average performance as in the four-vote ensemble, the random forest model from the seven-vote ensemble achieves an average recall or true positive rate of 0.60, a precision score of 0.78, and an F_1 score of 0.67. After calculating the recall and precision for the individual algorithms, the best algorithm (by F_1 score) was found to be THRASH with 90% fit, which achieved a recall of 0.47 and precision of 0.58 with an F_1 score of 0.52. Since there was an increase in the number of unassigned clusters (potentially false positives) in the four-vote ensemble (56,363) vs four-vote ensemble (45,117), it suggests that the seven-vote method learned to filter out false positives more accurately than the four-vote system (Table S6.3 and S6.4).

In summary, adding more deconvolution algorithms has the potential of increasing the identification of peaks potentially missed by other deconvolution algorithms and to improve the classification performance to filter out more false positives. Two additional deconvolution algorithms including pParseTD and UniDec, which are based on online support vector machine algorithm and a Bayesian algorithm, respectively, will be ideal for the continual development of this machine learning strategy due to the differences in algorithmic approaches compared to the four deconvolution algorithms used in this study. However, pParseTD currently only processes Thermo datasets, and the output peak list requires additional processing to assign charges for the isotopic distribution to locate the deconvoluted peaks in the spectrum. UniDec is optimized for native mass spectrometry where proteins and their fragment ions typically carry lower charges relative to mass compared to those under denatured conditions. Additional efforts are needed to incorporate these two algorithms into the machine learning strategy described in this study.

Unmatched clusters and missed expert annotations

When investigating the missed expert annotations (false negatives) and the unassigned clusters (false positives) from the machine learning strategy, two key observations surfaced. First, the unassigned clusters might actually be real isotopic distributions. Second, the corrected isotopic distributions may introduce both a false positive and false negative calls into the analysis.

There are cases where the unassigned clusters may actually be real isotopic distributions that the manual annotator could have missed due to low abundance. These low-abundance isotopic distributions might also suffer from an imperfect distribution due to the noise. Figure 6.5a gives two examples of low-abundance isotopic distributions that could be real annotations. This indicates that the method would be useful in proposing other annotations within data.

During manual annotation and correction, there are many instances where the annotator has to correct the charge and/or peak of the most abundant mass. Figure 6.5b provides an example of an annotation that has been corrected by an expert annotator. Annotations that have been corrected in this way may introduce both a false positive and a false negative into the method analysis. The false positive would arise from the original peak without the correction from the deconvolution method, and the false negative would come from the corrected peak in the expert annotations.

The ability to shift the charges and most abundant mass is an area of continual research in this project, in order to identify more expert annotations without incurring additional false positives. For example, generating the expert annotated results for the αTpm protein with ECD activation required the expert annotator to remove 52% (840 of 1,631 peaks), adjust the charge state for 7% (109 of 1,631 peaks), and shift the monoisotopic mass for 2% (38 of 1,631 peaks) from the deconvoluted peaks found by the THRASH algorithm with 60% fit. The machine learning strategy did succeed in reducing the false positive rate, but making additional modifications to identify and fix the annotations would further reduce the time spent on manual verification and peak correction.

Effects of improved deconvoluted peaks on database searching results

To investigate whether using the machine learning strategy can help with protein identification, we compared the MS-Align+ database search results from peak lists generated by different deconvolution algorithms and machine learning methods. Using the ECD spectrum of the αTpm proteoform, we evaluated and plotted the database search results using the deconvoluted results from expert annotation, TopFD, simple voting method, and random forest model (Figure 6.6 and Table S6.2). The E-value metric was utilized to evaluate the confidence of protein identification, with a lower E-value indicating high identification confidence. In the figure, the $-\log_{10}$ value of the E-value was used for visualization instead in the *y* axis, as a greater $-\log_{10}$(E-value) suggests higher protein identification confidence. The simple voting results were plotted by thresholding upon the number of votes. In the random forest model, the plot was generated at different thresholds of cross-validated probability of a correct expert annotation. For the four-vote ensemble, only a small fraction of probability from the simple voting and the random forest model could achieve higher confidence in protein identification compared to that from expert annotations (Figure 6.6a). In comparison, the confidence in protein identification from the seven-vote ensemble in most majority votes from the simple voting model and most probability thresholds from the random forest model exceeded the $-\log_{10}$(E-value) score obtained from the expert annotations (Figure 6.6b). The improvement in protein identification confidence from the four-vote ensemble to the seven-vote ensemble was also reflective of the observed increase in accuracy (in both limiting false positives while finding more peak clusters that match with an expert annotated peak, Figure 6.4c) when using a larger ensemble. Other proteoforms such as βTpm with CID activation showed a similar trend in the analysis (Figure S6.6 and Table S6.2), except for a few special cases. These results indicate that some of the lower intensity isotopic distributions which were identified using the machine learning strategy could help improve the identification confidence values.

The amount of true positive and false positive peaks that constitutes the consensus peak list has an impact on the database search when protein isoforms have a long homologous sequence. While evaluating the database search results for the ssTnT ECD spectrum using generated peak lists, several isoforms were identified including A0A5K1V8N4 (Troponin T, slow skeletal muscle isoform b, correct identification), H9FC02 (Troponin T, slow skeletal muscle isoform c), A0A1D5RIQ3 (Troponin T1, slow skeletal type), and F7HR11 (Troponin T1, slow skeletal type) (Table S6.5). Using a sequence alignment tool, it was observed that only the N-terminal sequence has variations among these four isoforms (Figure S6.7). Ideally, thresholding on higher probabilities should retain the true expert annotations while reducing the number of false positives. A lower threshold would also result in the inclusion of more false positive annotations. In this particular case, a simple voting method at low majority votes (less than three votes) yielded incorrect identification if the database search algorithm was given a set of peaks with many false positives. On the contrary, at higher thresholds for both the random forest algorithm and simple voting method, the omission of true positives led to either diminishing E-value of correct identification, meaning a less confident database search result, or an incorrect identification.

For the spectrum for ssTnC protein with ECD activation, none of the single algorithms, except for THRASH with 80% fit, were able to identify the target sequence (Table S6.2 and S6.6).

For the simple voting method, a majority vote (three or more votes) could correctly find the protein. This result indicates that utilizing a consensus peak list could help identify the proteoform in spectra, even when most of the deconvolution algorithms failed to find the correct identification. If at least one algorithm can find the correct identification, then theoretically the ensemble should also be able find the correct identification. Also, if there are several distinct false positive peaks (or no expert annotated peaks) from each algorithm, using a majority vote should help reduce the false positives (i.e., reduce the noise from each algorithm) to achieve a better identification rate.

On the basis of the database search results, both simple voting ensemble method and random forest machine learning algorithm were found to enhance both the accuracy and confidence in proteoform identification. For the simple voting ensemble method which utilized only clustering and simple voting, a majority vote (three votes in the seven-vote ensemble) yielded the best results. In the case of the random forest algorithm which required clustering and training a machine learning model, a probability threshold greater than 0.3 to 0.4 provided the optimal results.

Liquid chromatography-MS/MS data analysis

The results here are derived from targeted MS/MS data, and the machine learning strategy holds potential in improving the number of confident identifications with liquid chromatography (LC)-MS/MS runs. Further investigation needs to be done to determine whether models built using the expert annotations from MS/MS runs will improve the identification rate on a separate LC-MS/MS run or if other annotations are needed to improve performance. Annotating deconvoluted peaks from spectra with confident protein identification would be a good starting point. A simple voting model would be more easily applicable for the LC-MS/MS experiment, as other machine learning algorithms may require enough annotated top-down LC-MS/MS spectra in order to develop models for performance optimization.

Conclusion

We have designed and demonstrated a machine learning strategy that allows for the combination of deconvolution results from multiple algorithms into an accurate consensus peak list for downstream processing. With the detection of more real isotopic distributions while filtering out false positives, the process showed promise in reducing the time spent manually validating and correcting the ion annotations in top-down MS/MS protein identification. In both the simple voting ensemble method and random forest machine learning algorithm, the resulting consensus peak lists could improve on the accuracy and confidence in proteoform identification compared to a single deconvolution algorithm. This machine learning strategy shows promise for high-throughput protein identification and characterization in the LC-MS/MS data set for top-down proteomics. Integrating the tool into MASH Explorer will enable users to find more true positive deconvoluted peaks and consequently enhance the data analysis of the high-resolution top-down MS data set.

Acknowledgement

This work is dedicated to Professor John R. Yates III, the recipient of the 2019 ASMS John B. Fenn Award for a Distinguished Contribution in Mass Spectrometry. Financial support was provided by NIH R01 GM125085 (to Y. G.). Y. G. also would like to acknowledge the NIH grants, R01 GM117058 and S10 OD018475.

References

1. Zhang, H.; Ge, Y., Comprehensive Analysis of Protein Modifications by Top-Down Mass Spectrometry. *Circ. Cardiovasc. Genet.* **2011,** *4* (6), 711.

2. Smith, L. M.; Kelleher, N. L.; Proteomics, C. T. D., Proteoform: a single term describing protein complexity. *Nat. Methods* **2013,** *10* (3), 186-187.

3. Cai, W. X.; Tucholski, T. M.; Gregorich, Z. R.; Ge, Y., Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. *Expert Rev. Proteomic* **2016,** *13* (8), 717-730.

4. Chen, B. F.; Brown, K. A.; Lin, Z. Q.; Ge, Y., Top-Down Proteomics: Ready for Prime Time? *Anal. Chem.* **2018,** *90* (1), 110-127.

5. Aebersold, R.; Agar, J. N.; Amster, I. J.; Baker, M. S.; Bertozzi, C. R.; Boja, E. S.; Costello, C. E.; Cravatt, B. F.; Fenselau, C.; Garcia, B. A.; Ge, Y.; Gunawardena, J.; Hendrickson, R. C.; Hergenrother, P. J.; Huber, C. G.; Ivanov, A. R.; Jensen, O. N.; Jewett, M. C.; Kelleher, N. L.; Kiessling, L. L.; Krogan, N. J.; Larsen, M. R.; Loo, J. A.; Loo, R. R. O.; Lundberg, E.; MacCoss, M. J.; Mallick, P.; Mootha, V. K.; Mrksich, M.; Muir, T. W.; Patrie, S. M.; Pesavento, J. J.; Pitteri, S. J.; Rodriguez, H.; Saghatelian, A.; Sandoval, W.; Schluter, H.; Sechi, S.; Slavoff, S. A.; Smith, L. M.; Snyder, M. P.; Thomas, P. M.; Uhlen, M.; Van Eyk, J. E.; Vidal, M.; Walt, D. R.; White, F. M.; Williams, E. R.; Wohlschlager, T.; Wysocki, V. H.; Yates, N. A.; Young, N. L.; Zhang, B., How many human proteoforms are there? *Nat. Chem. Biol.* **2018,** *14* (3), 206-214.

6. Ge, Y.; Lawhorn, B. G.; ElNaggar, M.; Strauss, E.; Park, J. H.; Begley, T. P.; McLafferty, F. W., Top down characterization of larger proteins (45 kDa) by electron capture dissociation mass spectrometry. *J. Am. Chem. Soc.* **2002,** *124* (4), 672-678.

7. Shaw, J. B.; Li, W. Z.; Holden, D. D.; Zhang, Y.; Griep-Raming, J.; Fellers, R. T.; Early, B. P.; Thomas, P. M.; Kelleher, N. L.; Brodbelt, J. S., Complete Protein Characterization Using Top-Down Mass Spectrometry and Ultraviolet Photodissociation. *J. Am. Chem. Soc.* **2013,** *135* (34), 12646-12651.

8. 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. W.; Payne, S. H.; Sun, L. L.; Thomas, P. M.; Tucholski, T.; Wang, Z.; Wu, S.; Wu, Z. J.; Yu, D. H.; Shortreed, M. R.; Smith, L. M., Identification and Quantification of Proteoforms by Mass Spectrometry. *Proteomics* **2019,** *19* (10).

9. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W., Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* **2000,** *11* (4), 320-332.

10. Liu, X. W.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A., Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins. *Mol. Cell. Proteomics* **2010,** *9* (12), 2772-2782.

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

12. Sun, R. X.; Luo, L.; Wu, L.; Wang, R. M.; Zeng, W. F.; Chi, H.; Liu, C.; He, S. M., pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification. *Anal. Chem.* **2016,** *88* (6), 3082-3090.

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

14. Köster, C. Mass spectrometry method for accurate mass determination of unknown ions. US6188064B, 2001.

15. Cai, W. X.; Guner, H.; Gregorich, Z. R.; Chen, A. J.; Ayaz-Guner, S.; Peng, Y.; Valeja, S. G.; Liu, X. W.; Ge, Y., MASH Suite Pro: A Comprehensive Software Tool for Top-Down Proteomics. *Mol. Cell. Proteomics* **2016,** *15* (2), 703-714.

16. Kuncheva, L. I.; Whitaker, C. J., Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* **2003,** *51* (2), 181-207.

17. Sollich, P.; Krogh, A., Learning with ensembles: How over-fitting can be useful. *Adv. Neur. In.* **1996,** *8*, 190-196.

18. Geurts, P.; Fillet, M.; de Seny, D.; Meuwis, M. A.; Malaise, M.; Merville, M. P.; Wehenkel, L., Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics* **2005,** *21* (14), 3138-3145.

19. Ru, X. Q.; Li, L. H.; Zou, Q., Incorporating Distance-Based Top-n-gram and Random Forest To Identify Electron Transport Proteins. *J. Proteome Res.* **2019,** *18* (7), 2931-2939.

20. Tran, N. H.; Zhang, X. L. L.; Xin, L.; Shan, B. Z.; Li, M., De novo peptide sequencing by deep learning. *P. Natl. Acad. Sci. USA* **2017,** *114* (31), 8247-8252.

21. McIlwain, S. J.; Wu, Z.; Wenger, K.; Wetzel, M.; Tucholski, T.; Liu, X.; Sun, L.; Ong, I. M.; Ge, Y. In *MASH Explorer, a Universal, Comprehensive, and User-friendly Software Environment for Top-down Proteomics*, 67th ASMS Conference on Mass Spectrometry and Allied Topics, Atlanta, GA, June 3rd; Atlanta, GA, 2019.

22. Chen, B. F.; Lin, Z. Q.; Zhu, Y. L.; Jin, Y. T.; Larson, E.; Xu, Q. G.; Fu, C. X.; Zhang, Z. R.; Zhang, Q. Y.; Pritts, W. A.; Ge, Y., Middle-Down Multi-Attribute Analysis of Antibody-Drug Conjugates with Electron Transfer Dissociation. *Anal. Chem.* **2019,** *91* (18), 11661-11669.

23. Peng, Y.; Gregorich, Z. R.; Valeja, S. G.; Zhang, H.; Cai, W. X.; Chen, Y. C.; Guner, H.; Chen, A. J.; Schwahn, D. J.; Hacker, T. A.; Liu, X. W.; Ge, Y., Top-down Proteomics Reveals Concerted Reductions in Myofilament and Z-disc Protein Phosphorylation after Acute Myocardial Infarction. *Mol. Cell. Proteomics* **2014,** *13* (10), 2752-2764.

24. Chen, Y. C.; Ayaz-Guner, S.; Peng, Y.; Lane, N. M.; Locher, M. R.; Kohmoto, T.; Larsson, L.; Moss, R. L.; Ge, Y., Effective Top-Down LC/MS plus Method for Assessing Actin Isoforms as a Potential Cardiac Disease Marker. *Anal. Chem.* **2015,** *87* (16), 8399-8406.

25. Perez-Riverol, Y.; Csordas, A.; Bai, J. W.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Perez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaino, J. A., The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **2019,** *47* (D1), D442-D450.

26. Tibshirani, R.; Hastie, T.; Narasimhan, B.; Soltys, S.; Shi, G. Y.; Koong, A.; Le, Q. T., Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics* **2004,** *20* (17), 3034-3044.

27. Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; *al, e.*, SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python. *arXiv e-prints* **2019**, arXiv:1907.10121.

28. Liaw, A.; Wiener, M., Classification and Regression by randomForest. *R News* **2002,** *2* (3), 18-22.

29. Liu, X. W.; Hengel, S.; Wu, S.; Tolic, N.; Pasa-Tolic, L.; Pevzner, P. A., Identification of Ultramodified Proteins Using Top-Down Tandem Mass Spectra. *J. Proteome Res.* **2013,** *12* (12), 5830-5838.

30. Saxena, A.; Prasad, M.; Gupta, A.; Bharill, N.; Patel, O. P.; Tiwari, A.; Er, M. J.; Ding, W. P.; Lin, C. T., A review of clustering techniques and developments. *Neurocomputing* **2017,** *267*, 664-681.

31. Zhang, Y. Y.; Xin, Y.; Li, Q.; Ma, J. S.; Li, S.; Lv, X. D.; Lv, W. Q., Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. *Biomed. Eng. Online* **2017,** *16*.

32. Cortes, C.; Vapnik, V., Support-Vector Networks. *Mach. Learn.* **1995,** *20* (3), 273-297.

33. Schmidhuber, J., Deep learning in neural networks: An overview. *Neural Networks* **2015,** *61*, 85-117.

Table 6.1. Description of Features used in Machine Learning

Feature Name	**Data Type**	**Description**
Activation	ECD/CID	Activation used to generate spectra
Charge	Integer	Charge of the peaks within the cluster
Votes	Integer	Number of deconvolution algorithms that called a peak within that cluster
SumIntensity	Numeric	Sum of the intensity of peaks in the cluster
AverageIntensity	Numeric	Average intensity of peaks in the cluster
MSDeconv	Boolean	MS-Deconv called this peak
TopFD	Boolean	TopFD called this peak
THRASH60	Boolean	THRASH with 60% Fit called this peak
THRASH70	Boolean	THRASH with 70% Fit called this peak
THRASH80	Boolean	THRASH with 80% Fit called this peak
THRASH90	Boolean	THRASH with 90% Fit called this peak
SNAP	Boolean	SNAP called this peak
PrecursorCharge	Integer	Charge of the precursor
AvgMass	Numeric	Average Mass of the peaks within the cluster
StdDev	Numeric	Standard Deviation of the mass of the peaks within the cluster
PrecursorMass	Numeric	Monoisotopic mass of the precursor
PrecursorMZ	Numeric	m/z of the precursor

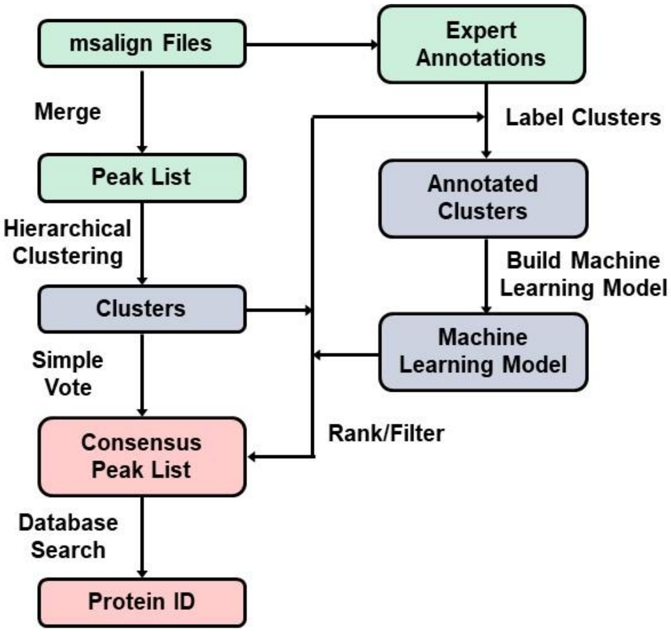


Figure 6.1. Flowchart for the machine learning strategy. This figure shows the steps taken to combine deconvolution results into a consensus peak list using either the simple voting method or a machine learning algorithm.

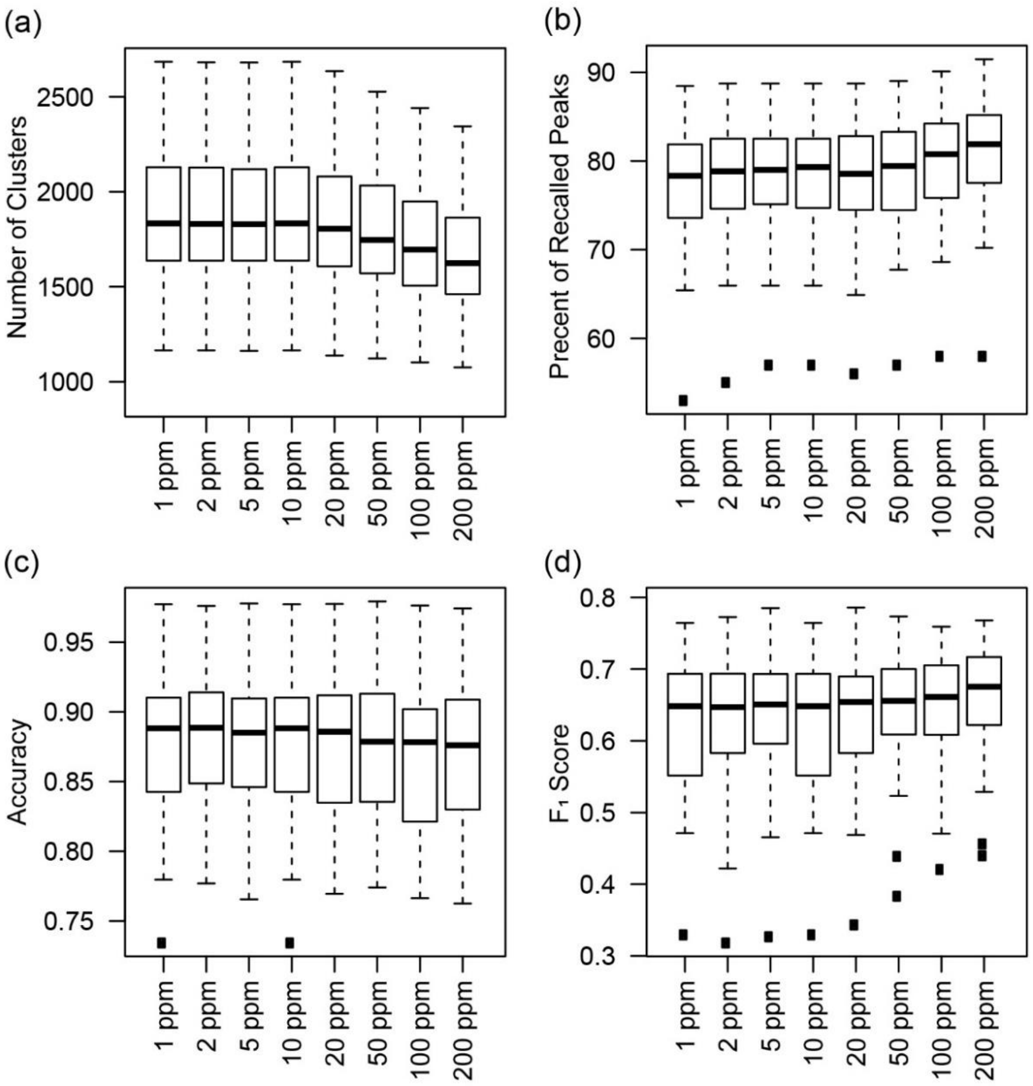


Figure 6.2. Cluster cutoff performance. Each plot is a boxplot that shows the spread of the metric measured from the 30 spectra versus different ppm cutoffs used in the hierarchical clustering step. (a) Number of clusters, (b) percent of recalled peaks versus ppm cutoff, (c) random forest accuracy versus ppm cutoff, and (d) random forest F_1 score versus ppm cutoff. The black squares in the figure represent outliers in the data set.

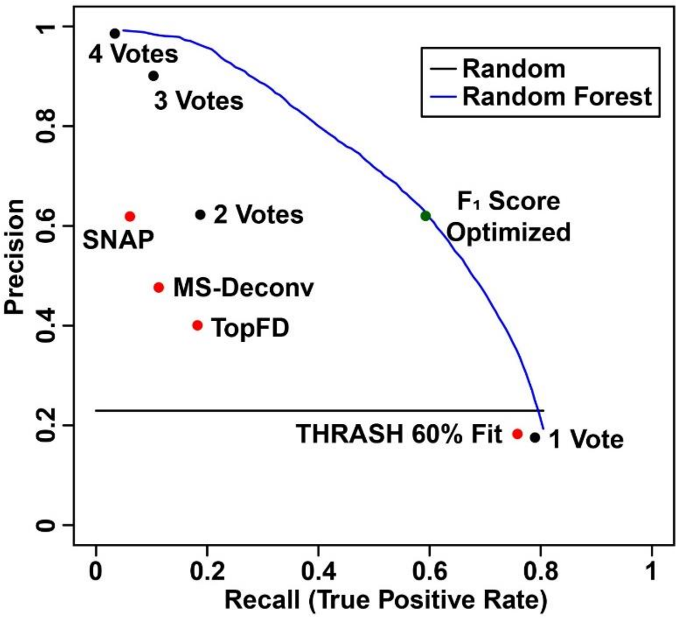


Figure 6.3. Precision-recall curves and points of the expert annotation prediction task. Plot displays the precision and recall performance of the deconvolution methods by themselves (red points), the simple voting (black points), and random forest (blue line). The green point represents random forest algorithm with the F_1 score optimized.

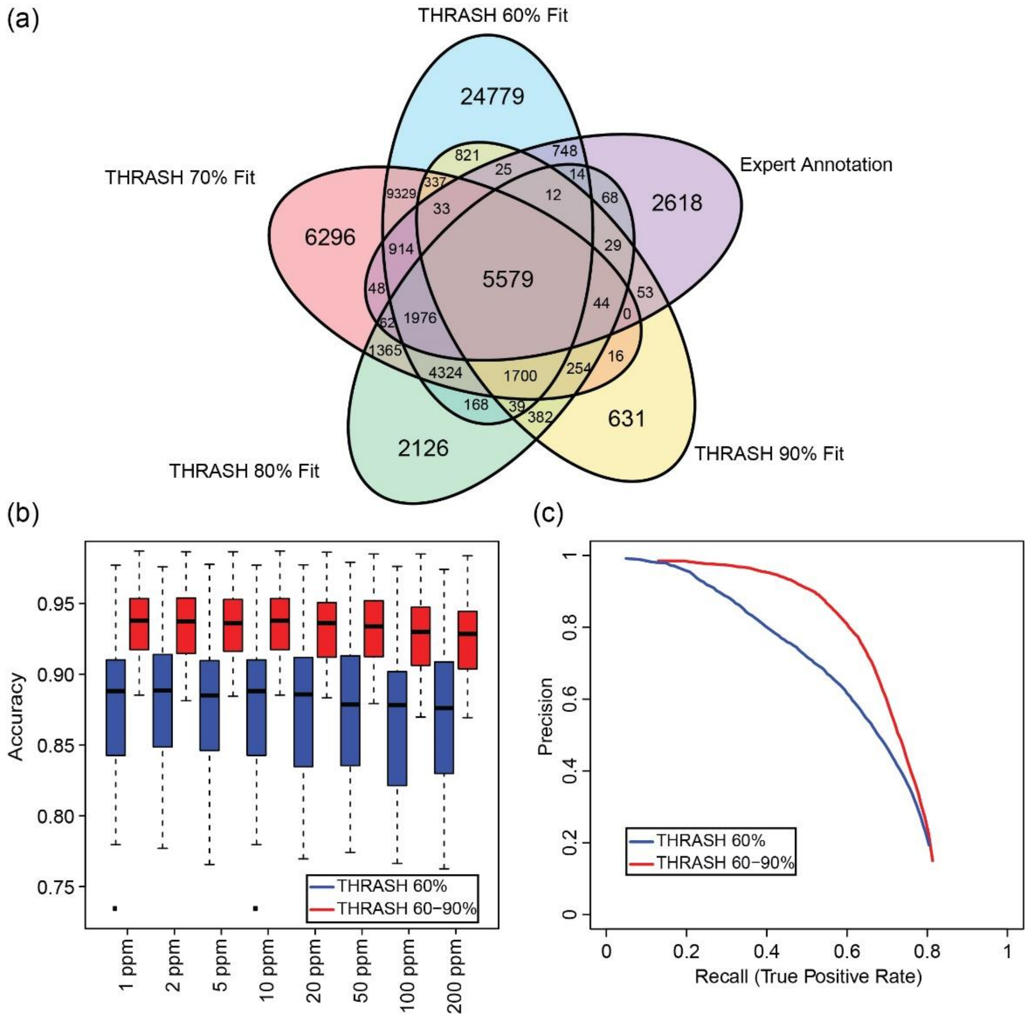


Figure 6.4. Performance comparison between four-vote ensemble and seven-vote ensemble. (a) Venn diagram of peaks found using THRASH with the fit parameter set at 60, 70, 80, or 90%. (b) Boxplot of random forest accuracy between the four-vote ensemble (red) and seven-vote ensemble (blue) with different cluster cutoffs. At all cluster cutoffs value, the seven-vote ensemble had better performance than four-vote ensemble. (c) Precision-recall curve using the four-vote ensemble (blue, THRASH 60%) and seven-vote ensemble (red, THRASH 60-90%). The seven-vote ensemble had improved performance compared to four-vote ensemble. The black squares in the figure represent outliers in the data set.

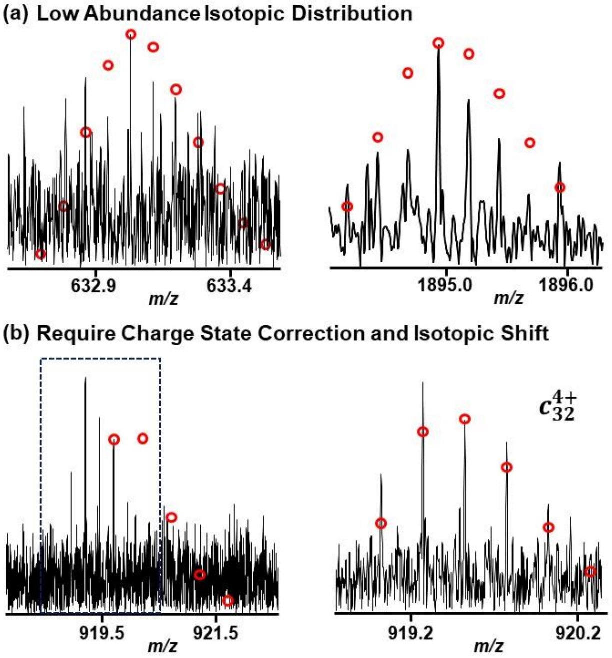


Figure 6.5. Example annotation of isotopic distributions. (a) Low-abundance isotopic distribution that could be found on the consensus peak list. These peaks were found only by the machine learning strategy. (b) Example isotopic distribution that has been manually corrected by shifting the charge and monoisotopic peak.

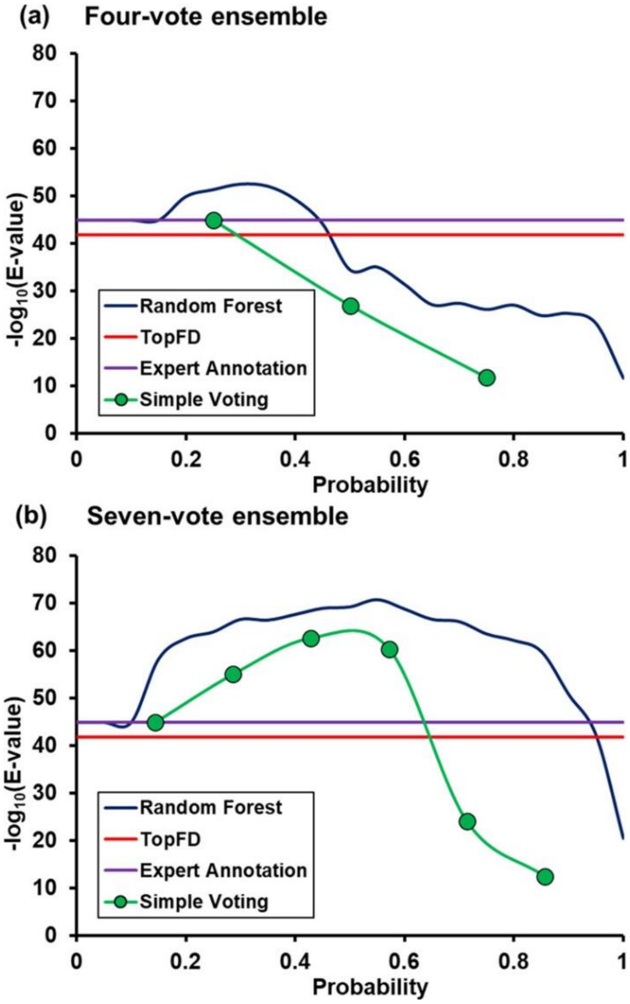


Figure 6.6. MS-Align+ database search results for αTpm. (a) Four-vote ensemble results. (b) Seven-vote ensemble results. Each plot has the -log$_{10}$(E-Value) for TopFD (red line), expert annotation (purple line), simple voting thresholding (no. of votes/max votes, green points/lines), and random forest probability thresholding (blue line).

Supplemental Information

Table S6.1. Summary of selected publicly available deconvolution algorithms

Algorithms	Description	Reference
THRASH	Uses a subtractive peak finding routine to locate possible isotopic clusters in the spectrum, using least-squares fits to theoretically derive isotopic abundance distributions.	Horn, D. M.; Zubarev, R. A.; McLafferty, F. W., Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J Am Soc Mass Spectr* **2000,** *11* (4), 320-332.
MS-Deconv	A combinatorial algorithm that uses graph theory to find the heaviest path in a largest set of potential candidate envelopes.	Liu, X. W.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A., Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins. *Mol Cell Proteomics* **2010,** *9* (12), 2772-2782.
TopFD	A successor to MS-Deconv, after grouping top-down spectral peaks into isotopomer envelopes, the algorithm converts isotopomer envelopes to monoisotopic neutral masses.	Kou, Q.; Xun, L. K.; Liu, X. W., TopPIC: a software tool for top-down mass spectrometry-based proteoform identification and characterization. *Bioinformatics* **2016,** *32* (22), 3495-3497.
SNAP	Fits a function of superimposed bell curves to the isotopic distributions within the spectrum.	Köster, C. Mass spectrometry method for accurate mass determination of unknown ions. US6188064B1, 2001.
UniDec	A bayesian approach to incorporate the charge-state distribution as a Bayesian prior to provide separation of the *m/z* spectrum into the	Marty, M. T.; Baldwin, A. J.; Marklund, E. G.; Hochberg, G. K. A.; Benesch, J. L. P.;

	corresponding physical mass and charge components.	Robinson, C. V., Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. *Anal Chem* **2015,** *87* (8), 4370-4376.
pParseTD	Utilizes a support vector machine (SVM) with radial basis functions that is trained online to incorporate a variety of features to detect the isotopic clusters and determine their charge states.	Sun, R. X.; Luo, L.; Wu, L.; Wang, R. M.; Zeng, W. F.; Chi, H.; Liu, C.; He, S. M., pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification. *Anal Chem* **2016,** *88* (6), 3082-3090.

Table S6.2. Summary of protein proteoforms used for expert annotation.

Protein ID	**Accession Number**	**Modifications**
αTpm	P04692[a]	+Ac
βTpm	P07951[b]	+Ac
ssTnT	A0A5K1V8N4[c]	+P+Ac-Met
fsTnT5	A0A2K5WPH1[d]	+P+Ac-Met
ssTnI	I2CW22	-Met
fsTnI	A0A1D5QTK6	+Ac-Met
fsTnC	F7HGA7	+Ac-Met
ssTnC	G7MV95	+Ac
MLC-1F	G7N8T7	+$(Me)_3$-Met
MLC-2F	F7EI96	+2P+$(Me)_3$-Met
MLC-2S	A0A1D5RDY5	+Dea+$(Me)_3$-Met
MLC-3F	F7B2B7	+Ac-Met
PDLIM5	F6Z147	+Me+Ac-Met
PDLIM7	NA	+Ac
LDB3	O75112-6[b]	+Ac-Met

Ac, acetylation; P, phosphorylation; Met, methionine; Me, methylation; Dea, deamidation.

[a] The accession number comes from the uniprot rat database.

[b] The accession number comes from the uniprot human database.

[c] The accession F7HR10 for the Troponin T isoforms in the publication by Jin et al. was changed to A0A5K1V8N4 in December, 2019 by authors who originally published the sequence.

[d] The accession number comes from the uniprot cynomolgus monkey database.

Table S6.3. Clusters classification for four-vote ensemble

Entry	**Protein**	**Activation**	**Cluster_True Positive**	**Cluster_False Positive**	**Cluster_False Negative**
1	αTpm	CID	124	1601	64
2	αTpm	ECD	467	1388	184
3	βTpm	CID	401	1703	131
4	βTpm	ECD	284	1348	77
5	ssTnT	CID	247	1854	73
6	ssTnT	ECD	254	1685	65
7	fsTnT5	CID	57	1102	43
8	fsTnT5	ECD	750	1919	152
9	ssTnI	CID	367	1812	70
10	ssTnI	ECD	346	1893	83
11	fsTnI	CID	204	1130	54
12	fsTnI	ECD	239	1424	86
13	fsTnC	CID	390	1519	145
14	fsTnC	ECD	221	959	31
15	ssTnC	CID	265	1741	77
16	ssTnC	ECD	80	1140	20
17	MLC-1F	CID	394	1100	58
18	MLC-1F	ECD	699	1653	127
19	MLC-2F	CID	133	1664	45
20	MLC-2F	ECD	662	1537	192
21	MLC-2S	CID	304	1640	98
22	MLC-2S	ECD	487	1637	125
23	MLC-3F	CID	242	1135	55
24	MLC-3F	ECD	323	1348	41
25	PDLIM5	CID	335	1447	77
26	PDLIM5	ECD	255	1361	54
27	PDLIM7	CID	357	1863	121
28	PDLIM7	ECD	427	1472	95
29	LDB3	CID	63	1673	30
30	LDB3	ECD	317	1369	63
Total			9694	45117	2536

Table S6.4. Clusters classification for seven-vote ensemble

Entry	**Protein**	**Activation**	**Cluster_True Positive**	**Cluster_False Positive**	**Cluster_False Negative**
1	αTpm	CID	129	2234	59
2	αTpm	ECD	485	1743	166
3	βTpm	CID	412	2196	120
4	βTpm	ECD	292	1771	69
5	ssTnT	CID	249	2599	71
6	ssTnT	ECD	265	2002	54
7	fsTnT5	CID	58	1328	42
8	fsTnT5	ECD	772	2417	130
9	ssTnI	CID	373	2273	64
10	ssTnI	ECD	356	2531	73
11	fsTnI	CID	210	1325	48
12	fsTnI	ECD	250	1707	75
13	fsTnC	CID	396	1827	139
14	fsTnC	ECD	226	1094	26
15	ssTnC	CID	268	2153	74
16	ssTnC	ECD	82	1348	18
17	MLC-1F	CID	402	1300	50
18	MLC-1F	ECD	721	2056	105
19	MLC-2F	CID	138	2106	40
20	MLC-2F	ECD	669	1919	185
21	MLC-2S	CID	314	2082	88
22	MLC-2S	ECD	501	2009	111
23	MLC-3F	CID	248	1319	49
24	MLC-3F	ECD	326	1550	38
25	PDLIM5	CID	344	1774	68
26	PDLIM5	ECD	263	1655	46
27	PDLIM7	CID	369	2370	109
28	PDLIM7	ECD	442	1754	80
29	LDB3	CID	64	2224	29
30	LDB3	ECD	326	1697	54
Total			9950	56363	2280

Table S6.5. Database search results for ssTnT with ECD activation from different peak lists

Input peak list	**E-value**	**Protein ID**
Expert Annotation (“True Positive”)	1.40E-45	tr|H9FC02|H9FC02_MACMU Troponin T, slow skeletal muscle isoform c (Fragment) OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1
THRASH 60% fit	1.40E-45	tr|A0A1D5RIQ3|A0A1D5RIQ3_MACMU Troponin T1, slow skeletal type OS=Macaca mulatta OX=9544 GN=TNNT1 PE=4 SV=2
THRASH 70% fit	**1.40E-45**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
THRASH 80% fit	**1.40E-45**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
THRASH 90% fit	**6.37E-54**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
MS-Deconv	28.1438675	tr|F7GAK8|F7GAK8_MACMU Uncharacterized protein OS=Macaca mulatta OX=9544 GN=SLC35A1 PE=4 SV=3
TopFD	**1.42E-06**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
SNAP	3.91E-22	tr|F7HR11|F7HR11_MACMU Troponin T1, slow skeletal type OS=Macaca mulatta OX=9544 GN=TNNT1 PE=4 SV=3
1 Vote	1.40E-45	tr|A0A1D5RIQ3|A0A1D5RIQ3_MACMU Troponin T1, slow skeletal type OS=Macaca mulatta OX=9544 GN=TNNT1 PE=4 SV=2
2 Votes	1.40E-45	tr|F7HR11|F7HR11_MACMU Troponin T1, slow skeletal type OS=Macaca mulatta OX=9544 GN=TNNT1 PE=4 SV=3
3 Votes	**1.40E-45**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
4 Votes	**3.15E-52**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**

5 Votes	**1.77E-17**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
6 Votes	8.431491443	tr|A0A5F7ZUT8|A0A5F7ZUT8_MACMU Zinc finger protein 239 OS=Macaca mulatta OX=9544 GN=ZNF239 PE=4 SV=1
Random Forest (⩾ 0 Probability)	1.40E-45	tr|A0A1D5RIQ3|A0A1D5RIQ3_MACMU Troponin T1, slow skeletal type OS=Macaca mulatta OX=9544 GN=TNNT1 PE=4 SV=2
Random Forest (≥ 0.05 Probability)	**1.40E-45**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.1 Probability)	**1.40E-45**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.15 Probability)	**4.72E-49**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.2 Probability)	**1.63E-49**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.25 Probability)	**2.69E-47**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.3 Probability)	**1.40E-45**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.35 Probability)	**7.65E-46**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.4 Probability)	**2.62E-62**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.45 Probability)	**4.79E-69**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**

Random Forest (≥ 0.5 Probability)	**4.72E-67**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.55 Probability)	**6.11E-62**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.6 Probability)	**1.69E-60**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.65 Probability)	**1.77E-59**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.7 Probability)	**1.80E-59**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.75 Probability)	**8.01E-60**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.8 Probability)	**1.62E-53**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.85 Probability)	**1.23E-45**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.9 Probability)	**6.96E-44**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**
Random Forest (≥ 0.95 Probability)	**6.34E-38**	**tr|A0A5K1V8N4|A0A5K1V8N4_MACMU Troponin T, slow skeletal muscle isoform b OS=Macaca mulatta OX=9544 GN=TNNT1 PE=2 SV=1**

The **bold** entries represent the input peak list from the methods that match with the targeted protein sequence.

Table S6.6. Database search results for ssTnC with ECD activation from different peak lists

Input peak list	**E-value**	**Protein ID**
Expert Annotation (“True Positive”)	**3.15E-12**	**tr|G7MV95|G7MV95_MACMU Troponin C, slow skeletal and cardiac muscles OS=Macaca mulatta OX=9544 GN=TNNC1 PE=2 SV=1**
THRASH 60% fit	4.444046	tr|A0A1Y2T072|A0A1Y2T072_9BIFI Cell division protein SepF OS=Alloscardovia macacae OX=1160091 GN=sepF PE=3 SV=1
THRASH 70% fit	300.8025	tr|G7NTX5|G7NTX5_MACFA BRO1 domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_01363 PE=4 SV=1
THRASH 80% fit	**3.67E-13**	**tr|G7MV95|G7MV95_MACMU Troponin C, slow skeletal and cardiac muscles OS=Macaca mulatta OX=9544 GN=TNNC1 PE=2 SV=1**
THRASH 90% fit	0.180059	tr|G7PLX3|G7PLX3_MACFA DUF4515 domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_04680 PE=4 SV=1
MS-Deconv	83.40762	tr|G7Q1W7|G7Q1W7_MACFA Uncharacterized protein OS=Macaca fascicularis OX=9541 GN=EGM_19286 PE=4 SV=1
TopFD	24.41666	tr|G7PLX3|G7PLX3_MACFA DUF4515 domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_04680 PE=4 SV=1
SNAP	N/A	No Identification
1 Vote	1.109976	tr|A0A1Y2T072|A0A1Y2T072_9BIFI Cell division protein SepF OS=Alloscardovia macacae OX=1160091 GN=sepF PE=3 SV=1
2 Votes	0.29619	tr|I2CUR0|I2CUR0_MACMU DENN domain-containing protein 1A isoform 1 OS=Macaca mulatta OX=9544 GN=DENND1A PE=2 SV=1
3 Votes	**1.01E-12**	**tr|G7MV95|G7MV95_MACMU Troponin C, slow skeletal and cardiac muscles OS=Macaca mulatta OX=9544 GN=TNNC1 PE=2 SV=1**
4 Votes	0.002308	tr|G7NUJ4|G7NUJ4_MACFA Uncharacterized protein OS=Macaca fascicularis OX=9541 GN=EGM_01143 PE=4 SV=1
5 Votes	0.000135	tr|G7PP36|G7PP36_MACFA U-box domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_06285 PE=4 SV=1
6 Votes	N/A	No Identification
Random Forest (≥ 0 Probability)	4.439906	tr|A0A1Y2T072|A0A1Y2T072_9BIFI Cell division protein SepF OS=Alloscardovia macacae OX=1160091 GN=sepF PE=3 SV=1

Random Forest (≥ 0.05 Probability)	**1.74E-09**	**tr|G7MV95|G7MV95_MACMU Troponin C, slow skeletal and cardiac muscles OS=Macaca mulatta OX=9544 GN=TNNC1 PE=2 SV=1**
Random Forest (≥ 0.1 Probability)	9.763027	tr|G7PE20|G7PE20_MACFA Homeobox domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_18395 PE=4 SV=1
Random Forest (≥ 0.15 Probability)	7.369076	tr|G7N411|G7N411_MACMU Tumor protein D54 isoform a OS=Macaca mulatta OX=9544 GN=TPD52L2 PE=2 SV=1
Random Forest (≥ 0.2 Probability)	5.825325	tr|G7N411|G7N411_MACMU Tumor protein D54 isoform a OS=Macaca mulatta OX=9544 GN=TPD52L2 PE=2 SV=1
Random Forest (≥ 0.25 Probability)	4.396271	tr|G7N411|G7N411_MACMU Tumor protein D54 isoform a OS=Macaca mulatta OX=9544 GN=TPD52L2 PE=2 SV=1
Random Forest (≥ 0.3 Probability)	1.953599	tr|G7N411|G7N411_MACMU Tumor protein D54 isoform a OS=Macaca mulatta OX=9544 GN=TPD52L2 PE=2 SV=1
Random Forest (≥ 0.35 Probability)	1.317487	tr|G7N411|G7N411_MACMU Tumor protein D54 isoform a OS=Macaca mulatta OX=9544 GN=TPD52L2 PE=2 SV=1
Random Forest (≥ 0.4 Probability)	1.070612	tr|G7N411|G7N411_MACMU Tumor protein D54 isoform a OS=Macaca mulatta OX=9544 GN=TPD52L2 PE=2 SV=1
Random Forest (≥ 0.45 Probability)	0.408344	tr|G7N411|G7N411_MACMU Tumor protein D54 isoform a OS=Macaca mulatta OX=9544 GN=TPD52L2 PE=2 SV=1
Random Forest (≥ 0.5 Probability)	2.780761	tr|G7N411|G7N411_MACMU Tumor protein D54 isoform a OS=Macaca mulatta OX=9544 GN=TPD52L2 PE=2 SV=1
Random Forest (≥ 0.55 Probability)	2.780761	tr|G7N411|G7N411_MACMU Tumor protein D54 isoform a OS=Macaca mulatta OX=9544 GN=TPD52L2 PE=2 SV=1
Random Forest (≥ 0.6 Probability)	1.25252	tr|G7PLX3|G7PLX3_MACFA DUF4515 domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_04680 PE=4 SV=1
Random Forest (≥ 0.65 Probability)	0.987189	tr|G7PLX3|G7PLX3_MACFA DUF4515 domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_04680 PE=4 SV=1
Random Forest (≥ 0.7 Probability)	232.97	tr|G7PP36|G7PP36_MACFA U-box domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_06285 PE=4 SV=1
Random Forest (≥ 0.75 Probability)	232.97	tr|G7PP36|G7PP36_MACFA U-box domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_06285 PE=4 SV=1

Random Forest (≥ 0.8 Probability)	188.9821	tr|G7PP36|G7PP36_MACFA U-box domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_06285 PE=4 SV=1
Random Forest (≥ 0.85 Probability)	178.5063	tr|G7PP36|G7PP36_MACFA U-box domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_06285 PE=4 SV=1
Random Forest (≥ 0.9 Probability)	91.52464	tr|G7PP36|G7PP36_MACFA U-box domain-containing protein OS=Macaca fascicularis OX=9541 GN=EGM_06285 PE=4 SV=1
Random Forest (≥ 0.95 Probability)	10.09021	tr|F7HHK1|F7HHK1_MACMU Myosin binding protein C, fast type OS=Macaca mulatta OX=9544 GN=MYBPC2 PE=4 SV=3

The **bold** entries represent the input peak list from the methods that match with the targeted protein sequence.

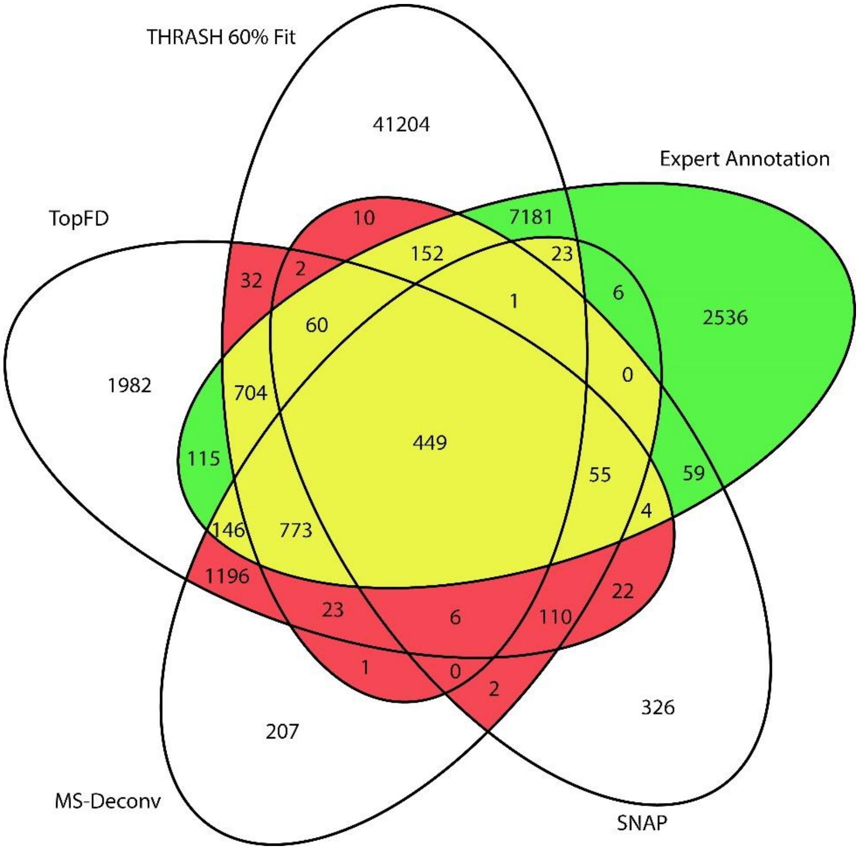


Figure S6.1. Venn diagram for the majority vote greater than two and its overlap with expert annotation ("true positive"). Three color codes were used in this figure. Red represents the unique peaks that is in the majority vote (2 or more votes), but not in the expert annotations. Green represents the unique peaks that is in the expert annotation, but not in the majority vote (2 or more votes). Yellow represents the peaks found by both expert annotations and the majority vote (2 or more votes).

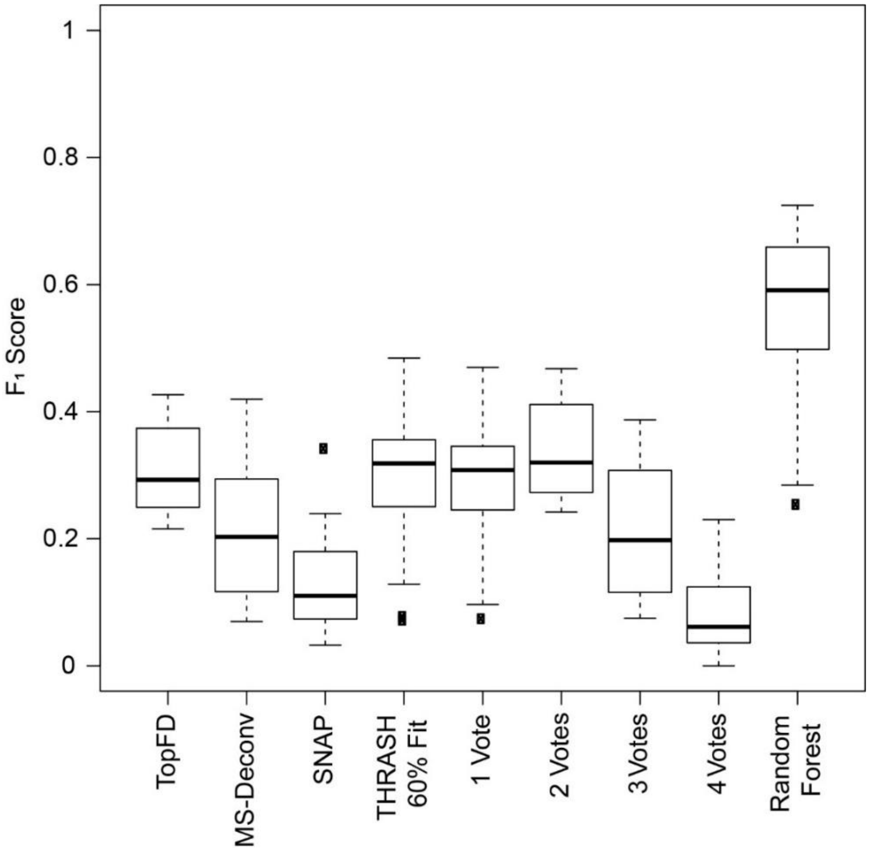


Figure S6.2. Boxplot of F_1 scores using cross-validated fold results from the four-vote ensemble comparison. The line in each box shows the median F_1 score across the testing folds from the leave-one-spectrum out cross validation. The edges of the box show the 1st and 3rd quartile, the whiskers are either the extreme value or 1.5 times the interquartile range (IQR), whichever is smaller. The dots are the extreme points that lie outside of the 1.5 x IQR range.

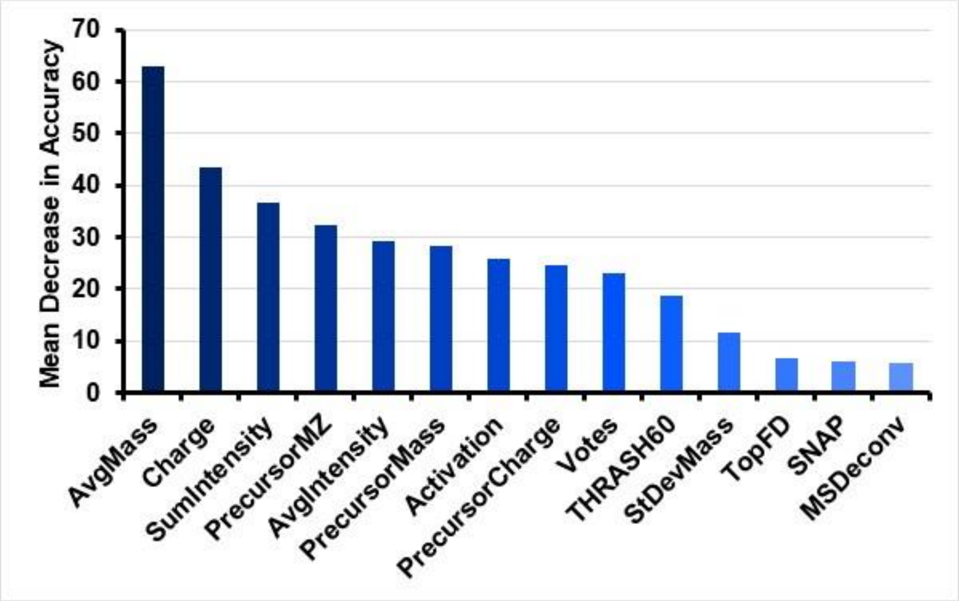


Figure S6.3. Feature ranking for mean decrease in accuracy. Each bar is a measure of the magnitude of the decrease in the random forest model's accuracy after the feature has been permuted. A higher value indicates the feature was "important" to the overall performance of the random forest model.

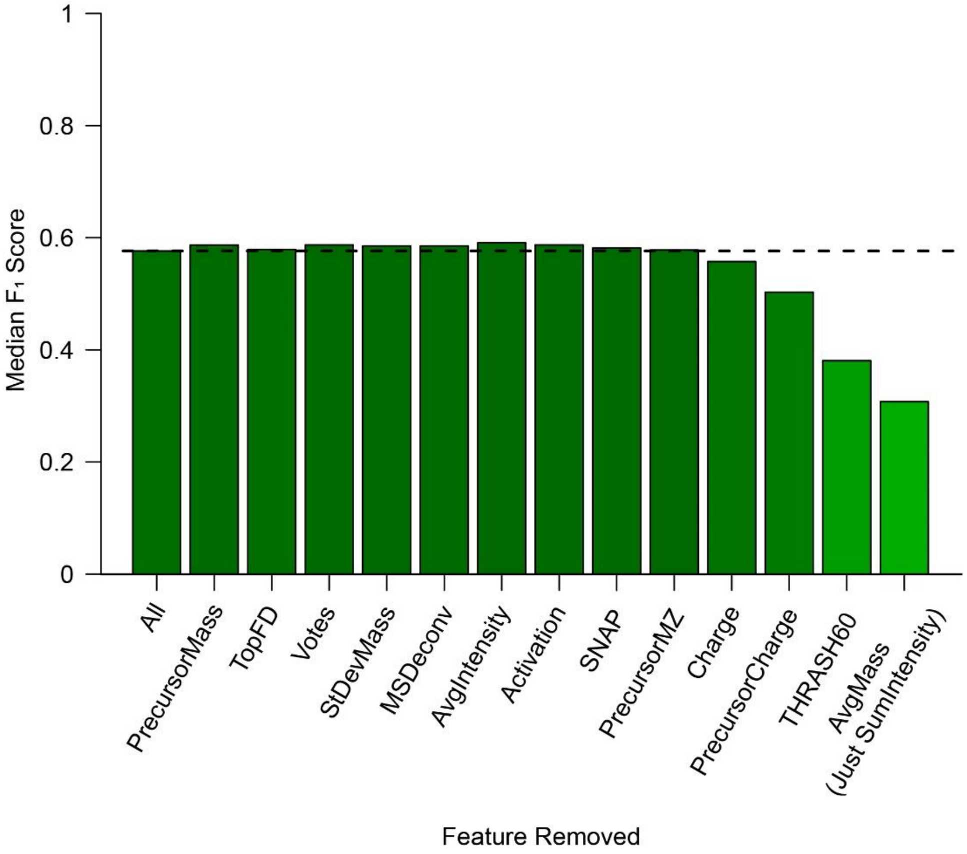


Figure S6.4. Backward selection of features for the random forest model. Backward selection was performed with leave-one-spectrum-out cross-validation, and optimization on the median F1-score. Each bar represents the feature removed iteratively from left to right during the backward selection process. Removal of features such as Charge, PrecursorCharge, THRASH60, AvgMass, and SumIntensity significantly impact the median of F_1 score.

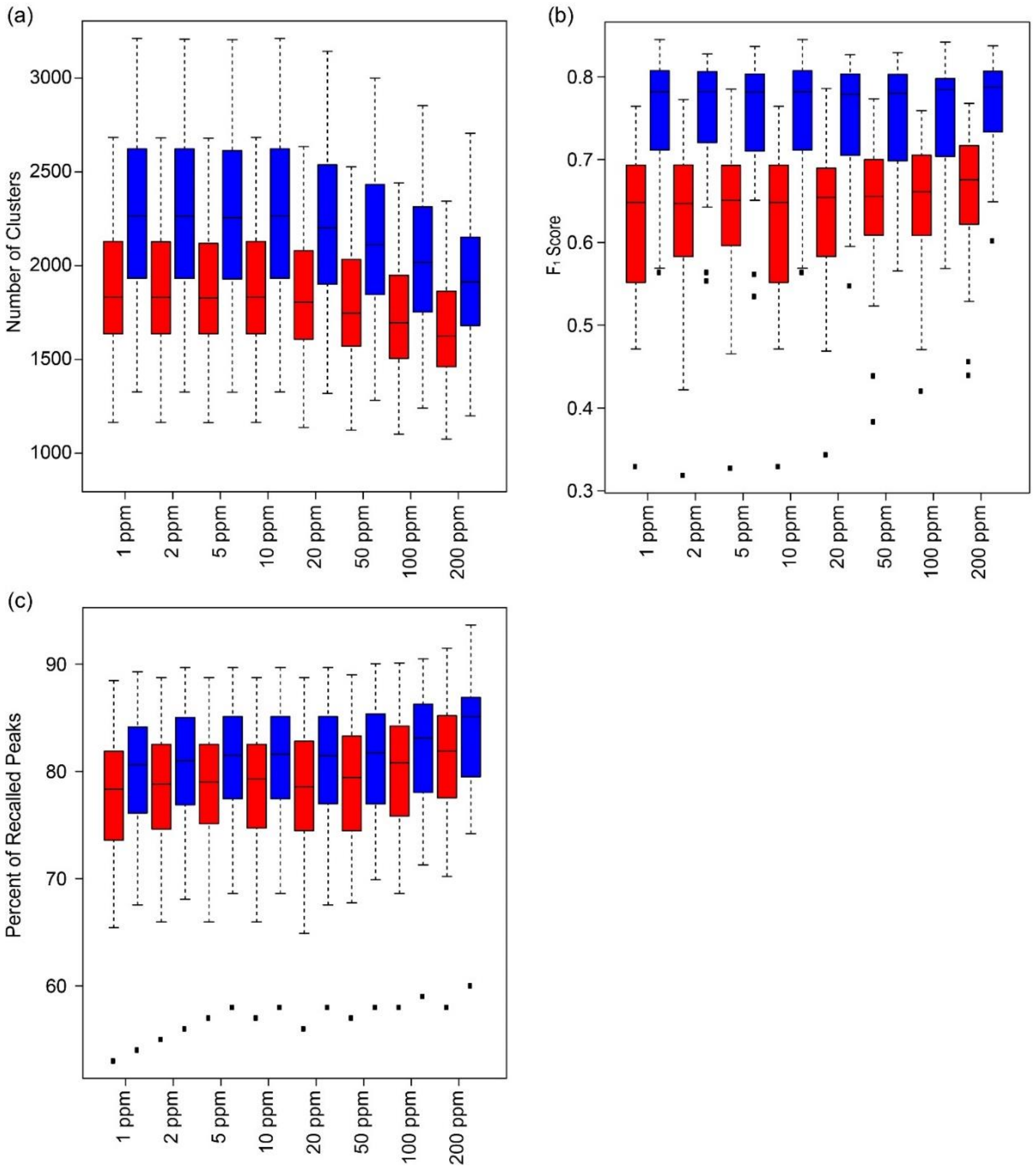


Figure S6.5. Statistical analysis between four-vote ensemble and seven-vote ensemble. Three boxplots display the spread of the metric measured from the 30 spectra versus different ppm cutoffs used in the hierarchical clustering step. (a) Number of clusters, (b) F_1 score, and (c) Percent of recalled peaks. Results from the four-vote and seven-vote ensembles are depicted in red and blue, respectively.

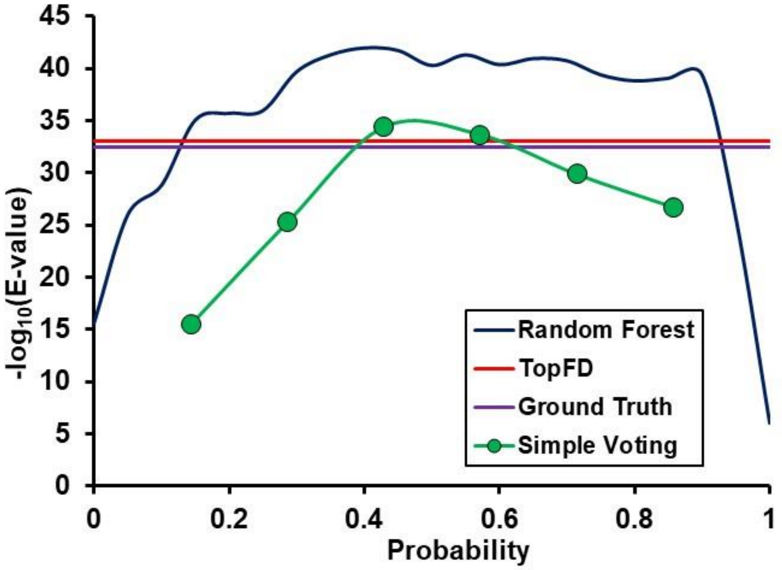
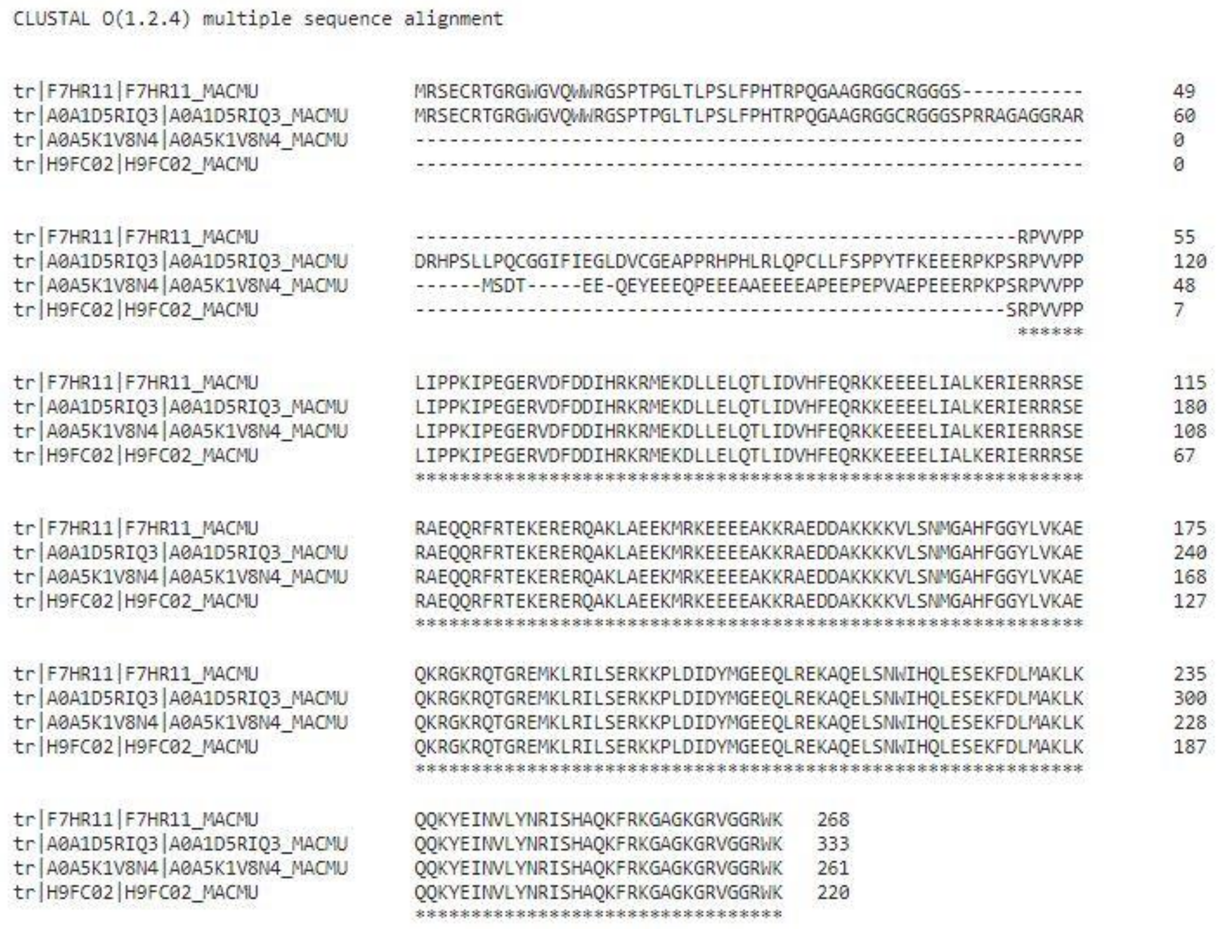


Figure S6.6. E-value evaluation for βTpm spectrum with CID activation. Each plot has the -$\log_{10}$(E-value) for TopFD (red line), Expert Annotation (purple line), simple voting thresholding (#votes/max votes, green points/lines), and random forest probability thresholding (blue line).



```
CLUSTAL O(1.2.4) multiple sequence alignment


tr|F7HR11|F7HR11_MACMU                MRSECRTGRGWGVQWWRGSPTPGLTLPSLFPHTRPQGAAGRGGCRGGGS-----------      49
tr|A0A1D5RIQ3|A0A1D5RIQ3_MACMU        MRSECRTGRGWGVQWWRGSPTPGLTLPSLFPHTRPQGAAGRGGCRGGGSPRRAGAGGRAR      60
tr|A0A5K1V8N4|A0A5K1V8N4_MACMU        ------------------------------------------------------------      0
tr|H9FC02|H9FC02_MACMU                ------------------------------------------------------------      0

tr|F7HR11|F7HR11_MACMU                ------------------------------------------------------RPVVPP      55
tr|A0A1D5RIQ3|A0A1D5RIQ3_MACMU        DRHPSLLPQCGGIFIEGLDVCGEAPPRHPHLRLQPCLLFSPPYTFKEEERPKPSRPVVPP      120
tr|A0A5K1V8N4|A0A5K1V8N4_MACMU        ------MSDT-----EE-QEYEEEQPEEEAAEEEEAPEEPEPVAEPEEERPKPSRPVVPP      48
tr|H9FC02|H9FC02_MACMU                -----------------------------------------------------SRPVVPP      7
                                                                                            ******

tr|F7HR11|F7HR11_MACMU                LIPPKIPEGERVDFDDIHRKRMEKDLLELQTLIDVHFEQRKKEEEELIALKERIERRRSE      115
tr|A0A1D5RIQ3|A0A1D5RIQ3_MACMU        LIPPKIPEGERVDFDDIHRKRMEKDLLELQTLIDVHFEQRKKEEEELIALKERIERRRSE      180
tr|A0A5K1V8N4|A0A5K1V8N4_MACMU        LIPPKIPEGERVDFDDIHRKRMEKDLLELQTLIDVHFEQRKKEEEELIALKERIERRRSE      108
tr|H9FC02|H9FC02_MACMU                LIPPKIPEGERVDFDDIHRKRMEKDLLELQTLIDVHFEQRKKEEEELIALKERIERRRSE      67
                                      ************************************************************

tr|F7HR11|F7HR11_MACMU                RAEQQRFRTEKERERQAKLAEEKMRKEEEEAKKRAEDDAKKKKVLSNMGAHFGGYLVKAE      175
tr|A0A1D5RIQ3|A0A1D5RIQ3_MACMU        RAEQQRFRTEKERERQAKLAEEKMRKEEEEAKKRAEDDAKKKKVLSNMGAHFGGYLVKAE      240
tr|A0A5K1V8N4|A0A5K1V8N4_MACMU        RAEQQRFRTEKERERQAKLAEEKMRKEEEEAKKRAEDDAKKKKVLSNMGAHFGGYLVKAE      168
tr|H9FC02|H9FC02_MACMU                RAEQQRFRTEKERERQAKLAEEKMRKEEEEAKKRAEDDAKKKKVLSNMGAHFGGYLVKAE      127
                                      ************************************************************

tr|F7HR11|F7HR11_MACMU                QKRGKRQTGREMKLRILSERKKPLDIDYMGEEQLREKAQELSNWIHQLESEKFDLMAKLK      235
tr|A0A1D5RIQ3|A0A1D5RIQ3_MACMU        QKRGKRQTGREMKLRILSERKKPLDIDYMGEEQLREKAQELSNWIHQLESEKFDLMAKLK      300
tr|A0A5K1V8N4|A0A5K1V8N4_MACMU        QKRGKRQTGREMKLRILSERKKPLDIDYMGEEQLREKAQELSNWIHQLESEKFDLMAKLK      228
tr|H9FC02|H9FC02_MACMU                QKRGKRQTGREMKLRILSERKKPLDIDYMGEEQLREKAQELSNWIHQLESEKFDLMAKLK      187
                                      ************************************************************

tr|F7HR11|F7HR11_MACMU                QQKYEINVLYNRISHAQKFRKGAGKGRVGGRWK   268
tr|A0A1D5RIQ3|A0A1D5RIQ3_MACMU        QQKYEINVLYNRISHAQKFRKGAGKGRVGGRWK   333
tr|A0A5K1V8N4|A0A5K1V8N4_MACMU        QQKYEINVLYNRISHAQKFRKGAGKGRVGGRWK   261
tr|H9FC02|H9FC02_MACMU                QQKYEINVLYNRISHAQKFRKGAGKGRVGGRWK   220
                                      *********************************
```

Figure S6.7. Sequence comparison among four slow skeletal Troponin T isoforms. For these four isoforms, the sequences differ in the N-terminal protein sequence, and the rest of the sequences to the C-terminus match. The sequence alignment was performed using Clustal Omega tool.

Chapter 7

Conclusion and Future Directions

Reversible phosphorylation including both phosphorylation and dephosphorylation is one of the most important PTMs in biological processes. MS-based proteomics has been extensively utilized to study the proteome, which includes analyzing protein sequences and PTMs. In particular, top-down MS-based proteomics is the method of choice to comprehensively study proteoforms arising from PTMs, alternative splicing and sequence variations. In this dissertation, I used top-down MS and developed methods to study phosphoprotein quantification and characterization. Furthermore, I established a platform using novel nanomaterials to perform effective kinase enrichment for future top-down analysis for intact kinases, which are crucial interactors in the process of phosphorylation. Finally, I developed both a software tool and novel deconvolution machine learning algorithms that enables researchers to study proteoforms using top-down mass spectrometry more effectively and efficiently.

In Chapter 2 and Chapter 3, the quantification and characterization of phosphoproteins are investigated. In Chapter 2, the impact of ESI quantification of phosphoproteins were examined using two model proteins, ENH2 and β-casein. In this study, it was revealed that monophosphorylation in the case of ENH2 had a minimal impact on the ESI quantification, whereas pentakisphosphorylation in the case of β-casein had a significant influence. Further investigation showed that the charge state envelope shifted for β-casein, suggesting that differences in physiochemical property between the unphosphorylated and phosphorylated β-casein were present. In Chapter 3, comprehensive characterization using top-down MS was conducted on PKA C-subunit, which is an important kinase in many biological processes. For the first time, the sequence variation and seven phosphorylation sites of the bacterially expressed PKA C-subunit were characterized simultaneously. Four of these seven phosphorylation sites were located at the 6xHis-tag used for affinity purification.

Top-down MS is a reliable method for relative quantification of proteoforms, owing to the belief that small modifications do not significantly impact the ionization/detection efficiency of intact proteins.[1-2] One limitation of this study is the availability of phosphoproteins whose phosphorylation status can be well controlled. I imagine the most ideal case of this study will be to have a phosphoprotein which has two phosphorylation sites that are controlled by two distinct kinases.[3] This system can then be used for studying possible structural changes during phosphorylation by comparing monophosphorylated proteoform of each site with unphosphorylated proteoform. It will also allow for comparison among three proteoforms to examine the impact of phosphorylation on ESI quantification. Additionally, β-casein with pentakisphosphorylation showed significant impact on the ESI quantification. It leads me to believe that larger modifications do have an impact compared to smaller modifications. Larger modifications such as glycosylation may need further interrogation.

Characterization of intact phosphoprotein has become more accessible through the rapid development of top-down MS. Our group has published numerous publication which characterized the sequence variation, alternative splicing, and phosphorylation site localization.[4-7] However, characterization using top-down MS still faces challenges. One of them is that the buffer conditions of phosphoprotein storage are generally not MS friendly. Surfactant and glycerol are often necessary maintain catalytic activities and stability in solution for phosphoprotein such as kinases. Additionally, large scale analysis of intact phosphoproteins has not been widely accessible. Characterization and identification of phosphoproteins through online LC-MS/MS is still limited by technical developments in top-down MS, in particular bioinformatics. Top-down analysis still requires manual work and experience from researchers.

In Chapter 4, a nanomaterial platform was developed for kinase enrichment. Using allene ligand coated on the iron oxide nanoparticle surface, kinase inhibitor modified with thiol functionality can be efficiently incorporated on the nanoparticle via "thiol-ene" chemistry. The functionalized nanoparticle was shown to capture kinases with minimal non-specific binding using a simple system. Using bottom-up proteomics, some kinases could be identified in the complex through kinase enrichment using functionalized nanoparticle.

While some success was demonstrated in the complex system, the kinase enrichment workflow still requires optimization. At the moment, the functionalized nanoparticle appeared to have significant nonspecific binding in the complex system. The intermediate goal after optimization is to use this platform to examine intact kinases. This platform can also be extended by incorporation other kinase inhibitors.[8-11] While most kinase inhibitors are very polar, the synthetic scheme using glutamic-cysteinamide moiety shown in Chapter 4 can be applicable to modify other amine functionalized kinase inhibitors by decreasing the polarity of the final products. A side experiment during graduate school has shown success to modified AX14596, which is a polar molecule, with the same glutamic-cysteinamide moiety from DMF reaction, worked up by extraction using ethyl acetate and water, and separated by flash column chromatography.

In Chapter 5 and 6, a software tool and a novel deconvolution machine learning strategy were developed to address the needs for comprehensive tools in top-down proteomics analysis. In Chapter 5, I developed MASH Explorer software, which is a universal, comprehensive, and user-friendly software environment for top-down proteomics. This software is able to process datasets from multiple vendor formats, as well as universal formats. MASH Explorer integrates multiple spectral deconvolution and database search algorithms into a unified platform. Software elements such as Configuration Wizard, Workflow Manager, and Ion Finder Tools were implemented to improve the user experience in analyzing top-down proteomics data. In Chapter 6, I designed a machine learning strategy to process and combine peak list results from multiple deconvolution algorithms. This strategy used hierarchical clustering methods and combined different peak list results into a consensus peak list. The optimized consensus peak list showed both improved accuracy and precision in deconvolution tasks than individual algorithms. This algorithm is effective in enhancing the throughput of deconvolution task by detecting true positive peaks while filtering out false positive peaks, and shows promising in improving the proteoform identification and characterization workflows for top-down proteomics.

Bioinformatics tools are still under-developed for intact protein analysis. As mentioned in the Chapter 1, spectral deconvolution includes both MS and MS/MS level deconvolution. This dissertation focuses primarily on optimizing and utilizing MS/MS level deconvolution and the workflow should be suitable for most intact protein analysis. However, challenges still exist in analyzing large proteins using bioinformatics. Using conventional proteomics analysis workflow, large proteins are directly fragmented by activation methods. In the case of CID, labile bonds are cleaved, resulting in limited sequence coverage. For electron-based activation such as ECD and ETD, the fragmentation efficiency may need to be optimized, and large ions are often hard to resolve without extended acquisition time. From the datasets, it appeared that a lot of real isotopic distributions could not be matched to the proposed sequence. A more comprehensive tool is needed to reveal the identity of these isotopic distributions, which may be arisen in different cases. First, internal fragments can be generated if a higher fragmentation energy is implemented. [12-14] For large proteins, this may provide additional information of the primary sequence. Additionally, these fragment ions can be a result of complex PTMs. For instance, if the protein presents in the spectrum in three proteoforms, such as unphosphorylated, monophosphorylated and bisphosphorylated forms, characterization of the monophosphorylated proteoform may be challenging. Isolation of the monophospohrylated proteoform may include a mixture of phosphorylated proteoforms with either one of the phosphorylation sites. Software tools that are able to isolate and quantify the ions between the two monophosphorylated proteoforms can significantly help with assigning isotopic distributions as well as addressing the PTM occupancy.[15]

Another area to explore for large intact protein analysis is the middle-down proteomics workflow. As mentioned in the Chapter 1, middle-down proteomics is normally used for large proteins analysis or proteoforms with complex modifications. Regarding large proteins, specialized enzymes have been developed to allow robust and high-throughput analysis such as that for monoclonal antibody using IdeS for digestion. However, analysis of other large proteins that are greater than 100 kDa is still under-development. One direction to address this problem is to utilize MS level deconvolution to identify intact masses of large polypeptide after digestion and map these intact masses to the protein sequence. MS level deconvolution has attracted significant interest, as it has been included in TopPIC Suite, Informed-Proteomics and recently FLASHDeconv.[16-18] These tools can collapse charge state distribution and provide intact mass information of large polypeptides, which could be used to compute generated polypeptide peak list from an imported sequence. Theoretical polypeptide masses can be calculated using information including the use of specialized enzyme such as Asp-N, Glu-C, and Lys-C, as well as the number of miscleavages. Additionally, possible mass differences such as +80 Da for phosphorylation and +42 Da for acetylation can be assigned for unmatched large polypeptides. This workflow might also be able to perform preliminary survey and output the results to guide users for in depth analysis, which may require users to verify using MS/MS analysis. This workflow has shown some success in analyzing RBM20, a protein that plays a role in titan splicing.[19]

References

1. Pesavento, J. J.; Mizzen, C. A.; Kelleher, N. L., Quantitative analysis of modified proteins and their positional isomers by tandem mass spectrometry: Human histone H4. *Anal. Chem.* **2006,** *78* (13), 4271-4280.

2. Wu, Z. J.; Tiambeng, T. N.; Cai, W. X.; Che, B. F.; Lin, Z. Q.; Gregoric, Z. R.; Ge, Y., Impact of Phosphorylation on the Mass Spectrometry Quantification of Intact Phosphoproteins. *Anal. Chem.* **2018,** *90* (8), 4935-4939.

3. Cohen, P., The regulation of protein function by multisite phosphorylation - a 25 year update. *Trends Biochem. Sci.* **2000,** *25* (12), 596-601.

4. Peng, Y.; Gregorich, Z. R.; Valeja, S. G.; Zhang, H.; Cai, W. X.; Chen, Y. C.; Guner, H.; Chen, A. J.; Schwahn, D. J.; Hacker, T. A.; Liu, X. W.; Ge, Y., Top-down Proteomics Reveals Concerted Reductions in Myofilament and Z-disc Protein Phosphorylation after Acute Myocardial Infarction. *Mol. Cell. Proteomics* **2014,** *13* (10), 2752-2764.

5. Jin, Y. T.; Peng, Y.; Lin, Z. Q.; Chen, Y. C.; Wei, L. M.; Hacker, T. A.; Larsson, L.; Ge, Y., Comprehensive analysis of tropomyosin isoforms in skeletal muscles by top-down proteomics. *J. Muscle Res. Cell M.* **2016,** *37* (1-2), 41-52.

6. Jin, Y.; Diffee, G. M.; Colman, R. J.; Anderson, R. M.; Ge, Y., Top-down Mass Spectrometry of Sarcomeric Protein Post-translational Modifications from Non-human Primate Skeletal Muscle. *J. Am. Soc. Mass Spectrom.* **2019,** *30* (12), 2460-2469.

7. Lin, Z. Q.; Guo, F.; Gregorich, Z. R.; Sun, R. X.; Zhang, H.; Hu, Y.; Shanmuganayagam, D.; Ge, Y., Comprehensive Characterization of Swine Cardiac Troponin T Proteoforms by Top-Down Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2018,** *29* (6), 1284-1294.

8. Wissing, J.; Jansch, L.; Nimtz, M.; Dieterich, G.; Hornberger, R.; Keri, G.; Wehland, J.; Daub, H., Proteomics analysis of protein kinases by target class-selective prefractionation and tandem mass spectrometry. *Mol. Cell. Proteomics* **2007,** *6* (3), 537-547.

9. Bantscheff, M.; Eberhard, D.; Abraham, Y.; Bastuck, S.; Boesche, M.; Hobson, S.; Mathieson, T.; Perrin, J.; Raida, M.; Rau, C.; Reader, V.; Sweetman, G.; Bauer, A.; Bouwmeester, T.; Hopf, C.; Kruse, U.; Neubauer, G.; Ramsden, N.; Rick, J.; Kuster, B.; Drewes, G., Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat. Biotechnol.* **2007,** *25* (9), 1035-1044.

10. Klaeger, S.; Heinzlmeir, S.; Wilhelm, M.; Polzer, H.; Vick, B.; Koenig, P. A.; Reinecke, M.; Ruprecht, B.; Petzoldt, S.; Meng, C.; Zecha, J.; Reiter, K.; Qiao, H. C.; Helm, D.; Koch, H.; Schoof, M.; Canevari, G.; Casale, E.; Depaolini, S. R.; Feuchtinger, A.; Wu, Z. X.; Schmidt, T.; Rueckert, L.; Becker, W.; Huenges, J.; Garz, A. K.; Gohlke, B. O.; Zolg, D. P.; Kayser, G.; Vooder, T.; Preissner, R.; Hahne, H.; Tonisson, N.; Kramer, K.; Gotze, K.; Bassermann, F.; Schlegl, J.; Ehrlich, H. C.; Aiche, S.; Walch, A.; Greif, P. A.; Schneider, S.; Felder, E. R.; Ruland, J.; Medard, G.; Jeremias, I.; Spiekermann, K.; Kuster, B., The target landscape of clinical kinase drugs. *Science* **2017,** *358* (6367).

11. Reinecke, M.; Ruprecht, B.; Poser, S.; Wiechmann, S.; Wilhelm, M.; Heinzlmeir, S.; Kuster, B.; Medard, G., Chemoproteomic Selectivity Profiling of PIKK and PI3K Kinase Inhibitors. *Acs. Chem. Biol.* **2019,** *14* (4), 655-664.

12. Cobb, J. S.; Easterling, M. L.; Agar, J. N., Structural Characterization of Intact Proteins Is Enhanced by Prevalent Fragmentation Pathways Rarely Observed for Peptides. *J. Am. Soc. Mass Spectrom.* **2010,** *21* (6), 949-959.

13. Durbin, K. R.; Skinner, O. S.; Fellers, R. T.; Kelleher, N. L., Analyzing Internal Fragmentation of Electrosprayed Ubiquitin Ions During Beam-Type Collisional Dissociation. *J. Am. Soc. Mass Spectrom.* **2015,** *26* (5), 782-787.

14. Lyon, Y. A.; Riggs, D.; Fornelli, L.; Compton, P. D.; Julian, R. R., The Ups and Downs of Repeated Cleavage and Internal Fragment Production in Top-Down Proteomics. *J. Am. Soc. Mass Spectrom.* **2018,** *29* (1), 150-157.

15. Tsai, C. F.; Wang, Y. T.; Yen, H. Y.; Tsou, C. C.; Ku, W. C.; Lin, P. Y.; Chen, H. Y.; Nesvizhskii, A. I.; Ishihama, Y.; Chen, Y. J., Large-scale determination of absolute phosphorylation stoichiometries in human cells by motif-targeting quantitative proteomics. *Nat. Commun.* **2015,** *6*, 6622.

16. 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.

17. Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Liu, T.; Petyuk, V. A.; Tolić, N.; Paša-Tolić, L.; Smith, R. D.; Payne, S. H.; Kim, S., Informed-Proteomics: open-source software package for top-down proteomics. *Nat. Methods* **2017,** *14*, 909.

18. Jeong, K.; Kim, J.; Gaikwad, M.; Hidayah, S. N.; Heikaus, L.; Schluter, H.; Kohlbacher, O., FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics. *Cell Syst.* **2020,** *10* (2), 213-218 e6.

19. Sun, M.; Jin, Y.; Zhu, C.; Rexiati, M.; Cai, H.; Chen, Z.; Ge, Y.; Guo, W., Mass Spectrometry Analysis of RBM20 Phosphorylation and Its Role in Titin Splicing. *The FASEB Journal* **2018,** *32* (S1), 791.13-791.13.